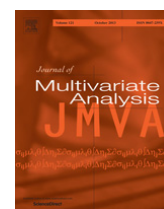


Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

## Journal of Multivariate Analysis

journal homepage: [www.elsevier.com/locate/jmva](http://www.elsevier.com/locate/jmva)A semiparametric factor model for CDO surfaces dynamics<sup>☆</sup>Barbara Choroś-Tomczyk<sup>a</sup>, Wolfgang Karl Härdle<sup>a</sup>, Ostap Okhrin<sup>b,\*</sup><sup>a</sup> *Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E.–Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany*<sup>b</sup> *Chair of Econometrics and Statistics esp. Transportation, Institute of Economics and Transport, Faculty of Transportation, Dresden University of Technology, Helmholtzstrasse 10, 01069 Dresden, Germany*

## ARTICLE INFO

## Article history:

Received 21 November 2014

Available online xxxx

## JEL classification:

C14

C51

G11

G17

## AMS subject classifications:

62Gxx

62H12

91G10

62M20

## Keywords:

CDO

Curve trade

Dynamic factor model

Semiparametric model

Surfaces dynamics

## ABSTRACT

Modelling the dynamics of credit derivatives is a challenging task in finance and economics. This work studies risk of collateralized debt obligations (CDOs) by investigating the evolution of tranche spread surfaces and base correlation surfaces using a dynamic semiparametric factor model (DSFM). The DSFM offers a combination of flexible functional data analysis and dimension reduction methods, where the change in time is linear but the shape is nonparametric. The study provides an empirical analysis based on a big data set of iTraxx Europe tranches and proposes an application to curve trading strategies. The DSFM allows us to describe the dynamics of all the tranches for all available maturities and series simultaneously which yields better understanding of the risk associated with trading CDOs and other structured products.

© 2015 Elsevier Inc. All rights reserved.

## 1. Introduction

This study proposes an empirical research of a large data set of iTraxx Europe indices and their tranches of Series 2–10. We investigate around 50 000 observations of iTraxx tranches over 1000 days between the year 2005 and 2009. To the best of our knowledge, this is the first study on CDOs that considers such an extensive data set. Moreover, the dynamics of the iTraxx tranches over time has not been investigated in literature so far.

The iTraxx Europe is the most widely traded credit index in Europe. Its reference portfolio consists of 125 equally weighted, most liquid credit default swaps (CDS) on European companies. For every index five standardized tranches of different risk profiles are traded. The cash-flows structure of iTraxx tranches is the same as of synthetic CDO tranches. Because of the regular index roll, every day we find on the market tranches with various times to expiration. By plotting

<sup>☆</sup> The financial support from the German Research Foundation, Project HA2229/7-2 and via CRC 649 *Economic Risk* at Humboldt-Universität zu Berlin is gratefully acknowledged.

\* Corresponding author.

E-mail address: [ostap.okhrin@tu-dresden.de](mailto:ostap.okhrin@tu-dresden.de) (O. Okhrin).

<http://dx.doi.org/10.1016/j.jmva.2015.09.002>

0047-259X/© 2015 Elsevier Inc. All rights reserved.

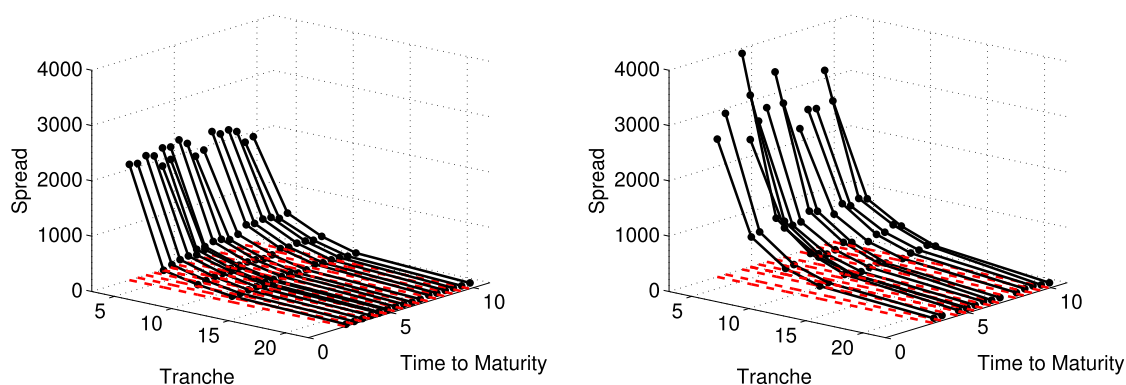


Fig. 1. Spreads of all tranches of all series observed on 20080909 (left) and 20090119 (right).

prices (base correlations) of all available tranches at one day as a function of the time to maturity and the tranche seniority, one gets a two-dimensional surface that represents the entire market information about spreads (base correlations), see Fig. 1. The tranches with 5 years maturity are the most liquid, unlike those with 3 years maturity that are rarely quoted. This makes the modelling of the surfaces a challenging task as each day one observes a different number of curves with not necessarily the same number of points on each curve. When we record these surfaces every day, we can follow how they change their shape and level. The dynamics over time of such surfaces is the main goal of this paper.

Mainly because of the high dimensionality of the CDO problem the vast majority of papers consider only CDOs of one particular maturity, see e.g. Hamerle et al. [17]. Up to our knowledge, the available literature do not look at the CDO market as a whole. Since CDOs are quoted for distinct maturities and with different liquidity, we should consider the effect of the CDO term structure.

From an investor's point of view, it is desirable to have an insight into the behaviour in the future of spreads and their main characteristics, namely base correlations. The forecasting has useful applications in hedging and trading CDOs, computation of risk measures, or construction of investment strategies. One of the simplest solutions would be to consider the classic time series analysis for each tranche of each series for every maturity. However, there are several reasons why this methodology is not applicable. Firstly, due to illiquidity of the tranche market we encounter multiple missing observations. Moreover, many iTraxx series issued during the financial crisis have too short data history. For the same reasons multivariate time series models could not find their application here. Thus, the major challenge we are facing in the analysis of iTraxx tranches is that every day only scattered observations of a two-dimensional surface are observable. This study proposes an estimation and forecasting method for CDO surfaces.

Modelling surfaces is one of the primary goals of the functional data analysis (FDA) where the data are functions, see [6]. Functional data sets naturally appear in many fields of science ranging from finance to genetics. Worth mentioning statistical approaches for handling complex high-dimensional problems are a structural analysis of curves by Kneip and Gasser [22], a functional regression with scalar (see [4]) or functional [5] response, a stochastic warping model by Liu and Müller [24], penalized splines by Kauer mann et al. [21], and a functional principal components approach by Gromenko et al. [15]. For recent advances in FDA we refer the reader to Ramsay and Silverman [27], Ferraty and Vieu [13], Ferraty and Romain [12], Horváth and Kokoszka [19] and Bongiorno et al. [3]. One of the most popular methods are factor type models as they effectively reduce the dimensionality. Factor models assume that the comovements of big number of variables are generated by a small set of latent factors. When data disclose a dynamic structure then one needs a technique that is able to correctly detect and describe the observed behaviour, e.g. [16].

In this study we employ a dynamic semiparametric factor model (DSFM). In the DSFM the observed variables are expressed as linear combinations of the factors. The factors and the factor loadings are estimated from the data. The first ones represent the spatial, time-invariant component. The latter ones form multidimensional time series that reflect the dynamics. The inference on the original variables reduces to the inference on the factors and the factor loadings. For advances in semiparametric functional data modelling we refer reader to Goia and Vieu [14].

The DSFM was introduced by Fengler et al. [11] for modelling the dynamics of implied volatility surfaces. Further, Härdle [18] use it for limit order book analysis, Detlefsen and Härdle [8] for variance swaps, and van Bömmel [30] for fMRI images. In this work we study the dynamics of CDO surfaces with the DSFM and propose an application to curve trading strategies.

The paper is structured as follows. Section 2 discusses the CDOs. Section 3 describes the DSFM. Section 4 shows results of the empirical modelling. Section 5 presents applications in CDO trading. Section 6 concludes.

## 2. Collateralized debt obligations

A collateralized debt obligation is a credit derivative used by financial institutions to repackage individual assets into a product that can be sold to investors on the secondary market. The assets may be mortgages, auto loans, credit card debt,

corporate debt or credit default swaps (CDS). CDOs were initially constructed for securitization of big portfolios. The entire portfolio risk is sliced into tranches and then transferred to investors. Prior to the credit crisis, CDOs provided outstanding investment opportunities to market participants. Tranching made it possible to create new securities of different risk classes that met the needs of a wide range of clients. The market observed an excess demand for senior CDO tranches because they were considered as safe and offered unusual high returns. As we know now, the rating agencies underestimated default risk of CDOs. Consequently, investors were exposed to more risk than the ratings of these CDOs implied. The CDO market has significantly shrunk since the beginning of the financial crisis. However, the methodology proposed in our study can be used in modelling and trading other financial instruments, especially non-standardized and bespoke structured products.

Consider a CDO with a maturity of  $T$  years,  $J$  tranches and a pool of  $d$  entities at the valuation day  $t_0$ . A tranche  $j = 1, \dots, J$  absorbs losses between  $l_j$  percent and  $u_j$  percent of the total portfolio loss.  $l_j$  and  $u_j$  are called an attachment and a detachment point respectively and  $l_j < u_j$ . For the iTraxx Europe, successive tranches have the following attachment points: 0%, 3%, 6%, 9%, 12%, 22%. The corresponding detachment points are 3%, 6%, 9%, 12%, 22%, 100%.

2.1. Valuation

We assume that there exists a risk-neutral measure  $P$  under which the discounted asset prices are martingales. The expectations in the formulas below are taken with respect to this measure.

The loss of the portfolio of  $d$  assets at time  $t$  is defined as

$$L(t) = \frac{\text{LGD}}{d} \sum_{i=1}^d \Gamma_i(t), \quad t \in [t_0, T],$$

where LGD is a common loss given default and  $\Gamma_i(t) = \mathbf{1}(\tau_i \leq t)$ ,  $i = 1, \dots, d$ , is a default indicator showing that the credit  $i$  defaults at time  $t$  within the period  $[t_0, T]$  if the time of default random variable  $\tau_i \leq t$ . The loss of a tranche  $j = 1, \dots, J$  at time  $t$  is expressed as  $L_j(t) = L^u(t, u_j) - L^l(t, l_j)$ , with  $L^u(t, x) = \min\{L(t), x\}$ ,  $x \in [0, 1]$ . The outstanding notional of the tranche  $j$  is given by  $F_j(t) = F^u(t, u_j) - F^l(t, l_j)$  with  $F^u(t, x) = x - L^u(t, x)$ ,  $x \in [0, 1]$ . At the predefined dates  $t = t_1, \dots, T$ ,  $t_1 > t_0$ , the protection seller and the protection buyer exchange the payments. The protection buyer pays to the protection seller a predetermined premium, called a spread on the outstanding tranche notional and is compensated for losses that occur within the range of the tranche. Each default in the portfolio reduces the outstanding tranche notional. This leads to a decline in the value of the periodic fee payment. The cash exchange takes place until  $T$  or until the portfolio losses exceed the detachment point.

The protection leg  $DL_j$  is defined as the present value of all expected payments made upon defaults

$$DL_j(t_0) = \sum_{t=t_1}^T \beta(t_0, t) E\{L_j(t) - L_j(t - \Delta t)\}, \quad j = 1, \dots, J, \tag{1}$$

where  $\beta$  is a discount factor and  $\Delta t$  is a time between  $t$  and the previous payment day. The premium leg  $PL_j$  is expressed as the present value of all expected premium payments

$$PL_j(t_0) = \sum_{t=t_1}^T \beta(t_0, t) s_j(t_0) \Delta t E\{F_j(t)\}, \quad j = 2, \dots, J, \tag{2}$$

where  $s_j$  denotes the spread of tranche  $j$ . The first tranche, called the equity is traded with an upfront payment  $\alpha$  and a fixed spread of 500 bp. Its premium leg (2) turns into

$$PL_1(t_0) = \alpha(t_0)(u_1 - l_1) + \sum_{t=t_1}^T \beta(t_0, t) \cdot 500 \cdot \Delta t E\{F_1(t)\}.$$

A spread  $s_j$  is calculated once, at  $t_0$  so that the marked-to-market value of the tranche is zero, i.e. the value of the premium leg equals the value of the protection leg

$$s_j(t_0) = \frac{\sum_{t=t_1}^T \beta(t_0, t) E\{L_j(t) - L_j(t - \Delta t)\}}{\sum_{t=t_1}^T \beta(t_0, t) \Delta t E\{F_j(t)\}}, \quad \text{for } j = 2, \dots, J. \tag{3}$$

The upfront payment of the equity tranche is computed as

$$\alpha(t_0) = \frac{100}{u_1 - l_1} \sum_{t=t_0}^T (\beta(t, t_0) [E\{L_1(t) - L_1(t - \Delta t)\} - 0.05 \Delta t E\{F_1(t)\}]).$$

For more details we refer to Bluhm and Overbeck [2] and Kakodkar et al. [20].

The main challenge in calculating the fair tranche spread (3) is the correct calculation of the expected losses. This task requires the analysis of how the portfolio entities are likely to default together. There are two main types of credit risk models: structural and reduced form models. The structural model is motivated from a Merton style approach where a default occurs when the value of an asset drops below a certain level. In the reduced form approach a default is modelled with an intensity process. A third class is based on copula theory and is connected with the first two approaches. For a comprehensive overview we refer to Bielecki and Rutkowski [1].

There has been a multitude of CDO risk models proposed that apply different dependency concepts. The market standard for pricing CDOs is the large pool Gaussian copula model that has been introduced to the valuation of multi-name credit derivatives by Li [23]. The main drawback of the Gaussian copula is that it exhibits no tail dependence and in consequence it cannot model the extreme events accurately. However, due to its analytical tractability and numerical simplicity, the large pool Gaussian copula model still remains the benchmark on the market.

2.2. Base correlation

In the Gaussian copula model the main driver of the tranche price is the correlation coefficient. The correlations can be computed from market data by inverting the pricing formula (3). If we keep the value of other parameters fixed, then the correlation parameter that matches the quoted tranche spread is called an implied compound correlation. It is observed that implied compound correlations are not constant across the tranches. This phenomenon is called an implied correlation smile. Still, the main disadvantage of the compound coefficient is that the mezzanine tranches are not monotonic in correlation and two parameters might result in the same spread value. The second problem that we might encounter is a nonexistence of the implied correlation. These disadvantages caused the enhanced popularity of base correlations proposed by McGinty and Ahluwalia [25].

The main idea behind the concept of the base correlation is that each tranche  $[l_j, u_j]$  can be represented as a difference of two, equity type tranches that have the lower attachment point zero:  $[0, u_j]$  and  $[0, l_j]$ . Here we use a property that the equity tranche is monotone in correlation. The base correlations can be implied from the market spreads using standard bootstrapping techniques. One needs the spread value of the tranche  $[l_j, u_j]$  and the base correlation of the tranche  $[0, l_j]$  in order to imply the base correlation  $[0, u_j]$ . In this approach, (3) is calculated as

$$s_j(t_0) = \frac{\sum_{t=t_1}^T \beta(t_0, t) [E_{\rho(0, u_j)}\{L_j^u(t, u_j) - L_j^u(t - \Delta t, u_j)\} - E_{\rho(0, l_j)}\{L_j^u(t, l_j) - L_j^u(t - \Delta t, l_j)\}]}{\sum_{t=t_1}^T \beta(t_0, t) \Delta t [E_{\rho(0, u_j)}\{F_j^u(t, u_j)\} - E_{\rho(0, l_j)}\{F_j^u(t, l_j)\}]} \tag{4}$$

for  $j = 2, \dots, J$ , where the expected value  $E_{\rho(0, u_j)}$  is calculated with respect to the loss distribution determined by the base correlation  $\rho(0, u_j)$  of the tranche  $[0, u_j]$ . In the Gaussian copula model the base correlations are nondecreasing with respect to the seniority of tranches and the implied correlation smile turns into a correlation skew.

3. Dynamic semiparametric factor model

Let  $Y_{t,k}$  be a data point, a tranche spread or a base correlation, observed on a day  $t, t = 1, \dots, T$ . The index  $k$  represents an intra-day numbering of observations on that day,  $k = 1, \dots, K_t$ . The observations  $Y_{t,k}$  are regressed on two-dimensional covariates  $X_{t,k}$  that contain the tranche seniority and the remaining time to maturity

$$Y_{t,k} = m_0(X_{t,k}) + \sum_{l=1}^L Z_{t,l} m_l(X_{t,k}) + \varepsilon_{t,k}, \tag{5}$$

where  $m_l : \mathbb{R}^2 \rightarrow \mathbb{R}, l = 0, \dots, L$ , are factor loading functions,  $Z_{t,l} \in \mathbb{R}$  are factors, and  $\varepsilon_{t,j}$  are error terms with zero means and finite variances.

The additive structure of (5) is a typical approach in regression models. Here, the functions  $m$  are estimated nonparametricly and represent the time-invariant, spatial component. The factors  $Z_t$  drive the dynamics of  $Y_t$ . The number of factors  $L$  is fixed and should be small relative to the number of observed data points so that we achieve a significant reduction in the dimension. The investigation of the dynamics of the entire system boils down to the analysis of the factors' variability. These arguments justify calling (5) a dynamic semiparametric factor model.

Fengler et al. [11] estimate  $m$  and  $Z_t$  iteratively using kernel smoothing methods, Härdle and Ritov [29] apply functional principal component analysis, Härdle and Borak [26] estimate  $m$  with a series based estimator. For numerical convenience we follow the last paper and define functions  $\psi_b : \mathbb{R}^2 \rightarrow \mathbb{R}, b = 1, \dots, B, B \geq 1$ , such that  $\int_{\mathbb{R}^2} \psi_b^2 dx = 1$ . Then, a tuple of functions  $(m_0, \dots, m_L)^\top$  may be approximated by  $A\psi$ , where  $A$  is a  $(L + 1 \times B)$  matrix of coefficients  $\{a_{l,b}\}_{l=0}^{L+1} \}_{b=1}^B$  and

$\psi = (\psi_1, \dots, \psi_B)^\top$ . We take  $\{\psi_b\}_{b=1}^B$  to be a tensor  $B$ -spline basis. For a survey over the mathematical foundations of splines we refer to de Boor [7]. With this parametrization (5) turns into

$$Y_{t,k} = Z_t^\top m(X_{t,k}) + \varepsilon_{t,k} = Z_t^\top A\psi(X_{t,k}) + \varepsilon_{t,k},$$

where  $Z_t = (Z_{t,0}, \dots, Z_{t,L})^\top$  with  $Z_{t,0} = 1$  and  $m = (m_0, \dots, m_L)^\top$ .

The estimates  $\widehat{Z}_t = (\widehat{Z}_{t,0}, \dots, \widehat{Z}_{t,L})^\top$  and  $\widehat{A}$  are obtained similarly to Ramsay and Silverman [27] by

$$(\widehat{Z}_t, \widehat{A}) = \arg \min_{Z_t, A} \sum_{t=1}^T \sum_{k=1}^{K_t} \{Y_{t,k} - Z_t^\top A\psi(X_{t,k})\}^2, \quad (6)$$

yielding estimated basis functions  $\widehat{m} = \widehat{A}\psi$ . The minimization is carried out using an iterative algorithm. However, the estimates of  $m$  and  $Z_t$  are not uniquely defined. Therefore, the final estimates of  $m$  are orthonormalized and  $Z_t$  are centered. Park et al. [26] also prove that the difference of the inference based on the estimated  $\widehat{Z}_{t,l}$  and the true, unobserved  $Z_{t,l}$  is asymptotically negligible. This result justifies fitting an econometric model, like a vector autoregressive to the estimated factors for further analysis of data.

The number of factors  $L$  as well as the numbers of spline knots in both maturity and tranche directions  $R_1, R_2$ , and the orders of splines  $r_1, r_2$  have to be chosen in advance. A common approach is to maximize a proportion of the variation explained by the model among the total variation. We propose a following criterion

$$EV(L, R_1, r_1, R_2, r_2) = 1 - \frac{\sum_{t=1}^T \sum_{k=1}^{K_t} \left\{ Y_{t,k} - \sum_{l=1}^L Z_{t,l} m_l(X_{t,k}) \right\}^2}{\sum_{t=1}^T \sum_{k=1}^{K_t} \{Y_{t,k} - \widetilde{m}_0(X_{t,k})\}^2}, \quad (7)$$

where

$$\widetilde{m}_0(X_\ell) = \frac{1}{T} \sum_{t=1}^T \frac{\sum_{k=1}^{K_t} Y_{t,k} \mathbf{1}\{X_\ell = X_{t,k}\}}{\sum_{k=1}^{K_t} \mathbf{1}\{X_\ell = X_{t,k}\}}, \quad \ell = 1, \dots, K_{\max}, \quad (8)$$

is an empirical mean surface and  $K_{\max}$  is the number of all different  $X_{t,k}$  observed during  $T$  days. The criterion (7) is a modified version of the one considered in [11] and other literature on the DSFM, where instead of the empirical mean surface, the overall mean of the observations is used. The mean surface (8) makes more sense, since our data reflect monotonous behaviour w.r.t. the tranche seniority.

The  $\widetilde{m}_0$  factor in (5) is usually interpreted as a mean function of the data. We propose to first subtract the estimate (8) from the data and then fit the DSFM. The extraction of the empirical mean  $\widetilde{m}_0$  leads to the following model

$$Y_{t,k} = \widetilde{m}_0(X_{t,k}) + \sum_{l=1}^L Z_{t,l} m_l(X_{t,k}) + \varepsilon_{t,k} = \widetilde{m}_0(X_{t,k}) + Z_t^\top A\psi(X_{t,k}) + \varepsilon_{t,k}, \quad (9)$$

where  $m_l$  are factor functions,  $l = 1, \dots, L$ ,  $Z_{t,l}$  are factor loadings, and  $A$  is a  $(L \times B)$  coefficient matrix. The representation (9) reduces the number of the factor functions estimated in the iterative algorithm (6). As the model (9) achieved a bit better performance in the empirical study we present only the results of this approach. For simplicity's sake the model (9) is hereafter called the DSFM.

## 4. Modelling the dynamics of CDO surfaces

### 4.1. Data description

The data set analysed in this study contains daily spreads of iTraxx tranches of Series 2–10 between 30 March 2005 (hereafter denoted 20050330) and 2 February 2009 (denoted 20090202) obtained from Bloomberg. We have in total  $\sum_{t=1}^T K_t = 49\,502$  data points over  $T = 1004$  days.

Twice a year, every March and September, a new series of iTraxx is issued. Therefore, every day one observes a bunch of indices from various series and different maturities. Here we analyse tranche spreads and also base correlations, both denoted  $Y_t$ , as a function of the tranche seniority  $\xi_t$  and the remaining time to maturity  $\tau_t$ . Each iTraxx index has 3, 5, 7, or 10 years maturity. The seniority of a tranche  $\xi_t$  is represented by its corresponding detachment point. The remaining time to maturity of a tranche is an actual time left till its expiration and takes values between zero and 10.25. For every day a separate surface representing the entire market information is available. The number of observed every day indices is low (minimum 4, maximum 17, median 12, see Fig. 2). This results in a string structure in the data. Each string corresponds to

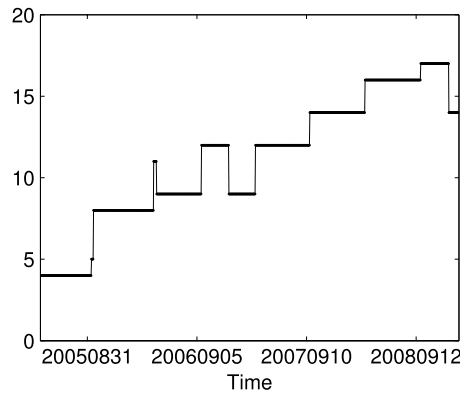


Fig. 2. Daily number of curves for every surface during the period 20050330–20090202.

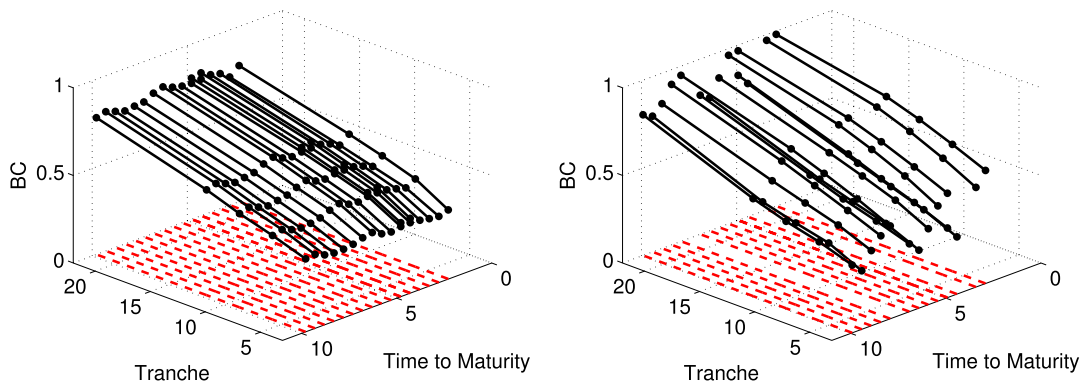


Fig. 3. Base correlations of all series observed on 20080909 (left) and 20090119 (right).

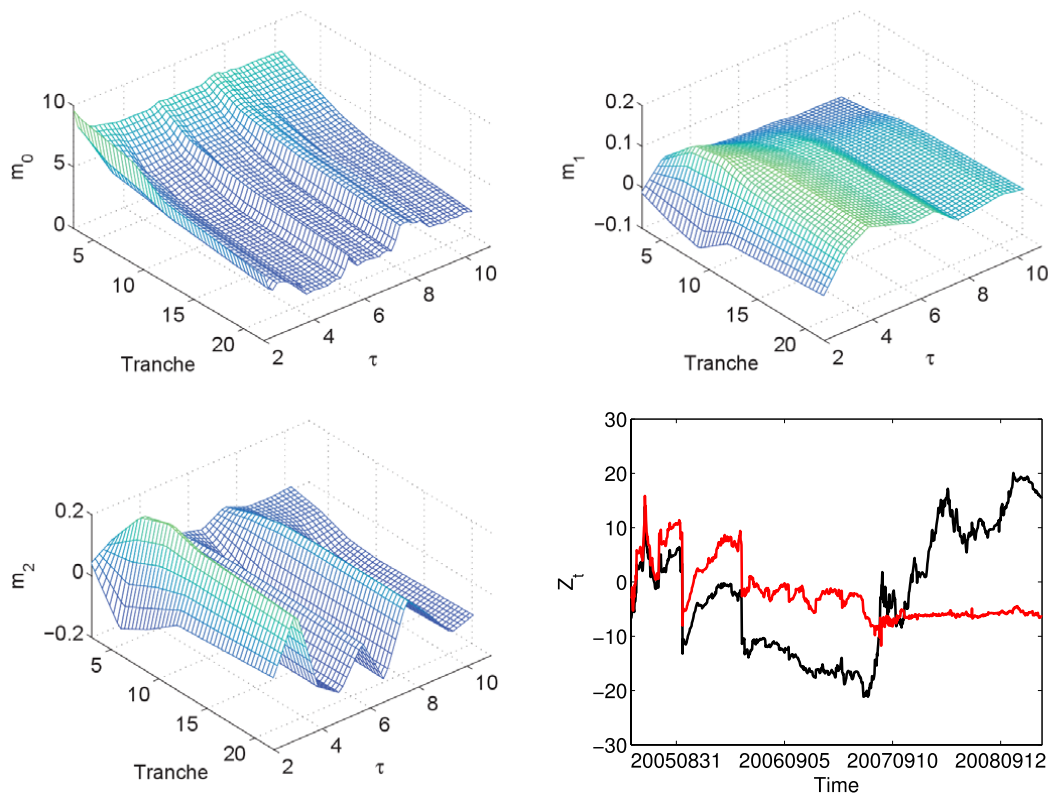
Table 1  
Percentage of missing values during the period 20050330–20090202.

Year	3Y					5Y					7Y					10Y				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
2005	100	100	100	100	100	34	34	34	34	48	5	5	5	5	5	35	34	35	34	35
2006	78	56	55	100	100	6	7	6	6	8	3	3	3	4	4	5	5	6	8	6
2007	88	99	99	100	100	3	2	2	3	3	2	2	3	2	3	3	2	3	2	2
2008	47	99	100	100	100	24	25	25	24	27	24	25	25	25	27	24	27	24	25	24
2009	100	100	100	100	100	42	42	47	42	42	42	43	43	42	42	42	43	42	42	43
All	72	93	93	100	100	16	17	17	16	20	13	14	13	13	14	16	17	16	17	16

one  $\tau_t \in [0, 10.25]$  and is composed out of at most five points. The market quotes five out of six tranches as the most senior tranche is usually not traded. Figs. 1 and 3 present the curves of market spreads and corresponding implied base correlations on 20080909 and 20090119. As time passes, the curves move through the space towards expiry and simultaneously change their skewness and level.

Since the shortest maturity is 3 years and every half a year new four indices are issued, the number of indices present on the market grows in time. Table 1 outlines a percentage of missing values for every maturity and for every tranche during the entire period considered and during the annual subperiods. We see that the CDO market was booming in 2006 and 2007. However, since the beginning of the financial crisis in 2008 the demand for credit derivatives had been shrinking meaningfully. In the first quarter of 2009 the iTraxx tranches became highly illiquid. As mentioned before, many missing data may create challenges to the econometric analysis. Because tranches with 3 years maturity were rarely traded, this maturity was excluded from our study.

Sometimes on a particular day, for a particular tranche and a particular remaining time to maturity we observe two different spreads. As an example consider a day  $t_0$  on which a new series with 3 years maturity is issued. If 5 years earlier a series with 7 years maturity was issued, then on day  $t_0$  this series has also 3 years remaining time to maturity. In this situation we include in our data set the observation that comes from the most actual series (in the example we take the series issued on  $t_0$ ).



**Fig. 4.** Sample mean, estimated factors and loadings ( $Z_{t,1}$  black,  $Z_{t,2}$  red) in the DSFM for the log-spreads. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The base correlations (4) are implied from the market spreads using the large pool Gaussian copula model (assuming the LGD of 60%). The common intensity parameters are derived from iTraxx indices. The discount factors are calculated from rates of Euribor and Euro Swaps.

The structure of the equity tranche is different from the other tranches. It is quoted as an upfront payment plus 500 bp spread paid quarterly. In order to include the equity tranche in the joint analysis of all the tranches, we convert its quotes to standard spreads with zero upfront fee using the large pool Gaussian copula model.

#### 4.2. DSFM estimation results

Since our data are positive and monotone, we convert spreads into log-spreads and for base correlations apply the Fisher's  $Z$ -transformation defined as

$$\mathcal{T}(u) = \operatorname{arctanh}(u) = \frac{1}{2} \log \frac{1+u}{1-u}.$$

It transforms the empirical Pearson's correlations between bivariate normal variables to a normally distributed variable. We will use it for the base correlations as it stabilizes their variance.

Since the design of the data in the tranche seniority dimension is fixed, we choose in this direction quadratic  $B$ -splines and five knots. Table 2 presents a proportion of the explained variation (7) for different numbers of factors, knots and different orders of splines in the maturity dimension. Similar to Park et al. [26], we find that the order of splines and the number of knots have a small influence on the proportion of the explained variation. We pick two factors and the quadratic  $B$ -splines placed on 10 knots in  $\tau$  dimension for both types of data. The number of knots is close to the median number of observed strings every day. Figs. 4 and 5 exhibit  $\hat{m}$  and  $\hat{Z}_t$  estimated in the DSFM for the log-spreads and the  $Z$ -transformed base correlations respectively.

In the DSFM for the log-spreads the first and the second factor can be interpreted as a shift function and a slope-curvature respectively. When we shift  $\hat{Z}_{t,1}$ , the whole surface shifts along the  $z$ -axis. Increasing  $\hat{Z}_{t,2}$  results in the enhancement of the surface's steepness, whereas, decreasing  $\hat{Z}_{t,2}$  implies its flattening. The interpretation of the DSFM factors of  $Z$ -transformed base correlations is not so clear. When varying  $\hat{Z}_{t,1}$  and  $\hat{Z}_{t,2}$  both the slope and the curvature change. The upward shift of the surface can be a result of a decrease in  $\hat{Z}_{t,1}$  or an increase in  $\hat{Z}_{t,2}$ .

Fig. 6 displays the in-sample fit of the models to data on 20080909 and 20090119. The convergence of the models is typically reached after 8 cycles. The mean squared error of the in-sample fit over all dates considered in this study is 0.045 for the log-spreads surfaces and equals 0.006 for the  $Z$ -transformed base correlations surfaces.

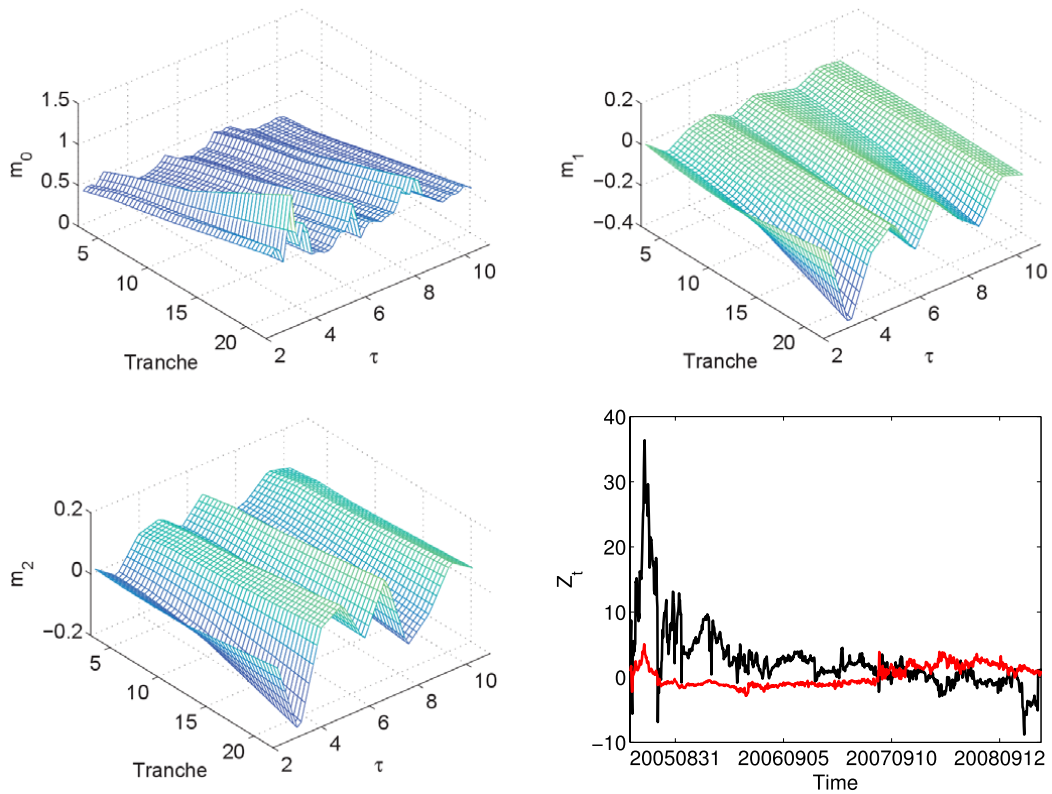


Fig. 5. Sample mean, estimated factors and loadings ( $Z_{t,1}$  black,  $Z_{t,2}$  red) in the DSFM for the Z-transformed base correlations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

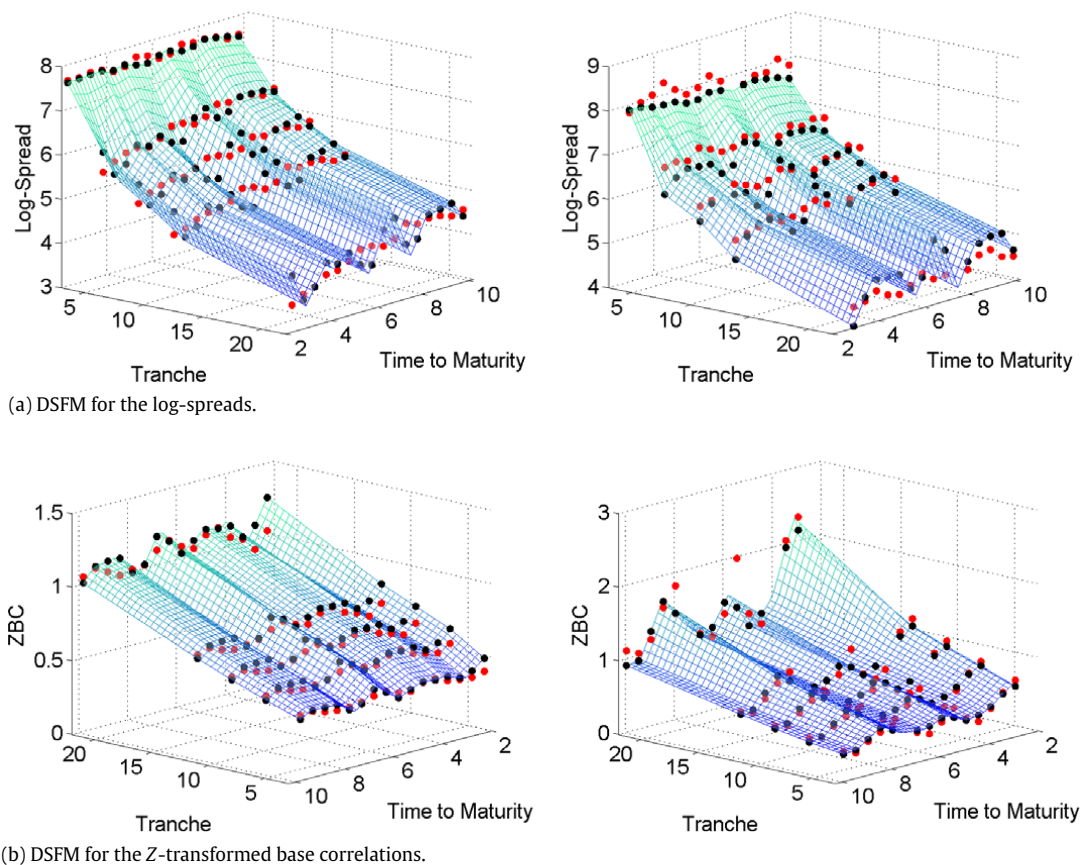
Table 2

Proportion of the explained variation by the DSFM for  $L = 1, 2, 3$ , different numbers of knots and different orders of splines in the maturity dimension. The values of the selected models marked with italic.

Number of factors	Spline Order	Log-spr				Z-BC			
		Knots							
		5	10	15	20	5	10	15	20
	1	0.797	0.876	0.897	0.898	0.629	0.640	0.660	0.660
	2	0.877	0.896	0.905	0.910	0.633	0.654	0.657	0.664
	3	0.867	0.898	0.906	0.908	0.638	0.650	0.662	0.664
	4	0.871	0.898	0.907	0.910	0.639	0.653	0.659	0.662
	1	0.842	0.925	0.940	0.945	0.730	0.835	0.860	0.869
	2	0.926	0.952	0.961	0.954	0.781	0.861	0.876	0.888
	3	0.911	0.952	0.941	0.950	0.763	0.867	0.883	0.887
	4	0.917	0.956	0.947	0.954	0.783	0.870	0.881	0.886
	1	0.858	0.940	0.959	0.973	0.746	0.854	0.888	0.898
	2	0.941	0.967	0.977	0.982	0.815	0.896	0.907	0.925
	3	0.927	0.967	0.975	0.979	0.805	0.901	0.922	0.930
	4	0.932	0.972	0.977	0.982	0.817	0.903	0.910	0.927

The covariance structure of the  $\hat{Z}_t$  time series is investigated by means of VAR analysis. The augmented Dickey–Fuller test indicates that the first differences of  $\hat{Z}_t$  are stationary. The check of the sample partial autocorrelation functions of the residuals of the estimated VAR(1) models for the factor loadings confirms that the VAR(1) process captures the autocorrelation structure of the factor loadings. Certainly, one may investigate more complex multivariate time series models that account for a dynamic structure of the conditional variance–covariance and of the conditional correlation like the BEKK-GARCH or the DCC-GARCH, see [9]. Since we are interested in the conditional mean process only, the VAR model appears to be sufficient. Moreover, a relatively simple out-of-sample VAR forecasting can be used in forecasting the evolution of the surfaces.





**Fig. 6.** In-sample fit (black points) of the models to data (red points) on 20080909 (left) and 20090119 (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

## 5. Applications in trading

### 5.1. Curve trades

The popularity of the iTraxx market led to increased liquidity in its standardized tranches allowing investors to implement complex credit positions. Here we present curve trades, namely flatteners and steepeners—strategies that combine tranches of different time to maturity, see also [20].

A flattener is a trade that involves a simultaneous sale of protection on a long-term tranche and a purchase of protection on a short-term tranche. An example would be: sell 10Y 3%–6% and buy 5Y 6%–9%. In this trade the investor expresses not only a bullish long-term outlook but also a bearish short-term view on the market. The opposite trade is called a steepener. It is achieved by selling the short-term protection and buying the long-term protection. Both strategies are popular in trading CDS, credit indices, and yield curves. Credit curves got a lot of attention in May 2012 when J.P. Morgan announced a loss of \$2 billion on its flattener trade on the CDX IG 9 index. The final loss reached \$6.2 billion.

In our study both long and short term tranches have equal notional amounts. However, by adjusting the notionals, a trade can be structured so that it is risky duration neutral, carry neutral, correlation neutral, or theta (sensitivity to implied correlation changes) neutral, see [28]. As recommended by Felsenheimer et al. [10] we consider trades that generate no or a positive carry, i.e. the spread of the sold protection does not exceed the spread of the bought protection.

It is important to remark that our trades are exposed to default risk. If one buys 6%–9% and sells 3%–6%, then these tranches provide protection of different portions of portfolio risk. If there is any default in 3%–6%, then we must deliver a payment obligation and incur a loss. Since we do not possess data of historical defaults in iTraxx, we cannot include the default payments in the further analysis. Consequently, in calculating the profit-and-loss (P&L) of the strategy we also do not account for the positive carry that we cumulate until the both positions are closed.

Felsenheimer et al. [10], Kakodkar et al. [20] and Roy [28] consider various scenarios of flattener trades. They also assume that we do not observe any defaults in the collateral. However, their examples are not based on real data and do not investigate the performance of the trades over time.

Assume that an investor enters a curve trade and sells protection at a spread of  $s_1(t_0)$  for the period  $[t_0, T_1]$  and buys protection at a spread of  $s_2(t_0)$  for the period  $[t_0, T_2]$ . If the trade is a flattener, then  $T_1 > T_2$ . The spreads of the tranches are

calculated in such a way that on the date of the trade  $t_0$  the marked-to-market (MTM) values of both positions are zero

$$MTM_\ell(t_0) = \sum_{t=t_1}^{T_\ell} \beta(t_0, t) [s_\ell(t_0) \Delta t E\{F_\ell(t)\} - E\{L_\ell(t) - L_\ell(t - \Delta t)\}] = 0, \quad \ell = 1, 2.$$

Since spread values constantly vary over time, immediately after initiation of the trade,  $\tilde{t} > t_0$ , the market trades the tranches at  $s_\ell(\tilde{t})$ . In consequence, we observe a change in the MTM value of our positions

$$MTM_\ell(\tilde{t}) = \{s_\ell(t_0) - s_\ell(\tilde{t})\} \sum_{t=\tilde{t}_1}^{T_\ell} \beta(\tilde{t}, t) \Delta t E\{F_\ell(t)\}, \quad \ell = 1, 2, \tag{10}$$

where  $\tilde{t}_1$  is the first payment day after  $\tilde{t}$ .

A positive MTM means that the contract has a positive value to the protection seller. If the protection seller closes the position  $\ell$  at time  $\tilde{t}$ , then receives from the protection buyer the amount  $MTM_\ell(\tilde{t})$ .

The aim of the curve trade investor is to maximize the P&L function that equals the total MTM value

$$PL(\tilde{t}) = MTM_1(\tilde{t}) - MTM_2(\tilde{t}). \tag{11}$$

### 5.2. Empirical results

The key decision in constructing a curve trade is which tranche to buy and which to sell. If an investor entered a flattener on 20080909, then the trade incorporated two tranches whose spreads are depicted on the left panel of Fig. 1. If the investor decided to close the positions on 20090119, then their MTM values (10) were calculated using the spread quotes exhibited on the right panel of Fig. 1 and using the base correlations (needed for  $E\{F_\ell\}$ ) shown on the right panel of Fig. 3. Having the data displayed on Figs. 1 and 3, we can compute the MTM values of all tranches that were quoted on both days. In consequence, we can easily recover those two tranches that maximize the P&L function (11). However, it is only possible if we possess the whole market information from these two points in time.

With an efficient forecasting technique, one can compute, for a given time horizon, a prediction of each point that is displayed on Figs. 1 and 3. By doing it using standard econometric methods, each tranche from every series has to be traded as an individual time series. Disregarding the fact that there are many missing values in our data, see Table 1, we have many series that do not have a long history. If an investor bought a tranche from Series 9 on 20080320, the day of its launch and decided to sell it a day or a week later, then we might not have enough past observations to fit and forecast the time series model.

In the DSFM modelling we do not differentiate the indices by their series number but by their remaining time to maturity. If in the past we already had observations with a very long remaining time to maturity, then we are able to price upcoming series even before they appear on the market. Moreover, we can forecast them using the DSFM.

We carry out the forecasting of log-spreads and Z-transformed base correlations in moving windows. A moving window procedure is used when only the most recent data are considered to be relevant for the estimation. We impose a static window of  $w = 250$  days. Then for every time  $t_0$  between the day  $w$  and the last day  $T$  in our data, we analyse  $\{\{Y_{t,k}\}_{k=1}^{K_t}\}_{t=t_0-w+1}^{t_0}$ . For each such set we estimated the DSFM model (9). As a result, we obtain  $T - w + 1$  times the estimated factor functions  $\hat{m} = (\hat{m}_0, \dots, \hat{m}_L)^\top$  and the series of the factor loadings  $\hat{Z}_t = (\hat{Z}_{t,0}, \dots, \hat{Z}_{t,L})^\top$  of length  $w$ . Since the factor functions are fixed, the forecasting is performed only on the factor loadings. As discussed in Section 4.2, we apply VAR(1) models to compute the predictions for a horizon  $h$  of one day, one week (five days), and one month (20 days). Due to the fixed scheme of issuing the iTraxx on the market, for every time  $t$ ,  $w+h \leq t \leq T$  we know which indices are traded. Therefore, the number of points that could be observed  $K_t$  and the possible remaining times to maturity  $\tau_t$  are known. Thus, the bivariate vector  $X_{t,k}$ ,  $k = 1, \dots, K_t$ , does not have to be forecasted. The forecast  $\hat{Y}_{t,k}$  is calculated from the  $\hat{Z}_t$  forecast. Finally, a proper inverse transformation is applied to  $\hat{Y}_{t,k}$  in order to recover the values of the spreads and the base correlations.

The calculation of the expected tranche losses using the large pool Gaussian copula model needs as an input a homogeneous default probability. Since the spread predictions are calculated out-of-sample, we also forecast the default probabilities. All predicted values of spreads and base correlations that lead to an arbitrage in prices, i.e. negative spreads, default probabilities and base correlations outside  $[0, 1]$ , were excluded.

Afterwards, for every predicted  $\{\hat{s}_k(t), \hat{\rho}_k(t)\}$ ,  $t = w+h, \dots, T$ ,  $k = 1, \dots, K_t$ , we compute  $\widehat{MTM}_k(t)$  according to (10) where the initial spread is the spread observed on  $t-h$ . Consequently, we create a surface of the predicted MTM values, see Fig. 7. Each surface has its extremes that indicate the tranches recommended for buying and selling.

The empirical analysis of the curve trades' performance is conducted using tranches 2–5 for all dates and indices considered in Section 4.2. Since the equity tranche is quoted in percent as an upfront fee, its corresponding spread is significantly higher than the spreads of other tranches. As it causes a large skew of our spread surfaces, we excluded it from the study. However, the calculation of the spread and the MTM value of the tranche 2 requires as an input a value of the base correlation of the equity tranche. Therefore, we first estimate and forecast the DSFM in moving windows using all

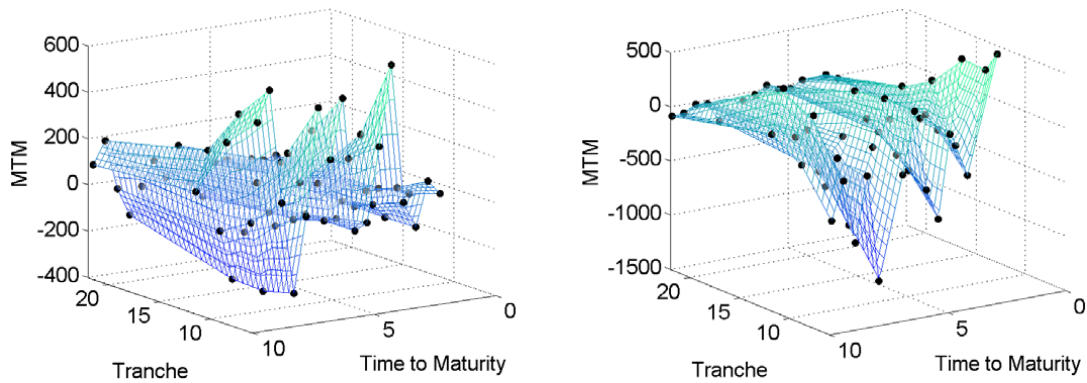


Fig. 7. MTM surfaces on 20080909 (left) and 20090119 (right) calculated using one-day spread and base correlation predictions.

Table 3

Average bid-ask spread excess over the mid spread as a percentage of the mid spread for tranches of Series 8 during the period 20070920–20090202.

Maturity	1	2	3	4	5
5Y	1.88	1.78	2.52	3.77	6.28
7Y	1.49	1.65	2.31	2.97	4.87
10Y	1.41	1.66	1.83	2.52	4.09

Table 4

Flatteners. Calculations based on predictions of log-spreads and Z-transformed BCs marked as LZ; based only on Z-transformed BCs marked as Z.

Tranches	Maturity	Mean of daily gains in %						Number of executed trades					
		1 day		1 week		1 month		1 day		1 week		1 month	
		LZ	Z	LZ	Z	LZ	Z	LZ	Z	LZ	Z	LZ	Z
2	All	0.28	0.32	0.12	0.11	0.05	0.04	751	750	744	742	728	729
3	All	0.16	0.20	0.06	0.07	0.02	0.01	754	754	750	750	734	735
4	All	0.10	0.15	0.03	0.04	0.01	0.01	754	754	750	750	735	734
5	All	0.08	0.10	0.03	0.03	0.01	0.01	741	739	733	734	735	733
All	10-5	0.19	0.21	0.06	0.08	0.02	0.02	698	754	700	750	701	727
All	10-7	0.25	0.25	0.08	0.08	0.03	0.03	736	754	730	749	726	730
All	7-5	0.14	0.15	0.04	0.05	0.01	0.00	724	750	714	748	717	733
2	10-5	0.26	0.26	0.12	0.13	0.04	0.04	476	549	483	566	495	622
3	10-5	0.12	0.17	0.05	0.05	0.01	-0.01	576	718	587	705	571	690
4	10-5	0.12	0.11	0.03	0.02	0.01	-0.01	573	731	582	722	582	695
5	10-5	0.07	0.07	0.03	0.02	0.01	0.00	587	696	572	689	577	660
2	10-7	0.34	0.29	0.14	0.12	0.06	0.05	555	542	560	571	545	635
3	10-7	0.17	0.22	0.06	0.05	0.02	0.00	635	704	635	703	616	712
4	10-7	0.12	0.13	0.03	0.03	0.01	0.00	596	721	610	717	602	719
5	10-7	0.08	0.07	0.03	0.02	0.01	0.00	604	707	590	702	595	709
2	7-5	0.17	0.18	0.06	0.06	0.02	0.00	587	721	573	708	587	708
3	7-5	0.09	0.11	0.04	0.03	0.01	-0.01	627	727	616	732	637	718
4	7-5	0.08	0.08	0.02	0.02	0.01	-0.00	592	704	595	724	623	711
5	7-5	0.06	0.05	0.02	0.02	0.01	0.00	650	704	645	718	644	703

tranches 1–5. From this analysis we obtain the forecast of the first tranche's parameter which we use in calculations of the final results.

Buying and selling tranches involve transaction charges. However, we do not have information on trading costs neither the entire history of the bid and ask prices. We only analyse the bid-ask spreads of Series 8. Table 3 shows an average distance of the bid spread and of the ask spread from the mid spread as a percentage of the mid spread. For the investigation of the trading strategies, the tranche spread data used in this study are adjusted in the following way. The protection buyer delivers an ask spread that is calculated as a mid spread increased by a proper percent listed in Table 3. The protection seller receives a bid spread which is calculated as a mid spread reduced by this percentage.

For every day  $w \leq t \leq T - h$  we construct a curve trade. Namely, we fit and forecast the DSFM model and calculate  $h$ -day forecasts of the MTM surfaces. From these surfaces we recover which two tranches and from which series optimize a given strategy. The accuracy of the predictions is evaluated by conducting a backtesting of the trades using the historical observations. For a given strategy and for tranches selected by the DSFM forecasting procedures we check the corresponding observed market spreads, calculate the resulting MTM values, and register the realized P&L. Tables 4 and 5 present the overall means of the daily gains in percent and the number of executed trades for the flattener and steeper trades

**Table 5**  
Steepeners. Calculations based on predictions of log-spreads and Z-transformed BCs marked as LZ; based only on Z-transformed BCs marked as Z.

Tranches	Maturity	Mean of daily gains in %						Number of executed trades					
		1 day		1 week		1 month		1 day		1 week		1 month	
		LZ	Z	LZ	Z	LZ	Z	LZ	Z	LZ	Z	LZ	Z
2	All	0.45	0.46	0.13	0.16	0.05	0.06	509	507	498	487	473	451
3	All	0.30	0.35	0.09	0.09	0.02	0.02	435	423	427	412	423	401
4	All	0.20	0.24	0.05	0.06	0.01	0.01	439	435	441	445	426	423
5	All	0.12	0.16	0.04	0.04	0.01	0.02	474	455	459	462	472	443
All	5–10	0.16	0.18	0.05	0.08	0.01	0.00	723	744	711	741	708	722
All	7–10	0.21	0.25	0.07	0.09	0.02	0.02	726	748	716	748	717	735
All	5–7	0.12	0.13	0.03	0.03	0.00	-0.01	747	749	740	746	717	727
2	5–10	0.33	0.48	0.13	0.35	0.05	0.12	80	80	76	76	61	61
3	5–10	0.02	0.11	-0.03	0.04	-0.05	-0.04	69	61	65	57	50	50
4	5–10	0.22	0.26	0.09	0.10	0.08	0.09	51	48	49	44	37	30
5	5–10	0.09	0.13	0.07	0.07	0.04	0.05	49	47	48	42	35	28
2	7–10	0.44	0.66	0.21	0.42	0.10	0.11	86	87	82	83	67	68
3	7–10	0.41	0.56	0.09	0.23	0.03	0.04	81	84	77	78	62	61
4	7–10	0.38	0.55	0.11	0.13	0.04	0.04	89	93	85	82	71	68
5	7–10	0.24	0.29	0.12	0.13	0.05	0.05	122	121	119	113	102	95
2	5–7	0.56	0.50	0.27	0.19	0.15	0.02	93	80	89	78	74	63
3	5–7	0.40	0.39	0.16	0.09	0.09	0.03	103	91	99	83	83	64
4	5–7	0.23	0.28	0.08	0.07	0.05	0.04	108	105	105	94	86	73
5	5–7	0.20	0.18	0.07	0.05	0.03	0.02	109	116	108	108	85	78

**Table 6**  
Joint flatteners and steepeners. Calculations based on predictions of log-spreads and Z-transformed BCs marked as LZ; based only on Z-transformed BCs marked as Z.

Tranches	Maturity	Mean of daily gains in %						Number of executed trades					
		1 day		1 week		1 month		1 day		1 week		1 month	
		LZ	Z	LZ	Z	LZ	Z	LZ	Z	LZ	Z	LZ	Z
All	All	0.30	0.30	0.11	0.13	0.04	0.03	754	754	750	750	735	735
2	All	0.33	0.28	0.12	0.13	0.05	0.04	752	753	745	748	729	735
3	All	0.18	0.23	0.07	0.07	0.02	0.02	754	754	750	750	735	735
4	All	0.12	0.18	0.04	0.05	0.01	0.01	754	754	750	750	735	734
5	All	0.08	0.11	0.03	0.04	0.01	0.01	741	744	735	739	735	733

respectively. For every trade the two tranches are selected either from a fixed seniority (e.g. choose always tranche 2) or always from all seniorities. Moreover, one can restrict the choice to fixed maturities (e.g. always buy 7Y maturity, sell 5Y) or choose from maturities. We also include a strategy that allows the investor to switch between flatteners and steepeners every day, see Table 6. If a strategy that combines flatteners and steepeners allows in addition choosing any tranche and any maturities, then the selected tranches are the maximum and the minimum of the forecasted MTM surface.

If for a particular day there are no tranches that for a given strategy return a positive P&L forecast, we assume that the investor decides not to take any action and we do not include this date in the overall summary of this strategy.

The spread predictions can alternatively be computed directly from the base correlations predictions by using (4). In consequence, it is not necessary to apply the DSFM to historical spreads. In Tables 4–6 the columns labelled with Z present the results obtained by modelling and forecasting the Z-transformed base correlations only.

The results show that the highest daily gains achieve the strategies that invest in tranche 2 and 3. Obviously, these tranches are quoted at the highest spreads but also carry the greatest risk. The steepeners for a fixed tranche and fixed maturities reveal a very good performance. However, these strategies were rarely carried out which means that the conditions of these strategies are difficult to meet. The models based entirely on the predictions of the base correlations achieve better results for one-day and one-week forecasting horizon. The models that combine the spread predictions and the base correlations predictions show better results for one-month forecasting horizon. Since the forecasting for the longer time horizons is less accurate, we observe a significantly better performance of the trades designed for short term periods.

## 6. Conclusions

This work investigates dynamics of collateralized debt obligations (CDOs) by modelling the evolution of tranche spread surfaces and base correlation surfaces using a dynamic semiparametric factor model (DSFM). The empirical study is conducted using an extensive data set of 49,502 observations of iTraxx Europe tranches of Series 2–10 for the time period between 30 March 2005 and 2 February 2009. The base correlations are implied from spreads using the large pool Gaussian copula model. The tranche spreads and the base correlations are represented as a function of the tranche seniority and the remaining time to maturity. Every day data appear in a small number of curves that form a surface in the three-dimensional

space. As time passes, the surfaces move through the space towards expiry and simultaneously change their shapes. The DSFM captures their evolution simultaneously in space and time dimensions by a small number of factors. We propose a modification of the classic DSFM and of the criterion of choosing the number of factors. The results show that the DSFM successfully reproduces the dynamics in data. The study is completed by presenting an application in trading strategies. We show how DSFM can be used in constructing the curve trades. Based on the DSFM predictions of the spread and base correlation surfaces we calculate the predictions of the marked-to-market (MTM) surfaces for different investment horizons. We analyse the performance of 43 strategies that combine different positions, tranches, and maturities. A backtesting using historical data shows that the curve trades achieve high daily gains.

## References

- [1] T.R. Bielecki, M. Rutkowski, *Credit Risk: Modelling Valuation and Hedging*, Springer Finance, 2004.
- [2] C. Bluhm, L. Overbeck, *Structured Credit Portfolio Analysis, Baskets and CDOs*, CRC Press LLC, 2006.
- [3] E.G. Bongiorno, A. Goia, E. Salinelli (Eds.), *Contributions in Infinite-Dimensional Statistics and Related Topics*, Societa Editrice Esculapio, Bologna, 2014.
- [4] H. Cardot, P. Sarda, Functional linear regression, in: F. Ferraty, Y. Romain (Eds.), *The Oxford Handbook of Functional Data Analysis*, in: Oxford Handbooks, Oxford University Press, 2011, pp. 21–46.
- [5] J.-M. Chiou, H.-G. Müller, J.-L. Wang, Functional response models, *Statist. Sinica* 14 (2004) 675–693.
- [6] A. Cuevas, A partial overview of the theory of statistics with functional data, *J. Statist. Plann. Inference* 147 (2014) 1–23.
- [7] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, 2001.
- [8] K. Detlefsen, W.K. Härdle, Variance swap dynamics, *Quant. Finance* 13 (5) (2013) 675–685.
- [9] R. Engle, Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *J. Bus. Econom. Statist.* 20 (3) (2002) 339–350.
- [10] J. Felsenheimer, P. Gisdakis, M. Zaiser, DJ iTraxx: Credit at its best!, *Credit derivatives special*, HVB Corporates & Markets, 2004.
- [11] M. Fengler, W.K. Härdle, E. Mammen, A semiparametric factor model for implied volatility surface dynamics, *J. Financ. Econ.* 5 (2) (2007) 189–218.
- [12] F. Ferraty, Y. Romain, *The Oxford Handbook of Functional Data Analysis*, in: Oxford Handbooks, Oxford University Press, 2011.
- [13] F. Ferraty, P. Vieu, *Nonparametric Functional Data Analysis: Theory and Practice*, in: Springer Series in Statistics, Springer, New York, 2006.
- [14] A. Goia, P. Vieu, Some advances on semi-parametric functional data modelling, in: E.G. Bongiorno, A. Goia, E. Salinelli (Eds.), *Contributions in Infinite-Dimensional Statistics and Related Topics*, Societa Editrice Esculapio, Bologna, 2014.
- [15] O. Gromenko, P. Kokoszka, L. Zhu, J. Sojka, Estimation and testing for spatially distributed curves with application to ionospheric and magnetic field trends, *Ann. Appl. Stat.* 6 (2) (2012) 669–696.
- [16] M. Hallin, R. Liška, Determining the number of factors in the general dynamic factor model, *J. Amer. Statist. Assoc.* 102 (478) (2007) 603–617.
- [17] A. Hamerle, A. Igl, K. Plank, Correlation smile, volatility skew, and systematic risk sensitivity of tranches, *J. Deriv.* 19 (3) (2012) 8–27.
- [18] W.K. Härdle, N. Hautsch, A. Mihoci, Modelling and forecasting liquidity supply using semiparametric factor dynamics, *J. Empir. Finance* 19 (4) (2012) 610–625.
- [19] L. Horváth, P. Kokoszka, *Inference for Functional Data with Applications*, Springer, New York, 2012.
- [20] A. Kakodkar, S. Galiani, J.G. Jónsson, A. Gallo, *Credit derivatives handbook*, Vol. 2: A guide to the exotics credit derivatives market, Technical Report, Merrill Lynch, 2006.
- [21] G. Kauermann, T. Krivobokova, L. Fahrmeir, Some asymptotic results on generalized penalized spline smoothing, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71 (2) (2009) 487–503.
- [22] A. Kneip, T. Gasser, Statistical tools to analyze data representing a sample of curves, *Ann. Statist.* 20 (3) (1992) 1266–1305.
- [23] D.X. Li, On default correlation: a copula function approach, *J. Fixed Income* 9 (4) (2000) 43–54.
- [24] X. Liu, H.-G. Müller, Functional convex averaging and synchronization for time-warped random curves, *J. Amer. Statist. Assoc.* 99 (467) (2004) 687–699.
- [25] L. McGinty, R. Ahluwalia, A model for base correlation calculation, Technical Report, JP Morgan, 2004.
- [26] B. Park, E. Mammen, W.K. Härdle, S. Borak, Time series modelling with semiparametric factor dynamics, *J. Amer. Statist. Assoc.* 104 (485) (2009) 284–298.
- [27] J.O. Ramsay, B.W. Silverman, *Functional Data Analysis*, second ed., Springer, 2005.
- [28] R. Roy, Trading credit tranches: Taking default correlation out of the black box, in: A. Rajan, G. McDermott, R. Roy (Eds.), *The Structured Credit Handbook*, John Wiley & Sons, 2007.
- [29] S. Song, W.K. Härdle, Y. Ritov, Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series, *Econom. J.* 17 (2) (2014) 101–131.
- [30] A. van Bömmel, S. Song, P. Majer, P. Mohr, H. Heekeren, W.K. Härdle, Risk patterns and correlated brain activities. Multidimensional statistical analysis of fMRI data in economic decision making study, *Psychometrika* 79 (3) (2014) 489–514.

# Reference Dependent Preferences and the EPK Puzzle \*

Maria Grith,<sup>†</sup> Wolfgang K. Härdle,<sup>‡</sup> Volker Krätschmer,<sup>§</sup>

## Abstract

Supported by several recent investigations, the empirical pricing kernel (EPK) puzzle might be considered a stylized fact. Based on an economic model with state dependent preferences for the financial investors, we want to emphasize a microeconomic view that succeeds in explaining the puzzle. We retain the expected utility framework in a one period model and illustrate the case when the state is defined with respect to a reference point. We further investigate how the model relates the shape of the EPK to the economic conditions.

KEYWORDS: Pricing kernel, aggregate agent, empirical pricing kernel, EPK puzzle, state dependent utilities, reference dependent utilities, reference points. JEL CLASSIFICATION: D04, D53, C02, G13  
AMS CLASSIFICATION: 15A29, 62G07, 62G35

---

\*This research was supported by Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

<sup>†</sup>Ladislaus von Bortkiewicz Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin; e-mail: gritmari@wiwi.hu-berlin.de.

<sup>‡</sup>Center for Applied Statistics and Economics, Ladislaus von Bortkiewicz Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin; e-mail: haerdle@wiwi.hu-berlin.de.

<sup>§</sup>Center for Applied Stochastics, Faculty of Mathematics, University of Duisburg-Essen, Forsthausweg 2, D-47057 Duisburg; e-mail: Volker.kraetschmer@uni-due.de.

# 1 Introduction

The empirical pricing kernel puzzle emerged as an empirical phenomenon in the financial markets, particularly with respect to the prices of European options written on the underlying stock index. Several authors have investigated if such patterns of the EPK can be justified in a general equilibrium setting and if the observed prices can be the outcome of investors' optimal behavior. The starting point for many of the investigations is settled within similar economic models that assume a representative agent in financial markets whose preferences have classical expected utility representation. Additionally, the risk neutral valuation principle is supposed to be valid for the financial markets by means of pricing kernels. If the pricing kernels represent state contingent equilibrium prices they might be identified with the v. Neumann-Morgenstern marginal utility indices of the representative agent.

Starting with Ait-Sahalia and Lo (2000), Jackwerth (2000), Engle and Rosenberg (2002), different econometric methods have been applied to estimate pricing kernels with varying underlying models for the financial markets. It turned out as a common result, that the estimates, the so called empirical pricing kernels (EPK), have non-monotonic shape regardless of the used data sets. Typically, we find either a U-shaped pricing kernel or a hump-shaped pricing kernel. In either cases the empirical kernels fail to be monotone, contrasting the standard theory of expected utility. This is what we shall call the *EPK puzzle*. Based on conditional estimates of the risk neutral and physical densities, it appears that periods of unusual low and stable realized and risk neutral volatility feature a hump shaped EPK, whereas during periods of high volatility the estimates look U-shaped. Several studies report the shape of the pricing kernel as being hump-shaped for most months between 2004 and 2007. This holds for both the German DAX 30 index Giacomini and Härdle (2008); Grith et al. (2012) and the American S&P 500 index Barone-Adesi et al. (2013); Beare and Schmidt (2012); Polkovnichenko and Zhao (2012).

Monotonicity tests for the EPK have been proposed by Golubev et al. (2008) who construct test for the local concavity of the utility function and Härdle et al. (2012) who build uniform confidence bands for the empirical pricing kernel; they apply the test to DAX 30 index EPK. Beare and Schmidt (2012) test the concavity of the ordinal dominance curve associated with the risk neutral and physical distributions associated with S&P 500 index. Typically, the null hypothesis of nonincreasing EPK was rejected.

Recent econometric models point at volatility as a state variable, that help explain the observed non-monotonicities in the pricing kernel. Chabi-Yo (2012); Song and Xiu (2012) find that, consistent with economic theory, the pricing kernel decreases in the market index return, conditional on the market volatility. As such, unconditional estimates of the PK may appear U-shaped. Christoffersen et al. (2012), propose an augmented Heston and Nandi (2000) model that allows for U-shaped pricing kernel in a one period model by introducing a variance preference parameter.

There is a large body of literature that investigates the mechanisms through which a locally increasing region in the pricing kernel can occur. Hens and Reichlin (2012) conduct a systematic analysis of the EPK puzzle by relaxing in turn the assumptions embedded in the standard expected utility models: complete markets, risk-averse investors and correct beliefs. They calibrate a hump-shaped pricing kernel and find that incomplete markets can alone explain the puzzle. The authors rule out local risk-proclivity, that works only as a 'pathological example with a few states'. With homogeneous agents, misestimation of objective probability in isolations misses some essential features of the data. This finding is in line with Ziegler (2007).

Closely related to the latter interpretation, heterogeneity in beliefs about the future realizations of the returns occurs in several papers as a possible interpretation for the EPK puzzle. Bakshi and Madan (2008); Bakshi et al. (2010) consider an equilibrium model with short and long equity investors that is able to explain U-shaped pricing kernel; in particular, the positively sloped regions in the pricing kernel occur when some investors are shorting equities. This model is able to explain some features of the option data: decreasing negative returns in strikes of the OTM calls and the even pronounced negative returns of put options, increasing in strike prices. However, it cannot capture the positive returns of call options for high strikes as reported in Bondarenko (2003). Ziegler (2007) considers three groups of heterogeneous agents with biased beliefs about the physical density but concludes that the estimates of the mean are not realistic for the pessimistic groups. Optimism and pessimism reflect biases in the first moment of the objective probabilities; Shefrin (2008) points out that one should consider higher order biases in order to explain the empirical findings and emphasizes the bias in the second moments that leads to risk neutral and physical distribution having different variance.



Some studies argue that modifications of standard preferences are needed to explain the data. Departing from the expected utility framework, Polkovnichenko and Zhao (2012) propose a rank dependent utility model and estimate probability weighting function nonparametrically. For most of the years the estimates are inverse S-shaped, consistent with a U-shaped PK but they become S-shaped in the years 2004-2007, suggesting a hump-shaped EPK. In line with experimental findings, inverse S-shaped weighting function imply that investors tend to overweight low-probability events while underweighting the likelihood of high-probability ones. The converse holds for the S-shaped probability weighting function but the authors do not make further investigations about the differences in these treatments. Hens and Reichlin (2012) show that a combination of reasonable pessimism and inverse S-shaped weighting function can explain the hump shaped EPK.

Shefrin (2008) rationalizes the EPK puzzle in a model with mixed expected utility maximizers and agents endowed with SP/A preferences - security, potential and aspiration theory, proposed by Lopes (1987) and developed in Lopes and Oden (1999). The idea that investors are endowed with utilities that mirror their concerns for portfolio maximization also pervades our paper.

Another stream of literature that tries to rationalize the EPK puzzle considers state dependence. State dependence has been traditionally used to explain the asset pricing puzzles in equilibrium models mainly based on two utility classes: habit formation, see Constantinides (1990), Campbell and Cochrane (1999), or recursive utilities, see Epstein and Zin (2001). In these papers, one typically assumes a Markov switching process for the evolution of states and derive asset related characteristics in a consumption based model. Garcia et al. (2003) investigate recursive utility functions with state dependency in the fundamentals. Melino and Yang (2003) disentangle the roles played by state dependent intertemporal substitution and time preference in explaining the risk aversion puzzle in a model with state dependent recursive preferences. Veronesi (2004) extends the state dependent utility by assuming that the agents possess a probability distribution over their state and introduces the concept of 'belief-dependent preferences'. A first explanation for the empirical pricing kernel puzzle via state dependence has been offered by Chabi-Yo et al. (2008), who generalize the setup of Melino and Yang (2003). The crucial idea of the authors is to suppose that regime switches are inherent of the price

process of the stock market. More specifically, within a discrete time period  $\{0, 1, \dots, T\}$ , there are two types of price processes  $(S_t^0)_{t \in \{0, \dots, T\}}$ ,  $(S_t^1)_{t \in \{0, \dots, T\}}$  for the risky asset which have joint continuous distributions, and constitute separately together with the riskless bond arbitrage free financial markets in the sense of section 2. Furthermore, they assume a latent regime switching variables in terms of an unobservable Markov-chain  $(U_t)_{t \in \{0, 1, \dots, T\}}$  of Bernoulli-distributed random variables. The observable price process  $(S_t)_{t \in \{0, 1, \dots, T\}}$  is then modeled by  $S_t = U_t S_t^1 + (1 - U_t) S_t^0$  for  $t \in \{0, \dots, T\}$ . Assuming the risk neutral valuation principle for the latent two basic financial markets and for the observable one, the authors drew a comparison of the associated pricing kernels via a simulation study. Indeed it turned out that the empirical pricing kernels in the separated financial market were nonincreasing whereas the empirical pricing kernel in the integrated financial market failed to have the property of monotonicity. Therefore the empirical pricing kernel might be explained by a switch of the price processes of the underlying in the financial market. The authors also investigate what type of conditioning - in preferences, economic fundamentals or beliefs - are more likely to explain the EPK puzzle over time. The time variant shape of the EPK is explained in Barone-Adesi et al. (2013) through optimism/pessimism and overconfidence/underconfidence defined as the difference in the first and second moments of the physical and risk neutral distribution. In this sense the authors find that the hump-shaped pricing kernel stems from a mix of optimistic overconfident and pessimistic underconfident agents. Grith et al. (2012) use the shape invariant model, a semi-parametric approach for multiple curves with shape-related nonlinear variation, to model the dynamics of the empirical pricing kernel (EPK) based on the hump feature. The approach allows to summarize the nonlinear variability with a few interpretable parameters that can be used to conduct a further analysis that links the shape of the pricing kernel to the business condition. They find that over periods of concerted negative evolution of the economic indicators, the EPK hump will move to the right in the returns space, increase its spread and shrink in vertical direction. Based on the initializing thought that regime switching is caused by changes of the investors' preferences our aim is to make the influence of these changes on the shape of the pricing kernels more explicit. We conjecture that the existing models with variance dependent component can be improved

by exploiting the time varying and possible nonmonotone relationship between returns and volatility. We apply the concept of reference points in a different context that it has been previously used in prospect theory, underlying another type of behavior that is not focused on loss aversion but performance comparative to a benchmark.

We propose a model that can accommodate both shapes of the EPK observed in the empirical literature while retaining the expected utility framework in a one period model and endow the financial investors with preferences that might be state sensitive. More technically, investors switch between two utility indexes - over terminal wealth sets - at a point that projected on the market index space we call 'reference point'. As a consequence, while the individual utility indices are concave, the market utility may have jumps in the aggregate wealth space. In equilibrium, this renders pricing kernel non-monotonic. Agents' heterogeneity with respect to their 'reference point' is summarized in the model by a distribution of the reference points. This, together with preference parameters will characterize the shape of PK.

## 2 Financial Market and Preferences

We consider a simple one period two-dates exchange economy model. Let  $[0, T]$  be the time interval of investment in the financial market, where  $t = 0$  denotes the present time and  $t = T \in ]0, \infty[$  the time of maturity. It is assumed that a riskless bond and a risky asset are traded in the financial market as basic securities. The price process of the riskless bond  $(B_t)_{t \in [0, T]}$  is defined by  $B_t = \exp(-\int_0^t r_x dx)$  via a deterministic Riemannian-integrable interest process  $(r_t)_{t \in [0, T]}$ . The price process of the risky asset  $(S_t)_{t \in [0, T]}$  is taken to be a nonnegative semimartingale with continuously distributed marginals  $S_t$ . Discrete time models may be also subsumed to this setting. Let us further suppose that the financial market is arbitrage free in the sense that there exists an equivalent martingale measure. We further assume that the risk neutral valuation principle is valid for nonnegative payoffs  $\psi(S_T)$ . Hence there is an unknown Radon-Nikodym density  $\pi$  of a martingale measure such that the price of any random payoffs  $\psi(S_T)$  is characterized by

$$\mathbb{E} [B_T^{-1} \psi(S_T) \pi]. \quad (1)$$

By factorization with some Borel-measurable  $\mathcal{K}_\pi$ , that we call  $\mathcal{K}_\pi$  pricing kernel (w.r.t.  $\pi$ ) with  $\mathbb{E}[\pi|S_T] = \mathcal{K}_\pi(S_T)$  we obtain

$$\int_0^\infty B_T^{-1} \psi(x) \mathcal{K}_\pi(x) p_{S_T}(x) dx, \quad (2)$$

where  $p_{S_T}$  denotes a density function of the distribution of  $S_T$ .

We will consider a portfolio choice problem that links risk attitudes of investors to the pricing rule of the financial markets. Within the classical framework, that assumes a representative agent, investor preferences may be represented by expected utilities  $\mathbb{E}[u\{\bar{w}(S_T)\}]$  depending on the aggregate final wealth  $\bar{w}(S_T)$ , with v. Neumann-Morgenstern utility index  $u$ . Under some further technical conditions one can show that there is some positive  $\beta$  such that

$$\left. \frac{du}{dx} \right|_{x=\bar{w}(s_T)} = \beta \mathcal{K}_\pi(s_T)$$

for every realization  $s_T$  of  $S_T$ . Within this framework the pricing kernel has to be nonincreasing due to concavity of the utility index  $u$ . We shall provide a simple economic model where the pricing kernel need not to be nonincreasing. The key idea is to consider the investors preferences representable by state dependent utilities. An axiomatic justification for this concept of state dependent preferences is provided by Karni et al. (1983).

### 3 A Microeconomic View on the EPK puzzle

#### 3.1 State Dependent Preferences

Let us assume that we have  $m$  investors who have exogenous initial wealth  $w_{10}, \dots, w_{m0} > 0$  and stochastic financial wealth in form of nonnegative random variables  $e_1(S_T), \dots, e_m(S_T)$ . Without loss of generality we assume that the numeraire bond equals one. This means that all the prices are discounted. The terminal wealth  $w_i(S_T)$  fulfills the individual budget constraint:

$$\int_0^\infty w_i(x) \mathcal{K}_\pi(x) p_{S_T}(x) dx \leq w_{i0} + \int_0^\infty e_i(x) \mathcal{K}_\pi(x) p_{S_T}(x) dx, \quad i = 1, \dots, m. \quad (3)$$

Financial wealth  $e_i(S_T)$  at  $t = T$  depends on the initial holdings of securities and the investment choice at  $t = 0$ . If we denote by  $\delta_i$  the fraction of the portfolio invested in the risky asset,  $e_i(S_T) = \delta_i(S_T - 1) + 1$  and  $\delta_i$  expresses the risk exposure given initial wealth  $w_{i0}$ .

The consumers are assumed to have state dependent utilities in terms of extended expected utility preferences within the terminology of Mas-Colell et al. (1995). In particular, this means that consumer  $i$  has numerical representation of her preferences as:

$$u^i\{S_T, w(S_T)\}$$

where  $u^i : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R} \cup \{-\infty\}$  denotes a state dependent v. Neumann-Morgenstern utility index satisfying:

$$u^i(x, y) \in \mathbb{R} \text{ for } x \geq 0, \quad y > 0, \quad (4)$$

$$u^i(x, \cdot) \text{ is strictly increasing and strictly concave for any } x \geq 0, \quad (5)$$

$$u^i(\cdot, y) \text{ is Borel-measurable for every } y \geq 0. \quad (6)$$

If  $u^i(x, \cdot)$  is continuously differentiable the usual *Inada conditions* are assumed to hold for  $i = 1, \dots, m$

$$\lim_{y \rightarrow 0} \frac{du^i(x, \cdot)}{dy} \Big|_y = \infty, \quad \lim_{y \rightarrow \infty} \frac{du^i(x, \cdot)}{dy} \Big|_y = 0. \quad (7)$$

Investors choose their optimal wealth  $(\bar{w}_1(S_T), \dots, \bar{w}_m(S_T))$  such that the following properties are fulfilled.

(ii) *individual optimization*: For each consumer  $i$ ,  $\bar{w}_i(S_T)$  solves

$$\begin{aligned} \max_{w_i(S_T)} \mathbb{E} \left[ u^i\{S_T, w_i(S_T)\} \right] \\ \text{s.t. } w_i(S_T) \text{ satisfies individual budget constraint (3).} \end{aligned} \quad (8)$$

(i) *market clearing*:

$$\sum_{i=1}^m \bar{w}_i(S_T) = \bar{w}(S_T). \quad (9)$$

The conditions (8) and (9) describe a weak version of a *contingent Arrow Debreu equilibrium* (Dana and Jeanblanc (2003), sect. 7.1). As a by product  $\bar{w}_1(S_T), \dots, \bar{w}_m(S_T)$  are Pareto optimum too, i.e. there are no  $w_1(S_T), \dots, w_m(S_T)$  with  $U^i\{w_i(S_T)\} \geq U^i\{\bar{w}_i(S_T)\}$  for every  $i$  and such that  $U^i\{w_i(S_T)\} > U^i\{\bar{w}_i(S_T)\}$  for at least one  $i$ . By *Negeishi method* cf. Dana and Jeanblanc (2003) we may find nonnegative weight vector  $\alpha$  s.t. the aggregate preferences have extended expected utility representation

$$E[u_\alpha\{S_T, \bar{w}(S_T)\}],$$

for the aggregate state dependent utility  $u_\alpha : \mathbb{R}_+^2 \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$  defined by

$$u_\alpha(x, y) \stackrel{\text{def}}{=} \sup_{\{y_i\}_{i=1}^m} \left\{ \sum_{i=1}^m \alpha_i u^i(x, y_i) \mid y_1, \dots, y_m \geq 0, \sum_{i=1}^m y_i \leq y \right\}.$$

These can be concluded from Lemma B.1, B.2 (cf. Appendix B). We impose a further condition on the asymptotic elasticity of the utilities that represents a minimal requirement to describe the optimal investment in terms of the marginal utilities and a pricing kernel.

$$\limsup_{y \rightarrow \infty} \frac{du^i(x, \cdot)}{dy} \Big|_y < 1 \quad \text{for any } x \geq 0 \text{ and every } i \in \{1, \dots, m\}. \quad (10)$$

The condition follows the guidelines of Kramkov and Schachermayer (1999); a similar condition appears in Dana and Jeanblanc (2003), Duffie (1996), Karatzas and Shreve (1998). We find this formulation more convenient to establish the following theorem.

**Theorem 3.1** *In addition to (4) – (10) let  $u^1(x, \cdot), \dots, u^m(x, \cdot)$  be twice continuously differentiable for  $x \geq 0$ . Then  $u_\alpha(x, \cdot)$  is continuously differentiable for every realization  $s_T$  of  $S_T$ . Furthermore for any  $\alpha_i > 0$  there exists some  $\beta_i > 0$  such that*

$$\frac{du_\alpha(s_T, \cdot)}{dy} \Big|_{y=\bar{w}(S_T)} = \alpha_i \frac{du^i(s_T, \cdot)}{dy} \Big|_{y=\bar{w}_i(S_T)} = \alpha_i \beta_i \mathcal{K}_\pi(s_T) = \beta \mathcal{K}_\pi(s_T)$$

for every realization  $s_T$ .

The proof of Theorem 3.1 is delegated to the end of Appendix A.

Theorem 3.1 is the corner stone for linking aggregated individual preferences to the market pricing kernel with its potential nonmonotonicities. If we assume that the initial aggregate wealth sums up to zero it is reasonable to conclude that market final wealth specializes to  $\bar{w}(S_T) = S_T$  if the bond is in zero net supply. Let  $R_T = \frac{S_T}{S_0}$  be the return at maturity. Theorem 3.1 reads as follows in terms of relative price.

**Corollary 3.2** *Let  $\bar{w}(R_T) = R_T$  and let  $u^1(x, \cdot), \dots, u^m(x, \cdot)$  be twice continuously differentiable for  $x \geq 0$ . Then under (4) – (10),  $u_\alpha(x, \cdot)$  is continuously differentiable for every realization  $r_T$ , of  $R_T$  and for any  $\alpha_i > 0$  there exists some  $\beta_i > 0$  such that*

$$\frac{du_\alpha(r_T, \cdot)}{dy} \Big|_{y=r_T} = \alpha_i \frac{du^i(r_T, \cdot)}{dy} \Big|_{y=\bar{w}_i(r_T)} = \beta \mathcal{K}_\pi(r_T) \stackrel{\text{def}}{=} \tilde{\mathcal{K}}_\pi(r_T),$$

for  $\bar{w}(R_T) = R_T$ . Without loss of generality we can assume that  $\beta = 1$ .

## 3.2 Reference Dependent Preferences

The framework of state dependent utilities of the investors allows us to describe a switching behavior of them when facing a threshold or a reference. We will consider a simple case when the reference is with respect to the future realization of the market return  $R_T$ . In more detail, let us assume that each investor  $i$  is disposed of two basic continuous, strictly increasing and strictly concave utility indices  $u_i^0, u_i^1 : [0, \infty[ \rightarrow \mathbb{R} \cup \{-\infty\}$  with  $u_i^0(y), u_i^1(y) \in \mathbb{R}$  for  $y > 0$ . She is changing between these indices dependent on a threshold  $x_i > 0$  in the space of future returns i.e.

$$u^i\{r_T, w_i(r_T)\} = u_i^0\{w_i(r_T)\} \mathbf{I}\{r_T \in [0, x_i]\} + u_i^1\{w_i(r_T)\} \mathbf{I}\{r_T \in (x_i, \infty)\} \quad (11)$$

for every realization  $r_T$  of  $R_T$ . The reader may think of  $u_i^0, u_i^1$  as utility indices representing bearish and bullish risk attitudes of investor  $i$ , and that her revealed attitudes are adapted to the prices of the financial market.

In order to simplify notations, let us assume that the thresholds are ordered by  $x_1 \leq \dots \leq x_m$ . There exist different competing potential representative agent groups in the market with representations of aggregate utility indices defined by

$$u_\alpha^j\{\bar{w}(R_T)\} = \sum_{k=1}^m \alpha_k u_k^0\{\bar{w}_k(R_T)\} \mathbf{I}\{k \geq j\} + \sum_{k=1}^m \alpha_k u_k^1\{\bar{w}_k(R_T)\} \mathbf{I}\{k < j\} \quad (12)$$

In view of Lemma B.1, B.2 in Appendix B they have expected utility representations

$$\mathbb{E}\left[u_\alpha^j\{\bar{w}(R_T)\}\right],$$

$j = 1, \dots, m+1$ . It is now a routine exercise to verify that

$$u_\alpha(x, y) = u_\alpha^1(y) \mathbf{I}\{x \in [0, x_1]\} + \sum_{i=1}^{m-1} u_\alpha^{i+1}(y) \mathbf{I}\{x \in (x_i, x_{i+1}]\} + u_\alpha^{m+1}(y) \mathbf{I}\{x \in (x_m, \infty)\} \text{ for } x, y \geq 0.$$

As a consequence the aggregate utility index might be interpreted as expressing the hegemony of different potential representative agents. Moreover, via Corollary 3.2 we obtain for some  $\beta > 0$  and any realisation  $r_T$  of  $R_T$  the expression for  $\tilde{\mathcal{K}}_\pi(r_T)$  is

$$\frac{du_\alpha^1(y)}{dy} \Big|_{y=r_T} \mathbf{I}\{r_T \in [0, x_1]\} + \sum_{i=1}^{m-1} \frac{du_\alpha^{i+1}(y)}{dy} \Big|_{y=r_T} \mathbf{I}\{r_T \in (x_i, x_{i+1}]\} + \frac{du_\alpha^{m+1}(y)}{dy} \Big|_{y=r_T} \mathbf{I}\{r_T \in (x_m, \infty)\}$$

From this observation it becomes clear that the pricing kernel is nonincreasing separately on the intervals  $[0, x_1[, ]x_1, x_2[, \dots, ]x_m, \infty[,$  but it might fail to be monotone just at the switching points  $x_1, \dots, x_m$ .

### 3.3 Reference Points and Pricing Kernel

To illustrate this point let us assume that the distribution of  $R_T$  has  $[0, \infty[$  as support, and that the investors have an identical switching point say  $x_1$ ; the market pricing kernel has the following representation

$$\frac{du_\alpha^1(y)}{dy} \Big|_{y=r_T} \mathbf{I}\{r_T \in [0, x_1]\} + \frac{du_\alpha^{m+1}(y)}{dy} \Big|_{y=r_T} \mathbf{I}\{r_T \in (x_1, \infty)\} = \tilde{\mathcal{K}}_\pi(r_T) \quad (13)$$

for every realization  $r_T$  of  $R_T$ .

From (12) one can show that  $u_\alpha^j$  inherits the properties of utility indices  $u_i^0$  and  $u_i^1$ : it is continuous, strictly increasing and strictly concave and fulfills the Inada conditions. Its first derivative has an inverse  $F_\alpha^j$  that is continuously differentiable and strictly decreasing. The application of Lemma B.1 and Proposition B.3 in Appendix B yields

$$r_T = F^1\left(\frac{du_\alpha^1(y)}{dy} \Big|_{y=r_T}\right) = F^{m+1}\left(\frac{du_\alpha^{m+1}(y)}{dy} \Big|_{y=r_T}\right)$$



for any positive realization  $r_T$ .

For example, let us suppose that each investor  $i$  switches between CRRA utilities  $u_i^j(y) = \frac{y^{1-\gamma_i^j}}{1-\gamma_i^j}$  with  $y > 0$  and Arrow-Pratt coefficients of relative risk aversion  $\gamma_i^j$  ( $j = 0, 1; 1 > \gamma_i^0 > \gamma_i^1 > 0$ ). It follows that  $u_1^0, \dots, u_m^0$  represent more risk averse attitudes than  $u_1^1, \dots, u_m^1$ . In particular for stock returns lower or equal  $x_1$  we have a bullish market, whereas we obtain a bearish market when stock returns exceed  $x_1$ . For this parametrization of the utility indices, the mappings  $F^j : [0, \infty) \rightarrow [0, \infty)$  are defined

$$F^j(z) = \sum_{\substack{i=1 \\ \alpha_i > 0}}^{m+1} \left( \frac{z}{\alpha_i} \right)^{\frac{1}{\gamma_i^j}} \quad (j = 0, 1)$$

If  $x_1$  is larger than the intersection of  $F^1$  and  $F^{m+1}$  then

$$F^1 \left( \frac{du_\alpha^1(y)}{dy} \Big|_{y=x_1} \right) = x_1 = F^{m+1} \left( \frac{du_\alpha^{m+1}(y)}{dy} \Big|_{y=x_1} \right) > F^1 \left( \frac{du_\alpha^{m+1}(y)}{dy} \Big|_{y=x_1} \right).$$

for any realization  $r_T \geq x_1$ . Therefore

$$\frac{du_\alpha^{m+1}(y)}{dy} \Big|_{y=x_1} > \frac{du_\alpha^1(y)}{dy} \Big|_{y=x_1}$$

That means that  $\widetilde{\mathcal{K}}_\pi$  is not monotone at  $x_1$ .

We illustrate the case of a single reference point for the following cases.

**Example 1.** Market utility indexes have  $u_\alpha^1$  and  $u_\alpha^{m+1}$  have power representation with different aggregate constant coefficients of relative risk aversion  $\gamma_\alpha^0$  and  $\gamma_\alpha^1$ .

$$r_T^{-\gamma_\alpha^0} \mathbf{I}\{r_T \in [0, x_1]\} + r_T^{-\gamma_\alpha^1} \mathbf{I}\{r_T \in (x_1, \infty)\} = \widetilde{\mathcal{K}}_\pi(r_T)$$

**Example 2.** Market utility indexes  $u_\alpha^1$  and  $u_\alpha^{m+1}$  have power representation with equal aggregate constant coefficients of relative risk aversion  $\gamma_\alpha$  but differ by a multiplicative constant  $b > 1$ .

$$r_T^{-\gamma_\alpha} \mathbf{I}\{r_T \in [0, x_1]\} + br_T^{-\gamma_\alpha} \mathbf{I}\{r_T \in (x_1, \infty)\} = \widetilde{\mathcal{K}}_\pi(r_T)$$

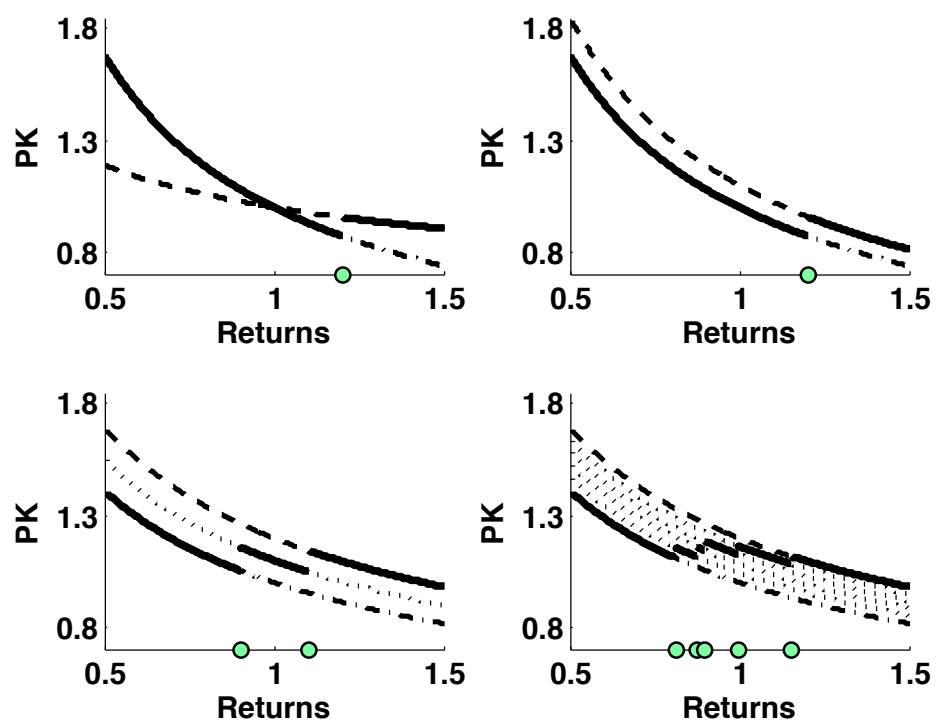


Figure 1:  $\frac{du_{\alpha}(r_T)}{dr_T}$  (solid),  $\frac{du_{\alpha}^j(r_T)}{dr_T}$  (dotted),  $\frac{du_{\alpha}^l(r_T)}{dr_T}$  (dashed-dotted) and  $\frac{du_{\alpha}^{m+1}(r_T)}{dr_T}$  (dashed)

A graphical illustration for these example is in figure 1: left panel top for  $\gamma_\alpha^0 = 0.75$  and  $\gamma_\alpha^1 = 0.25$  and  $x_1 = 1.2$ ; a jump of similar size is depicted in the right upper panel of the same figure for the case when utilities differ just by a constant  $u_i^1 = bu_i^0$  with  $b = 1.2$  and  $\gamma_\alpha = 0.75$ .

Next, we exemplify the case of investors with heterogeneous reference points  $x_i$ . For exposition purposes we will assume that the investors are equally important, that is  $\alpha_1 = \alpha_2 = \dots = \alpha_m = \alpha$ . In a simple case, we assume that all agents switch between the same two utility indices  $u_i^j(y) = u^j(y)$ , ( $j = 0, 1$ ) for all  $i = 1, \dots, m$ . Let us denote

$$F(r_T) = \frac{1}{m} \sum_{i=1}^m \mathbf{I}\{r_T \in (0, x_i)\}$$

the cumulative distribution function of the reference points;  $F$  is basically the share of agents that have preferences described by  $u^1$  at the realization  $r_T$ . The interpretation of the ordered reference points is the following: for  $x_1 < x_2$  we will say the investor 1 is more optimistic than the agent 2. The degree of heterogeneity of the agents with respect to their reference points is an indicator for market uncertainty. This point will be extended upon in section 6.

**Example 3.** We exemplify with the individual utility functions  $u^j$ ,  $j = 0, 1$ .

$$u^j(y) = \begin{cases} b_j \frac{y^{1-\gamma}}{1-\gamma} & \text{if } \gamma > 0 \text{ and } \gamma \neq 1 \\ b_j \log(y) & \text{if } \gamma = 1 \end{cases}$$

The positive constants  $b_0 < b_1$  retain the relationship between  $u^0$  and  $u^1$  in the previous example; in that sense  $b_1$  represent bullish attitudes. Given our parametric specifications for the utility indices and  $F$  we can rewrite the formulas for  $\tilde{\mathcal{K}}_\pi(r_T)$  developed in section 3.2 as

$$\tilde{\mathcal{K}}_\pi(r_T) = \left[ \frac{r_T}{\{1 - F(r_T)\}b_0^{\frac{1}{\gamma}} + F(r_T)b_1^{\frac{1}{\gamma}}} \right]^{-\gamma}, \quad (14)$$

for every possible realization  $r_T$  of  $R_T$ . We illustrate the results in Figure 1 for  $\gamma^0 = \gamma^1 = 0.5$ ,  $b_0 = 1$ ,  $b_1 = 1.2$  and  $m = 2$  (lower panel left) and  $m = 5$  respectively (lower panel right).

**Example 4.** If agents have homogeneous, state dependent CRRA preferences

$$u^j(y) = \begin{cases} \frac{y^{1-\gamma^j}}{1-\gamma^j} & \text{if } \gamma^j > 0 \text{ and } \gamma^j \neq 1 \\ \log(y) & \text{if } \gamma^j = 1 \end{cases}$$

the market pricing kernel can be written as a power function

$$\widetilde{\mathcal{K}}_\pi(r_T) = b(r_T) r_T^{-\gamma_\alpha(r_T)} \quad (15)$$

with non-constant coefficient of relative risk aversion  $\gamma_\alpha(r_T)$

$$\gamma_\alpha(r_T) = r_T \left[ \{1 - F(r_T)\} \frac{\bar{w}^0}{\gamma^1} + F(r_T) \frac{\bar{w}^1}{\gamma^0} \right]^{-1}$$

and

$$b(r_T) = \left[ \left\{ 1 - F(r_T) b_0^{\frac{1}{\gamma_0}} \right\} \frac{\bar{w}^0}{r_T} + F(r_T) b_1^{\frac{1}{\gamma_1}} \frac{\bar{w}^1}{r_T} \right]^{\gamma_\alpha(r_T)}$$

for  $\bar{w}^j$  the optimal wealth path in state  $j$ ,  $j = 0, 1$ .

**Example 5.** Introducing state dependence in both  $b$  and  $\gamma$  results in a pricing kernel of the form (15) with  $b(r_T) = \left[ \{1 - F(r_T)\} \frac{\bar{w}^0}{r_T} + F(r_T) \frac{\bar{w}^1}{r_T} \right]^{\gamma_\alpha(r_T)}$ .

A further generalization of the previous examples is possible if we consider heterogeneity of agents in CRRA,  $\gamma_i^j$  and/or constants  $b_i^j$ . However, then the link to  $F$  is lost. We will use notations  $\mathcal{K}_{\theta, F} = \widetilde{\mathcal{K}}_\pi(r_T)$  for the models described in Examples 3 through 5, for  $\theta = (b, \gamma) \top$  a parameters vector describing preferences.

## 4 Investors' Portfolio Choice

From Corrolary 3.2 and Appendix A we can establish the relationship between the optimal terminal wealth of investor  $i$  and the market pricing kernel

$$\bar{w}_i(r_T) = I_i\left\{r_T, \frac{1}{\alpha_i} \tilde{\mathcal{K}}_\pi(r_T)\right\} \text{ for } i = 1, \dots, m \quad (16)$$

More explicitly, given the reference dependent utility specification in equation (11)

$$I_i\left\{r_T, \frac{1}{\alpha_i} \tilde{\mathcal{K}}_\pi(r_T)\right\} = \bar{w}_i^0(r_T) \mathbf{I}\{r_T \in [0, x_i]\} + \bar{w}_i^1(r_T) \mathbf{I}\{r_T \in (x_i, \infty)\} \quad (17)$$

where  $\bar{w}_i^j(r_T) = I_i^j\left\{\frac{1}{\alpha_i} \tilde{\mathcal{K}}_\pi(r_T)\right\}$ , for  $I_i^j(\cdot)$  continuously differentiable, strictly decreasing on  $]0, \infty[$ , the inverse functions of  $\frac{du_i^j(y)}{dy}$ ,  $j = 0, 1$ .

At the same time, the optimal wealth  $\bar{w}_i(r_T)$  also satisfies

$$\bar{w}_i(r_T) = w_{i0} + \delta_i(r_T - 1) + 1. \quad (18)$$

for every realization  $r_T$  of  $R_T$ . Equating the right hand side of equations (16) and (18), and taking expectations we can derive the optimal weight invested in the risky asset

$$\delta_i^* = \frac{\mathbb{E}[\bar{w}_i^0(r_T) \mathbf{I}\{r_T \in [0, x_i]\}] + \mathbb{E}[\bar{w}_i^1(r_T) \mathbf{I}\{r_T \in (x_i, \infty)\}] - w_{i0} - 1}{\mathbb{E}(r_T) - 1} \quad (19)$$

For  $u_i^0$  denoting bearish and  $u_i^1$  bullish attitudes, in the sense that there exists a threshold  $x$  so that for

$$\frac{du_i^1(y)}{dy} > \frac{du_i^0(y)}{dy} \text{ for } y \geq x,$$

the investors invest a higher fraction of wealth in the risky assets when  $x_i \geq x$  is small.

This is because  $\bar{w}_i^1(r_T) > \bar{w}_i^0(r_T)$  for  $r_T \geq x$ . The risk attitudes induced by a relatively smaller reference point  $x_i$  we will call 'optimism'. Obviously, the higher  $\delta_i^*$  is the higher is investors's expected wealth  $\mathbb{E}[\bar{w}_i(r_T)]$ . These are typically the agents that will take a long position in the risky assets, while short selling might occur for agents that have their reference points further to the right. Bakshi et al. (2010)

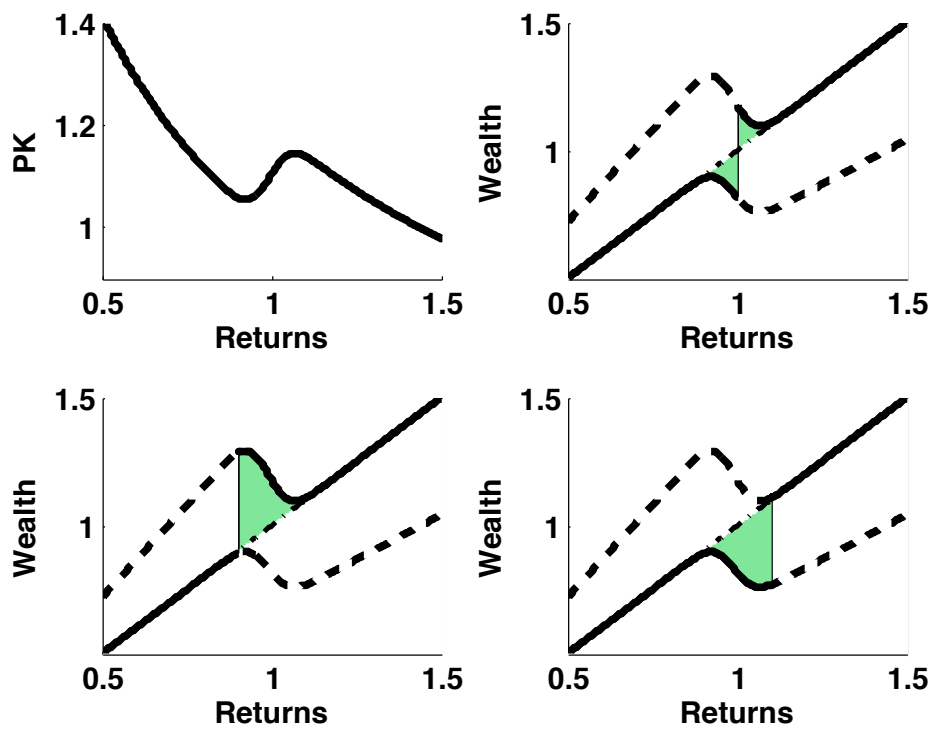


Figure 2: Market pricing kernel and (scaled) final wealth of three type of agents: mixed agent (upper right); optimistic agent (lower left) and pessimistic agents (lower right);  $m\bar{w}_i(r_T)$  (solid),  $m\bar{w}_i^0(r_T)$  and  $m\bar{w}_i^1(r_T)$  (dotted)

suggest that investors shorting equities possibly generate a positively sloped region in the pricing kernel.

Terminal wealth for three types of agents is illustrated in figure 2. The 45 degree line depicts the wealth of the aggregate agent contrasting to the optimal wealth allocated to the individual investors. The portfolio of an 'optimistic' investor 'beats' the market for realizations of  $r_T$  at the right of its reference point for the increasing region of the pricing kernel, whereas the portfolio of a pessimistic agents underperforms compared to the benchmark at the left of the reference points for the mixed and pessimistic type.

## 5 Simulation Study

### 5.1 Comparative Statics

According to section 2, the price of the risky asset at  $t = 0$  is given by

$$S_0 = \int_0^\infty s_T \tilde{\mathcal{K}}_\pi(s_T) p_{s_T}(s_T) ds_T. \quad (20)$$

For a fixed probability density function  $p_{s_T}$  the pricing kernel  $\mathcal{K}_\pi$  has a direct effect on the price at  $t = 0$  through the way it weights the possible realizations of  $s_T$ . For  $\mathcal{K}_{\theta,F} = \tilde{\mathcal{K}}_\pi$  we analyze the effects that the model's  $F$  and  $\theta$  have on the price  $S_0$ . The baseline model given by equation (14) for  $b = b_1/b_0$  and  $b_0 = 1$  is marked with solid line in figure (3).

We parametrize  $F$  to be  $N(1, 0.05)$  and we investigate the effect that the change in the mean and variance of the distribution has on the price in the upper panels of figure (3). A decrease in the mean results in higher weights associated with higher realizations for nonzero values of  $dF(\cdot)$ , while a decrease in the variance makes the hump more pronounced by simultaneously lowering the weights of lower realizations and increasing those of higher realizations around the nonzero values of  $dF(\cdot)$ . In the first case this is due to the prevalence of optimistic investors that tilt their portfolios towards the risky asset, triggering an increase in price  $S_0$ ; in the second case, the heterogeneity of investors' reference points

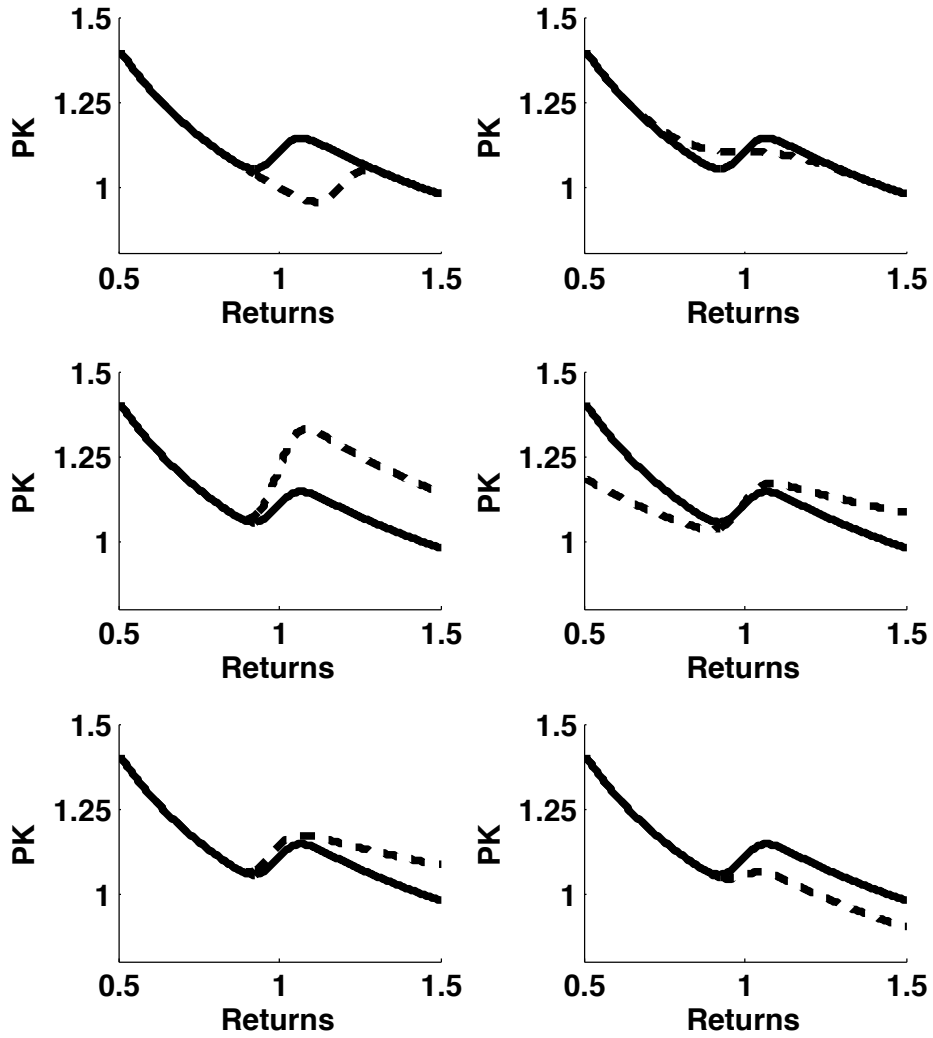


Figure 3: Impact of model parameters on the shape of PK: baseline model (solid):  $\gamma = 0.5$ ,  $b = 1.2$ ,  $F = N(1, 0.05)$ ; comparative models (dashed) left panel up  $F = N(1.2, 0.05)$ ; right panel up  $F = N(1, 0.15)$ ; left panel middle  $b = 1.4$ ; right panel middle  $\gamma = 0.25$ ; left panel down  $\gamma_1 = 0.25$ ; right panel down  $F = 1/2N(1, 0.05)$



$x_i$ -s is lower; this increases the slope of the upward region without significant effects on the price. We also observe that for small mean and large variance of  $F$  the humped feature disappears.

The next two panels depict the shape of the pricing kernel under various  $b$ -s and  $\gamma$ -s. We notice that for higher  $b$  the weights associated with higher returns are higher and hence large price  $S_0$ . In this example, varying  $\gamma$  makes pricing kernel 'rotate' around the value of  $r_T$  corresponding to the mean of  $F$ . Lower CRRA results in higher weights for higher returns and lower rates for lower returns realizations, over all domain of  $r_T$ . The overall effect is an increase in the price in a similar fashion it produces in state independent preferences case, by reducing the price per unit of probability of bad states and conversely for the good states. If we let CRRA to vary between the two states and apply pricing kernel specification in equation (15) we can see how the divergence between  $\gamma_1$  and  $\gamma_2$  affect the shape of the pricing kernel and consequently the price  $S_0$ .

Finally, in the lower panel right, we allow for a ratio of investors to have state independent preferences of type  $u^0$  (as specified in the baseline model). These influence the price  $S_0$  in a negative and this effect is more pronounced the higher the ratio of agents with preferences  $u^0$  is. Obviously, the predictions for the change in  $S_0$  will be in the opposite direction for state independent preferences of type  $u^1$ .

## 5.2 Identifiability

In this subsection, we discuss some aspects related to the applicability of the model proposed in the previous section in practice, when we try to fit it to empirical pricing kernel  $\widehat{\mathcal{K}}$ . If we denote  $\widehat{\mathcal{K}}(s_j) = y_j$  the estimates of the pricing kernel at observation points  $s_j$ , for  $j = 1, \dots, n$  and assume that

$$y_j = \mathcal{K}_{\theta, F} + \varepsilon_j, \text{ with } \varepsilon_j \sim (0, \sigma^2) \quad (21)$$

the fitting problem involves finding  $\theta^*$  and  $F^*$  that minimize

$$\sum_{j=1}^n \{y_j - \mathcal{K}_{\theta, F}(s_j)\}^2, \quad (22)$$

or a weighted version of it. We demonstrate the inverse problem in a simulation exercise, for  $\mathcal{K}_{\theta, F}$

given by (14) and zero error term. The pricing kernel in figure 4 was generated for parameters  $\gamma^0 = \gamma^1 = 0.5$ ,  $b = 1.2$  and  $F$  a *edf* of 400 random reference points from a normal distribution  $N(1, 1.2)$ . The two panels on the left depict the pricing kernel and  $F$ ; the dashed line marks the regions where  $F$  takes values 0 or 1. These are the regions that allow us to identify parameters  $b$  and  $\gamma$ , and consequently  $F$ . However, if the probability density function associated with  $F$  doesn't have compact support on the observed domain, these components can not be identified without further restrictions. The right panel up in figure 4 zooms in the pricing kernel at its left side so that the dashed lines are no more visible. This allows us to illustrate the case of non-identifiability of the model; underneath this panel we plot different combinations for  $\gamma$ ,  $b$  and  $F$  that give a perfect fit of the PK above. For instance, the top fascicle of dotted curves depicts  $F$  for  $b = 1.2$  and  $\gamma = (0.46, 0.47, 0.48, 0.49, 0.50, 0.52)$ , and for the next two bundles of curves we vary  $b$  to 1.3 and 1.5 respectively. Obviously, these combinations of parameters will determine the shape of the pricing kernel in the tails, where they diverge from the true pricing kernel in various degrees.

This exercise is relevant in practice; in particular, observations in the tails are sparse and the pointwise confidence intervals (or confidence bands) for the EPK are wider in the tails regions. This means that when trying to fit the model to the real data there will be a set of possible solutions that minimize the objective function (22). The characterization of these solutions are beyond the scope of this paper and constitutes the object of future work.

## 6 Real Data Analysis

Due to the identification problems explained in section 5.2, a quantitative analysis in terms of  $\theta$  and  $F$  over time is not feasible due to the multiplicity of solutions. The authors are investigating possible solutions under suitable constraints in a concurring study. However, the comparative statics analysis in subsection 5.1 allows us to make a qualitative evaluation of the model for dynamically estimated PK. Further on, we refer to the results of Grith et al. (2012), GHP as of now. Their EPK estimates relate to the European call and put options written on the German DAX 30 index, between June 2003 and May

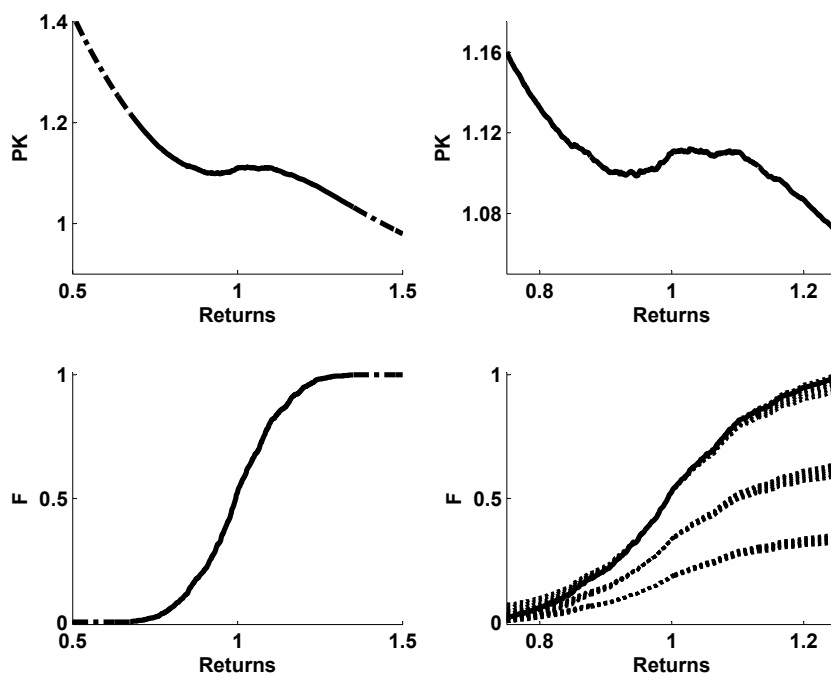


Figure 4: Parameters identifiably: compact support (left) and non-compact support (right) for the pdf of F on the observed domain

2006, at a monthly frequency. The authors assume that the conditional physical density is stationary, that is,  $p_{S_t}$  evolves slowly and most of the variation in the pricing kernel is due to  $q_{S_t}$ . If we extend the equation 20 to the contingent claims, we can explain the time variable patterns of the option prices through the changes in the pricing kernel. GHP relate the time variability of the pricing kernel hump to the economic conditions; in table 4 they report significant correlations between the changes in the shape of the EPK and the business indicators.

The changes in the height of the hump varies positively with the return on the index. The increase in the 'peak' in our model can be induced either via  $F$  or through a larger  $b$  ( $b_1$ ) or lower  $\gamma$  ( $\gamma_1$ ). The later causes an increase in the hump's spread, which is at odds with another finding of the GHP paper that suggests that the spread and the height of the peak are negatively related. It means, that in terms of our model, the mechanism that triggers an increase in the peak works through  $b$  and/or  $F$ . This suggestion is supported in the model proposed by Basak and Pavlova (2012), who add to the utility function of their institutional investors a state dependent component that is directly related to the performance of the index; while the retail investors have standard preferences. The fraction of institutional investors is a key parameter in their model and its increase exercises pressure on the stock index pushing it up; the same effect is present in our model by increasing the number of agents that have  $u^1$  type of preferences (or have reference dependent preferences).

The height of the EPK hump might respond to the business conditions as well, as suggested by the correlations with the credit spread - the difference between the yield on the corporate bond, based on the German CORPTOP Bond maturing in 3-5 years, and the government bond maturing in 5 years. Its countercyclical relation to the economic conditions and the negative relation to the height of the peak imply that its decrease pushes up the level of the peak. It is not yet clear how co-movements between  $b$  and  $F$  happen in the dynamics but the evidence so far seems to suggest that  $b$  may be interpreted as a magnitude parameter, that is increasing in  $S_t$  over time, while the overall economic conditions impact  $F$ .

The scale and shape parameters that modify the PK in the horizontal direction respond to changes in the yield term slope. The slope, computed as the difference between the 30-year government bond

yield and three-month interbank rate, has been shown to be pro-cyclical in Estrella and Hardouvelis (1991). A smaller slope shifts the increasing region of the PK to the right and widens its spread. These effects become effective in our model through the positive changes in the first two moments of  $F$ , meaning an increase in the pessimism and diversion of investors' reference points on the domain of future returns.

The arguments above suggest that our model delivers sensible mechanisms of PK's dynamics. We observe that at least what the changes in the EPK shape are concerned, they do not necessarily involve  $\gamma$ . It is possible that through this parameter, models that mimic other features of the pricing kernels, that are not consistent with the PK puzzle - e.g. generalized disappointment aversion model in Routledge and Zin (2010) - be reproduced; such generalizations necessitate further efforts and constitute material for new studies. On the other hand, it is possible that the mechanism that we suggest only manifests in certain circumstances while agents have permanent structural biases; explanation of inverse S-shaped weighting function Polkovnichenko and Zhao (2012) may practically hold for all periods but cease to capture some features in the data during some economic conditions. We do not rule out the possibility that the asset prices depend on investors' subjective beliefs regarding future realizations of  $S_T$  and our model can incorporate such extensions. Based on our analysis, we find that the investors incorporate information from the other part of the economy when making investment decisions. Our explanation of reference dependent preferences seems a plausible explanation for the time varying shape of the EPK.

## 7 Conclusions

Based on our specification for the marginal investors' preferences, the v. Neumann-Morgenstern utility index of the aggregate agent might switch between different 'regimes', meaning possible jumps in the pricing kernel. We empirically investigate its switching behavior in a simulation study and interpret the time varying patterns of real data in connection to our model. The theoretical model encompasses a fixed investment horizon, since we are only taking a snapshot of the market and try to explain

the observed shape in the pricing kernel. The natural extension for building a dynamic equilibrium model, starting from the static approach is to endogenize the formation of reference points. 'Keeping up with the Joneses' or status concerns Hong et al. (2012), the history of previous gains and losses Barberis et al. (2001), learning Benzoni et al. (2011), performance relative to a benchmark Basak and Pavlova (2012); Tang and Xiong (2012) are further possible explanations and extensions that need to be investigated and that come close to our approach. The model can be extended to other markets: commodities, interest rate and credit derivatives, in order to investigate if similar behavior occurs.

## A Appendix

The aim of this section is to provide a proof for Theorem 3.1. We continue with the model of section 3, retaking all assumptions and notations. Firstly, we characterize the optimal terminal wealth  $\bar{w}_1(S_T), \dots, \bar{w}_m(S_T)$  of the individual investor.

The Inada conditions together with (5) imply that for any  $i \in \{1, \dots, m\}$  and every  $x \geq 0$  the mapping  $\frac{du^i(x, \cdot)}{dy} | ]0, \infty[$  is injective onto  $]0, \infty[$  with continuously differentiable, strictly decreasing inverse say  $I_i(x, \cdot)$ . This enables us to apply the dominated convergence theorem to show

(A1) continuity of mappings

$$g_{s_T}^i : ]0, \infty[ \rightarrow \mathbb{R}, y \mapsto I_i\{s_T, y \mathcal{K}_\pi(s_T)\} \cdot \mathcal{K}_\pi(s_T) \quad (s_T \geq 0, i \in \{1, \dots, m\}).$$

$$(A2) \quad \lim_{y \rightarrow 0} g_{s_T}^i(y) = \infty \text{ and } \lim_{y \rightarrow \infty} g_{s_T}^i(y) = 0.$$

We are now ready to extend the classical characterization of the optimal terminal wealth to the case of extended expected utility preferences.

**Theorem A.1** *Assuming (4) – (10), there exists  $y_i > 0$  such that*

$$\bar{w}_i(S_T) = I_i\{S_T, y_i \mathcal{K}_\pi(S_T)\} \text{ for every } i = 1, \dots, m$$

**Proof:**

Let us fix  $i \in \{1, \dots, m\}$  and denote  $z_i \stackrel{\text{def}}{=} w_0^i + \mathbb{E}[e_i(S_T) \mathcal{K}_\pi(S_T)]$ . Since  $z_i > 0$  we may find in view of (A1), (A2) some  $y_i > 0$  with  $g(y_i) = x_i$ .

Let  $w(S_T)$  be a nonnegative random variable with  $\mathbb{E}[w(S_T) \mathcal{K}_\pi(S_T)] \leq z_i$ . Then

$$\begin{aligned} \mathbb{E}[u(S_T, w(S_T))] + y_i \{z_i - \mathbb{E}[w(S_T) \mathcal{K}_\pi(S_T)]\} &= y_i z_i + \mathbb{E}[u(S_T, w(S_T)) - y_i w(S_T) \mathcal{K}_\pi(S_T)] \leq \\ y_i z_i + \sup_{x \geq 0} \mathbb{E}[u(S_T, x) - y_i x \mathcal{K}_\pi(S_T)] &= \\ y_i z_i + \mathbb{E}[u(S_T, I_i(S_T, y_i \mathcal{K}_\pi(S_T))) - y_i I_i(S_T, y_i \mathcal{K}_\pi(S_T)) \mathcal{K}_\pi(S_T)] &= \mathbb{E}[u(S_T, I_i(S_T, y_i \mathcal{K}_\pi(S_T)))] \end{aligned}$$

Therefore  $I_i(S_T, y_i \mathcal{K}_\pi(S_T))$  solves the optimization problem of investor  $i$ . Moreover, the numerical representation  $U_i$  of investor's  $i$  preferences is strictly concave in view of strict concavity of  $u^i(x, \cdot)$  for every  $x \geq 0$ . In particular  $I_i(S_T, y_i \mathcal{K}_\pi(S_T))$  is the unique solution, hence being identical with  $\bar{w}_i(S_T)$ .  $\square$

Before starting with the proof of Theorem 3.1 let us consider for purposes of reference the classical case of the investor being expected utility maximizer. Indeed as an additional corollary of Theorem 3.1, we may retain the folk result concerning the risk neutral price valuation and the v. Neumann-Morgenstern utility index of the representative agent. More precisely, let us assume that there exist mappings  $u_1, \dots, u_r$  from  $\mathbb{R}_+$  into  $\mathbb{R} \cup \{-\infty\}$  satisfying  $u^1(x, \cdot) = u_1, \dots, u^m(x, \cdot) = u_m$  for  $x \geq 0$ , and

$$(A3) \quad u_1(y), \dots, u_m(y) \in \mathbb{R} \text{ for } y > 0,$$

$$(A4) \quad u_1, \dots, u_m \text{ are continuous, strictly increasing as well as strictly concave.}$$

Then

$$u(y) \stackrel{\text{def}}{=} \sup \left\{ \sum_{i=1}^m \alpha_i u_i(y_i) \mid y_1, \dots, y_m \geq 0, \sum_{i=1}^m y_i \leq y \right\} = u_\alpha(x, y) \text{ for } x, y \geq 0.$$

We shall impose the so called *Inada conditions* on the state independent utility indices  $u_1, \dots, u_m$ , i.e.

$$(A5) \quad u_1|]0, \infty[, \dots, u_m|]0, \infty[ \text{ are assumed to be continuously differentiable satisfying}$$

$$\lim_{e \rightarrow 0} \frac{du_i}{dy} \Big|_{y=e} = \infty, \quad \lim_{e \rightarrow \infty} \frac{du_i}{dy} \Big|_{y=e} = 0 \quad (i = 1, \dots, m).$$

(A6)  $E[I_1(y\mathcal{K}_\pi(S_T))], \dots, E[I_m(y\mathcal{K}_\pi(S_T))] < \infty$  for any  $y > 0$ , where  $I_1, \dots, I_r$  denote the inverses of  $\frac{du_1}{dy}, \dots, \frac{du_m}{dy}$  respectively.

We may conclude immediately from Theorem 3.1 the announced result.

**Proof of Theorem 3.1:**

Without loss of generality let us set  $\{1, \dots, r\} \stackrel{\text{def}}{=} \{i \in \{1, \dots, m\} \mid \alpha_i > 0\}$ . Then, defining  $g_i \stackrel{\text{def}}{=} \alpha_i u_i$ , we have  $u_\alpha = \sum_{i=1}^r g_i$ , and we may apply Lemma B.1, B.2 and Proposition B.3 (cf. Appendix B). Then, in view of Lemma B.1, B.2 and B.3, we obtain

$$u_\alpha(s_T, \bar{w}(s_T)) = \sum_{i=1}^r \alpha_i u^i(s_T, \bar{w}^i(s_T))$$

for every realization  $s_T$  of  $S_T$ .

On one hand by Theorem A.1, there exist  $y_1, \dots, y_m > 0$  such that

$$\bar{w}_i(s_T) = I_i(s_T, y_i \mathcal{K}_\pi(s_T)) > 0 \text{ for } i = 1, \dots, r.$$

On the other hand, due to Proposition B.3,  $u_\alpha(s_T, \cdot) \mid ]0, \infty[$  is differentiable for every realization  $s_T$ , satisfying

$$\alpha_i \frac{du^i(s_T, \cdot)}{dy} \Big|_{y=\bar{w}^i(s_T)} = \frac{du_\alpha(s_T, \cdot)}{dy} \Big|_{y=\bar{w}(s_T)}$$

for  $i \in \{1, \dots, r\}$  and any realization  $s_T$ . Notice that by construction the random variable  $\bar{w}(S_T)$  has strictly positive outcomes only. Now, the statement of Theorem 3.1 is clear.

## B Appendix

Throughout this section let the mappings  $g_1, \dots, g_r : \mathbb{R}_+^2 \rightarrow \mathbb{R} \cup \{-\infty\}$  satisfy the following conditions:

(B0)  $g_1(x, y), \dots, g_r(x, y) \in \mathbb{R}$  for  $x \geq 0, y > 0$ ;

(B1)  $g_1(x, \cdot), \dots, g_r(x, \cdot)$  are continuous, strictly increasing and strictly concave for  $x \geq 0$ ;

(B2)  $g_1(\cdot, y), \dots, g_r(\cdot, y)$  are Borel-measurable for  $y \geq 0$ .



Furthermore, let  $g : \mathbb{R}_+^2 \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$  be defined by

$$g(x, y) = \sup \left\{ \sum_{i=1}^r g_i(x, y_i) \mid y_1, \dots, y_r \geq 0, \sum_{i=1}^r y_i \leq y \right\}.$$

Indeed  $g(x, 0) = \sum_{i=1}^r g_i(x, 0) \in \mathbb{R} \cup \{-\infty\}$  for  $x \geq 0$ , and

$$-\infty < \sum_{i=1}^r g_i(x, \frac{y}{r}) \leq g(x, y) \leq \sum_{i=1}^r g_i(x, y) < \infty$$

for  $x \geq 0, y > 0$  due to (B0), (B1).

**Lemma B.1** For any  $x, y \geq 0$  there is some unique  $\phi(x, y) = (\phi_1(x, y), \dots, \phi_r(x, y)) \in \mathbb{R}_+^r$  such that  $\sum_{i=1}^r \phi_i(x, y) \leq y$  and

$$\sum_{i=1}^r g_i(x, \phi_i(x, y)) = g(x, y).$$

Furthermore,  $\sum_{i=1}^r \phi_i(x, y) = y$ .

**Proof:**

Let  $x, y \geq 0$ . For  $y = 0$  the statement of Lemma B.1 is obvious. So let  $y > 0$ , which means  $g(x, y) \in \mathbb{R}$ .

Due to (B1), the mapping

$$f : \left\{ (y_1, \dots, y_r) \in \mathbb{R}_+^r \mid \sum_{i=1}^r y_i \leq y, \sum_{i=1}^r g_i(x, y_i) \geq g(x, y) - 1 \right\} \rightarrow \mathbb{R}, (y_1, \dots, y_r) \mapsto \sum_{i=1}^r g_i(x, y_i)$$

is continuous, strictly concave, and defined on a nonvoid convex compact set. Therefore  $f$  attains its maximum at a unique  $\phi(x, y)$ . Obviously,  $\sum_{i=1}^r \phi_i(x, y) = y$  because  $f$  is strictly increasing too by (B1).

The proof is complete.

Lemma B.1 defines a mapping  $\phi = (\phi_1, \dots, \phi_r) : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+^r$ . It is Borel-measurable as will be shown now.

**Lemma B.2**  $\phi$  is Borel-measurable.

**Proof:**

It suffices to show that  $\phi^{-1} \left( \prod_{i=1}^r [0, a_i] \right)$  is a Borel-subset of  $\mathbb{R}_+^2$ . For this purpose define for any  $(a_1, \dots, a_r)$  from  $\mathbb{R}_+^r$  the mapping  $g_{a_1 \dots a_r} : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R} \cup \{-\infty\}$  by

$$g_{a_1 \dots a_r}(x, y) = \sup \left\{ \sum_{i=1}^r g_i(x, y_i) \mid (y_1, \dots, y_r) \in \prod_{i=1}^r [0, a_i], \sum_{i=1}^r y_i \leq y \right\}.$$

Notice that  $g_{a_1 \dots a_r}(x, y) \in \mathbb{R}$  for  $x \geq 0, y > 0$ , analogously to  $g(x, y) \in \mathbb{R}$  for  $x \geq 0, y > 0$ . Furthermore  $g_1(x, \cdot), \dots, g_r(x, \cdot)$  are continuous for any  $x \geq 0$ . Hence, setting  $\mathcal{R}_{a_1 \dots a_r} = \prod_{i=1}^r [0, a_i] \times \mathbb{Q}^m$ ,

$$g_{a_1 \dots a_r}^{-1}(]z, \infty[) = \bigcup_{(y_1, \dots, y_r) \in \mathcal{R}_{a_1 \dots a_r}} \left( \sum_{i=1}^r \alpha_i g_i(\cdot, y_i) \right)^{-1} (]z, \infty[) \times \left[ \sum_{i=1}^r y_i, \infty[ \quad (z \in \mathbb{R}).$$

Thus  $g_{a_1 \dots a_r}^{-1}(]z, \infty[)$  is a Borel-subset of  $\mathbb{R}_+^2$  for every  $z \in \mathbb{R}$  by assumption (B2). Then we may conclude that

$$\phi^{-1} \left( \prod_{i=1}^r [0, a_i] \right) = \left( \sup_{(b_1, \dots, b_r) \in \mathbb{Q}_+^r} g_{b_1 \dots b_r} - g_{a_1 \dots a_r} \right)^{-1} (\{0\})$$

is a Borel subset of  $\mathbb{R}_+^2$  for any  $(a_1, \dots, a_r) \in \mathbb{R}_+^r$ , which completes the proof.

In order to characterize the mapping  $\phi$  in terms of derivatives of the functions  $g_1(x, \cdot), \dots, g_r(x, \cdot)$ , it is customary to impose the *Inada conditions*, i.e.

(B3) for any  $x \geq 0$  the mappings  $g_1(x, \cdot)|]0, \infty[, \dots, g_r(x, \cdot)|]0, \infty[$  are assumed to be continuously differentiable satisfying

$$\lim_{\epsilon \rightarrow 0} \frac{\partial g^i(x, \cdot)}{\partial y} \Big|_{y=\epsilon} = \infty, \quad \lim_{\epsilon \rightarrow \infty} \frac{\partial g^i(x, \cdot)}{\partial y} \Big|_{y=\epsilon} = 0, \quad i = 1, \dots, r.$$

The Inada conditions together with condition (B1) imply that for any  $i \in \{1, \dots, r\}$  and every  $x \geq 0$  the mapping  $\frac{\partial g^i(x, \cdot)}{\partial y} |]0, \infty[$  is injective onto  $]0, \infty[$  with continuously differentiable, strictly decreasing inverse say  $I_i(x, \cdot)$ .

**Proposition B.3** *Let the assumptions (B0) - (B3) be fulfilled, and let  $g_1(x, \cdot)|]0, \infty[, \dots, g_r(x, \cdot)|]0, \infty[$  be twice continuously differentiable.*

*Then for any  $x \geq 0$  the mapping  $g(x, \cdot) |]0, \infty[$  is differentiable satisfying*

$$\phi(x, y) = \left[ I_1 \left\{ x, \frac{\partial g(x, \cdot)}{\partial y} \Big|_y \right\}, \dots, I_r \left\{ x, \frac{\partial g(x, \cdot)}{\partial y} \Big|_y \right\} \right] \text{ for } y > 0.$$

**Proof:**

Let for  $x \geq 0$  the mapping  $F_x : ]0, \infty[ \times ]0, \infty[ \rightarrow \mathbb{R}$  be defined by  $F_x(y, z) = \sum_{i=1}^r I_i(x, z) - y$ .

Since the mappings  $g_1(x, \cdot)|]0, \infty[, \dots, g_r(x, \cdot)|]0, \infty[$  are assumed to be strictly concave and twice continuously differentiable, their second derivatives are strictly negative. Then by local inverse theorem

the mappings  $I_1(x, \cdot), \dots, I_r(x, \cdot)$  are continuously differentiable, having strictly negative derivatives. In particular  $F_x$  is continuously differentiable, satisfying

$$\frac{\partial F_x}{\partial z} \Big|_{(y,z)} \neq 0 \text{ for } y, z > 0.$$

Furthermore, since  $I_1(x, \cdot), \dots, I_r(x, \cdot)$  are continuous and strictly decreasing mappings onto  $]0, \infty[$ , we may find for any  $y > 0$  a unique  $\varphi(y) > 0$  with  $F(y, \varphi(y)) = 0$ . Drawing on the implicit function theorem,  $y \mapsto \varphi(y)$  defines a differentiable mapping  $\varphi : ]0, \infty[ \rightarrow ]0, \infty[$ .

Moreover, for  $y > 0$  and  $y_1, \dots, y_r \geq 0$  with  $\sum_{i=1}^r y_i \leq y$ , we may conclude

$$\begin{aligned} \sum_{i=1}^r g_i(x, y_i) + \varphi(y)(y - \sum_{i=1}^r y_i) &= \varphi(y)y + \sum_{i=1}^r \{g_i(x, y_i) + \varphi(y)y_i\} \leq \\ &\varphi(y)y + \sum_{i=1}^r \sup_{z \geq 0} \{g_i(x, z) + \varphi(y)z\} = \\ &\varphi(y)y + \sum_{i=1}^r [g_i\{x, I_i(x, \varphi(y))\} + \varphi(y)I_i\{x, \varphi(y)\}] = \\ &\sum_{i=1}^r g_i[x, I_i\{x, \varphi(y)\}] - F_x\{y, \varphi(y)\} = \sum_{i=1}^r g_i[x, I_i\{x, \varphi(y)\}]. \end{aligned}$$

This means

$$g(x, y) = \sum_{i=1}^r g_i[x, I_i\{x, \varphi(y)\}],$$

and hence by Lemma B.1

$$(*) \quad \varphi(x, y) = (I_1[x, \varphi(y)], \dots, I_r[x, \varphi(y)]).$$

As a further consequence  $g(x, \cdot) | ]0, \infty[$  is differentiable satisfying

$$\frac{dg(x, \cdot)}{dy} \Big|_y = \sum_{i=1}^r \varphi(y) \frac{dI_i(x, \cdot) \circ \varphi}{dy} \Big|_y = \varphi(y) \frac{d\left(\sum_{i=1}^r I_i(x, \cdot) \circ \varphi\right)}{dy} \Big|_y = \varphi(y).$$

For the last equation notice that  $\sum_{i=1}^r I_i(x, \cdot) \circ \varphi$  is just the identity on  $]0, \infty[$ . In view of (\*) the proof is complete.

## References

- Ait-Sahalia, Y. and Lo, A. W. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94(1-2):9–51.
- Bakshi, G. and Madan, D. (2008). Investor heterogeneity and the non-monotonicity of the aggregate marginal rate of substitution in the market index. Working Paper, University of Maryland.
- Bakshi, G., Madan, D., and Panayotov, G. (2010). Returns of claims on the upside and the viability of u-shaped pricing kernels. *Journal of Financial Economics*, 97:130 – 154.
- Barberis, N., Huang, M., and Santos, T. (2001). Prospect theory and assets prices. *The Quarterly Journal of Economics*, 116:1 – 53.
- Barone-Adesi, G., Mancini, L., and Shefrin, H. M. (2013). A tale of two investors: Estimating risk aversion, optimism, and overconfidence. Swiss Finance Institute Research Paper No. 12-21. Available at SSRN: <http://ssrn.com/abstract=2060983>.
- Basak, S. and Pavlova, A. (2012). Asset prices and institutional investors. *American Economic Review*. Forthcoming.
- Beare, B. K. and Schmidt, L. D. W. (2012). An empirical test of pricing kernel monotonicity. University of California at San Diego, Economics Working Paper Series.
- Benzoni, L., Collin-Dufresne, P., and Goldstein, R. S. (2011). Can standard preferences explain the prices of out-of-the-money s&p 500 put options? Working Paper No. 2011-11, Federal Reserve Bank of Chicago.
- Bondarenko, O. (2003). Why are putoptions so expensive? Working paper. University of Illinois, Chicago, IL.
- Campbell, J. Y. and Cochrane, J. H. (1999). By force of habit: A consumption based explanation of aggregate stock market behaviorl. *The Journal of Political Economy*, 107:205–251.

- Chabi-Yo, F. (2012). Pricing kernels with stochastic skewness and volatility risk. *Management Science*, 58:624–640.
- Chabi-Yo, Y., Garcia, R., and Renault, E. (2008). State dependence can explain the risk aversion puzzle. *Review of Financial Studies*, 21:973–1011.
- Christoffersen, P., Heston, S., and Jacobs, K. (2012). Capturing option anomalies with a variance-dependent pricing kernel. *Review of Financial Studies*, *Revised and Resubmitted*.
- Constantinides, G. (1990). Habit formation: A resolution of the equity puzzle. *The Journal of Political Economy*, 98(3):519–543.
- Dana, R.-A. and Jeanblanc, M. (2003). *Financial Markets in Continuous Time*. Springer, Berlin.
- Duffie, D. (1996). *Dynamic Asset Pricing Theory, 2nd ed.* Princeton University Press.
- Engle, R. F. and Rosenberg, J. V. (2002). Empirical pricing kernels. *Journal of Financial Economics*, 64(3):341–372.
- Epstein, L. and Zin, S. (2001). The independence axiom and asset returns. *Journal of Empirical Finance*, 8:537–572.
- Estrella, A. and Hardouvelis, G. (1991). The term structure as a predictor of real economic activity. *Journal of Finance*, 46:555–576.
- Garcia, R., Luger, R., and Renault, E. (2003). Empirical assessment of an intertemporal option pricing model with latent variables. *Journal of Econometrics*, 116:49–83.
- Giacomini, E. and Härdle, W. (2008). Dynamic semiparametric factor models in pricing kernel estimation. In Dabo-Niang, S. and Ferraty, F., editors, *Functional and Operational Statistics, Contributions to Statistics*, pages 181–187. Springer Verlag.
- Golubev, Y., Härdle, W. K., and Timofeev, R. (2008). Testing monotonicity of pricing kernels. Working paper, Université de Provence, Humboldt University Berlin, 2008.

- Grith, M., Härdle, W. K., and Park, J. (2012). Shape invariant modeling of pricing kernels and risk aversion. *Journal of Financial Econometrics*. In press, doi: 10.1093/jfinec/nbs019.
- Härdle, W. K., Okhrin, Y., and Wang, W. (2012). Uniform confidence bands for pricing kernels. Revised and resubmitted to *Journal of Financial Econometrics*.
- Hens, T. and Reichlin, C. (2012). Three solutions to the pricing kernel puzzle. *Review of Finance*. <http://dx.doi.org/10.1093/rof/rfs008>.
- Heston, S. and Nandi, S. (2000). A closed-form garch option pricing model. *Review of Financial Studies*, 13:585 – 626.
- Hong, H. G., Jiang, W., and Zhao, B. (2012). Trading for status. Available at SSRN: <http://ssrn.com/abstract=1961833>.
- Jackwerth, J. (2000). Recovering risk aversion from option prices and realized returns. *Review of Financial Studies*, 13:433–451.
- Karatzas, I. and Shreve, S. E. (1998). *Methods of Mathematical Finance*. Springer, New York.
- Karni, E., Schmeidler, D., and Vind, K. (1983). On state dependent preferences and subjective probabilities. *Econometrica*, 51(4):1021–1031.
- Kramkov, D. and Schachermayer, W. (1999). The asymptotic elasticity of utility functions and optimal investment in incomplete markets. *The Annals of Applied Probability*, 9(3):904–950.
- Lopes, L. L. (1987). Between hope and fear: The psychology of risk. *Advances in Experimental Social Psychology*, 20:255–295.
- Lopes, L. L. and Oden, G. C. (1999). The role of aspiration level in risk choice: A comparison of cumulative prospect theory and sp/a theory. *Journal of Mathematical Psychology*, 43(2):286–313.
- Mas-Colell, A., Whinston, M. D., and Greene, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

- Melino, A. and Yang, A. X. (2003). State dependent preferences can explain the equity premium puzzle. *Review of Economic Dynamics*, 6(4):806–830.
- Polkovnichenko, V. and Zhao, F. (2012). Probability weighting functions implied by option prices. *Journal of Financial Economics*. In press, doi: 10.1016/j.jfineco.2012.09.008.
- Routledge, B. R. and Zin, S. E. (2010). Generalized disappointment aversion and asset prices. *Journal of Finance*, 65(4):1303–1332.
- Shefrin, H. (2008). *A behavioral approach to asset pricing*. Amsterdam ; Boston : Academic Press/Elsevier, 2nd edition.
- Song, Z. and Xiu, D. (2012). A tale of two option markets: Pricing kernels and volatility risk. Chicago Booth Research Paper No. 12-10; Fama-Miller Working Paper. Available at SSRN: <http://ssrn.com/abstract=2013381>.
- Tang, K. and Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68:54 – 74.
- Veronesi, P. (2004). Belief-dependent utilities, aversion to state-uncertainty, and asset prices. Working paper. University of Chicago.
- Ziegler, A. (2007). Why does implied risk aversion smile? *Review of Financial Studies*, 20(3):859 – 904.

# Statistical inference for generalized additive models: simultaneous confidence corridors and variable selection

Shuzhuan Zheng · Rong Liu · Lijian  
Yang · Wolfgang K.Härdle

Received: date

**Abstract** In spite of widespread use of generalized additive models (GAMs) to remedy the “curse of dimensionality”, there is no well-grounded methodology developed for simultaneous inference and variable selection for GAM in existing literature. However both are essential in enhancing the capability of statistical models. To this end, we establish simultaneous confidence corridors (SCCs) and a type of Bayesian information Criterion (BIC) through the spline-backfitted kernel smoothing techniques proposed in recent articles. To characterize the global features of each nonparametric components, SCCs are constructed for testing their overall trends and entire shapes. By extending the BIC in additive models with identity/trivial link, an asymptotically consistent BIC approach for variable selection is built up in GAM to improve the parsimony of model without loss of prediction accuracy. Simulations and a real example corroborate the above findings.

---

Shuzhuan Zheng  
Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou  
215006, China  
Department of Economics, Columbia University, New York, NY 10027, USA

Rong Liu  
Department of Mathematics and Statistics, University of Toledo, Toledo OH 43606, USA

Lijian Yang  
Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou  
215006, China  
E-mail: yanglijian@suda.edu.cn

Wolfgang K. Härdle  
C.A.S.E. – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin,  
Unter den Linden 6,10099 Berlin, Germany  
Lee Kong Chian School of Business, Sim Kee Boon Institute for Financial Economics, Sin-  
gapore Management University, Singapore



**Mathematics Subject Classification (2000)** 62G08 · 62G15 · 62G32

**Keywords** BIC · Confidence corridor · Extreme value · Generalized additive mode · Spline-backfitted kernel

## 1 Introduction

Generalized additive model (GAM) has gained popularity on addressing the curse of dimensionality in multivariate nonparametric regressions with non-Gaussian responses. GAM was developed by Hastie and Tibshirani (1990) for blending generalized linear model with nonparametric additive regression, which stipulates that a data set  $\{Y_i, \mathbf{X}_i^T\}_{i=1}^n$  consists of iid copies of  $\{Y, \mathbf{X}^T\}$  that satisfies

$$\begin{aligned} \mathbb{E}(Y|\mathbf{X}) &= b' \{m(\mathbf{X})\}, \text{var}(Y|\mathbf{X}) = a(\phi) b'' \{m(\mathbf{X})\}, \\ m(\mathbf{X}) &= c + \sum_{l=1}^d m_l(X_l), \end{aligned} \quad (1)$$

$$Y = b' \{m(\mathbf{X})\} + \sigma(\mathbf{X}) \varepsilon, \sigma(\mathbf{X}) = \{\text{var}(Y|\mathbf{X})\}^{1/2}$$

where the response  $Y$  is one of certain types, such as Bernoulli, Poisson and so forth, the vector  $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$  consists of the predictors,  $m_l(\cdot), 1 \leq l \leq d$  are unknown smooth functions, the white noise  $\varepsilon$  satisfies that  $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$  and  $\mathbb{E}(\varepsilon^2|\mathbf{X}) = 1$ , while  $c$  is an unknown constant,  $a(\phi)$  is a nuisance parameter that quantifies overdispersion, and the known inverse link function  $b'$  satisfies that  $b' \in C^2(\mathbb{R}), b''(\theta) > 0, \theta \in \mathbb{R}$ , see Assumption (A2) in the Appendix. In particular, if one takes the identity/trivial link, model (1) becomes a common additive model, see Huang and Yang (2004).

The joint density  $f(\mathbf{x})$  of  $(X_1, \dots, X_d)$  is assumed to be continuous and

$$0 < c_f \leq \inf_{\mathbf{x} \in [0,1]^d} f(\mathbf{x}) \leq \sup_{\mathbf{x} \in [0,1]^d} f(\mathbf{x}) \leq C_f < \infty,$$

see Assumption (A4) in the Appendix. Furthermore, for each  $1 \leq l \leq d$ , the marginal density function  $f_l(x_l)$  of  $X_l$  has continuous derivatives on  $[0, 1]$  and the same uniform bounds  $C_f$  and  $c_f$ . There exists a  $\sigma$ -finite measure  $\lambda$  on  $\mathbb{R}$  such that the distribution of  $Y_i$  conditional on  $X_i$  has a probability density function  $f_{Y|\mathbf{X}}(y; b' \{m(\mathbf{x})\})$  relative to  $\lambda$  whose support for  $y$  is a common  $\Omega$ , and is continuous in both  $y \in \Omega$  and  $x \in [0, 1]^d$ .

It is often the case that in model (1) the probability density function of  $Y_i$  conditional on  $\mathbf{X}_i$  with respect to a fixed  $\sigma$ -finite measure forms an exponential family:

$$f(Y_i|\mathbf{X}_i, \phi) = \exp[\{Y_i m(\mathbf{X}_i) - b \{m(\mathbf{X}_i)\}\} / a(\phi) + h(Y_i, \phi)]. \quad (2)$$

Nonetheless, such an assumption is not necessary in this paper. Instead, we only stipulate that the conditional variance and conditional mean are linked by

$$\text{var}(Y|\mathbf{X} = \mathbf{x}) = a(\phi) b'' \left[ (b')^{-1} \{\mathbb{E}(Y|\mathbf{X} = \mathbf{x})\} \right].$$

For identifiability, one needs

$$E\{m_l(X_l)\} = 0, 1 \leq l \leq d$$

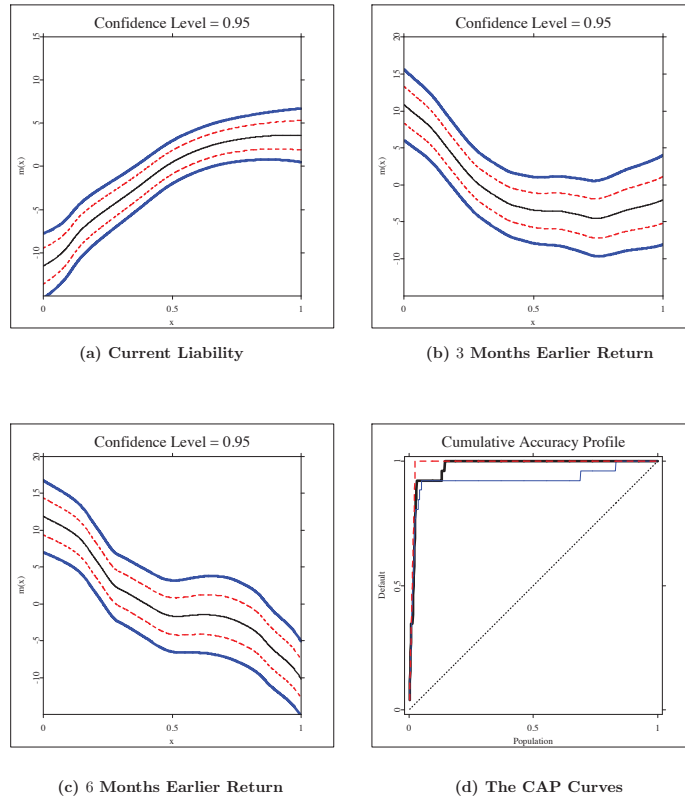
that leads to unique additive representations of  $m(\mathbf{x}) = c + \sum_{l=1}^d m_l(x_l)$ . Without loss of generality,  $\mathbf{x}$  take values in  $\chi = [0, 1]^d$ .

Model (1) has numerous applications. In corporate credit rating, for instance, one is interested in modelling how the default or non-default of a given corporate or company depends on the additive effects of the covariates in financial statements, i.e., the response  $Y = 0, 1$  with 1 indicating default, 0 indicating non-default, and the predictors are selected from financial statements with a logit-link  $(b')^{-1}(x) = \log\{x/(1-x)\}$ . Our method has been applied to 3,472 companies in Japan within a 5-year default horizon (2005-2010), and it has been discovered that the current liabilities and stock market returns of current, 3 months and 6 months prior to default are very significant as rating factors, and the default impact of the selected factors are examined via the simultaneous confidence corridors (SCCs) in Figure 1 (a)-(c). More details of this example are contained in Section 6.

The smooth functions  $\{m_l(x_l)\}_{l=1}^d$  in (1) can be estimated by, for instance, kernel methods in Linton and Härdle (1996), Linton (1997) and Yang, Sperlich and Härdle (2003), B-spline methods in Stone (1986) and Xue and Liang (2010), and two-stage methods in Horowitz and Mammen (2004). To make statistical inference on these functions individually and collectively, however, the proper tools are nonparametric simultaneous confidence corridors (SCCs) and consistent variable selection criteria, both of which are absent in the literature.

Nonparametric SCCs methodology has become increasingly important in statistical literature, see Xia (1998), Fan and Zhang (2000), Wu and Zhao (2007), Zhao and Wu (2008), Ma, Yang and Carroll (2012), Wang et al. (2014), Zheng, Yang and Härdle (2014), Gu et al. (2014), Cai and Yang (2015) and Gu and Yang (2015) for recent theoretical works on nonparametric SCCs. Capturing global shape properties by SCCs of the functions  $\{m_l(x_l)\}_{l=1}^d$  in GAM (1) is of prime importance. A nonparametric component can be replaced by a parametric one covered entirely within the SCCs, significantly decreasing the estimation variance, see He, Zhu and Fung (2002), He, Fung and Zhu (2005) for discussions. As far as we know, SCCs has not been established for functions  $\{m_l(x_l)\}_{l=1}^d$  in GAM (1) due to the lack of estimators that fit in Gaussian process extreme value theory. Using the spline-backfitted kernel (SBK) smoothing of Liu, Yang and Härdle (2013), we extend the SCCs works of univariate nonparametric regression in Bickel and Rosenblatt (1973) and Härdle (1989) to those of GAM. The SBK smoothing has been well developed in Wang and Yang (2007), Wang and Yang (2009), Liu and Yang (2010) and Ma and Yang (2011) for the much simpler additive model (i.e., GAM with  $b'(x) \equiv x$ ) including the construction of SCCs, but ours is the first work on SCCs on GAM with nonlinear link.

While variable selection for nonparametric additive model has been investigated under different settings, see Wang, Li and Huang (2008), there is



**Fig. 1** Plots of the rating factors in (a)-(c): SBK estimators (thin), 95% CIs (dashed) and 95% SCCs (thick). Plot of the CAPs defined as (24) in (d): Perfect (dashed), GAM (thick solid), GLM (thin solid), noninformative (dotted).

lack of theoretically-reliable variable selection approach for GAM. To the best of our knowledge, only Zhang and Lin (2006) proposed a sounding method named “COSSO”, which stands for components (CO) LASSO using penalized likelihood method, for selecting components in nonparametric regression with exponential families, but it leaves the asymptotic distributions and variable selection consistency to be desired. Instead, we tackle this issue by building a BIC type criterion based on spline pre-smoothing (first stage in the SBK), which is asymptotically consistent and easy to compute. Our work extends the BIC criterion for additive models (trivial link) in Huang and Yang (2004). Such an extension is challenging since a much more complicated quasi-likelihood is used in GAM with possibly nonlinear link instead of the log mean squared error for trivial link, see the Appendix for details.

The rest of paper is organized as follows. The SBK estimator and its oracle property are briefly described in Section 2. Asymptotic extreme value distribution of the SBK estimator is investigated in Section 3, which is used

to construct the SCCs of component functions. Section 4 introduces a BIC criterion in the GAM setting and provides results on consistent component selection as well as the implementation, followed by the Monte Carlo simulations in Section 5. Section 6 illustrates the application of our SCCs and BIC methods to predict default of nearly 3,500 listed companies in Japan. Technical assumptions and proofs are presented in the Appendix.

## 2 Spline-backfitted kernel smoothing in GAM

In this section we briefly describe the spline-backfitted kernel (SBK) estimator for GAM (1) and its oracle properties obtained in Liu, Yang and Härdle (2013). Let  $\{\mathbf{X}_i, Y_i\}_{i=1}^n$  be i.i.d. observations following model (1). Without loss of generality, one denotes  $\mathbf{x}_{\cdot 1} = (x_2, \dots, x_d)$  and  $m_{\cdot 1}(\mathbf{x}_{\cdot 1}) = c + \sum_{l=2}^d m_l(x_l)$  and estimates  $m_1(x_1)$ .

As a benchmark of efficiency, we introduce the ‘‘oracle smoother’’ by treating the constant  $c$  and the last  $d - 1$  components  $\{m_l(x_l)\}_{l=2}^d$  as known, then the only unknown component  $m_1(x_1)$  may be estimated by the following procedure. Although the exponential family Equation (2) does not necessarily hold, one still defines, as in Severini and Staniswalis (1994), for each  $x_1 \in [h, 1 - h]$  a local log-likelihood function  $\tilde{l}(a) = \tilde{l}(a, x_1)$  as

$$\tilde{l}(a, x_1) = n^{-1} \sum_{i=1}^n [Y_i \{a + m_{\cdot 1}(\mathbf{X}_{i, \cdot 1})\} - b \{a + m_{\cdot 1}(\mathbf{X}_{i, \cdot 1})\}] K_h(X_{i1} - x_1), \quad (3)$$

where  $a \in A$ , a set whose interior contains  $m_1([0, 1])$ . The oracle smoother of  $m_1(x_1)$  is

$$\tilde{m}_{K,1}(x_1) = \operatorname{argmax}_{a \in A} \tilde{l}(a, x_1).$$

Although  $\tilde{m}_{K,1}(x_1)$  is not a statistic since  $c$  and  $\{m_l(x_l)\}_{l=2}^d$  are actually unknown, its asymptotic properties serve as a benchmark for estimators of  $m_1(x_1)$  to achieve.

To define the SBK, we introduce the linear B spline basis for smoothing:  $b_J(x) = (1 - |x - \xi_J|/H)_+$ ,  $0 \leq J \leq N + 1$  where  $0 = \xi_0 < \xi_1 < \dots < \xi_N < \xi_{N+1} = 1$  are a sequence of equally spaced points, called interior knots, on interval  $[0, 1]$ . Denote by  $H = (N + 1)^{-1}$  the width of each subinterval  $[\xi_J, \xi_{J+1}]$ ,  $0 \leq J \leq N$  and the degenerate knots by  $\xi_{-1} = 0, \xi_{N+2} = 1$ . The space of  $l$ -empirically centered linear spline functions on  $[0, 1]$  is

$$G_{n,l}^0 = \left\{ g_l : g_l(x_l) \equiv \sum_{J=0}^{N+1} \lambda_{Jl} b_J(x_l), E_n \{g_l(X_l)\} = 0 \right\}, 1 \leq l \leq d, \quad (4)$$

with empirical expectation  $E_n \{g_l(X_l)\} = n^{-1} \sum_{i=1}^n g_l(X_{li})$ . The space of additive spline functions on  $\chi = [0, 1]^d$  is

$$G_n^0 = \left\{ g(\mathbf{x}) = c + \sum_{l=1}^d g_l(x_l); c \in \mathbb{R}, g_l \in G_{n,l}^0 \right\}.$$

The SBK method is defined in two steps. One first pre-estimates the unknown functions  $\{m_l(x_l)\}_{l=2}^d$  and constants  $c$  by linear spline smoothing. We define the log-likelihood function  $\widehat{L}(g)$  as

$$\widehat{L}(g) = n^{-1} \sum_{i=1}^n [Y_i g(\mathbf{X}_i) - b\{g(\mathbf{X}_i)\}], g \in G_n^0. \quad (5)$$

According to Lemma 14 of Stone (1986), (5) has a unique maximizer with probability approaching 1. Therefore, the multivariate function  $m(\mathbf{x})$  can be estimated by an additive spline function:

$$\widehat{m}(\mathbf{x}) = \operatorname{argmax}_{g \in G_n^0} \widehat{L}(g). \quad (6)$$

The spline estimator is asymptotically consistent, and can be solved efficiently via generalized linear models. However, as stated in Wang and Yang (2007) and Liu, Yang and Härdle (2013), spline methods only provide convergence rates but no asymptotic distributions, so no measures of confidence can be assigned to the estimators. To overcome this problem, we adapt the SBK estimator, which combines the strength of kernel smoothing with regression spline. One then rewrites  $\widehat{m}(\mathbf{x}) = \widehat{c} + \sum_{l=2}^d \widehat{m}_l(x_l)$  for  $\widehat{c} \in \mathbb{R}$  and  $\widehat{m}_l(x_l) \in G_{n,l}^0$  and defines a univariate quasi-likelihood function similar to  $\widetilde{l}(a, x_1)$  in (3) as

$$\widehat{l}(a, x_1) = n^{-1} \sum_{i=1}^n [Y_i \{a + \widehat{m}_{\cdot 1}(\mathbf{X}_{i,\cdot 1})\} - b\{a + \widehat{m}_{\cdot 1}(\mathbf{X}_{i,\cdot 1})\}] K_h(X_{i1} - x_1)$$

with  $\widehat{m}_{\cdot 1}(\mathbf{x}_{\cdot 1}) = \widehat{c} + \sum_{l=2}^d \widehat{m}_l(x_l)$  being the pilot spline estimator of  $m_{\cdot 1}(\mathbf{x}_{\cdot 1})$ . Consequently, the spline-backfitted kernel (SBK) estimator of  $m_1(x_1)$  is

$$\widehat{m}_{\text{SBK},1}(x_1) = \operatorname{argmax}_{a \in A} \widehat{l}(a, x_1). \quad (7)$$

We now introduce some useful results and definitions from Liu, Yang and Härdle (2013), under Assumptions (A1)-(A7) in appendix, as  $n \rightarrow \infty$ ,

$$\sup_{x_1 \in [0,1]} |\widehat{m}_{\text{SBK},1}(x_1) - \widetilde{m}_{\text{K},1}(x_1)| = \mathcal{O}_{a.s.}(n^{-1/2} \log n), \quad (8)$$

$$\begin{aligned} \widetilde{m}_{\text{K},1}(x_1) - m_1(x_1) &= \operatorname{bias}_1(x_1) h^2 / D_1(x_1) \\ &+ n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \sigma(\mathbf{X}_i) \varepsilon_i / D_1(x_1) + r_{\text{K},1}(x_1) \end{aligned} \quad (9)$$

in which the higher order remainder  $r_{\text{K},1}(x_1)$  satisfies

$$\sup_{x_1 \in [h, 1-h]} |r_{\text{K},1}(x_1)| = \mathcal{O}_{a.s.}(n^{-1/2} h^{1/2} \log n). \quad (10)$$

The scale function  $D_1(x_1)$  and bias function  $\operatorname{bias}_1(x_1)$  are defined in Liu, Yang and Härdle (2013) as:

$$\begin{aligned} \sigma_b^2(x_1) &= \mathbb{E}[b''\{m(\mathbf{X})\} | X_1 = x_1], \quad \sigma^2(x_1) = \mathbb{E}\{\sigma^2(\mathbf{X}) | X_1 = x_1\} \\ D_1(x_1) &= f_1(x_1) \sigma_b^2(x_1), \quad v_1^2(x_1) = \|K\|_2^2 f_1(x_1) \sigma^2(x_1). \end{aligned} \quad (11)$$

$$\begin{aligned} \text{bias}_1(x_1) &= \mu_2(K) \times \{m_1''(x_1) D_1(x_1) + m_1'(x_1) f(x_1) \sigma_b^2(x_1)' \\ &\quad - \{m_1'(x_1)\}^2 f(x_1) E[b''' \{m(\mathbf{X})\} | X_1 = x_1]\} \end{aligned}$$

where  $\|K\|_2^2 = \int K^2(u) du$ ,  $\mu_2(K) = \int K(u) u^2 du$ . The above equations (8), (9) and (10) lead one to a simplifying decomposition of the estimation error  $\widehat{m}_{\text{SBK},1}(x_1) - m_1(x_1)$

$$\begin{aligned} \sup_{x_1 \in [h, 1-h]} \left| \widehat{m}_{\text{SBK},1}(x_1) - m_1(x_1) - n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \sigma(\mathbf{X}_i) \varepsilon_i / D_1(x_1) \right| \\ = \mathcal{O}_{a.s.} \left( n^{-1/2} h^{1/2} \log n + n^{-1/2} \log n + h^2 \right). \end{aligned} \quad (12)$$

The decomposition in (12) is fundamental for constructing SCCs in section 3, and it follows from Theorems 1 and 4 of Liu, Yang and Härdle (2013), which were proved under weak dependence. A similar Theorem 2 in Horowitz and Mammen (2004) for the two-stage estimator was established only for a fixed  $x_1$ , not uniformly for  $x_1$  in the growing interval  $[h, 1-h]$ , and exclusively for iid data, not dependent data, see detailed discussion on page 621 of Liu, Yang and Härdle (2013).

### 3 GAM inference via simultaneous confidence corridor

In this section, we propose SCCs for GAM components based on the SBK smoothing, extending the works for univariate nonparametric function estimation in Bickel and Rosenblatt (1973) and Härdle (1989).

#### 3.1 Main Results

Denote  $a_h = \sqrt{-2 \log h}$ ,  $C(K) = \|K'\|_2^2 \|K\|_2^{-2}$  and for any  $\alpha \in (0, 1)$ , the quantile

$$Q_h(\alpha) = a_h + a_h^{-1} \left[ \log \left\{ \sqrt{C(K)} / (2\pi) \right\} - \log \left\{ -\log \sqrt{1-\alpha} \right\} \right]. \quad (13)$$

Also with  $D_1(x_1)$  and  $v_1^2(x_1)$  given in (11), we define

$$\sigma_n(x_1) = n^{-1/2} h^{-1/2} v_1(x_1) D_1^{-1}(x_1). \quad (14)$$

**Theorem 1** Under Assumptions (A1)-(A7), as  $n \rightarrow \infty$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \sup_{x_1 \in [h, 1-h]} |\widehat{m}_{\text{SBK},1}(x_1) - m_1(x_1)| / \sigma_n(x_1) \leq Q_h(\alpha) \right\} = 1 - \alpha.$$

A  $100(1-\alpha)\%$  simultaneous confidence corridor for  $m_1(x_1)$  is

$$\widehat{m}_{\text{SBK},1}(x_1) \pm \sigma_n(x_1) Q_h(\alpha). \quad (15)$$

The above SCC for component function  $m_1(x_1)$  resembles the SCCs in Bickel and Rosenblatt (1973) and Härdle (1989) for estimating unknown univariate nonparametric function, although it is for multivariate nonparametric regression.

### 3.2 Implementation

To satisfy Assumption (A4), one could use the transformed  $U_{il} = F_{nl}(X_{il})$  instead of  $X_{il}$  as predictors for each  $l = 1, \dots, d$  and  $i = 1, \dots, n$ , where  $F_{nl}$  is the empirical distribution of  $(X_{1l}, \dots, X_{nl})$ . We still use symbol  $X$  instead of  $U$  to avoid involving new symbols, but the  $X$  variates have been transformed in simulation study and applications.

To construct the SCC for  $m_1(x_1)$  in (15), one needs to select the bandwidth  $h$  and the number of knots  $N$  to evaluate  $m_{\text{SBK},1}(x_1)$ ,  $Q_h(\alpha)$  and  $\sigma_n(x_1)$  given in (7), (13) and (14).

Assumption (A6) requires that the bandwidth for SCCs be different from the mean square optimal bandwidth  $h_{\text{opt}} \sim n^{-1/5}$  (minimizing AMISE) in Liu, Yang and Härdle (2013). This is due to the two conflicting goals in SCCs construction: coverage of the true curve and narrowness of the corridor, are not quantifiable in a single measure to minimize, such as the mean integrated squared error. We therefore take  $h = h_{\text{opt}}(\log n)^{-1/4}$ , as a data-driven undersmoothing bandwidth for SCCs construction to fulfill Assumption (A6), where  $h_{\text{opt}}$  is computed as in Liu, Yang and Härdle (2013), page 623-624. Recent articles on SCCs for time series, such as Wu and Zhao (2007), Zhao and Wu (2008), have used similar undersmoothing bandwidths.

For a given  $l$  and a chosen bandwidth  $h$ , one can easily estimate  $m_{\text{SBK},1}(x_1)$  and  $Q_h(\alpha)$  as in (7), (13). To evaluate  $\sigma_n(x_1)$ , one needs to estimate  $v_1(x_1)$  and  $D_1^{-1}(x_1)$  given in (11), i.e., estimating  $f(x_1)$ ,  $\sigma_b^2(x_1)$  and  $\sigma^2(x_1)$ . The density function  $f(x_1)$  is estimated by  $\hat{f}(x_1) = n^{-1} \sum_{i=1}^n K_{h_{\text{ROT}}}(X_{i1} - x_1)$ , where  $h_{\text{ROT}}$  is the rule-of-thumb bandwidth in equation (5.8), page 200 of Fan and Yao (2003), namely  $h_{\text{ROT}} = (8\sqrt{\pi}/3)^{1/5} \mu_2(K) \|K\|_2^{2/5} n^{-1/5} \hat{\sigma}$ , in which  $\hat{\sigma}$  is the sample standard deviation of  $\{X_{i1}\}_{i=1}^n$ . We further illustrate the spline estimates of  $\sigma_b^2(x_1)$  and  $\sigma^2(x_1)$  below:

One partitions  $\min_i X_{i1} = t_{1,0} < \dots < t_{1,N+1} = \max_i X_{i1}$  where to satisfy Assumption (A7) in the Appendix, the number of spline interior knots equals

$$\max\left(1, \min\left(\left\lfloor n^{1/4} \log n + 1 \right\rfloor, \left\lfloor n/4d - 1/d \right\rfloor - 1\right)\right), \quad (16)$$

which ensures that the number of parameters in equation (6),  $1 + d(N + 2)$ , does not exceed  $n$ . An estimator for  $\sigma_b^2(x_1)$  is  $\sum_{k=0}^3 \hat{a}_{1,k} x_1^k + \sum_{k=4}^{N+3} \hat{a}_{1,k} (x_1 - t_{l,k-3})_+^3$  where  $\{\hat{a}_{1,k}\}_{k=0}^{N+3}$  minimize

$$\sum_{i=1}^n \left[ b'' \{\hat{m}(\mathbf{X}_i)\} - \left\{ \sum_{k=0}^3 a_{1,k} X_{i1}^k + \sum_{k=4}^{N+3} a_{1,k} (X_{i1} - t_{k-3})_+^3 \right\} \right]^2, \quad (17)$$

and  $\sigma^2(x_1)$  can be estimated as  $\sum_{k=0}^3 \hat{a}_{1,k} x_1^k + \sum_{k=4}^{N+3} \hat{a}_{1,k} (x_1 - t_{l,k-3})_+^3$  where  $\{\hat{a}_{1,k}\}_{k=0}^{N+3}$  minimize

$$\sum_{i=1}^n \left[ Y_i - b' \{\hat{m}(\mathbf{X}_i)\} \right]^2 - \left\{ \sum_{k=0}^3 a_{l,k} X_{i1}^k + \sum_{k=4}^{N+3} a_{l,k} (X_{i1} - t_{k-3})_+^3 \right\} \right]^2. \quad (18)$$

The resulted estimate  $\hat{\sigma}_n(x_1)$  of  $\sigma_n(x_1)$ , using (17) and (18) satisfies  $\sup_{x_1 \in [h, 1-h]} |\hat{\sigma}_n(x_1) - \sigma_n(x_1)| = \mathcal{O}_p(n^{-\gamma})$  for some  $\gamma > 0$ , see Liu, Yang and Härdle (2013) Section 5 for details. This consistency and Slutsky's theorem ensure that

$$P \left\{ \sup_{x_1 \in [h, 1-h]} |\widehat{m}_{\text{SBK},1}(x_1) - m_1(x_1)| / \hat{\sigma}_n(x_1) \leq Q_h(\alpha) \right\} \rightarrow 1 - \alpha$$

as  $n \rightarrow \infty$ , and therefore  $\widehat{m}_{\text{SBK},1}(x_1) \pm \hat{\sigma}_n(x_1) Q_h(\alpha)$  is a  $100(1 - \alpha)\%$  simultaneous confidence corridor for  $m_1(x_1)$ . The SCCs constructions of other components  $m_2(x_2), \dots, m_d(x_d)$  are similar. It is worth while to emphasize that, based on extensive simulation experiments, the estimators  $\widehat{m}_{\text{SBK},1}(x_1)$ ,  $\widehat{Q}_h(\alpha)$ ,  $\widehat{f}(x_1)$  and  $\hat{\sigma}_n(x_1)$  remain stable if  $h$  and  $N$  slightly vary.

#### 4 Variable selection in GAM

In this section, we propose a Bayesian Information Criterion (BIC) for component function selection based on spline smoothing in step one of the SBK estimation for GAM and an efficient implementation follows.

##### 4.1 Main Results

According to Stone (1985), p.693, the space of  $l$ -centered square integrable functions on  $[0, 1]$  is defined as

$$\mathcal{H}_l^0 = \{g : E\{g(X_l)\} = 0, E\{g^2\{X_l\}\} < \infty, 1 \leq l \leq d\}, \quad (19)$$

and the model space  $\mathcal{M}$  is

$$\mathcal{M} = \left\{ g(\mathbf{x}) = c + \sum_{l=1}^d g_l(\mathbf{x}_l); c \in \mathbb{R}, g_l \in \mathcal{H}_l^0, 1 \leq l \leq d \right\}. \quad (20)$$

To introduce the proposed BIC, let  $\{1, \dots, d\}$  denote the complete set of indices of  $d$  tuning variables  $(X_1, \dots, X_d)$ . For each subset  $S \subset \{1, \dots, d\}$ , define a corresponding model space  $\mathcal{M}_S$  for  $S$  as

$$\mathcal{M}_S = \left\{ g(\mathbf{x}) = c + \sum_{l \in S} g_l(\mathbf{x}_l); c \in \mathbb{R}, g_l \in \mathcal{H}_l^0, l \in S \right\},$$

with  $\mathcal{H}_l^0$  given in (19), and the space of the additive spline functions as

$$G_{n,S}^0 = \left\{ g(\mathbf{x}) = c + \sum_{l \in S} g_l(x_l); c \in \mathbb{R}, g_l \in G_{n,l}^0, l \in S \right\},$$

with  $G_{n,l}^0$  given in (4). Following Definition 1 of Huang and Yang (2004), the set  $S_0$  of significant variables is defined as the minimal set  $S \subset \{1, \dots, d\}$  such that  $m \in \mathcal{M}_S$ . According to Lemma 1 of Huang and Yang (2004), the set  $S_0$  is uniquely defined. Standard theory of Hilbert space and subspace



projection implies that the set  $S_0$  is also the minimal set  $S \subset \{1, \dots, d\}$  such that  $E\{m(\mathbf{X}) - m_S(\mathbf{X})\}^2 = 0$  in which the least squares projection of function  $m$  in  $\mathcal{M}_S$  is

$$m_S = \operatorname{argmin}_{g \in \mathcal{M}_S} E\{m(\mathbf{X}) - g(\mathbf{X})\}^2. \quad (21)$$

To identify  $S_0$ , one computes for an index set  $S$  the BIC as

$$\text{BIC}_S = -2\widehat{L}(\widehat{m}_S) + \frac{N_S}{n} (\log n)^3 \quad (22)$$

where  $\widehat{L}(\cdot)$  is given in (5),  $\widehat{m}_S(\mathbf{x}) \in G_{n,S}^0$  is the pilot spline estimator as in (6),  $N_S = 1 + (N+1)\#(S)$  with  $N$  the number of interior knots as defined in (16),  $\#(S)$  the cardinality of  $S$ .

Our variable selection rule takes the subset  $\widehat{S} \subset \{1, \dots, d\}$  that minimizes  $\text{BIC}_S$ .

**Theorem 2** *Under Assumptions (A1)-(A5), (A7),  $\lim_{n \rightarrow \infty} P(\widehat{S} = S_0) = 1$ .*

According to Theorem 2, the variable selection rule based on the BIC in (22) is consistent. The nonparametric version BIC was firstly established in Huang and Yang (2004) for additive autoregression model, and adapted to additive coefficient model by Xue and Yang (2006), to single index model by Wang and Yang (2009). Our proposed BIC differs from all of the above as it is based on quasi-likelihood rather than mean squared error, which makes the technical proof of consistency much more challenging. To the best of our knowledge, it is the first theoretically reliable information criterion in this setting.

## 4.2 Implementation

We have not implemented the BIC variable selection by a greedy search through all possible subsets. Instead, a forward stepwise procedure is used with minimizing BIC as the criterion since it is more common that only a few variables are significant among many variables. We have also experimented with backward as well as forward-backward stepwise procedures which have yielded similar outcomes in simulation examples.

## 5 Simulation

This section studies under simulated setting the performance of the proposed procedures including the computational cost of the SBK, the consistency of selecting variables via BIC and the coverage frequency of the SCCs. The data are generated from

$$P(Y = 1 | \mathbf{X} = \mathbf{x}) = b' \left\{ c + \sum_{l=1}^d m_l(X_l) \right\}, b'(x) = \frac{e^x}{1 + e^x} \quad (23)$$

$d$	$r$	$n$	Computing Time			Accuracy					
			BIC	COSSO	ratio	BIC			COSSO		
10	0	250	0.17	1.85	10.9	25	441	34	98	327	75
		500	0.41	4.33	10.6	6	476	28	42	414	44
		1000	0.66	20.14	30.5	2	491	7	26	455	19
	0.5	250	0.18	1.91	10.6	165	298	37	204	221	75
		500	0.42	4.43	10.5	11	452	37	89	359	52
		1000	0.67	20.64	30.8	1	493	6	67	401	32
50	0	250	1.00	—	—	312	78	110	—	—	—
		500	1.43	59.77	41.8	106	327	67	124	207	169
		1000	3.32	268.24	80.8	2	465	33	20	426	54
	0.5	250	1.04	—	—	319	65	116	—	—	—
		500	1.55	60.87	39.2	297	174	29	203	145	152
		1000	3.48	274.25	78.8	47	428	25	52	356	92

**Table 1** Simulation comparison of the proposed BIC method and COSSO with  $d = 10, 50$ . Computing Time is in seconds and the ratio is the computing time of COSSO over that of BIC. For  $d = 50$  and  $n = 250$ , COSSO becomes unstable to the point of crashing. Accuracy (the last 6 columns) gives for BIC and COSSO the numbers of underfitting, correct fitting, and overfitting out of 500 replications.

with  $d = 10, c = 0, m_3(x) = \sin(4\pi x), m_4(x) = m_5(x) = \sin(\pi x), m_6(x) = x, m_7(x) = e^x - (e - e^{-1})$  and  $m_l(x) = 0$  for  $l = 1, 2, 8, 9, 10$ . The predictors are generated by

$$X_{il} = 2\Phi(Z_{il}) - 1, \mathbf{Z}_i = (Z_{i1}, \dots, Z_{id}) \sim N(0, \Sigma), 1 \leq i \leq n, 1 \leq l \leq d,$$

where  $\Phi$  is the standard normal c.d.f. and  $\Sigma = (1 - r)\mathbf{I}_{d \times d} + r\mathbf{1}_d\mathbf{1}_d^T$ . The parameter  $r$  ( $0 \leq r < 1$ ) controls the correlation between  $Z_{il}, 1 \leq l \leq d$ . To examine the computing advantage of BIC for large  $d$ , we have also included results for  $d = 50$  with  $m_3, \dots, m_7$  as above and all the other component functions are 0.

COSSO is a penalized likelihood method proposed in Zhang and Lin (2006) for LASSO type component selection and nonparametric regression in exponential families. In what follows, the performance of BIC and COSSO is firstly compared, followed by a computational comparison between the SBK and a kernel method in GAM, and it ends with a report on the SCCs coverage frequency for components function (the frequency that SCCs covering the entire curve on the domain). We have tried numbers of knots different from the one in (16) with similar results, so our conclusion is that the performance of BIC is rather insensitive to the number of knots.

Table 1 shows the simulation results from 500 replications, where the outcome is defined in accuracy as correct fitting, if  $\hat{S} = S_0$ ; overfitting, if  $S_0 \subset \hat{S}$ ; and underfitting, if  $S_0 \not\subset \hat{S}$ . It is clear that the performance of BIC on selecting 5 significant variables  $m_l(X_l), l = 3, \dots, 7$ , is quite satisfactory. The selection accuracy becomes higher as the sample size increases and/or the correlation decreases; it is poorer with higher dimension  $d (= 50)$  but still high when sample size  $n = 1000$ . The accuracy and computing time of COSSO are also listed

$r$	$n$	$l$						
		1	2	3	4	5	6	7
0.0	250	0.9305	0.9250	0.9235	0.9250	0.9235	0.9240	0.9230
	500	0.9455	0.9475	0.9430	0.9405	0.9425	0.9440	0.9530
	1000	0.9515	0.9520	0.9475	0.9455	0.9480	0.9510	0.9485
0.5	250	0.9215	0.9185	0.9120	0.9145	0.9205	0.9210	0.9185
	500	0.9420	0.9405	0.9330	0.9325	0.9375	0.9385	0.9415
	1000	0.9485	0.9505	0.9420	0.9475	0.9455	0.9430	0.9445

**Table 2** The 95% SCCs coverage frequency for  $m_l(x)$ ,  $l = 1, 2, \dots, 7$  from 2000 replications

for comparison (Platform: R; PC: Intel 3.1 GHz processor and 8 GB RAM). It is shown in Table 1 that the BIC significantly outperforms the COSSO in terms of accuracy and computing time, and the advantage in computing time widens significantly for  $d = 50$ .

In addition to the above comparison for model selection, we have also conducted numerical comparison between COSSO and our proposed SBK estimation method in terms of probability prediction. The proposed SBK method has higher prediction accuracy in almost all cases, see Table 4 in the Supplement. Comparison regarding SCC has not been made against COSSO because it does not produce one.

The SCCs coverage frequency for  $m_l(x_l)$ ,  $l = 1, \dots, 7$  is reported in Table 2. Among the zero functions, we have omitted the results for  $m_8, m_9$  and  $m_{10}$  because the results are very similar to  $m_1$  and  $m_2$ . The empirical coverage approaches the nominal confidence levels as  $n$  increases, and better coverage occurs when the correlation is lower. The coverage frequencies vary slightly when  $d$  increases, the numerical results of which have not been included for brevity. We have also compared the coverage frequency of SCC and method VOT (Volume of Tube) in the same setup of the simulation 1 in Wiesenfarth et al. (2012), which considered only the case of trivial link function. The performance of our proposed SCC is quite similar to the VOT method Wiesenfarth et al. (2012), see Table 3 in the Supplement.

The above studies evidently indicate the reliability of our methodology, such as high selection accuracy of the BIC and desired coverage frequency of the SCCs. It ensures their applications for credit rating modelling in the following section.

## 6 Application

We now return to forecast default probabilities of the listed companies in Japan. The data taken from the Risk Management Institute, National University of Singapore include the comprehensive financial statements and the credit events (default or bankruptcy) from 2005 to 2010 of 3583 Japanese firms.

Berg (2007) found that the liability status was important to indicate the creditworthiness of a company, while Bernhardsen (2001) and Ryser and Denzler (2009) proposed to consider the “leverage effect” expressed by the financial

statement ratios. Therefore, we have pooled two situations by considering  $X_1$ : Current liability,  $X_2$ : Current stock return,  $X_3$ : Long term borrow,  $X_4$ : Short term borrow,  $X_5$ : Total asset,  $X_6$ : Non-current liability,  $X_7$ : 3 months earlier (stock) return,  $X_8$ : 6 months earlier (stock) return,  $X_9$ : Current ratio,  $X_{10}$ : Net liability to shareholder equity,  $X_{11}$ : Shareholder equity to total liability and equity,  $X_{12}$ : TCE ratio,  $X_{13}$ : Total debt to total asset,  $X_{14}$ : Quick ratio.

Selecting the rating factors via the BIC given in (22), we have found that  $X_1$ : Current liabilities,  $X_7$ : 3 months earlier return,  $X_8$ : 6 months earlier return are significant. Similar rating covariates were also discovered in China and Moore (2003), Berg (2007) and Ryser and Denzler (2009). However, Berg (2007) selected 23 variables which led to a non-parsimonious GAM. In contrast, Ryser and Denzler (2009) had found that 3 financial ratios (capital turnover, long-term debt ratio, return on total capital) were significant based on the blockwise cross-validation (CV) method which is nonetheless extremely time consuming in comparison to the proposed BIC.

Figure 1 (a)-(c) depicts the SBK estimator of the factor's default impact curve on domain, while a shoal of 95% CIs and the 95% SCCs present respectively the pointwise and global uncertainty of the whole curve. The SBK estimators indicate overall monotonicities of each rating factors, and the SCCs turn out to be fairly narrow to warrant the global nonlinearities of the factors' curves which reveal the underlying nonlinear features in different segments of domain.

As for the model evaluations, the Cumulative Accuracy Profile (CAP) is plotted in Figure 1 (d). For any score function  $S$ , one defines its alarm rate  $F(s) = P(S \leq s)$  and the hit rate  $F_D(s) = P(S \leq s | D)$  where  $D$  represents the conditioning event of "default". One then defines the CAP curve as

$$\text{CAP}(u) = F_D(F^{-1}(u)), u \in (0, 1), \quad (24)$$

which is the percentage of default-infected obligators that are found among the first (according to their scores) 100u% of all obligators. A satisfactory model's CAP would be expected to approach to that of the perfect model (i.e.,  $\text{CAP}_P(u) = \min(u/p, 1)$ ,  $u \in (0, 1)$  where  $p$  is the unconditional default probability) and always better than the noninformative. In contrast, a noninformative rating method with zero discriminatory power displays a diagonal line  $\text{CAP}_N(u) \equiv u$ ,  $u \in (0, 1)$ . See details of the CAP in Engelmann, Hayden and Tasche (2003).

The AR is the ratio of two areas  $a_R$  and  $a_P$ . The area between the given CAP curve and the noninformative diagonal  $\text{CAP}_N(u) \equiv u$  is  $a_R$ , whereas  $a_P$  is the area between the perfect CAP curve  $\text{CAP}_P(u)$  and the noninformative diagonal  $\text{CAP}_N(u)$ . Thus

$$\text{AR} = \frac{a_R}{a_P} = \frac{2 \int_0^1 \text{CAP}(u) du - 1}{1 - p}, \quad (25)$$

where  $\text{CAP}(u)$  is given in (24). The AR takes value in  $[0, 1]$ , with value 0 corresponding to the noninformative scoring, and 1 the perfect scoring method, a higher AR indicates an overall higher discriminatory power of a method.

Using both GAM and GLM obtained from first 2000 companies to predict the default rate of the rest 1583 companies, the accuracy ratio is 97.56% for GAM, much higher than the 89.76% for GLM. We have also applied the COSO method to the same data, and the following error message has appeared “Error in solve.QP(GH\$H, GH\$H %\*% old.theta - GH\$G, t(Amat), bvec): matrix D in quadratic function is not positive definite!”, which once again has illustrated the advantage of the proposed BIC procedure over the existing method.

**Acknowledgements** This work is supported in part by the Jiangsu Specially-Appointed Professor Program SR10700111, the Jiangsu Key Discipline Program (Statistics) ZY107002, ZY107992 National Natural Science Foundation of China award 11371272, Research Fund for the Doctoral Program of Higher Education of China award 20133201110002, United States NSF awards DMS 0706518, DMS 1007594, an Michigan State University Dissertation Continuation Fellowship, funding from the National University of Singapore, the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 “Economic Risk”, and the International Research Training Group (IRTG) 1792. The helpful comments from two Reviewers and an Associate Editor are gratefully acknowledged.

## Appendix

In what follows, we take  $\|\cdot\|$  and  $\|\cdot\|_\infty$  as the Euclidean and supremum norms, respectively, i.e., for any  $\mathbf{x} = (x_1, x_2, \dots, x_d)^T \in \mathbb{R}^d$ ,  $\|\mathbf{x}\| = \left(\sum_{l=1}^d x_l^2\right)^{1/2}$  and  $\|\mathbf{x}\|_\infty = \max_{1 \leq l \leq d} |x_l|$ . For any interval  $[a, b]$ , denote the space of  $p$ -th order smooth function by  $C^{(p)}[a, b] = \{g \mid g^{(p)} \in C[a, b]\}$ , and the class of Lipschitz continuous functions by

$$\text{Lip}([a, b], C) = \{g \mid |g(x) - g(x')| \leq C|x - x'|, \forall x, x' \in [a, b]\}$$

for constant  $C > 0$ . Lastly, define the following latent regression errors

$$\xi_i = Y_i - b' \{m(\mathbf{X}_i)\} = \sigma(\mathbf{X}_i) \varepsilon_i, 1 \leq i \leq n. \quad (26)$$

### A.1 Technical assumptions

We need the following technical assumptions:

- (A1) The additive component functions  $m_l \in C^{(1)}[0, 1]$ ,  $1 \leq l \leq d$ :  $m_1 \in C^{(2)}[0, 1]$ ,  $m'_l \in \text{Lip}([0, 1], C_m)$ ,  $2 \leq l \leq d$  for some constant  $C_m > 0$ .
- (A2) The inverse link function  $b'$  satisfies that  $b' \in C^2(\mathbb{R})$ ,  $b''(\theta) > 0$ ,  $\theta \in \mathbb{R}$ . For a compact interval  $\Theta$  whose interior contains  $m([0, 1]^d)$ ,  $C_b > \max_{\theta \in \Theta} b''(\theta) \geq \min_{\theta \in \Theta} b''(\theta) > c_b$  for constants  $0 < c_b < C_b < \infty$ .
- (A3) The conditional variance function  $\sigma^2(\mathbf{x})$  is continuous and positive for  $\mathbf{x} \in [0, 1]^d$ . The errors  $\{\varepsilon_i\}_{i=1}^n$  satisfy that  $E(\varepsilon_i \mid \mathbf{X}_i) = 0$ ,  $E(|\varepsilon_i|^{2+\eta}) \leq C_\eta$  for some  $\eta \in (1/2, 1]$ .

(A4) The joint density  $f(\mathbf{x})$  of  $(X_1, \dots, X_d)$  is continuous and

$$0 < c_f \leq \inf_{\mathbf{x} \in [0,1]^d} f(\mathbf{x}) \leq \sup_{\mathbf{x} \in [0,1]^d} f(\mathbf{x}) \leq C_f < \infty.$$

For each  $1 \leq l \leq d$ , the marginal density function  $f_l(x_l)$  of  $X_l$  has continuous derivatives on  $[0, 1]$  and the same uniform bounds  $C_f$  and  $c_f$ . There exists a  $\sigma$ -finite measure  $\lambda$  on  $\mathbb{R}$  such that the distribution of  $Y_i$  conditional on  $\mathbf{X}_i$  has a probability density function  $f_{Y|\mathbf{X}}(y; b' \{m(\mathbf{x})\})$  relative to  $\lambda$  whose support for  $y$  is a common  $\Omega$ , and is continuous in both  $y \in \Omega$  and  $\mathbf{x} \in [0, 1]^d$ .

(A5)  $\{\mathbf{Z}_i = (\mathbf{X}_i^T, \varepsilon_i)\}_{i=1}^n$  are independent and identically distributed.

(A6) The kernel function  $K(x)$  is a symmetric probability density function supported on  $[-1, 1]$  and  $\in C^1[-1, 1]$ . The bandwidth  $h = h_n$  satisfies that  $h = o(n^{-1/5}(\log n)^{-1/5})$ ,  $h^{-1} = \mathcal{O}(n^{1/5}(\log n)^\delta)$  for some constant  $\delta > 1/5$ .

(A7) The number of interior knots  $N$  satisfies  $c_N n^{1/4} \log n \leq N \leq C_N n^{1/4} \log n$  for some constants  $c_N, C_N > 0$ .

Assumptions (A1)-(A7) are standard in GAM, see Stone (1986), Xue and Yang (2006). The i.i.d. feature is technically acceptable if the data are collected across a large number of sections, for instance, our real example in Section 6. Assumptions (A5), (A6) are more restrictive than in Liu, Yang and Härdle (2013) for the purpose of constructing simultaneous confidence corridor, but are unnecessary for Theorem 2 on the consistency of BIC. All these assumptions are satisfied by the simulation example in Section 5.

## A.2 Preliminaries

Throughout this section,  $C$  denotes some generic positive constant unless stated otherwise. Define

$$M_h(t) = h^{-1/2} \int_0^1 K\{(x-t)/h\} dW(x) \quad (27)$$

where  $W(x)$  is a Wiener process defined on  $(0, \infty)$  and denote

$$d_h = (-2 \log h)^{1/2} + (-2 \log h)^{-1/2} \left\{ \sqrt{C(K)} / (2\pi) \right\}$$

with  $C(K)$  given in (13).

**Lemma 1** Under Assumption (A6). for any  $x \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[ (-2 \log h)^{1/2} \left\{ \sup_{t \in [h, 1-h]} |M_h(t)| / \|K\|_2^2 - d_h \right\} < x \right] = e^{-2e^{-x}}.$$

*Proof* One simply applies the same steps in proving Lemma 2.2 of Härdle (1989).

Denote by  $T_i$  the random variable  $b' \{m(\mathbf{X}_i)\}$ , and the Lebesgue measure on  $\mathbb{R}^d$  as  $\mu^{(d)}$ . By Assumption (A4),  $\mathbf{X}_i$  has pdf wrt the Lebesgue measure

$\mu^{(d)}$ , and Assumptions (A1) and (A2) ensure that functions  $b'$  and  $m$  are at least  $C^1$ , thus the random vector  $(T_i, X_{i1})$  has a joint pdf wrt the Lebesgue measure  $\mu^{(2)}$ , which one denotes as  $f_{T, X_1}(t, x_1)$ .

**Lemma 2** *Under Assumptions (A1)-(A5), for  $\xi_i$  in (26), the distribution of  $(\xi_i, X_{i1})$  has joint pdf wrt  $\mu^{(2)}$  as*

$$f_{\xi, X_1}(z, x_1) = \int_{\Omega} f_{Y|\mathbf{X}}(y; y-z) f_{T, X_1}(y-z, x_1) d\lambda(y).$$

*Proof* The joint pdf of  $(Y_i, T_i, X_{i1})$  wrt  $\lambda \times \mu^{(2)}$  is  $f_{Y|\mathbf{X}}(y; t) f_{T, X_1}(t, x_1)$ . For any  $(z, x_1) \in \mathbb{R} \times [0, 1]$ , and  $\Delta z, \Delta x_1 > 0$ , one has

$$\begin{aligned} & P[(\xi_i, X_{i1}) \in (z - \Delta z, z + \Delta z) \times (x_1 - \Delta x_1, x_1 + \Delta x_1)] = \\ & P[(Y_i - T_i, X_{i1}) \in (z - \Delta z, z + \Delta z) \times (x_1 - \Delta x_1, x_1 + \Delta x_1)] = \\ & \int_{\Omega} d\lambda(y) \int_{y-\tau \in (z-\Delta z, z+\Delta z)} d\tau \int_{\chi_1 \in (x_1-\Delta x_1, x_1+\Delta x_1)} f_{Y|\mathbf{X}}(y; \tau) f_{T, X_1}(\tau, \chi_1) d\chi_1. \end{aligned}$$

Applying dominated convergence theorem, one has as  $\max(\Delta z, \Delta x_1) \rightarrow 0$ ,

$$\begin{aligned} & P[(\xi_i, X_{i1}) \in (z - \Delta z, z + \Delta z) \times (x_1 - \Delta x_1, x_1 + \Delta x_1)] \\ & = \left\{ \int_{\Omega} f_{Y|\mathbf{X}}(y; y-z) f_{T, X_1}(y-z, x_1) d\lambda(y) \right\} \times \\ & \mu^{(2)}[(z - \Delta z, z + \Delta z) \times \{(x_1 - \Delta x_1, x_1 + \Delta x_1) \cap [0, 1]\}] + o(\Delta z \Delta x_1) \end{aligned}$$

hence the joint pdf of  $(\xi_i, X_{i1})$  wrt  $\mu^{(2)}$  is  $\int_{\Omega} f_{Y|\mathbf{X}}(y; y-z) f_{T, X_1}(y-z, x_1) d\lambda(y)$ .

For theoretical analysis, we write  $c_{J,l} = \mathbb{E} b_J(X_l) = \int b_J(x_l) f_l(x_l) dx_l$  and define the centered B spline basis  $b_{J,l}(x_l)$  and the standardized B spline basis  $B_{J,l}(x_l)$  respectively as

$$\begin{aligned} b_{J,l}(x_l) &= b_J(x_l) - \frac{c_{J,l}}{c_{J-1,l}} b_{J-1}(x_l), \\ B_{J,l}(x_l) &= \frac{b_{J,l}(x_l)}{\left\{ \int b_{J,l}^2(x_l) f_l(x_l) dx_l \right\}^{1/2}}, 1 \leq J \leq N+1, \end{aligned} \quad (28)$$

so that  $\mathbb{E} B_{J,l}(X_l) \equiv 0$ ,  $\mathbb{E} B_{J,l}^2(X_l) \equiv 1$ .

With slight abuse of notations the log-likelihood  $\widehat{L}(g)$  in (5) is

$$\widehat{L}(g) = \widehat{L}(\boldsymbol{\lambda}) = n^{-1} \sum_{i=1}^n \left[ Y_i \boldsymbol{\lambda}^T \mathbf{B}(\mathbf{X}_i) - b \left\{ \boldsymbol{\lambda}^T \mathbf{B}(\mathbf{X}_i) \right\} \right],$$

with  $g(\mathbf{X}_i) = \boldsymbol{\lambda}^T \mathbf{B}(\mathbf{X}_i) \in G_n^0$ ,  $\boldsymbol{\lambda} = (\lambda_0, \lambda_{J,l})_{1 \leq J \leq N+1, 1 \leq l \leq d}^T \in \mathbb{R}^{N_d}$  with  $N_d = (N+1)d + 1$ ,  $\mathbf{B}(\mathbf{x}) = \{1, B_{1,1}(x_1), \dots, B_{N+1,d}(x_d)\}^T$  and  $B_{J,l}(x_l)$  as given in (28). It is straightforward to verify that the gradient and Hessian of  $\widehat{L}(\boldsymbol{\lambda})$  are

$$\begin{aligned} \nabla \widehat{L}(\boldsymbol{\lambda}) &= n^{-1} \sum_{i=1}^n \left[ Y_i \mathbf{B}(\mathbf{X}_i) - b' \left\{ \boldsymbol{\lambda}^T \mathbf{B}(\mathbf{X}_i) \right\} \mathbf{B}(\mathbf{X}_i) \right], \\ \nabla^2 \widehat{L}(\boldsymbol{\lambda}) &= -n^{-1} \sum_{i=1}^n b'' \left\{ \boldsymbol{\lambda}^T \mathbf{B}(\mathbf{X}_i) \right\} \mathbf{B}(\mathbf{X}_i) \mathbf{B}(\mathbf{X}_i)^T. \end{aligned} \quad (29)$$

**Proposition 1** Under Assumptions (A1)-(A5) and (A7), for  $m \in M$  with  $M$  given in (20) and  $\hat{m}$  as in (6), as  $n \rightarrow \infty$ ,  $\|m - \hat{m}\|_{2,n} + \|m - \hat{m}\|_2 = \mathcal{O}_{a.s.}(N^{1/2}n^{-1/2} \log n)$  and  $\|m - \hat{m}\|_\infty = \mathcal{O}_{a.s.}(Nn^{-1/2} \log n)$ . With probability approaching 1, the Hessian matrix  $\nabla^2 \hat{L}(\boldsymbol{\lambda})$  satisfies that  $\nabla^2 \hat{L}(\boldsymbol{\lambda}) < \mathbf{0}, \forall \boldsymbol{\lambda}$  and  $\nabla^2 \hat{L}(\boldsymbol{\lambda}) \leq -c_b c_V \mathbf{I}$  if  $\boldsymbol{\lambda}^T \mathbf{B}(\mathbf{X}_i) \in \Theta, 1 \leq i \leq n$ .

*Proof* See Lemma A.13 of Liu, Yang and Härdle (2013), Assumption (A2), equation (29) and Lemma A.11 of Liu, Yang and Härdle (2013).

### A.3 Proof of Theorem 1

Define a stochastic process  $\hat{\varepsilon}_n(x_1) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \xi_i, x_1 \in [0, 1]$  with  $\xi_i$  given in (26), then (9) and (10) show that

$$\sup_{x_1 \in [h, 1-h]} |\tilde{m}_{K,1}(x_1) - m_1(x_1) - D_1^{-1}(x_1) \hat{\varepsilon}_n(x_1)| = \mathcal{O}_{a.s.}(h^2 + n^{-1/2} h^{1/2} \log n),$$

which, together with (8), lead to

$$\begin{aligned} & \sup_{x_1 \in [h, 1-h]} |\hat{m}_{\text{SBK},1}(x_1) - m_1(x_1) - D_1^{-1}(x_1) \hat{\varepsilon}_n(x_1)| \\ &= \mathcal{O}_{a.s.}(h^2 + n^{-1/2} h^{1/2} \log n + n^{-1/2} \log n) = \mathcal{O}_{a.s.}(h^2 + n^{-1/2} \log n). \end{aligned} \quad (30)$$

Using  $v_1(x_1)$  given in (11), one can standardize  $\hat{\varepsilon}_n(x_1)$  to obtain

$$\begin{aligned} \hat{\zeta}_n(x_1) &= (nh)^{1/2} v_1^{-1}(x_1) \hat{\varepsilon}_n(x_1) \\ &= (nh)^{1/2} v_1^{-1}(x_1) \left\{ n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \xi_i \right\}. \end{aligned} \quad (31)$$

Assumptions (A5), (A8) imply that the following Rosenblatt transformation to the 2-dimensional sequence  $\{X_{i1}, \xi_i\}_{i=1}^n$  produces  $\{X'_{i1}, \xi'_i\}_{i=1}^n$  with  $(X'_{i1}, \xi'_i)$  uniformly distributed on  $[0, 1]^2$ :

$$(X'_{i1}, \xi'_i) = T(X_{i1}, \xi_i) = \{F_{X_1}(X_{i1}), F_{\xi|X_1}(\xi_i|X_{i1})\}.$$

Denote  $Z_n(x_1, \xi) = \sqrt{n} \{F_n(x_1, \xi) - F(x_1, \xi)\}$  where  $F_n(x_1, \xi)$  is the empirical distribution of  $\{X_{i1}, \xi_i\}_{i=1}^n$ , one can rewrite  $\hat{\zeta}_n(x_1)$  as

$$\hat{\zeta}_n(x_1) = h^{-1/2} v_1^{-1}(x_1) \int \int K\{(u - x_1)/h\} \xi dZ_n(u, \xi).$$

By the strong approximation theorem in Tusnady (1977), there exists a version of the 2-dimensional Brownian Bridge  $B_n(x'_1, \xi')$  such that

$$\sup_{x_1, \xi} |Z_n(x_1, \xi) - B_n\{T(x_1, \xi)\}| = \mathcal{O}_{a.s.}(n^{-1/2} \log^2 n).$$



Applying standard techniques used in Bickel and Rosenblatt (1973), Härdle (1989), one can show that

$$\sup_{t \in [h, 1-h]} \left| \widehat{\zeta}_n(t) - M_h(t) / \|K\|_2^2 \right| = o_p \left\{ (\log n)^{-1/2} \right\}, \quad (32)$$

for a version of the  $M_h(t)$  given in (27). Similar result can be found in Xia (1998). Furthermore, (30) and (31) imply that

$$\begin{aligned} & \sup_{x_1 \in [h, 1-h]} \left| \sigma_n^{-1}(x_1) \{ \widehat{m}_{\text{SBK},1}(x_1) - m_1(x_1) \} - \widehat{\zeta}_n(x_1) \right| \\ &= \mathcal{O}_{a.s.} \left( n^{1/2} h^{5/2} + h^{1/2} \log n \right), \end{aligned} \quad (33)$$

with  $\sigma_n(x)$  given in (14). Under Assumption (A6), which entails that  $(-2 \log h)^{1/2}$  is of the same order as  $(\log n)^{1/2}$ , (32) and (33) can show that

$$\begin{aligned} & \sup_{x_1 \in [h, 1-h]} (-2 \log h)^{1/2} \left| \sigma_n^{-1}(x_1) | \widehat{m}_{\text{SBK},1}(x_1) - m_1(x_1) | - |M_h(x_1)| / \|K\|_2^2 \right| \\ &= \mathcal{O}_{a.s.} \left\{ (\log n)^{1/2} \times \left( n^{1/2} h^{5/2} + h^{1/2} \log n \right) \right\} + o_p(1) = o_p(1). \end{aligned}$$

Finally, Theorem 1 follows from Lemma 1 and Slutsky's Theorem.

#### A.4 Proof of Theorem 2

See the Supplement.

#### References

1. Berg D (2007) Bankruptcy prediction by generalized additive models. *Appl. Stoch. Models Bus. Ind.* 23: 129–143
2. Bernhardsen E (2001) A model of bankruptcy prediction. Norges Bank, WP
3. Bickel PJ, Rosenblatt M (1973) On some global measures of the deviations of density function estimates. *Ann. Statist.* 1: 1071–1095
4. Cai L, Yang L (2015) A smooth simultaneous confidence band for conditional variance function. *TEST* 24: 632–655
5. Engelmann B, Hayden E, Tasche D (2003) Testing rating accuracy. *Risk* 16: 82–86
6. Fan J, Yao Q (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, Berlin
7. Fan J, Zhang WY (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.* 27: 715–731
8. Gu L, Wang L, Härdle W, Yang L (2014) A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *TEST* 23: 806–843
9. Gu, L, Yang L (2015) Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electronic Journal of Statistics* 9: 1540–1561
10. Härdle W (1989) Asymptotic maximal deviation of M-smoothers. *J. Multivariate Anal.* 29: 163–179
11. Hastie TJ, Tibshirani RJ (1990) *Generalized additive models*. Chapman and Hall, London
12. He X, Fung W, Zhu Z (2005) Robust estimation in generalized partial linear models for clustered data. *J. Amer. Statist. Assoc.* 100: 1176–1184

13. He X, Zhu Z, Fung, W (2002) Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* 89: 579–590
14. Horowitz J, Mammen E (2004) Nonparametric estimation of an additive model with a link function. *Ann. Statist.* 32: 2412–2443
15. Huang JZ, Yang L (2004) Identification of nonlinear additive autoregression models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 66: 463–477
16. Linton OB (1997) Efficient estimation of additive nonparametric regression models. *Biometrika* 84: 469–473
17. Linton OB, Härdle W (1996) Estimation of additive regression models with known links. *Biometrika* 83: 529–540
18. Liu R, Yang L (2010) Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory* 26: 29–59
19. Liu R, Yang L, Härdle W (2013) Oracally efficient two-step estimation of generalized additive model. *J. Amer. Statist. Assoc.* 108: 619–631
20. Ma S, Yang L (2011) Spline-backfitted kernel smoothing of partially linear additive model. *J. Statist. Plann. Inference* 141: 204–219
21. Ma S, Yang L, Carroll RJ (2012) Simultaneous confidence band for sparse longitudinal regression. *Statist. Sinica* 22: 95–122
22. Ryser M, Denzler S (2009) Selecting credit rating models: a cross-validation-based comparison of discriminatory power. *Financ. Mark. Portf. Manag.* 23: 187–203
23. Severini T, Staniswalis J (1994) Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* 89: 501–511.
24. Shina Y, Moore W (2003) Explaining credit rating differences between Japanese and U.S. agencies. *Rev. Finan. Econ.* 12: 327–344
25. Stone CJ (1985) Additive regression and other nonparametric models. *Ann. Statist.* 13: 689–705
26. Stone CJ (1986) The dimensionality reduction principle for generalized additive models. *Ann. Statist.* 14: 590–606
27. Tusnady G (1977) A remark on the approximation of the sample distribution function in the multidimensional case. *Period. Math. Hungar.* 8: 53–55
28. Wang J, Liu R, Cheng F, Yang L (2014) Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. *Ann. Statist.* 42: 654–668
29. Wang L, Yang L (2007) Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* 35: 2474–2503
30. Wang L, Yang L (2009) Spline estimation of single index model. *Statist. Sinica* 19: 765–783
31. Wang L, Li H, Huang J (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* 103: 1556–1569
32. Wiesenfarth M, Krivobokova T, Klases S, Sperlich S (2012) Direct Simultaneous Inference in Additive Models and its Application to Model Undernutrition. *J. Amer. Statist. Assoc.* 107: 1286–1296
33. Wu W, Zhao Z (2007) Inference of trends in time series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69: 391–410
34. Xia Y (1998) Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 60: 797–811
35. Xue L, Yang L (2006) Additive coefficient modeling via polynomial spline. *Statist. Sinica* 16: 1423–1446
36. Yang L, Sperlich S, Härdle W (2003) Derivative estimation and testing in generalized additive models. *J. Statist. Plann. Inference* 115: 521–542
37. Zhang H, Lin Y (2006) Component selection and smoothing for nonparametric regression in exponential families. *Statist. Sinica* 16: 1021–1042
38. Zhao Z, Wu W (2008) Confidence bands in nonparametric time series regression. *Ann. Statist.* 36: 1854–1878
39. Zheng S, Yang L, Härdle W (2014) A smooth simultaneous confidence corridor for the mean of sparse functional data. *J. Amer. Statist. Assoc.* 109: 661–673





## Confidence Corridors for Multivariate Generalized Quantile Regression

Shih-Kang Chao, Katharina Proksch, Holger Dette & Wolfgang Karl Härdle


To cite this article: Shih-Kang Chao, Katharina Proksch, Holger Dette & Wolfgang Karl Härdle (2015): Confidence Corridors for Multivariate Generalized Quantile Regression, Journal of Business & Economic Statistics, DOI: [10.1080/07350015.2015.1054493](https://doi.org/10.1080/07350015.2015.1054493)

To link to this article: <http://dx.doi.org/10.1080/07350015.2015.1054493>

 View supplementary material [↗](#)

 Accepted author version posted online: 11 Jun 2015.

 Submit your article to this journal [↗](#)

 Article views: 24

 View Crossmark data [↗](#)

Full Terms & Conditions of access and use can be found at  
<http://amstat.tandfonline.com/action/journalInformation?journalCode=ubes20>

# Confidence Corridors for Multivariate Generalized Quantile Regression

Shih-Kang Chao\*      Katharina Proksch†      Holger Dette‡  
 Wolfgang Karl Härdle†‡

## Abstract

We focus on the construction of confidence corridors for multivariate nonparametric generalized quantile regression functions. This construction is based on asymptotic results for the maximal deviation between a suitable nonparametric estimator and the true function of interest, which follow after a series of approximation steps including a Bahadur representation, a new strong approximation theorem and exponential tail inequalities for Gaussian random fields.

As a byproduct we also obtain multivariate confidence corridors for the regression function in the classical mean regression. In order to deal with the problem of slowly decreasing error in coverage probability of the asymptotic confidence corridors, which results in meager coverage for small sample sizes, a simple bootstrap procedure is designed based on the leading term of the Bahadur representation. The finite sample properties of both procedures are investigated by means of a simulation study and it is demonstrated that the bootstrap procedure considerably outperforms the asymptotic bands in terms of coverage accuracy. Finally, the bootstrap confidence corridors are used to study the efficacy of the National Supported Work Demonstration, which is a randomized employment enhancement program launched in the 1970s. This article has supplementary materials online.

*Keywords:* Bootstrap; Expectile regression; Goodness-of-fit tests; Quantile treatment effect; Smoothing and nonparametric regression.

*JEL:* C2, C12, C14

---

\*Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. email: shih-kang.chao@cms.hu-berlin.de; haerdle@wiwi.hu-berlin.de.

†Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany. email: katharina.proksch@rub.de; holger.dette@rub.de.

‡Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore.

## 1. Introduction

Mean regression analysis is a widely used tool in statistical inference for curves. It focuses on the center of the conditional distribution, given  $d$ -dimensional covariates with  $d \geq 1$ . In a variety of applications though the interest is more in tail events, or even tail event curves such as the conditional quantile function. Applications with a specific demand in tail event curve analysis include finance, climate analysis, labor economics and systemic risk management.

Tail event curves have one thing in common: they describe the likeliness of extreme events conditional on the covariate  $X$ . A traditional way of defining such a tail event curve is by translating "likeliness" with "probability" leading to conditional quantile curves. Extreme events may alternatively be defined through conditional moment behaviour leading to more general tail descriptions as studied by [Newey and Powell \(1987\)](#) and [Jones \(1994\)](#). We employ this more general definition of generalized quantile regression (GQR), which includes, for instance, expectile curves and study statistical inference of GQR curves through confidence corridors.

In applications parametric forms are frequently used because of practical numerical reasons. Efficient algorithms are available for estimating the corresponding curves. However, the "monocular view" of parametric inference has turned out to be too restrictive. This observation prompts the necessity of checking the functional form of GQR curves. Such a check may be based on testing different kinds of variation between a hypothesized (parametric) model and a smooth alternative GQR. This approach though involves either an explicit estimate of the bias or a pre-smoothing of the "null model". In this paper we pursue the Kolmogorov-Smirnov type of approach, that is, employing the maximal deviation between the null and the smooth GQR curve as a test statistic. Such a model check has the advantage that it may be displayed graphically as confidence corridors (CC; also called "simultaneous confidence band" or "uniform confidence band/region") but has been considered so far only for univariate covariates. The basic technique for constructing CC of this type is extreme value theory for the sup-norm of an appropriately centered nonparametric estimate

of the quantile curve.

Confidence corridors with one-dimensional predictor were developed under various settings. Classical one-dimensional results are confidence bands constructed for histogram estimators by [Smirnov \(1950\)](#) or more general one-dimensional kernel density estimators by [Bickel and Rosenblatt \(1973\)](#). The results were extended to a univariate nonparametric mean regression setting by [Johnston \(1982\)](#), followed by [Härdle \(1989\)](#) who derived CCs for one-dimensional kernel  $M$ -estimators. [Claeskens and Van Keilegom \(2003\)](#) proposed uniform confidence bands and a bootstrap procedure for regression curves and their derivatives.

In recent years, the growth of the literature body shows no sign of decelerating. In the same spirit of [Härdle \(1989\)](#), [Härdle and Song \(2010\)](#) and [Guo and Härdle \(2012\)](#) constructed uniform confidence bands for local constant quantile and expectile curves. [Fan and Liu \(2013\)](#) proposed an integrated approach for building simultaneous confidence band that covers semiparametric models. [Giné and Nickl \(2010\)](#) investigated adaptive density estimation based on linear wavelet and kernel density estimators and [Lounici and Nickl \(2011\)](#) extended the framework of [Bissantz et al. \(2007\)](#) to adaptive deconvolution density estimation. Bootstrap procedures are proposed as a remedy for the poor coverage performance of asymptotic confidence corridors. For example, the bootstrap for the density estimator is proposed in [Hall \(1991\)](#) and [Mojirsheibani \(2012\)](#), and for local constant quantile estimators in [Song et al. \(2012\)](#).

However, only recently progress has been achieved in the construction of confidence bands for regression estimates with a multivariate predictor. [Hall and Horowitz \(2013\)](#) derived an expansion for the bootstrap bias and established a somewhat different way to construct confidence bands without the use of extreme value theory. Their bands are uniform with respect to a fixed but unspecified portion (smaller than one) of points in a possibly multidimensional set in contrast to the classical approach where uniformity is achieved on the complete set considered. [Proksch et al. \(2015\)](#) proposed multivariate confidence bands for convolution type inverse regression models with fixed design.

# ACCEPTED MANUSCRIPT

To the best of our knowledge, the classical Smirnov-Bickel-Rosenblatt type confidence corridors are not available for multivariate GQR or mean regression with random design.

In this work we go beyond the earlier studies in three aspects. First, we extend the applicability of the CC to  $d$ -dimensional covariates with  $d > 1$ . Second, we present a more general approach covering not only quantile or mean curves but also GQR curves that are defined via a minimum contrast principle. Third, we propose a bootstrap procedure and we show numerically its improvement in the coverage accuracy as compared to the asymptotic approach.

Our asymptotic results, which describe the maximal absolute deviation of generalized quantile estimators, can not only be used to derive a goodness-of-fit test in quantile and expectile regression, but they are also applicable in testing the quantile treatment effect and stochastic dominance. We apply the new method to test the quantile treatment effect of the National Supported Work Demonstration program, which is a randomized employment enhancement program launched in the 1970s. The data associated with the participants of the program have been widely applied in the field of treatment effect research since the pioneering study of [LaLonde \(1986\)](#). More recently, [Delgado and Escanciano \(2013\)](#) found that the program is beneficial for individuals of over 21 years of age. In our study, we find that the treatment tends to do better at raising the upper bounds of the earnings growth than raising the lower bounds. In other words, the program tends to increase the potential for high earnings growth but does not reduce the risk of negative earnings growth. The finding is particularly evident for those individuals who are older and spent more years at school. We should note that the tests based on the unconditional distribution cannot unveil the heterogeneity in the earnings growth quantiles in treatment effects.

The remaining part of this paper is organized as follows. In [Section 2](#) we present our model, describe the estimators and state our asymptotic results. [Section 3](#) is devoted to the bootstrap and we discuss its theoretical and practical aspects. The finite sample properties of both methods are investigated by means of a simulation study in [Section 4](#), where we also compare the numerical performance of our method with the method proposed in [Hall and Horowitz \(2013\)](#) via simula-

ACCEPTED MANUSCRIPT

tions. The application of our new method is illustrated by a real data example in Section 5. The assumptions for our asymptotic theory are listed and discussed after the references. All detailed proofs are available in the supplement material.

## 2. Asymptotic confidence corridors

In Section 2.1 we present the prerequisites such as the precise definition of the model and a suitable estimate. The results on constructing confidence corridors (CCs) based on the distribution of the maximal absolute deviation are given in Section 2.2. In Section 2.3 we describe how to estimate the scaling factors, which appear in the limit theorems, using residual based estimators. Section 3.1 introduce a new bootstrap method for constructing CCs, while Section 3.2 is devoted to specific issues related to bootstrap CCs for quantile regression. Assumptions are listed and discussed after the references.

### 2.1. Prerequisites

Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be a sequence of independent identically distributed random vectors in  $\mathbb{R}^{d+1}$  and consider the nonparametric regression model

$$Y_i = \theta_0(X_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where  $\theta_0$  is an aspect of  $Y$  conditional on  $X$ , such as the  $\tau$ -quantile, the  $\tau$ -expectile or the mean regression curve, and the model errors  $\varepsilon_1, \dots, \varepsilon_n$  are i.i.d. with  $\tau$ -quantile,  $\tau$ -expectile or mean equal to 0, respectively, depending on which  $\theta_0$  is in the model. The function  $\theta(x)$  can be estimated by:

$$\hat{\theta}(x) = \arg \min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \rho(Y_i - \theta), \quad (2)$$



where  $K_h(\mathbf{u}) = h^{-d}K(\mathbf{u}/h)$  for some kernel function  $K : \mathbb{R}^d \rightarrow \mathbb{R}$ , and a loss function  $\rho_\tau : \mathbb{R} \rightarrow \mathbb{R}$ . In this paper we are concerned with the construction of uniform confidence corridors for quantile as well as expectile regression curves when the predictor is multivariate, that is, we focus on the loss functions

$$\rho_\tau(u) = |\mathbf{1}(u < 0) - \tau||u|^k,$$

for  $k = 1$  and  $2$  associated with quantile and expectile regression. We derive the asymptotic distribution of the properly scaled maximal deviation  $\sup_{\mathbf{x} \in \mathcal{D}} |\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x})|$  for both cases, where  $\mathcal{D} \subset \mathbb{R}^d$  is a compact subset. We use strong approximations of the empirical process, concentration inequalities for general Gaussian random fields and results from extreme value theory. To be precise, we show that

$$\mathbb{P} \left[ (2\delta \log n)^{1/2} \left\{ \sup_{\mathbf{x} \in \mathcal{D}} |r_n(\mathbf{x})[\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x})]| / \|K\|_2 - d_n \right\} < a \right] \rightarrow \exp \{ -2 \exp(-a) \}, \quad (3)$$

as  $n \rightarrow \infty$ , where  $r_n(\mathbf{x})$  is a scaling factor which depends on  $\mathbf{x}$ ,  $n$  and the loss function under consideration.

## 2.2. Asymptotic results

In this section we present our main theoretical results on the distribution of the uniform maximal deviation of the quantile and expectile estimator. The proofs of the theorems at their full lengths are deferred to the appendix. Here we only give a brief sketch of proof of Theorem 2.1 which is the limit theorem for the case of quantile regression.

**THEOREM 2.1.** *Let  $\hat{\theta}_n(\mathbf{x})$  and  $\theta_0(\mathbf{x})$  be the local constant quantile estimator and the true quantile function, respectively and suppose that assumptions (A1)-(A6) in Section A hold. Let further*

# ACCEPTED MANUSCRIPT

$\text{vol}(\mathcal{D}) = 1$  and

$$d_n = (2d \cdot \kappa \log n)^{1/2} + \{2d\kappa(\log n)\}^{-1/2} \left[ \frac{1}{2}(d-1) \log \log n^\kappa + \log \{(2\pi)^{-1/2} H_2(2d)^{(d-1)/2}\} \right],$$

where  $d$  is the dimension of covariate  $\mathbf{X}$ ,  $h \asymp n^{-\kappa}$ ,  $H_2 = (2\pi\|K\|_2^2)^{-d/2} \det(\Sigma)^{1/2}$ ,  $\Sigma = (\Sigma_{ij})_{1 \leq i, j \leq d} = \left( \int \frac{\partial K(\mathbf{u})}{\partial u_i} \frac{\partial K(\mathbf{u})}{\partial u_j} d\mathbf{u} \right)_{1 \leq i, j \leq d}$ ,

$$r_n(\mathbf{x}) = \sqrt{\frac{nh^d f_{\mathbf{X}}(\mathbf{x})}{\tau(1-\tau)}} f_{Y|X}\{\theta_0(\mathbf{x})|\mathbf{x}\},$$

Then the limit theorem (3) holds.

**Sketch of proof.** A major technical difficulty is imposed by the fact that the loss function  $\rho_\tau$  is not smooth which means that standard arguments such as those based on Taylor's theorem do not apply. As a consequence the use of a different, extended methodology becomes necessary. In this context [Kong et al. \(2010\)](#) derived a uniform Bahadur representation for an  $M$ -regression function in a multivariate setting (see appendix). It holds uniformly for  $\mathbf{x} \in \mathcal{D}$ , where  $\mathcal{D}$  is a compact subset of  $\mathbb{R}^d$ :

$$\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x}) = \frac{1}{nS_{n,0,0}(\mathbf{x})} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{X}_i) \psi_\tau\{Y_i - \theta_0(\mathbf{x})\} + O\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}, \quad a.s. \quad (4)$$

Here  $S_{n,0,0}(\mathbf{x}) = \int K(\mathbf{u})g(\mathbf{x} + h\mathbf{u})f_{\mathbf{X}}(\mathbf{x} + h\mathbf{u})d\mathbf{u}$ ,  $\psi_\tau(u) = \mathbf{1}(u < 0) - \tau$  is the piecewise derivative of the loss function  $\rho_\tau$  and

$$g(\mathbf{x}) = \left. \frac{\partial}{\partial t} \mathbb{E}[\psi_\tau(Y - t)|\mathbf{X} = \mathbf{x}] \right|_{t=\theta_0(\mathbf{x})}.$$

Notice that the error term of the Bahadur expansion does not depend on the design  $\mathbf{X}$  and it converges to 0 with rate  $(\log n/nh^d)^{\frac{3}{4}}$  which is much faster than the convergence rate  $(nh^d)^{-\frac{1}{2}}$  of the stochastic term.

# ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

Rearranging (4), we obtain

$$S_{n,0,0}(\mathbf{x})\{\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x})\} = \frac{1}{n} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{X}_i) \psi_\tau\{Y_i - \theta_0(\mathbf{x})\} + O\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}. \quad (5)$$

Now we express the leading term on the right hand side of (5) by means of the centered empirical process

$$Z_n(y, \mathbf{u}) = n^{1/2}\{F_n(y, \mathbf{u}) - F(y, \mathbf{u})\}, \quad (6)$$

where  $F_n(y, \mathbf{x}) = n^{-1} \sum_{i=1}^n \mathbf{1}(Y_i \leq y, X_{i1} \leq x_1, \dots, X_{id} \leq x_d)$ . This yields, by Fubini's theorem,

$$S_{n,0,0}(\mathbf{x})\{\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x})\} - b(\mathbf{x}) = n^{-1/2} \int \int K_h(\mathbf{x} - \mathbf{u}) \psi_\tau\{y - \theta_0(\mathbf{x})\} dZ_n(y, \mathbf{u}) + O\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}, \quad (7)$$

where

$$b(\mathbf{x}) = -\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{X}_i) \psi\{Y_i - \theta_0(\mathbf{x})\}\right]$$

denotes the bias which is of order  $O(h^s)$  by Assumption (A3) in the Appendix. The variance of the first term of the right hand side of (7) can be estimated via a change of variables and Assumption (A5), which gives

$$\begin{aligned} & (nh^d)^{-2} n \mathbb{E}[K^2\{(\mathbf{x} - \mathbf{X}_i)/h\} \psi^2\{Y_i - \theta_0(\mathbf{x})\}] \\ &= (nh^d)^{-2} nh^d \int \int K^2(\mathbf{v}) \psi^2\{y - \theta_0(\mathbf{x})\} f_{Y|X}(y|\mathbf{x} - h\mathbf{v}) f_X(\mathbf{x} - h\mathbf{v}) dy d\mathbf{v} \\ &= (nh^d)^{-1} \int \int K^2(\mathbf{v}) \psi^2\{y - \theta_0(\mathbf{x})\} f_{Y|X}(y|\mathbf{x}) f_X(\mathbf{x}) dy d\mathbf{v} + O((nh^{d-1})^{-1}) \\ &= (nh^d)^{-1} f_X(\mathbf{x}) \sigma^2(\mathbf{x}) \|K\|_2^2 + O\{(nh^d)^{-1} h\}, \end{aligned}$$

where  $\sigma^2(\mathbf{x}) = \mathbb{E}[\psi^2\{Y - \theta_0(\mathbf{x})\} | X = \mathbf{x}]$ . The standardized version of (5) can therefore be approxi-

# ACCEPTED MANUSCRIPT

mated by

$$\begin{aligned} & \frac{\sqrt{nh^d}}{\sqrt{f_X(\mathbf{x})\sigma(\mathbf{x})\|K\|_2}} S_{n,0,0}(\mathbf{x})\{\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x})\} \\ &= \frac{1}{\sqrt{h^d f_X(\mathbf{x})\sigma(\mathbf{x})\|K\|_2}} \int \int K\left(\frac{\mathbf{x}-\mathbf{u}}{h}\right) \psi\{Y_i - \theta_0(\mathbf{x})\} dZ_n(y, \mathbf{u}) + \mathcal{O}(\sqrt{nh^d}h^s) + \mathcal{O}\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}. \end{aligned} \quad (8)$$

The dominating term is defined by

$$Y_n(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{h^d f_X(\mathbf{x})\sigma(\mathbf{x})}} \int \int K\left(\frac{\mathbf{x}-\mathbf{u}}{h}\right) \psi\{y - \theta_0(\mathbf{x})\} dZ_n(y, \mathbf{u}). \quad (9)$$

Involving strong Gaussian approximation and Bernstein-type concentration inequalities, this process can be approximated by a stationary Gaussian field:

$$Y_{5,n}(\mathbf{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\mathbf{x}-\mathbf{u}}{h}\right) dW(\mathbf{u}), \quad (10)$$

where  $W$  denotes a Brownian sheet. The supremum of this process is asymptotically Gumbel distributed, which follows, e.g., by Theorem 2 of [Rosenblatt \(1976\)](#). Since the kernel is symmetric and of order  $s$ , we can estimate the term

$$S_{n,0,0} = f_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x})f_X(\mathbf{x}) + \mathcal{O}(h^s)$$

if (A5) holds. On the other hand,  $\sigma^2(\mathbf{x}) = \tau(1-\tau)$  in quantile regression. Therefore, the statements of the theorem hold. □

**Corollary 2.2** (CC for multivariate quantile regression). Under the assumptions and notations of

# ACCEPTED MANUSCRIPT

Theorem 2.1, an approximate  $(1 - \alpha) \times 100\%$  confidence corridor is given by

$$\hat{\theta}_n(\mathbf{t}) \pm (nh^d)^{-1/2} \{ \tau(1 - \tau) \|K\|_2 / \hat{f}_X(\mathbf{t}) \}^{1/2} \hat{f}_{\varepsilon|X}\{0|\mathbf{t}\}^{-1} \{ d_n + c(\alpha)(2\kappa d \log n)^{-1/2} \},$$

where  $\alpha \in (0, 1)$  and  $c(\alpha) = \log 2 - \log |\log(1 - \alpha)|$  and  $\hat{f}_X(\mathbf{t}), \hat{f}_{\varepsilon|X}\{0|\mathbf{t}\}$  are consistent estimates for  $f_X(\mathbf{t}), f_{\varepsilon|X}\{0|\mathbf{t}\}$  with convergence rate in sup-norm faster than  $o_p((\log n)^{-1/2})$ .

**Remark 2.3.** Note that under the conditions of Corollary 2.2 we find

$$\sup_{\mathbf{x} \in \mathcal{D}} |r_n(\mathbf{x})(\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x}))| = O_p(\sqrt{\log n}),$$

where

$$r_n(\mathbf{x}) = \sqrt{\frac{nh^d f_X(\mathbf{x})}{\tau(1 - \tau)}} f_{Y|X}\{\theta_0(\mathbf{x})|\mathbf{x}\}.$$

For kernel estimators  $\hat{f}_{\varepsilon|X}(0, \cdot)$  and  $\hat{f}_X(\cdot)$  converging in sup-norm with rate  $o_p(\log(n)^{-1/2})$  to  $f_{\varepsilon|X}(0, \cdot)$  and  $f_X(\cdot)$ , respectively, the quantity  $\hat{r}_n(\mathbf{x})$ , defined by

$$\hat{r}_n(\mathbf{x}) = \sqrt{\frac{nh^d \hat{f}_X(\mathbf{x})}{\tau(1 - \tau)}} \hat{f}_{\varepsilon|X}(0, \mathbf{x}),$$

inherits this rate. Furthermore, since we consider an additive error model, the conditional density  $f_{Y|X}\{\theta_0(\mathbf{x})|\mathbf{x}\}$  can be replaced by  $f_{\varepsilon|X}(0, \mathbf{x})$  (see Section 2.3 below for more details and the definition of suitable estimators). This yields

$$\sup_{\mathbf{x} \in \mathcal{D}} |\hat{r}_n(\mathbf{x})(\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x}))| = o_p(1) + \sup_{\mathbf{x} \in \mathcal{D}} |r_n(\mathbf{x})(\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x}))|.$$

Hence, by Slutsky's Lemma, the quantities  $\sup_{\mathbf{x} \in \mathcal{D}} |\hat{r}_n(\mathbf{x})(\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x}))|$  and  $\sup_{\mathbf{x} \in \mathcal{D}} |r_n(\mathbf{x})(\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x}))|$  have the same asymptotic distribution.

# ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

The expectile confidence corridor can be constructed in an analogous manner as the quantile confidence corridor. The two cases differ in the form and hence the properties of the loss function. Therefore we find for expectile regression:

$$S_{n,0,0}(\mathbf{x}) = -2[F_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x})(2\tau - 1) - \tau]f_X(\mathbf{x}) + O(h^\nu).$$

Through similar approximation steps as the quantile regression, we derive the following theorem.

**THEOREM 2.4.** *Let  $\hat{\theta}_n(\mathbf{x})$  be the local constant expectile estimator and  $\theta_0(\mathbf{x})$  the true expectile function. If Assumptions (A1), (A3)-(A6) and (EA2) of Section A hold with a constant  $b_1$  satisfying*

$$n^{-1/6}h^{-d/2-3d/(b_1-2)} = O(n^{-\nu}), \quad \nu > 0.$$

Then the limit theorem (3) holds with a scaling factor

$$r_n(\mathbf{x}) = \sqrt{nh^d f_X(\mathbf{x})} \sigma^{-1}(\mathbf{x}) \{2[\tau - F_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x})(2\tau - 1)]\}$$

with quantities  $d$ ,  $h \asymp n^{-\kappa}$ ,  $H_2$  and  $d_n$  as defined in Theorem 2.1, where  $\sigma^2(\mathbf{x}) = \mathbb{E}[\psi_\tau^2(Y - \theta_0(\mathbf{x}))|\mathbf{X} = \mathbf{x}]$  and  $\psi_\tau(u) = 2(\mathbf{1}(u \leq 0) - \tau)|u|$  is the derivative of the expectile loss function  $\rho_\tau(u) = |\tau - \mathbf{1}(u < 0)||u|^2$ .

The proof of this result is deferred to the appendix. In the next corollary, the explicit form of the CCs for expectiles is given.

**Corollary 2.5** (CC for multivariate expectile regression). Under the same assumptions of Theorem 2.4, an approximate  $(1 - \alpha) \times 100\%$  confidence corridor is given by

$$\hat{\theta}_n(\mathbf{t}) \pm (nh^d)^{-1/2} \{\hat{\sigma}^2(\mathbf{t}) \|K\|_2 / \hat{f}_X(\mathbf{t})\}^{1/2} \left\{ -2[\hat{F}_{\varepsilon|X}\{0|\mathbf{t}\}(2\tau - 1) - \tau] \right\}^{-1} \left\{ d_n + c(\alpha)(2kd \log n)^{-1/2} \right\},$$

# ACCEPTED MANUSCRIPT

where  $\alpha \in (0, 1)$   $c(\alpha) = \log 2 - \log |\log(1 - \alpha)|$  and  $\hat{f}_X(\mathbf{t})$ ,  $\hat{\sigma}^2(\mathbf{t})$  and  $\hat{F}_{\varepsilon|X}(0|\mathbf{x})$  are consistent estimates for  $f_X(\mathbf{t})$ ,  $\sigma^2(\mathbf{t})$  and  $F_{\varepsilon|X}(0|\mathbf{x})$  with convergence rate in sup-norm faster than  $o_p((\log n)^{-1/2})$ .

A further immediate consequence of Theorem 2.4 is a similar limit theorem in the context of local least squares estimation of the regression curve in classical mean regression.

**Corollary 2.6** (CC for multivariate mean regression). Consider the loss function  $\rho(u) = u^2$  corresponding to  $\psi(u) = 2u$ . Under the assumptions and notations of Theorem 2.4, with the same constants  $H_2$  and  $d_n$ , (3) holds for the local constant estimator  $\hat{\theta}$  and the regression function  $\theta(\mathbf{x}) = E[Y|X = \mathbf{x}]$  with scaling factor  $r(\mathbf{x}) = \sqrt{nh^d f_X(\mathbf{x})} \sigma^{-1}(\mathbf{x})$  and  $\sigma^2(\mathbf{x}) = \text{Var}[Y|X = \mathbf{x}]$ .

**Remark 2.7.** We would like to stress that our purely nonparametric approach offers flexibility and reasonable results in moderate dimensions  $d = 2, d = 3$ , but it is not suitable for inference in high dimensional models due to the curse of dimensionality. The case of high dimensional regressors may be handled via a semi-parametric specification of the regression curve, such as, for instance, a partial linear model. Such a model was considered in Song et al. (2012) with a one-dimensional nonparametric component. Our approach allows to adapt their ideas and, as an extension, to consider a nonparametric component which is multivariate. Hence, our approach then offers higher flexibility in semi-parametric modeling. This semi-parametric approach is not pursued further in this paper but it clearly deserves future research.

### 2.3. Estimating the scaling factors

The performance of the confidence bands is greatly influenced by the scaling factors  $\hat{f}_{\varepsilon|X}(v|\mathbf{x})$ ,  $F_{\varepsilon|X}(v|\mathbf{x})$  and  $\hat{\sigma}(\mathbf{x})^2$ . The purpose of this subsection is thus to propose a way to estimate these factors and investigate their asymptotic properties.

As pointed out by our referee, estimating  $f_{\varepsilon|X}(0)$  is not a trivial task. The application of a rank test described in Chapter 3.5 of Koenker (2005) is an alternative to avoid estimating  $f_{\varepsilon|X}(0)$  in parametric quantile regression. However, it is a challenging task to apply this technique to

kernel smoothing quantile regression. For pointwise nonparametric inference, it may be possible to construct a test by adding weights (given by  $h^{-1}K((\mathbf{x} - \mathbf{X}_i)/h)$ , where  $h$  is the bandwidth and  $K$  is the kernel function) in the linear programming problem and therefore its dual can also be computed. However, a global shape test like the one investigated in this paper cannot be derived from the rank test. Hence, it seems inevitable to estimate the nuisance parameters and plug them into the test statistics.

Since we consider the additive error model (1), the conditional distribution function  $F_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x})$  and the conditional density  $f_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x})$  can be replaced by  $F_{\varepsilon|X}(0|\mathbf{x})$  and  $f_{\varepsilon|X}(0|\mathbf{x})$ , respectively, where  $F_{\varepsilon|X}$  and  $f_{\varepsilon|X}$  are the conditional distribution and density functions of  $\varepsilon$ . Similarly, we have

$$\sigma^2(\mathbf{x}) = \mathbb{E}[\psi_\tau(Y - \theta_0(\mathbf{x}))^2 | \mathbf{X} = \mathbf{x}] = \mathbb{E}[\psi_\tau(\varepsilon)^2 | \mathbf{X} = \mathbf{x}]$$

where  $\varepsilon$  may depend on  $\mathbf{X}$  due to heterogeneity. It should be noted that the kernel estimators for  $f_{\varepsilon|X}(0|\mathbf{x})$  and  $f_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x})$  are asymptotically equivalent, but show different finite sample behavior. We explore this issue further in the following section.

Introducing the residuals  $\hat{\varepsilon}_i = Y_i - \hat{\theta}_n(\mathbf{X}_i)$ , we propose to estimate  $F_{\varepsilon|X}$ ,  $f_{\varepsilon|X}$  and  $\sigma^2(\mathbf{x})$  by

$$\hat{F}_{\varepsilon|X}(v|\mathbf{x}) = n^{-1} \sum_{i=1}^n G\left(\frac{v - \hat{\varepsilon}_i}{h_0}\right) L_{\bar{h}}(\mathbf{x} - \mathbf{X}_i) / \hat{f}_X(\mathbf{x}), \quad (11)$$

$$\hat{f}_{\varepsilon|X}(v|\mathbf{x}) = n^{-1} \sum_{i=1}^n g_{h_0}(v - \hat{\varepsilon}_i) L_{\bar{h}}(\mathbf{x} - \mathbf{X}_i) / \hat{f}_X(\mathbf{x}), \quad (12)$$

$$\hat{\sigma}^2(\mathbf{x}) = n^{-1} \sum_{i=1}^n \psi^2(\hat{\varepsilon}_i) L_{\bar{h}}(\mathbf{x} - \mathbf{X}_i) / \hat{f}_X(\mathbf{x}), \quad (13)$$

where  $\hat{f}_X(\mathbf{x}) = n^{-1} \sum_{i=1}^n L_{\bar{h}}(\mathbf{x} - \mathbf{X}_i)$ ,  $G$  is a given continuously differentiable cumulative distribution function and  $g$  is its derivative. The construction of estimators in (11) and (12) follows from the estimator for general conditional distribution and density functions discussed in Chapter 5 and 6 of Li and Racine (2007). The same bandwidth  $\bar{h}$  is applied to the three estimators, but the choice



of  $\bar{h}$  will make the convergence rate of (13) sub-optimal. More details on the choice of  $\bar{h}$  are given in section 3.2 below. Nevertheless, the rate of convergence of (13) is of polynomial order in  $n$ . The theory developed in this subsection can be generalized to the case of different bandwidth for different direction without much difficulty.

The estimators (11) and (12) belong to the family of residual-based estimators. The consistency of residual-based density estimators for errors in a regression model are explored in the literature in various settings. It is possible to obtain an expression for the residual based kernel density estimator as the sum of the estimator with the true residuals, the partial sum of the true residuals and a term for the bias of the nonparametrically estimated function, as shown in [Muhsal and Neumeyer \(2010\)](#), among others. The residual based *conditional* kernel density case is less considered in the literature. [Kiwitt and Neumeyer \(2012\)](#) consider the residual based kernel estimator for conditional distribution function conditioning on a one-dimensional variable.

Below we give consistency results for the estimators defined in (11), (12) and (13). The proof can be found in the appendix.

**Lemma 2.8.** Under conditions (A1), (A3)-(A5), (B1)-(B3) in Section A, we have

- 1)  $\sup_{v \in I} \sup_{\mathbf{x} \in \mathcal{D}} |\hat{F}_{\varepsilon|X}(v|\mathbf{x}) - F_{\varepsilon|X}(v|\mathbf{x})| = O_p(t_n)$ ,
- 2)  $\sup_{v \in I} \sup_{\mathbf{x} \in \mathcal{D}} |\hat{f}_{\varepsilon|X}(v|\mathbf{x}) - f_{\varepsilon|X}(v|\mathbf{x})| = O_p(t_n)$ ,
- 3)  $\sup_{\mathbf{x} \in \mathcal{D}} |\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})| = O_p(u_n)$ ,

where  $t_n = O\{h_0^s + h^s + \bar{h}^s + (n\bar{h}^d)^{-1/2} \log n + (nh^d)^{-1/2} \log n\} = O(n^{-\lambda})$ , and  $u_n = O\{h^s + \bar{h}^s + (n\bar{h}^d)^{-1/2} \log n + (nh^d)^{-1/2} \log n\} = O(n^{-\lambda_1})$  for some constants  $\lambda, \lambda_1 > 0$ .

The factor  $\log n$  shown in the convergence rate is the price which we pay for the sup norm deviation. Since these estimators uniformly converge in a polynomial rate in  $n$ , the asymptotic distributions in Theorem 2.1 and 2.4 remain the same if we plug these estimators into the formulae.

### 3. Bootstrap confidence corridors

#### 3.1. Asymptotic theory

In the case of the suitably normed maximum of independent standard normal variables, it is shown in Hall (1979) that the speed of convergence in limit theorems of the form (3) is of order  $1/\log n$ , that is, the coverage error of the asymptotic CC decays only logarithmically. This leads to unsatisfactory finite sample performance of the asymptotic methods, especially for small sample sizes and dimensions  $d > 1$ . However, Hall (1991) suggests that the use of a bootstrap method, based on a proper way of resampling, can increase the speed of shrinking of coverage error to a polynomial rate of  $n$ . In this section we therefore propose a specific bootstrap technique and construct a confidence corridor for the objects to be analysed.

Given the residuals  $\hat{\varepsilon}_i = Y_i - \hat{\theta}_n(\mathbf{X}_i)$ , the bootstrap observations  $(\mathbf{X}_i^*, \varepsilon_i^*)$  are sampled from

$$\hat{f}_{\varepsilon, \mathbf{X}}(v, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n g_{h_0}(\hat{\varepsilon}_i - v) L_{\bar{h}}(\mathbf{x} - \mathbf{X}_i), \quad (14)$$

where  $g$  and  $L$  are kernel functions with bandwidths  $h_0, \bar{h}$  satisfying assumptions (B1)-(B3). In particular, in our simulation study, we choose  $L$  to be a product Gaussian kernel. In the following discussion  $\mathbb{P}^*$  and  $\mathbb{E}^*$  stand for the probability and expectation conditional on the data  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ .

We introduce the notation

$$A_n^*(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_h(\mathbf{x} - \mathbf{X}_i^*) \psi_\tau(\varepsilon_i^*),$$

# ACCEPTED MANUSCRIPT

and define the so-called "one-step estimator"  $\hat{\theta}^*(\mathbf{x})$  from the bootstrap sample by

$$\hat{\theta}^*(\mathbf{x}) - \hat{\theta}_n(\mathbf{x}) = \hat{S}_{n,0,0}^{-1}(\mathbf{x}) \{A_n^*(\mathbf{x}) - E^*[A_n^*(\mathbf{x})]\}, \quad (15)$$

where

$$\hat{S}_{n,0,0}(\mathbf{x}) = \begin{cases} \hat{f}_{\varepsilon|X}(0|\mathbf{x})\hat{f}_X(\mathbf{x}), & \text{quantile case;} \\ 2\{\tau - \hat{F}_{\varepsilon|X}(0|\mathbf{x})(2\tau - 1)\}\hat{f}_X(\mathbf{x}), & \text{expectile case.} \end{cases} \quad (16)$$

note that  $E^*[\hat{\theta}^*(\mathbf{x}) - \hat{\theta}_n(\mathbf{x})] = 0$ , so  $\hat{\theta}^*(\mathbf{x})$  is unbiased for  $\hat{\theta}_n(\mathbf{x})$  under  $E^*$ . As a remark, we note that undersmoothing is applied in our procedure for two reasons: first, the theory we developed so far is based on undersmoothing; secondly, it is suggested in [Hall \(1992\)](#) that undersmoothing is more effective than oversmoothing given that the goal is to achieve coverage accuracy.

Note that the bootstrap estimate (15) is motivated by the smoothed bootstrap procedure proposed in [Claeskens and Van Keilegom \(2003\)](#). In contrast to these authors we make use of the leading term of the Bahadur representation. [Mammen et al. \(2013\)](#) also use the leading term of a Bahadur representation proposed in [Guerre and Sabbah \(2012\)](#) to construct bootstrap samples. [Song et al. \(2012\)](#) propose a bootstrap for quantile regression based on oversmoothing, which has the drawback that it requires iterative estimation, and oversmoothing is in general less effective in terms of coverage accuracy.

For the following discussion define

$$Y_n^*(\mathbf{x}) = \frac{1}{\sqrt{h^d \hat{f}_X(\mathbf{x}) \sigma_*(\mathbf{x})}} \int \int K\left(\frac{\mathbf{x} - \mathbf{u}}{h}\right) \psi_\tau(v) dZ_n^*(v, \mathbf{u}) \quad (17)$$

as the bootstrap analogue of the process (9), where

$$Z_n^*(y, \mathbf{u}) = n^{1/2} \{F_n^*(v, \mathbf{u}) - \hat{F}(v, \mathbf{u})\}, \quad \sigma_*(\mathbf{x}) = \sqrt{E^*[\psi_\tau(\varepsilon_i^*)^2|\mathbf{x}]} \quad (18)$$

# ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

and

$$F_n^*(v, \mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1} \{ \varepsilon_i^* \leq v, X_1^* \leq u_1, \dots, X_d^* \leq u_d \}.$$

The process  $Y_n^*$  serves as an approximation of a standardized version of  $\hat{\theta}_n^* - \hat{\theta}_n$ , and similar to the previous sections the process  $Y_n^*$  is approximated by a stationary Gaussian field  $Y_{n,5}^*$  under  $P^*$  with probability one, that is,

$$Y_{5,n}^*(\mathbf{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\mathbf{x} - \mathbf{u}}{h}\right) dW^*(\mathbf{u}).$$

Finally,  $\sup_{\mathbf{x} \in \mathcal{D}} |Y_{5,n}^*(\mathbf{x})|$  is asymptotically Gumbel distributed conditional on samples.

**THEOREM 3.1.** *Suppose that assumptions (A1)-(A6), (C1) in Section A hold, and  $\text{vol}(\mathcal{D}) = 1$ , let*

$$r_n^*(\mathbf{x}) = \sqrt{\frac{nh^d}{\hat{f}_{\mathbf{X}}(\mathbf{x})\sigma_*^2(\mathbf{x})}} \hat{S}_{n,0,0}(\mathbf{x}),$$

where  $\hat{S}_{n,0,0}(\mathbf{x})$  is defined in (16) and  $\sigma_*^2(\mathbf{x})$  is defined in (18). Then

$$P^* \left\{ (2d \cdot \kappa \log n)^{1/2} \left( \sup_{\mathbf{x} \in \mathcal{D}} [r_n^*(\mathbf{x})|\hat{\theta}^*(\mathbf{x}) - \hat{\theta}_n(\mathbf{x})|] / \|K\|_2 - d_n \right) < a \right\} \rightarrow \exp \{ -2 \exp(-a) \}, \quad a.s. \quad (19)$$

as  $n \rightarrow \infty$  for the local constant quantile regression estimate, with quantities  $d, h \asymp n^{-\kappa}, H_2$  and  $d_n$  as defined in Theorem 2.1. If (A1)-(A6) and (EC1) hold with a constant  $b \geq 4$  satisfying

$$n^{-\frac{1}{6} + \frac{4}{b^2} - \frac{1}{b}} h^{-\frac{d}{2} - \frac{6d}{b}} = O(n^{-\nu}), \quad \nu > 0,$$

then (19) also holds for expectile regression with corresponding  $\sigma_*^2(\mathbf{x})$ .

The proof can be found in the appendix. The following lemma suggests that we can replace  $\sigma_*^2(\mathbf{x})$  in the limiting theorem by  $\hat{\sigma}(\mathbf{x})$ .

# ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

**Lemma 3.2.** If assumptions (B1)-(B3), and (EC1) in Section A are satisfied with  $b > 2(2s' + d + 1)/(2s' + 3)$ , then

$$\|\sigma_*^2(\mathbf{x}) - \hat{\sigma}^2(\mathbf{x})\| = o_p^*((\log n)^{-1/2}), \quad a.s.$$

The following corollary is a consequence of Theorem 3.1.

**Corollary 3.3.** Under the same conditions as stated in Theorem 3.1, the (asymptotic) bootstrap confidence set of level  $1 - \alpha$  is given by

$$\left\{ \theta : \sup_{\mathbf{x} \in \mathcal{D}} \left| \frac{\hat{S}_{n,0,0}(\mathbf{x})}{\sqrt{\hat{f}_X(\mathbf{x})\hat{\sigma}^2(\mathbf{x})}} [\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})] \right| \leq \xi_\alpha^* \right\}, \quad (20)$$

where  $\xi_\alpha^*$  satisfies

$$\lim_{n \rightarrow \infty} \mathbb{P}^* \left( \sup_{\mathbf{x} \in \mathcal{D}} \left| \frac{\hat{S}_{n,0,0}(\mathbf{x})}{\sqrt{\hat{f}_X(\mathbf{x})\hat{\sigma}^2(\mathbf{x})}} [\hat{\theta}^*(\mathbf{x}) - \hat{\theta}_n(\mathbf{x})] \right| \leq \xi_\alpha^* \right) = 1 - \alpha, \quad a.s. \quad (21)$$

where  $\hat{S}_{n,0,0}$  is defined in (16).

Note that it does not create much difference to standardize the  $\hat{\theta}_n(\mathbf{x}) - \theta_0(\mathbf{x})$  in (19) with  $\hat{f}_X$  and  $\hat{\sigma}^2(\mathbf{x})$  constructed from original samples or  $\hat{f}_X^*$  and  $\hat{\sigma}^2(\mathbf{x})$  from the bootstrap samples. The simulation results of [Claeskens and Van Keilegom \(2003\)](#) show that the two ways of standardization give similar coverage probabilities for confidence corridors of kernel ML estimators.

## 3.2. Implementation

In this section, we discuss issues related to the implementation of the bootstrap for quantile regression.

Note that the *width* of the CC is determined by the variance and the *location* is affected by the bias of the quantile function estimator, and both depend on the bandwidth used for estimation.

ACCEPTED MANUSCRIPT

Hence, the choice of bandwidth needs to balance the bias (location) and the variance (size). It is chosen such that the bias is only just negligible after normalization, that is, slightly smaller than the  $L^2$ -optimal bandwidth. Therefore, it is enough to take an undersmoothed  $h = O(n^{-1/(2s+d)-\delta})$ , given that  $s > d$  and  $\delta > 0$ , where  $s$  is the order of Hölder continuity of the function  $\theta_0$  and  $\delta$  is the degree of undersmoothing. We may use the methods proposed by [Yu and Jones \(1998\)](#) for nonparametric quantile regression to choose the bandwidth before undersmoothing, namely

$$h_{\tau,j} = h_{1,j} \{ \tau(1-\tau) / \phi(\Phi^{-1}(\tau))^2 \}^{1/5}, \quad j = 1, 2, \quad (22)$$

where  $h_{1,j}$  is chosen by common methods like the rule-of-thumb or cross-validation for mean regression or density estimation and  $\Phi$  is the CDF of the standard Gaussian distribution. In our simulation study, we select  $h_{1,j}$  in (22) by the rule-of-thumb, implemented with the `np` package in R. In our application analysis,  $h_{1,j}$  in (22) is chosen by the cross-validated bandwidth for the conditional distribution smoother of  $Y$  given  $\mathbf{X}$ , implemented with the `np` package in R. This package is based on the paper of [Li et al. \(2013\)](#).

For expectile regression, we use the rule-of-thumb bandwidth for the conditional distribution smoother of  $Y$  given  $\mathbf{X}$ , chosen with the `np` package in R.

The choice of  $h_0$  and  $\bar{h}$  for estimating the scaling factors in Section 2.3 should minimize the uniform convergence rate of the residual based estimators. Hence, observing that the terms related to  $h_0$  and  $\bar{h}$  are similar to those in usual  $(d+1)$ -dimensional density estimators, it is reasonable to choose  $h_0 \sim \bar{h} \sim n^{-1/(5+d)}$ , given that  $L, g$  are second order kernels. We choose the rule-of-thumb bandwidths for conditional densities with the R package `np` in our simulation and application studies.

The one-step estimator for quantile regression defined in (15) depends sensitively on the estimator of  $\hat{S}_{n,0,0}(\mathbf{x})$ . Unlike in the expectile case, the function  $\psi(\cdot)$  in the quantile case is bounded, and, as a result, the bootstrapped density based on (20) is very easily influenced by the factor

$\hat{S}_{n,0,0}(\mathbf{x})$ ; in particular,  $\hat{f}_{\varepsilon|\mathbf{X}}(0|\mathbf{x})$ . As pointed out by Feng et al. (2011), the residual of quantile regression tends to be less dispersed than the model error; thus  $\hat{f}_{\varepsilon|\mathbf{X}}(0|\mathbf{x})$  tends to over-estimate the true  $f_{\varepsilon|\mathbf{X}}(0|\mathbf{x})$  for each  $\mathbf{x}$ .

The way of getting around this problem is based on the following observation: An additive error model implies the equality  $f_{Y|\mathbf{X}}\{v + \theta_0(\mathbf{x})|\mathbf{x}\} = f_{\varepsilon|\mathbf{X}}(v|\mathbf{x})$ , but this property does not hold for the kernel estimators

$$\hat{f}_{\varepsilon|\mathbf{X}}(0|\mathbf{x}) = n^{-1} \sum_{i=1}^n g_{h_0}(\hat{\varepsilon}_i) L_{\tilde{h}}(\mathbf{x} - \mathbf{X}_i) / \hat{f}_{\mathbf{X}}(\mathbf{x}), \quad (23)$$

$$\hat{f}_{Y|\mathbf{X}}(\hat{\theta}_n(\mathbf{x})|\mathbf{x}) = n^{-1} \sum_{i=1}^n g_{h_1}(Y_i - \hat{\theta}_n(\mathbf{x})) L_{\tilde{h}}(\mathbf{x} - \mathbf{X}_i) / \hat{f}_{\mathbf{X}}(\mathbf{x}), \quad (24)$$

of the conditional density functions. In general  $\hat{f}_{\varepsilon|\mathbf{X}}(0|\mathbf{x}) \neq \hat{f}_{Y|\mathbf{X}}(\hat{\theta}_n(\mathbf{x})|\mathbf{x})$  in  $\mathbf{x}$  although both estimates are asymptotically equivalent. In applications the two estimators can differ substantially due to the bandwidth selection because we usually have  $h_0 \neq h_1$  when they are chosen based on data. For example, if a common method for bandwidth selection such as a rule-of-thumb is used,  $h_1$  will tend to be larger than  $h_0$  since the sample variance of  $Y_i$  tends to be larger than that of  $\hat{\varepsilon}_i$ . Given that the same kernels are applied, it happens often that  $\hat{f}_{Y|\mathbf{X}}(\hat{\theta}_n(\mathbf{x})|\mathbf{x}) > \hat{f}_{\varepsilon|\mathbf{X}}(0|\mathbf{x})$ , even if  $\hat{\theta}_n(\mathbf{x})$  is usually very close to  $\theta_0(\mathbf{x})$ . To correct such abnormality, we are motivated to set  $h_1 = h_0$  which is the rule-of-thumb bandwidth of  $\hat{f}_{\varepsilon|\mathbf{X}}(v|\mathbf{x})$  in (24). As the result, it leads to a more rough estimate for  $\hat{f}_{Y|\mathbf{X}}(\hat{\theta}_n(\mathbf{x})|\mathbf{x})$ .

In order to exploit the roughness of  $\hat{f}_{Y|\mathbf{X}}(\hat{\theta}_n(\mathbf{x})|\mathbf{x})$  while making the CC as narrow as possible, we develop a trick depending on

$$\frac{\hat{f}_{Y|\mathbf{X}}\{\hat{\theta}_n(\mathbf{x})|\mathbf{x}\}}{\hat{f}_{\varepsilon|\mathbf{X}}(0|\mathbf{x})} = \frac{h_0 \sum_{i=1}^n g_{h_1}(\{Y_i - \hat{\theta}_n(\mathbf{x})\}/h_1) L_{\tilde{h}}(\mathbf{x} - \mathbf{X}_i)}{h_1 \sum_{i=1}^n g_{h_0}(\hat{\varepsilon}_i/h_0) L_{\tilde{h}}(\mathbf{x} - \mathbf{X}_i)}. \quad (25)$$

As  $n \rightarrow \infty$ , (25) converges to 1. If we impose  $h_0 = h_1$ , as the multiple  $h_0/h_1$  vanishes, (25) captures the deviation of the two estimators without the difference of the bandwidth in the way.

In particular, the bandwidth  $h_0 = h_1$  is selected as the rule-of-thumb bandwidth for  $\hat{f}_{\varepsilon|X}(y|\mathbf{x})$ . This makes  $\hat{f}_{\varepsilon|X}(y|\mathbf{x})$  larger and thus leads to a narrower CC, as will be more clear below.

We propose the alternative bootstrap confidence corridor for quantile estimator:

$$\left\{ \theta : \sup_{\mathbf{x} \in \mathcal{D}} \left| \sqrt{\hat{f}_X(\mathbf{x}) \hat{f}_{Y|X}\{\hat{\theta}_n(\mathbf{x})|\mathbf{x}\}} [\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})] \right| \leq \xi_\alpha^\dagger \right\},$$

where  $\xi_\alpha^\dagger$  satisfies

$$P^* \left( \sup_{\mathbf{x} \in \mathcal{D}} \left| \hat{f}_X(\mathbf{x})^{-1/2} \frac{\hat{f}_{Y|X}\{\hat{\theta}_n(\mathbf{x})|\mathbf{x}\}}{\hat{f}_{\varepsilon|X}(0|\mathbf{x})} [A_n^*(\mathbf{x}) - E^* A_n^*(\mathbf{x})] \right| \leq \xi_\alpha^\dagger \right) = 1 - \alpha. \quad (26)$$

Note that the probability on the left-hand side of (26) can again be approximated by a Gumbel distribution function asymptotically, which follows by Theorem 3.1.

## 4. A simulation study

In this section we investigate the methods described in the previous sections by means of a simulation study. We construct confidence corridors for quantiles and expectiles for different levels  $\tau$  and use the quartic (product) kernel. The performance of our methods is compared to the performance of the method proposed by Hall and Horowitz (2013) at the end of this section. For the confidence based on asymptotic distribution theory, we use the rule of thumb bandwidth chosen from the R package np, and then rescale it as described in Yu and Jones (1998), finally multiply it by  $n^{-0.05}$  for undersmoothing. The sample sizes are given by  $n = 100, 300$  and  $500$ , so the undersmoothing multiples are 0.794, 0.752 and 0.733 respectively. We take  $20 \times 20$  equally distant grids in the square  $[0.1, 0.9]^2$  and estimate quantile or expectile functions pointwisely on this set of grids. In the quantile regression bootstrap CC, the bandwidth  $h_1$  used for estimating  $\hat{f}_{Y|X}(y|\mathbf{x})$  is chosen to be the rule-of-thumb bandwidth of  $\hat{f}_{\varepsilon|X}(0|\mathbf{x})$  and multiplied by a multiple 1.5. This would give slightly wider CCs.



# ACCEPTED MANUSCRIPT

The data are generated from the normal regression model

$$Y_i = f(X_{1,i}, X_{2,i}) + \sigma(X_{1,i}, X_{2,i})\varepsilon_i, \quad i = 1, \dots, n \quad (27)$$

where the independent variables  $(X_1, X_2)$  follow a joint uniform distribution taking values on  $[0, 1]^2$ ,  $\text{Cov}(X_1, X_2) = 0.2876$ ,  $f(X_1, X_2) = \sin(2\pi X_1) + X_2$ , and  $\varepsilon_i$  are independent standard Gaussian random variables. For both quantile and expectile, we look at three quantiles of the distribution, namely  $\tau = 0.2, 0.5, 0.8$ . The set of grid point is  $H \times H$  where  $H$  is the set of 20 equidistant grids on univariate interval  $[0.1, 0.9]$ . Thus, the grid size is  $|H \times H| = 400$ .

In the homogeneous model, we take  $\sigma(X_1, X_2) = \sigma_0$ , for  $\sigma_0 = 0.2, 0.5, 0.7$ . In the heterogeneous model, we take  $\sigma(X_1, X_2) = \sigma_0 + 0.8X_1(1 - X_1)X_2(1 - X_2)$ . 2000 simulation runs are carried out to estimate the coverage probability.

The upper part of Table 1 shows the coverage probability of the asymptotic CC for nonparametric quantile regression functions. It can be immediately seen that the asymptotic CC performs very poorly, especially when  $n$  is small. A comparison of the results with those of one-dimensional asymptotic simultaneous confidence bands derived in [Claeskens and Van Keilegom \(2003\)](#) or [Fan and Liu \(2013\)](#), shows that the accuracy in the two-dimensional case is much worse. Much to our surprise, the asymptotic CC performs better in the case of  $\tau = 0.2, 0.8$  than in the case of  $\tau = 0.5$ . On the other hand, it is perhaps not so amazing to see that asymptotic CCs behave similarly under both homogeneous and heterogeneous models. As a final remark about the asymptotic CC we mention that it is highly sensitive with respect to  $\sigma_0$ . Increasing values of  $\sigma_0$  yields larger CC, and this may lead to greater coverage probability.

The lower part of Table 1 shows that the bootstrap CCs for nonparametric quantile regression functions yield a remarkable improvement in comparison to the asymptotic CC. For the bootstrap CC, the coverage probabilities are in general close to the nominal coverage of 95%. The bootstrap CCs are usually wider, and getting narrower when  $n$  increases. Such phenomenon can also

be found in the simulation study of [Claeskens and Van Keilegom \(2003\)](#). Bootstrap CCs are less sensitive than asymptotic CCs with respect to the choice  $\sigma_0$ , which is also considered as an advantage. Finally, we note that the performance of bootstrap CCs does not depend on which variance specification is used too.

The upper part of [Table 2](#) shows the coverage probability of the CC for nonparametric expectile regression functions. The results are similar to the case of quantile regression. The asymptotic CCs do *not* give accurate coverage probabilities. For example in some cases like  $\tau = 0.2$  and  $\sigma_0 = 0.2$ , not a single simulation in the 2000 iterations yields a case where surface is completely covered by the asymptotic CC.

The lower part of [Table 2](#) shows that bootstrap CCs for expectile regression give more accurate approximates to the nominal coverage than the asymptotic CCs. One can see in the parenthesis that the volumes of the bootstrap CCs are significantly larger than those of the asymptotic CCs, especially for small  $n$ .

[Table 3](#) presents the proportion in the 2000 iterations which covers 95% of the 400 grid points, using the bootstrap method proposed in [Hall and Horowitz \(2013\)](#) (abbreviated as HH) for nonparametric mean regression at  $d = 2$ . HH derived an expansion for the bootstrap bias and established a somewhat different way to construct confidence bands without the use of extreme value theory. It is worth noting that their bands are uniform with respect to a fixed but unspecified portion of  $(1 - \xi) \cdot 100\%$  (smaller than 100%) of grid points, while in our approach the uniformity is achieved on the whole set of grids.

The simulation model is [\(27\)](#) with the same homogeneous and heterogeneous variance specifications as before. We choose three levels  $\xi = 0.005, 0.05$  and  $0.1$ . It is suggested in HH that  $\xi = 0.1$  is usually sufficient in univariate nonparametric mean regression  $d = 1$ . Note that  $\xi = 0.005$  corresponds to the second smallest pointwise quantile  $\hat{\beta}(x, 0.05)$  in the notation of HH, given that our grid size is 400. This is close to the uniform CC in our sense. The simulation model associated with the [Table 3](#) is the same with that of the case  $\tau = 0.5$  in the bootstrap part of [Table 1](#) and [Table](#)

2, because in case of the normal distribution the median equals the mean and  $\tau = 0.5$  expectile is exactly the mean. However, one should be aware that our coverage probabilities are more stringent because we check the coverage at every point in the set of grids, rather than only 95% of the points (we refer it as *complete coverage*). Hence, the complete coverage probability of HH will be lower than the proportion of 95% coverage shown in Table 3. The proportion of 95% coverage should therefore be viewed as an upper bound for the complete coverage.

We summarize our findings as follows. Firstly the proportion of 95% coverage in general present similar patterns as shown in Table 1 and 2. The coverage improves when  $n$  and  $\sigma_0$  get larger, and the volume of the band decreases as  $n$  increases and increases when  $\sigma_0$  increases. The homogeneous and heterogeneous model yield similar performance. Comparing with the univariate result in HH, it is found that the proportion of coverage tends to perform worse than that in HH under the same sample size. This is due to the curse of dimensionality, the estimation of a bivariate function is less accurate than that of an univariate function. As the result, a more conservative  $\xi$  has to be applied. If we compare Table 3 to the bootstrap part of Table 1 with  $\tau = 0.5$ , it can be seen that our complete coverage probabilities are comparable to the proportion of 95% coverage at the case  $\xi = 0.005$ , though in the case of  $\sigma_0 = 0.2$  our CC does not perform very well. However, the volumes of our CC are much less than that of HH in the cases of small  $n$  and moderate and large  $\sigma_0$ . This suggests that our CC is more efficient. Finally, the proportion of 95% coverage at  $\xi = 0.005$  in Table 3 is similar to the complete coverage probability in bootstrap part of Table 2 with  $\tau = 0.5$ , but when sample size is small, the volume of our CC is smaller.

## 5. Application: a treatment effect study

The classical application of the proposed method is testing the hypothetical functional form of the regression function. Nevertheless, the proposed method can also be applied to test for a quantile treatment effect (see [Koenker, 2005](#)) or to test for conditional stochastic dominance (CSD)

# ACCEPTED MANUSCRIPT

as investigated in [Delgado and Escanciano \(2013\)](#). In this section we shall apply the new method to test these hypotheses for data collected from a real government intervention.

The estimation of the quantile treatment effect (QTE) recovers the heterogeneous impact of intervention on various points of the response distribution. To define QTE, given vector-valued exogenous variables  $\mathbf{X} \in \mathcal{X}$  where  $\mathcal{X} \subset \mathbb{R}^d$ , suppose  $Y_0$  and  $Y_1$  are response variables associated with the control group and treatment group, and let  $F_{0|\mathbf{X}}$  and  $F_{1|\mathbf{X}}$  be the conditional distribution for  $Y_0$  and  $Y_1$ , the QTE at level  $\tau$  is defined by

$$\Delta_\tau(\mathbf{x}) \stackrel{\text{def}}{=} Q_{1|\mathbf{X}}(\tau|\mathbf{x}) - Q_{0|\mathbf{X}}(\tau|\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}, \quad (28)$$

where  $Q_{0|\mathbf{X}}(y|\mathbf{x})$  and  $Q_{1|\mathbf{X}}(y|\mathbf{x})$  are the conditional quantile of  $Y_0$  given  $\mathbf{X}$  and  $Y_1$  given  $\mathbf{X}$ , respectively. This definition corresponds to the idea of horizontal distance between the treatment and control distribution functions appearing in [Doksum \(1974\)](#) and [Lehmann \(1975\)](#).

A related concept in measuring the efficiency of a treatment is the so called "conditional stochastic dominance".  $Y_1$  conditionally stochastically dominates  $Y_0$  if

$$F_{1|\mathbf{X}}(y|\mathbf{x}) \leq F_{0|\mathbf{X}}(y|\mathbf{x}) \quad \text{a.s. for all } (y, \mathbf{x}) \in (\mathcal{Y}, \mathcal{X}), \quad (29)$$

where  $\mathcal{Y}, \mathcal{X}$  are domains of  $Y$  and  $\mathbf{X}$ . For example, if  $Y_0$  and  $Y_1$  stand for the income of two groups of people  $G_0$  and  $G_1$ , (29) means that the distribution of  $Y_1$  lies on the right of that of  $Y_0$ , which is equivalent to saying that at a given  $0 < \tau < 1$ , the  $\tau$ -quantile of  $Y_1$  is greater than that of  $Y_0$ . Hence, we could replace the testing problem (29) by

$$Q_{1|\mathbf{X}}(\tau|\mathbf{x}) \geq Q_{0|\mathbf{X}}(\tau|\mathbf{x}) \quad \text{for all } 0 < \tau < 1 \text{ and } \mathbf{x} \in \mathcal{X}. \quad (30)$$

Comparing (30) and (28), one would find that (30) is just a uniform version of the test  $\Delta_\tau(\mathbf{x}) \geq 0$  over  $0 < \tau < 1$ .

# ACCEPTED MANUSCRIPT

The method that we introduced in this paper is suitable for testing a hypothesis like  $\Delta_\tau(\mathbf{x}) = 0$  where  $\Delta_\tau(\mathbf{x})$  is defined in (28). One can construct CCs for  $Q_{1|X}(\tau|\mathbf{x})$  and  $Q_{0|X}(\tau|\mathbf{x})$  respectively, and then check if there is overlap between the two confidence regions. One can also extend this idea to test (30) by building CCs for several selected levels  $\tau$ .

We use our method to test the effectiveness of the National Supported Work (NSW) demonstration program, which was a randomized, temporary employment program initiated in 1975 with the goal to provide work experience for individuals who face economic and social problems prior to entering the program. The data have been widely applied to examine techniques which estimate the treatment effect in a nonexperimental setting. In a pioneer study, [LaLonde \(1986\)](#) compares the treatment effect estimated from the experimental NSW data with that implied by nonexperimental techniques. [Dehejia and Wahba \(1999\)](#) analyse a subset of Lalonde's data and propose a new estimation procedure for nonexperimental treatment effect giving more accurate estimates than Lalonde's estimates. The paper that is most related to our study is [Delgado and Escanciano \(2013\)](#). These authors propose a test for hypothesis (29) and apply it to Lalonde's data, in which they choose "age" as the only conditional covariate and the response variable being the increment of earnings from 1975 to 1978. They cannot reject the null hypothesis of nonnegative treatment effect on the earnings growth.

The previous literature, however, has not addressed an important question. We shall depict this question by two pictures. In [Figure 1](#), it is obvious that  $Y_1$  stochastically dominates  $Y_0$  in both pictures, but significant differences can be seen between the two scenarios. For the left one, the 0.1 quantile improves more dramatically than the 0.9 quantile, as the distance between  $A$  and  $A'$  is greater than that between  $B$  and  $B'$ . In usual words, the gain of the 90% lower bound of the earnings growth is more than that of the 90% upper bound of the earnings growth after the treatment. "90% lower bound of the earnings growth" means the probability that the earnings growth is above the bound is 90%. This suggests that the treatment induces greater reduction in downside risk but less increase in the upside potential in the earnings growth. For the right picture the interpretation is

ACCEPTED MANUSCRIPT

just the opposite.

To see which type of stochastic dominance the NSW demonstration program belongs to, we apply the same data as [Delgado and Escanciano \(2013\)](#) for testing the hypothesis of positive quantile treatment effect for several quantile levels  $\tau$ . The data consist of 297 treatment group observations and 423 control group observations. The response variable  $Y_0$  ( $Y_1$ ) denotes the difference in earnings of control (treatment) group between 1978 (year of postintervention) and 1975 (year of preintervention). We first apply common statistical procedures to describe the distribution of these two variables. Figure 2 shows the unconditional densities and distribution function. The cross-validated bandwidth for  $\hat{f}_0(y)$  is 2.273 and 2.935 for  $\hat{f}_1(y)$ . The left figure of Figure 2 shows the unconditional densities of the income difference for treatment group and control group. The density of the treatment group has heavier tails while the density of the control group is more concentrated around zero. The right figure shows that the two unconditional distribution functions are very close on the left of the 50% percentile, and slight deviation appears when the two distributions are getting closer to 1. Table 4 shows that, though the differences are small, but the quantiles of the unconditional cdf of treatment group are mildly greater than that of the control group for each chosen  $\tau$ . The two-sample Kolmogorov-Smirnov and Cramér-von Mises tests, however, yield results shown in the Table 5 which cannot reject the null hypothesis that the empirical cdfs for the two groups are the same with confidence levels 1% or 5%.

Next we apply our test on quantile regression to evaluate the treatment effect. In order to compare with [Delgado and Escanciano \(2013\)](#), we first focus on the case of a one-dimensional covariate. The first covariate  $X_{1i}$  is the age. The second covariate  $X_{2i}$  is the number of years of schooling. The sample values of schooling years lie in the range of [3, 16] and age lies between [17, 55]. In order to avoid boundary effect and sparsity of the samples, we look at the ranges [7, 13] for schooling years and [19, 31] for age. We apply the bootstrap CC method for quantiles  $\tau = 0.1, 0.2, 0.3, 0.5, 0.7, 0.8$  and 0.9. We apply the quartic kernel. The cross-validated bandwidths are chosen in the same way as for conditional densities with the R package `np`. The resulting band-

widths are (2.2691, 2.5016) for the treatment group and (2.7204, 5.9408) for the control group. In particular, for smoothing the data of the treatment group, for  $\tau = 0.1$  and  $0.9$ , we enlarge the cross-validated bandwidths by a constant of 1.7; for  $\tau = 0.2, 0.3, 0.7, 0.8$ , the cross-validated bandwidths are enlarged by constant factor 1.3. These inflated bandwidths are used to handle violent roughness in extreme quantile levels. The bootstrap CCs are computed with 10,000 repetitions. The level of the test is  $\alpha = 5\%$ .

The results of the two quantile regressions with one-dimensional covariate, and their CCs for various quantile levels are presented in Figure 3 and 4. We observe that for all chosen quantile levels the quantile estimates associated to the treatment group lie above that of the control group when age is over certain levels, and particularly for  $\tau = 10\%, 50\%, 80\%$  and  $90\%$ , the quantile estimates for treatment group exceeds the upper CCs for the quantile estimates of the control group. On the other hand, at  $\tau = 10\%$ , the quantile estimates for the control group drop below the CC for treatment group for age greater than 27. Hence, the results here show a tendency that both the downside risk reduction and the upside potential enhancement of earnings growth are achieved, as the older individuals benefit the most from the treatment. Note that we observe a heterogeneous treatment effect in age and the weak dominance of the conditional quantiles of the treatment group with respect to those of the control group, i.e., (30) holds for the chosen quantile levels, which are in line with the findings of Delgado and Escanciano (2013).

We now turn to Figure 4, where the covariate is the years of schooling. The treatment effect is not significant for conditional quantiles at levels  $\tau = 10\%, 20\%$  and  $30\%$ . This suggests that the treatment does little to reduce the downside risk of the earnings growth for individuals with various degrees of education. Nonetheless, we constantly observe that the regression curves of the treatment group rise above that of the control group after a certain level of the years of schooling for quantile levels  $\tau = 50\%, 70\%, 80\%$  and  $90\%$ . Notice that for  $\tau = 50\%$  and  $80\%$  the regression curves associated to the treatment group reach the upper boundary of the CC of the control group. This suggests that the treatment effect tends to raise the upside potential of the earnings growth, in

# ACCEPTED MANUSCRIPT

particular for those individuals who spent more years in the school. It is worth noting that we also see a heterogeneous treatment effect in schooling years, although the heterogeneity in education is less strong than the heterogeneity in age.

The previous regression analyses separately conditioning on covariates age and schooling years only give a limited view on the performance of the program, we now proceed to the analysis conditioning on the two covariates  $(X_{1i}, X_{2i})$  jointly. The estimation settings are similar to the case of univariate covariate. Figure 5 shows the quantile regression CCs. From a first glance of the pictures, the  $\tau$ -quantile CC of the treatment group and that of the control group overlap extensively for all  $\tau$ . We could not find sufficient evidence to reject the null hypothesis that the conditional distribution of treatment group and control group are equivalent.

The second observation obtained from comparing subfigures in Figure 6, we find that the treatment has larger impact in raising the upper bound of the earnings growth than improving the lower bound. For lower quantile levels  $\tau = 10\%, 20\%$  and  $30\%$  the solid surfaces uniformly lie inside the CC of the control group, while for  $\tau = 50\%, 70\%, 80\%$  and  $90\%$ , we see several positive exceedances over the upper boundary of the CC of the control group. Hence, the program tends to do better at raising the upper bound of the earnings growth but does worse at improving the lower bound of the earnings growth. In other words, the program tends to increase the potential for high earnings growth but does little in reducing the risk of negative earnings growth.

Our last conclusion comes from inspecting the shape of the surfaces: conditioning on different levels of years of schooling (age), the treatment effect is heterogeneous in age (years of schooling). The most interesting cases occur when conditioning on high age and high years of schooling. Indeed, when considering the cases of  $\tau = 80\%$  and  $90\%$ , when conditioning on the years of schooling at 12 (corresponding to finishing the high school), the earnings increment of the treatment group rises above the upper boundary of the CC of the control group. This suggests that the individuals who are older and have more years of schooling tend to benefit more from the treatment.

ACCEPTED MANUSCRIPT



## Supplementary Materials

Section A contains the detailed proofs of Theorems 2.1, 2.3, 3.1 and Lemmas 2.6 and 3.2, as well as intermediate results. Section B contains some results obtained by other authors, which we use in our study. We incorporate them here for the sake of completeness.

## Acknowledgements

We thank the Co-Editor, Associate Editor, and two anonymous referees for their insightful comments and suggestions which have helped improve this article. The financial support from the Deutsche Forschungsgemeinschaft (DFG) via SFB 649 "Economic Risk" (Teilprojekt B1) and SFB 823 "Statistical modeling of nonlinear dynamic processes" (Teilprojekt A1, C1, C4) is gratefully acknowledged. Shih-Kang Chao is partially supported by the Einstein Foundation Berlin via the Berlin Doctoral Program in Economics and Management Science (BDPEMS).

## References

- Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *The Annals of Statistics* **1**(6): 1071–1095.
- Bissantz, N., Dümbgen, L., Holzmann, H. and Munk, A. (2007). Nonparametric confidence bands in deconvolution density estimation, *Journal of the Royal Statistical Society: Series B* **69**(3): 483–506.
- Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives, *The Annals of Statistics* **31**(6): 1852–1884.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American Statistical Association* **94**(448): 1053–1062.
- Delgado, M. A. and Escanciano, J. C. (2013). Conditional stochastic dominance testing, *Journal of Business & Economic Statistics* **31**(1): 16–28.
- Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case, *The Annals of Statistics* **2**(2): 267–277.
- Fan, Y. and Liu, R. (2013). A direct approach to inference in nonparametric and semiparametric quantile regression models, Preprint.
- Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression, *Biometrika* **98**(4): 995–999.
- Giné, E. and Nickl, R. (2010). Confidence bands in density estimation, *The Annals of Statistics* **38**(2): 1122–1170.

# ACCEPTED MANUSCRIPT

- Guerre, E. and Sabbah, C. (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function, *Econometric Theory* **28**(1): 87–129.
- Guo, M. and Härdle, W. (2012). Simultaneous confidence bands for expectile functions, *AStA Advances in Statistical Analysis* **96**(4): 517–541.
- Hall, P. (1979). On the rate of convergence of normal extremes, *Journal of Applied Probability* **16**(2): 433–439.
- Hall, P. (1991). On convergence rates of suprema, *Probability Theory and Related Fields* **89**(4): 447–455.
- Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density, *The Annals of Statistics* **20**(2): 675–694.
- Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions, *The Annals of Statistics* **41**(4): 1892–1921.
- Härdle, W. (1989). Asymptotic maximal deviation of  $M$ -smoothers, *Journal of Multivariate Analysis* **29**(2): 163–179.
- Härdle, W. and Song, S. (2010). Confidence bands in quantile regression, *Econometric Theory* **26**(4): 1180–1200.
- Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates, *Journal of Multivariate Analysis* **12**(3): 402–414.
- Jones, M. C. (1994). Expectiles and  $M$ -quantiles are quantiles, *Statistics & Probability Letters* **20**(2): 149–153.
- Kiwitt, S. and Neumeier, N. (2012). Estimating the conditional error distribution in non-parametric

ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

regression, *Scandinavian Journal of Statistics* **39**(2): 259–281.

Koenker, R. (2005). *Quantile Regression*, Econometric Society Monographs, Cambridge University Press, New York.

Kong, E., Linton, O. and Xia, Y. (2010). Uniform Bahadur representation for local polynomial estimates of  $M$ -regression and its application to the additive model, *Econometric Theory* **26**(5): 1529–1564.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data, *The American Economic Review* **76**(4): 604–620.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Models Based on Ranks*, Springer, San Francisco, CA.

Li, Q., Lin, J. and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions, *Journal of Business & Economic Statistics* **31**(1): 57–65.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton university press, New Jersey.

Lounici, K. and Nickl, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators, *The Annals of Statistics* **39**(1): 201–231.

Mammen, E., Van Keilegom, I. and Yu, K. (2013). Expansion for moments of regression quantiles with applications to nonparametric testing, *ArXiv e-prints* .

Mojirsheibani, M. (2012). A weighted bootstrap approximation of the maximal deviation of kernel density estimates over general compact sets, *Journal of Multivariate Analysis* **112**: 230–241.

Muhsal, B. and Neumeier, N. (2010). A note on residual-based empirical likelihood kernel density

ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

estimator, *Electronic Journal of Statistics* **4**: 1386–1401.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing, *Econometrica* **55**(4): 819–847.

Proksch, K., Bissantz, N. and Dette, H. (2015). Confidence bands for multivariate and time dependent inverse regression models, *Bernoulli* **21**(1): 144–175.

Rosenblatt, M. (1976). On the maximal deviation of  $k$ -dimensional density estimates, *The Annals of Probability* **4**(6): 1009–1015.

Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables, *Doklady Akad. Nauk SSSR* **74**: 189–191.

Song, S., Ritov, Y. and Härdle, W. (2012). Partial linear quantile regression and bootstrap confidence bands, *Journal of Multivariate Analysis* **107**: 244–262.

Yu, K. and Jones, M. C. (1998). Local linear quantile regression, *Journal of the American Statistical Association* **93**(441): 228–237.

## Appendices

### A. Assumptions

(A1)  $K$  is of order  $s - 1$  (see (A3)), has bounded support  $[-A, A]^d$  for  $A > 0$  a positive real scalar, and is continuously differentiable up to order  $d$  with bounded derivatives, i.e.  $\partial^\alpha K \in L^1(\mathbb{R}^d)$  exists and is continuous for all multi-indices  $\alpha \in \{0, 1\}^d$

(A2) Let  $a_n$  be an increasing sequence,  $a_n \rightarrow \infty$  as  $n \rightarrow \infty$ , and the marginal density  $f_Y$  be such

ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

that

$$(\log n)h^{-3d} \int_{|y|>a_n} f_Y(y)dy = O(1) \quad (31)$$

and

$$(\log n)h^{-d} \int_{|y|>a_n} f_{Y|X}(y|\mathbf{x})dy = O(1), \text{ for all } \mathbf{x} \in \mathcal{D}$$

as  $n \rightarrow \infty$  hold.

- (A3) The function  $\theta_0(\mathbf{x})$  is continuously differentiable and is in Hölder class with order  $s > d$ .
- (A4)  $f_X(\mathbf{x})$  is bounded, continuously differentiable and its gradient is uniformly bounded. Moreover,  $\inf_{\mathbf{x} \in \mathcal{D}} f_X(\mathbf{x}) > 0$ .
- (A5) The joint probability density function  $f(y, \mathbf{u})$  is bounded, positive and continuously differentiable up to  $s$ th order (needed for Rosenblatt transform). The conditional density  $f_{Y|X}(y|\mathbf{x})$  exists and is bounded and continuously differentiable with respect to  $\mathbf{x}$ . Moreover,  $\inf_{\mathbf{x} \in \mathcal{D}} f_{Y|X}(\theta_0(\mathbf{x})|\mathbf{x}) > 0$ .
- (A6)  $h$  satisfies  $\sqrt{nh^d}h^s \sqrt{\log n} \rightarrow 0$  (undersmoothing), and  $nh^{3d}(\log n)^{-2} \rightarrow \infty$ .
- (EA2)  $\sup_{\mathbf{x} \in \mathcal{D}} \left| \int v^{b_1} f_{\varepsilon|X}(v|\mathbf{x})dv \right| < \infty$ , for some  $b_1 > 0$ .
- (B1)  $L$  is a Lipschitz, bounded, symmetric kernel.  $G$  is Lipschitz continuous cdf, and  $g$  is the derivative of  $G$  and is also a density, which is Lipschitz continuous, bounded, symmetric and five times continuously differentiable kernel.
- (B2)  $F_{\varepsilon|X}(v|\mathbf{x})$  is in  $s' + 1$  order Hölder class with respect to  $v$  and continuous in  $\mathbf{x}$ ,  $s' > \max\{2, d\}$ .  $f_X(\mathbf{x})$  is in second order Hölder class with respect to  $\mathbf{x}$  and  $v$ .  $E[\psi^2(\varepsilon_i)|\mathbf{x}]$  is second order continuously differentiable with respect to  $\mathbf{x} \in \mathcal{D}$ .
- (B3)  $nh_0\bar{h}^d \rightarrow \infty$ ,  $h_0, \bar{h} = O(n^{-\nu})$ , where  $\nu > 0$ .

ACCEPTED MANUSCRIPT

# ACCEPTED MANUSCRIPT

(C1) There exist an increasing sequence  $c_n, c_n \rightarrow \infty$  as  $n \rightarrow \infty$  such that

$$(\log n)^3 (nh^{6d})^{-1} \int_{|v| > c_n/2} f_\varepsilon(v) dv = O(1), \quad (32)$$

as  $n \rightarrow \infty$ .

(EC1)  $\sup_{\mathbf{x} \in \mathcal{D}} \left| \int v^b f_{\varepsilon|\mathbf{X}}(v|\mathbf{x}) dv \right| < \infty$ , for some  $b > 0$ .

The assumptions (A1)-(A5) are assumptions frequently seen in the papers of confidence corridors, such as [Härdle \(1989\)](#), [Härdle and Song \(2010\)](#) and [Guo and Härdle \(2012\)](#). (EA2) and (EC1) essentially give the uniform bound on the 2nd order tail variation, which is crucial in the sequence of approximations for expectile regression. (B1)-(B3) are similar to the assumptions listed in chapter 6.1 of [Li and Racine \(2007\)](#). (A6) characterizes the two conflicting conditions: the undersmoothing of our estimator and the convergence of the strong approximation. To make the condition hold for large  $d$ , sometimes we need large  $s$ , which is the smoothness of the true function. (C1) and (EC1) are relevant to the theory of bootstrap, where we need bounds on the tail probability and 2nd order variation.

Method	$n$	Homogeneous			Heterogeneous			
		$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$	$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$	
$\sigma_0 = 0.2$								
Asympt.	100	.000(0.366)	.109(0.720)	.104(0.718)	.000(0.403)	.120(0.739)	.122(0.744)	
	300	.000(0.304)	.130(0.518)	.133(0.519)	.002(0.349)	.136(0.535)	.153(0.537)	
	500	.000(0.262)	.117(0.437)	.142(0.437)	.008(0.296)	.156(0.450)	.138(0.450)	
	$\sigma_0 = 0.5$							
	100	.070(0.890)	.269(1.155)	.281(1.155)	.078(0.932)	.300(1.193)	.302(1.192)	
	300	.276(0.735)	.369(0.837)	.361(0.835)	.325(0.782)	.380(0.876)	.394(0.877)	
	500	.364(0.636)	.392(0.711)	.412(0.712)	.381(0.669)	.418(0.743)	.417(0.742)	
	$\sigma_0 = 0.7$							
	100	.160(1.260)	.381(1.522)	.373(1.519)	.155(1.295)	.364(1.561)	.373(1.566)	
300	.438(1.026)	.450(1.109)	.448(1.110)	.481(1.073)	.457(1.155)	.472(1.152)		
500	.533(0.888)	.470(0.950)	.480(0.949)	.564(0.924)	.490(0.984)	.502(0.986)		
$\sigma_0 = 0.2$								
Bootst.	100	.325(0.676)	.784(0.954)	.783(0.954)	.409(0.717)	.779(0.983)	.778(0.985)	
	300	.442(0.457)	.896(0.609)	.894(0.610)	.580(0.504)	.929(0.650)	.922(0.649)	
	500	.743(0.411)	.922(0.502)	.921(0.502)	.839(0.451)	.950(0.535)	.952(0.536)	
	$\sigma_0 = 0.5$							
	100	.929(1.341)	.804(1.591)	.818(1.589)	.938(1.387)	.799(1.645)	.773(1.640)	
	300	.950(0.920)	.918(1.093)	.923(1.091)	.958(0.973)	.919(1.155)	.923(1.153)	
	500	.988(0.861)	.968(0.943)	.962(0.942)	.990(0.902)	.962(0.986)	.969(0.987)	
	$\sigma_0 = 0.7$							
	100	.976(1.811)	.817(2.112)	.808(2.116)	.981(1.866)	.826(2.178)	.809(2.176)	
300	.986(1.253)	.919(1.478)	.934(1.474)	.983(1.308)	.930(1.537)	.920(1.535)		
500	.996(1.181)	.973(1.280)	.968(1.278)	.997(1.225)	.969(1.325)	.962(1.325)		

Table 1: *Nonparametric quantile model coverage probabilities. The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic quantile regression CC and bootstrap method corresponds to quantile regression bootstrap CC.*



Method	$n$	Homogeneous			Heterogeneous			
		$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$	$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$	
$\sigma_0 = 0.2$								
Asympt.	100	.000(0.428)	.000(0.333)	.000(0.333)	.000(0.463)	.000(0.362)	.000(0.361)	
	300	.049(0.341)	.000(0.273)	.000(0.273)	.079(0.389)	.001(0.316)	.002(0.316)	
	500	.168(0.297)	.000(0.243)	.000(0.243)	.238(0.336)	.003(0.278)	.002(0.278)	
	$\sigma_0 = 0.5$							
	100	.007(0.953)	.000(0.776)	.000(0.781)	.007(0.997)	.000(0.818)	.000(0.818)	
	300	.341(0.814)	.019(0.708)	.017(0.709)	.355(0.862)	.017(0.755)	.018(0.754)	
	500	.647(0.721)	.067(0.645)	.065(0.647)	.654(0.759)	.061(0.684)	.068(0.684)	
	$\sigma_0 = 0.7$							
	100	.012(1.324)	.000(1.107)	.000(1.107)	.010(1.367)	.000(1.145)	.000(1.145)	
300	.445(1.134)	.021(1.013)	.013(1.016)	.445(1.182)	.017(1.062)	.016(1.060)		
500	.730(1.006)	.062(0.928)	.078(0.929)	.728(1.045)	.068(0.966)	.066(0.968)		
$\sigma_0 = 0.2$								
Bootst.	100	.686(2.191)	.781(2.608)	.787(2.546)	.706(2.513)	.810(2.986)	.801(2.943)	
	300	.762(0.584)	.860(0.716)	.876(0.722)	.788(0.654)	.877(0.807)	.887(0.805)	
	500	.771(0.430)	.870(0.533)	.875(0.531)	.825(0.516)	.907(0.609)	.904(0.615)	
	$\sigma_0 = 0.5$							
	100	.886(5.666)	.906(6.425)	.915(6.722)	.899(5.882)	.927(6.667)	.913(6.571)	
	300	.956(1.508)	.958(1.847)	.967(1.913)	.965(1.512)	.962(1.866)	.969(1.877)	
	500	.968(1.063)	.972(1.322)	.972(1.332)	.972(1.115)	.971(1.397)	.974(1.391)	
	$\sigma_0 = 0.7$							
	100	.913(7.629)	.922(8.846)	.935(8.643)	.929(8.039)	.935(9.057)	.932(9.152)	
300	.969(2.095)	.969(2.589)	.971(2.612)	.974(2.061)	.972(2.566)	.979(2.604)		
500	.978(1.525)	.976(1.881)	.967(1.937)	.981(1.654)	.978(1.979)	.974(2.089)		

Table 2: Nonparametric expectile model coverage probability. The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic expectile regression CC and bootstrap method corresponds to expectile regression bootstrap CC.

$n$	Homogeneous			Heterogeneous		
	$\xi = 0.005$	$\xi = 0.05$	$\xi = 0.1$	$\xi = 0.005$	$\xi = 0.05$	$\xi = 0.1$
$\sigma_0 = 0.2$						
100	.693(3.027)	.529(1.740)	.319(1.040)	.680(3.452)	.546(2.051)	.332(1.224)
300	.891(0.580)	.748(0.365)	.642(0.323)	.907(0.667)	.798(0.414)	.698(0.364)
500	.886(0.335)	.770(0.265)	.678(0.244)	.896(0.379)	.789(0.298)	.699(0.274)
$\sigma_0 = 0.5$						
100	.720(7.264)	.611(4.489)	.394(2.686)	.729(7.594)	.616(4.676)	.414(2.829)
300	.945(1.423)	.849(0.859)	.755(0.746)	.940(1.511)	.854(0.912)	.760(0.791)
500	.944(0.795)	.846(0.600)	.750(0.548)	.937(0.833)	.839(0.632)	.751(0.577)
$\sigma_0 = 0.7$						
100	.730(10.183)	.634(6.411)	.430(3.853)	.752(10.657)	.658(6.577)	.441(3.923)
300	.936(1.995)	.854(1.197)	.751(1.037)	.951(2.091)	.875(1.256)	.772(1.086)
500	.933(1.098)	.854(0.831)	.774(0.758)	.938(1.145)	.853(0.865)	.770(0.789)

Table 3: Proportion in 2000 iteration that the coverage of  $\geq 95\%$  grid points for nonparametric mean model, using the bootstrap method of [Hall and Horowitz \(2013\)](#). The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor.

$\tau(\%)$	10	20	30	50	70	80	90
Treatment	-4.38	-1.55	0.00	1.40	5.48	8.50	11.15
Control	-4.91	-1.73	-0.17	0.74	4.44	7.16	10.56

Table 4: The unconditional sample quantiles of treatment and control groups.

Type of test	Statistics	$p$ -value
Kolmogorov-Smirnov	0.0686	0.3835
Cramér-von Mises	0.2236	0.7739

Table 5: The two sample empirical cdf tests results for treatment and control groups.

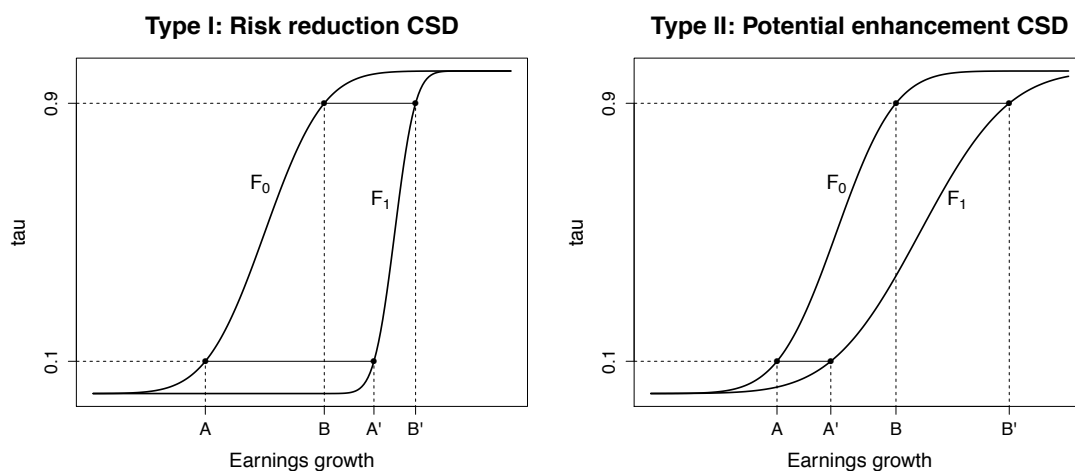


Figure 1: The illustrations for the two possible types of stochastic dominance. In the left figure, the 0.1 quantile improves (downside risk reduction) more dramatically than the 0.9 quantile (upside potential increase), as the distance between  $A$  and  $A'$  is greater than that between  $B$  and  $B'$ . For the right picture the interpretation is just the opposite.

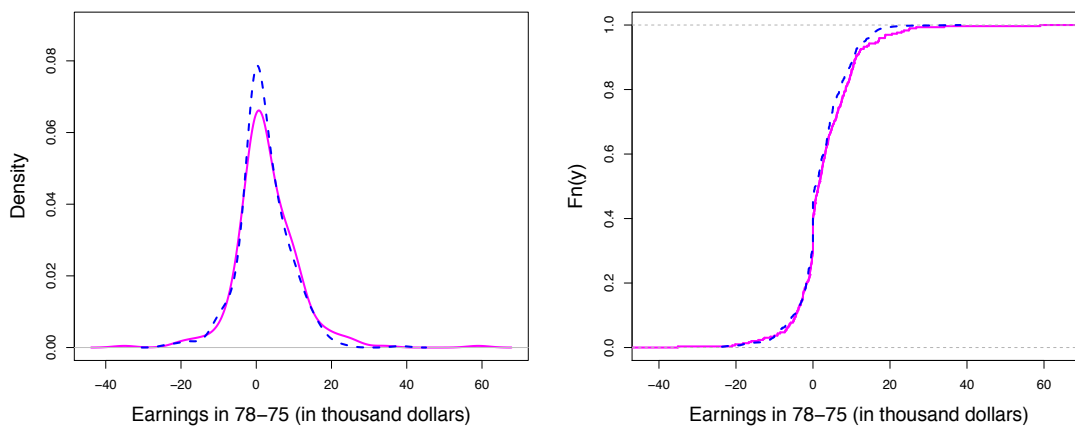


Figure 2: Unconditional empirical density function (left) and distribution function (right) of the difference of earnings from 1975 to 1978. The dashed line is associated with the control group and the solid line is associated with the treatment group.

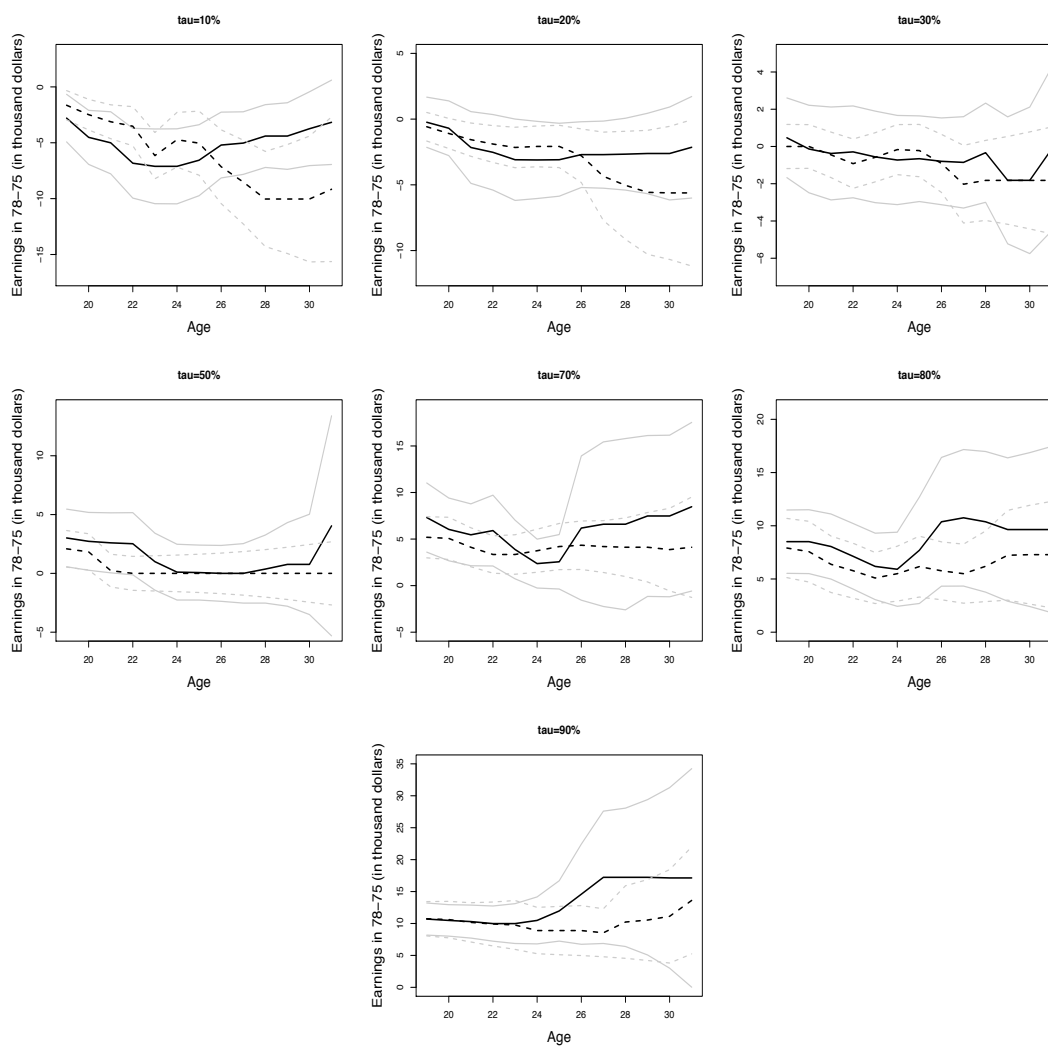


Figure 3: Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of age. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC.

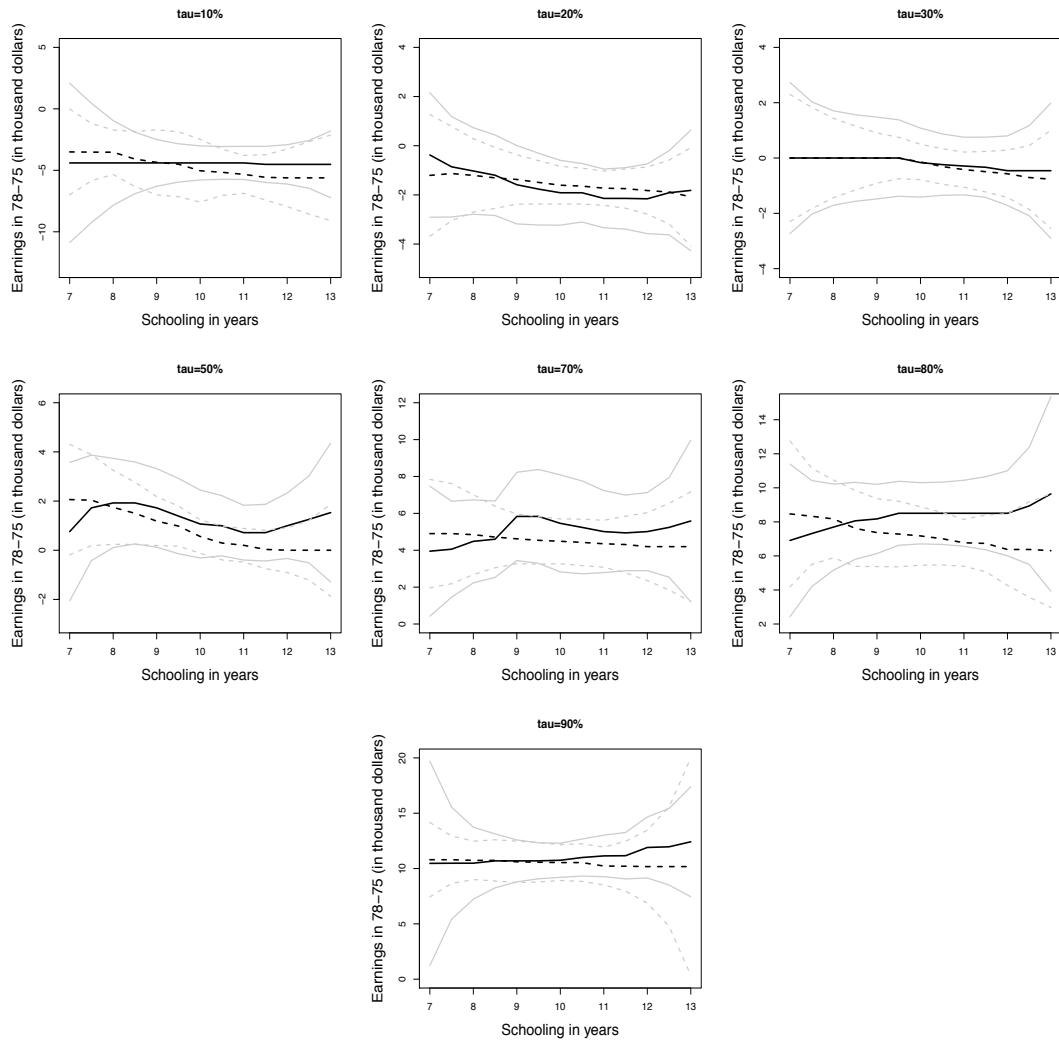


Figure 4: Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of years of schooling. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC.

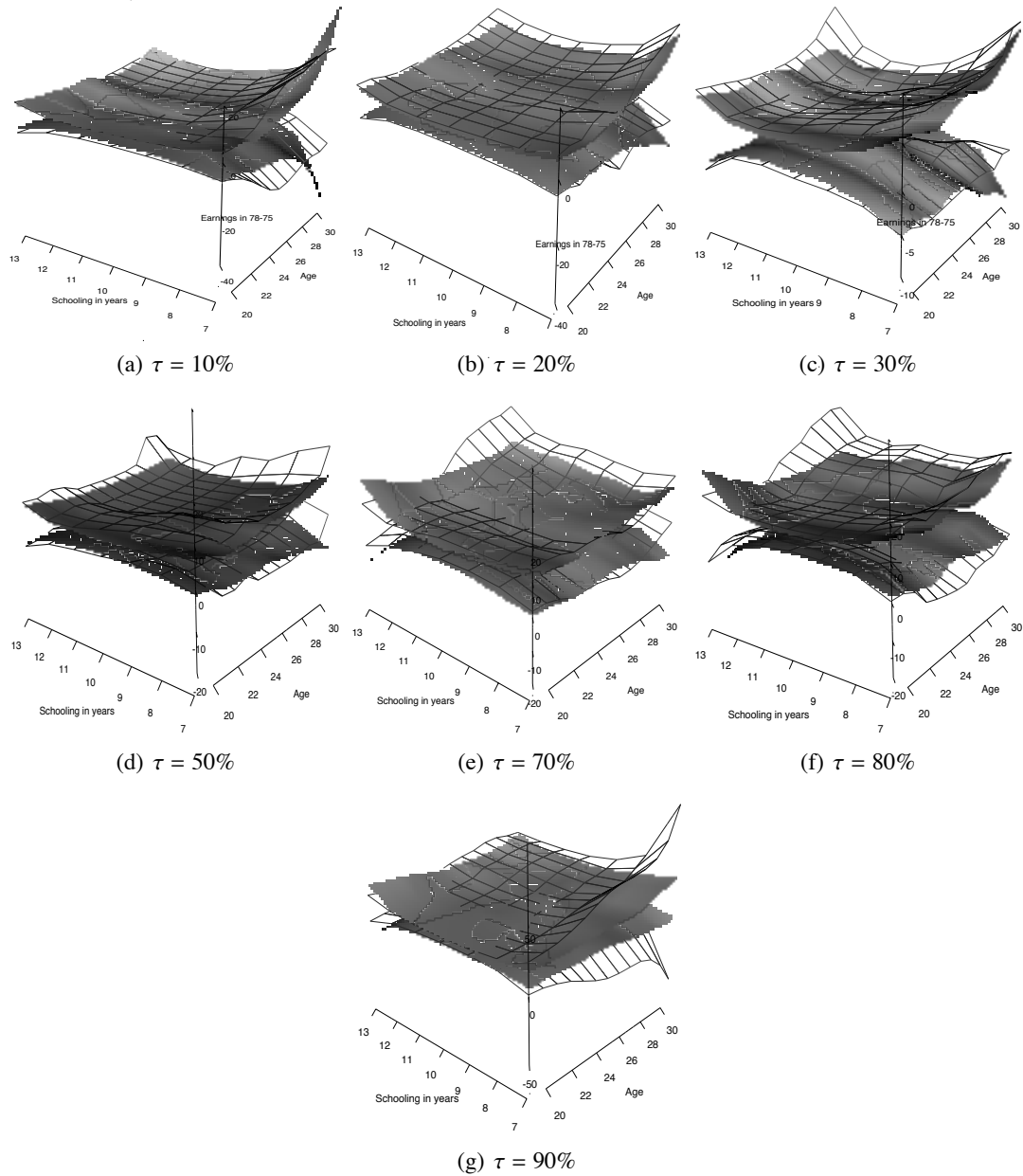


Figure 5: The CCs for the treatment group and the control group. The net surface corresponds to the control group quantile CC and the solid surface corresponds to the treatment group quantile CC.

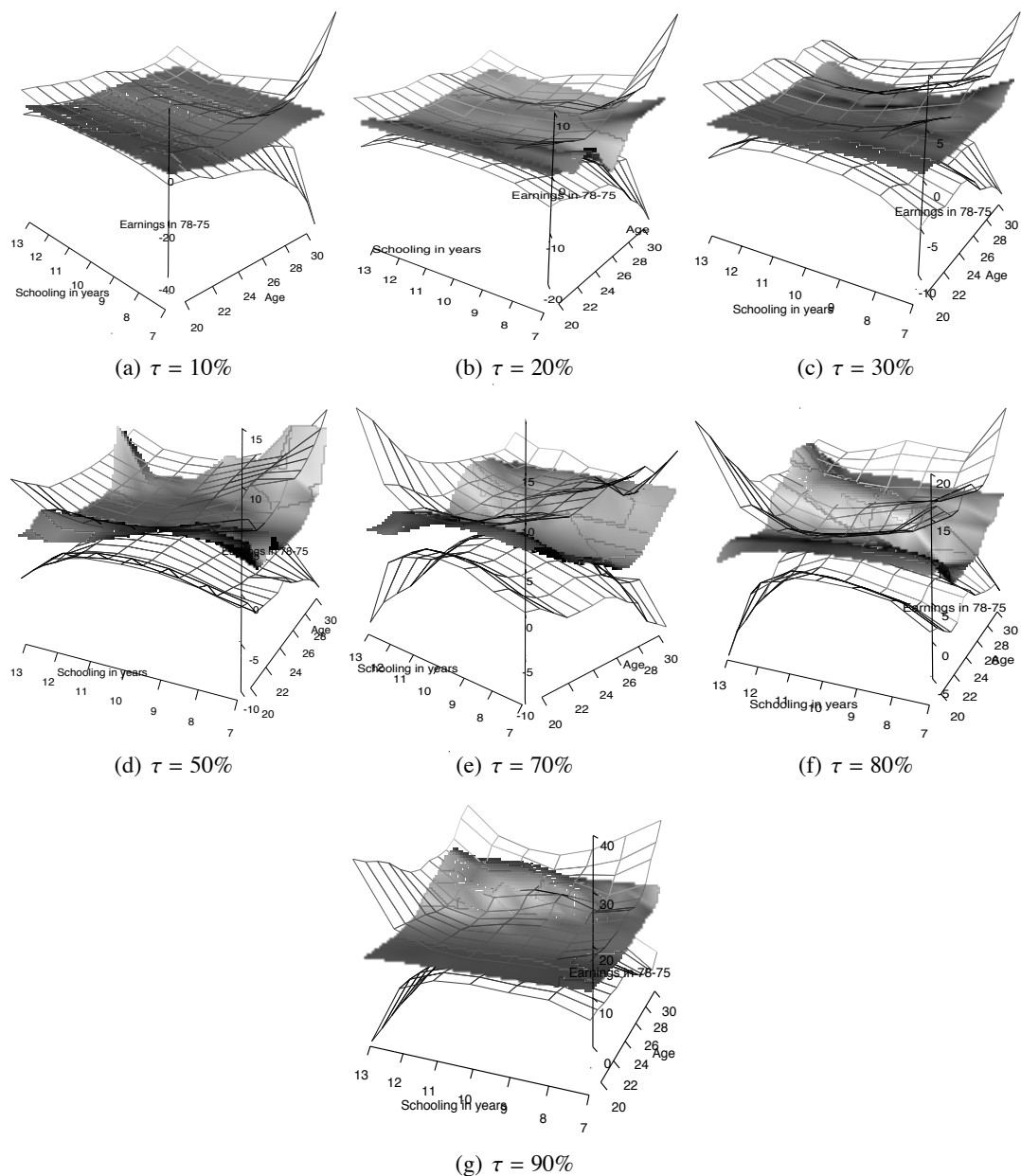


Figure 6: The conditional quantiles (solid surfaces) for the treatment group and the CCs (net surfaces) for the control group.

## LOCAL ADAPTIVE MULTIPLICATIVE ERROR MODELS FOR HIGH-FREQUENCY FORECASTS

WOLFGANG K. HÄRDLE<sup>a,b</sup>, NIKOLAUS HAUTSCH<sup>c,d</sup> AND ANDRIJA MIHOCI<sup>a,\*</sup>

<sup>a</sup> CASE, Humboldt-Universität zu Berlin, Germany

<sup>b</sup> School of Business, Singapore Management University, Singapore

<sup>c</sup> Department of Statistics and Operations Research, University of Vienna, Austria

<sup>d</sup> Center for Financial Studies (CFS), Frankfurt, Germany

### SUMMARY

We propose a local adaptive multiplicative error model (MEM) accommodating time-varying parameters. MEM parameters are adaptively estimated based on a sequential testing procedure. A data-driven optimal length of local windows is selected, yielding adaptive forecasts at each point in time. Analysing 1-minute cumulative trading volumes of five large NASDAQ stocks in 2008, we show that local windows of approximately 3 to 4 hours are reasonable to capture parameter variations while balancing modelling bias and estimation (in)efficiency. In forecasting, the proposed adaptive approach significantly outperforms a MEM where local estimation windows are fixed on an ad hoc basis. Copyright © 2014 John Wiley & Sons, Ltd.

### 1. INTRODUCTION

Recent research in econometrics and statistics shows that modelling and forecasting of high-frequency financial data is a challenging task. Researchers strive to understand the dynamics of processes when all single events are recorded while accounting for external shocks as well as structural shifts on financial markets. The fact that high-frequency dynamics are not stable over time but are subject to regime shifts is hard to capture by standard time series models. This is particularly true whenever it is unclear where the time-varying nature of the data actually comes from and how many underlying regimes there might be.

This paper addresses the phenomenon of time-varying dynamics in high-frequency data, such as (cumulative) trading volumes, trade durations, market depth or bid–ask spreads. The aim is to adapt and to implement a local parametric framework for multiplicative error processes and to illustrate its usefulness when it comes to out-of-sample forecasting under possibly non-stable market conditions. We propose a flexible statistical approach allowing adaptive selection of a data window over which a local constant-parameter model is estimated and forecasts are computed. The procedure requires (re-)estimating models on windows of evolving lengths and yields an optimal local estimation window. As a result, we provide insights into the time-varying nature of parameters and of local window lengths.

The so-called multiplicative error model (MEM), introduced by Engle (2002), serves as a workhorse for the modelling of positive-valued, serially dependent high-frequency data. It is successfully applied to financial duration data, where it was originally introduced by Engle and Russell (1998) in the context of an autoregressive conditional duration (ACD) model. Likewise, it is applied to model intra-day trading volumes, see, among others, Manganelli (2005); Brownlees *et al.* (2011); Hautsch *et al.* (2014). MEM parameters are typically estimated over long estimation windows in order to increase estimation efficiency. However, empirical evidence makes parameter constancy in high-frequency models over long time intervals questionable. Possible structural breaks in MEM parameters have been addressed, for instance, by Zhang *et al.* (2001), who identify regime shifts in trade durations and suggest a thresh-

---

\* Correspondence to: Andrija Mihoci, CASE—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany. E-mail: mihociax@cms.hu-berlin.de



old ACD (TACD) specification in the spirit of threshold ARMA models, see, for example, Tong (1990). To capture smooth transitions of parameters between different states, Meitz and Teräsvirta (2006) propose a smooth transition ACD (STACD) model. Whereas in STACD models parameter transitions are driven by observable variables, Hujer *et al.* (2002) allow for an underlying (hidden) Markov process governing the underlying state of the process.

Regime-switching MEM approaches have the advantage of allowing for changing parameters on possibly high frequencies (in the extreme case from observation to observation) but require imposition of a priori structures on the form of the transition, the number of underlying regimes and (in the case of transition models) on the type of the transition variable. Moreover, beyond short-term fluctuations, parameters might also reveal transitions on lower frequencies governed by the general (unobservable) state of the market. Such regime changes might be captured by adaptively estimating a MEM based on a window of varying length and thus providing updated parameter estimates at each point in time. The main challenge of the latter approach, however, is the selection of the estimation window. From a theoretical perspective, the length of the window should, on the one hand, be maximal to increase the precision of parameter estimates and, on the other, sufficiently short to capture structural changes. This observation is also reflected in the well-known result that aggregations over structural breaks (caused by too long estimation windows) can induce spurious persistence and long range dependence.

This paper suggests a data-driven length of (local) estimation windows. The key idea is to implement a sequential testing procedure to search for the longest time interval with given right end for which constancy of model parameters cannot be rejected. This mechanism is carried out by re-estimating (local) MEMs based on data windows of increasing lengths and sequentially testing for a change in parameter estimates. By controlling the risk of false alarm, the algorithm selects the longest possible window for which parameter constancy cannot be rejected at a given significance level. Based on this data interval, forecasts for the next period are computed. By repeating these steps in every period, variations in parameters are thus automatically captured.

The proposed framework builds on the *local parametric approach* (LPA) originally proposed by Spokoiny (1998). The presented methodology has been gradually introduced into the time series literature; see, for example, Mercurio and Spokoiny (2004) for an application to daily exchange rates and Čížek *et al.* (2009) for an adaptation of the approach to generalized autoregressive conditional heteroskedasticity (GARCH) models. In realized volatility analysis, LPA has been applied by Chen *et al.* (2010) to daily stock index returns.

The contributions of this paper are to introduce local adaptive calibration techniques into the class of multiplicative error models, to provide valuable empirical insights into the (non-)homogeneity of high-frequency processes and to show the usefulness of the approach in the context of out-of-sample forecasting. Though we specifically focus on 1-minute cumulative trading volumes of five highly liquid stocks traded at NASDAQ, our findings may be carried over to other high-frequency series, as the stochastic properties of high-frequency volumes are quite similar to those of, e.g., trade counts, squared midquote returns, market depth or bid–ask spreads.

We aim at answering the following research questions: (i) How strong is the variation of MEM parameters over time? (ii) What are typical interval lengths of parameter homogeneity implied by the adaptive approach? (iii) How good are out-of-sample short-term forecasts compared to adaptive procedures where the length of the estimation windows is fixed on an ad hoc basis?

Implementing the proposed framework requires re-estimating and re-evaluating the model based on rolling windows of different lengths which are moved forward from minute to minute. This proceeding yields extensive insights into the time-varying nature of high-frequency trading processes. Based on NASDAQ trading volumes, we show that parameter estimates and estimation quality clearly change over time and provide researchers valuable rule of thumbs for the choice of local intervals. In particular, we show that, on average, precise adaptive estimates require local estimation windows of approximately 3 to 4 hours. Moreover, it turns out that the proposed adaptive method yields

significantly better short-term forecasts than competing approaches using fixed-length rolling windows of comparable sizes. Hence it is not only important to use local windows but also to adaptively adjust their length in accordance with prevailing (market) conditions. This is particularly true in periods of market distress where forecasts utilizing too much historical information perform clearly worse.

The remainder of the paper is structured as follows. After the data description in Section 2, the multiplicative error model and the local parametric approach are introduced in Sections 3 and 4, respectively. Empirical results on forecasts of trading volumes are provided in Section 5. Section 6 concludes.

## 2. DATA

We use transaction data of five large companies traded at NASDAQ—Apple Inc. (AAPL), Cisco Systems, Inc. (CSCO), Intel Corporation (INTC), Microsoft Corporation (MSFT) and Oracle Corporation (ORCL)—which account for approximately one third of the market capitalization within the technology sector. Our variable of interest is the 1-minute cumulative trading volume covering the period from 2 January to 31 December 2008. To remove effects due to market opening, the first 30 minutes of each trading session are discarded. Hence, at each trading day, we analyse data from 10:00 to 16:00.

Descriptive statistics (not shown in the paper) indicate right-skewed distributions, whereas the Ljung–Box test statistics show a strong serial dependence as the null hypothesis of no autocorrelation (among the first 10 lags) is clearly rejected. Autocorrelation functions indicate that high-frequency volumes are strongly and persistently clustered over time.

Denote the 1-minute cumulative trading volume at time point  $i$  by  $\check{y}_i$ . Assuming a multiplicative impact of intra-day periodicity effects, we compute seasonally adjusted volumes by

$$y_i = \check{y}_i s_i^{-1} \quad (1)$$

with  $s_i$  representing the intra-day periodicity component at time point  $i$ . Seasonality components are typically assumed to be constant over time. However, to capture slowly moving ('long-term') components in the spirit of Engle and Rangel (2008), we estimate the periodicity effects on the basis of 30-day rolling windows. Alternatively, seasonal effects could be captured directly within the local adaptive framework presented below. As our focus is on (pure stochastic) short-term variations in parameters rather than on deterministic periodicity effects, we decide to remove the former beforehand. This leaves us with non-homogeneity in the processes, which is not straightforwardly taken into account and allows us evaluating the potential of a local parametric approach even more convincingly. The intra-day component  $s_i$  is specified via a flexible Fourier series approximation as proposed by Gallant (1981):

$$s_i = \delta \cdot \bar{t} + \sum_{m=1}^M \{ \delta_{c,m} \cos(\bar{t} \cdot 2\pi m) + \delta_{s,m} \sin(\bar{t} \cdot 2\pi m) \} \quad (2)$$

Here,  $\delta$ ,  $\delta_{c,m}$  and  $\delta_{s,m}$  are coefficients to be estimated, and  $\bar{t} \in (0, 1]$  denotes a normalized intra-day time trend defined as the number of minutes from opening until  $i$  divided by the length of the trading day, i.e.  $\bar{t} = i/360$ . The order  $M$  is selected according to the Bayes information criterion (BIC) within each 30-day rolling window. To avoid forward-looking biases, the periodicity component is estimated using previous data only. The sample of seasonally standardized cumulative 1-minute trading volumes thus covers the period from 14 February to 31 December 2008. The estimated daily seasonality factors change mildly in their level, reflecting slight long-term movements.

Figure 1 displays the intra-day periodicity components associated with the lowest and largest monthly volumes, respectively, observed through the sample period. We observe the well-known

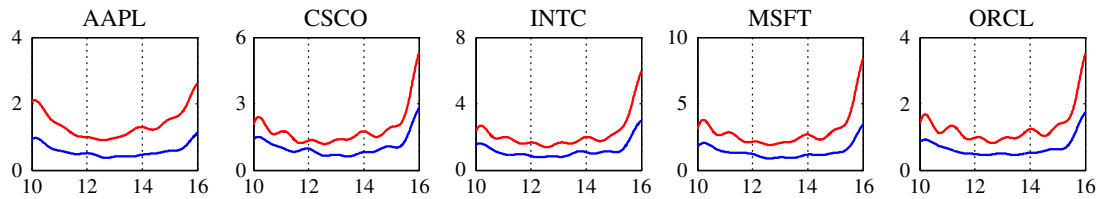


Figure 1. Estimated intra-day periodicity components for cumulative one-minute trading volumes (in units of 100,000 and plotted against the time of the day) of selected companies at NASDAQ on 2 September (lower, lowest 30-day trading volume) and 30 October 2008 (upper, highest 30-day volume)

(asymmetric) U-shaped intra-day pattern with high volumes at the opening and before market closure. Particularly before closure, it is evident that traders intend to close their positions, creating high market activity.

### 3. LOCAL MULTIPLICATIVE ERROR MODELS

The multiplicative error model (MEM), as discussed by Engle (2002), has become a workhorse for analysing and forecasting positive valued financial time series, such as trading volumes, trade durations, bid–ask spreads, price volatilities, market depth or trading costs. The idea of a multiplicative error structure originates from the structure of the autoregressive conditional heteroskedasticity (ARCH) model introduced by Engle (1982). In high-frequency financial data analysis, a MEM was first proposed by Engle and Russell (1998) to model the dynamic behaviour of the time between trades and has been referred to as autoregressive conditional duration (ACD) model. The ACD model is thus a special type of MEM applied to financial durations. During the remainder of the paper, we use both labels as synonyms. For a comprehensive literature overview, see Hautsch (2012).

#### 3.1. Model Structure

The principle of a MEM is to model a non-negative valued process  $y = \{y_i\}_{i=1}^n$ , e.g., the trading volume time series in our context, in terms of the product of its conditional mean process  $\mu_i$  and a positive valued error term  $\varepsilon_i$  with unit mean:

$$y_i = \mu_i \varepsilon_i, \quad E[\varepsilon_i | \mathcal{F}_{i-1}] = 1 \quad (3)$$

conditional on the information set  $\mathcal{F}_i$  up to observation  $i$ . The conditional mean process of order  $(p, q)$  is given by an ARMA-type specification:

$$\mu_i = \mu_i(\theta) = \omega + \sum_{j=1}^p \alpha_j y_{i-j} + \sum_{j=1}^q \beta_j \mu_{i-j} \quad (4)$$

with parameters  $\omega$ ,  $\alpha = (\alpha_1, \dots, \alpha_p)^\top$  and  $\beta = (\beta_1, \dots, \beta_q)^\top$ . The model structure resembles the conditional variance equation of a GARCH( $p, q$ ) model, as soon as  $y_i$  denotes the squared (de-measured) log return at observation  $i$ .

Natural choices for the distribution of  $\varepsilon_i$  are the (standard) exponential distribution and the Weibull distribution. The former distribution allows for quasi maximum likelihood estimation and consistent estimates of EACD parameters even in the case of distributional misspecification. The latter is a simple but powerful generalization being sufficiently flexible in most applications. Define  $I = [i_0 - n, i_0]$  as a (right-end) fixed interval of  $(n + 1)$  observations at observation  $i_0$ . Then, local ACD models are given as follows:

(i) *Exponential-ACD model (EACD)*:  $\varepsilon_i \sim \exp(1)$ ,  $\theta_E = (\omega, \alpha^\top, \beta^\top)^\top$ , with (quasi) log-likelihood function over  $I = [i_0 - n, i_0]$  given  $i_0$ :

$$\ell_I(y; \theta_E) = \sum_{i=\max(p,q)+1}^n \left( -\log \mu_i - \frac{y_i}{\mu_i} \right) \mathbf{I}(i \in I) \tag{5}$$

(ii) *Weibull-ACD model (WACD)*:  $\varepsilon_i \sim \mathcal{G}(s, 1)$ ,  $\theta_W = (\omega, \alpha^\top, \beta^\top, s)^\top$ , with log-likelihood function over  $I = [i_0 - n, i_0]$  given  $i_0$ :

$$\ell_I(y; \theta_W) = \sum_{i=\max(p,q)+1}^n \left[ \log \frac{s}{y_i} + s \log \frac{\Gamma(1 + 1/s)y_i}{\mu_i} - \left\{ \frac{\Gamma(1 + 1/s)y_i}{\mu_i} \right\}^s \right] \mathbf{I}(i \in I) \tag{6}$$

Correspondingly, the (quasi-)maximum likelihood estimates ((Q)MLEs) of  $\theta_E$  and  $\theta_W$  over the data interval  $I$  are given by

$$\tilde{\theta}_I = \arg \max_{\theta \in \Theta} \ell_I(y; \theta) \tag{7}$$

### 3.2. Local Parameter Dynamics

The idea behind the local parametric approach (LPA) is to select at each time point an optimal length of data window over which a constant parametric model cannot be rejected by a test to be described below. The resulting *interval of homogeneity* is used to locally estimate the model and to compute out-of-sample predictions. Since the approach is implemented on a rolling window basis, it naturally captures time-varying parameters and allows identifying breakpoints where the length of the locally optimal estimation window has to be adjusted.

The implementation of the LPA requires estimating the model at each point in time using estimation windows with sequentially varying lengths. We consider data windows with lengths of 1 hour, 2 hours, 3 hours, 1 trading day (6 hours), 2 trading days (12 hours) and 1 trading week (30 hours). As non-trading periods (i.e. overnight periods, weekends or holidays) are removed, the estimation windows contain data potentially covering several days. Applying (local) EACD(1, 1) and WACD(1, 1) models based on five stocks, we estimate in total 4,644,000 parameter vectors. It turns out that estimated MEM parameters substantially change over time, with the variations depending on the lengths of underlying local (rolling) windows. As an illustration, Figure 2 shows EACD parameters employing 1-day (6 trading hours) and 1-week (30 trading hours) estimation windows for Intel Corporation (INTC). Note that the first 30 days are used for the estimation of intra-day periodicity effects, whereas an additional 5 days are required to obtain the first ‘weekly’ estimate (i.e. an estimate using 1 trading week of data).

We observe that estimated parameters  $(\tilde{\omega}, \tilde{\alpha}$  and  $\tilde{\beta})$  and persistence levels  $(\tilde{\alpha} + \tilde{\beta})$  clearly vary over time. As expected, estimates are less volatile if longer estimation windows (such as 1 week of data) are used. Conversely, estimates based on local windows of 6 hours are less stable. This might be induced either by high (true) local variations which are smoothed away if the data window becomes larger, or by an obvious loss of estimation efficiency as fewer data points are employed. These differences in estimates’ variations are also reflected in the empirical time series distributions of MEM parameters. Table I provides quartiles of the estimated persistence  $(\tilde{\alpha} + \tilde{\beta})$  (pooled across all five stocks) in dependence of the length of the underlying data window. We associate the first quartile (25% quantile) with a ‘low’ persistence level, whereas the second quartile (50% quantile) and third quartile

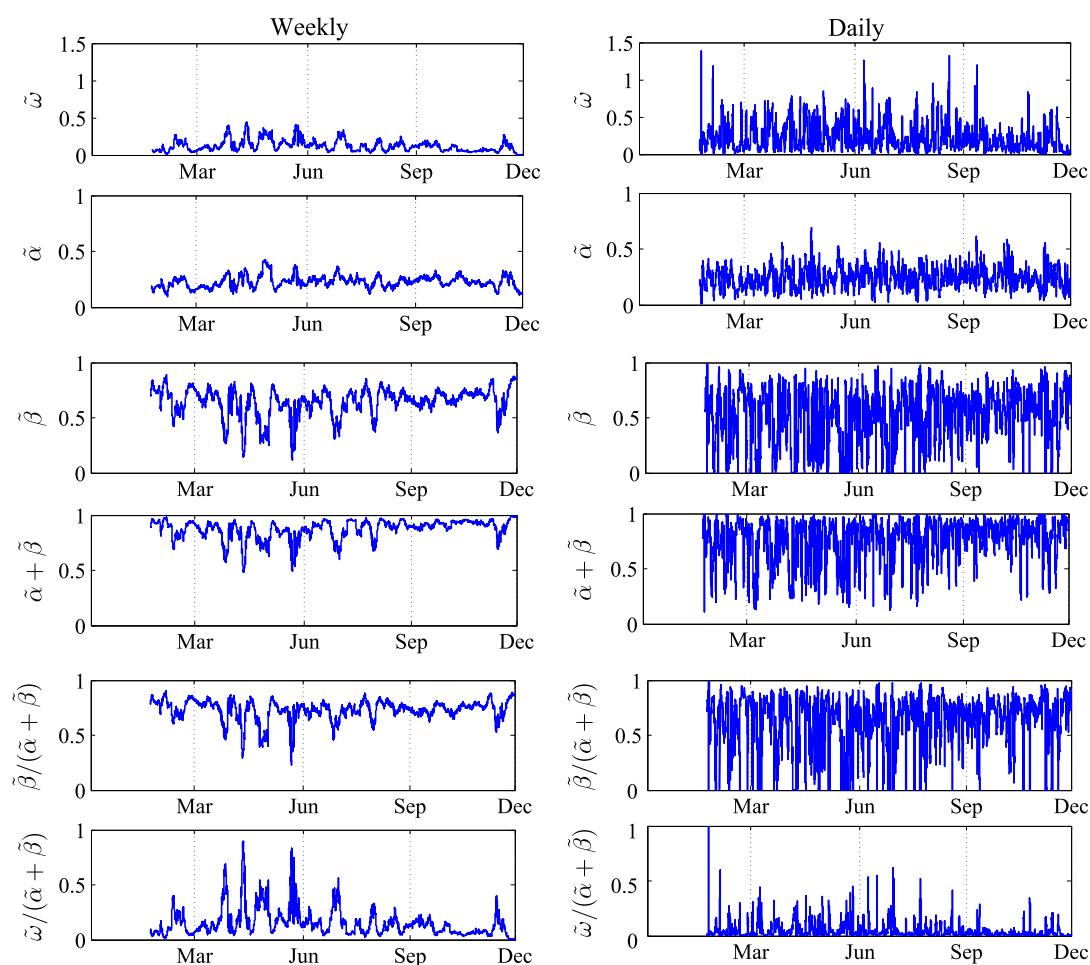


Figure 2. Time series of estimated 'weekly' (left panel, rolling windows covering 1800 observations) and 'daily' (right panel, rolling windows covering 360 observations) EACD(1, 1) parameters and functions thereof based on seasonally adjusted 1-minute trading volumes for Intel Corporation (INTC) at each minute from 22 February to 31 December 2008

(75% quantile) are associated with 'moderate' and 'high' persistence levels, respectively. It is shown that the estimated persistence increases with the length of the estimation window. Again, this result might reflect that the 'true' persistence of the process can only be reliably estimated over sufficiently long sampling windows. Alternatively, it might indicate that the revealed persistence is just a spurious effect caused by aggregations over underlying structural changes.

Summarizing these first pieces of empirical evidence on local variations of MEM parameters, we can conclude: (i) MEM parameters, their variability and their distribution properties change over time and are obviously dependent on the length of the underlying estimation window; (ii) longer local estimation windows increase the estimation precision but also enlarge the risk of misspecifications (due to averaging over structural breaks) and thus increase the modelling bias. Standard time series approaches would strive to obtain precise estimates by selecting large estimation windows, inflating, however, at the same time the bias. Conversely, the LPA aims at finding a balance between parameter variability and modelling bias. By controlling estimation risk, the procedure accounts for the possible

Table I. Quartiles of estimated persistence levels  $(\tilde{\alpha} + \tilde{\beta})$  for all five stocks at each minute from 22 February to 31 December 2008 (215 trading days) and six lengths of local estimation windows based on EACD and WACD specifications. We label the first quartile as ‘low’, the second quartile as ‘moderate’ and the third quartile as ‘high’

Estimation window	EACD(1,1)			WACD(1,1)		
	Low	Moderate	High	Low	Moderate	High
1 week	0.85	0.89	0.93	0.82	0.88	0.92
2 days	0.77	0.86	0.92	0.74	0.84	0.91
1 day	0.68	0.82	0.90	0.63	0.79	0.89
3 hours	0.54	0.75	0.88	0.50	0.72	0.87
2 hours	0.45	0.70	0.86	0.42	0.67	0.85
1 hour	0.33	0.58	0.80	0.31	0.57	0.80

Table II. Quartiles of 774,000 estimated ratios  $\tilde{\beta} / (\tilde{\alpha} + \tilde{\beta})$  (based on estimation windows covering 1800 observations) for all five stocks at each minute from 22 February to 31 December 2008 (215 trading days) and both model specifications (EACD and WACD) conditional on the persistence level (low, moderate or high). We label the first quartile as ‘low’, the second quartile as ‘mid’ and the third quartile as ‘high’

Model	Low persistence			Moderate persistence			High persistence		
	Low	Mid	High	Low	Mid	High	Low	Mid	High
EACD, $\tilde{\alpha}$	0.28	0.22	0.18	0.30	0.23	0.19	0.31	0.24	0.20
EACD, $\tilde{\beta}$	0.56	0.62	0.67	0.59	0.66	0.71	0.62	0.68	0.73
WACD, $\tilde{\alpha}$	0.28	0.21	0.17	0.30	0.23	0.18	0.32	0.24	0.19
WACD, $\tilde{\beta}$	0.54	0.60	0.65	0.58	0.65	0.70	0.60	0.68	0.74

trade-off between (in)efficiency and the coverage of local variations by finding the longest possible interval over which parameter homogeneity cannot be rejected.

An important ingredient of the sequential testing procedure in the LPA is a set of critical values. The critical values have to be calculated for reasonable parameter constellations. Therefore, we aim at parameters which are most likely to be estimated from the data. As a first criterion we distinguish between different levels of persistence,  $\tilde{\alpha} + \tilde{\beta}$ . This is performed by classifying the estimates into three persistence groups (low, medium or high persistence) according to the first row of Table I. Then, within each persistence group, we distinguish between different magnitudes of  $\tilde{\alpha}$  relative to  $\tilde{\beta}$ . This naturally results into groups according to the quartiles of the ratio  $\tilde{\beta} / (\tilde{\alpha} + \tilde{\beta})$ , yielding again three categories (low, mid or high ratio). As a result, we obtain nine groups of parameter constellations, see Table II, which are used below to simulate critical values for the sequential testing procedure.

### 3.3. Estimation Quality

Addressing the inherent trade-off between estimation (in)efficiency and local flexibility requires controlling the estimation quality. In the proposed LPA framework, the so-called *pseudo true* parameter changes over time (see, for example, Spokoiny, 2009). The key idea is to approximate this process by a model with parameters which are constant over an interval with optimized length. Denote the *pseudo true* (time-varying) parameter vector by  $\theta^*$  associated with a fixed interval  $I$ , where, for convenience, we omit the time subscript and only keep an asterisk (\*) through the text. The quality of the (Q)MLE

$\tilde{\theta}_I$  of the *pseudo true*  $\theta^*$  is assessed by the Kullback–Leibler (KL) divergence. In particular, for a fixed interval  $I$ , we consider the (positive) difference  $\ell_I(\tilde{\theta}_I) - \ell_I(\theta^*)$  with log-likelihood expressions for the EACD and WACD models given by equations (5) and (6), respectively. Denote the corresponding loss function by  $L_I(\tilde{\theta}_I, \theta^*) = |\ell_I(\tilde{\theta}_I) - \ell_I(\theta^*)|$ .

By introducing the  $r$ th power of the loss function, i.e. for any  $r > 0$ , there is a constant  $\mathcal{R}_r(\theta^*)$  satisfying

$$E_{\theta^*} \left| L_I(\tilde{\theta}_I, \theta^*) \right|^r \leq \mathcal{R}_r(\theta^*) \quad (8)$$

and denoting the (parametric) risk bound depending on  $r > 0$  and  $\theta^*$  (see, for example, Spokoiny (2009); Čížek *et al.* (2009)). The risk bound (8) allows the construction of non-asymptotic confidence sets and testing the validity of the (local) parametric model. For the construction of critical values, we exploit equation (8) to show that the random set  $\mathcal{S}_I(z_\alpha) = \{\theta : L_I(\tilde{\theta}_I, \theta^*) \leq z_\alpha\}$  is an  $\alpha$ -confidence set in the sense that  $P_{\theta^*}(\theta^* \notin \mathcal{S}_I(z_\alpha)) \leq \alpha$ .

The parameter  $r$  drives the tightness of the risk bound. Accordingly, different values of  $r$  lead to different risk bounds, critical values and thus adaptive estimates. Higher values of  $r$  lead to, *ceteris paribus*, a selection of longer intervals of homogeneity and more precise estimates, however, increase the modelling bias. It might be chosen in a data-driven way, e.g. by minimizing forecasting errors. Here, we follow Čížek *et al.* (2009) and consider  $r = 0.5$  and  $r = 1$ , a ‘modest risk case’ and a ‘conservative risk case’, respectively.

#### 4. LOCAL PARAMETRIC MODELLING

The local parametric approach requires a time series to be locally, i.e. over short periods of time, approximated by a parametric model. Though local approximations are obviously more accurate than global ones, this proceeding raises the question of the optimal size of the local interval.

##### 4.1. Statistical Framework

Including more observations in an estimation window reduces the variability, but obviously enlarges the bias. The algorithm presented below strikes a balance between bias and parameter variability and yields an *interval of homogeneity*. Our goal is to well approximate the ‘true’ model over an interval  $I_k$  by the parametric model with constant parameter  $\theta$ . The quality of approximation is measured by the KL divergence. Consider the KL divergence  $\mathcal{K}(v, v')$  between probability distributions induced by  $v$  and  $v'$ . Then, define  $\Delta_{I_k}(\theta) = \sum_{i \in I_k} \mathcal{K}\{\mu_i, \mu_i(\theta)\}$ , where  $\mu_i(\theta)$  denotes the model described by equation (4) and  $\mu_i$  is the true (unknown) data-generating process. The entity  $\Delta_{I_k}(\theta)$  measures the distance between the underlying process and the assumed parametric model and thus allows us to control the modelling bias.

Let, for some  $\theta \in \Theta$ ,

$$E[\Delta_{I_k}(\theta)] \leq \Delta \quad (9)$$

where  $\Delta \geq 0$  denotes the *small modelling bias* (SMB) for an interval  $I_k$ . The SMB condition implies that, for some parameter  $\theta$ , the random quantity  $\Delta_{I_k}(\theta)$  is bounded by a small constant with a high probability. Therefore, on the interval  $I_k$ , the ‘true’ model can be well approximated by the parametric model with parameter  $\theta$  while keeping the modelling bias ‘small’ according to equation (9). The best parametric fit (4) on  $I_k$  is obtained by minimizing  $E[\Delta_{I_k}(\theta)]$  over  $\theta \in \Theta$ . Here, the KL concept is used for theoretical underpinning, but we do not estimate it in practice.

Čížek *et al.* (2009) show that under the SMB condition (9), estimation loss scaled by the parametric risk bound  $\mathcal{R}_r(\theta^*)$  is stochastically bounded. In particular, in the case of (Q)ML estimation with loss function  $L_I(\tilde{\theta}_I, \theta^*)$ , the SMB condition implies

$$\mathbb{E} \left[ \log \left\{ 1 + \left| L_I(\tilde{\theta}_I, \theta^*) \right|^r / \mathcal{R}_r(\theta^*) \right\} \right] \leq 1 + \Delta \quad (10)$$

The proposed framework captures dependent data given a linear specification of the conditional mean process. The methodology, however, can be generalized to nonlinear structures, assuming that, locally, a nonlinear model approximates the ‘true’ (unknown) conditional mean process. Then the KL divergence considers the probability measures induced by the ‘true’ model and that of the nonlinear data structure, yielding, however, different (and more complex) risk bounds.

Consider  $(K + 1)$  nested intervals (with fixed right-end point  $i_0$ )  $I_k = [i_0 - n_k, i_0]$  of length  $n_k$ ,  $I_0 \subset I_1 \subset \dots \subset I_K$ . Then, the ‘oracle’ (i.e. theoretically optimal) choice  $I_{k^*}$  of the interval sequence is defined as the largest interval for which the SMB condition holds:

$$\mathbb{E} [\Delta_{I_{k^*}}(\theta)] \leq \Delta \quad (11)$$

This ‘oracle’ choice provides the ‘best’ local fit but not necessarily the best out-of-sample forecast. Optimizing the procedure in terms of out-of-sample forecasting performance, however, is beyond the scope of this paper. This task may appear infeasible in the case of high-frequency data modelling due to the increased computational burden, unless very restrictive assumptions are imposed. It is therefore our major research question to what extent an ‘optimal’ local fit is beneficial for out-of-sample forecasts.

So far, there has been limited attention devoted to the selection of optimal window lengths in the econometric forecasting literature. As stressed by Čížek *et al.* (2009), time-varying coefficients are typically assumed as smooth functions (of time) or, alternatively, as piecewise constant functions. For instance, Pesaran and Timmermann (2007) consider a linear regression framework subject to structural breaks under the assumption of the presence of sudden jumps in the parameter values. Clark and McCracken (2009) extend this work and allow for conditional heteroskedasticity and serial correlation in the regression error terms. The LPA approach, however, includes both scenarios as special cases: parameters can vary over time as the interval changes with  $i$  and, at the same time, can reveal discontinuities and jumps as a function of time. In both cases, the observed data are described by an unobserved process which, at each point  $i$ , can be described by a historical interval in which the process (approximately) follows a parametric specification. This local assumption enables us to apply well-developed parametric methods to estimate the underlying parameter.

In practice,  $\Delta_{I_k}$  is unknown and therefore the oracle choice  $k^*$  cannot be implemented. Consequently, the aim is to mimic the oracle choice using a sequential testing procedure for the different intervals  $k = 1, \dots, K$ . The resulting interval  $I_{\hat{k}}$  is then used to construct the local estimator. Čížek *et al.* (2009) and Spokoiny (2009) show that the estimation errors induced by the adaptive estimation during steps  $k \leq k^*$  are not larger than those induced by (Q)ML estimation directly using  $k^*$ . Hence the sequential estimation and testing procedure does not incur a larger estimation error compared to the situation where  $k^*$  is known; see equation (10).

In applications, the lengths of the underlying intervals evolve on a geometric grid with initial length  $n_0$  and a multiplier  $c > 1$ ,  $n_k = \lceil n_0 c^k \rceil$ . In the present study, we select  $n_0 = 60$  observations (i.e. minutes) and consider two schemes with  $c = 1.50$  and  $c = 1.25$  and  $K = 8$  and  $K = 13$ , respectively:

- (i)  $n_0 = 60$  min,  $n_1 = 90$  min,  $\dots$ ,  $n_8 = 1$  week (9 estimation windows,  $K = 8$ ); and
- (ii)  $n_0 = 60$  min,  $n_1 = 75$  min,  $\dots$ ,  $n_{13} = 1$  week (14 estimation windows,  $K = 13$ ).

The latter scheme bears a slightly finer granulation than the first one.



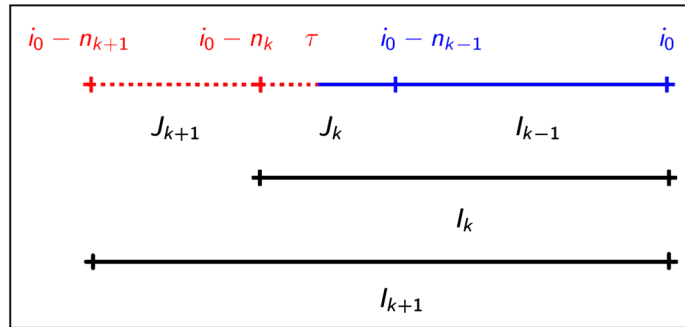


Figure 3. Graphical illustration of sequential testing for parameter homogeneity in interval  $I_k$  with length  $n_k = |I_k|$  ending at fixed time point  $i_0$ . Suppose we have not rejected homogeneity in interval  $I_{k-1}$ , we search within the interval  $J_k = I_k \setminus I_{k-1}$  for a possible change point  $\tau$ . In the top figure, the dotted region marks interval  $A_{k,\tau}$  and the blue region marks interval  $B_{k,\tau}$  splitting the interval  $I_{k+1}$  into two parts depending upon the position of the unknown change point  $\tau$

**4.2. Local Change Point (LCP) Detection Test**

Selecting the optimal length of the interval builds on a sequential testing procedure where at each interval  $I_k$  one tests the null hypothesis on parameter homogeneity against the alternative of a change point at unknown location  $\tau$  within  $I_k$ .

The test statistic is given by

$$T_{k-1,k} = \sup_{\tau \in J_k} \left\{ \ell_{A_{k,\tau}}(\tilde{\theta}_{A_{k,\tau}}) + \ell_{B_{k,\tau}}(\tilde{\theta}_{B_{k,\tau}}) - \ell_{I_{k+1}}(\tilde{\theta}_{I_{k+1}}) \right\} \tag{12}$$

where  $J_k$  and  $B_k$  denote intervals  $J_k = I_k \setminus I_{k-1}$ ,  $A_{k,\tau} = [i_0 - n_{k+1}, \tau]$  and  $B_{k,\tau} = (\tau, i_0]$  utilizing only a part of the observations within  $I_{k+1}$ . As the location of the change point is unknown, the test statistic considers the supremum of the corresponding likelihood ratio statistics over all  $\tau \in I_k$ .

Figure 3 illustrates the underlying idea graphically: assume that, for a given time point  $i_0$ , parameter homogeneity in interval  $I_{k-1}$  has been established. Then, homogeneity in interval  $I_k$  is tested by considering any possible breakpoint  $\tau$  in the interval  $J_k = I_k \setminus I_{k-1}$ . This is performed by computing the log-likelihood values over the intervals  $A_{k,\tau} = [i_0 - n_{k+1}, \tau]$  dotted area and  $B_{k,\tau} = (\tau, i_0]$  solid area in the top figure for given  $\tau$ . Computing the supremum of these two likelihood values for any  $\tau \in J_k$  and relating it to the log-likelihood associated with  $I_{k+1}$  ranging from  $i_0$  to  $i_0 - n_{k+1}$  results in the test statistic (12). For instance, in our setting based on  $(K + 1) = 14$  intervals, we test for a breakpoint, e.g. in interval  $I_1 = 75$  min, by searching only within the interval  $J_1 = I_1 \setminus I_0$ , containing observations from  $y_{i_0-75}$  up to  $y_{i_0-60}$ . Then, for any observation within this interval, we sum equations (5) and (6) for the EACD and WACD model, respectively, over  $A_{1,\tau}$  and  $B_{1,\tau}$  and subtract the likelihood over  $I_2$ . Then, the test statistic (12) corresponds to the largest obtained likelihood ratio.

Comparing the test statistic (12) for given  $i_0$  at every step  $k$  with the corresponding (simulated) critical value, we search for the longest *interval of homogeneity*  $I_k^{\wedge}$  for which the null is not rejected. Then, the *adaptive estimate*  $\hat{\theta}$  is the (Q)MLE at the *interval of homogeneity*, i.e.  $\hat{\theta} = \tilde{\theta}_{I_k^{\wedge}}$ . If the null is already rejected at the first step, then  $\hat{\theta}$  equals the (Q)MLE at the shortest interval  $I_0$ . Conversely, if no breakpoint can be detected within  $I_K$ , then  $\hat{\theta}$  equals the (Q)MLE of the longest window  $I_K$ .

**4.3. Critical Values**

Under the null hypothesis of parameter homogeneity, the correct choice in the pure parametric situation is the largest considered interval  $I_K$ . In the case of selecting  $k < K$  and thus choosing  $\hat{\theta} = \tilde{\theta}_{I_k}$  instead of  $\tilde{\theta}_{I_K}$ , the loss is  $L_{I_K}(\tilde{\theta}_{I_k}, \hat{\theta}) = \ell_{I_K}(\tilde{\theta}_{I_k}) - \ell_{I_K}(\hat{\theta})$  and is stochastically bounded:

$$E_{\theta^*} \left| L_{I_K}(\tilde{\theta}_{I_k}, \hat{\theta}) \right|^r \leq \rho \mathcal{R}_r(\theta^*) \tag{13}$$

Critical values must ensure that the loss associated with ‘false alarm’ (i.e. selecting  $k < K$ ) is at most a  $\rho$ -fraction of the parametric risk bound of the ‘oracle’ estimate  $\tilde{\theta}_{I_K}$ . For  $r \rightarrow 0$ ,  $\rho$  can be interpreted as the false alarm probability. We select the minimal critical values ensuring a small probability of such a false alarm.

Accordingly, an estimate  $\hat{\theta}_{I_k}$ ,  $k = 1, \dots, K$ , should satisfy

$$E_{\theta^*} \left| L_{I_k}(\tilde{\theta}_{I_k}, \hat{\theta}_{I_k}) \right|^r \leq \rho_k \mathcal{R}_r(\theta^*) \tag{14}$$

with  $\rho_k = \rho k / K \leq \rho$ . Condition (14) is fulfilled with the choice

$$z_k = a_0 r \log(\rho^{-1}) + a_1 r \log(n_K / n_{k-1}) + a_2 \log(n_k), \quad k = 1, \dots, K \tag{15}$$

with constants  $a_0, a_1$  and  $a_2$ . Since the number of selected intervals  $\{I_k\}_{k=1}^K$  and their corresponding lengths  $\{n_k\}_{k=1}^K$  are fixed, Čížek *et al.* (2009) show that the critical values are of the form  $z_k = C + D \log(n_k)$  for  $k = 1, \dots, K$  with some constants  $C$  and  $D$ . A relevant choice of these constants has to be selected by Monte Carlo simulation on the basis of the assumed data-generating process (4) and the assumption of parameter homogeneity over the interval sequence  $\{I_k\}_{k=1}^K$ . The procedure is run for fixed values  $C$  and  $D$  using simulated data, allowing to evaluate its performance and to monitor if the condition (14) is fulfilled. Then, for a fixed value of  $C$ , one finds the minimal value  $D(C) < 0$  ensuring a decreasing pattern (with  $k$ ) of the critical values. Therefore, a false alarm at an early stage is more crucial since it is associated with a comparably variable estimate. After fixing the false alarm probability at the first step, one determines the constant  $C$  (see, for example, Čížek *et al.*, (2009). The authors note that, alternatively, the constants  $C$  and  $D$  could be found by minimizing the related prediction errors.

To simulate the data-generating process, we use the parameter constellations underlying the nine groups described in Section 3.2. and shown in Table II for nine different parameters  $\theta^*$ . The Weibull parameter  $s$  is set to its median value  $\tilde{s} = 1.57$  in all cases. Moreover, we consider two risk levels ( $r = 0.5$  and  $r = 1$ ), two interval granulation schemes ( $K = 8$  and  $K = 13$ ) and two significance levels ( $\rho = 0.25$  and  $\rho = 0.50$ ) underlying the test.

The resulting critical values satisfying equation (14) for the nine possibilities of ‘true’ parameter constellations of the EACD(1, 1) model for  $K = 13$ ,  $r = 0.5$  (‘moderate risk case’) and  $\rho = 0.25$  are displayed in Figure 4. We observe that the critical values are virtually invariable with respect to  $\theta^*$  across the nine scenarios. The largest difference between all cases appears for interval lengths up to 90 minutes. Beyond that, the critical values are robust across the range of parameters also for the conservative risk case ( $r = 1$ ), other significance levels and interval selection schemes.

In the sequential testing procedure, we employ parameter-specific critical values. In particular, at each minute  $i_0$ , we estimate a local MEM over a given interval length and choose the critical values (for given levels of  $\rho$  and  $r$ ) simulated for those parameter constellations (according to Table II) which are closest to our local estimates. For instance, suppose that at some point  $i_0$  we have  $\tilde{\alpha} = 0.32$  and  $\tilde{\beta} = 0.53$ . Then, we select the curve associated with the low persistence  $(\tilde{\alpha} + \tilde{\beta})$  and the low ratio

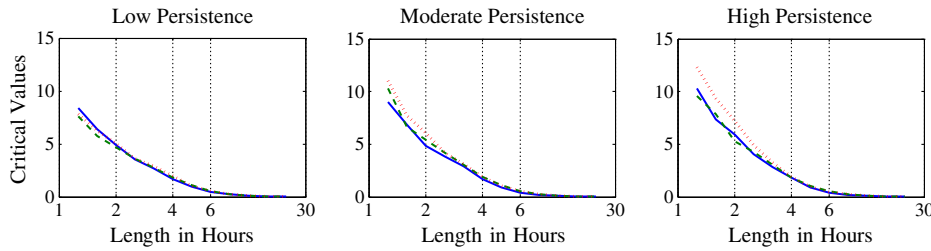


Figure 4. Simulated critical values of an EACD(1, 1) model for the ‘moderate risk case’ ( $r = 0.5$ ),  $\rho = 0.25$ ,  $K = 13$  and chosen parameter constellations according to Table II. The low (solid), middle (dashed) and upper (dotted) curves are associated with the corresponding ratio levels  $\tilde{\beta}/(\tilde{\alpha} + \tilde{\beta})$

Table III. Summary of the local change point (LCP) detection test and adaptive estimation at fixed observation  $i_0$ . Here  $\tau$  denotes the unknown change point and  $n_k$  represents the length of the interval  $I_k$

---

*LCP: step 1*

- Select intervals:  $I_2, I_1, J_1 = I_1 \setminus I_0, A_{1,\tau} = [i_0 - n_2, \tau]$  and  $B_{1,\tau} = (\tau, i_0]$
- Compute the test statistic (12) at step 1:  $T_{0,1} = \sup_{\tau \in J_1} \{ \ell_{A_{1,\tau}}(\tilde{\theta}_{A_{1,\tau}}) + \ell_{B_{1,\tau}}(\tilde{\theta}_{B_{1,\tau}}) - \ell_{I_2}(\tilde{\theta}_{I_2}) \}$

*LCP: step k*

- Select intervals:  $I_{k+1}, I_k, J_k = I_k \setminus I_{k-1}, A_{k,\tau} = [i_0 - n_{k+1}, \tau]$  and  $B_{k,\tau} = (\tau, i_0]$
- Compute the test statistic (12) at step  $k$ :  $T_{k-1,k} T_{k,k-1} = \sup_{\tau \in J_k} \{ \ell_{A_{k,\tau}}(\tilde{\theta}_{A_{k,\tau}}) + \ell_{B_{k,\tau}}(\tilde{\theta}_{B_{k,\tau}}) - \ell_{I_{k+1}}(\tilde{\theta}_{I_{k+1}}) \}$

*Testing procedure*

- Select the set of critical values  $\{\tilde{\beta}_k\}_{k=1}^K$  according to the ‘persistence’ level  $(\tilde{\alpha} + \tilde{\beta})$  and ‘smoothness’ level  $\tilde{\beta}/(\tilde{\alpha} + \tilde{\beta})$  of the ‘weekly’ estimate  $\tilde{\theta}_K$  and the desired tuning parameter constellation
- Compare  $T_{k-1,k}$  with the simulated critical value  $\tilde{\beta}_k$  at step  $k$
- Decision: reject the null of parameter homogeneity if  $T_{k-1,k} > \tilde{\beta}_k$

*Adaptive estimation*

- Interval of homogeneity  $I_{\hat{k}}$ : the null has been first rejected at step  $\hat{k} + 1$
- Adaptive estimate:  $\hat{\theta} = \tilde{\theta}_{\hat{k}}^k$  (i.e. (Q)MLE at the interval of homogeneity)

---

$\tilde{\beta}/(\tilde{\alpha} + \tilde{\beta})$ . The key steps of the LCP detection test and the adaptive estimation are for convenience summarized in Table III.

For illustration, the resulting adaptive choice of intervals at each minute on 22 February 2002 is shown by Figure 5. Adopting the EACD specification (for  $\rho = 0.25$  and  $K = 13$ ) in the modest risk case ( $r = 0.5$ , solid curve), one would select the length of the adaptive estimation interval lying between 1.5 and 3.5 hours over the course of the selected day. Likewise, in the conservative risk case ( $r = 1$ , dashed curve), the approach would select longer time windows with smaller variability and thus larger modelling bias.

The time series of the chosen length of the intervals of homogeneity for Intel Corporation is shown in Figure 6. The length of intervals ranges between 1 and 4 hours in the modest risk case ( $r = 0.5$ ) and between 2.5 and 5 hours in the conservative risk case ( $r = 1$ ). The results indicate a larger variability over shorter interval lengths in the modest risk case.

#### 4.4. Empirical Findings

We apply the LPA to seasonally adjusted 1-minute aggregated trading volumes for all five stocks at each minute from 22 February to 31 December 2008 (215 trading days; 77,400 trading minutes). We use the EACD and WACD models as the two (local) specifications, two risk levels (modest,  $r = 0.5$ ;

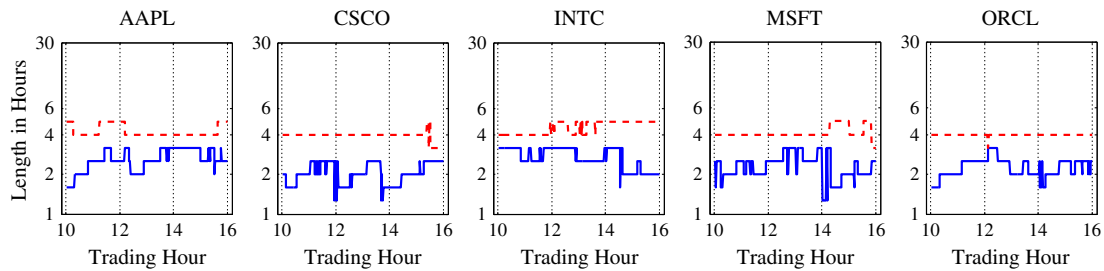


Figure 5. Estimated length of intervals of homogeneity  $n_{\hat{k}}$  (in hours) for seasonally adjusted 1-minute cumulative trading volumes of selected companies in the case of a modest ( $r = 0.5$ , solid line) and conservative ( $r = 1$ , dashed line) modelling risk level. We use the interval scheme with  $K = 13$  and  $\rho = 0.25$ . Underlying model: EACD(1, 1). NASDAQ trading on 22 February 2008

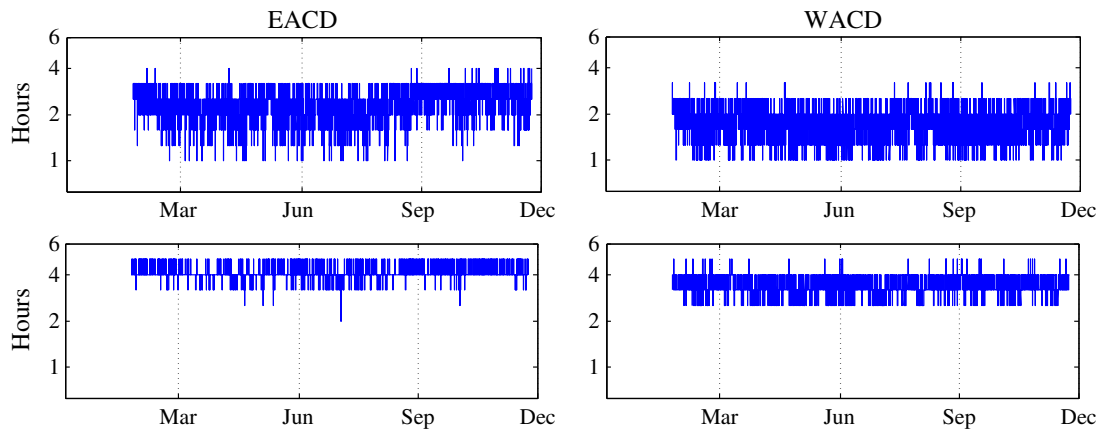


Figure 6. Estimated length of intervals of homogeneity  $n_{\hat{k}}$  (in hours) for seasonally adjusted 1-minute cumulative trading volumes for Intel Corporation (INTC) in case of a modest ( $r = 0.5$ , upper panel) and conservative ( $r = 1$ , lower panel) modelling risk level. We use the interval scheme with  $K = 13$  and  $\rho = 0.25$ . Underlying models: EACD(1, 1) (left) and WACD(1, 1) (right). NASDAQ trading from 22 February to 22 December 2008 (210 trading days)

and conservative,  $r = 1$ ) and two significance levels ( $\rho = 0.25$  and  $\rho = 0.50$ ). Furthermore, interval length schemes with (i)  $K = 8$  and (ii)  $K = 13$  are employed.

Figure 7 depicts the time series distributions of selected oracle interval lengths. First, as expected, the chosen intervals are shorter in the modest risk case ( $r = 0.5$ ) than in the conservative case ( $r = 1$ ). Practically, if a trader aims at obtaining more precise volume estimates, it is advisable to select longer estimation periods, such as 4–5 hours. By doing so, the trader increases the modelling bias, but can still control it according to equation (8). Hence this risk level allows for more controlled flexibility in modelling the data. Conversely, setting  $r = 1$  implies a smaller modelling bias and thus lower estimation precision. Consequently, it yields smaller local intervals ranging between 2 and 3 hours in most cases.

Secondly, our results provide guidance on how (a priori) to choose the length of a local window in practice. Interestingly, the procedure never selects the longest possible interval according to our interval scheme (1 week of data), but chooses a maximum length of 6 hours. This finding suggests that even a week of data is clearly too long to capture parameter inhomogeneity in high-frequency variables. As a rough rule of thumb, a horizon of up to 1 trading day seems to be reasonable. This result is remarkably robust across the individual stocks, suggesting that the stochastic properties of

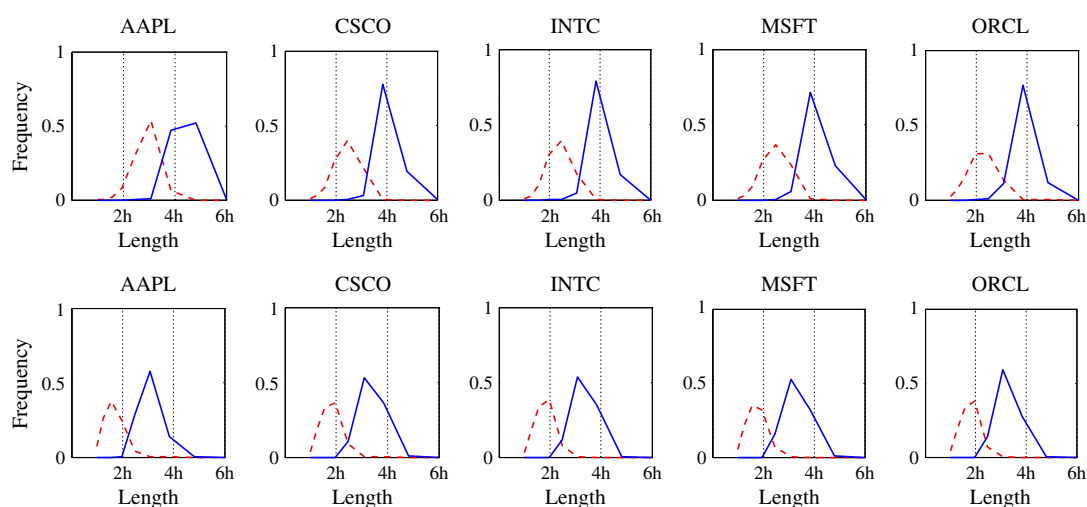


Figure 7. Distribution of estimated interval length  $n_{\hat{k}}$  (in hours) for seasonally adjusted trading volumes of selected companies in the case of modest ( $r = 0.5$ , dashed) and conservative modelling risk ( $r = 1$ , solid), using an EACD (upper panel) and a WACD model (lower panel) from 22 February to 31 December 2008 (215 trading days). We select 13 estimation windows based on significance level  $\rho = 0.25$

Table IV. Average daily number of changes of the adaptively selected interval of homogeneity for five stocks at NASDAQ from 22 February to 22 December 2008 (210 trading days) across different tuning parameter constellations

	EACD					WACD				
	AAPL	CSCO	INTC	MSFT	ORCL	AAPL	CSCO	INTC	MSFT	ORCL
$r = 0.5, \rho = 0.25$	17.8	27.2	27.2	26.7	29.2	39.1	36.4	35.8	37.1	34.5
$r = 0.5, \rho = 0.50$	18.1	26.7	27.2	26.6	29.3	39.1	36.4	36.2	37.2	34.7
$r = 1.0, \rho = 0.25$	8.4	9.6	10.3	11.0	9.8	17.5	18.1	17.6	17.1	17.1
$r = 1.0, \rho = 0.50$	8.7	9.7	10.4	10.9	9.7	18.3	17.8	18.0	16.9	17.0

high-frequency trading volumes are quite similar, at least across (heavily traded) blue chip stocks. Nevertheless, as also illustrated in Figure 5, our findings show that the selected interval lengths clearly vary across time. Hence a priori fixing the length of a rolling window can be still problematic and suboptimal—even over the course of a day.

Thirdly, the optimal length of local windows does obviously also depend on the complexity of the underlying (local) model. In fact, we observe that local EACD specifications seem to better approximate the data over longer estimation windows than in the case of WACD specifications. This is true for nearly all stocks. Furthermore, from the average daily number of changes of the ‘optimal’ window, as reported in Table IV, one observes that the WACD results in roughly twice as many changes as the EACD model. Hence more complex (local) modelling specifications obviously yield more changes of the ‘optimal’ window. Interestingly, this (distributional) effect is more pronounced in the conservative risk approach ( $r = 1$ ), where one expects around 10 (EACD) or 20 (WACD) changes per day. In the modest risk case ( $r = 0.5$ ) we observe more changes with a moderate difference between the underlying models, i.e. between 30 (EACD) and 40 (WACD) changes per day. All stocks reveal quite similar patterns across the scenarios.

Finally, in Figure 8, we show time series averages of selected interval lengths in dependence of the time of the day. Even after removing the intra-day seasonality component, we observe slightly shorter

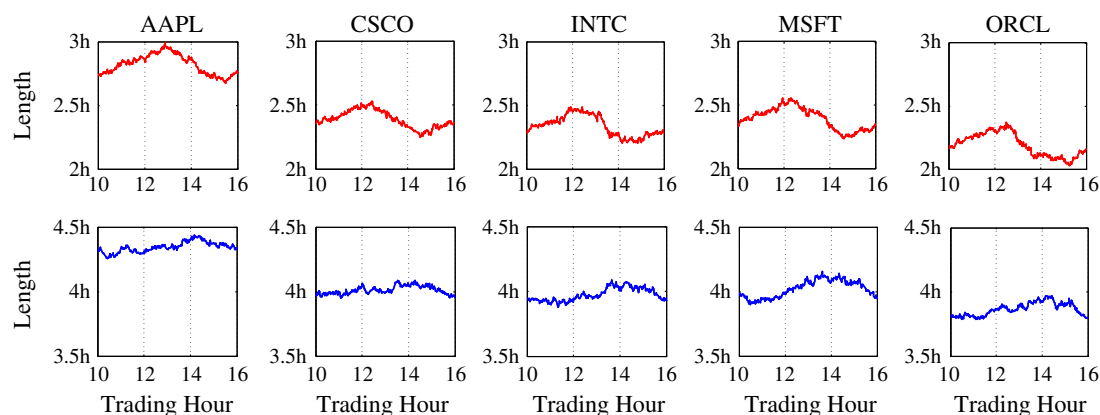


Figure 8. Average estimated interval length  $n_{\hat{k}}$  (in hours) over the course of a trading day for seasonally adjusted trading volumes of selected companies in the case of modest ( $r = 0.5$ , upper panel) and conservative modelling risk ( $r = 1$ , lower panel), using an EACD model from 22 February to 31 December 2008 (215 trading days). We select  $K = 13$  windows and set the significance level to  $\rho = 0.25$

intervals after opening and before closure. This is obviously induced by the fact that the local estimation window during the morning still includes significant information from the previous day. This effect is strongest at the opening, where estimates are naturally based on previous-day information solely and becomes weaker as time moves on and the proportion of current-day information is increasing. Consequently, we observe the longest intervals around mid-day, where most information in the local window stems from the current day. Hence the LPA automatically accounts for the effects arising from concatenated time series omitting non-trading periods. During the afternoon, interval lengths further shrink as trading becomes more active (and obviously less time homogeneous) before closure.

#### 4.5. Drivers of the ‘Optimal’ Window Length

To identify potential (observable) determinants influencing the stability of parameter estimates, we analyse the impact of key market variables on the selected length of the interval of homogeneity. In particular, we study to what extent the locally selected window length is predictable based on variables potentially causing inhomogeneity in trading processes, namely market volatility, the occurrence of outliers and of news announcements.

Analysing the impact of market volatility on the average daily selected ‘optimal’ window length, we distinguish between three regimes (low, moderate and high) of the daily volatility index (VIX). The low (high) is defined in terms of VIX realizations lower (higher) than the corresponding first (third) quartile. We report the correlation between the average daily length of the local estimation window and the daily VIX series in the different regimes in Table V.

The strongest dependence is observed in the high-volatility regime. Here, abrupt increases of market volatility significantly change the length of the selected intervals. Focusing on significant coefficients only, the EACD model reveals positive correlations between the volatility and length of intervals. In contrast, the WACD specification mostly induces a negative relationship. The results are quite robust across all five stocks and surprisingly stable for different risk (power) levels. Hence, in summary, we can conclude that market volatility has some impact on parameter homogeneity in trading volume models but the direction of this dependence is not clearly identifiable and obviously depends on the flexibility of the underlying local approximation.

Moreover, we analyse the effect of the occurrence of an outlier on the window length selection. The latter is defined as a realization of cumulative trading volumes exceeding the 99% percentile. We

Table V. Correlation coefficients between the average daily length of the interval of homogeneity and the daily VIX for five stocks at NASDAQ from 22 February to 22 December 2008 (210 trading days) across different tuning parameter constellations and three volatility regimes (low, moderate and high). The low (high) regime considers positive changes of the VIX that are lower (higher) than the corresponding first (third) quartile. We set  $\rho = 0.25$

	EACD					WACD				
	AAPL	CSCO	INTC	MSFT	ORCL	AAPL	CSCO	INTC	MSFT	ORCL
$r = 0.5$										
Low	0.10	-0.02	0.03	0.01	-0.02	-0.03	-0.07	0.10	0.01	-0.11
Moderate	-0.02	0.03	-0.03	-0.03	0.03	-0.03	-0.01	-0.09	-0.02	-0.02
High	0.26*	0.31*	0.23*	0.25*	0.30*	0.19*	-0.02	-0.07	-0.17*	-0.12
$r = 1$										
Low	0.19*	-0.07	-0.03	0.01	-0.12	0.04	0.00	0.08	0.01	-0.11
Moderate	-0.02	0.11	0.03	0.01	0.04	-0.08	0.05	-0.01	-0.02	-0.05
High	0.22*	0.26*	0.26*	0.19*	0.31*	0.19*	-0.11	0.09	-0.20*	-0.22*

Note: \*5% significance.

Table VI. Percentage change of the average length of the interval of homogeneity after a large outlier has been observed for five stocks at NASDAQ from 22 February to 22 December 2008 (210 trading days) across different tuning parameter constellations. We set  $\rho = 0.25$

	AAPL	CSCO	INTC	MSFT	ORCL
EACD, $r = 0.5$	-1.55	-3.06*	-2.78*	-2.45*	-2.09
EACD, $r = 1.0$	-0.37	-1.12	-1.42*	-1.04	-0.94
WACD, $r = 0.5$	-4.98*	-4.59*	-3.04*	-4.54*	-3.62*
WACD, $r = 1.0$	-1.88*	-1.60	-1.96*	-2.09*	-1.92*

Note: \*5% significance.

compute the average length of intervals of homogeneity at the time point of an outlier's appearance and 5 minutes thereafter.

As shown in Table VI, the selected interval of homogeneity becomes smaller after observing a large outlier. On average, the estimation window becomes on average shorter by 1% and 5% across all stocks as well as across the different modelling frameworks. In most cases, the effect is statistically significant at the 5% level. Interestingly, the changes are more pronounced based on a WACD specification and based on a modest risk level ( $r = 0.5$ ). These results confirm our finding that a more complex modelling approach or less conservative risk level yields a higher variability in 'optimal' window lengths.

Finally, we analyse to what extent daily news arrivals cause structural instability and thus changes of local window lengths. For this purpose we utilize pre-processed company-relevant news data from a news analytics tool of Reuters: the Reuters NewsScope Sentiment Engine. Here, firm-specific news is processed based on an automated linguistic analysis of news stories and is classified according to news direction and relevance; for details, see, for example, Groß-Klußmann and Hautsch (2011). As reported in Table VII, the number of 'relevant' company-specific news per day has only a minor impact on the lengths of local intervals of parameter homogeneity. In fact, the corresponding correlations are not significantly different from zero. Only for one stock (Microsoft) we find significant (negative) relationship in the modest risk case ( $r = 0.5$ ). Here, the length of the interval of homogeneity varies stronger if news arrive.

Table VII. Correlation coefficients between the average daily length of the interval of homogeneity and the daily number of relevant company-specific news for five stocks at NASDAQ from 22 February to 22 December 2008 (210 trading days). We consider the modest ( $r = 0.5$ ) and the conservative risk case  $r = 1$  and set  $\rho = 0.25$

	EACD					WACD				
	AAPL	CSCO	INTC	MSFT	ORCL	AAPL	CSCO	INTC	MSFT	ORCL
$r = 0.5$	0.01	0.00	0.01	-0.12**	0.03	-0.03	0.01	-0.10	-0.13*	-0.06
$r = 1.0$	0.02	0.06	0.03	-0.03	0.00	-0.05	0.08	0.02	-0.01	-0.06

Note: \*10% significance; \*\*5% significance.

### 5. FORECASTING TRADING VOLUMES

Besides providing empirical evidence on the time (in)homogeneity of high-frequency data, our aim is to analyse the potential of the LPA when it comes to out-of-sample forecasts. The most important question is whether the proposed adaptive approach yields better predictions than a (rolling window) approach where the length of the estimation window is fixed on an a priori basis. To set up the forecasting framework as realistic as possible, at each trading minute from 22 February to 22 December 2008, we predict the trading volume over all horizons  $h = 1, 2, \dots, 60$  minutes during the next hour. The predictions are computed using multi-step-ahead forecasts using the currently prevailing MEM parameters and initialized based on the data from the current local window.

The local window is selected according to the LPA approach using  $r \in \{0.5, 1\}$  and  $\rho \in \{0.25, 0.5\}$ . Denoting the corresponding  $h$ -step prediction by  $\hat{y}_{i+h}$ , the resulting prediction error is  $\hat{\varepsilon}_{i+h} = \check{y}_{i+h} - \hat{y}_{i+h}$ , with  $\check{y}_{i+h}$  denoting the observed trading volume. As a competing approach, we consider predictions based on a fixed estimation window covering 1 hour (i.e. 60 observations), 2 hours (i.e. 120 observations), 1 day (i.e. 360 observations) and, alternatively, 1 week (i.e. 1800 observations) yielding predictions  $\tilde{y}_{i+h}$  and prediction errors  $\tilde{\varepsilon}_{i+h} = \check{y}_{i+h} - \tilde{y}_{i+h}$ . To account for the multiplicative impact of intra-day periodicities according to equation (1), we multiply the corresponding forecasts by the estimated seasonality component associated with the previous 30 days.

To test for the significance of forecasting superiority, we apply the Diebold and Mariano (1995) test. Define the loss differential  $d_h$  between the squared prediction errors stemming from both methods given horizon  $h$  and  $n$  observations as  $d_h = \{d_{i+h}\}_{i=1}^n$ , with  $d_{i+h} = \hat{\varepsilon}_{i+h}^2 - \tilde{\varepsilon}_{i+h}^2$ . Then, testing whether one forecasting model yields qualitatively lower prediction errors is performed based on the statistic

$$T_{ST,h} = \left\{ \sum_{i=1}^n \mathbf{I}(d_{i+h} > 0) - 0.5n \right\} / \sqrt{0.25n} \tag{16}$$

which is approximately  $N(0, 1)$  distributed. Our sample covers  $n = 75,600$  trading minutes (corresponding to 210 trading days). To test for quantitative forecasting superiority, we test the null hypothesis  $H_0 : E[d_h] = 0$  using the test statistic

$$T_{DM,h} = \bar{d}_h / \sqrt{2\pi \hat{f}_{d_h}(0)/n} \xrightarrow{\mathcal{L}} N(0, 1) \tag{17}$$

Here,  $\bar{d}_h$  denotes the average loss differential  $\bar{d}_h = n^{-1} \sum_{i=1}^n d_{i+h}$  and  $\hat{f}_{d_h}(0)$  is a consistent estimate of the spectral density of the loss differential at frequency zero. As shown by Diebold and Mariano (1995), the latter can be computed by



$$\hat{f}_{d_h}(0) = (2\pi)^{-1} \sum_{m=-(n-1)}^{n-1} \mathbf{I}\left(\left|\frac{m}{h-1}\right| \leq 1\right) \hat{\gamma}_{d_h}(m) \tag{18}$$

$$\hat{\gamma}_{d_h}(m) = n^{-1} \sum_{i=|m|+1}^n (d_{i+h} - \bar{d}_h) (d_{i+h-|m|} - \bar{d}_h) \tag{19}$$

Figures 9 and 10 display the Diebold–Mariano test statistics  $T_{DM,h}$  against the forecasting horizon  $h$ . The underlying LPA is based on the EACD model with significance level  $\rho = 0.25$ . Negative

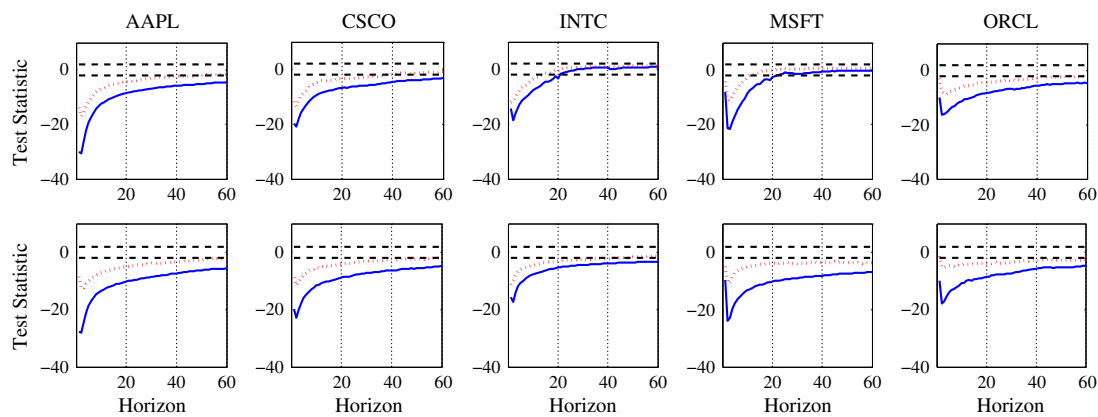


Figure 9. Test statistic  $T_{DM,h}$  across all 60 forecasting horizons for five large companies traded at NASDAQ from 22 February to 22 December 2008 (210 trading days). The dotted curve depicts the statistic based on a test of the LPA against a fixed-window scheme using 360 observations (6 trading hours). The solid curve depicts the statistic based on a test of the LPA against a fixed-window scheme using 1800 observations (30 trading hours). Upper panel: results for the ‘modest risk case’ ( $r = 0.5$ ); lower panel: results for the ‘conservative risk case’ ( $r = 1$ ) given a significance level of  $\rho = 0.25$

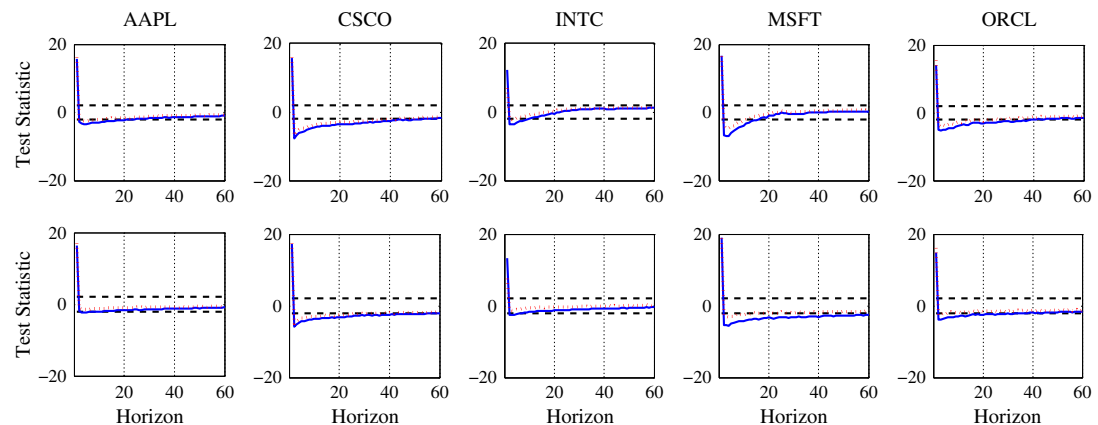


Figure 10. Test statistic  $T_{DM,h}$  across all 60 forecasting horizons for five large companies traded at NASDAQ from 22 February to 22 December 2008 (210 trading days). The dotted curve depicts the statistic based on a test of the LPA against a fixed-window scheme using 60 observations (1 trading hour). The solid curve depicts the statistic based on a test of the LPA against a fixed-window scheme using 120 observations (2 trading hours). Upper panel: results for the ‘modest risk case’ ( $r = 0.5$ ); lower panel: results for the ‘conservative risk case’ ( $r = 1$ ) given a significance level of  $\rho = 0.25$

statistics indicate that the LPA provides smaller forecasting errors. We observe that, in all cases, the fixed-window based forecast is worse than the LPA. The fixed-window approach performs particularly poorly if it utilizes windows covering 1 week or even 1 day of data. These windows seem to be clearly too long to cover local variations in parameters and thus yield estimates which are too strongly smoothed. Our results show that these misspecifications of (local) dynamics result in qualitatively significantly worse predictions. Conversely, shorter (fixed) windows provide clearly better forecasts. Nevertheless, even in this case, the LPA significantly outperforms the fixed-window setting, reflecting the importance of time-varying window lengths.

Analysing the prediction performance in dependence of the forecasting horizon, we observe that LPA-based predictions are particularly powerful over short horizons. The highest LPA overperformance is achieved at horizons of approximately 3–4 minutes. This is not surprising as the local adaptive estimates and thus corresponding forecasts are most appropriate in periods close to the local interval. Conversely, over longer prediction horizons, the advantage of local modelling vanishes as the occurrence of further breakpoints is more likely. We show that the best forecasting accuracy is achieved over horizons of up to 20 minutes. Finally, an important finding is that the results are quite robust with respect to the choice of the modelling risk level  $r$ . This makes the method quite general and not critically dependent on the selection of tuning parameters.

Table VIII summarizes the test statistics  $T_{ST,h}$ . The table reports the correspondingly largest (i.e. least negative) statistics across 30 forecasting horizons. These results clearly confirm the findings reported in Figure 9: the LPA produces significantly smaller (squared) forecasting errors in almost all cases. Moreover, Table VIII confirms the findings above that the forecasting accuracy is widely unaffected by the selection of LPA tuning parameters.

Table VIII. Largest (in absolute terms) test statistic  $T_{ST,h}$  across 30 forecasting horizons as well as EACD and WACD specifications for five companies traded at NASDAQ from 22 February to 22 December 2008 (210 trading days). We compare LPA-implied forecasts with those based on rolling windows using a priori fixed lengths of 1 week, 1 day, 2 hours and 1 hour, respectively. Negative values indicate lower squared prediction errors resulting from the LPA. According to the Diebold–Mariano test (17), the average loss differential is significantly negative in almost all cases (significance level 5%)

	EACD					WACD				
	AAPL	CSCO	INTC	MSFT	ORCL	AAPL	CSCO	INTC	MSFT	ORCL
<i>1 week</i>										
$r = 0.5, \rho = 0.25$	-38.9	-28.6	-24.1	-33.8	-31.4	-22.6	-25.7	-20.2	-26.7	-26.6
$r = 0.5, \rho = 0.50$	-38.7	-28.7	-24.2	-33.8	-31.4	-22.7	-25.5	-20.3	-26.7	-26.6
$r = 1.0, \rho = 0.25$	-40.5	-31.4	-23.3	-39.1	-32.8	-27.9	-30.8	-21.5	-31.3	-29.8
$r = 1.0, \rho = 0.50$	-40.4	-31.3	-23.3	-39.0	-32.9	-28.1	-30.8	-21.5	-31.5	-29.7
<i>1 day</i>										
$r = 0.5, \rho = 0.25$	-10.8	-6.0	-13.1	-5.7	-15.1	-6.4	-3.5	-6.1	-4.9	-12.6
$r = 0.5, \rho = 0.50$	-10.6	-6.0	-12.8	-5.5	-15.0	-6.3	-3.2	-6.2	-4.8	-12.7
$r = 1.0, \rho = 0.25$	-6.9	-8.6	-8.7	-4.4	-12.9	-4.1	-5.1	-6.5	-4.2	-11.5
$r = 1.0, \rho = 0.50$	-7.1	-8.6	-8.8	-4.4	-13.0	-3.9	-5.2	-6.5	-4.1	-11.4
<i>2 hours</i>										
$r = 0.5, \rho = 0.25$	-11.3	-3.4	-14.1	-11.8	-24.0	-5.6	-5.9	-11.5	-11.2	-20.3
$r = 0.5, \rho = 0.50$	-11.2	-3.5	-14.1	-11.7	-23.9	-5.6	-5.8	-11.4	-11.2	-20.4
$r = 1.0, \rho = 0.25$	-5.9	2.0	-13.4	-5.0	-22.4	-5.0	-1.1	-12.5	-7.6	-20.6
$r = 1.0, \rho = 0.50$	-5.9	2.1	-13.5	-5.0	-22.4	-5.1	-1.1	-12.5	-7.6	-20.5
<i>1 hour</i>										
$r = 0.5, \rho = 0.25$	-9.3	-6.6	-10.5	-2.0	-27.2	-4.9	-8.5	-10.4	-0.5	-24.7
$r = 0.5, \rho = 0.50$	-9.2	-6.6	-10.4	-2.0	-27.1	-4.8	-8.6	-10.4	-0.4	-24.7
$r = 1.0, \rho = 0.25$	-3.3	-0.9	-8.7	4.5	-27.7	-3.4	-3.0	-9.4	4.7	-25.1
$r = 1.0, \rho = 0.50$	-3.3	-0.7	-8.7	4.5	-27.7	-3.4	-2.9	-9.7	4.9	-25.0

By depicting the ratio of root mean squared errors

$$\sqrt{n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_{i+h}^2} / \sqrt{n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_{i+h}^2}$$

In Figure 11, we provide deeper insights into the forecasting performance of the two competing approaches over time and over the sample. In most cases, the ratio is clearly below one and thus also indicates a better forecasting performance of the LPA method. This is particularly true during the last months and thus the height of the financial crisis in 2008. During this period, market uncertainty has been high and trading activity has been subject to various information shocks. Our results show that the flexibility offered by the LPA is particularly beneficial in such periods, whereas fixed-window approaches tend to perform poorly.

Figure 12 shows the ratio of root mean squared errors in dependence of the length of the forecasting horizon (in minutes). It turns out that the LPA's overperformance is strongest over horizons between

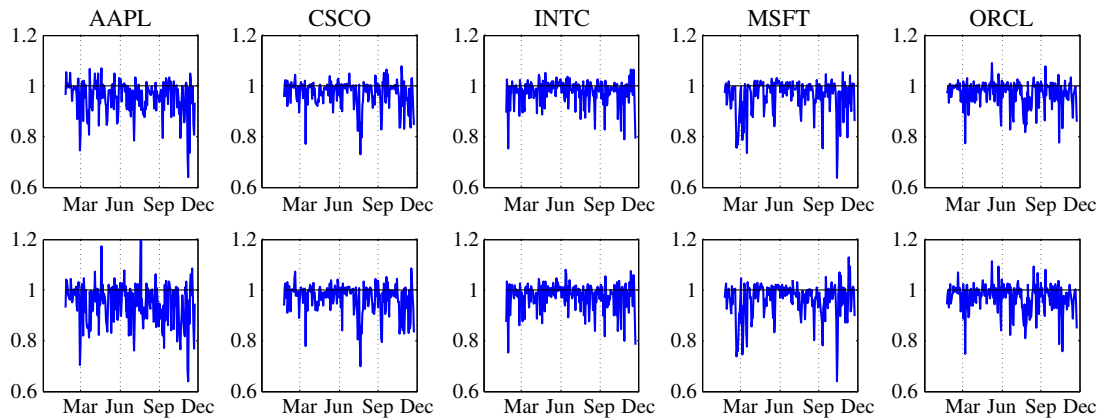


Figure 11. Ratio between the RMSPEs of the LPA and of a fixed-window approach (covering 6 trading hours) over the sample from 22 February to 22 December 2008 (210 trading days). Upper panel: results for the underlying (local) EACD model; lower panel: results for the underlying (local) WACD model

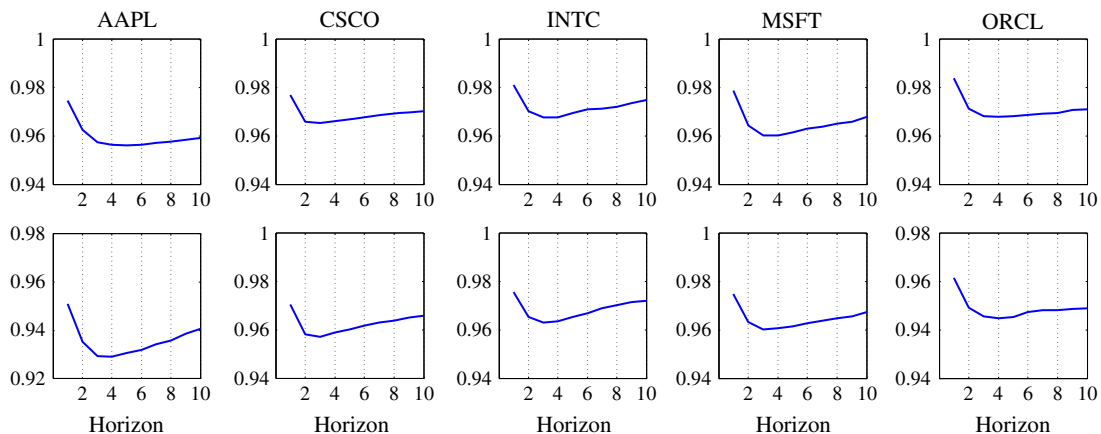


Figure 12. Ratio between the RMSPEs of the LPA and of a fixed-window approach (covering 6 trading hours) over the sample from 22 February to 22 December 2008 (210 trading days). Upper panel: EACD model; lower panel: WACD model

2 and 4 minutes. Over these intervals, the effects of superior (local) estimates of MEM parameters fully pay out. Over longer horizons, differences in prediction performance naturally shrink as forecasts converge to unconditional averages.

## 6. CONCLUSIONS

We propose a local adaptive multiplicative error model for financial high-frequency variables. The approach addresses the inherent inhomogeneity of parameters over time and is based on local window estimates of MEM parameters. Adapting the local parametric approach (LPA) by Spokoiny (1998) and Mercurio and Spokoiny (2004), the length of local estimation intervals is chosen by a sequential testing procedure. Balancing modelling bias and estimation (in)efficiency, the approach provides the longest interval of parameter homogeneity which is used for modelling and forecasting.

Applying the proposed approach to the high-frequency series of 1-minute cumulative trading volumes based on several NASDAQ blue chip stocks, we can conclude as follows. First, MEM parameters reveal substantial variations over time. Second, the optimal length of local intervals varies between 1 and 6 hours. Nevertheless, as a rule of thumb, local intervals of around 4 hours are suggested. Third, the local adaptive approach provides significantly better out-of-sample forecasts than competing approaches using a priori fixed lengths of estimation intervals. This result demonstrates the importance of an adaptive approach. Finally, we show that the findings are robust with respect to the choice of LPA steering parameters controlling modelling risk.

As the stochastic properties of cumulative trading volumes are similar to those of other (persistent) high-frequency series, our findings are likely to be carried over to, for instance, the time between trades, trade counts, volatilities, bid–ask spreads and market depth. Adaptive techniques thus constitute a powerful device to improve high-frequency forecasts and to gain deeper insights into local variations of model parameters.

## ACKNOWLEDGEMENTS

Financial support from the Deutsche Forschungsgemeinschaft via CRC 649 ‘Economic Risk’, Humboldt-Universität zu Berlin, is gratefully acknowledged. Hautsch acknowledges research support by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF). We thank the reviewers for their constructive comments and helpful suggestions.

## REFERENCES

- Brownlees CT, Cipollini F, Gallo GM. 2011. Intra-daily volume modeling and prediction for algorithmic trading. *Journal of Financial Econometrics* **9**(3): 489–518.
- Chen Y, Härdle W, Pigorsch U. 2010. Localized realized volatility. *Journal of the American Statistical Association* **105**(492): 1376–1393.
- Čížek P, Härdle WK, Spokoiny V. 2009. Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models. *Econometrics Journal* **12**: 248–271.
- Clark TE, McCracken MW. 2009. Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review* **50**(2): 363–395.
- Diebold F, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**(3): 253–263.
- Engle RF. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4): 987–1007.
- Engle RF. 2002. New frontiers for ARCH models. *Journal of Applied Econometrics* **17**: 425–446.
- Engle RF, Rangel JG. 2008. The spline-GARCH model for low-frequency volatility and its global macroeconomic causes. *Review of Financial Studies* **21**: 1187–1222.

- Engle RF, Russell JR. 1998. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* **66**(5): 1127–1162.
- Gallant AR. 1981. On the bias of flexible functional forms and an essentially unbiased form. *Journal of Econometrics* **15**: 211–245.
- Groß-Klußmann A, Hautsch N. 2011. When machines read the news: using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance* **18**: 321–340.
- Hautsch N. 2012. *Econometrics of Financial High-Frequency Data*. Springer: Berlin.
- Hautsch N, Malec P, Schienle M. 2014. Capturing the zero: a new class of zero-augmented distributions and multiplicative error processes. *Journal of Financial Econometrics* **12**(1): 89–121.
- Hujer R, Vuletić S, Kokot S. 2002. The Markov switching ACD model. Working paper—finance and accounting, no. 90, Johann Wolfgang Goethe-University, Frankfurt.
- Manganelli S. 2005. Duration, volume and volatility impact of trades. *Journal of Financial Markets* **8**: 377–399.
- Meitz M, Teräsvirta T. 2006. Evaluating models of autoregressive conditional duration. *Journal of Business and Economic Statistics* **24**(1): 104–124.
- Mercurio D, Spokoiny V. 2004. Statistical inference for time-inhomogeneous volatility models. *Annals of Statistics* **32**(2): 577–602.
- Pesaran M, Timmermann A. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137**: 134–161.
- Spokoiny V. 1998. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Annals of Statistics* **26**(4): 1356–1378.
- Spokoiny V. 2009. Multiscale local change point detection with applications to value-at-risk. *Annals of Statistics* **37**(3): 1405–1436.
- Tong H. 1990. *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press: Oxford.
- Zhang MY, Russell JR, Tsay RS. 2001. A nonlinear autoregressive conditional duration model with applications to financial transaction data. *Journal of Econometrics* **104**: 179–207.



Contents lists available at ScienceDirect

## Insurance: Mathematics and Economics

journal homepage: [www.elsevier.com/locate/ime](http://www.elsevier.com/locate/ime)

## State price densities implied from weather derivatives

Wolfgang Karl Härdle<sup>a,b</sup>, Brenda López-Cabrera<sup>a</sup>, Hwei-Wen Teng<sup>c,\*</sup><sup>a</sup> C.A.S.E Center for Applied Statistics and Economics and Ladislaus von Bortkiewicz Chair of Statistics, Humboldt-Universität zu Berlin, Germany<sup>b</sup> Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore<sup>c</sup> Graduate Institute of Statistics, National Central University, Taoyuan City, Taiwan

## ARTICLE INFO

## Article history:

Received July 2014

Received in revised form

April 2015

Accepted 4 May 2015

Available online 14 May 2015

## Keywords:

Weather derivatives

Temperature derivatives

HDD

CDD

State Price Density

Quadrature

Bayesian

Data sparsity

## ABSTRACT

A State Price Density (SPD) is the density function of a risk neutral equivalent martingale measure for option pricing, and is indispensable for exotic option pricing and portfolio risk management. Many approaches have been proposed in the last two decades to calibrate a SPD using financial options from the bond and equity markets. Among these, non and semiparametric methods were preferred because they can avoid model mis-specification of the underlying. However, these methods usually require a large data set to achieve desired convergence properties. One faces the problem in estimation by e.g., kernel techniques that there are not enough observations locally available. For this situation, we employ a Bayesian quadrature method because it allows us to incorporate prior assumptions on the model parameters and hence avoids problems with data sparsity. It is able to compute the SPD of both call and put options simultaneously, and is particularly robust when the market faces the data sparsity issue. As illustration, we calibrate the SPD for weather derivatives, a classical example of incomplete markets with financial contracts payoffs linked to non-tradable assets, namely, weather indices. Finally, we study related weather derivatives data and the dynamics of the implied SPDs.

© 2015 Elsevier B.V. All rights reserved.

## 1. Introduction

A State Price Density (SPD) is the density function of a Risk Neutral (RN) equivalent martingale measure for option pricing, and it is a measure more tied to uncertainty than to volatility and it is indispensable for (exotic) option pricing and portfolio risk management. It does not only reflect a risk-adaptive behavior of investors based on historical assessment of the futures market, but it also gives insights about the preferences and risk aversion of a representative agent, see for example [Ait-Sahalia and Lo \(2000\)](#), [Jackwerth and Rubinstein \(1996\)](#) and [Rosenberg and Engle \(2002\)](#).

Consider a European call option with maturity date  $T$  and strike price  $K$ . Under the non-arbitrage principle, its price at  $t$  can be given as:

$$C(K) = e^{-r\tau} \int \max(x - K, 0) f(x) dx \quad (1)$$

where  $r$  is the risk-free interest rate,  $\tau$  time to maturity and  $f(x)$  is the defined SPD. The advantage of extracting the SPD directly from market prices is that volatility and other moments can easily

be calculated using this SPD independent of any particular pricing model.

There are many approaches to calibrate the SPD using financial options from the bond and equity markets. Assuming a Black and Scholes (B&S) model implies that the RN measure is a lognormal distribution which may result in severe bias of the SPD estimation since certain volatility properties are not correctly reflected. As observed by [Breedon and Litzenberger \(1978\)](#), the SPD of any risky asset can be derived as the second derivative with respect to the strike price of an estimate of the pricing function  $C$ . A number of econometric techniques have been developed to address this calibration issue. The most notable examples include the stochastic volatility models and the GARCH models. [Derman and Kani \(1994\)](#), [Dupire \(1994\)](#) and [Rubinstein \(1994\)](#) implied SPDs using binomial trees, hence avoiding too strong stochasticity assumption like e.g., Geometric Brownian motion. Others like [Abadir and Rockinger \(2003\)](#) use hypergeometric distributions. Although useful in a variety of contexts, these (parametric) models are still susceptible to model specification.

Various non-parametric models have been employed to overcome this problem. [Ait-Sahalia and Lo \(1998\)](#) introduce a semi-parametric alternative where the volatility of the B&S formulation is modeled non-parametrically. From a statistical point of view, estimating the SPD becomes estimating the second derivative of a

\* Corresponding author.

E-mail address: [venteng@gmail.com](mailto:venteng@gmail.com) (H.-W. Teng).

regression function, but the SPD needs to be a proper density function (non negative and integrates to one). This dictates that the price is decreasing and convex in terms of the strike price. How to impose these constraints presents the main difficulties of direct applications of nonparametric regression. Ait-Sahalia and Duarte (2003), Yatchew and Härdle (2006) and Härdle and Hlávka (2009) stress the importance of enforcing such shape constraints. Fan and Mancini (2009) use a non-parametric technique to estimate the state price distribution but not the density because the former is easier to estimate. Giacomini et al. (2008) use mixtures of scales and shifted  $t$ -distributions, while Yuan (2009) uses a mixture of lognormals. Curve fitting method have been presented in Rubinstein (1994) and Jackwerth and Rubinstein (1996). Liechty and Teng (2009) introduce the Bayesian quadrature model, where both the locations and weights of the support points for approximating the SPD are random variables. Most nonparametric methods require a rich body of data to achieve desired convergence properties. The main goal of this paper is to infer the SPD from markets, where trading activities are less frequently occurred.

For this purpose, we employ a Bayesian quadrature method as a calibration method for the SPD from option prices, because it allows us to incorporate prior assumptions on the model parameters and hence avoids problems with data sparsity. This approach takes a prior distribution which can be parametric (e.g. lognormal) or a uniform density. The posterior distribution of the SPD is calibrated to market data. This method is a special case of a mixture model, where the component densities are point measures.

The novelty of the Bayesian quadrature approach relies on the fact that it uses unequal weights and is in a Bayesian framework. Approximating the state price density with weighted sum of  $\delta$ -functions produces good model fitting by using a parsimonious model. Bayesian inference gives a straightforward probabilistic framework and provides reasonable credible regions for the implied state price density, which can be further used for various purposes such as hedging and pricing.

We show that the proposed method has some advantages over other nonparametric methods: (1) it considers the locations and weights of the support points in the finite representation of the SPD as random variables, (2) it is parsimonious and allows for statistical inference, it enables us to construct credible regions for the current value of the SPD (3) it is computationally efficient in the sense that a Markov chain Monte Carlo algorithm with Gibbs sampler can be adopted, so that no additional tuning procedures are required for exploring the posterior distribution and (4) it is robust even if the market faces data sparsity issues. (5) These classes of Risk Neutral probabilities do not stem from market-risk-price assumptions.

We conduct our empirical analysis based on weather derivative (WD) data traded at the Chicago Mercantile Exchange (CME). WDs are newly developed financial instruments. Key features of weather derivatives are that the underlying process, i.e., temperature or rainfall index is not tradable and cannot be replicated by other risk factors (Benth et al., 2007; Härdle and López-Cabrera, 2012; López-Cabrera et al., 2013). Consequently, the Black–Scholes formula is unsuitable since an essential element of it is the tradability of the underlying. In addition, the temperature index shows apparent seasonality and it is determined by physical phenomena. An interesting feature is that weather futures and options are rarely traded and traded only at a few strike prices compared with other more frequently traded equity markets. The CME (the official WD platform) provides closing prices, which are however not the real trading prices negotiated by the market participants. The SPD enables to price options with complicated payoff functions simply by numerical integration of the payoff with respect to this density. However, data sparsity makes the SPD estimation a statistical challenge. In addition, we study the dynamics of the SPD which

provides useful insight into the economic behavior of agents sensitive to weather conditions and the time inhomogeneity of the market.

This paper is structured as follows. Section 2 describes the quadrature approach and its comparison to other popular SPD density estimation methods. Section 3 conducts the empirical analysis of SPDs from CME weather option data, studies the dynamics of the SPD weather type, and gives economic interpretations from the implied SPD. In Section 4, we address the data sparsity issue by addressing why other nonparametric methods fail particularly when options with only a few strike prices are traded. Section 5 concludes the paper. All quotations of currency in this paper will be in USD and therefore we will omit the explicit notion of the currency. All the SPDs computations were carried out in Matlab version 7.6. The option data on temperature indices were obtained from CME and are also available from the research data center of the CRC 649 “Economic Risk”.

## 2. The Bayesian quadrature method

Options are contingent claims on an underlying asset. Plain vanilla option is of either put or call type with a fixed maturity, i.e., the value of the underlying is compared to a strike price  $K$  at maturity  $T$ . Let  $x$  denote the underlying asset's price at maturity (in our application this will be equivalent to futures prices on weather indexes). For a call option, one has the payoff  $\max(x - K, 0)$  and for a put  $\max(K - x, 0)$ . If we denote a put as  $i = 1$  and a call with  $i = 2$ , and observed strike prices  $E_{ij}$  for  $i = 1, 2$  and  $j = 1, \dots, N_i$  indexing all possible strike prices on any given day  $t$ , then the payoff function at maturity, denoted by  $\wp_{ij}(x)$ , can be represented by one formula,

$$\wp_{ij}(x) = (-1)^i (x - E_{ij}) \mathbf{I} \{ (-1)^i (x - E_{ij}) > 0 \} (x),$$

where  $\mathbf{I}\{A\}$  is an indicator function for a set  $A$ . Let  $t$  be the current time. The fair option price is given as (1) as the discounted value of the expected payoff function:

$$C_{ij} = \exp(-r\tau) E^Q[\wp_{ij}(x)],$$

where  $\tau = T - t$  is the time to maturity and  $E^Q[\cdot]$  is the expectation operator taken under the risk-neutral measure. The density  $f(x)$  under this risk-neutral measure is the defined SPD. When the SPD  $f(x)$  exists, this equals:

$$C_{ij} = \exp(-r\tau) \int \wp_{ij}(x) f(x) dx. \quad (2)$$

The left hand side of (2) is observed on the market for different payoff types depending on put/call ( $i = 1, 2$ ), strike price  $E_{ij}$ , and time to maturity  $\tau$ . The interest of statistical calibration is to infer the SPD  $f(x)$  from a set of observed option prices.

### 2.1. The quadrature method

The word “quadrature” means a numerical method to approximate an integral either analytically or numerically, see Ueberhuber (1997) for example. In this research, we work the adverse way, since the interest is to infer the unknown density from the observed integrals (option prices). Define the  $\delta$ -function  $\delta_w(\cdot)$  as a unit point measure at the location  $s$  by

$$\delta_s(x) = \mathbf{I}\{s = x\}.$$

The basic idea of the quadrature method is to approximate the SPD  $f(x)$  by  $f_N(x|w, \theta)$ , a weighted sum of  $\delta$ -functions:

$$f_N(x|w, \theta) = w_1 \delta_{\theta_1}(x) + \dots + w_N \delta_{\theta_N}(x), \quad (3)$$

with unknown locations  $\theta = (\theta_1, \dots, \theta_N)^\top$  and weights  $w = (w_1, \dots, w_N)^\top$ . Here,  $N$  is a non-negative integer (smoothing)

parameter. To produce a legitimate probability density, the locations  $\theta$  are constrained to be non-negative quantities, and the weights  $w$  are constrained to be nonnegative quantities and sum up to one. From a modeling perspective, the quadrature method (3) can be seen as a finite mixture distribution with the point measure as the component density. Fig. 1 illustrates (3) for  $N = 5$ .

The option price (2) under  $f_N(x|w, \theta)$  is:

$$C_{ij}^N(w, \theta) = \exp(-r\tau) \sum_{n=1}^N w_n \phi_{ij}(\theta_n). \tag{4}$$

Note that (4) is an approximation to (2) and the aim of calibration is to extract  $(w, \theta)$  by matching  $C_{ij}^N(w, \theta)$  to the observed option prices. More specifically, a call option price calculated with (3) is:

$$C_{2j}^N(w, \theta) = \exp\{-r\tau\} \sum_{n=1}^N w_n \max(\theta_n - E_{ij}, 0), \tag{5}$$

whereas a put option price under the quadrature method is:

$$C_{1j}^N(w, \theta) = \exp\{-r\tau\} \sum_{n=1}^N w_n \max(E_{ij} - \theta_n, 0). \tag{6}$$

2.2. Bayesian modeling and computation

Empirical observations show that options having higher prices usually have higher price variation, see Ghysels et al. (1995) and Ghysels et al. (1997). Hence for the calibration task as a variance stabilizing transformation, we consider the logarithm of option prices. The observations  $y_{ijk}$  are perturbations of the model option price  $C_{ij}^N(w, \theta)$ :

$$\log y_{ijk} = \log C_{ij}^N(w, \theta) + \varepsilon_{ijk} \tag{7}$$

for  $i = 1, 2, j = 1, \dots, N_i, k = 1, \dots, N_{ij}$ , where the error  $\varepsilon_{ijk} \sim N(0, \sigma^2)$ .  $\varepsilon_{ijk}$  is attributed to market friction and the approximation discrepancy (Garcia et al., 2010; Renault, 1997). In Section 3, residual analysis of our empirical studies will support this error assumption.

These parameters,  $w, \theta$ , and  $\sigma^2$ , are estimated in a Bayesian framework instead of a maximum likelihood method. Following (7), the likelihood is

$$L(y|w, \theta, \sigma^2) = \prod_{i=1}^2 \prod_{j=1}^{N_i} \prod_{k=1}^{N_{ij}} (2\pi\sigma^2)^{-\frac{1}{2}} \times \exp\left[-\frac{\{\log y_{ijk} - \log C_{ij}^N(w, \theta)\}^2}{2\sigma^2}\right]. \tag{8}$$

A natural prior distribution for the weights  $w$  is the Dirichlet distribution, which ensures  $w$  being positive and summing up to one. The Dirichlet distribution with parameter  $\gamma = (\gamma_1, \dots, \gamma_N)^T$  has the density function,

$$f(w|\gamma) = \frac{1}{B(\gamma)} \prod_{n=1}^N w_n^{\gamma_n-1} \tag{9}$$

for  $w_n > 0, n = 1, \dots, N$ , and  $w_1 + \dots + w_N = 1$ . The normalizing constant  $B(\gamma)$  is defined as

$$B(\gamma) = \frac{\prod_{n=1}^N \Gamma(\gamma_n)}{\Gamma\left(\sum_{n=1}^N \gamma_n\right)}$$

where  $\Gamma(\cdot)$  is the gamma function (Chen and Shao, 1997).

Let  $K_{\min}$  and  $K_{\max}$  denote the minimum and maximum of the observed strike prices  $E_{ij}$ , respectively. To avoid label switching

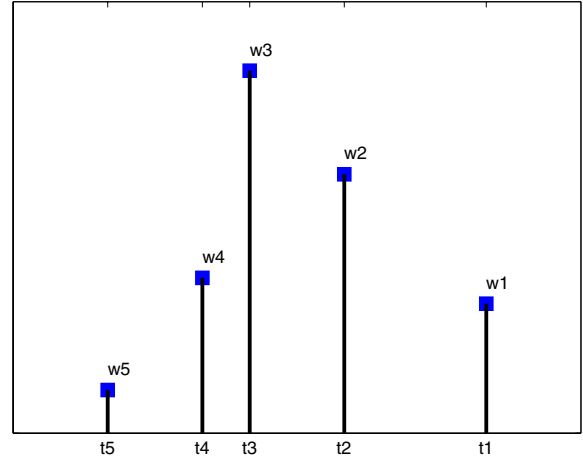


Fig. 1. The SPD  $f_N(x|w, \theta)$  from (3) for  $N = 5$ .

problems for  $\theta$ , we assume that the locations are ordered, i.e.,  $\theta_1 \leq \dots \leq \theta_N$ . Moreover, to avoid model option prices in (4) being zeros, assume a priori that the smallest location,  $\theta_1$ , is less than the minimum of the observed strike price, and that the largest location,  $\theta_N$ , is larger than the maximum of the observed strike prices. Therefore, we assume that the distribution of the locations  $\theta$  is uniformly distributed over the set  $\{\theta_1 \leq \theta_2 \leq \dots \leq \theta_N, \theta_1 < K_{\min}, \theta_N > K_{\max}\}$ :

$$f(\theta|K_{\min}, K_{\max}) \propto \mathbf{I}\{\theta_1 \leq \dots \leq \theta_N, \theta_1 < K_{\min}, \theta_N > K_{\max}\}(\theta). \tag{10}$$

For simplicity, we consider an inverse-gamma distribution with shape parameter  $\alpha$  and scale parameter  $\beta$  as a prior distribution for  $\sigma^2$ , denoted by  $\sigma^2 \sim IG(\alpha, \beta)$ . The prior density of  $\sigma^2$  is

$$f(\sigma^2|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right). \tag{11}$$

Putting things together allows a conjugate prior for  $\sigma^2$ , as described in Casella and Berger (2001).

Note that (9)–(11) can be changed in cases where appropriate information is available. Bayesian inference for the parameters of interest is based on the posterior distribution of  $w, \theta$ , and  $\sigma^2$ :

$$f(w, \theta, \sigma^2|y, \alpha, \beta, \gamma, K_{\min}, K_{\max}) \propto L(y|w, \theta, \sigma^2) f(w|\gamma) f(\theta|K_{\min}, K_{\max}) f(\sigma^2|\alpha, \beta). \tag{12}$$

Because of the complexity of (12), it is difficult to derive a closed-form formula for the posterior distribution (Liechty and Teng, 2009). The Markov chain Monte Carlo (MCMC) simulation is therefore used to sample  $w, \theta$ , and  $\sigma^2$ . Because of the monotonicity of parameters  $w$  and  $\theta$  in (4), an MCMC algorithm with slice samplers can be used to avoid manual tuning procedures in the MCMC simulation. In the following, we summarize major steps to run the MCMC algorithm. Let  $U(A)$  denote the uniform distribution on the set  $A$ .

1. Start  $w, \theta$ , and  $\sigma^2$  randomly.
2. At each iteration, repeat the following steps until the samples appear to converge.
  - (a) Sample  $w_n \sim U(T_n)$  for  $n = 1, \dots, N - 1$ , where  $T_n$  is a properly derived open interval. Set  $w_N = 1 - w_1 - \dots - w_{N-1}$ .
  - (b) Sample  $\theta_n \sim U(S_n)$  for  $n = 1, \dots, N$ , where  $S_n$  is a properly derived open interval.



**Table 1**  
The volume for HDD–CDD monthly, seasonal strips and average temperature products in each US city.

Index	City	Future				Option					Avg	MS (%)	Rank
		HDD monthly	CDD monthly	HDD strips	CDD strips	HDD monthly	CDD monthly	HDD strips	CDD strips				
1	Atlanta	49621	35567	14400	2150	50	56431	11647	117165	71950	0	9.44	3
2	Baltimore	6633	3545	600	700	0	2600	100	12500	1100	0	0.73	16
3	Boston	24178	19066	2200	1150	0	11029	550	42174	19450	0	3.15	13
4	Chicago	90585	54950	3975	2800	0	39676	19300	107616	67725	0	10.17	2
5	Cincinnati	50155	38035	2967	1700	455	29280	28910	73255	74975	0	7.89	4
6	Colorado Springs	1936	1450	0	0	0	15025	8750	0	0	0	0.71	17
7	Dallas	27206	55540	3700	1961	200	13085	39775	47450	94850	0	7.47	5
8	Des Moines	40929	30510	3190	1450	50	38631	4460	64790	60900	0	6.44	6
9	Detroit	2185	351	50	50	0	0	0	0	0	0	0.07	23
10	Houston	16901	18229	1400	1700	0	3700	5000	52950	33950	0	3.52	12
11	Jacksonville	100	1600	0	0	0	0	16575	0	0	0	0.48	19
12	Kansas City	36513	23145	1325	1350	1100	11025	7200	45050	33750	0	4.22	10
13	Las Vegas	12680	26635	325	1650	0	3100	14200	34600	76650	0	4.47	9
14	Little Rock	120	105	0	0	0	0	12250	0	0	0	0.33	20
15	Los Angeles	100	400	0	0	0	0	50	0	0	0	0.01	24
16	Minneapolis	50085	27955	2150	1500	0	18206	3850	63350	34000	0	5.29	8
17	New York	187264	154605	6700	4860	0	90620	35175	141850	136350	0	19.93	1
18	Philadelphia	16441	34449	2300	2250	150	6408	18210	56000	76150	0	5.59	7
19	Portland	10329	10855	725	450	0	1720	450	48200	76450	0	3.92	11
20	Raleigh Durham	550	1500	0	0	0	23700	0	0	0	0	0.68	18
21	Sacramento	6383	23401	550	750	0	2850	1675	16200	48000	0	2.63	14
22	Salt Lake City	739	504	150	0	0	0	0	4500	3500	0	0.25	22
23	Tucson	7283	16965	350	750	0	2700	3010	28750	27800	0	2.30	15
24	Washington	550	25	250	0	0	650	1500	6650	2000	0	0.31	21

(c) Sample

$$\sigma^2 \sim IG \left( \sum_{i=1}^2 \sum_{j=1}^{N_i} \alpha + M/2, \beta + \sum_{i=1}^2 \sum_{j=1}^{N_i} \sum_{k=1}^{N_{ij}} (\log y_{ijk} - \log C_{ij}^N(w, \theta))^2 \right) / 2,$$

where  $M = \sum_{i=1}^2 \sum_{j=1}^{N_i} \sum_{k=1}^{N_{ij}} 1$  is the number of observed options.

The derivations of open intervals  $T_n$  and  $S_n$  are rather lengthy and are hence omitted here for brevity. Please refer to Liechty and Teng (2009) for details.

2.3. Kernel smoothing density estimate of the quadrature method

The density  $\hat{f}_N(x|w, \theta)$  from (3) is a weighted sum of  $\delta$  functions and hence is not a continuous density. However, in many cases, it is interesting to visualize the SPD as a smoothed density. The kernel density for a set of  $M$  observed points  $\vartheta = (\vartheta_1, \dots, \vartheta_M)^T$  is:

$$\hat{f}(x|\vartheta) = \int \hat{g}(u)K_h(x - u)du = \frac{1}{M} \sum_{m=1}^M K_h(x - \vartheta_m) \tag{13}$$

where  $K_h(\cdot) = h^{-1}K(\cdot/h)$  is a kernel function with a bandwidth  $h$  and  $\hat{g}(u)$  the sum  $\delta$ -functions

$$\hat{g}(u) = M^{-1} \sum_{m=1}^M \delta_{\vartheta_m}(u)$$

with locations  $\vartheta$ . Obviously, different values of  $h$  will change the appearance of  $\hat{f}(x|\vartheta)$ . Silverman’s rule of thumb suggests a bandwidth

$$h_C = 1.06\hat{\sigma}M^{-1/5} \tag{14}$$

where  $\hat{\sigma}$  is the sample standard deviation of  $\vartheta$  and a normal kernel  $K = \varphi$  the pdf of  $N(0, 1)$  (Silverman, 1986).

Note that each  $\vartheta_m$  for  $m = 1, \dots, M$  appears with equal probability  $1/M$ . However, in the Bayesian quadrature method,  $\theta_n$

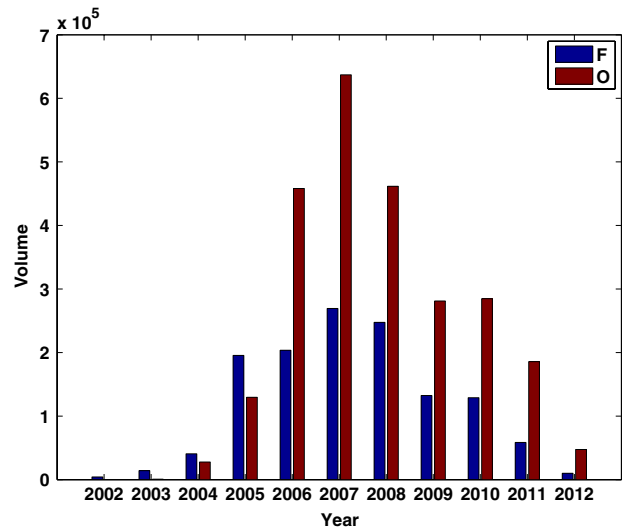
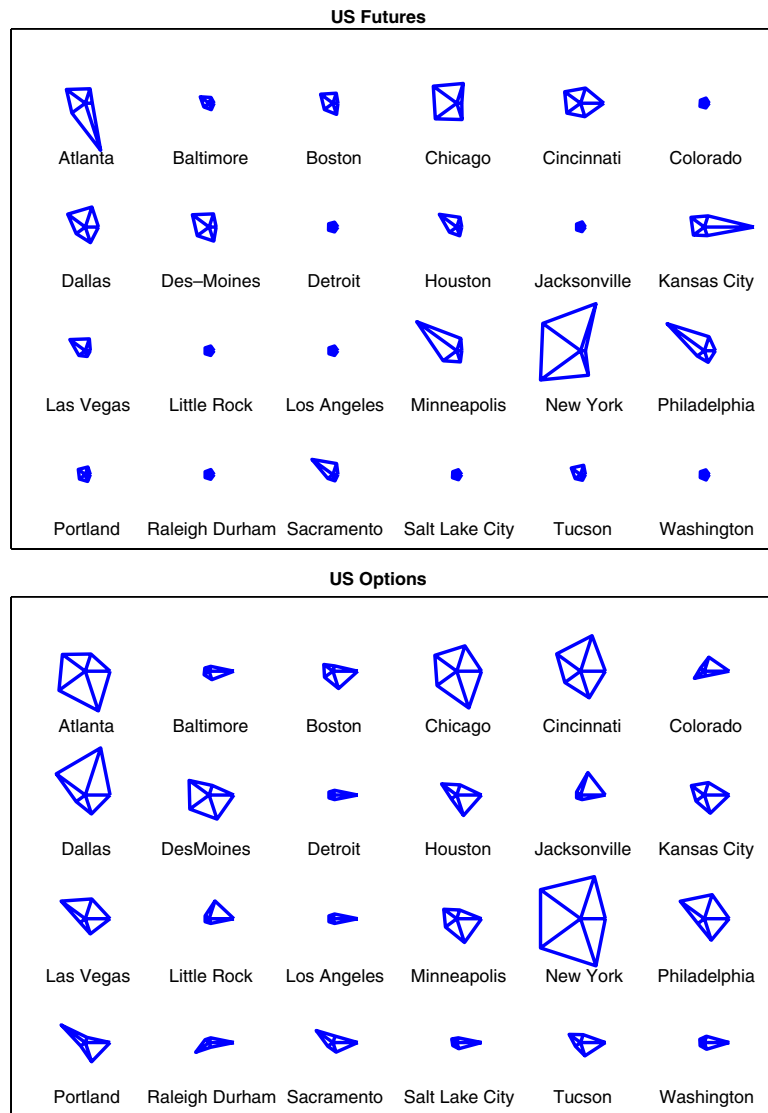


Fig. 2. The volume for US temperature futures (F) and options (O).

appears with probability  $w_n$ , for  $n = 1, \dots, N$ . Therefore, we need to adjust the sample size and use  $\hat{g}(u) = \sum_{n=1}^N w_n \delta_{\theta_n}(u)$  instead. The smoothed density version of (3) becomes

$$f_N^s(x|w, \theta) = \sum_{n=1}^N w_n K_h(x - \theta_n). \tag{15}$$

To apply Silverman’s rule in the case of unequal weights in (15), we round off each  $w_n$  to the second decimal and adjust the sample size to be 100. The smoothed SPD appears to be reasonable. Indeed, it is possible to consider a more precise approximation: Round off each  $w_n$  to the  $q$ -th decimal, and set the sample size  $M$  to be  $10^q$ . In the  $i$ th swipe of the MCMC algorithm, we obtain  $w^{(i)}$  and  $\theta^{(i)}$  and the smoothed SPD  $f_N^s(x|w^{(i)}, \theta^{(i)})$ . We then report the posterior mean and 90% credible regions of the smoothed SPD based on  $\hat{f}(x|w^{(i)}, \theta^{(i)})$  point-wisely.



**Fig. 3.** Star plots representing the volume for US temperature contracts (HDD–CDD monthly, HDD–CDD seasonal strips, and weekly average) futures (upper panel) and options (lower panel) for each city. Each city is represented as a star whose *i*th spoke is proportional in length to the volume size of *i*th product (HDD Monthly, CDD Monthly, HDD Strips, CDD strips, Average) of the observed city.

As a remark, the bandwidth can be adjusted to other kernels by a canonical bandwidth, Härdle et al. (2004). For example for the quartic kernel:

$$K(u) = \frac{15}{16}(1 - u^2)^2 \mathbf{1}\{|u| \leq 1\}, \tag{16}$$

Silverman's rule of thumb  $h_G$  transforms into:

$$h_{QUA} = 2.62 \cdot h_G \tag{17}$$

### 3. Empirical analysis

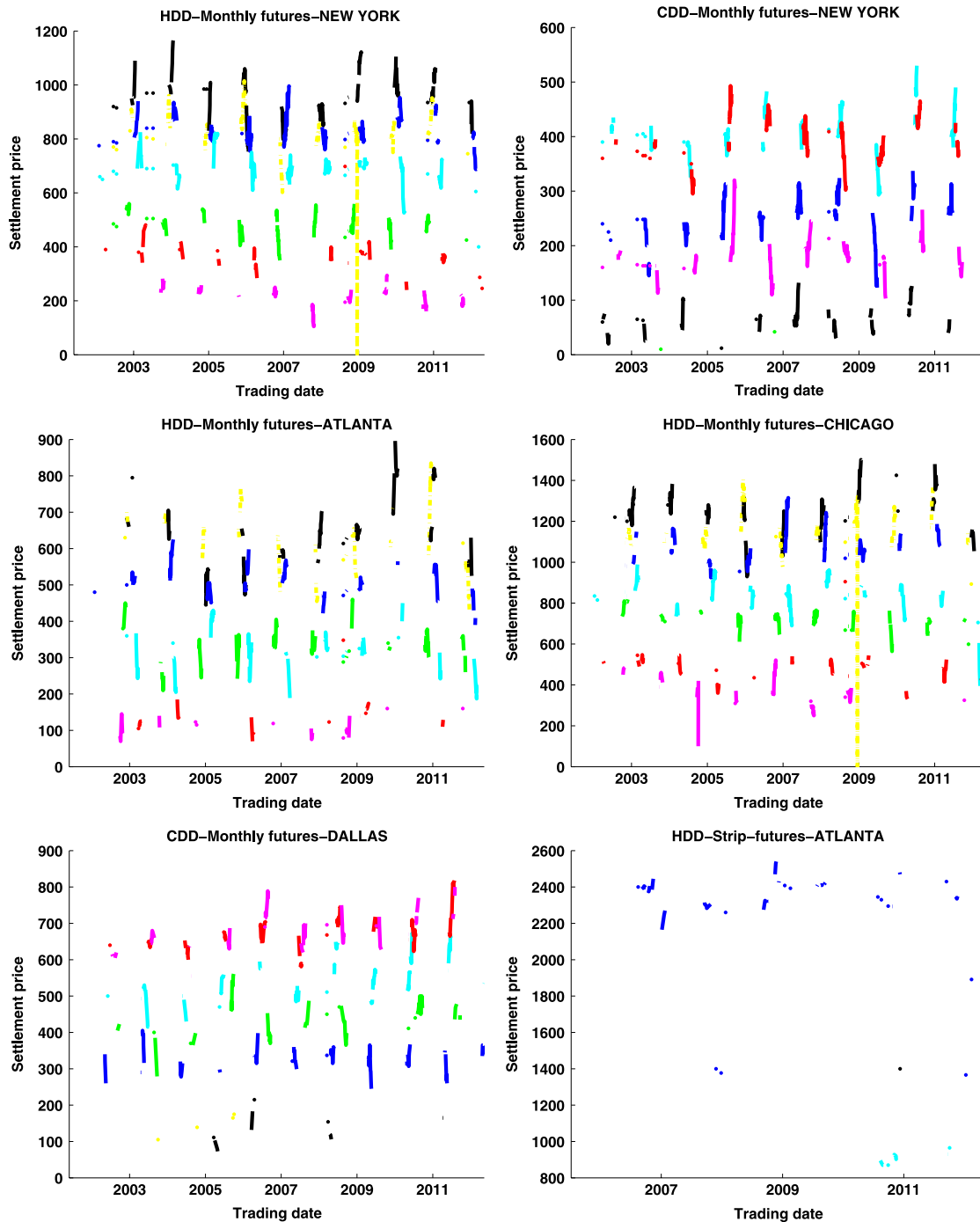
This section introduces the weather derivatives (WD) market and presents an overview on WD data. One major feature of the WD market is data sparsity, which makes most existing methods for estimating the SPD challenging and difficult. We then apply the described Bayesian quadrature technique to estimate the implied SPD on WD data, conduct an out-of-sample analysis, and study its dynamics.

#### 3.1. Weather derivatives

WDs are financial contracts designed to hedge weather risk. The most common contracts traded at CME are based on temperature indices linked to the temperature at time  $t$ , denoted by  $T_t$ . These are the Heating Degree Days (HDD), the Cooling Degree Days (CDD), and the cumulative average temperature (CAT):

$$\begin{aligned} HDD(\tau_1, \tau_2) &= \sum_{t=\tau_1}^{\tau_2} \max(c - T_t, 0) \\ CDD(\tau_1, \tau_2) &= \sum_{t=\tau_1}^{\tau_2} \max(T_t - c, 0) \\ CAT(\tau_1, \tau_2) &= \sum_{t=\tau_1}^{\tau_2} T_t \end{aligned} \tag{18}$$

where  $c$  is a threshold (usually 65°F or 18 °C) and  $[\tau_1, \tau_2]$  with  $\tau_1 < \tau_2$  is the temperature measurement period. The standard is



**Fig. 4.** Time series plots of New York, Atlanta, Chicago, Dallas HDD/CDD monthly and seasonal strips futures prices. HDD monthly futures with the measurement period of January (Black), February (Blue), March (Cyan), September (Red), October (Magenta), November (Yellow) and December (Green). CDD monthly futures with the measurement period of May (Black), June (Blue), July (Cyan), August (Red), and September (Magenta). HDD seasonal strip with the measurement period in January (Blue), March (Cyan) and December (Cyan). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that  $[\tau_1, \tau_2]$  denotes a month of the year or as seasonal strips. The futures in question are delivering over a period  $[\tau_1, \tau_2]$ , and not at a fixed delivery time  $\tau$ . The HDD index measures the demand for heating, while the CDD index measures the demand for cooling. Consequently, temperature indices are the underlying and not the temperature by itself.

Financial mathematical tools given in [Benth et al. \(2007, 2011\)](#) and [Härdle and López-Cabrera \(2012\)](#) allow the pricing of the

non-tradable underlying by risk adjusted conditional expectation. Hereby, the futures temperature contract price on the sum of temperature  $I(\tau_1, \tau_2) = \sum_{t=\tau_1}^{\tau_2} T_t$  with accumulation period  $[\tau_1, \tau_2]$  is given by:

$$F(t, \tau_1, \tau_2) = E^Q [I(\tau_1, \tau_2) | \mathcal{F}_t] \tag{19}$$

where  $E^Q[\cdot]$  is any equivalent martingale measure and  $\mathcal{F}_t$  is a filtration information set.

**Table 2**

The number of transactions—trading days (TD) and volume (vol) of New York/Atlanta/Chicago/Dallas HDD and CDD monthly and HDD seasonal strip options with respect to time to maturity ( $\tau$ ) in month and the number of strike prices.

		HDD—New York					HDD—Atlanta					HDD—Chicago					
		Number of strike prices															
$\tau$		1	2	3	4	Total	1	2	3	4	5	Total	1	2	3	4	Total
$\leq 1$	TD	71	23	7	1	102	56	12	4	1	1	74	50	10	-	-	60
	vol	17 495	12 650	9900	1400	41 445	12 861	4 700	2950	700	1250	22 461	10961	4975	-	-	15 936
(1, 2]	TD	54	26	3	4	87	39	26	2	1	-	68	32	13	2	2	49
	vol	12 450	21 700	1075	5400	40 625	10 245	19 825	2800	1000	-	33 870	50	2000	-	-	2 050
(2, 3]	TD	3	1	-	-	4	1	-	-	-	-	1	2	-	-	-	2
	vol	1 000	1 000	-	-	2 000	100	-	-	-	-	100	2 000	-	-	-	2 000
(3, 4]	TD	2	1	-	-	3	-	-	-	-	-	-	1	-	-	-	1
	vol	300	2 000	-	-	2 300	-	-	-	-	-	-	2 000	-	-	-	2 000
(4, 5]	TD	1	1	-	-	2	-	-	-	-	-	-	1	-	-	-	1
	vol	250	2 000	-	-	2 250	-	-	-	-	-	-	2 000	-	-	-	2 000
(5, 6]	TD	-	1	-	-	1	-	-	-	-	-	-	1	-	-	-	1
	vol	-	2 000	-	-	2 000	-	-	-	-	-	-	2 000	-	-	-	2 000

		CDD—New York					HDD strips—Atlanta								CDD—Dallas						
		Number of strike prices																			
$\tau$		1	2	3	4	Total	1	2	3	4	5	6	7	8	Total	1	2	3	4	Total	
$\leq 1$	TD	43	3	1	-	47	1	6	-	6	-	-	-	-	13	40	-	-	90	1	131
	vol	17 425	2000	600	-	20 025	200	3 100	-	2700	-	-	-	-	6 000	12 400	9150	1250	-	-	-
(1, 2]	TD	34	13	2	-	49	6	6	1	-	-	-	-	-	13	30	-	-	10	1	41
	vol	8 200	5750	1200	-	15 150	4700	6 250	1875	-	-	-	-	-	-	9 125	5300	750	-	-	-
(2, 3]	TD	-	-	-	-	-	3	9	-	3	-	-	-	-	15	-	1	-	-	-	1
	vol	-	-	-	-	-	2240	7 500	-	5500	-	-	-	-	-	-	450	-	-	-	450
(3, 4]	TD	-	-	-	-	-	-	9	-	-	-	-	-	-	9	-	1	-	-	-	1
	vol	-	-	-	-	-	-	10 400	-	-	-	-	-	-	10 400	-	450	-	-	-	450
(4, 5]	TD	-	-	-	-	-	1	11	1	4	-	1	1	-	19	-	1	-	-	-	1
	vol	-	-	-	-	-	1750	10 500	1000	9500	-	6500	6500	-	-	-	450	-	-	-	450
(5, 6]	TD	-	-	-	-	-	1	11	-	3	-	-	1	-	16	-	-	-	-	-	-
	vol	-	-	-	-	-	250	11 700	-	6750	-	-	-	4250	-	-	-	-	-	-	-
(6, 7]	TD	-	-	-	-	-	10	-	-	-	-	-	-	-	1	-	-	-	-	-	-
	vol	-	-	-	-	-	9000	-	-	-	-	-	-	-	9000	-	-	-	-	-	-
(7, 8]	TD	-	-	-	-	-	3	-	-	-	-	-	-	-	3	-	-	-	-	-	-
	vol	-	-	-	-	-	6000	-	-	-	-	-	-	-	6000	-	-	-	-	-	-

Consequently, the European temperature call option price written on the futures price is defined as:

$$C(K) = \exp\{-r\tau\} \int \max\{F(t, \tau_1, \tau_2) - K, 0\} f(x) dx. \tag{20}$$

In order to compute (19), (20), it is necessary to know the stochastic properties of the temperature process  $T_t$  under the “physical measure”  $P$  and then adjust the risk measure  $Q$ , see Härdle and López-Cabrera (2012). In other words, the temperature derivative price is given by finding a model for the daily weather process consisting of a trend, a seasonality, an autoregressive part, seasonal variance and normally distributed residuals. Then one could specify a class of probability measures using the Radon–Nikodym derivative determined by the Esscher transform, see López-Cabrera et al. (2013). Another way is to model the index directly, see Dorfleitner and Wimmer (2010).

Here we estimate the SPD, different to the afore mentioned approach, directly under the risk neutral measure  $Q$  from real option data. Note that (20) is exactly (5) for  $f = f_N$ .

The options at CME are cash settled, i.e., the owner of a future receives 20 times the Degree Day Index at the end of the measurement period, in return for a fixed price. At time  $t$ , CME trades contracts with different measurement periods  $t \leq \tau_1 < \tau_2$  or different maturities  $\tau = \tau_2 - t$ . The measurement period for CAT/HDD futures is typically during April–November, while CDD futures are measured during November–April.

3.2. Overview on the WD data

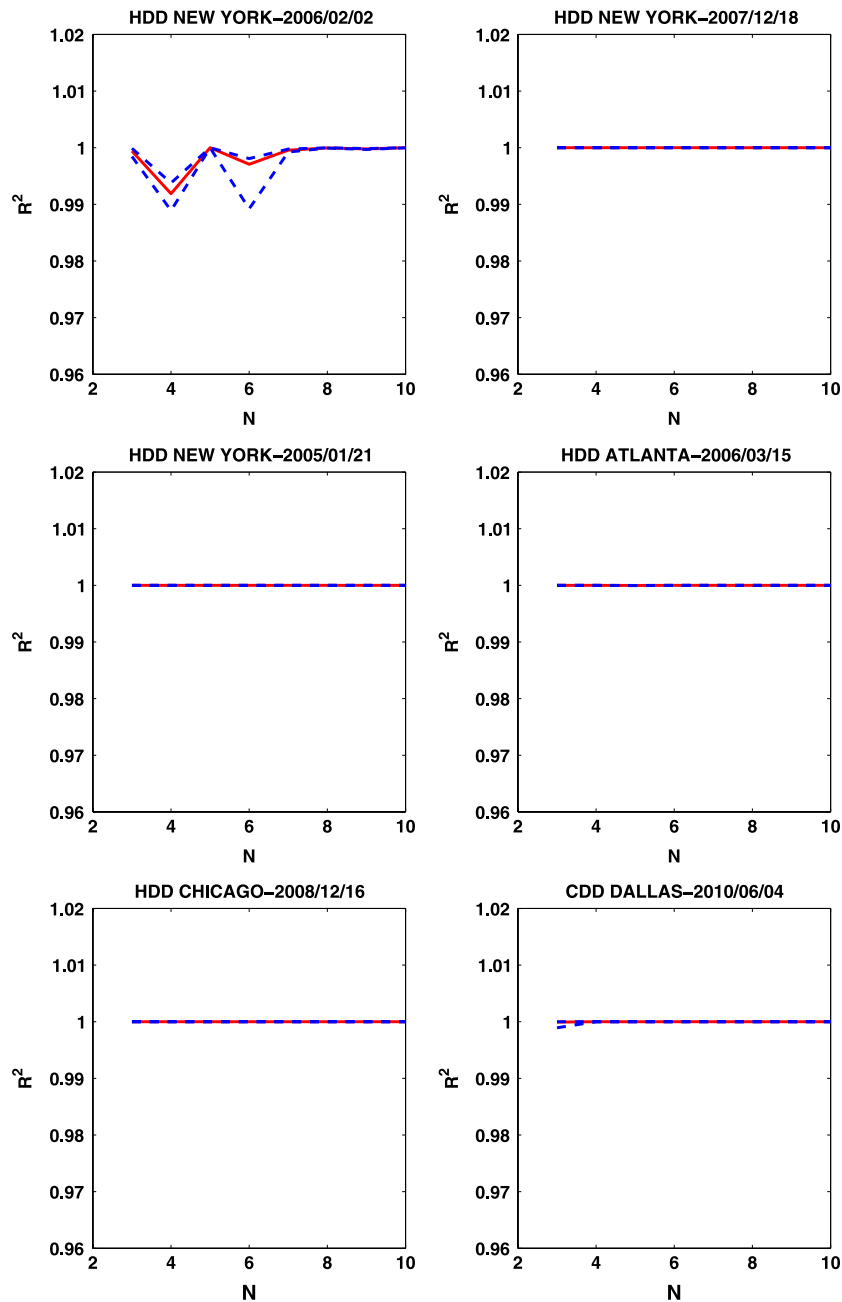
The WD data was purchased from CME for the study period from 2002/01/02 to 2012/05/11. The reported price is the settlement

price for the future or option contract, and the volume is the number of contracts traded.

Depending on the measurement period, temperature products in the US market are further categorized into monthly, seasonal strips, and average products. HDD monthly products have seven contract months: October, November, December, January, February, March, April, and CDD monthly products have seven contract months: April, May, June, July, August, September, and October. For HDD seasonal strips, the contract period covers from October to April, and for CDD seasonal strips, the contract period covers from April to October. Contract for weekly average products covers all five weeks. Table 1 gives an overview of the volume of the temperature market.

Fig. 2 illustrates the volume for US temperature futures and options in the study period. The trading activity increased dramatically since 2002 but declined after the 2008 financial crisis. This is surprising since one could expect that these markets are uncorrelated with financial markets. However we believe that the decline is because the temperature market is not yet well known as a intermediary for diversification of weather risk. Star plots in Fig. 3 divide the volume into HDD–CDD monthly, HDD–CDD seasonal strips, and weekly average for futures and options for each US city. A star plot represents each city as a star whose  $i$ th spoke is proportional in length to the volume size of  $i$ th product (HDD Monthly, CDD Monthly, HDD Strips, CDD strips, Average) of the observed city. Clearly, monthly products are the most popular traded products, followed by seasonal strips. Nevertheless, no weekly average products are really traded in the US temperature market.

The volumes of HDD, CDD, Average monthly and seasonal strips futures and options for all US cities are reported in Table 1. New



**Fig. 5.** The 90% credible regions (in blue dashed lines) and posterior means (in red lines) of  $R^2$  when fitting New York/Atlanta/Chicago/Dallas HDD–CDD monthly options using the Bayesian quadrature approach against different number of support points  $N$ . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

York is to be the biggest temperature market and takes about 20% of the market volume, followed by Chicago (10%), Atlanta (9%), Cincinnati (8%), and Dallas (7%). The market share of these five cities exceeds 50% of the US temperature market. Following this, we took these cities as the most representative cities. Fig. 4 gives time series plots for New York, Atlanta, Chicago and Dallas monthly HDD, CDD monthly and seasonal strip future prices. The futures market is more liquid but also more volatile than option prices. In addition, most HDD and CDD futures are traded only with time to maturities less than a year. These features of future prices make the pricing mechanism for weather derivatives unique and challenging.

We further divide the volume of HDD and CDD monthly and HDD seasonal strip options with respect to strike prices and time

to maturities, as summarized in Table 2. It is shown that most options are traded with only a few number of strike prices and of a short time to maturity (within one month and less than a year). Because of the fact that options are only traded with a few number of strike prices, this data sparsity problem makes most existing nonparametric methods (such as mixture of lognormal models or kernel methods) very difficult.

### 3.3. Implementation of the technique

As depicted in Fig. 8, we calibrate the SPD for HDD–CDD monthly and Seasonal strip options. These four plots present a typical pattern of option prices of weather options: options were

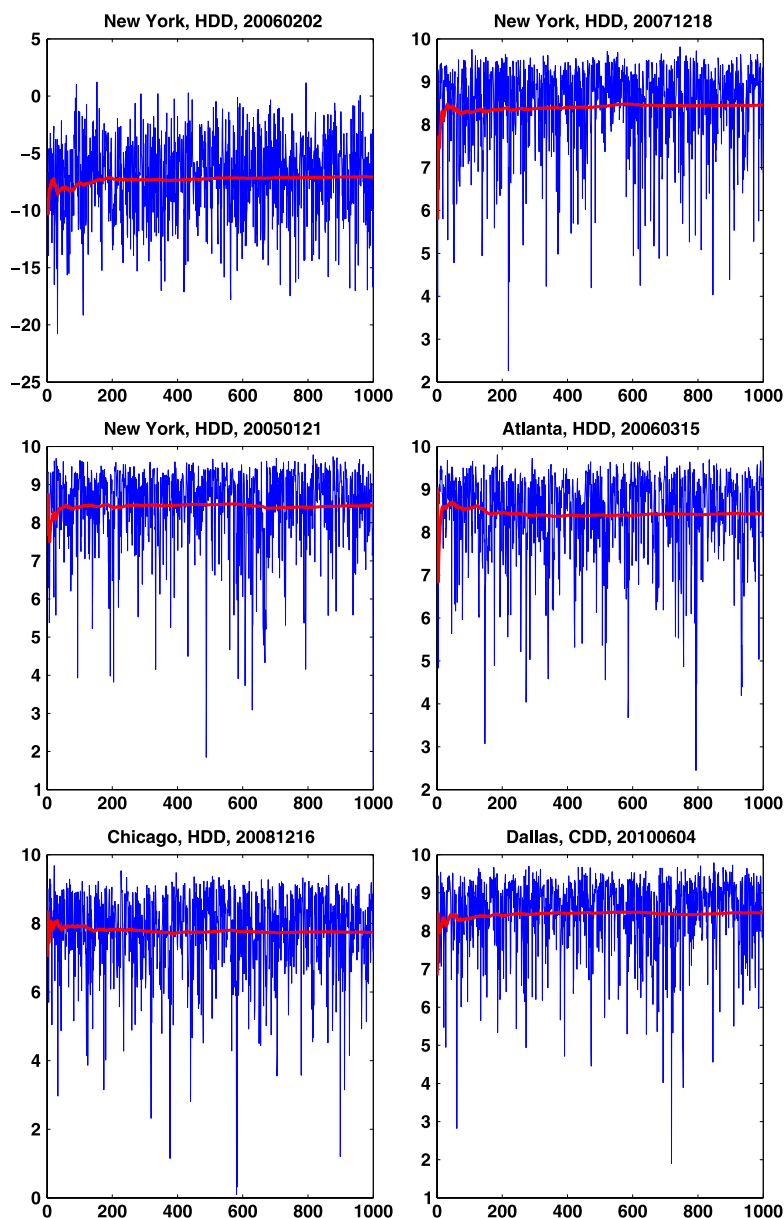


Fig. 6. Trace plots (in blue lines) and cumulative averages (in red lines) of the log-likelihood in the MCMC algorithm of different HDD/CDD monthly products. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

traded only with a very few number of strike prices, sometimes only call options or put options were traded or the both of them. The Bayesian quadrature method allows us to incorporate prior assumptions on the model parameters and hence avoid problems with data sparsity. It is able to compute the SPD of both call and put options simultaneously, and is particularly robust.

There is a trade-off in the selection of  $N$ . When  $N$  is larger, one produces better fit because there exist more free parameters in the model, but drawbacks of model complexity and computational demanding come along with. We provide more information on the sensitivity of the Bayesian quadrature approach with respect to  $N$ : Fig. 5 depicts posterior means and 90% credible regions of  $R^2$  in fitting HDD/CDD options using the Bayesian quadrature method versus different number of support points  $N$  from three to ten. As shown in Fig. 5 that all  $R^2$ 's are close to one, we conclude that the Bayesian quadrature approach provides good model fit with small  $N$ . Based on Fig. 5, we select  $N = 5$  in our following

analysis because it gives a simple model yet producing good model fit.

To calibrate (5) and (6), we implement an MCMC algorithm to explore the posterior distribution in (12). Because of the employment of unequal weights and the adoption of a Bayesian framework, inferring these parameters is computationally challenging. For this reason, we use an MCMC algorithm with slice samplers for making statistical inference. In our analysis, we discard the first 500 iterates in the MCMC algorithm (the burn-in period), and use the following 1000 iterates. Trace plots of the loglikelihood in Fig. 6 show that the MCMC algorithm appears to converge very fast, and autocorrelation plots of the loglikelihood in Fig. 7 indicate that samples in the MCMC algorithm are efficient.

Fig. 8 imposes model prices of the last swipe of the MCMC algorithm and demonstrates the fit, because model prices are close to market prices.

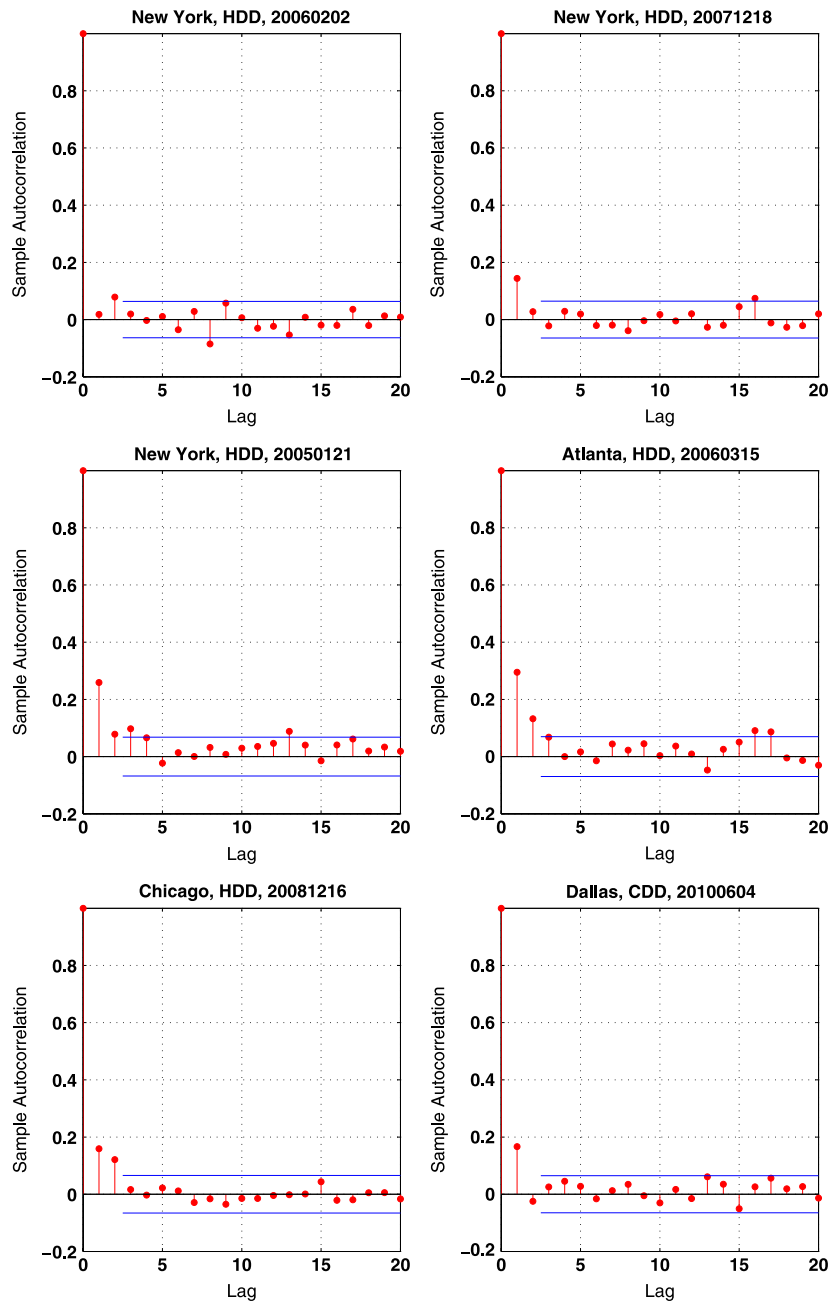


Fig. 7. Autocorrelation plots of the log-likelihood in the MCMC algorithm of different HDD/CDD monthly products.

For residual analysis, we calculate the residual at each swipe of the MCMC algorithm by

$$r_{ijk} = \log y_{ijk} - \log C_{ij}^N(w, \theta) \quad (21)$$

and provide kernel smooth density plots of the posterior distribution of these residuals in Fig. 9. All these four panel plots demonstrate that the residuals have mean zero, and are symmetric about zero when comparing with the normal KDE. This visual presentation supports our error assumption as a normal distribution in (8).

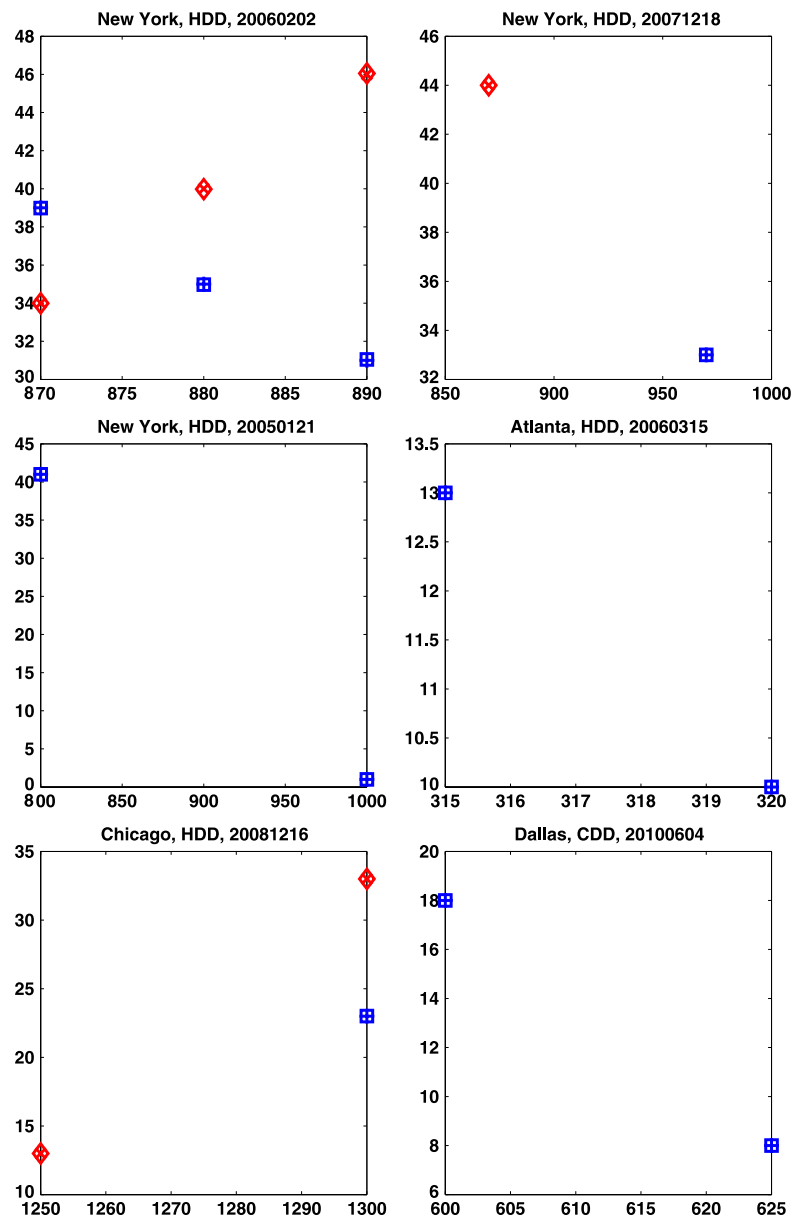
The density (3) approximates the SPD by weighted sum of  $\delta$  functions and is discontinuous by its nature. As described earlier, to produce a smoothed SPD for visualization, we round off each  $w_n$  to the second decimal, and set the adjusted sample size as 100. Then we employ the kernel density estimation with a Gaussian kernel

$K(\cdot) = \varphi(\cdot)$  and a bandwidth selected using the rule of thumb in (14) to calculate a smoothed SPD at each swipe in the MCMC algorithm.

Thus, it is clear that the smoothed density version (15) becomes:

$$\begin{aligned} f_N^s(x|w, \theta) &= \sum_{n=1}^N w_n K_h(x - \theta_n) \\ &= \sum_{n=1}^N w_n \frac{1}{h} \varphi\left(\frac{x - \theta_n}{h}\right) \\ &= \sum_{n=1}^N w_n \varphi(x; \theta_n, h) \end{aligned}$$

where  $\varphi(x; \theta_n, h)$  is the pdf of  $N(\theta_n, h^2)$  distribution.



**Fig. 8.** Plots of market prices of New York/Atlanta/Chicago/Dallas HDD–CDD monthly options and model prices of the last swipe in the MCMC algorithm using the Bayesian quadrature method with  $N = 5$ . For market prices, a call option is indicated with a blue plus and a put is indicated with a red cross. For model prices, a call option is indicated with a blue diamond, and a put option is indicated with a red square. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

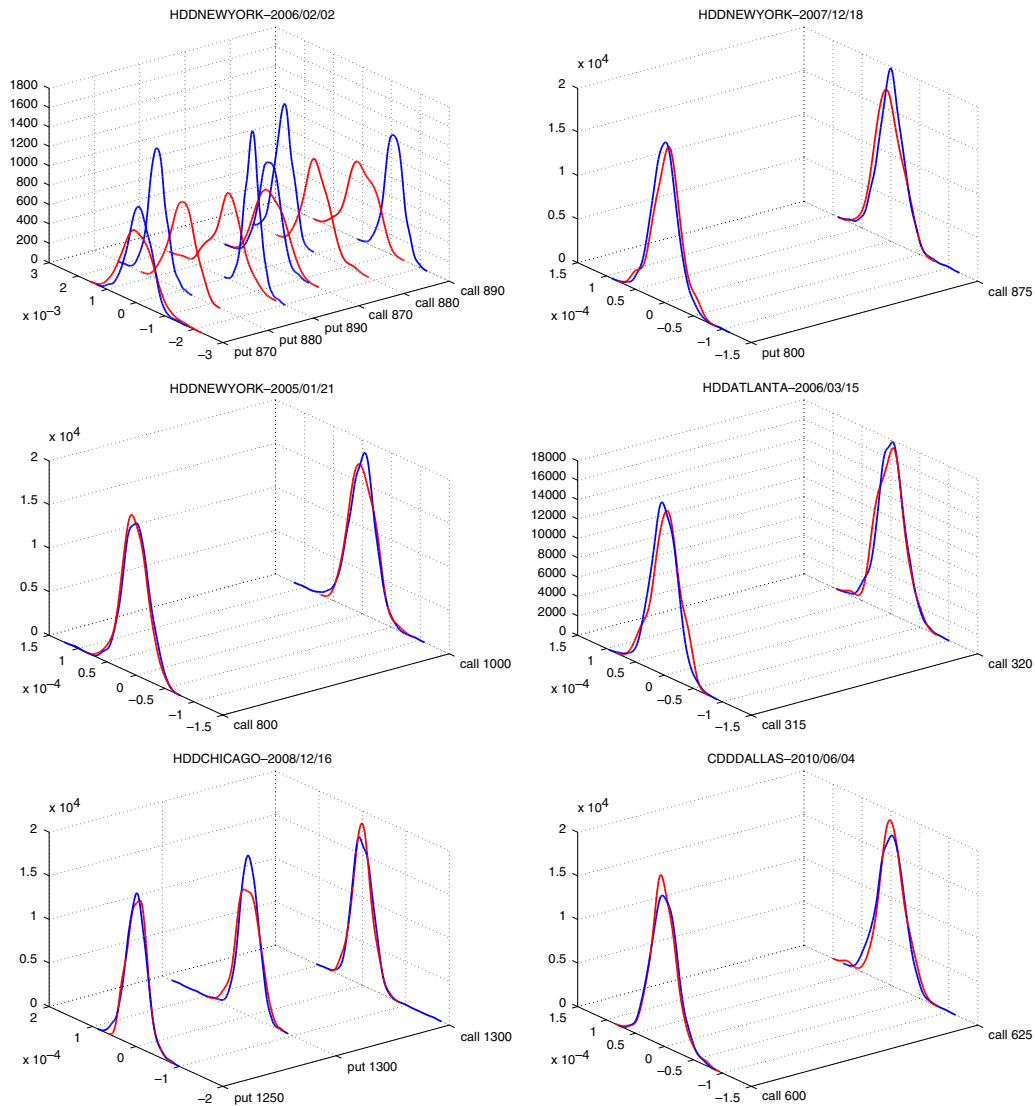
Collecting these smoothed SPD, Fig. 10 gives the posterior mean (red line) and 90% credible regions (blue dotted lines) of the implied SPD. The right-upper and left-lower pictures show that the 90% credible regions are tight to the posterior mean of the smoothed SPD, whereas the other pictures depict that the 90% credible regions are wide. The cluster of star points in the horizontal axis denotes the future prices.

In Bayesian analysis, the 90% credible region for the smoothed SPD provides a region where 90% of the posterior distribution of the smoothed SPD will fall into. In the case of HDD New York options with maturity in 2 months traded at 20050121 and the case of CDD Dallas Option one month to maturity traded at 20100604, the left tail of 90% credible region appears to be extremely wide. This feature is not surprising though, because the data set for calibration consists of call options with only two strike prices, namely, 800 and 1000 and 600 and 625 respectively. Indeed, a call option price is

simply the expected future price larger than the strike price under the SPD. Therefore, an option price only provides information for the right tail of the SPD. Once a few quadrature points in the right tail have achieved a high likelihood, points of the quadrature in the left tail (in this case, smaller than 800) do not affect the likelihood. As a result, these points are influenced only by its prior distributions. The prior assumptions in (9) and (10) put simply vague information for the weights and locations in the quadrature method. Such an assumption allows points in the left tail of the quadrature method moving freely, and causes a wider credible region in the left tail, as demonstrated in the left-upper panel in Fig. 10.

Similarly, for the right-lower panel, the 90% credible region is wider around strike 1000 but is tight in two side tails. This is because the data set consists of one call with strike 970 and one put with strike 870. As a result, the call option price gives





**Fig. 9.** Kernel density estimate (KDE) plots for the posterior distribution of residuals for HDDs and CDDs products (in blue lines) versus the normal KDE plots (in red lines). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

information of the right tail of the SPD, whereas the put option price gives information of the left tail of the SPD. When some points of both right and left tails in the quadrature method have achieved a high likelihood, points of the quadrature around 920 would not affect the likelihood. These points are determined by their prior assumptions again, and provide a wider credible region around 920.

Selecting prior distributions for the quadrature method is critical. In this research, we choose vague prior assumptions for the parameters and the analysis successfully reveals the fact that the width of the 90% credible region depends highly on the information provided by option prices and the prior assumptions on the parameters in the quadrature method. One may adopt more sophisticated prior distributions based on experience and knowledge. This flexibility may be considered as a technical advantage of the Bayesian quadrature method.

### 3.4. Out-of-sample analysis

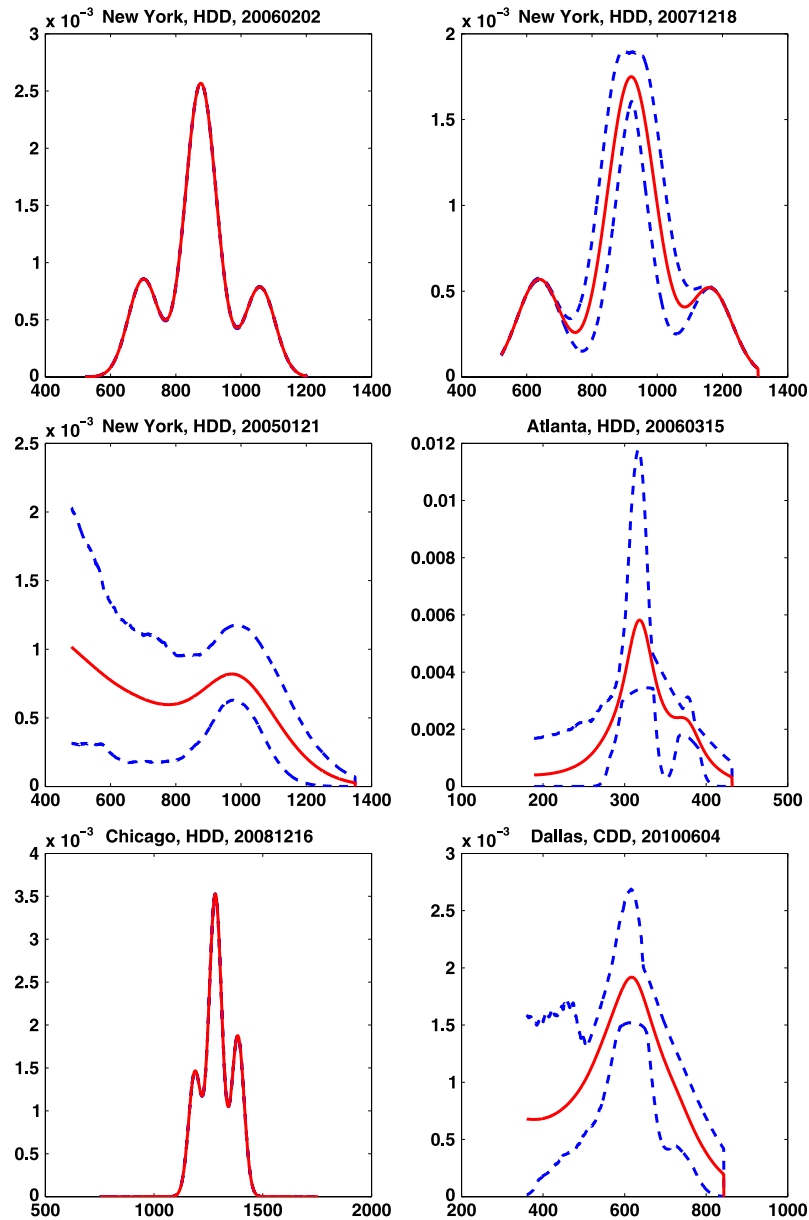
In incomplete markets, no unique martingale measure exists. As a consequence, this may have a negative effect that the parameters

estimated fit well the in-sample data, but they are inaccurate with out-of-sample data. In the following, we modify the quadrature model to forecast out-of-sample data and provide an empirical analysis confirming that our quadrature model preforms well for the out-of-sample data.

Recall that  $t$  is the current time,  $\tau$  is the time to maturity, and  $F(t, \tau_1, \tau_2)$  is the underlying temperature index future price at time  $t$  with accumulation period  $[\tau_1, \tau_2]$ . Let  $\tilde{t}$  be the forecast time and  $\tilde{\tau}$  be the time to maturity from time  $\tilde{t}$  to maturity. Similar to [Dumas et al. \(1998\)](#) and [Fan and Mancini \(2009\)](#), we consider a one-week-ahead forecast horizon. Both  $F(t, \tau_1, \tau_2)$  and  $F(\tilde{t}, \tau_1, \tau_2)$  are known in our out-of-sample analysis.

To begin with, the option prices at time  $t$  are used to calibrate the Bayesian quadrature model, where  $w$  and  $\theta$  are parameters. To provide a plausible yet simple quadrature model for forecast at time  $\tilde{t}$  with parameters  $\tilde{w}$  and  $\tilde{\theta}$ , we first set  $\tilde{w} = w$ . Recall that  $r$  is the risk-free interest rate. To adjust the current underlying temperature index future price and the discounted factor, define  $\bar{\theta} = (\bar{\theta}_1, \dots, \bar{\theta}_N)'$  as the normalized location parameters extracted from the quadrature model by

$$\theta_i = F(t, \tau_1, \tau_2) e^{r\tau} \bar{\theta}_i,$$



**Fig. 10.** Posterior means (in red solid lines) and the 90% credible regions (in blue dashed lines) of the smoothed SPD implied from New York HDD monthly options with respect to trading dates and time to maturity. The cluster of star points in the horizontal axis denotes the future prices. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for  $i = 1, \dots, N$ . Likewise,  $\tilde{\theta}$  is linked to  $\theta$  via the normalized location parameters by

$$\tilde{\theta}_i = F(\tilde{t}, \tau_1, \tau_2) e^{r\tilde{t}} \theta_i = \frac{F(\tilde{t}, \tau_1, \tau_2)}{F(t, \tau_1, \tau_2)} e^{r(\tilde{t}-t)} \theta_i, \quad (22)$$

for  $i = 1, \dots, N$ . As a result, the forecast quadrature model at time  $t_1$  is

$$w_1 \delta_{\tilde{\theta}_1}(x) + \dots + w_N \delta_{\tilde{\theta}_N}(x),$$

where  $\tilde{\theta}$  is given in (22), and can be used to calculate option prices forecast at time  $\tilde{t}$  directly.

In the Bayesian out-of-sample analysis, we report the 90% Bayesian prediction intervals for option prices at time  $\tilde{t}$  and averaged  $R^2$  in the MCMC algorithm. Table 3 provides detailed

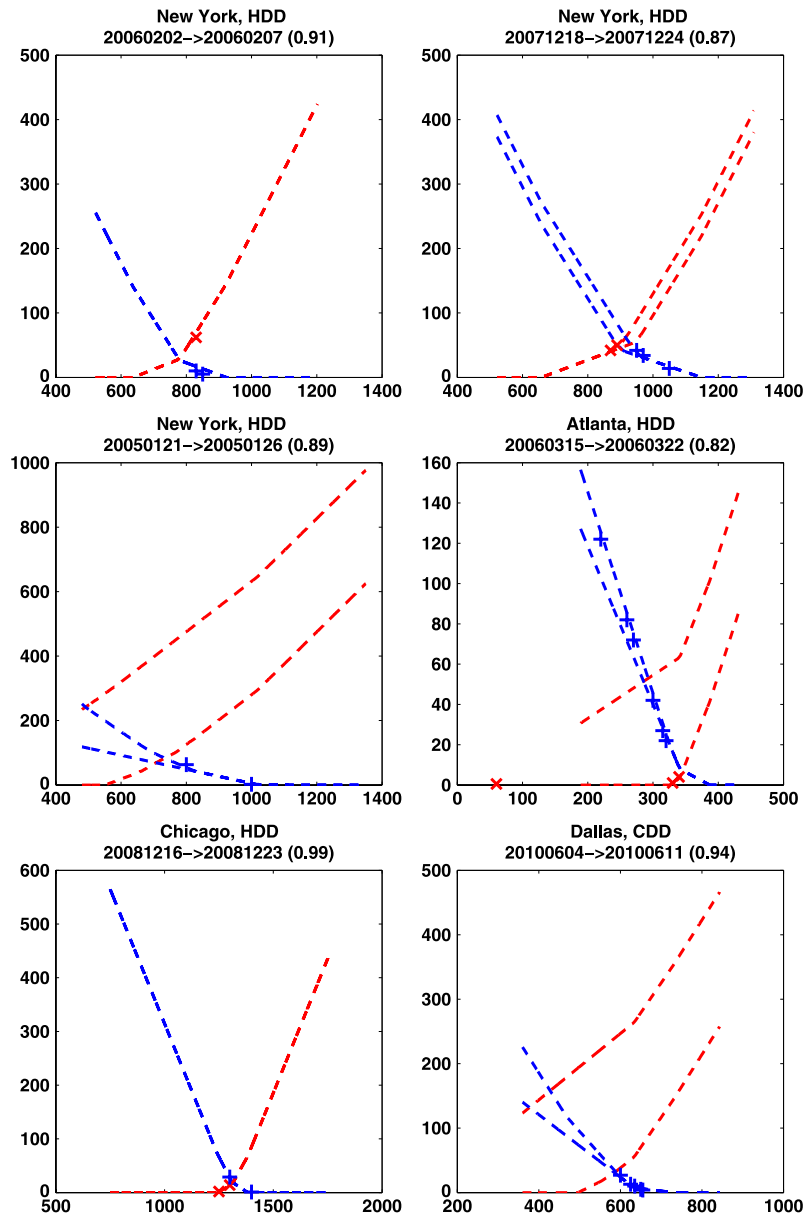
information on the data used for the out-of-sample analysis: For each case of weather derivatives, Table 3 lists its current time  $t$ , forecast time  $\tilde{t}$ , maturity, underlying temperature index future prices at time  $t$  and  $\tilde{t}$ , and averaged  $R^2$  in the MCMC algorithm. We remark that because the weather derivatives markets are less frequently traded, given  $t$ , if there is no settlement prices in the one-week-ahead horizon, we use options traded nearest to the one-week-ahead horizon as the out-of-sample data.

Fig. 11 depicts the 90% Bayesian prediction intervals for the forecast and realized market option prices traded at time  $\tilde{t}$ . It is shown that realized market option prices are within or close to the 90% Bayesian prediction intervals. Together with the numerical results that the averaged  $R^2$  in Table 3 ranges from 0.82 to 0.99, we conclude that our quadrature method empirically performs well in this out-of-sample analysis.

**Table 3**

Data and averaged  $R^2$  in the out-of-sample analysis. This table lists the type of weather derivatives, current time  $t$ , forecast time  $\tilde{t}$ , maturity, underlying temperature index future prices at times  $t$  and  $\tilde{t}$ , and averaged  $R^2$  in the MCMC algorithm.

Product	$t$	$\tilde{t}$	Maturity	$F(t, \tau_1, \tau_2)$	$F(\tilde{t}, \tau_1, \tau_2)$	Averaged $R^2$
New York, HDD	20060202	20060207	Feb. 2006	875	778	0.91
New York, HDD	20071218	20071224	Jan. 2008	905	910	0.87
New York, HDD	20050121	20050126	Feb. 2005	805	845	0.89
Atlanta, HDD	20060315	20060322	Mar. 2006	320	342	0.82
Chicago, HDD	20081216	20081223	Dec. 2008	1290	1315	0.99
Dallas, CDD	20100604	20100611	Jun. 2010	600	617	0.94



**Fig. 11.** Out-of-sample analysis for different contracts. For realized market prices, a call option is indicated with a blue plus and a put is indicated with a red cross. The 90% Bayesian prediction intervals of call and put option are in blue and red dashed lines, respectively. Averaged  $R^2$  is recorded in parentheses in the title of each panel plot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 3.5. Dynamics of SPD

Table 4 records the number of trading days with respect to the number of strike prices and trading months for the New York HDD/CDD monthly options with time to maturity  $\tau$  less than one month and Atlanta HDD Seasonal Strips with  $\tau = 6$ . Again, the

data sparsity remains an issue in the biggest weather market, the New York market. For illustration, we implement (4) for every trading day in March 2006. For other months, most of the number of strike in each trading day is simply one, and such case makes the SPD estimation very difficult in the sense that option price only provides information for one side of the SPD.

**Table 4**

Number of trading days with respect to the trading month and the number of strike prices for HDD/CDD monthly options with time to maturity less  $\tau$  than one month and Atlanta HDD Seasonal Strips with  $\tau = 6$  (time of measurement period of 5 months).

Year	Type	City	Month	Nr. of strike prices						Year	Type	City	Month	Nr. of strike prices						Total			
				$\tau$	1	2	3	4	5					Total	$\tau$	1	2	3	4		5	6	Total
2002	HDD	NY	11	1	1	-	-	-	-	1	2006	HDD-Strip	Atlanta	10	6	-	-	1	-	-	-	1	
2004	HDD	NY	1	1	2	-	-	-	-	2	2007	HDD-Strip	Atlanta	10	6	-	-	3	1	2	1	1	8
2004	HDD	NY	2	1	1	-	-	-	-	1	2008	HDD-Strip	Atlanta	10	6	-	-	4	-	1	1	1	5
2004	HDD	NY	3	1	1	-	-	-	-	1	2009	HDD-Strip	Atlanta	10	6	-	-	1	-	1	-	-	3
2005	HDD	NY	1	1	1	-	-	-	-	1	2010	HDD-Strip	Atlanta	10	6	-	-	2	-	-	-	-	2
2005	HDD	NY	2	1	2	-	-	-	-	2	2002	HDD	Chicago	11	1	1	-	-	-	-	-	-	1
2005	HDD	NY	3	1	1	-	-	-	-	1	2002	HDD	Chicago	12	1	2	-	-	-	-	-	-	2
2005	HDD	NY	12	1	1	1	-	-	-	2	2004	HDD	Chicago	2	1	1	-	-	-	-	-	-	1
2006	HDD	NY	1	1	2	-	-	-	-	2	2005	HDD	Chicago	2	1	2	-	-	-	-	-	-	2
2006	HDD	NY	2	1	5	3	1	-	-	9	2005	HDD	Chicago	12	1	4	-	-	-	-	-	-	4
2006	HDD	NY	3	1	3	5	4	1	-	13	2006	HDD	Chicago	3	1	5	2	-	-	-	-	-	7
2006	HDD	NY	10	1	1	1	-	-	-	2	2006	HDD	Chicago	10	1	2	1	-	-	-	-	-	3
2006	HDD	NY	11	1	1	2	-	-	-	3	2006	HDD	Chicago	11	1	2	-	-	-	-	-	-	2
2006	HDD	NY	12	1	1	-	-	-	-	1	2007	HDD	Chicago	1	1	-	3	-	-	-	-	-	3
2007	HDD	NY	1	1	2	-	-	-	-	2	2007	HDD	Chicago	2	1	1	1	-	-	-	-	-	1
2007	HDD	NY	2	1	4	-	-	-	-	4	2007	HDD	Chicago	3	1	1	-	-	-	-	-	-	1
2007	HDD	NY	3	1	1	-	-	-	-	1	2007	HDD	Chicago	12	1	-	1	-	-	-	-	-	1
2007	HDD	NY	11	1	1	-	-	-	-	1	2008	HDD	Chicago	1	1	-	1	-	-	-	-	-	1
2007	HDD	NY	12	1	1	2	-	-	-	3	2008	HDD	Chicago	2	1	3	-	-	-	-	-	-	3
2008	HDD	NY	1	1	6	1	2	-	-	9	2008	HDD	Chicago	3	1	2	1	-	-	-	-	-	3
2008	HDD	NY	2	1	3	-	-	-	-	3	2008	HDD	Chicago	12	1	2	2	-	-	-	-	-	4
2008	HDD	NY	12	1	2	1	-	-	-	3	2009	HDD	Chicago	1	1	2	-	-	-	-	-	-	2
2009	HDD	NY	1	1	4	-	-	-	-	4	2009	HDD	Chicago	12	1	1	-	-	-	-	-	-	1
2009	HDD	NY	2	1	2	-	-	-	-	2	2010	HDD	Chicago	3	1	1	-	-	-	-	-	-	1
2009	HDD	NY	3	1	1	-	-	-	-	1	2010	HDD	Chicago	11	1	1	-	-	-	-	-	-	1
2009	HDD	NY	11	1	3	-	-	-	-	3	2010	HDD	Chicago	12	1	2	-	-	-	-	-	-	2
2009	HDD	NY	12	1	1	-	-	-	-	1	2011	HDD	Chicago	1	1	3	-	-	-	-	-	-	3
2010	HDD	NY	3	1	1	-	-	-	-	1	2011	HDD	Chicago	2	1	2	-	-	-	-	-	-	2
2010	HDD	NY	11	1	5	-	-	-	-	5	2011	HDD	Chicago	11	1	2	-	-	-	-	-	-	2
2010	HDD	NY	12	1	2	-	-	-	-	2	2011	HDD	Chicago	12	1	2	-	-	-	-	-	-	2
2011	HDD	NY	1	1	4	1	-	-	-	5	2004	CDD	Dallas	9	1	1	-	-	-	-	-	-	1
2011	HDD	NY	2	1	5	-	-	-	-	5	2005	CDD	Dallas	8	1	1	-	-	-	-	-	-	1
2011	HDD	NY	3	1	1	1	-	-	-	2	2006	CDD	Dallas	6	1	1	-	-	-	-	-	-	1
2011	HDD	NY	11	1	1	-	-	-	-	1	2006	CDD	Dallas	9	1	1	-	-	-	-	-	-	1
2011	HDD	NY	12	1	2	-	-	-	-	2	2007	CDD	Dallas	7	1	1	-	-	-	-	-	-	1
2012	HDD	NY	2	1	1	-	-	-	-	1	2008	CDD	Dallas	5	1	6	-	-	-	-	-	-	6
2004	CDD	NY	9	1	-	1	-	-	-	1	2008	CDD	Dallas	6	1	1	-	-	-	-	-	-	1
2005	CDD	NY	5	1	1	1	-	-	-	1	2008	CDD	Dallas	7	1	2	-	-	-	-	-	-	2
2005	CDD	NY	6	1	8	1	-	-	-	9	2008	CDD	Dallas	8	1	1	-	-	-	-	-	-	1
2005	CDD	NY	7	1	-	-	-	-	-	1	2009	CDD	Dallas	5	1	4	1	-	-	-	-	-	5
2005	CDD	NY	8	1	4	-	-	-	-	4	2009	CDD	Dallas	6	1	1	-	1	-	-	-	-	2
2006	CDD	NY	6	1	4	-	-	-	-	4	2009	CDD	Dallas	8	1	1	-	-	-	-	-	-	1
2006	CDD	NY	7	1	1	-	-	-	-	1	2009	CDD	Dallas	9	1	1	-	-	-	-	-	-	1
2006	CDD	NY	8	1	3	-	-	-	-	3	2010	CDD	Dallas	5	1	1	-	-	-	-	-	-	1
2006	CDD	NY	9	1	3	-	-	-	-	2	2010	CDD	Dallas	6	1	2	3	-	-	-	-	-	5
2007	CDD	NY	7	1	1	-	-	-	-	1	2010	CDD	Dallas	7	1	-	2	-	-	-	-	-	2
2007	CDD	NY	8	1	1	-	-	-	-	1	2010	CDD	Dallas	8	1	-	2	-	-	-	-	-	2
2007	CDD	NY	9	1	1	-	-	-	-	1	2010	CDD	Dallas	9	1	3	-	1	-	-	-	-	4
2008	CDD	NY	6	1	5	-	-	-	-	5	2011	CDD	Dallas	5	1	-	-	1	-	-	-	-	1
2008	CDD	NY	7	1	1	-	-	-	-	1	2011	CDD	Dallas	6	1	3	1	-	-	-	-	-	4
2008	CDD	NY	8	1	1	-	-	-	-	1	2011	CDD	Dallas	7	1	3	-	-	-	-	-	-	3
2009	CDD	NY	8	1	2	-	-	-	-	2	2011	CDD	Dallas	8	1	2	-	-	-	-	-	-	2
2010	CDD	NY	5	1	2	-	-	-	-	2	2012	CDD	Dallas	2	1	2	-	-	-	-	-	-	2
2010	CDD	NY	5	1	2	-	-	-	-	2													
2010	CDD	NY	6	1	1	-	-	-	-	1													
2010	CDD	NY	7	1	1	-	-	-	-	1													
2010	CDD	NY	9	1	1	-	-	-	-	1													
2011	CDD	NY	9	1	1	-	-	-	-	1													
2011	CDD	NY	6	1	3	-	-	-	-	3													

Fig. 12 plots give the evolution of New York-HDD, Atlanta HDD-Seasonal and option prices with time to maturities in one and six months respectively, against strike prices and trading days in March 2006 and October 2007. It is clear that options were traded with very few strike prices (from one to four strike prices) during this month. We used five support points ( $N = 5$ ) in the quadrature

method, and calculate  $R^2$  in a logarithmic scale:

$$R^2 = 1 - \frac{\sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \sum_{k=1}^{N_{ij}} \{\log y_{ijk} - \log C_{ij}^N(w, \theta)\}^2}{\sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \sum_{k=1}^{N_{ij}} \log y_{ijk}^2}$$

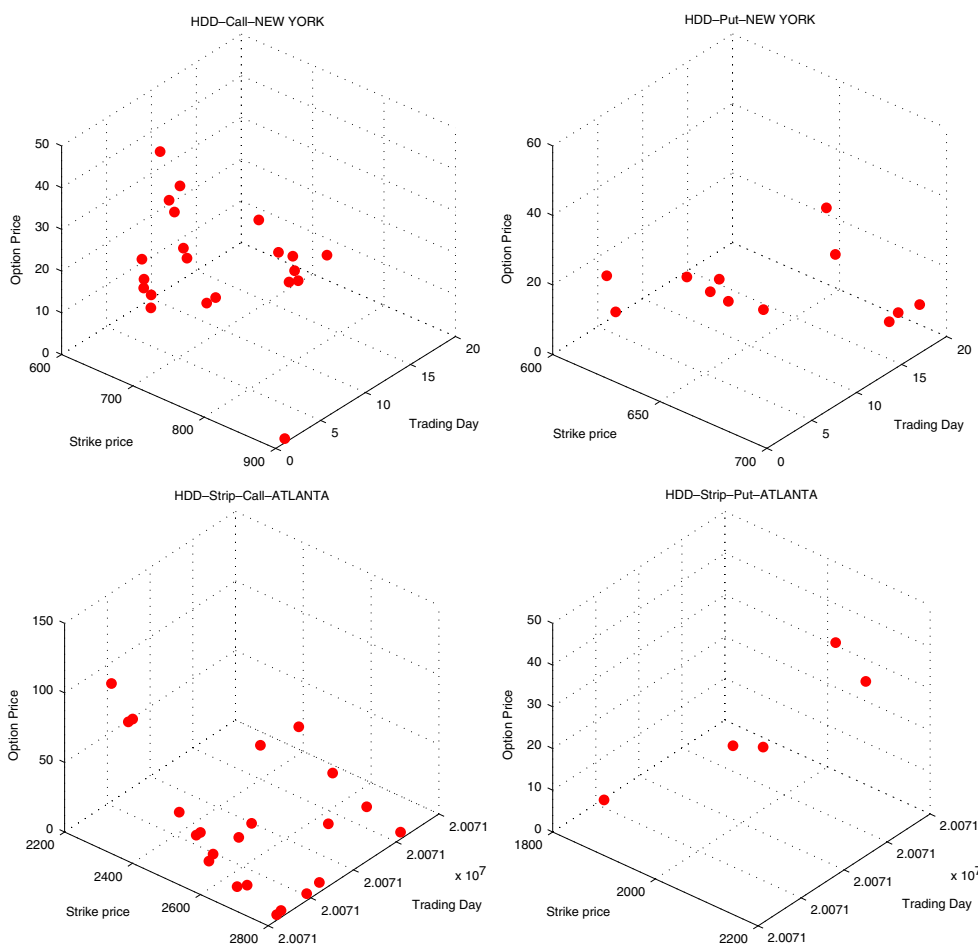


Fig. 12. New York monthly HDD option prices with time to maturity in one month against strike prices and trading days in March 2006, HDD Atlanta Seasonal option prices with time to maturity in 6 months and trading days in October 2007.

Recall that  $N_1$  and  $N_2$  are numbers of different strike prices for call and put options, respectively.  $N_i$  relies on the given data set.  $N_{ij}$  is the number of repeated options given a strike price and a type of option. In our empirical analysis, because we just take one daily closing price, we set  $N_{ij} = 1$ .

When  $R^2$  is close to one, model prices are close to market prices and the model produces nice fit. In the Bayesian quadrature method, we calculate  $R^2$  at each swipe of the Markov chain Monte Carlo algorithm, and summarize its posterior mean and quantiles for inference. Because all the mean, median, and the 2.5% and 97.5% quantiles of the  $R^2$  calculated in the MCMC algorithm are close to one, we conclude that the Bayesian quadrature method produces an almost perfect fit for all these trading days. Fig. 13 presents dynamics of the smoothed implied SPD, all of which deviate from lognormality.

A simple way to investigate the dynamics of the implied SPD at each trading day is to calculate moments based on the quadrature method. In each swipe of the Markov chain Monte Carlo algorithm, we calculate the mean ( $\mu$ ), volatility ( $v$ ), skewness ( $s$ ), and kurtosis ( $\kappa$ ) of the quadrature method, by the following formulas,

$$\mu = \sum_{n=1}^N w_n \theta_n$$

$$v = \sqrt{\sum_{n=1}^N w_n (\theta_n - \mu)^2}$$

$$s = \sum_{n=1}^N w_n (\theta_n - \mu)^3 / v^3$$

$$\kappa = \sum_{n=1}^N w_n (\theta_n - \mu)^4 / v^4.$$

Fig. 14 gives the dynamics of the posterior means of the SPDs. Table 5 shows the posterior means of these four quantities of the quadrature method. However, skewness and kurtosis of weather options can be either positively or negatively skewed depending on futures maturity.

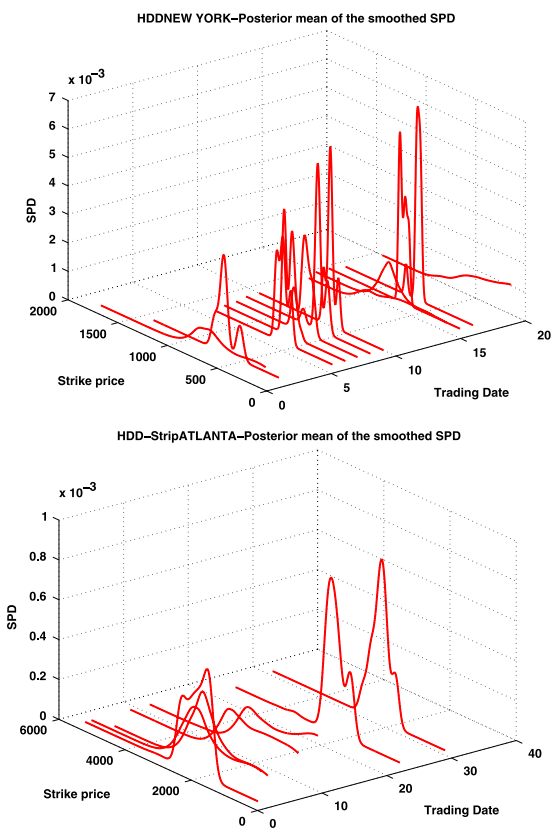
WD-SPD's tend to be positively skewed for short maturity contracts indicating that the tail on the right side is longer or fatter than the left side as the call-options only provide information in the right tail of the SPD. Conversely, negative skew indicates that the tail on the left side of the SPD, provided by put-options, is longer or fatter than the right side.

The heterogeneity beliefs on investors (hedgers versus speculators), reflected by the weather sensitivity preferences among agents, lead the SPD spread to the tails and even becomes bimodal. As shown in Jackwerth and Rubinstein (1996) and Rubinstein (1994), it is common to get in incomplete markets, like the stock index options, multimodal risk neutral densities. The presence of severe modes might be caused due to nonlinear relationship between the seasonal variance of the temperature process  $T_t$  and the underlying temperature index futures in (19), see Härdle and López-Cabrera (2012), Benth et al. (2007). Temperature

**Table 5**

Posterior mean of the mean ( $\mu$ ), volatility ( $v$ ), skewness ( $s$ ), and kurtosis ( $\kappa$ ) of the quadrature method at each trading day (TD) calibrated from New York HDD monthly options in March 2006, Atlanta HDD seasonal strip options in October 2007 and Dallas CDD monthly options in June 2010.

	TD	2	3	6	7	8	9	10	14	15	16	17	20
HDD	$\mu$	511.16	676.00	618.80	676.31	660.00	645.00	660.00	422.78	554.10	708.58	698.31	380.26
NY	$v$	248.98	81.57	57.02	73.88	61.38	59.89	45.07	240.72	217.04	41.54	42.11	205.93
	$s$	0.09	-0.92	0.25	4.30	0.48	0.21	0.38	0.43	-0.42	1.47	1.65	0.60
	$\kappa$	1.54	2.70	2.79	56.28	2.26	3.28	2.98	2.16	1.70	5.66	7.59	2.90
	TD	2	3	4	9	12	24	31					
HDD	$\mu$	2225.54	1741.96	1865.99	1630.19	1470.13	2335.12	2340.70					
Strips	$v$	324.45	818.35	757.39	751.52	835.64	287.44	258.52					
Atlanta	$s$	0.58	-0.05	-0.33	0.21	0.59	0.64	0.07					
	$\kappa$	3.44	1.42	1.73	2.15	2.44	7.59	3.00					
	TD	4	10	14									
CDD	$\mu$	489.96	390.42	596.63									
Dallas	$v$	175.79	195.06	155.40									
	$s$	-0.83	0.18	-0.57									
	$\kappa$	2.72	1.98	2.76									



**Fig. 13.** Quadrature method and smoothed SPDs implied from New York monthly HDD option prices with time to maturity in one month against strike prices and trading days in March 2006, HDD Atlanta Seasonal option prices with time to maturity in 6 months and trading days in October 2007, Dallas monthly CDD options with time to maturity in one month traded in June 2010.

tends to stay stable during periods with low seasonal variance. Thus, SPDs are depending on the conditional volatility: the SPD is wider when the conditional volatility is high. This is different with what documented on index options market (unimodal densities) (Bakshi et al., 2010), interest rate derivatives market (log-normal densities) (Li and Zhao, 2007) and temperature markets (unimodal normal densities) (Benth et al., 2007), but similar to rainfall markets (skewed densities) (López-Cabrera et al., 2013). This is also explained by the economic behavior of agents

sensitive to weather conditions. Investors expect that temperature variations, that affect their cash flows, will occur with high probability in winter times than in summer times (conversely for WDs in Australia). Hence some investors will use these option contracts for hedging purposes in presence of negative expected payoffs to eliminate their risk, while others will act as speculators from bearing hedgers' risk. The results show, as expected, that the option temperature market offers a much greater premium than the futures temperature market (Härdle and López-Cabrera, 2012).

**4. The infeasibility of other nonparametric methods**

Here, we compare the feasibility of our method with other known nonparametric approaches, which are popular tools avoiding risk of misspecification. Some of these methods estimate the SPD by differentiating an interpolation of smoothing of option prices. In this context, data sparsity makes the estimation of the SPD a statistical challenge. Let us now explain why the kernel regression method and mixtures of lognormals do not work well in the context of WD implied SPDs.

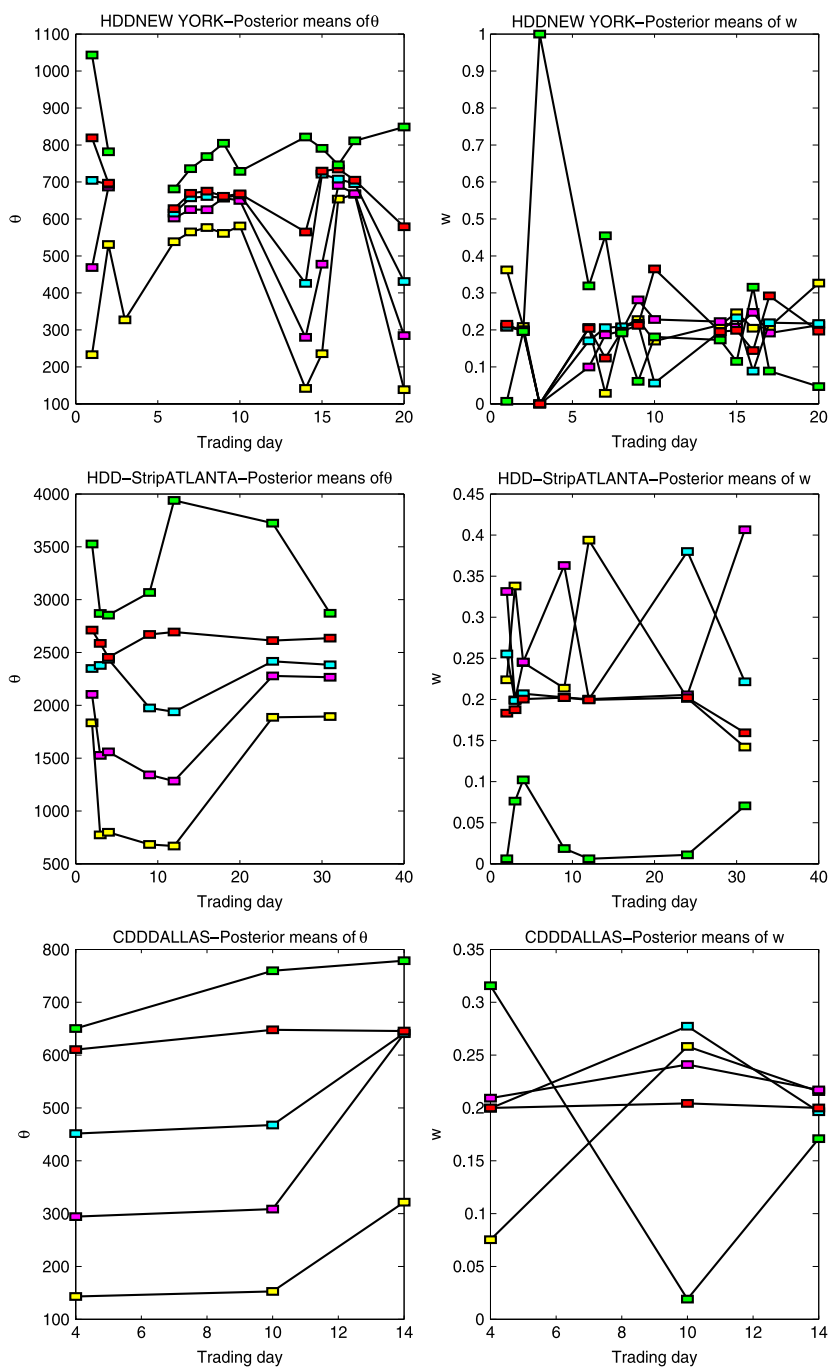
The cross-section of the call and put prices is given in Fig. 12. Different grids correspond to different contracts with different times to measurement periods and consequently one can argue that option prices can be extrapolated as a smoothed function of the strike.

**4.1. Kernel regression**

The kernel regression method (KRM) takes advantage of differentiating twice (1):

$$f(K) = e^{r\tau} \frac{\partial^2}{\partial K^2} C(K). \tag{23}$$

In order to employ (23), one needs more observations as the option price function is treated as a continuous function of strikes and therefore relies on the put–call parity to transfer put option prices to call option prices. When the market is rarely traded, it is not promising to employ the put–call parity though. In our empirical data analysis most options are traded with only a few strike prices. Very often, an option was traded only with one or two strike prices. When the kernel method is applied to a data set of such a case, it is even difficult to find an option function  $C(K)$ , not to mention to find its second derivatives and interpolated version, may not yield a density estimate that guarantee to be positive and integrate to one. Consequently, KRM is sensitive to data sparsity.



**Fig. 14.** Dynamics of the parameter of the quadrature method implied from New York monthly HDD option prices with time to maturity in one month against strike prices and trading days in March 2006, HDD Atlanta Seasonal option prices with time to maturity in 6 months and trading days in October 2007, Dallas monthly CDD options with time to maturity in one month traded in June 2010.

4.2. Mixture of lognormals

When applying the mixture of lognormal methods, it is necessary to specify the range of the variances of the lognormal density. The selection is objective and influences the estimated SPD dramatically. When the data set consists of a few data point, it is possible to produce two totally different densities (particularly in terms of variances) which produce the same quality of model fit.

We estimate the SPD using a mixture of lognormal for New York HDD monthly options, given in Fig. 8. For mixture of lognormals, we will show that two different SPD using mixture of lognormal

produce the same model fit, but they have quite different higher moments.

Yuan (2009) proposed a function class:

$$\mathcal{F} = \left\{ f(\cdot) : f(x) = \int f(x|\mu, \sigma^2)dG(\mu, \sigma), \right. \\ \left. \text{supp}(g) \subset [-M, M] \times [\underline{\sigma}, \bar{\sigma}] \right\}$$

where  $M < \infty$  and  $0 < \underline{\sigma} \leq \bar{\sigma} < \infty$ ,  $f(x|\mu, \sigma^2)$  is the pdf of the lognormal distribution with location  $\mu$  and scale  $\sigma$  and

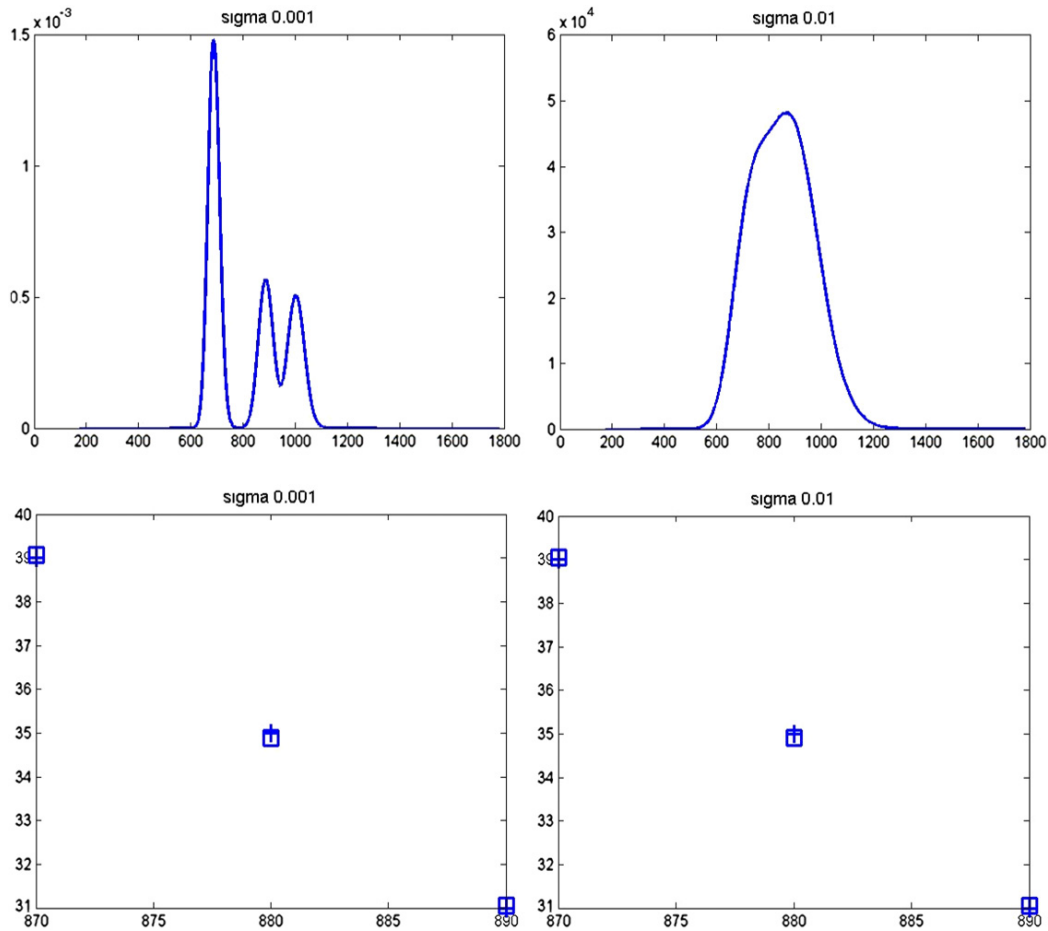


Fig. 15. Estimated SPD using mixtures of lognormal. The left-upper and right-upper panels give estimated SPD using  $\sigma$  equal to 0.01 and 0.001, respectively. The left-lower and right-lower panels give market prices indicated by a blue cross and model prices indicated by a blue square. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$G$  determines the mixing distribution. The corresponding pricing function in this case is similar to (2):

$$C(X; \mu, \sigma^2) = \exp(-r\tau) \int_0^\infty \wp_{ij}(x) f(x) dx$$

$$C(X; G) = \int C(X; \mu, \sigma^2) dG(\mu, \sigma) \tag{24}$$

as the SPD  $f(x)$  is defined in the previous family  $\mathcal{F}$ . The least squares estimate of the pricing function can be written as:

$$\hat{G}(\cdot) = \arg \min_{G \in \mathcal{G}} n^{-1} \sum_{i=1}^n \{y_{ijk} - C(X; G)\}^2 \tag{25}$$

where  $\mathcal{G}$  is the collection of all probability measures on  $\mu$  and  $\sigma^2$ . Note that the minimization is taken over a function space of infinite dimensions, however the solution can be represented in a finite dimensional space. In particular, all solutions can be expressed as a convex combination of at most  $n+1$  Black–Scholes type of pricing functions.

This model has several nice theoretical properties. For example, as the sample size  $n$  increases, the pricing functions can be recovered with squared error converging to zero at the rate of  $\log^2 n/n$ , which is close to the parametric rate of convergence  $1/n$ . However, practical difficulties arise when fitting mixtures of lognormal distributions (or other mixtures models) to real data. The feature that weather options are traded with a few number of

strike prices make mixture models inapplicable, because mixture models need to select corresponding scale parameters and the number of components. For example, when options are traded with  $n$  different strike prices, maximum likelihood suggests to use  $n/2$  support points. When  $n$  is large, this leads a very complicated model and possible over fitting problem. When  $n$  is small, the resulting model may be inappropriate. In addition, numerical procedure for searching the maximum likelihood estimate is particularly difficult for large  $n$ .

We apply mixture of lognormals by Yuan (2009) to the New York monthly HDD call options traded on 2006/02/02, with two different manually selected variances. Fig. 15 shows that these two estimated SPD are quite different in shapes, although they produce similar quality of model fit. Therefore, this illustration shows that the estimated SPD is very sensitive to the selection of  $\sigma$ . In practical implementation, Yuan (2009) suggests to determine  $\sigma$  by cross-validation. This however is very computationally demanding.

### 5. Conclusions

We estimate SPDs for WDs using the Bayesian quadrature method. The WD market is characterized by its incompleteness and less frequently traded activities. This makes the estimation of the SPD a statistical challenge. However, the quadrature method, in advantage to the parametric and other non-semiparametric techniques, avoids model miss-specification and allows the SPD estimation by a parsimonious model. The technique is



computationally fast and robust. The obtained SPD do not stem from market-risk-price assumptions. We present empirical results on real CME temperature derivatives data, which help us to understand the dynamics of SPD. The results suggest that the SPD of weather derivatives exhibits a non-normal behavior type.

## Acknowledgments

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 “Ökonomisches Risiko”, Humboldt-Universität zu Berlin is gratefully acknowledged. Huei-Wen Teng’s research is supported by the Ministry of Science and Technology, Taiwan (ROC), under grant 103-2633-M-008-001.

## References

- Abadir, K., Rockinger, M., 2003. Density functionals, with an option-pricing application. *Econometric Theory* 19 (5), 778–811.
- Aït-Sahalia, Y., Duarte, J., 2003. Nonparametric option pricing under shape restrictions. *J. Econometrics* 116, 9–47.
- Aït-Sahalia, Y., Lo, A.W., 1998. Nonparametric estimation of state-price densities implicit in financial asset prices. *J. Finance* 53, 499–547.
- Aït-Sahalia, Y., Lo, A.W., 2000. Nonparametric risk management and implied risk aversion. *J. Econometrics* 94, 9–51.
- Bakshi, G., Madan, D., Panayotov, G., 2010. Returns of claims on the upside and the viability of u-shaped pricing kernels. *J. Financ. Econ.* 97 (1), 130–154.
- Benth, F., Benth, S., Koekebakker, S., 2007. Putting a price on temperature. *Scand. J. Statist.* 34, 746–767.
- Benth, F., Härdle, W.K., López-Cabrera, B., 2011. Pricing Asian temperature risk. In: Cizek, , Härdle, , Weron, (Eds.), *Statistical Tools for Finance and Insurance*, second ed.. Springer Verlag, Heidelberg.
- Breeden, D., Litzenberger, R., 1978. Price of state-contingent claims implicit in option prices. *J. Bus.* 51, 621–651.
- Casella, G., Berger, R.L., 2001. *Statistical Inference*. Duxbury Press.
- Chen, M., Shao, Q., 1997. On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* 25 (4), 1563–1594.
- Derman, E., Kani, I., 1994. Riding on the smile. *Risk* 7, 32–39.
- Dorflleitner, G., Wimmer, M., 2010. The pricing of temperature futures at the Chicago Mercantile exchange. *J. Banking Finance* 34 (6), 1360–1370.
- Dumas, B., Fleming, J., Whaley, R.E., 1998. Implied volatility functions: Empirical tests. *J. Finance* 53 (6), 2059–2106.
- Dupire, B., 1994. Pricing with a smile. *Risk* 7, 18–20.
- Fan, J., Mancini, L., 2009. Option pricing with model-guided nonparametric methods. *J. Amer. Statist. Assoc.* 104 (488), 1351–1372.
- García, R., Ghysels, E., Renault, E., 2010. Econometrics of option pricing models. In: Ait-Sahalia, Y., Hansen, L.P. (Eds.), *Handbook of Financial Econometrics*, Vol. 1. North Holland, Amsterdam.
- Ghysels, E., Harvey, A., Renault, R., 1995. Stochastic volatility. In: Maddala, G.S., Rao, C.R. (Eds.), *Handbook of Statistics 14*, Statistical Methods in Finance. North Holland, Amsterdam.
- Ghysels, E., Patilea, V., Renault, E., Torres, O., 1997. Nonparametric methods and option pricing. In: Hand, D., Jacka, S. (Eds.), *Statistics in Finance*. Edward Arnold, London.
- Giacomini, R., Gottschling, A., Haefke, C., White, H., 2008. Mixtures of *t* distributions for finance and forecasting. *J. Econometrics* 144, 175–192.
- Härdle, W.K., Hlávka, Z., 2009. Dynamics of state price densities. *J. Econometrics* 150, 1–15.
- Härdle, W.K., López-Cabrera, B., 2012. Inferring the market price of weather risk. *Appl. Math. Finance* 19 (1), 59–95.
- Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. *Nonparametric and Semiparametric Models*. Springer Verlag, Heidelberg.
- Jackwerth, J., Rubinstein, M., 1996. Recovering probabilities distributions from option prices. *J. Finance* 51, 1611–1631.
- Li, H., Zhao, F., 2007. Nonparametric estimation of state-price densities implicit in interest rate cap prices. *Rev. Financ. Stud.* 22 (11), 4335–4376.
- Liechty, J., Teng, H.-W., 2009. Bayesian quadrature approaches to state price density estimation. working paper. Available at: <http://www.personal.psu.edu/jcl12/Bayesian%20Quadrature.pdf>.
- López-Cabrera, B., Odening, M., Ritter, M., 2013. Pricing rainfall futures at the CME. *J. Banking Finance* 37 (11), 4286–4298.
- Renault, E., 1997. Econometric models of options pricing errors. In: Kreps, D., Wallis, K. (Eds.), *Advances in Economics and Econometrics: Theory and Applications \* Seventh World Congress*, Vol. III. Cambridge University Press, Cambridge.
- Rosenberg, J., Engle, R., 2002. Empirical pricing kernels. *J. Financ. Econ.* 64, 341–372.
- Rubinstein, M., 1994. Implied binomial trees. *J. Finance* 49, 771–818.
- Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.
- Ueberhuber, C.W., 1997. *Numerical Computation 2: Methods, Software, and Analysis*. Springer-Verlag.
- Yatchew, A., Härdle, W., 2006. Nonparametric state price density estimation using constrained least squares and the bootstrap. *J. Econometrics* 133 (2), 579–599.
- Yuan, M., 2009. State price density estimation via nonparametric mixtures. *Ann. Appl. Stat.* 3 (3), 963–984.



## Distillation of News Flow into Analysis of Stock Reactions

Junni L. Zhang, Wolfgang K. Härdle, Cathy Y. Chen & Elisabeth Bommers

To cite this article: Junni L. Zhang, Wolfgang K. Härdle, Cathy Y. Chen & Elisabeth Bommers (2015): Distillation of News Flow into Analysis of Stock Reactions, Journal of Business & Economic Statistics, DOI: [10.1080/07350015.2015.1110525](https://doi.org/10.1080/07350015.2015.1110525)

To link to this article: <http://dx.doi.org/10.1080/07350015.2015.1110525>



Accepted author version posted online: 09 Nov 2015.



Submit your article to this journal [↗](#)



Article views: 22



View related articles [↗](#)



View Crossmark data [↗](#)

Full Terms & Conditions of access and use can be found at  
<http://www.tandfonline.com/action/journalInformation?journalCode=ubes20>

# Distillation of News Flow into Analysis of Stock Reactions\*

Junni L. Zhang

Guanghua School of Management and Center for Statistical Science  
Peking University  
Beijing, 100871, China

Wolfgang K. Härdle

Humboldt-Universität zu Berlin  
Unter den Linden 6, Berlin 10099, Germany  
and

Sim Kee Boon Institute for Financial Economics  
Singapore Management University

Administration Building, 81 Victoria Street, Singapore 188065

Cathy Y. Chen

Chung Hua University  
707, Sec.2, WuFu Rd., Hsinchu, Taiwan 30012  
and

Humboldt-Universität zu Berlin  
Unter den Linden 6, Berlin 10099, Germany

Elisabeth Bommers

Humboldt-Universität zu Berlin  
Unter den Linden 6, Berlin 10099, Germany

October 5, 2015

---

\* This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic Risk', Humboldt-Universität zu Berlin. We like to thank the Research Data Center (RDC) for the data used in this study. We would also like to thank the International Research Training Group (IRTG) 1792.

## Abstract

The gargantuan plethora of opinions, facts and tweets on financial business offers the opportunity to test and analyze the influence of such text sources on future directions of stocks. It also creates though the necessity to distill via statistical technology the informative elements of this prodigious and indeed colossal data source. Using mixed text sources from professional platforms, blog fora and stock message boards we distill via different lexica sentiment variables. These are employed for an analysis of stock reactions: volatility, volume and returns. An increased sentiment, especially for those with negative prospectation, will influence volatility as well as volume. This influence is contingent on the lexical projection and different across Global Industry Classification Standard (GICS) sectors. Based on review articles on 100 S&P 500 constituents for the period of October 20, 2009 to October 13, 2014, we project into BL, MPQA, LM lexica and use the distilled sentiment variables to forecast individual stock indicators in a panel context. Exploiting different lexical projections to test different stock reaction indicators we aim at answering the following research questions:

- (i) Are the lexica consistent in their analytic ability?
- (ii) To which degree is there an asymmetric response given the sentiment scales (positive v.s. negative)?
- (iii) Are the news of high attention firms diffusing faster and result in more timely and efficient stock reaction?
- (iv) Is there a sector specific reaction from the distilled sentiment measures?

We find there is significant incremental information in the distilled news flow and the sentiment effect is characterized as an asymmetric, attention-specific and sector-specific response of stock reactions.

*Keywords: Investor Sentiment, Attention Analysis, Sector Analysis, Volatility Simulation, Trading Volume, Returns*  
*JEL Classifications: C81, G14, G17*

# 1 Introduction

News are driving financial markets. News are nowadays massively available on a variety of modern digital platforms with a wide spectrum of granularity scales. It is exactly this combination of granularity and massiveness that makes it virtually impossible to process all the news relevant to certain financial assets. How to distinguish between “noise” and “signal” is also here the relevant question. With a few exceptions the majority of empirical studies on news impact work has therefore been concentrated on specific identifiable events like scheduled macroeconomic announcements, political decisions, or asset specific news. Recent studies have looked at continuous news flow from an automated sentiment machine and it has been discovered to be relevant to high frequency return, volatility and trading volume. Both approaches have limitations since they concentrate on identifiable indicators (events) or use specific automated linguistic algorithms.

This paper uses text data of different granularity from blog fora, news platforms and stock message boards. Using several lexical projections, we define pessimistic (optimistic) sentiment with specific meaning as the average proportions of negative (positive) words in articles published in specific time windows before the focal trading day, and examine their impacts on stock trading volume, volatility and return. We analyze those effects in a panel data context and study their influence on stock reactions. These reactions might be interesting since large institutions, more sophisticated investors, usually express their views on stock prospective or prediction through published analyst forecasts. However, analysts’ recommendations may be contaminated by their career concerns and compensation scheme; they may also be in alliance with other financial institutions such as investment banks, brokerage houses or target companies (Hong and Kubik, 2003; Liu, 2012). Due to the possible conflicts of interest from analysts and their powerful influence on naive small investors, the opinions from individual small investors may be trustworthy since their personal opinions hardly create any manipulation that governs stock reactions. The advent of social media such as *Seeking Alpha* enables small investors to share and express their opinions frequently, real time and responsively.

We show that small investors’ opinions contribute to stock markets and create a “news-driven” stock reaction. The conversation in the internet or social media is valuable since the

introduction of conversation among a subset of market participants may have large effects on the stock price equilibrium (Cao et al., 2001). Other literature such as Antweiler and Frank (2004), Das and Chen (2007), Chen et al. (2014) demonstrate the value of individual opinions on financial market. They show that small investor opinions predict future stock returns and earnings surprises even after controlling the financial analyst recommendation.

The projections (of a text into sentiment variables) we employ are based on three sentiment lexica: the BL, LM and MPQA lexica. They are used to construct sentiment variables that feed into the stock reaction analysis. Exploiting different lexical projections, and using different stock reaction indicators we aim at answering the following research questions:

- (i) Are the lexica consistent in their analytic ability to produce stock reaction indicators, including volatility, detrended log trading volume and return?
- (ii) To which degree is there an asymmetric response given the sentiment scales (positive v.s. negative)?
- (iii) Are the news of high attention firms diffusing faster and result in more timely and efficient stock reaction?
- (iv) Is there a sector specific reaction from the distilled sentiment measures?

Question (i) addresses the variation of news content across different granularity and lexica. Whereas earlier literature focuses on numerisized input indices like ReutersNews-Content or Google Search Volume Index, we would like to investigate the usefulness of automated news inputs for e.g. statistical arbitrage algorithms. Question (ii) examines the effect of different sentiment scales on stock reactions like volatility, trading volume and returns. Three lexica are employed that are producing different numerical values and thus raise the concern of how much structure is captured in the resulting sentiment measure. An answer to this question will give us insight into whether the well known asymmetric response (bad vs. good news) is appropriately reflected in the lexical projections. Question (iii) and (iv) finally analyze whether stylized facts play a role in our study. This is answered via a panel data scheme using GICS sector indicators and attention ratios.

Groß-Klußmann and Hautsch (2011) analyze in a high frequency context market reactions to the intraday stock specific “Reuters NewsScope Sentiment” engine. Their findings

support the hypothesis of news influence on volatility and trading volume, but are in contrast to our study since they are based on a single news source and confined to a limited number of assets for which high frequency data are available.

Antweiler and Frank (2004) analyze text contributions from stock message boards and find that the amount and bullishness of messages have predictive value for trading volume and volatility. On message boards, the self-disclosed sentiment to hold a stock position is not bias free, as indicated in Zhang and Swanson (2010). Tetlock (2007) concludes that negative sentiment in a Wall Street Journal column has explanatory power for downward movement of the Dow Jones. Bollen et al. (2011) classify messages from the micro-blogging platform Twitter in six different mood states and find that public mood helps to predict changes in daily Dow Jones values. Zhang et al. (2012) extends this by filtering the Twitter messages (tweets) for keywords indicating a financial context and they consider different markets such as commodities and currencies. Si et al. (2013) use a refined filtering process to obtain stock specific tweets and conclude that topic based Twitter sentiment improves day-to-day stock forecast accuracy. Sprenger et al. (2014) also use tweets on stock level and conclude that the number of retweets and followers may be used to assess the quality of investment advice. Chen et al. (2014) use articles and corresponding comments on *Seeking Alpha*, a social media platform for investment research, and show predictive value of negative sentiment for stock returns and earnings surprises. According to Wang et al. (2014), the correlation of Seeking Alpha sentiment and returns is higher than between returns and sentiment in Stocktwits, messages from a micro-blogging platform specialized in finance.

Using either individual lexical projections or a sentiment index comprising the common component of the three lexical projections, we find that the text sentiment shows an incremental influence on the stocks collected from S&P 500 constituents. An asymmetric response of the stock reaction indicators to the negative and positive sentiments is confirmed and supports the leverage effect, that is, the stocks react to negative sentiment more. The reaction to the distilled sentiment measures is attention-specific and sector-specific as well. Due to the advent of social media, the opinions of small traders that have been ignored from past till now, do shed some light on stock market activity. The rest of

the paper is organized as follows. Section 2 describes the data gathering process, summarizes definitions of variables and introduces the different sentiment lexica. In Section 3, we present the regression and simulation results using the entire sample and samples grouped by attention ratio and sectors. The conclusion follows in Section 4.

## 2 Data

### 2.1 Text Sources and Stock Data

While there are many possible sources of financial articles on the web, there are also legal and practical obstacles to clear before obtaining the data. The text source *Seeking Alpha*, as used in Chen et al. (2014), prohibits any application of automatic programs to download parts of the website (web scraper) in their Terms of Use (TOS). While the usage of web scrapers for non-commercial academic research is principally legal, these TOS are still binding as stated in Truyens and Eecke (2014). For messages on Yahoo! Finance, another popular source of financial text data used in Antweiler and Frank (2004); Zhang and Swanson (2010), the TOS are not a hindrance but only limited message history is provided. As of December 2014, only the last 10,000 messages are shown in each stock specific message board and this roughly corresponds to a two-month-period for stocks that people talk frequently about like Apple. In opposition to these two examples, NASDAQ offers a platform for financial articles by selected contributors including social media websites such as *Seeking Alpha* and *Motley Fool*, investment research firms such as Zacks. Neither do the TOS prohibit web scraping nor is the history of shown articles limited. We have collected 116,691 articles and corresponding stock symbols, spanning roughly five years from October 20, 2009 to October 13, 2014. The data is downloaded by using a self-written web scraper to automate the downloading process.

The process of gathering and processing the article data and producing the sentiment scores can be seen in Figure 1. Firstly, the URLs of all articles on NASDAQ are gathered and every webpage containing an article is downloaded. Each URL can be used in the next steps as unique identifier of individual articles to ensure that one article is not used twice due to real-time updates of the NASDAQ webpage. In the pre-processing step,



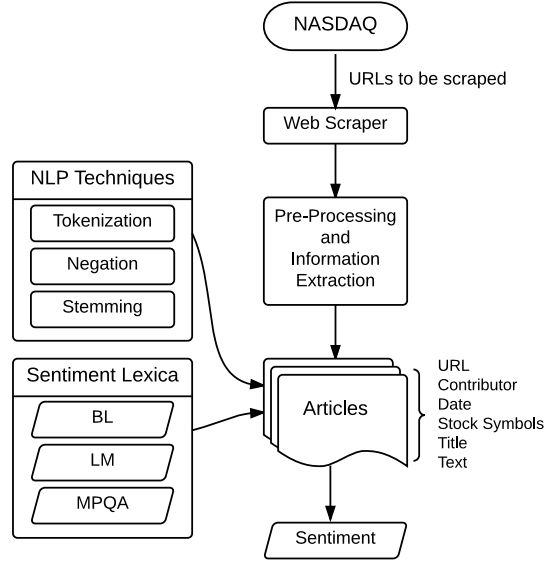


Figure 1: Flowchart of data gathering process

the page navigation and design elements of NASDAQ are removed. The specifics of each article, namely contributor, publication date, mentioned stock symbols, title and article text, are identified and read out. In case of the article text, the results are stored in individual text files. This database is available for research purposes at [RDC](#), CRC 649, Humboldt-Universität zu Berlin.

Furthermore, we collected stock specific financial data. Daily prices and trading volume, defined as number of shares traded, of all stock symbols that are S&P 500 constituents are collected from Datastream while Compustat is used to gather the GICS sector for these stocks.

We consider three stock reaction indicators: log volatility, detrended log trading volume and return. For stock symbol  $i$  and trading day  $t$ , we first compute the Garman and Klass (1980) range-based measure of volatility defined as:

$$\sigma_{i,t}^2 = 0.511(u - d)^2 - 0.019\{c(u + d) - 2ud\} - 0.383c^2 \quad (1)$$

$$\text{with } u = \log(P_{i,t}^H) - \log(P_{i,t}^O),$$

$$d = \log(P_{i,t}^L) - \log(P_{i,t}^O),$$

$$c = \log(P_{i,t}^C) - \log(P_{i,t}^O),$$

with  $P_{i,t}^H$ ,  $P_{i,t}^L$ ,  $P_{i,t}^O$ ,  $P_{i,t}^C$  as the daily highest, lowest, opening and closing stock prices, respectively. Chen et al. (2006) and Shu and Zhang (2006) show that the Garman and Klass range-based measure of volatility essentially provides equivalent results to high-frequency realized volatility. For example, Shu and Zhang (2006) find that an empirical test with S&P 500 index return data shows that the range-based variances are quite close to the high-frequency realized variance computed using the sum of 15-minute squared returns. Andersen and Bollerslev (1997) show that the high-frequency realized volatility is very sensitive to the selected interval. In addition, it is also affected by the bid/ask spread. The range-based measure of volatility, on the other hand, avoids the problems caused by microstructure effects. However, Alizadeh et al. (2002) argue that range based measures such as the Garman-Klass estimator do not make use of the log-normality of volatility. As shown by Andersen et al. (2001), log realized volatility is less skewed and less leptokurtic in comparison to raw realized volatility. Therefore, we use  $\log \sigma_{i,t}$  instead, which also avoids regressing on a strictly positive variable in the subsequent analysis.

Following Girard and Biswas (2007), we estimate the detrended log trading volume for each stock by using a quadratic time trend equation:

$$V_{i,t}^* = \alpha + \beta_1(t - t_0) + \beta_2(t - t_0)^2 + V_{i,t}, \quad (2)$$

where  $t_0$  is the starting point of the time window in consideration,  $V_{i,t}^*$  is the raw daily log trading volume and the residual  $V_{i,t}$  is the detrended log trading volume. In order to avoid imposing a look-ahead bias, for each trading day  $t$ , we use a rolling window of past 120 observations,  $V_{i,t-120}^*, \dots, V_{i,t-1}^*$  with  $t_0 = t - 120$ , to estimate the coefficients and get a one-step ahead out-of-sample forecast  $\hat{V}_{i,t}^*$ , and then calculate  $V_{i,t} = V_{i,t}^* - \hat{V}_{i,t}^*$ . Furthermore, we calculate the returns as  $R_{i,t} = \log P_{i,t}^C - \log P_{i,t-1}^C$ .

We focus on 100 stock symbols that are S&P 500 constituents on all 1,255 trading days between October 20, 2009 and October 14, 2014, that belong to one of nine major GICS sectors for stock symbols that are S&P 500 constituents on at least one trading day during this period, and that have the most trading days with articles. The distribution of GICS sectors among these 100 symbols are given in Table 1. Out of the 116,691 articles collected, there are 43,459 articles associated with these 100 stock symbols; the number of articles for these stocks range from 340 to 5,435, and the number of trading days with articles

ranges from 271 to 1,039. Most of the articles are not about one single symbol but contain references to several stocks.

GICS Sector	No. Stocks
Consumer Discretionary	21
Consumer Staples	9
Energy	6
Financials	12
Health Care	15
Industrials	10
Information Technology	21
Materials	4
Telecommunication Services	2

Table 1: Distribution of GICS sectors among the 100 stock symbols

## 2.2 Sentiment Lexica and Sentiment Variables

To distill sentiment variables from each article, we use and compare three sentiment lexica. The first lexicon (BL) is a list of 6,789 sentiment words (2,006 positive and 4,783 negative) compiled over many years starting from Hu and Liu (2004) and maintained by [Bing Liu at University of Chicago, Illinois](#). We filter each article with this lexicon and calculate the proportions of positive and negative words. The second lexicon (LM) is based on Loughran and McDonald (2011) which is specifically designed for financial applications, and contains 354 positive words, 2,329 negative words, 297 uncertainty words, 886 litigious words, 19 strong modal words and 26 weak modal words. To be consistent with the usage of the other lexica, we only consider the list of positive and negative words and calculate the proportions of positive and negative words for each article.

The third lexicon is the MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon by Wilson et al. (2005) which we later refer to as the MPQA lexicon. This lexicon

contains 8,222 entries. In order to show the rather tedious distillation process let us look at six example entries:

```
type=weaksubj len=1 word1=abandoned pos1=adj stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandonment pos1=noun stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandon pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abase pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abasement pos1=anypos stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abash pos1=verb stemmed1=y priorpolarity=negative
```

Here **type** refers to whether the word is classified as strongly subjective, indicating that the word is subjective in most contexts, or weakly subjective, indicating that the word only has certain subjective usages; **len** denotes the length of the word; **word1** is the spelling of the word; **pos1** is part-of-speech tag of the word, which could take values **adj** (adjective), **noun**, **verb**, **adverb**, or **anypos** (any part-of-speech tag); **stemmed1** is an indicator for whether this word is stemmed, where stemming refers to the process of reducing inflected (or sometimes derived) words to their word stem, base or root form; and **priorpolarity** refers to polarity of the word, which could take values **negative**, **positive**, **neutral**, or **both** (both negative and positive). The MPQA lexicon contains 4913 entries with negative polarity, 2718 entries with positive polarity, 570 entries with neutral polarity, and 21 entries with both polarity. To be consistent with the usage of the other two lexica, we only consider positive and negative polarity.

We first use the NLTK package in Python to tokenize sentences and (un-stemmed) words in each article, and derive the part-of-speech tagging for each word. We filter each tokenized article with the list of entries with **stemmed1=n** in the MPQA lexicon to count the number of positive and negative word. We then use the Porter Stemmer in the NLTK package to stem each word and filter each article with the list of entries with **stemmed1=y** in the MPQA lexicon. If a word has been assigned polarity in the first filtering step, it will no longer be counted in the second filtering step. For each article, we can thus count the numbers of negative and positive words, and divide them by the length of the article to get the proportions of negative and positive words.

Regardless of which lexicon is used, we use a variation of the approach in Hu and Liu (2004) to account for sentiment negation. If the word distance between a negation word (“not”, “never”, “no”, “neither”, “nor”, “none”, “n’t”) and the sentiment word is no larger

than 5, the positive or negative polarity of the word is changed to be the opposite of its original polarity.

Among the words that appear at least three times in our list of articles, there are 470 positive and 918 negative words that are unique to the BL lexicon, 267 positive and 916 negative words that are unique to the LM lexicon, and 512 positive and 181 negative words that are unique to the MPQA lexicon. The LM lexicon contains less unique positive words than the other two lexica, and the MPQA lexicon contains less unique negative words than the other two lexica. Table 2 presents the lists of ten most frequent positive words and ten most frequent negative words that are unique to these three lexica. Since the BL and MPQA lexica are designed for general purpose and the LM lexicon is designed specifically for financial applications, the unique words under the BL and MPQA lexica indeed look more general.

Words in the general-purpose lexica may be misclassified for financial applications; for example, the word “proprietary” in the negative list of the BL lexicon may refer to things like “a secure proprietary operating system that no other competitor can breach” and hence have a positive tone in financial applications, and the word “division” in the negative list of the MPQA lexicon may only refer to divisions of companies. However, financial analysis using textual information is unavoidably noisy, and words in the LM lexicon can also be misclassified; for example, the word “closing” in the negative list of the LM lexicon may actually refer to a positive event of closing a profitable deal. Also, the LM lexicon does not take into account financial words such as “debt” and “risks” in the BL lexicon.

We next investigate the pairwise relationship among the above three lexica. Among the words that appear at least three times in our list of articles, there are 131 positive and 322 negative words that are shared only by the BL and LM lexica, 971 positive and 1,164 negative words that are shared only by the BL and MPQA lexica, and 32 positive and 30 negative words that are shared only by the LM and MPQA lexica. It is not surprising that the two general-purpose lexica, BL and MPQA, share the most positive and negative words. Out of the two general-purpose lexica, BL lexicon shares more positive and negative words with the special-purpose LM lexicon. Table 3 presents the lists of ten most frequent positive words and ten most frequent negative words that are shared only by two of these

BL		LM		MPQA	
Positive (470)	Negative (918)	Positive (267)	Negative (916)	Positive (512)	Negative (181)
Available (5,836)	Debt (12,540)	Opportunities (4,720)	Declined (9,809)	Just (17,769)	Low (12,739)
Led (5,774)	Fell (9,274)	Strength (4,393)	Dropped (4,894)	Help (17,334)	Division (5,594)
Lead (4,711)	Fool (5,473)	Profitability (4,174)	Late (4,565)	Profit (15,253)	Least (5,568)
Recovery (4,357)	Issues (3,945)	Highest (3,409)	Claims (3,785)	Even (13,780)	Stake (4,445)
Work (3,808)	Risks (2,850)	Greater (3,321)	Closing (3,604)	Deal (13,032)	Slightly (3,628)
Helped (3,631)	Issue (2,821)	Surpassed (2,464)	Closed (3,378)	Interest (12,237)	Close (3,105)
Enough (3,380)	Falling (2,768)	Enable (2,199)	Challenges (2,574)	Above (12,203)	Trial (2,544)
Pros (2,841)	Aggressive (1,796)	Strengthen (2,157)	Force (2,157)	Accord (11,760)	Decrease (2,205)
Integrated (2,652)	Hedge (1,640)	Alliance (1,842)	Unemployment (2,062)	Natural (10,135)	Disease (2,001)
Savings (2,517)	Proprietary (1,560)	Boosted (1,831)	Question (1,891)	Potential (9,905)	Little (1,775)

Table 2: Lists of ten most frequent positive words and ten most frequent negative words that are unique to the BL, MPQA or LM lexica, along with their frequencies given in parentheses.

three lexica. Words shared by the two general-purpose lexica (BL and MPQA) may be misclassified for financial applications; for example, the word “gross” shared by the negative lists of these two lexica may refer to “the annual gross domestic product” and have a neutral tone. However, words shared by the LM lexicon and one of the general-purpose lexica may also be misclassified; for example, the word “critical” shared by the negative lists of the BL and LM lexica may appear in sentences such as “mobile devices are becoming critical tools in the worlds of advertising and market research” and have a positive tone.

The above discussion shows that projections using the three lexica are all noisy, therefore it is worthwhile to compare results from these projections. For each stock symbol  $i$  and

BL and LM		BL and MPQA		LM and MPQA	
Positive (131)	Negative (322)	Positive (971)	Negative (1164)	Positive (32)	Negative (30)
Gains (7,604)	Losses (5,938)	Free (133,395)	Gross (8,228)	Despite (7,413)	Against (8,877)
Gained (7,493)	Missed (3,165)	Well (3,0270)	Risk (7,471)	Able (5,246)	Cut (3,401)
Improved (7,407)	Declining (3,053)	Like (24,617)	Limited (5,884)	Opportunity (4,398)	Challenge (1,042)
Improve (5,726)	Failed (2,421)	Top (14,899)	Motley (5,165)	Profitable (3,580)	Serious (1,022)
Restructuring (3,210)	Concerned (1,991)	Guidance (11,715)	Crude (5,109)	Efficiency (2,615)	Contrary (401)
Gaining (3,150)	Declines (1,654)	Significant (10,576)	Cloud (4,906)	Popularity (1,588)	Severely (348)
Enhance (2,753)	Suffered (1,435)	Worth (10,503)	Fall (4,732)	Exclusive (1,225)	Despite (342)
Outperform (2,518)	Weaker (1,288)	Gold (9,303)	Mar (3,190)	Tremendous (611)	Argument (324)
Stronger (1,657)	Critical (1,131)	Support (9,120)	Hard (2,957)	Dream (581)	Seriously (240)
Win (1,491)	Drag (1,095)	Recommendation (8,993)	Cancer (2,521)	Satisfaction (410)	Staggering (209)

Table 3: Lists of ten most frequent positive words and ten most frequent negative words that are shared only by BL and LM lexica, only by BL and MPQA lexica, or only by LM and MPQA lexica, along with their frequencies given in parentheses.

each trading day  $t$ , we derive the sentiment variables listed in Table 4 based on articles associated with symbol  $i$  and published on or after trading day  $t$  and before trading day  $t + 1$ .

Sentiment Variable	Description
$I_{i,t}$	Indicator for whether there is an article.
$Pos_{i,t}$ (BL)	The average proportion of positive words using the BL lexicon.
$Neg_{i,t}$ (BL)	The average proportion of negative words using the BL lexicon.
$Pos_{i,t}$ (LM)	The average proportion of positive words using the LM lexicon.
$Neg_{i,t}$ (LM)	The average proportion of negative words using the LM lexicon.
$Pos_{i,t}$ (MPQA)	The average proportion of positive words using the MPQA lexicon.
$Neg_{i,t}$ (MPQA)	The average proportion of negative words using the MPQA lexicon.

Table 4: Sentiment variables for articles published on or after trading day  $t$  and before trading day  $t + 1$ .

### 3 Empirical Results

#### 3.1 Entire Sample Results

##### 3.1.1 Descriptive Statistics and Comparison of the Lexical Projections

Table 5 presents summary statistics of the sentiment variables derived using the BL, LM and MPQA lexical projections for 43,569 symbol-day combinations with  $I_{i,t} = 1$ , where  $I_{i,t}$  is defined in Table 4 and indicates whether there is an article associated with symbol  $i$  and published on or after trading day  $t$  and before trading day  $t + 1$ . This number is slightly different from the number of articles associated with the 100 selected symbols (43,459), since an article can be associated with multiple symbols. The positive proportion is the largest under the MPQA projection, and the smallest under the LM projection. The negative proportions under the three projections are similar. Polarity in Table 5 measures the relative dominance between positive sentiment and negative sentiment. For example, the situation,  $Pos_{i,t}$  (BL) >  $Neg_{i,t}$  (BL), accounts for 88.04% of the 43,569 observations. Note that under each projection, there are a small percentage of the observations for which  $Pos_{i,t} = Neg_{i,t}$ . Under both the BL and MPQA projections, positive sentiment is more dominant and widespread than negative sentiment. The LM projection, however, results in a relative balance between positive and negative sentiment.

To check whether the sentiment polarity actually reflects the sentiment of the articles,



Variable	$\hat{\mu}$	$\hat{\sigma}$	Max	Q1	Q2	Q3	Polarity
$Pos_{i,t}$ (BL)	0.033	0.012	0.134	0.025	0.032	0.040	88.04%
$Neg_{i,t}$ (BL)	0.015	0.010	0.091	0.008	0.014	0.020	10.51%
$Pos_{i,t}$ (LM)	0.014	0.007	0.074	0.009	0.013	0.018	55.70%
$Neg_{i,t}$ (LM)	0.012	0.009	0.085	0.006	0.011	0.016	40.17%
$Pos_{i,t}$ (MPQA)	0.038	0.012	0.134	0.031	0.038	0.045	96.26%
$Neg_{i,t}$ (MPQA)	0.013	0.008	0.133	0.007	0.012	0.017	2.87%

Note: Sample mean, sample standard deviation, maximum value, 1st, 2nd and 3rd quartiles, and polarity. These descriptive statistics are conditional on  $I_{i,t} = 1$ .

Table 5: Summary Statistics for Text Sentiment Variables

Manual Label	BL Label			LM Label			MPQA Label			Total
	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	
Pos	56	4	1	41	12	8	61	0	0	61
Neg	9	2	1	0	9	3	9	2	1	12
Neu	22	5	0	10	15	2	26	0	1	27
Total	87	11	2	51	36	13	96	2	2	100

Table 6: Sentiment Classification Results for 100 Randomly Selected Articles

we actually carefully checked and read the contents of 100 randomly selected articles and manually classified their polarity (positive, negative and neutral), and also use the lexical projections to automatically classify these articles as follows. If the proportion of positive words for an article is larger than (or small than, or equal to) the proportion of negative words for the same article, then this article is automatically classified as positive (or negative, or neutral). Table 6 reports the results. It appears that the BL and MPQA projections put too much weight on positive sentiment, and are not powerful in detecting negative sentiment. In contrast, the LM sentiment is powerful in detecting negative sentiment, but is not so good in detecting positive sentiment.

Figure 2 and 3 respectively show the monthly correlation between positive and negative proportions under two of the three projections. In general, the negative proportions are

more correlated than positive proportions. Also, the correlation between the BL and LM projections and that between the BL and MPQA projections are larger than the correlation between the LM and MPQA projections, which is consistent with the discussion about the list of words shared by two of the three projections (see Table 3).

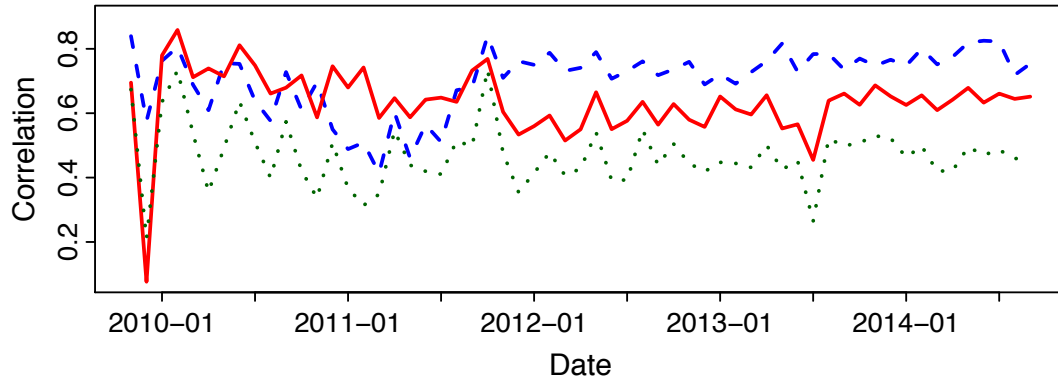


Figure 2: Monthly Correlation between Positive Sentiment: BL and LM (solid), BL and MPQA (dashed), LM and MPQA (dotted)

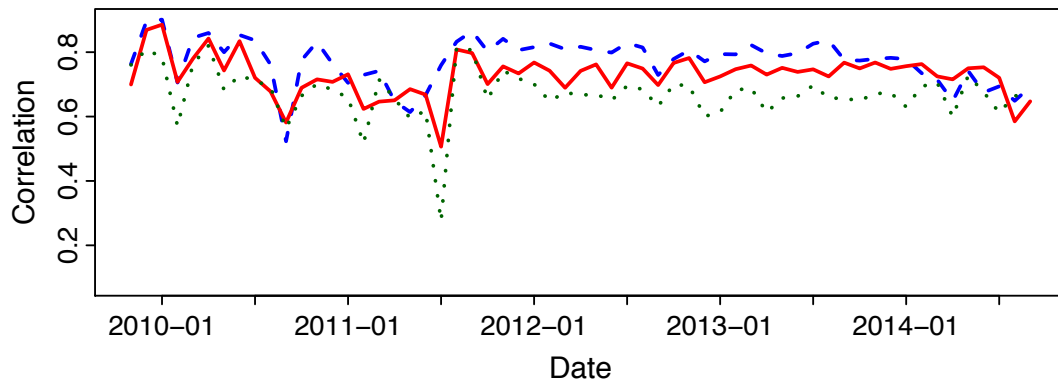


Figure 3: Monthly Correlation between Negative Sentiment: BL and LM (solid), BL and MPQA (dashed), LM and MPQA (dotted)

### 3.1.2 Main Results

Recall from Section 2.1 that we focus on three stock reaction indicators: log volatility  $\log \sigma_{i,t}$ , where  $\sigma_{i,t}^2$  is defined in (1), detrended log trading volume  $V_{i,t}$  as in (2) and returns  $R_{i,t}$ . We first consider analyzing these three indicators with one trading day into the future, and use the following (separate) panel regressions.

$$\log \sigma_{i,t+1} = \alpha + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t}, \quad (3)$$

$$V_{i,t+1} = \alpha + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t}, \quad (4)$$

$$R_{i,t+1} = \alpha + \beta_1 I_{i,t} + \beta_2 Pos_{i,t} + \beta_3 Neg_{i,t} + \beta_4^\top X_{i,t} + \gamma_i + \varepsilon_{i,t}. \quad (5)$$

where  $\gamma_i$  is the fixed effect for stock symbol  $i$  satisfying  $\sum_i \gamma_i = 0$ .  $X_{i,t}$  is a vector of control variables that includes a set of market variables to control for systematic risk such as (1) S&P 500 index return ( $R_{M,t}$ ) to control for general market returns; (2) the CBOE VIX index on date  $t$  to measure the generalized risk aversion ( $VIX_t$ ); and a set of firm idiosyncratic variables such as (3) the lagged log volatility ( $\log \sigma_{i,t}$ ); (4) the lagged return ( $R_{i,t}$ ); (5) the lagged detrended log trading volume ( $V_{i,t}$ ), where the lagged dependent variable is used to capture the persistence and omitted variables. These three indicators essentially have a triple dynamic correlation, and they have been modeled as a trivariate vector autoregressive (VAR) model, see Chen et al. (2001) and Chen et al. (2002). Our indicators in Eqs.(4) to (5) not only have themselves dynamic relationship with their lagged values, but also are impacted by the other lagged indicators. We incorporate clustered standard errors by Arellano (1987) as they allow for both time and cross-sectional dependence in the residuals. Petersen (2009) concludes that standard errors clustered on both dimensions are unbiased and achieve correctly sized confidence intervals while ordinary least squares standard errors might be biased in a panel data setting.

To answer our research question (i), if the three lexica are not consistent in their analytic ability to produce stock reaction indicators, we would expect that the value and the significance of  $\beta_1$ ,  $\beta_2$  or  $\beta_3$  varies across three lexical projections. For question (ii), if the positive and negative sentiments have asymmetric impacts, we would expect that  $\beta_2$  and  $\beta_3$  have different signs or significance. To address question (iii), we would expect that the value and the significance of  $\beta_1$ ,  $\beta_2$  or  $\beta_3$  varies with different attention levels and in

Variable	BL		LM		MPQA		PCA	
Panel A: Future Log Volatility $\log \sigma_{i,t+1}$								
$I_{i,t}$	-0.005	(0.009)	-0.019***	(0.007)	-0.004	(0.010)	-0.014	(0.010)
$Pos_{i,t}$	-0.396*	(0.228)	0.156	(0.378)	-0.517**	(0.217)	-0.210	(0.201)
$Neg_{i,t}$	0.905***	(0.257)	0.942***	(0.271)	1.464***	(0.325)	1.041***	(0.247)
$R_{M,t}$	-1.507***	(0.217)	-1.501***	(0.216)	-1.500***	(0.216)	-1.505***	(0.216)
$VIX_t$	2.329***	(0.085)	2.335***	(0.085)	2.331***	(0.086)	2.330***	(0.085)
$\log \sigma_{i,t}$	0.242***	(0.010)	0.242***	(0.010)	0.242***	(0.010)	0.242***	(0.010)
$R_{i,t}$	1.652***	(0.196)	1.653***	(0.196)	1.651***	(0.196)	1.653***	(0.196)
$V_{i,t}$	0.065***	(0.006)	0.065***	(0.006)	0.065***	(0.006)	0.065***	(0.006)
Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$								
$I_{i,t}$	0.040***	(0.008)	0.027***	(0.005)	0.046***	(0.009)	0.035***	(0.008)
$Pos_{i,t}$	-0.496***	(0.188)	0.051	(0.275)	-0.483**	(0.194)	-0.274*	(0.166)
$Neg_{i,t}$	0.726***	(0.257)	0.563**	(0.251)	0.548*	(0.290)	0.590**	(0.232)
$R_{M,t}$	-3.625***	(0.181)	-3.620***	(0.181)	-3.617***	(0.181)	-3.622***	(0.181)
$VIX_t$	-0.492***	(0.027)	-0.487***	(0.027)	-0.487***	(0.027)	-0.489***	(0.027)
$\log \sigma_{i,t}$	0.132***	(0.004)	0.132***	(0.004)	0.132***	(0.004)	0.132***	(0.004)
$R_{i,t}$	1.164***	(0.126)	1.166***	(0.126)	1.164***	(0.126)	1.166***	(0.126)
Panel C: Future Returns $R_{i,t+1}$								
$I_{i,t}$	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	-0.000	(0.000)
$Pos_{i,t}$	0.019***	(0.007)	0.030***	(0.011)	0.014*	(0.008)	0.018***	(0.006)
$Neg_{i,t}$	-0.004	(0.008)	-0.000	(0.010)	-0.009	(0.010)	-0.003	(0.008)
$R_{M,t}$	-0.050***	(0.006)	-0.050***	(0.006)	-0.050***	(0.006)	-0.050***	(0.006)
$VIX_t$	0.011***	(0.001)	0.011***	(0.001)	0.011***	(0.001)	0.011***	(0.001)
$\log \sigma_{i,t}$	-0.001***	(0.000)	-0.001***	(0.000)	-0.001***	(0.000)	-0.001***	(0.000)
$R_{i,t}$	-0.018***	(0.007)	-0.018***	(0.007)	-0.018***	(0.007)	-0.018***	(0.007)
$V_{i,t}$	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)

\*\*\* refers to a  $p$  value less than 0.01, \*\* refers to a  $p$  value more than or equal to 0.01 and smaller than 0.05, and \* refers to a  $p$  value more than or equal to 0.05 and less than 0.1. Values in parentheses are clustered standard errors.

Table 7: Entire Panel Regression Results

particular that the coefficient size is larger for higher attention firms. As to question (iv), we would expect that the coefficients of sentiment variables are sector-specific.

We will discuss the analysis of different attention levels and different sectors respectively in Sections 3.2 and 3.3, and focus now on the entire sample. The regression results are given in Table 7. Results in Panel A indicate that the arrival of articles ( $I_{i,t}$ ) distilled using the LM method is strongly negatively related to future log volatility, and that contingent on arriving articles, the negative sentiment distilled using the three methods is significantly

positively related to future log volatility, whereas the positive sentiment distilled using the BL and MPQA methods is significantly negatively related to future log volatility. Results in Panel B show that contingent on arriving articles, the positive and negative sentiment have asymmetric strong impacts on future detrended log trading volume: the negative sentiment across three lexica strongly drives up future detrended log trading volume, whereas the positive sentiment distilled using the BL and MPQA methods is strongly negatively related to future detrended log trading volume. The arrival of articles also strongly drives up future detrended log trading volume across three lexica. These findings support the mixture of distribution hypothesis originated by Clark (1973). As to future returns in Panel C, across three lexica and contingent on arriving articles, the positive sentiments are strongly positively related to future returns whereas the negative sentiment is unrelated to future returns. This finding sheds light on case against one unpleasant finding from Antweiler and Frank (2004) in which bullishness is not statistically significant for future return. It is interesting to note that the coefficients for the control variables do not vary much across lexical projections, which indicates that the sentiment measures are not so much correlated with the control variables and indeed provide incremental information.

It is difficult to diagnose a consensual performance from Table 7 because each lexicon may not fully reflect the complete sentiment and may have its own idiosyncratic nature as being evident from Table 2. To overcome this problem that none of the lexica is perfectly complete, we design an artificial sentiment index: the first principal component, to capture a common component of three lexica and to take into account the fact from Figures 2 and 3 that they reveal the shared sentiment. The positive (negative) sentiment index explains 94.14% (92.33%) of the total sample variance. As seen in the last column of Table 7, these general positive and negative sentiment indices are beneficial to achieve more consistent and interpretable results. The negative sentiment index spurs the future stock volatility and trading volume. However, the positive sentiment index has very restrictive influence on future volatility, and suppresses the trading volume but increases stock returns.

### 3.1.3 Sentiment Effect with Larger Lags and Neutral Sentiment

Based on the sequential arrival of information hypothesis (hereafter SAIH, Copeland, 1976, 1977), information arrives to traders at different times and hence relationship with lags larger than one can exist. Hence, we extend the length of lag under investigation to be two to five trading days and run regressions using the entire sample. From Table 8, volatility still reacts to the news in lagged two days but no more earlier than it: lagged two day negative sentiment extracted by BL and LM are influential, indicating that the SAIH has been observed here but lagged relationship is restricted to past one and two day while article was posting. In this sense, the market seems efficient to incorporate information no longer than two days. Likewise, we find the negative sentiment in lagged two day still has an influence on future return. The coefficients across three lexicon projections are significant but positive. The coefficients of negative sentiment projected by the BL and LM methods are significant but positive. The negative sign, even insignificant, in lagged one day turns positive in lagged two day to reflect that stock returns revert to mean value, which is consistent with Antweiler and Frank (2004). Although not significant, the coefficients' sign for lag one indicates a slight negative influence on tomorrow's stock returns, but return will revert to its mean value in two days later shown by positive sign as negative news vanish. The sooner reversion is the more efficient market is. For the detrended log trading volume, the lagged effect is relatively insignificant.

Financial market is characterized by the clustering of information (news) arrival, so that we will see the volatility clustering (Engle, 2004). The clustering of arrival of sentimental information motivates us to accumulate the sentiment variables from past trading days. Let  $I_{i,t:(t+h-1)}$ ,  $Pos_{i,t:(t+h-1)}$  and  $Neg_{i,t:(t+h-1)}$  denote the indicator of arrival of articles, the average proportion of positive words and the average proportion of negative words based on articles published on or after trading day  $t$  and before trading day  $t+h$ . Strikingly, the accumulated sentiment effect projected by BL and LM method on future volatility shown in Table 9 is very clear and keeps asymmetric, that is, only reacts to negative not to positive sentiment. Sometimes the sentiment news arrive consecutively and its accumulated influence lasts up to five trading days (one week). The accumulative sentiment effect can be also observed on the detrended log trading volume while accumulating to lagged four

and five days, and on the future return while accumulating to lagged two days.

We also tried to consider the proportion of neutral words and examine its impact. Based on the neutral proportion defined by MPQA method, in general we find the neutral words have no influence in stock indicators. The results can be provided upon the request.

Lag $h$	BL			LM			MPQA		
	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$
Panel A: Future Volatility $\log \sigma_{i,t+h}$									
$h = 2$	-0.000 (0.000)	0.000 (0.002)	0.005* (0.003)	-0.000 (0.000)	0.001 (0.003)	0.005* (0.003)	-0.000 (0.000)	0.001 (0.002)	0.004 (0.003)
$h = 3$	-0.000 (0.000)	-0.001 (0.002)	0.003 (0.003)	-0.000 (0.000)	0.001 (0.003)	0.004 (0.003)	-0.000 (0.000)	0.000 (0.002)	0.003 (0.003)
$h = 4$	-0.000 (0.000)	0.000 (0.002)	0.002 (0.003)	-0.000 (0.000)	0.002 (0.003)	0.004 (0.003)	-0.000 (0.000)	0.001 (0.002)	0.000 (0.003)
$h = 5$	0.000 (0.000)	-0.002 (0.002)	0.002 (0.003)	0.000 (0.000)	-0.002 (0.003)	0.003 (0.003)	-0.000 (0.000)	-0.000 (0.002)	0.001 (0.003)
Panel B: Future Detrended Log Trading Volume $V_{i,t+h}$									
$h = 2$	0.003 (0.006)	0.112 (0.140)	-0.198 (0.174)	0.004 (0.005)	0.079 (0.227)	-0.158 (0.183)	0.003 (0.007)	0.006 (0.140)	-0.414 (0.219)
$h = 3$	0.001 (0.006)	-0.011 (0.140)	-0.082 (0.174)	0.001 (0.005)	-0.003 (0.227)	-0.125 (0.183)	0.002 (0.007)	-0.170 (0.140)	-0.188 (0.219)
$h = 4$	-0.001 (0.006)	0.064 (0.140)	-0.539 (0.488)	0.004 (0.005)	-0.324 (0.227)	-0.556 (0.536)	0.001 (0.007)	-0.020 (0.140)	-0.811 (0.479)
$h = 5$	0.008 (0.006)	-0.208 (0.140)	-0.410 (0.301)	-0.004 (0.005)	-0.022 (0.227)	-0.096 (0.183)	0.001 (0.007)	-0.069 (0.140)	-0.416 (0.278)
Panel C: Future Returns $R_{i,t+h}$									
$h = 2$	-0.000 (0.000)	0.000 (0.007)	0.016* (0.009)	-0.000 (0.000)	-0.003 (0.012)	0.024** (0.010)	-0.000 (0.000)	0.001 (0.008)	0.026** (0.012)
$h = 3$	0.000 (0.000)	-0.001 (0.008)	-0.001 (0.009)	0.000 (0.000)	-0.010 (0.012)	0.005 (0.010)	0.001 (0.000)	-0.011 (0.008)	0.003 (0.012)
$h = 4$	-0.000 (0.000)	0.001 (0.007)	0.016* (0.009)	-0.000 (0.000)	0.010 (0.012)	0.006 (0.010)	-0.000 (0.000)	-0.003 (0.008)	0.011 (0.012)
$h = 5$	0.000 (0.000)	-0.011 (0.007)	0.009 (0.009)	0.000 (0.000)	-0.018 (0.012)	0.002 (0.010)	0.000 (0.000)	-0.013 (0.009)	0.014 (0.012)

\*\*\* refers to a  $p$  value less than 0.01, \*\* refers to a  $p$  value more than or equal to 0.01 and smaller than 0.05, and \* refers to a  $p$  value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

Table 8: Entire Panel Regression Results with Larger Lags (Noncumulative Articles)

Lag $h$	BL			LM			MPQA		
	$I_{i,t:(t+h-1)}$	$Pos_{i,t:(t+h-1)}$	$Neg_{i,t:(t+h-1)}$	$I_{i,t:(t+h-1)}$	$Pos_{i,t:(t+h-1)}$	$Neg_{i,t:(t+h-1)}$	$I_{i,t:(t+h-1)}$	$Pos_{i,t:(t+h-1)}$	$Neg_{i,t:(t+h-1)}$
Panel A: Future Volatility $\log \sigma_{i,t+h}$									
$h = 2$	-0.000 (0.000)	-0.001 (0.002)	0.006** (0.002)	-0.000 (0.000)	-0.001 (0.003)	0.007** (0.003)	-0.000 (0.000)	-0.001 (0.002)	0.004 (0.003)
$h = 3$	-0.000 (0.000)	-0.002 (0.002)	0.006*** (0.002)	-0.000* (0.000)	-0.001 (0.003)	0.008*** (0.003)	-0.000 (0.000)	-0.001 (0.002)	0.005 (0.003)
$h = 4$	-0.000 (0.000)	-0.001 (0.002)	0.006*** (0.002)	-0.000** (0.000)	-0.000 (0.003)	0.008*** (0.003)	-0.000 (0.000)	-0.001 (0.002)	0.003 (0.003)
$h = 5$	0.000 (0.000)	-0.003 (0.002)	0.006** (0.002)	-0.000 (0.000)	-0.002 (0.003)	0.008*** (0.003)	-0.000 (0.000)	-0.002 (0.002)	0.003 (0.003)
Panel B: Future Detrended Log Trading Volume $V_{i,t+h}$									
$h = 2$	0.006 (0.006)	-0.016 (0.125)	-0.133 (0.156)	0.008* (0.004)	-0.148 (0.203)	-0.187 (0.169)	0.002 (0.006)	0.006 (0.126)	-0.253 (0.198)
$h = 3$	0.006 (0.005)	-0.072 (0.123)	-0.111 (0.153)	0.005 (0.004)	-0.189 (0.198)	-0.078 (0.167)	0.001 (0.006)	-0.063 (0.124)	-0.174 (0.193)
$h = 4$	0.008 (0.005)	-0.310** (0.152)	-0.096 (0.124)	0.010** (0.004)	-0.486** (0.200)	-0.293* (0.168)	0.004 (0.006)	-0.138 (0.125)	-0.473 (0.327)
$h = 5$	0.014** (0.006)	-0.242* (0.126)	-0.408* (0.246)	0.008* (0.004)	-0.428** (0.202)	-0.228 (0.171)	0.009 (0.007)	-0.193 (0.126)	-0.646 (0.493)
Panel C: Future Returns $R_{i,t+h}$									
$h = 2$	-0.001* (0.000)	0.013** (0.007)	0.009 (0.008)	-0.000 (0.000)	0.019* (0.011)	0.010 (0.009)	-0.001** (0.000)	0.013* (0.007)	0.009 (0.010)
$h = 3$	-0.000 (0.000)	0.009 (0.007)	0.004 (0.008)	-0.000 (0.000)	0.013 (0.011)	0.007 (0.009)	-0.000 (0.000)	0.004 (0.007)	0.007 (0.010)
$h = 4$	-0.000 (0.000)	0.008 (0.007)	0.012 (0.008)	-0.000 (0.000)	0.017 (0.011)	0.009 (0.009)	-0.001 (0.000)	0.005 (0.007)	0.016 (0.010)
$h = 5$	-0.000 (0.000)	0.000 (0.007)	0.010 (0.008)	-0.000 (0.000)	0.004 (0.011)	0.006 (0.009)	-0.000 (0.000)	-0.003 (0.007)	0.019 (0.010)

\*\*\* refers to a  $p$  value less than 0.01, \*\* refers to a  $p$  value more than or equal to 0.01 and smaller than 0.05, and \* refers to a  $p$  value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

Table 9: Entire Panel Regression Results with Larger Lags (Cumulative Articles)

### 3.1.4 Monte Carlo Simulation based on Entire Sample Results

The text sentiment effects, as reported in Table 7, allow us deeper insights and analysis. More precisely we may address the important question of asymmetric reactions to the given sentiment scales. In order to do so we employ Monte Carlo techniques to investigate different facets of the sentiment effects. The components of this Monte Carlo study are: (1) to simulate the appearance of articles with presumed probabilities; (2) to provide a realistic set of scenarios regarding the frequency and content (positive v.s. negative) of articles; (3) to obtain volatility induced by the generated article (using Table 7); (4) to demonstrate the impact of synthetic text on future volatility; (5) to visualize and test an asymmetry



effect as formulated in research question (ii).

The simulation scenarios (for each variable involved) are summarized briefly as follows. We employ a Bernoulli random variable  $I_{i,t}$  indicating that articles arrive at a specific frequency  $p_i$ , where for each individual stock symbol  $i$ ,  $p_i$  is estimated by the fraction of days with at least one relevant article. Given the outcome of this article indicator, we generate the corresponding positive and negative proportions through a copula approach using the conditional inversion method as described in Frees and Valdez (1998). We follow the two-step approach that is widely mentioned in literature such as Patton (2006), Hotta et al. (2006) and Di Clemente and Romano (2004). In the first step, the marginal distributions are modeled by their corresponding empirical distribution function (edf) to avoid imposing a parametric distribution; in the second step, a Gaussian copula is estimated to take the inherent dependence among variables into account. For the sentiment variables, this approach is applied to each firm separately since each firm has a different  $p_i$  and only days with at least one article relevant to the firm are included in the estimation. To simulate market returns  $R_{M,t}$  and individual returns  $R_{i,t}$  for all 100 symbols, we first filter these variables by estimated MA(1)-GARCH(1,1) processes and standardize the residuals by dividing them by estimated standard deviations. We then apply the copula approach to the standardized residuals, and the simulated standardized residuals are transformed into simulated values of  $R_{M,t}$  or  $R_{i,t}$  by multiplying them by the median of the priorly estimated standard deviations for the market or the specific firm  $i$ . The company specific fixed effects  $\gamma_i$  are not incorporated as the simulated volatility for different firms is otherwise not graphically comparable. For the other control variables, CBOE VIX index  $VIX_t$  is fixed at its mean value over the sample period, and past log volatility and past detrended log trading volume are not used in the simulation.

Figure 4 demonstrates, for one simulation, the association between the negative and positive proportions as distilled via our three projection methods and their simulated future volatility outcomes. We estimate a local linear regression model (solid line) and corresponding 95% uniform confidence bands based on Sun and Loader (1994). Both are estimated using Locfit by Loader (1999) in the R environment. Loader and Sun (1997) discuss the robustness of this approach and conclude that the results are conservative but

reasonable for heavy tailed error distributions. The bandwidth is automatically chosen by using the plug-in selector according to Ruppert et al. (1995). We limit the visible area to sentiment values between 0 and 0.04 as well as volatility values between 1.45 and 1.65 to make the different lexica visually comparable. Nevertheless, all simulated values are utilized in the estimation of the regression curve and confidence bands. The clustered points lying on the vertical axis indicate that there is absence of articles. The range of this cluster from 0.77 to 2.57 is caused by the impact from the simulated control variables as well as the idiosyncratic impact captured by the residual term.

Apparently, an asymmetry effect becomes visible. One observes that the slopes of the volatility curves given negative sentiment is mainly positive while the curves for positive sentiment seem to be rather flat and even go down in the case of BL and MPQA methods. One can also compare the confidence bands to address the question whether negative sentiment has a significantly higher effect on the volatility than positive sentiment. The confidence bands of *Pos* and *Neg* do not overlap for sentiment values between 0.023 and 0.056 for BL, between 0.017 and 0.039 for LM and between 0.023 and 0.05 for MPQA.

This asymmetry effect parallels the well known imbalance of future volatility given good v.s. bad news. The leverage effect depicts a negative relation between the lagged return and the risk resulting from bad news that causes higher volatility. Black (1976) and Christie (1982) find that bad news in the financial market produce such an asymmetric effect on future volatility relative to good news. This leverage effect has also been shown by Bekaert and Wu (2000) and Feunou and Tédongap (2012). In the same vein, Glosten et al. (1993) introduce GARCH with differing effects of negative and positive shocks taking into account the leverage effect.

### 3.2 Does Attention Ratio matter?

While people post their text to express their opinions, or the comments to other articles, they are undoubtedly paying attention to the firm mentioned by their articles. In this respect article posting is a revealed attention measure. In fact, in our collected 43,459 articles across 100 stocks, it is obvious that not every firm shares the attention equivalently. To reflect these differences, we define the attention ratio for a symbol as the number of

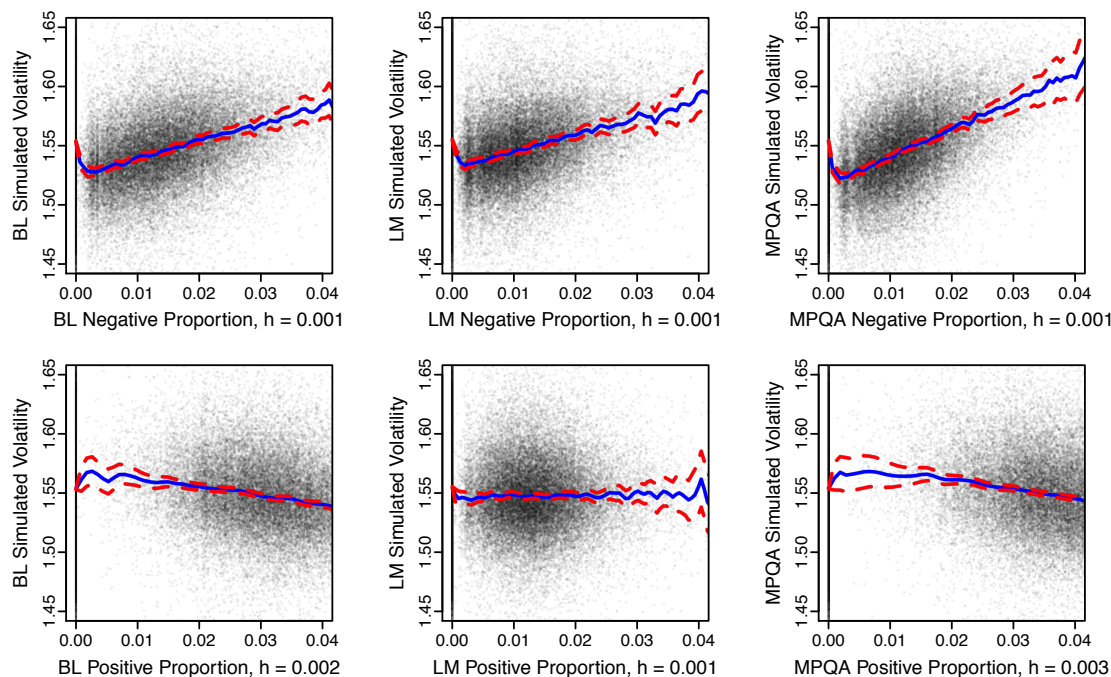


Figure 4: Monte Carlo Simulation based on Entire Sample Results

days with articles divided by the total number of days in the sample period, 1,255. The symbol “AAPL” (Apple Computer Inc.) attracts the most attention with an attention ratio of 0.818. Articles involving AAPL arrive in social media almost every day (81.8 days over 100 days). However, the symbol “TRV” (Travelers Companies, Inc.) has the lowest attention ratio, 0.204, which means that one finds a related article every five trading days, i.e. one week. Different from the “indirect” attention measures from stock indicators such as trading volumes, extreme returns or price limits, this attention measure is a kind of “direct” measure of investor attention, and shares the same idea as the Search Volume Index (SVI) constructed by Google. Beyond the SVI, our attention can be further projected to “Positive” or “Negative” attention. In our main research question (ii), we are interested in whether the well known asymmetric response (bad vs. good news) is appropriately reflected in the lexical projections. Assuming that investors are more risk-averse, they should be more aware of negative articles and pay more attention to them.

Attention is one of the basic elements in traditional asset pricing models. The conventional asset pricing models assume that information is instantaneously incorporated into

asset prices when it arrives. The basic assumption behind this argument is that investors pay “sufficient” attention to the asset. Under this condition, the market price of asset should be very efficient in incorporating any relevant news. In this aspect, the high attention firms should be more responsive to the text sentiment distilled from the articles, and their market prices should reflect this efficiency. As such, the high attention samples stand on the side of the traditional asset pricing models, and the findings from them are expected to support the efficient market hypothesis. However, attention in reality is a scarce cognitive resource, and investors have limited attention instead (Kahneman, 1973). Further research on this topic from Merton (1987), Sims (2003) and Peng and Xiong (2006) confirms that the limited attention can affect asset pricing. The low attention firms with very limited attention may ineffectively or insufficiently reflect the text sentiment information, so that their corresponding stock reactions could be greatly bounded. This argument is in accordance with the fact that the limited attention causes stock prices to deviate from the fundamental values (Hong and Stein, 1999), implying a potential arbitrage opportunity.

### 3.2.1 Descriptive Statistics for the Firms with different Attention Ratios

Grouping the samples by their attention ratios and examining the responses from different attention groups may offer a clue to the aforementioned conjectures. The criterion used to group the sample firms is based on the quantiles of the attention ratio. Firms whose attention ratios are above the 75% quantile (0.3693) are grouped as “extremely high”, between 50% (0.3026) and 75% quantiles as “high”, between 25% (0.2455) and 50% quantiles as “median”, and lower than 25% quantile as “low”. For each attention group, Table 10 reports across lexical projections the mean values of positive ( $\mu_{Pos}$ ) and negative ( $\mu_{Neg}$ ) sentiment proportions, calculated by averaging  $Pos_{i,t}$  or  $Neg_{i,t}$  over all relevant symbol-day combinations, the proportion of relevant symbol-day combinations with  $Neg_{i,t} > Pos_{i,t}$ , the average attention ratio, and the average number of days with articles, calculated by averaging the number of days with articles over all relevant symbols. The “extreme high” groups receive an average attention ratio of 55.14%, indicating on average these firms have been looked at every two days. By contrast, the low attention group with an average attention ratio of 21.97% receives attention at weekly frequency (5 trading days). By comparing the

magnitude of  $\mu_{Neg}$ , one observes that investors are inclined to express negative sentiments in the “extreme high” group. One may conclude therefore that higher attention is coming with a “negative text”, or inversely speaking: the negative article creates higher attention. This is evident for example in the case of the LM method, where the proportion of symbol-day combinations with dominance of negative sentiment is 46% in the “extremely high” group. For the constituents in this particular attention group, we find on average 691 days with articles observed over a total of 1255 sample days (5 years), which is almost three times the average number of days with articles for the low attention group.

Attention	BL			LM			MPQA			Attention Ratio	Number of Days with Articles
	$\mu_{Pos}$	$\mu_{Neg}$	$Neg > Pos$	$\mu_{Pos}$	$\mu_{Neg}$	$Neg > Pos$	$\mu_{Pos}$	$\mu_{Neg}$	$Neg > Pos$		
Extremely high	0.032	0.016	0.119	0.013	0.014	0.460	0.038	0.013	0.027	0.551	691
High	0.032	0.015	0.113	0.013	0.012	0.403	0.038	0.013	0.031	0.343	430
Median	0.035	0.014	0.083	0.014	0.011	0.339	0.039	0.012	0.027	0.273	356
Low	0.036	0.014	0.086	0.015	0.011	0.333	0.040	0.012	0.031	0.220	264

Table 10: The Summary Statistics for different Attention Ratio Groups

### 3.2.2 The Results of Attention Analysis

The central interest of this research focuses on understanding to which extent distilled news flow and its derived parameters (like attention) impacts the relation between text sentiment and stock reactions. We employ panel regression designed for the given attention groups, and therefore each panel regression equally comprises of 25 sample firms. The results are displayed in Table 11. For the “extremely high” group, the text sentiment carries a major and highly significant influence on future volatility consistently across the three lexical projections. As a caveat though please note that the sentiment effect on volatility shown in Panel A is exclusive for negative news contingent on arriving articles, the stock volatility rarely reacts to positive or optimistic news. Panel B summarizes the attention analysis on the detrended log trading volume. For the “extremely high” group, in the LM and MPQA projection methods, arrival of articles ( $I_{i,t}$ ) brings relevant information, and creates a growing trading volume, especially when it comes with negative news. The corresponding analysis for stock returns are also reasonable. The stock returns of “high” group react clearly to the sentiments, contingent on arriving articles, they rise for optimistic news

and decline for pessimistic consensus. In the case of LM method, the significant positive coefficient of  $Neg_{i,t}$  for the “extremely high” group suggests that the market participants act according to the uncertain market hypothesis developed by Brown et al. (1988) and based on the overreaction hypothesis by Bondt and Thaler (1985). Here, the market participants set new prices before the full range of the news content is resolved. In case of unfavorable news, the investors set stock prices significantly below their conditional expected values and thus, react risk-averse. On the subsequent day, the mispriced stock price will revert to its true value.

The collected empirical evidence so far suggests that the distilled news of high attention firms effectively drive their stock volatilities, trading volumes and returns. They are highly responsive to the sentiment across lexical projections.

Attention	BL			LM			MPQA		
	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$
Panel A: Future Volatility $\log \sigma_{i,t+1}$									
Low	0.020 (0.025)	-0.736 (0.666)	-0.074 (0.766)	0.010 (0.016)	-1.027 (1.027)	-0.195 (0.788)	0.016 (0.029)	-0.655 (0.633)	0.275 (0.866)
Median	0.004 (0.016)	-0.690 (0.449)	1.107** (0.446)	-0.012 (0.016)	-0.308 (0.778)	1.126* (0.630)	0.008 (0.019)	-0.872* (0.515)	1.767** (0.707)
High	-0.016 (0.017)	-0.460 (0.442)	1.324*** (0.475)	-0.046*** (0.013)	0.967 (0.724)	1.806*** (0.615)	-0.019 (0.016)	-0.636** (0.315)	2.548*** (0.662)
Extremely High	-0.010 (0.014)	0.027 (0.257)	0.784** (0.371)	-0.013 (0.013)	0.483 (0.457)	0.747** (0.300)	-0.002 (0.017)	-0.182 (0.284)	0.909** (0.433)
Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$									
Low	0.054** (0.024)	-0.817 (0.502)	0.312 (0.665)	0.044*** (0.014)	-0.923 (0.657)	-0.109 (0.556)	0.049* (0.029)	-0.433 (0.567)	-0.197 (0.796)
Median	0.052*** (0.014)	-0.851** (0.398)	1.116* (0.600)	0.032*** (0.010)	-0.199 (0.535)	0.861 (0.601)	0.062*** (0.013)	-0.754** (0.342)	0.449 (0.689)
High	0.036*** (0.009)	-0.198 (0.299)	0.554 (0.459)	0.021* (0.011)	0.815* (0.487)	0.447 (0.451)	0.046*** (0.016)	-0.358 (0.385)	0.419 (0.559)
Extremely High	0.023 (0.014)	-0.242 (0.336)	0.958** (0.416)	0.017* (0.008)	0.299 (0.521)	0.796* (0.429)	0.032** (0.014)	-0.408 (0.299)	1.084** (0.427)
Panel C: Future Returns $R_{i,t+1}$									
Low	0.000 (0.001)	0.012 (0.022)	0.009 (0.023)	0.000 (0.000)	0.021 (0.030)	-0.001 (0.023)	0.000 (0.001)	0.010 (0.021)	-0.016 (0.032)
Median	-0.001 (0.001)	0.024* (0.012)	0.009 (0.018)	0.000 (0.000)	0.035* (0.019)	-0.022 (0.024)	-0.001 (0.001)	0.034* (0.018)	0.007 (0.024)
High	0.000 (0.000)	0.028** (0.012)	-0.034*** (0.011)	0.001** (0.000)	0.038* (0.022)	-0.046** (0.018)	0.000 (0.001)	0.024** (0.011)	-0.044*** (0.016)
Extremely High	0.000 (0.000)	0.017 (0.012)	0.004 (0.012)	-0.000 (0.000)	0.031 (0.021)	0.033** (0.013)	0.001* (0.000)	-0.006 (0.011)	0.009 (0.016)

\*\*\* refers to a  $p$  value less than 0.01, \*\* refers to a  $p$  value more than or equal to 0.01 and smaller than 0.05, and \* refers to a  $p$  value more than or equal to 0.05 and less than 0.1. Values in parentheses are clustered standard errors.

Table 11: Attention Analysis: The Impact on future Volatility, Trading Volume and Returns

Given the high attention received, any relevant information including the articles made by individual traders has been fully incorporated into their asset prices and dynamics. Due to their efficiency, the article posting and discussing today can predict stock reactions tomorrow. For lower attention firms, one cannot make such a strong claim. Investors may think those firms are negligible and may therefore underreact to the available information. The underreaction from limited attention is likely to cause stock prices to deviate from the fundamental values, and an arbitrage opportunity may emerge. Our evidence is in line with Da et al. (2011) in which they support the attention-induced price pressure hypothesis.

By using the SVI from Google as attention measure, they find stronger attention-induced price pressure among stocks in which individual investor attention matters most. Beyond their study, we find that high attention is usually accompanied with negative articles, and negative articles contribute more to attention and cause more stock reactions, supporting an asymmetric response.

It is interesting to note that the coefficients for the control variables do not vary much across lexical projections in each attention group (results not shown here), which indicates that for each attention group, the sentiment measures are not so much correlated with the control variables and provide incremental information.

### 3.2.3 Monte Carlo Simulation based on Attention Analysis

Like Section 3.1.4, we present a realistic Monte Carlo scenario for different attention groups using the results from Table 11. We keep the parameter settings of the data generation and the calculation of confidence bands as before. Figure 5 summarizes the associations between the negative proportions and the simulated future volatilities across different attention groups. The scatter plots of the high attention panel are quite dense, whereas those of the low attention group are sparser due to its lower frequency of articles. Interestingly, the higher volatilities of high attention firms are prominently driven by negative text sentiment, but have an inverse relationship with positive sentiment. Through comparison of the confidence bands we can conclude for all three lexica that the effect of negative sentiment significantly differs from that of positive sentiment. The regions where the bands do not overlap are quite large for BL (0.022 - 0.056) and MPQA (0.020 - 0.053) but much smaller for LM (0.019 - 0.024). The associations in the low attention panel are somewhat ambiguous. Indeed, we can note that the confidence bands for positive and negative sentiment overlap over the whole range of sentiment value and across all three lexica. These simulations support the estimations in Table 11 with a strong link found in the “extremely high” and “high” attention groups and a preeminent asymmetric response. The firms that have been paid high attentions are more sensitive to the text sentiment than negligible firms. The sentiment effect together with the observable asymmetry are highly influential on stock returns, volatilities and trading volumes. In this sense, their stock reactions



are more responsive to the opinions in social media. In other words, they are also more vulnerable to signals from small investors.

### 3.3 Sector Analysis

The stock reactions that we analyze in relation to text sentiment can be further segmented into sector specific responses. Given a growing body of literature that has suggested that industry plays a role in stock reactions (see Fama and French (1997), Chen et al. (2007), Hong et al. (2007)), we investigate whether this relation is industry-specific in nature. A detailed analysis of sector specific reactions would go far beyond the scope of this paper and is in fact the subject of research by Chen et al. (2015). We therefore only highlight a few insights from lexical sentiment for the business sectors. We ignore the “Telecommunication Services” sector since it only contains two stock symbols. Descriptive statistics for the other 8 sectors are displayed in Table 12 across the three lexical projections. It is of interest to study the variation of the proportion of negative over positive sentiments across the 8 sectors. One observes that consistently over all lexical projections the financial sector has the highest average discrepancy in negative and positive proportion. By contrast the health care sector has (except for MPQA) the lowest average discrepancy. Investors show their discrepant opinions or disagreement in a very extreme case of  $Neg > Pos = 0.5$ , implying that 50% of investors stand on one side and the rest of 50% stand on the opposite side. Table 12 indicates that the financial sector related texts are more divergent in opinions than others and that apparently the health care sector does not receive such adverse opinion positions as the other sectors do. The investors who invest the stocks in health care sector are more likely to reach their shared consensus or convergent agreement.

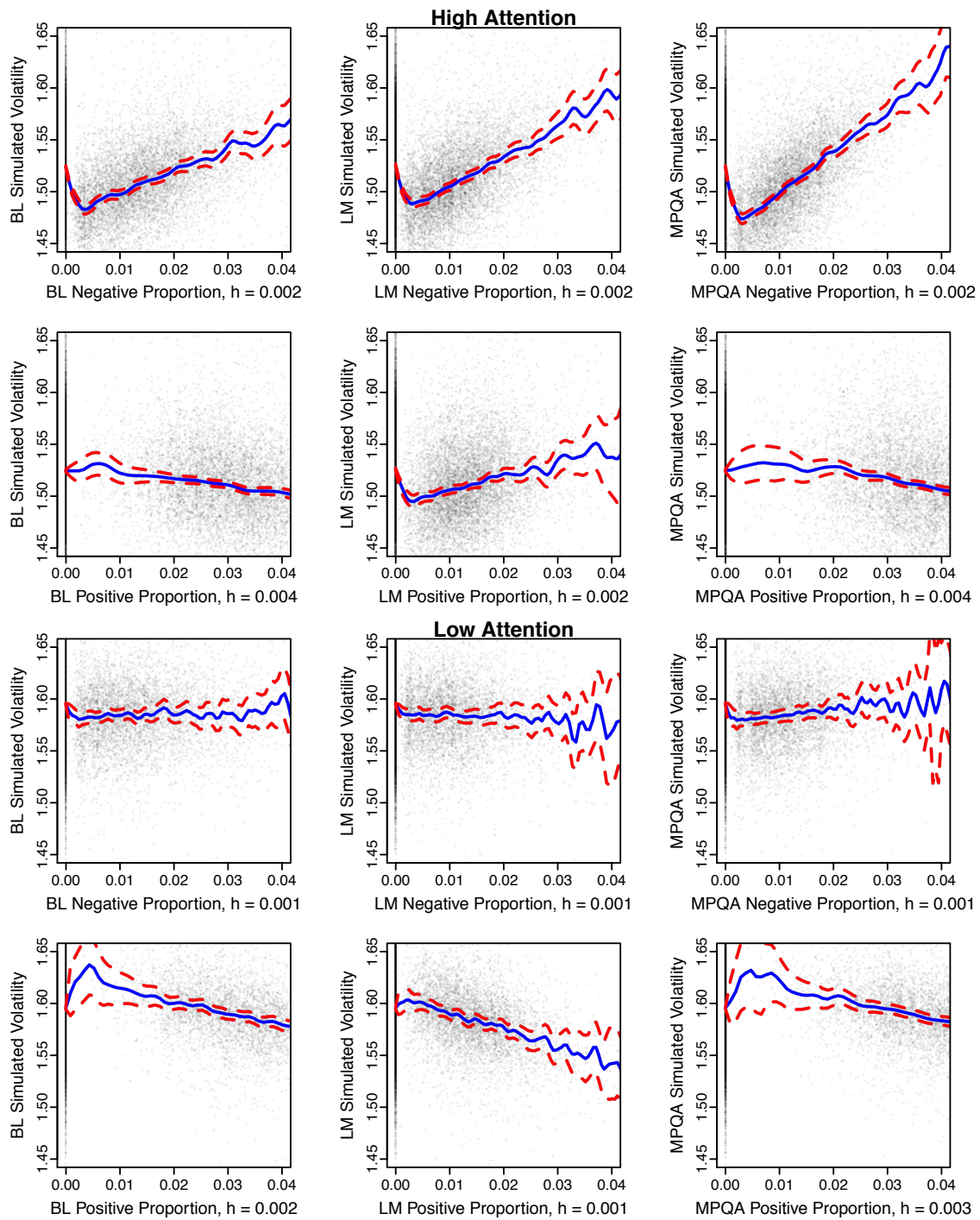


Figure 5: Monte Carlo Simulation based on Attention Analysis Results

Sector	BL			LM			MPQA			Attention
	$\mu_{Pos}$	$\mu_{Neg}$	$Neg > Pos$	$\mu_{Pos}$	$\mu_{Neg}$	$Neg > Pos$	$\mu_{Pos}$	$\mu_{Neg}$	$Neg > Pos$	Ratio
Consumer Discretionary	0.034	0.014	0.088	0.014	0.011	0.346	0.038	0.012	0.030	0.332
Consumer Staples	0.034	0.014	0.099	0.014	0.012	0.365	0.037	0.013	0.025	0.324
Energy	0.028	0.015	0.152	0.011	0.011	0.467	0.038	0.014	0.033	0.370
Financials	0.032	0.019	0.195	0.013	0.018	0.594	0.038	0.015	0.045	0.413
Health Care	0.035	0.014	0.059	0.014	0.011	0.344	0.039	0.014	0.031	0.287
Industrials	0.035	0.012	0.069	0.013	0.011	0.355	0.041	0.011	0.018	0.336
Information Technology	0.033	0.015	0.101	0.014	0.012	0.373	0.038	0.023	0.012	0.364
Materials	0.034	0.014	0.097	0.013	0.013	0.498	0.039	0.031	0.013	0.287

Note: This table reports, for the BL, LM and MPQA methods, the mean values of positive ( $\mu_{Pos}$ ) and ( $\mu_{Neg}$ ) negative sentiment proportions as well as the proportion of relevant symbol-day combinations with dominance of negative sentiment. For each sector, an article is accumulated only if a firm appeared in this article belongs to this sector. The attention ratio for each sector is calculated as the number of days with articles related to this sector divided by the total number of days in the sample period.

Table 12: Summary statistics in each sector

The attention also vary with the sectors. The evidence that financials sector has attracted the highest attention with an attention ratio of 0.413 may be attributed to (1) the investors' widespread involvement in this industry because we all need to keep a relationship with banks to deposit our money, trade for securities or some financial reasons; (2) the outbreak of the US subprime crisis and the European sovereign debt crisis have brought the highest attention to this sector; (3) their sensitivity on changes in the economy, monetary policy and regulatory policy. The health care sector, however, is much less attractive and this could be explained by a stable demand and reduced sensitivity to economic cycles. Given these observations we will now continue our analysis of stock reactions for these two sectors only, and leave a bundle of interesting issues to further research.

To address the important question of whether there is a sector dependent stock reactions, we further analyze how the text sentiment affects, as reported in Table 13, the future volatility, trading volume and return. In order to do so we employ the panel regression (as described in (4)-(5)) and report the results in Table 13. The variable  $I_{i,t}$  was used to indicate arrival of articles on this sector. Contingent on arriving articles, the three sentiment projections in financial sectors yielded significant and positive effects on future log volatility from negative proportions, meaning that increasing the negative text sentiments will result in higher volatility. The exclusive response to negative sentiment in financial

sector indeed is in line with our entire panel evidence. However, the finding in the health care sector is too insignificant to claim it. Potentially, investor inattention for the health care sector may cause a significant mispricing on the stocks. Investors possibly neglect the news of this sector posted on social media, or this sector has a slow information diffusion that could lead to a delayed reaction.

Sector	BL			LM			MPQA		
	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$
Panel A: Future Volatility $\log \sigma_{i,t+1}$									
Financials	-0.023 (0.026)	-0.052 (0.319)	1.075** (0.435)	-0.025 (0.027)	0.275 (0.924)	1.027*** (0.259)	-0.025 (0.029)	-0.143 (0.503)	1.816*** (0.586)
Health Care	0.031 (0.026)	-0.426 (0.522)	-0.509 (0.891)	0.009 (0.023)	0.052 (1.138)	-0.130 (0.921)	0.001 (0.024)	-0.118 (0.595)	0.854 (0.783)
Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$									
Financials	0.037* (0.020)	-0.334 (0.494)	0.015 (0.527)	0.017 (0.015)	1.110 (0.766)	-0.313 (0.476)	0.054*** (0.015)	-0.747** (0.305)	0.049 (0.536)
Health Care	0.031 (0.023)	0.110 (0.436)	-0.314 (0.846)	0.022 (0.018)	0.603 (0.863)	-0.042 (0.837)	0.037 (0.025)	-0.104 (0.443)	-0.211 (0.873)
Panel C: Future Returns $R_{i,t+1}$									
Financials	-0.001 (0.001)	0.034* (0.017)	0.028** (0.014)	-0.000 (0.001)	0.030 (0.033)	0.042** (0.016)	0.001 (0.001)	0.003 (0.020)	0.013 (0.019)
Health Care	0.000 (0.000)	-0.000 (0.008)	0.008 (0.018)	0.000 (0.000)	0.006 (0.019)	0.015 (0.018)	0.000 (0.001)	0.006 (0.012)	-0.011 (0.022)

\*\*\* refers to a  $p$  value less than 0.01, \*\* refers to a  $p$  value more than or equal to 0.01 and smaller than 0.05, and \* refers to a  $p$  value more than or equal to 0.05 and less than 0.1. Values in parentheses are standard errors.

Table 13: Sector analysis: The Impact on future Volatility, Trading Volume and Returns

The trading volume is another stock reaction we may attribute to text sentiments. Using the BL and the MPQA projection method, we find that the arrival of article brings relevant information and therefore stimulates the trading volume. It is interesting to note that contingent on arriving articles, the negative sentiment distilled using the BL and LM methods is significantly positively related to stock returns on the next trading day. To investigate the reason for this, we also run a contemporaneous regression for the financials sector (results not shown) and found a significantly negative impact of the negative sentiment distilled using the BL and MPQA methods on contemporaneous returns  $R_{i,t}$ , and the size of the coefficients is about twice of that in lagged regression in Table 13. This might suggest that the market participants monitor financial companies quite carefully and overreact in case of bad news. On the next day, the participants fully recognize the scope

of the news and reverse part of their prior decisions, and hence the negative sentiment on trading day  $t$  has positive impact on returns on trading day  $t + 1$ . This is also in line with the finding in Kuhnen (2015) which suggests that that being in a negative domain leads people to form overly pessimistic beliefs about stocks. After the 2008 financial crisis and the bankruptcy of some major financial companies, this might be the case for the financials sector.

From these analysis, we know that investors indeed pay different attentions to sectors they are of interest, and their attentions effectively govern the equity's variation. Attention constraints in some sectors may affect investors' trading decisions and the speed of price adjustments.

## 4 Conclusion

In this paper, to analyze the reaction of stocks' future log volatility, future detrended log trading volume and future returns to social media news, we distill sentiment measures from news using two general-purpose lexica (BL and MPQA) and a lexicon specifically designed for financial applications (LM). We demonstrate that these sentiment measures carry incremental information for future stock reactions. Such information varies across lexical projections, across groups of stocks that attract different level of attention, and across different sectors. The positive and negative sentiments also have asymmetric impact on future stock reaction indicators. A detailed summary of the results is given in Table 14 in the Supplementary Material. There is no definite picture for which lexicon is the best. This is an important contribution of our paper to the line of research on textual analysis for financial market. Besides, the advanced statistical tools that we have utilized, including panel regression and confidence bands, are novel contributions to this line of research.

## References

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *The Journal of Finance*, 57(3):1047–1091.

- Andersen, T. G. and Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2-3):115–158.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76.
- Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.
- Arellano, M. (1987). Computing Robust Standard Errors for Within-Groups Estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–34.
- Bekaert, G. and Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies*, 13:1–42.
- Black, F. (1976). Studies of stock price volatility changes. In *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section, American Statistical Association*, pages 177–181.
- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.
- Bondt, W. F. M. D. and Thaler, R. (1985). Does the stock market overreact? *Journal of Finance*, 40(3):793.
- Brown, K. C., Harlow, W., and Tinic, S. M. (1988). Risk aversion, uncertain information, and market efficiency. *Journal of Financial Economics*, 22(2):355–385.
- Cao, H. H., Coval, J. D., and Hirshleifer, D. A. (2001). Sideline investors, trading-generated news, and security returns. *Dice Working Paper No. 2000-2*.
- Chen, G. M., Firth, M., and Rui, O. M. (2001). The dynamic relation between stock returns, trading volume, and volatility. *The Financial Review*, 36(3):153–174.
- Chen, G. M., Firth, M., and Rui, O. M. (2002). The dynamic relationship between stock returns and trading volume: Domestic and cross-country evidence. *Journal of Banking and Finance*, 36(3):51–78.

- Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5):1367–1403.
- Chen, L., Lakonishok, J., and Swaminathan, B. (2007). Industry classifications and return comovement. *Financial Analysts Journal*, 63:56–70.
- Chen, Y. H. C., Bommers, E., Härdle, W. K., and Zhang, J. (2015). News and big news: a text sentiment analysis for GICS specific stock reactions. *SFB 649, discussion paper*.
- Chen, Z., Daigler, R. T., and Parhizgari, A. M. (2006). Persistence of volatility in futures markets. *Journal of Futures Markets*, 26:571–594.
- Christie, A. A. (1982). The stochastic behavior of common stock variance. *Journal of Financial Economics*, 10:407–432.
- Clark, P. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–55.
- Copeland, T. E. (1976). A model of asset trading under the assumption of sequential information arrival. *Journal of Finance*, 41:135–156.
- Copeland, T. E. (1977). A probability model of asset trading. *Journal of Financial and Quantitative Analysis*, 12:563–578.
- Da, Z., Engelberg, J., and Gao, P. (2011). In search of attention. *The Journal of Finance*, 66(5):1461–1499.
- Das, S. and Chen, M. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, pages 1375–1388.
- Di Clemente, A. and Romano, C. (2004). Measuring and Optimizing Portfolio Credit Risk: A Copula-based Approach\*. *Economic Notes*, 33(3):325–357.
- Engle, R. F. (2004). Risk and volatility: economic models and financial practice. *The American Economic Review*, 94:405–420.

- Fama, E. F. and French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics*, 43:153–193.
- Feunou, B. and Tédongap, R. (2012). A stochastic volatility model with conditional skewness. *Journal of Business and Economic Statistics*, 30:576–591.
- Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial journal*, 2(1):1–25.
- Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *The Journal of Business*, 53(1):67–78.
- Girard, E. and Biswas, R. (2007). Trading volume and market volatility: developed versus emerging stock markets. *Financial Review*, 42(3):429–459.
- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). Relationship between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801.
- Groß-Klußmann, A. and Hautsch, N. (2011). When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, 18(2):321–340.
- Hong, H. and Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *The Journal of Finance*, 58(1):313–351.
- Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184.
- Hong, H., Torous, W., and Valkanov, R. (2007). Do industries lead stock markets? *Journal of Financial Economics*, 83:367–396.
- Hotta, L. K., Lucas, E. C., and Palaro, H. P. (2006). Estimation of VaR Using Copula and Extreme Value Theory. *SSRN Electronic Journal*.



- Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.
- Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall, Englewood Cliffs, NJ.
- Kuhnen, C. M. (2015). Asymmetric Learning from Financial Information: Asymmetric Learning from Financial Information. *The Journal of Finance*, 70(5):2029–2062.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool.
- Loader, C. (1999). *Local regression and likelihood*. Statistics and computing. Springer, New York.
- Loader, C. and Sun, J. (1997). Robustness of tube formula based confidence bands. *Journal of Computational and Graphical Statistics*, 6:242.
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.
- Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, 42:483–510.
- Patton, A. J. (2006). Modelling asymmetric exchange rate dependence\*. *International economic review*, 47(2):527–556.
- Peng, L. and Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80:563–602.
- Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of financial studies*, 22(1):435–480.
- Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90:1257–1270.
- Shu, J. and Zhang, J. E. (2006). Testing range estimators of historical volatility. *Journal of Futures Markets*, 26:297–313.

- Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50:665–690.
- Sprenger, T. O., Tumasjan, A., Sandner, P. G., and Welpe, I. M. (2014). Tweets and trades: the information content of stock microblogs: Tweets and trades. *European Financial Management*, 20(5):926–957.
- Sun, J. and Loader, C. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, pages 1328–1345.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.
- Truyens, M. and Eecke, P. V. (2014). Legal aspects of text mining. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., and Zhao, B. Y. (2014). Crowds on wall street: Extracting value from social investing platforms. *Working Paper*.
- Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of HLT-EMNLP-2005*.
- Zhang, X., Fuehres, H., and Gloor, P. A. (2012). Predicting asset value through twitter buzz. In *Advances in Collective Intelligence 2011*, pages 23–34. Springer.
- Zhang, Y. and Swanson, P. E. (2010). Are day traders bias free?—evidence from internet stock message boards. *Journal of Economics and Finance*, 34(1):96–112.

## 5 Supplementary Material

Table 14 summarizes all the results from entire panel sample analysis, attention analysis and sector analysis. Take the “BL” row in Panel A as an example. Arrival of articles ( $I_{i,t}$ ) and the positive sentiment distilled using the BL method ( $Pos_{i,t}$ ) has no significant impact on future volatility  $\log \sigma_{i,t+1}$  in entire sample analysis, attention analysis or sector analysis; the negative sentiment distilled using the BL method ( $Neg_{i,t}$ ) is significantly positively related to future volatility in entire sample analysis and for the “Extremely High” group in attention analysis, and is significantly negatively related to future volatility for the “Health Care” sector in sector analysis.

Lexicon	Type of Analysis	$I_{i,t}$	$Pos_{i,t}$	$Neg_{i,t}$
Panel A: Future Volatility $\sigma_{i,t+1}$				
BL	Entire Sample	/	Negative	Positive
	Attention Analysis	/	/	Positive for “Median”, “High” and “Extremely High”
	Sector Analysis	/	/	Positive for “Financials”
LM	Entire Sample	Negative	/	Positive
	Attention Analysis	Negative for “High”	/	Positive for “Median”, “High” and “Extremely High”
	Sector Analysis	/	/	Positive for “Financials”
MPQA	Entire Sample	/	Negative	Positive
	Attention Analysis	/	Negative for “Median” and “High”	Positive for “Median”, “High” and “Extremely High”
	Sector Analysis	/	/	Positive for “Financials”
Panel B: Future Detrended Log Trading Volume $V_{i,t+1}$				
BL	Entire Sample	Positive	Negative	Positive
	Attention Analysis	Positive for “low”, “Median” and “High”	Negative for “Median”	Positive for “Median” and “Extremely High”
	Sector Analysis	Positive for “Financials”	/	/
LM	Entire Sample	Positive	/	Positive
	Attention Analysis	Positive for all groups	Positive for “High”	Positive for “Extremely High”
	Sector Analysis	/	/	/
MPQA	Entire Sample	Positive	Negative	Positive
	Attention Analysis	Positive for all groups	Negative for “Median”	Positive for “Extremely High”
	Sector Analysis	Positive for “Financials”	Negative for “Financials”	/
Panel C: Future Returns $R_{i,t+1}$				
BL	Entire Sample	/	Positive	/
	Attention Analysis	/	Positive for “Median” and “High”	Negative for “High”
	Sector Analysis	/	Positive for “Financials”	Positive for “Financials”
LM	Entire Sample	/	Positive	/
	Attention Analysis	Positive for “High”	Positive for “Median” and “High”	Negative for “High”, positive for “Extremely High”
	Sector Analysis	/	/	Positive for “Financials”
MPQA	Entire Sample	/	Positive	/
	Attention Analysis	Positive for “Extremely High”	Positive for “Median” and “High”	Negative for “High”
	Sector Analysis	/	/	/

The signs of the significant coefficients are given, with a significance level of 0.1.

Table 14: Summary of the Results

# Common factors in credit defaults swap markets

Cathy Yi-Hsuan Chen<sup>1</sup> · Wolfgang Karl Härdle<sup>2,3</sup>

Received: 22 April 2014 / Accepted: 27 March 2015 / Published online: 20 April 2015  
© Springer-Verlag Berlin Heidelberg 2015

**Abstract** We examine what are the common factors that determine systematic credit risk, and estimate and interpret these factors. We also compare the contributions of common factors in explaining the changes of credit default swap spreads during the pre-crisis, the crisis and the post-crisis period; there is evidence to suggest that the eigenstructures across these three sub-periods are distinct. Furthermore, we examine whether the observable economic variables are in fact the underlying latent factors and analyze the predictability in the factors that capture the time-variation of credit default swap spreads.

**Keywords** Credit default swaps · Common factors · Credit risk · Factor model

## 1 Introduction

The fact that investors holding fixed income portfolios to diversify risk or enhance investment returns have suffered from systematic credit risk of different entities has

---

The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through SFB 649 “Economic Risk” and IRTG 1792 “High Dimensional Non Stationary Time Series”.

---

✉ Cathy Yi-Hsuan Chen  
cathy1107@gmail.com

<sup>1</sup> Department of Finance, Chung Hua University, No. 707, Sec. 2, WuFu Rd., Hsinchu 300, Taiwan

<sup>2</sup> Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

<sup>3</sup> Sim Kee Boon Institute for Financial Economics, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

been observed more recently. This raises the question whether there are common factors determining systematic credit risk across different entities, different credit ratings, different countries and different maturities. In fact, an increase in systematic credit risk will harm the benefit from a well-diversified bond portfolio. An examination into common credit risk factors explores the nature of correlated defaults. Several illustrations for correlated defaults were proposed by [Das et al. \(2007\)](#). Firstly, firms may be exposed to common or correlated risk factors. Secondly, the event of default by one firm may be contagious. Thirdly, learning from default may generate default correlation. This study examines what are the common factors that determine systematic credit risk, and estimates and interprets the common risk factors. In further steps, we estimate the market prices of risk factors and test their significance. Based on factor models, we propose various time series properties for common factors and idiosyncratic components, and examine which one can produce the best forecasting to the dynamics of credit default swap (CDS) spreads.

Understanding how corporate defaults are correlated is particularly important for the risk management of corporate debt portfolio, since banks have to retain greater capital to survive default losses if defaults are heavily clustered in time. An investigation of the sources and degree of default clustering is also crucial for the rating and risk analysis of structured credit products, such as collateralized debt obligations (CDOs) and options on portfolios of default swaps that are exposed to correlated default. Several attempts have been made in the literature to address this issue. The first one incorporates correlated default into the reduce-form credit risk modeling ([Das et al. 2006, 2007](#)). The second research stream assumes that default probabilities depend on firm-specific and market-wide factors. Typically, portfolio loss distributions are based on the correlating influence from such observable market-wide factors. A number of potentially observable factors from macroeconomic fundamentals have been proposed to analyze correlated defaults ([Collin-Dufresne et al. 2001](#); [Benkert 2004](#); [Ericsson et al. 2009](#)). The third research stream, however, extracts some latent/unobservable factors mainly from the principal components analysis (PCA) method to avoid a possible downward bias from estimating tail loss ([Duffie et al. 2009](#); [Cesare and Guazzarotti 2010](#); [Anderson 2008](#)). As we know, not all relevant risk factors are potentially observable by econometricians ([Duffie et al. 2009](#)).

Recent research claims that common latent factors increasingly and apparently explain the time-variation of credit risk, especially during the financial crisis. [Anderson \(2008\)](#) finds that a very high fraction of weekly variations in the implied default intensity is explained by a single common factor. [Cesare and Guazzarotti \(2010\)](#) found that CDS spread changes were increasingly driven by a common factor during the US subprime crisis. This paper goes beyond these two studies by additionally interpreting the common latent factors and modelling their time-variation patterns. We demonstrate this by using a very extensive CDS data set, encompassing different maturities, different credit ratings, different entities and different countries, and produce robust common factors with a convincing interpretation.

We compare the contributions of common factors in explaining the CDS spreads changes during the pre-crisis, the crisis and the post-crisis period. We find that the fraction of CDS variation explained by the first principal component increases from 58.7 to 72.3 % during the crisis period, and then declines to 47 % after the crisis. The

results suggest that during the crisis, the changes of CDS spreads are increasingly driven by common factors and less by idiosyncratic components. Furthermore, the eigenstructures across three sub-periods are distinct based on the result of a likelihood ratio test that compares the common principal components model against the unrestricted model indicates. To interpret the estimated factors, we investigate the association between the latent factors and the observed economic variables.

Having applied the factor model to CDS spreads, we model the time-variation of common factors to examine the predictability of CDS spreads. This prediction will certainly benefit investors to hedge, speculate and arbitrage in the credit markets. We propose various factor models and compare their out-of-sample forecasting performance. Testing their equal predictive ability is also required to show whether relatively outperformance is statistically significant.

The remainder of this research is organized as follows. The next section describes the data we have used. Section 3 presents the factor models used in this study, and provides an economic interpretation for the estimated factors. In Sect. 4, we propose several factor specifications to predict the times-variation of CDS spreads; evaluating their out-of-sample forecasting performances and testing their equal predicting ability are both conducted in this section.

## 2 Data description

Credit default swap data are collectable from Markit, an aggregator of CDS pricing data from the leading broker-dealers. In terms of our focus on the commonality of CDS spreads, we are interested in the CDS indices rather than single name reference entity CDS contracts to mitigate the idiosyncratic components and liquidity risk. Our concern coincides with [Driessen et al. \(2003\)](#) in studying the common factors in international bond returns. They suggest that bond portfolio data is the preferred method to clear idiosyncratic risk embedded in individual bonds. Markit provides a detailed CDS index series, for example, the Markit CDX indices comprise the most liquid baskets of names covering North American Investment Grade and High Yield single name credit default swaps with various maturities, while the Markit iTraxx indices comprise of the most liquid names in the rest of regions such as Europe, Asia, Australia and Japan. Each index rolls biannually in March and September. Credit events that trigger settlement for individual components are bankruptcy and failure to pay, and are subsequently settled via credit event auctions. For traders, trading CDS indices is more attractive since they are allowed to trade large sizes and confirm all trades electronically. Stronger support from dealers and industry participants has prominently enhanced liquidity in all market conditions. The transparency of CDS markets has gradually improved since the default of Lehman ([Avellaneda and Cont 2010](#)). Central clearing and increased reporting of CDS trades to data repositories are important steps towards increased transparency, which regulators intend to use for monitoring and enhancing market stability. As such, they are quite acceptable as a representative benchmark of the overall market credit risk.

The indices quoted on a spread basis are selected by its regions: North American (CDX), Europe (iTraxx EU), by maturities: 5 and 10-year, by credit ratings:

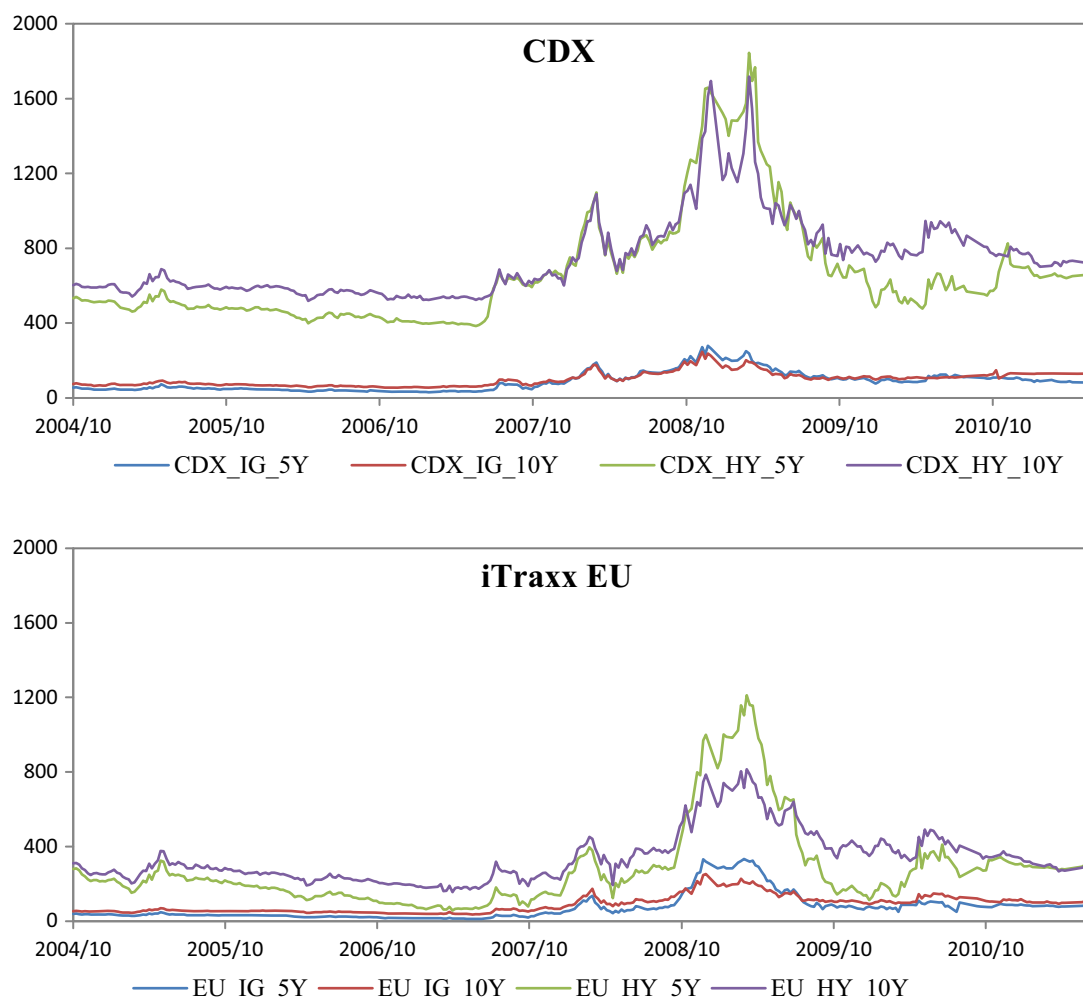
**Table 1** Summary statistics for entire sample period, pre-, during and post-crisis period

	Entire		Pre-crisis		Crisis		Post-crisis	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CDX.IG.5Y	0.47	18.68	-0.21	2.51	1.71	16.65	-0.01	32.43
CDX.IG.10Y	0.17	7.02	-0.16	2.64	0.83	11.58	-0.15	2.98
CDX.HY.5Y	0.46	53.34	-1.06	10.62	3.54	83.20	-1.05	46.44
CDX.HY.10Y	0.56	60.96	-0.49	11.90	1.25	98.24	-0.60	44.91
EU.IG.5Y	0.17	10.21	-0.19	1.63	0.42	15.14	0.48	10.74
EU.IG.10Y	0.35	8.62	-0.11	2.06	1.02	13.22	0.24	7.86
EU.HY.5Y	0.86	38.60	-1.65	11.86	4.43	58.11	0.43	36.13
EU.HY.10Y	1.06	29.35	-1.08	13.15	4.93	43.30	-0.44	26.18

The entire sample period covers from Oct 2004 to June 2011. The indices are selected by its regions: North American (CDX), Europe (iTraxx EU), by maturities: 5- and 10-year, by credit rating: investment-grade (IG) and high-yield grade (HY). We have 134 weekly observations in the pre-crisis period (from Oct 2004 to May 2007), 104 observations in the crisis period (from June 2007 to July 2009) and 76 observations in the post-crisis period (from Aug 2009 to June 2011). The changes of CDS indices are quoted as basis points and their mean and standard deviation are reported

investment-grade (IG) and high-yield grade (HY). From October 2004 to June 2011, these eight indices with different regions, maturities and credit ratings will be analyzed in the subsequent sections. The US subprime crisis period is emphasized since the function of money markets in the U.S. was severely impaired in the summer of 2007, and then even further following the collapse of Bear Sterns in mid-March 2008 and the bankruptcy of Lehman Brother in September 2008. The turmoil from June 2007 to July 2009 is referred to a crisis period. After mapping the trading date among eight CDS indices, each index has 315 weekly observations: 134 in the pre-crisis period (from October 2004 to May 2007), 104 in the crisis period (from June 2007 to July 2009) and 76 in the post-crisis period (from August 2009 to June 2011). Table 1 summarizes the descriptive statistics for the entire sample period, the pre-crisis, the crisis and the post-crisis period. During the crisis period, the average changes of CDS spreads are all apparently positive, and are extremely volatile.

The time-variations of CDS indices as displayed in Fig. 1 exhibit a changing dynamic. One noticeable feature is a high level of comovement across various maturities and credit ratings, which motivates the study of common factors. Specially, in Fig. 1 the apparent spike during the outbreak of the U.S. subprime crisis shows an inversion of the risk structure. For a given maturity, a high-yield (HY) index should be higher than an investment-grade (IG) one to compensate for a higher default risk taken by investors. The default risk premium between a HY and an IG may expand during the financial crisis to reflect a shift in investor risk appetite. Due to this changing risk attitude in a distressed time, risk-averse investors require a higher default risk premium. Pan and Singleton (2008) claimed that a comovement effect in the CDS markets is partly caused by a shift in investor risk appetite, especially for the turbulent period.



**Fig. 1** Time series plots of CDX index and iTraxx EU index

Figure 1 also shows the term structure of CDS markets. Normally, the slope of CDS term structure is upward in which the longer-term CDS spreads are higher than the respective shorter-term ones due to a greater risk-taking in longer maturity contracts. In this regard, the term structure should never be inverted. But, the term structure did occasionally invert, especially during the financial crisis (Pan and Singleton 2008). For an upcoming crisis, the demand for short-term CDS contracts is appealing. To cover a higher hedging cost faced by protection sellers, the bid-ask spreads of short-term contracts should be comparable to those of longer-dated contracts. As shown in Fig. 1, we have consistent evidence in the CDS term structure of an inverted slope in the crisis period and an upward slope in the rest of periods.

### 3 Factor representation of CDS spreads change

#### 3.1 Model specifications

Let  $S_{it}$  be the observed change of CDS spreads for the  $i$ th cross-section unit at time  $t$ , for  $i = 1, \dots, N$ , and  $t = 1, \dots, T$ . The factor model for given  $i$ th unit is:



$$S_{it} = F_t \lambda_i + e_{it} \quad (1)$$

where  $F_t$  is a vector of common factors and is not observable,  $\lambda_i$  is a vector of factor loadings associated with  $F_t$ , and  $e_{it}$  is the idiosyncratic component of  $S_{it}$ . It is assumed that factors and idiosyncratic disturbances are mutually uncorrelated,  $E(F_t, e_{it}) = 0$ . Obviously, Eq. (1) is the static factor representation of the change of the CDS spreads. For the forecasting exercise in subsequent sections, we will invoke the assumptions about the cross-sectional and temporal dependence in the idiosyncratic components.

The asymptotic principal components technique established by [Stock and Watson \(2002\)](#) and [Bai and Ng \(2002\)](#) can be used to consistently estimate the common factors. One starts with an arbitrary number of factors  $k$  ( $k < \min\{N, T\}$ ) and estimates  $\lambda^k$  and  $F^k$  by solving:

$$(\lambda^k, F^k) = \arg \min_{\Lambda^k, F^k} (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T (S_{it} - F_t^k \lambda_i^k)^2 \quad (2)$$

subject to the normalization of either  $\Lambda^{kT} \Lambda^k / N = I_k$  with  $\Lambda^k = [\lambda_1^k \dots \lambda_N^k]^T$  or  $F^{kT} F^k / T = I_k$ . One solution of this optimization is given by  $(\hat{\Lambda}^k, \hat{F}^k)$ , where  $\hat{\Lambda}^k$  is  $\sqrt{N}$  times the eigenvectors corresponding to the  $k$  largest eigenvalues of the  $N \times N$  matrix  $S^T S$  where  $S$  is a  $T$  by  $N$  dimension matrix comprising  $N$  units until time  $T$ , and  $\hat{F}^k = S \hat{\Lambda}^k / N$ .

### 3.2 Common principal components in the different sub-periods

In [Table 2](#) we present the results for the factor model using the CDS index data, and find that a four-factor model in general explains up to 90.5% of the variance in the changes of CDS spreads. The first factor explains 63% of the variance of the change of CDS spreads, the explained variance of the second, third and fourth factors are 12.1, 8, and 7.4%. When turning to three sub-periods, the first factor explains 58.7% of the variance in the pre-crisis period, 72.3% of the variance in the crisis period and 47% of the variance in the post-crisis period. The fraction of CDS variation explained by the

**Table 2** Explained variance by principal component analysis

	% variance explained				Total variance explained (%)
	Factor 1 (%)	Factor 2 (%)	Factor 3 (%)	Factor 4 (%)	
Entire	63.0	12.0	8.0	7.5	90.5
Pre-crisis	58.7	13.3	9.0	7.6	88.6
Crisis	72.3	12.4	5.4	4.0	94.1
Post-crisis	47.0	16.5	12.6	10.2	86.5

For entire sample period and three sub-periods, this table presents the proportion of the total variance of the changes of CDS spreads explained by the variation of a given factor

first principal component increases from 58.7 % before the crisis to 72.3 % during the crisis period, but declines to 47 % after the crisis. The CDS spreads during the crisis are increasingly driven by common factors and less by idiosyncratic components, which is evident by an increased explanatory power up to 94.1 %.

To formally test whether the eigenstructures across three sub-periods are distinct, we perform a likelihood ratio test comparing a restricted (the Common Principal Components (CPC) model) against the unrestricted model (the model where all covariances are treated separately). The likelihood ratio statistic is given by

$$T_{(n_1, n_2, \dots, n_h)} = -2 \log \frac{L(\hat{\Sigma}_1, \dots, \hat{\Sigma}_h)}{L(S_1, \dots, S_h)} \tag{3}$$

where  $\Sigma_i = \Gamma \Lambda_i \Gamma^T, i = 1, \dots, h$ , is a positive definite  $N \times N$  covariance matrix for every  $i$ ,  $\Gamma = (\gamma_1, \dots, \gamma_N)$  is an orthogonal  $N \times N$  transformation matrix and  $\Lambda_i = \text{diag}(\vartheta_{i1}, \dots, \vartheta_{iN})$  is the matrix of eigenvalues where all  $\vartheta_i$  are assumed to be distinct. The CPC is motivated by the similarity of the covariance matrices in the  $h$ -sample problem. The basic assumption of CPC is that the space spanned by the eigenvectors is identical across several groups, whereas variances associated with the components are allowed to vary (Flury 1988).

Let  $S$  be the sample covariance matrix of an underlying  $N$ -variate normal distribution with sample size  $n$ . Then the distribution of  $nS$  has  $n - 1$  degree of freedom and is known as the Wishart distribution.

$$nS \sim W_N(\Sigma, n - 1)$$

Hence, for Wishart covariance matrices  $S_i, i = 1, \dots, h$  with sample size  $n_i$ , the likelihood function can be expressed as

$$L(\Sigma_1, \dots, \Sigma_h) = C \prod_{i=1}^h \exp \left[ \text{tr} \left\{ -\frac{1}{2} (n_i - 1) \Sigma_i^{-1} S_i \right\} \right] |\Sigma_i|^{-\frac{1}{2}(n_i-1)} \tag{4}$$

where  $C$  is a constant independent of the parameters  $\Sigma_i$ . See Härdle and Simar (2011), inserting (4) to (3), the likelihood ratio statistic is obtained and has a  $\chi^2$  distribution as  $\min(n_i)$  tends to infinity with

$$h \left\{ \frac{1}{2} N(N - 1) + 1 \right\} - \left\{ \frac{1}{2} N(N - 1) + hN \right\} = \frac{1}{2} (h - 1) N(N - 1)$$

degree of freedom. Using  $h = 3$  sub-periods sample covariance matrix data, the calculation yields 897.54 for the likelihood ratio statistic, which corresponds to a zero  $p$ -value for the  $\chi^2$  (56) distribution. Hence, the CPC model is rejected against the unrestricted model, where the PCA model is applied to each sub-period separately. The finding indicates that the eigenstructures across three sub-periods, pre-, during and post-crisis, are dramatically distinct. There is no common eigenstructures (e.g. of CPC type) for these periods. Indeed, the outbreak of subprime credit crisis has led to a structure change in the commonality of CDS markets.

### 3.3 Interpreting the factor loadings

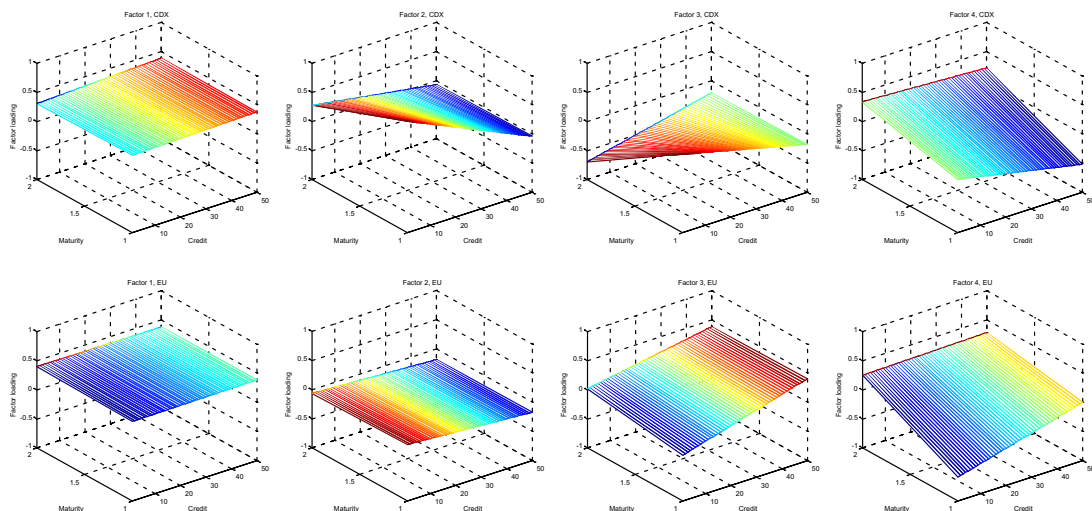
To get a better feel from the estimated factor loadings in Table 3, we plot the estimated factor loadings against credit rating and maturity in Fig. 2. The characteristics of factors seem intuitive and interpretable. For factor 1, the factor loadings all have the same sign and same magnitude across maturities and ratings. It can therefore be interpreted as a *level effect*. The CDS spreads, resembled in bond spreads, are sensitive to the level and movement of the interest rate. As pointed out by Longstaff and Schwartz (1995), the static effect of a higher spot rate increases the risk-neutral drift of the firm value process, which reduces the probability of default and in turn, reduces the CDS spreads. Further empirical evidence is supported by Duffie (1998) and the above references.

Factor 2 can be interpreted as a *region effect*. The factor loadings of CDX series are higher than those of iTraxx Europe family. Since the PCA technology joins the U.S. and European CDS indices, at least one factor should capture the fundamental

**Table 3** Estimated factor loadings

	Entire				Crisis			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
CDX.IG5Y	0.337	0.921	0.353	-0.079	0.267	0.666	0.481	0.148
CDX.IG10Y	0.308	0.278	-0.697	0.431	0.305	0.518	-0.391	-0.581
CDX.HY5Y	0.379	-0.039	-0.178	-0.522	0.384	-0.127	-0.389	0.153
CDX.HY10Y	0.389	-0.066	0.002	0.221	0.376	-0.136	-0.454	0.118
EU.IG5Y	0.372	-0.025	-0.208	-0.585	0.377	0.032	-0.004	0.590
EU.IG10Y	0.401	-0.063	0.017	0.251	0.382	0.014	0.207	0.086
EU.HY5Y	0.385	-0.175	0.406	-0.003	0.362	-0.360	0.339	-0.148
EU.HY10Y	0.380	-0.184	0.387	0.285	0.351	-0.347	0.315	-0.475

This table reports the estimated factor loadings for the entire sample and for the crisis period



**Fig. 2** The association between factor loadings, credit ratings and maturities

or economic differences between the regions. It is not so straightforward to interpret factor 3 in the CDX case, but factor 3 in the iTraxx Europe case may be related to a *volatility effect*. In Table 3 and Fig. 2, we find that for iTraxx Europe, the factor loadings of HY are higher than those of IG. The evidence that the HY spreads are more sensitive to volatility than IG ones is well documented in the literature. The contingent-claims approach implies that the debt claim has features similar to a short position in a put option. Since option values increase with volatility, increased volatility increases the probability of default. Finally, we interpret factor 4 as a *term structure effect*. This is certainly clear because in Table 3 and Fig. 2, the sign of loading of 5-year CDS spreads is always negative while that of 10-year CDS spreads is positive. This is in accordance with Pan and Singleton (2008) who found that the term structure of CDS spreads is associated with a default risk premium. An increase in the default risk premium pushes up the long-term CDS spreads more than the short-term CDS spreads, leading to a steeper term structure of CDS spreads.

We admit that the information from Fig. 2 is insufficient to label the latent factors, therefore we have regressed the latent factors on the economic variables and find that it's not easy to label the factors by the chosen economic variables.<sup>1</sup> The difficulty is attributable to that the chosen economic variables such as the change of interest rate, the change of yield curve, the credit spread change and the change of VIX level generally exhibit the indistinguishable contributions or explanatory powers for the latent factors. In our findings, the latent factors are linear combination of the economic variables. These economic variables are highly correlated since they are governed by the same latent factors. Applying them together into the regression may result in a collinearity problem and bias our interpretation. For instance, in our case the change of VIX level almost dominates across the four factors. Eichengreen et al. (2012) claim that the exact association of a economic variable with any one of the latent factors is hard to define due to non-uniqueness of the factor estimates. Although our interpretation for Fig. 2. is not testable, the information from Fig. 2 helps to propose the observed economic variables in the subsequent analysis.

### 3.4 Connecting latent factors with observed variables

To realize the degree of association between the unobservable factors and observable economic variables, and to answer the question of interest; whether some of the observables are in fact underlying latent factors, we apply the method developed by Bai and Ng (2006) to determine if the observed and the latent are identical. The observed indicator with a stronger coherence with the latent factors is a good proxy. Two statistical criteria, the  $R^2$  and the noise-to-signal ratio, are used to examine whether any of the economic series yields the same information that is contained in the factors.

Let  $G_t$  be an  $J$ -dimensional vector of observed economic variables. The basic idea behind the test developed by Bai and Ng (2006) is to investigate whether any of the economic series can be represented as a linear combination of the latent factors by

---

<sup>1</sup> We appreciate the suggestion from the reviewer and the editor.

permitting a limited degree of noise in this association, thus

$$G_{j,t} = \beta_j^T F_t + \varsigma_{j,t} \quad (5)$$

where  $\beta_j$  is estimated by the OLS regression, and  $\varsigma_{j,t}$  is denoted as the error term. The above equation yields the predicted value  $\hat{G}_{j,t} = \hat{\beta}_j^T \hat{F}_t$ .  $R^2(j)$  is designed to measure the association between  $G_{j,t}$  and  $\hat{G}_{j,t}$ , and defined as:

$$R^2(j) = \frac{\widehat{\text{var}}(\hat{G}_j)}{\widehat{\text{var}}(G_j)} \quad (6)$$

where  $\widehat{\text{var}}(\cdot)$  denotes the sample variance and  $\widehat{\text{var}}(\hat{G})$  is computed by using the sample analog of the factors' asymptotic covariance matrix.  $R^2(j)$  is bounded between zero and one. It is equal to one if they have a high association, and is close to zero in the absence of correlation. A second measure  $NS(j)$ , called the noise-to-signal ratio, is constructed as:

$$NS(j) = \frac{\widehat{\text{var}}(\hat{\varsigma}_j)}{\widehat{\text{var}}(G_j)} \quad (7)$$

A larger  $NS(j)$  thus indicates an important departure of  $G_j$  from the latent factors. Normally, the magnitude of  $R^2(j)$  is reverse to that of  $NS(j)$  since the sum of  $R^2(j)$  and  $NS(j)$  should be equal to one.

As further observed economic variables in Eq. (5), one may include the change of the interest rate level, change of the credit spread, change of the interest rate term structure and the change of the stock index volatility. These variables are suggested by [Collin-Dufresne et al. \(2001\)](#), [Benkert \(2004\)](#) and [Ericsson et al. \(2009\)](#) since they are important determinants of credit assets. We limit our attention to the U.S. variables because the corresponding European variables are highly correlated with the U.S. series. The 1-year Treasury bond rate represents the level of the risk-free interest rate in the U.S. The difference between the 10-year Treasury bond rate and the 1-year Treasury bond rate is used to evaluate the slope of the yield curve in the U.S. The credit spread in the U.S. is the difference between the average Moody's Baa yield and the average Moody's Aaa yield of U.S. corporate bonds. We also employ the CBOE VIX index to measure generalized risk aversion.

Table 4 shows the association of the first four factors with the chosen economic variables. For the entire sample period, the  $R^2$  criterion gives a value of 0.3 and 0.375 on the credit spread and VIX index, respectively. The four factors are more correlated with the credit spread and VIX, and less correlated with the level and the term structure of the interest rate. This finding is accordance with [Cao et al. \(2010\)](#), [Cremers et al. \(2008\)](#) and [Collin-Dufresne et al. \(2001\)](#). The implication is that perceptions of credit risk were shaped by the common factors that are best summarized by credit spread and a generalized risk aversion. In other words, the result suggests that a higher credit spread or a higher generalized risk aversion does actually translate into systematic credit risk. Analogically, the sub-period analysis reports that credit spread and VIX

**Table 4** The association between the latent factors and the economic variables

	Entire	Pre-crisis	During crisis	Post-crisis
Level	0.156	0.082	0.252	0.122
Credit spread	0.300	0.294	0.418	0.286
Yield curve	0.132	0.009	0.160	0.245
VIX	0.375	0.267	0.350	0.590

The  $R^2$  criterion defined in Eq. (5) is calculated and reported. The observed economic variables include the 1-year Treasury bond rate that represents level of the risk-free interest rate in U.S., the credit spread measured as the difference between the average Moody's Baa yield and the average Moody's Aaa yield of U.S. corporate bonds, the slope of the yield curve as the difference between the 10-year treasury bond rate and the 1-year treasury bond rate, CBOE VIX index to measure the generalized risk aversion

are relatively correlated with the latent factors prior to the crisis. During the crisis, the  $R^2$  criterion even gives a value of 0.418 on credit spread, implying that the latent factors are best summarized by credit spread. The post-crisis analysis reveals that a generalized risk aversion with 0.59  $R^2$  criterion is highly associated to common factors.

### 3.5 Factor risk prices

How the market prices the factor risk inherent in the CDS spreads is of interest, since one can deduce how the market compensates investors, often referred to as the protection sellers, for bearing credit risk. If we fit the factor model into the framework of the arbitrage pricing theory (Ross 1976), the factor model for an  $N$ -dimensional returns on CDS indices of different credit ratings, maturities and regions,  $R_t$ , at time  $t$  can be presented as

$$R_t = \lambda\gamma + \lambda F_t + e_t \quad (8)$$

The arbitrage pricing theory states that the cross-section returns,  $R_t$ , are determined by  $K$  common factors  $F_t$  through the  $N \times K$  factor loading matrix  $\lambda$ . Given the assumption that the unobservable common factor  $F_t$  and error term  $e_t$  are *i.i.d.* distributed, the elements of the  $K$ -dimensional vector  $\gamma$  can be interpreted as the market prices of factor risk. Eq. (8) implies that the expected CDS returns satisfy

$$E(R_t) = \lambda\gamma \quad (9)$$

Given the estimated factor loadings  $\lambda$ , we can estimate the prices of factor risk  $\gamma$  by the generalized methods of moments (GMM) (Hansen 1982) on the moment restrictions in Eq. (9). This is equivalent to a GLS regression of the average changes of CDS indices on the factor loading matrix  $\lambda$ . Since we adopted a four-factor model in the previous sections, the GMM method enables us to estimate the prices of factor risk in this model and test their significance. As shown in Table 5, the market prices of a four-factor model are all significant, and the first two factors exhibit a promising size

**Table 5** Estimation of factor risk prices

	Four-factor model	Five-factor model
Factor 1	-0.0521 (-3.873)	-0.0498 (-4.957)
Factor 2	0.0121 (4.023)	0.0156 (4.940)
Factor 3	0.0055 (2.902)	0.0052 (4.393)
Factor 4	0.0009 (2.575)	0.0009 (4.240)
Factor 5		0.0005 (0.895)
<i>J</i> -statistic	1.206 (0.876)	1.445 (0.842)
$R^2$ of GLS	95.42%	95.89%

The market price of factor risk is estimated using the GMM, and the value in parentheses is *t*-statistic. The GMM *J*-statistics and the associated *p* values in parentheses are also presented to test the over identifying restrictions. The  $R^2$  of GLS regression evaluates the goodness-of-fit of the factor models

in their risk prices. If we consider a five-factor model, the risk prices are significant in the first four factors but insignificant in the fifth factor.

Table 5 also contains the GMM *J*-statistic, a test statistic for testing the over identifying restrictions in Eq. (9), and the corresponding *p* value. The *J*-statistic acts as an omnibus test statistic for model miss-specification. In a well specified over identifying model with valid moment conditions, the *J*-statistic behaves like a Chi-square random variable with degrees of freedom equal to the number of over identifying restrictions. Typically, a large *J*-statistic indicates a miss-specified model. In Table 5, the *J*-statistics in the both four- and five-factor models cannot reject the null hypothesis, implying that both models are well-specified. Furthermore, the four- and five-factor models provide a good fit, as measured by the  $R^2$  of the GLS regression, which is equal to 95.42 and 95.89 %, respectively. The results from *J*-statistic,  $R^2$  of the GLS and the significance of factor prices suggest that the four-factor model is efficient enough to measure the CDS returns.

## 4 Method of asymptotic principal components and forecast performance

### 4.1 Competing factor models

According to this study and previous literature, the common latent factors extracted from factor models have proven their representative ability for systematic credit risk. This motivates us to examine whether modelling the time series properties of the factors can improve our ability to forecast the time-variation of CDS index changes. Acting as the benchmark model, the static model in Eq. (1) is too restricted to accommodate the realistic time-variation. The latent factors it produces can only follow one of the few plausible, realistic patterns that do actually appear in the credit markets. The generalized models in which the factors could be defined in a general way are developed to minimize the gap, and should entail less restrictions.

*The dynamic factor model*, a simple vector autoregressive (VAR) specification, is the first shown to achieve a remarkable fit of the factors' dynamics. By permitting a

VAR specification in the factors with autoregressive parameters  $\mathbf{B}$ , this model captures the common dynamics in the cross-sectional analysis. Additionally, the error term,  $\mathbf{u}_t$ , from a VAR equation in Eq. (10) is conditionally heteroscedastic and follows a GARCH( $p, q$ ) process.

$$S_{it} = \mathbf{F}_t \lambda_{it} + e_{it}$$

$$\left(\mathbf{I} - \mathbf{B}_1 L - \dots - \mathbf{B}_h L^h\right) \mathbf{F}_t = \mathbf{u}_t \tag{10}$$

$$\mathbf{u}_t = \mathbf{H}_t^{1/2} \boldsymbol{\eta}_t \tag{11}$$

$$vech(\mathbf{H}_t) = \mathbf{c} + \sum_{j=1}^q \mathbf{A}_j vech\left(\mathbf{u}_{t-j} \mathbf{u}_{t-j}^T\right) + \sum_{j=1}^p \mathbf{D}_j vech(\mathbf{H}_{t-j}) \tag{12}$$

where  $i = 1, \dots, N, t = 1, \dots, T, \mathbf{F}_t$  is  $T \times k$  and  $\lambda_{it}$  is  $k \times 1$ .  $\boldsymbol{\eta}_t$  is white noise.

To take into account the possibility that the idiosyncratic errors in Eq. (1) may entail serial and cross-section correlation, the dynamic factor with dependent error model is built with additional assumptions on the idiosyncratic components shown in Eqs. (13), (14) and (15).

$$S_{it} = \mathbf{F}_t \lambda_{it} + e_{it}$$

$$\left(\mathbf{I} - \mathbf{B}_1 L - \dots - \mathbf{B}_h L^h\right) \mathbf{F}_t = \mathbf{u}_t$$

$$\mathbf{u}_t = \mathbf{H}_t^{1/2} \boldsymbol{\eta}_t$$

$$vech(\mathbf{H}_t) = \mathbf{c} + \sum_{j=1}^q \mathbf{A}_j vech\left(\mathbf{u}_{t-j} \mathbf{u}_{t-j}^T\right) + \sum_{j=1}^p \mathbf{D}_j vech(\mathbf{H}_{t-j})$$

$$(1 - \alpha L) e_{it} = v_{it} + \theta_1 v_{i+1,t} + \theta_2 v_{i-1,t} \tag{13}$$

$$v_{it} = \sigma_{it} \eta_{it} \tag{14}$$

$$\sigma_{it}^2 = \delta_0 + \delta_1 \sigma_{i,t-1}^2 + \delta_2 v_{i,t-1}^2 \tag{15}$$

The idiosyncratic components,  $e_{it}$ , in Eq. (13) are serially correlated, with an AR(1) coefficient  $\alpha$ , and weakly cross-section correlated with the coefficients  $\theta_1$  and  $\theta_2$ . The innovations  $v_{it}$  are conditionally heteroscedastic and follow a GARCH(1,1) process with parameters  $\delta_0, \delta_1$ , and  $\delta_2$  in Eq. (15).

In practice, when factors are constructed over a long period, some degree of temporal instability is inevitable. Following Stock and Watson (2002), we model this instability as stochastic drift in the factor loadings, and the factor loading evolves through time with a serial correlation  $\rho_i$  shown in Eq. (16).

$$\lambda_{it} = \rho_i \lambda_{i,t-1} + (c/T) \zeta_{it} \tag{16}$$

where  $\zeta_{it}$  is white noise. Equation (16) implies that factor loadings for the  $i$ th variable shift by an amount,  $(c/T) \zeta_{it}$ , in time period  $t$ . In addition, it keeps a relationship with its previous level which is measured by  $\rho_i$ . The time-varying factor loading model ideally incorporates all of the features covering from Eqs. (10) to (16). Whether



this model is more superior due to its abundant generalization will be examined with respect to its predictive ability, and will be analyzed in the subsequent section.

## 4.2 Out-of-sample forecasting performance

Having proposed the competing models developed by more general ways, we take an explicit out-of-sample forecasting approach to evaluate their predicting performance regarding the CDS dynamics. Using the previous 1-year weekly data, we estimate the parameters and produce a 1-week ahead forecast. After estimation, we find that the dynamic of the CDS index captured by these factor models exhibits significant time-variation and persistence, and we summarize their forecasting performance in Table 6. The most outperformed one can be potentially applied to price credit risk accurately and achieve a better credit risk management.

To assess an out-of-sample forecasting performance, for each proposed model we compute each day  $t$ , the following four measures (a) mean squared error (MSE) between the observed change of CDS spreads and the predicted change of CDS spreads from the competing factor models; (b) mean absolute error (MAE); (c) mean correct prediction (MCP) of the direction of change in CDS spreads. The MCP exhibits the average numbers from  $N$  CDS indices are correctly forecast based on their signs of changes; (d) the trace of  $R^2$  of the multivariate regression of  $\hat{S}$  onto  $S$ ,

$$R_{\hat{S}, S}^2 = \hat{E} \| P_S \hat{S} \|^2 / \hat{E} \| \hat{S} \|^2 = \hat{E} \text{tr} \left( \hat{S}^T P_S \hat{S} \right) / \hat{E} \text{tr} \left( \hat{S}^T \hat{S} \right), \quad (17)$$

where  $S$  is a  $T \times N$  matrix comprising  $N$  units until time  $T$ ,  $\hat{E}$  denotes the expectation estimated by averaging the relevant statistic and  $P_S = S (S^T S)^{-1} S^T$ . As shown in Table 6, the time-varying factor loading model exhibits the best 1-week ahead point-forecast performance with the lowest MSE, MAE and the highest MCP, trace of  $R^2$ . For each model, we measure the forecasting performances under different numbers of factors that range from one to seven. Table 6 indicates that the dynamic factor model and the time-varying factor loading model constitute a promising improvement over the static factor model. A poorest forecast performance in the static factor model implies that the factors exhibit persistency, predictability and temporal instability, and these characteristics contribute to the prediction on the changes of CDS spreads. We further conduct a test for their equal predictive ability against the static factor model in Sect. 4.3.

Determining the number of factors can be regarded as a model selection problem, which is a trade-off between goodness-of-fit and parsimony. Following Bai and Ng (2002), the number of factors is estimated by an information criteria function (IC):

$$k = \arg \min_{0 \leq k \leq k_{max}} IC(k) \quad (18)$$

where  $IC(k) = \log \left( V(k, \hat{F}^k) \right) + kg(N, T)$ .  $V(k, \hat{F}^k) = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T (S_{it} - \hat{F}_t^k \lambda_i^k)^2$  is simply the average residual variance, and  $g(N, T)$  is a penalty func-

**Table 6** Forecasting performance

	MSE	MAE	MCP	Trace $R^2$	$IC_{p1}$	$IC_{p2}$	$IC_{p3}$
<i>A. Static factor model</i>							
$k = 1$	837.196	14.479	4.184	0.079	7.014	7.041	6.989
$k = 2$	935.015	15.225	4.113	0.090	7.409	7.464	7.360
$k = 3$	980.284	15.649	4.113	0.095	7.741	7.823	7.667
$k = 4$	994.165	15.797	4.067	0.096	8.040	8.149	7.941
$k = 5$	1011.411	15.915	4.166	0.098	8.341	8.478	8.218
$k = 6$	1011.353	16.002	4.083	0.098	8.626	8.790	8.478
$k = 7$	1014.162	16.074	4.067	0.098	8.913	9.105	8.741
<i>B. Dynamic factor model</i>							
$k = 1$	512.226	11.061	4.127	0.123	6.523	6.550	6.498
$k = 2$	515.263	11.387	4.109	0.108	6.813	6.876	6.812
$k = 3$	521.053	11.530	4.072	0.106	7.109	7.191	7.035
$k = 4$	527.623	11.547	3.949	0.105	7.406	7.516	7.308
$k = 5$	518.325	11.604	4.040	0.109	7.673	7.810	7.550
$k = 6$	521.404	11.634	4.149	0.112	7.963	8.128	7.816
$k = 7$	521.863	11.618	4.189	0.110	8.249	8.440	8.076
<i>C. Dynamic factor with dependent errors model</i>							
$k = 1$	725.655	13.458	4.069	0.082	6.871	6.898	6.847
$k = 2$	540.526	12.439	4.125	0.098	6.861	6.876	6.812
$k = 3$	534.201	11.844	4.127	0.110	7.134	7.721	7.060
$k = 4$	526.395	11.672	4.109	0.115	7.404	7.513	7.305
$k = 5$	524.747	11.628	4.021	0.113	7.685	7.822	7.562
$k = 6$	527.945	11.575	4.076	0.105	7.976	8.140	7.828
$k = 7$	521.499	11.568	4.123	0.110	8.248	8.440	8.076
<i>D. Time-varying factor loading model</i>							
$k = 1$	784.773	13.293	3.985	0.036	6.949	6.977	6.925
$k = 2$	509.891	12.079	4.101	0.129	6.803	6.858	6.754
$k = 3$	493.244	11.744	4.090	0.114	7.054	7.136	6.980
$k = 4$	479.815	11.443	4.105	0.151	7.311	7.421	7.213
$k = 5$	479.944	11.415	4.061	0.155	7.596	7.733	7.473
$k = 6$	481.839	11.384	4.130	0.148	7.885	8.049	7.737
$k = 7$	479.683	11.383	4.185	0.156	8.165	8.356	7.992

The information criteria function  $IC_{p1}$ ,  $IC_{p2}$  and  $IC_{p3}$  can be referred to (20), (21) and (22) in the text

tion for overfitting. Bai and Ng (2002) have proposed three specific formulations of  $g(N, T)$  that depend on both  $N$  and  $T$ .

$$IC_{p1}(k) = \log \left( V \left( k, \hat{F}^k \right) \right) + k \left( \frac{N+T}{NT} \right) \log \left( \frac{NT}{N+T} \right) \quad (19)$$

$$IC_{p2}(k) = \log \left( V \left( k, \hat{F}^k \right) \right) + k \left( \frac{N+T}{NT} \right) \log (\min \{N, T\}) \quad (20)$$

$$IC_{p3}(k) = \log \left( V \left( k, \hat{F}^k \right) \right) + k \left( \frac{\log (\min \{N, T\})}{\min \{N, T\}} \right) \quad (21)$$

Table 6 summarizes the results of the  $IC$  function and shows that for both the static factor model and the dynamic factor model, the one-factor model with the minimized information criteria is the best one to model the common factors in the changes of CDS spreads. However, for both the dynamic factor with dependent errors model and for the time-varying factor loading model, the two-factor model is relatively adequate.

### 4.3 Testing equal predictive ability

To formally assess the statistical significance of the superior out-of-sample performance of the dynamic factor models over the static factor model, we employ the equal predictive ability test of Diebold and Mariano (1995) and report the testing results in Table 7. Diebold and Mariano (1995) propose a method for measuring and assessing the significance of divergences between two competing forecasts, and allow for forecast errors that are potentially non-Gaussian, serially correlated and contemporaneously correlated.

To be specific, let  $d_t$  be the loss differential between two forecast errors. The null hypothesis is no difference in the accuracy of two forecasts, that is  $Ed_t = 0$ . The asymptotic distribution of the sample mean loss differential is:

$$\sqrt{T} (\bar{d} - \mu) \sim N(0, 2\pi f_d(0)) \quad (22)$$

where  $f_d(0)$  is the spectral density of the loss differential at frequency 0.

The statistical significance of the difference in forecast errors between the models is summarized in Table 7. The tabulated  $p$  values indicate that we can reject the null hypothesis of equal forecasting ability between the static factor model and the time-varying factor model. We also reject the equal predicting ability between the static factor model and the dynamic factor with dependent errors model. With the exception in CDX 5-year IG and 10-year HY indices, the equal predictive ability between the static factor model and the dynamic factor model is rejected. Furthermore, to claim that the time-varying factor model is the best one, we compare its forecast ability with the dynamic factor model, and the dynamic factor with dependent errors model. We find that significant differences exist in their predicting ability in both cases.

In summary, the results in Table 6 together with Table 7 indicate that the time-varying factor model reveals a statistically significant outperformance for most of the cases, suggesting that common factors drive the time-variation of CDS spreads and that the dynamics in the factors exhibit moderate predictability in the short-run. As evident, the temporal instability in the common factors is inevitable and contributes to forecasting. However, the serial or cross correlation in the idiosyncratic components only have little effect on the forecasts, implying that the common factors dominate the

**Table 7** Comparing predictive accuracy

	Static factor versus		Dynamic factor versus		Dynamic factor versus	
	Dynamic factor	Dynamic factor dependent errors	Time-varying factor loading	Dynamic factor dependent errors	Time-varying factor loading	Time-varying factor loading
CDX.IG.5Y	1.345 (0.089)	26.337 (0.000)	7.588 (0.000)	29.714 (0.000)	9.135 (0.000)	37.224 (0.000)
CDX.IG.10Y	3.985 (0.000)	6.801 (0.000)	13.870 (0.000)	5.669 (0.000)	17.019 (0.000)	11.719 (0.000)
CDX.HY.5Y	2.479 (0.006)	3.118 (0.000)	6.188 (0.000)	1.930 (0.026)	6.887 (0.000)	5.568 (0.000)
CDX.HY.10Y	1.567 (0.058)	8.736 (0.000)	7.136 (0.000)	16.304 (0.000)	14.266 (0.000)	3.399 (0.000)
EU.IG.5Y	2.175 (0.014)	9.721 (0.000)	10.397 (0.000)	16.590 (0.000)	16.910 (0.000)	2.490 (0.006)
EU.IG.10Y	8.376 (0.000)	7.283 (0.000)	17.643 (0.000)	1.625 (0.052)	23.472 (0.000)	24.876 (0.000)
EU.HY.5Y	1.808 (0.035)	4.587 (0.000)	0.892 (0.186)	15.280 (0.000)	7.392 (0.000)	13.696 (0.000)
EU.HY.10Y	5.070 (0.000)	7.032 (0.000)	12.389 (0.000)	6.983 (0.000)	14.079 (0.000)	8.240 (0.000)

This table reports the statistics and  $p$  values (parentheses) of the [Diebold and Mariano \(1995\)](#) test of equal predictive ability

predicting performance. The predictability of CDS spreads changes, certainly benefits the hedging, speculating and arbitraging activities in the credit markets.

## 5 Conclusion

The commonalities in CDS spreads and their factor loadings are analyzed in this study. We collect CDS indices in North American and Europe with 5- and 10-year maturities, and with different credit ratings (IG and HY) from October 2004 to June 2011. The estimated risk factors can be interpreted as the *level*, the *region*, the *volatility* and the *term structure* effect. By conducting a test if there are common principal components, we find that the eigenstructures are distinct for the pre-, during and post-crisis periods. The first factor explains 58.7% of the variance in the pre-crisis period, 72.3% of the variance in the crisis period and 47% of the variance in the post-crisis period, indicating that during the crisis, CDS spreads are increasingly driven by common factors and less by idiosyncratic components. We also find that during the crisis the latent factors are more correlated with the credit spread and VIX, and less correlated with the level and the term structure of the interest rate.

The time-variation of CDS spreads changes is modelled via various dynamic factor models. We apply the asymptotic principal component technique to extract the common factors, and then determine the number of factors by information criteria functions. The out-of-sample forecasting performance and the results of equal predictive ability indicate that the common factors drive the time-variation of CDS spreads and the dynamics in the factors exhibit moderate predictability in the short-run. In addition, the temporal instability in the common factors is inevitable and contributes to forecasting, but the serial or cross correlation in the idiosyncratic components have little effect on the forecasts.

## References

- Anderson RW (2008) What accounts for time variation in the price of default risk? Working paper
- Avellaneda M, Cont R (2010) Transparency in credit default swap markets. *Finance Concepts* 1–23. <http://www.finance-concepts.com/images/fc/CDSMarketTransparency.pdf>
- Bai J, Ng S (2002) Determining the number of factors in approximate factor models. *Econometrica* 70:191–221
- Bai J, Ng S (2006) Evaluating latent and observed factors in macroeconomics and finance. *J Econom* 131:507–537
- Benkert C (2004) Explaining credit default swap premia. *J Fut Mark* 24:71–92
- Cao C, Yu F, Zhong Z (2010) The information content of option-implied volatility for credit default swap valuation. *J Financ Mark* 13:321–343
- Cesare AD, Guazzarotti G (2010) An analysis of the determinants of credit default swap spread changes before and during the subprime financial turmoil. *Bank of Italy Temi di Discussione* 749:1–45
- Collin-Dufresne P, Goldstein RS, Martin JS (2001) The determinants of credit spread changes. *J Finance* 56:2177–2207
- Cremers M, Driessen J, Maenhout P, Weinbaum D (2008) Individual stock-option prices and credit spreads. *J Bank Finance* 32:2706–2715
- Das SR, Freed L, Geng G, Kapadia N (2006) Correlated default risk. *J Fixed Income* 16:7–32
- Das SR, Duffie F, Kapadia N, Saita L (2007) Common failings: how corporate defaults are correlated. *J Finance* 62:93–117
- Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *J Bus Econ Stat* 13:253–263

- Driessen J, Melenberg B, Nijman T (2003) Common factors in international bond returns. *J Int Money Finance* 22:629–656
- Duffie D (1998) Credit swap valuation. *Financ Anal J* 55:73–87
- Duffie D, Eckner A, Horel G, Saita L (2009) Frailty correlated default. *J Finance* 64:2089–2123
- Eichengreen B, Mody A, Nedeljkovic M, Sarno L (2012) How the subprime crisis went global: evidence from bank credit default swap spreads. *J Int Money Finance* 31:1299–1318
- Ericsson J, Jacobs K, Oviedo R (2009) The determinants of credit default swap premia. *J Financ Quant Anal* 44:109–132
- Flury B (1988) *Common principle components analysis and related multivariate models*. Wiley, New York
- Härdle W, Simar L (2011) *Applied multivariate statistical analysis*, 3rd edn. Springer, Berlin
- Hansen L (1982) Large sample properties of generalized methods of moments estimators. *Econometrica* 50:1029–1054
- Longstaff FA, Schwartz ES (1995) A simple approach to valuing risky fixed and floating rate debt. *J Finance* 50:789–819
- Pan J, Singleton KJ (2008) Default and recovery implicit in the term structure of sovereign CDS spreads. *J Finance* 63:2345–2384
- Ross SA (1976) The arbitrage theory of capital. *J Financ Econ* 13:341–360
- Stock JH, Watson MW (2002) Forecasting using principal components from a large number of predictors. *J Am Stat Assoc* 97:1167–1179



## An Application of Principal Component Analysis on Multivariate Time-stationary Spatio-temporal Data

Stephan Stahlschmidt, Wolfgang K. Härdle & Helmut Thome

To cite this article: Stephan Stahlschmidt, Wolfgang K. Härdle & Helmut Thome (2015) An Application of Principal Component Analysis on Multivariate Time-stationary Spatio-temporal Data, *Spatial Economic Analysis*, 10:2, 160-180, DOI: [10.1080/17421772.2015.1023339](https://doi.org/10.1080/17421772.2015.1023339)

To link to this article: <http://dx.doi.org/10.1080/17421772.2015.1023339>



Published online: 27 Mar 2015.



Submit your article to this journal [↗](#)



Article views: 123



View related articles [↗](#)



View Crossmark data [↗](#)

Full Terms & Conditions of access and use can be found at  
<http://www.tandfonline.com/action/journalInformation?journalCode=rsea20>

## An Application of Principal Component Analysis on Multivariate Time-stationary Spatio-temporal Data

STEPHAN STAHLSCHMIDT, WOLFGANG K. HÄRDLE & HELMUT THOME

(Received February 2014; accepted November 2014)

**ABSTRACT** *Principal component analysis (PCA) denotes a popular algorithmic technique to dimension reduction and factor extraction. Spatial variants have been proposed to account for the particularities of spatial data, namely spatial heterogeneity and spatial autocorrelation, and we present a novel approach which transfers PCA into the spatio-temporal realm. Our approach, named spatio-temporal principal component analysis (stPCA), allows for dimension reduction in the attribute space while striving to preserve much of the data's variance and maintaining the data's original structure in the spatio-temporal domain. Additionally to spatial autocorrelation stPCA exploits any serial correlation present in the data and consequently takes advantage of all particular features of spatio-temporal data. A simulation study underlines the superior performance of stPCA if compared to the original PCA or its spatial variants and an application on indicators of economic deprivation and urbanism demonstrates its suitability for practical use.*

### Une application de l'Analyse de Composante principale sur des données spatio-temporelles à temps stationnaire multivarié

**RÉSUMÉ** *L'analyse en composante principale (ACP) dénote une technique algorithmique populaire pour la réduction de dimensions et l'extraction de facteurs. Des variantes spatiales ont été proposées pour tenir compte des particularités des données spatiales, à savoir l'hétérogénéité spatiale et l'autocorrélation spatiale, et nous présentons une nouvelle méthode transférant l'analyse en composante principale dans le contexte spatio-temporel. Notre méthode, dénommée ACPst, tient compte de la réduction des dimensions dans l'attribut espace, tout en s'efforçant de conserver une grande partie de la variance des données et en maintenant la structure originale des données dans le contexte spatio-temporel. En plus de l'autocorrélation spatiale, l'ACPst exploite toute corrélation série présente dans les données, et tient compte, en conséquence, de toutes les particularités des données spatio-temporelles. Une étude de simulation souligne le rendement supérieur de l'ACPst lorsqu'on le compare à l'ACP*

Stephan Stahlschmidt (to whom correspondence should be sent), School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. Email: [stahlschmidt@wiwi.hu-berlin.de](mailto:stahlschmidt@wiwi.hu-berlin.de). Wolfgang K. Härdle, C.A.S.E Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. Email: [haerdle@wiwi.hu-berlin.de](mailto:haerdle@wiwi.hu-berlin.de). Helmut Thome, Institute of Sociology, Martin-Luther-Universität Halle-Wittenberg, Institute of Sociology, Adam-Kuckhoff-Straße 41, 06108 Halle (Saale), Germany. Email: [helmut.thome@soziologie.uni-halle.de](mailto:helmut.thome@soziologie.uni-halle.de)



*original ou ses variantes spatiales, et une application sur les indicateurs du dénuement économique et de l'urbanisme démontre sa convenance pour des applications pratiques.*

### **Una aplicación del análisis de componentes principales sobre datos multivariantes espacio-temporales y estacionarios en el tiempo**

RESUMEN *el análisis de componentes principales indica una técnica algorítmica conocida para la reducción dimensional y la extracción factorial. Se han propuesto variables espaciales para tener en cuenta las particularidades de los datos espaciales, específicamente la heterogeneidad espacial y la autocorrelación espacial, y presentamos un nuevo enfoque que transfiere el análisis de componentes principales al dominio espaciotemporal. Nuestro enfoque, que se denomina stPCA, da cabida a la reducción dimensional en el espacio de los atributos, además de preservar una gran parte de la varianza de los datos y de mantener la estructura original de los datos en el dominio espaciotemporal. Además de la autocorrelación espacial, el método stPCA explota cualquier correlación serial presente en los datos y, en consecuencia, aprovecha todas las características particulares de los datos espaciotemporales. Un estudio de simulación destaca el rendimiento superior del método stPCA si se compara con el PCA o sus variantes espaciales y una aplicación sobre los indicadores de privación económica y urbanismo, demuestra su idoneidad para el uso práctico.*

#### **关于多变量短时平稳时空数据的主成分分析应用**

##### **摘要**

主成分分析法是指一种流行的降维和因素抽取算法技术。已有人提出用空间变异来解释空间数据的特殊性，即空间异质性和空间自相关。我们则提出了一种将主成分分析转移到时空境界的新的做法。我们采用的名为 stPCA 的方法，可在属性空间降低维度，同时尽量保持更多数据的方差，并维护时空域中数据的原始结构。除了空间自相关之外，stPCA 还利用存在于数据中的序列相关性。因此，利用了所有时空数据的特殊功能。与原来的PCA或其空间变异相比，模拟研究强调了 stPCA 的卓越性能，而经济贫困和城市化指标的应用则表明了其实际应用的适用性。

KEYWORDS: *dimension reduction; economic deprivation; factor extraction; PCA; spatio-temporal analysis; urbanism*

JEL CLASSIFICATION: C31; C33; R11

### **1. Introduction**

Factor extraction refers to the process of concentrating several variables into a set of factors with lower cardinality and has been applied in virtually any field of statistical analysis. It denotes a dimension reduction technique, as well as a vehicle to disclose latent factors. Because of the reduction factor, extraction relieves the computational burden in any subsequent analysis, might help to avoid the curse of dimensionality and most importantly presents measurements of theoretical interest which would otherwise remain hidden due to incomplete knowledge on the subject matter or

due to the latent nature of the variable of interest. Consequently, factor extraction might be understood as an analysis tool, which helps to identify the relevant factors of interest.

Principal component analysis (PCA; Pearson, 1901; Hotelling, 1933), which is also known as discrete Karhunen–Loève transformation (Karhunen, 1947; Loève, 1948), Hotelling transformation (Hotelling, 1933) or the method of empirical orthogonal functions (Lorenz, 1956) among others, is frequently applied to extract factors from a set of variables (e.g. Jolliffe, 2002, chap. 4). It is in fact based on a transformation of the data, in which the orthogonal coordinates are rotated in order to load as much variance as possible on the first components and less and less variance on subsequent components. Consequently the first components, formed by a linear combination of the original variables, represent an essential information content of the data and might be understood as factors. By contrast the final components, presenting little residual variance, might be ignored in the analysis and allow thus for dimension reduction. In a strict implementation without any additional rotation and based on standardized variables, PCA resembles more an algorithm than a model and restricts the researcher's influence on choosing the appropriate number of latent factors. This feature distinguishes PCA from other factor extraction techniques, most notably the model-based factor analysis (Spearman, 1904).

However, the application of the PCA algorithm is not exclusively restricted to the attribute subspace, but in case of spatio-temporal data might also be used on the geographical or temporal subspace and consequently reduce either the geographical or the temporal dimension. Demšar et al. (2013) review the application of PCA in the context of spatial data and Richman (1986) proposes a classification of PCA for spatio-temporal data into six modes, where each mode describes exclusive combinations of two subspaces. For example, the application of PCA on multivariate spatial entities is *labelled* R-mode and several spatial PCA variants have been proposed (Wartenberg, 1985; Thioulouse et al., 1995; Fotheringham et al., 2002; Jombart et al., 2008). Contrary to the original PCA, these techniques incorporate either spatial autocorrelation or spatial heterogeneity into the PCA approach to factor extraction and the authors demonstrate the superior performance of these spatial PCA variants to disclose any spatial factor if compared to the original PCA.

On the other hand, these spatial PCA variants only address spatial cross-sectional data and do not apply to spatio-temporal data. In order to allow for a truly spatio-temporal analysis, we propose a novel PCA approach, that not only accounts for the spatial peculiarities, but also incorporates serial correlation over time. This spatio-temporal PCA variant (henceforth spatio-temporal principal component analysis [stPCA]) allows for dimension reduction on the attribute space, while preserving the geographical and temporal space, that is, it extracts spatio-temporal factors from several spatio-temporal variables while maintaining the geographical and temporal structure of the original variables.

In the framework of Richman (1986) stPCA can be understood as the combined PR-mode of PCA on spatio-temporal data and the technique describes a transfer of the original PCA to the spatio-temporal realm of geographical and serial correlation. Consequently the proposed technique shares some features with the three-mode PCA of Kroonenberg & de Leeuw (1980), which however relies on independent and identically distributed (i.i.d.) observations and has not been studied for correlated observations. Furthermore three-mode PCA includes a dimension reduction in every subspace, whereas stPCA focuses exclusively on the attribute subspace.

The inclusion of latent factors in models for spatio-temporal data is also facilitated by Bayesian hierarchical models (Gelman & Hill, 2006). Recent examples include Tzala & Best (2007), Lawson et al. (2008) and Choi et al. (2012) in public health studies and Hogan & Tchernis (2004) in economics. These models rely on latent factors to regress some explanatory variables on a dependent variable and the latent factors consequently serve as an intermediate step and are not of particular interest in the respective analysis. stPCA consequently represents a novel attempt to incorporate spatial and temporal correlation into a PCA framework and hence facilitates the inclusion of latent factors into spatio-temporal models.

In order to illustrate the performance of stPCA, we present a simulation study and apply stPCA to a data-set of economic deprivation and urbanism indicators in Germany. In the Monte Carlo simulation, stPCA improves the ordinary and spatial PCA approaches if a non-negligible spatial structure is present in the spatio-temporal data. The reported difference is substantial and significant. A large gain is made on small  $n$ , high  $t$  samples, whereas the additional value for large  $n$  data seems less pronounced.

The application of stPCA on the indicators of economic deprivation and urbanism in Germany illustrates the additional value of a combined spatio-temporal approach if compared with a cross-sectional spatial approach. Only stPCA allows for time specific projections, which highlight the west-east and internal north-south divide in economic deprivation and reliably indicates the big German metropolitan areas.

The following Section 2 presents the proposed stPCA approach, which is afterwards evaluated via a simulation in Section 3. An actual implementation of stPCA is presented in Section 4 and Section 5 concludes with a discussion.

## 2. The stPCA

The original PCA of Pearson (1901) and Hotelling (1933) describes a rotation of the  $p$ -dimensional coordinate system. The rotated coordinates present the best orthogonal fit of the data, in which the first coordinate is aligned in the direction of the data's maximum variance. Any subsequent coordinate is afterwards orientated to contain as much of the residual variance as possible conditioned on being orthogonal to all former coordinates.

In this new coordinate system, the coordinates possessing much variance contain most of the data's information, whereas coordinates with a relative small amount of variance contribute little additional information and consequently can be ignored at little cost. This advantage of PCA is facilitated by the orthogonal rotation and allows for dimension reduction in multivariate data while preserving the general structure of the individual data points.

Upon obtaining the new coordinates the  $p$ -dimensional and centred random variables  $X \in R^p$  are projected onto this new coordinates system via a linear combination. The projections  $\phi$  onto the first coordinate are obtained via  $\phi = Xu$ , where  $u$  denotes a weight vector which can be identified via the aforementioned variance characteristics of the rotated coordinates. In detail, PCA maximizes the variance in the rotated coordinates, that is, the variance of the projected data points  $\phi$ :

$$\max_u \text{Var}(\phi) = \max_u \text{Var}(Xu) = \max_u u^\top n^{-1} X^\top Xu = \max_u u^\top \Sigma u \quad (1)$$

where  $\Sigma$  denotes the covariance matrix of the centred  $\mathbf{X}$  and the maximization is subject to some identification restriction, like  $\|u\| = 1$ .

An eigendecomposition of  $\Sigma$  resolves the maximization requirement (1), as the eigenvector corresponding to the largest eigenvalue constitutes the optimal  $u$  (Härdle & Simar, 2012). Likewise the projection onto subordinate components is conducted via the remaining eigenvectors, where the corresponding eigenvalues describe the variance explained by this component, and consequently its rank.

The just described original PCA does not address the particularities of spatial data, like spatial autocorrelation or spatial heterogeneity. Spatial extensions to PCA have been proposed, which explicitly account for either heterogeneity (Fotheringham et al., 2002) or autocorrelation (Wartenberg, 1985; Thioulouse et al., 1995; Jombart et al., 2008). In this paper we concentrate on the second type, but would like to note that the suggested spatio-temporal approach might also be adapted to the spatial heterogeneity case.

The suggested extensions amplify the maximization criterion by incorporating the spatial autocorrelation of the projected data points  $\phi$ . Consequently, the proposed methods seek to project the observations onto a new coordinate system, while preserving the spatial relation between the observations and this second objective differentiates the spatial approaches from the ordinary PCA.

Moran's I describes a frequently used statistic of spatial autocorrelation (Moran, 1950) which defines the spatial autocorrelation for some random variable  $X$  with mean  $\bar{X}$  as

$$I(X) = \frac{N}{\sum_{i=1}^N \sum_{j=1}^N w_{ij}} \frac{\sum_{i=1}^N \sum_{j=1}^N w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2}, \quad (2)$$

where  $w_{i,j}$ , drawn from a spatial weight matrix  $W$ , describes the spatial weight imposed by observation  $j$  onto observation  $i$ . Choosing an appropriate spatial weight matrix for either point or areal data and a suitable standardization is up to the subject-matter researcher and might simplify the computation (2).

Indeed the differences of the above-mentioned spatial PCA approaches can be attributed to the particular spatial weight matrix and specific transformation of the original variables  $\mathbf{X}$  chosen by the authors. In detail, Multivariate Spatial Correlation (MSC) (Wartenberg, 1985) standardizes the original variables and the distance based spatial weight matrix, whereas the Global Structure by Thioulouse et al. (1995) relies on a standardized binary connection matrix and transforms the original data by a mean which is based on assigning weights according to the number of individual neighbours. Finally, spatial principal component analysis (sPCA) (Jombart et al., 2008) applies a row standardization on the binary connection matrix and, because of its specific application to alleles does not standardize the data, but subtracts only the mean.

All these spatial extensions to PCA seek to maximize the product of the variance and spatial autocorrelation of  $\phi$ :

$$\max_v \text{Var}(\phi)I(\phi) = \max_v \text{Var}(\mathbf{X}v)I(\mathbf{X}v) = \max_v v^\top n^{-1} \mathbf{X}^\top W \mathbf{X} v = \max_v v^\top \Omega v$$

where  $\Omega = n^{-1} \mathbf{X}^\top W \mathbf{X}$  denotes a spatial correlation matrix and the optimal  $v$  are found via an eigendecomposition of  $\Omega$ . Wartenberg (1985) and Thioulouse et al.

(1995) point out, that  $\Omega$  might not be positive definite and state that the resulting negative eigenvalues represent local structure. In case of a non-symmetrical spatial weight matrix  $W$ , Jombart et al. (2008) observe, that the optimal  $\nu$  is given by the eigenvector corresponding to the largest eigenvalue of  $(2n)^{-1}X^\top(W + W^\top)X$ .

Spatio-temporal data add another subspace to the attribute and geographical space of spatial data and present measurements of the same multivariate spatial entities over time. Consequently, PCA or any spatial PCA variant could be applied at every  $t$ , and  $T$  eigendecompositions of the time dependent (spatial) covariance matrix could be computed. Hence any serial correlation over time would be ignored and at every  $t$  we would have a separate cross-sectional analysis.

Contrarily, stPCA forms a truly spatio-temporal technique. Instead of conducting an analysis at every  $t$  separately stPCA proposes to calculate a time average of the spatial covariance matrix and apply an eigendecomposition on this average. Consequently, stPCA exploits any serial correlation and makes use of the fact that the repeated measurements on the time stable spatial entities represent the same information content, whereas any additional noise might vary over time. Hence the time-averaged spatial covariance matrix will include a higher signal-to-noise ratio and present time stable eigenvectors.

This feature of stPCA, contrary to the repeated application of PCA or its spatial variants, will result in consistent signs and order of the components across  $t$  and consequently facilitates the interpretation and further use of the findings. Finally, stPCA is faster, as a function of  $t$ , than any repeated application of its non-temporal siblings, as the time-consuming eigendecomposition has to be conducted only once instead of  $t$  times.

In detail stPCA maximizes the time average of the product between the variance and spatial autocorrelation of the projected data points  $\phi$ :

$$\begin{aligned} \max_{\mu} T^{-1} \sum_{t=1}^T \text{Var}(\phi_t) I(\phi_t) &= \max_{\mu} T^{-1} \sum_{t=1}^T \text{Var}(\mathbf{X}_t \mu) I(\mathbf{X}_t \mu) \\ &= \max_{\mu} \mu^\top \left( T^{-1} n^{-1} \sum_{t=1}^T \mathbf{X}_t^\top W \mathbf{X}_t \right) \mu \quad (3) \\ &= \max_{\mu} \mu^\top \Theta \mu, \end{aligned}$$

where  $\Theta = T^{-1} n^{-1} \sum_{t=1}^T \mathbf{X}_t^\top W \mathbf{X}_t$  denotes a time average of the spatial correlation matrix. If  $W$  is symmetric, the optimal weight vector  $\mu$  may be extracted as before from a direct eigendecomposition of  $\Theta$ . Otherwise, and along the reasoning of Jombart et al. (2008), the optimal  $\mu$  may be found by the eigendecomposition of  $(2Tn)^{-1} \sum_{t=1}^T \mathbf{X}_t^\top (W + W^\top) \mathbf{X}_t$ . As in the ordinary PCA and its spatial variants the projections  $\phi_t$  of stPCA are obtained via multiplying the original data  $X_t$  with the time stable principal eigenvectors  $\mu$ .

### 3. Simulation

We present two simulations which compare the performance of the original PCA, its spatial variants and the novel stPCA approach to detect spatio-temporal factors. In a first step, we apply the distinct PCA variants to an artificial data-set of a single,

hidden and stable spatio-temporal factor, which is observed via three noisy variables and is obscured by three additional random noise variables. As in Wartenberg (1985) a ratio between the first eigenvalue and the sum of the absolute value of all eigenvalues is presented to reveal the sensitivity of these techniques in detecting the spatio-temporal factor. Obviously, a high ratio indicates that the respective PCA procedure correctly identifies the single predetermined factor present in the simulated data.

In a second simulation we apply PCA, its spatial variants and the stPCA approach to a data-set with two different spatio-temporal factors in order to learn the accuracy of these principal component approaches. Each factor exhibits a distinct and stable spatial-temporal pattern, which affects three noisy variables each and is furthermore masked by six additional random noise variables. We check whether the PCA variants identify the correct number of factors, how the PCA variants weigh the original variables in the computation of the projections and compare to what extent the diverse projections match the original factors.

We start the first simulation by generating a spatial structure  $S^{(1)}$  which takes the form of a square grid of size  $\sqrt{n} \times \sqrt{n}$ . The  $n$  observations  $S_{i \in \{1, \dots, n\}}^{(1)}$  follow a normal distribution with a mean depending on the grid's column index  $c$ :

$$S_{i,c}^{(1)} \sim \mathcal{N}(0, 1), \text{ for } c \leq \frac{C}{2} \quad S_{i,c}^{(1)} \sim \mathcal{N}(\delta, 1), \text{ for } c > \frac{C}{2},$$

where  $\delta$  defines an increment and  $C = \sqrt{n}$  denotes the number of columns. Consequently we simulate a patch, which differentiates between the left and right side of the grid by the expectation  $\mathbb{E}\left[S_{i,c}^{(1)} | c \leq \frac{C}{2}\right] = 0$  and  $\mathbb{E}\left[S_{i,c}^{(1)} | c > \frac{C}{2}\right] = \delta$ . This spatial structure is standardized and subsequently introduced as a constant in the AR(1) process of the spatio-temporal factor  $F_{i,t}^{(1)}$ . Switching to vector notation, the factor  $F_t^{(1)} = \text{vec}(F_{1,t}^{(1)}, \dots, F_{n,t}^{(1)})$  is generated via

$$F_t^{(1)} = S^{(1)} + 0.5F_{t-1}^{(1)} + \varepsilon_t, \quad (4)$$

where we define the error vector by  $\varepsilon_t \sim \mathcal{N}(0_n, 0.75I_n)$ . This simulated factor produces  $n \times t$  observations, which exhibit a stable spatial pattern over time defined by the size of the increment  $\delta$ . A high value of  $\delta$  will result in a more pronounced spatial pattern and, due to the standardization, will not automatically increase the factor's variance, which is instead defined by the coefficient and error vector in the AR(1) process (4).

In our simulation the standardized spatial factor affects  $p_1 = 3$  dependent variables  $X_{t,p_1 \in \{1, \dots, 3\}}$ , which are defined by the sum of the factor  $F_t^{(1)}$  and an individual AR(1) noise process  $Z_{t,p_1}$ :

$$\mathbf{X}_{t,p_1} = \mathbf{F}_t^{(1)} + \mathbf{Z}_{t,p_1}.$$

The noise process  $\mathbf{Z}_{t,p_1}$  differentiates the three variables  $\mathbf{X}_{t,p_1}$  via its error component:

$$\mathbf{Z}_{t,p_1} = 0.5\mathbf{Z}_{t-1,p_1} + \varepsilon_{Z,t},$$

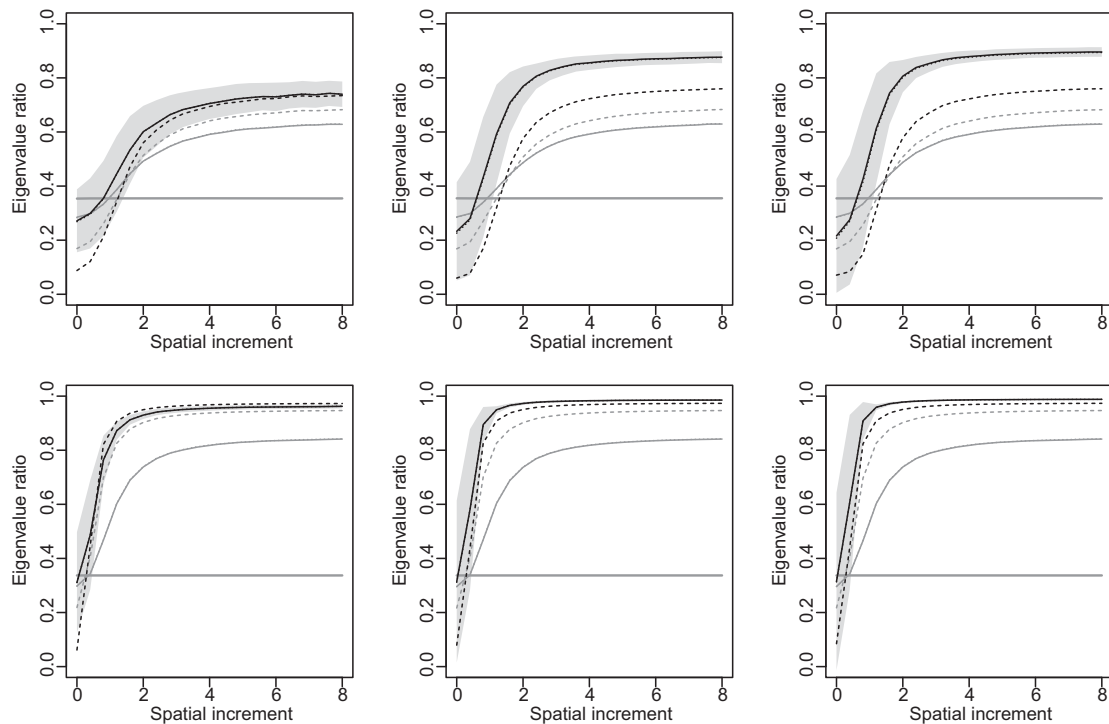
where  $\varepsilon_{Z,t} \sim \mathcal{N}(0_n, 0.75I_n)$  denotes white noise. Consequently, we separate the variables by their specific errors drawn from the same normal distribution.

Apart from the dependent variables  $X_{t,p_1}$ , we also add  $p_2 = 3$  random noise variables  $X_{t,p_2 \in \{4, \dots, 6\}}$ , which are independent of the factors and follow an ordinary AR(1) process:

$$\mathbf{X}_{t,p_2} = 0.5\mathbf{X}_{t-1,p_2} + \varepsilon_{X,t},$$

where  $\varepsilon_{X,t} \sim N(\mathbf{0}_n, 0.75\mathbf{I}_n)$  denotes white noise. Hence these three variables possess the same mean and variance as the spatio-temporal factor and interfere with its disclosure.

We run this simulation in two settings to cover small and large  $n$  applications. At first we set  $n_1 = 49$  and  $t_1 \in \{5, 50, 100\}$ . This specification allows for all possible combinations of  $n$  and  $t$ :  $t < n$ ,  $t \approx n$  and  $t > n$ . The same holds for the second setting, where  $n_2=400$  and  $t_2 \in \{40, 400, 800\}$ . In order to observe the impact of the variable scale in  $X$  and the specific weight matrix  $W$ , we apply three different combinations of  $X$  and  $W$  which the authors of the spatial PCA variants have proposed in their original paper described above. The spatial increment is evaluated by increasing  $\delta$  gradually via steps of 0.4 in the interval  $[0,8]$  for all combinations of  $n$  and  $t$ . At  $\delta=0$  obviously no spatial factor is produced, as this particular parametrization describes an i.i.d. scenario. Finally, we run each combination of the parameters 1,000 times and present the respective mean ratio between the first eigenvalue and the sum of all eigenvalues in Figure 1.



**Figure 1.** Mean ratio (with standard deviation for temporal sPCA) of the first eigenvalue to the sum of all eigenvalues as assigned by PCA (thick solid grey line), sPCA (solid grey line), MSC (dashed grey line), Global Structure PCA (dotted grey line), temporal sPCA (solid black line), temporal MSC (dashed black line) and temporal Global Structure PCA (dotted black line) for  $n_1=49$  (first row) with  $t_1=5$  (left graph),  $t_1=50$  (middle graph) and  $t_1=100$  (right graph), and  $n_2=400$  (second row) with  $t_2=40$  (left graph),  $t_2=400$  (middle graph) and  $t_2=800$  (right graph).

We observe that PCA present a constant ratio between the largest eigenvalue and the sum of all eigenvalues. This ratio remains unaffected by an increase in the spatial increment  $\delta$  and hence PCA fails to clearly identify the increasingly pronounced spatio-temporal factor. On the other hand all spatial PCA variants, and especially the MSC approach, gain strongly from an increase in  $\delta$ . The initial ratio at  $\delta=0$  is increased more than twofold at  $\delta=8$  and the spatial PCA variants cause higher ratios than the original PCA for  $\delta \geq 1.2(n_1 = 49)$ , respectively  $\delta \geq 0.4(n_2=400)$ . Consequently we can verify the results of Wartenberg (1985), Thioulouse et al. (1995) and Jombart et al. (2008), and observe that extending PCA by a spatial component improves the sensibility of the spatial PCA variants to identify a spatial factor.

However, as can be observed in Figure 1, the general stPCA approach is even more responsive to an increase in  $\delta$  than the spatial PCA variants. For example at  $\delta \geq 0.8(n_1=49)$ , respectively  $\delta \geq 0.4(n_2=400)$  the ratio reported by the temporal sPCA variant is larger than the ratio of any other PCA variant including the purely spatial PCA variants. In detail, all stPCA implementations not only reports larger ratios for any given  $n$  and  $t$  than the spatial PCA variants at non-negligible levels of  $\delta$ , but also exploit an increase in  $n$  much stronger. Furthermore, only the stPCA implementations makes use of the time dimension and reports higher ratios for an increase in  $t$ . Unsurprisingly the solely spatial PCA variants do not gain on such an increase in  $t$ , but only on an increase in  $n$  and this superior performance of the stPCA variants also holds if the spatio-temporal factor consists of a spatial trend instead of a patch.

As stated before, any spatial principal component approach will also try to identify local structure and report this structure as a negative eigenvalue. In the current simulation, which does not explicitly include local structure, this feature appears twice. At first, the stPCA and the purely spatial PCA variants perform worse than the original PCA on non-spatial or only slightly spatial data, as can be observed by ratio which correspond to  $\delta=0$ . Second, this search for local structure causes the stPCA approach to possess a pronounced standard deviation as depicted in Figure 1, which however disappears as the spatial increment grows.

In a second simulation we apply PCA, its spatial and stPCA variants to a dataset with two different spatio-temporal factors in order to learn the accuracy of these principal component approaches. Hence, we extend the preceding simulation by an additional spatio-temporal factor, which is based on the spatial structure  $S^{(2)}$ :

$$S_{i,r}^{(2)} \sim N(0, 1), \text{ for } r \leq \frac{R}{2} \quad S_{i,r}^{(2)} \sim N(\delta, 1), \text{ for } r > \frac{R}{2},$$

where  $\delta$  describes the spatial increment,  $r$  denotes a row indicator and  $R = \sqrt{n}$  indicates the number of rows. Consequently the spatial structure  $S^{(2)}$  describes a spatial patch, which differentiates between the upper and lower half of the grid.

The resulting structure defines the spatial distribution of the second spatio-temporal factor  $F_{i,t}^{(2)}$  by serving as a constant in the respective AR(1) process:

$$\mathbf{F}_t^{(2)} = \mathbf{S}^{(2)} + 0.5\mathbf{F}_{t-1}^{(2)} + \varepsilon_t,$$

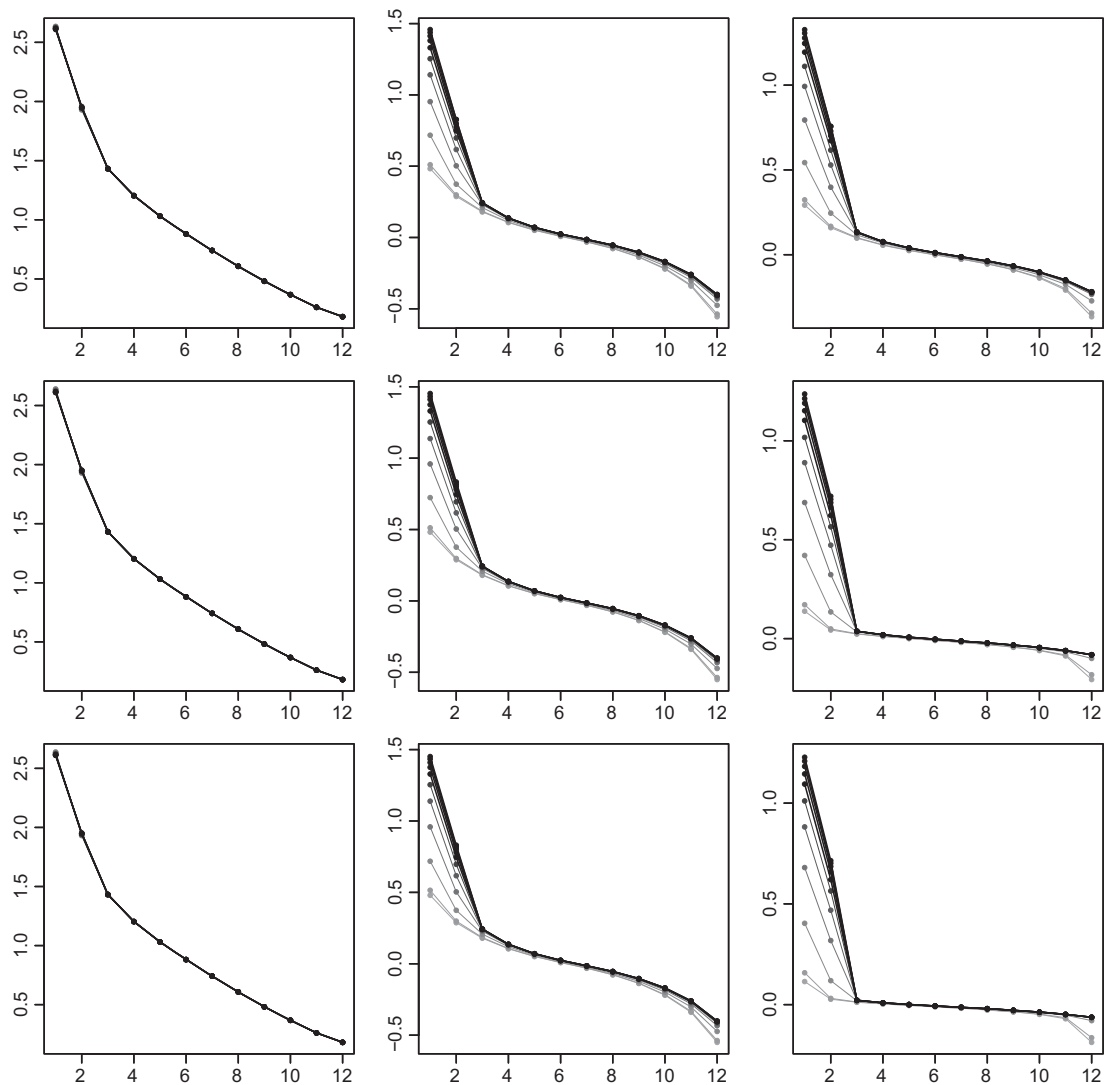
where  $\varepsilon_t$  is defined as above. As in the preceding simulation, the hidden spatio-temporal factor  $F_t^{(2)}$  is observed via three noisy variables, which differ, as before, in their error components. In this context, we set the white noise of variables affected by the first factor to  $\varepsilon_{Z,t}^{(1)} \sim N(0_n, 0.375\mathbf{I}_n)$  and the error vector of the variables defined by the second factor to  $\varepsilon_{Z,t}^{(2)} \sim N(0_n, 1.125\mathbf{I}_n)$ .



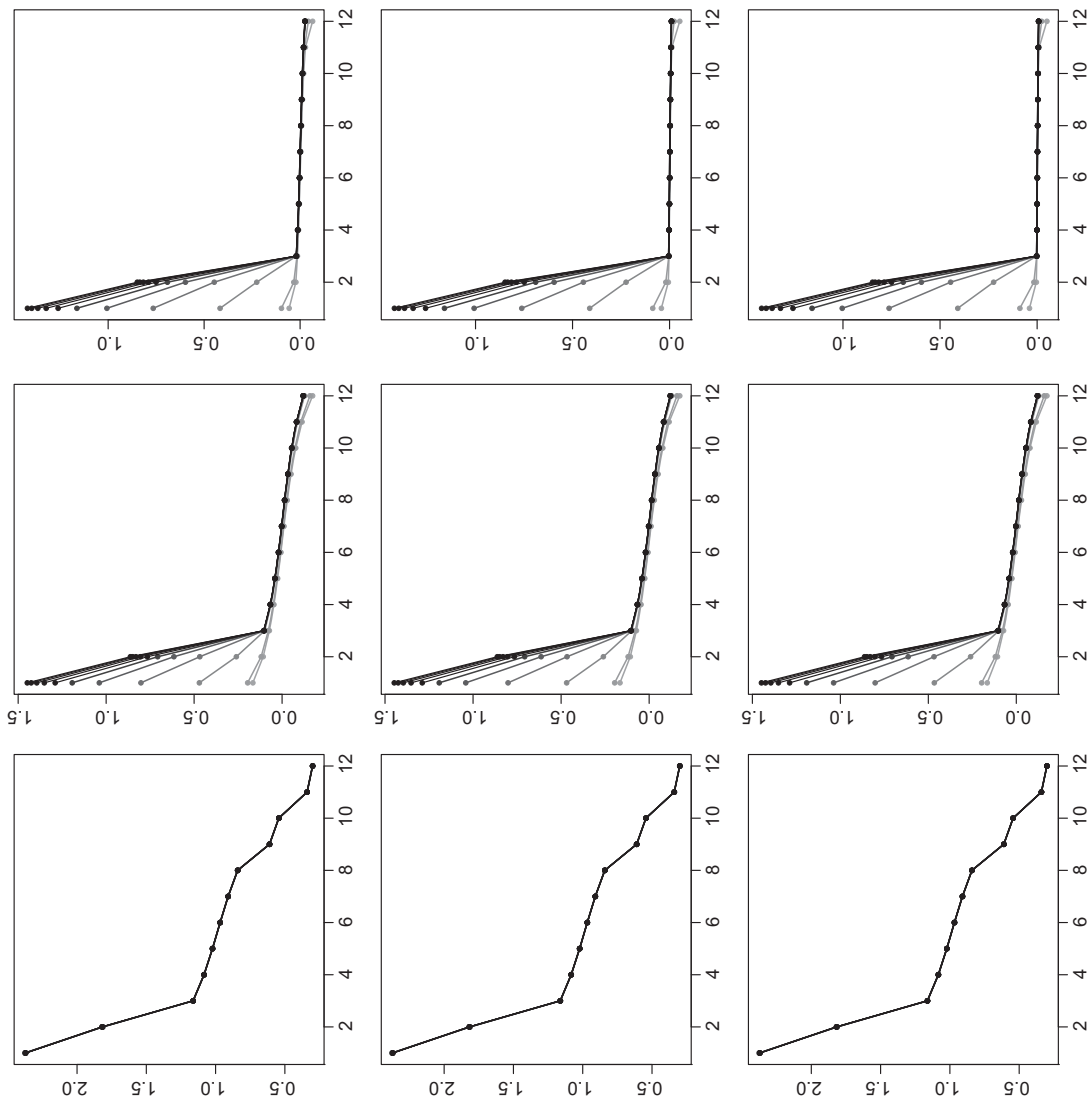
Furthermore, we add three additional i.i.d. noise processes and consequently observe three variables affected by the spatio-temporal factor  $F_t^{(1)}$ , three variables influenced by spatio-temporal factor  $F_t^{(2)}$  and six additional noise variables, which complicate the disclosure of the two spatio-temporal factors.

As before we run this simulation in two settings to account for small ( $n_1=49$ ) and large ( $n_2=400$ ) data-sets and allow for varying time dimensions:  $t_1 \in \{5, 50, 100\}$  and  $t_2 \in \{40, 400, 800\}$ . We increase the spatial increment gradually in the interval  $[0,8]$  to observe its effect, make use of the aforementioned spatial weight matrices and transformations of  $X$ , and run each combination of parameters 1,000 times.

We begin our inspection of the simulation results by assessing the power of the diverse principal component approaches to identify the correct number of factors. Figures 2 and 3 present modified scree plots for PCA, sPCA and temporal sPCA in



**Figure 2.** Scree plots for PCA (first column), sPCA (second column) and temporal sPCA (third column) depicted for  $n_1=49$ , the time frame  $t_1=5$  (first row),  $t_1=50$  (second row),  $t_1=100$  (third row) and the spatial increments  $\delta \in [0,8]$  indicated by an increase in the grey strength.



**Figure 3.** Scree plots for PCA (first column), sPCA (second column) and temporal sPCA (third column) depicted for  $n_2=400$ , the time frame  $t_2=40$  (first row),  $t_2=400$  (second row),  $t_2=800$  (third row) and the spatial increments  $\delta \in [0, 8]$  indicated by an increase in the grey strength.

which the mean eigenvalues are interpolated to allow for several spatial increments  $\delta$  to be shown in the same graph.

At first, we observe that the distinct principal component approaches return different eigenvalues. The original PCA does not react to an increase in  $\delta$ , as this non-spatial approach presents nearly the same eigenvalues for all levels of  $\delta$ . On the other hand, sPCA and the temporal sPCA do respond to an increase in the spatial increment. Especially, the two largest and the smallest eigenvalues increase with  $\delta$  and their reaction is amplified by more observations.

At  $n_1=49$  the scree plot of PCA presents a slight decrease in the gradient starting at the third eigenvalue and consequently indicates the presence of two factors. These factors arise due to the constant in the simulation's AR(1) process. At  $n_2=400$ , PCA presents three obvious changes in the gradient resulting in two or more factors and consequently the scree plot does not clearly indicate the correct number of factors.

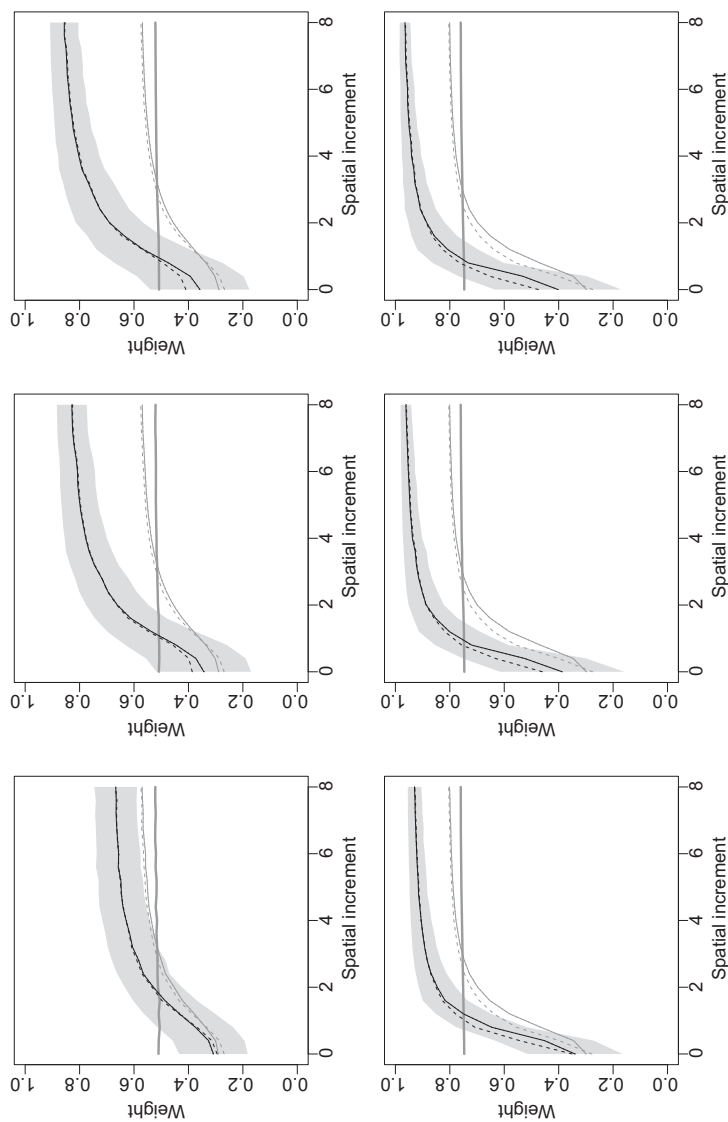
In contrast sPCA and temporal sPCA react explicitly to an increase in  $\delta$ . At  $\delta \leq 0.8$ , the eigenvalues returned by sPCA do not indicate any obvious change in the gradient, but rather suggest a smooth curve. Only at  $\delta > 0.8$  one can make out a clear change in the curvature after the second eigenvalue and consequently sPCA identifies the two spatio-temporal factors. This finding is even more apparent as  $\delta$  and  $n$  are increased. However, sPCA also returns large negative eigenvalues, which erroneously indicate a high level of local structure.

The application of temporal sPCA results in eigenvalues, which possess a similar structure as sPCA, but which is much more pronounced. At the low level of  $\delta \leq 0.8$  the respective eigenvalues also suggest a smooth curve without any clear indication of the number of factors. At  $\delta > 0.8$  stPCA increasingly indicates the presence of the two factors. However, the difference between the second and third eigenvalue is much wider in the case of temporal sPCA than ordinary sPCA, e.g. the ratio at  $\delta=4$ ,  $n_2=400$  and  $t_2=400$  for temporal sPCA (1:151.071) surpasses the ratio of sPCA (1:7.163) more than 20-fold and consequently temporal sPCA indicates the two factors more evidently than sPCA. Furthermore, temporal sPCA indicates the presence of local structure only at very low levels of  $\delta$  and the clarity of its scree plot is not only amplified by an increase in the spatial increment  $\delta$  and the sample size  $n$ , but also by the time dimension  $t$ .

The scree plots of the spatial or spatio-temporal variants of MSC or Global Structure present a very similar pattern in the respective eigenvalues and we consequently do not present them here.

In a second evaluation step, we examine the eigenvectors which correspond to the two largest eigenvalues. These specify the weights in the linear combination of the variables to obtain the projections. In our simulated data the first spatio-temporal factor affects only the first three variables, whereas the second spatio-temporal factor defines the fourth, fifth and sixth variables. Consequently, we would expect the first eigenvector to carry large weights on the first, second and third variable and the second eigenvector to accentuate the fourth, fifth and sixth variable. Furthermore, an optimal principal component approach would assign zero weights to all the other variables, as these do not contain any information on the two spatio-temporal factors.

Figure 4 presents the mean weight the two eigenvectors assign to the respective variables, that is the mean of the weight assigned by the first eigenvector to the first, second and third variable and the weight given by the second eigenvector to the fourth, fifth and sixth variable. Obviously a high ratio indicates that the



**Figure 4.** Mean weight (with standard deviation for temporal sPCA assigned by PCA (thick solid grey line), sPCA (solid grey line), MSC (dashed grey line), Global Structure PCA (dotted grey line), temporal sPCA (solid black line), temporal MSC (dashed black line) and temporal Global Structure PCA (dotted black line) in the two eigenvectors to the respective variables for  $n_1 = 49$  (first row) with  $t_1 = 5$  (left graph),  $t_1 = 50$  (middle graph) and  $t_1 = 100$  (right graph), and  $n_2 = 400$  (second row) with  $t_2 = 40$  (left graph),  $t_2 = 400$  (middle graph) and  $t_2 = 800$  (right graph)).

respective principal component approach does not erroneously highlight variables, which do not incorporate any information on the spatio-temporal factors.

At first we observe, that the original PCA attributes substantial weights to the respective variables, and the weights increase with  $n$ . However, PCA does not react to an increase in either the spatial increment  $\delta$  or the time dimension  $t$  and produces constant weights in this respect. Contrary to this indifference the spatial variants do react to an increase in  $\delta$  and the corresponding weights surpass the original PCA's weights at  $\delta \geq 3.2$  for  $n_1=50$ , respectively  $\delta \geq 2.8$  for  $n_2=400$ . A further increase in the spatial increment widens the margin between PCA and the spatial PCA variants even more. But the spatial PCAs do not react to an increase in  $t$  and due to their cross-sectional approach they fail to exploit the serial correlation in order to improve the weights even further.

This is instead accomplished by the new stPCA procedure. The mean weights presented by this spatio-temporal technique improve upon an increase of either the spatial increment  $\delta$ , the number of observations  $n$  and also upon an increase in the time dimension  $t$ . The stPCA approach exceeds the weights of the alternative non-temporal approaches at a spatial increment between  $\delta \geq 2$  ( $n_1=49$  and  $t_1=5$ ) and  $\delta \geq 0.8$  ( $n_2=400$  and  $t_2=800$ ), and this difference is amplified as  $t$  is increased. For example, at a parametrization of  $n_2=400$ ,  $t_2=400$  and  $\delta=4$  the mean weight of the eigenvectors presented by the stPCA approach exceeds the weights of second-best procedure, namely the solely spatial PCA variants, by 19.5%.

In a third evaluation step, we directly compare the spatio-temporal factors with its projections. In detail, we compute the mutual information  $MI(.,.)$  between the artificially created factor values  $F_t^{(1,2)}$  and the projections  $\widehat{F}_t^{(1,2)}$  identified by the diverse principal component approaches. The computation is based on the two largest positive eigenvalues of every principal component approach. Figure 5 reports on the average mutual information over time, which we define as

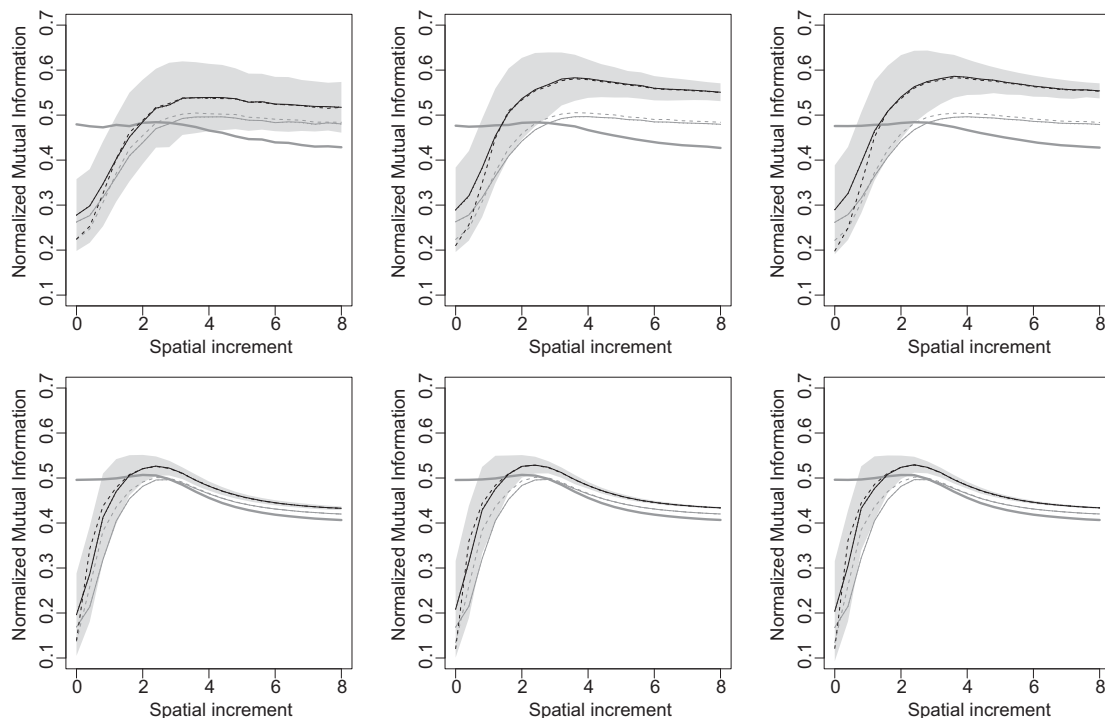
$$\frac{1}{2T} \sum_{t=1}^T \max \left\{ MI(\widehat{\mathbf{F}}_t^{(1)}, \mathbf{F}_t^{(1)}) + MI(\widehat{\mathbf{F}}_t^{(2)}, \mathbf{F}_t^{(2)}), MI(\widehat{\mathbf{F}}_t^{(1)}, \mathbf{F}_t^{(2)}) + MI(\widehat{\mathbf{F}}_t^{(2)}, \mathbf{F}_t^{(1)}) \right\}$$

to address issues arising from factor switching.

It might first be noted from Figure 5, that all PCA approaches struggle with a more and more pronounced spatial patch. All techniques report a decrease in the mutual information, if the spatial increment  $\delta$  is increased up to its maximum. This observation can be accredited to the particular simulation setting in which the spatial patch is obscured by normally distributed noise. Consequently this error component prohibits a clearer identification of the increasing spatial structure, as the diverse PCA procedures weight the key variables and hence their accompanying noise stronger, as  $\delta$  is increased. However, this effect is not observed if a spatial trend is modelled instead of a patch and might ultimately be explained by the particular spatial structure.

The original PCA approach explains nearly half of the entropy of the factor values. It performs only marginally better if the number of observation  $n$  is increased and does not react to the time dimension. As explained above its performance worsens as the spatial increment  $\delta$  is raised.

The spatial variants of PCA surpass its ordinary cousin at  $\delta \geq 2.8$ , but its performance depends on the level of the spatial increment. An increase in  $\delta$  at low levels will strengthen its performance up to a maximum and a further increase in  $\delta$  will afterwards worsen the mutual information between the factor values and its



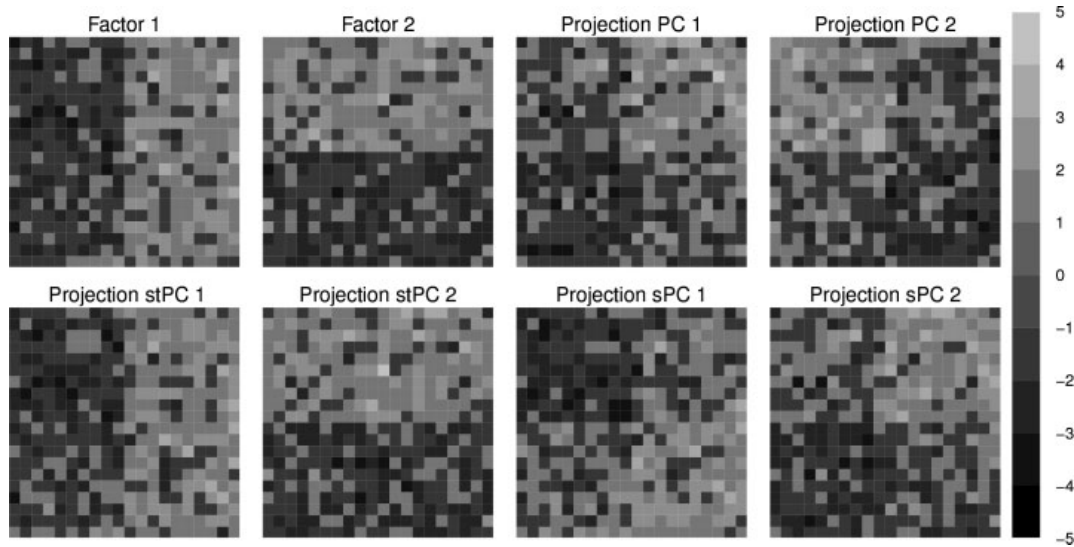
**Figure 5.** Normalized mutual information (with standard deviation for temporal sPCA) between the simulated factor values and the factor scores indicated by PCA (thick solid grey line), sPCA (solid grey line), MSC (dashed grey line), Global Structure PCA (dotted grey line), temporal sPCA (solid black line), temporal MSC (dashed black line) and temporal Global Structure PCA (dotted black line) for  $n_1=49$  (first row) with  $t_1=5$  (left graph),  $t_1=50$  (middle graph) and  $t_1=100$  (right graph), and  $n_2=400$  (second row) with  $t_2=40$  (left graph),  $t_2=400$  (middle graph) and  $t_2=800$  (right graph).

projections. The same observation holds for an increase in the number of observations  $n$  and as before the spatial PCA approaches do not respond to an increase in  $t$ .

In contrast to the ordinary and spatial PCA variants the stPCA approach exploits an increase in the time dimension and the corresponding mutual information exceeds the alternative ones at  $\delta \geq 2(n_1=49)$ , respectively  $\delta \geq 1.6(n_2=400)$ . However, like the spatial PCA approaches its performance on an increase in  $\delta$  and  $n$  is mixed. On low levels of the spatial increments the mutual information rises steeply, if either  $\delta$  or  $n$  are increased, but on high levels of  $\delta$  a further increase of either the spatial increment or the number of observation worsen the mutual information of the stPCA procedure.

However, the implementations of stPCA outperform the ordinary and spatial PCA approaches on all parameter values apart from very low levels of  $\delta$  and the difference is significant. Obviously the largest gain is made on small  $n$ , high  $t$  samples, whereas the additional value for large  $n$  data seems less pronounced.

In order to visualize the observed difference in the mutual information, Figure 6 presents the simulated factor values and corresponding projections returned by the distinct principal component approaches for  $n_2=400$ ,  $t_2=40$  and  $\delta=4$  at an exemplary point in time. In order to get a better contrast in the graph, we added, respectively subtracted, 1 from the original values and projections.



**Figure 6.** Simulated factor values and projections by the respective principal component approaches for  $n_2=400$ ,  $t_2=40$  and  $\delta=4$  for an exemplary point in time.

The original factor values present the aforementioned patch between either the left and right or upper and lower half of the grid. Whereas the single projections by stPCA do not resemble the original factors perfectly, as a whole they clearly present the same basic structure of two diverse spatial structures. This observation does not hold for either the original PCA or its spatial variants, as both techniques present a spatial structure which clearly discriminates between four disjunct parts in every corner, but fail to present the original spatial structure of one north–south and one west–east patch. Consequently and as noted above both techniques incorporate lower mutual information values.

#### 4. Application on Urbanism and Economic Deprivation

Apart from case-by-case specific impact factors, criminological theory and research based on data for areal units have persistently and mainly in the USA identified two broad dimensions of social structure that have proven to be robust predictors of violent crime rates: (1) economic well-being/relative deprivation and (2) population structure/urbanism (McCall et al., 2010). In a recent study, Messner et al. (2013) have confirmed the explanatory power of these two factors with regard to German assault and robbery rates collected in 413 geographic-administrative districts called ‘Kreise’ (counties) and aggregated over the years 2005–2007.

In this study, the factor Urbanism was constructed via PCA performed on four indicators: (1) the average population size across the 3-year period; (2) population density, i.e. population per square kilometre; (3) the proportion of the workforce employed in agriculture or forestry and (4) percent of the foreign-born population, which tends to be concentrated in urban areas.

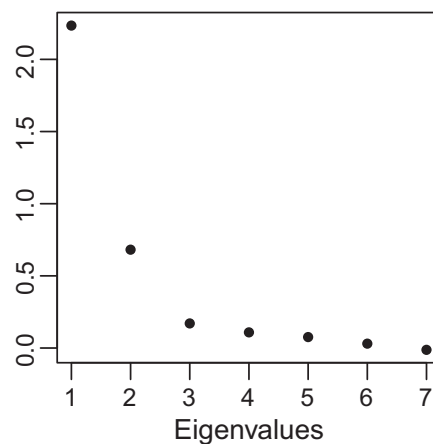
The factor economic well-being/deprivation was constructed via PCA performed on five indicators: (1) the percentage of the civilian labour force that is unemployed; (2) the percentage of those persons who receive social assistance; (3) average monthly household income; (4) the proportion of households receiving housing assistance and (5) a measure of high school dropouts.

In our own analysis, we have removed the fifth indicator from the economic deprivation factor, since in a subsequent study using roughly the same database. Thome & Stahlschmidt (2013) have presented theoretical arguments and empirical evidence demonstrating that high school attendance should be included, together with three additional indicators, in another factor labelled ‘disintegrative individualism’. Since our present article does not deal with theoretical issues we have limited our analysis, i.e. the application of the stPCA approach, to the two factors most commonly applied in criminological research based on areal units, economic deprivation and urbanism. Furthermore, we exclude the percentage of foreign-born population, as this variable is closely related to both factors, urbanism and relative deprivation, and therefore interferes with our aim to construct two clearly distinguished factors.

Whereas Messner et al. (2013) averaged their data over a 3-year period to generate the factor scores in a cross-sectional approach, stPCA allows to include every year separately in the analysis. Hence, the presented projections are based on more data points and take into account serial correlation.

Figure 7 presents the scree plot generated by stPCA and based on a row-weighted binary spatial weight matrix indicating direct neighbouring counties and standardized data. An aggregation of the seven indicators into two factors seems reasonable, as the first two eigenvalues clearly stand out if compared to the remaining ones.

The corresponding variable weights for each factor are detailed in Table 1. The variables are grouped as expected by criminological theory. The unemployment



**Figure 7.** Scree plot for the indicators on urbanism and economic deprivation.

**Table 1.** The first and second eigenvector corresponding to the two largest eigenvalues

	Eco. deprivation	Urbanism
Unemployment rate	0.563	0.222
Social welfare rate	0.446	0.339
Household income	-0.465	0.293
Housing allowance rate	0.503	-0.092
Population size	-0.013	0.396
Population density	-0.060	0.554
Employment in agriculture	0.105	-0.527

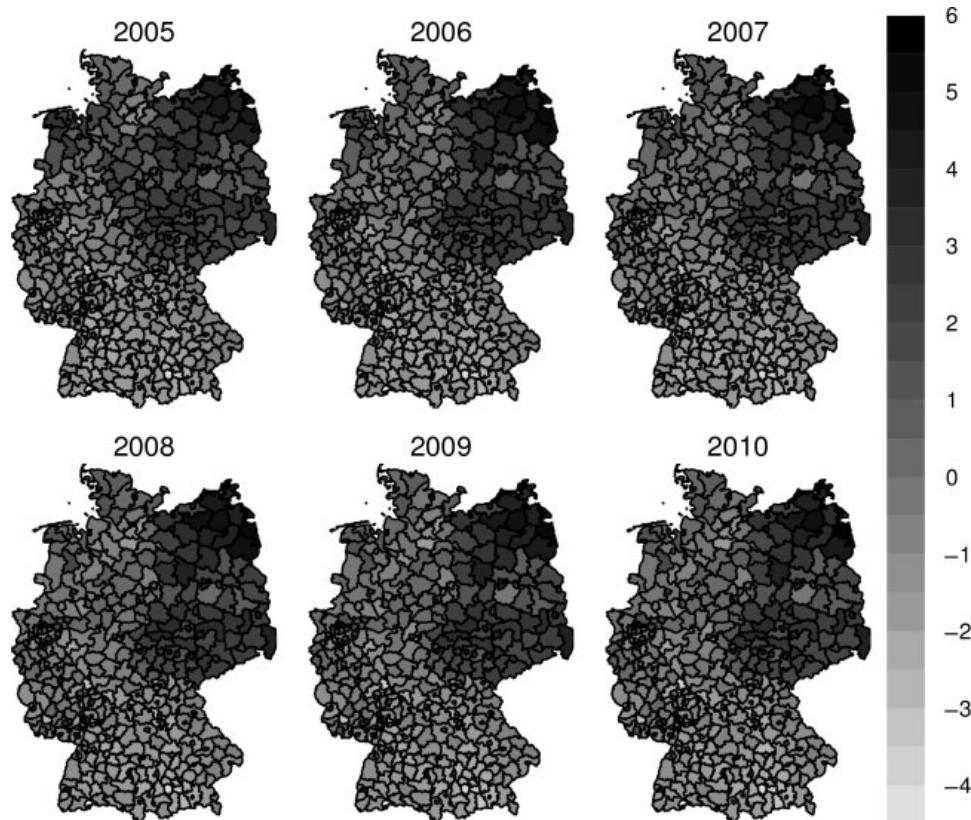


rate, social welfare rate, household income and housing allowance rate together describe a factor interpreted as relative economic deprivation, whereas population size, population density and employment in agriculture jointly specify the level of urbanism. Most of the single variables can clearly be attributed to one of the two factors. Only the social welfare rate takes on similar weights in both factors since the need for such payments arises less often in rural areas.

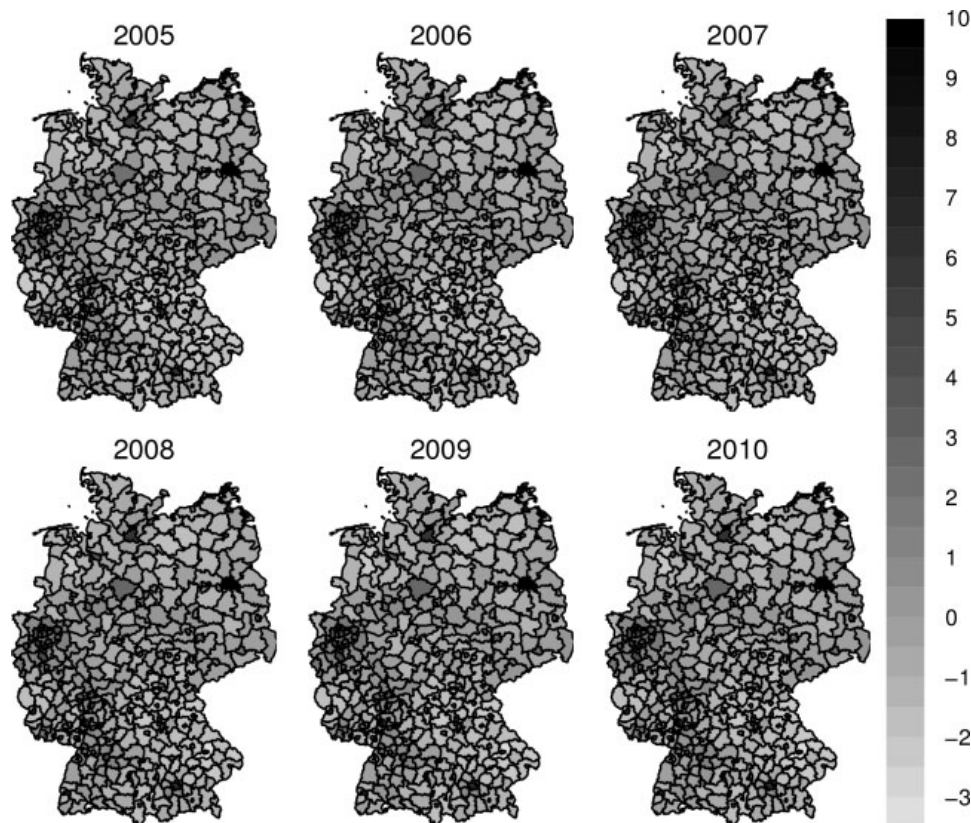
Finally, [Figures 8](#) and [9](#) display the projections resulting from the application of the weights to the single counties. The projections of the economic deprivation factor depict two spatial patterns. First, there remains a clear disparity between East and West Germany, as counties of the former German Democratic Republic possess less economic means than their western counterparts. Second, inside these two blocks there arises a North–South divide as the southern part of Germany has, during the last two decades, achieved a more advanced, higher-level balance between traditional and modernized sectors of economic development. These two patterns remain stable over the inspected time horizon. This stationarity over time is also observed in case of the urbanism factor. The corresponding projections clearly highlight the big urban hubs of Berlin, Munich, Hamburg or the Rhine–Ruhr metropolitan region and the sparsely populated north–eastern part of Germany.

## 5. Conclusion

The analyses of purely spatial data are confronted with the implied peculiarities of such data, namely spatial correlation and heterogeneity. Consequently, such



**Figure 8.** Projections of the factor ‘Economic Deprivation’ on German Kreise. Digital map provided and copyrighted by GeoBasis–DE/BKG 2013.



**Figure 9.** Projections of the factor ‘Urbanism’ on German Kreise. Digital Map provided and copyrighted by GeoBasis–DE/BKG 2013.

analyses entail a need for a high amount of data in order to obtain reliable estimates of any parameter of interest. But due to natural limits, which are especially obvious in the case of areal data, sample sizes cannot be enlarged indefinitely over space. However, a feasible solution in dealing with this problem can be forged ahead by the extension of such data over time resulting in spatio-temporal analyses. This approach requires the transformation of well-known instruments and techniques from the i.i.d. or spatial environment to the spatio-temporal one.

To our knowledge, stPCA offers a first attempt to transfer the original PCA to the spatio-temporal realm of geographical and serial correlation. The proposed technique allows for dimension reduction on the attribute space while preserving the geographical and temporal structure and it offers a promising approach to generate consistent factors from spatio-temporal data. It differs from any explicit factor modelling approach by its algorithmic nature, which can be viewed as a welcomed feature or a drawback depending on substantive issues given in a particular research context.

In any case stPCA possesses a superior performance in terms of sensibility to detect and of accuracy to disclose spatio-temporal factors if compared to the original PCA and the proposed spatial variants thereof. Especially the original PCA lacks power to correctly identify spatial factors and its spatial variants fail to exploit any serial correlation to improve their results due to their static nature. Furthermore stPCA is much faster than its archetypes, as the time-consuming eigendecomposition has to be calculated only once instead of  $t$  times.

As PCA and the spatial variants have to be applied separately for every  $t$ , they are prone to sign and factor switching over  $t$ . This behaviour complicates an analysis, as such instances have to be detected and resolved before the projections can be analysed or supplied for further use. stPCA avoids such issues, as it presents time stable weights for the linear combinations and consequently allows for a direct and consistent interpretation of these values. Summing up, stPCA seems to be better suited than its static forerunners to address the specific requirements arising from spatio-temporal analyses.

However, the projections resulting from stPCA obviously depend on the variable scale and on the appropriateness of the spatial weight matrix, although our simulation indicates only minor effects of the diverse parametrisations. Furthermore the performance of stPCA might be amplified by a suitable orthogonal or oblique rotation, which will consequently restrict the algorithmic nature of stPCA and increase the researcher's influence.

Finally, we would like to mention two modifications of stPCA, from which certain applications might benefit. First, the spatial weight matrix could be understood as time dependent and, upon availability, a time-specific  $W_t$  could be introduced into the optimisation (3) to refine the technique. Second, stPCA could also be developed into an adaptive approach, in which the projections for  $t$  are not obtained via the time average of the spatial covariance matrix over the whole time frame  $T$ , but only over the interval  $[t - t^*, t + t^*]$ , where  $t^* < T/2$  denotes a tuning parameter. An appropriate weighting schema over this interval could furthermore enhance the flexibility. This moving window approach resides between the spatial variants of PCA and stPCA, as it exploits serial correlation over time, but foregoes any computational advantage, as an eigendecomposition has to be obtained at every  $t$  separately. Yet this adaptive stPCA might allow for time trends in the data and consequently account for not only the spatial peculiarities in spatial-temporal data, but also for the temporal ones.

## Funding

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 'Ökonomisches Risiko', Humboldt-Universität zu Berlin, and IRTG 1792 'High Dimensional Non Stationary Time Series', Humboldt-Universität zu Berlin, is gratefully acknowledged.

## References

- Choi, J., Lawson, A. B., Cai, B., Hossain, M. M., Kirby, R. S. & Liu, J. (2012) A Bayesian latent model with spatio-temporally varying coefficients in low birth weight incidence data, *Statistical Methods in Medical Research*, 21, 445–456.
- Demšar, U., Harris, P., Brunson, C., Fotheringham, A. S. & McLoone, S. (2013) Principal component analysis on spatial data: an overview, *Annals of the Association of American Geographers*, 103, 106–128.
- Fotheringham, A. S., Brunson, C. & Charlton, M. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, Wiley.
- Gelman, A. & Hill, J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, Cambridge University Press.
- Härdle, W. K. & Simar, L. (2012) *Applied Multivariate Statistical Analysis* (3rd ed.). Berlin, Springer.
- Hogan, J. W. & Tchernis, R. (2004) Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data, *Journal of the American Statistical Association*, 99, 314–324.

- Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24, 417–441.
- Jolliffe, I. T. (2002) *Principal Component Analysis*. New York, Springer.
- Jombart, T., Devillard, S., Dufour, A.-B. & Pontier, D. (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method, *Heredity*, 101, 92–103.
- Karhunen, K. (1947) *Über lineare Methoden in der Wahrscheinlichkeitsrechnung* [On linear methods in probability and statistics]. Helsinki, Suomalainen Tiedeakatemia
- Kroonenberg, P. M. & de Leeuw, J. (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms, *Psychometrika*, 45, 69–97.
- Lawson, A. B., Song, H.-R., Cai, B., Hossain, M. M. & Huang, K. (2008) Space-time latent component modeling of geo-referenced health data, *Statistics in Medicine*, 29, 2012–2027.
- Loève, M. (1948) Fonctions Aléatoires de second order, in: P. Lévy (ed), *Processus Stochastique et Movement Brownien*, Paris, Hermann.
- Lorenz, E. (1956) Empirical orthogonal functions and statistical weather prediction. Statistical Forecasting Project, Scientific Report 1.
- McCall, P. L., Land, K. C. & Parker, K. F. (2010) An empirical assessment of what we know about structural covariates of homicide rates: a return to a classic 20 years later, *Homicide Studies*, 14, 219–243.
- Messner, S. F., Teske, R. H. C., Baller, R. D. & Thome, H. (2013) Structural covariates of violent crime rates in Germany: exploratory spatial analyses of Kreise, *Justice Quarterly*, 30, 1015–1041.
- Moran, P. A. P. (1950) Notes on continuous stochastic phenomena, *Biometrika*, 37, 17–23.
- Pearson, K. (1901) On lines and planes of closest fit to systems of points in space, *Philosophical Magazine, Series 6*, 2, 559–572.
- Richman, M. B. (1986) Rotation of principal components, *Journal of Climatology*, 6, 293–335.
- Spearman, C. (1904) “General Intelligence,” objectively determined and measured, *The American Journal of Psychology*, 15, 201–293.
- Thioulouse, J., Chessel, D. & Champely, S. (1995) Multivariate analysis of spatial patterns: a unified approach to local and global structures, *Environmental and Ecological Statistics*, 2, 1–14.
- Thome, H. & Stahlschmidt, S. (2013) Ost und West, Nord und Süd: Zur räumlichen Verteilung und theoretischen Erklärung der Gewaltkriminalität in Deutschland, *Berliner Journal für Soziologie*, 23, 441–470.
- Tzala, E. & Best, N. (2007) Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality, *Statistical Methods in Medical Research*, 17, 97–118.
- Wartenberg, D. (1985) Multivariate spatial correlation: a method for exploratory geographical analysis, *Geographical Analysis*, 17, 263–283.



## PORTFOLIO DECISIONS AND BRAIN REACTIONS VIA THE CEAD METHOD

PIOTR MAJER

HUMBOLDT-UNIVERSITÄT ZU BERLIN

PETER N. C. MOHR · HAUKE R. HEEKEREN

FREIE UNIVERSITÄT BERLIN

WOLFGANG K. HÄRDLE

HUMBOLDT-UNIVERSITÄT ZU BERLIN  
SINGAPORE MANAGEMENT UNIVERSITY

Decision making can be a complex process requiring the integration of several attributes of choice options. Understanding the neural processes underlying (uncertain) investment decisions is an important topic in neuroeconomics. We analyzed functional magnetic resonance imaging (fMRI) data from an investment decision study for stimulus-related effects. We propose a new technique for identifying activated brain regions: cluster, estimation, activation, and decision method. Our analysis is focused on clusters of voxels rather than voxel units. Thus, we achieve a higher signal-to-noise ratio within the unit tested and a smaller number of hypothesis tests compared with the often used General Linear Model (GLM). We propose to first conduct the brain parcellation by applying spatially constrained spectral clustering. The information within each cluster can then be extracted by the flexible dynamic semiparametric factor model (DSFM) dimension reduction technique and finally be tested for differences in activation between conditions. This sequence of Cluster, Estimation, Activation, and Decision admits a model-free analysis of the local fMRI signal. Applying a GLM on the DSFM-based time series resulted in a significant correlation between the risk of choice options and changes in fMRI signal in the anterior insula and dorsomedial prefrontal cortex. Additionally, individual differences in decision-related reactions within the DSFM time series predicted individual differences in risk attitudes as modeled with the framework of the mean-variance model.

Key words: risk, risk attitude, fMRI, decision making, neuroeconomics, semiparametric model, factor structure, brain imaging, spatial clustering, inference on clusters, CEAD method.

**JEL Classification** C3, C6, C9, C14, D8

### 1. Introduction

Economic decision making takes place when, for example, an individual buys beverages in a supermarket, purchases a car, or chooses an investment fund. Some of these choices are made when the outcome is uncertain and hard to anticipate, which is particularly true for an investment decision. The decision-making process builds on different mechanisms such as representation and integration of relevant evidence for and a comparison process of different choice options. This mechanism has attracted considerable attention in many different fields, from cognitive psychology, behavioral economics to neuroscience, see, e.g., [Glimcher and Fehr \(2013\)](#). Economic decisions are usually explained in a value-based scheme, where different choice options are evaluated and the option with the highest value is chosen. The values attributed to different

Correspondence should be made to Piotr Majer, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany. Email: majerpio@cms.hu-berlin.de

incarnations of options may be generated by a nonobservable utility function. It was first formalized by [Bernoulli \(1738\)](#) and further developed by [Neumann and Morgenstern \(1953\)](#) and [Kahneman and Tversky \(1979\)](#) to address the uncertainty of outcomes. In this case individual risk preferences are attributed to the curvature of the utility function. Alternatively, decision making can be explained in a framework of risk-return models, which incorporate the risk attitude as a weighting factor, see, e.g., [Weber and Milliman \(1997\)](#).

Research in the field of Decision Neuroscience (as well as its sub-field Neuroeconomics) attempts to address human economic behavior (i.e., decisions) by looking at neural systems that underlie decision making (e.g., [Camerer, 2007](#); [Heekeren, Marrett, & Ungerleider, 2008](#)). In practice one measures changes in brain activity using methods such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), see, e.g., [Ruff and Huettel \(2013\)](#). fMRI is based on measuring the blood-oxygen-level-dependent (BOLD) signal and captures parameters related to changes in blood flow and blood oxygenation. fMRI data are recorded over time, for example during multiple investment decisions. The captured changes in fMRI BOLD signal are indirectly related to neural firing rates ([Logothetis, 2008](#)). The acquired images are high-dimensional, and detecting stimulus-related effects is a non-trivial task. Changes in brain activation in response to decision making may be of a modest size (i.e., in comparison to reactions to visual or auditorial stimuli) and possible hemodynamic responses may be subtle and hardly detectable in the BOLD signal. It poses a genuine challenge to all existing methods and may require some extraordinary techniques.

A benchmark method to detect brain regions activated by the stimulus is the general linear model (GLM). GLM is a single-voxel technique which tests each voxel separately and results in a 3-D map of changes in fMRI signal. The test is done in a linear regression setup, where the voxel time series are modeled according to the hypothesized and predefined regressors (design matrix), which correspond to the experimental paradigm and potential confounds. This simple methodology has proved to be extremely successful in practice and has led to a wealth of important findings (e.g., [Kable & Glimcher, 2007](#)), also regarding investment decisions ([Mohr, Biele, Krugel, Li, & Heekeren, 2010](#); [Mohr, Biele, & Heekeren, 2010](#)). Nevertheless, it has several limitations. Firstly, all neural activity not predefined in the design is neglected and cannot be identified by the model. In contrast to this model-based approach, recently introduced model-free approaches ([Beckmann & Smith, 2005](#); [van Bömmel et al., 2013](#)) offer to identify effects without any a priori hypothesis. Secondly, possible information reflected in variability and higher moments of the BOLD signal ([Mohr & Nagel, 2010](#); [Garrett et al., 2013](#)) is disregarded by the GLM approach. Moreover, activation maps derived by the single-voxel approach may be “inherently limited” by a typically low signal-to-noise ratio of individual voxel data, as reported by [Heller, Stanley, Yekutieli, Rubin, and Benjamini \(2006\)](#). Alternatively, a simultaneous analysis of multi-voxel data that co-vary with the experimental design may increase the signal without adding noise.

To overcome these shortcomings, we follow the idea of [Heller et al. \(2006\)](#) and focus our analysis on the cluster rather than voxel unit. This leads in fact to an alternative technique for analyzing fMRI data, where the brain parcellation serves as a starting point. The fMRI clustering is done by the normalized cut spectral algorithm ([Shi & Malik, 2000](#)) which became very popular in neuroscience, see, e.g., [Craddock, James, Holtzheimer, Hu, and Mayberg \(2012\)](#). The algorithm makes use of a correlation between neighboring voxels which defines their proximity. Thus, a possible co-movement (i.e., simultaneous hemodynamic response) plays a key role in defining a homogeneous cluster. The shape and spatial structure is data driven, and clusters are contiguous volumes of voxels, ensuring interpretability. After functional connectivity maps are constructed, one needs to investigate neural activity displayed by the cluster unit. Our approach is model-free, the signal carried within a cluster is extracted by the dynamic semiparametric factor model (DSFM). The DSFM, proposed by [Park, Mammen, Härdle, and Borak \(2009\)](#), is employed here

as a dimension reduction technique (van Bömmel et al., 2013). It filters the noise and extracts only the common temporal information (i.e., joint reaction by neighboring voxels to the stimulus). The resulting simple, denoised temporal representation of cluster dynamics may be tested for activation within the GLM framework or using a model-free approach. Our technique: Cluster, Estimation, Activation, and Decision (CEAD) method combines parcellation based on functional connectivity and DSFM. Thus, it greatly simplifies the complexity of the data while preserving the high accuracy of the representation. Particularly this high spatiotemporal accuracy is of great importance, when stimulus-related effects may be subtle and local (such as in investment decisions under risk).

The presented methodology is applied to investigate a possible relationship between individual differences in risk preferences and dynamics in the BOLD response. In the first step, the extracted temporal information from clusters is tested for changes in brain activation. These, possibly few, activated clusters correlated with risk are further investigated with respect to risk attitudes estimated from subject responses to investment decision (ID) tasks. Here, we establish a link between changes in BOLD signal and individuals' risk weights in a risk-return model. Based on this analysis, we identify bilateral anterior insula (aINS) activity as a correlate of risk (standard deviation). The risk attitudes derived from the subject's investment decisions are successfully predicted based only on the underlying brain activity in aINS.

In the upcoming Section (2), we describe the experimental procedures, our methodology, and derivation of risk attitudes. At the end of that part, a short simulation study of testing performance is shown. In the next Section (3), our modeling parameters and empirical findings are reported. We show and exploit the relation between risk preferences and temporal information extracted from clusters. Our conclusions are detailed in the discussion section.

## 2. Materials and Methods

In this section, our experimental and fMRI data acquisition setup is presented. In the next step, we describe our methodology and employed statistical tools. It begins with an introduction to the normalized cut spectral clustering. Secondly, the advanced dimension reduction technique, DSFM, is discussed. It shows how to extract a temporal information (i.e., hemodynamic response) from entire clusters. We briefly sketch our activation testing procedure which is similar to the voxelwise GLM approach. The testing performance is evaluated in a simulation study. Finally, we introduce the risk-return model and estimate the subjects' risk attitudes based on their investment decisions.

### 2.1. Experimental Procedures

Subjects,  $I = 19$ , performed an adjusted version of the Risk Perception in Investment Decisions Task (Mohr et al., 2010). In this task subjects see past returns of either one single investment or two investments that form a portfolio (50 % of the money invested in each). While they see the past returns they have to make a choice between, if they would prefer to invest in a bond with 5 % fixed return or the investment that is displayed (either single risky investment or risky portfolio). The choice situations differed in three within-subject conditions: (A) choices between 5 % fixed return and a single risky investment, (B) choices between 5 % fixed return and a risky portfolio of 2 single investments with perfectly ( $\rho = 1$ ) correlated returns, and (C) choices between 5 % fixed return and a risky portfolio of 2 single investments with uncorrelated returns ( $\rho = 0$ ). Importantly, the return history of the risky options (either single investment or portfolio) was exactly the same in all 3 conditions. All displayed returns were gaussian with different set of parameters  $\mu$  and  $\sigma$ , where  $\mu = 5, 7, 9, 11$  % and  $\sigma = 2, 4, 6, 8$  %. Each of the choices regarding single investments

was repeated once to hold the number of choices between the bond and a single investment and the bond and a portfolio constant. In total subjects made 256 choices in two blocks of 128 choices each. Subjects had a maximum of 7 s to enter their choices via a response box with two buttons. The location of the choice options on the screen was counterbalanced between left and right to avoid order effects.

## 2.2. fMRI Data

MRI data were acquired on a 3 T scanner (Trio; Siemens) using a 12-channel head coil. Functional images were acquired with a gradient echo T2\*-weighted echo-planar sequence (TR = 2,000 ms, TE = 30 ms, flip angle = 70, 64 × 64 matrix, field of view = 192 mm, voxel size = 3 × 3 × 3 mm<sup>3</sup>). A total of 37 axial slices (3-mm thick, no gap) were sampled for whole-brain coverage. Imaging data were acquired in two functional runs with 695 and 705 volumes, respectively. A high-resolution T2-weighted anatomical scan of the whole brain was acquired (256 × 256 matrix, voxel size = 2 × 2 × 2 mm<sup>3</sup>).

The data were initially pre-processed with FSL 4.0 (FMRIB's Software Library). Pre-processing included motion correction and slice time correction. Additionally, images were normalized into a standard stereotaxic space [Montreal Neurological Institute (MNI), Montreal, Quebec, Canada]. As a result high-dimensional data were obtained 91 × 109 × 91 × 1400, where  $t = 1, \dots, 1400$  for each subject  $i = 1, \dots, 19$ .

## 2.3. fMRI Analysis

The key idea of this study is to use data-driven, contiguous clusters as the units of the analysis. The clustering is done by a Spatially Constrained Spectral Clustering algorithm which became extremely successful in neuroscience, see, e.g., Craddock et al. (2012). In the second step, temporal information contained in each cluster is extracted by the DSFM approach, as an alternative to averaging over voxels in the clusters proposed by Heller et al. (2006). Comparison with the latter approach is presented in a simulation study (see Section 2.4) and our empirical results. After the cluster temporal information is extracted, activated regions of interest (ROIs) are found by the GLM testing procedure.

*2.3.1. Spatially Constrained Spectral Clustering* The brain parcellation results from normalized cut spectral clustering (NCUT). This technique, first proposed by Shi and Malik (2000), is reported to be robust to outliers (Luxburg, 2007) and computationally efficient. It also allows for a simple incorporation of constraints, i.e., a spatial contiguity, which can be exploited in the human brain mapping. The method was introduced to the field of cognitive neuroscience by (van den Heuvel, Mandl, & Hulshoff Pol, 2008; Shen, Papademetris, & Constable, 2010; Craddock et al., 2012). Shen et al. (2010) reported that task-related fMRI data may be analyzed with this algorithm and that the resulting brain parcellation is highly consistent with the resting-state fMRI. The NCUT approach is closely related to the graph theoretic formulation of clustering. The set of voxels  $Y = (Y_1, \dots, Y_J)$  is represented as a weighted undirected graph, where the nodes of the graph are the voxels and an edge is given between every pair of voxels  $Y_j$  and  $Y_{j'}$ . The weight on each edge, denoted by  $w(j, j')$ , is a proximity measure between voxels (nodes)  $j$  and  $j'$  and is defined as in the previous paper:

$$w(j, j') = \begin{cases} \max \{ \text{Corr}_t(Y_j, Y_{j'}), 0 \}, & \text{for } \|X_j - X_{j'}\| < d, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$



where  $\|\cdot\|$  denotes the Euclidean norm in  $\mathbb{R}^3$  space,  $X_j \in \mathbb{R}^3$  are  $j$ th voxel coordinates. The radius  $d$  is selected in such a way that only the 26 nearest neighbors (face and edge touching; 3-D neighborhood of a single voxel) are included. Such a constraint ensures a contiguous shape of each cluster (Xu, desJardins, & Wagstaff, 2005; Kamvar, Klein, & Manning, 2003). Moreover, the similarity matrix  $W = \{w(j, j')\}_{j, j'=1, \dots, J}$  (of size  $J \times J$ ) derived by (1) is sparse and thus computational complexity is reduced. The similarity between voxels in 3-D neighborhood is given by correlation coefficient of the voxels time series with a threshold to make it non-negative. By applying the correlation as a similarity measure, we ensure the temporal homogeneity within a cluster, which is further exploited in the next Section (2.3.2). Once a proximity measure is chosen, a group-building algorithm for creating a functional connectivity map needs to be specified. The NCUT algorithm is a hierarchical procedure, it starts with the coarsest partition possible: one cluster contains all of the voxels. It proceeds by splitting the single cluster up into smaller sized clusters until a pre-specified number of groups  $C$  is achieved. The partition of an initial set is done such that the similarity between voxels within the proposed group is greater than the similarity between voxels in different groups. For example, for two disjoint groups  $P$  and  $Q$ , one computes the normalized cut cost by

$$N_{\text{cut}}(P, Q) = \frac{\sum_{Y_j \in P, Y_{j'} \in Q} w(j, j')}{\sum_{Y_j \in P, Y_{j'} \in R} w(j, j')} + \frac{\sum_{Y_j \in Q, Y_{j'} \in R} w(j, j')}{\sum_{Y_j \in Q, Y_{j'} \in R} w(j, j')}, \quad (2)$$

where  $R = Q + P$  is the initial set that has to be partitioned. The denominators in the formula (2) may be seen as a sum of all similarities between sets  $P$  and  $Q$  that are neglected in this division. The nominators stand for all the similarities between the proposed groups ( $P$  and  $Q$ ) and the initial set  $R$ , thus a size of a group has an influence on the normalized cut cost. Finding an optimal division of set  $R$  might be found by minimizing the normalized cut criterion:

$$(P^*, Q^*) = \arg \min_{R=P+Q} N_{\text{cut}}(P, Q). \quad (3)$$

Therefore we ensure that, simultaneously, similarities within each cluster are maximized and similarities between clusters are minimized. This approach leads to balanced sizes of clusters and reduces the likelihood of obtaining singletons as a result. Shi and Malik (2000) showed that minimizing (2) is equivalent to minimizing the Rayleigh quotient denoted by

$$Q(y) = \frac{y^\top \mathcal{L} y}{y^\top D y}, \quad (4)$$

under the constraint that  $y$  is a piecewise (discrete) vector  $J \times 1$  and  $y^\top \text{diag}(D)1_J = 0$ . Matrix  $\text{diag}(D)$  is defined by  $D = (d_1, \dots, d_J)$  a degree vector,  $d_j = \sum_{j'=1}^J w(j, j')$  and  $\mathcal{L}$  is the Laplacian of the graph given by

$$\mathcal{L}(j, j') = \begin{cases} d_j & , j = j', \\ -w(j, j') & , w(j, j') > 0, \\ 0 & , \text{elsewhere.} \end{cases} \quad (5)$$

Minimizing the formula (4) is closely related to spectral clustering, where the first nontrivial eigenvector of the graph Laplacian matrix  $\mathcal{L}$  is used. The authors showed that the problem is NP-complete, an approximate discrete solution can be found efficiently.

2.3.2. *Dynamic Semiparametric Factor Model* The clusters are constructed to maximize the temporal homogeneity between voxels. Their similar time evolution (i.e., reflected in joint hemodynamic response after stimuli) explicitly suggest possible low-dimensional representation of the multidimensional time series. The temporal variability in the cluster series, that may be related to investment decisions and possibly individual differences in risk attitude, is captured by a dynamic semiparametric factor model (DSFM), proposed by [Park et al. \(2009\)](#). DSFM serves here as a dimension reduction technique, which is able to extract temporal dynamics from the functional connectivity brain maps by corresponding low-dimensional time series (factor loadings) in only one estimation step. Due to a subject-specific spatial structure of the brain functional connectivity maps, we model each cluster separately.

The BOLD signal of all voxels in a single cluster  $c$ ,  $c = 1, \dots, C$  during the entire experiment is a multi-dimensional time series. The stated below DSFM is designed to model such high-dimensional time series:

$$Y_{t,j} = m_0(X_{t,j}) + \sum_{l=1}^L Z_{t,l} m_l(X_{t,j}) + \varepsilon_{t,j}, \quad 1 \leq j \leq J_c, \quad 1 \leq t \leq T.$$

$$\stackrel{\text{def}}{=} Z_t^\top m(X_{t,j}) + \varepsilon_{t,j} = Z_t^\top A^* \Psi_{t,j} + \varepsilon_{t,j}, \quad (6)$$

where  $Z_t = (\mathbf{1}, Z_{t,1}, \dots, Z_{t,L})^\top$  is a latent  $(L + 1)$ -dimensional stochastic process and  $m$  is an  $(L + 1)$ -tuple  $(m_0, \dots, m_L)$  of unknown real-valued functions  $m_l$ . More precisely, the voxel's coordinates  $(x_1, x_2, x_3) \in \mathbb{R}^3$  that belong to an analyzed cluster  $c$  is the covariate  $X_{t,j}$  (in this setup it is time-invariant  $X_{t,j} = X_j$ ) and the normalized BOLD signal is the dependent variable  $Y_{t,j}$ ;  $j = 1, \dots, J_c$ ;  $t = 1, \dots, T$ . We assume  $\varepsilon_{t,j} \perp Z_{t,j}$ ,  $\mathbb{E} \varepsilon_{t,j} = 0$  and  $\mathbb{E} \varepsilon_{t,j}^2 < \infty$ . The functions  $m_l$  are given as a linear combination of space basis functions  $\Psi_{t,j} = [\psi_1(X_{t,j}), \dots, \psi_K(X_{t,j})]^\top$  and corresponding  $(L + 1) \times K$  matrix of unknown coefficients  $A^*$ . In our setup,  $[\psi_1(X_{t,j}), \dots, \psi_K(X_{t,j})]^\top$  are quadratic tensor B-splines on  $K$  equidistant knots. To find the estimates of  $Z_t^\top$  and  $A^*$ , one solves

$$(\widehat{Z}_t, \widehat{A}^*) = \arg \min_{Z_t, A^*} \sum_{t=1}^T \sum_{j=1}^J \{Y_{t,j} - Z_t A^* \Psi_{t,j}\}^2. \quad (7)$$

A solution to the problem stated in (7) may be found by the Newton–Raphson method. Time dynamics are represented by  $\widehat{Z}_t$ , while  $\widehat{A}^*$  captures the smooth, nonparametrically estimated spatial structure of clusters.

In the formula (6) the time frame is constant over all clusters and equals  $T = 1,400$ . Due to varying spatial structure and size of each cluster  $c$ ,  $c = 1, \dots, C$ , we denote the dimension  $J_c$  as the  $c$  cluster size. The statistical inference of the each cluster is then based on the low-dimensional time series analysis for  $Z_t$ . As shown by [Park et al. \(2009\)](#), the inference based on the estimates  $\widehat{Z}_t^\top$  holds for “true” unobserved time series  $Z_t^\top$ , as the difference between  $Z_t^\top$  and  $\widehat{Z}_t^\top$  is asymptotically negligible.

2.3.3. *General Linear Model and Testing Procedure* In practice, the analysis of BOLD fMRI data is conducted using voxelwise General Linear Model, see, e.g, [Friston et al. \(1994\)](#) and [Worsley et al. \(2002\)](#), where the magnetic resonance signal at voxel  $j$  is modeled by

$$Y_j = \widetilde{X} \beta_j + e_j, \quad (8)$$

where  $\tilde{X}$  denotes the  $T \times p$  design matrix,  $\beta_j$  is the  $p \times 1$  vector of regression coefficients, and  $e_j$  is a (often serially correlated) measurement error. The matrix  $\tilde{X}$  is constructed as a convolution of hemodynamic response function (HRF)  $h(t)$  and the stimulus time signal and might also incorporate additional elements (i.e., temporal derivatives) when required by a specific experiment setup. It is common practice to model the HRF by a difference of two gamma functions, i.e.,

$$h(t) = \left(\frac{t}{5.4}\right)^6 \exp\{-(t - 5.4)/0.9\} - 0.35 \left(\frac{t}{10.8}\right)^{12} \exp\{-(t - 10.8)/0.9\},$$

see, e.g., [Worsley et al. \(2002\)](#). Inference focuses on the estimates  $\hat{\beta}_j$  and the hypothesis  $H_0 : \beta_j = 0$  is tested voxelwise (first-level analysis).  $\hat{\beta}_j$  being significantly different from 0 is interpreted as activation at the voxel  $j$ . Group analysis is usually done in the mixed-effects framework, where the activation pattern for  $i$  subject at  $j$  voxel  $\hat{\beta}_j^i$  serves as an input for the model (higher-level analysis). This standard technique implemented in FSL's FLAME (FMRIB's local analysis of mixed effects) is used here to test whether regression coefficients are significant and activation can be reported at the group level. The region of interest is reported to be significantly activated for clusters reaching uncorrected threshold of  $Z$ -score  $> 3.09$  and consisting of at least 20 neighboring voxels. For more details, we refer here to the technical reports of the FMRIB Analysis Group, see, e.g., [Beckmann, Jenkinson, and Smith \(2003\)](#) and [Beckmann and Smith \(2004\)](#).

**2.3.4. Cluster, Estimation, Activation, and Decision (CEAD) Method** The resulting cluster representation by  $\hat{Z}_t^\top$  serves as the unit of analysis for the relevant signals related to the ID tasks and decisions. Profiting from higher signal-to-noise ratio present on the group level ([Heller et al. 2006](#)) clusters are tested for activation. For analysis of all participated subjects  $i = 1, \dots, I$ , our multivariate scheme may be summarized in the following steps:

1. **Cluster step:** for each subject  $i$  construct the brain parcellation into  $C$  groups using spectral clustering NCUT algorithm.
2. **Estimation step:** given the subject-specific clustering results, for subject  $i$  take the  $c$  cluster and fit the DSFM, given in (6). Repeat this estimation procedure for all clusters  $c = 1 \dots, C$  and all subjects  $i = 1, \dots, I$ . The DSFM approach is thus applied  $C \times I$  times separately.
3. **Activation step:** representing  $(i, c)$ ,  $i = 1, \dots, I$ ,  $c = 1 \dots, C$  cluster dynamics by low-dimensional representation  $\hat{Z}_t^{(i,c)}$  test the time series activation in the GLM framework. Select the activated clusters that are related to neural processes underlying (risky) investment decisions.
4. **Decision step:** investigate the activated factor loadings  $\hat{Z}_t^{(i,c)}$ . Is the subjects investment behavior represented in any of the activated clusters? Is there any relation between the risk attitude and the low-dimensional time series?

#### 2.4. Simulation Study

This part of our study is designed to investigate the performance of the proposed method in a simulation study. Our approach is evaluated against the benchmark, voxelwise GLM and the averaging technique introduced by [Heller et al. \(2006\)](#) (in each cluster take average over voxels and test for activation). We simulated data at one, exemplary cluster on the  $6 \times 7 \times 6$  grid that mimics the average cluster obtained in our empirical analysis:  $Y_t = Z_t^\top m(X) + \varepsilon_t$ , where  $Y_t$  is a  $6 \times 7 \times 6 \times 1400$  BOLD signal,  $m(X) = m(x, y, z) = \|(x, y, z) - (6, 8, 6)\|$  is a smooth spatial structure,  $Z_t$  is a (perfect) stimulus time series (HRF  $\times 64$ , see [Figure 10](#)) and  $\varepsilon_t$  is noise. The (single) factor  $m(\cdot)$  is a smooth, non-linear function that decreases in the direction

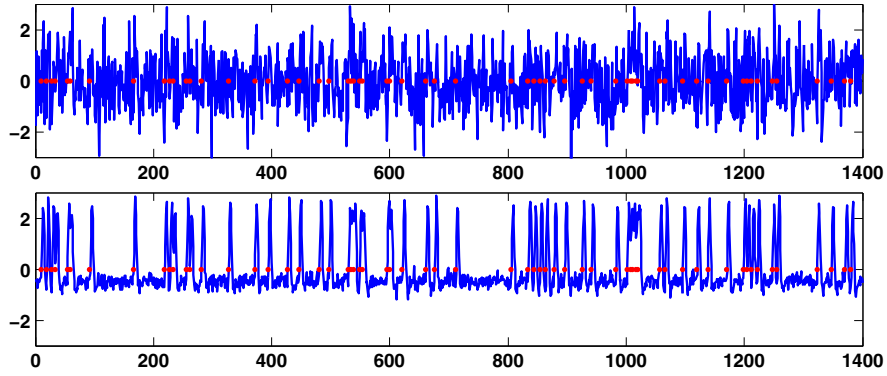


FIGURE 1.

Setup (a): the simulated  $(1, 1, 1)$  voxel  $Y_{t,1}$  (*top*) and the estimated  $\hat{Z}_t$  (*bottom*) plotted against time (each 2 s); *red dots* denote stimulus;  $\text{Corr}_t(\hat{Z}_t, \text{stimulus}) = 0.98$ .

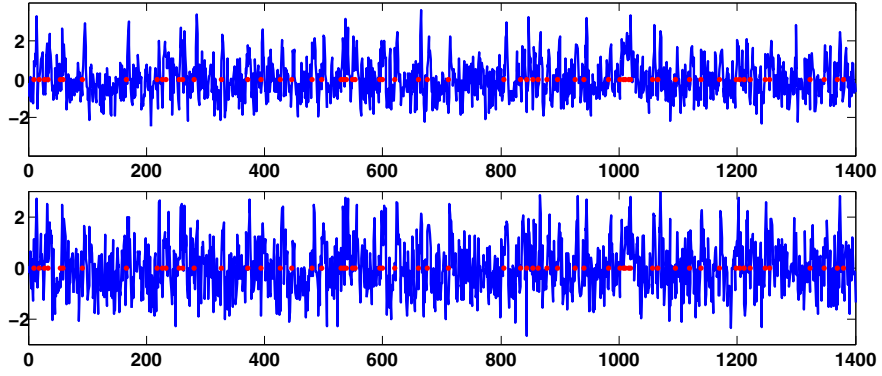


FIGURE 2.

Setup (b): the simulated  $(1, 1, 1)$  voxel  $Y_{t,1}$  (*top*) and the estimated  $\hat{Z}_t$  (*bottom*) plotted against time (each 2 s); *red dots* denote stimulus;  $\text{Corr}_t(\hat{Z}_t, \text{stimulus}) = 0.60$ .

of the point  $(6,8,6)$ , that is not present on the grid, thus  $m(\cdot) > 0$ . The  $Z_t$  is the simplest design matrix (here  $1 \times 1400$ ) from GLM setup and in this case stands for all stimuli corresponding to the correlated portfolio from our experiment. Therefore, we assume that only one true neural process is present in this cluster. We investigate two possible cases for  $\varepsilon_t$  ( $6 \times 7 \times 6 \times 1400$ ): (a)  $\varepsilon_t$  is i.i.d. Gaussian and (b)  $\varepsilon_t$  is spatially correlated Gaussian;  $\mu = 0$  and  $\sigma = 1$ . The spatially correlated noise time series  $\varepsilon_{sc,t}$  is derived (independently at each  $t$ ,  $t = 1, \dots, 1400$ ) as a convolution of i.i.d. Gaussian noise from (a) with a spatial Gaussian kernel (FWHM 8 mm) and depicted in Figure 11. Examples of simulated BOLD signals are shown in Figures 1 and 2. The performance for all three techniques: DSFM with  $L = 1$ , GLM (pre-smoothed with FWHM 8 mm) and averaging over voxels in the cluster (with and without pre-smoothing) for the setup (a) is remarkably good and all statistics are higher than 100. The (b) study is summarized in Table 1. Firstly, all investigated techniques discover a significant activation and yield similar results. Secondly, the maximum  $Z$ -score in the GLM approach is the highest test statistics in all cases. When the  $Z$ -scores are averaged over all voxels, the DSFM approach yields the best result. Moreover, the simple averaging approach is outperformed by the DSFM. We conclude that DSFM might serve as an interesting alternative to the benchmark GLM method, especially if the analysis goes beyond an identification of activation patterns (i.e., higher moments, time series analysis of voxels in a neighborhood).

TABLE 1.

Test statistics  $Z$ -scores derived in simulation setup (b) for GLM, DSFM, averaging, and averaging for smoothed (FHW 8mm) data denoted by Average(s).

	GLM	DSFM	Average(s)	Average
Max $Z$ -score	30.54	27.96	27.14	27.48
Mean $Z$ -score	26.34	27.96	27.14	27.48

TABLE 2.

Test statistics  $Z$ -scores derived in simulation setup (c)-upper and (d)-lower panel, respectively, for GLM, DSFM, averaging, and averaging for smoothed (FHW 8mm) data denoted by Average(s).

	GLM	DSFM	Average(s)	Average
Max $Z$ -score	1.90	0.38	0.68	0.62
Mean $Z$ -score	0.61	0.38	0.68	0.62
Max $Z$ -score	1.66	1.10	1.07	1.10
Mean $Z$ -score	0.99	1.10	1.07	1.10

The performance of the proposed method is also studied, when the exemplary cluster does not exhibit stimulus-related effects. In particular, we simulated the  $6 \times 7 \times 6 \times 1,400$  BOLD signal  $Y_t = \tilde{Z}_t^\top m(X) + \varepsilon_{sc,t}$ , where: (c)  $\tilde{Z}_t = 1_{1,400}$  is a constant series of ones and (d)  $\tilde{Z}_t$  is a simulated autoregressive process of order 2, where  $\tilde{Z}_t = 0.5\tilde{Z}_{t-1} + 0.2\tilde{Z}_{t-2} + \varepsilon_{AR,t}$ ,  $\varepsilon_{AR,t}$  is a white noise independent of  $\varepsilon_{sc,t}$  and  $\text{Corr}_t(\tilde{Z}_t, \text{stimulus}) = 0.04$ , see Figure 12. Therefore, the setup (c) corresponds to a case, when only the (spatially correlated) noise is present in the cluster and there is no common neural signal. The setup (d) assumes a common neural process which is not related to the stimulus. The results of all 3 techniques are summarized in Table 2. The resulting  $Z$ -scores are remarkably smaller than a typical threshold 3.09 and the stimulus-related effects are not identified. Furthermore, all approaches yield similar results.

## 2.5. Behavioral Modeling

The subject-specific risk attitudes can be directly derived from subject responses to ID tasks. Following (Markowitz, 1952 and Caraco, 1981), we apply the benchmark mean-variance model to reflect the subject's decision-making process:

$$V_i(x) = \bar{x} - \phi_i S(x), \quad (9)$$

where  $V_i(x)$  is the *value* a subject  $i$  assigns to an investment  $x$ ,  $\bar{x}$  is an empirical mean and represents the *expected return*,  $S(x)$  stands for a standard deviation and represents the subject's *risk*, and  $\phi_i$  is the individual risk weight: risk attitude. Therefore, in line with the portfolio theory introduced by Markowitz (1952), we follow the common mean-variance approach.

The risk attitude can be estimated based on subject responses (risky choice vs. sure, 5% return) by the logistic model:

$$P \{\text{risky choice}|x\} = \frac{1}{1 + \exp\{\bar{x} - \phi S(x) - 5\}}. \quad (10)$$

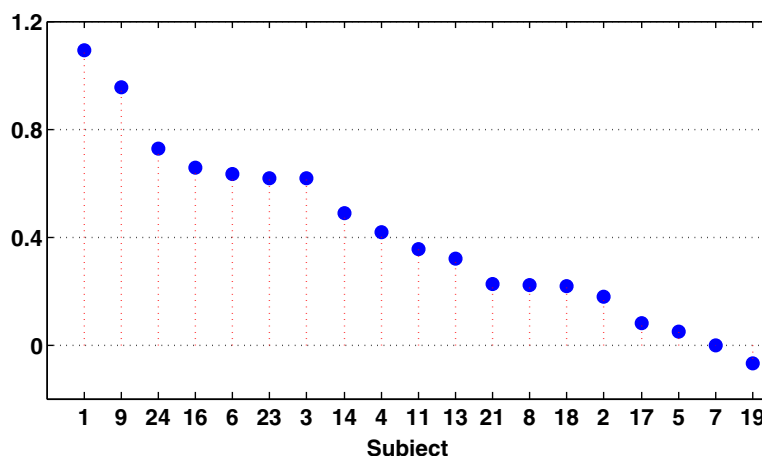


FIGURE 3.  
Risk attitudes of 19 subjects (indexed on  $x$ -axis) derived by the (10).

Negative values of  $\hat{\phi}_i$  indicate a risk-seeking behavior,  $\hat{\phi}_i \approx 0$  relates to risk neutrality and  $\hat{\phi}_i > 0$  to risk aversion. The estimated risk attitudes are shown in Figure 3 and additional analysis in Figures 13 and 14. For simplicity of presentation, in the subsequent part of the analysis we show data for two most extreme subjects: 19th, risk-seeking: risk weight =  $-0.0699$  and 1st, risk-averse: risk weight =  $1.092$ .

### 3. Results

Choice of the model parameters is described and the clustering results together with the estimated factor loadings are presented. This 2-step dimension reduction technique simplifies the brain dynamics into  $C$ -dimensional time series. The activated clusters are selected in the Activation step and the subjects' risk aversion is modeled and predicted based only on the fMRI data.

#### 3.1. Model Parameters

Selection of the number of clusters plays of course a role in our analysis. Choosing only few regions of interest (i.e., 50 parcellations) leads to over-generalized and condensed regions that are anatomically distinctive, see, e.g., Craddock et al. (2012). Increasing the division into 200 clusters is reflecting the anatomical brain atlases (Talairach & Tournoux, 1988; Desikan et al., 2006) and an approach based on the brain identified atlas zones is often used. When a more precise parcellation is called for, practitioners then select 1,000 clusters as discussed by Craddock et al. (2012). Our study aims to find activated brain regions related to the investment decisions, where the possible HRF may be subtle. Moreover, a successful implementation of the dynamic semiparametric factor model and conducted testing procedure requires highly accurate and homogenous inputs, we thus select  $C = 1,000$  clusters and ensure thereby the high accuracy of the representation. In the next step, each (homogenous) cluster is represented by the DSFM technique with 1 dynamic factor,  $L = 1$  for all cluster  $c = 1, \dots, 1,000$ . Inclusion of higher number, though yielding a better fit, does not allow for a simple interpretation.

The parcellation technique is based on (1) as a proximity measure. In order to check stability of (1) over the entire experiment, we conduct a moving window exercise. Figure 15 shows the correlation between 3 neighboring voxels derived by a rolling window exercise (for past  $250 \approx 8$

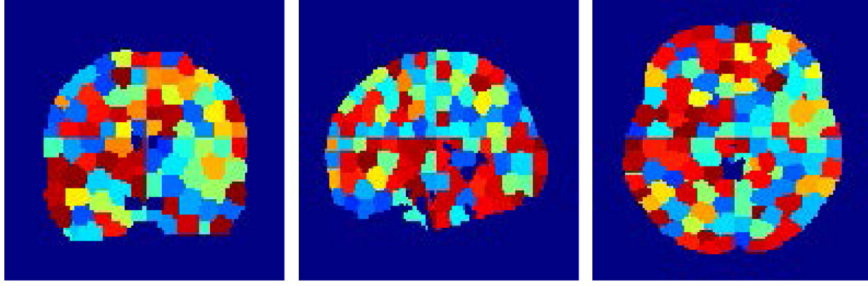


FIGURE 4.

Illustration of the clustering results for subject 1 derived by the NCUT algorithm,  $C = 1,000$ . The parcellation is represented as an orthogonal view, and color-coding is arbitrarily used to capture the clusters' boundaries.

min and  $500 \approx 17$  min). One observes a stable, stationary behavior over time which stands in favor of our modeling setup.

### 3.2. Clustering Results

Clustering results are illustrated in Figure 4. The subject-specific parcellation, though computationally extensive, addresses inter-subject functional variability. Therefore, we derive spatially coherent regions of homogenous functional connectivity, that are present at a voxel scale. The clusters are contiguous sets of neighboring voxels and a distinction between network nodes and large-scale network of nodes is ensured, see Smith et al. (2009). The neuroscientific interpretability is preserved and further elaborated on in the modeling and testing part of our study. An average cluster is of a size 207 voxels, which might be compared to a  $6 \times 6 \times 6 = 216$  (12 mm) cube. The smallest cluster is a singleton and the largest consists of 353 voxels. Clusters have a data-driven shape and vary with respect to the size and spatial structure as shown in Figure 16.

### 3.3. Factor Loadings $\widehat{Z}_t$

The clustering spatial maps serve as a basis for further exploratory analysis. The information carried in time evolution of the derived clusters is extracted by the DSFM technique. More precisely, all voxels belonging to cluster  $c$  of subject  $i$ :  $Y_{c,1}^i, \dots, Y_{c,J_c^i}^i$ , where  $J_c^i$  is the size of  $c$  cluster for subject  $i$ , are jointly modeled by (6). For simplicity of representation and as a natural consequence of cluster (homogenous) construction, we employ the DSFM with  $L = 1$ . Thus, each cluster's dynamics are captured by the univariate time series  $\widehat{Z}_t^{i,c}, i = 1, \dots, I; c = 1, \dots, 1,000$ , and the complete brain representation consists of 1,000 processes. The derived brain model significantly simplifies the complexity of the data, while ensuring the interpretability and a good quality fit. For a demonstration two extreme subjects: 1 (with the smallest risk attitude) and 19 (with the largest risk attitude) are selected, see Figure 3. Figure 5 shows the estimated  $\widehat{Z}_t^1$  and  $\widehat{Z}_t^{19}$  for anterior insula (aINS; left and right) and dorsomedial prefrontal cortex (DMPFC) clusters. All factor loadings exhibit stationary behavior, high persistency, and a high fluctuation around their mean value (see Figure 17; Table 4), which may be related to the underlying investment decision stimulus.

### 3.4. Activation Results $\widehat{Z}_t$

The derived low-dimensional representation of each cluster  $\widehat{Z}_t$  serves as a principal unit of this study and is tested for activation. We compare our method with both the standard voxelwise GLM technique and the approach proposed by Heller et al. (2006) (average over voxels and use it

PSYCHOMETRIKA

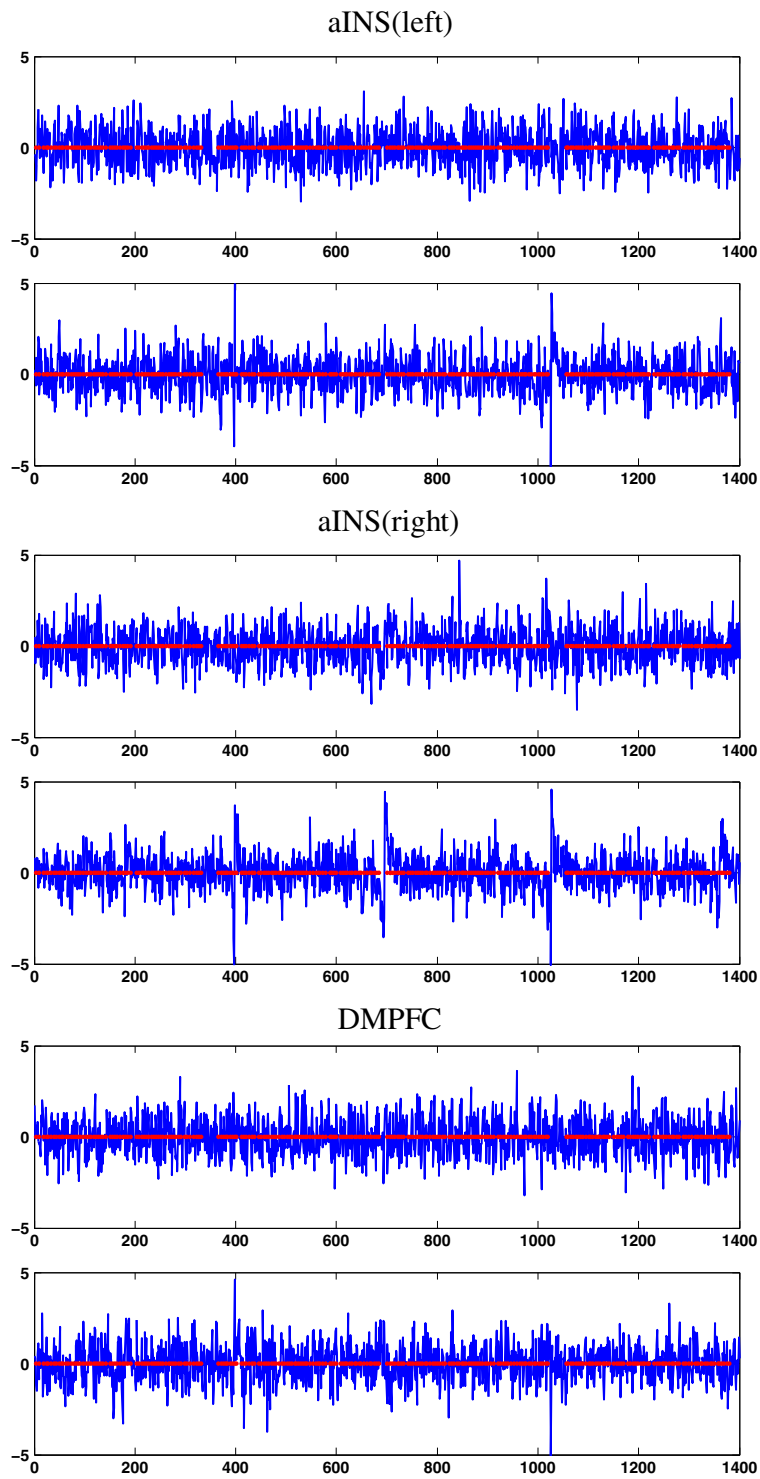


FIGURE 5.

Factor loadings  $\hat{Z}_t$  for clusters aINS(left), aINS(right), and DMPFC (upper, middle lower panel) for risk-averse subject 1 (top) and weakly risk-seeking subject 19 (bottom) plotted against time (each 2 s). Red points correspond to the time points of stimuli.



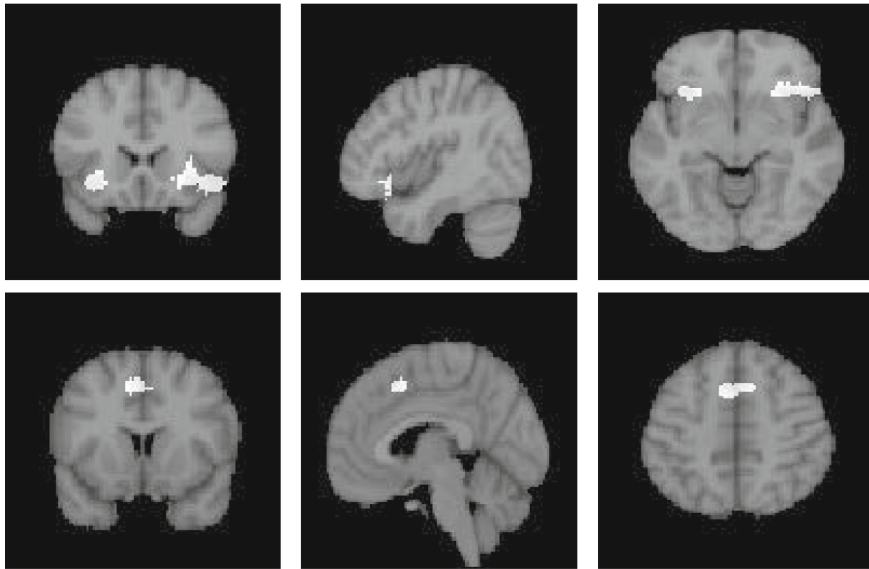


FIGURE 6.

Results of the higher-level analysis (mixed-effects model) associated with decision making;  $Z$ -scores  $> 3.09$ . *Upper panel:* the bilateral aINS, *lower panel:* DMPFC.

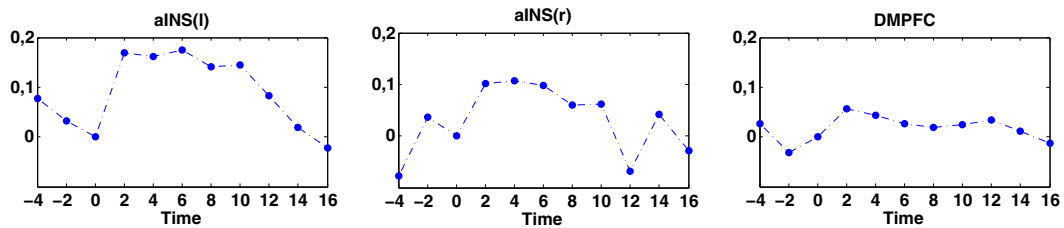


FIGURE 7.

Average reaction to the ID stimulus over all 19 subjects for bilateral aINS and DMPFC regions plotted against time (from  $-4$  s before the stimuli up to 16 s afterwards).

as a cluster temporal representation). Four separate analyses were conducted (single, correlated, and uncorrelated, jointly all types of portfolio). For each type of investment, we reported the same activation pattern, thus only the joint analysis (all portfolios) is reported here.

Figure 6 presents significant brain correlates of the ID task: aINS and DMPFC associated with decision making. These activation results are in line with findings by Mohr et al. (2010), Mohr, Biele, and Heekeren (2010) and contribute to the neural foundations of risk-return model. Altogether 9 activated clusters were detected which survived statistical thresholding at  $Z$ -scores  $> 3.09$  and had a cluster size of at least 20 voxels. Besides aINS and DMPFC factors corresponding to decision making, we identified other brain regions previously associated with visual perception and motoric responses. These factors are most likely not connected to the decision-making process but confirm the activity of regions which were necessary to give the answer by pushing the button. Average reactions to the ID stimuli over all 19 subjects are depicted in Figure 7. Reported maximum  $Z$ -scores for aINS and DMPFC are shown in Table 5. One observes that all approaches yield very similar results, although the highest maximum  $Z$ -score is achieved by the GLM technique for all 3 ROIs. Secondly, the DSFM outperforms the simple averaging over voxels. The non-parametric estimation pays off in terms of the quality of the representation.

TABLE 3.  
Risk attitude regressed on the average response for all 19 subjects;  $R^2 = 0.47$ , adjusted  $R^2 = 0.36$ .

	Estimate	SE	$t$ Statistic	$p$ value
$\alpha_0$	0.097	0.115	0.861	0.403
$\overline{\Delta \widehat{Z}}_{\text{DMPFC}}$	0.851	0.526	1.619	0.126
$\overline{\Delta \widehat{Z}}_{\text{aINS}(r)}$	-1.506	0.550	-2.737	0.015
$\overline{\Delta \widehat{Z}}_{\text{aINS}(l)}$	-1.126	0.379	-2.967	0.001

#### 4. Risk Attitude \ Stimulus Response

The key goal in neuroeconomics is to “(...) ground economic theory in detailed neural mechanisms which are expressed mathematically and make behavioral predictions.” as [Camerer \(2007, 2013\)](#) states. Motivated by that, we investigated a connection between the neural processes underlying decision making and risk perception. Without prior knowledge of the subjects’ answers, based only on the activated cluster dynamics, represented by  $\widehat{Z}_t$  a simple model is proposed to predict the risk attitude  $\phi_i$ . As described in Section (3.4), three activated (see Table 5) clusters are associated with decision making under risk. Therefore only cluster dynamics of bilateral aINS and DMPFC are considered here as regressors for the risk attitude  $\phi_i$ . These loadings (brain regions) respond to the stimulus and thus mimic neural processes present in a whole cluster during investment decisions under risk in our study. The hemodynamic response function usually peaks around 6 s after the stimulus. Therefore, we focus on an average reaction to  $r$ ,  $r = 1, \dots, 256$ , stimulus for the  $i$ th subject:  $\Delta \widehat{Z}_r^i = \frac{1}{4} \sum_{\tau=1}^4 \widehat{Z}_{r+\tau}^i - \widehat{Z}_r^i$ .  $\Delta \widehat{Z}_r^i$  covers a period up to 8 s afterward and ensures that the HRF maximum is captured. An average reaction to all stimuli (entire experiment) for a single cluster is defined as  $\overline{\Delta \widehat{Z}}^i = \frac{1}{256} \sum_{r=1}^{256} \Delta \widehat{Z}_r^i$ . Our model-free methodology closely follows the statistics proposed by [van Bömmel et al. \(2013\)](#), [Brown, Lazar, Datta, Jang, and McDowell \(2014\)](#).

Understanding which among the variables:  $\overline{\Delta \widehat{Z}}_{\text{aINS}(l)}$ ,  $\overline{\Delta \widehat{Z}}_{\text{aINS}(r)}$ ,  $\overline{\Delta \widehat{Z}}_{\text{DMPFC}}$  are related to the  $\phi$  and an exploration of the forms of these relationships is done via regression analysis. More precisely,

$$\phi_i = \alpha_0 + \alpha_1 \cdot \overline{\Delta \widehat{Z}}_{\text{DMPFC}}^i + \alpha_2 \cdot \overline{\Delta \widehat{Z}}_{\text{aINS}(l)}^i + \alpha_3 \cdot \overline{\Delta \widehat{Z}}_{\text{aINS}(r)}^i + \tilde{\varepsilon}^i, \quad (11)$$

where  $\alpha_0$  is an intercept,  $\alpha = (\alpha_1, \alpha_2, \alpha_3)^\top$  is a vector of regression coefficients and  $\tilde{\varepsilon}$  stands for the error term. In other words, (spatially constrained, local) information extracted from the BOLD signal serves as regressors for the subject’s risk weights.

Summary statistics of the model defined in (11) are reported in Table 3. Surprisingly, we report that the DMPFC factor, though significantly activated, does not carry explanatory power for risk preferences. This finding, among others, goes far beyond classical fMRI analysis done within the GLM framework and highlights the flexibility and advantages of our approach. Furthermore, the aINS, both left and right regions, are picked up by the model and reported  $p$ -values are remarkably smaller than 0.05. Overall, the explanatory power is satisfactory despite the simplicity of linear relation and the noisy nature of the studied panel data (for both, BOLD signal and risk weights). We obtain  $R^2 = 0.47$  and adjusted  $R^2 = 0.36$ . The regression fit is depicted in Figure 8.

Dropping out of the insignificant terms in (11) yields

$$\phi_i = \alpha_2 \cdot \overline{\Delta \widehat{Z}}_{\text{aINS}(l)}^i + \alpha_3 \cdot \overline{\Delta \widehat{Z}}_{\text{aINS}(r)}^i + \tilde{\varepsilon}^i. \quad (12)$$

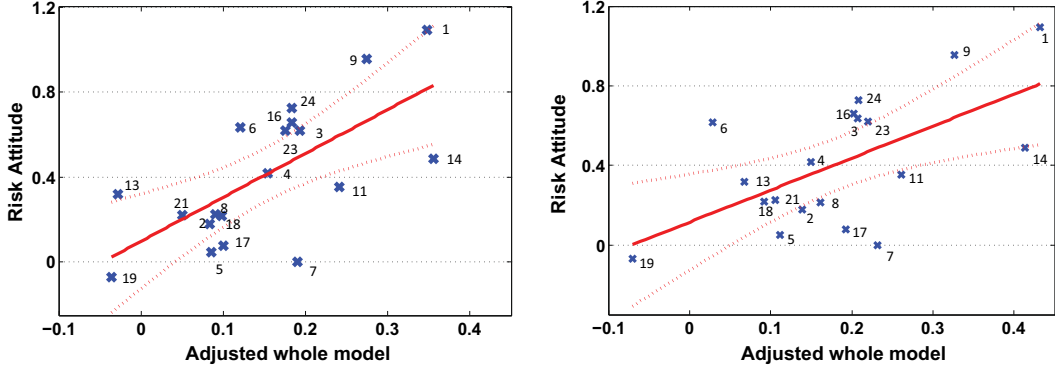


FIGURE 8.

Added variable plot for models given in (11) left and (12) right panel, respectively. Horizontal axis denotes the (rescaled) best linear combination of regressors  $\overline{\Delta\widehat{Z}}$  that fit  $\phi$ .

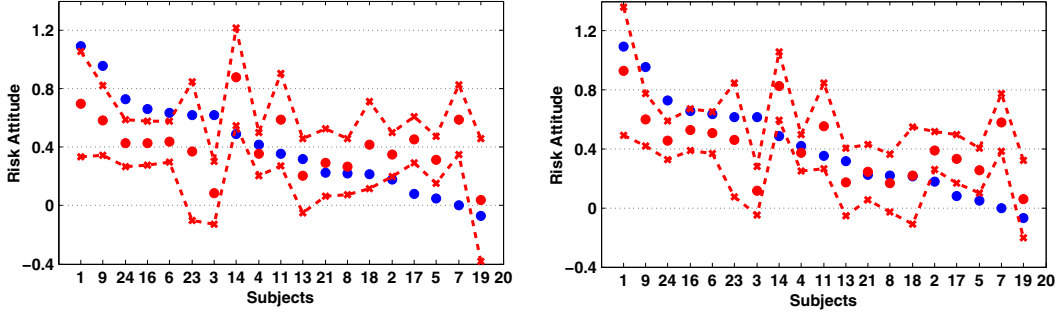


FIGURE 9.

Predicted risk preferences by the model given in (12) for the average  $\overline{\Delta\widehat{Z}}$  and the weighted average  $\overline{\Delta_w\widehat{Z}}$ : left and right panel, respectively. Information extracted from the aINS BOLD signal;  $w = (0.38, 0.41, 0.16, 0.05)^\top$ .

The simplified model achieves  $R^2 = 0.37$ , adjusted  $R^2 = 0.30$ , and the  $p$ -values are 0.03 and 0.02 for  $\overline{\Delta\widehat{Z}}_{\text{aINS}(r)}$  and  $\overline{\Delta\widehat{Z}}_{\text{aINS}(l)}$ , respectively. Figure 8 shows the regression fit. In this setup subject risk aversion depends only on the average reaction to the stimulus in the aINS regions. This setup, consisting only of activated (see Table 5) and significant BOLD cluster statistics, is kept in the remainder of the analysis.

#### 4.1. Risk Attitude Forecasting

The regression results presented in Table 3 indicate that the DMPFC factor is not significant and does not carry explanatory power for  $\phi_i$ . Thus, the regression setup, stated in (12), is used to predict the subject risk attitude based only on the information extracted from BOLD signal in aINS. For each subject  $i = 1, \dots, 19$ , its information is excluded from the regression analysis and the model (12) is re-estimated. Plugging in the neural low-dimensional representation,  $\overline{\Delta\widehat{Z}}_{\text{aINS}(l)}^i$  and  $\overline{\Delta\widehat{Z}}_{\text{aINS}(r)}^i$ , to the new model predicts the risk weight  $\phi_i$ , and the out-of-sample performance is shown in Figure 9. Seven predicted risk attitudes, out of 19, lie out of 95% prediction confidence intervals and the absolute average forecasting error is 0.257. One could expect that the proposed statistics  $\overline{\Delta\widehat{Z}}$  is not the best univariate projection of the hemodynamic response to the stimulus. To overcome some possible deviations in the HRF peak's location, we apply the weighted average reaction to the stimulus denoted by a weighted average reaction:  $\Delta_w\widehat{Z}_r^i = \sum_{\tau=1}^4 w_\tau (\widehat{Z}_{r+\tau}^i - \widehat{Z}_r^i)$ , with  $\sum_{\tau=1}^4 w_\tau = 1$ .

Thus, observations after stimulus are weighted with unknown weights  $w_\tau$ . The procedure introduced before is repeated for  $\Delta_w \widehat{Z}^i = \frac{1}{256} \sum_{r=1}^{256} \Delta_w \widehat{Z}_r^i$  and the weights are found by minimizing the absolute average forecasting error. The optimal weights  $w = (0.38, 0.41, 0.16, 0.05)^\top$  are derived by Monte Carlo simulation with 10,000 iterations and the new absolute average prediction error is 0.202. The prediction fit is reported in Figure 9. In this setup the first three observations (up to 6 s after stimuli) exhibit a remarkably higher impact than the 4th one.

The neural predictions of risk attitudes, though satisfactory, do not perfectly match risk weights derived from subjects' investment decisions. A plausible explanation from a statistical point of view would be the simplicity of linear relation, inhomogeneity of studied subjects, and above all, the noisy nature of the data. Nevertheless, we are convinced that the neural processes underlying investment decisions and corresponding risk preferences are a far more complex phenomenon and go beyond the aINS and DMPFC only. Our statistical methodology is constrained here by the experiment setup that, naturally, cannot capture all brain reactions and allows only to estimate a proxy of "true" risk preferences by risk-return model. Although the activation is reported by the benchmark testing procedure, we suspect additional brain regions to contribute to investment decisions (e.g., Mohr, Biele, and Heekeren, 2010) not identified in this fMRI study. This goes beyond the scope of this paper and deserves further research.

## 5. Discussion

We have presented a novel method for analyzing fMRI data based on cluster units: CEAD. In the first step, the clusters are derived via the NCUT algorithm as contiguous groups of voxels and there are no further constraints concerning the shape and spatial structure. This data-driven approach makes use of the correlation between neighboring voxels and therefore ensures a co-movement of the BOLD signals within cluster. This property of "anatomic" homogeneity pays off when temporal information carried by each cluster has to be extracted. Derived functional connectivity maps are a starting point of analysis. In the estimation step, the DSFM method is applied on each cluster and serves here as a dimension reduction technique. It serves as a filter of the noise and only extracts the common temporal information: the signal (i.e., joint reaction to the stimulus). This semiparametric approach can handle various specifications of noise observed at the voxel level and yields favorable results in comparison to simple averaging over voxels (Heller et al., 2006). It is a model-free technique that derives complete spatiotemporal information from brain regions. In the activation step, the extracted signal is further studied for experimental responses. Our local-dynamic representation yields similar results as traditional GLM analyses. The high accuracy of the model plays an important role when possible task-related effects are subtle and local. Our approach ensures a simplicity of neural interpretation and addresses the key limitations of the benchmark method GLM. In the decision step, the CEAD method allows for any model-free analysis of spatiotemporal ROI's information.

We apply the CEAD methodology to study neural systems that underlie decision making under risk. In particular, investment decision is a complex process of valuation and comparison of possible choices with unknown outcomes. Risk attitude is a crucial metric that influences the subjective value of investment. In this paper, we analyzed an fMRI experiment with 19 subjects. Each subject was scanned during multiple ID tasks and a series of 1,400 images of  $91 \times 109 \times 91$  voxels are investigated here. Using our methodology, we decomposed individual brains into sets of 1,000 spatially disjoint factors and factor loadings  $\widehat{Z}_t^{i,c}$ ,  $i = 1, \dots, 19$  and  $c = 1, \dots, 1,000$ . Derived spatiotemporal representation is

subject specific and possible variations in functional brain structure are addressed. Therefore we ensure high accuracy and interpretability of the results. Extracted  $\widehat{Z}_t$  are tested for activation in the GLM (mixed-effects model) framework. For the studied population, we detect significant activation at aINS and DMPFC regions as correlates for risk, already reported in Mohr, Biele, and Heekeren (2010). Our approach yields similar results to the benchmark and is complimentary.

To deepen our understanding of changes in neural activity underlying risk preferences, we conducted a model-free analysis. The focus is on those ROIs that show ID-related effects: aINS (left and right) and DMPFC (see Table 5) which have previously been associated with decision making. More precisely, we explore the relation between average reaction to the stimulus in subject-specific loadings  $\widehat{Z}_t$  representing selected regions. Following Brown et al. (2014) we construct simple, model-free statistics that capture the peak of HRF:  $\overline{\Delta\widehat{Z}}_{\text{aINS}(l)}$ ,  $\overline{\Delta\widehat{Z}}_{\text{aINS}(r)}$ ,  $\overline{\Delta\widehat{Z}}_{\text{DMPFC}}$  and explore their explanatory power on the risk attitude  $\phi_i$ . The resulting regression model with brain dynamics as regressors achieves  $R^2 = 0.47$ . Changes in brain activity represented by  $\overline{\Delta\widehat{Z}}_{\text{DMPFC}}$  did not carry informative power for risk attitude. Simultaneously, both aINS regions are picked up to be statistically significant and reported  $p$ -values are  $\approx 0.01$ . We conclude that DMPFC, though activated by the risk of the investment, is not significantly correlated to risk attitudes. Dropping off all irrelevant terms and reestimating the regression model (12) yields  $R^2 = 0.37$ . This parsimonious and informative setup is used to predict the risk attitudes based only on fMRI information. The analysis is further refined adjusting for possible variation of hemodynamic response by adding the weights to the sequence of observations after stimulus.

We report, that neural predictions of risk attitudes, though satisfactory, do not mimic perfectly risk weights derived from subject investment decisions. One may claim that the applied mean-variance model does not reflect true risk attitudes adequately well and additional measures for subjective expected returns and perceived risk than mean and standard deviation should be introduced. Secondly, the risk preferences and neural responses identified in this study may not cover all the effects and brain reactions. Risk attitude is far more complex and may not be only localized in aINS. Therefore we plan to apply our methodology to a wide spectrum of similar studies for further investigations.

### Acknowledgments

The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through SFB 649 ‘‘Economic Risk’’ and IRTG 1792 ‘‘High Dimensional Non Stationary Time Series’’.

### Appendix

See (Figures 10–17 and Tables 4, 5).

#### *Simulation Study*

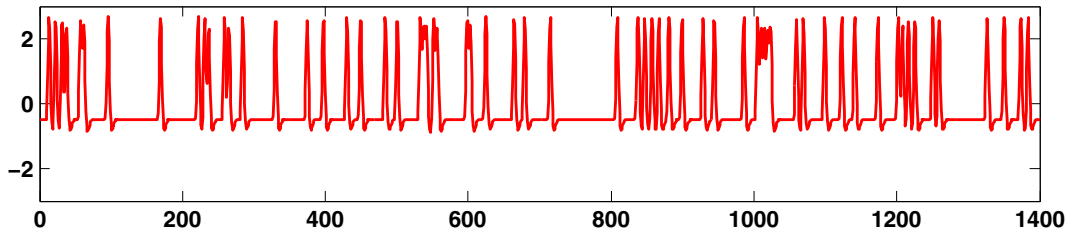


FIGURE 10.

Stimulus time series derived as a convolution of double Gamma hemodynamic response function and uncorrelated portfolio stimulus  $\times 64$  plotted against time (each 2 s).

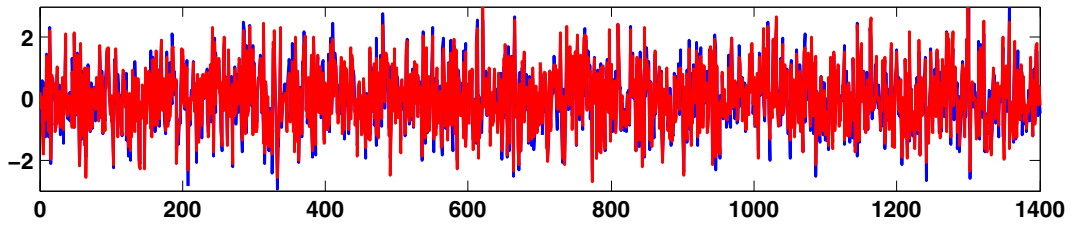


FIGURE 11.

Simulated spatially correlated Gaussian noise for 2 vertical neighbor voxels (*red* and *blue*) plotted against time (each 2 s);  $\text{Corr}_t(\varepsilon_{t,1}, \varepsilon_{t,2}) = 0.97$ .

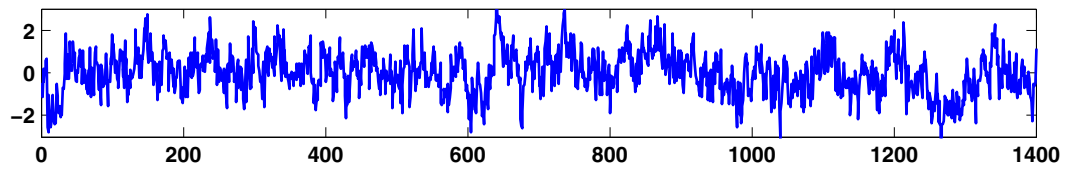


FIGURE 12.

Simulated stimulus time series as the AR(2) process:  $\tilde{Z}_t = 0.5\tilde{Z}_{t-1} + 0.2\tilde{Z}_{t-2} + \varepsilon_{AR,t}$ , plotted against time (each 2 s).

*Clustering and Sensitivity Analysis*

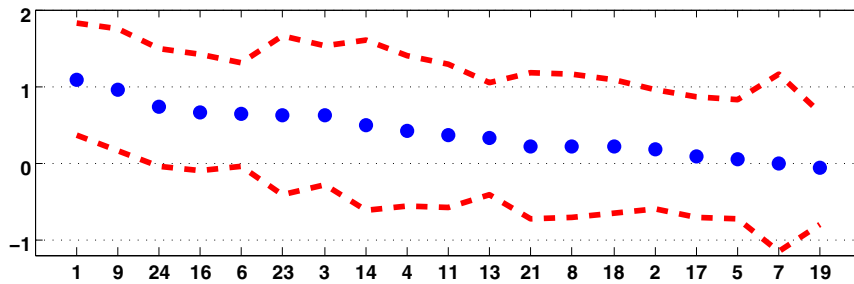


FIGURE 13.

Sensitivity analysis of the risk attitude  $\phi$ : estimates  $\hat{\phi}_i, i = 1, \dots, 19$  with 95 % confidence intervals.

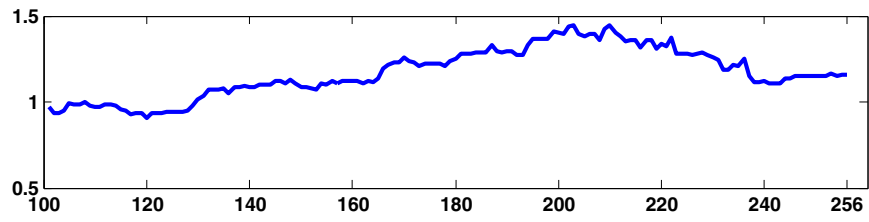


FIGURE 14.  
The derived risk attitude of subject 1 in a rolling window exercise ( $\hat{\phi}_i$  estimated from past 100 ID answers).

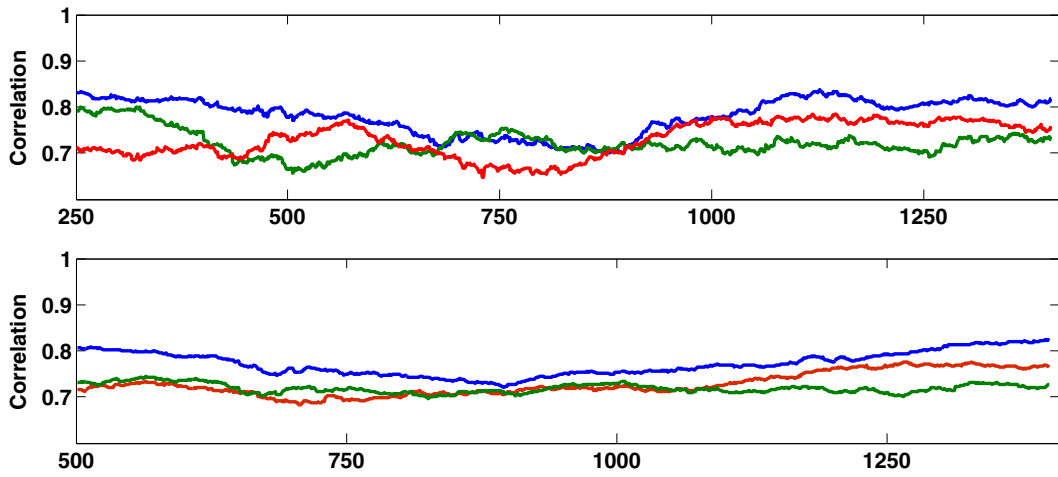


FIGURE 15.  
Time series of the correlation coefficient derived by the rolling window (250 top, 500 bottom) for the *center voxel* and: *horizontal*, *vertical diagonal* neighboring voxel for aNS(right) of subject 1.

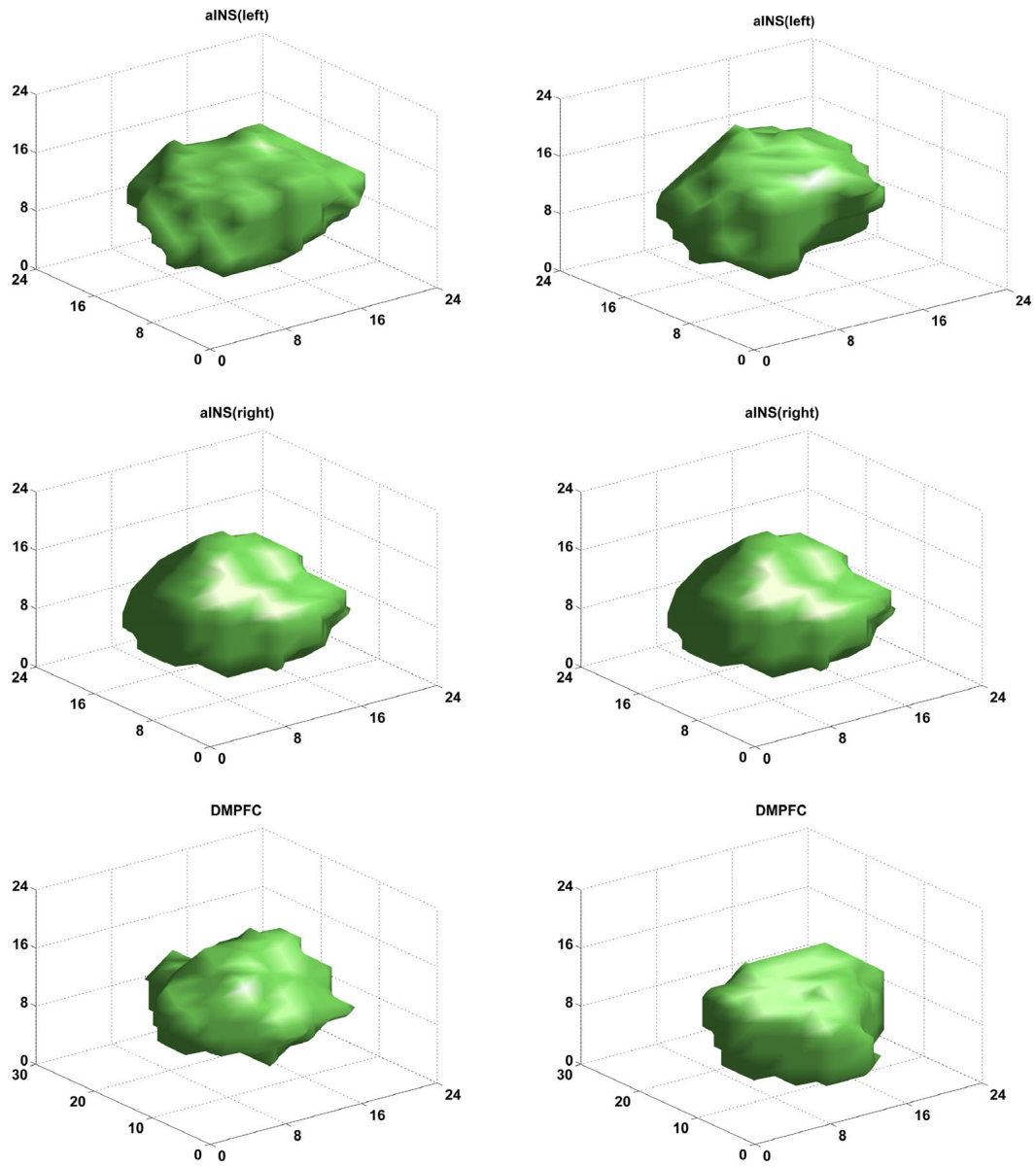


FIGURE 16.

Contour plots of derived aINS(*left*), aINS(*right*) and DMPFC (*upper, middle lower panel*) clusters for subjects 1 (*left*) and 19 (*right*), respectively; derived by the NCUT algorithm with  $C = 1,000$ .  $x, y, z$  axis denote the 3D space given in millimeters.



*Factor Loadings*

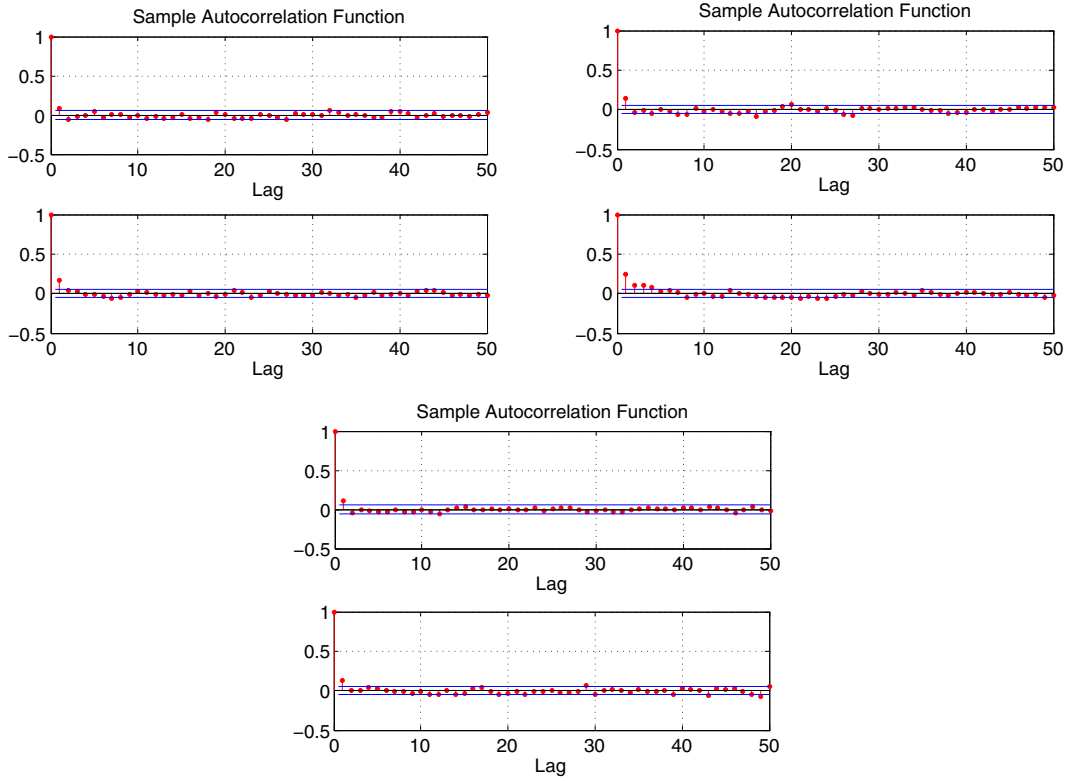


FIGURE 17.

Sample autocorrelation function of aINS(*left*), aINS(*right*), and DMPFC  $\widehat{Z}_t$  (*top left, top right, bottom panel*, respectively) for subjects 1 (*top*) and 19 (*bottom*), respectively.

TABLE 4.

KPSS, ADF test statistics for estimated factor loadings aINS(*left*), aINS(*right*), and DMPFC  $\widehat{Z}_t$ ; subject 1 (*left panel*), subject 19 (*right panel*) (KPSS:  $H_0$ : weak stationarity, critical values at 0.10, 0.05, 0.01 are 0.119, 0.146, and 0.216; ADF:  $H_0$ : unit root, critical values at 0.01, 0.05, 0.10 are  $-1.61$ ,  $-1.94$ , and  $-2.58$ ).

	aINS(l)	aINS(r)	DMPFC	aINS(l)	aINS(r)	DMPFC
KPSS	0.035	0.063	0.038	0.044	0.051	0.044
ADF	-0.128	-0.137	-0.110	-0.185	-0.207	-0.159

PSYCHOMETRIKA

TABLE 5.

The position of the cluster local maximum, denoted in the Montreal Neurological Institute (MNI) standard at 2mm resolution, corresponding Z-score (middle) and  $p$  value (bottom) of activated “risk” clusters during the ID stimuli.

	DSFM	Average	GLM
aINS(l)	(-34, 18, -8) 4.13 $3 \times 10^{-4}$	(-36, 18, -8) 4.08 $4 \times 10^{-4}$	(-32, 22, -12) 4.58 $3 \times 10^{-3}$
aINS(r)	(34, 24, -4) 4.39 $6 \times 10^{-6}$	(36, 18, -6) 4.21 $6 \times 10^{-7}$	(40, 22, -16) 5.24 $3 \times 10^{-7}$
DMPFC	(6, 24, 42) 4.43 $2 \times 10^{-9}$	(4, 24, 42) 3.88 $1 \times 10^{-8}$	(4, 24, 24) 4.56 $3 \times 10^{-7}$

Average stands for a mean value over voxels in each cluster (results of the NCUT parcellation with  $C = 1, 000$ ). Analysis done in the FSL (FEAT/FLAME) software.

References

- Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, 20(2), 1052–1063. doi:10.1016/S1053-8119(03)00435-X.
- Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2), 137–152. doi:10.1109/TMI.2003.822821.
- Beckmann, C. F., & Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage*, 25(1), 294–311. doi:10.1016/j.neuroimage.2004.10.043.
- Bernoulli, D. (1738). Specimen Theoriae Novae de Mensura Sortis. *Papers of the Imperial Academy of Sciences in Petersburg*, 5, 172–192.
- Brown, D. A., Lazar, N. A., Datta, G. S., Jang, W., & McDowell, J. E. (2014). Incorporating spatial dependence into Bayesian multiple testing of statistical parametric maps in functional neuroimaging. *NeuroImage*, 84(1), 97–112. doi:10.1016/j.neuroimage.2013.08.024.
- Camerer, C. F. (2007). Neuroeconomics: Using neuroscience to make economic predictions. *The Economic Journal*, 117(519), C26–C42. doi:10.1111/j.1468-0297.2007.02033.x.
- Camerer, C. F. (2013). Goals, methods, and progress in neuroeconomics. *Annual Review of Economics*, 5(1), 425–455. doi:10.1146/annurev-economics-082012-123040.
- Caraco, T. (1981). Energy budgets, risk and foraging preferences in dark-eyed juncos (*Junco hyemalis*). *Behavioral Ecology and Sociobiology*, 8(3), 213–217. doi:10.1007/BF00299833.
- Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8), 1914–1928. doi:10.1002/hbm.21333.
- Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. R., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3), 968–980. doi:10.1016/j.neuroimage.2006.01.021.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4), 189–210. doi:10.1002/hbm.460020402.
- Garrett, D. D., Samanez-Larkin, G. R., MacDonald, S. W., Lindenberger, U., McIntosh, A. R., & Grady, C. L. (2013). Moment-to-moment brain signal variability: A next frontier in human brain mapping. *Neuroscience and Biobehavioral Reviews*, 37(4), 610–624. doi:10.1016/j.neubiorev.2013.02.015.
- Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain* (2nd ed.). London: Academic Press. ISBN: 9780124160088.
- Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, 9(6), 467–479. doi:10.1038/nrn2374.
- Heller, R., Stanley, D., Yekutieli, D., Rubin, N., & Benjamini, Y. (2006). Cluster-based analysis of FMRI data. *NeuroImage*, 33(2), 599–608. doi:10.1016/j.neuroimage.2006.04.233.
- Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Reviews Neuroscience*, 10, 1625–1633.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292. doi:10.2307/1914185.
- Kamvar, S. D., Klein, D., & Manning, C. D. (2003). Spectral Learning. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence, IJCAI'03* (pp. 561–566). San Francisco: Morgan Kaufmann Publishers Inc. ISBN: 9780127056616.

- Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, 453(7197), 869–878. doi:[10.1038/nature06976](https://doi.org/10.1038/nature06976).
- Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4), 395–416. doi:[10.1007/s11222-007-9033-z](https://doi.org/10.1007/s11222-007-9033-z).
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77–91. doi:[10.1111/j.1540-6261.1952.tb01525.x](https://doi.org/10.1111/j.1540-6261.1952.tb01525.x).
- Mohr, P., Biele, G., & Heekeren, H. (2010). Neural processing of risk. *The Journal of Neuroscience*, 30(19), 6613–6619. doi:[10.1523/JNEUROSCI.0003-10.2010](https://doi.org/10.1523/JNEUROSCI.0003-10.2010).
- Mohr, P. N. C., Biele, G., Krugel, L. K., Li, S.-C., & Heekeren, H. R. (2010). Neural foundations of risk-return trade-off in investment decisions. *NeuroImage*, 49(3), 2556–2563. doi:[10.1016/j.neuroimage.2009.10.060](https://doi.org/10.1016/j.neuroimage.2009.10.060).
- Mohr, P. N. C., & Nagel, I. E. (2010). Variability in brain activity as an individual difference measure in neuroscience? *The Journal of Neuroscience*, 30(23), 7755–7757. doi:[10.1523/JNEUROSCI.1560-10.2010](https://doi.org/10.1523/JNEUROSCI.1560-10.2010).
- Park, B. U., Mammen, E., Härdle, W. K., & Borak, S. (2009). Time series modelling with semiparametric factor dynamics. *Journal of the American Statistical Association*, 104(485), 284–298. doi:[10.1198/jasa.2009.0105](https://doi.org/10.1198/jasa.2009.0105).
- Ruff, C. C., & Huettel, S. A. (2013). Chapter Experimental methods in cognitive neuroscience. *Neuroeconomics: Decision making and the brain* (2nd ed.). London: Academic Press. ISBN: 9780124160088.
- Shen, X., Papademetris, X., & Constable, R. T. (2010). Graph-theory based parcellation of functional subunits in the brain from resting-state fMRI data. *NeuroImage*, 50(3), 1027–1035. doi:[10.1016/j.neuroimage.2009.12.119](https://doi.org/10.1016/j.neuroimage.2009.12.119).
- Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888–905. doi:[10.1109/34.868688](https://doi.org/10.1109/34.868688).
- Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31), 13040–13045. doi:[10.1073/pnas.0905267106](https://doi.org/10.1073/pnas.0905267106).
- Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-D proportional system: An approach to cerebral imaging (Thieme Classics)*. Stuttgart: Thieme.
- van Bömmel, A., Song, S., Majer, P., Mohr, P. N. C., Heekeren, H. R., & Härdle, W. K. (2013). Risk patterns and correlated brain activities. Multidimensional statistical analysis of fMRI data in economic decision making study. *Psychometrika*. doi:[10.1007/s11336-013-9352-2](https://doi.org/10.1007/s11336-013-9352-2).
- van den Heuvel, M., & Mandl, R. (2008). Normalized cut group clustering of resting-state fMRI data. *PLoS ONE*, 3(4), e2001. doi:[10.1371/journal.pone.0002001](https://doi.org/10.1371/journal.pone.0002001).
- von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton: Princeton University Press.
- Weber, E. U., & Milliman, R. A. (1997). Perceived risk attitudes: Relating risk perception to risky choice. *Management Science*, 43(2), 123–144. doi:[10.1287/mnsc.43.2.123](https://doi.org/10.1287/mnsc.43.2.123).
- Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., et al. (2002). A general statistical analysis for fMRI data. *NeuroImage*, 15(1), 1–15. doi:[10.1006/nimg.2001.0933](https://doi.org/10.1006/nimg.2001.0933).
- Xu, Q., desJardins, M., & Wagstaff, K. (2005). Constrained spectral clustering under a local proximity structure assumption. In: *Proceedings of the 18th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 866–867). Palo Alto: AAAI Press. ISBN: 9781577352341.

# COPICA—independent component analysis via copula techniques

Ray-Bing Chen · Meihui Guo · Wolfgang K. Härdle ·  
Shih-Feng Huang

Received: 7 January 2012 / Accepted: 18 October 2013 / Published online: 22 March 2014  
© Springer Science+Business Media New York 2014

**Abstract** Independent component analysis (ICA) is a modern computational method developed in the last two decades. The main goal of ICA is to recover the original independent variables by linear transformations of the observations. In this study, a copula-based method, called COPICA, is proposed to solve the ICA problem. The proposed COPICA method is a semiparametric approach, the marginals are estimated by nonparametric empirical distributions and the joint distributions are modeled by parametric copula functions. The COPICA method utilizes the estimated copula parameter as a dependence measure to search the optimal rotation matrix that achieves the ICA goal. Both simulation and empirical studies are performed to compare the COPICA method with the state-of-art methods of ICA. The results indicate that the COPICA attains higher signal-to-noise

ratio (SNR) than several other ICA methods in recovering signals. In particular, the COPICA usually leads to higher SNRs than FastICA for near-Gaussian-tailed sources and is competitive with a nonparametric ICA method for two dimensional sources. For higher dimensional ICA problem, the advantage of using the COPICA is its less storage and less computational effort.

**Keywords** Blind source separation · Canonical maximum likelihood method · Givens rotation matrix · Signal/noise ratio · Simulated annealing algorithm

## 1 Introduction

Independent component analysis (ICA) is a recently developed multivariate statistical method, and can be treated as a generalization of principal component analysis (PCA). PCA is based on the eigenvalue decomposition of the covariance matrix, and projects data onto the eigenvectors of the covariance matrix. Although an eigenvalue decomposition of covariance yields only uncorrelated factors, together with Gaussian distributional assumption, the principal components are independent. However, the “independent” property will not hold if Gaussianity is violated. Non-Gaussianity of the independent components is a fundamental restriction of ICA, since one can only estimate the ICA model of Gaussian data up to an orthogonal transformation and the mixing matrix is not identifiable if there are more than two Gaussian independent components. Thus ICA targets on non-Gaussian samples. The main goal of ICA is to find linear transformations that map the observed multivariate time series into independent components (ICs). To accomplish the ICA goal, unlike the eigenvalue decomposition approach in PCA, ICs are estimated via an optimization problem, in

---

R.-B. Chen  
Dept. of Statistics, National Cheng Kung University, Tainan 701,  
Taiwan  
e-mail: [rbchen@stat.ncku.edu.tw](mailto:rbchen@stat.ncku.edu.tw)

M. Guo (✉)  
Dept. of Applied Math., National Sun Yat-sen University,  
Kaohsiung 804, Taiwan  
e-mail: [guomh@math.nsysu.edu.tw](mailto:guomh@math.nsysu.edu.tw)

W.K. Härdle  
Center for Applied Statistics and Economics,  
Humboldt-Universität zu Berlin, Berlin, Germany  
e-mail: [haerdle@wiwi.hu-berlin.de](mailto:haerdle@wiwi.hu-berlin.de)

W.K. Härdle  
Lee Kong Chian School of Business, Singapore Management  
University, Singapore, Singapore

S.-F. Huang  
Dept. of Applied Math., National University of Kaohsiung,  
Kaohsiung 811, Taiwan  
e-mail: [huangsf@nuk.edu.tw](mailto:huangsf@nuk.edu.tw)

which the statistical cross dependency among the extracted ICs is minimized. In practice, ICA has been successfully applied in blind source separation (Comon 1994), image denoising (Hyvärinen 1999b), natural image patch (Bell and Sejnowski 1995), single-trial EEG records (Tsai et al. 2006) and many other applications (see for example Lee 1998; Hyvärinen and Oja 2000; Abayomi et al. 2011).

There has been a wide development of interest in the computational technique of ICA in the past two decades. The ICA method can be formulated as optimization of an objective function which minimizes the cross-dependency among the components. The performance of the ICA method depends on the choice of objective function and the algorithm used for implementation of the optimization problem determines the speed of the ICA method. Various objective functions used in ICA include maximum likelihood, negentropy, higher order cumulants, kurtosis and mutual information. Several procedures and algorithms were proposed to search the independent components based on different objective functions and searching algorithms. The well known FastICA proposed by Hyvärinen and Oja (1997) was based on maximization of non-Gaussianity via measurements such as kurtosis and negentropy. Since the negentropy is always nonnegative and vanishes if and only if the signal is Gaussian, it can be used as a measure of distance to normality. And an approximative Newton iteration fixed-point algorithm is used to improve the computational efficiency of the FastICA which is faster than the gradient based methods. The details of FastICA can be found in Hyvärinen and Oja (2000). Bell and Sejnowski (1995) proposed a natural gradient ICA algorithm by minimizing the mutual information among the outputs, which can be considered as the Kullback-Leibler divergence (KL divergence) between the current joint density and the product of marginal densities. Their approach can also be treated as a maximum likelihood approach. Comon (1994) gave a contrast function for ICA by approximating the mutual information in terms of third-order and fourth-order cumulants. CuBICA, proposed by Blaschke and Wiskott (2004), improved Comon's algorithm by simplifying the corresponding contrast function. Bach and Jordan (2002) proposed the kernel independent component analysis which uses flexible kernels to model the dependence between the variables. Gretton et al. (2005) proposed another kernel independent criterion, the Hilbert-Schmidt Independent Criterion (HSIC), and the HSIC-based ICA contrast has a diagonal Hessian at independence. Then Shen et al. (2009) introduce an optimization method for HSIC, named FastKICA, and RADICAL (Learned-Miller and Fisher 2003) used an estimate of univariate entropies to find Jacobi rotations that make pairs of signals as independent as possible. Kirshner and Póczos (2008) used Schweizer-Wolff measure of dependence to search the independent components.

In this research, a new procedure called COPICA is proposed for ICA. In the COPICA procedure, the joint distribution of the components is modeled by copula functions. For better modeling of non-Gaussianity and other empirical facts such as heavy tail behavior of financial data, copulae have been introduced into the quantitative finance practice. The copula technique is based on the thought that every multivariate distribution can be seen as a coupling of a distribution function (on the unit cube) operating on the marginal distribution functions of each variable. This coupling function has been coined the name "copula" (Sklar 1959 and 1996). Copulae can be parameterized with low dimensional parameters and fitted to multivariate data by a variety of optimization techniques (Nelsen 2006). Copulae also provide a flexible family for modeling dependencies and include the product copula as the family element representing independence. An important property of the copula parameters is that in some cases they are also the tail dependence parameters. Hence, the estimates of the copula parameters provide direct parametric estimates of the tail dependence. In the proposed COPICA approach, we use the deviation between the fitted copula parameters and the copula parameters at independence as a measure of dependence, then define the corresponding divergence function used as the objective function in ICA. Thus the COPICA procedure combines ICA ideas from the engineering literature with the copula based research in quantitative finance. For parameter estimation, we use the historical empirical distribution in the estimation of marginal distributions then use the canonical maximum likelihood (CML) to estimate the copula parameters. A simulated annealing algorithm is used to minimize our divergence function to find the best recovered matrix. Since the marginal distribution is estimated by a nonparametric empirical estimate and the joint distribution is modeled by a parametric copula function, the proposed COPICA method can be viewed as a semiparametric ICA approach.

We took the advantage of copula to separate the parameter space of the full likelihood function into the copula parameter space and the marginal parameter space. If the margins are well fitted, then an estimator on the joint part (i.e. the copula parameters) can recover independence. In COPICA, we estimate the marginal distribution by the nonparametric empirical distribution. An advantage of estimating marginals using empirical distributions is that this procedure is relatively free of assumptions. And the empirical distribution has nice asymptotic properties including consistency and asymptotic normality. Since marginal distributions are estimated nonparametrically, the copula parameters are the only unknown parameters in COPICA. Based on the whitening data, our goal is to find the proper rotation matrix to recover the independent sources. To accomplish this goal, the divergence function is defined via the copula parameters. Given a rotation matrix,  $R$ , the estimations of copula parameters in divergence function are obtained via CML

approach based on the current empirical marginal distributions of the rotated data. Thus the copula parameter estimators and our divergence function are function of the rotation matrix,  $R$ . However, it is difficult to express the divergence function explicitly in terms of the rotation matrix (or rotation angles). Hence to solve our optimization problem w.r.t. rotation angles, the gradient based optimization approach cannot be used. Simulated annealing algorithm is a stochastic optimization method which does not need the gradient information. Of course, SA is not the only optimization approach to solve our target problem. Other possible approaches are pattern search, Gold search and other stochastic optimization approach, for example, genetic algorithm.

In addition to our COPICA method, copula based independent component analysis approach has also been proposed in Ma and Sun (2007), Abayomi et al. (2008, 2011). Abayomi et al. (2008, 2011) considered the objective function based on the mutual information via copula, which measures a norm between the estimator and the oracular value. Specifically, Abayomi et al. (2008) provided a theoretical foundation of mutual information based approach and a version of their norm was utilized in Abayomi et al. (2011). Their rotation matrix is obtained by minimizing the mutual information (distance) between parametric copula and independent marginals. In addition to the full parametric approach, they also proposed a semiparametric approach by using the empirical distributions for marginals. Two numerical approaches were introduced to obtain their rotation matrix. In their full model method, the mutual information is used as the objective function and the gradient type approach is applied to obtain the rotation matrix numerically. In their partite model approach, they use Singular Value Decomposition of the bivariate mutual information matrix, which is constructed via pairwise copula, to find the orthogonal transformation matrix.

Although, the COPICA method and Abayomi et al.'s approach both use copula to model the joint distributions of the components, the objective function and optimization algorithm are different, which are the two major components determining the performance and speed of the ICA method. In Abayomi et al. (2008, 2011), mutual information was used as the dependent measurement. For the ICA problem when independent signals are obtained, the joint density function is equal to the product of the marginal densities and the mutual information is zero. And the copula parameter (no matter which copula is fitted) equals to its independent parameter, consequently our COPICA objective function equals to zero which is the same as the mutual information. Hence although our norm is not generated directly from the mutual information, yet it achieves the same optimal point when independence are obtained.

In the next section, the detail procedure of COPICA is introduced. In Sect. 3, blind source separation examples are

demonstrated to illustrate the performance of our method. In Sect. 4, we compare the performance of COPICA with FastICA in terms of their signal to noise ratios (SNR) on the recovered signals for blind source separation problems. In Sect. 5, we compare COPICA method with nonparametric rank-based approaches. Both simulation and empirical studies will be performed to compare the COPICA method with the state-of-art methods of ICA. Our numerical results and empirical study also support the applicability of the proposed COPICA method. In summary, the comparison results show that:

- (1) The computational burden in determining the ICA transformation are the same for the COPICA and the FastICA.
- (2) The COPICA method attains higher SNR than the FastICA for near-Gaussian-tail sources on the recovered signals for the blind source separation problems. We also noted that the FastICA method sometimes fails to converge for near-Gaussian-tail sources.
- (3) The COPICA method is competitive with the ICA method via a nonparametric measure, Schweizer-Wolff  $\sigma_{SW}$  for bivariate sources. For higher dimensional case, the COPICA method attains higher SNR than the ICA method via Schweizer-Wolff  $\sigma_{SW}$  on the average and reduces significantly the storage space.

Finally conclusion is given in Sect. 6.

## 2 COPICA procedure

Assume we observe the  $n$  linear mixtures

$$X = (x_1, x_2, \dots, x_n)^T$$

of the  $n$  independent components  $S = (s_1, s_2, \dots, s_n)^T$ , that is  $X = AS$ , where  $A = (a_{ij})$  is the  $n \times n$  mixing matrix. Here we assume that  $A$  is full rank. The independent components  $s_j$ 's are latent random variables with zero mean which cannot be observed directly and the mixing matrix  $A$  is unknown. The goal of ICA is to find linear combination of the observed data  $X$ ,  $Y = BX$  such that the components of  $Y$ ,  $y_i$ 's, are as independent as possible. Here unlike PCA to obtain uncorrelated linear combination of  $x_i$ , to achieve the independence among  $y_i$ 's, the possible measurements are related to nonlinear transformations of  $y_i$ , for example, nonlinear correlation,  $E(f(y_i)g(y_j))$ , where  $f$  and  $g$  are two function and at least one is nonlinear (Hyvärinen and Oja 2000). Thus ICA can be treated as to remove the nonlinear dependence by using the linear transformation of data.

In addition to centralize the observed data, most of the ICA procedures, such as FastICA, whiten the observations first by the matrix  $W = \Sigma^{-1/2}$ , where  $\Sigma$  is the covariance matrix of  $X$ . That is, the components of  $Z = WX$  are uncorrelated with unit norm, i.e.  $\text{Cov}(Z) = I_n$ . The independent

components are obtained by multiplying the pre-whitened observations with an orthogonal matrix  $R$  such that the outputs  $Y = RZ$  are nearly statistically independent.

In this section, we first introduce the copula modeling of the joint dependence structure of the transformed components, then define the copula parameters as a measure of dependence. A rotation matrix representation of the orthogonal matrix  $R$  is also given. Finally, the COPICA procedure is introduced.

### 2.1 Copula model

According to Nelsen (2006), an  $n$ -dimensional copula is defined as follows.

**Definition 1** An  $n$ -dimensional copula  $C(\mathbf{u})$ , where  $\mathbf{u} = (u_1, \dots, u_n)$ , is a function from  $[0, 1]^n \rightarrow [0, 1]$  with the following properties:

1.  $C(\mathbf{u})$  is grounded, that is,

$$C(u_1, \dots, u_{i-1}, 0, u_{i+1}, \dots, u_n) = 0,$$

which means that the copula is zero if one of the arguments is zero, and  $C(1, \dots, 1, u, 1, \dots, 1) = u$ , which means that the copula is equal to  $u$  if one argument is  $u$  and all others are 1.

2.  $C(\mathbf{u})$  is  $n$ -increasing, that is, for each hyperrectangle  $B = \prod_{i=1}^n [x_i, y_i] \subseteq [0, 1]^n$ ,

$$\int_B dC(\mathbf{u}) = \sum_{\mathbf{z} \in \times_{i=1}^n \{x_i, y_i\}} (-1)^{N(\mathbf{z})} C(\mathbf{z}) \geq 0,$$

where  $\mathbf{z} = (z_1, \dots, z_n)$ ,  $\times_{i=1}^n \{x_i, y_i\}$  denotes the set of the vertices of  $B$ , and  $N(\mathbf{z})$  is the number of  $\{k : z_k = x_k\}$ .

Copula has recently become the most significant new tool to handle co-movement between markets in the field of finance and the analysis of current status data in biostatistics, because it provides a flexible way to connect the marginal distributions of individual component to their multivariate joint distribution. Sklar’s theorem provides the theoretical foundation for the application of copulae. Let  $F_{X_j}(x_j)$  denote the marginal distribution of  $X_j$ ,  $j = 1, \dots, n$ . Based on the work of Sklar (1959), there exists a copula function  $C$  such that

$$F_X(x_1, \dots, x_n) = C\{F_{X_1}(x_1), \dots, F_{X_n}(x_n); \theta\}, \tag{1}$$

where  $F_X(x_1, \dots, x_n)$  is the joint distribution of  $X = (X_1, \dots, X_n)$  and  $\theta = (\theta_1, \dots, \theta_d)$  denotes the copula parameters. In the case of independence, the joint distribution is the product of the marginal distributions, that is  $F_X(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$ . This corresponds to the product (independence) copula  $C(\mathbf{u}) = u_1 \cdots u_n$ . In the following, we introduce some well-known copula families which will be considered in this work as an illustration.

- (1) Gumbel copula:

$$C(u_1, \dots, u_n; \theta) = \exp\left[-\left\{\sum_{j=1}^n (-\log u_j)^\theta\right\}^{\frac{1}{\theta}}\right],$$

$$\theta \geq 1.$$

When the Gumbel parameter  $\theta = 1$ , it is the independence copula. The Gumbel copula is motivated by limit theorems for joint extremes (Kotz and Nadarajah 2000) and has for long played an important role in modeling distributions of extremes. The Gumbel copula can model upper tail dependence. For instance the bivariate Gumbel copula, the upper tail dependence of two random variables  $X_1$  and  $X_2$  is defined as

$$\begin{aligned} \lambda_U &= \lim_{v \rightarrow 1^-} P(F_{X_2}(X_2) > v \mid F_{X_1}(X_1) > v) \\ &= \lim_{v \rightarrow 1^-} (1 - 2v + C(v, v))/(1 - v) = 2 - 2^{1/\theta}, \end{aligned} \tag{2}$$

which is always positive for  $\theta > 1$ . In addition, the Gumbel copula can be rotated to change the direction of the tail dependence. For example, in the 2-dimensional case, the survival Gumbel copula, denoted by  $\hat{C}(u_1, u_2; \theta)$ , can be obtained by rotating a Gumbel copula by 180 degrees, that is,

$$\hat{C}(u_1, u_2; \theta) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2; \theta),$$

where  $C(u_1, u_2; \theta)$  is the 2-dimensional Gumbel copula. Thus, the survival Gumbel copula can be used to model lower tail dependence.

- (2) Clayton copula:

$$C(u_1, \dots, u_n; \theta) = \left(\sum_{j=1}^n u_j^{-\theta} - n + 1\right)^{-1/\theta}, \quad \theta > 0.$$

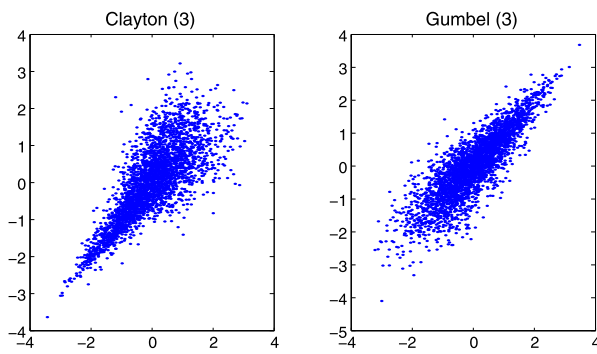
As the copula parameter  $\theta \rightarrow 0$ , the Clayton copula approaches to the independence copula. The Clayton copula can model multivariate lower tail dependence. For instance the bivariate Clayton copula, the lower tail dependence of two random variables  $X_1$  and  $X_2$  is defined as

$$\begin{aligned} \lambda_L &= \lim_{v \rightarrow 0^+} P(F_{X_2}(X_2) \leq v \mid F_{X_1}(X_1) \leq v) \\ &= \lim_{v \rightarrow 0^+} C(v, v)/v = 2^{-1/\theta}, \end{aligned} \tag{3}$$

which is positive for all  $\theta > 0$ . Similar to the Gumbel copula, the Clayton copula can also be used to depict the upper tail dependence by rotation.

- (3) Gaussian copula: for a given correlation matrix  $\Sigma \in R^{n \times n}$ , the Gaussian copula with parameter matrix  $\Sigma$  can be written as

$$C(u_1, \dots, u_n; \Sigma) = \Phi_\Sigma\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)\},$$



**Fig. 1** Bivariate plots of Clayton and Gumbel copulae with  $\theta = 3$  and  $N(0, 1)$  marginals (Color figure online)

where  $\Phi_{\Sigma}$  is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix  $\Sigma$  and  $\Phi^{-1}$  is the inverse cumulative distribution function of  $N(0, 1)$ . In particular, if the correlation matrix is the identity matrix, then Gaussian copula is the independence copula. Furthermore, if the  $X_j$ 's are normally distributed, then the Gaussian copula is correspond to the multivariate normal distribution. Gaussian copula is a popular and convenient type of copula, especially when the dimension is large. Since Gaussian copula depends only on the pairwise rank correlations between the marginals when the marginal are continuous (Mardia 1970), it continues to capture the dependence structure of the Normal-To-Anything (NORTA) distribution with arbitrary continuous marginal distributions (Ghosh and Henderson 2003). The bivariate Gaussian copula can model neither upper nor lower tail dependence, unless the correlation coefficient  $\rho = 1$ , since  $\lambda_U = \lambda_L = 0$  for  $\rho < 1$  and  $\lambda_U = \lambda_L = 1$  for  $\rho = 1$ .

In Fig. 1, we give the bivariate plots of random samples generated from Clayton and Gumbel copulae with parameter  $\theta = 3$  and  $N(0, 1)$  marginal distributions, respectively. Although the marginal distributions of the two cases are the same, different tail dependencies are displayed. Yet there are modeling limitations for the Gumbel and Clayton copulae (in general the family of Archimedean copulae) in higher-dimensions, as they imply exchangeability and hence equicorrelated ranks, which is obviously untenable in real application. For general reference of copulae, please refer to Nelsen (2006). For generalization of Archimedean copula models, see for example, McNeil and Nešlehová (2010) and Genest et al. (2011). For applications of copula in data mining, see Yu et al. (2011). For more discussion of the tail dependence parameters and the Gaussian copula, we refer to Schweizer and Wolff (1981) and Genest et al. (2011).

Let  $\{\phi(\cdot|\theta)\}_{\theta \in \Theta}$  be a family of copula densities, where  $\Theta \subset \mathfrak{R}^q$  is the parameter space. In this work, we use the

canonical maximum likelihood (CML) estimator to estimate the copula parameter  $\theta$  defined as below. Let  $\{X_t = (x_{1t}, x_{2t}, \dots, x_{nt})^\top\}_{t=1}^T$  is a realization of length  $T$  of the linear mixture  $X$ .

Step 1: Obtain  $\hat{F}_{X_i}(\cdot)$ ,  $i = 1, \dots, n$  are the empirical marginal distributions, and then

Step 2:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \sum_{t=1}^T \ln(\phi((\hat{F}_{X_i}(x_{it}), i = 1, 2, \dots, n)|\theta)), \quad (4)$$

Therefore, the fitted copula  $\phi(\cdot|\hat{\theta})$  with the CML estimator can be treated as the best approximation for the true copula  $\psi$  in the copula family  $\{\phi(\cdot|\theta)\}_{\theta \in \Theta}$  based on  $\{X_t = (x_{1t}, x_{2t}, \dots, x_{nt})^\top\}$ ,  $t = 1, \dots, T$ . In the proposed COP-ICA method, we consider the best approximations of the transformed data in the three copula families, Gumbel, Clayton and Gaussian copulae to capture its different dependence structure.

### 2.2 Representation of orthogonal matrices

In an ICA model, the following two ambiguities are well known to hold. Firstly because we can freely change the order of the components  $s_i$ 's, and call any of the independent components the first one, we cannot determine the order of the independent components. This ambiguity is insignificant in most applications though. Secondly we cannot determine the variances of the independent components. Thus, without loss of generality, we assume that each component of  $Y = RZ$  has unit variance. Then by independence assumption of  $Y$ , we have  $\text{Cov}(Y) = I_n$ . Therefore the transformation matrix  $R$  satisfies

$$RR^\top = R \text{Cov}(Z) R^\top = \text{Cov}(RZ) = \text{Cov}(Y) = I_n.$$

That is, the transformation matrix  $R$  is an orthogonal matrix which can be represented as the following product of the Givens rotation matrices,

$$R = \prod_{1 \leq i < j \leq n} G_{ij}(\beta_{ij}).$$

The matrix  $G_{ij}(\beta_{ij})$  is an  $n$ -dimensional Givens rotation matrix which represents a rotation in the plane spanned by the axes  $x_i$  and  $x_j$ ,  $i < j$ , with angle  $\beta_{ij}$ . Specifically,  $G_{ij}(\beta_{ij})$  is obtained by modifying the identity matrix so that the  $(i, i)$ ,  $(i, j)$ ,  $(j, i)$  and  $(j, j)$  elements of this matrix are respectively  $\cos \beta_{ij}$ ,  $\sin \beta_{ij}$ ,  $-\sin \beta_{ij}$ , and  $\cos \beta_{ij}$ , where  $\beta_{ij} \in [0, 2\pi)$ . This Givens matrix representation of  $R$  has been used in ICA algorithms, such as Comon (1994), Blaschke and Wiskott (2004), Kirshner and Póczos (2008) and so on. The product of the orthogonal matrix  $R$  and the whitening matrix  $W$ ,  $B = RW$ , is our objective transformation matrix of the observed data  $X$  to achieve independence.



The major task is to search the rotation angles,  $\beta_{ij}$ , to make the components of

$$Y = RZ = (RW)X = (RW)(AS)$$

nearly independent. In the bivariate case,  $n = 2$ , the Givens matrix is derived in the following proposition.

**Theorem 1** Assume  $S = (s_1, s_2)^\top$  is a random vector of two independent random variables with unit variance. Let  $X = AS$  where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is a non-degenerated mixing matrix. Let  $Z = WX$  and  $W = (AA^\top)^{-1/2}$  is the whitening matrix of  $X$ . Then the following Givens matrix of order 2,

$$G(\beta) = \begin{pmatrix} \cos \beta_{12} & \sin \beta_{12} \\ -\sin \beta_{12} & \cos \beta_{12} \end{pmatrix} \tag{5}$$

is the objective rotation matrix, that is

$$G(\beta)Z = \begin{cases} (s_1, s_2)^\top, & \text{if } ad - bc > 0, \\ (s_2, s_1)^\top, & \text{if } ad - bc < 0, \end{cases}$$

where

$$\begin{cases} \cos \beta_{12} = \frac{(a+d) \text{sign}(ac+bd)}{\sqrt{(a+d)^2+(b-c)^2}}, \\ \sin \beta_{12} = \frac{(-b+c) \text{sign}(ac+bd)}{\sqrt{(a+d)^2+(b-c)^2}}, \end{cases} \text{ if } ad - bc > 0,$$

or

$$\begin{cases} \cos \beta_{12} = \frac{(b+c) \text{sign}(ac+bd)}{\sqrt{(a-d)^2+(b+c)^2}}, \\ \sin \beta_{12} = \frac{(-a+d) \text{sign}(ac+bd)}{\sqrt{(a-d)^2+(b+c)^2}}, \end{cases} \text{ if } ad - bc < 0.$$

*Proof* First, consider the case  $ad - bc > 0$ . Since

$$G(\beta)Z = G(\beta)WX = G(\beta)(AA^\top)^{-1/2}AS = S,$$

it implies  $G(\beta) = A^{-1}(AA^\top)^{1/2}$ . Let  $U = \begin{pmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{pmatrix}$  be the positive definite matrix satisfying  $U^2 = AA^\top$ . We have

$$\begin{cases} u_{11}^2 + u_{12}^2 = a^2 + b^2, \\ u_{11}u_{12} + u_{12}u_{22} = ac + bd, \\ u_{12}^2 + u_{22}^2 = c^2 + d^2. \end{cases} \tag{6}$$

Combining with the constrains of  $ad - bc > 0$  and  $U$  be the positive definite matrix, the solutions of (6) are

$$\begin{aligned} u_{11} &= \frac{(a^2 + b^2 + ad - bc) \text{sign}(ac + bd)}{\sqrt{(a + d)^2 + (b - c)^2}}, \\ u_{12} &= \frac{|ac + bd|}{\sqrt{(a + d)^2 + (b - c)^2}}, \\ u_{22} &= \frac{(c^2 + d^2 + ad - bc) \text{sign}(ac + bd)}{\sqrt{(a + d)^2 + (b - c)^2}}. \end{aligned}$$

Thus,

$$\begin{aligned} G(\beta) &= A^{-1}(AA^\top)^{1/2} = A^{-1}U \\ &= \begin{pmatrix} \frac{(a+d) \text{sign}(ac+bd)}{\sqrt{(a+d)^2+(b-c)^2}} & \frac{(-b+c) \text{sign}(ac+bd)}{\sqrt{(a+d)^2+(b-c)^2}} \\ \frac{(b-c) \text{sign}(ac+bd)}{\sqrt{(a+d)^2+(b-c)^2}} & \frac{(a+d) \text{sign}(ac+bd)}{\sqrt{(a+d)^2+(b-c)^2}} \end{pmatrix}. \end{aligned}$$

Similarly, if  $ad - bc < 0$ , then  $G(\beta)Z = JS$ , where  $J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ . The solutions of (6) are

$$\begin{aligned} u_{11} &= \frac{(a^2 + b^2 + bc - ad) \text{sign}(ac + bd)}{\sqrt{(a - d)^2 + (b + c)^2}}, \\ u_{12} &= \frac{|ac + bd|}{\sqrt{(a - d)^2 + (b + c)^2}}, \\ u_{22} &= \frac{(c^2 + d^2 + bc - ad) \text{sign}(ac + bd)}{\sqrt{(a - d)^2 + (b + c)^2}}. \end{aligned}$$

Thus,

$$\begin{aligned} G(\beta) &= JA^{-1}(AA^\top)^{1/2} = JA^{-1}U \\ &= \begin{pmatrix} \frac{(b+c) \text{sign}(ac+bd)}{\sqrt{(a-d)^2+(b+c)^2}} & \frac{(-a+d) \text{sign}(ac+bd)}{\sqrt{(a-d)^2+(b+c)^2}} \\ \frac{(a-d) \text{sign}(ac+bd)}{\sqrt{(a-d)^2+(b+c)^2}} & \frac{(b+c) \text{sign}(ac+bd)}{\sqrt{(a-d)^2+(b+c)^2}} \end{pmatrix}. \end{aligned}$$

This completes the proof.  $\square$

Geometrically speaking, the rotation angle  $\beta_{12}$  represents the angle between one of the column vectors in the matrix  $(AA^\top)^{-1/2}A$  and the  $x_1$ -axis. In general for higher dimensional case, we have  $G(\beta) = A^\top(AA^\top)^{-1/2}$ , where  $\beta$  is the vector of the Givens rotation angles,  $\beta_{ij}$ . However the formula is not practically applicable, due to the fact that the matrix  $A$  is unknown in real applications. In order to determine the rotation angles of the Givens matrix, we will adopt a criterion based on copula parameter.

### 2.3 Divergence function based on copula parameter

Suppose  $X = (X_1, \dots, X_n)$  comes from the joint distribution,  $F_X$ . Then according to Eq. (1), we have that  $dC(x) = \frac{dF_X(x)}{\prod_i \{dF_{X_i}(x_i)\}}$ , where  $F_{X_i}$  is the marginal distribution of  $X_i$ . That is that the derivative of the copula is the ration of the joint density function and the product of the marginal density functions. Therefore, the copula parameters contain the information of the dependence among  $X$ . Furthermore, the mutual information for  $X$  can be re-presented via given copula  $C$  and its copula density  $\phi$  by

$$\begin{aligned} MI(X) &= \int \log \frac{dF_X(x)}{\prod_i \{dF_{X_i}(x_i)\}} dF_X(x) \\ &= \int_{I^n} \log(dC(u)) dC(u) \\ &= \int_{I^n} \phi(u|\theta) \log(\phi(u|\theta)) du, \end{aligned}$$

where  $I^n = [0, 1]^n$ . Once the independent copula parameters are obtained, the value of the mutual information is

zero. Thus the copula parameter  $\theta$  could be used as the measurement of the dependency. The similar idea was also mentioned in Abayomi et al. (2008). Another point comes from the relation of the tail dependence and copula parameters. As shown by (2) and (3), both the upper tail dependence  $\lambda_U$  of the bivariate Gumbel copula and the lower tail dependence  $\lambda_L$  of the bivariate Clayton copula are monotonic function of their copula parameters  $\theta$ . Therefore the copula parameters of the Gumbel copula and of the Clayton copula are also their tail dependence parameters. This is also supported us to use the copula parameters as the measure of dependence.

In ICA approach, we need to define an objective function for the source separation such that if the minimal value of this objective function is attended, then the recovered sources are independent. Since the copula parameters are used as the dependence measurement, we illustrate by the bivariate case in the following about how to choose the corresponding objective function for ICA problem. For a given (demean) realization  $\{X_t = (x_{1t}, x_{2t})^\top\}_{t=1}^T$ , select a copula function and a rotation angle  $\beta_{12}$ . Transform the whitening data  $Z = WX$  by the Givens rotation matrix  $R = G$  of the form given in (5), then compute the CML estimator  $\hat{\theta}$  based on transformed data. Let  $\theta_0$  denote the copula parameter at independence of the selected copula, for example  $\theta_0 = 1$  for the Gumbel copula. The magnitude

$$o(\hat{\theta}|\beta) = \|\hat{\theta} - \theta_0\|$$

is used as a measure of deviation from independence between  $x_{1t}$  and  $x_{2t}$  for this rotation angle  $\beta_{12}$ . Then search the angle  $\beta_{12}$  to minimize  $o(\hat{\theta}|\beta)$  which is regarded as the optimal solution of the Givens rotation matrix to make  $RZ = R(WX)$  nearly independent. In brief, we first find the best approximation of the true copula of the transformed data in a copula family, then measure the deviation from independence by the fitted copula (dependence) parameter. The objective rotation angle is obtained by minimizing the deviation from independence defined via copula parameters. In this study, we consider the best approximations of the true copula in the three copula families: Gumbel, Clayton and Gaussian. The three families are used to model upper tail, lower tail dependence structure and pairwise rank correlation between the marginals of the transformed components, respectively. Accordingly, the divergence function based on the three copula families is defined by the following weighted sum,

$$O(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3|\beta) = \sum_{i=1}^3 w_i o_i(\hat{\theta}_i|\beta), \tag{7}$$

where  $o_i(\hat{\theta}_i) = \|\hat{\theta}_i - \theta_{i0}\|$ ,  $\hat{\theta}_i$  and  $\theta_{i0}$  are respectively the fitted copula parameter and the independent parameter value of the  $i$ -th copula model, and  $w_i$ 's are the positive weights. In our simulation and empirical studies, we set the weights  $w_i$ 's to be inverse proportional to the standard deviations of

the CML estimators. In the implementation we will rotate the transformed components by the angles  $i\pi/2, i = 1, 2, 3$  to identify possible dependent structure, and include the corresponding measure of deviation from independence in the divergence function  $O$ .

The idea can be extended to higher dimensional case. For  $n$ -dimensional random variables  $Y_1, \dots, Y_n$  ( $n \geq 2$ ), mutual independence implies that any subset random variables of  $Y_1, \dots, Y_n$  are also mutually independent. Therefore, the divergence function measuring multivariate dependence of a selected copula function can be defined as

$$O(\hat{\theta}|\beta) = \sum_{\{i,j\} \subset \mathcal{N}} w_{ij} o(\hat{\theta}_{ij}|\beta) + \sum_{\{i,j,k\} \subset \mathcal{N}} w_{ijk} o(\hat{\theta}_{ijk}|\beta) + \dots + w_{\mathcal{N}} o(\hat{\theta}_{\mathcal{N}}|\beta), \tag{8}$$

where the vector parameter

$$\hat{\theta} = (\hat{\theta}_{12}, \dots, \hat{\theta}_{(n-1)n}, \hat{\theta}_{123}, \dots, \hat{\theta}_{\mathcal{N}}),$$

denote the copula parameter estimations of the transformed (pre-whitening and rotated via the rotation angle vector  $\beta$ ) data of

$$(Y_1, Y_2), \dots, (Y_{n-1}, Y_n), (Y_1, Y_2, Y_3), \dots, (Y_1, \dots, Y_n),$$

respectively. The divergence function  $O(\hat{\theta}|\beta)$  measure all multivariate dependence of dimensionality greater than or equal to 2. The components of the data are deemed nearly independent when  $O(\hat{\theta}|\beta)$  is close to zero. Similarly, multiple copula families can be included in the divergence function (8) as in (7).

Based on the chosen copulae and the pre-whitening data  $Z$ , the magnitude of the divergence function  $O(\hat{\theta}|\beta)$  given the rotation angle vector  $\beta$ , is computed in the following steps:

- (1) Rotate the data according to the rotation angle vector  $\beta$ ;
- (2) Find the currently empirical marginal distributions,  $\hat{F}_{Y_i}$ ;
- (3) Obtain CML estimator  $\hat{\theta}$  by minimizing Eq. (4) based on  $\hat{F}_{Y_i}$ ;
- (4) Compute  $O(\hat{\theta}|\beta)$ .

Thus the CML estimator  $\hat{\theta}$  is the function of the rotation angles  $\beta_{ij}, 1 \leq i < j \leq n$ . And the independent components are identified once  $O(\hat{\theta}|\beta)$  attains its minimum value in the rotation angle vector  $\beta$ . For brevity, we use  $O(\beta_{12}, \dots, \beta_{(n-1)n})$  to denote the objective function  $O(\hat{\theta}|\beta)$ , and then our ICA problem is equivalent to the minimization problem

$$\min_{\beta_{ij}, 1 \leq i < j \leq n} O(\beta_{12}, \dots, \beta_{(n-1)n}), \tag{9}$$

which means that we find the rotation angles  $\beta'_{ij}$ 's to minimize the divergence function  $O(\hat{\theta}|\beta)$  at the CML estimator  $\hat{\theta}$  with respect to  $\beta$ .

**Algorithm 1** COPICA by simulated annealing algorithm

- 
- (I) **[Initialization]**
- (1) Center the data  $X$  to make its mean zero, and obtain its sample covariance matrix  $\hat{\Sigma}$ .
  - (2) Whiten the data by setting  $Z = WX$  where  $W = \hat{\Sigma}^{-1/2}$ .
  - (3) Choose the copula families and define the objective function  $O(\beta)$ .
- (II) **[Optimization by simulated annealing algorithm]**
- (1) Select initial angles,  $\beta_i^{(0)}$ ; Choose the decreasing function  $T(t)$  and set  $t = 1$ .
  - (2) Repeat until  $t$  is large enough
    - (2.1) Run  $N_t$  iterations of the Gibbs sampler with  $\pi_{T(t)}(\beta)$  as its target distribution. Pass the final sample as  $\beta^{(t)}$ .
    - (2.2)  $t = t + 1$ .
  - (3) Identify the optimal angle vector  $\beta^*$  for  $O(\beta)$ .
- (III) **[Transformation matrix]**
- (1) Compute  $R = \prod_{1 \leq i < j \leq n} G_{ij}(\beta_{ij}^*)$
  - (2) The optimal transformation matrix  $B = RW$ .
- 

## 2.4 The details of the COPICA procedure

The COPICA procedure used to find the independent components of a given data set  $X$  is given in Algorithm 1, in which the optimization step is by the simulated annealing method.

There are two crucial steps in Algorithm 1: selection of the copulae models, and estimation of the copula parameter and search the best rotation angles. Discussion of the two steps are given below.

**Copula selection** It is well known that copulae are invariant with respect to strictly increasing transformations but not necessary to general linear transformation. Hence even if we can find the “true” copula family of the whiten random vector  $Z$ , the “true” copula after rotation might still fall in another family. Therefore, the key is not to find the “true” copula but to choose the proper copulae whose best approximations are useful in providing dependence measure under mis-specification situation. In general, prior information of the data or selection criteria are helpful to choose the copulae, for example, the Copula Information Criterion (CIC in short) proposed by Grønneberg and Hjort (2008). Due to the characteristic of the signals/sources in engineering or finance applications, heavy-tailed source is a widely used assumption in blind source separation (BSS) problems, for example, Kidmose (2001) and Chen and Wu (2007). In addition to BSS, the heavy-tailed assumption is also popular in the analysis of EEG signal (Tsai et al. 2006), natural image representation (Olshausen and Field 1996) and so on. The existence of tail dependence is a special feature of heavy-tailed signals/sources as well as an indication of non-independence. In this study, we utilize the tail dependence

feature of the Gumbel and Clayton copulae to estimate the dependency of the transformed data. The Gaussian copula is also considered to capture pairwise rank correlations between the marginals. In an extensive simulation study, we compare the copula based dependence measure with the nonparametric Kendall’s  $\tau$  in various settings of misspecified models. For the reason of concision, we only report the summary results here without detailed description. The results show that the dependence measured by the copula parameters is in good accordance with the Kendall’s  $\tau$ . And under misspecified models, the copula-based measure still provide valuable dependence information. The advantages of the copula based criterion over the Kendall’s  $\tau$  is its faster convergence rate and higher SNR values in ICA application. In addition, we will demonstrate the effectiveness of the three copula based dependency measure by blind source separation (BSS) examples in next section.

*Optimization procedure for identifying best rotation angles*

Recall from Eq. (4), we estimate the copula parameters by the CML estimator which gives the best approximation to the “true” copula in its family based on the transformed data. Due to the constraints of copula parameter estimation method, it is in general difficult to have closed form solution of the CML estimators. As a result there is no explicit form of the objective function  $O(\beta_{12}, \dots, \beta_{(n-1)n})$  even for low dimensional case. Derivative-free optimization methods, such as genetic algorithm, simulated annealing algorithm, direct search method and so on, can aid to solve the minimization problem defined in Eq. (9). Herein, the simulated annealing (SA) algorithm, proposed by Metropolis et al. (1953) and introduced as an optimization technique by Kirkpatrick et al. (1983), is used as an illustration to search these optimal angles. For simplicity of notation, we denote the rotation angles  $\beta_{12}, \dots, \beta_{(n-1)n}$  by  $\beta_1, \dots, \beta_q$ , where  $q = n(n-1)/2$ . First define a density

$$\pi_{T(t)}(\beta) \propto \exp\{-O(\beta)/T(t)\},$$

where  $O(\beta)$  is the objective function,  $\beta = (\beta_1, \dots, \beta_q)^\top$  and  $T(t)$  is the “temperature” at time  $t$  which is a decreasing function from initial temperature,  $T(0) > 0$ , to  $0^+$ . The key step of the SA algorithm is that for  $t$ , we run  $N_t$  iterations of the Gibbs sampler with  $\pi_{T(t)}(\beta)$  as its target distribution, and then choose the final sample as  $\beta^{(t)} = (\beta_1^{(t)}, \dots, \beta_q^{(t)})^\top$  that denotes  $\beta$  at time  $t$ . In order to speed up our optimization process, we use  $\exp(O(\beta))$  in the SA algorithm instead of  $O(\beta)$  directly. For more details about the SA algorithm, please refer to Liu (2001).

In implementing the COPICA with SA algorithm, the data is rotated by each sampled angles  $\beta_{ij}$ , and the copula parameter vector  $\theta$  are re-estimated based on the rotated data to compute the magnitude of the divergency function. And in the Gibbs sampler, the simple inversion method is

employed by using discretization of the continuous cumulative distribution function. Note here for this discretization method, suppose that we approximate the cumulative distribution function of  $\pi_{T(t)}(\beta_i)$  by  $K$  points,  $\beta_i^j, j = 1, \dots, K$ . Then for each point,  $\beta_i^j$ , CML approach is used to obtain the copula parameter estimator based on the rotated data with respect to  $\beta_i^j$ , and evaluate the corresponding objective function values. Finally we can have the approximated cumulative distribution function for  $\beta_i$ .

When we implement SA algorithm, we need to set the initial values of rotation angles,  $\beta$ , and the temperature,  $T(t)$ . For the initial  $\beta$ , we can simply set the  $\beta = \mathbf{0}$  for the initial angles or we can set  $\beta$  based on our prior information, for example, the angles obtained by the FastICA. Consider the temperature  $T(t)$ . In order to get the global optimal point, the temperature  $T(t)$  of the SA should decrease slowly such as  $O(\log(t)^{-1})$ , for details we refer to Liu (2001). However in practice, it is too slow to get the global optimal point and instead the linear or exponential temperature decreasing is used. In our COPICA, the temperature is chosen as  $O(t^{-1/4})$  which would lead to a reasonable convergent area quickly. From our simulations and real example results, it seems that this  $T(t)$  works well in our approach.

The complexity of Algorithm 1 can be analyzed as follows. First we consider the 2-dimensional situation and there is only one angle  $\beta_1$  needed to identify via SA algorithm. Then in each iteration of Gibbs sampler in SA algorithm, the complexity of the inversion method is  $O(KSCT \log(T))$ , where  $K$  is the number of points to obtain the approximation cumulative distribution function,  $T \log(T)$  is the complexity to sort each marginal source,  $C$  is the cost for maximization in CML method, and  $S$  is the number of the copula parameters used in our divergence function. Finally for general  $n$ -dimensional problem, the complexity of sweeping  $q = n(n - 1)/2$  angles in each iteration of the Gibbs sample is  $O(n^2KSCT \log(T))$ .

### 3 COPICA for blind source separation

We illustrate the performance of the proposed COPICA method by solving blind source separation (BSS) problems. Recently, blind source separation by ICA has received lots of attention because of its potential applications in signal processing such as in speech recognition systems, telecommunications and medical signal processing. In BSS problems, the observations  $\mathbf{x}_t = (x_{1t}, \dots, x_{nt})^T$  are assumed to be mixtures of  $n$  mutually independent sources,  $\mathbf{s}_t = (s_{1t}, \dots, s_{nt})^T$  at time  $t$ , that is

$$\mathbf{x}_t = A\mathbf{s}_t, \quad t = 1, \dots, T, \tag{10}$$

where  $A$  is an  $n \times n$  invertible mixing matrix. The goal of the BSS problem is to estimate the mixing matrix  $A$  and

recover the original sources  $\mathbf{s}_t$ , for given mixtures,  $\mathbf{x}_t, t = 1, \dots, T$  simultaneously. If the matrix  $A$  is invertible and known, then the independent sources can be recovered by  $A^{-1}\mathbf{x}_t, t = 1, \dots, T$ . While applying ICA methods to solve the BSS problems, the optimal transformation matrix will be the inverse matrix,  $A^{-1}$ , multiplying by a permutation matrix or a scaler.

In the following examples, we generate the sources  $\{\mathbf{s}_t\}_{t=1}^T$  independently from a mixture normal distribution or from natural sound signals. The observation vectors are then generated by Eq. (10) for a given mixing matrix,  $A$ . In order to measure the performance of the COPICA method, we consider the following signal/noise ratio (SNR) value

$$\text{SNR}_{s_i}(\hat{s}_i)[B] = 10 \log_{10}(\|s_i\|^2 / \|s_i - \hat{s}_i\|^2) \tag{11}$$

where  $s_i = (s_{it}, t = 1, \dots, T), i = 1, \dots, n$ , are the original signals from the sources,  $\hat{s}_i = (\hat{s}_{it}, t = 1, \dots, T), i = 1, \dots, n$ , are the recovered signals transformed by the matrix  $B$  found by the COPICA method, and  $\|\cdot\|$  denote the  $L^2$ -norm. Note that the columns of the inverse of the transformation matrix,  $B$ , will be proportional to the true mixing matrix  $A$ , and the signals are normalized for SNR computation. By the definition of SNR in Eq. (11), larger value of SNR indicates better performance. We consider  $\text{SNR} \geq 10$  as a threshold of high SNR value, see also Sodoyer et al. (2003). By its definition,  $\text{SNR} \geq 10$  is equivalent to  $\|s_i - \hat{s}_i\|^2 / \|s_i\|^2 \leq 10\%$ , which implies approximately at least 90 % of the signals are recovered by  $\hat{s}_i$ . Statistical reasoning of using 10 as high SNR value is also given below. If under independent and normal assumptions, we have roughly  $\|s_i\|^2 \sim \chi_T^2$  and  $\|s_i - \hat{s}_i\|^2 \sim \chi_{T-m}^2$ , where  $T \gg m$ , hence  $\frac{\|s_i\|^2/T}{\|s_i - \hat{s}_i\|^2/(T-m)} \sim F_{T, T-m}$ . Since the probability of the event  $\{\|s_i - \hat{s}_i\|^2 / \|s_i\|^2 \leq 10\% \} (\equiv \text{SNR} \geq 10)$  is very small for large  $T$ , it is reasonable to consider 10 as a large SNR value.

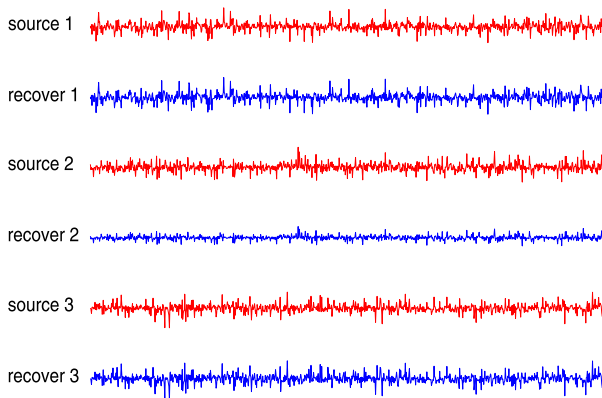
*Example 1* Three sources are generated independently from the following mixture normal density,

$$f(s_i) = 0.7 f_{N(0,1)}(s_i) + 0.3 f_{N(0,3^2)}(s_i),$$

where  $f_{N(\mu, \sigma^2)}$  is the density function of the normal distribution with mean  $\mu$  and variance  $\sigma^2$ . That is each sample is generated from  $N(0, 1)$  with probability 0.7 and from  $N(0, 3^2)$  with probability 0.3. The mixing matrix  $A$  is set to be

$$\begin{pmatrix} 1.0000 & -2.0000 & 1.0000 \\ -1.0000 & 1.0000 & 2.0000 \\ -1.0000 & 1.0000 & 1.0000 \end{pmatrix}. \tag{12}$$

Two copulae, Gumbel and Clayton are used to measure the tail dependence. The objective function is set to be



**Fig. 2** The simulation results for a three dimensional blind source separation problem with mixture normal sources. The red lines are the original sources, and the blue lines are the recovered signals (Color figure online)

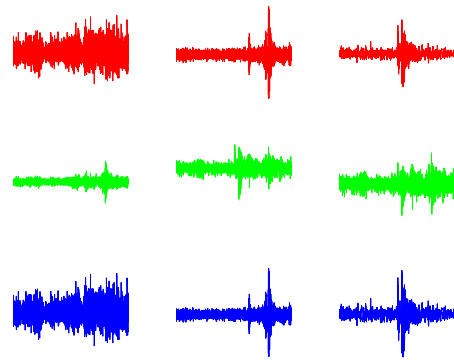
$$\begin{aligned}
 O(\hat{\theta}) = & \omega_1 * |\hat{\theta}_{123,Gumbel} - 1| \\
 & + \omega_2 * \sum_{i < j} |\hat{\theta}_{ij,Gumbel} - 1| \\
 & + \omega_3 * \sum_{i < j} |\hat{\theta}_{ij,Clayton}| \\
 & + \omega_4 * \sum_{i < j} |\hat{\theta}_{ij,Gaussian}|,
 \end{aligned} \tag{13}$$

where the weights  $(\omega_1, \omega_2, \omega_3, \omega_4) = (200, 300, 200, 500)$  are chosen to be inverse proportional to the standard deviations of the CML estimators of the copula parameters. After 100 iterations of the COPICA algorithm, the inverse transformation matrix found by the COPICA procedure is

$$B^{-1} = \begin{pmatrix} 1.9974 & -3.5685 & 1.7266 \\ -1.9178 & 1.8719 & 3.6426 \\ -1.9304 & 1.8337 & 1.8368 \end{pmatrix}.$$

Note that each column of this matrix is approximately proportional to the corresponding column of the genuine mixing matrix  $A$ , and the three recovered signals give high SNR values, 25.3238, 26.0529 and 32.8434. Figure 2 shows the original source signals and the recovered signals, which also illustrates high similarity between the two signals. The results show that the COPICA method successfully solve this simulated BSS problem.

**Example 2** In this example we demonstrate a real case with one near-Gaussian-tail signal. Three natural sounds of thunder, water and fire each containing 5000 sample points are used as the original signals. The sample kurtosises of these three natural sounds are 3.5323, 29.4978 and 16.6685 respectively. Note that the  $p$ -values of the Jarque-Bera test for these natural sounds are all less than  $10^{-3}$ , which indicates non-Gaussianity. The first source (thunder sound) is a near-Gaussian-tail sample since its sample kurtosis is close to 3,



**Fig. 3** The numerical results for three dimensional blind source separation problem with three natural sounds (thunder, water and fire). The red lines are the original sounds, green lines are their mixtures, and the blue lines are the recovered signals (Color figure online)

while the other two sources (water and fire sounds) are of heavy-tailed distributions. Using the same mixing matrix in Eq. (12) and the objective function defined in Eq. (13), after 100 iterations of the COPICA method, we obtained

$$B^{-1} = \begin{pmatrix} 1.8946 & -2.0933 & 0.7112 \\ -1.8875 & 0.9933 & 1.5947 \\ -1.8888 & 1.0118 & 0.8202 \end{pmatrix},$$

and the corresponding SNR values are 35.6150, 35.2765 and 32.3912. We also found great similarity in the original natural sounds and the recovered signals shown in Fig. 3.

**Example 3** In this example we demonstrate a real case with three near-Gaussian-tail signals. Three sounds with 10000 sample points, boat engine, rain and wind, are used as the original signals. The values of their sample kurtosis are 3.20, 3.23 and 3.71, respectively. Note that the  $p$ -values of the Jarque-Bera test for the signals are all less than  $10^{-3}$ , which indicates non-Gaussianity.

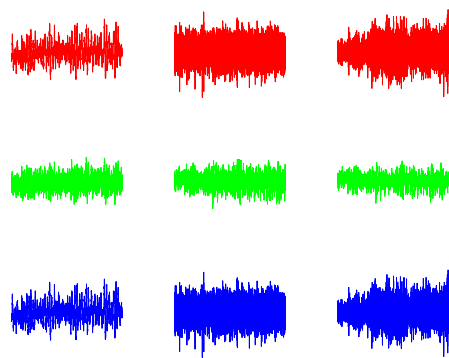
The mixing matrix  $A$  and the objective function are the same as in Example 2. After 100 iterations of the COPICA method, we obtain

$$B^{-1} = \begin{pmatrix} 5.6205 & -3.3402 & 2.1439 \\ -6.0068 & 1.6872 & 3.4948 \\ -5.8871 & 1.7050 & 1.6141 \end{pmatrix},$$

and the SNR values are 25.8270, 35.3203 and 23.8432 respectively. The time plots of the original natural sounds and the recovered signals are given in Fig. 4, again the result show that the COPICA method successfully separate the original natural sounds from their mixtures.

#### 4 Comparisons with the FastICA

The FastICA (Hyvärinen and Oja 1997; Hyvärinen 1999a) is one widely used and efficient method for identifying independent components. The FastICA is a two-step method.



**Fig. 4** The numerical results for three dimensional blind source separation problem with three natural sounds (boat engine, rain and wind). The red lines are the original sounds, green lines are their mixtures, and the blue lines are the recovered signals (Color figure online)

**Table 1** The kurtoses of the mixture normal distributions with  $\sigma_1 = 1$  and  $\sigma_2 = 3$

	Near-Gaussian-tailed		Heavy-tailed		
	$p = 0.1$	$p = 0.2$	$p = 0.4$	$p = 0.6$	$p = 0.7$
Kur.	3.2570	3.5610	4.3698	5.6122	6.4879

After whitening the data at the first step, the FastICA find the independent components based on a fixed-point iteration scheme for finding a maximum of the non-Gaussianity of a linear projection. And the kurtosis or negentropy is used as the measure of non-Gaussianity. The computer program of the FastICA is available at the web-site,

<http://www.cis.hut.fi/projects/ica/fastica/>.

In this section, we compare the performance of COPICA and FastICA for the BSS problems via simulation study. The original independent sources are generated from mixture normal distributions with the pre-specified parameters  $\sigma_1, \sigma_2$  and  $p$ . The corresponding kurtosis is  $3\{p\sigma_1^4 + (1-p)\sigma_2^4\}/\{p\sigma_1^2 + (1-p)\sigma_2^2\}^2$ . Thus we can generate samples with different kurtosis by choosing proper values of  $p, \sigma_1$  and  $\sigma_2$ . In the following,  $(\sigma_1, \sigma_2)$  is set to be  $(1, 3)$  and  $p = 0.1, 0.2, 0.4, 0.6, 0.7$  respectively, and the corresponding kurtoses are given in Table 1.

Three dimensional BSS problem is considered for comparison. At each replication, three original sources are generated independently from the same mixture normal distribution with sample size  $T$ , and the observations are obtained by mixing the original sources with the matrix  $A$  given in Eq. (12). Two sample sizes  $T = 1000$  and  $T = 5000$  are considered and 100 replications are performed. Since there are three signals, 300 SNR values are obtained for each sample size. For each value of  $p = 0.1, 0.2, 0.4, 0.6, 0.7$ , we report the medians and standard deviations of the 300 SNR values obtained respectively by the two methods. The results

**Table 2** Medians and standard deviations of the SNR for the BBS problem with three mixture-normal sources mixed by a fixed matrix, where  $N_1$  denotes the number of sources whose COPICA SNR values are larger than the FastICA SNR values,  $N_2$  is a  $3 \times 1$  vector whose components denote the non-recovery numbers of FastICA for each source, and  $N_3$  (or  $N_4$ ) is a  $3 \times 1$  vector with each component representing the number that the COPICA (or FastICA) SNR values are less than 10 (including the number of non-recovery)

		Median	std.	$N_1$ $N_2$	$N_3$ $N_4$
$T = 1000$	$p = 0.1$	28.05 (7.56)	4.78 (6.14)	297 (15, 16, 11)	(0,0,0) (74, 64, 75)
	$p = 0.2$	27.67 (14.62)	4.44 (6.59)	290 (5, 4, 2)	(0, 0, 0) (19, 19, 26)
	$p = 0.4$	27.73 (18.63)	4.66 (5.77)	274 (0, 0, 0)	(0, 0, 0) (1, 3, 5)
	$p = 0.6$	27.39 (22.62)	4.19 (6.98)	219 (0, 0, 0)	(0, 0, 0) (1, 0, 1)
	$p = 0.7$	27.05 (22.82)	4.08 (6.25)	225 (0, 0, 0)	(0, 0, 0) (1, 1, 0)
	$T = 5000$	$p = 0.1$	31.81 (15.09)	4.51 (6.52)	290 (1, 1, 2)
$p = 0.2$		32.30 (21.44)	4.15 (6.34)	280 (0, 0, 0)	(0, 0, 0) (0, 0, 1)
$p = 0.4$		31.43 (26.27)	4.90 (6.11)	225 (0, 0, 0)	(0, 0, 0) (0, 0, 0)
$p = 0.6$		30.45 (29.35)	4.21 (5.54)	162 (0, 0, 0)	(0, 0, 0) (0, 0, 0)
$p = 0.7$		30.98 (29.38)	4.93 (5.69)	184 (0, 0, 0)	(0, 0, 0) (0, 0, 0)

are given in the first two columns of Table 2. The reason for reporting the medians instead of the means is to avoid the case of non-recovery (the FastICA method sometimes cannot recover the original sources for near-Gaussian-tailed case). We also compute the number of sources whose COPICA SNR values are larger than the FastICA SNR values denoted by  $N_1$ . And let  $N_2$  be a  $3 \times 1$  vector whose components denote the non-recovery numbers of FastICA for each source, and let  $N_3$  (or  $N_4$ ) be a  $3 \times 1$  vector with each component representing the number that the COPICA (or FastICA) SNR values are less than 10 (including the number of non-recovery). The results of  $N_1$ – $N_4$  are given in the third and fourth columns of Table 2.

We summarized the results by the tail type of the original sources. The distribution is referred to “near-Gaussian-tailed” if the kurtosis is less than 4, to “heavy-tailed” if the kurtosis is greater than or equal to 4. In all cases, the COPICA method gives larger SNR medians and smaller standard deviations than the FastICA method. Note that there are 300 original sources for each pair  $(p, T)$ , since all the values of  $N_1 > 150$ , the COPICA method attains higher SNR

values more than half of all time. Significant dominance in the SNR medians and  $N_1$  of the COPICA over the FastICA is apparent for smaller sample size ( $T = 1000$ ) and near-Gaussian-tail case  $p = 0.1, 0.2$ . The non-recovered number of the FastICA method,  $N_2$ , are noted when  $p = 0.1, 0.2$ ,  $T = 1000$  and  $p = 0.1$ ,  $T = 5000$ , which indicates the FastICA method might fail to recover the near-Gaussian-tailed signals. All the values of  $N_3$  are equal to zero, implies the SNR values obtained by the COPICA are greater than 10 for all cases. Moreover, there are significant times ( $N_4$ ) that the FastICA attains small SNR ( $\leq 10$ ) values for the near-Gaussian-tailed case  $p = 0.1, 0.2$ ,  $T = 1000$  and  $p = 0.1$ ,  $T = 5000$ . Based on the above, we conclude that for all generated sources, COPICA successfully identifies the three independent components, while FastICA works well for heavy-tailed sources, but may fail for the near-Gaussian-tailed sources. The reason might be due to the criterion of the FastICA is based on the kurtosis and negentropy which is not sensitive to near Gaussian-tailed distributions. However, the signals with kurtosis close to 3 do exist in real application. Recall the sample kurtosis of thunder, boat engine, rain and wind sounds in Examples 2 and 3 are all close to 3. We further applied the FastICA method to these two real sound examples. For the case with one near-Gaussian-tailed and two heavy-tailed signals (Example 2), the inverse matrix found by the FastICA is

$$B_{\text{FastICA}}^{-1} = \begin{pmatrix} 1.8889 & -2.0828 & 0.7548 \\ -1.8882 & 1.0274 & 1.5716 \\ -1.8880 & 1.0311 & 0.7970 \end{pmatrix},$$

and the SNR's are 37.0375, 51.6663 and 46.4812 which are all larger than those obtained by the COPICA method. While for the case with three near-Gaussian-tail signals (Example 3), the inverse transformation matrix found by the FastICA of is

$$B_{\text{FastICA}}^{-1} = \begin{pmatrix} 5.7296 & -3.3709 & 1.7740 \\ -5.5537 & 2.1166 & 3.9765 \\ -5.6351 & 2.0033 & 2.0973 \end{pmatrix},$$

and the SNR's are 18.3674, 18.3564 and 22.6044 which are all smaller than those found by the COPICA method. The results of the real sound examples also support the aforementioned simulation findings. Finally from Table 2, one can see that both methods improve their SNR median values when the sample size increases from  $T = 1000$  to  $T = 5000$ .

In addition, we also compare the performance of the two methods by using random mixing matrix. The original sources are generated independently from a mixture-normal distribution with  $\sigma_1 = 1$ ,  $\sigma_2 = 3$  and  $p \in \{0.1, 0.2, 0.4, 0.6, 0.7\}$ . The size of each source is set to be 1000. However, in each replication, each component of the mixing matrix,  $A$ , is generated from  $[-5, 5]$  uniformly such that  $A$  is invertible. That is the mixing matrix is different for each replication. The 100 simulation results are shown in

**Table 3** Medians and standard deviations of the SNR for the BBS problem with three mixture-normal sources mixed by a random matrix, where  $N_1, \dots, N_4$  are defined the same as in Table 2

	Median	std.	$N_1$ $N_2$	$N_3$ $N_4$
$p = 0.1$	27.97 (7.07)	4.82 (6.32)	298 (20, 18, 12)	(0, 0, 0) (72, 63, 66)
$p = 0.2$	27.95 (13.97)	4.31 (6.15)	288 (0, 1, 1)	(0, 0, 0) (19, 21, 23)
$p = 0.4$	27.51 (19.64)	4.31 (6.34)	260 (0, 0, 0)	(0, 0, 0) (4, 1, 4)
$p = 0.6$	27.90 (21.64)	4.16 (6.37)	236 (0, 0, 0)	(0, 0, 0) (0, 0, 0)
$p = 0.7$	27.67 (22.64)	4.46 (6.39)	241 (0, 0, 0)	(0, 0, 0) (1, 1, 0)

Table 3. From Table 3 similar conclusions are obtained as from Table 2. That is COPICA recovers all original sources from their mixtures but FastICA might be fail for some near-Gaussian-tail sources, and overall COPICA attains higher SNR than FastICA, especially for the cases of near-Gaussian-tail sources.

The infomax principle, maximizing the output entropy of a neural network with nonlinear outputs, has been applied to develop ICA algorithm in Bell and Sejnowski (1995), and this principle is closely related to the maximum likelihood approach. Hyvärinen (1999a) pointed out that the fixed-point scheme in FastICA can be directly applied to infomax type ICA algorithm by choosing the corresponding nonlinearity  $g$ , for example,  $g(y) = -2 \tanh(y)$  for heavy-tailed sources. We also studied the performance of the FastICA using the infomax principle with  $g(y) = -2 \tanh(y)$  for the three dimensional BSS problem with different mixture normal sources and the mixing matrix  $A$  given by Eq. (12). Since the results are similar to Table 2, thus details are omitted here.

### 5 COPICA vs. nonparametric rank-based approach

In this section, we compare the COPICA method with several nonparametric rank-based ICA approaches via simulation studies. Many non-linear dependence measures for a pair of continuous random variables  $(X, Y)$  are based on ranks. Among most commonly used are Kendall's  $\tau$  and Spearman's  $\rho$ . Kendall's  $\tau$ , is defined as the difference between probability of concordance and probability of discordance. Spearman's  $\rho$  is defined as the Pearson's correlation coefficient between the ranks of the two samples and for a given copula model. More details of these two measures can be found in Nelsen (2006). Another nonparametric approach for measuring the dependence is the Blomqvist's  $\beta$  (Schmid

and Schmidt 2007). Recently, Kirshner and Póczos (2008) suggested using the Schweizer-Wolff  $\sigma_{SW}$ , defined as

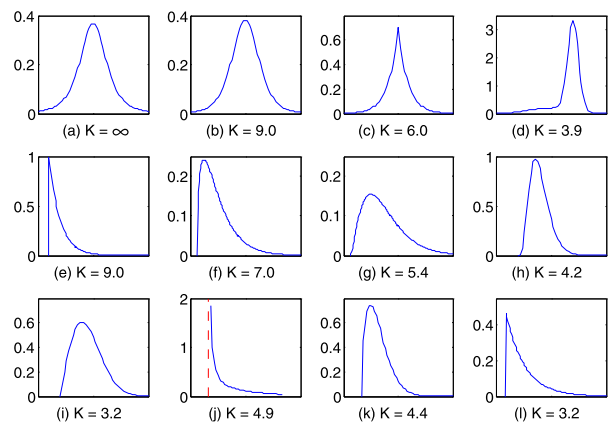
$$\sigma_{SW} = 12 \int_{[0,1]^2} |C(u, v) - uv| dudv, \tag{14}$$

to measure the pairwise dependence (Schweizer and Wolff 1981). They proposed an algorithm for ICA by replacing the copula function in (14) with empirical copula.

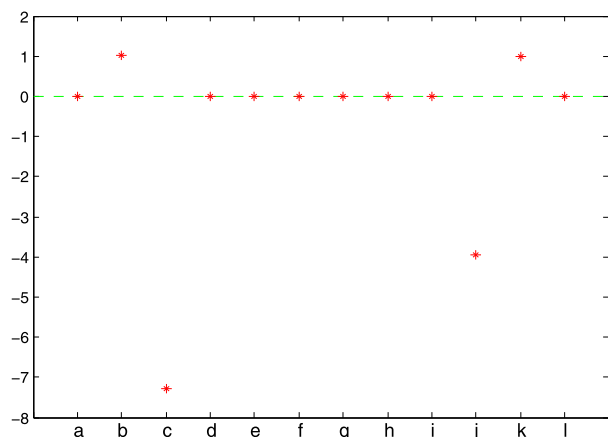
Most of the ICA algorithms use an approximation to mutual dependence as their objective functions. And the performance of an ICA algorithm depends on how accurate the approximate dependence measure is. The above four non-parametric dependence measures are all zero if  $X$  and  $Y$  are independent. However, the converse is not necessary true. Kirshner and Póczos (2008) showed that  $\sigma_{SW}$  is more robust than Kendall’s  $\tau$  and Spearman’s  $\rho$  with added outliers and noise. However, to obtain the nonparametric multivariate empirical distribution requires an intensive computational effort when sample size is large or dimensionality is high. It might even collapse when dimension is too high say higher than 4. A semiparametric approach such as COPICA, which estimates the joint distribution via copula function and one dimensional empirical distribution, can provide an alternative to greatly relieve the computational burden.

We conduct several simulation studies to compare the performance of COPICA with nonparametric ICA methods based on Kendall’s  $\tau$ , Spearman’s  $\rho$ , Blomqvist’s  $\beta$ , and Schweizer-Wolff  $\sigma_{SW}$ . Basically COPICA method attains higher SNRs than the ICA methods based on Kendall’s  $\tau$ , Spearman’s  $\rho$  or Blomqvist’s  $\beta$ , and is competitive with the ICA method via Schweizer-Wolff  $\sigma_{SW}$ . To save the space, we only show comparison between COPICA and ICA method via Schweizer-Wolff  $\sigma_{SW}$  (ICA\_SW). Note that the ICA-SW used here is similar to a ICA algorithm proposed by Kirshner and Póczos (2008). In the simulation study to compare the ICA performance of COPICA and ICA-SW, various types of heavy-tailed sources are used to generating original independent sources. Similar to the experimental setting of Bach and Jordan (2002), we consider 12 different one-dimensional densities with kurtosis greater than 3, shown in Fig. 5, including those densities commonly used in finance (a)–(d), in reliability and lifetime modeling (e)–(h) and (k) and in communications (i), (j) and (l).

For the bivariate case, we generate two independent sources each of size 1,000 from the same density, normalize the sources, and then mix them by a matrix whose elements are randomly sampled from  $[-5, 5]$ . We compute the SNR’s of the COPICA and ICA\_SW for the 12 heavy-tailed sources, respectively. Figure 6 plots the medians of the differences in the SNRs of COPICA and ICA\_SW (COPICA-ICA\_SW) based on 100 replications. Since most of the medians in Fig. 6 are around zero, the results indicate the ICA performance of the semiparametric COPICA method is competitive with the nonparametric ICA\_SW method. Also,



**Fig. 5** Probability density functions of heavy-tailed sources. (a) Student  $t$  with 3 degrees of freedom (d.f.); (b) Student  $t$  with 5 d.f.; (c) double exponential distribution; (d) mixture of two Gaussians, where the density is  $f(x) = 0.5\phi(x + 0.5) + \phi(2x - 1)$ ; (e) exponential distribution; (f) Chi-square distribution with 3 d.f.; (g) Chi-square distribution with 5 d.f.; (h) gamma distribution; (i) Rayleigh distribution; (j) Nakagami distribution; (k) Weibull distribution; (l) Rician distribution (Color figure online)



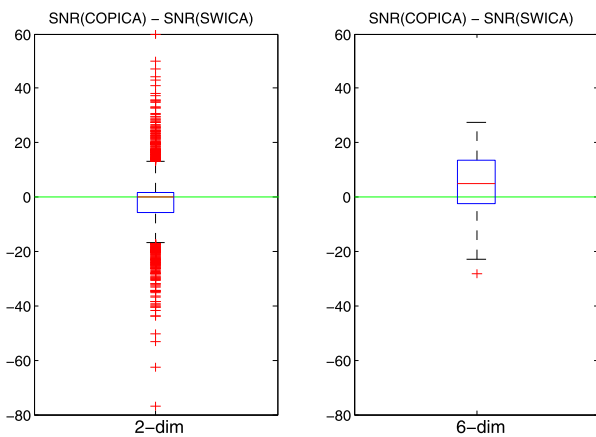
**Fig. 6** The medians of SNR(COPICA)-SNR(ICA\_SW) abased on 100 replications, where the two original independent sources are generated from the same density in (a)–(l) with sample size 1,000 (Color figure online)

in Table 4 we list the numbers of SNRs greater than 10 (or 15) in recovering the 200 mixed signals (of two sources and 100 replications) for COPICA and ICA\_SW. The results show both methods recover almost all the mixed signals with SNRs greater 10. The SNRs of both methods increase as the sample size increases from 1,000 to 5,000, and the result based on 5,000 samples are similar to the ones shown in Fig. 6 and Table 4. However, the computational time of ICA\_SW increases quadratically (i.e., 25 times) while the computational time of COPICA only increases linearly (i.e., 5 times). In the semiparametric COPICA approach, sorting is only needed for one dimensional marginal empirical distributions for each source and the joint distribution is



**Table 4** The number of SNRs greater than 10 (or 15) of COPICA and ICA\_SW in the 2-dimensional cases

Case	#{SNR ≥ 10}		#{SNR ≥ 15}	
	COPICA	ICA_SW	COPICA	ICA_SW
(a)	200	200	197	198
(b)	200	193	188	176
(c)	199	200	191	200
(d)	200	200	200	200
(e)	200	200	200	200
(f)	200	200	200	200
(g)	200	200	200	200
(h)	200	200	198	200
(i)	200	200	195	200
(j)	200	200	200	200
(k)	200	200	200	200
(l)	200	200	197	200



**Fig. 7** Box-plots of the SNR difference (COPICA - ICA\_SW) for 2-dimensional (left) and 6-dimensional (right) sources generated randomly from the 12 heavy tailed densities (a)–(l) (Color figure online)

linked by parametric copulae. While for the nonparametric approach ICA\_SW, sorting is required for both one dimensional and two dimensional joint copulae. In sum this means that our method is computationally lighter than their technique.

Another scenario of the experiments is to generate each independent source randomly from the 12 heavy tailed densities (a)–(l) (therefore the sources are not necessary of the same distribution) and then mix the sources by a matrix with elements sampled randomly from  $[-5, 5]$ . We show the SNR differences (COPICA-ICA\_SW) of 2-dimensional and 6-dimensional cases in Fig. 7. The SNR’s are obtained after 90 iterations, the sample size of each source is 1,000, and the numbers of replication are 1,000 and 100 for the 2-dimensional and 6-dimensional cases, respectively. The results show that the COPICA method is still competitive

with the ICA\_SW method for the random mixing bivariate cases. Nevertheless, in the 6-dimensional case the COPICA method attains higher SNR than the ICA\_SW method on the average.

### 6 Conclusions and discussions

In this article, a new ICA method, COPICA, is proposed. Similar to the FastICA, the COPICA method is also a two-step procedure. After whitening the data, COPICA projects the whitened data into the  $n$ -dimensional plane simultaneously, and this projection is chosen in terms of the parameters of the pre-specified copulae. Thus in COPICA, ICA problem is transformed to a minimization problem whose objective function is defined by the weighted combination of the divergence functions of copula parameters. The weights in the objective function are chosen to be inverse proportional to the standard deviations of the parameter estimators. Here the copula parameters are estimated via CML approach. Thus given a rotation matrix,  $R$  and the current recovered data,  $Y = R(WX)$ , the empirical marginal distributions of  $Y_i$  are obtained first and then the copula parameter vector,  $\theta$ , is found by maximizing Eq. (4). Hence we only have parameterized copula model assumption and do not have other assumptions on the marginal distribution. That is why we treat our COPICA as a semiparametric approach.

By comparing COPICA with the commonly used FastICA method and the nonparametric ICA methods, we find that the copula parameter based divergence function of the three copulae Gumbel, Clayton and Gaussian provide useful dependency measures when the observations come from a linear mixing model. The simulation and real data studies indicate that COPICA attains higher SNR than FastICA in BSS problems, especially when the original sources come from near-Gaussian-tailed distributions. Also, the COPICA is shown to have higher SNRs than the ICA\_SW on the average in the 6-dimensional case. Another interesting problem is to study the COPICA method for multi-modes densities, which is referred to our future work.

We investigate the sensitivity of COPICA w.r.t. weights via the BSS problem. A preliminary study is conducted here. In addition to  $(\omega_1, \omega_2, \omega_3, \omega_4) = (200, 300, 200, 500)$  in Example 1, six more weight combinations  $(\omega_1, \omega_2, \omega_3, \omega_4)$  are considered and five mixture-normal distributions with  $p = 0.1, 0.2, 0.4, 0.6, 0.7$  are considered. In each replication three independent sources of length  $T = 1000$ , generated from a mixture-normal distribution, are mixed by the matrix  $A$  defined in (12). For each  $p$ , the average SNRs of  $3 \times 100$  independent copies are obtained for each weight combination. The highest average SNR among the seven weight combinations is taken as the benchmark value. The ratios of the average SNR of each weight combination to the benchmark SNR are reported in Table 5. The

**Table 5** The average SNR ratios of COPICA for different weight combinations of the objective function (13)

Weights	$p$				
	0.1	0.2	0.4	0.6	0.7
(200, 300, 200, 500)	0.84	0.87	0.82	0.86	0.83
(200, 200, 200, 200)	1.00	1.00	1.00	1.00	1.00
(0, 200, 200, 200)	0.97	0.95	0.98	0.89	0.99
(200, 200, 0, 200)	0.94	0.89	0.75	0.74	0.79
(200, 200, 0, 0)	0.84	1.00	0.84	0.89	0.96
(200, 0, 200, 0)	0.94	0.87	0.76	0.78	0.83
(200, 0, 0, 200)	0.78	0.61	0.31	0.25	0.30
Highest ave. SNR	6.25	7.56	14.83	17.79	17.98

initial rotation angles are set to be zero and the number of iterations in the SA algorithm are set to be 100. The SNR ratios of the first six weight combinations range from 0.74 to 1, which indicates the COPICA is only slightly sensitive to these six weight combinations. The weight combinations (200, 300, 200, 500) has relatively robust performance among 7 weight cases. The weight combinations (200, 200, 200, 200) and (0, 200, 200, 200) are the best two obtaining the high SNRs, while the combination (200, 0, 0, 200) has the poorest performance in this scenario. This suggest the necessity of including the Gumbel and Clayton copulae in the objective function. Moreover, the highest average SNR of the COPICA increases as  $p$  increases (equivalently the kurtosis increases, see Table 1) after 100 iterations in the SA algorithm. To find general rules for weight selection of high SNR further studies are still needed.

**Acknowledgements** The authors gratefully acknowledge the National Science Council in Taiwan, National Center for Theoretical Sciences (South), Tainan, Taiwan and the Deutsche Forschungsgemeinschaft through the SFB 649 “Economic Risk”, Humboldt-Universität zu Berlin. This work was supported in part by National Science Council under grants NSC 96-2118-M-390-002- (Chen), NSC 100-2118-M-110-001-003- (Guo) and NSC 101-2118-M-390-002- (Huang).

## References

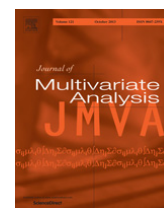
- Abayomi, K., Lall, U., de la Pena, V.: Copula Based Independent Component Analysis. Working paper (2008)
- Abayomi, K., de la Pena, V., Lall, U., Levy, M.: Quantifying sustainability: methodology for and determinants of an environmental sustainability index. In: Luo, Z.W. (ed.) Green Finance and Sustainability: Environmentally-Aware Business Models and Technologies, pp. 74–89 (2011)
- Bach, F.R., Jordan, M.I.: Kernel independent component analysis. *J. Mach. Learn. Res.* **3**, 1–38 (2002)
- Bell, A.J., Sejnowski, T.J.: An information maximization approach to blind source separation and blind deconvolution. *Neural Comput.* **7**, 1129–1159 (1995)
- Blaschke, T., Wiskott, L.: CuBICA: independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Trans. Signal Process.* **52**, 1250–1256 (2004)
- Chen, R.-B., Wu, Y.N.: A null space method for over-complete blind source separation. *Comput. Stat. Data Anal.* **51**, 5519–5536 (2007)
- Comon, P.: Independent component analysis—a new concept? *Signal Process.* **36**(3), 287–314 (1994)
- Genest, C., Nešlehová, J., Ziegel, J.: Inference in multivariate Archimedean copula models. *Test* **20**, 223–256 (2011)
- Ghosh, S., Henderson, S.: Behaviour of the NORTA method for correlated random vector generation as the dimension increases. *ACM Trans. Model. Comput. Simul.* **13**, 276V294 (2003)
- Gretton, A., Bousquet, O., Smola, A.J., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: ALT, pp. 63–78. Springer, Heidelberg (2005)
- Grønneberg, S., Hjort, N.L.: The Copula Information Criterion. Technical report, Department of Math., University of Oslo, Norway (2008)
- Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**, 626–634 (1999a)
- Hyvärinen, A.: Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. *Neural Comput.* **11**, 1739–1768 (1999b)
- Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9**, 1483–1492 (1997)
- Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and application. *Neural Netw.* **13**, 411–430 (2000)
- Kidmose, P.: Blind Separation of Heavy Tail Signals. Ph.D. Thesis, Technical University of Denmark, Lyngby (2001)
- Kirkpatrick, S., Gelatt, C.D. Jr., Vecchi, M.P.: Optimization by simulated annealing. *Science* **220**, 671–680 (1983)
- Kirshner, S., Póczos, B.: ICA and ISA using Schweizer-Wolff measure of dependence. In: International Conference on Machine Learning (ICML-2008) (2008). <http://icml2008.cs.helsinki.fi/papers/551.pdf>
- Kotz, S., Nadarajah, S.: Extreme Value Distributions. Theory and Applications. Imperial College Press, London (2000)
- Learned-Miller, E.G., Fisher, J.W.: ICA using spacings estimates of entropy. *J. Mach. Learn. Res.* **4**, 1271–1295 (2003)
- Lee, T.-W.: Independent Component Analysis. Theory and Applications. Kluwer Academic, Dordrecht (1998)
- Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, New York (2001)
- Ma, J., Sun, Z.: Copula component analysis. In: Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation, pp. 73–80. ACM, London (2007)
- Mardia, K.V.: A translation family of bivariate distributions and Fréchet’s bounds. *Sankhya A* **32**, 119–122 (1970)
- McNeil, A.J., Nešlehová, J.: From Archimedean to Liouville copulas. *J. Multivar. Anal.* **101**, 1772–1790 (2010)
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H.: Equation of state calculation by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953)
- Nelsen, R.B.: An Introduction to Copulas. Springer, Berlin (2006)
- Olshausen, B.A., Field, D.J.: Natural image statistics and efficient coding. *Network* **7**, 333–339 (1996)
- Schmid, F., Schmidt, R.: Nonparametric inference on multivariate versions of Blomqvist’s beta and related measures of tail dependence. *Metrika* **66**, 323–354 (2007)
- Schweizer, B., Wolff, E.F.: On nonparametric measures of dependence for random variables. *Ann. Stat.* **9**, 879–885 (1981)
- Shen, H., Jegelka, S., Gretton, A.: Fast kernel-based independent component analysis. *IEEE Trans. Signal Process.* **57**, 3498–3511 (2009)

- Sklar, A.: Fonctions de Repartition à  $n$  Dimensions et Leurs Marges. Publ. Inst. Stat. Univ. Paris **8**, 229–231 (1959)
- Sklar, A.: Random variables, distribution functions, and copulas—a personal look backward and forward. In: Ruschendorf, L., Schweizer, B., Taylor, M.D. (eds.) *Distributions with Fixed Marginals and Related Topics*, pp. 1–14. Institute of Mathematical Statistics, Hayward (1996)
- Sodoyer, D., Girin, L., Jutten, C., Schwartz, J.-L.: Speech extraction based on ICA and audio-visual coherence. In: *Proceedings of the 7th International Symposium on Signal Processing and Its Applications (ISSPA'03)*, vol. 2, pp. 65–68 (2003)
- Tsai, A.C., Liou, M., Jung, T.-P., Onton, J.A., Cheng, P.E., Huang, C.-C., Duann, J.-R., Makeig, S.: Mapping single-trial EEG records on the cortical surface through a spatiotemporal modality. *NeuroImage* **32**, 195–207 (2006)
- Yu, L., Verducci, J.S., Blower, P.E.: The tau-path test for monotone association in an unspecified subpopulation. In: *Application to Chemogenomic Data Mining Statistical Methodology*, vol. 8, pp. 97–111 (2011)



Contents lists available at ScienceDirect

## Journal of Multivariate Analysis

journal homepage: [www.elsevier.com/locate/jmva](http://www.elsevier.com/locate/jmva)

# Tie the straps: Uniform bootstrap confidence bands for semiparametric additive models<sup>☆,☆☆</sup>

Wolfgang Karl Härdle<sup>a,b</sup>, Ya'acov Ritov<sup>c</sup>, Weining Wang<sup>a,\*</sup><sup>a</sup> C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany<sup>b</sup> Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore<sup>c</sup> Department of Statistics Hebrew University of Jerusalem Mount Scopus, Jerusalem 91905, Israel

## ARTICLE INFO

## Article history:

Received 9 November 2013

Available online 20 November 2014

## AMS subject classifications:

62G08

62G09

62G20

62H12

62P20

## Keywords:

Nonparametric regression

Bootstrap

Quantile regression

Confidence bands

Additive model

Robust statistics

## ABSTRACT

This study considers the theoretical bootstrap “coupling” techniques for nonparametric robust smoothers and quantile regression, and we verify the bootstrap improvement. To handle the curse of dimensionality, a variant of “coupling” bootstrap techniques is developed for additive models both in a robust mean regression and in a quantile regression framework. Our bootstrap method can be used in many situations such as constructing confidence intervals and bands. We demonstrate the bootstrap improvement over the theoretical asymptotic band in simulations and in applications to firm expenditures and the interaction of economic sectors and the stock market.

© 2014 Elsevier Inc. All rights reserved.

## 1. Introduction

Confidence bands are important tools for model specifications. However, it is difficult to construct precise confidence bands for nonparametric curves since a supreme norm is usually involved in the statistics. Traditional methods based on the asymptotic theory have natural drawbacks in their finite sample performance, and this motivates bootstrap methods to attain more precise bands. In this article, we deal with bootstrap bands construction for a general class of nonparametric  $M$ - and  $L$ -estimates; moreover, we adopt additive models to handle the multivariate covariates case. We believe that the developed technique is essential for componentwise shape inspection of additive models in empirical economics. Applications can be found in the work of [6] or [4].

Consider  $Y, X \in \mathbb{R}^{d+1}$  with variable  $Y$  and  $X \in \mathbb{R}^d$ .

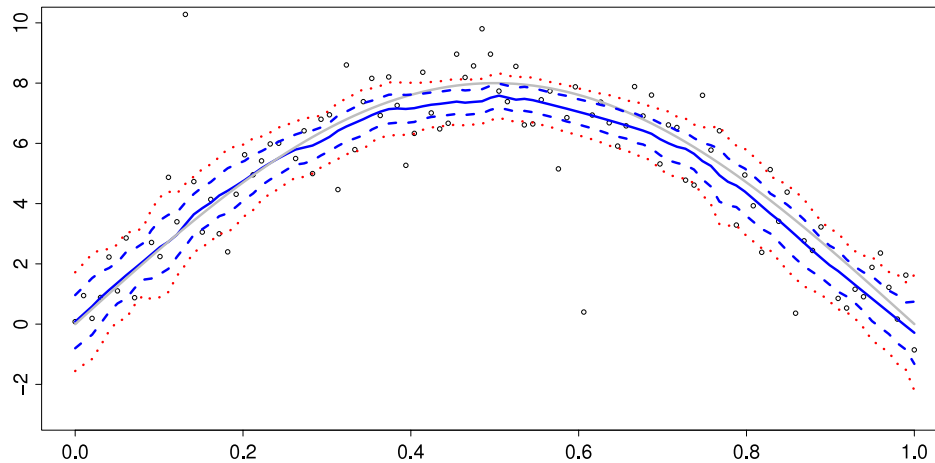
$$l(x) = \arg \min_{\theta} E_{Y|X=x} \rho(Y - \theta), \quad (1)$$

<sup>☆</sup> We thank the Editors and two referees for helpful comments. The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 “Ökonomisches Risiko”, Humboldt-Universität zu Berlin is gratefully acknowledged.

<sup>☆☆</sup> We appreciate the friendly working environment and atmosphere of the MFO Mathematical Research Institute of Oberwolfach during our Research in pairs stay.

\* Corresponding author.

E-mail addresses: [haerdle@wiwi.hu-berlin.de](mailto:haerdle@wiwi.hu-berlin.de) (W.K. Härdle), [yaacov.ritov@gmail.com](mailto:yaacov.ritov@gmail.com) (Y. Ritov), [wangwein@cms.hu-berlin.de](mailto:wangwein@cms.hu-berlin.de) (W. Wang).



**Fig. 1.** Plot of true curve (gray), robust estimate with Tukey biweight loss and 5% confidence bands (blue dashed), local polynomial estimate (black), bootstrap band (red dotted),  $n = 150$ .

with  $\rho(\cdot)$  as a loss function described in detail in Section 2.1. For the confidence bands construction, one stream of literature using empirical process theory follows the asymptotic results of [2], which is further extended by [19], and has recently been studied by [13] for  $L$ -smoothers. However, it has also been shown by [31] that such an asymptotic confidence band has a much lower coverage probability in a finite sample than what it is supposed to have. The poor performance of such a band in a finite sample has been well attributed to its slow convergence, see [7]. To address improvement of a finite sample performance, there is a class of literature on the bootstrap confidence band, see [3,12], among others. Fig. 1 shows a bootstrap confidence band and an asymptotic band for an  $M$ -smoother with  $\rho(\cdot)$  being Tukey's bisquare loss. One can see that the asymptotic band is narrower than the bootstrap one. Moreover, the asymptotic band does not cover the true curve while the bootstrap band does.

Many different resampling techniques have been developed, among which the permutation tests are very popular and can be adapted to many application scenarios, see [28,1,22,23]. The bootstrap is a class of data driven resampling techniques that provide non-asymptotic approximations of finite sample distributions of different statistics. In a location model (more generally a regression model), resampling is done from the estimated residuals, and a typical theoretical analysis leads to the conclusion that "bootstrap works" in the sense that a suitably centered bootstrap estimator converges to the same asymptotic normal distribution as the original estimator under consideration. A large literature body has focused on showing bootstrap improvements and refinements of approximations via bootstrap resampling, see [8,21,15,11], which discuss the conditions for bootstrap consistency, and also prove the bootstrap accuracy as an approximation to the exact finite sample distribution for special types of statistics in a nonparametric framework. However, very few articles have focused on nonlinear statistics (e.g. maximum) in nonparametric regression. [31] proposes a bootstrap procedure and shows its improvement properties.

This stimulates the current research on finding common properties that loss functions have to share in order to attain such an improvement. Accordingly, we prove a generalized version in the univariate case for a class of loss function with bounded influence. The bootstrap becomes difficult when the dimension  $d$  of the regressors gets large. One way to avoid this problem is to impose a structure, such as an additive model, on the multivariate nonparametric function. The additive structure assumes that the covariates' effects are separable, and this effect is presented in many economic applications, [10]. Specifically, we consider the regression function

$$m(x_1, \dots, x_d) = \sum_{j=0}^d m_j(x_j), \quad (2)$$

with  $m_0(x_0)$  a constant. It is worth noting that the additive structure implicitly assumes that the covariates effects are separable, and of course this assumption needs to be tested in advance. [32] develops a test on testing the interaction effects; correspondingly, our method can also be extended to implement similar tests.

It is well known that (2) avoids the "curse the dimensionality" in the sense that one-dimensional convergence rates are achieved for the estimation of  $m(x_1, \dots, x_d)$ , but keeps enough flexibility of the marginal influence of the different variables. See [14,17,16] among many others. [14] focuses on generalized additive models with unknown link functions, [17] proposes a two-stage estimation for quantile regression in additive models, [16] shows the equivalence between spline, kernel and other methods in terms of optimal minimax rate in the additive model estimation. The resulting estimate  $\hat{m}_j(x_j)$  in (2) though needs to be screened for closeness to  $m_j(x_j)$ . This requires construction of confidence intervals and bands as a function of  $x_j$ . For such screening tests, our tightened bootstrap techniques will be verified. Namely, the bootstrap-based confidence bands are shown to be very close to the true finite sample distribution-based ones.

In summary, we investigate a coupling technique that allows us to “tie the straps” even a little tighter for a class of bounded influence estimators. Theoretically speaking, confidence band construction is made more precise in a variety of the estimation problems that we consider. The coupling idea is based on mimicking the distribution of the original data via a controllable random mechanism. Similar results like (9) will be derived for additive models.

The remainder of this paper is organized as follows. In Section 2, we explain the model setup and the bootstrap method in more detail. Section 3 presents the main results. In Section 4, a small simulation study is presented. Moreover, we show in Section 5 applications on managerial compensation and the impact on stock markets. Finally, Section 6 concludes with some comments and directions for future research.

## 2. Models and bootstrap confidence sets

This section describes the estimator and our coupling techniques, which motivate the obtainable theoretical results, and discusses some of the assumptions.

### 2.1. Univariate case

Let us describe the coupled bootstrap in the simple case of nonparametric minimum contrast curve estimation. Here  $(X, Y)^T \in \mathbb{R}^2$ . The object of estimation is identified via:

$$l(x) = \arg \min_{\theta} E_{(Y|X=x)} \rho(Y - \theta), \tag{3}$$

where  $\rho(\cdot)$  is a loss function of e.g. Hampel/Huber type or more generally (up to a constant) a negative (pseudo) log likelihood. In the quantile regression case,  $\rho(x) = |x|\{\tau - \mathbf{1}(x \leq 0)\}$  is the check function. Another example for  $\rho(\cdot)$  includes the trimmed mean, [18]

$$\rho(x) = \begin{cases} x^2/2, & |x| \leq k, \\ -k^2/2 + k|x|, & |x| > k \end{cases} \tag{4}$$

or a form of Winsorized mean:

$$\rho(x) = \begin{cases} x^2, & |x| \leq k, \\ k^2, & |x| > k. \end{cases} \tag{5}$$

A sample based version of (3) is:

$$\hat{l}_h(x) = \operatorname{argmin}_{\theta} n^{-1} \sum_{i=1}^n \rho(Y_i - \theta) K_h(x - X_i), \tag{6}$$

where  $K_h(u) = K(u/h)/h$  is a kernel function with bandwidth  $h$ . Now we generate a bootstrap sample using i.i.d. uniform random variables  $U_1, \dots, U_n \in U(0, 1)$ , and then generate:

$$Y_i^* = \hat{l}_g(X_i) + \varepsilon_i^*, \quad i = 1, \dots, n, \tag{7}$$

where  $\varepsilon_i^* \sim \hat{F}_{(\varepsilon|X=X_i)}^{-1}(U_i)$  (discussed in detail in (24)) and  $g = \mathcal{O}(n^{-1/9})$  a slightly larger bandwidth than  $h$ . The basic idea of coupling is based on comparing this sample to the pseudo observations:

$$Y_i^\sharp = l(X_i) + \varepsilon_i^\sharp, \quad i = 1, \dots, n, \tag{8}$$

where  $\varepsilon_i^\sharp = F_{(Y|X=X_i)}^{-1}(U_i)$ . Note that given  $\{X_i\}_{i=1}^n$ , the distribution of  $Y_i^\sharp$  and  $Y_i$  is the same. We will show that for a class of loss functions the following approximation holds:

$$\sup_{x \in B} \left[ \hat{l}_h^\sharp(x) - l(x) - \{\hat{l}_{h,g}^*(x) - \hat{l}_g(x)\} \right] = \mathcal{O}_p(h^2 \Gamma_n), \tag{9}$$

where  $B$  is a closed compact set in  $[0, 1]$ ,  $\Gamma_n$  a slowly increasing sequence (a sequence  $a_n$  is slowly increasing if  $n^{-\alpha} a_n \rightarrow 0$  for any  $\alpha > 0$ ),  $\hat{l}_h^\sharp(\cdot)$  is the nonparametric estimate calculated from  $\{(X_i, Y_i^\sharp)\}$ ,  $\hat{l}_{h,g}^*(X_i)$  is an estimate calculated from the bootstrap sample  $\{(X_i, Y_i^*)\}$  with bandwidth  $h$ ,

$$\hat{l}_{h,g}^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta} \sum_{i=1}^n \rho(Y_i^* - \theta) K_h(x - X_i) \tag{10}$$

and  $\hat{l}_g(X_i)$  is calculated as in (6) from the original sample with bandwidth  $g$ . The basic elements in proving (9) are smoothness of  $F_{\varepsilon|X=x}(\cdot)$  and bounded influence of  $\rho(\cdot)$  in (3).

## 2.2. Additive models and bootstrap confidence sets

For any  $x \in \mathbb{R}^d$  ( $d > 1$ ), the nonparametric approach in (6) is not appropriate when  $d$  is large, as the standard nonparametric optimal convergence rate,  $\mathcal{O}_p(n^{-4/(4+d)})$ , would be too slow when  $d$  is large. Additive models were suggested to remedy the problems posed by the dimensionality. Recall (2), and impose the additive structure:

$$Y_i = \sum_{j=0}^d m_j(x_{i,j}) + \varepsilon_i. \quad (11)$$

Further, approximate the additive model via a basis function approach:

$$m_j(x_{i,j}) \approx \sum_{l=1}^{L_j+1} a_{l,j} v_l(x_{i,j}),$$

where the  $v_1(\cdot), v_2(\cdot), \dots$  could be any sequence of functions spanning the  $L^2$  space. Our implementation uses  $B$ -splines, for example, linear  $B$ -splines: consider a sequence of  $H^{-1}$  equally spaced knots on the interval  $[0, 1]$ , which defines the width  $H$  subintervals  $[lH, (l+1)H], 0 \leq l \leq (H^{-1} - 1)$ . The linear  $B$ -spline basis is:

$$v_l(x) = \begin{cases} H^{-1}x - l + 1 & (l-1)H \leq x \leq lH, \\ l + 1 - H^{-1}x & lH \leq x \leq (l+1)H, \\ 0 & \text{otherwise.} \end{cases}$$

Denote the theoretical standardized  $B$  spline basis  $\phi_l(\cdot)$ , for the  $j$ th variable,  $j = 1, \dots, d$ ,

$$\begin{aligned} \phi_{l,j}(x_j) &\stackrel{\text{def}}{=} v_l(x_j) - v_{l-1}(x_j)c_{l,j}/c_{l-1,j}, \\ B_{j,l}(x_j) &= \phi_{l,j}(x_j)/\|\phi_{l,j}(x_j)\|_2, \end{aligned}$$

where  $l = 0, \dots, H^{-1}$ ,  $c_{l,j} \stackrel{\text{def}}{=} \int \phi_{l,j}(u)f_j(u)du$  with  $f_j(u)$  is the density for  $x_j$ , so that  $EB_{j,l}(x_j) = 0, EB_{j,l}(x_j)^2 = 1$ .

The additive estimate can then be obtained as follows. Define the vectors in  $(\mathbb{R}^{H^{-1}d+1})$ ,

$$\begin{aligned} A &= (a_0, \mathbf{a}_1^\top, \dots, \mathbf{a}_d^\top)^\top, \\ \Phi(X_i) &= \{1, \mathbf{g}(x_{i,1})^\top, \dots, \mathbf{g}(x_{i,d})^\top\}^\top, \end{aligned}$$

where

$$\begin{aligned} \mathbf{a}_j &= (a_{1,j}, \dots, a_{H,j})^\top, \\ \mathbf{g}(x_{i,j})^\top &= \{v_1(x_{i,j}), \dots, v_H(x_{i,j})\}^\top. \end{aligned}$$

Finally, let  $\hat{A}$  be the estimation of  $A$ :

$$\hat{A} = \arg \min_A \sum_{i=1}^n \rho\{Y_i - A^\top \Phi(X_i)\}, \quad (12)$$

and

$$\hat{m}_j(x_{i,j}) = \hat{\mathbf{a}}_j^\top \mathbf{g}(x_{i,j}), \quad (13)$$

with  $j = 1, \dots, d$ .

## 2.3. Coupled bootstrap for cases of high dimensional covariates

The additive structure in (11) is one solution to the curse of dimensionality, however, the bootstrap approach in (7) does not work for this modeling scenario as nonparametric estimation of  $F_{\varepsilon|X}(\cdot)$  would again run into the ‘‘curse of dimensionality’’ problem. We suggest another bootstrap technique, and prove that it strongly approximates a model with the same asymptotic properties as the original model.

Define

$$Z_i = \begin{cases} 1 & \text{with prob } \tau \\ -1 & \text{with prob } 1 - \tau, \end{cases} \quad i = 1, \dots, n. \quad (14)$$

This includes the special case for symmetric error distributions with  $\tau = 1/2$ , which is the usual assumption for mean or robust  $M$ -smoothers. Moreover, it generally adapts to the case of quantile regression for asymmetric error distributions. The bootstrap couple  $(\varepsilon^*, \varepsilon^\#)$ , the bootstrap residuals and its associate theoretical couple respectively are:

$$\varepsilon_i^* \stackrel{\text{def}}{=} Z_i |\hat{\varepsilon}_i| \quad (15)$$

$$\varepsilon_i^\# \stackrel{\text{def}}{=} Z_i \eta_i, \quad i = 1, \dots, n, \quad (16)$$

where

$$\eta_i \stackrel{\text{def}}{=} F_{i,Z_i}^{-1}\{F_{i,\text{sgn}(\varepsilon_i)}(|\varepsilon_i|)\}, \quad i = 1, \dots, n, \tag{17}$$

and

$$F_{i,s}(t) \stackrel{\text{def}}{=} P(|\varepsilon_i| \leq t | s\varepsilon_i > 0), \quad i = 1, \dots, n, \quad s \in \{1, -1\}. \tag{18}$$

Note that the same  $Z_i$  appears both in (15) and in (16).

Recall that  $F_{Y|X=X_i}\{I(X_i)\} = \tau$  and, hence,  $F_{\varepsilon|X=X_i}(0) = \tau$ . Now, it is easy to see that  $V_i \stackrel{\text{def}}{=} F_{i,\text{sgn}(\varepsilon_i)}(|\varepsilon_i|)$  has a standard uniform distribution, and if  $Z_i$  is as above, then  $\varepsilon_i$  and  $Z_i F_{i,Z_i}^{-1}(V_i)$  have the same distribution. Formally, note that

$$F_{i,+1}(t) = \frac{F_i(t) - 1 + \tau}{\tau},$$

$$F_{i,-1}(t) = \frac{1 - \tau - F_i(-t)}{1 - \tau},$$

where  $F_i(\cdot)$  is the cdf of  $\varepsilon_i$ . Hence, for  $t > 0$ :

$$\begin{aligned} P(\varepsilon_i^\# < t) &= \tau P[F_{i,+1}^{-1}\{F_{i,\text{sgn}(\varepsilon_i)}(|\varepsilon_i|)\} < t] + 1 - \tau \\ &= \tau P\{F_{i,\text{sgn}(\varepsilon_i)}(|\varepsilon_i|) < F_{i,+1}(t)\} + 1 - \tau \\ &= \tau P\{\varepsilon_i < 0, F_{i,-1}(-\varepsilon_i) < F_{i,+1}(t)\} + \tau P\{\varepsilon_i > 0, F_{i,+1}(\varepsilon_i) < F_{i,+1}(t)\} + 1 - \tau \\ &= \tau P\left\{\varepsilon_i < 0, \frac{1 - \tau - F_i(\varepsilon_i)}{1 - \tau} < \frac{F_i(t) - 1 + \tau}{\tau}\right\} + \tau P(0 < \varepsilon_i < t) + 1 - \tau \\ &= \tau P\left[1 - \tau > F_i(\varepsilon_i) > \frac{1 - \tau}{\tau}\{1 - F_i(t)\}\right] + \tau P(0 < \varepsilon_i < t) + 1 - \tau \\ &= \tau \left[1 - \frac{1 - \tau}{\tau}\{1 - F_i(t)\} - \tau\right] + \tau\{F_i(t) - 1 + \tau\} + 1 - \tau \\ &= F_i(t). \end{aligned}$$

The case  $t < 0$  is dealt similarly. It follows

$$\mathcal{L}(\varepsilon_i^\#) = \mathcal{L}(\varepsilon_i). \tag{19}$$

Our confidence “ideal” interval is conditional on  $\{V_i\}_{i=1}^n$  which has a direct link to the absolute value of errors  $\{|\varepsilon_i|\}_{i=1}^n$ . Note, however, that the estimator is asymptotically consistent and its bias does not depend on these absolute values. Moreover, by the law of large numbers, the pointwise width of the conditional confidence interval is within a factor of  $1 + \mathcal{O}_p(1)$  of the unconditional one.

### 2.4. How does the coupling work?

The basic idea of our approach is to construct an empirically feasible bootstrap sample that strongly approximates a sample from the true distribution. One example of the coupled bootstrap approach has already been explained in (7) and (8). It, however, relies on estimators of the conditional distribution  $F_{Y|X=X}(\cdot)$ , which become very imprecise when  $d$  is large.

Another approach proposed in Section 2.3 motivated as the wild bootstrap is based on randomizing the obtained residuals and using the same random source to mimic the stochastic of the unobservable errors. To get the basic idea, let us assume for a moment that the distributions of  $\varepsilon_i$  are symmetric. Then the coupling may be performed via a Rademacher randomized variable  $Z_i$  with

$$P(Z_i = 1) = P(Z_i = -1) = 1/2$$

and generation of the couple  $\varepsilon_i^*$  (the bootstrapped residuals),  $\varepsilon_i^\#$  (the theoretical coupling), where  $\{\varepsilon_i^*, \varepsilon_i^\#\}$  is:

$$\varepsilon_i^* \stackrel{\text{def}}{=} Z_i |\hat{\varepsilon}_i|$$

$$\varepsilon_i^\# \stackrel{\text{def}}{=} Z_i \eta_i. \tag{20}$$

With this construction, we are able to establish a result similar to (9).

In a non symmetric distribution (required for quantile regression), one defines  $Z_i$  with  $P(Z_i = 1) = \tau$  and  $P(Z_i = -1) = 1 - \tau$  assuming the centering  $F_{Y|X_i}\{I(X_i)\} = \tau$ , and the couple  $(\varepsilon_i^*, \varepsilon_i^\#)$  is given by (15) and (16). It was argued that the distributions of  $\varepsilon_i^\#$  and  $\varepsilon_i$  are identical and also the conditional distributions given  $\{V_i\}_{i=1}^n$  are the same.

The resampling technique will be applied to a nonparametric estimation of an additive quantile regression model. The reanalysis of the data used by [17] provides us with sharper bands that have not been calculated in that paper.



2.5. Extension to generalized linear models

The model in Section 2 can be extended to generalized linear models, with the relation  $g\{E(Y|X)\} = l(X)$ , where  $g(\cdot)$  is the known prespecified link function. For continuous random variable  $Y$ , the extension to the above bootstrap procedure is trivial. Moreover, it can be also generalized to the models with discrete  $Y_i$ . For example, for the binary logistic model with  $Y_i$  as binary data, define

$$\hat{\varepsilon}_i = Y_i - g\{\hat{l}_h(X_i)\}, \tag{21}$$

then  $\hat{\varepsilon}_i$  will be bounded in  $[-1, 1]$ , and one can apply the same bootstrap procedure as in (20).

3. Main results

The section gives asymptotic results for the estimators described in Section 2. To establish the asymptotic property, some assumptions are needed:

Assumptions

- A.1 The solution  $l(\cdot)$  of (1) is two-times differentiable and bounded. We abbreviate  $\psi(Y, \cdot)$  to  $\psi(\cdot)$ , and  $\psi(\cdot) = \rho'(\cdot)$  (or subgradient in the quantile regression case) is a.s. bounded by  $M < \infty$ .  $E\{\psi(\varepsilon)|X\} = 0$  w.p. 1. Also  $\psi(\cdot)$  is Lipschitz continuous except for a finite number of points in the compact set  $B$ .
- A.2 Assume the support of  $X$  is  $[0, 1]^d$ , the conditional density of  $\varepsilon|_{X=x}(\cdot|\cdot)$ , and is bounded by  $C_1$  and is twice differentiable in the interior point of  $[0, 1]^d$ , and bounded away from 0 by  $c_1 > 0$ .
- A.3 The kernel function  $K_h(\cdot)$  is a product kernel composed from one dimensional kernels with bandwidth  $h$ :

$$K_h(s) = \prod_{j=1}^d K(s_j/h)/h, \quad s = (s_1, \dots, s_d)^\top \in \mathbb{R}^d. \tag{22}$$

- A.4 The kernel bandwidth satisfies  $h \sim n^{-1/5}$ . Let  $g$  be another bandwidth sequence  $g \gg h$ , e.g.,  $g = \mathcal{O}(n^{-1/9})$ .
- A.5 For each  $j$ , the true regression function  $m_j(\cdot)$ ,  $j \in 1, \dots, d$ , is at least one time continuous differentiable function on  $[0, 1]$ .
- A.6  $E\{g_l^2(X_j)\} = 1$  for  $j \in 1, \dots, d$ .  $\|\Phi_l(X_j)\|_\infty \leq C_3H$ , a.s., where  $\Phi_l(X_j) \stackrel{\text{def}}{=} \{\phi_l^2(x_{1,j}), \dots, \phi_l^2(x_{n,j})\}^\top$ , with  $j \in 1, \dots, d$ .
- A.7 The number of regressors in (12) is of  $\mathcal{O}(p)$ , where  $p = dH^{-1} + 1$  with  $H^{-1} \sim n^{1/5}$ , and  $d = \mathcal{O}(n^{2/3})$ .

A.1 is about the continuity and the bounded influence structure of the loss function, it is quite essential for proving the bootstrap improvement. Also, A.1 includes a very general class of regression loss functions although it does not include the usual mean regression loss. A.2 assumes W.L.O.G., the covariates are on bounded support and impose assumption on the conditional density of the error term. A.3 is a standard assumption on the kernel function. A.4 is about theoretical rates of the bandwidths  $h$  and  $g$ .  $h$  is of the standard optimal rate in nonparametric regression, while  $g$  is required to be larger as we need to mimic the bias of estimation in the original nonparametric regression, see [12]. A.5–A.7 are assumptions on additive models. A.5 assumes that additive components behave properly. A.6 imposes conditions on the basis functions, and the linear  $B$ -spline satisfies A.6. The above assumptions are all very mild or standard assumptions according to the literature.

We show first convergence results for the bootstrap methods in (7) and (8). The resampling step has been defined in (7), where the smooth estimate of the conditional distribution is:

$$\tilde{F}_{(\varepsilon|X=x)}(t) \stackrel{\text{def}}{=} \sum_{i=1}^n W_{h,i}(x) \mathbf{1}[\{Y_i - \hat{l}_h(X_i)\} \leq t], \tag{23}$$

with  $W_{h,i}(x) = n^{-1}K_h(x - X_i)/\hat{f}_h(x)$  and  $\hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$  the kernel density estimator. To have a correctly centered estimation for  $F_{(\varepsilon|X=x)}(t)$ , we define

$$d\hat{F}_{(\varepsilon|X=x)}(t) = \begin{cases} \frac{d\tilde{F}_{(\varepsilon|X=x)}(t)}{\tilde{F}_{(\varepsilon|X=x)}(0) + C_0(1 - \tilde{F}_{(\varepsilon|X=x)}(0))} & \text{if } t < 0, \\ \frac{C_0 d\tilde{F}_{(\varepsilon|X=x)}(t)}{\tilde{F}_{(\varepsilon|X=x)}(0) + C_0(1 - \tilde{F}_{(\varepsilon|X=x)}(0))} & \text{otherwise} \end{cases} \tag{24}$$

where  $C_0$  is a constant defined as

$$C_0 \stackrel{\text{def}}{=} \frac{\int_{-\infty}^0 \psi(u) d\tilde{F}_{(\varepsilon|X=x)}(u)}{\int_0^\infty \psi(u) d\tilde{F}_{(\varepsilon|X=x)}(u)}.$$

Note that the estimator in (24) is centered so that

$$E_{\hat{F}_{(\varepsilon|X=x_i)}} \psi(\varepsilon_i^*) = 0 = E_{F_{(\varepsilon|X=x_i)}} \psi(\varepsilon_i^\#). \tag{25}$$

The influence function of the estimator is proportional to  $\psi(\cdot) = \rho'(\cdot)$ . If it is bounded with bounded derivatives a.e. and  $\mathcal{L}(\varepsilon|X)$  is such that  $\|\hat{F}_{(\varepsilon|X_i=x)}(\cdot) - F_{(\varepsilon|X=x)}(\cdot)\|_\infty = \mathcal{O}(h^2 \Gamma_n)$ , then a similar coupling argument as in (7) can be used. Recall  $\varepsilon_i^* = \hat{F}_{(\varepsilon|X_i=x)}^{-1}(F_{(\varepsilon|X_i=x)})$ , then

$$|\psi(\varepsilon_i^*) - \psi(\varepsilon_i^\#)| = \mathcal{O}_p(h^2 \Gamma_n). \tag{26}$$

This ensures that

$$n^{-1} \sum_{i=1}^n \{\psi(\varepsilon_i^*) - \psi(\varepsilon_i^\#)\} K_h(x - X_i) = \mathcal{O}_p(h^2 \Gamma_n). \tag{27}$$

The argument is based on two facts. First from (25) the means are zero and second that (26) holds.

Once the  $Y_i^*$ s are generated, one applies (10) to the bootstrap data  $\{(X_i, Y_i^*)\}_{i=1}^n$  to obtain  $\hat{l}_{h,g}^*(x)$ . Summarizing, we have:

**Theorem 3.1.** *Let assumptions A.1–A.4 be fulfilled and define  $\hat{l}_h^\#(\cdot)$  as in (9). Then*

$$\sup_{x \in B} |A_n(x)| = \mathcal{O}_p(h^2 \Gamma_n), \tag{28}$$

where

$$A_n(x) \stackrel{\text{def}}{=} (\hat{l}_h^\# - l)(x) - \{(\hat{l}_{h,g}^* - \hat{l}_g)(x)\}. \tag{29}$$

Let assumptions A.1–A.7 be fulfilled and consider the additive model of (11) with the estimator (12), and the resampling scheme is considered as in (15) and (16), then

$$\sup_{x \in B} |(\hat{m}_j^\# - m_j)(x) - \{(\hat{m}_j^* - \hat{m}_j)(x)\}| = \mathcal{O}_{a.s.}(H^2 \Gamma_n).$$

**Remark 1.** The aforementioned strong approximation results mean that the stochastic behavior of  $\hat{l}_h^\#(x) - l(x) ((\hat{m}_j^\# - m_j)(x))$  is well approximated by its bootstrap counterpart  $\hat{l}_{h,g}^*(x) - \hat{l}_g(x) ((\hat{m}_j^* - \hat{m}_j)(x))$ . This implies in particular that the distribution of  $\sup_x |\hat{l}_h^\#(x) - l(x)| (\sup_{x \in B} |(\hat{m}_j^\# - m_j)(x)|)$  is consistently approximated by that of  $\sup_x |\hat{l}_{h,g}^*(x) - \hat{l}_g(x)| (\sup_{x \in B} |(\hat{m}_j^* - \hat{m}_j)(x)|)$ . Also the rate  $H^2 \Gamma_n$  is sufficient for the validity of the bootstrap for supremum-functionals, see [24]. Therefore, the bootstrap confidence band is a direct consequence of the results.

**Remark 2.** A similar result was proved by [31] for quantile regression. There the centering ensures that the bootstrap error distribution has the proper quantile. Here it is generalized to a wider class of centering, and to additive models.

### 4. Simulation

This section is divided into two parts. First, we concentrate on the univariate  $x \in [0, 1]$  case and the bootstrap procedure (7), (8), check the validity of the bootstrap procedure, and compare it with asymptotic uniform confidence bands. Second, we adopt the bootstrap procedure for the additive model as in (20), and check the validity of the bootstrap band in the same fashion.

#### 4.1. Univariate case

The simulation setup in the univariate case is:

- (1) Simulate  $(X_i, Y_i)$ ,  $i = 1, \dots, n$  according to the predefined joint probability density function (pdf)  $f(x, y)$ . In order to compare with [9], we set the joint pdf of  $(X, Y)$  as,

$$f(x, y) = g\{y - \sin(\pi x)\} \mathbf{1}(x \in [0, 1]), \tag{30}$$

$$g(u) = 9\varphi(u)/10 + \varphi(u/9)/90. \tag{31}$$

- (2) Compute  $\hat{l}_h(x)$  as in (6), with  $\rho(\cdot)$  as a biweight loss,  $\hat{\varepsilon}_i \stackrel{\text{def}}{=} Y_i - \hat{l}_h(X_i)$ .
- (3) Compute the estimated conditional edf as in (A.2) with a Gaussian kernel

$$K_h(u) = (\sqrt{2\pi})^{-1} \exp\{-u^2/2h\}/h,$$

and  $h = 0.06$  is selected by cross validation.

**Table 1**

Relative errors and areas (estimated as averaged width of the confidence bands over the grid points) of asymptotic and bootstrap with 1000 repetitions per sample, and 5000 samples. Relative error  $\stackrel{\text{def}}{=} |(\text{true coverage} - \text{nominal coverage})|/\text{nominal coverage}$ . Standard deviations measuring variability over samples are shown in brackets.

	$n$	95%		90%	
		Rel. Err.	Area	Rel. Err.	Area
Boot.	100	0.02(0.015)	2.52(0.547)	0.02(0.014)	2.21(0.746)
	200	0.01(0.010)	2.01(0.387)	0.02(0.010)	1.82(0.679)
	400	0.01(0.004)	1.33(0.201)	0.01(0.003)	1.13(0.301)
Asym.	100	0.09(0.034)	1.22(0.596)	0.11(0.070)	1.02(0.711)
	200	0.06(0.026)	0.91(0.602)	0.10(0.051)	0.67(0.518)
	400	0.05(0.018)	0.72(0.314)	0.09(0.043)	0.68(0.325)

(4) For each  $i = 1, \dots, n$ , generate random variable  $\varepsilon_i^* \sim \hat{F}_{(\varepsilon|X)}(t), i = 1, \dots, n$ :

$$Y_i^* = \hat{l}_g(X_i) + \varepsilon_i^*,$$

with  $g = 0.2$ .

(5) For each sample  $\{X_i, Y_i^*\}_{i=1}^n$ , compute  $\hat{l}_{h,g}^*(\cdot)$  and the random variable

$$d_i^* \stackrel{\text{def}}{=} \sup_{x \in B} [|\hat{l}_{h,g}^*(x) - \hat{l}_g(x)| \sqrt{\hat{f}_X(x) \{\hat{f}_{(\varepsilon|X=x_i)}(\varepsilon_i^*)\}} / \sqrt{\hat{E}_{Y|X} \{\psi^2(\varepsilon_i^*)\}}].$$

(6) Calculate the  $1 - \alpha$  quantile  $d_\alpha^*$  of  $d_1, \dots, d_n^*$ .

(7) Construct the bootstrap uniform band centered around  $\hat{l}_h(x)$

$$\hat{l}_h(x) \pm [\sqrt{\hat{f}_X(x) \{\hat{f}_{(\varepsilon|X=x_i)}(\varepsilon_i^*)\}} / \sqrt{\hat{E}_{Y|X} \{\psi^2(\varepsilon_i^*)\}}]^{-1} d_\alpha^*.$$

Fig. 1 shows the theoretical signal curve, the robust estimate using Huber loss function with a corresponding 95% uniform confidence band from the asymptotic theory and the confidence band using the bootstrap method. The real curve is marked as the gray solid line. We then compute the confidence band based on asymptotic theory according to [9]. Here, we notice that the asymptotic band is narrower than the bootstrap band. The width of the bands has not been affected by outliers since we adopt a robust estimation. To compare which method is more precise, Table 1 presents respectively the simulated coverage probabilities together with the calculated area of the 95% and 90% confidence band, for sample sizes  $n = 100, 200, 400$ . (Practically we would like to keep the sample size larger than 100 to achieve a reasonably precise estimation.) 5000 simulation runs are carried out and 1000 bootstrap samples are generated for each simulation. From Table 1, we observe that, for the asymptotic method, coverage probabilities improve with increased sample size and the bootstrap method (shown in brackets) obtains a larger coverage probability than the asymptotic one, as it has smaller relative errors. It is also observed that the sizes of the bands decrease with increased sample sizes. Overall, the bootstrap method displays a better convergence rate, while not sacrificing much on the width of the bands.

4.2. Additive model

We now extend the study to the case of multivariate covariates, where we use an additive model for the estimation. The bootstrap procedure is as follows:

(1) Simulate  $(X_i, Y_i), i = 1, \dots, n$  following model (11). The variables  $x_1, x_2, x_3, x_4 \sim U(-2.5, 2.5)$ ,

$$m_1(x_1) = \sin(\pi x_1), \quad m_2(x_2) = \Phi(3x_2), \quad m_3(x_3) = x_3^3, \quad m_4(x_4) = x_4^4,$$

and  $\varepsilon_i$  is simulated from a mixture normal density function with density  $\varphi(u/9)/90 + \varphi(u)/10$ .

(2) Compute the estimation  $\hat{m}_1(x_1), \hat{m}_2(x_2), \hat{m}_3(x_3), \hat{m}_4(x_4)$  via (13) and  $\hat{\varepsilon}_i = Y_i - \sum_{j=1}^4 \hat{m}_j(x_{i,j})$ .

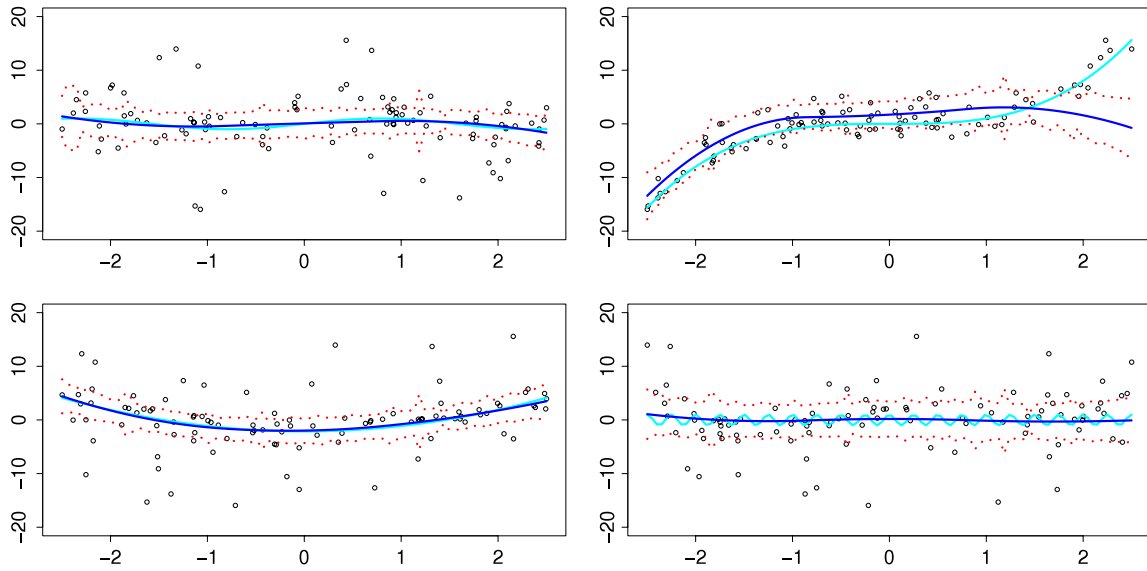
(3) For each  $i = 1, \dots, n$ , generate random variable  $\varepsilon_i^*, i = 1, \dots, n$  as in (15):

$$Y_{i,i^*} = \sum_{j=1}^4 \hat{m}_j(x_{i,j}) + \varepsilon_i^*.$$

(4) For each sample  $\{x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, y_i^*\}$ , compute  $\hat{m}_j^*(\cdot)$  and the random variable

$$d_i^* \stackrel{\text{def}}{=} \sup_{x \in [-2.5, 2.5]} \{ \sqrt{\hat{f}_{X_j}(x) \{\hat{f}_{(\varepsilon|X_j=x_{i,j})}(\varepsilon_i^*)\}} / \sqrt{\hat{E}_{Y|X_j=x_{i,j}} \{\psi^2(\varepsilon_i^*)\}} |\hat{m}_j^*(x) - \hat{m}_j(x)| \}.$$

(5) Calculate the  $1 - \alpha$  quantile  $d_\alpha^*$  of  $d_1, \dots, d_n^*$ .



**Fig. 2.** Plot of true curve (dark blue), robust estimates and bands (cyan), bootstrap band (red dotted). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**

Relative errors and areas of bootstrap bands for  $\hat{m}_1(\cdot)$ ,  $\hat{m}_2(\cdot)$ ,  $\hat{m}_3(\cdot)$ ,  $\hat{m}_4(\cdot)$  with 1000 repetitions per sample, and 5000 samples. Standard deviations measuring variability over samples are shown in brackets.

		$n$	$\hat{m}_1(\cdot)$	$\hat{m}_2(\cdot)$	$\hat{m}_3(\cdot)$	$\hat{m}_4(\cdot)$	
Rel. Err.	95%	100	0.08(0.051)	0.05(0.027)	0.08(0.037)	0.09(0.020)	
		200	0.07(0.042)	0.03(0.015)	0.02(0.018)	0.04(0.019)	
		400	0.04(0.023)	0.02(0.012)	0.01(0.010)	0.03(0.012)	
	90%	100	0.06(0.050)	0.09(0.079)	0.06(0.066)	0.04(0.022)	
		200	0.04(0.038)	0.05(0.026)	0.03(0.048)	0.02(0.014)	
		400	0.03(0.028)	0.02(0.013)	0.02(0.021)	0.01(0.010)	
	Area	95%	100	6.76(1.291)	6.99(1.213)	6.87(1.241)	6.69(1.546)
			200	5.54(1.112)	4.84(1.007)	4.98(1.115)	4.79(1.134)
			400	4.56(0.988)	3.78(1.001)	3.21(0.943)	3.76(1.019)
		90%	100	5.99(1.667)	5.45(1.472)	5.75(1.987)	6.01(1.654)
			200	4.84(1.331)	4.38(1.112)	4.13(1.198)	4.11(1.219)
			400	3.51(0.969)	3.23(0.989)	2.98(0.823)	3.09(1.012)

(6) Construct the bootstrap uniform band centered around  $\hat{m}_j(x_j)$

$$\hat{m}_j(x_{i,j}) \pm \left[ \sqrt{\hat{f}_{x_{i,j}}(x_j) \{ \hat{f}_{(\varepsilon|X_j=x_{i,j})}(\varepsilon_i^*) \}} / \sqrt{\hat{E}_{Y|X_j=x_{i,j}} \{ \psi^2(\varepsilon_i^*) \}} \right] d_\alpha^*$$

The estimation of  $\hat{m}_j(x_j)$ s ( $j = 1, \dots, 4$ ) and their bootstrap confidence bands is shown in Fig. 2.

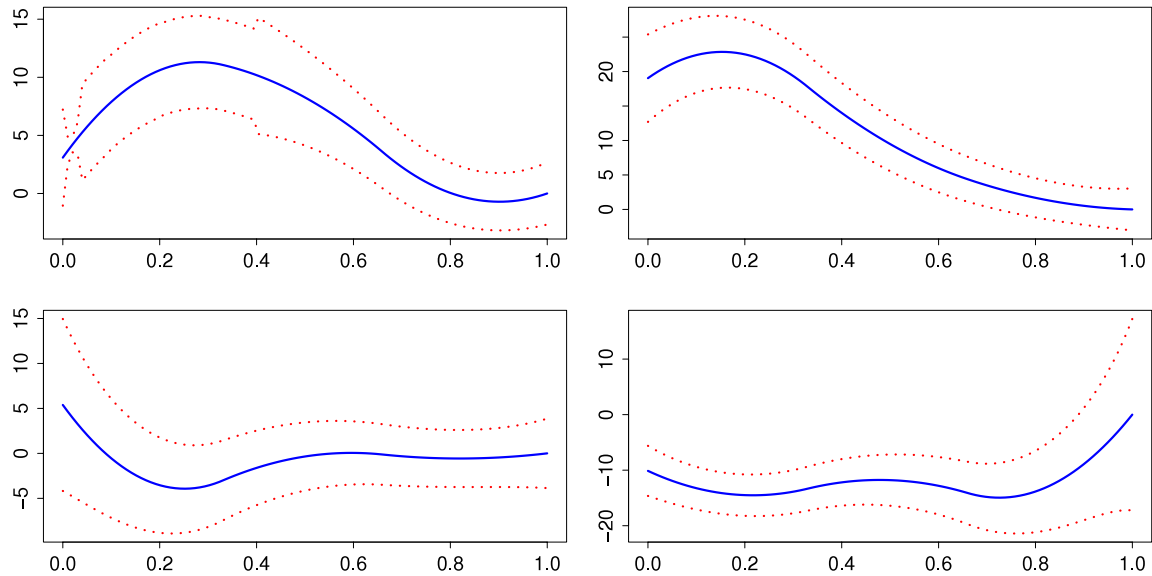
The simulated coverage probabilities are shown in Table 2. The coverage probabilities are close to the nominal levels and the widths of the bands are clearly shrinking w.r.t. the sample sizes.

## 5. Empirical analysis

### 5.1. Firm expenses analysis

[35] uses a sample of Japanese firms in the chemical industry to examine whether a concentrated shareholding is associated with lower expenditure on activities with scope for managerial private benefits. We focus on the same regression problem as proposed in [17]. The dependent variable  $Y$  is: general sales and administrative expenses deflated by sales (denoted by MH5), which is one of five measures of expenditures on activities with scope for managerial private benefits considered. The covariates are: ownership concentration (denoted by TOPTEN, cumulative shareholding by the largest ten shareholders), and firm characteristics: the log of assets, firm age, and leverage (the ratio of debt to debt plus equity), the sample size is 185. The regression model we consider here is:

$$MH5 = m_0 + m_1(TOPTEN) + m_2\{\log(Assets)\} + m_3(Age) + m_4(Leverage) + error.$$



**Fig. 3.** Robust estimates (blue), bootstrap bands (red dotted), left up:  $\text{Log}(\text{Asset})$ , right up: Leverage, left below: Age, right below: TOPTEN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The estimated additive components and its bootstrap confidence bands are shown in Fig. 3. Similarly, it can be seen that the nonlinear effects are  $\text{log}(\text{asset})$  and TOPTEN, and the firm age effects are minor compared to the other three. The effect of leverage is also a little bit nonlinear, and the shape of the curves deviates from what [17] presents, especially for the effect of TOPTEN. This may be due to the different subjects studied: in our case, robust estimation with Tukey biweight loss, while in their case the conditional median curve.

## 5.2. The impact on the stock market

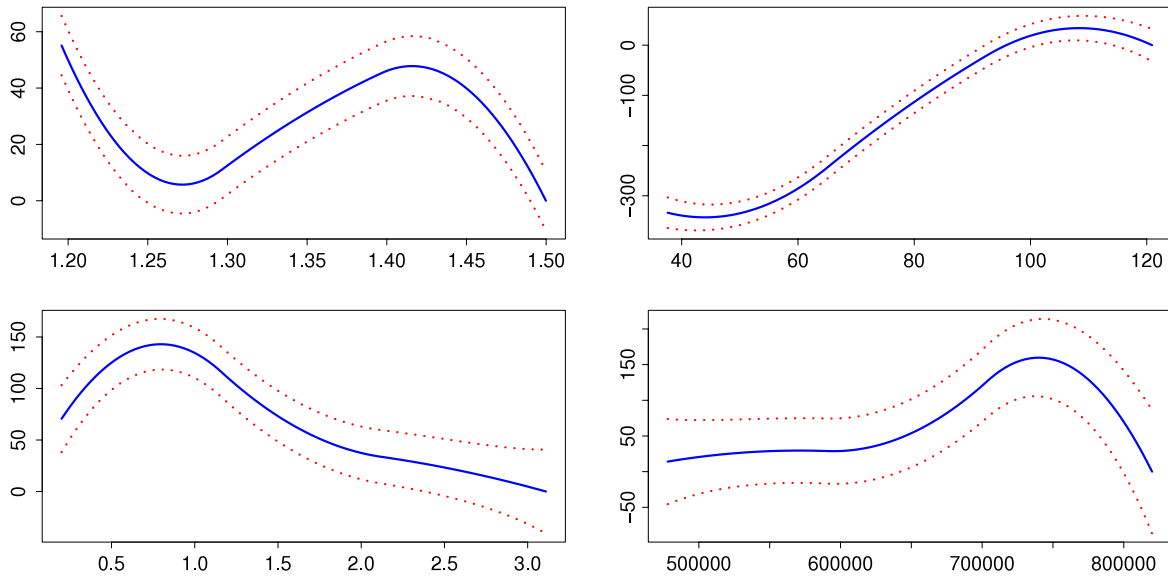
We analyze how four assets (oil, currency, bonds, real estate) affect the stock market. This study would give implications on the interactions of the economic conditions among different sectors. The data source is ProQuest Statistical Datasets, and we focus on the United States market. Therefore, the covariates are taken as: the crude oil index, EUR–USD exchange rate, the 10-year treasury constant maturity inflation index %, the real estate index, and the Y variable is S&P 500 index returns. The data are synchronized to a weekly frequency. We selected the data during the period of 09/03/2008–11/28/2011.

It can be observed that all of the four markets have nonlinear effects on the stock indices values, Fig. 4, but only the exchange rate EUR–USD and crude oil prices affect the stock indices returns nonlinearly, Fig. 5. It is not difficult to interpret the relationships: In Fig. 4, for the exchange rate EUR–USD, the weakness of EUR up to a certain level ( $< 1.27$ ) is negatively correlated with the stock indices, and then a positive correlation follows, but this relationship is again reversed when the EUR is too high ( $> 1.43$ ). Oil prices have a negative impact on stock indices at every level, but the effects decrease when the prices rise. As for the inflation index, when the inflation rate is high, interest rates are typically high, this may reduce the consumption and investments in the stock market. One can see a negative correlation there when the inflation index is larger than 0.7. Finally, an increasing real estate index can be a sign of booming economic conditions, therefore the stock indices rise when the real estate index gets higher. However, when the real estate index is too high, it is likely that there exists a bubble, so a drop in the market indices is seen.

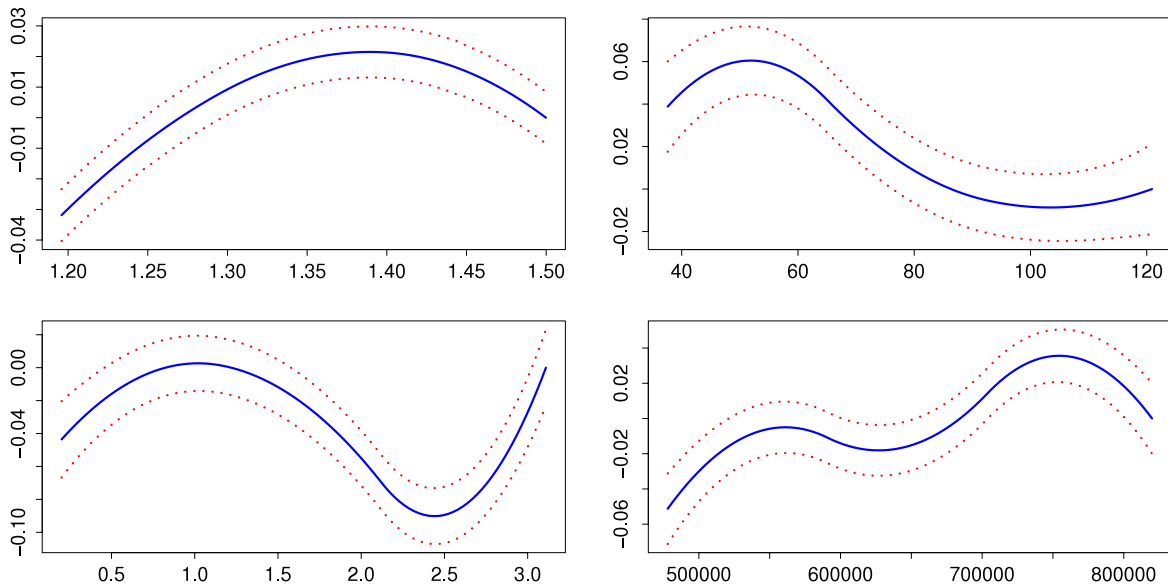
In Fig. 5, we see a difference in effects on S&P log returns, exchange rate EUR–USD is positively correlated with returns until a high level ( $> 1.40$ ), the crude oil has major negative effects on stock returns. More nonlinearity is presented in the plots for the inflation index and the real estate index.

## 6. Conclusion

We have developed and proved the bootstrap improvement for a wide class of smoothers with bounded influence functions. Moreover, we extended our results to additive models in order to cope with the curse of dimensionality. Our bootstrap method can be further extended to serve various purposes in additive models. It may be used for componentwise hypothesis testing, testing the separability or interaction and bias correction, see [32]. Furthermore, similar bootstrap improvement results can be extended to other semiparametric models, e.g. partial linear models, and single index models. Moreover, different types of resampling procedures can be adopted to directly resample  $(Y_i, X_i)$ , such as in [27,22]. And, according to [1], we are able to deal with cases with discrete covariates.



**Fig. 4.** Robust estimate (blue), bootstrap bands (red dotted),  $Y$ : S&P index, left up: exchange rates EUR–USD, right up: crude oil price, left below: inflation index, right below: real estate index. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



**Fig. 5.** Robust estimate (blue), bootstrap bands (red dotted),  $Y$ : S&P index log return, left up: exchange rates EUR–USD, right up: crude oil price, left below: inflation index, right below: real estate index. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Acknowledgment**

The financial support from ISF (Complex statistical models and time (ISF, 2010–2014)) is greatly acknowledged.

**Appendix. Proof**

*A.1. Proof of Theorem 3.1 in the  $d = 1$  case*

To prove Theorem 3.1, we show finally, with (optimal rate)  $h = \mathcal{O}(n^{-1/5})$ :

$$\max_i [(\hat{l}_h^i - l)(X_i) - \{(\hat{l}_{h,g}^* - \hat{l}_g)(X_i)\}] = \mathcal{O}_p(h^2 \Gamma_n). \tag{A.1}$$

This is extended to any point  $x$  in a compact set  $B$  in (A.12).

Proving (A.1) can be done by showing first:

$$\max_i |\psi(\varepsilon_i^*) - \psi(\varepsilon_i^\#)| = \mathcal{O}_p(h^2 \Gamma_n). \tag{A.2}$$

Recall the definition of  $\tilde{F}_{\psi|X}(\cdot)$  in (24), it is the induced conditional cdf of  $\{\psi(\varepsilon_i^*)\}_{i=1}^n$ . By [5] we have for a small  $b$ :

$$\sup_{|t| \leq b, i=1, \dots, n} |\tilde{F}_{\psi|X}(t) - F_{(\varepsilon|X)}(t)| = \mathcal{O}_p(h^2 \Gamma_n). \tag{A.3}$$

Also recall (24)  $\hat{F}_{(\varepsilon|X)}$  is a scaled version of  $\tilde{F}_{(\varepsilon|X)}$ . Then according to A.2, the inverse function of  $\hat{F}_{(\varepsilon|X)}(\cdot)$  and  $F_{(\varepsilon|X)}(\cdot)$  will also be close in the sense that

$$\max_i |\varepsilon_i^\# - \varepsilon_i^*| = \max_i |F_{(\varepsilon|X)}^{-1}(U_i) - \hat{F}_{(\varepsilon|X)}^{-1}(U_i)| = \mathcal{O}_p(h^2 \Gamma_n). \tag{A.4}$$

Moreover, for a  $\delta_i = \mathcal{O}_p(h^2 \Gamma_n)$  uniformly, define  $F_{\psi|X}^{-1}(\cdot)$  as the conditional c.d.f. of  $\psi(\varepsilon)$  on  $X$ ,

$$\begin{aligned} |\psi(\varepsilon_i^\#) - \psi(\varepsilon_i^*)| &= |F_{\psi|X}^{-1}[\hat{F}_{\psi|X}\{\psi(\varepsilon_i^*)|X_i\}|X_i] - \psi(\varepsilon_i^*)|, \\ &= |F_{\psi|X}^{-1}[F_{\psi|X}\{\psi(\varepsilon^*)|X_i\} + \delta_i|X_i] - \psi(\varepsilon_i^*)|, \\ &= |F_{\psi|X}^{-1}[F_{\psi|X}\{\psi(\varepsilon^*)|X_i\} + \delta_i|X_i] - F_{\psi|X}^{-1}[F_{\psi|X}\{\psi(\varepsilon^*)|X_i\}|X_i]|, \\ &\leq \frac{1}{c_1} \delta_i = \mathcal{O}_p(h^2 \Gamma_n), \end{aligned}$$

by A.2. Therefore, (A.2) is proved. As  $\psi(\cdot)$  plays a role in the estimation via the zero functions defined below, to prove (A.1), we first want to write the estimation difference  $(\hat{l}_n^\# - l)(X_i)$  and  $(\hat{l}_{h,g}^* - \hat{l}_g)(X_i)$  written w.r.t. to their zero functions defined as follows.

$$\begin{aligned} G_n^*(\theta, X_i) &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) [\psi\{\varepsilon_j^* - \theta + \hat{l}_g(X_j)\}], \\ &= \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) \{\psi(Y_j^* - \theta)\}, \\ G_n^\#(\theta, X_i) &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) [\psi\{\varepsilon_j^\# - \theta + l(X_j)\}], \\ &= \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) \{\psi(Y_j^\# - \theta)\}. \end{aligned}$$

Note that, for the first moment, we have the natural equality as follows, so we can focus on the difference in the second moment,

$$E_{\hat{F}_{\varepsilon|X_i=x}} \psi(\varepsilon_i^*) = 0 = E_{F_{\varepsilon|X_i=x}} \psi(\varepsilon_i^\#). \tag{A.5}$$

We abbreviate  $E_{F_{\varepsilon|X_i=x}}$  as  $E$  and  $E_{\hat{F}_{\varepsilon|X_i=x}}$  as  $E^*$ , define

$$\tilde{l}(x) \stackrel{\text{def}}{=} \underset{\theta}{\operatorname{argmin}} E \rho(Y - \theta) K_h(X - x), \tag{A.6}$$

$$\tilde{l}_g(x) \stackrel{\text{def}}{=} \underset{\theta}{\operatorname{argmin}} E^* \rho(Y^* - \theta) K_h(X - x), \tag{A.7}$$

$$T_n(X_i) \stackrel{\text{def}}{=} G_n^*(\tilde{l}_g(X_i), X_i) - G_n^\#(\tilde{l}(X_i), X_i) \tag{A.8}$$

as the unbiased versions of the true function. It is not hard to derive that the bias has the order of  $\mathcal{O}_p(h^2)$ .

$$\begin{aligned} \tilde{l}(X_i) - l(X_i) &= \frac{EG_n^\#(l(X_i), X_i)}{\partial EG_n^\#(l(X_i), X_i)} + \mathcal{O}(\tilde{l}(X_i) - l(X_i)), \\ &= \{h^2 l''(X_i)/2 + f'_X(X_i) l'(X_i) h^2 / f_X(X_i)\} \|K\|_s^2 + \mathcal{O}_p(h^2). \end{aligned}$$

Similarly for  $\tilde{l}_g(X_i) - \hat{l}_g(X_i)$ , we have

$$\begin{aligned} \tilde{l}_g(X_i) - \hat{l}_g(X_i) &= \frac{E^* G_n^*(l(X_i), X_i)}{\partial E E^* G_n^*(l(X_i), X_i)} + \mathcal{O}(\tilde{l}_g(X_i) - \hat{l}_g(X_i)), \\ &= \{h^2 \hat{l}_g''(X_i)/2 + f'_X(X_i) \hat{l}_g'(X_i) h^2 / f_X(X_i)\} \|K\|_s^2 + \mathcal{O}_p(h^2), \end{aligned}$$

where  $\|K\|_s^2 \stackrel{\text{def}}{=} \int s^2 K^2(s) ds$ . A balance between bias and variance term would lead to the choice of the rate  $h$  as  $\mathcal{O}(n^{-1/5} \Gamma_n)$ , and we have  $|\tilde{l}(X_i) - l(X_i) - \{\tilde{l}_g(X_i) - \hat{l}_g(X_i)\}| = \mathcal{O}_p(h^2)$ , so we can write the difference of the bias term as  $\mathcal{O}_p(h^2)$  in the following derivation.

**Remark.** In the case when  $\psi(\cdot)$  is not differentiable, in particular for the quantile case, the above stochastic expansion is still valid, see the stochastic expansion in [20].

According to A.1, excluding the non-differentiable  $\psi(\cdot)$  case, one can use the Lipschitz condition of the function  $\psi(\cdot)$ , and  $\psi(\cdot)$  is bounded, we have,  $\exists$  a constant  $C$ , such that,

$$\begin{aligned} \max_i |T_n(X_i)| &= \max_i \left| \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) \{ \psi(\varepsilon_j^* - \tilde{l}_g(X_i) + \hat{l}_g(X_j)) - \psi(\varepsilon_j^\# - \tilde{l}(X_i) + l(X_j)) \} \right|, \\ &\leq \max_i \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) |C[\{\varepsilon_j^* - \tilde{l}_g(X_i) + \hat{l}_g(X_j)\} - \{\varepsilon_j^\# - \tilde{l}(X_i) + l(X_j)\}]|. \end{aligned}$$

So we can break the upper bound of  $\max_i |T_n(X_i)|$  into two terms, the first term  $\max_i T_{n,1}(X_i)$  involves the bootstrapped error and its theoretical couple, which has been proved by (A.3), and the second term  $\max_i T_{n,2}(X_i)$  is concerning the convergence rate of  $\hat{l}_g(\cdot)$ ,

$$\begin{aligned} \max_i |T_n(X_i)| &\leq \max_i \frac{1}{n} C \sum_{j=1}^n W_{h,j}(X_i) |\varepsilon_j^* - \varepsilon_j^\#| + \max_i \frac{1}{n} C \sum_{j=1}^n W_{h,j}(X_i) |[\hat{l}_g(X_j) - \tilde{l}_g(X_i)] - [l(X_j) - \tilde{l}(X_i)]|, \\ &\leq \max_i T_{n,1}(X_i) + \max_i T_{n,2}(X_i). \end{aligned}$$

$\max_i |T_{n,1}(X_i)|$  is known to have the rate  $\mathcal{O}_p(n^{-1/2} h^{3/2} \Gamma_n)$ , and

$$\max_i T_{n,2}(X_i) = \max_i \frac{1}{n} C \sum_{j=1}^n W_{h,j}(X_i) |\{\hat{l}_g(X_{i,j,0}) - l'(X_{i,j,0})\}(X_i - X_j)| + \mathcal{O}_p(h^2),$$

where  $X_{i,j,0}$  is a point between  $X_i$  and  $X_j$ , and  $C$  is a constant, according to the mean value theorem.  $\sup_{x \in B} |\hat{l}'_g(x) - l'(x)|$  is of the rate  $\mathcal{O}_p(g^{-1}(ng)^{-1/2} \Gamma_n + g^3)$ , see [33]. Therefore the optimal rate for  $g$  would be  $\mathcal{O}(n^{-1/9})$  in our case (as in A.4), this rate is slower than the choice of  $h$ , which confirms the results in [12]. Then we can achieve

$$\max_i T_{n,2}(X_i) = \mathcal{O}_p(h^2 \Gamma_n).$$

As the second derivative of the loss function  $\psi(\cdot)$  does not exist at zero, we use the local version of equicontinuity lemma from chapter VII.1 in [29]. Our loss function has the following expansion with the remainder term defined as  $r\{y, \theta(x)\}$

$$\rho\{y - \theta(x)\} = \rho\{y - l(x)\} + \{\theta - l(x)\} \psi\{y - l(x)\} + |\theta - l(x)| r\{y, \theta(x)\}, \tag{A.9}$$

define  $R_n^*(Y^*, \tilde{l}_g(X_i)) \stackrel{\text{def}}{=} n^{-1} \sum_{j=1}^n r(Y_j^*, \tilde{l}_g(X_i))$ ,  $R_n(Y, \tilde{l}(X_i)) \stackrel{\text{def}}{=} n^{-1} \sum_{j=1}^n r(Y_j, \tilde{l}(X_i))$  are the high order function, we need to prove it satisfies the equicontinuity condition around the point  $\tilde{l}_g(X_i)$  and  $\tilde{l}(X_i)$ , namely, for each  $\eta$  and  $\epsilon$ ,  $\exists \delta$  such that,

$$\limsup_n P\left( \sup_{l': \|\tilde{l}'_g - \tilde{l}_g\| < \delta} \|R_n^*(l'_g) - R_n^*(\tilde{l}_g)\| > \eta \right) < \epsilon, \tag{A.10}$$

$$\limsup_n P\left( \sup_{l': \|\tilde{l}' - \tilde{l}\| < \delta} \|R_n(l') - R_n(\tilde{l})\| > \eta \right) < \epsilon. \tag{A.11}$$

According to the equicontinuity lemma in [29], first of all we can prove that  $|r(y, \theta)|$  has an envelope, by having the function  $r(y, \theta)$  satisfying certain conditions.

So we can achieve the estimations are linked to the zero functions around the point  $\tilde{l}_g(X_i)$  and  $\tilde{l}(X_i)$  respectively, as follows:

$$\begin{aligned} \hat{l}_{h,g}^*(X_i) - \tilde{l}_g(X_i) &= -\frac{G_n^* \{\tilde{l}(X_i), X_i\}}{\partial E E^* G_n^* \{\tilde{l}_g(X_i), X_i\}} + \mathcal{O}_p((nh)^{-1/2}), \\ \hat{l}_h^\#(X_i) - \tilde{l}(X_i) &= -\frac{G_n^\# \{\tilde{l}(X_i), X_i\}}{\partial E G_n^\# \{\tilde{l}(X_i), X_i\}} + \mathcal{O}_p((nh)^{-1/2}), \end{aligned}$$

where  $\partial E E^* G_n^* \{\tilde{l}_g(X_i), X_i\}$  denotes the partial derivative of  $E E^* G_n^* \{\theta, X_i\}$  w.r.t.  $\theta$  taking value at the point  $\tilde{l}_g(X_i)$ .



This means,

$$\begin{aligned} |\hat{l}_{h,g}^*(X_i) - \hat{l}_g(X_i) - \tilde{l}_h^\sharp(X_i) + l(X_i)| &= -\frac{G_n^* \{\tilde{l}_g(X_i), X_i\}}{\partial E E^* G_n^* \{\tilde{l}_g(X_i), X_i\}} + \frac{G_n^\sharp \{\tilde{l}(X_i), X_i\}}{\partial E G_n^\sharp \{\tilde{l}(X_i), X_i\}} + \mathcal{O}_p(h^2), \\ &= -\frac{\partial E G_n^\sharp \{\tilde{l}(X_i), X_i\} [G_n^* \{\tilde{l}_g(X_i), X_i\} - G_n^\sharp \{\tilde{l}(X_i), X_i\}]}{\partial E E^* G_n^* \{\tilde{l}_g(X_i), X_i\} \partial E G_n^\sharp \{\tilde{l}(X_i), X_i\}} \\ &\quad + \frac{G_n^\sharp \{\tilde{l}(X_i), X_i\} [\partial E E^* G_n^* \{\tilde{l}_g(X_i), X_i\} - \partial E G_n^\sharp \{\tilde{l}(X_i), X_i\}]}{\partial E E^* G_n^* \{\tilde{l}_g(X_i), X_i\} \partial E G_n^\sharp \{\tilde{l}(X_i), X_i\}} + \mathcal{O}_p(h^2). \end{aligned}$$

Therefore, since  $G_n^* \{\hat{l}_g(X_i), X_i\}$  and  $G_n^\sharp \{l(X_i), X_i\}$  are known by S.L.L.N. to have strong consistency to  $E^* G_n^* \{\hat{l}_g(X_i), X_i\}$  and  $E G_n^\sharp \{l(X_i), X_i\}$ , we have,

$$\max_i |\hat{l}_{h,g}^*(X_i) - \hat{l}_g(X_i) - \tilde{l}_h^\sharp(X_i) + l(X_i)| = \mathcal{O}(\max_i T_n(X_i)) + \mathcal{O}_p(h^2 \Gamma_n) + \mathcal{O}_p(h^2 \Gamma_n),$$

and (A.1) is proved.

Define the order statistics  $X_{(1)}, \dots, X_{(n)}$  of  $X_1, \dots, X_n$ , so the claim (28) can be proved from (A.1) using the fact that

$$\sup |A_n(x)| \leq \max_i |A_n(X_{(i)})| + \max_i \sup_{x \in [X_{(i)}, X_{(i+1)}]} |A_n(X_{(i)}) - A(x)|, \tag{A.12}$$

it suffices to consider the speed of the last term. With Lipschitz continuity of  $A_n(\cdot)$ :

$$\max_i \sup_{x \in [X_{(i)}, X_{(i+1)}]} |A_n(X_{(i)}) - A(x)| \leq c_2 \max_i \sup_x |X_i - x|, \tag{A.13}$$

where  $c_2 > 0$  is a constant, this upper random bound is of order  $\mathcal{O}_p(n^{-1/d} \log n) = \mathcal{O}_p(h^2 \Gamma_n)$ . The uniform bound for  $\|X_i - x\|$  results from the uniform law of large numbers over a ball of size  $n^{-1/d}$ , see [26, Theorem 1.1].

**Remark.** For the non-differentiable  $\psi(\cdot)$  cases, in particular the quantile regression case, one can still establish similar inequality, as

$$\begin{aligned} \max_i \left| \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) \{ \psi(\varepsilon_j^* - \tilde{l}_g(X_i) + \hat{l}_g(X_j)) - \psi(\varepsilon_i^\sharp - \tilde{l}(X_i) + l(X_j)) \} \right| \\ \leq \max_i \frac{1}{n} \sum_{j=1}^n W_{h,j}(X_i) | \psi\{\varepsilon_j^* - \tilde{l}_g(X_i) + \hat{l}_g(X_j)\} - \psi\{\varepsilon_i^\sharp - \tilde{l}(X_i) + l(X_j)\} |. \end{aligned}$$

Define  $\psi_{i,j} \stackrel{\text{def}}{=} | \psi\{\varepsilon_j^* - \tilde{l}_g(X_i) + \hat{l}_g(X_j)\} - \psi\{\varepsilon_i^\sharp - \tilde{l}(X_i) + l(X_j)\} | = \mathcal{O}_p(h^2)$  because, so

$$\begin{aligned} P(\psi_{i,j} > c_3 h^2) &\leq E(|\psi_{i,j}|) / (c_3 h^2) \\ &= \mathcal{O}_p(\{ \tilde{l}_g(X_i) + \hat{l}_g(X_j) - \tilde{l}(X_i) + l(X_j) \} / h^2). \end{aligned}$$

Then the argument follows from the above proof.

Moreover, one can also use the strong consistency of  $G_n^\sharp, G_n^*$  to  $E(G_n^\sharp)$  and  $E^* G_n^*$  respectively based on A.1 and Lemma 2.4 of [25].

### A.2. Proof of second part of Theorem 3.1

The number of regressors in (11) is of order  $p = H^{-1}d + 1$ . [30] shows that as long as  $n^{-1}(p \log n)^{3/2} \rightarrow 0$  then the estimators of the regression parameters are consistent and have the standard variance. In our situation,

$$n^{-1} n^{1/5*2/3} \log n = \mathcal{O}(1), \tag{A.14}$$

and therefore the condition is satisfied.

Now we have a look at the behavior of the design matrix in (12).

$$\begin{aligned} \hat{\mathcal{L}}(A) &\stackrel{\text{def}}{=} -n^{-1} \sum_{i=1}^n \rho\{Y_i - A^\top \Phi(X_i)\}, \\ \nabla \hat{\mathcal{L}}(A) &\stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n \psi\{Y_i - A^\top \Phi(X_i)\} \Phi(X_i), \\ \nabla^2 E \hat{\mathcal{L}}(A) &\stackrel{\text{def}}{=} -\nabla E \psi\{Y_i - A^\top \Phi(X_i)\} \Phi(X_i) \Phi(X_i)^\top. \end{aligned}$$

Recall that

$$\hat{A} = \underset{A}{\operatorname{argmin}} \hat{\mathcal{L}}(A).$$

Lemma 14 of [34] ensures that with probability approaching 1,  $\hat{A}$  exists uniquely and that  $\nabla \hat{\mathcal{L}}(\hat{A}) = 0$ . In addition, there exists  $\bar{m}(x) = \bar{A}^\top \Phi(x)$ , such that

$$\sup_{x \in B} |\bar{m}(x) - m(x)| \leq C_\infty H^2. \tag{A.15}$$

According to the bounded influence condition A.1, the  $j$ th term,  $j \in 1, 2, \dots, H^{-1}d + 1$ .

$$\begin{aligned} | \{ \nabla \hat{\mathcal{L}}(\bar{A}) \}_j | &= \left| \left\{ -n^{-1} \sum_{i=1}^n [\psi \{ m(X_i) + \varepsilon_i - \bar{A}^\top \Phi(X_i) \} - \psi(0)] \Phi(X_i) \right\}_j \right|, \\ &\leq \left[ n^{-1} \sum_{i=1}^n C_3 | \{ A^\top \Phi(X_i) + \varepsilon_i - \bar{A}^\top \Phi(X_i) \} \Phi(X_i) | \right]_j, \\ &\leq \left[ n^{-1} \sum_{i=1}^n C_3 | \{ A^\top \Phi(X_i) + \varepsilon_i - \bar{A}^\top \Phi(X_i) \} \Phi(X_i) | \right]_j. \end{aligned}$$

We know first that,

$$E[ \{ m(X_i) - \bar{m}(X_i) \} \Phi(X_i) ]_j = \mathcal{O}(H^2).$$

Let  $\xi_{i,j} \stackrel{\text{def}}{=} \{ |m(X_i) - \bar{m}(X_i) \Phi(X_i)| - E |m(X_i) - \bar{m}(X_i) \Phi(X_i)| \}_j$ , by Bernstein's Lemma:

**Lemma A.1.** Let  $Z_1, \dots, Z_n$  be independent r.v.s.

$$\log E \exp(tZ_i) \leq E(Z_i^2)t^2/2$$

for all  $t \in [0, \infty]$ . Then

$$P \left( \left| \sum_{i=1}^n Z_i \right| \geq t \sqrt{2 \sum_{i=1}^n E Z_i^2} \right) \leq 2 \exp(-t^2).$$

Finally, we can derive that,

$$n^{-1} \sum_{i=1}^n \xi_{i,j} = \mathcal{O}_{a.s.}(H^2 n^{-1/2} \Gamma_n), \quad j \neq 1. \tag{A.16}$$

The last term

$$n^{-1} \left| \left\{ \sum_{i=1}^n \varepsilon_i \Phi(X_i) \right\}_j \right| = \mathcal{O}_{a.s.}(n^{-1/2} \Gamma_n). \tag{A.17}$$

Therefore, one has a collective term from (A.16) and (A.17),

$$\| \nabla \hat{\mathcal{L}}(\bar{A}) \| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2} n^{-1/2} \Gamma_n),$$

where  $\| \cdot \|$  denotes the  $L_2$  norm.

By assumptions A.5, A.7,  $\forall l = 1, \dots, H^{-1}$ , the  $d$  dimensional vector  $\Phi_l^\top(X_i)$  satisfies,

$$\beta \|b\|^2/d \geq E b^\top \Phi_l^\top(X_i) \Phi_l(X_i) b \geq \alpha \|b\|^2/d,$$

where  $\alpha$  and  $\beta$  are two constants.

**Lemma A.2.** Assume A.1 and A.5, as  $n \rightarrow \infty$ ,

$$\begin{aligned} \|\hat{A} - A\| &= \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2} n^{-1/2} \Gamma_n), \\ \max_{i \in 1, \dots, n} \|\hat{m}(X_i) - \bar{m}(X_i)\| &= \mathcal{O}_{a.s.}(H + H^{-1} n^{-1/2} \Gamma_n). \end{aligned}$$

**Proof.** According to similar equicontinuity arguments, exists an  $(H^{-1}d + 1) \times (H^{-1}d + 1)$  matrix  $\bar{A}$ , such that

$$\|\hat{A} - \bar{A}\| = \mathcal{O}(\|\{\nabla^2 \mathbb{E} \hat{\mathcal{L}}(\bar{A}_0)\}^{-1} \{-\nabla \hat{\mathcal{L}}(\bar{A})\}\|),$$

since

$$\nabla^2 \mathbb{E} \hat{\mathcal{L}}(X_i) = \Phi(X_i) \Phi(X_i)^\top \nabla \mathbb{E} \psi(Y_i - \hat{A}_0^\top \Phi(X_i)).$$

According to assumption A.7,

$$c_3 I_{H^{-1}d+1} \geq \nabla^2 \mathbb{E} \hat{\mathcal{L}}(X_i) \geq c_4 I_{H^{-1}d+1},$$

therefore

$$\|\hat{A} - \bar{A}\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2} n^{-1/2} \Gamma_n).$$

Moreover, by Cauchy–Schwarz inequality

$$\begin{aligned} \max_{i \in 1, \dots, n} |\hat{m}(X_i) - \bar{m}(X_i)| &\leq \max_i \|\hat{A} - \bar{A}\| \|\Phi(X_i)\|, \\ &= \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2} n^{-1/2} \Gamma_n) \mathcal{O}_{a.s.}(H^{-1/2}), \\ &= \mathcal{O}_{a.s.}(H + H^{-1} n^{-1/2} \Gamma_n). \end{aligned}$$

We would like to check for the pseudo observations  $Y_i^\# = m(X_i) + \varepsilon_i^\#$ .

$$\begin{aligned} &|\{\hat{m}_k^*(X_{i,k}) - \hat{m}_k(X_{i,k})\} - \{\hat{m}_k^\#(X_{i,k}) - m_k(X_{i,k})\}| \\ &\leq |(\hat{A}^{*\top} - \hat{A}^\top - \hat{A}^{\#\top} + A^\top) \Phi(X_{i,k})|, \\ &= \mathcal{O}(\|\{\nabla^2 \mathbb{E} \hat{\mathcal{L}}^*(\hat{A})\}^{-1} \{-\nabla \hat{\mathcal{L}}^*(\hat{A})\} - \nabla^2 \mathbb{E} \hat{\mathcal{L}}(\bar{A})^{-1} \{-\nabla \hat{\mathcal{L}}(\bar{A})\}\| \|\Phi(X_i)\| + H^2), \end{aligned}$$

as  $\nabla^2 \mathbb{E} \hat{\mathcal{L}}^*(\hat{A})^{-1}$  and  $\nabla^2 \mathbb{E} \hat{\mathcal{L}}(\bar{A})^{-1}$  are both bounded,

$$\begin{aligned} &|\{\hat{m}_k^*(X_{i,k}) - \hat{m}_k(X_{i,k})\} - \{\hat{m}_k^\#(X_{i,k}) - m_k(X_{i,k})\}| \\ &= \mathcal{O} \left( \left| n^{-1} \sum_{i=1}^n \{\psi(\varepsilon_i^*) - \psi(\varepsilon_i^\# + \bar{A}^\top \Phi(X_i) - A^\top \Phi(X_i))\} \Phi(X_i)^\top \Phi(X_i) \right| + H^2 \right), \\ &= \mathcal{O} \left( \left| n^{-1} \sum_{i=1}^n \{\psi(\varepsilon_i^*) - \psi(\varepsilon_i^\#)\} \Phi(X_i)^\top \Phi(X_i) \right| + H^2 \right), \\ &= \mathcal{O} \left( \left| n^{-1} \sum_{i=1}^n \{\varepsilon_i^* - \varepsilon_i^\#\} \Phi(X_i)^\top \Phi(X_i) \right| + H^2 \right), \\ &= \mathcal{O} \left( \left| n^{-1} \sum_{i=1}^n \{Z_i |\hat{\varepsilon}_i| - Z_i \eta_i\} \Phi(X_i)^\top \Phi(X_i) \right| + H^2 \right). \end{aligned}$$

One can derive using a coupling argument,

$$\begin{aligned} |\{\hat{m}_k^*(X_{i,k}) - \hat{m}_k(X_{i,k})\} - \{\hat{m}_k^\#(X_{i,k}) - m_k(X_{i,k})\}| &= \mathcal{O}_{a.s.}(n^{-1/2}(H^2 + n^{-1/2} \Gamma_n)H^{-1/2} + H^2) \\ &= \mathcal{O}_{a.s.}(H^2). \end{aligned}$$

## References

- [1] D. Basso, M. Chiarandini, L. Salmaso, Synchronized permutation tests in replicated  $i \times j$  designs, *J. Statist. Plann. Inference* 137 (8) (2007) 2564–2578.
- [2] P.J. Bickel, M. Rosenblatt, On some global measures of the deviations of density function estimates, *Ann. Statist.* 1 (6) (1973) 1071–1095.
- [3] G. Claeskens, I. Van Keilegom, Bootstrap confidence bands for regression curves and their derivatives, *Ann. Statist.* 31 (6) (2003) 1852–1884.
- [4] A. Deaton, J. Muellbauer, An almost ideal demand system, *Amer. Econ. Rev.* 40 (3) (1980) 312–326.
- [5] J. Franke, P. Mwita, W. Wang, Nonparametric estimates for conditional quantiles of time series, SFB 649 Discussion Paper.
- [6] M. Fuss, D. McFadden, Y. Mundlak, A survey of functional forms in the economic analysis of production, Vol. 1, McMaster University Archive for the History of Economic Thought, 1978.
- [7] P. Hall, Edgeworth expansions for nonparametric density estimators, with applications, *Statistics* 22 (2) (1991) 215–232.
- [8] P. Hall, *The Bootstrap and Edgeworth Expansion*, Springer Verlag, 1992.
- [9] W.K. Härdle, Asymptotic maximal deviation of  $M$ -smoothers, *J. Multivariate Anal.* 29 (2) (1989) 163–179.
- [10] W.K. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, 1990.
- [11] W.K. Härdle, J. Horowitz, J.-P. Kreiss, Bootstrap method for time series, *Int. Statist. Rev.* 71 (2) (2003) 435–459.
- [12] W.K. Härdle, S.J. Marron, Bootstrap simultaneous error bars for nonparametric regression, *Ann. Statist.* 19 (2) (1991) 778–796.
- [13] W.K. Härdle, S. Song, Confidence bands in quantile regression, *Econometric Theory* 26 (4) (2010) 1180–1200.
- [14] J.L. Horowitz, Nonparametric estimation of a generalized additive model with an unknown link function, *Econometrica* 69 (2) (2001) 499–513.
- [15] J.L. Horowitz, The bootstrap, in: *Handbook of Econometrics*, Vol. 5, Elsevier, 2001.

- [16] J.L. Horowitz, J. Klemelä, E. Mammen, Optimal estimation in additive regression models, *Bernoulli* 12 (2) (2006) 271–298.
- [17] J.L. Horowitz, S. Lee, Nonparametric estimation of an additive quantile regression model, *J. Amer. Statist. Assoc.* 100 (472) (2005) 1238–1249.
- [18] P.J. Huber, Robust estimation of a location parameter, *Ann. Math. Statist.* 35 (1) (1964) 73–101.
- [19] G.J. Johnston, Probabilities of maximal deviations for nonparametric regression function estimates, *J. Multivariate Anal.* 12 (3) (1982) 402–414.
- [20] E. Kong, O. Linton, Y. Xia, Uniform Bahadur representation for local polynomial estimates of  $M$ -regression and its application to the additive model, *Econometric Theory* 26 (05) (2010) 1529–1564.
- [21] E. Mammen, *When Does Bootstrap Work?: Asymptotic Results and Simulations*, Springer Verlag, 1992.
- [22] M. Marozzi, Applications in business, medical and industrial statistics of bi-aspect nonparametric tests for location problems, *Stat. Methods Appl.* 12 (2) (2003) 187–194.
- [23] M. Marozzi, Multivariate tri-aspect non-parametric testing, *J. Nonparametr. Stat.* 19 (6–8) (2007) 269–282.
- [24] M.H. Neumann, J.-P. Kreiss, Regression-type inference in nonparametric autoregression, *Ann. Statist.* 26 (4) (1998) 1570–1613.
- [25] W.K. Newey, D.L. McFadden, Large sample estimation and hypothesis testing, in: *Handbook of Econometrics*, Vol. 4, 1986, pp. 2111–2245.
- [26] M.D. Penrose, A strong law for the largest nearest-neighbour link between random points, *J. Lond. Math. Soc.* 60 (3) (1999) 951–960.
- [27] F. Pesarin, L. Salmaso, Finite-sample consistency of combination-based permutation tests with application to repeated measures designs, *J. Nonparametr. Stat.* 22 (5) (2010) 669–684.
- [28] F. Pesarin, L. Salmaso, *Permutation Tests for Complex Data: Theory, Applications and Software*, John Wiley & Sons, 2010.
- [29] D. Pollard, *Convergence of Stochastic Processes*, Springer Verlag, 1984.
- [30] S. Portnoy, Local asymptotics for quantile smoothing splines, *Ann. Statist.* 25 (1) (1997) 414–434.
- [31] S. Song, Y. Ritov, W.K. Härdle, Bootstrap confidence bands and partial linear quantile regression, *J. Multivariate Anal.* 107 (2012) 244–262.
- [32] S. Sperlich, D. Tjøstheim, L. Yang, Nonparametric estimation and testing of interaction in additive models, *Econometric Theory* 18 (02) (2002) 197–251.
- [33] C.J. Stone, Optimal global rates of convergence for nonparametric regression, *Ann. Statist.* 10 (4) (1982) 1040–1053.
- [34] C.J. Stone, Additive regression and other nonparametric models, *Ann. Statist.* 13 (2) (1985) 689–705.
- [35] Y. Yafeh, O. Yosha, Large shareholders and banks: who monitors and how? *Econ. J.* 113 (2003) 128–146.

# HIDDEN MARKOV STRUCTURES FOR DYNAMIC COPULAE

WOLFGANG KARL HÄRDLE, OSTAP OKHRIN, AND WEINING WANG  
*Humboldt-Universität zu Berlin*

Understanding the time series dynamics of a multi-dimensional dependency structure is a challenging task. Multivariate covariance driven Gaussian or mixed normal time varying models have only a limited ability to capture important features of the data such as heavy tails, asymmetry, and nonlinear dependencies. The present paper tackles this problem by proposing and analyzing a hidden Markov model (HMM) for hierarchical Archimedean copulae (HAC). The HAC constitute a wide class of models for multi-dimensional dependencies, and HMM is a statistical technique for describing regime switching dynamics. HMM applied to HAC flexibly models multivariate dimensional non-Gaussian time series.

We apply the expectation maximization (EM) algorithm for parameter estimation. Consistency results for both parameters and HAC structures are established in an HMM framework. The model is calibrated to exchange rate data with a VaR application. This example is motivated by a local adaptive analysis that yields a time varying HAC model. We compare its forecasting performance with that of other classical dynamic models. In another, second, application, we model a rainfall process. This task is of particular theoretical and practical interest because of the specific structure and required untypical treatment of precipitation data.

## 1. INTRODUCTION

Modeling multi-dimensional time series is often an underestimated exercise of routine econometrical and statistical work. This slightly pejorative attitude towards day to day statistical analysis is unjustified since actually the calibration of time series models in several dimensions for standard data sizes is not only difficult on the numerical side but also on the mathematical side. Computationally speaking, integrated models for multi-dimensional time series become more involved when the parameter space is too large. Consequently the mathematical and econometrical aspects become more difficult since the parameter space becomes too complex, especially when their time variation is allowed. An example is the multivariate GARCH(1,1) BEKK model, which for even two dimensions

Our special thanks go to Oliver Linton, Peter Phillips, Cheng-Der Fuh and referees for helpful comments. We remain responsible for errors and omission.

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 “Ökonomisches Risiko”, Humboldt-Universität zu Berlin and IRTG 1972 “High Dimensional Non Stationary Time Series” is gratefully acknowledged. Address correspondence to Weining Wang, Hermann-Otto-Hirschfeld Junior Professor in Nonparametric Statistics and Dynamic Risk Management at the Ladislaus von Bortkiewicz Chair of Statistics of Humboldt-Universität zu Berlin, Spandauer Straße 1, 10178 Berlin, Germany; e-mail: wangwein@cms.hu-berlin.de.

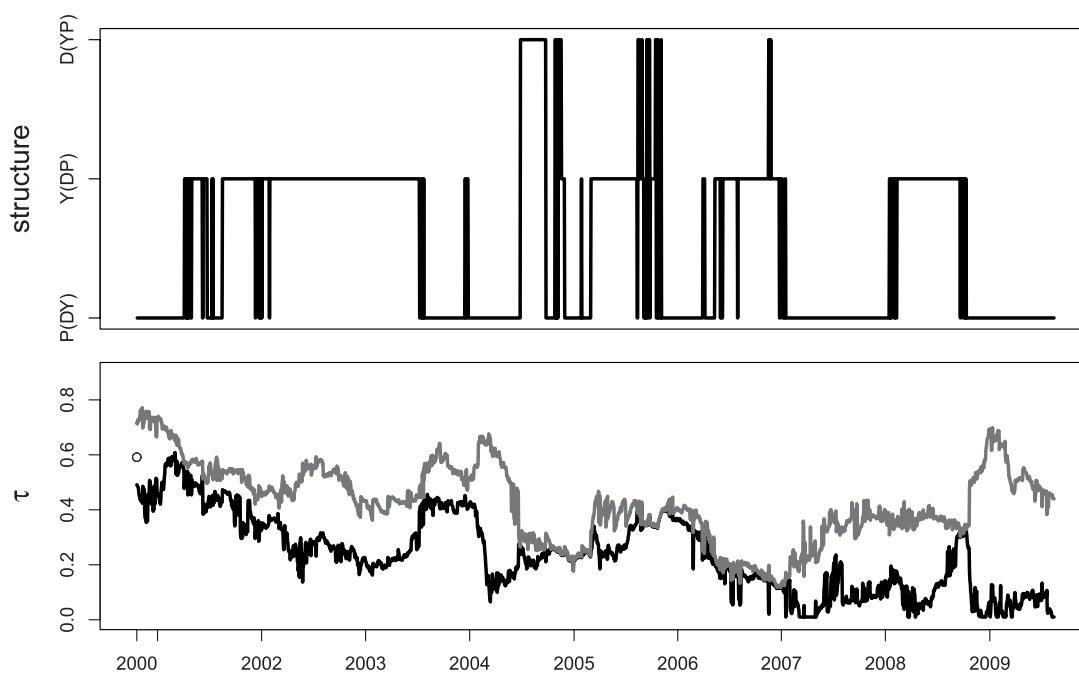
has an associated parameter space of dimension 12. For moderate sample sizes, the parameter space dimension might be in the range of the sample size or even bigger. This data situation has evoked a new strand of the literature on dimension reduction via penalty methods.

In this paper we take a different route, by calibrating an integrated dynamic model with unknown dependency structure among the  $d$ -dimensional time series variables. More precisely, the unknown dependency structure may vary within a set of given dependencies. These dependency structures might have been selected via a preliminary study, as described in, e.g., Härdle, Herwartz, and Spokoiny (2003). The specific dependence at each time  $t$  is unknown to the data analyst, and depends on the dependency pattern at time  $t - 1$ . Therefore, hidden Markov models (HMM) naturally come into play. This leaves us with the task of specifying the set of dependencies.

An approach based on assuming a multivariate Gaussian or mixed normals is inappropriate in the presence of important types of data features such as heavy tails, asymmetry, and nonlinear dependencies. Such a simplification is certainly in practical questions concerning too restrictive tails and might lead to biased results. The use of copulae is one possible approach to solving these problems. Moreover, copulae allow us to separate the marginal distributions and the dependency model, see Sklar (1959). In recent decades, copula-based models have gained popularity in various fields like finance, insurance, biology, hydrology, etc. Nevertheless, many basic multivariate copulae are still too restrictive and the extension to more parameters leads initially to a nonparametric density estimation problem that suffers of course from the curse of dimensionality. A natural compromise is the class of hierarchical Archimedean copulae (HAC). An HAC allows a rich copula structure with a finite number of parameters. Recent research has demonstrated their flexibility (see McNeil and Nešlehová, 2009; Okhrin, Okhrin, and Schmid, 2013; Whelan, 2004).

Insights into the dynamics of copulae have been offered by Chen and Fan (2005), who assume an underlying Markovian structure, and a specific class of copulae functions for the temporal dependence; Patton (2004) considers an asset-allocation problem with a time-varying parameter of bivariate copulae; and Rodriguez (2007) studies financial contagion using switching-parameter bivariate copulae. Similarly, Okimoto (2008) provides strong empirical evidence that a Markov switching multivariate normal model is not appropriate for the dependence structures in international equity markets.

Moreover, an adaptive method isolating a time varying dependency structure via a local change point method (LCP) has been proposed in Giacomini, Härdle, and Spokoiny (2009) and Härdle, Okhrin, and Okhrin (2013). Figure 1 presents an analysis of HAC for exchange rate data using LCP on a moving window, where the window sizes are adaptively selected by the LCP algorithm. It plots the changes of estimated structure (upper panel) and parameters (lower panel) in each window over time. In particular, in the upper panel, the  $y$ -axis corresponds to the dependency structures picked by estimation of three-dimensional copulae; in the



**FIGURE 1.** LCP for exchange rates: structure (upper) and parameters (lower,  $\theta_1$  (gray) and  $\theta_2$  (black)) for Gumbel HAC.  $m_0 = 40$  (starting value for the window size in the algorithm).

lower panel, the  $y$ -axis shows the two estimated dependency parameters (value converted to Kendall's  $\tau$ ) corresponding to the estimated structure. In more detail, we have three exchange rates series: P (GBP/EUR), Y (JPY/EUR), D (USD/EUR); the label P(DY) means that the pair D and Y have a stronger dependency than other possible pairs. For a more detailed introduction to HAC and their structures, see Section 2.1. One observes that the structure very often remains the same for a long time, the parameters only varying slowly over time. This indicates that the dynamics of HAC functions is likely to be driven by a Markovian sequence seemingly determining the structures and parameter values. This observation motivates us to pursue a different path of modeling the dynamics. Instead of taking a local point of view, we adopt a global dynamic model HMM for the change of both the tree structure and the parameters of the HAC over time. In this situation, the not directly observable underlying Markov process  $X$  determines the state of distributions of  $Y$ .

HMM has been widely applied to speech recognition, see Rabiner (1989), molecular biology, and digital communications over unknown channels. Markov switching models were introduced to the economics literature by Hamilton (1989), where the trend component of a univariate nonstationary time series changes according to an underlying Markov chain. Later, it was extended and combined with many different time series models, see, e.g., Pelletier (2006). For estimation and inference issues in HMM, see Bickel, Ritov, and Rydén (1998) and Fuh (2003), among others.

In this paper, we propose a new type of dynamic model, called HMM HAC, which incorporates HAC into an HMM framework. The theoretical problems,

such as parameter consistency and structure consistency, are solved. The expectation maximization (EM) algorithm is developed in this framework for parameter estimation. See Section 2 for a description of the model, and Section 3 for theorems about its consistency and asymptotic normality. Issues as to the EM algorithm and computation are in Section 4. Section 5 treats a simulation study, and Section 6 is the applications. The technical details are put into the Appendix.

## 2. MODEL DESCRIPTION

In this section, we introduce our model and estimation method. Section 2.1 briefly introduces the definition and properties of HAC, and Section 2.2 introduces the HMM HAC. In the last subsection, we describe the estimation and algorithm we use.

### 2.1. Copulae

Let  $Z_1, \dots, Z_d$  be r.v. with continuous cumulative distribution function (cdf)  $F(\cdot)$ . The Sklar theorem guarantees the existence and uniqueness of copula functions:

**THEOREM 2.1 (Sklar's theorem).** *Let  $F$  be a multivariate distribution function with margins  $F_1^m, \dots, F_d^m$ , then a copula  $C$  exists such that*

$$F(z_1, \dots, z_d) = C\{F_1^m(z_1), \dots, F_d^m(z_d)\}, \quad z_1, \dots, z_d \in \mathbb{R}.$$

*If  $F_i^m(\cdot)$  are continuous for  $i = 1, \dots, d$  then  $C(\cdot)$  is unique. Otherwise  $C(\cdot)$  is uniquely determined on  $F_1^m(\mathbb{R}) \times \dots \times F_d^m(\mathbb{R})$ .*

*Conversely, if  $C(\cdot)$  is a copula and  $F_1^m, \dots, F_d^m$  are univariate distribution functions, then the function  $F$  defined above is a multivariate distribution function with margins  $F_1^m, \dots, F_d^m$ .*

The family of Archimedean copulae is very flexible: it captures tail dependency, has an explicit form, and is simple to estimate,

$$C(u_1, \dots, u_d) = \phi\{\phi^{-1}(u_1) + \dots + \phi^{-1}(u_d)\}, \quad u_1, \dots, u_d \in [0, 1], \tag{1}$$

where  $\phi(\cdot)$  is defined as the generator of the copula and depends on a parameter  $\theta$ , see Nelsen (2006).  $\phi(\cdot)$  is  $d$  monotone, and  $\phi(\cdot) \in \mathcal{L} = \{\phi(\cdot) : [0; \infty) \rightarrow (0, 1] \mid \phi(0) = 1, \phi(\infty) = 0; (-1)^j \phi^{(j)} \geq 0; j = 1, \dots, d - 2\}$ . As an example, the Gumbel generator is given by  $\phi(x) = \exp(-x^{1/\theta})$  for  $0 \leq x < \infty, 1 \leq \theta < \infty$ .

In the present paper we consider less restrictive compositions of simple Archimedean copulae leading to a Hierarchical Archimedean Copula (HAC)  $C(u_1, \dots, u_d; \boldsymbol{\theta}, s)$ , where  $s = \{(\dots(i_1 \dots i_{j_1}) \dots (\dots) \dots)\}$  denotes the structure of HAC, with  $i_\ell \in \{1, \dots, d\}$  being a reordering of the indices of the variables and  $s_j$  the structure of the subcopulae with  $s_d = s$ , and  $\boldsymbol{\theta}$  is the set of copula parameters.



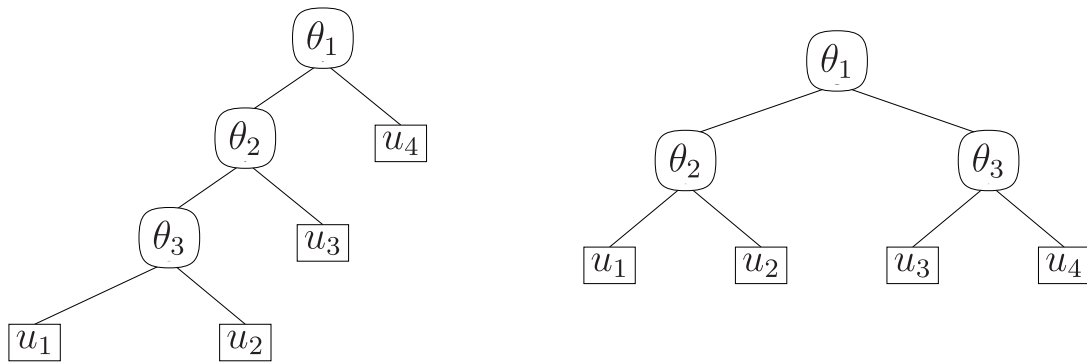


FIGURE 2. Fully and partially nested copulae of dimension  $d = 4$  with structures  $s = (((12)3)4)$  on the left and  $s = ((12)(34))$  on the right.

For example, the fully nested HAC (see Figure 2, left) can be expressed by

$$\begin{aligned}
 C(u_1, \dots, u_d; \boldsymbol{\theta}, s = s_d) &= C\{u_1, \dots, u_d; (\theta_1, \dots, \theta_{d-1})^\top, ((s_{d-1})d)\} \\
 &= \phi_{d-1, \theta_{d-1}}^{-1} \left( \phi_{d-1, \theta_{d-1}}^{-1} \circ C \left\{ u_1, \dots, u_{d-1}; (\theta_1, \dots, \theta_{d-2})^\top, ((s_{d-2})(d-1)) \right\} \right. \\
 &\quad \left. + \phi_{d-1, \theta_{d-1}}^{-1}(u_d) \right),
 \end{aligned}$$

where  $s = \{(\dots(12)3)\dots d\}$ . On the RHS of Figure 2 we have the partially nested HAC with  $s = ((12)(34))$  in dimension  $d = 4$ .

For more details about HAC, see Joe (1997), Whelan (2004), Savu and Trede (2010), and Okhrin, Okhrin, and Schmid (2013).

It is worth noting that not all generator functions can be mixed within one HAC. We therefore concentrate on one single generator family within one HAC. This boils down to binary structures, i.e., at each level of the hierarchy only two variables are joined together. In fact, this makes the architecture very flexible and yet parsimonious.

Note that not only are the parameters unknown for each HMM HAC, but also the structure has to be determined. We adopt the modified computational steps of Okhrin et al. (2013) to estimate the HAC structure and parameters. One estimates the marginal distributions either parametrically or nonparametrically. Then (assuming that the marginal distributions are known) one selects the couple of variables with the strongest fit and denotes the corresponding estimator of the parameter at the first level by  $\hat{\theta}_1$  and the set of indices of the variables by  $I_1$ . The selected couple is joined together to define the pseudo-variables  $z_1 = C\{(I_1); \hat{\theta}_1, \phi_1\}$ . Next, one proceeds in the same way by considering the remaining variables and the new pseudovariable. At every level, the copula parameter is estimated by assuming that the margins as well as the copula parameters at lower levels are known. This algorithm allows us to determine the estimated structure of the copula recursively.

### 2.2. Incorporating HAC into HMM

A hidden Markov model is a parameterized time series model with an underlying Markov chain viewed as missing data, as in Leroux (1992), Bickel et al. (1998), and Gao and Song (2011). More specifically, in the HMM HAC framework, let  $\{X_t, t \geq 0\}$  be a stationary Markov chain of order one on a finite state space  $D = \{1, 2, \dots, M\}$ , with transition probability matrix  $P = \{p_{ij}\}_{i,j=1,\dots,M}$  and initial distribution  $\pi = \{\pi_i\}_{i=1,\dots,M}$ .

$$P(X_0 = i) = \pi_i, \tag{2}$$

$$\begin{aligned} P(X_t = j | X_{t-1} = i) &= p_{ij} \tag{3} \\ &= P(X_t = j | X_{t-1} = i, X_{t-2} = x_{t-2}, \dots, X_1 = x_1, X_0 = x_0), \\ &\quad i, j = 1, \dots, M \end{aligned}$$

Let  $\{Y_t, t \geq 0\}$  be the associated observations, and they are adjoined with  $\{X_t, t \geq 0\}$  in such a way that given  $X_t = i, i = 1, \dots, M$ , the distribution of  $Y_t$  is fixed:

$$(X_t | X_{0:(t-1)}, Y_{0:(t-1)}) \stackrel{\mathcal{L}}{=} (X_t | X_{t-1}), \tag{4}$$

$$(Y_t | Y_{0:(t-1)}, X_{(0:t)}) \stackrel{\mathcal{L}}{=} (Y_t | X_t), \tag{5}$$

where  $Y_{0:(t-1)}$  stands for  $\{Y_0, \dots, Y_{t-1}\}, t < T$ .

Let  $f_j\{\cdot\}$  be the conditional density of  $Y_t$  given  $X_t = j$  with  $\theta \in \Theta, s \in S, j = 1, \dots, M$  being the unknown parameters. Here,  $\{X_t, t \geq 0\}$  is the Markov chain, and given  $X_0, X_1, \dots, X_T$ , the  $Y_0, Y_1, \dots, Y_T$  are independent. Note that  $\theta = (\theta^{(1)}, \dots, \theta^{(M)}) \in \mathbb{R}^{(d-1)M}$  are the unknown dependency parameters,  $s = (s^{(1)}, \dots, s^{(M)})$  are the unknown HAC structures. Denote their true values by  $\theta^*$  and  $s^*$  respectively.

For the time series  $y_1, \dots, y_T \in \mathbb{R}^d$  ( $y_t = (y_{1t}, y_{2t}, y_{3t}, \dots, y_{dt})^\top$ ) and the unobservable (or missing)  $x_1, \dots, x_T$  from the given hidden Markov model, define  $\pi_{x_0}$  as the  $\pi_i$  for  $x_0 = i, i = 1, \dots, M$ , and  $p_{x_{t-1}x_t} = p_{ji}$  for  $x_{t-1} = j$  and  $x_t = i$ . The full likelihood for  $\{x_t, y_t\}_{t=1}^T$  is then:

$$p_T(y_{0:T}; x_{0:T}) = \pi_{x_0} f_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t), \tag{6}$$

and the likelihood for the observations  $\{y_t\}_{t=1}^T$ , only is calculated by marginalization:

$$p_T(y_{0:T}) = \sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} f_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t). \tag{7}$$

The HAC is a parameterization of  $f_{x_t}(y_t)(x_t = i)$ , which helps properly understand the dynamics of a multivariate distribution. Up to now, typical parameterizations have been mixtures of log-concave or elliptical symmetric densities,

such as those from Gamma or Poisson families, which are not flexible enough to model multi-dimensional time series. The advantage of the copula is that it splits the multivariate distribution into its margins and a pure dependency component. In other words, it captures the dependency between variables, eliminating the impact of the marginal distributions as introduced in the previous subsection.

Furthermore, we incorporate this procedure within an HMM framework. We denote the underlying Markov variable  $X_t$  as a dependency type variable. If  $x_t = i$ , the parameters  $(\boldsymbol{\theta}^{(i)}, s^{(i)})$  determined by state  $i = 1, \dots, M$  take values on  $\Theta \times S$ , where  $S$  is a set of discrete candidate states corresponding to different dependency structures of the HAC, and  $\Theta$  is a compact subset of  $\mathbb{R}^{d-1}$  in which the HAC parameters take their values. Therefore,

$$f_i(\cdot) = c \left\{ F_1^m(y_1), F_2^m(y_2), \dots, F_d^m(y_d), \boldsymbol{\theta}^{(i)}, s^{(i)} \right\} f_1^m(y_1) f_2^m(y_2) \cdots f_d^m(y_d), \tag{8}$$

with  $f_i^m(y_i)$  being the marginal densities,  $F_i^m(y_i)$  the marginal cdf and  $c(\cdot)$  the copula density, which is defined in assumption A.4 in Section 3.

Let  $\boldsymbol{\theta}^{(i)} = (\theta_{i1}, \dots, \theta_{i,d-1})^\top$  be the dependency parameters of the copulae starting from the lowest up to the highest level connected with a fixed state  $x_t = i$  and corresponding density  $f_i(\cdot)$ . Refining the algorithm of Okhrin et al. (2013), the multistage maximum likelihood estimator  $(\hat{\boldsymbol{\theta}}^{(i)}, \hat{s}^{(i)})$  solves the system

$$\left( \frac{\partial \mathcal{L}_1}{\partial \theta_{i1}}, \dots, \frac{\partial \mathcal{L}_{d-1}}{\partial \theta_{i,d-1}} \right)^\top = \mathbf{0}, \tag{9}$$

where

$$\begin{aligned} \mathcal{L}_j &= \sum_{t=1}^T w_{it} l_{ij}(Y_t), \quad \text{for } j = 1, \dots, d-1, \\ l_{ij}(Y_t) &= \log \left( c \left[ \{ \hat{F}_m^m(y_{tm}) \}_{m \in \{1, \dots, j\}}; \{ \theta_{i\ell} \}_{\ell=1, \dots, j-1}, s_m^{(i)} \right] \prod_{m \in \{1, \dots, j\}} \hat{f}_m^m(y_{tm}) \right) \\ &\text{for } t = 1, \dots, T. \end{aligned}$$

where  $j$  denotes the layers of the tree structure, and  $\hat{F}_m^m(\cdot)$  is an estimator (either nonparametric with  $\hat{F}_m^m(x) = (T+1)^{-1} \sum_{t=1}^T \mathbf{1}(Y_{tm} \leq x)$  or parametric  $\hat{F}_m^m(x) = F_m^m(x, \hat{\boldsymbol{\alpha}}_m)$ ) of the marginal cdf  $F_m^m(\cdot)$ , where  $\hat{\boldsymbol{\alpha}}_m$  stand for estimated parameters of a marginal distribution. Note that a nonparametric estimation of the margins would lead to our estimation's having a semiparametric nature. The marginal densities  $\hat{f}_m^m(\cdot)$  are estimated parametrically or nonparametrically (kernel density estimation) corresponding to the estimation of the marginal distribution functions, and  $w_{it}$  is the weight associated with state  $i$  and time  $t$ , see (14). Chen and Fan (2006) and Okhrin et al. (2013) provide the asymptotic behavior of the estimates.

### 2.3. Likelihood estimation

For the estimation of the HMM HAC model, we adopt the EM algorithm, see Dempster, Laird, and Rubin (1977), also known as the Baum–Welch algorithm in the context of HMM. Recall the full likelihood  $p_T(y_{0:T}; x_{0:T})$  in (6) and the partial likelihood  $p_T(y_{0:T})$  in (7), and the log likelihood:

$$\log\{p_T(y_{0:T})\} = \log \left\{ \sum_{x_0=1}^M \cdots \sum_{x_n=1}^M \pi_{x_0} f_{x_0}(y_0; \boldsymbol{\theta}^{(x_0)}) \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t; \boldsymbol{\theta}^{(x_t)}, s^{(x_t)}) \right\}. \quad (10)$$

The EM algorithm suggests estimating a sequence of parameters  $\mathbf{g}_{(r)} \stackrel{\text{def}}{=} (P_{(r)}, \boldsymbol{\theta}_{(r)}, \mathbf{s}_{(r)})$  (for the  $r$ th iteration) by iterative maximization of  $\mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)})$  with

$$\mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)}) \stackrel{\text{def}}{=} E_{\mathbf{g}_{(r)}} \{\log p_T(Y_{0:T}; X_{0:T}) | Y_{0:T} = y_{0:T}\}.$$

That is, one carries out the following two steps:

- (a) E-step: compute  $\mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)})$ ,
- (b) M-step: choose the update parameters  $\mathbf{g}_{(r+1)} = \arg \max_{\mathbf{g}} \mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)})$ .

The essence of the EM algorithm is that  $\mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)})$  can be used as a substitute for  $\log p_T(y_{0:T}; x_{0:T}; \theta)$ , see Cappé, Moulines, and Rydén (2005).

In our setting, we may write  $\mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)})$  as:

$$\begin{aligned} \mathcal{Q}(\mathbf{g}; \mathbf{g}_{(r)}) &= \sum_{i=1}^M E_{\mathbf{g}_{(r)}} [\mathbf{1}\{X_0 = i\} \log\{\pi_i f_i(y_0)\} | Y_{0:T} = y_{0:T}] \quad (11) \\ &+ \sum_{t=1}^T \sum_{i=1}^M E_{\mathbf{g}_{(r)}} [\mathbf{1}\{X_t = i\} \log f_i(y_t) | Y_{0:T} = y_{0:T}] \\ &+ \sum_{t=1}^T \sum_{i=1}^M \sum_{j=1}^M E_{\mathbf{g}_{(r)}} [\mathbf{1}\{X_t = j\} \mathbf{1}\{X_{t-1} = i\} \log\{p_{ij}\} | Y_{0:T} = y_{0:T}] \\ &= \sum_{i=1}^M P_{\mathbf{g}_{(r)}}(X_0 = i | Y_{0:T} = y_{0:T}) \log\{\pi_i f_i(y_0)\} \\ &+ \sum_{t=1}^T \sum_{i=1}^M P_{\mathbf{g}_{(r)}}(X_t = i | Y_{0:T} = y_{0:T}) \log f_i(y_t) \\ &+ \sum_{t=1}^T \sum_{i=1}^M \sum_{j=1}^M P_{\mathbf{g}_{(r)}}(X_{t-1} = i, X_t = j | Y_{0:T} = y_{0:T}) \log\{p_{ij}\}, \quad (12) \end{aligned}$$

where  $f_i(\cdot)$  is as in (8). The E-step, in which  $P_{\mathbf{g}_{(r)}}(X_t = i | Y_{0:T})$ ,  $P_{\mathbf{g}_{(r)}}(X_{t-1} = i, X_t = j | Y_{0:T})$  are evaluated, is carried out by the forward–backward algorithm

and the M-step is explicit in the  $p_{ij}$ s and the  $\pi_i$ s. Adding constraints to (12) yields

$$\mathcal{L}(\mathbf{g}, \lambda; \mathbf{g}') = \mathcal{Q}(\mathbf{g}; \mathbf{g}') + \sum_{i=1}^M \lambda_i \left( 1 - \sum_{j=1}^M p_{ij} \right). \tag{13}$$

For the M-step, we need to take the first order partial derivatives, and plug into (13). So the dependency parameters  $\boldsymbol{\theta}$  and the structure parameters  $\mathbf{s}$  need to be estimated iteratively, for  $\boldsymbol{\theta}^{(i)}$  ( $\boldsymbol{\theta}^{(i)} = \{\theta_{i1}, \dots, \theta_{i(d-1)}\}$ ):

$$\frac{\partial \mathcal{L}(\mathbf{g}, \lambda; \mathbf{g}')}{\partial \theta_{ij}} = \sum_{t=1}^T P_{\mathbf{g}'}(X_t = i | Y_{0:T}) \partial \log f_i(y_t) / \partial \theta_{ij}. \tag{14}$$

To simplify the procedure, we adopt the HAC estimation method (9) with weights  $w_{it} \stackrel{\text{def}}{=} P_{\mathbf{g}'}(X_t = i | Y_{0:T})$ . We also fix  $\pi_i, i = 1, \dots, M$ , as this influences only the first observation  $x_0$  which may be considered also as given and fixed. Maximizing (12) w.r.t.  $\pi_i$  would return the first derivative with one observation  $y_0$ . Also as the previous information for the distribution of  $x_0$  is not available, our EM algorithm would not involve updating  $\pi_i$ . The estimation of the transition probabilities  $p_{ij}$  follows:

$$\frac{\partial \mathcal{L}(\mathbf{g}, \lambda; \mathbf{g}')}{\partial p_{ij}} = \sum_{t=1}^T \frac{P_{\mathbf{g}'}(X_{t-1} = i, X_t = j | Y_{0:T})}{p_{ij}} - \lambda_i, \tag{15}$$

$$\frac{\partial \mathcal{L}(\mathbf{g}, \lambda; \mathbf{g}')}{\partial \lambda_i} = 1 - \sum_{j=1}^M p_{ij}. \tag{16}$$

Equating (15) and (16) yields

$$\hat{p}_{ij} = \frac{\sum_{t=1}^T P_{\mathbf{g}'}(X_{t-1} = i, X_t = j | Y_{0:T})}{\sum_{t=1}^T \sum_{l=1}^M P_{\mathbf{g}'}(X_{t-1} = i, X_t = l | Y_{0:T})}. \tag{17}$$

### 3. THEORETICAL RESULTS

In this section, we discuss the conditions needed to derive the consistency and the asymptotic properties of our estimates. We then state our main theoretical theorems. Throughout the theory we concentrate on the most interesting case: a semi-parametric estimation with nonparametric margins.

#### Assumptions.

A.1  $\{X_t\}$  is a stationary, strictly irreducible, and aperiodic Markov process of order one with final discrete state, and  $\{Y_t\}_{t=1}^T$  are conditionally independent given  $\{X_t\}_{t=1}^T$  and generated from an HAC HMM model with parameters  $\{s^{*(i)}, \theta^{*(i)}, \pi^*, \{p_{ij}^*\}_{i,j}\}, i, j = 1, \dots, d$ .

It is worth noting that A.1 is the basic assumption on the evolution of a hidden Markov chain. One key property is that given one realization of the path of  $\{X_t\}$ , the conditional distributions of  $\{Y_t\}_{t=1}^T$  are totally fixed. But  $\{Y_t\}$  will be dependent and will even have a finite mixture distribution from the given parametric family. The evolution of  $\{X_t\}$  will later be linked to the dependency parameters of the state space distribution of  $\{Y_t\}$ .

A.2 The family of mixtures of at most  $M$  elements  $\{f(y; \boldsymbol{\theta}^{(i)}, s^{(i)}) : \boldsymbol{\theta}^{(i)} \in \Theta^{(i)}, s^{(i)} \in S\}$  is identifiable w.r.t. the parameters and structures:

$$\sum_{i=1}^M \alpha_i f(y; \boldsymbol{\theta}^{(i)}, s^{(i)}) = \sum_{i=1}^M \alpha'_i f(y; \boldsymbol{\theta}'^{(i)}, s'^{(i)}) \quad a.e. \tag{18}$$

$$\text{then, } \sum_{i=1}^M \alpha_j \delta_{\boldsymbol{\theta}^{(i)}, s^{(i)}} = \sum_{i=1}^M \alpha'_i \delta_{\boldsymbol{\theta}'^{(i)}, s'^{(i)}}, \tag{19}$$

defining  $\delta_{\boldsymbol{\theta}^{(i)}, s^{(i)}}$  as the distribution function for a point mass in  $\Theta$  associated with the structure  $s^{(i)}$ , noting that  $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}'^{(i)}$  is only meaningful when  $s^{(i)} = s'^{(i)}$ .

The property of identifiability is nothing else than the construction of a finite mixture model, see McLachlan and Peel (2000). As a copula is a special form of a multivariate distribution, similar techniques may be applied to get identifiability also in the case of copulae. The family of copula mixtures has been thoroughly investigated in Caia, Chen, Fan, and Wang (2006) while developing estimation techniques. In that general case, one should be careful, as the general copula class is very wide and its mixture identification may cause some problems because of the different forms of the densities. The very construction of the HAC narrows this class. Imposing the same generator functions on all levels of the HAC, we restrict the family to the vector of parameters and the tree structure, see also Okhrin et al. (2013). Moreover, we restrict the classes to only binary trees with distinct parameters to avoid identifiability issues induced by the case of the same parameter values on each layer of a tree. Our preliminary numerical analysis shows that the HAC fulfills the identifiability property for all the structures and parameters used in this study.

A.3 The true marginal distribution  $f_m^m(\cdot) \in C^2$ , and the derivatives up to a second order are bounded for all  $m = 1, \dots, d$ . Also  $\sqrt{f^m}$  is absolute continuous. In the case of a nonparametric estimation for  $f_i^m(\cdot) \in C^2$ , one needs also to ensure that the kernel function  $K(\cdot) \in C^2$  subject to  $\int_B K(u)du = 1$ , has support on a compact set  $B$ , is symmetric, and has integrable first derivative.

We would like to focus on the dependency parameter, therefore in the following setting, we simply assume that the marginal processes  $y_{t1}, y_{t2}, \dots, y_{td}$  are identically distributed.

A.4  $E\{|\log f_i(y)|\} < \infty$ , for  $i = 1, \dots, M, \forall s^{(i)} \in S$ . Define the copulae density function  $c(u_1, u_2, \dots, u_d, \boldsymbol{\theta}^{(i)}, s^{(i)}) \stackrel{\text{def}}{=} \partial^d C(u_1, u_2, \dots, u_d, \boldsymbol{\theta}^{(i)}, s^{(i)}) / \partial u_1 \partial u_2 \cdots \partial u_d$ , then  $\log c(u_1, u_2, \dots, u_d, \boldsymbol{\theta}^{(i)}, s^{(i)})$  as well as its first and second

partial derivatives w.r.t.  $u_i$ s and  $\theta^{(i)}$  are well defined for  $((0, 1)^d \times \Theta^{(i)})$ . Also, their suprema in a compact set  $((E^d) \times \Theta^{(i)})$  ( $E^d \in [0, 1]^d$ ) has finite moments up to the order four.

A.5 For every  $\theta^{(i)} \in \Theta$ , and any particular structure  $s \in S$ ,

$$E \left[ \sup_{\|\theta^{(i)} - \theta^{(i)}\| < \delta} \{f_i(Y_1, \theta^{(i)}, s)\}^+ \right] < \infty,$$

for some  $\delta > 0$ .

A.6 The true point  $\theta^*$  is an interior point of  $\Theta$ .

A.7 There exists a constant  $\delta^0$ , such that  $P(\sup_{\|\theta^{(i)} - \theta^{(i)}\| < \delta^0} \max_{i,j} E \frac{\{f_i(Y_1, \theta', s)\}}{\{f_j(Y_1, \theta', s)\}} = \infty | X_1 = i) < 1$ .

Denote by  $p_T(y_{0:T}; v, \omega)$  the density in (7) with parameters  $\{v, \omega\} \in \{V, \Omega\}$  as described in the Appendix 7.2. Define  $\hat{\theta}^{(i)}, \hat{s}^{(i)}$  as  $\hat{\theta}^{(i)}(\hat{v}, \hat{\omega})$ , and  $\hat{s}^{(i)}(\hat{v}, \hat{\omega})$  with  $(\hat{v}, \hat{\omega})$  being the point where  $p_T(y_{0:T}; v, \omega)$  achieves its maximum value over the parameter space  $\{V, \Omega\}$ .

It is known that HMM is not itself identifiable, as a permutation of states would yield the same value for  $p_T(y_{0:T}; v, \omega)$ . We assume therefore  $\theta^{*(j)}$ s and  $s^{*(j)}$ s to be distinct in the sense that for any  $s^{*(i)} = s^{*(j)}, i \neq j$  we have  $\theta^{*(i)} \neq \theta^{*(j)}$ .

**THEOREM 3.1.** *Under A.1–A.7, we find the corresponding structure:*

$$\lim_{T \rightarrow \infty} \min_{i \in 1, \dots, M} P(\hat{s}^{(i)} = s^{*(i)}) = 1. \tag{20}$$

Moreover,

**THEOREM 3.2.** *Assume that A.1–A.7 hold then the parameter  $\hat{\theta}^{(i)}$  satisfies,  $\forall \varepsilon > 0$ :*

$$\lim_{T \rightarrow \infty} \max_{i \in 1, \dots, M} P(|\hat{\theta}^{(i)} - \theta^{*(i)}| > \varepsilon | \hat{s}^{(i)} = s^{*(i)}) = 0. \tag{21}$$

In addition, we can also establish asymptotic normality results for parameters.

**THEOREM 3.3.** *Assume that A.1–A.7 hold, and given that  $s^{*(i)}$  is correctly estimated, which is an event with probability tending to 1, we have*

$$\sqrt{T} \{ \hat{\theta} - \theta \} \rightarrow N(0, \Sigma^*), \tag{22}$$

where  $\Sigma^*$  is the asymptotic covariance function, defined as  $\Sigma^* \stackrel{\text{def}}{=} B^{-1} \text{Var}(\sqrt{T} A) B^{-1}$ , where  $B, A$  are defined in the Appendix in (A.19).

The proofs are presented in the Appendix.

#### 4. SIMULATION

The estimation performance of HMM HAC is evaluated in this section: subsection I aims to investigate whether the performance of the estimation is affected

by 1) adopting a nonparametric or parametric margins; 2) introducing a GARCH dependency in the marginal time series. Subsection II presents results for a five-dimensional time series model. In subsection III we compare the DCC method and our HMM HAC method. All the simulations show that our algorithm converges after a few iterations with moderate estimation errors, and the results are robust with respect to different estimation methods for the margins. Moreover our method dominates the DCC one.

Regarding the selection of the orders, in both the simulations and the applications, we have started with a model with three states, which is suggested by the initial moving window analysis described later. In the applications, the number of states will even be degenerated to two or one for some windows. This suggests that three states are sufficient for model estimations. However, one can consider general BIC or AIC criteria for selecting the number of states.

### 4.1. Simulation I

In this subsection, a three-dimensional generating process has fixed marginal distributions:  $Y_{t1}, Y_{t2}, Y_{t3} \sim N(0, 1)$ . To study the effect of deGARCH step in our application (DeGARCH is meant by prefitting marginal time series with a GARCH model, and take the residuals for estimation in later steps.), we simulated also according to a GARCH(1,1) model,

$$Y_{tj} = \mu_{tj} + \sigma_{tj}\varepsilon_{tj} \text{ with } \sigma_{tj}^2 = \omega_j + \alpha_j\sigma_{t-1j}^2 + \beta_j(Y_{t-1j} - \mu_{t-1j})^2, \tag{23}$$

with parameters  $\omega_j = 10^{-6}, \alpha_j = 0.8, \beta_j = 0.1$ , with standard normal residuals  $\varepsilon_{t1}, \varepsilon_{t2}, \varepsilon_{t3} \sim N(0, 1)$ . The dependence structure is modeled through HAC with Gumbel generators. Let us consider now a Monte Carlo setup where the setting employs realistic models. The three states with  $M = 3$  are as follows:

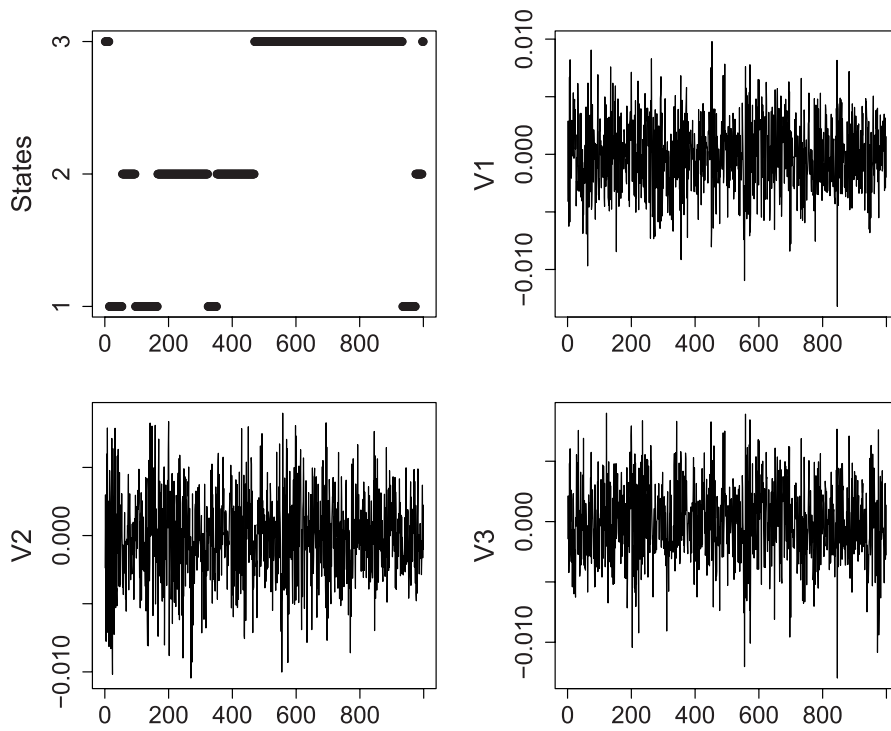
$$\begin{aligned} C\{u_1, C(u_2, u_3; \theta_1^{(1)} = 1.3); \theta_2^{(1)} = 1.05\} & \text{ for } i = 1, \\ C\{u_2, C(u_3, u_1; \theta_1^{(2)} = 2.0); \theta_2^{(2)} = 1.35\} & \text{ for } i = 2, \\ C\{u_3, C(u_1, u_2; \theta_1^{(3)} = 4.5); \theta_2^{(3)} = 2.85\} & \text{ for } i = 3, \end{aligned}$$

where the dependency parameters correspond to Kendall's  $\tau$ s ranging between 0.05 and 0.78, which is typical for financial data. The transition matrix is chosen as:

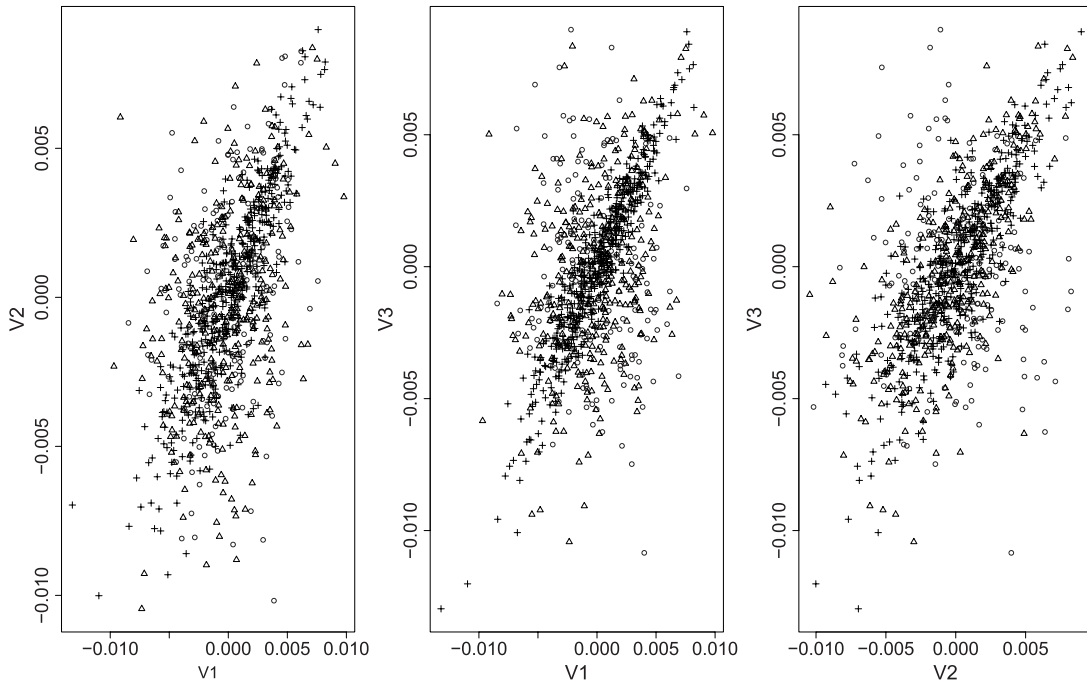
$$P = \begin{pmatrix} 0.982 & 0.010 & 0.008 \\ 0.008 & 0.984 & 0.008 \\ 0.003 & 0.002 & 0.995 \end{pmatrix},$$

with initial probabilities as  $\pi = (0.2, 0.1, 0.7)$  and sample size  $T = 2000$ . Figure 3 presents the underlying states and a marginal plot of the generated three-dimensional time series. No state switching patterns are evident from the marginal plots. Figure 4, however, clearly displays the switching of dependency

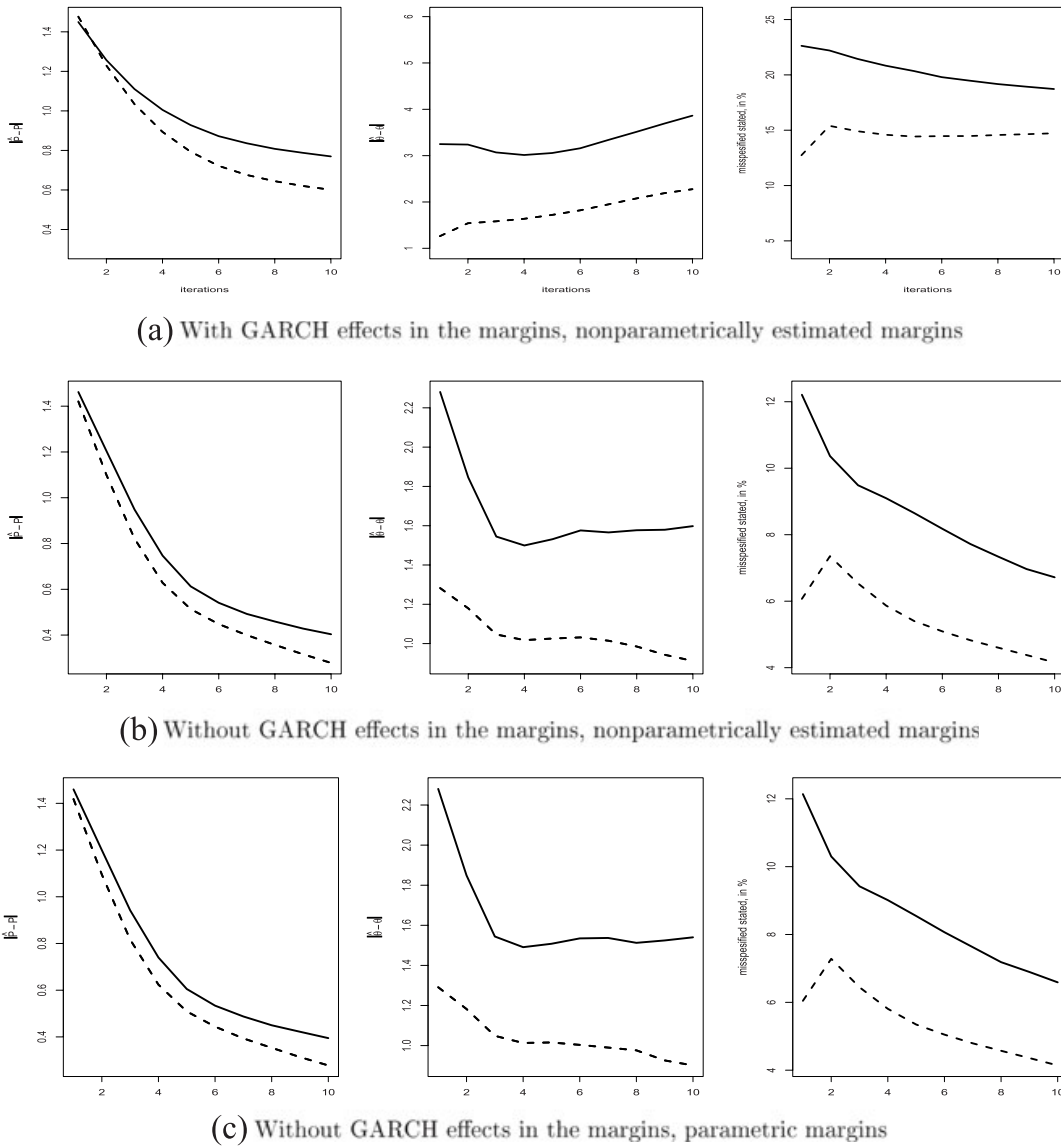




**FIGURE 3.** The underlying sequence  $x_t$  (upper left panel), marginal plots of  $(y_{t1}, y_{t2}, y_{t3})(t = 0, \dots, 1000)$ .



**FIGURE 4.** Snapshots of pairwise scatter plots of dependency structures ( $t = 0, \dots, 1000$ ), the  $(y_{t1})$  vs.  $(y_{t2})$  (left), the  $(y_{t1})$  vs.  $(y_{t3})$  (middle), and the  $(y_{t2})$  vs.  $(y_{t3})$ (right).



**FIGURE 5.** The averaged estimation errors for the transition matrix (left panel), parameters (middle panel), and convergence of states (right panel). Estimation starts from near true value (dashed); starts from values obtained by rolling window (solid). x-axis represents iterations. Number of repetitions is 1000.

patterns. The circles, triangles, and crosses correspond to the observations from states  $i = 1, 2, 3$ , respectively.

Generally, the iteration procedure stops after around ten steps. Figure 5 presents the deviations from their true values of the estimated states, the transition matrix, and the parameters for the first ten iterations of one sample. Since the starting values may influence the results, a moving window estimation is proposed to decide the initial parameters. The dashed black and solid black lines show, respectively, how the estimators behave with the initial values close to the true (dashed) and initial values obtained from the proposed rolling window algorithm

(solid). By “close to the true initial states”, we mean true structures with parameters all shifted up by 0.5 from the true ones. For “rolling window algorithm” we estimate HAC for overlapping windows of width 100, and then take the  $M$  most frequent structures with averaged parameters as initial states. The left panel of Figure 5 shows the ( $L_1$ ) difference ( $\sum_{i,j=1}^d |\hat{p}_{ij} - p_{ij}|$ ) of the true transition matrix from the estimated ones at each iteration, we see that for the three particular samples, the values all converge to around 0.4, which are moderately small; the middle panel is the sum of the estimated parameter errors of the four states with the correctly estimated states, we see that the accumulated errors are different depending on the different starting values; the right panel presents the percentage of wrongly estimated states, in all cases the percentage of wrongly estimated states is smaller than 8%. One can see that our choice of initial values can perform as well as the true ones through showing small differences, and our results from more iterations further confirm this.

Generally, the iteration procedure stops after around ten steps. Figure 5 presents the deviations from their true values of the estimated states, the transition matrix, and the parameters for the first ten iterations of one sample. Since the starting values may influence the results, a moving window estimation is proposed to decide the initial parameters. The dashed black and solid black lines show, respectively, how the estimators behave with the initial values close to the true (dashed) and initial values obtained from the proposed rolling window algorithm (solid). By “close to the true initial states”, we mean true structures with parameters all shifted up by 0.5 from the true ones. For “rolling window algorithm” we estimate HAC for overlapping windows of width 100, and then take the  $M$  most frequent structures with averaged parameters as initial states. The left panel of Figure 5 shows the ( $L_1$ ) difference ( $\sum_{i,j=1}^d |\hat{p}_{ij} - p_{ij}|$ ) of the true transition matrix from the estimated ones at each iteration, we see that for the three particular samples, the values all converge to around 0.4, which are moderately small; the middle panel is the sum of the estimated parameter errors of the four states with the correctly estimated states, we see that the accumulated errors are different depending on the different starting values; the right panel presents the percentage of wrongly estimated states, in all cases the percentage of wrongly estimated states is smaller than 8%. One can see that our choice of initial values can perform as well as the true ones through showing small differences, and our results from more iterations further confirm this.

Finally, we summarize our estimation results over 1000 repetitions. In Tables 1–2, we report the averaged estimation values with standard deviations (in brackets) and MSE (in brackets) for the estimated states, the transition matrix, and the parameters. Table 1 presents the results with the marginal time series being generated as just identically distributed data, while Table 2 presents the results with the marginal DGPs being GARCH(1,1). For the impact of estimating the copula model on estimated standardized residuals (after GARCH fitting, for example), we have also included a comparison of the estimation on the deGARCHed residuals (nonparametrically estimated margins).

**TABLE 1.** Simulation results for the marginal time series being generated as identically distributed data, sample size  $T = 2000$ , 1000 repetitions, standard deviations and MSEs are provided in brackets

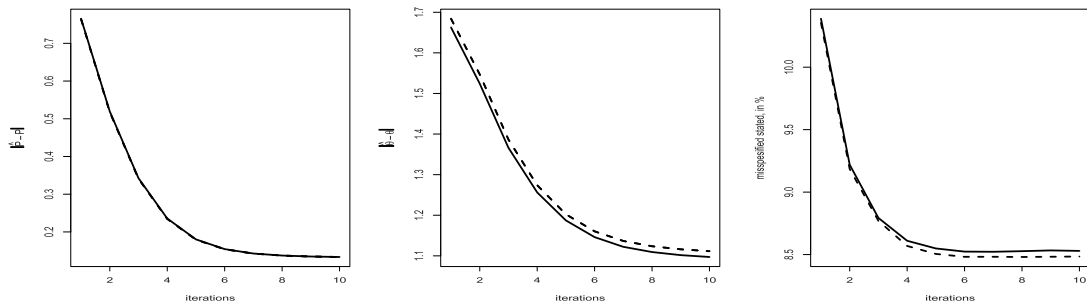
			True	Rol. Win.	True Str.	
Nonparametric Margins	C <sub>1</sub>	$\theta_1^{(1)}$	1.05	1.030 (0.046, 0.003)	1.057 (0.068, 0.005)	
		$\theta_2^{(1)}$	1.30	1.313 (0.156, 0.025)	1.308 (0.083, 0.007)	
	C <sub>2</sub>	$\theta_1^{(2)}$	1.35	1.366 (0.121, 0.015)	1.346 (0.182, 0.033)	
		$\theta_2^{(2)}$	2.00	2.556 (1.052, 1.416)	3.212 (1.991, 5.433)	
	C <sub>3</sub>	$\theta_1^{(3)}$	2.85	2.854 (0.073, 0.005)	2.854 (0.073, 0.005)	
		$\theta_2^{(3)}$	4.50	4.497 (0.133, 0.018)	4.496 (0.130, 0.017)	
	rat. of correct states				0.958 (0.029)	0.933 (0.056)
	$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $				0.278 (0.230)	0.404 (0.307)
	rat. of correct structures				0.949	0.918
Parametric Margins	C <sub>1</sub>	$\theta_1^{(1)}$	1.05	1.030 (0.041, 0.002)	1.056 (0.066, 0.004)	
		$\theta_2^{(1)}$	1.30	1.310 (0.154, 0.024)	1.306 (0.087, 0.008)	
	C <sub>2</sub>	$\theta_1^{(2)}$	1.35	1.365 (0.130, 0.017)	1.344 (0.173, 0.030)	
		$\theta_2^{(2)}$	2.00	2.544 (0.962, 1.221)	3.157 (1.906, 4.971)	
	C <sub>3</sub>	$\theta_1^{(3)}$	2.85	2.855 (0.074, 0.006)	2.854 (0.074, 0.005)	
		$\theta_2^{(3)}$	4.50	4.513 (0.133, 0.018)	4.513 (0.132, 0.018)	
	rat. of correct states				0.959 (0.029)	0.934 (0.056)
	$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $				0.278 (0.232)	0.395 (0.297)
	rat. of correct structures				0.955	0.921
deGARCHing	C <sub>1</sub>	$\theta_1^{(1)}$	1.05	1.030 (0.045, 0.002)	1.056 (0.067, 0.005)	
		$\theta_2^{(1)}$	1.30	1.320 (0.264, 0.070)	1.307 (0.081, 0.007)	
	C <sub>2</sub>	$\theta_1^{(2)}$	1.35	1.367 (0.123, 0.015)	1.345 (0.166, 0.028)	
		$\theta_2^{(2)}$	2.00	2.577 (1.273, 1.953)	3.180 (1.976, 5.297)	
	C <sub>3</sub>	$\theta_1^{(3)}$	2.85	2.852 (0.074, 0.005)	2.852 (0.074, 0.005)	
		$\theta_2^{(3)}$	4.50	4.489 (0.133, 0.018)	4.488 (0.130, 0.017)	
	rat. of correct states				0.958 (0.029)	0.933 (0.056)
	$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $				0.280 (0.234)	0.399 (0.299)
	rat. of correct structures				0.950	0.919

Also the estimation for different ways of deciding starting values are shown: “close to the true initial states” (True str), rolling window algorithm (Rol. Win.). Apparently, nonparametric or parametric estimation of the margins does not make big differences; this is also true for the prewhitening step. Regarding the precision

**TABLE 2.** Simulation results for the marginal DGPs (data generating processes) being GARCH(1,1), sample size  $T = 2000$ , 1000 repetitions, standard deviations and MSEs are provided in brackets

		True	Rol. Win.	True Str.	
Nonparametric Margins	$C_1$	$\theta_1^{(1)}$	1.05	1.100 (0.888, 0.791)	1.138 (0.080, 0.014)
		$\theta_2^{(1)}$	1.30	1.407 (0.888, 0.800)	1.246 (0.080, 0.009)
	$C_2$	$\theta_1^{(2)}$	1.35	1.403 (1.473, 2.173)	1.436 (2.608, 6.089)
		$\theta_2^{(2)}$	2.00	3.288 (1.473, 3.829)	5.106 (2.608, 16.449)
	$C_3$	$\theta_1^{(3)}$	2.85	2.772 (0.936, 0.882)	2.790 (0.941, 0.889)
		$\theta_2^{(3)}$	4.50	4.570 (0.936, 0.881)	4.606 (0.941, 0.897)
	rat. of correct states			0.853 (0.054)	0.813 (0.061)
	$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $			0.601 (0.217)	0.770 (0.242)
rat. of correct structures			0.853	0.757	
Parametric Margins	$C_1$	$\theta_1^{(1)}$	1.05	1.205 (1.261, 1.614)	1.107 (0.079, 0.009)
		$\theta_2^{(1)}$	1.30	1.843 (1.261, 1.885)	1.145 (0.079, 0.030)
	$C_2$	$\theta_1^{(2)}$	1.35	1.577 (1.381, 1.959)	1.838(1.612, 2.837)
		$\theta_2^{(2)}$	2.00	3.150 (1.381, 3.230)	3.480 (2.270, 7.343)
	$C_3$	$\theta_1^{(3)}$	2.85	3.879 (1.453, 3.170)	3.906 (1.523, 3.435)
		$\theta_2^{(3)}$	4.50	6.390 (1.453, 5.683)	6.592 (1.523, 6.696)
	rat. of correct states			0.732 (0.080)	0.747 (0.053)
	$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $			0.761 (0.179)	0.760 (0.156)
rat. of correct structures			0.358	0.323	
deGARCHing	$C_1$	$\theta_1^{(1)}$	1.05	1.030 (0.736, 0.542)	1.067 (0.141, 0.020)
		$\theta_2^{(1)}$	1.30	1.333 (0.736, 0.543)	1.305 (0.141, 0.020)
	$C_2$	$\theta_1^{(2)}$	1.35	1.356 (1.059, 1.122)	1.333 (1.755, 3.080)
		$\theta_2^{(2)}$	2.00	2.579 (1.059, 1.457)	3.351 (1.755, 4.905)
	$C_3$	$\theta_1^{(3)}$	2.85	2.835 (0.816, 0.666)	2.833 (0.816, 0.666)
		$\theta_2^{(3)}$	4.50	4.452 (0.816, 0.668)	4.451 (0.816, 0.668)
	rat. of correct states			0.958 (0.028)	0.925 (0.058)
	$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $			0.299 (0.235)	0.460 (0.325)
rat. of correct structures			0.938	0.916	

of the estimation, one sees that when the true GDP is GARCH(1,1), the prewhitening step is necessary to guarantee the quality of estimation. Also we see that for the parameter  $\theta_2^{(2)}$  the estimation errors are larger. The standard deviations of the design matrix are also relatively high. This is due to our selected design matrix



**FIGURE 6.** The averaged estimation errors for transition matrix (left panel), parameters (middle panel), convergence of states (right panel). Estimation starts from near true value (dashed); starts from values obtained by rolling window(solid). x-axis represents iterations. Number of repetitions is 1000.

having very small off-diagonal values, so for some realizations we have too few observations for state 2 to achieve accurate estimates. One sees in our simulation II nicer results with a different transition matrix.

### 4.2. Simulation II

In this subsection, we consider a five-dimensional model. The marginal distributions are taken as:  $Y_{t1}, Y_{t2}, Y_{t3}, Y_{t4}, Y_{t5} \sim N(0, 1)$ . The dependence structure is modeled through an HAC with Gumbel generators as well. We set also three states ( $M = 3$ ) :

$$C(u_1, C[u_2, C\{u_3, C(u_5, u_4; \theta_1 = 3.15); \theta_2 = 2.45]; \theta_3 = 1.75]; \theta_4 = 1.05) \quad \text{for } i = 1,$$

$$C(u_3, C[u_5, C\{u_2, C(u_1, u_4; \theta_1 = 3.15); \theta_2 = 2.45]; \theta_3 = 1.75]; \theta_4 = 1.05) \quad \text{for } i = 2,$$

$$C(u_5, C[u_4, C\{u_3, C(u_1, u_2; \theta_1 = 3.15); \theta_2 = 2.45]; \theta_3 = 1.75]; \theta_4 = 1.05) \quad \text{for } i = 3,$$

the transition matrix is chosen as:

$$P = \begin{pmatrix} 0.82 & 0.10 & 0.08 \\ 0.08 & 0.84 & 0.08 \\ 0.03 & 0.02 & 0.95 \end{pmatrix},$$

and the initial probabilities are  $\pi = (0.2, 0.1, 0.7)$  and  $T = 2000$ . Figure 7 shows the pairwise scatterplots of the observations generated from the above mentioned model. Similarly, Figure 6 and Table 3 present the estimation accuracy for this model. Although the computation is more demanding when the dimension is higher, we still can achieve the same degree of accuracy as in the three-dimensional case.

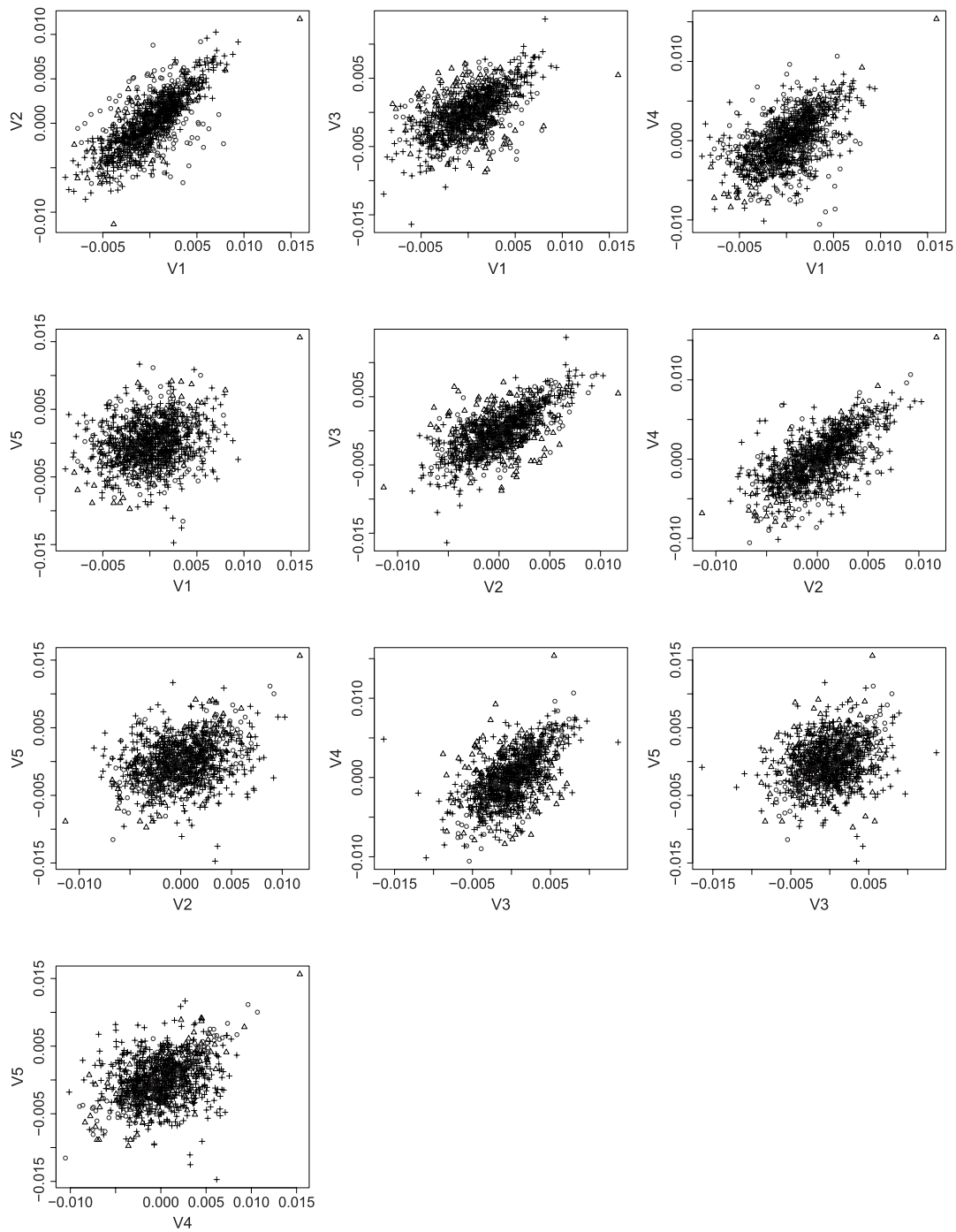
**TABLE 3.** The summary of estimation accuracy in five dimensional model, standard deviations and MSEs are provided in brackets. The case of deGARCHing is with nonparametrically estimated margins

	True	Param. Margins	deGARCHing	
$C_1$	$\theta_1^{(1)}$	1.05	1.019 (0.020, 0.001)	1.019 (0.020, 0.001)
	$\theta_2^{(1)}$	1.75	1.739 (0.077, 0.006)	1.741 (0.078, 0.006)
	$\theta_3^{(1)}$	2.45	2.584 (0.126, 0.034)	2.583 (0.126, 0.034)
	$\theta_4^{(1)}$	3.15	3.328 (0.194, 0.069)	3.318 (0.194, 0.066)
$C_2$	$\theta_1^{(2)}$	1.05	1.017 (0.021, 0.002)	1.017 (0.021, 0.002)
	$\theta_2^{(2)}$	1.75	1.795 (0.084, 0.009)	1.797 (0.084, 0.009)
	$\theta_3^{(2)}$	2.45	2.499 (0.120, 0.017)	2.499 (0.122, 0.017)
	$\theta_4^{(2)}$	3.15	3.381 (0.216, 0.100)	3.369 (0.215, 0.094)
$C_3$	$\theta_1^{(3)}$	1.05	1.044 (0.017, 0.000)	1.045 (0.018, 0.000)
	$\theta_2^{(3)}$	1.75	1.745 (0.041, 0.002)	1.747 (0.041, 0.002)
	$\theta_3^{(3)}$	2.45	2.492 (0.065, 0.006)	2.492 (0.065, 0.006)
	$\theta_4^{(3)}$	3.15	3.189 (0.094, 0.010)	3.185 (0.095, 0.010)
rat. of correct states		0.915 (0.011)	0.915 (0.011)	
$\sum_{i,j=1}^d  \hat{p}_{ij} - p_{ij} $		0.133 (0.054)	0.133 (0.054)	
rat. of correct structures		1	1	

### 4.3. Simulation III

To compare the forecasting performances of the different models, we simulate data from different true models: HMM GARCH, HMM id, and DCC, from which we simulate three-dimensional time series with  $T - 1$  observations. Then we fit different models (HMM GARCH, HMM id, HAC GARCH, HAC id, and DCC) with the  $T - 1$  observations at hand, and compare the one-step ahead distribution forecasts for the true and the estimated models. More specifically, for the distribution forecast comparison, we calculate the sum  $y_{T1} + y_{T2} + y_{T3}$  (which may be thought of as the returns of an equally weighted portfolio).

Simulation of 1000 observations  $y_{T1} + y_{T2} + y_{T3}$  allows us to compare the forecast distribution between the true model and the estimated models. Furthermore, we calculate Kolmogorov–Smirnov (KS) test statistics to measure the difference between the forecast distribution of observations from the true and the estimated model. The comparison has been done with  $T = 250, 500, 1000$  Table 4 reports the means and the standard deviations of the KS test statistics for different models w.r.t. to the true one. We see obvious advantages of our method over the DCC model in the sense that our HMM GARCH model



**FIGURE 7.** Snapshots of pairwise scatter plots of dependency structures ( $t = 0, \dots, 1000$ ).

is in all cases closer on average to the forecast distribution of the true model than is the DCC model. Especially when the data generating processes are HMM GARCH or HMM ID. We use nonparametric estimated margins in this subsection.



**TABLE 4.** The estimated mean KS test statistics (standard deviation) of the forecast distribution from the true model and the estimated model. Number of repetitions is 1000

True\Estimated	Sample size	HMMGARCH	HMM ID	DCC
HMM GARCH	250	<b>0.0899 (0.0353)</b>	0.1243 (0.0571)	0.1949 (0.1112)
DCC		<b>0.0607 (0.0241)</b>	0.0723 (0.0320)	0.0782 (0.0309)
HMM ID		0.0908 (0.0359)	<b>0.0867 (0.0345)</b>	0.1424 (0.0271)
HMMGARCH	500	<b>0.0889 (0.0338)</b>	0.1203 (0.0556)	0.2117 (0.0782)
DCC		<b>0.0541 (0.0194)</b>	0.0672 (0.0325)	0.0774 (0.0254)
HMM ID		<b>0.0936 (0.0331)</b>	0.0924 (0.0326)	0.1515 (0.0239)
HMM GARCH	1000	<b>0.0869 (0.0321)</b>	0.1237 (0.0605)	0.3703 (0.1366)
DCC		<b>0.0494 (0.0166)</b>	0.0659 (0.0320)	0.0823 (0.0392)
HMM ID		<b>0.0919 (0.0331)</b>	0.0907 (0.0322)	0.1509 (0.0213)

## 5. APPLICATIONS

To see how HMM HAC performs on a real data set, applications to financial and rainfall data are offered. A good model for the dynamics of exchange rates gives insights into exogenous economic conditions, such as the business cycle. It is also helpful for portfolio risk management and decisions on asset allocation. We demonstrate the performance of our proposed technique by applying it to forecasting the VaR of a portfolio and compare it with multivariate GARCH models (DCC, BEKK, etc.) The backtesting results show that the VaR calculated from HMM HAC performs significantly better.

The second application is on modeling a rainfall process. HMM is a conventional model for rainfall data, however, bringing HMM and HAC together for modeling the multivariate rainfall process is an innovative modeling path.

### 5.1. Application I

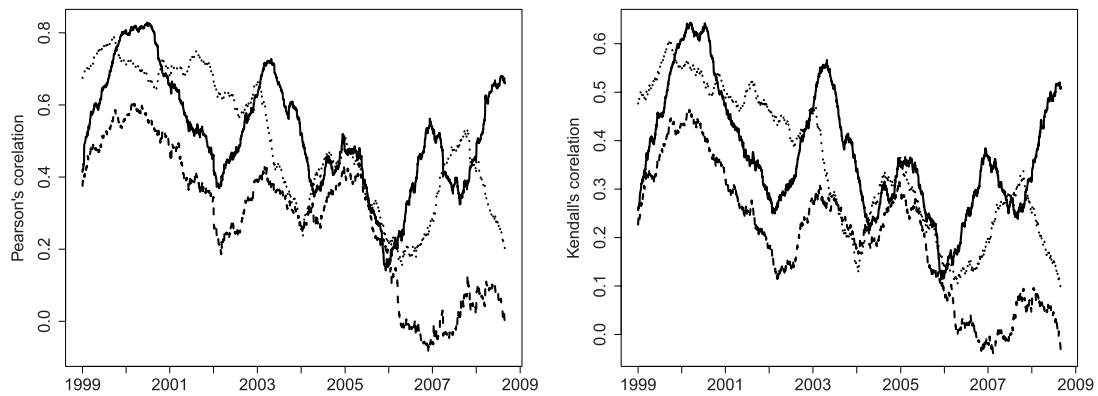
*5.1.1. Data.* The data set consists of the daily values for the exchange rates JPY/EUR, GBP/EUR, and USD/EUR. The covered period is [4.1.1999; 14.8.2009], resulting in 2771 observations.

To eliminate intertemporal conditional heteroscedasticity, we fit a univariate GARCH(1,1) process to each marginal time series of log-returns

$$Y_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t} \quad \text{with } \sigma_{j,t}^2 = \omega_j + \alpha_j \sigma_{j,t-1}^2 + \beta_j (Y_{j,t-1} - \mu_{j,t-1})^2 \quad (24)$$

and  $\omega > 0, \alpha_j \geq 0, \beta_j \geq 0, \alpha_j + \beta_j < 1$ .

The residuals exhibit the typical behavior: they are not normally distributed, which motivates nonparametric estimation of the margins. From the results of



**FIGURE 8.** Rolling window estimators of Pearson's (left) and Kendall's (right) correlation coefficients between the GARCH(1,1) residuals of exchange rates: JPY and USD (solid line), JPY and GBP (dashed line), GBP and USD (dotted line). The width of the rolling window is set to 250 observations.

the Box–Ljung test, whose  $p$ -values are 0.73, 0.01, and 0.87 for JPY/EUR, GBP/EUR, and USD/EUR, we conclude that the autocorrelation of the residuals is strongly significant only for the GBP/EUR rate. After this intertemporal correction, we work only with the residuals.

The dependency variation is measured by Kendall's and Pearson's correlation coefficients: Figure 8 shows the variation of both coefficients calculated in a rolling window of width  $r = 250$ . Their dynamic behavior is similar, but not identical. This motivates once more a time varying copula based model.

*5.1.2. Fitting a HMM model.* Figures 1, 9, and 10 summarize the analysis using three methods: moving window, LCP, and HMM HAC. LCP uses moving windows, with varying sizes. To be more specific, LCP is a scaling technique which determines a local homogeneous window at each time point, see Härdle et al. (2013). In contrast to LCP, HMM HAC is based on a global modeling concept rather than a local one. One observes relatively smooth changes of the parameters, see Figures 1 and 9. HMM HAC is as flexible as LCP, as can be seen from Figures 1, 9, and 10, since the estimated structure also takes three values and is confirmed by the variations of structures estimated from LCP. Moreover, the moving window analysis or LCP can serve as a guideline for choosing the initial values for our HMM HAC. Figure 11 displays the number of states for HMM HAC for rolling windows with a length of 500 observations.

A VaR estimation example is undertaken to show the good performance of HMM HAC. We generate  $N = 10^4$  paths with  $T = 2219$  observations, and  $|W| = 1000$  combinations of different portfolios, where  $W = \{(1/3, 1/3, 1/3)^\top \cup [w = (w_1, w_2, w_3)^\top]\}$ , with  $w_i = w'_i / \sum_{i=1}^3 w'_i$ ,  $w'_i \in U(0, 1)$ . The Profit Loss (P&L) function of a weighted portfolio based on assets  $y_{td}$  is  $L_{t+1} \stackrel{\text{def}}{=} \sum_{d=1}^3 w_i (y_{t+1d} - y_{td})$ , with weights  $w = (w_1, w_2, w_3) \in W$ . The VaR of a particular portfolio at

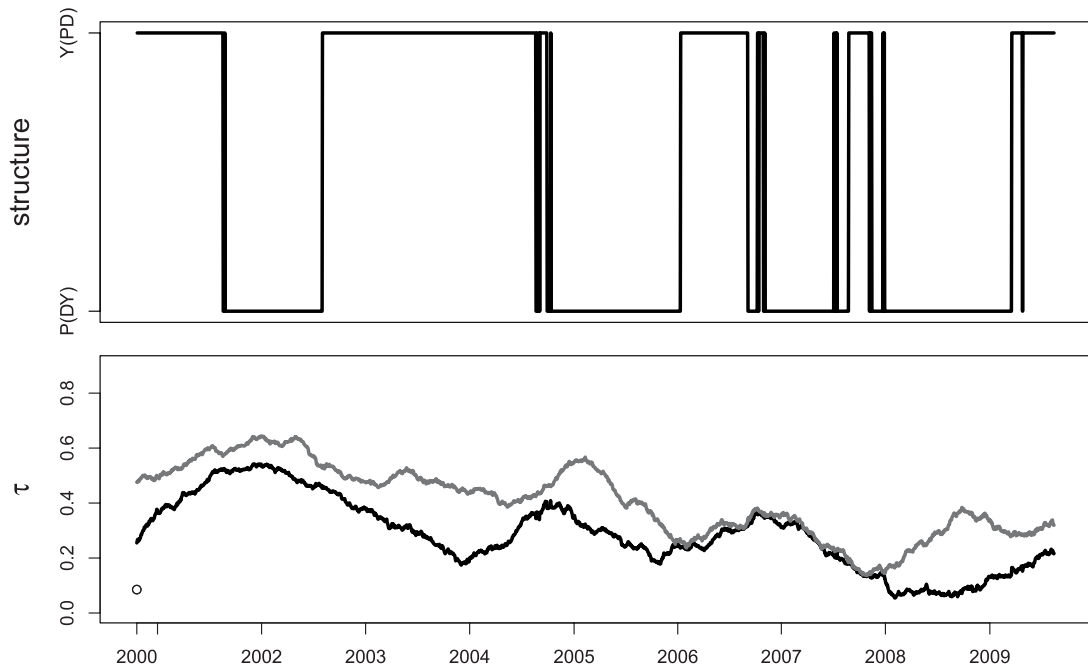


FIGURE 9. Rolling window for exchange rates: structure (upper) and dependency parameters (lower,  $\theta_1$  (gray) and  $\theta_2$  (black)) for Gumbel HAC. Rolling window size  $win = 250$ .

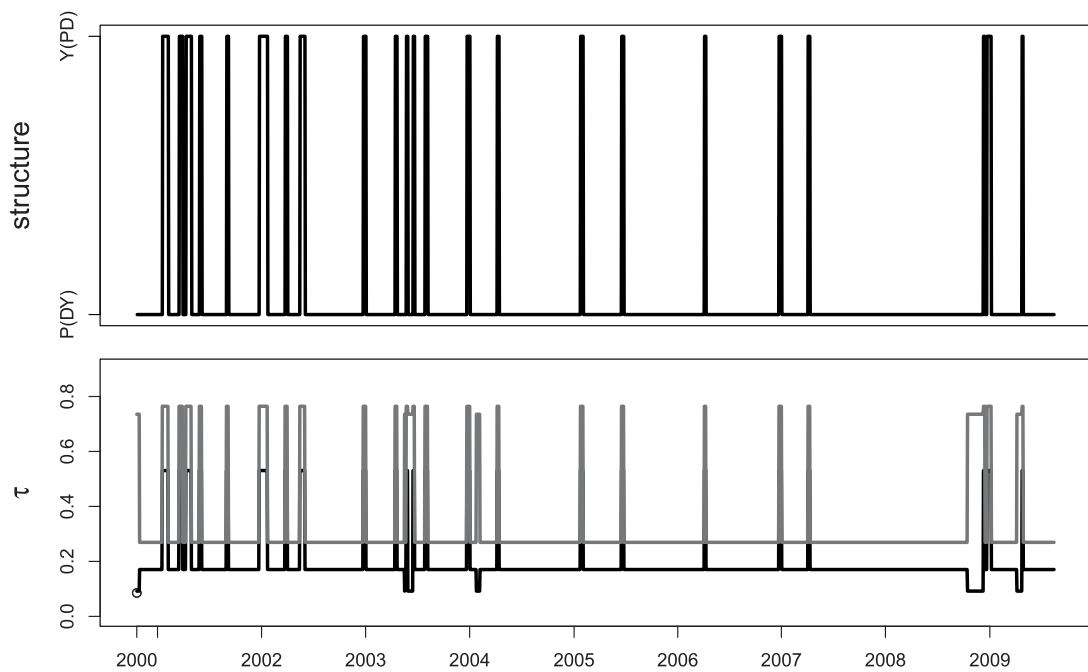


FIGURE 10. HMM for exchange rates: structure (upper) and dependency parameters (lower,  $\theta_1$  (gray) and  $\theta_2$  (black)) for Gumbel HAC.

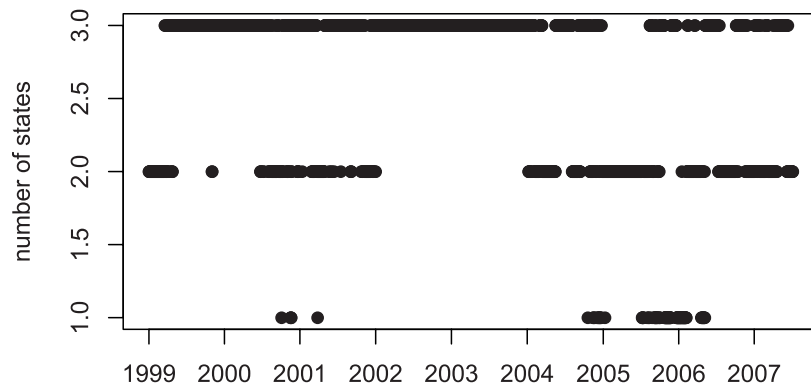


FIGURE 11. Plot of estimated number of states for each window.

level  $0 < \alpha < 1$  is defined as  $VaR(\alpha) \stackrel{\text{def}}{=} F_L^{-1}(\alpha)$ , where the  $\hat{\alpha}_w$  is estimated as a relative fraction of violations, see Table 5:

$$\hat{\alpha}_w \stackrel{\text{def}}{=} T^{-1} \sum_{t=1}^T \mathbf{I}\{L_t < \widehat{VaR}_t(\alpha)\},$$

and the distance between  $\hat{\alpha}_w$  and  $\alpha$  is

$$e_w \stackrel{\text{def}}{=} (\hat{\alpha}_w - \alpha)/\alpha.$$

If the portfolio distribution is i.i.d., and a well calibrated model properly mimicks the true underlying asset process,  $\hat{\alpha}_w$  is close to its nominal level  $\alpha$ . The performance is measured by averaging  $\alpha_w$  over all  $|W|$  portfolios, see Table 5.

We consider four main models: HMM HAC for 500 observation windows for Gumbel and rotated Gumbel; multiple rolling window with 250 observations windows; LCP with  $m_0 = 20$  and  $m_0 = 40$  with Gumbel copulae (the LCP finds the optimal length of window in the past by a sequence of tests on windows of increasing sizes,  $m_0$  is a starting window size); and DCC, see Engle (2002), based

TABLE 5. VaR backtesting results,  $\bar{\hat{\alpha}}$ , where “Gum” denotes the Gumbel copula and “RGum” the rotated survival Gumbel one

Window\alpha	0.1	0.05	0.01	
HMM, RGum	500	0.0980	<b>0.0507</b>	<b>0.0128</b>
HMM, Gum	500	<b>0.0981</b>	0.0512	0.0135
Rolwin, RGum	250	0.1037	0.0529	0.0151
Rolwin, Gum	250	0.1043	0.0539	0.0162
LCP, $m_0 = 40$	468	0.0973	0.0520	0.0146
LCP, $m_0 = 20$	235	0.1034	0.0537	0.0169
DCC	500	0.0743	0.0393	0.0163

**TABLE 6.** Robustness relative to  $A_W(D_W)$

	Window\alpha	0.1	0.05	0.01
HMM, RGum	500	-0.0204 (0.013)	<b>0.0147</b> (0.012)	<b>0.2827</b> (0.064)
HMM, Gum	500	<b>-0.0191</b> (0.008)	0.0233 (0.018)	0.3521 (0.029)
Rolwin, RGum	250	0.0375 (0.009)	0.0576 (0.012)	0.5076 (0.074)
Rolwin, Gum	250	0.0426 (0.009)	0.0772 (0.030)	0.6210 (0.043)
LCP, $m_0 = 40$	468	-0.0270 (0.010)	0.0391 (0.018)	0.4553 (0.037)
LCP, $m_0 = 20$	235	0.0344 (0.009)	0.0735 (0.026)	0.6888 (0.050)
DCC	500	-0.2573 (0.015)	-0.2140 (0.015)	0.6346 (0.091)

on 500 observation windows. For each model we make an out of sample forecast. To better evaluate the performance, we calculated the average and SD of  $e_W$ :

$$A_W = \frac{1}{|W|} \sum_{w \in W} e_w, \quad D_W = \left\{ \frac{1}{|W|} \sum_{w \in W} (e_w - A_W)^2 \right\}^{1/2}.$$

Tables 5 and 6 show the backtesting performance for the described models. One concludes that HMM HAC performs better than the concurring moving window, LCP, or DCC, as  $A_w$  and  $D_w$  are typically smaller in absolute value.

### 5.2. Application II

Rainfall models are used to forecast, simulate, and price weather derivatives. The difficulty in precipitation data is the nonzero point mass at zero and spatial relationships, see Ailliot, Thompson, and Thomson (2009) for Gaussian dependency among locations with HMM application.

In this application we extend it to a copula framework. Unlike application I, the marginal distribution here vary over states. We propose two methods for modeling the marginal distributions: one is to take  $y_{tk}$  to be censored normal distributions, with the following equation:

$$f_k^m\{y_{tk}\} = \begin{cases} 1 - p_k^{x_t} & y_{tk} = 0, \\ p_k^{x_t} \varphi\left\{\frac{y_{tk} - \mu^{x_t}(k)}{\sigma^{x_t}(k)}\right\} / \sigma^{x_t}(k) & y_{tk} > 0; \end{cases}$$

with  $k = 1, \dots, d$  as the location,  $\varphi(\cdot)$  as the standard normal density,  $p_k^{x_t}$  as the rainfall occurrence probability for the location  $k$  and state  $x_t$ , and  $\mu^{x_t}(k), \sigma^{x_t}(k)$  the mean and standard deviation parameters at time  $t$  for location  $k$ .

A second proposal for the marginal distributions are the gamma distributions:

$$f_k^m\{y_{tk}\} = \begin{cases} 1 - p_k^{x_t} & y_{tk} = 0, \\ p_k^{x_t} \gamma\{y_{tk}; \alpha(k)^{x_t}, \beta(k)^{x_t}\} & y_{tk} > 0; \end{cases}$$

where again the  $\alpha(k)^{x_t}, \beta(k)^{x_t}$  are the shape and scale parameters for state  $x_t$  and location  $k$ . We take the joint distribution function to be a truncated version of a continuous copula function, with the copula density  $c_d(\cdot)$  denoted by

$$c_d(\mu, \theta) = \begin{cases} c_c(\mu, \theta), & y_{tk} > 0, \forall k, \\ \partial C_c(\mu, \theta) / \partial \mu_{k_1} \dots \partial \mu_{k_B}, & k_i \in \{y_{tk_i} > 0\}, i \in 1, \dots, E; \end{cases} \quad (25)$$

where  $E$  denotes the number of wet places among the  $d$  locations, the  $C_c$  are the continuous copula functions, and  $c_c$  are the continuous copula densities.

Assume that the daily rainfall observations from the same month are yearly independent realizations of a common underlying hidden Markov model, whose states represent different weather types. As an example, we take every June’s daily rainfall.

$$\begin{aligned} & \log p_T(y_{0:T}, x_{0:T}; v \times \omega) \\ &= \sum_{i=1}^M \mathbf{1}\{x_0 = i\} \log\{\pi_i f_i(y_0)\} + \sum_{t=1}^T \sum_{i=1}^M \sum_{j=1}^M \mathbf{1}\{x_t = j\} \mathbf{1}\{x_{t-1} = i\} \log\{p_{ij} f_j(y_t)\} \\ &+ \sum_{t \in B} \sum_{i=1}^M \left[ \mathbf{1}\{x_t = i\} \{\log(\pi_i)\} - \sum_{j=1}^M \mathbf{1}\{x_t = i\} \mathbf{1}\{x_{t-1} = j\} \log(p_{ji}) \right], \end{aligned}$$

with  $B$  is the set of days which are the first day of June for each year. We use here 50 years of rainfall data from three locations in China: Guangxi, Guangdong, and Fujian (Figure 12). The graphical correlation can naturally be captured by the fitting of different copulae state parameters.

Table 7 presents (with a truncated Gumbel) the estimated three states, the corresponding different marginal distributions and copula parameters, with estimated initial probability:  $\hat{\pi}_{X_t} = (0.298, 0.660, 0.042)$  and estimated transition probability matrix:

$$\hat{P} = \begin{pmatrix} 0.590 & 0.321 & 0.089 \\ 0.188 & 0.742 & 0.080 \\ 0.329 & 0.271 & 0.400 \end{pmatrix}.$$

In the case of our data, gamma distributions fit better as marginals. The states filtered out represent different weather types. The third states are the most humid states, with high rainfall occurrence probabilities, while the second states are drier, and the first are the driest. From the parameters of the gamma distributions, one sees that the variance increases from the first to the third states, which indicates a higher chance for heavy rainfall for the humid states.

To validate our model, 1000 samples of artificial time series of 1500 observations were generated from the fitted model and compared with the original data. Table 8 presents the true Pearson correlation compared with the estimated ones from the generated time series. The 5% confidence intervals of the estimators cover the true correlation, which implies that the simulated rainfall can describe the real correlation of the data quite well. Figure 13 shows a marginal plot



FIGURE 12. Map of Guangxi, Guangdong, Fujian in China.

TABLE 7. Rainfall occurrence probability and shape, scale parameters estimated from HMM (data 1957–2006)

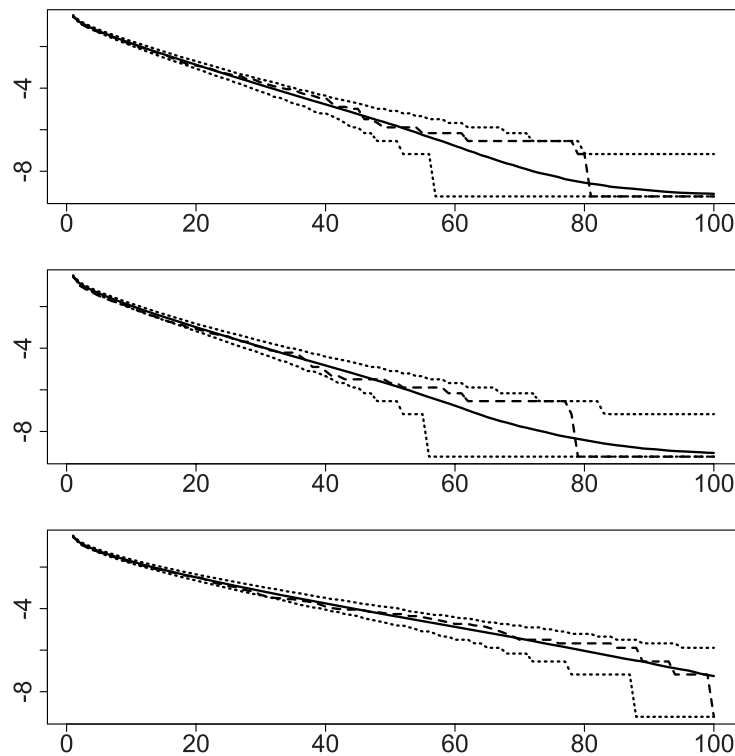
$X_t$	Shape	Scale	Occur Prob
1	(0.442, 0.429, 0.552)	(139.33, 116.70, 169.66)	(0.252, 0.256, 0.439)
2	(0.671, 0.618, 0.561)	(273.83, 253.25, 427.46)	(0.806, 0.786, 0.683)
3	(0.636, 1.125, 0.774)	(381.09, 264.83, 514.08)	(0.667, 1.000, 0.944)

TABLE 8. True correlations, simulated averaged correlations from 1000 samples and their 5% confidence intervals. 1 Fujian, 2 Guangdong, 3 Guangxi

Location	True	$\widehat{\text{Corr}}(Y_{t,1}, Y_{t,2})$
1 – 2	0.308	0.300 (0.235, 0.373)
2 – 3	0.261	0.411 (0.256, 0.586)
1 – 3	0.203	0.130 (0.058, 0.215)

of the log survival function derived from the empirical cdf of the real data and generated data. The log survival function is a transformation of the marginal cdf  $F_k^m(y_{tk})$ :

$$\log\{1 - F_k^m(y_{tk})\}. \tag{26}$$



**FIGURE 13.** Log-survivor-function (black solid) and 95% prediction intervals (gray dotted) of the simulated distribution for the fitted model with sample log-survivor-function superimposed (black dashed).

Again we see that the 95% confidence interval can cover the true curve fairly well.

Table 8 contains the autocorrelations and cross-correlations of the real data and the generated time series. Unfortunately, our generated time series does not show a similar autocorrelation or cross-correlation. Since there is usually more than one significant lag of autocorrelation or cross-correlation, the simulated time series mostly only have one lag. This is an issue also observed in Ailliot et al. (2009). The precipitation can be modeled first by a vector autoregressive (VAR) type model, adjusted for zero observations. An alternative could be to impose an additional dependency structure on  $\{Y_t\}$ .

## 6. CONCLUSION

We propose a dynamic model for multivariate time series with non-Gaussian dependency. Applying an HMM for general copulae leads to a rich clan of dynamic dependency structures. The proposed methodology is helpful in studying financial contagion at an extreme level over time, and it can naturally help in deriving conditional risk measures, such as CoVaR, see Adrian and Brunnermeier (2011). We have shown that dynamic copula models fit financial markets well, and rainfall patterns too.



In the financial application, we performed deGARCHing to remove the second order dependencies in the marginal time series. As this is a  $\sqrt{n}$  step, it will not contaminate the final estimation, and our simulation study confirms this. In the rainfall application, we extend our model to allow the marginal distribution's parameters to also vary over states. Typically it will adapt to nonstationary marginal time series with trend.

## REFERENCES

- Adrian, T. & M.K. Brunnermeier (2011) CoVaR, *Staff Reports 348*, Federal Reserve Bank of New York.
- Ailliot, P., C. Thompson, & P. Thomson (2009) Space-time modeling of precipitation by using a hidden Markov model and censored Gaussian distributions. *Journal of the Royal Statistical Society* 58, 405–426.
- Bickel, P.J., Y. Ritov, & T. Rydén (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics* 26(4), 1614–1635.
- Bickel, P.J. & M. Rosenblatt (1973) On some global measures of the deviations of density function estimates. *The Annals of Statistics* 1, 1071–1095.
- Bradley, R. (1986) Basic properties of strong mixing conditions. In E. Eberlein & M.S. Taquq (eds.), *Dependence in Probability and Statistics*, pp. 165–192. Birkhauser.
- Caia, Z., X. Chen, Y. Fan, & X. Wang (2006) Selection of Copulas with Applications in Finance. Working paper. Available at <http://www.economics.smu.edu.sg/femes/2008/papers/219.pdf>.
- Cappé, O., E. Moulines, & T. Rydén (2005) *Inference in Hidden Markov Models*. Springer-Verlag.
- Chen, X. & Y. Fan (2005) Estimation of copula-based semiparametric time series models. *Journal of Econometrics* 130(2), 307–335.
- Chen, X. & Y. Fan (2006) Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics* 135, 125–154.
- Dempster, A., N. Laird, & D. Rubin (1977) Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society B* 39, 1–38.
- Engle, R. (2002) Dynamic conditional correlation. *Journal of Business and Economic Statistics* 20(3), 339–350.
- Fuh, C.-D. (2003) SPRT and CUSUM in hidden Markov models. *Annals of Statistics* 31(3), 942–977.
- Gao, X. & P.X.-K. Song (2011) Composite likelihood EM algorithm with applications to multivariate hidden Markov model. *Statistica Sinica* 21, 165–185.
- Giacomini, E., W.K. Härdle, & V. Spokoiny (2009) Inhomogeneous dependence modeling with time-varying copulae. *Journal of Business and Economic Statistics* 27(2), 224–234.
- Hamilton, J. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57(2), 357–384.
- Härdle, W., H. Herwartz, & V. Spokoiny (2003) Time inhomogeneous multiple volatility modeling. *Journal of Financial econometrics* 1(1), 55–95.
- Härdle, W.K., O. Okhrin, & Y. Okhrin (2013) Dynamic structured copula models. *Statistics & Risk Modeling* 30(4), 361–388.
- Joe, H. (1997) *Multivariate Models and Dependence Concepts*. Chapman & Hall.
- Leroux, B.G. (1992) Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications* 40, 127–143.
- Liu, W. & W. Wu (2010) Simultaneous nonparametric inference of time series. *The Annals of Statistics* 38, 2388–2421.
- McLachlan, G. & D. Peel (2000) *Finite Mixture Models*. Wiley.
- McNeil, A.J. & J. Nešlehová (2009) Multivariate Archimedean copulas,  $d$ -monotone functions and  $l_1$  norm symmetric distributions. *Annals of Statistics* 37(5b), 3059–3097.

- Nelsen, R.B. (2006) *An Introduction to Copulas*. Springer-Verlag.
- Okhrin, O., Y. Okhrin, & W. Schmid (2013) On the structure and estimation of hierarchical Archimedean copulas. *Journal of Econometrics* 173, 189–204.
- Okimoto, T. (2008) Regime switching for dynamic correlations. *Journal of Financial and Quantitative Analysis* 43(3), 787–816.
- Patton, A.J. (2004) On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics* 2, 130–168.
- Pelletier, D. (2006) Regime switching for dynamic correlations. *Journal of Econometrics* 131, 445–473.
- Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE* 77(2), 257–286.
- Rodriguez, J.C. (2007) Measuring financial contagion: A copula approach. *Journal of Empirical Finance* 14, 401–423.
- Savu, C. & M. Tiede (2010) Hierarchical Archimedean copulas. *Quantitative Finance* 10, 295–304.
- Sklar, A. (1959) Fonctions de répartition à  $n$  dimension et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris* 8, 299–231.
- Whelan, N. (2004) Sampling from Archimedean copulas. *Quantitative Finance* 4, 339–352.

## APPENDIX

### A.1. Proof of Theorems 3.1 and 3.2

In the HMM HAC framework, let  $\{X_t, t \geq 0\}$  with transition probability matrix  $P^{v,\omega} = [p_{ij}^{v,\omega}]_{i,j=1,\dots,M}$  and initial distribution  $\pi^{v,\omega} = \{\pi_i^{v,\omega}\}_{i=1,\dots,M}$ , where  $\{v,\omega\} \in \{V,\Omega\}$  denotes an element in the parameter space  $\{V,\Omega\}$  which parametrizes this model, and  $q$  is the number of continuous parameters (note that our parameter space is partly discrete ( $V$ ) and partly continuous ( $\Omega$ )). We introduce the event  $\{v,\omega\}$  because  $\Omega$  correspond to events induced by continuous parameters  $\theta, s_j, p_{ij}, \pi_i$ . Suppose that  $B_{t,j}$  is a real-valued additive component equal to  $\sum_{k=0}^t Y_{k,j}, j \in 1, \dots, d$ , with  $B_t = (B_{t,1}, B_{t,2}, \dots, B_{t,d})^\top$  and with  $Y_k = (Y_{k,1}, Y_{k,2}, \dots, Y_{k,d})^\top$  a r.v. taking values on  $\mathbb{R}^d$ . Suppose further that  $B_{t,j}$  is adjoined to the chain in such a way that  $\{(X_t, B_t), t \geq 0\}$  is a Markov chain on  $D \times \mathbb{R}^d$  and

$$\begin{aligned} & \mathbb{P}\{(X_t, B_t) \in A \times (B+b) | (X_{t-1}, B_{t-1}) = (i, b)\} \\ &= \mathbb{P}\{(X_1, B_1) \in A \times B | (X_0, B_0) = (i, 0)\} \\ &= \mathbb{P}(i, A \times B) = \sum_{j \in A} \int_{b \in B} p_{ij}^{v \times \omega} f_j \left\{ b; \theta^{(j)}(v \times \omega), s^{(j)}(v \times \omega) \right\} \mu(db), \end{aligned} \tag{A.1}$$

where  $B, b \subseteq \mathbb{R}^d$ ,  $A \subseteq D$ ,  $f_j\{b; \theta^{(j)}(v, \omega), s^{(j)}(v, \omega)\}$  is the conditional density of  $Y_t$  given  $X_{t-1}, X_t$  with respect to a  $\sigma$ -finite measure  $\mu$  on  $\mathbb{R}^d$ , and  $\theta(v, \omega) \in \Theta, s(v, \omega) \in S, j = 1, \dots, M$  are the unknown parameters. That is,  $\{X_t, t \geq 0\}$  is a Markov chain, given  $X_0, X_1, \dots, X_T$ , with  $Y_1, \dots, Y_T$  being independent. In this situation,  $\{B_t, t \geq 0\}$  is called a *hidden Markov model* if there is a Markov chain  $\{X_t, t \geq 0\}$  such that the process  $\{(X_t, B_t), t \geq 0\}$  satisfies (A.1). Note that in (A.1), the usual parameterization  $\theta^{(j)}(v, \omega) = \theta^{(j)}$ , and  $s^{(j)}(v, \omega) = s^{(j)}$ .

Recall the associated parameter space  $\{V, \Omega\}$ , where  $V$  consists of a set of discrete finite elements and  $\Omega$  is associated with the parameters  $\theta, [p_{ij}]_{i,j}$ . Define  $\mathbf{s}^*$  and  $\theta^*$  associated

with the point  $\{v^0, \omega^0\}$  in the parameter space, as follows.

$$q_T(Y_{0:T}; v^0, \omega^0) \stackrel{\text{def}}{=} \max_{j \in 0, \dots, M} p_T(Y_{0:T} | x_1 = j; v^0, \omega^0) \tag{A.2}$$

$$H(v^0, \omega^0) \stackrel{\text{def}}{=} E_{v^0, \omega^0} \{-\log p(Y_0 | Y_{-1}, Y_{-2}, \dots; v^0, \omega^0)\}$$

Here, the  $Y_{-1}, \dots, Y_{-T}$  are a finite number of past values of the process.

$$H(v^0, \omega^0, v, \omega) \stackrel{\text{def}}{=} E_{v^0, \omega^0} \{\log p_T(Y_{0:T}; v, \omega)\}$$

THEOREM A.1 (Leroux (1992)). *Under A.1–A.5,*

$$\begin{aligned} \lim_{T \rightarrow \infty} T^{-1} E_{v^0, \omega^0} \{\log p_T(Y_{0:T}; v^0, \omega^0)\} &= -H(v^0, \omega^0) \\ \lim_{T \rightarrow \infty} T^{-1} \log p_T(Y_{0:T}; v^0, \omega^0) &= -H(v^0, \omega^0), \end{aligned}$$

with probability 1, under  $(v^0, \omega^0)$ , and

$$\begin{aligned} \lim_{T \rightarrow \infty} T^{-1} E_{v^0, \omega^0} \{\log p_T(Y_{0:T}; v, \omega)\} &= H(v^0, \omega^0, v, \omega) \\ \lim_{T \rightarrow \infty} T^{-1} \log p_T(Y_{0:T}; v, \omega) &= H(v^0, \omega^0, v, \omega), \end{aligned}$$

with probability 1, under  $(v^0, \omega^0)$ .

LEMMA A.2.  $\forall v_i, u_j, i, j \in 1, \dots, M$  as weights, the difference between  $M$  linear combination of states leads to

$$\sum_{i=1}^M v_i f(y, \theta_{s^{(i)}}, s^{(i)}) \neq \sum_{j=1}^M \mu_j f(y, \theta_{s^{(j)}}, s^{(j)}). \tag{A.3}$$

**Proof.** For each  $s^{(i)}, i \in 1, \dots, M$  associated with dependency parameter  $\theta_{s^{(i)}} \in \mathbb{R}_+^d$ . So

$$\sum_{i=1}^M v_i \delta_{s^{(i)}} \neq \sum_{j=1}^M \mu_j \delta_{s^{(j)}}, a.e. \tag{A.4}$$

implies

$$\sum_{i=1}^M v_i \delta_{s^{(i)}} \delta \theta_{s^{(i)}} \neq \sum_{j=1}^M \mu_j \delta_{s^{(j)}} \delta \theta_{s^{(j)}}, a.e. \tag{A.5}$$

■

Furthermore, if (A.4), then the corresponding point in the parameter space  $(v, \omega)$  leads to  $\mathcal{K}(v^0, \omega^0; v, \omega)$ , and  $(v, \omega)$  would not be in the equivalent class of  $(v^0, \omega^0)$  as long as the points  $v$  and  $v^0$  are different as (A.4) (the equivalence class of  $v^0$  is defined in Leroux (1992)), and the divergence between  $(v, \omega)$  and  $(v^0, \omega^0)$  is defined as

$\mathcal{K}(v^0, \omega^0; v, \omega) \stackrel{\text{def}}{=} H(v^0, \omega^0, v^0, \omega^0) - H(v^0, \omega^0, v, \omega)$ . This is connected with the log likelihood ratio process, and one can prove that if either (A.4) or (A.5), and provided that (A.2) holds, then (A.3) will hold, and so  $\mathcal{K}(v^0, \omega^0; v, \omega) > 0$ . Namely, the divergence can distinguish between points from different equivalent classes.

Next, we study whether plugging in nonparametric estimated margins would affect the consistency results by analyzing the uniform convergence of  $\hat{f}_i(y)$ .

Recall  $\hat{f}_i(y) \stackrel{\text{def}}{=} c\{\hat{F}_1^m(y_1), \hat{F}_2^m(y_2), \dots, \hat{F}_d^m(y_d), \hat{\theta}^{(i)}, \hat{s}^{(i)}\} \hat{f}_1^m(y_1) \hat{f}_2^m(y_2) \cdots \hat{f}_d^m(y_d)$ . We have, according to the uniform consistency of copulae density, for all  $t \in 1, \dots, T$ ,  $i \in 1, \dots, M$ ,

$$\max_{s^{(i)}} \sup_{y_{t1}, \dots, y_{td} \in B^d, \theta^{(i)} \in \Theta} \left| \hat{c}(\hat{F}_1^m(y_{t1}), \hat{F}_2^m(y_{t2}), \dots, \hat{F}_d^m(y_{td}), \theta^{(i)}, s^{(i)}) - c(F_1^m(y_{t1}), F_2^m(y_{t2}), \dots, F_d^m(y_{td}), \theta^{(i)}, s^{(i)}) \right| \tag{A.6}$$

$$\leq \sum_{j=1}^d \left| c(F_{1, \eta_1}^m(y_{t1}), F_{2, \eta_2}^m(y_{t2}), \dots, F_{d, \eta_d}^m(y_{td})) \{ \hat{F}_j^m(y_{tj}) - F_j^m(y_{tj}) \} \right|, \tag{A.7}$$

where  $F_{j, \eta_j}^m(\cdot) \stackrel{\text{def}}{=} F_j^m(\cdot) + \eta_j[F(\cdot) - F_j^m(\cdot)]$ ,  $\eta_j = [0, 1]$ , and  $F_{j, \eta_j}^m(\cdot)$  lies in the set of admission functions for  $F_j^m$ .

Bickel et al. (1998) states that as  $\{X_t\}$  is ergodic, then it follows that  $\{Y_t\}$  is also ergodic. It is known that any strictly irreducible and aperiodic Markov chain is  $\beta$ -mixing, Bradley (1986). Then the marginal distribution of  $Y_{tm}, m = 1, \dots, M$  follows a process that is  $\beta$ -mixing with an exponential decay rate, namely  $\beta_t = \mathcal{O}\{t^{-b}\}$  for some constant  $a$ . The temporal dependence of the marginal univariate time series  $Y_{tm}$  is inherited simply from the underlying Markov chain as it is a measurable transformation of  $X_t$ . Since  $\{Y_t\}$  follows HMM HAC, then the marginal distribution of  $Y_{tm}$  follows a process that is  $\beta$ -mixing with decay rate  $\beta_t = \mathcal{O}(b^{-t})$  for some constant  $b$ . Then it follows from the results of Liu and Wu (2010), under assumptions A1–A5, that the marginal kernel density estimation has a Bickel and Rosenblatt (1973)-type of uniform consistency.

$$\sup_{y \in B} |\hat{f}_i^m(y) - f_i^m(y)| = \mathcal{O}_p(1) \tag{A.8}$$

Also according to Chen and Fan (2005),

$$\sqrt{T} \sup_{y \in B} |\hat{F}_m^m(y) - F_m^m(y)| = \mathcal{O}_p(1). \tag{A.9}$$

Finally, we have

$$\max_s \sup_{y_1, \dots, y_d \in B^d, \theta \in E} \left| \hat{c}(\hat{F}_1^m(y_1), \hat{F}_2^m(y_2), \dots, \hat{F}_d^m(y_d), \theta^{(i)}, s^{(i)}) - c(F_1^m(y_1), F_2^m(y_2), \dots, F_d^m(y_d), \theta^{(i)}, s^{(i)}) \right| = \mathcal{O}_p(1).$$

Therefore, the multivariate distribution at each state satisfies

$$\sup_{y \in B^d} |\hat{f}_j(y) - f_j(y)| = \mathcal{O}_p(1),$$

where  $B, B^d$  are compact sets. So the plug in estimation does not contaminate the consistency results.

To prove the consistency of our estimation of this parameter, we restate the theorems of consistency in Leroux (1992) for our parameter space. One needs to show that for a discrete subspace  $V^c$  which does not contain any point of the equivalence class of  $v^0$ , for  $v \in V^c$  and an arbitrary value of  $\omega \in \Omega$ , that, with probability 1,

$$\lim_{T \rightarrow \infty} \left[ \min_{v \in V^c} \log \sup_{\omega \in \Omega} p_T(Y_{0:T}; v, \omega) - \log p_T(Y_{0:T}; v^0, \omega^0) \right] \rightarrow -\infty. \tag{A.10}$$

This follows directly from Lemma A.2 (the identifiability of the state parameters) and its consequence  $\mathcal{K}(v^0, \omega^0; v, \omega) > 0$ . Theorem 3.1 is proved.

To prove Theorem 3.2, note that  $\lim_{T \rightarrow \infty} \max_{i \in 1, \dots, M} \mathbb{P}(|\hat{\theta}^{(i)} - \theta^{*(i)}| > \varepsilon | \hat{s}^{(i)} = s^{*(i)})$  is conditioned on the event  $\{\hat{s}^{(i)} = s^{*(i)}\}$  which asymptotically holds with probability 1. Therefore it suffices to prove, for any  $\hat{s}^{(i)} = s^{(i)}$

$$\lim_{T \rightarrow \infty} \min_{i \in 1, \dots, M} \mathbb{P}(|\hat{\theta}^{(i)} - \theta^{*(i)}| > \varepsilon) = 0. \tag{A.11}$$

To show (A.11), one needs to show that for a  $(V^c, \Omega^c)$  which does not contain any point of the equivalence class of  $(v^0, \omega^0)$ , we have, with probability 1,

$$\lim_{T \rightarrow \infty} \left\{ \log \sup_{\omega \in \Omega^c} p_T(Y_{0:T}; v^0, \omega) - \log p_T(Y_{0:T}; v^0, \omega^0) \right\} \rightarrow -\infty, \tag{A.12}$$

which is implied from the following statement: for any closed subset  $C$  of  $\Omega^c$ , there exists a sequence of open subsets of  $\mathcal{O}_{\omega_h}$  with  $h = 1, \dots, H$  with  $C \subseteq \cup_{h=1}^H \mathcal{O}_{\omega_h}$ , such that

$$\lim_{T \rightarrow \infty} \left\{ \max_h \log \sup_{\omega \in \mathcal{O}_{\omega_h}} p_T(Y_{0:T}; v^0, \omega) - \log p_T(Y_{0:T}; v^0, \omega^0) \right\} \rightarrow -\infty. \tag{A.13}$$

To prove (A.13), we have the modified definition:

$$H(v^0, \omega^0, v^0, \omega; \mathcal{O}_{\omega_h}) \stackrel{\text{def}}{=} \lim_T \log \sup_{\omega' \in \omega^0} q_T(Y_{0:T}, v^0, \omega')/T. \tag{A.14}$$

It can be derived that

$$H(v^0, \omega^0, v^0, \omega) < H(v^0, \omega^0, v^0, \omega^0), \tag{A.15}$$

when  $(v^0, \omega)$  and  $(v^0, \omega^0)$  do not lie in the same equivalence class. Then (A.15) is a consequence of the identifiability condition A.2, and this leads to:  $\exists \varepsilon > 0, T_\varepsilon$  and  $\mathcal{O}_\omega$  such that

$$E \log \sup_{\omega' \in \mathcal{O}_\omega} q_{T_\varepsilon}(v^0, \omega')/T_\varepsilon < E \log q_{T_\varepsilon}(v^0, \omega)/T_\varepsilon + \varepsilon < H(v^0, \omega^0, v^0, \omega^0) - \varepsilon.$$

Also because  $\log \sup_{\omega' \in \mathcal{O}_\omega} p_T(Y_{0:T}, v^0, \omega')/T$  and  $\log \sup_{\omega' \in \mathcal{O}_\omega} q_T(Y_{0:T}, v^0, \omega')/T$  have the same limit value, there exists a constant  $\varepsilon > 0$ ,

$$\lim_{T \rightarrow \infty} \log \sup_{\omega' \in \mathcal{O}_{\omega_h}} p_T(y_{0:T}, v^0, \omega')/T = H(v^0, \omega^0, v^0, \omega; \mathcal{O}_{\omega_h}) \leq H(v^0, \omega^0, v^0, \omega^0) - \varepsilon.$$

Now (A.13) follows.

### A.2. Proof of Theorem 3.3

Recall from the last subsection, under A.3,

$$\sup_{y \in B} |\hat{f}_i^m(y) - f_i^m(y)| = \mathcal{O}_p(1) \tag{A.16}$$

$$\sqrt{T} \sup_{y \in B} |\hat{F}_m^m(y) - F_m^m(y)| = \mathcal{O}_p(1). \tag{A.17}$$

Let  $U_{tm} \stackrel{\text{def}}{=} F_m^m(Y_{tm})$ ,  $\tilde{U}_{tm} \stackrel{\text{def}}{=} \hat{F}_m^m(Y_{tm})$ , and  $\mathbf{U}_t \stackrel{\text{def}}{=} (U_{t1}, \dots, U_{td})$ . Define the log likelihood  $L_T(\boldsymbol{\theta}) = L_T(\boldsymbol{\theta}, \mathbf{U}_{0:T}) \stackrel{\text{def}}{=} \log p_T(y_{0:T})$ ; in our case, we will work with  $L_T(\boldsymbol{\theta}, \tilde{\mathbf{U}}_{0:T})$ . Relying on the LAN property proved in Bickel et al. (1998), under A.1–A.7, we have

$$\begin{aligned} L_T(\boldsymbol{\theta}^* + T^{-1/2}\boldsymbol{\theta}, \mathbf{U}_{0:T}) - L_T(\boldsymbol{\theta}^*, \mathbf{U}_{0:T}) \\ = T^{-1/2}\boldsymbol{\theta}^\top \partial L_T(\boldsymbol{\theta}^*) + T^{-1}\boldsymbol{\theta}^\top \partial^2 L_T(\boldsymbol{\theta}^*)\boldsymbol{\theta}/2 + R_T(\boldsymbol{\theta}), \end{aligned} \tag{A.18}$$

where  $R_T(\boldsymbol{\theta})$  tends to zero in probability, uniformly on compact subsets of the parameter space of  $\boldsymbol{\theta}$ .

Next we need to prove that, uniformly over  $\boldsymbol{\theta}$ ,

$$\begin{aligned} L_T(\boldsymbol{\theta}^* + T^{-1/2}\boldsymbol{\theta}, \mathbf{U}_{0:T}) - L_T(\boldsymbol{\theta}^*, \mathbf{U}_{0:T}) - L_T(\boldsymbol{\theta}^* + n^{-1/2}\boldsymbol{\theta}, \tilde{\mathbf{U}}_{0:T}) + L_T(\boldsymbol{\theta}^*, \tilde{\mathbf{U}}_{0:T}) \\ - T^{-1/2}\boldsymbol{\theta}^\top \sum_t \sum_m W_m(U_{tm}) = \mathcal{O}_p\{R_T(\boldsymbol{\theta})\}, \end{aligned}$$

where

$$\begin{aligned} W_m(U_{tm}) \stackrel{\text{def}}{=} \int_{v_1, \dots, v_d} \{\mathbf{1}(U_{tm} \leq v_m) - v_m\} (\mathbb{E} \partial \tilde{a}_t \tilde{b}_m / \partial \boldsymbol{\theta} | \boldsymbol{\theta} = \boldsymbol{\theta}^*) \\ \times c(v_1, \dots, v_d, \boldsymbol{\theta}^{*(m)}, s^{*(m)}) dv_1 \cdots dv_d. \end{aligned}$$

$\tilde{a}_t(\cdot)$  and  $\tilde{b}_m(\cdot)$  are functions defined later in the proof.

Similarly, we have

$$\begin{aligned} L_T(\boldsymbol{\theta}^*, \tilde{\mathbf{U}}_{0:T}) - L_T(\boldsymbol{\theta}^*, \mathbf{U}_{0:T}) \\ = \log \left( \frac{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \tilde{f}_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} \tilde{f}_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} f_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)} \right) \\ = \frac{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \tilde{f}_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} \tilde{f}_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)} \\ - \frac{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} f_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)} + \mathcal{O}_p(1) \\ \stackrel{\text{def}}{=} \sum_t \sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \tilde{a}_t(\boldsymbol{\theta}^*) \{ \tilde{f}_{x_t}(y_t) - f_{x_t}(y_t) \} + \mathcal{O}_p(1), \end{aligned}$$

where  $\tilde{a}_{t_0}(\boldsymbol{\theta}^*) = \frac{\pi_{x_0} \tilde{f}_{x_0}(y_0) \prod_{t=1}^{t_0} p_{x_{t-1}x_t} \tilde{f}_{x_t}(y_t) \prod_{t=t_0+1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}{\sum_{x_0=1}^M \cdots \sum_{x_T=1}^M \pi_{x_0} f_{x_0}(y_0) \prod_{t=1}^T p_{x_{t-1}x_t} f_{x_t}(y_t)}$ .

As

$$\begin{aligned} \tilde{f}_{x_t}(y_t) - f_{x_t}(y_t) &= c\left(\tilde{\mathbf{U}}_{0:T}, \boldsymbol{\theta}^{*(x_t)}, s^{*(x_t)}\right) \prod_{m=1}^d f_m^m - c\left(\mathbf{U}_{0:T}, \boldsymbol{\theta}^{*(x_t)}, s^{*(x_t)}\right) \prod_{j=1}^d f_j^m \\ &= \sum_m c_{u_m} \left\{ F_1^m(y_{1t}), F_2^m(y_{2t}), \dots, F_d^m(y_{dt}), \boldsymbol{\theta}^{*(x_t)}, s^{*(x_t)} \right\} \\ &\quad \times \left\{ \hat{F}_m^m(y_{mt}) - F_m^m(y_{mt}) \right\} \prod_{j=1}^d f_j^m + \mathcal{O}_p(1) \\ &\stackrel{\text{def}}{=} \sum_m \tilde{b}_m(\boldsymbol{\theta}^{(x_t)}) \left\{ \hat{F}_m^m(y_{mt}) - F_m^m(y_{mt}) \right\} + \mathcal{O}_p(1), \end{aligned}$$

where  $\tilde{b}_m(\boldsymbol{\theta}^{(x_t)}) \stackrel{\text{def}}{=} c_{u_m} \{ F^m(y_{1t}), F^m(y_{2t}), \dots, F^m(y_{dt}), \boldsymbol{\theta}^{(x_t)}, s^{(x_t)} \} \prod_{j=1}^d f_j^m$ , and  $c_{u_m}$  denotes the partial derivative of the copulae density w.r.t.  $u_m$ .

Then it follows that

$$\begin{aligned} &L_T(\boldsymbol{\theta}^* + T^{-1/2}\boldsymbol{\theta}, \mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T}) - L_T(\boldsymbol{\theta}^* + T^{-1/2}\boldsymbol{\theta}, \tilde{\mathbf{U}}_{1:T}) + L_T(\boldsymbol{\theta}^*, \tilde{\mathbf{U}}_{1:T}) \\ &= T^{-1/2}\boldsymbol{\theta}^\top \sum_{x_0=1}^M \dots \sum_{x_T=1}^M \sum_t \left[ \sum_m \partial \tilde{a}_t \tilde{b}_m / \partial \boldsymbol{\theta} \{ \hat{F}_m^m(y_{mt}) - F_m^m(y_{mt}) \} \right] + \mathcal{O}_p(T^{-1/2}) \\ &= T^{-1/2}\boldsymbol{\theta}^\top \sum_t \sum_m W_m(U_{tm}) + \mathcal{O}_p(T^{-1/2}) \end{aligned}$$

So, let

$$\begin{aligned} B &\stackrel{\text{def}}{=} \mathbb{E} \{ \partial^2 L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T}) \} \\ A &\stackrel{\text{def}}{=} \left\{ \partial L_T(\boldsymbol{\theta}^*, \mathbf{U}_{1:T}) + \sum_t \sum_m W_m(U_{tm}) \right\}, \end{aligned} \tag{A.19}$$

Finally, we have that the estimated  $\hat{\boldsymbol{\theta}}$  can be represented by  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = B^{-1}A + \mathcal{O}_p(T^{-1/2})$  coming from Bickel et al. (1998).



# Ladislaus von Bortkiewicz—Statistician, Economist and a European Intellectual

Wolfgang Karl Härdle<sup>1,2</sup> and Annette B. Vogt<sup>3</sup>

<sup>1</sup>*Ladislaus von Bortkiewicz Chair of Statistics, C. A. S. E.—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany*

<sup>2</sup>*Lee Kong Chian School of Business, Singapore Management University, Singapore*

*E-mail: haerdle@hu-berlin.de*

<sup>3</sup>*Max Planck Institute for the History of Science, Boltzmannstraße 22, 14195 Berlin, Germany*

*E-mail: vogt@mpiwg-berlin.mpg.de*

## Summary

Ladislaus von Bortkiewicz (1868–1931) was a European statistician. His scientific work covered theoretical economics, stochastics, mathematical statistics and radiology; today, we would call him a cross-disciplinary scientist. With his clear views on mathematical principles with their applications in these fields, he stood in conflict with the mainstream economic schools in Germany at the dawn of the 20th century. He had many prominent students (Gumbel, Leontief and Freudenberg among them), and he carved out the path of modern statistical thinking. He was a true European intellectual with a career path from St. Petersburg via Göttingen to Straßburg and finally the Berliner Universität, now Humboldt-Universität zu Berlin. He is known for the precise calibration of insurance claims applying the—at that time hardly known—Poisson distribution to Prussian horse kick and child suicide data. He proposed a simple solution to the Marxian transformation problem and wrote numerous articles and books on the mathematical treatment of statistical (including radiological physical) data. In this article, we sketch his life and work and point out the prominent role that he has in today's statistical thinking.

*Key words:* History of science; statistics; horse kicks; Bortkiewicz.



## 1 Introduction

On 16 July 1931 in an obituary for the Berlin newspaper ‘Vossische Zeitung’, Ock (1931) wrote:

Bortkiewicz war ein Meister der Wahrscheinlichkeitsrechnung. Die Beherrschung dieser Wahrscheinlichkeitsrechnung, die er auf Versicherungswissenschaft und Bevölkerungslehre genau so anwandte wie auf naturwissenschaftliche Gebiete, besonders auf die radioaktive Strahlung, trug dazu bei, Bortkiewicz den Ruf als einen der fähigsten Statistiker der Welt zu schaffen. (Bortkiewicz was a master in theory of probability. The containment of this theory of probability allowed him to apply it to the insurance science and the theory of population as well as on topics in science such as radioactivity; this led him to have the reputation as one of the most capable (competent) statistician in the world.)

About 30 years later, one of Bortkiewicz’s students, Emil Julius Gumbel (1891–1966), underlined that ‘He was one of the few representatives of mathematical statistics in Germany and as such a lonely figure, highly respected but rarely understood’ (Gumbel, 1968, p. 128) and pointed out that ‘His writings stimulated numerous scientists in Germany, in the northern European countries and in Italy, but not in England.’ (Gumbel, 1968, p. 130) These quotations demonstrate that Ladislaus von Bortkiewicz (LvB) was in fact one of the founders of statistical science as we know it today. What was this topic—statistical science—in the early 20th century? How was it linked with the development of scientific disciplines like economy, political sciences and mathematics? How and what did LvB contribute to this development? These are questions that we would like to answer and thereby shed some light on the development of statistics in the early 20th century. Statistics as a scientific discipline like physics or medicine, meaning statistics as a science with its own fundamental laws, its own technology and methods, was defined at the end of the 19th century. The first International Statistical Congress was organised by Adolphe Quetelet (1796–1874) and held in Brussels in September 1853. The first Session of the International Statistical Institute (ISI) was held in Rome in 1887, and the second in Paris in 1889. LvB was elected to the ISI in 1903, and his friend A. A. Chuprov in 1913. In central Europe, it took a little while longer to establish statistics firmly in the curriculum *universitas*. The importance of this science was made evident when the German Statistical Association was founded in June 1911. Georg von Mayr (1841–1925) became its first president, and he was the president until his death in 1925. When he was honoured in 1911, it was written about statistics:

Die Statistik nimmt heute auf weiten Gebieten des öffentlichen Lebens eine Achtung gebietende, einflussreiche Stellung ein. Reich, Staat, Kommune, Allgemeinheit, Privatwirtschaft, Wissenschaft bedienen sich ihrer Hilfe in ausgedehntem Maße. Die Statistik ist selbst zu einer Wissenschaft geworden. (quoted in Steger (2011), p. 17)

(Statistics today occupies an influential and an imposing position in many public spheres. The (German) Federation, counties and communities, the general public, industry and science make use of it extensively. Moreover, statistics itself has become a science in itself.)

One of the driving forces for the establishment of statistics as a science at this time was LvB, as he realised that the introduction of mathematical concepts into the analysis of statistical data created a new quality. As S. Hertz rightly mentioned, ‘Of his generation, Bortkiewicz was the main representative of mathematical statistics in Germany’ (Hertz, 2001, p. 274). Unfortunately, LvB died too early to see the fruits of his thoughts ripen in the work of his brilliant students who shared his view on applicability of statistical concepts to science in general. On 15 January 1901, the Russian citizen Vladislav Josephovich Bortkievich (1868–1931)—known as Dr. habil. Ladislaus von Bortkiewicz (also transliterated as Bortkewitsch)—was appointed as

‘*ausserordentlicher Professor*’ at the Friedrich-Wilhelms-Universität Berlin (after 1945 named Humboldt-Universität zu Berlin) by the Prussian Ministry for Culture and Education. How was it possible that the Prussian administration for science and education appointed an official staff member of the civil service in an Imperial Russian ministry? Who was this young scientist, where was he trained and what had he done of importance? The letters of the administration related to LvB and other relevant documents are saved in the archive of Berlin University (Archive HU), containing the aforementioned appointment letter from the Prussian Ministry for Education. But the personal papers of LvB, including hundreds of letters, his class notes and his manuscripts, are not in Berlin. The Bortkiewicz papers are saved in Uppsala (Archive Uppsala). We will later explain why this happened. Independently from archival sources, the life, destiny and fate of LvB were always linked to Europe. LvB, by his training, his mind and his vision, was a true European scholar and one of the most respectable founders of modern statistical science. In this paper, we would like to demonstrate how LvB has contributed to the development of statistics by a cross-disciplinary view of the sciences. We start with a sketch of his life and scientific growth in Section 2 and continue by describing his courses in Section 3. His network of friends and colleagues is described in Section 4. After giving an overview on his numerous publications in Section 4.5, we discuss LvB and the transformation problem in Section 4.6 and his stochastic thinking and his influence on modern statistics in Section 5. Finally, we summarise our results in Section 6.

## 2 From St. Petersburg to Berlin—The Ways of Education of LvB

Ladislaus von Bortkiewicz was born in the Russian imperial city of St. Petersburg (see [1]) on 7 August 1868 into a Polish family. In the Russian language, his name reads, Vladislav Josephovich Bortkievich (BSE, vol. 5, p. 605). His father was Joseph Ivanovitch Bortkiewicz, a military man and an instructor teaching artillery and mathematics. His mother was Helene von Rokicka. LvB had two sisters, and his beloved Helene had the same name as their mother (Sheynin, 1970, p. 318). He studied law at St. Petersburg University for eight semesters. At that time, the education system of Imperial Russia successfully applied the principle of ‘*komandirovka za granicu*’ (travelling abroad, to foreign countries), that is, the mission of young academic researchers to universities outside of Russia, in most cases to western European countries, especially to France and Germany. This principle was described by Pelageja Jakovlevna Kochina (1899–1999) who studied the career paths of the Russian disciples of the mathematician Karl Weierstrass (1815–1897) in her biography about LvB (Kochina, 1985). The same principle was also applied to ophthalmologists and physicists who studied with Hermann von Helmholtz (1821–1894). After studying at Western European universities, most of these post-doc students obtained doctoral degrees there and later became professors at Russian universities. LvB chose the University in Göttingen, to study with Wilhelm Lexis (1837–1914) who was one of the most prominent economists and statisticians at that time. He finished his dissertation on 2 August 1892 and received his Doktor-Diplom the following year on 6 February 1893 after his thesis was published (LvB, 1893; Figure 1).

Further studies in economics and statistics led him to Straßburg, Alsace (today Strasbourg), where he worked together with Georg Friedrich Knapp (1842–1926), another prominent statistician at that time. Between 1871 and 1918, Alsace, and therefore Straßburg, belonged to Prussian Germany and consequently had a Prussian university. The Habilitation (see [2]) of LvB was finished on 2 March 1895, and LvB became a ‘*Privatdozent*’. In Straßburg, he was a contemporary of A. A. Chuprov (1874–1926) who was also a disciple of G. F. Knapp. However, just being a *Privatdozent* does not pay well, is not really exciting and, even today in Germany, is not

1894. 4837

.10

# Die Mittlere Lebensdauer.

Die Methoden ihrer Bestimmung und ihr Verhältnis  
zur Sterblichkeitsmessung.

---

## Inaugural-Dissertation

zur

Erlangung der philosophischen Doktorwürde

eingereicht bei der

Philosophischen Fakultät der Georg-August-Universität zu Göttingen

von

### Ladislaus von Bortkewitsch

aus St. Petersburg.



Jena.

Gustav Fischer.

(1893).

Figure 1. The dissertation of LvB, 1893.

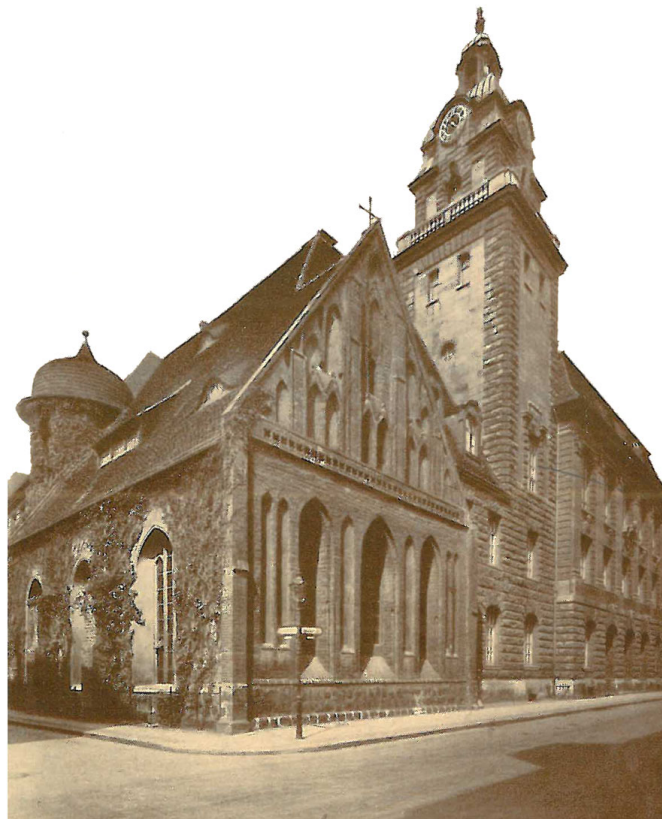
equivalent to a real professorship. Other prominent examples of this non-incentive-based German university system include Johann von Neumann (1903–1957), who was a Privatdozent at the Berliner Universität. In brief: It was not Straßburg, but it was his Privatdozent status that made LvB change locations. After 7 years in Germany, LvB returned to St. Petersburg to look for an academic position. He was offered an appointment in the civil service of the Russian ministry of transport, and thanks to Aleksandr Ivanovich Chuprov (1842–1908)—the father of Aleksandr Aleksandrovich Chuprov (1874–1926)—he also taught statistics at the Aleksandrovskij Lyceum (Oscar Sheynin, 1996, p. 38; Sheynin, 2005). LvB achieved excellent work in St. Petersburg, and A. I. Chuprov tried again to find an academic position for him in 1905/1906

but failed and then finally LvB accepted the offer by the Berlin Universität and made the decision to stay in Berlin for the rest of his life.

### 3 Teaching Modern Statistics

LvB lectured statistics, insurance science, mortality and fertility forecasting, mathematics, quantitative economics and mathematical statistics at two academic institutions in Berlin. From 1901 until 1931, he was a professor, and from 1920 onwards, a full professor ('Professor ad personam', i.e. Personal Chair), at Berlin University (Friedrich-Wilhelms-Universität zu Berlin), (von Bortkiewicz (1930)). In addition, from 1906 until 1923, he taught at the newly founded Berlin School of Economics (Handels-Hochschule, located in Spandauer Str. 1), where a higher income from teaching was possible (Figure 2).

At Berlin University, each semester, LvB normally offered (the winter semester started in October each year and ended in February the following year, and the summer semester was from April to July) two courses, one lecture and one training seminar ('Übung'). Over a period of 30 years, he offered approximately 120 courses (see [3]), most of the lectures and Übungen he held more on general statistics (once an introduction to statistics); all in all, he offered them 28 times. Another big issue (19 lectures and five training courses) he offered was on population theory and population statistics. Sometimes, he offered these lectures with special consideration to Malthus's theory. He held seven lectures on the mathematical and statistical foundations of insurance science, and 11 special training courses on insurance science and insurance mathematics. He also offered similar lectures and courses at the Berlin School of Economics, but the students there protested against the high mathematical content. Only four



**Figure 2.** *Handels-Hochschule (now School of Business and Economics) with the Heiliggeist Kapelle.*

times, between 1917 and 1920, did he offer courses on mathematical statistics and one training course on mathematical statistics. In winter semester 1915/1916, the class list of his seminar (he called it ‘Statistisches Konservatorium’) contained the young post-doc student from Munich, Emil Julius Gumbel (1891–1966) . Figure 3 displays the original handwritten table of LvB (in Archive Uppsala) on class room attendance. Gumbel (1958) wrote later that this class motivated him to work on extreme value theory.

LvB rarely taught jointly with colleagues, and when he did, it was with the close friend Carl (Karl) Ballod (1864–1931). They offered eight tutorials on socio-economic and business statistics between 1902 and 1914. The Faculty of Economics, Humboldt–Universität zu Berlin, where LvB also taught, is a traditional academic teaching and research institution. It was founded in 1886 as the Economic-Statistical Seminar of the Friedrich-Wilhelms-Universität zu Berlin. The early years were dominated, besides LvB, by economists and statisticians such as Richard Boeckh (1824–1907), Gustav von Schmoller (1838–1917) and Adolph Wagner (1835–1917). In 1904, the chamber of commerce of Berlin decided to establish and to build a school of economics in Spandauer Straße 1, between St. Wolfgang Str. and Anna Louisa Karsch Str. The Heiliggeist Kapelle (Holy Spirit Chapel)—built around 1300—is one of the oldest buildings in Berlin and was integrated into the new building; see Figure 2. On 27 October 1906, the Berlin School of Economics was inaugurated. Ever since that time, economics has been permanently taught at Spandauer Straße 1. Because of the Nazi’s policy, many prominent scientists were forced to leave the university and the school of economics. Among the many emigrants were the rector of the School of Economics (1931–1933) Moritz Julius Bonn (1873–1965), one of the founders of modern insurance science Alfred Maues (1877–1963) and the philosopher Arthur Liebert (1878–1946). The statistician and economist Franz Eulenburg (1867–1943), rector from 1929 to 1930, became a victim of the Nazi persecution.

*Statistisches Konservatorium*  
W.S. 1915/16.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	29.10.	9.11	15.11	22.11	29.11	6.12	13.12	10.1	17.1	24.1	31.1	7.2	14.2	21.2	28.2
1. Baum	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2. Frau Boss	.	.	.	.	.	e	f	.	.	.	.	.	.	.	.
3. ✓ Buske	-	.	.	.	.	.	.	.	.	.	f	.	.	.	.
4. Sil. Geiger	-	.	e	.	.	.	.	.	.	f?	.	.	.	.	e
5. Gumbel, Jr.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6. Sil. v. Harnack	.	.	.	.	.	e	.	.	.	.	.	.	.	.	.
7. Hirschke	-	-	.	.	.	.	.	.	.	.	.	.	f	f	.
8. ✓ Horsten	-	-	.	.	.	.	.	.	.	.	.	.	f	f	.
9. Sil. Hoffmann	.	.	.	.	.	.	.	.	.	.	.	.	f	.	.
10. Kautsky	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11. Klett	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12. Mundt	-	-	.	.	e	f	f	<hr/>							
13. ✓ Sil. Reichardt	-	.	.	.	.	f	.	f	f	f	f	f	<hr/>		
14. Tillmann	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15. Farokilli	-	.	f	f	.	f	.	f	f	f?	f	f	f	f	f

**Figure 3.** Class list of ‘Statistisches Konservatorium’ (i.e. seminar) 1915/1916—‘e’ means excused, and ‘f’ means not in class without excuse (In: Archive Uppsala, Bortkiewicz Papers, box 36).

#### 4 The Network of LvB—Family, Friends and Colleagues

Thanks to different primary and secondary sources, we were able to reconstruct the network of the many relationships LvB had with different scientists and scholars from many European countries. The correspondence in the Bortkiewicz papers in Uppsala altogether contain 991 letters from more than 60 colleagues (63 in total). After his death, several obituaries were written, by, for instance, E. J. Gumbel in the journal ‘Statistisches Zentralblatt’, by colleagues such as Thor Andersson (1869–1935), Hermann Schumacher (1868–1952), Oskar Anderson (1887–1960) and Ferdinand Tönnies (1855–1936), (O. Anderson (1932), T. Andersson (1931), H. Schumacher (1931), F. Tönnies (1931)). The leading intellectual newspaper at that time, the ‘Vossische Zeitung’, published an obituary (Figure 4). An excerpt of this text appears at the beginning of this article.

Some authors reviewing his life have pointed out that LvB was a very ambitious and strong university teacher as well as an intensive reader of the publications of his colleagues. Hermann Schumacher stressed the stereotype that LvB lived together with his sister, living without a

**Ladislaus von Bortkiewicz †.** Gestern abend starb unerwartet der ordentliche Professor und Direktor des Staatswissenschaftlich-Statistischen Seminars an der Universität Berlin, Professor Ladislaus von Bortkiewicz. Bortkiewicz wurde 1868 in Petersburg geboren und studierte dort Jurisprudenz. Nach bestandem Staatsexamen wurde er vom russischen Unterrichtsministerium zur weiteren Ausbildung ins Ausland geschickt. Schon frühzeitig hatte sich der junge Jurist mathematischen und wirtschaftlichen Studien zugewandt und sich schließlich im Laufe der Jahre auf dem Gebiet der mathematischen Statistik und der volkswirtschaftlichen Theorie einen internationalen Ruf erworben. Als Schüler von Legis in Göttingen und Knapp in Straßburg konnte er diese Fähigkeiten soweit fördern, daß er nach kurzem Studium an der Straßburger Universität zum Dr. phil. promovieren konnte. Nach kurzem Aufenthalt in Rußland, wo er am russischen Verkehrsministerium tätig war, wurde er nach Deutschland zurückberufen und übernahm im Jahre 1901 eine außerordentliche Professur an der Berliner Universität. 1920 wurde er dann zum ordentlichen Professor ernannt. Eine ganze Reihe bedeutender statistischer Arbeiten sind aus der Feder von Bortkiewicz geflossen. Zu den bekanntesten gehört die 1898 in Leipzig erschienene Schrift über „Das Gesetz der kleinen Zahl“ und die 1893 in Jena veröffentlichte Abhandlung über die „Mittlere Lebensdauer“. Bortkiewicz war ein Meister der Wahrscheinlichkeitsrechnung. Die Beherrschung dieser Wahrscheinlichkeitsrechnung, die er auf Versicherungswissenschaft und Bevölkerungslehre genau so anwandte, wie auch auf gewisse naturwissenschaftliche Gebiete, besonders auf die radioaktive Strahlung, trug dazu bei, Bortkiewicz den Ruf als einen der fähigsten Statistiker der Welt zu schaffen. Seine Berufung zum Mitglied zahlreicher Akademien der Wissenschaften des In- und Auslandes zeugte davon. K. O c k.

Figure 4. Obituary, in the ‘Vossische Zeitung’, 16 July 1931.

family because he wanted to serve the sciences only. On his characteristic way of working, Gumbel wrote:

He presented each problem from all sides with extreme thoroughness and patience after an extensive study of literature. This multiple foundation makes the solution unassailable, but the reader can trace no single line from premises to conclusion: the central line of thought is entwined with numerous sidelines and extensive polemics, especially on matters of scientific priority.’ (Gumbel, 1968, p. 130)

#### 4.1 *Family*

LvB was born in a wealthy Polish family in St. Petersburg, in that time the capital of the Russian Empire. His father Joseph von Bortkiewicz died in summer 1914, shortly before the outbreak of World War I (WWI). It was during this summer of 1914 that LvB visited St. Petersburg for the last time; he never returned because of the political changes in Russia. The Russian intelligentsia made sure that children learned French and German as second languages to read and write fluently; consequently, his family motivated LvB and his younger sister Helene to study in Germany. LvB’s sister Olga married and died of cancer later in Russia in December 1917. Her death was announced by the second sister Helene who had, because of WWI, communicated with LvB in Berlin from St. Petersburg via Thor Andersson in Sweden. Helene von Bortkiewicz (3 August 1870 St. Petersburg to 29 October 1939 Berlin) was a remarkable woman; she was one of the first female students of mathematics to attend Women’s Courses (Vysshiiye zhenskiiye kursy—Higher Women’s Courses) in St. Petersburg, where she received a very good scientific training. These Women’s Courses were opened in 1878 with support from scientists of the University and of the Russian Academy of Science (Kochina, 1988, pp. 45–47). Helene von Bortkiewicz then became one of the Russian mathematics students at the University of Göttingen. She took classes from David Hilbert (1862–1943) and Felix Klein (1849–1925) (see Archive of the Göttingen University, Tobies, 1991/1992, pp. 156, 158, 165–166). Back in Russia, Helene von Bortkiewicz published papers in Russian journals, but the situation was not comfortable for her because the only widely accepted professions for women were as a physician or a teacher in a girls school. From 1910 until summer 1914, she lived with her brother. She travelled together with him to St. Petersburg in the summer of 1914 but did not return with him to Berlin. After the outbreak of WWI, she stayed in St. Petersburg and then in fact worked as a teacher of mathematics and languages. After the first Russian revolution, in February 1917, she became a staff member in a St. Petersburg bank. After the October revolution 1917, she moved, with the help of Thor Andersson, to Berlin where she lived from 1919 until 1931 in her brother’s apartment in Berlin-Halensee (Johann Sigismundstr. 2). We can only speculate whether she worked scientifically with LvB, or whether she ran her brother’s household. After the death of her brother in 1931, she encountered serious financial problems (see [4]) and finally had to give up the apartment and to move to Berlin-Steglitz (see Archive HU, personal file LvB, Bd. 1, Bl. 22R, Bl. 24). This move to a more modest accommodation must also have been the reason that the Bortkiewicz papers have found shelter in Sweden.

#### 4.2 *Friends*

Among LvB’s friends, one of the first, was the Swedish economist and statistician Thor Andersson (1869–1935). Thor Andersson was not only an economist and statistician, but he was also an entrepreneur and publisher. He founded and edited the journals ‘Nordisk statistisk tidskrift’ and later the ‘Nordisk Statistical Journal’ (Sjöström, 2002, p. 195). He invited his friend and colleague LvB to publish in his journals and often visited Berlin. It was his

idea to start the ‘LvB Collected Papers’ project, which he unfortunately could not finish. His oldest friend, dating back to LvB’s student years in St. Petersburg, was the mathematician, statistician and economist Aleksandr Aleksandrovich Chuprov (1874–1926). Aleksandr Chuprov received his doctoral degree in 1896, and as a post-doc, he stayed until 1902 at several German universities. First, he went to Berlin where he visited Adolph Wagner (1835–1917) and then moved to Straßburg, where he studied with Georg Friedrich Knapp. In 1902, he defended his dissertation in economics (Staatswissenschaft) and also passed his master examination at the Faculty of Law of Moscow University. From 1902 to 1917, he taught statistics at the St. Petersburg Polytechnical Institute and was always in close contacts with LvB. In May 1917, he visited Sweden and, with the October Revolution, became an emigré. First, he lived in Stockholm and later in Oslo. In 1920, he moved to Germany, desperately looking for an academic position. For 5 years, he lived in Dresden and taught in Prague, at the Russian Institute. In the LvB Papers in the Uppsala Archive, 125 letters are kept, detailing concerning these almost parallel career paths (Sheynin, 2005). One of the disciples of A. A. Chuprov was the statistician Oskar (Nikolaevich) Anderson (2 August 1887 Minsk to 12 February 1960 Munich). He studied mathematics and physics first in Kazan, and then statistics in St. Petersburg, where in 1912 he defended his thesis (on correlation analysis). From 1910 onwards, he worked with Chuprov and LvB. From September 1912 to October/November 1917, he was employed as an instructor (teacher) in the Higher Commercial School in Lesnoe (near St. Petersburg), where he taught political economy, commercial geography and jurisprudence (Sheynin, 1996, p. 59). After the revolution in October 1917, he emigrated, first teaching in Kiev, and then in the Higher Commercial School in Varna (Bulgaria), where he lived from 1924 to 1942. During WWII, in 1942, he became a full professor of statistics in Kiel, and after 1947, he was a professor in Munich (see ‘*Metrika*’ 3 (1960), pp. 89–94). Among his papers, the one titled ‘Über die Anwendung der Differenzenmethode bei Reihenausgleichungen, Stabilitätsuntersuchungen und Korrelationsmessungen’, (Anderson, O. (1926, 1927)), had great influence. He propagated cross-disciplinary links between humanities and mathematics (Anderson, O. (1935)). This perception was that of LvB and more of his friends and was a rather rare position among statisticians at that time, and not only at that time! Carl Ballod (1864–1931) was also a close friend of LvB. He was a statistician and an expert on Russian economy and taught at Berlin University from winter 1900/1901 until summer 1919. Ballod received his doctoral degree in 1892 at the University of Jena, and he made the Habilitation at Berlin University where he became Privatdozent in December 1899. From 1905 until 1914/1919, he was ausserordentlicher Professor at Berlin University and at the same time staff member at the Prussian Statistical Office. The couple Wladimir Savel’evich Woytinsky (12 November 1885 St. Petersburg to 11 June 1960 Washington, DC) and Emma Shadkhan Woytinsky (19 April 1893 Witebsk to April 1968 Washington, DC) became close friends of LvB and Helene vB. This friendship was an unusual one. In their autobiographies (Wl. Woytinsky, 1961, pp. 452–453; Emmy Woytinsky, 1965, pp. 108–110), both described the history of this relationship. Their famous publication ‘*Die Welt in Zahlen*’ (The World in Figures) in seven volumes was published in Berlin between 1925 and 1928. Originally, the series should have been published in both Russian and German, by the publishing house Rudolf Mosse in Berlin. In fact, only two volumes were published in Russian, in 1924 and in 1925, and then it was halted because of the developing situation in the Soviet Union. However, all seven German volumes were published, edited by LvB. Emma Woytinsky called it a ‘marvelous feature of this project’, that he

played (a part) in it. We learned later that he had been the terror of all German publishers, most of whom had ceased to send him their statistical publications for comment. Not that he was mean—actually, he was just the opposite. He was the embodiment of scientific integrity



and honesty; ... The only trouble was that it was extremely difficult to satisfy him, to reach his level. He was called the ‘Pope of Statistics’, also ‘Die Leuchte’ (The Luminary). (Emma Woytinsky, 1965, p. 109)

She finished her description about the collaboration with LvB and the later friendship with him with the observation:

Nobody who knew Bortkiewicz from his behaviour at the university or from his writings, so highly technical that he could never distribute all ten of the reprints he received from a journal, could realize how much wit and fun he had in him when he let the bars down. (Emma Woytinsky, 1965, p. 110)

After the success, Wladimir S. Woytinsky became the head of the small statistical department of the leading trade union organisation (Allgemeiner Deutscher Gewerkschaftsbund) in Berlin. Here, he worked together with a young colleague, Bruno Gleitze (1903–1980), who later in 1946 became the first dean of the newly established economics faculty at Berlin University. Emma and Wladimir Woytinsky described LvB as a warm, friendly, helpful and very generous person, quite the opposite of other descriptions of him (as dry, cold, very strict and dangerous). It is worth mentioning that Wladimir S. Woytinsky also published a remarkable article ‘Limits of Mathematics in Statistics’ (1954).

### 4.3 Colleagues

Among the colleagues of LvB, we have to first mention his teachers Wilhelm Lexis in Göttingen and Georg Friedrich Knapp in Straßburg. LvB was in close and regular contact with both of them and kept almost all of their letters that they had written to him (see Archive Uppsala). When LvB was appointed in 1901, statistics was taught by Richard Boeckh (1824–1907)—the co-founder of the Economic-Statistical Seminar (Staatswissenschaftlich-Statistisches Seminar), established in 1886, and the ‘Altmeister der Berliner Statistik’ (master of Berlin statistics)—and the agricultural statistician August Meitzen (1822–1910). He also had contact with Adolph Wagner (1835–1917) and Gustav von Schmoller (1838–1917). Boeckh and Meitzen had enormous practical experience from their work in statistical offices; the Royal Prussian Statistical Bureau, the Imperial Statistical Office and the Statistical Office of the city of Berlin, where Boeckh was the director from 1875 to 1902. Studying the lecture schedules, we found that until 1910, LvB taught special courses on statistics. After the death of Boeckh and Meitzen, LvB became the only expert on statistics, and he offered the introductory courses. From 1907 until 1922, he was the only professor of statistics at Berlin University. From 1922 to 1928, Rudolf Meerwarth (1883–1946) joined him in teaching economic and business statistics. One of the very few female colleagues of LvB was Charlotte Lorenz (1895–1979). In 1919, she received her doctoral degree on a thesis about the economic situation in Turkey. Later, she became interested in statistics, and she was employed in the Imperial Statistical Office. Her thesis for ‘Habilitation’ was on price indices, and her work was highly acknowledged by LvB who was one of her advisers in 1927 (see Archive HU, Phil. Fak. Nr. 1242, pp. 217–237). LvB underlined in his review that Charlotte Lorenz was willing to learn the mathematical basis and was able to study recent mathematical literature on price indices. LvB highly acknowledged her work. Only in 1937 did Charlotte Lorenz become a professor at Berlin University, teaching mainly economic and business statistics. After 1945, she taught at Göttingen University. Scientific activities received momentum in 1920 when Richard von Mises (1883–1953) was appointed as the first professor of applied mathematics and director of the Institute for Applied Mathematics and Mechanics at Berlin University. Both von Mises and

LvB belonged to the Philosophical Faculty then, and von Mises took over the classes that had a more mathematical touch from LvB. Soon, they had a joint doctoral student. Both belonged to the Berlin Mathematical Society too, and they had much in common. Among the colleagues of the Berlin School of Economics (Handels-Hochschule), we have to mention the founder of modern insurance science, Alfred Manes (1877–1963), and the economist and statistician Franz Eulenburg (1867–1943). Alfred Manes was an economist and editor of a series on insurance mathematics, and in 1919, he published the book ‘Staatsbankrotte’ (national bankruptcy; the third edition came out in 1923). LvB and Manes had common interests in insurance calculations, and LvB published some articles in the journal of the German Association of Insurance Science (Deutsche Gesellschaft für Versicherungswissenschaft) where A. Manes was one of the heads (Koch, 1990). It was the statistician Franz Eulenburg, from 1929 until 1930, rector of the School of Economics, who first had the idea to publish a series ‘Collected Papers’ of LvB (see the letter written by F. Eulenburg to Thor Andersson, spring 1933 in: von Bortkiewicz Papers, Uppsala). Because of the world economic crisis (1929) and the Nazis’ rise, this project failed. In 1933, both A. Manes and F. Eulenburg were dismissed from the School of Economics, and A. Manes also from Berlin University. Whereas A. Manes successfully managed to emigrate in 1935, first to South America, and then later to the USA (University of Chicago), Franz Eulenburg stayed in Berlin. In December 1943, he was arrested by the Gestapo in Berlin and died on 28 December 1943 in a Gestapo prison.

#### 4.4 Disciples and Doctoral Students

In 1926, the Austrian physician Karl Freudenberg [11 October 1892 Berlin to 14 January 1966 Berlin (West)] defended his thesis on statistics in medicine at Berlin University (see Archive HU, Phil. Fak. Nr. 646, Bl. 373–397, Diss. Karl Freudenberg, 12 October 1926). His advisors were LvB and Richard von Mises. LvB also supported the ‘Habilitation’ of Karl Freudenberg, which followed 2 years later (see Archive HU, Habil. Med. Fak. Nr. 1359, Bl. 153–167, Habilitation Karl Freudenberg, 9 June 1928). From 1928 until 1935, Karl Freudenberg was a Privatdozent at the Medical Faculty of Berlin University, and he was the only teacher in medical statistics (Medizinalstatistik). In 1935, he was dismissed (see Archive HU, personal file Karl Freudenberg), and in 1938, he was arrested by the Gestapo; but in 1939, he was able to emigrate to the Netherlands, where he escaped Nazi persecution. As one of the very few German-Jewish emigrants in mathematics and statistics, he returned to Berlin in 1947 and taught medical statistics at Free University Berlin. One of the most famous disciples was Emil Julius Gumbel (18 July 1891 Munich to 10 September 1966 New York), who later followed the ideas of LvB on distributions (Gumbel, 1958). Obviously, E. J. Gumbel met LvB often in Berlin, between 1920 and 1932, and he lived in Berlin regularly during the semester breaks. Gumbel was not only a successful mathematician and statistician at Heidelberg University but also politically very active. As a member of the German League for Human Rights, he became one of the leading individuals to fight against the Nazis before 1933. He published two books against them in the Weimar Republic (Jansen, 1991; Vogt, 1991; Brenner, 2001). As a result, he was forced to leave Germany, emigrating first to France and in 1940 to the USA. In the early 1950s, Gumbel was a guest professor at Free University Berlin, where he again met, among others, Karl Freudenberg. Gumbel recalled LvB and his work repeatedly (Gumbel, 1931, 1968). Thanks to the documents in the Archive of Berlin University, we were able to analyze all the instances where LvB was the advisor of doctoral students. At the Philosophical Faculty, two different doctoral degrees were possible, Dr. phil. and Dr. rer. pol. Between 1920 and 1931, LvB was the advisor of 11 PhD projects, which led to Dr. phil. In addition, he was the advisor of 29 students who received the degree Dr. rer. pol. Six of his 11 PhD students had to go into

exile because of the Nazi regime. We have already mentioned Karl Freudenberg, and another student was Karl Kost who received a degree in 1926, who also then emigrated to Argentina where he became a novelist. Raimund Goldschmidt (b. in 1904) who received a doctoral degree in 1928 emigrated later to the USA where he published, as Raymond W. Goldsmith, many articles and papers. Another famous doctoral student of LvB was Wassilij Leontief (5 August 1905 Munich to 5 February 1999 New York). He received his doctoral degree in December 1928 (see Archive HU, Phil. Fak. Nr. 678, Bl. 135–197), and his advisors were LvB and Werner Sombart. LvB wrote a long reference about Leontief's thesis, at six pages (see Archive HU, Phil. Fak. Nr. 678, Bl. 156–158R), where LvB strongly argued that this young man from St. Petersburg (at the time Leningrad) was highly talented in statistics as well as in economics. After his studies in Berlin, Leontief obtained an assistant position in Kiel, and thanks to a fellowship, he was able to go to the USA after 1933, where he worked very successfully. In 1973, he was awarded a Nobel Prize in Economics. Another student was Miron Kantorowitsch (b. 1895 Minsk) who lived in Germany from 1919 to 1933. He received his doctoral degree in 1930, and from 1934 to 1938, he worked as a demographer in London and published in the *Journal of the Royal Statistical Society*. In 1938, he emigrated to the USA where he became an acknowledged statistician and an expert on the Soviet Union and Soviet demography (Tolts, 2012), and he changed his name to Myron Kantorowicz, later Myron K. Gordon. Also, Harald von Waldheim received his doctoral degree in 1930. He was a disciple and an assistant of Alfred Manes. Like him, Harald von Waldheim also had to go into exile. LvB was a member of the German Society (Association) of Insurance Sciences, and he belonged to the founding members of the German Statistical Society (Grohmann *et al.* (2011), pp. 227–229). He was also a member of the ISI in Brussels. He was less visible in the UK, possibly because of a controversy with Karl Pearson (1857–1936) (Porter, 2005). Shortly before his death, he was invited to give a key lecture of the Annual Meeting of the American Statistical Society in 1930. He cancelled this trip because of health complications.

#### 4.5 The Publication Activities of LvB

LvB's list of publications is long, and thanks to T. Anderson (1931) and Gumbel (1931, 1968), three bibliographies exist. The publications of LvB can be separated into three groups. The first group includes articles in journals like 'Allgemeines Statistisches Archiv' (the journal of the German Statistical Society), where he published four papers until 1915, and the 'Nordisk Statistical Journal', where he published articles between 1922 and 1930. The second group of publications includes articles in encyclopaedias and the most famous is his article 'Anwendungen der Wahrscheinlichkeitsrechnung auf Statistik' (Application of the Theory of Probability on Statistics) in the 'Mathematische Encyclopedie' (1901). The third group includes his reviews, among others in the 'Deutsche Literaturzeitung'. As Emma S. Woytinsky remembered (1965, p. 109): 'We learned later that he had been the terror of all German publishers, most of whom had ceased to send him their statistical publications for comment.' In the mid-1920s, several publishers were afraid of LvB and his critical reviews, and consequently, he did not publish reviews any more. When one re-reads his various publications, one should do so with the following aspects in mind: the role of mathematics in his work, the under-representation of mathematical statistics in Germany and the extraordinary role that LvB and E. J. Gumbel had in this field (Grohmann *et al.*, 2011, p. 16, 23f, 81, 138–139 and p. 141), LvB's publications on theory of probability and radioactivity, his discussions on Karl Pearson (Quine and Seneta, 1987) and finally the links to his disciple E. J. Gumbel and his book 'Statistics of Extremes' (1958). Gumbel, who followed LvB in studying the Poisson distribution, had well described the different fields that LvB was working on:

Besides classical economics, the work of Bortkiewicz covered population statistics and theory, actuarial science, mathematical statistics, probability theory, mathematical economics, and physical statistics—fields separate in content but analogous in methodology. He contributed to the process of consolidating each of these disciplines and did classic work in mathematical statistics. (Gumbel, 1968, p. 128)

As S. Hertz suggested, ‘A systematic study of Bortkiewicz as a statistician would throw valuable light on this crucial period in the history of mathematical statistics.’ (Hertz, 2001, p. 276)

#### 4.6 LvB and the Transformation Problem

The transformation problem deals with the question of how to transform the value of goods into prices of production. The value of goods is measured for instance in units of time of labour and is in its simplest form decomposable as

$$W = c + v + m,$$

where  $W$  denotes the value,  $c$  the capital invested in production,  $v$  the circulating or variable capital and  $m$  the surplus value (Marx, 1867, *Das Kapital*, Band 1). In the third volume of ‘*Das Kapital*’ (containing thoughts actually prior to 1867, the publication date of the first volume), an additional assumption on the profit rate  $m/(c + v)$  is added: it is assumed to be constant across all industry branches. At this point, we would like to invoke a famous quote by Schumpeter:

By far LvB’s most important achievement is his analysis of the theoretical framework of the Marxian system, much the best thing ever written on it and, incidentally, on its other critics. (Schumpeter, 1932, 2nd. ed. 1956, p. 303)

What, in terms of mathematical symbols, is the transformation problem? What was LvB’s contribution to it? Let us consider as in LvB (1907) a simple three-sector economy producing three goods: investment goods, foodstuffs and luxury goods. The original Marx presentation is based on writing the price of good  $W_i$  as  $P(W_i) = l_i \cdot x_i$ , with  $l_i$  the labour costs and  $x_i$  (unknown) coefficients. The price of  $W$  is a sum of the price of capital and that of the surplus:

$$P(W) = P(C) + P(M)$$

If  $p$  is the average profit rate, then  $P(M) = p \cdot P(C)$  and with  $x_0 = 1/(1 + p)$

$$P(C) = x_0 \cdot P(W)$$

Combining the equations leads to

$$P(C_i) = x_0 \cdot l_i \cdot x_i, \quad i = 1, \dots, 3 \quad (1)$$

If we denote the proportion of goods  $j$  to produce  $i$  as  $q_{ji}$ , then (1) can also be written as

$$P(C_j) = \sum q_{ji} \cdot x_0 \cdot l_i \cdot x_i, \quad i = 1, \dots, 3, \text{ for all } j \quad (2)$$

Putting (1) and (2) together will yield three equations:

$$x_0 \cdot l_j \cdot x_j = \sum_{i=1}^3 q_{ji} \cdot x_0 \cdot l_i \cdot x_i, \quad \text{for all } j \quad (3)$$

with four unknowns  $(x_0, x_1, x_2, x_3)$ . Marx left open how to tackle this simple algebraic problem. Among the first ideas to complement this set of equations was given by Mühlpfordt (1893) who proposed that the sum of the values (sum of  $l_i$ ) should equal the sum of the prices (sum of  $l_i \cdot x_i$ ). Although by writing it down like this, he was unable to express it mathematically (Quaas, 1991). It was LvB later in 1907 who proposed a solution to the transformation problem. Let us follow the outline of Quaas (1991) and define the  $(3 \times 3)$  matrix  $A$  as  $[c \ v \ 0]$ , where  $c = (c_1, c_2, c_3)^\top$  denotes fixed capital and  $v = (v_1, v_2, v_3)^\top$  denotes variable capital. In addition, we have the surplus  $m = (m_1, m_2, m_3)^\top$ . Then the first three equations of (3) are as follows:

$$(1 + p) \cdot A \cdot x = \text{diag}(1^\top c, 1^\top v, 1^\top m) \cdot x, \quad x \in \mathbb{R}^3 \quad (4)$$

We can see this by setting  $c_i = q_{i1} \cdot l_1$ ,  $v_i = q_{i2} \cdot l_2$ ,  $m_i = q_{i3} \cdot l_3$  and observing that because the three-sector economy is circular:  $l_1 = 1^\top c$ ,  $l_2 = 1^\top v$ ,  $l_3 = 1^\top m$ . These transformation conditions are under-determined. Thus, Bortkiewicz simply set the sum of surplus equal to the sum of profits by setting  $x_3 = 1$ . This way, the sum over the output prices on the right-hand side is equal to the costs plus the average profit rate  $p$ . Moreover, all costs of production are considered with cost of capital, and all prices are determined. Herein, Bortkiewicz provided ‘the’ solution to the transformation problem.

## 5 v. Bortkiewicz and his Influence on Modern Statistics

At the dawn of the 20th century, the mindset about statistics and probability theory as applied to natural or social sciences can be described in the clever way that the Viennese mathematician Emanuel Czuber (1851–1925) used in 1898:

An der Schwelle der Wahrscheinlichkeitstheorie steht eine Reihe von Begriffen, welche der Mathematik fremd sind, und über deren Deutung die Discussion nicht abgeschlossen ist, ja heute lebhafter geführt wird denn je.

(At the border of theory of probability we find a number of concepts which are alien to mathematics and their interpretation has not been finished yet and even needs to be discussed more than ever.)

The allocation of the statistical science as a non-mathematical discipline has also been underlined by David Hilbert (1862–1943). In 1900 in Paris, he presented his 23 ‘open problems’. Problem number 6 was as follows:

Mathematische Behandlung der Axiome der Physik. Durch die Untersuchungen über die Grundlagen der Geometrie wird uns die Aufgabe nahegelegt, nach diesem Vorbilde diejenigen physikalischen Disziplinen axiomatisch zu behandeln, in denen schon heute die Mathematik eine hervorragende Rolle spielt; dies sind in erster Linie die Wahrscheinlichkeitsrechnung und die Mechanik. Was die Axiome der Wahrscheinlichkeitsrechnung angeht, so scheint es mir wünschenswert, daß mit der logischen Untersuchung derselben zugleich eine

strenge und befriedigende Entwicklung der Methode der mittleren Werte in der mathematischen Physik, speziell in der kinetischen Gastheorie Hand in Hand gehe. (Hilbert, 1971, p. 47)

Hilbert classified here the theory of probability as a part of physics that was to be seen as a future mathematical sub-discipline. As the probabilistic tools of the statistical discipline were mostly used in physics (promoted by papers of Einstein, Maxwell and Boltzmann), Hilbert more likely classified statistics not as a mathematical discipline but rather as a part of physics. In fact, it was the British School around Karl Pearson (1857–1936), William S. Gosset (1876–1937), Ronald A. Fisher (1890–1962), Jerzy Neyman (1894–1981) and Egon Sharpe Pearson (1895–1980) who developed the branch of mathematical statistics. In contradiction, statistics in Germany leaned more towards a descriptive analysis of data with a preference for a social economic context. It was Ladislaus von Bortkiewicz who in several books and papers promoted the probabilistic approach, for example, on the Poisson and exponential distributions and on the distributions of runs (iterations). In his book on ‘Die Iterationen’ (von Bortkiewicz (1917)) in the second chapter on ‘Grundsätzliches aus der Wahrscheinlichkeitstheorie’, he gave a clear exposition of the mathematical foundations of probability theory. In a somewhat sneering tone, he comments that Marbe (1916–1919) was ‘touchingly clumsy’ in his quantitative description of a simple coin-flipping experiment (Figure 5).

In mathematischer Hinsicht ist Marbe auch sonst von einer, man möchte beinahe sagen, rührenden Unbeholfenheit. (see figure inset)

Karl Marbe (1869–1953), a professor of psychology, argued that a run of male births leads *per se* to an increased probability of a female birth. He employed a ‘nature argument’ on equalising the sex ratio. Bortkiewicz showed, however, that Marbe’s mathematics was wrong.

LvB became well known also for his precise calibration of real data. In 1898, he published the book ‘Das Gesetz der kleinen Zahlen’ (The Law of Small Numbers) in which he first noted that events with low frequency in a large population follow a Poisson distribution (Quine and Seneta, 1987; Haight, 1967). The two data sets he considered were the Prussian horse-kick data and child suicides. The horse-kick data give the number of soldiers killed by being kicked by

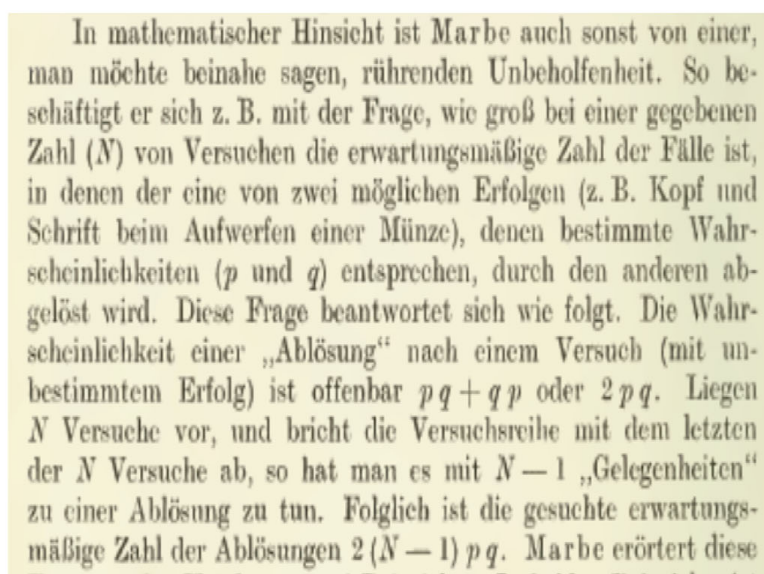
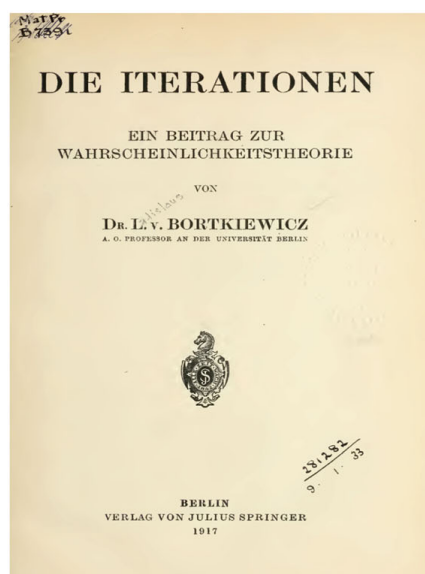
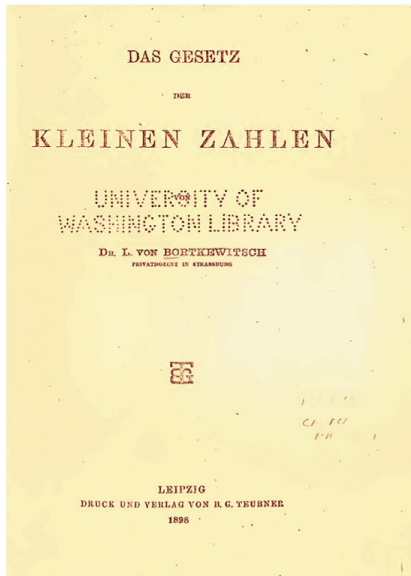


Figure 5. ‘Die Iterationen’ and an excerpt of ‘Die Iterationen’.



	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	—	2	1	—	2
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

Figure 6. Law of small numbers (left) and Prussian horse-kick data (right). The columns are the years 1875–1894, and rows are the corps numbers.

Table 1. Results for a subset of Prussian horse-kick data.

$k$	$nk$	$p$	$\hat{p}$	exp	$\chi^2$
0	109	0.545	0.54335	108.670	0.00100
1	65	0.325	0.33144	66.289	0.02506
2	22	0.110	0.10109	20.218	0.15705
3	3	0.015	0.02056	4.111	0.30025
4	1	0.005	0.00313	0.627	0.22201
	200			199.915	0.70537 $\sim \chi^2_4$

a horse each year in each of the 14 cavalry corps of the Prussian Army over a 20-year period (Figure 6).

The Poisson distribution was first derived in 1837 by Siméon D. Poisson (1781–1840) who applied it to the decisions of juries. Yet, Poisson’s analysis was not regarded as a central piece of statistical data analysis. It was not until LvB’s publication in 1898 with his convincing analysis of the Prussian horse-kick data that this distribution entered the standard canon. As a consequence, it was suggested that the Poisson distribution should have been named the ‘Bortkiewicz distribution’. Let us just check his analysis. For simplicity, we will take a subset of 200 observations as it is presented on the Internet. The maximum-likelihood estimator is  $\lambda = 0.61$  and with the Bortkiewicz distribution: we arrive at Table 1 showing a remarkably good fit indeed. LvB inspired his students to use mathematical techniques for data calibration. His work on Prussian horse kicks and child suicide data promoted in his book on the law of small numbers was trend-setting not only in Germany. LvB can therefore be seen as a founder of modern econometric and statistical education in Germany and beyond.

## 6 Conclusion

As we have seen, the life, destiny and fate of LvB were always linked to Europe, from the North, including the Scandinavian countries, to the South, where in Varna in Bulgaria a friend

and colleague from St. Petersburg worked. LvB, by his training, his mind and his vision, was a European intellectual, a European scholar and one of the respectable founders of modern statistical science. Wladimir Woytinsky, who had the luck and the honour to be trained by LvB in private courses, underlined that LvB had his own philosophy on statistics and measurement and that LvB highly acknowledged the role of mathematics (Wl. Woytinsky, 1961, p. 453). E. J. Gumbel, another disciple and follower of LvB working on distributions, evaluated the work and research results, LvB succeeded, in his posthumously published article in 1968, and he concluded it with the words:

Four of his contributions are decisive: the proof that the Poisson distribution corresponds to a statistical reality; the introduction of mathematical statistics into the study of radioactivity; the inception of the statistical theory of extreme values; and the lonely effort to construct a Marxian econometry. (Gumbel, 1968, p. 130)

Another 45 years later, one could argue like E. J. Gumbel, but furthermore, one should request that people should re-read and study the classic papers, written by LvB, again, at least to get a great degree of stimulation.

## Bibliography

### *Archives*

Archive HU—Archive of the Humboldt University, Berlin.

*Personal files (PA) of LvB, Karl Freudenberg, Alfred Manes, and others; documents related to doctoral degrees and 'Habilitation'; lecture schedules (Vorlesungs-Verzeichnisse)*

Archive of the Göttingen University.

*Documents related to Helene and Ladislaus von Bortkewitsch (sic)*

Archive Uppsala—Department of Manuscripts and Music, Uppsala University Library.

*Bortkiewicz Papers (41 Boxes)*

### *Annotations*

- (1) *Until WWI, August 1914, the city was named St. Petersburg, created by Zar Peter I. (the Great) (1672–1725). From 1914 onwards, the town was named Petrograd; in 1924, it obtained the name Leningrad. In 1991, it was given its original name back.*
- (2) *The 'Habilitation' was introduced at German universities in about 1830. The procedure consists of three elements/steps: a thesis (usually a book), a talk ('Probevortrag') and a lecture ('Probevorlesung'). When a chosen committee at a Faculty agreed on all three steps, the candidate was nominated as a 'Privatdozent'. This entailed the right to teach at the Faculty. A Privatdozent (PD) position was the first and also lowest position in the staff hierarchy of a Faculty. When a PD wanted to move to another university, he (until 1919, only male PDs were allowed at German universities) had to obtain permission from the new Faculty. See the description by Richard Goldschmidt (1960), pp. 52, 66-67.*
- (3) *The sources to analyze his teaching activities are the printed schedules of lectures (in German Vorlesungs-Verzeichnisse), university calendars or course catalogues, of Berlin University, which were published every semester. These schedules of lectures allowed for detailed reconstruction of the teaching activities between 1901 and 1931.*



- (4) *The file of the Humboldt University Archive contains several letters of Helene v. B., to the Prussian Ministry when she claimed some money after the death of LvB. As she was not LvB's wife, but rather his sister, the claims could not be adequately responded. But the Prussian Minister of Education admitted her once a small sum (see Archive HU, personal file LvB, Bd. 1, Bl. 15–30).*

## Acknowledgements

The research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 'Economic risk'. We would like to give sincere thanks to our colleagues who discussed with us several aspects of our article, and thanks to the reviewers for their work. We want to give special thanks to the archives, especially Uppsala, Berlin and Göttingen. Warm thanks to Leslie Udvarhelyi and Awdesch Melzer for the practical and technical help that brought this article to its final version for publication.

## References

- Anderson, O. (1926, 1927). Über die Anwendung der Differenzenmethode bei Reihenausgleichungen, Stabilitätsuntersuchungen und Korrelationsmessungen. *Biometrika*, **18**, 19.
- Anderson, O. (1935). *Einführung in die mathematische Statistik*. Wien: Julius Springer Verlag.
- Anderson, O. (1932). Ladislaus von Bortkiewicz (1868–1931) (obituary). *Zeitschrift für Nationalökonomie*, **3**, 242–250.
- Andersson, T. (1931). Obituary of Ladislaus Josephowitsch Bortkiewicz. *Nordic Statistical Journal*, **3**, 9–26.
- Brenner, A.D. (2001). Emil J. Gumbel. Weimar German Pacifist and Professor. Boston: Brill.
- BSE—Bol'shaja Sovjetskaja Encyclopedia (Great Soviet Encyclopedia, in Russian). (1949), 2nd ed., Vol. 5. p. 605 Moscow: Goz. Izd. (State Publisher).
- Goldschmidt, R.B. (1960). *In and Out of the Ivory Tower*. Seattle: Univ. of Washington Press.
- Grohmann, H., Krämer, W. & Steger, A. (eds). (2011). *100 Jahre Deutsche Statistische Gesellschaft*. Berlin, Heidelberg: Springer Verlag.
- Gumbel, E.J. (1931). Nachruf auf Ladislaus von Bortkiewicz. *Deutsches statistisches Zentralblatt*, **23**(8), 231–236. (bibliography of LvB, Columns 233–236).
- Gumbel, E.J. (1958). *Statistics of Extremes*. New York: Columbia Univ. Press. (A Russian translation came out in Moscow in 1965 with a foreword by Boris V. Gnedenko (1912–1995)).
- Gumbel, E.J. (1968). Bortkiewicz, Ladislaus von. In *International Encyclopedia of the Social Sciences* 2, pp. 128–131. New York: Macmillan Publishers. (Bibliography of LvB, 130–131).
- Härdle, W.K. & Vogt, A.B. (2014). Ladislaus von Bortkiewicz statistician, economist, and a European intellectual. *SFB649 Discussion Papers 2014-015* (With full set of figures) Available at <http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2014-015.pdf>. Accessed 14 February 2014.
- Haight, F.A. (1967). *Handbook of the Poisson Distribution*. New York: Wiley.
- Hertz, S. (2001). Ladislaus von Bortkiewicz. In *Statisticians of the Centuries*, Eds. C.C. Heyde & E. Seneta, pp. 273–277. New York: Springer.
- Hilbert, D. (1971). Mathematische Probleme. In *Die Hilbertschen Probleme von D. Hilbert*, Ed. P.S. Alexandrov, pp. 22–80. Leipzig: Akademische Verlagsgesellschaft Geest & Portig.
- Jansen, C. (1991). *Emil Julius Gumbel. Porträt eines Zivilisten*. Heidelberg: Verlag Das Wunderhorn.
- Koch, P. (1990). Alfred Manes. In *Neue Deutsche Biographie* 16 22f. Berlin: Duncker & Humblot.
- Kochina, P.J. (1985). *Karl Wejerstrass*. Moskva: Nauka, (Russian).
- Kochina, P.J. (1988). *Nauka, Ljudi, gody. Vospominanija i vystupenija*. Moskva: Nauka, (Russian). (memories of the woman mathematician P. Ja. Kochina (1899–1999)).
- Marbe, K. (1916–1919). *Die Gleichförmigkeit in der Welt: Untersuchungen zur Philosophie und positiven Wissenschaft*. München: Beck'sche Verlagsbuchhandlung.
- Marx, K. (1867). *Das Kapital*, 39th ed., Vol. 1 Berlin: Karl Dietz Verlag. 2008.
- Mühlpfordt, W. (1893). Preis und Einkommen in der privatkapitalistischen Gesellschaft. (*Diss.*) Königsberg, Hartungsche Buchdruckerei.

- Necrologes, obituaries, written by: O. Anderson, Th. Andersson, E. J. Gumbel, H. Schumacher, F. Tönnies Bibliographies of his work are published in: Gumbel, columns 233–236, (1931), Gumbel, 130–131, (1968) and Anderson (1931), 2, 478.*
- Ock, K. (1931). (obituary on LvB) in: *Vossische Zeitung*, (16. Juli 1931) Newspaper clipping in: *Archive HU, UK 347, PA LvB, Bd. 2, Bl. 54.*
- Porter, T.M. (2005). *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton: Princeton Univ. Press.
- Quaas, F. (1991). Wolfgang Mühlpfordt-ein Vorgänger von Bortkiewicz? Zu den theoretischen Quellen des sogenannten Transformationsproblems. *Munich Personal RePEc Archive # 20348.*
- Quine, M.P. & Seneta, E. (1987). Bortkiewicz's Data and the Law of Small Numbers. *Int. Stat. Rev.*, **55**, 173–181.
- Schumacher, H. (1931). Ladislaus von Bortkiewicz (obituary). *Allgemeines Statistisches Archiv*, **21**, 573–576.
- Schumpeter, J.A. (1932). (1956) Ladislaus von Bortkiewicz: 1868–1931. In *Joseph A. Schumpeter, Ten Great Economists from Marx to Keynes*, pp. 302–305. (1932) New York: Oxford Univ. Press. First published in Volume 42 of the *Economic Journal*, 338–340.
- Sheynin, O.B. (1970). Bortkiewicz (or Bortkewitsch), Ladislaus (or Vladislav) Josephowitsch. In *Dictionary of Scientific Biography*, Vol. 2, pp. 318–319. New York: Scribner's.
- Sheynin, O. (1996). *A. A. Chuprov: Life, Work, Correspondence*. Göttingen: Vandenhoeck & Ruprecht, (Reihe Angewandte Statistik und Ökonometrie, Heft 38). (Transl. from the publ., Moscow 1990 (in Russian), by O. Sheynin).
- Sheynin, O. (2005). V. I. Bortkevich, A. A. Chuprov. *Perepiska (1895–1926) (in Russian)*. *Sostavitel' O. B. Sheynin. (Correspondence, compiled by O. B. Sheynin)*, Berlin.
- Sjöström, O. (2002). *Svensk statistik historia. En undanskynd kritisk tradition*. Ö-rlinge: Hedemora Gidlunds Förlag. (thanks to Per Wisselgren for this literature).
- Steger, A. (2011). Wie alles begann. In *100 Jahre Deutsche Statistische Gesellschaft*, Eds. H. Grohmann *et al.*, pp. 3–18. Berlin, Heidelberg: Springer, 2011, (Quote p. 3, originally in: Zahn, F. (ed.) *Die Statistik in Deutschland nach ihrem heutigen Stand. Ehrengabe für Georg von Mayr*. München, Berlin: Schweitzer, 1911, vol. 1 (2 vols.))
- Tobies, R. (1991/1992). Zum Beginn des mathematischen Frauenstudiums in Preußen. *NTM*, **28**, 151–172. (Zeitschrift für Geschichte der Naturwissenschaften, Technik und Medizin).
- Tönnies, F. (1931). Ladislaus von Bortkiewicz (1868-1931). *Kölner Vierteljahreshefte für Soziologie (KVS)*, **10**(1931/32), 433–446.
- Tolts, M. (2012). A Demographer in Spite of Himself: The Migrant's Destiny of Miron Kantorowicz (Myron K. Gordon). *Paper*, Moscow, (13.11.2012) Available at [https://www.academia.edu/2159718/A\\_Demographer\\_in\\_Spite\\_of\\_Himself\\_The\\_Migrants\\_Destiny\\_of\\_Miron\\_Kantorowicz\\_Myron\\_K.\\_Gordon\\_](https://www.academia.edu/2159718/A_Demographer_in_Spite_of_Himself_The_Migrants_Destiny_of_Miron_Kantorowicz_Myron_K._Gordon_). Accessed 5 December 2013.
- Vogt, A. (1991). *Emil Julius Gumbel. Auf der Suche nach Wahrheit*. Berlin: Dietz Verlag.
- von Bortkiewicz, L. (1893). *Die mittlere Lebensdauer; Die Methoden ihrer Bestimmung und ihr Verhältnis zur Sterblichkeitsmessung* Jena: Fischer Verlag.
- von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen* Leipzig: Teubner Verlag.
- von Bortkiewicz, L. (1952). Value and Price in the Marxian System. *Intern. Economic Papers*, **2**, 5–60. (first in German: Wertrechnung und Preisrechnung im Marx'schen System. In: *Archiv für Sozialwissenschaft und Socialpolitik*, Bd. 23, 1906 und Bd. 25, 1907).
- von Bortkiewicz, L. (1907). On the Correction of Marx's Fundamental Theoretical Construction in the Third Volume of Capital. In *von Böhm-Bawerk, Eugen. Karl Marx and the Close of His System*. Kelley. 1949 (first in German: Zur Berichtigung der grundlegenden theoretischen Konstruktion von Marx im 3. Band des "Kapital". In: *Jahrbücher für Nationalökonomie und Statistik, F. 3 (Folge 3)*, Bd. 34, 1907, 319–335); 197–221.
- von Bortkiewicz, L. (1917). *Die Iterationen. Ein Beitrag zur Wahrscheinlichkeitstheorie* Berlin: Springer Verlag.
- von Bortkiewicz, L. (1930). *Biography (authorized)*. *Reichshandbuch der Deutschen Gesellschaft* Bd. I, Berlin: Deutscher Wirtschaftsverlag, 188 with photo.
- Woytinsky, W.S. (1961). *Stormy Passage. A Personal History Through Two Russian Revolutions to Democracy and Freedom: 1905–1960* New York: Vanguard Press.
- Woytinsky, E.S. (Shadkhan). (1965). *Two Lives in One* New York, Washington: Frederick A. Praeger Publ.
- Woytinsky, W.S. (1954). Limits of mathematics in statistics. *The American Statistician*, **8**(1), 6–10 + 18. (18 Feb. 1954).

[Received February 2014, accepted July 2014]

## A simultaneous confidence corridor for varying coefficient regression with sparse functional data

Lijie Gu · Li Wang · Wolfgang K. Härdle ·  
Lijian Yang

Received: 19 January 2014 / Accepted: 27 June 2014 / Published online: 22 July 2014  
© Sociedad de Estadística e Investigación Operativa 2014

**Abstract** We consider a varying coefficient regression model for sparse functional data, with time varying response variable depending linearly on some time-independent covariates with coefficients as functions of time-dependent covariates. Based on spline smoothing, we propose data-driven simultaneous confidence corridors for the coefficient functions with asymptotically correct confidence level. Such confidence corridors are useful benchmarks for statistical inference on the global shapes of coefficient functions under any hypotheses. Simulation experiments corroborate with the theoretical results. An example in CD4/HIV study is used to illustrate how inference is made with computable  $p$  values on the effects of smoking, pre-infection CD4 cell percentage and age on the CD4 cell percentage of HIV infected patients under treatment.

**Keywords** B spline · Confidence corridor · Karhunen–Loève  $L^2$  representation · Knots · Functional data · Varying coefficient

**Mathematics Subject Classification (2000)** 62G08 · 62G15 · 62G32

---

L. Gu · L. Yang (✉)  
Center for Advanced Statistics and Econometrics Research,  
Soochow University, Suzhou 215006, China  
e-mail: yanglijian@suda.edu.cn

L. Wang  
Department of Statistics, Iowa State University, Ames, IA 50011, USA

W. K. Härdle  
Center for Applied Statistics and Economics (C.A.S.E.),  
Humboldt-Universität zu Berlin, 10099 Berlin, Germany

W. K. Härdle  
Lee Kong Chian School of Business, Singapore Management University, Singapore, Singapore

## 1 Introduction

Functional data, known also as “curve data”, are commonly encountered in biomedical studies, epidemiology and social science, where information is collected over a time period for each subject. Conceptually, such data can be viewed as a simple random sample from the abstract space of functions, see for instance, [Ferraty and Vieu \(2006\)](#), [Manteiga and Vieu \(2007\)](#). For functional data analysis (FDA) approach without nonparametric smoothing, see [Gabrys et al. \(2010\)](#), and the recent comprehensive review in [Horváth and Kokoszka \(2012\)](#). In this, paper we have taken from [Ramsay and Silverman \(2005\)](#) the more convenient view of functional data as discretely recorded observations of independent stochastic processes contaminated with measurement errors.

In many longitudinal studies, repeated measurements are often collected at finite number of time points. If the time points of observation for every subjects are dense and regular, the data are termed dense functional data, see [Cao et al. \(2012a, b\)](#), and [Zhu et al. \(2012\)](#) for theoretical development and real examples of dense functional data. If, on the other hand, data collection is made at few and irregular time points for each subject, the data are frequently referred to as sparse longitudinal or sparse functional data, see [James et al. \(2000\)](#), [James and Sugar \(2003\)](#), [Yao et al. \(2005a\)](#), [Hall et al. \(2006\)](#), [Zhou et al. \(2008\)](#), [Ma et al. \(2012\)](#) for works on sparse functional data. It should be emphasized especially that by “sparse” we mean that the covariate is observed sparsely over a compact interval, not having anything to do with sparsity used in variable selection context such as the popular LASSO method. A crucial condition for sparse FDA is that the time points from all subjects are dense in the data range despite sparsity for any individual subject, see Assumption (A3) in Appendix A that the design density  $f(t)$  has a positive lower bound  $c_f$ , which implies that the sampling frequency is almost uniform for the time covariate.

In longitudinal study, often, interest lies in studying the association between the covariates and the response variable. In recent years, there has been an increasing interest in nonparametric analysis of longitudinal data to enhance flexibility, see e.g., [Yao and Li \(2013\)](#). The varying coefficient model (VCM) proposed by [Hastie and Tibshirani \(1993\)](#) strikes a delicate balance between the simplicity of linear regression and the flexibility of multivariate nonparametric regression and has been widely applied in various settings, for instance, the Cobb–Douglas model for GDP growth in [Liu and Yang \(2010\)](#), and the longitudinal model for CD4 cell percentages in AIDS patients in [Wu and Chiang \(2000\)](#), [Fan and Zhang \(2000\)](#) and [Wang et al. \(2008\)](#). See [Fan and Zhang \(2008\)](#) for an extensive literature review of VCM.

To examine whether the association changes over time, [Hoover et al. \(1998\)](#) proposed the following VCM

$$Y(t) = \beta_0(t) + \mathbf{X}(t)^\top \boldsymbol{\beta}(t) + \varepsilon(t), \quad t \in \mathcal{T}, \quad (1)$$

where  $\mathbf{X}(t) = (X_1(t), \dots, X_d(t))^\top$  are covariates at time  $t$ ,  $\varepsilon(t)$  is a mean zero process, and  $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_d(t))^\top$  are functions of  $t$ . Model (1) is a special case of functional linear models, see [Ramsay and Silverman \(2005\)](#) and [Wu et al. \(2010\)](#).

The coefficient functions  $\beta_l(t)$ s in model (1) can be estimated by, for example, kernel method in Hoover et al. (1998), basis function approximation method in Huang et al. (2002), polynomial spline method in Huang et al. (2004) and smoothing spline method in Brumback and Rice (1998). Fan and Zhang (2000) proposed a two-step method to overcome the computational burden of the smoothing spline method.

For some longitudinal studies, the covariates are independent of time, and their observations are cross-sectional. Take for instance the longitudinal CD4 cell percentage data among HIV seroconverters. This dataset contains 1,817 observations of CD4 cell percentages on 283 homosexual men infected with the HIV virus. Three of the covariates are observed at the time of HIV infection and hence by nature independent of the measurement time and frequency:  $X_{i1}$ , the  $i$ th patient's smoking status;  $X_{i2}$ , the  $i$ th patient's centered pre-infection CD4 percentage; and  $X_{i3}$  the  $i$ th patient's centered age at the time of HIV infection. A fourth predictor, however, is time dependent:  $T_{ij}$ , the time (in years) of the  $j$ th measurement of CD4 cell on the  $i$ th patient after HIV infection; while the response  $Y_{ij}$  is also time dependent: the  $j$ th measurement of the  $i$ th patient's CD4 cell percentage at time  $T_{ij}$ . Wu and Chiang (2000), Fan and Zhang (2000) and Wang et al. (2008) all contain detailed descriptions and analysis of this dataset.

A feasible VCM for multivariate functional data such as the above takes the form

$$Y_{ij} = \sum_{l=1}^d \eta_{il}(T_{ij}) X_{il} + \sigma(T_{ij}) \varepsilon_{ij}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq N_i, \quad (2)$$

where the measurement errors  $(\varepsilon_{ij})_{i=1, j=1}^{n, N_i}$  satisfy  $\mathbf{E}(\varepsilon_{ij}) = 0$ ,  $\mathbf{E}(\varepsilon_{ij}^2) = 1$ , and  $\{\eta_{il}(t), t \in \mathcal{T}\}$  are i.i.d copies of a  $L^2$  process  $\{\eta_l(t), t \in \mathcal{T}\}$ , i.e.,  $\mathbf{E} \int_{\mathcal{T}} \eta_l^2(t) dt < +\infty$ ,  $l = 1, \dots, d$ . The common mean function of processes  $\{\eta_l(t), t \in \mathcal{T}\}$  is denoted as  $m_l(t) = \mathbf{E}\{\eta_l(t)\}$ ,  $l = 1, \dots, d$ . The actual data set consists of  $\{\mathbf{X}_i, T_{ij}, Y_{ij}\}$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq N_i$ , in which the  $i$ th subject is observed  $N_i$  times, the time-independent covariates for the  $i$ th subject are  $\mathbf{X}_i = (X_{il})_{l=1}^d$ ,  $1 \leq i \leq n$ , and the random measurement time  $T_{ij} \in \mathcal{T} = [a, b]$ . The aforementioned data example is called sparse functional as the number of measurements  $N_i$  for the  $i$ th subject is relatively low. (In the above CD4 example actually at most 14). In contrast, for a dense functional data  $\lim_{n \rightarrow \infty} \min_{1 \leq i \leq n} N_i = \infty$ .

For the CD4 cell percentage data, we introduce a fourth time-independent covariate, the baseline  $X_{i0} \equiv 1$ , and denote by  $m_l(t)$ ,  $l = 0, 1, 2, 3$ , the coefficient functions for baseline CD4 percentage, smoking status, centered pre-infection CD4 percentage and centered age, respectively. Figures 2, 3, 4, 5 contain spline estimates of the  $m_l(t)$ ,  $0 \leq l \leq 3$ , and simultaneous confidence corridors (SCC) at various confidence levels.

In previous works the theoretical focus has mainly been on consistency and asymptotic normality of the estimators of the coefficient functions of interest, and the construction of pointwise confidence intervals. However, as demonstrated in Fan and Zhang (2000), this is unsatisfactory as investigators are often interested in testing whether some coefficient functions are significantly nonzero or varying, for which a SCC is needed. Take for instance, Fig. 3, which shows both the 95 and 20.277 % SCC

of  $m_1(t)$  contain the zero line completely, thus with a very high  $p$  value of 0.79723 the null hypothesis of  $m_1(t) \equiv 0, t \in \mathcal{T}$  is not rejected. More details are in Sect. 6.

Construction of computationally simple SCCs with exact coverage probability is known to be difficult even with independent cross-sectional data; see, Wang and Yang (2009) and related earlier work Härdle and Luckhaus (1984) on uniform consistency. Most earlier methods proposed in the literature restrict to asymptotic conservative SCCs. Wu et al. (1998) developed asymptotic SCCs for the unknown coefficients based on local polynomial methods, which are computationally intensive, as the kernel estimator requires solving an optimization problem at every point. Huang et al. (2004) proposed approximating each coefficient function by a polynomial spline and developed spline SCCs, which are simpler to construct, while Xue and Zhu (2007) proposed maximum empirical likelihood estimators and constructed SCCs for the coefficient functions. All these SCCs are Bonferroni-type variability bands according to Hall and Titterington (1988). The idea is to invoke pointwise confidence intervals on a very fine grid of  $[a, b]$ , then adjust the level of these confidence intervals by the Bonferroni method to obtain uniform confidence bands, and finally bridge the gaps between the grid points via smoothness conditions on the coefficient curve. However, to use these bands in practice, one must have a priori bounds on the magnitude of the bias on each subinterval as well as a choice for the number of grid points. Chiang et al. (2001) proposed a bootstrap procedure to construct confidence intervals. However, theoretical properties of their procedures have not yet been developed.

In this paper, we derive SCCs with exact coverage probability for the coefficient functions  $m_l(t)$ ,  $l = 1, \dots, d$ , in (3) via extreme value theory of Gaussian processes and approximating coefficient functions by piecewise-constant splines. The results represent the first attempt at developing exact SCCs for the coefficient functions in VCM for sparse functional data. Our simulation studies indicate the proposed SCCs are computationally efficient and have the right coverage probability for finite samples. Our work parallels Zhu et al. (2012) which established asymptotic theory of SCC in the case of VCM for dense functional data. It is important to mention as well that the linear covariates in Zhu et al. (2012) are time dependent, which does not complicate the problem as they work with dense data instead of the sparse data we concentrate on. Our work can also be viewed as an extension of the univariate longitudinal regression in Ma et al. (2012) to varying coefficient regression, the latter corresponds exactly to the special case of  $d = 1, X_{i1} \equiv 1$ . Theorem 1 of Ma et al. (2012) follows from Theorems 1 and 2 in this paper with some slight modifications.

We organize our paper as follows. Section 2 describes spline estimators, and establish their asymptotic properties for sparse longitudinal data. Section 3.1 proposes asymptotic pointwise confidence intervals and SCCs constructed from piecewise constant splines. Section 3.2 describes actual steps to implement the proposed SCCs. In Sect. 4, we provide further insights into the estimation error structure of spline estimators. Section 5 reports findings from a simulation study. A real data example appears in Sect. 6. Proofs of technical lemmas are in Appendix A.

## 2 Spline estimation and asymptotic properties

For a functional data  $\{\mathbf{X}_i, T_{ij}, Y_{ij}\}$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq N_i$ , denote the eigenvalues and eigenfunctions sequences of its covariance operator  $G_l(s, t) = \text{cov}\{\eta_l(s), \eta_l(t)\}$  as  $\{\lambda_{k,l}\}_{k=1}^\infty$ ,  $\{\psi_{k,l}(t)\}_{k=1}^\infty$ , in which  $\lambda_{1,l} \geq \lambda_{2,l} \geq \dots \geq 0$ ,  $\sum_{k=1}^\infty \lambda_{k,l} < \infty$ , and  $\{\psi_{k,l}\}_{k=1}^\infty$  form an orthonormal basis of  $L^2(\mathcal{T})$ . It follows from spectral theory that  $G_l(s, t) = \sum_{k=1}^\infty \lambda_{k,l} \psi_{k,l}(s) \psi_{k,l}(t)$ . For any  $l = 1, \dots, d$ , the  $i$ th trajectory  $\{\eta_{il}(t), t \in \mathcal{T}\}$  allows the Karhunen–Loève  $L^2$  representation (Yao et al. 2005b):  $\eta_{il}(t) = m_l(t) + \sum_{k=1}^\infty \xi_{ik,l} \phi_{k,l}(t)$ , where the random coefficients  $\xi_{ik,l}$  are uncorrelated with mean 0 and variances 1, and the functions  $\phi_{k,l} = \sqrt{\lambda_{k,l}} \psi_{k,l}$ , thus  $G_l(s, t) = \sum_{k=1}^\infty \phi_{k,l}(s) \phi_{k,l}(t)$ , and the response measurements (2) can be represented as follows:

$$Y_{ij} = \sum_{l=1}^d m_l(T_{ij}) X_{il} + \sum_{l=1}^d \sum_{k=1}^\infty \xi_{ik,l} \phi_{k,l}(T_{ij}) X_{il} + \sigma(T_{ij}) \varepsilon_{ij}. \quad (3)$$

Without loss of generality, we take  $\mathcal{T} = [a, b]$  to be  $[0, 1]$ . Following Xue and Yang (2006), we approximate each coefficient function by the spline smoothing method. To describe the spline functions, one can divide the finite interval  $[0, 1]$  into  $(N_s + 1)$  equal subintervals  $\chi_J = [v_J, v_{J+1})$ ,  $J = 0, \dots, N_s - 1$ ,  $\chi_{N_s} = [v_{N_s}, 1]$ . A sequence of equally spaced points  $\{v_J\}_{J=1}^{N_s}$ , called interior knots, are given as  $v_0 = 0 < v_1 < \dots < v_{N_s} < 1 = v_{N_s+1}$ . Let  $v_J = Jh_s$  for  $0 \leq J \leq N_s + 1$ , where  $h_s = 1/(N_s + 1)$  is the distance between neighboring knots. We denote by  $G^{(-1)} = G^{(-1)}[0, 1]$  the space of functions that are constant on each subinterval  $\chi_J$ , and the B-spline basis of  $G^{(-1)}$ , as  $\{b_J(t)\}_{J=0}^{N_s}$ , which are simply indicator functions of intervals  $\chi_J$ ,  $b_J(t) = I_{\chi_J}(t)$ ,  $J = 0, 1, \dots, N_s$ . For any  $t \in [0, 1]$ , define its location index as  $J(t) = J_n(t) = \min\{\lceil t/h_s \rceil, N_s\}$  so that  $t \in \chi_{J(t)}$ .

Next we define the space of spline coefficient functions on  $\mathcal{T} \times \mathbb{R}^d$  as

$$\mathcal{M} = \left\{ g(t, \mathbf{x}) = \sum_{l=1}^d g_l(t) x_l : g_l(t) \in G^{(-1)}, t \in \mathcal{T}, \mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d \right\},$$

and propose estimating the multivariate function  $\sum_{l=1}^d m_l(t) x_l$  by

$$\hat{m}(t, \mathbf{x}) = \sum_{l=1}^d \hat{m}_l(t) x_l = \underset{g \in \mathcal{M}}{\text{argmin}} \sum_{i=1}^n \sum_{j=1}^{N_i} \{Y_{ij} - g(T_{ij}, \mathbf{X}_i)\}^2. \quad (4)$$

Let  $\sigma_Y^2(t, \mathbf{x})$  be the conditional variance of  $\mathbf{Y}$  given  $T = t$  and  $\mathbf{X} = \mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$

$$\sigma_Y^2(t, \mathbf{x}) = \text{Var}(Y | T = t, \mathbf{X} = \mathbf{x}) = \sum_{l=1}^d G_l(t, t) x_l^2 + \sigma^2(t).$$

Next for any  $t \in [0, 1]$ , let

$$\Gamma_n(t) = c_{J(t),n}^{-2} \{n\mathbf{E}(N_1)\}^{-1} \mathbf{E}\mathbf{X}\mathbf{X}^\top \left[ \int_{\chi_{J(t)}} \sigma_Y^2(u, \mathbf{X}) f(u) du + \frac{\mathbf{E}\{N_1(N_1-1)\}}{\mathbf{E}N_1} \sum_{l=1}^d X_l^2 \int_{\chi_{J(t)} \times \chi_{J(t)}} G_l(u, v) f(u) f(v) dudv \right], \tag{5}$$

where

$$c_{J,n} = \mathbf{E}b_J^2(T) = \int_0^1 b_J^2(t) f(t) dt, \quad J = 0, \dots, N_s. \tag{6}$$

Further denote

$$\Sigma_n(t) = \mathbf{H}^{-1} \Gamma_n(t) \mathbf{H}^{-1} = \left\{ \sigma_{n,ll'}^2(t) \right\}_{l,l'=1}^d, \tag{7}$$

where  $\sigma_{n,ll'}^2(t)$  are later shown to be the asymptotic covariances between  $\hat{m}_l(t)$  and  $\hat{m}_{l'}(t)$ .

**Theorem 1** Under Assumptions (A1)–(A6) in Appendix A, for any  $t \in [0, 1]$ , as  $n \rightarrow \infty$ ,

$$\Sigma_n^{-1/2}(t) \{\hat{\mathbf{m}}(t) - \mathbf{m}(t)\} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{d \times d}),$$

where  $\hat{\mathbf{m}}(t) = (\hat{m}_1(t), \dots, \hat{m}_d(t))^\top$  is the estimate of  $\mathbf{m}(t) = (m_1(t), \dots, m_d(t))^\top$ . Furthermore, for any  $l = 1, \dots, d$  and  $\alpha \in (0, 1)$ ,

$$\lim_{n \rightarrow \infty} P \left\{ \sigma_{n,ll}^{-1}(t) |\hat{m}_l(t) - m_l(t)| \leq Z_{1-\alpha/2} \right\} = 1 - \alpha.$$

*Remark 1* Note that  $\Sigma_n(t) = \left\{ \sigma_{n,ll'}^2(t) \right\}_{l,l'=1}^d$  in (7) is complicated to compute in practice. The next proposition suggests that, for any  $t \in [0, 1]$ ,  $\Gamma_n(t)$  in (5) can be simplified by

$$\tilde{\Gamma}_n(t) \equiv \mathbf{E} \left[ \mathbf{X}\mathbf{X}^\top \frac{\sigma_Y^2(t, \mathbf{X})}{f(t)h_s n \mathbf{E}(N_1)} \left\{ 1 + \frac{\mathbf{E}N_1(N_1-1)}{\mathbf{E}N_1} \frac{\sum_{l=1}^d X_l^2 G_l(t, t) f(t) h_s}{\sigma_Y^2(t, \mathbf{X})} \right\} \right]. \tag{8}$$

Denote the supremum of a function  $\phi$  on  $[a, b]$  by  $\|\phi\|_\infty = \sup_{t \in [a,b]} |\phi(t)|$ . For any matrix  $\mathbf{A} = (a_{ij})$ , define  $\|\mathbf{A}\|_\infty = \max |a_{ij}|$ , where the maximum is taken over all the elements of  $\mathbf{A}$ , while for a matrix function  $\mathbf{A}(t) = (a_{ij}(t))$ ,  $\|\mathbf{A}\|_\infty = \sup_{t \in [a,b]} \|\mathbf{A}(t)\|_\infty$ .

**Proposition 1** Under Assumptions (A2)–(A6) in Appendix A, there exists a constant  $c > 0$  such that as  $n \rightarrow \infty$ ,  $\|\Gamma_n(t) - \tilde{\Gamma}_n(t)\|_\infty = \mathcal{O}(n^{-1}h_s^{r-1}) = \mathcal{O}(n^{-c})$ .



To derive the maximal deviation distribution of estimators  $\hat{m}_l(t)$ ,  $l = 1, \dots, d$ , let

$$Q_{N_s+1}(\alpha) = b_{N_s+1} - a_{N_s+1}^{-1} \log \left\{ -\frac{1}{2} \log(1 - \alpha) \right\}, \quad \alpha \in (0, 1) \quad (9)$$

$$a_{N_s+1} = \{2 \log(N_s + 1)\}^{1/2}, \quad b_{N_s+1} = a_{N_s+1} - \frac{\log(2\pi a_{N_s+1}^2)}{2a_{N_s+1}}. \quad (10)$$

**Theorem 2** *Under Assumptions (A1)–(A6) in Appendix A, for  $l = 1, \dots, d$  and any  $\alpha \in (0, 1)$ ,*

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{t \in [0, 1]} \sigma_{n, ll}^{-1}(t) |\hat{m}_l(t) - m_l(t)| \leq Q_{N_s+1}(\alpha) \right\} = 1 - \alpha,$$

where  $\sigma_{n, ll}(t)$  and  $Q_{N_s+1}(\alpha)$  are given in (7) and (9), respectively.

One reviewer has pointed out that the use of constant instead of higher order spline is not optimal, which we completely agree. Further research involving sophisticated nonstationary Gaussian process extreme value theory is needed to extend our present work to splines of any order, such as the popular cubic spline. To be precise, analog of Proposition 4 for higher order spline concerns the maximum of a standardized continuous Gaussian process over interval  $[0, 1]$ , whereas for constant spline, the Gaussian process breaks down to  $N_s+1$  weakly correlated standard Gaussian variables.

### 3 Asymptotic confidence regions

In this section, we construct the confidence regions for functions  $m_l(t)$ ,  $l = 1, \dots, d$ .

#### 3.1 Asymptotic confidence intervals and SCCs

Theorems 1 and 2 allow one to construct pointwise confidence intervals and SCCs for components  $\hat{m}_l(t)$ ,  $l = 1, \dots, d$ . The next corollary provides the theoretical underpinning upon which SCCs can be actually implemented, see Sect. 3.2.

**Corollary 1** *Under Assumptions (A1)–(A6) in Appendix A, for any  $l = 1, \dots, d$  and  $\alpha \in (0, 1)$ , as  $n \rightarrow \infty$ ,*

- (i) *an asymptotic  $100(1 - \alpha)\%$  pointwise confidence interval for  $m_l(t)$ ,  $t \in [0, 1]$ , is  $\hat{m}_l(t) \pm \sigma_{n, ll}(t) Z_{1-\alpha/2}$ , with  $\sigma_{n, ll}(t)$  given in (7), while  $Z_{1-\alpha/2}$  is the  $100(1 - \alpha/2)\%$ th percentile of the standard normal distribution.*
- (ii) *an asymptotic  $100(1 - \alpha)\%$  SCC for  $m_l(t)$ , with  $Q_{N_s+1}(\alpha)$  given in (9), is  $\hat{m}_l(t) \pm \sigma_{n, ll}(t) Q_{N_s+1}(\alpha)$ ,  $t \in [0, 1]$ .*

One reviewer has raised the interesting question whether our SCC would significantly improve by some form of bootstrapping. The answer is negative for now due to the lack of convincing procedures that simultaneously resample from the unknown

distributions of both the unobserved error  $\varepsilon_{ij}$ s and the unobserved functional principal components  $\xi_{ik,l}$ s. On the other hand, further investigation in FDA will lead to theoretically sound resampling methods analogous to the smoothed bootstrap for nonparametric regression in [Claeskens and Van Keilegom \(2003\)](#).

### 3.2 Implementation

In the following, we describe procedures to construct the SCCs and the pointwise intervals given in Corollary 1. For any data set  $(T_{ij}, Y_{ij}, X_{il})_{i=1, j=1, l=1}^{n, N_i, d}$  from model (3), the spline estimators  $\hat{m}_l(t), l = 1, \dots, d$ , are obtained by (4), and the number of interior knots is taken to be  $N_s = \lceil cN_T^{1/3}(\log(n)) \rceil$ , in which  $N_T = \sum_{i=1}^n N_i$  is the total sample size,  $\lceil a \rceil$  denotes the integer part of  $a$ , and  $c$  is a positive constant.

To construct the SCCs, one needs to evaluate the functions  $\sigma_{n,ll}^2(t), l = 1, \dots, d$ , which are the diagonal elements of matrix  $\Sigma_n(t)$  in (7). Based on Proposition 1, one can estimate each unknowns  $f(t), \sigma_Y^2(t, \mathbf{x}), G_l(t, t)$  and matrix  $\mathbf{H}$  and then plug these estimators into the formula of the SCCs; see [Wang and Yang \(2009\)](#).

The number of interior knots for pilot estimation of  $f(t), \sigma_Y^2(t, \mathbf{x})$ , and  $G_l(t, t)$  is taken to be  $N_s^* = \lceil 0.5n^{1/3} \rceil$ , and  $h_s^* = 1 / (1 + N_s^*)$ . The histogram estimator of the density function  $f(t)$  is  $\hat{f}(t) = N_T^{-1} h_s^{*-1} \sum_{i=1}^n \sum_{j=1}^{N_i} b_{J(t)}(T_{ij})$ .

To estimate the covariance matrix  $\Gamma_n(t)$  in (5), define the raw covariance term  $R_{ij} = \left( Y_{ij} - \sum_{l=1}^d \hat{m}(T_{ij}) X_{il} \right)^2, 1 \leq j \leq N_i, 1 \leq i \leq n$ , the estimator of  $\sigma_Y^2(t, \mathbf{x})$  is

$$\hat{\sigma}_Y^2(t, \mathbf{x}) = \sum_{l=1}^d \sum_{J=0}^{N_s^*} \hat{\rho}_{J,l} b_J(t) x_l^2 + \sum_{J=0}^{N_s^*} \hat{\mu}_J b_J(t) = \sum_{l=1}^d \hat{G}_l(t, t) x_l^2 + \hat{\sigma}^2(t),$$

where  $\{\hat{\rho}_{0,1}, \dots, \hat{\rho}_{N_s^*,d}, \hat{\mu}_0, \dots, \hat{\mu}_{N_s^*}\}^\top$  are solutions of the following least squares problem:

$$\begin{aligned} & (\hat{\rho}_{0,1}, \dots, \hat{\rho}_{N_s^*,d}, \hat{\mu}_0, \dots, \hat{\mu}_{N_s^*})^\top \\ &= \underset{\substack{\text{argmin} \\ (\rho_{0,1}, \dots, \mu_{N_s^*})^\top \in \mathbb{R}^{(N_s^*+1)(d+1)}}}{\sum_{i=1}^n \sum_{j=1}^{N_i} \left\{ R_{ij} - \sum_{l=1}^d \sum_{J=0}^{N_s^*} \rho_{J,l} b_J(T_{ij}) X_{il}^2 - \sum_{J=0}^{N_s^*} \mu_J b_J(T_{ij}) \right\}^2}. \end{aligned}$$

The matrix  $\Gamma_n(t)$  is estimated by substituting  $f(t), G_l(t, t)$  and  $\sigma_Y^2(t, \mathbf{x})$  with  $\hat{f}(t), \hat{G}_l(t, t)$  and  $\hat{\sigma}_Y^2(t, \mathbf{x})$ . Define

$$\begin{aligned} \hat{\Gamma}_n(t) \equiv & \left[ n^{-1} \sum_{i=1}^n X_{il} X_{il'} \hat{\sigma}_Y^2(t, \mathbf{X}_i) \left\{ \hat{f}(t) h_s N_T \right\}^{-1} \right. \\ & \left. \times \left\{ 1 + \left( \frac{\sum_{i=1}^n N_i^2}{N_T} - 1 \right) \frac{\sum_{l=1}^d \hat{G}_l(t, t) X_{il}^2}{\hat{\sigma}_Y^2(t, \mathbf{X}_i)} \hat{f}(t) h_s \right\} \right]_{l,l'=1}^d. \end{aligned}$$

The following proposition provides the consistent rate of  $\hat{\Gamma}_n(t)$  to  $\Gamma_n(t)$ .

**Proposition 2** *Under Assumptions (A1)–(A6) in Appendix A, there exists a constant  $c > 0$  such that as  $n \rightarrow \infty$ ,  $\|\hat{\Gamma}_n(t) - \Gamma_n(t)\|_\infty = \mathcal{O}_p(n^{-c})$ .*

Proposition 2 implies that  $\Gamma_n(t)$  can be replaced by  $\hat{\Gamma}_n(t)$  with a negligible error. Define a  $d \times d$  matrix  $\hat{\mathbf{H}} = \{n^{-1} \sum_{i=1}^n X_{il} X_{il'}\}_{l,l'=1}^d$ , then  $\Sigma_n(t)$  can be estimated well by  $\hat{\Sigma}_n(t) = \{\hat{\sigma}_{n,ll'}^2(t)\}_{l,l'=1}^d = \hat{\mathbf{H}}^{-1} \hat{\Gamma}_n(t) \hat{\mathbf{H}}^{-1}$ . Therefore, as  $n \rightarrow \infty$ ,  $l = 1, \dots, d$ , the SCCs

$$\hat{m}_l(t) \pm \hat{\sigma}_{n,ll}(t) Q_{N_s+1}(\alpha), \quad (11)$$

with  $Q_{N_s+1}(\alpha)$  given in (9), and the pointwise intervals  $\hat{m}_l(t) \pm \hat{\sigma}_{n,ll}(t) Z_{1-\alpha/2}$  have asymptotic confidence level  $1 - \alpha$ .

#### 4 Decomposition

In this section, we describe the representation of the spline estimators  $\hat{m}_l(t)$ ,  $l = 1, \dots, d$ , in (4), then break the estimation error  $\hat{m}_l(t) - m_l(t)$  into three terms by the decomposition of  $Y_{ij}$  in model (3). Although such representation is not needed for applying the procedure described in Sect. 3.2 to analyze data, it provides insights into the proof of the main theoretical results in Sect. 2.

We consider the following rescaled B-spline basis  $\{B_J(t)\}_{J=0}^{N_s}$  for  $G^{(-1)}$ :

$$B_J(t) \equiv b_J(t) (c_{J,n})^{-1/2}, \quad J = 0, \dots, N_s. \quad (12)$$

It is easily verified that  $\mathbf{E}\{B_J(T)\}^2 = 1$  for  $J = 0, 1, \dots, N_s$ , and  $B_J(t) B_{J'}(t) \equiv 0$  for  $J \neq J'$ . By simple linear algebra, the spline estimator  $\hat{m}_l(t)$  defined in (4) equals

$$\hat{m}_l(t) = \sum_{J=0}^{N_s} \hat{\gamma}_{J,l} B_J(t) = c_{J(t),n}^{-1/2} \hat{\gamma}_{J(t),l}, \quad l = 1, \dots, d, \quad (13)$$

where the coefficients  $\hat{\gamma} = (\hat{\gamma}_0^\top, \dots, \hat{\gamma}_{N_s}^\top)^\top$  with  $\hat{\gamma}_J = (\hat{\gamma}_{J,1}, \dots, \hat{\gamma}_{J,d})^\top$  being the solution of the following least squares problem

$$\hat{\gamma} = \underset{\gamma = (\gamma_{0,1}, \dots, \gamma_{N_s,d})^\top \in \mathbb{R}^{d(N_s+1)}}{\operatorname{argmin}} \sum_{i=1}^n \sum_{j=1}^{N_i} \left\{ Y_{ij} - \sum_{l=1}^d \sum_{J=0}^{N_s} \gamma_{J,l} B_J(T_{ij}) X_{il} \right\}^2. \quad (14)$$

In the following, let  $\mathbf{Y} = (Y_{11}, \dots, Y_{1N_1}, \dots, Y_{n1}, \dots, Y_{nN_n})^\top$  be the collection of all the  $Y_{ij}$ s. Let  $\mathbf{B}(t) = (B_0(t), \dots, B_{N_s}(t))^\top$  and  $\mathbf{X}_i = (X_{i1}, \dots, X_{id})^\top$  be two

vectors of dimension  $(N_s + 1)$  and  $d$ , respectively. Denote

$$\mathbf{D} = (\mathbf{B}(T_{11}) \otimes \mathbf{X}_1, \dots, \mathbf{B}(T_{1N_1}) \otimes \mathbf{X}_1, \dots, \mathbf{B}(T_{n1}) \otimes \mathbf{X}_n, \dots, \mathbf{B}(T_{nN_n}) \otimes \mathbf{X}_n)^\top, \tag{15}$$

a  $N_T \times ((N_s + 1)d)$  matrix, where “ $\otimes$ ” denotes the Kronecker product. Solving the least squares problem in (14), we obtain

$$\hat{\boldsymbol{\gamma}} = (\mathbf{D}^\top \mathbf{D})^{-1} (\mathbf{D}^\top \mathbf{Y}). \tag{16}$$

Denote  $\mathbf{x} = (x_1, \dots, x_d)^\top$ , thus Eq. (4) can be rewritten as

$$\sum_{l=1}^d \hat{m}_l(t)x_l = (\mathbf{B}(t) \otimes \mathbf{x})^\top (\mathbf{D}^\top \mathbf{D})^{-1} (\mathbf{D}^\top \mathbf{Y}). \tag{17}$$

According to (15), one has  $\mathbf{D}^\top \mathbf{D} = \sum_{i=1}^n \sum_{j=1}^{N_i} \{\mathbf{B}(T_{ij})\mathbf{B}(T_{ij})^\top \otimes \mathbf{X}_i \mathbf{X}_i^\top\}$ , in which matrix  $\mathbf{B}(T_{ij})\mathbf{B}(T_{ij})^\top = \text{diag}\{B_0^2(T_{ij}), \dots, B_{N_s}^2(T_{ij})\}$ . So matrix  $\mathbf{D}^\top \mathbf{D}$  should be a block diagonal matrix, and  $N_T^{-1} \mathbf{D}^\top \mathbf{D} = \text{diag}\{\hat{\mathbf{V}}_0, \dots, \hat{\mathbf{V}}_{N_s}\}$ , where

$$\hat{\mathbf{V}}_J = \left\{ N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J^2(T_{ij}) X_{il} X_{il'} \right\}_{l,l'=1}^d. \tag{18}$$

On the other hand, we have  $\mathbf{D}^\top \mathbf{Y} = \sum_{i=1}^n \sum_{j=1}^{N_i} \{\mathbf{B}(T_{ij}) \otimes \mathbf{X}_i\} Y_{ij}$ . Thus,  $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0^\top, \dots, \hat{\gamma}_{N_s}^\top)^\top$  can be easily calculated using

$$\hat{\gamma}_J = \hat{\mathbf{V}}_J^{-1} \left\{ N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J(T_{ij}) X_{il} Y_{ij} \right\}_{l=1}^d, \quad J = 0, \dots, N_s. \tag{19}$$

Then the functions  $\mathbf{m}(t) = (m_1(t), \dots, m_d(t))^\top$  can be simply estimated by

$$\hat{\mathbf{m}}(t) = (\hat{m}_1(t), \dots, \hat{m}_d(t))^\top = c_{J(t),n}^{-1/2} (\hat{\gamma}_{J(t),1}, \dots, \hat{\gamma}_{J(t),d})^\top = c_{J(t),n}^{-1/2} \hat{\gamma}_{J(t)}. \tag{20}$$

Projecting the relationship in model (3) onto the space of spline coefficient functions on  $\mathcal{T} \times \mathbb{R}^d$  as  $\mathcal{M}$ , we obtain the following important decomposition:

$$\sum_{l=1}^d \hat{m}_l(t)x_l = \sum_{l=1}^d \tilde{m}_l(t)x_l + \sum_{l=1}^d \tilde{\xi}_l(t)x_l + \sum_{l=1}^d \tilde{\varepsilon}_l(t)x_l, \tag{21}$$

where for any  $l = 1, \dots, d$ ,

$$\tilde{m}_l(t) = \sum_{J=0}^{N_s} \tilde{\gamma}_{J,l} B_J(t) = c_{J(t),n}^{-1/2} \tilde{\gamma}_{J(t),l}, \quad (22)$$

$$\tilde{\xi}_l(t) = \sum_{J=0}^{N_s} \tilde{\alpha}_{J,l} B_J(t) = c_{J(t),n}^{-1/2} \tilde{\alpha}_{J(t),l}, \quad \tilde{\varepsilon}_l(t) = \sum_{J=0}^{N_s} \tilde{\theta}_{J,l} B_J(t) = c_{J(t),n}^{-1/2} \tilde{\theta}_{J(t),l}, \quad (23)$$

where  $(\tilde{\gamma}_{J,l}, J = 0, \dots, N_s, l = 1, \dots, d)^\top$ ,  $(\tilde{\alpha}_{J,l}, J = 0, \dots, N_s, l = 1, \dots, d)^\top$ , and  $(\tilde{\theta}_{J,l}, J = 0, \dots, N_s, l = 1, \dots, d)^\top$  are solutions to (14) with  $Y_{ij}$  replaced by  $\sum_{l=1}^d m_l(T_{ij}) X_{il}$ ,  $\sum_{l=1}^d \sum_{k=1}^{\infty} \xi_{ik,l} \phi_{k,l}(T_{ij}) X_{il}$ , and  $\sigma(T_{ij}) \varepsilon_{ij}$ , respectively.

Furthermore, under Assumption (A5) we can decompose  $\hat{m}_l(t)$  as

$$\hat{m}_l(t) = \tilde{m}_l(t) + \tilde{\xi}_l(t) + \tilde{\varepsilon}_l(t), \quad l = 1, \dots, d. \quad (24)$$

The next two propositions concern the functions  $\tilde{m}_l(t)$ ,  $\tilde{\xi}_l(t)$ ,  $\tilde{\varepsilon}_l(t)$ ,  $l = 1, \dots, d$ , given in (22) and (23). Proposition 3 gives the uniform convergence rate of  $\tilde{m}_l(t)$  to  $m_l(t)$ . Proposition 4 provides the asymptotic distribution for the maximum of the normalized error terms.

**Proposition 3** Under Assumptions (A1), (A2) and (A4)–(A6) in Appendix A, the functions  $\tilde{m}_l(t)$ ,  $l = 1, \dots, d$  satisfy  $\sup_{t \in [0,1]} \sup_{1 \leq l \leq d} |\tilde{m}_l(t) - m_l(t)| = \mathcal{O}_p(h_s)$ .

**Proposition 4** Under Assumptions (A2)–(A6) in Appendix A, for  $\tau \in \mathbb{R}$ , and  $\sigma_{n,ll}(t)$ ,  $a_{N_s+1}$ ,  $b_{N_s+1}$  as given in (7) and (9),

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) \left| \tilde{\xi}_l(t) + \tilde{\varepsilon}_l(t) \right| \leq \tau/a_{N_s+1} + b_{N_s+1} \right\} = \exp(-2e^{-\tau}).$$

## 5 Simulation

To illustrate the finite-sample performance of the spline approach, we generate data from the following model

$$Y_{ij} = \left\{ m_1(T_{ij}) + \sum_{k=1}^2 \xi_{ik,1} \phi_{k,1}(T_{ij}) \right\} X_{i1} + \left\{ m_2(T_{ij}) + \sum_{k=1}^3 \xi_{ik,2} \phi_{k,2}(T_{ij}) \right\} X_{i2} \\ + \sigma(T_{ij}) \varepsilon_{ij}, \quad 1 \leq i \leq n, 1 \leq j \leq N_i,$$

where  $T \sim U[0, 1]$ ,  $X_1 \sim N(0, 1)$ ,  $X_2 \sim \text{Binomial}[1, 0.5]$ ,  $\xi_{k,1} \sim N(0, 1)$ ,  $k = 1, 2$ ,  $\xi_{k,2} \sim N(0, 1)$ ,  $k = 1, 2, 3$ ,  $\varepsilon \sim N(0, 1)$ , and  $N_i$  is generated from a discrete uniform distribution from  $2, \dots, 14$ , for  $1 \leq i \leq n$ . For the first component, we take  $m_1(t) = \sin\{2\pi(t - 1/2)\}$ ,  $\phi_{1,1}(t) = -2 \cos\{\pi(t - 1/2)\}/\sqrt{5}$ ,  $\phi_{2,1}(t) = \sin\{\pi(t - 1/2)\}/\sqrt{5}$ , thus  $\lambda_{1,1} = 2/5$ ,  $\lambda_{2,1} = 1/10$ . For the second

**Table 1** Coverage percentages of the SCCs for functions  $m_1$  (left) and  $m_2$  (right), based on 500 replications

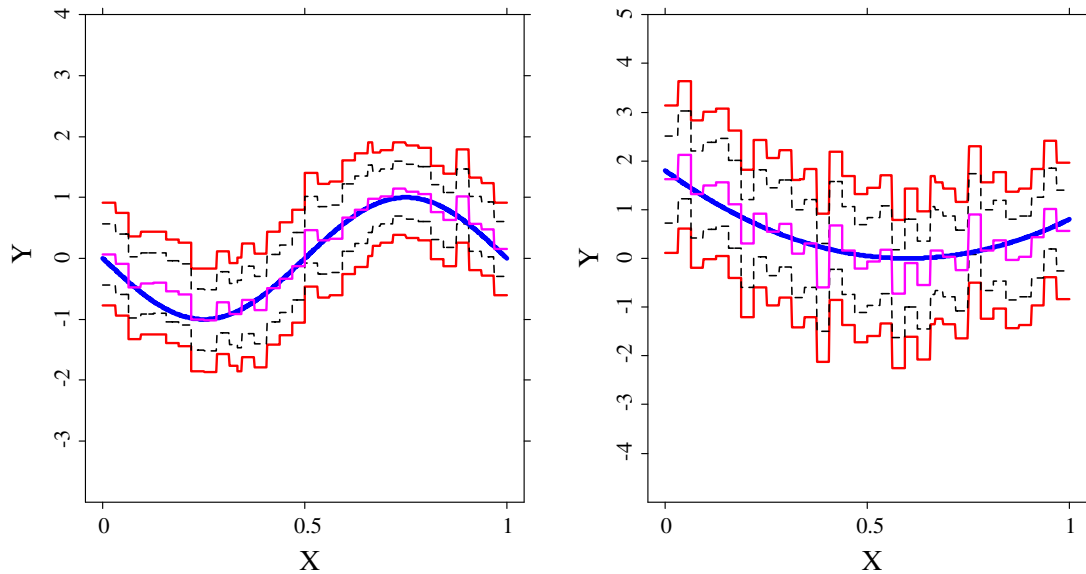
$\sigma$	$n$	$1 - \alpha$	$c = 0.3$	$c = 0.5$	$c = 0.8$	$c = 1$
1.0	200	0.950	0.950, 0.952	0.944, 0.948	0.920, 0.904	0.886, 0.884
		0.990	0.990, 0.998	0.990, 0.990	0.976, 0.984	0.968, 0.974
	400	0.950	0.944, 0.948	0.950, 0.930	0.922, 0.912	0.908, 0.904
		0.990	0.996, 0.984	0.990, 0.988	0.984, 0.988	0.974, 0.966
	600	0.950	0.934, 0.962	0.954, 0.946	0.930, 0.952	0.930, 0.924
		0.990	0.992, 0.996	0.992, 0.986	0.988, 0.990	0.984, 0.990
	800	0.950	0.936, 0.934	0.960, 0.966	0.950, 0.964	0.956, 0.934
		0.990	0.998, 0.996	0.994, 0.994	0.986, 0.992	0.988, 0.988
0.5	200	0.950	0.936, 0.948	0.952, 0.942	0.916, 0.900	0.912, 0.890
		0.990	0.988, 0.994	0.992, 0.990	0.972, 0.974	0.972, 0.972
	400	0.950	0.916, 0.930	0.936, 0.932	0.928, 0.916	0.904, 0.898
		0.990	0.994, 0.984	0.992, 0.988	0.996, 0.988	0.978, 0.976
	600	0.950	0.924, 0.948	0.952, 0.954	0.926, 0.958	0.936, 0.938
		0.990	0.996, 0.994	0.994, 0.986	0.984, 0.990	0.990, 0.994
	800	0.950	0.942, 0.900	0.950, 0.960	0.942, 0.962	0.960, 0.938
		0.990	0.996, 0.998	0.996, 0.994	0.990, 0.996	0.992, 0.988

component, we take  $m_2(t) = 5(t - 0.6)^2$ ,  $\phi_{1,2}(t) = 1$ ,  $\phi_{2,2}(t) = \sqrt{2} \sin(2\pi t)$ ,  $\phi_{3,2}(t) = \sqrt{2} \cos(2\pi t)$ , thus  $\lambda_{1,2} = \lambda_{2,2} = \lambda_{3,2} = 1$ . The noise level is chosen to be  $\sigma = 0.5, 1.0$ , and the number of subjects  $n$  is taken to be 200, 400, 600, 800.

We consider the confidence levels  $1 - \alpha = 0.95$  and  $0.99$ . Table 1 reports the coverage of the SCCs as the percentage out of the total 500 replications for which the true curve was covered by (11) at the 101 points  $\{k/100, k = 0, \dots, 100\}$ .

In the above SCC construction, the number of interior knots  $N_s$  is determined by the sample size  $n$  and a tuning constant  $c$  as described in Sect. 3.2. We have experimented with  $c = 0.3, 0.5, 0.8, 1.0$  in this simulation study. The simulation results in Table 1 reflect that the coverage percentages depend on the choice of  $c$ ; however, the dependency becomes weaker when sample sizes increase. For large sample sizes  $n = 600, 800$ , the effect of the choice of  $c$  on the coverage percentages is insignificant. Because  $N_s$  varies with  $N_i$ , for  $1 \leq i \leq n$ , the data-driven selection of an “optimal”  $N_s$  remains an open problem. At all noise levels, the coverage percentages for the SCC (11) are very close to the nominal confidence levels 0.95 and 0.99 for  $c = 0.5$ . Note that since  $EN_1 = 8$ , the total sample size  $N_T \approx 8 \times 200, 8 \times 400, 8 \times 600, 8 \times 800$  which explains the closeness of coverage percentages in Table 1 to the nominal levels. These large  $N_T$ s are realistic as we believe they are common for real data. For instance, the CD4 cell percentage data in Sect. 6 has  $N_T = 1,817$ .

For visualization of actual function estimates, Fig. 1 shows the true curve, the estimated curve, the asymptotic 95 % SCC and the pointwise confidence intervals at  $\sigma = 0.5$  with  $n = 200$ . The same plot for  $n = 600$  has shown significantly



**Fig. 1** Plots of 95 % SCC (11) (upper and lower solid), pointwise confidence intervals (dashed), the spline estimator (thin), and the true function (middle thick) at  $\sigma = 0.5$ ,  $n = 200$  for  $m_1$  (left) and  $m_2$  (right)

narrower SCC and pointwise confidence intervals as expected, but is not included to save space.

## 6 Real data analysis

To illustrate our method, we return to the CD4 cell percentage data discussed in Sect. 1 for further analysis. Since the actual visit times  $T_{ij}$  are irregularly spaced and vary from year 0 to year 6, we first transform the times by  $Z_{ij} = F_{N_T}(T_{ij})$ , where  $F_{N_T}$  is the empirical cdf of times  $\{T_{ij}\}_{i=1, j=1}^{n, N_i}$ . Then the  $Z_{ij}$  values are distributed fairly uniformly. We have set a slightly smaller number of interior knots  $N_s = \lceil 0.3N_T^{1/3}(\log(n)) \rceil$  to avoid singularity in solving the least squares problem.

The left plots of Figs. 2, 3, 4 and 5 depict the spline estimates, the asymptotic 95 % SCCs, the pointwise confidence intervals for  $m_l(t)$ ,  $l = 0, 1, 2, 3$ , respectively. The horizontal solid line represents zero. Based on the shape of the SCCs, we are interested in testing the following hypotheses:

$H_{00} : m_0(t) \equiv a + bt$ , for some  $a, b \in \mathbb{R}$  v.s.  $H_{10} : m_0(t) \neq a + bt$ , for any  $a, b \in \mathbb{R}$ ;

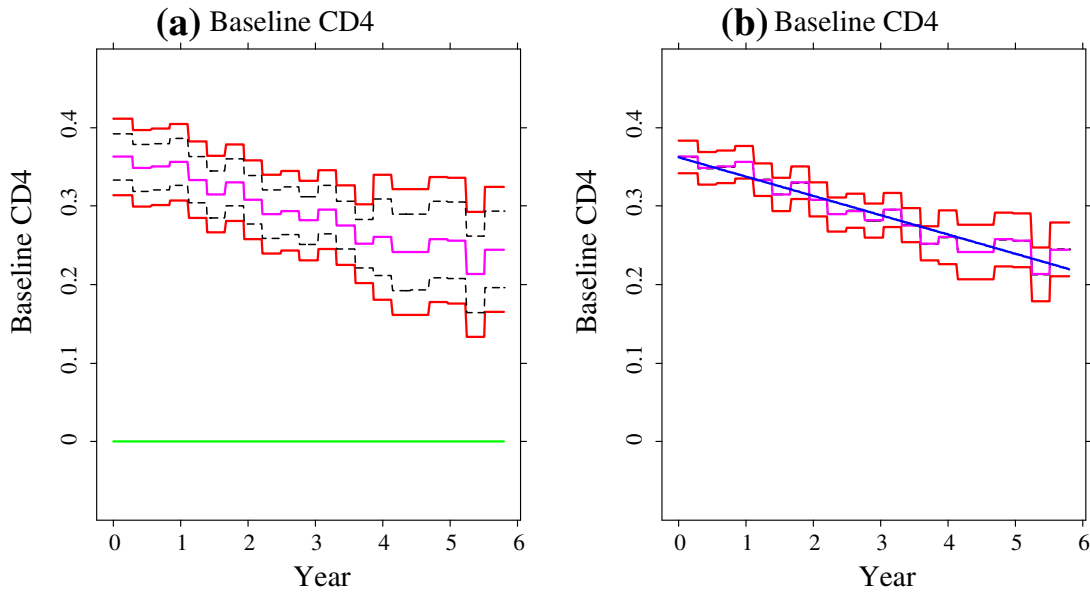
$H_{01} : m_1(t) \equiv 0$  v.s.  $H_{11} : m_1(t) \neq 0$ , for some  $t \in [0, 6]$ ;

$H_{02} : m_2(t) \equiv c$  for some  $c > 0$  v.s.  $H_{12} : m_2(t) \neq c$ , for any  $c > 0$ ;

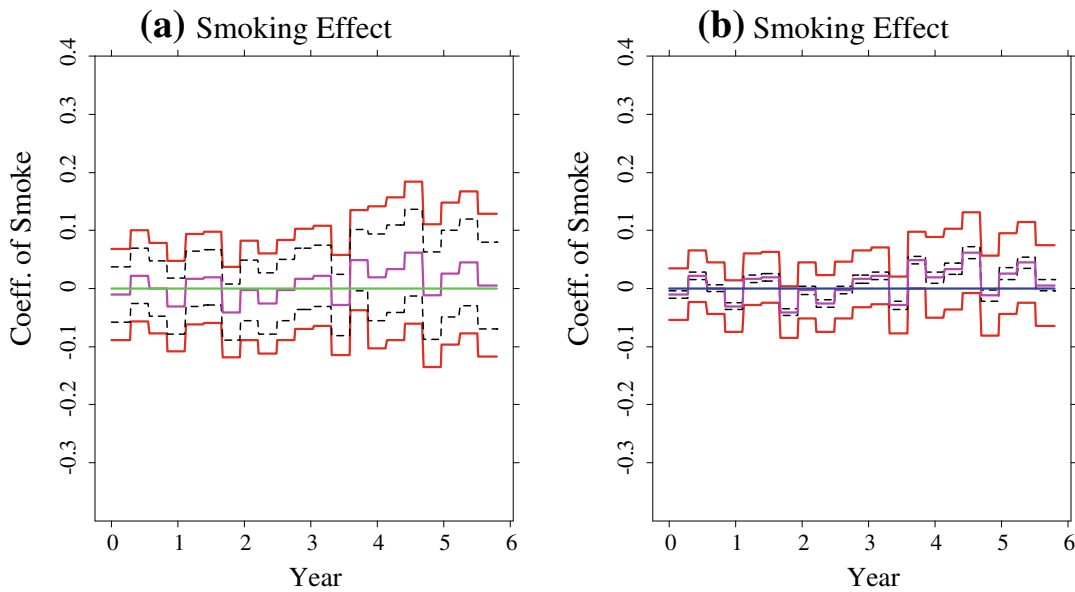
$H_{03} : m_3(t) \equiv 0$  v.s.  $H_{13} : m_3(t) \neq 0$ , for some  $t \in [0, 6]$ .

Asymptotic  $p$  values are calculated for each pair of hypotheses as  $\hat{\alpha}_0 = 0.99072$ ,  $\hat{\alpha}_1 = 0.79723$ ,  $\hat{\alpha}_2 = 0.25404$ ,  $\hat{\alpha}_3 = 0.10775$ . Apparently, none of the null hypothesis is rejected. The  $p$  values are calculated as, for example

$$\hat{\alpha}_0 = 1 - \exp \left[ -2 \exp \left( -a_{N_s+1} \left\{ \max_{k=0}^{400} \left| \frac{\hat{m}_0(t_k) - (\hat{a} + \hat{b}t_k)}{\hat{\sigma}_{n,0}(t_k)} \right| - b_{N_s+1} \right\} \right) \right],$$



**Fig. 2** Plots of **a** 95 % SCC (*upper and lower solid*), pointwise confidence intervals (*dashed*) and the spline estimator  $\hat{m}_0$  (*middle solid*) for baseline effect; and **b** the same except with confidence level  $1 - \hat{\alpha}_0$  and the estimated  $m_0$  under  $H_{00}$  (*solid linear*)

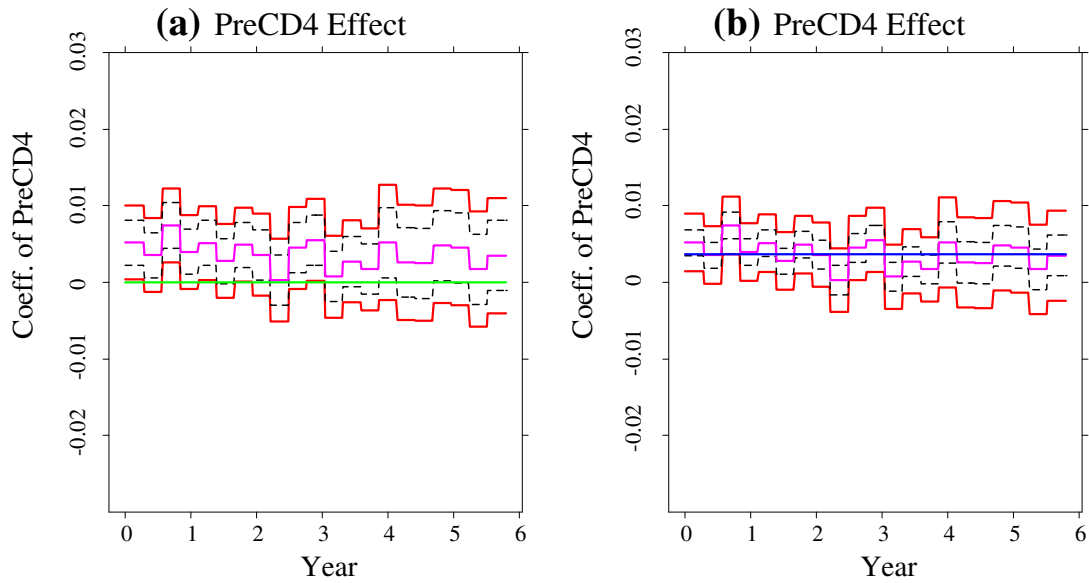


**Fig. 3** Plots of **a** 95 % SCC (*upper and lower solid*), pointwise confidence intervals (*dashed*) and the spline estimator  $\hat{m}_1$  (*middle solid*) for smoking effect; and **b** the same except with confidence level  $1 - \hat{\alpha}_1$  and the estimated  $m_1$  under  $H_{01}$  (*solid linear*)

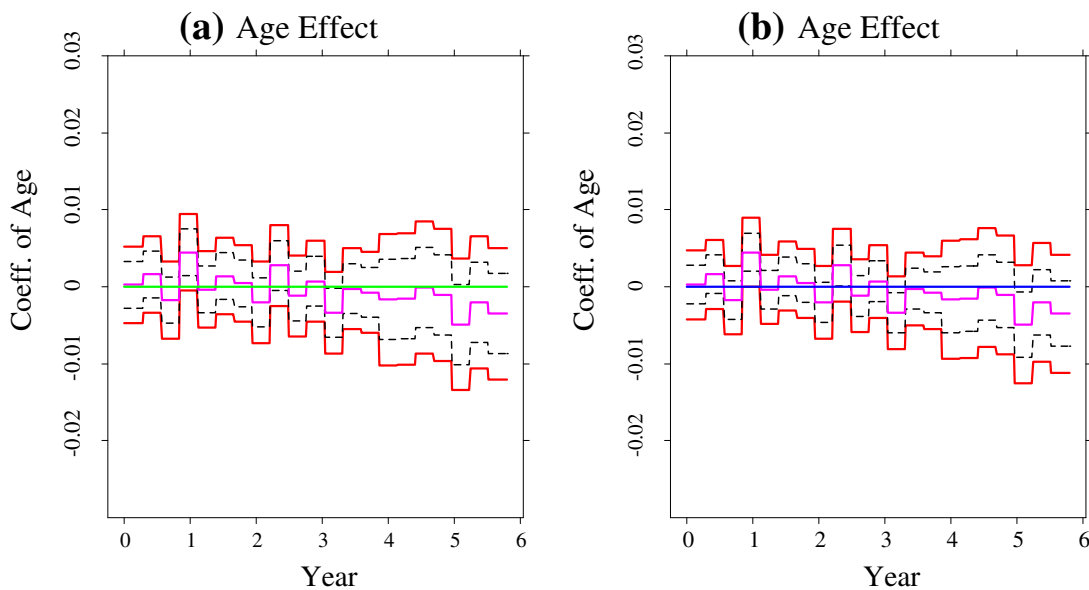
where  $t_k, k = 0, \dots, 400$  are equally spaced grid points over the range of the actual visit times, while  $\hat{a} + \hat{b}t$  is a least squares linear approximation to  $\hat{m}_0(t)$ . In other words, the  $p$  value  $\hat{\alpha}_0$  is a solution of

$$\max_{k=0}^{400} \left| \frac{\hat{m}_0(t_k) - (\hat{a} + \hat{b}t_k)}{\hat{\sigma}_{n,ll}(t_k)} \right| = b_{N_s+1} - a_{N_s+1}^{-1} \log \left\{ -\frac{1}{2} \log(1 - \hat{\alpha}_0) \right\}.$$





**Fig. 4** Plots of **a** 95 % SCC (*upper and lower solid*), pointwise confidence intervals (*dashed*) and the spline estimator  $\hat{m}_2$  (*middle solid*) for pre-infection CD4 effect; and **b** the same except with confidence level  $1 - \hat{\alpha}_2$  and the estimated  $m_2$  under  $H_{02}$  (*solid linear*)



**Fig. 5** Plots of **a** 95 % SCC (*upper and lower solid*), pointwise confidence intervals (*dashed*) and the spline estimator  $\hat{m}_3$  (*middle solid*) for age effect; and **b** the same except with confidence level  $1 - \hat{\alpha}_3$  and the estimated  $m_3$  under  $H_{03}$  (*solid linear*)

The right plots of Figs. 2, 3, 4 and 5 show the spline estimates, the  $100(1 - \hat{\alpha}_l)\%$  SCCs and the pointwise confidence intervals, and estimates of  $m_l(t)$  under  $H_{0l}$ ,  $l = 0, 1, 2, 3$ . From these figures, one can see the baseline CD4 percentage of the population is a decreasing linear function of time and greater than zero over the range of time. The effects of smoking status and age at HIV infection are insignificant, while the pre-infection CD4 percentage is positively proportional to the post-infection CD4 percentage. These findings are consistent with the observations in Wu and Chiang

(2000), Fan and Zhang (2000) and Wang et al. (2008), but are put on rigorous standing due to the quantification of type I errors by computing asymptotic  $p$  values relative to the SCCs.

**Acknowledgments** This work is part of Lijie Gu’s dissertation and has been supported in part by the Deutsche Forschungsgemeinschaft through the CRC 649 “Economic Risk”, the US National Science Foundation awards DMS 0905730, 1007594, 1106816, 1309800, Jiangsu Specially Appointed Professor Program SR10700111, Jiangsu Province Key-Discipline Program (Statistics) ZY107002, National Natural Science Foundation of China award 11371272, and Research Fund for the Doctoral Program of Higher Education of China award 20133201110002.

### Appendix A

Throughout this section,  $a_n \sim b_n$  means  $\lim_{n \rightarrow \infty} b_n/a_n = c$ , where  $c$  is some nonzero constant. For functions  $a_n(t)$ ,  $b_n(t)$ ,  $a_n(t) = \mathcal{U}\{b_n(t)\}$  means  $a_n(t)/b_n(t) \rightarrow 0$  as  $n \rightarrow \infty$  uniformly for  $t \in [0, 1]$ , and  $a_n(t) = \mathcal{U}\{b_n(t)\}$  means  $a_n(t)/b_n(t) = \mathcal{O}(1)$  as  $n \rightarrow \infty$  uniformly for  $t \in [0, 1]$ . We use  $\mathcal{U}_p(\cdot)$  and  $\mathcal{U}_p(\cdot)$  if the convergence is in the sense of uniform convergence in probability.

#### A.1 Technical assumptions

We define the modulus of continuity of a continuous function  $\phi$  on  $[a, b]$  by  $\omega(\phi, \delta) = \max_{t, t' \in [a, b], |t-t'| \leq \delta} |\phi(t) - \phi(t')|$ . For any  $r \in (0, 1]$ , denote the collection of order  $r$  Hölder continuous function on  $[0, 1]$  by

$$C^{0,r}[0, 1] = \left\{ \phi : \|\phi\|_{0,r} = \sup_{t \neq t', t, t' \in [0,1]} \frac{|\phi(t) - \phi(t')|}{|t - t'|^r} < +\infty \right\},$$

in which  $\|\phi\|_{0,r}$  is the  $C^{0,r}$ -seminorm of  $\phi$ . Let  $C[0, 1]$  be the collection of continuous function on  $[0, 1]$ . Clearly,  $C^{0,r}[0, 1] \subset C[0, 1]$  and, if  $\phi \in C^{0,r}[0, 1]$ , then  $\omega(\phi, \delta) \leq \|\phi\|_{0,r} \delta^r$ .

The following regularity assumptions are needed for the main results.

- (A1) The regression functions  $m_l(t) \in C^{0,1}[0, 1]$ ,  $l = 1, \dots, d$ .
- (A2) The set of random variables  $(T_{ij}, \varepsilon_{ij}, N_i, \xi_{ik,l}, X_{il})_{i=1, j=1, k=1, l=1}^{n, N_i, \infty, d}$  is a subset of variables  $(T_{ij}, \varepsilon_{ij}, N_i, \xi_{ik,l}, X_{il})_{i=1, j=1, k=1, l=1}^{\infty, \infty, \infty, d}$  consisting of independent random variables, in which all  $T_{ij}$ ’s i.i.d with  $T_{ij} \sim T$ , where  $T$  is a random variable with probability density function  $f(t)$ ;  $X_{il}$ ’s i.i.d for each  $l = 1, \dots, d$ ;  $N_i$ ’s i.i.d with  $N_i \sim N$ , where  $N > 0$  is a positive integer-valued random variable with  $\mathbf{E}\{N^{2r}\} \leq r!c_N^r$ ,  $r = 2, 3, \dots$ , for some constant  $c_N > 0$ . Variables  $(\xi_{ik,l})_{i=1, k=1, l=1}^{\infty, \infty, d}$  and  $(\varepsilon_{ij})_{i=1, j=1}^{\infty, \infty}$  are i.i.d  $N(0, 1)$ .
- (A3) The functions  $f(t)$ ,  $\sigma(t)$  and  $\phi_{k,l}(t) \in C^{0,r}[0, 1]$  for some  $r \in (0, 1]$  with  $f(t) \in [c_f, C_f]$ ,  $\sigma(t) \in [c_\sigma, C_\sigma]$ ,  $t \in [0, 1]$ , for constants  $0 < c_f \leq C_f < +\infty$ ,  $0 < c_\sigma \leq C_\sigma < +\infty$ .

- (A4) For  $l = 1, \dots, d, \sum_{k=1}^{\infty} \|\phi_{k,l}\|_{\infty} < +\infty$ , and  $G_l(t, t) \in [c_{G,l}, C_{G,l}], t \in [0, 1]$ , for constants  $0 < c_{G,l} \leq C_{G,l} < +\infty$ .
- (A5) There exist constants  $0 < c_{\mathbf{H}} \leq C_{\mathbf{H}} < +\infty$  and  $0 < c_{\eta} \leq C_{\eta} < +\infty$ , such that  $c_{\mathbf{H}}I_{d \times d} \leq \mathbf{H} = \{H_{ll'}\}_{l,l'=1}^d = \mathbf{E}(\mathbf{X}\mathbf{X}^T) \leq C_{\mathbf{H}}I_{d \times d}$ . For some  $\eta > 4$ ,  $l = 1, \dots, d, c_{\eta} \leq \mathbf{E}|X_l|^{8+\eta} \leq C_{\eta}$ .
- (A6) As  $n \rightarrow \infty$ , the number of interior knots  $N_s = \mathcal{O}(n^{\vartheta})$  for some  $\vartheta \in (1/3, 1/2)$  while  $N_s^{-1} = \mathcal{O}\{n^{-1/3}(\log(n))^{-1/3}\}$ . The subinterval length  $h_s \sim N_s^{-1}$ .

Assumptions (A1)–(A3) are common conditions used in the literature; see for example, [Ma et al. \(2012\)](#). Assumption (A1) controls the rate of convergence of the spline approximation  $\hat{m}_l, l = 1, \dots, d$ . The requirement of  $N_i$  in Assumption (A2) ensures that the observation times are randomly scattered, reflecting sparse and irregular designs. Assumption (A4) guarantees that the random variable  $\sum_{k=1}^{\infty} \xi_{ik,l}\phi_{k,l}(t)$  absolutely uniformly converges. Assumption (A5) is analog to Assumption (A2) in [Liu and Yang \(2010\)](#), ensuring that the  $X_{il}$ s are not multicollinear. Assumption (A6) describes the requirement of the growth rate of the dimension of the spline spaces relative to the sample size.

### A.2 Preliminaries

**Lemma 1** ([Bosq \(1998\)](#), Theorem 1.2). *Suppose that  $\{\xi_i\}_{i=1}^n$  are i.i.d with  $\mathbf{E}(\xi_1) = 0, \sigma^2 = \mathbf{E}\xi_1^2$ , and there exists  $c > 0$  such that for  $r = 3, 4, \dots, \mathbf{E}|\xi_1|^r \leq c^{r-2}r!\mathbf{E}\xi_1^2 < +\infty$ . Then for each  $n > 1, t > 0, P(|S_n| \geq \sqrt{n}\sigma t) \leq 2 \exp(-t^2(4 + 2ct/\sqrt{n}\sigma)^{-1})$ , in which  $S_n = \sum_{i=1}^n \xi_i$ .*

**Lemma 2** *Under Assumptions (A2)–(A6), we have*

$$A_{n,1} = \sup_{0 \leq J \leq N_s, 1 \leq l, l' \leq d} \frac{|\langle B_J X_l, B_J X_{l'} \rangle_{N_T} - \langle B_J X_l, B_J X_{l'} \rangle|}{\sqrt{\langle B_J X_l, B_J X_l \rangle} \sqrt{\langle B_J X_{l'}, B_J X_{l'} \rangle}} = \mathcal{O}_p\left(\sqrt{\frac{\log(n)}{nh_s}}\right),$$

where for any  $J = 0, \dots, N_s$  and  $l, l' = 1, \dots, d$ ,

$$\begin{aligned} \langle B_J X_l, B_J X_{l'} \rangle_{N_T} &= N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J^2(T_{ij}) X_{il} X_{il'}, \\ \langle B_J X_l, B_J X_{l'} \rangle &= \mathbf{E} \left\{ B_J^2(T_{ij}) X_{il} X_{il'} \right\} = H_{ll'}. \end{aligned}$$

*Proof* Let  $\omega_{i,J} = \omega_{i,J,l,l'} = \sum_{j=1}^{N_i} B_J^2(T_{ij}) X_{il} X_{il'}$ , then  $\mathbf{E}\omega_{i,J} = \mathbf{E}N_1 H_{ll'} \sim 1$  and  $\mathbf{E}(\omega_{i,J})^2 = \mathbf{E} \left\{ \sum_{j=1}^{N_i} B_J^2(T_{ij}) \right\}^2 \mathbf{E}(X_{il} X_{il'})^2 \sim h_s^{-1}$ . Next define a sequence  $D_n = n^{\alpha}$  with  $\alpha(4 + \eta/2) > 1$  and  $\sqrt{\log(n)} D_n n^{-1/2} h_s^{-1/2} \rightarrow 0, n^{1/2} h_s^{1/2} D_n^{-(3+\eta/2)} \rightarrow 0$ , which necessitates  $\eta > 2$  according to Assumption (A5). We make use of the following truncated and tail decomposition

$$X_{ill'} = X_{il} X_{il'} = X_{ill',1}^{D_n} + X_{ill',2}^{D_n},$$

where  $X_{ill',1}^{D_n} = X_{il}X_{il'}I\{|X_{il}X_{il'}| > D_n\}$ ,  $X_{ill',2}^{D_n} = X_{il}X_{il'}I\{|X_{il}X_{il'}| \leq D_n\}$ . Correspondingly, the truncated and tail parts of  $\omega_{i,J}$  are  $\omega_{i,J,m} = B_J^2(T_{ij})X_{ill',m}^{D_n}$ ,  $m = 1, 2$ . According to Assumption (A5), for any  $l, l' = 1, \dots, d$ ,

$$\sum_{n=1}^{\infty} P\{|X_{nl}X_{nl'}| > D_n\} \leq \sum_{n=1}^{\infty} \frac{E|X_{nl}X_{nl'}|^{4+\eta/2}}{D_n^{4+\eta/2}} \leq C_\eta \sum_{n=1}^{\infty} D_n^{-(4+\eta/2)} < \infty.$$

By Borel–Cantelli Lemma, one has  $\sum_{j=1}^{N_i} B_J^2(T_{ij})X_{ill',1}^{D_n} = 0, a.s..$  So we obtain

$$\sup_{J,l,l'} \left| n^{-1} \sum_{i=1}^n \omega_{i,J,1} \right| = \mathcal{O}_{a.s.}(n^{-k}), \quad k \geq 1,$$

and

$$\begin{aligned} E\omega_{i,J,1} &= E\left(X_{ill',1}^{D_n}\right) E\left\{\sum_{j=1}^{N_i} B_J^2(T_{ij})\right\} \\ &\leq D_n^{-(3+\eta/2)} E|X_{il}X_{il'}|^{4+\eta/2} EN_1 EB_J^2(T_{ij}) \leq cD_n^{-(3+\eta/2)}. \end{aligned}$$

Next we considerate the truncated part  $\omega_{i,J,2}$ . For large  $n$ ,  $E(\omega_{i,J,2}) = E(\omega_{i,J}) - E(\omega_{i,J,1}) \sim 1$ ,  $E(\omega_{i,J,2})^2 = E(\omega_{i,J})^2 - E(\omega_{i,J,1})^2 \sim h_s^{-1}$ . Define  $\omega_{i,J,2}^* = \omega_{i,J,2} - E(\omega_{i,J,2})$ , then  $E\omega_{i,J,2}^* = 0$ , and

$$\begin{aligned} E(\omega_{i,J,2}^*)^2 &= E(\omega_{i,J,2})^2 - (E\omega_{i,J,2})^2 = E\left\{\sum_{j=1}^{N_i} B_J^2(T_{ij})X_{ill',2}^{D_n}\right\}^2 - \mathcal{U}(1) \\ &= E\left(X_{ill',2}^{D_n}\right)^2 E\left\{\sum_{j=1}^{N_i} B_J^2(T_{ij})\right\}^2 - \mathcal{U}(1). \end{aligned}$$

Note that

$$\begin{aligned} E\left(X_{ill',2}^{D_n}\right)^2 E\left\{\sum_{j=1}^{N_i} B_J^2(T_{ij})\right\}^2 &\geq \left\{E(X_{ill'})^2 - E\left(X_{ill',1}^{D_n}\right)^2\right\} E\left\{\sum_{j=1}^{N_i} B_J^4(T_{ij})\right\} \\ &\geq \left\{E(X_{ill'})^2 - \mathcal{U}(1)\right\} EN_1 EB_J^4(T_{ij}). \end{aligned}$$

Thus, there exists  $c_\omega$  such that for large  $n$ ,  $E(\omega_{i,J,2}^*)^2 \geq c_\omega E(X_{ill'})^2 h_s^{-1}$ . Next for any  $r > 2$

$$\begin{aligned}
\mathbf{E} |\omega_{i,J,2}^*|^r &= \mathbf{E} |\omega_{i,J,2} - \mathbf{E}(\omega_{i,J,2})|^r \leq 2^{r-1} (\mathbf{E} |\omega_{i,J,2}|^r + |\mathbf{E}(\omega_{i,J,2})|^r) \\
&= 2^{r-1} \left\{ \mathbf{E} |X_{ill',2}^{D_n}|^r \mathbf{E} \left| \sum_{j=1}^{N_i} B_J^2(T_{ij}) \right|^r + \mathcal{U}(1) \right\} \\
&= 2^{r-1} \left[ \mathbf{E} |X_{ill',2}^{D_n}|^r \mathbf{E} \left\{ \sum_{0 \leq r_1, \dots, r_{N_i} \leq r}^{r_1 + \dots + r_{N_i} = r} \binom{r}{r_1 \dots r_{N_i}} \prod_{j=1}^{N_i} \mathbf{E} B_J^{2r_j}(T_{ij}) \right\} + \mathcal{U}(1) \right],
\end{aligned}$$

then there exists  $C_\omega > 0$  such that for any  $r > 2$  and large  $n$ ,

$$\begin{aligned}
\mathbf{E} |\omega_{i,J,2}^*|^r &\leq 2^{r-1} \left[ D_n^{r-2} \mathbf{E} (X_{ill'})^2 \mathbf{E} \left\{ N_1^r \max_{j=1}^{N_i} \prod_{j=1}^{N_i} \mathbf{E} B_J^{2r_j}(T_{ij}) \right\} + \mathcal{U}(1) \right] \\
&\leq 2^{r-1} \left[ D_n^{r-2} \mathbf{E} (X_{ill'})^2 (\mathbf{E} N_1^r) C_B h_s^{1-r} + \mathcal{U}(1) \right] \\
&\leq 2^r D_n^{r-2} (c_N^r r!)^{1/2} C_B h_s^{2-r} c_\omega^{-1} \mathbf{E} (\omega_{i,J,2}^*)^2 \\
&\leq (C_\omega D_n h_s^{-1})^{r-2} r! \mathbf{E} (\omega_{i,J,2}^*)^2,
\end{aligned}$$

which implies that  $\{\omega_{i,J,2}^*\}_{i=1}^n$  satisfies Cramér's condition with constant  $C_\omega D_n h_s^{-1}$ . Applying Lemma 1 to  $\sum_{i=1}^n \omega_{i,J,2}^*$ , for  $r > 2$  and any large enough  $\delta > 0$ ,  $P \left\{ \left| n^{-1} \sum_{i=1}^n \omega_{i,J,2}^* \right| \geq \delta (nh_s)^{-1/2} (\log(n))^{1/2} \right\}$  is bounded by

$$2 \exp \left\{ \frac{-\delta^2 (\log(n))}{4 + 2C_\omega D_n h_s^{-1} \delta (\log(n))^{1/2} n^{-1/2} h_s^{1/2}} \right\} \leq 2n^{-8}.$$

Hence

$$\sum_{n=1}^{\infty} P \left\{ \sup_{0 \leq J \leq N_s, 1 \leq l, l' \leq d} \left| n^{-1} \sum_{i=1}^n \omega_{i,J,2}^* \right| \geq \delta (nh_s)^{-1/2} (\log(n))^{1/2} \right\} < \infty.$$

Thus,  $\sup_{J,l,l'} \left| n^{-1} \sum_{i=1}^n \omega_{i,J,2}^* \right| = \mathcal{O}_{a.s.} \{ (nh_s)^{-1/2} (\log(n))^{1/2} \}$  as  $n \rightarrow \infty$  by Borel–Cantelli Lemma. Furthermore,

$$\begin{aligned}
&\sup_{J,l,l'} \left| n^{-1} \sum_{i=1}^n \omega_{i,J} - \mathbf{E} \omega_{i,J} \right| \\
&\leq \sup_{J,l,l'} \left| n^{-1} \sum_{i=1}^n \omega_{i,J,1} \right| + \sup_{J,l,l'} \left| n^{-1} \sum_{i=1}^n \omega_{i,J,2}^* \right| + \sup_{J,l,l'} |\mathbf{E} \omega_{i,J,1}|
\end{aligned}$$

$$\begin{aligned}
 &= \mathcal{U}_{a.s.} \left( n^{-k} \right) + \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\} + \mathcal{U} \left( D_n^{-(3+\eta/2)} \right) \\
 &= \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}.
 \end{aligned}$$

Finally, we notice that

$$\begin{aligned}
 \sup_{J,l,l'} \left| \langle B_J X_l, B_J X_{l'} \rangle_{N_T} - \langle B_J X_l, B_J X_{l'} \rangle \right| &= \sup_{J,l,l'} \left| \left( n N_T^{-1} \right) n^{-1} \sum_{i=1}^n \omega_{i,J} - (\mathbf{E} N_1)^{-1} \mathbf{E} \omega_{i,J} \right| \\
 &\leq \sup_{J,l,l'} (\mathbf{E} N_1)^{-1} \left| (n \mathbf{E} N_1) N_T^{-1} - 1 \right| \left| n^{-1} \sum_{i=1}^n \omega_{i,J} \right| + \sup_{J,l,l'} (\mathbf{E} N_1)^{-1} \left| n^{-1} \sum_{i=1}^n \omega_{i,J} - \mathbf{E} \omega_{i,J} \right| \\
 &= \mathcal{O}_p \left( n^{-1/2} \right) + \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\} = \mathcal{O}_p \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\},
 \end{aligned}$$

and  $\langle B_J X_l, B_J X_l \rangle = H_{ll} = \mathcal{U}(1)$ . Hence,  $A_{n,1} = \mathcal{O}_p \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}$ .  $\square$

For the random matrix  $\hat{\mathbf{V}}_J$  defined in (18), the lemma below shows that its inverse can be approximated by the inverse of a deterministic matrix  $\mathbf{H} = \mathbf{E}(\mathbf{X}\mathbf{X}^T)$ .

**Lemma 3** Under Assumptions (A2) and (A4)–(A6), for any  $J = 0, \dots, N_s$ , we have

$$\hat{\mathbf{V}}_J^{-1} = \mathbf{H}^{-1} + \mathcal{O}_p \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}. \tag{25}$$

*Proof* By Lemma 2, we have

$$\left\| \hat{\mathbf{V}}_J - \mathbf{H} \right\|_{\infty} = \mathcal{O}_p \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}.$$

Using the fact that for any matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,

$$(\mathbf{A} + h\mathbf{B})^{-1} = \mathbf{A}^{-1} - h\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + \mathcal{O}(h^2),$$

we obtain (25).  $\square$

The next lemma implies that the difference between  $\tilde{\xi}(t)$  and  $\hat{\xi}(t)$  and the difference between  $\tilde{\epsilon}(t)$  and  $\hat{\epsilon}(t)$  are both negligible uniformly over  $t \in [0, 1]$ .

**Lemma 4** Under Assumption (A2)–(A6), for  $\tilde{\xi}(t)$ ,  $\tilde{\epsilon}(t)$  given in (36), (37) and  $\hat{\xi}(t)$ ,  $\hat{\epsilon}(t)$  given in (38), (39), as  $n \rightarrow \infty$ , we have

$$\sup_{t \in [0,1]} \left\| \tilde{\xi}(t) - \hat{\xi}(t) \right\|_{\infty} = \mathcal{O}_p \left\{ n^{-1} h_s^{-3/2} \log(n) \right\}, \tag{26}$$

$$\sup_{t \in [0,1]} \left\| \tilde{\epsilon}(t) - \hat{\epsilon}(t) \right\|_{\infty} = \mathcal{O}_p \left\{ n^{-1} h_s^{-3/2} \log(n) \right\}. \tag{27}$$

*Proof* Comparing the equations of  $\tilde{\xi}(t)$  and  $\hat{\xi}(t)$  given in (A.2) and (A.4), we let

$$\frac{1}{N_T} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J(T_{ij}) X_{il} \sum_{l''=1}^d \sum_{k=1}^{\infty} \xi_{ik,l''} \phi_{k,l''}(T_{ij}) X_{il''} = \frac{n}{N_T} \sum_{l''=1}^d \sum_{i=1}^n \Omega_{i,J,l'',l}.$$

where  $\Omega_{i,J,l'',l} = \Omega_i = n^{-1} \left[ X_{il} X_{il''} \sum_{k=1}^{\infty} \left\{ \sum_{j=1}^{N_i} B_J(T_{ij}) \phi_{k,l''}(T_{ij}) \right\} \xi_{ik,l''} \right]$ . Note that  $E\Omega_i = 0$  and

$$\begin{aligned} \sigma_{\Omega_i,n}^2 &= E \left( \Omega_i^2 \mid (T_{ij}, N_i, X_{il})_{i=1,j=1,l=1}^{n,N_i,d} \right) \\ &= n^{-2} \left[ X_{il} X_{il''} \sum_{k=1}^{\infty} \left\{ \sum_{j=1}^{N_i} B_J(T_{ij}) \phi_{k,l''}(T_{ij}) \right\}^2 \right] \\ &\leq n^{-2} \left\{ X_{il}^2 X_{il''}^2 \sum_{k=1}^{\infty} N_i \sum_{j=1}^{N_i} B_J^2(T_{ij}) \phi_{k,l''}^2(T_{ij}) \right\} \\ &= n^{-2} \left\{ X_{il}^2 X_{il''}^2 N_i \sum_{j=1}^{N_i} B_J^2(T_{ij}) G_{l''}(T_{ij}, T_{ij}) \right\} \\ &\leq C n^{-2} h_s^{-1} X_{il}^2 X_{il''}^2 N_i^2. \end{aligned}$$

Given  $(T_{ij}, N_i, X_{il})_{i=1,j=1,l=1}^{n,N_i,d}$ ,  $\{\sigma_{\Omega_i,n}^{-1} \Omega_i\}_{i=1}^n$  are i.i.d  $N(0, 1)$ . It is easy to show that for any large enough  $\delta > 0$ ,

$$\begin{aligned} P \left\{ \frac{|\sum_{i=1}^n \Omega_i|}{\sqrt{\sum_{i=1}^n \sigma_{\Omega_i,n}^2}} \geq \delta \sqrt{\log(n)} \mid (T_{ij}, N_i, X_{il})_{i=1,j=1,l=1}^{n,N_i,d} \right\} \\ \leq 2 \exp \left\{ -\frac{1}{2} \delta^2 \log(n) \right\} \leq 2n^{-8}, \end{aligned}$$

$$P \left[ \left| \sum_{i=1}^n \Omega_i \right| \geq \delta \left\{ \frac{C \log(n)}{nh_s} n^{-1} \sum_{i=1}^n X_{il}^2 X_{il''}^2 N_i^2 \right\}^{1/2} \mid (T_{ij}, N_i, X_{il})_{i=1,j=1,l=1}^{n,N_i,d} \right] \leq 2n^{-8}.$$

Note that  $n^{-1} \sum_{i=1}^n X_{il}^2 X_{il''}^2 N_i^2 = \mathcal{O}_p(1)$ , hence

$$\sum_{n=1}^{\infty} P \left\{ \sup_{0 \leq J \leq N_s, 1 \leq l, l'' \leq d} \left| \sum_{i=1}^n \Omega_{i,J,l'',l} \right| \geq \delta (nh_s)^{-1/2} (\log(n))^{1/2} \right\} < \infty.$$

Thus,  $\sup_{J,l,l''} |\sum_{i=1}^n \Omega_{i,J,l'',l}| = \mathcal{O}_p \{ (nh_s)^{-1/2} (\log(n))^{1/2} \}$  as  $n \rightarrow \infty$  by Borel–Cantelli Lemma. It follows that  $\sup_{J,l} \left| n N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \Omega_{i,J,l'',l} \right| =$

$\mathcal{O}_p\{(nh_s)^{-1/2}(\log(n))^{1/2}\}$ . Finally, according to Lemma 25, we obtain (26). (27) is proved similarly.  $\square$

Denote the inverse matrix of  $\mathbf{H}$  by  $\mathbf{H}^{-1} = \{z_{ll'}\}_{l,l'=1}^d$ . For any  $l = 1, \dots, d$ , we rewrite the  $l$ th element of  $\hat{\xi}_l(t)$  and  $\hat{\varepsilon}_l(t)$  in (38) and (39) as the following

$$\hat{\xi}_l(t) = c_{J(t),n}^{-1/2} N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J(t),l'',l} \xi_{ik,l''}, \tag{28}$$

$$\hat{\varepsilon}_l(t) = c_{J(t),n}^{-1/2} N_T^{-1} \sum_{i=1}^n \sum_{j=1}^N R_{ij,\varepsilon,J(t),l} \varepsilon_{ij}, \tag{29}$$

where for any  $0 \leq J \leq N_s$ ,

$$R_{ik,\xi,J,l'',l} = \left( \sum_{l'=1}^d z_{ll'} X_{il'} X_{il''} \right) \left\{ \sum_{j=1}^{N_i} B_J(T_{ij}) \phi_{k,l''}(T_{ij}) \right\}, \tag{30}$$

$$R_{ij,\varepsilon,J,l} = \left( \sum_{l'=1}^d z_{ll'} X_{il'} \right) B_J(T_{ij}) \sigma(T_{ij}). \tag{31}$$

Further denote

$$S_{ill''} = \left( \sum_{l'=1}^d z_{ll'} X_{il'} X_{il''} \right)^2, \quad s_{ill''} = \mathbf{E}(S_{ill''}), \quad 1 \leq l, l'' \leq d. \tag{32}$$

**Lemma 5** Under Assumptions (A2)–(A6), for  $R_{ik,\xi,J,l'',l}$ ,  $R_{ij,\varepsilon,J,l}$  in (30), (31),

$$\begin{aligned} \mathbf{E} \left( \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 \right) &= c_{J,n}^{-1} s_{ill''} \left[ (\mathbf{E}N_1) \int b_J(u) G_{l''}(u, u) f(u) du \right. \\ &\quad \left. + \mathbf{E}\{N_1(N_1 - 1)\} \int b_J(u) b_J(v) G_{l''}(u, v) f(u) f(v) dudv \right], \end{aligned}$$

$$\mathbf{E} R_{ij,\varepsilon,J,l}^2 = c_{J,n}^{-1} z_{ll} \int b_J(u) \sigma^2(u) f(u) du,$$

for  $0 \leq J \leq N_s$  and  $0 \leq l, l'' \leq d$ . In addition, there exist  $0 < c_R < C_R < \infty$ , such that

$$c_R s_{ill''} \leq \mathbf{E} \left( \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 \right) \leq C_R s_{ill''}, \quad c_R \leq \mathbf{E} R_{ij,\varepsilon,J,l}^2 \leq C_R,$$



for  $0 \leq J \leq N_s$ ,  $0 \leq l, l'' \leq d$ , and as  $n \rightarrow \infty$

$$\begin{aligned} A_{n,\xi} &= \sup_{J,l'',l} \left| n^{-1} \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 - \mathbb{E} \left( \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 \right) \right| \\ &= \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}, \\ A_{n,\varepsilon} &= \sup_{J,l} \left| N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} R_{ij,\varepsilon,J,l}^2 - \mathbb{E} R_{ij,\varepsilon,J,l}^2 \right| = \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}. \end{aligned}$$

*Proof* By independence of  $\{T_{ij}\}_{j=1}^{\infty}$ ,  $\{X_{il}\}_{l=1}^d$ ,  $N_i$ , the definition of  $B_J$  and (32),

$$\begin{aligned} \mathbb{E} \left( \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 \right) &= \mathbb{E} (S_{ill''}) \mathbb{E} \sum_{k=1}^{\infty} \left\{ \sum_{j=1}^{N_i} B_J (T_{ij}) \phi_{k,l''} (T_{ij}) \right\}^2 \\ &= s_{ll''} \mathbb{E} \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J (T_{ij}) B_J (T_{ij'}) G_{l''} (T_{ij}, T_{ij'}) \\ &= s_{ll''} c_{J,n}^{-1} \left\{ (\mathbb{E} N_1) \int b_J (u) G_{l''} (u, u) f (u) du \right. \\ &\quad \left. + \mathbb{E} \{N_1(N_1 - 1)\} \int b_J (u) b_J (v) G_{l''} (u, v) f (u) f (v) dudv \right\}, \end{aligned}$$

thus there exist constants  $0 < c_R < C_R < \infty$  such that  $c_R s_{ll''} \leq \mathbb{E} \left( \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 \right) \leq C_R s_{ll''}$ ,  $0 \leq J \leq N_s$ ,  $0 \leq l, l'' \leq d$ .

If  $s_{ll''} = 0$ , one has  $S_{ill''} = 0$ , almost surely. Hence  $n^{-1} \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2 = 0$ , almost surely. In the case of  $s_{ll''} > 0$ , let  $\zeta_{i,J} = \zeta_{i,J,l'',l} = \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l}^2$  for brevity. Under Assumption (A5), it is easy to verify that

$$0 < s_{ll''}^2 \leq \mathbb{E} (S_{ill''})^2 \leq d^3 \sum_{l'=1}^d \mathbb{E} |z_{ll'} X_{il'} X_{il''}|^4 \leq d^3 \sum_{l'=1}^d z_{ll'} \left\{ \mathbb{E} |X_{il'}|^8 \mathbb{E} |X_{il''}|^8 \right\}^{1/2} < \infty.$$

So for large  $n$ ,

$$\begin{aligned} \mathbb{E} (\zeta_{i,J})^2 &= \mathbb{E} \left\{ (S_{ill''})^2 \left( \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J (T_{ij}) B_J (T_{ij'}) G_{l''} (T_{ij}, T_{ij'}) \right)^2 \right\} \\ &\geq \mathbb{E} (S_{ill''})^2 \frac{1}{4} c_{G,l''}^2 \mathbb{E} \left\{ \sum_{j=1}^{N_i} B_J (T_{ij}) \right\}^4 \geq c \mathbb{E} \sum_{j=1}^{N_i} B_J^4 (T_{ij}) \geq ch_s^{-1}, \end{aligned}$$

and

$$\begin{aligned} E(\zeta_{i,J})^2 &\leq E(S_{ill''})^2 4C_{G,l''}^2 E\left\{\sum_{j=1}^{N_i} B_J(T_{ij})\right\}^4 \\ &\leq cE\left[N_1^3 \sum_{j=1}^{N_i} EB_J^4(T_{ij}) \middle| N_1\right] \leq cEN_1^4 EB_J^4(T_{ij}) \leq ch_s^{-1}. \end{aligned}$$

Define a sequence  $D_n = n^\alpha$  that satisfies  $\alpha(2 + \eta/4) > 1$ ,  $D_n n^{-1/2} h_s^{-1/2} (\log(n))^{1/2} \rightarrow 0$ ,  $n^{1/2} h_s^{1/2} D_n^{-(1+\eta/4)} \rightarrow 0$ , which requires  $\eta > 4$  provided by Assumption (A5). We make use of the following truncated and tail decomposition

$$S_{ill''} = \sum_{l'=1}^d \sum_{l'''=1}^d z_{ll'} z_{ll''} X_{il'} X_{il''} X_{il''}^2 = S_{ill'',1}^{D_n} + S_{ill'',2}^{D_n},$$

where

$$\begin{aligned} S_{ill'',1}^{D_n} &= \sum_{l'=1}^d \sum_{l'''=1}^d z_{ll'} z_{ll''} X_{il'} X_{il''} X_{il''}^2 I\left\{\left|X_{il'} X_{il''} X_{il''}^2\right| > D_n\right\}, \\ S_{ill'',2}^{D_n} &= \sum_{l'=1}^d \sum_{l'''=1}^d z_{ll'} z_{ll''} X_{il'} X_{il''} X_{il''}^2 I\left\{\left|X_{il'} X_{il''} X_{il''}^2\right| \leq D_n\right\}. \end{aligned}$$

Define correspondingly the truncated and tail parts of  $\zeta_{i,J}$  as

$$\zeta_{i,J,m} = S_{ill'',m}^{D_n} \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J(T_{ij}) B_J(T_{ij'}) G_{l''}(T_{ij}, T_{ij'}), \quad m = 1, 2.$$

According to Assumption (A5), for any  $l', l'', l''' = 1, \dots, d$ ,

$$\sum_{n=1}^{\infty} P\left\{\left|X_{nl'} X_{nl''} X_{nl''}^2\right| > D_n\right\} \leq \sum_{n=1}^{\infty} \frac{E\left|X_{nl'} X_{nl''} X_{nl''}^2\right|^{2+\eta/4}}{D_n^{2+\eta/4}} \leq C_\eta \sum_{n=1}^{\infty} D_n^{-(2+\eta/4)} < \infty.$$

Borel–Cantelli Lemma implies that

$$\begin{aligned} P\left\{\omega \mid \exists N(\omega), \left|X_{nl'} X_{nl''} X_{nl''}^2(\omega)\right| \leq D_n \text{ for } n > N(\omega)\right\} &= 1, \\ P\left\{\omega \mid \exists N(\omega), \left|X_{il'} X_{il''} X_{il''}^2(\omega)\right| \leq D_n, i = 1, \dots, n \text{ for } n > N(\omega)\right\} &= 1, \\ P\left\{\omega \mid \exists N(\omega), I\left\{\left|X_{il'} X_{il''} X_{il''}^2(\omega)\right| > D_n\right\} = 0, i = 1, \dots, n \text{ for } n > N(\omega)\right\} &= 1. \end{aligned}$$

Furthermore, one has

$$n^{-1} \sum_{i=1}^n \left\{ S_{ill'',1}^{D_n} \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J(T_{ij}) B_J(T_{ij'}) G_{l''}(T_{ij}, T_{ij'}) \right\} = 0, \quad a.s.$$

Therefore, one has

$$\sup_{J,l,l''} \left| n^{-1} \sum_{i=1}^n \zeta_{i,J,1} \right| = \mathcal{O}_{a.s.} \left( n^{-k} \right), \quad k \geq 1.$$

Notice that

$$\begin{aligned} \mathbb{E} \left( S_{ill'',1}^{D_n} \right) &= \mathbb{E} \left[ \sum_{l'=1}^d \sum_{l''=1}^d z_{ll'} z_{ll''} X_{il'} X_{il''} X_{il''}^2 I \left\{ \left| X_{il'} X_{il''} X_{il''}^2 \right| > D_n \right\} \right] \\ &\leq D_n^{-(1+\eta/4)} \sum_{l'=1}^d \sum_{l''=1}^d z_{ll'} z_{ll''} \mathbb{E} \left| X_{il'} X_{il''} X_{il''}^2 \right|^{2+\eta/4} \\ &\leq c D_n^{-(1+\eta/4)}. \end{aligned}$$

So for large  $n$ ,

$$\begin{aligned} \mathbb{E} \left( \zeta_{i,J,1} \right) &= \mathbb{E} \left( S_{ill'',1}^{D_n} \right) \mathbb{E} \left\{ \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J(T_{ij}) B_J(T_{ij'}) G_{l''}(T_{ij}, T_{ij'}) \right\} \\ &\leq c D_n^{-(1+\eta/4)} 2C_{G,l''} \mathbb{E} \left\{ \sum_{j=1}^{N_i} B_J(T_{ij}) \right\}^2 \\ &\leq c D_n^{-(1+\eta/4)} \mathbb{E} \left( N_1^2 \right) \mathbb{E} B_J^2(T_{ij}) \\ &\leq c D_n^{-(1+\eta/4)}. \end{aligned}$$

Next we considerate the truncated part  $\zeta_{i,J,2}$ . For large  $n$ ,  $\mathbb{E}(\zeta_{i,J,2}) = \mathbb{E}(\zeta_{i,J}) - \mathbb{E}(\zeta_{i,J,1}) \sim 1$ ,  $\mathbb{E}(\zeta_{i,J,2})^2 = \mathbb{E}(\zeta_{i,J})^2 - \mathbb{E}(\zeta_{i,J,1})^2 \sim h_s^{-1}$ . Define  $\zeta_{i,J,2}^* = \zeta_{i,J,2} - \mathbb{E}(\zeta_{i,J,2})$ , then  $\mathbb{E}\zeta_{i,J,2}^* = 0$ , and there exist  $c_\zeta, C_\zeta > 0$  such that for  $r > 2$  and large  $n$ ,

$$\begin{aligned} \mathbb{E}(\zeta_{i,J,2}^*)^2 &= \mathbb{E} \left| S_{ill'',2}^{D_n} \right|^2 \mathbb{E} \left| \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J(T_{ij}) B_J(T_{ij'}) G_{l''}(T_{ij}, T_{ij'}) \right|^2 - \left( \mathbb{E}\zeta_{i,J,2} \right)^2 \\ &\geq \left\{ \mathbb{E} |S_{ill'',2}^{D_n}|^2 - \mathbb{E} |S_{ill'',1}^{D_n}|^2 \right\} \frac{1}{4} c_{G,l''}^2 \mathbb{E} \left\{ \sum_{j=1}^{N_i} B_J(T_{ij}) \right\}^4 - \mathcal{U}(1) \end{aligned}$$

$$\begin{aligned} &\geq \left\{ \mathbf{E} |S_{ill''}|^2 - \mathcal{U}(1) \right\} \frac{1}{4} c_{G,l''}^2 \mathbf{E} \left\{ \sum_{j=1}^{N_i} B_J^4(T_{ij}) \right\} - \mathcal{U}(1) \\ &\geq \frac{1}{2} \mathbf{E} |S_{ill''}|^2 \frac{1}{4} c_{G,l''}^2 \mathbf{E} N_1 \mathbf{E} B_J^4(T_{ij}) - \mathcal{U}(1) \\ &\geq c_\zeta \mathbf{E} |S_{ill''}|^2 h_s^{-1}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{E} |\zeta_{i,J,2}^*|^r &= \mathbf{E} |\zeta_{i,J,2} - \mathbf{E}(\zeta_{i,J,2})|^r \leq 2^{r-1} (\mathbf{E} |\zeta_{i,J,2}|^r + |\mathbf{E}(\zeta_{i,J,2})|^r) \\ &= 2^{r-1} \left\{ \mathbf{E} |S_{ill''}^{D_n}|^r \mathbf{E} \left| \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J(T_{ij}) B_J(T_{ij'}) G_{l''}(T_{ij}, T_{ij'}) \right|^r + \mathcal{U}(1) \right\} \\ &\leq 2^{r-1} \left[ (cD_n)^{r-2} \mathbf{E} |S_{ill''}|^2 (2C_{G,l''})^r \mathbf{E} \left\{ \sum_{j=1}^{N_i} B_J(T_{ij}) \right\}^{2r} + \mathcal{U}(1) \right] \\ &\leq 2^{r-1} \left[ (cD_n)^{r-2} \mathbf{E} |S_{ill''}|^2 (2C_{G,l''})^r (\mathbf{E} N_1^{2r}) C_B h_s^{1-r} + \mathcal{U}(1) \right] \\ &\leq 2^r (cD_n)^{r-2} (2C_{G,l''})^r c_N^r r! C_B h_s^{2-r} c_\zeta^{-1} \mathbf{E} (\zeta_{i,J,2}^*)^2 \\ &\leq (C_\zeta D_n h_s^{-1})^{r-2} r! \mathbf{E} (\zeta_{i,J,2}^*)^2, \end{aligned}$$

which implies that  $\{\zeta_{i,J,2}^*\}_{i=1}^n$  satisfies Cramér’s condition. Applying Lemma 1 to  $\sum_{i=1}^n \zeta_{i,J,2}^*$ , for  $r > 2$  and any large enough  $\delta > 0$ ,

$$\begin{aligned} &P \left\{ \left| n^{-1} \sum_{i=1}^n \zeta_{i,J,2}^* \right| \geq \delta (nh_s)^{-1/2} (\log(n))^{1/2} \right\} \\ &\leq 2 \exp \left\{ \frac{-\delta^2 \log(n)}{4 + 2C_\zeta D_n h_s^{-1} \delta (\log(n))^{1/2} n^{-1/2} h_s^{1/2}} \right\} \leq 2n^{-8}. \end{aligned}$$

Hence

$$\sum_{n=1}^\infty P \left\{ \sup_{J,l'',l} \left| n^{-1} \sum_{i=1}^n \zeta_{i,J,2}^* \right| \geq \delta (nh_s)^{-1/2} (\log(n))^{1/2} \right\} < \infty.$$

Thus,  $\sup_{J,l'',l} \left| n^{-1} \sum_{i=1}^n \zeta_{i,J,2}^* \right| = \mathcal{O}_{a.s.} \{ (nh_s)^{-1/2} (\log(n))^{1/2} \}$  as  $n \rightarrow \infty$  by the Borel–Cantelli lemma. Furthermore, we have

$$A_{n,\xi} \leq \sup_{J,l,l''} \left| n^{-1} \sum_{i=1}^n \zeta_{i,J,1} \right| + \sup_{J,l'',l} \left| n^{-1} \sum_{i=1}^n \zeta_{i,J,2}^* \right| + \sup_{J,l'',l} |\mathbf{E}(\zeta_{i,J,1})|$$

$$\begin{aligned}
&= \mathcal{U}_{a.s.} \left( n^{-k} \right) + \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\} + \mathcal{U} \left( D_n^{-(1+\eta/4)} \right) \\
&= \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}.
\end{aligned}$$

The properties of  $R_{ij,\varepsilon,J,l}$  are obtained similarly.  $\square$

Next define two  $d \times d$  matrices

$$\begin{aligned}
\Gamma_{\xi,n}(t) &= c_{J(t),n}^{-1} N_T^{-2} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} \left\{ \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) \phi_{k,l''}(T_{ij}) \right\}^2 X_{il''}^2 \mathbf{X}_i \mathbf{X}_i^{\top}, \\
\Gamma_{\varepsilon,n}(t) &= c_{J(t),n}^{-1} N_T^{-2} \sum_{i=1}^n \sum_{j=1}^{N_i} B_{J(t)}^2(T_{ij}) \sigma^2(T_{ij}) \mathbf{X}_i \mathbf{X}_i^{\top}.
\end{aligned}$$

**Lemma 6** For any  $t \in \mathbb{R}$ , the conditional covariance matrices of  $\hat{\xi}(t)$  and  $\hat{\varepsilon}(t)$  on  $(T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d}$  are

$$\begin{aligned}
\Sigma_{\xi,n}(t) &= \mathbb{E} \left\{ \hat{\xi}(t) \hat{\xi}^{\top}(t) \mid (T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d} \right\} = \mathbf{H}^{-1} \Gamma_{\xi,n}(t) \mathbf{H}^{-1}, \\
\Sigma_{\varepsilon,n}(t) &= \mathbb{E} \left\{ \hat{\varepsilon}(t) \hat{\varepsilon}^{\top}(t) \mid (T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d} \right\} = \mathbf{H}^{-1} \Gamma_{\varepsilon,n}(t) \mathbf{H}^{-1},
\end{aligned}$$

and with  $\Sigma_n(t)$  defined in (7),

$$\sup_{t \in [0,1]} \left\| \left\{ \Sigma_{\xi,n}(t) + \Sigma_{\varepsilon,n}(t) \right\} - \Sigma_n(t) \right\|_{\infty} = \mathcal{O}_{a.s.} \left\{ n^{-3/2} h_s^{-3/2} (\log(n))^{1/2} \right\}. \quad (33)$$

*Proof* Note that

$$\begin{aligned}
\hat{\xi}(t) \hat{\xi}^{\top}(t) &= c_{J(t),n}^{-1} \mathbf{H}^{-1} \left\{ \frac{1}{N_T^2} \sum_{i=1}^n \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) X_{il} \sum_{l''=1}^d \sum_{k=1}^{\infty} \xi_{ik,l''} \phi_{k,l''}(T_{ij}) X_{il''} \right. \\
&\quad \left. \times \sum_{i=1}^n \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) X_{il'} \sum_{l''=1}^d \sum_{k=1}^{\infty} \xi_{ik,l''} \phi_{k,l''}(T_{ij}) X_{il''} \right\}_{l,l'=1}^d \mathbf{H}^{-1}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\Sigma_{\xi,n}(t) &= \mathbb{E} \left\{ \hat{\xi}(t) \hat{\xi}^{\top}(t) \mid (T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d} \right\} = c_{J(t),n}^{-1} \mathbf{H}^{-1} \\
&\quad \times \left[ N_T^{-2} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} \left\{ \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) \phi_{k,l''}(T_{ij}) \right\}^2 X_{il''}^2 \mathbf{X}_i \mathbf{X}_i^{\top} \right] \mathbf{H}^{-1} \\
&= \mathbf{H}^{-1} \Gamma_{\xi,n}(t) \mathbf{H}^{-1}.
\end{aligned}$$

Similarly, we can derive the conditional covariance matrix of  $\hat{\boldsymbol{\epsilon}}(t)$ . Next let

$$\Psi_{ik,\xi,J,l,l',l''} = \left\{ \sum_{j=1}^{N_i} B_J(T_{ij})\phi_{k,l''}(T_{ij}) \right\}^2 X_{il''}^2 X_{il} X_{il'},$$

$$\Psi_{ij,\varepsilon,J,l,l'} = B_J^2(T_{ij})\sigma^2(T_{ij}) X_{il} X_{il'}.$$

Similar to the proof of Lemma 5,

$$\begin{aligned} \mathbb{E}\left(\sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''}\right) &= c_{J,n}^{-1} \mathbb{E}\left(X_{il''}^2 X_{il} X_{il'}\right) \left[ (\mathbb{E}N_1) \int_{\chi_J} G_{l''}(u, u) f(u) du \right. \\ &\quad \left. + \mathbb{E}\{N_1(N_1 - 1)\} \int_{\chi_J \times \chi_J} G_{l''}(u, v) f(u) f(v) dudv \right], \\ \mathbb{E}\Psi_{ij,\varepsilon,J,l,l'} &= c_{J,n}^{-1} \mathbb{E}(X_{il} X_{il'}) \int_{\chi_J} \sigma^2(u) f(u) du, \end{aligned}$$

and as  $n \rightarrow \infty$ ,

$$\begin{aligned} \sup_{J,l,l',l''} \left| n^{-1} \sum_{i=1}^n \sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''} - \mathbb{E}\left(\sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''}\right) \right| \\ = \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}, \end{aligned}$$

$$\sup_{J,l,l'} \left| N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} \Psi_{ij,\varepsilon,J,l,l'} - \mathbb{E}\Psi_{ij,\varepsilon,J,l,l'} \right| = \mathcal{O}_{a.s.} \left\{ (nh_s)^{-1/2} (\log(n))^{1/2} \right\}.$$

Furthermore,

$$\begin{aligned} \sup_{J,l,l',l''} \left| N_T^{-2} \sum_{i=1}^n \sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''} - n^{-1} (\mathbb{E}N_1)^{-2} \mathbb{E}\left(\sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''}\right) \right| \\ \leq \sup_{J,l,l',l''} n^{-1} (\mathbb{E}N_1)^{-2} \left\{ \left| \left(\frac{n\mathbb{E}N_1}{N_T}\right)^2 - 1 \right| \left| n^{-1} \sum_{i=1}^n \sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''} \right| \right. \\ \left. + \left| n^{-1} \sum_{i=1}^n \sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''} - \mathbb{E}\left(\sum_{k=1}^{\infty} \Psi_{ik,\xi,J,l,l',l''}\right) \right| \right\} \\ = \mathcal{O}_{a.s.} \left\{ n^{-3/2} h_s^{-1/2} (\log(n))^{1/2} \right\}, \end{aligned}$$

and

$$\begin{aligned} & \sup_{J,l,l'} \left| N_T^{-2} \sum_{i=1}^n \sum_{j=1}^{N_i} \Psi_{ik,\varepsilon,J,l,l'} - (n\mathbf{E}N_1)^{-1} \mathbf{E} \Psi_{ik,\varepsilon,J,l,l'} \right| \\ & \leq \sup_{J,l,l'} (n\mathbf{E}N_1)^{-1} \left\{ \left| \frac{n\mathbf{E}N_1}{N_T} - 1 \right| \left| N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} \Psi_{ik,\varepsilon,J,l,l'} \right| \right. \\ & \quad \left. + \left| N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} \Psi_{ik,\varepsilon,J,l,l'} - \mathbf{E} \Psi_{ik,\varepsilon,J,l,l'} \right| \right\} \\ & = \mathcal{O}_{a.s.} \left\{ n^{-3/2} h_s^{-1/2} (\log(n))^{1/2} \right\}. \end{aligned}$$

Notice that

$$\begin{aligned} \Sigma_n(t) &= \mathbf{H}^{-1} c_{J(t),n}^{-1} (n\mathbf{E}N_1)^{-1} \\ & \quad \times \left\{ (\mathbf{E}N_1)^{-1} \mathbf{E} \left( \sum_{l''=1}^d \sum_{k=1}^{\infty} \Psi_{ik,\xi,J(t),l,l',l''} \right) + \mathbf{E} \Psi_{ij,\varepsilon,J(t),l,l'} \right\}_{l,l'=1}^d \mathbf{H}^{-1}, \\ \Sigma_{\xi,n}(t) + \Sigma_{\varepsilon,n}(t) &= \mathbf{H}^{-1} c_{J(t),n}^{-1} N_T^{-2} \left\{ \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} \Psi_{ik,\xi,J(t),l,l',l''} + \sum_{i=1}^n \sum_{j=1}^{N_i} \Psi_{ij,\varepsilon,J(t),l,l'} \right\}_{l,l'=1}^d \mathbf{H}^{-1}, \end{aligned}$$

and (35) implies  $\sup_{t \in [0,1]} |c_{J(t),n}| = \mathcal{O}(h_s)$ . Hence (33) holds.  $\square$

Given  $(T_{ij}, N_i, X_{il})_{i=1,j=1,l=1}^{n,N_i,d}$ , let  $\sigma_{\xi_l,n}^2(t)$  and  $\sigma_{\varepsilon_l,n}^2(t)$  be the conditional variances of  $\hat{\xi}_l(t)$  and  $\hat{\varepsilon}_l(t)$  defined in (38) and (39), respectively. Lemma 6 implies that

$$\sup_{t \in [0,1]} \left| \sigma_{\xi_l,n}^2(t) + \sigma_{\varepsilon_l,n}^2(t) - \sigma_{n,ll}^2(t) \right| = \mathcal{O}_{a.s.} \left\{ n^{-3/2} h_s^{-3/2} (\log(n))^{1/2} \right\}. \quad (34)$$

**Lemma 7** Under Assumptions (A2)–(A6), for  $l = 1, \dots, d$ ,  $\eta_l(t)$  defined in (40) is a Gaussian process consisting of  $(N_s + 1)$  standard normal variables  $\{\eta_{J,l}\}_{J=0}^{N_s}$  such that  $\eta_l(t) = \eta_{J(t),l}$  for  $t \in [0, 1]$ , and there exists a constant  $C > 0$  such that for large  $n$ ,  $\sup_{0 \leq J \neq J' \leq N_s} |\mathbf{E} \eta_{J,l} \eta_{J',l}| \leq Ch_s$ .

*Proof* For any fixed  $l = 1, \dots, d$  and  $0 \leq J \leq N_s$ ,  $\mathcal{L} \left\{ \eta_{J,l} \mid (T_{ij}, N_i, X_{il})_{i=1,j=1,l=1}^{n,N_i,d} \right\} = N(0, 1)$  by Assumption (A2), so  $\mathcal{L} \{ \eta_{J,l} \} = N(0, 1)$ , for  $0 \leq J \leq N_s$ .

Next we derive the upper bound for  $\sup_{0 \leq J \neq J' \leq N_s} |\mathbf{E} \eta_{J,l} \eta_{J',l}|$ . Let

$$\bar{R}_{\xi,J(t),l} = N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J(t),l'',l}^2, \quad \bar{R}_{\varepsilon,J(t),l} = N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} R_{ij,\varepsilon,J(t),l}^2,$$

then we have

$$\sigma_{\hat{\xi}_l,n}(t) = \left\{ c_{J(t),n}^{-1} N_T^{-2} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J(t),l'',l}^2 \right\}^{1/2} = \left\{ c_{J(t),n}^{-1} N_T^{-1} \bar{R}_{\xi,J(t),l} \right\}^{1/2},$$

$$\sigma_{\hat{\varepsilon}_l,n}(t) = \left\{ c_{J(t),n}^{-1} N_T^{-2} \sum_{i=1}^n \sum_{j=1}^{N_i} R_{ij,\varepsilon,J(t),l}^2 \right\}^{1/2} = \left\{ c_{J(t),n}^{-1} N_T^{-1} \bar{R}_{\varepsilon,J(t),l} \right\}^{1/2}.$$

For  $J \neq J'$ , by (31) and the definition of  $B_J$ ,

$$R_{ij,\varepsilon,J,l} R_{ij,\varepsilon,J',l} = \left( \sum_{l'=1}^d z_{ll'} X_{il'} \right)^2 B_J(T_{ij}) B_{J'}(T_{ij}) \sigma^2(T_{ij}) = 0,$$

along with the conditional independence of  $\hat{\xi}_l(t), \hat{\varepsilon}_l(t)$  on  $(T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d}$ , and independence of  $\xi_{ik,l}, T_{ij}, N_i, \{X_{il}\}_{l=1}^d, 1 \leq j \leq N_i, 1 \leq i \leq n, k = 1, 2, \dots,$

$$\begin{aligned} E(\eta_{J,l} \eta_{J',l}) &= E \left[ (\bar{R}_{\xi,J,l} + \bar{R}_{\varepsilon,J,l})^{-1/2} (\bar{R}_{\xi,J',l} + \bar{R}_{\varepsilon,J',l})^{-1/2} \right. \\ &\quad \times N_T^{-1} E \left\{ \left( \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l} \xi_{ik,l''} \right) \left( \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J',l'',l} \xi_{ik,l''} \right) \right. \\ &\quad \left. \left. + \left( \sum_{i=1}^n \sum_{j=1}^{N_i} R_{ij,\varepsilon,J,l} \varepsilon_{ij} \right) \left( \sum_{i=1}^n \sum_{j=1}^{N_i} R_{ij,\varepsilon,J',l} \varepsilon_{ij} \right) \middle| (T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d} \right\} \right] \\ &= EC_{n,J,J',l}, \end{aligned}$$

in which

$$C_{n,J,J',l} = (\bar{R}_{\xi,J,l} + \bar{R}_{\varepsilon,J,l})^{-1/2} (\bar{R}_{\xi,J',l} + \bar{R}_{\varepsilon,J',l})^{-1/2} \times \left\{ N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l} R_{ik,\xi,J',l'',l} \right\}.$$

Note that according to definitions of  $R_{ik,\xi,J,l'',l}, R_{ij,\varepsilon,J,l}$ , and Lemma 5, for  $0 \leq J \leq N_s$

$$\begin{aligned} &\bar{R}_{\xi,J(t),l} + \bar{R}_{\varepsilon,J(t),l} \geq \bar{R}_{\varepsilon,J(t),l} \geq ER_{ij,\varepsilon,J,l}^2 - A_{n,\varepsilon} \geq c_R - A_{n,\varepsilon}, \\ P \left[ \inf_{0 \leq J \neq J' \leq N_s} \{ (\bar{R}_{\xi,J,l} + \bar{R}_{\varepsilon,J,l}) (\bar{R}_{\xi,J',l} + \bar{R}_{\varepsilon,J',l}) \} \geq \left( c_R - \delta \sqrt{\frac{\log(n)}{nh_s}} \right)^2 \right] &\geq 1 - 2n^{-8}. \end{aligned}$$



Thus for large  $n$ , with probability  $\geq 1 - 2n^{-8}$ , the denominator of  $C_{n,J,J',l}$  is uniformly greater than  $c_R^2/4$ . On the other hand, we consider the numerator of  $C_{n,J,J',l}$ .

$$\begin{aligned} \mathbb{E} \left( N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l} R_{ik,\xi,J',l'',l} \right) &= \mathbb{E} \left\{ N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \left( \sum_{l'=1}^d z_{ll'} X_{il'} X_{il''} \right)^2 \right. \\ &\times \left. \left( \sum_{j=1}^{N_i} \sum_{j'=1}^{N_i} B_J(T_{ij}) B_{J'}(T_{ij'}) G_{l''}(T_{ij}, T_{ij'}) \right) \right\} \sim h_s. \end{aligned}$$

Applying Bernstein's inequality, there exists  $C_0 > 0$  such that, for large  $n$ ,

$$P \left( \sup_{0 \leq J \neq J' \leq N_s} \left| N_T^{-1} \sum_{l''=1}^d \sum_{i=1}^n \sum_{k=1}^{\infty} R_{ik,\xi,J,l'',l} R_{ik,\xi,J',l'',l} \right| \leq C_0 h_s \right) \geq 1 - 2n^{-8}.$$

Putting the above together, for large  $n$ ,  $C_1 = C_0 (c_R^2/4)^{-1}$ ,

$$P \left( \sup_{0 \leq J \neq J' \leq N_s} |C_{n,J,J',l}| \leq C_1 h_s \right) \geq 1 - 4n^{-8}.$$

Note that as a continuous random variable,  $\sup_{0 \leq J \neq J' \leq N_s} |C_{n,J,J',l}| \in [0, 1]$ , thus

$$\mathbb{E} \left( \sup_{0 \leq J \neq J' \leq N_s} |C_{n,J,J',l}| \right) = \int_0^1 P \left( \sup_{0 \leq J \neq J' \leq N_s} |C_{n,J,J',l}| > u \right) du.$$

For large  $n$ ,  $C_1 h_s < 1$  and then  $\mathbb{E} \left( \sup_{0 \leq J \neq J' \leq N_s, l} |C_{n,J,J',l}| \right)$  is

$$\begin{aligned} &\int_0^{C_1 h_s} P \left\{ \sup_{0 \leq J \neq J' \leq N_s, l} |C_{n,J,J',l}| > u \right\} du + \int_{C_1 h_s}^1 P \left\{ \sup_{0 \leq J \neq J' \leq N_s, l} |C_{n,J,J',l}| > u \right\} du \\ &\leq \int_0^{C_1 h_s} 1 du + \int_{C_1 h_s}^1 4n^{-8} du \leq C_1 h_s + 4n^{-8} \leq C h_s \end{aligned}$$

for some  $C > 0$  and large enough  $n$ . The lemma now follows from

$$\sup_{0 \leq J \neq J' \leq N_s} |\mathbb{E}(C_{n,J,J',l})| \leq \mathbb{E} \left( \sup_{0 \leq J \neq J' \leq N_s} |C_{n,J,J',l}| \right) \leq C h_s.$$

This completes the proof of the lemma.  $\square$

**Lemma 8** Under Assumptions (A2)–(A6), for  $\eta_l(t), \sigma_{n,ll}(t), l = 1, \dots, d$ , defined in (40) and (7), one has  $|\sigma_{n,ll}(t)^{-1} \{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \} - \eta_l(t)| = |r_{n,l}(t) - 1| |\eta_l(t)|$ , where  $r_{n,l}(t) = \sigma_{n,ll}^{-1}(t) \{ \sigma_{\xi_l,n}^2(t) + \sigma_{\varepsilon_l,n}^2(t) \}^{1/2}$ , and as  $n \rightarrow \infty$ ,

$$\sup_{t \in [0,1]} \{ a_{N_s+1} |r_{n,l}(t) - 1| \} = \mathcal{O}_{a.s.} \{ (nh_s)^{-1/2} (\log(N_s + 1) \log(n))^{1/2} \}.$$

*Proof* By Lemma 5,  $\sigma_{n,ll}^2(t)$  in (7) can be rewritten as

$$\begin{aligned} \sigma_{n,ll}^2(t) &= c_{J(t),n}^{-1} (n\mathbf{E}N_1)^{-1} \left\{ (\mathbf{E}N_1)^{-1} \sum_{l''=1}^d \mathbf{E} \left( \sum_{k=1}^{\infty} R_{ik,\xi,J(t),l'',l}^2 \right) + \mathbf{E}R_{ij,\varepsilon,J(t),l}^2 \right\} \\ &\sim n^{-1} h_s^{-1}. \end{aligned}$$

Hence, according to (34) and (10),

$$\begin{aligned} \sup_{t \in [0,1]} \{ a_{N_s+1} |r_{n,l}(t) - 1| \} &= \sup_{t \in [0,1]} \left\{ a_{N_s+1} \left| \sigma_{n,ll}^{-1}(t) \{ \sigma_{\xi_l,n}^2(t) + \sigma_{\varepsilon_l,n}^2(t) \}^{1/2} - 1 \right| \right\} \\ &\leq \sup_{t \in [0,1]} \left\{ a_{N_s+1} \left| \sigma_{n,ll}^{-2}(t) \{ \sigma_{\xi_l,n}^2(t) + \sigma_{\varepsilon_l,n}^2(t) \} - 1 \right| \right\} \\ &= \sup_{t \in [0,1]} \left\{ a_{N_s+1} \sigma_{n,ll}^{-2}(t) \left| \sigma_{\xi_l,n}^2(t) + \sigma_{\varepsilon_l,n}^2(t) - \sigma_{n,ll}^2(t) \right| \right\} \\ &= \mathcal{O}_{a.s.} \{ (nh_s)^{-1/2} (\log(N_s + 1) \log(n))^{1/2} \}. \end{aligned}$$

This completes the proof. □

### A.3 Proofs of Propositions 1–4

*Proof of Proposition 1* By Assumption (A3) on the continuity of functions  $\phi_{k,l}(t), \sigma^2(t)$  and  $f(t)$  on  $[0, 1]$  and Assumption (A4), for any  $t, u \in [0, 1]$  satisfying  $|t - u| \leq h_s$ ,

$$|G_l(t, t) - G_l(u, u)| \leq \sum_{k=1}^{\infty} \left| \phi_{k,l}^2(t) - \phi_{k,l}^2(u) \right| \leq 2 \sum_{k=1}^{\infty} \|\phi_{k,l}\|_{\infty} \omega(\phi_{k,l}, h_s) \leq Ch_s^r.$$

Furthermore,

$$\begin{aligned} \left| \int_{\chi_{J(t)}} \{ G_l(t, t) f(t) - G_l(u, u) f(u) \} du \right| &\leq Ch_s^{1+r} = \mathcal{O}(h_s^{1+r}), \\ \left| \int_{\chi_{J(t)} \times \chi_{J(t)}} \{ G_l(t, t) f^2(t) - G_l(u, v) f(u) f(v) \} dudv \right| &\leq Ch_s^{2+r} = \mathcal{O}(h_s^{2+r}), \end{aligned}$$

$$\left| \int_{\chi_{J(t)}} \left\{ \sigma^2(t) f(t) - \sigma^2(u) f(u) \right\} du \right| \leq C h_s^{1+r} = \mathcal{O} \left( h_s^{1+r} \right).$$

According to the definition of  $C_{J,n}$  in (6),

$$C_{J,n} = \int_{[v_J, v_{J+1}]} f(x) dx = f(v_J) h_s + \int_{[v_J, v_{J+1}]} \{f(x) - f(v_J)\} dx, \quad (35)$$

thus,  $|C_{J,n} - f(v_J) h_s| \leq w(f, h_s) h_s$  for all  $J = 0, \dots, N_s$ . Therefore,

$$\begin{aligned} \Gamma_n(t) &= \left\{ f(t) h_s + \mathcal{U} \left( h_s^{1+r} \right) \right\}^{-2} (n \mathbf{E} N_1)^{-1} \mathbf{E} \left[ \left\{ \sigma_Y^2(t, \mathbf{X}) f(t) h_s + \mathcal{U}_p \left( h_s^{1+r} \right) \right. \right. \\ &\quad \left. \left. + \frac{\mathbf{E} \{N_1(N_1 - 1)\}}{\mathbf{E} N_1} \sum_{l=1}^d X_l^2 G_l(t, t) f^2(t) h_s^2 + \mathcal{U}_p \left( h_s^{2+r} \right) \right\} \mathbf{X} \mathbf{X}^\top \right] \\ &= \mathbf{E} \left[ \mathbf{X} \mathbf{X}^\top \sigma_Y^2(t, \mathbf{X}) \left\{ f(t) h_s n \mathbf{E} N_1 \right\}^{-1} \left\{ 1 + \frac{\mathbf{E} \{N_1(N_1 - 1)\}}{\mathbf{E} N_1} \right. \right. \\ &\quad \left. \left. \times \frac{\sum_{l=1}^d X_l^2 G_l(t, t) f(t) h_s}{\sigma_Y^2(t, \mathbf{X})} \right\} \left\{ 1 + \mathcal{U}_p \left( h_s^r \right) \right\} \right] = \tilde{\Gamma}_n(t) + \mathcal{U} \left( n^{-1} h_s^{r-1} \right), \end{aligned}$$

establishing the proposition.  $\square$

*Proof of Proposition 2* The result follows from standard theory of kernel and spline smoothing, as in Wang and Yang (2009), thus omitted.  $\square$

*Proof of Proposition 3* According to the result on page 149 of de Boor (2001), there exist functions  $g_l \in G^{(-1)} [0, 1]$  that satisfies  $\|m_l - g_l\|_\infty = \mathcal{O}(h_s)$  for  $l = 1, \dots, d$ . By the definition of  $\tilde{m}_l(t)$  in (22),

$$\tilde{\mathbf{m}}(t) = (\tilde{m}_1(t), \dots, \tilde{m}_d(t))^\top = c_{J(t),n}^{-1/2} (\tilde{\gamma}_{J(t),1}, \dots, \tilde{\gamma}_{J(t),d})^\top = c_{J(t),n}^{-1/2} \tilde{\gamma}_{J(t)},$$

where  $\tilde{\gamma}_J = \hat{\mathbf{V}}_J^{-1} \left\{ N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J(T_{ij}) X_{il} \sum_{l'=1}^d m_{l'}(T_{ij}) X_{il'} \right\}_{l=1}^d$  for  $\hat{\mathbf{V}}_J$  defined in (18).

Let  $\tilde{\mathbf{g}}(t) = (\tilde{g}_1(t), \dots, \tilde{g}_d(t))^\top$ , then  $\tilde{\mathbf{m}}_l(t) - \tilde{\mathbf{g}}_l(t)$  equals to

$$c_{J(t),n}^{-1/2} \hat{\mathbf{V}}_J^{-1} \left[ \frac{1}{N_T} \sum_{i=1}^n \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) X_{il} \sum_{l'=1}^d \{m_{l'}(T_{ij}) - g_{l'}(T_{ij})\} X_{il'} \right]_{l=1}^d.$$

Observing that  $\tilde{g}_l \equiv g_l$  as  $g_l \in G^{(-1)} [0, 1]$ , there is a decomposition similar to (24),  $\tilde{m}_l(t) = \tilde{m}_l(t) - \tilde{g}_l(t) + g_l(t)$ ,  $l = 1, \dots, d$ .

By (35),  $\sup_{t \in [0,1]} |c_{J(t),n}| = \mathcal{O}(h_s)$ . Next  $\mathbf{E}|B_J(T_{ij})| = c_{J,n}^{-1/2} \int b_J(x)f(x)dx \sim h_s^{1/2}$ , thus  $\sup_{t \in [0,1]} |B_{J(t)}(T_{ij})| = \mathcal{O}_p(h_s^{1/2})$ . Then it is easy to show that  $\|\tilde{m}_l - \tilde{g}_l\|_\infty = \mathcal{O}_p(h_s^{-1/2}h_s^{1/2}h_s) = \mathcal{O}_p(h_s)$ . Hence, for  $l = 1, \dots, d$ ,

$$\|\tilde{m}_l - m_l\|_\infty \leq \|\tilde{m}_l - \tilde{g}_l\|_\infty + \|m_l - g_l\|_\infty = \mathcal{O}_p(h_s),$$

which completes the proof. □

Note that  $B_J(t) \equiv b_J c_{J,n}^{-1/2}$ ,  $t \in [0, 1]$ , so the terms  $\tilde{\xi}_l(t)$  and  $\tilde{\varepsilon}_l(t)$ ,  $l = 1, \dots, d$ , defined in (23) are

$$\tilde{\xi}(t) = (\tilde{\xi}_1(t), \dots, \tilde{\xi}_d(t))^T = c_{J(t),n}^{-1/2} (\tilde{\alpha}_{J(t),1}, \dots, \tilde{\alpha}_{J(t),d})^T = c_{J(t),n}^{-1/2} \tilde{\alpha}_{J(t)}, \tag{36}$$

$$\tilde{\varepsilon}(t) = (\tilde{\varepsilon}_1(t), \dots, \tilde{\varepsilon}_d(t))^T = c_{J(t),n}^{-1/2} (\tilde{\theta}_{J(t),1}, \dots, \tilde{\theta}_{J(t),d})^T = c_{J(t),n}^{-1/2} \tilde{\theta}_{J(t)}, \tag{37}$$

where

$$\tilde{\alpha}_J = \hat{\mathbf{V}}_J^{-1} \left\{ N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J(T_{ij}) X_{il} \sum_{l''=1}^d \sum_{k=1}^\infty \xi_{ik,l''} \phi_{k,l''}(T_{ij}) X_{il''} \right\}_{l=1}^d,$$

$$\tilde{\theta}_J = \hat{\mathbf{V}}_J^{-1} \left\{ N_T^{-1} \sum_{i=1}^n \sum_{j=1}^{N_i} B_J(T_{ij}) X_{il} \sigma(T_{ij}) \varepsilon_{ij} \right\}_{l=1}^d.$$

According to Lemma 3, the inverse of the random matrix  $\hat{\mathbf{V}}_J$  can be approximated by that of a deterministic matrix  $\mathbf{H} = \mathbf{E}(\mathbf{X}\mathbf{X}^T)$ . Substituting  $\hat{\mathbf{V}}_J$  with  $\mathbf{H}$  in (36) and (37), we define the random vectors

$$\hat{\xi}(t) = c_{J(t),n}^{-1/2} \mathbf{H}^{-1} \left\{ \frac{1}{N_T} \sum_{i=1}^n \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) X_{il} \sum_{l''=1}^d \sum_{k=1}^\infty \xi_{ik,l''} \phi_{k,l''}(T_{ij}) X_{il''} \right\}_{l=1}^d, \tag{38}$$

$$\hat{\varepsilon}(t) = c_{J(t),n}^{-1/2} \mathbf{H}^{-1} \left\{ \frac{1}{N_T} \sum_{i=1}^n \sum_{j=1}^{N_i} B_{J(t)}(T_{ij}) X_{il} \sigma(T_{ij}) \varepsilon_{ij} \right\}_{l=1}^d. \tag{39}$$

*Proof of Proposition 4* Given  $(T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d}$ , let  $\sigma_{\hat{\xi}_l, n}^2(t)$  and  $\sigma_{\hat{\varepsilon}_l, n}^2(t)$  be the conditional variances of  $\hat{\xi}_l(t)$  and  $\hat{\varepsilon}_l(t)$  defined in (38) and (39), respectively. Define

$$\eta_l(t) = \left\{ \sigma_{\hat{\xi}_l, n}^2(t) + \sigma_{\hat{\varepsilon}_l, n}^2(t) \right\}^{-1/2} \left\{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right\}. \tag{40}$$

By Lemma 7,  $\eta_l(t)$  is a Gaussian process consisting of  $(N_s + 1)$  standard normal variables  $\{\eta_{J,l}\}_{J=0}^{N_s}$  such that  $\eta_l(t) = \eta_{J(t),l}$  for  $t \in [0, 1]$ . Thus, for any  $\tau \in \mathbb{R}$ ,

$$\begin{aligned} & P\left(\sup_{t \in [0,1]} |\eta_l(t)| \leq \tau/a_{N_s+1} + b_{N_s+1}\right) \\ &= P\left(|\max\{\eta_{0,l}, \dots, \eta_{N_s,l}\}| \leq \tau/a_{N_s+1} + b_{N_s+1}\right). \end{aligned}$$

By Theorem 1.5.3 in Leadbetter et al. (1983), if  $\xi_0, \dots, \xi_{N_s}$  are i.i.d. standard normal r.v.'s, then for  $\tau \in \mathbb{R}$

$$P\left(|\max\{\xi_0, \dots, \xi_{N_s}\}| \leq \tau/a_{N_s} + b_{N_s}\right) \rightarrow \exp(-2e^{-\tau}).$$

Next by Lemma 11.1.2 in Leadbetter et al. (1983),

$$\begin{aligned} & P\left(|\max\{\eta_{0,l}, \dots, \eta_{N_s,l}\}| \leq \tau/a_{N_s+1} + b_{N_s+1}\right) \\ & \quad - P\left(|\max\{\xi_0, \dots, \xi_{N_s}\}| \leq \tau/a_{N_s+1} + b_{N_s+1}\right) \\ & \leq \frac{4}{2\pi} \sum_{0 \leq J < J' \leq N_s} |\mathbf{E}\eta_{J,l}\eta_{J',l}| (1 - |\mathbf{E}\eta_{J,l}\eta_{J',l}|^2)^{-1/2} \exp\left\{\frac{-(\tau/a_{N_s+1} + b_{N_s+1})^2}{1 + \mathbf{E}\eta_{J,l}\eta_{J',l}}\right\}. \end{aligned}$$

According to Lemma 7, there exists a constant  $C > 0$  such that  $\sup_{0 \leq J \neq J' \leq N_s} |\mathbf{E}\eta_{J,l}\eta_{J',l}| \leq Ch_s$  for large  $n$ . Thus, as  $n \rightarrow \infty$ ,

$$\begin{aligned} & P\left(|\max\{\eta_{0,l}, \dots, \eta_{N_s,l}\}| \leq \tau/a_{N_s+1} + b_{N_s+1}\right) \\ & \quad - P\left(|\max\{\xi_0, \dots, \xi_{N_s}\}| \leq \tau/a_{N_s+1} + b_{N_s+1}\right) \rightarrow 0. \end{aligned}$$

Therefore, for any  $\tau \in \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} P\left(\sup_{t \in [0,1]} |\eta_l(t)| \leq \tau/a_{N_s+1} + b_{N_s+1}\right) = \exp(-2e^{-\tau}). \quad (41)$$

By Lemma 8, we have

$$\begin{aligned} & a_{N_s+1} \left( \sup_{t \in [0,1]} \sigma_{n,l}^{-1}(t) \left| \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right| - \sup_{t \in [0,1]} |\eta_l(t)| \right) \\ &= \mathcal{O}_p \left\{ \log(N_s + 1) (nh_s)^{-1/2} (\log(n))^{1/2} \right\} = \mathcal{O}_p(1). \end{aligned}$$

On the other hand, Lemma 4 ensures that

$$\begin{aligned} & a_{N_s+1} \left( \sup_{t \in [0,1]} \sigma_{n,tl}^{-1}(t) \left| \tilde{\xi}_l(t) + \tilde{\varepsilon}_l(t) \right| - \sup_{t \in [0,1]} \sigma_{n,tl}^{-1}(t) \left| \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right| \right) \\ &= \mathcal{O}_p \left\{ (\log(N_s + 1) nh_s)^{1/2} n^{-1} h_s^{-3/2} \log(n) \right\} \\ &= \mathcal{O}_p \left\{ n^{-1/2} h_s^{-1} (\log(N_s + 1))^{1/2} \log(n) \right\} = \mathcal{O}_p(1). \end{aligned}$$

Then the proof follows from (41) and Slutsky’s Theorem. □

#### A.4 Proof of Theorem 1

For any vector  $\mathbf{a} = (a_1, \dots, a_d)^\top \in \mathbb{R}^d$ ,  $\mathbf{E} \left[ \sum_{l=1}^d a_l \left\{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right\} \right] = 0$ . Using the conditional independence of  $\hat{\xi}_l(t), \hat{\varepsilon}_l(t)$  on  $(T_{ij}, N_i, X_{il})_{i=1, j=1, l=1}^{n, N_i, d}$ , we have

$$\begin{aligned} & \text{Var} \left[ \sum_{l=1}^d a_l \left\{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right\} \middle| (T_{ij}, N_i, X_{il})_{j=1, i=1, l=1}^{N_i, n, d} \right] \\ &= \sum_{l=1}^d \sum_{l'=1}^d a_l a_{l'} \mathbf{E} \left\{ \hat{\xi}_l(t) \hat{\xi}_{l'}(t) + \hat{\varepsilon}_l(t) \hat{\varepsilon}_{l'}(t) \middle| (T_{ij}, N_i, X_{il})_{j=1, i=1, l=1}^{N_i, n, d} \right\} \\ &= \mathbf{a}^\top \left\{ \Sigma_{\xi, n}(t) + \Sigma_{\varepsilon, n}(t) \right\} \mathbf{a}. \end{aligned}$$

Meanwhile, Assumption (A2) entails that for any  $t \in [0, 1]$ , given  $(T_{ij}, N_i, X_{il})_{j=1, i=1, l=1}^{N_i, n, d}$ , the conditional distribution of  $\left[ \mathbf{a}^\top \left\{ \Sigma_{\xi, n}(t) + \Sigma_{\varepsilon, n}(t) \right\} \mathbf{a} \right]^{-1/2} \sum_{l=1}^d a_l \left\{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right\}$  is a standard normal distribution. So we have

$$\left[ \mathbf{a}^\top \left\{ \Sigma_{\xi, n}(t) + \Sigma_{\varepsilon, n}(t) \right\} \mathbf{a} \right]^{-1/2} \sum_{l=1}^d a_l \left\{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right\} \sim N(0, 1).$$

Using (33), we have as  $n \rightarrow \infty$

$$\left[ \mathbf{a}^\top \Sigma_n(t) \mathbf{a} \right]^{-1/2} \sum_{l=1}^d a_l \left\{ \hat{\xi}_l(t) + \hat{\varepsilon}_l(t) \right\} \xrightarrow{\mathcal{L}} N(0, 1).$$

Therefore,  $\left[ \mathbf{a}^\top \Sigma_n(t) \mathbf{a} \right]^{-1/2} \sum_{l=1}^d a_l \left\{ \hat{m}_l(t) - m_l(t) \right\} \xrightarrow{\mathcal{L}} N(0, 1)$  follows from (24), Proposition 3, Lemma 4 and Slutsky’s Theorem. Applying Cramér–Wold’s device, we obtain  $\Sigma_n^{-1/2}(t) \left\{ \hat{m}_l(t) - m_l(t) \right\}_{l=1}^d \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}_{d \times d})$ , and consequently,  $\sigma_{n,tl}^{-1}(t) \left\{ \hat{m}_l(t) - m_l(t) \right\} \xrightarrow{\mathcal{L}} N(0, 1)$  for any  $t \in [0, 1]$  and  $l = 1, \dots, d$ . □

### A.5 Proof of Theorem 2

By Proposition 3,  $\|\tilde{m}_l - m_l\|_\infty = \mathcal{O}_p(h_s)$ ,  $l = 1, \dots, d$ , so

$$a_{N_s+1} \left\{ \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) |\tilde{m}_l(t) - m_l(t)| \right\} = \mathcal{O}_p \left\{ (nh_s)^{1/2} (\log(N_s+1))^{1/2} h_s \right\} = \mathcal{O}_p(1).$$

According to (24), it is easy to show that

$$a_{N_s+1} \left\{ \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) |\hat{m}_l(t) - m_l(t)| - \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) |\tilde{\xi}_l(t) + \tilde{\varepsilon}_l(t)| \right\} = \mathcal{O}_p(1).$$

Meanwhile, Proposition 4 entails that, for any  $\tau \in \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} P \left\{ a_{N_s+1} \left( \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) |\tilde{\xi}_l(t) + \tilde{\varepsilon}_l(t)| - b_{N_s+1} \right) \leq \tau \right\} = \exp(-2e^{-\tau}).$$

Thus Slutsky's Theorem implies that

$$\lim_{n \rightarrow \infty} P \left\{ a_{N_s+1} \left( \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) |\hat{m}_l(t) - m_l(t)| - b_{N_s+1} \right) \leq \tau \right\} = \exp(-2e^{-\tau}).$$

Let  $\tau = -\log\{-\frac{1}{2} \log(1-\alpha)\}$ , the definition of  $Q_{N_s+1}(\alpha)$  in (9) entails

$$\begin{aligned} & \lim_{n \rightarrow \infty} P \{ m_l(t) \in \hat{m}_l(t) \pm \sigma_{n,ll}(t) Q_{N_s+1}(\alpha), \forall t \in [0, 1] \} \\ &= \lim_{n \rightarrow \infty} P \left\{ \sup_{t \in [0,1]} \sigma_{n,ll}^{-1}(t) |\hat{m}_l(t) - m_l(t)| \leq Q_{N_s+1}(\alpha) \right\} = 1 - \alpha. \end{aligned}$$

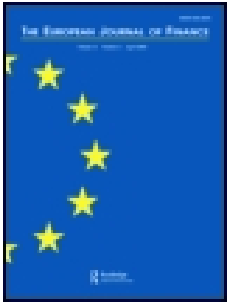
Theorem 2 is proved.  $\square$

### References

- Bosq D (1998) Nonparametric statistics for stochastic processes. Springer, New York
- Brumback B, Rice JA (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with Discussion). *J Am Stat Assoc* 93:961–994
- Cao G, Yang L, Todem D (2012) Simultaneous inference for the mean function based on dense functional data. *J Nonparametr Stat* 24:359–377
- Cao G, Wang J, Wang L, Todem D (2012) Spline confidence bands for functional derivatives. *J Stat Plan Inference* 142:1557–1570
- Chiang CT, Rice JA, Wu CO (2001) Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J Am Stat Assoc* 96:605–619
- Claeskens G, Van Keilegom I (2003) Bootstrap confidence bands for regression curves and their derivatives. *Ann Stat* 31:1852–1884
- de Boor C (2001) A practical guide to splines. Springer, New York
- Fan J, Zhang JT (2000) Functional linear models for longitudinal data. *J R Stat Soc Ser B* 62:303–322

- Fan J, Zhang WY (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand J Stat* 27:715–731
- Fan J, Zhang WY (2008) Statistical methods with varying coefficient models. *Stat Interface* 1:179–195
- Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, New York
- Gabrys R, Horváth L, Kokoszka P (2010) Tests for error correlation in the functional linear model. *J Am Stat Assoc* 105:1113–1125
- Hall P, Müller HG, Wang JL (2006) Properties of principal component methods for functional and longitudinal data analysis. *Ann Stat* 34:1493–1517
- Hall P, Titterton DM (1988) On confidence bands in nonparametric density estimation and regression. *J Mult Anal* 27:228–254
- Härdle W, Luckhaus S (1984) Uniform consistency of a class of regression function estimators. *Ann Stat* 12:612–623
- Hastie T, Tibshirani R (1993) Varying-coefficient models. *J R Stat Soc Ser B* 55:757–796
- Hoover DR, Rice JA, Wu CO, Yang LP (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85:809–822
- Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer, New York
- Huang JZ, Wu CO, Zhou L (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* 89:111–128
- Huang JZ, Wu CO, Zhou L (2004) Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Stat Sin* 14:763–788
- James GM, Hastie T, Sugar C (2000) Principal component models for sparse functional data. *Biometrika* 87:587–602
- James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. *J Am Stat Assoc* 98:397–408
- Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York
- Liu R, Yang L (2010) Spline-backfitted kernel smoothing of additive coefficient model. *Econ Theory* 26:29–59
- Ma S, Yang L, Carroll RJ (2012) A simultaneous confidence band for sparse longitudinal regression. *Stat Sin* 22:95–122
- Manteiga W, Vieu P (2007) Statistics for functional data. *Comput Stat Data Anal* 51:4788–4792
- Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York
- Wang L, Li H, Huang JZ (2008) Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J Am Stat Assoc* 103:1556–1569
- Wang L, Yang L (2009) Polynomial spline confidence bands for regression curves. *Stat Sin* 19:325–342
- Wu CO, Chiang CT (2000) Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Stat Sin* 10:433–456
- Wu CO, Chiang CT, Hoover DR (1998) Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J Am Stat Assoc* 93:1388–1402
- Wu Y, Fan J, Müller HG (2010) Varying-coefficient functional linear regression. *Bernoulli* 16:730–758
- Xue L, Yang L (2006) Additive coefficient modelling via polynomial spline. *Stat Sin* 16:1423–1446
- Xue L, Zhu L (2007) Empirical likelihood for a varying coefficient model with longitudinal data. *J Am Stat Assoc* 102:642–654
- Yao W, Li R (2013) New local estimation procedure for a non-parametric regression function for longitudinal data. *J R Stat Soc Ser B* 75:123–138
- Yao F, Müller HG, Wang JL (2005a) Functional linear regression analysis for longitudinal data. *Ann Stat* 33:2873–2903
- Yao F, Müller HG, Wang JL (2005b) Functional data analysis for sparse longitudinal data. *J Am Stat Assoc* 100:577–590
- Zhou L, Huang J, Carroll RJ (2008) Joint modelling of paired sparse functional data using principal components. *Biometrika* 95:601–619
- Zhu H, Li R, Kong L (2012) Multivariate varying coefficient model for functional responses. *Ann Stat* 40:2634–2666





## Yield curve modeling and forecasting using semiparametric factor dynamics

Wolfgang K. Härdle & Piotr Majer

To cite this article: Wolfgang K. Härdle & Piotr Majer (2014): Yield curve modeling and forecasting using semiparametric factor dynamics, The European Journal of Finance, DOI: [10.1080/1351847X.2014.926281](https://doi.org/10.1080/1351847X.2014.926281)

To link to this article: <http://dx.doi.org/10.1080/1351847X.2014.926281>



Published online: 11 Jun 2014.



Submit your article to this journal [↗](#)



Article views: 39



View related articles [↗](#)



View Crossmark data [↗](#)

## Yield curve modeling and forecasting using semiparametric factor dynamics

Wolfgang K. Härdle<sup>a,b</sup> and Piotr Majer<sup>a\*</sup>

<sup>a</sup>C.A.S.E. – Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany; <sup>b</sup>School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore

(Received 21 September 2012; final version received 8 May 2014)

Using a dynamic semiparametric factor model (DSFM) we investigate the term structure of interest rates. The proposed methodology is applied to monthly interest rates for four southern European countries: Greece, Italy, Portugal and Spain from the introduction of the Euro to the recent European sovereign-debt crisis. Analyzing this extraordinary period, we compare our approach with the standard market method – dynamic Nelson–Siegel model. Our findings show that two nonparametric factors capture the spatial structure of the yield curve for each of the bond markets separately. We attributed both factors to the slope of the yield curve. For panel term structure data, three nonparametric factors are necessary to explain 95% variation. The estimated factor loadings are unit root processes and reveal high persistency. In comparison with the benchmark model, the DSFM technique shows superior short-term forecasting in times of financial distress.

**Keywords:** yield curve; term structure of interests rates; semiparametric model; factor structure; prediction; European sovereign debt crisis

*JEL Classification:* G12; G17; C5; C4

### 1. Introduction

Modeling and forecasting the term structure of interest rates are important in financial economics. Pricing financial assets and their derivatives, allocating portfolios, managing financial risk, conducting monetary policy are the essential challenges which involve interest rates and dynamic evolution of the yield curve. For that reason researchers have developed a large toolbox of models and techniques. The most popular approaches are equilibrium and no-arbitrage models. The no-arbitrage models follow the Black–Scholes framework and ensure correct pricing of derivatives; the main contributions for no-arbitrage models are Hull and White (1990) and Heath, Jarrow, and Morton (1992). The equilibrium framework provides exact fits to the observed term structure (Longstaff and Schwartz 1992). However both approaches do not provide a good predictive performance, since forecasting is not the main goal of these approaches. To this end, Diebold and Li (2006) proposed the Nelson–Siegel (NS) curve with time varying parameters. The dynamic NS model has gained popularity among financial market practitioners and central banks. This relatively new dynamic factor model provides a remarkably good fit to the term structure of interest rates, where the given factors of the exponential form have a standard interpretation of level, slope and curvature. Parametric structure of dynamic NS model leads to easy estimation and

---

\*Corresponding author. Email: [majerpio@cms.hu-berlin.de](mailto:majerpio@cms.hu-berlin.de)

displays empirical tractability. In the same spirit generalizations of the NS approach were introduced by [Svensson \(1995\)](#) and [Christensen, Diebold, and Rudebusch \(2009\)](#). Dynamic factor models for yield curve modeling are reported to be extremely useful in practice (e.g. Federal Reserve Board [Gürkaynak, Sack, and Wright 2010](#); European Central Bank [Coroneo, Nyhlon, and Vidova-Koleva 2008](#)).

In this paper we go beyond the NS structure by proposing a dynamic semiparametric factor model (DSFM). The paper's major idea is to capture the shape of the yield curve by a lower-dimensional factor representation. The latent factors are estimated non-parametrically by tensor B-splines avoiding specification issues (e.g. exponential form imposed in the NS model). The choice of the B-splines series expansion is motivated by [Vasicek and Fong \(1982\)](#), who first implemented it in a term structure model. Since that time B-splines series has attracted much research attention and serves as flexible yield curve modeling approach ([Krivobokova, Kauer- mann, and Archontakis 2006](#); [Bowsher and Meeks 2008](#)). Similarly to parametric NS models and functional principal component analysis (FPCA, [Ramsay and Silverman 1997](#)), yield curve is represented as a linear combination of latent factors. The evolution in time is driven by time-varying factor loadings (in FPCA defined as scores), which are modeled parametrically employing a multivariate autoregressive approach. The factor decomposition is obtained by the DSFM also analyzed in [Fengler, Härdle, and Mammen \(2007\)](#), [Brüggemann et al. \(2008\)](#) and [Park et al. \(2009\)](#). Accordingly, the term structure of interest rates is modeled in terms of underlying latent factors, which are defined on the time to maturity grid space and may depend on additional explanatory variables. The inclusion of additional regressors is motivated by Taylor's rule ([Taylor 1992](#)) and was also picked up by [Diebold, Rudebusch, and Aruoba \(2006\)](#), [Ang and Piazzesi \(2003\)](#). The main idea is to incorporate the macroeconomic activity as a determinant of the yield curve. The connection between yield curve dynamics and contemporaneous macroeconomic fundamentals is investigated in terms of the extracted loadings. We analyze the effect of the harmonized consumer price index (INF), the manufacturing capacity utilization (CU), the unemployment rate (EMP), industrial production (IP) and the real gross domestic product ( $\Delta$ GDP). We evaluate the short- and long-run prediction power of the underlying macroeconomic fundamentals for the extracted time series.

We focus on the recent European sovereign-debt crisis. The last few years have challenged all the standard models and have revealed an urge for alternative statistical tools. Our attention is drawn by the bond markets of southern European countries, the epicenter of the recent European sovereign-debt crisis. The DSFM approach is first applied as a domestic term structure model for each yield curve separately. Yield curve factor models differ with respect to the number of latent factors. Increasing the number of factors leads to better in-sample fit but might weaken the forecasting performance and parsimony of the model. The NS model assumes three factors whereas its extension proposed by [Svensson \(1995\)](#) consists of four factors. To this end we investigate the number of factors required to model the yield curve reasonably well, particularly in times of financial turmoil. We select the optimal complexity of the model by statistical criteria. Flexibility of our model allows us to investigate the spatial structure of factors in dependence of additional explanatory variables. In the next step we extend our analysis to the panel data. Modeling the joint term structure of interest rates is a task of extreme importance nowadays, when financial markets have become increasingly globalized. Moreover all the countries share the same currency and monetary policy. They are members of one economic bloc and often grouped together as Euro-zone peripheral states. The joint yield curves are modeled by the panel DSFM (PDSFM) technique, presented in Appendix A.1.

This paper is structured as follows: in Section 2 we describe the data set. The DSFM and the dynamic NS model are introduced in Section 3. Empirical results and comparison of forecasting performance are provided in Sections 4, 5 and 6. In Section 7 we summarize the main contribution of the paper.

## 2. Data

In this section, we provide summary statistics on the term structure data. Our primary data sample consists of the monthly end-of-day government zero-coupon bond prices of Greece (GR), Italy (IT), Portugal (PT) and Spain (ES). We focus our analysis on the south-European states starting from the introduction of the European currency, the Euro. Our data set covers the period from January 1999 through to March 2012. Specifically, we consider the interest rates with 11 different times to maturity  $X_{t,j}$  ranging from 1 year to 15 years. In Figure 1 we provide a time series plots of Italian and Spanish zero-coupon yield curves. The summary statistics for all zero-curves are shown in Tables A2 and A3. The interest rate data set consists of 160 observations for each country.

To investigate the relation between term structure and macroeconomic activity we study the harmonized consumer price index (INF), CU, unemployment rate (EMP), IP and the real GDP ( $\Delta$ GDP), observed monthly. This data are from Ecwin.

## 3. Factor models

Factor models describe fluctuations over time in high-dimensional objects by a small set of factors. For analytical tractability and asymptotic properties a sub-additive structure of the model is assumed. In this framework factors are characterized up to scale and rotation transformations and contain the most underlying information. For instance,  $Y_t = (Y_{t,1}, Y_{t,2}, \dots, Y_{t,J}) \in \mathbb{R}^J$  can be represented as an (orthogonal)  $L$ -factor model

$$Y_{t,j} = m_{0,j} + Z_{t,1}m_{1,j} + \dots + Z_{t,L}m_{L,j} + \varepsilon_{t,j}, \tag{1}$$

where  $m_{l,j}$  are common factors,  $Z_{t,l}$  are factor loadings and  $\varepsilon_{t,j}$  are specific errors which explain the residual part. The time evolution of  $Y_t$  is represented by  $Z_t, t = 1, \dots, T$ . The factors  $m_l$  may be represented as a function of explanatory variable  $X_{t,j}$ . In the context of yield curve modeling,  $Y_{t,j}, j = 1, \dots, J$ , denotes the observed term structure of interests rates observed on day  $t = 1, \dots, T$ .  $X_{t,j}$  denotes the maturity time of the rate  $Y_{t,j}$ . Factor models have gained popularity in the 1990s and the prominent example is the dynamic NS model.

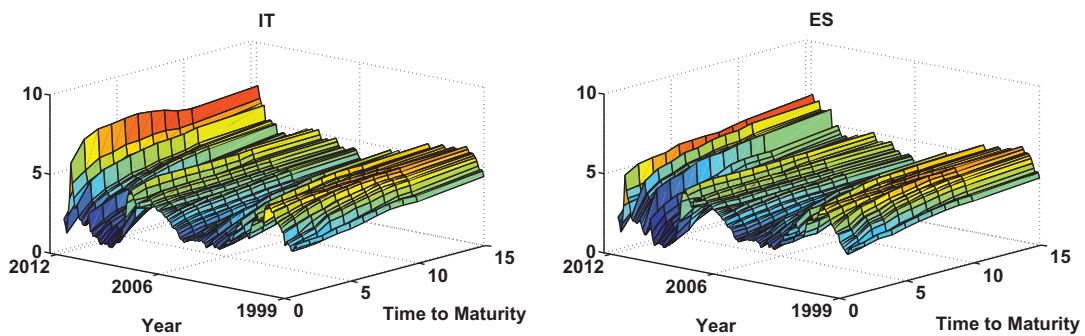


Figure 1. Zero-curves of Italy (left panel) and Spain (right panel) from 1 January 1999 to 31 March 2012.

### 3.1 Dynamic NS model

The NS model fits the yield curve with

$$Y_{t,j} = \beta_0 + \beta_1 \left\{ \frac{1 - \exp(-\lambda X_{t,j})}{\lambda X_{t,j}} \right\} + \beta_2 \left\{ \frac{1 - \exp(-\lambda X_{t,j})}{\lambda X_{t,j}} - \exp(-\lambda X_{t,j}) \right\} + \varepsilon_{t,j}, \quad (2)$$

where  $X_{t,j}$  denotes the time to maturity and  $\beta_0, \beta_1, \beta_2$  and  $\lambda$  are parameters. Parsimonious structure and an ability to provide a good fit to the cross-section of yields at a given point in time is a key reason for its popularity. To understand the evolution of the interest rates over time, a dynamic representation was proposed by [Diebold and Li \(2006\)](#), replacing the above parameters with time-varying ones

$$Y_{t,j} = L_t + S_t \left\{ \frac{1 - \exp(-\lambda X_{t,j})}{\lambda X_{t,j}} \right\} + C_t \left\{ \frac{1 - \exp(-\lambda X_{t,j})}{\lambda X_{t,j}} - \exp(-\lambda X_{t,j}) \right\} + \varepsilon_{t,j} \quad (3)$$

$$= Z_t^\top m(X_{t,j}) + \varepsilon_{t,j}, \quad (4)$$

where  $Z_t = (L_t, S_t, C_t)^\top$  are the loadings,  $m(\cdot) = (\mathbf{1}, (1 - \exp(-\lambda(\cdot)))/\lambda(\cdot), (1 - \exp(-\lambda(\cdot)))/\lambda(\cdot) - \exp(-\lambda(\cdot)))$  common factors and  $X_{t,j}$  – time to maturity. Note that the decay factor  $\lambda_t = \lambda$  is tied down to a constant, since time variability of  $\lambda_t$  has a negligible impact on the model fit and forecasting performance. The NS factors, with country-specific  $\lambda$ , stemming from our estimation results, are plotted in [Figures 4 and 5](#). The yield latent factors  $L_t, S_t$  and  $C_t$  correspond to a level, slope and curvature of the yield curve, respectively. A first-order vector autoregressive (VAR) process models the time evolution of a vector of latent factor loadings  $Z_t$

$$Z_t = \mu + \mathcal{A}Z_{t-1} + \eta_t, \quad (5)$$

where  $\mathcal{A}$  is  $(3 \times 3)$  parameter matrix,  $\mu$  denotes a  $(3 \times 1)$  parameter vector and the  $(3 \times 1)$  vector  $\eta_t \sim N(0, H)$ ,  $H$  is the conditional variance which is assumed to be diagonal and constant over time. The estimation of the NS model follows a two-step procedure. Fixing the  $\lambda$  to predetermined value, the latent factor loadings  $Z_t$  are estimated separately at each time point using ordinary least squares (OLS). Then, in a second step, the estimated factors can be used in a autoregressive (AR) models as represented in Equation (5).

### 3.2 Dynamic semiparametric factor model

The DSFM generalizes the factor models given in Equations (1) and (4) to functions of the covariates  $X_{t,j}$ . Therefore the model takes the form

$$Y_{t,j} = \sum_{l=0}^L Z_{t,l} m_l(X_{t,j}) + \varepsilon_{t,j}. \quad (6)$$

We assume that the processes  $X_{t,j}$ ,  $\varepsilon_{t,j}$  and  $Z_t$  are independent. The number of underlying factors  $L$  should be smaller than the number of grid (maturity) points. The functions  $m_l(\cdot)$  are nonparametric, while the factors  $Z_{t,l}$  represent the parametric part. Following [Vasicek and Fong \(1982\)](#), [Krivobokova, Kauermann, and Archontakis 2006](#) and [Lin \(2002\)](#), we select a tensor B-spline basis to approximate  $m_l(\cdot)$ ,  $l = 0, \dots, L$ . More formally, the factors  $m_l(\cdot)$  are represented by  $A\psi(\cdot)$ , where  $A = (a_{l,k}) \in \mathbb{R}^{(L+1)K}$  denotes a coefficient matrix and  $\psi(\cdot) = (\psi_1, \dots, \psi_K)^\top$  is a vector of selected basis functions.  $K$  stands for the number of knots of the tensor B-splines functions. The number of knots  $K$  corresponds to a bandwidth parameter if compared to the kernel

Table 1. Explained variation in percent of the model with different numbers of factors  $L$  for the DSFM.

	$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$	$L = 6$
<i>Separated DSFM</i>						
GR	0.9349	0.9872	0.9985	0.9990	0.9995	0.9999
IT	0.7544	0.9899	0.9952	0.9968	0.9994	0.9995
PT	0.7936	0.9763	0.9961	0.9987	0.9990	0.9999
ES	0.8329	0.9874	0.9934	0.9963	0.9978	0.9983

smoothing technique of [Fengler, Härdle, and Mammen \(2007\)](#). Moreover, the functions  $m_l(\cdot)$  are orthonormalized ( $\|m_l(\cdot)\| = 1$  and  $\langle m_l, m_k \rangle = 0$  for  $l \neq k$ ) and identifiable up to scale and rotation transformation.

The model (6) can be rewritten in terms of B-splines basis and coefficient matrix  $A$  as follows:

$$\sum_{l=0}^L Z_{t,l} m_l(X_{t,j}) = \sum_{l=0}^L Z_{t,l} \sum_{k=1}^K a_{l,k} \psi_k(X_{t,j}) = Z_t^\top A \psi(X_{t,j}). \tag{7}$$

Estimation of B-splines coefficient matrix  $A$  and low-dimensional factor loadings  $Z_t$  is achieved via least-squares method. Thus, the estimates  $\hat{A}$  and  $\hat{Z}_t$  are given by the following formula:

$$(\hat{Z}_t, \hat{A}) = \arg \min_{Z_t, A} \sum_{t=1}^T \sum_{j=1}^J \{Y_{t,j} - Z_t^\top A \psi(X_{t,j})\}^2. \tag{8}$$

The non-linear optimization problem stated in Equation (8) might be solved by a Newton–Raphson iterative algorithm. Some weak conditions on the initial choice of  $\{\text{vec}(A^{(0)}), Z_t^{(0)}\}$  ensure the convergence to the true unknown parameters matrix  $A$  and factor loadings  $Z_t$ . It was proved by [Park et al. \(2009\)](#) that the differences between the estimates  $\hat{Z}_t$  and the true, unobserved loadings  $Z_t$  can be asymptotically neglected. This fact allows us to model the dynamics of factor loadings based on estimated time series and therefore study the dynamics of the main, high-dimensional object of interest (Table 1).

### 3.3 DSFM $L$ selection

An important parameter in our model is the number of factors (and corresponding factor loadings)  $L$ . The choice of  $L$  here is based on the explained variance by factors

$$EV(L) = 1 - \frac{\sum_{t=1}^T \sum_{j=1}^J \{Y_{t,j} - \sum_{l=0}^L Z_{t,l} m_l(X_{t,j})\}^2}{\sum_{t=1}^T \sum_{j=1}^J \{Y_{t,j} - \bar{Y}\}^2}. \tag{9}$$

In the PDSFM the number of factors is based on the model’s explained variance  $\overline{EV}$  which is an average of  $EV$  of all analyzed countries. We evaluate the model’s goodness-of-fit by the root mean squared error (RMSE) criterion

$$RMSE = \sqrt{\frac{1}{TJ} \sum_{t=1}^T \sum_{j=1}^J \left\{ Y_{t,j} - \sum_{l=0}^L Z_{t,l} m_l(X_{t,j}) \right\}^2}. \tag{10}$$

#### 4. Estimation results

To model the yield curve dynamics we implement both DSFM as a domestic model and the panel version PDSFM applied to all states simultaneously. We model first the term structure as a function of time to maturity solely. Second, following [Diebold, Rudebusch, and Aruoba \(2006\)](#), [Ang and Piazzesi \(2003\)](#) and [Hautsch and Ou \(2012\)](#) we include macroeconomic variables such as the inflation rate and the IP, which may have an impact on the term structure.

##### 4.1 Domestic yield curve modeling

In a first step, the DSFM was calibrated to the data set comprising the entire period for the term structures domestically (for Greece the period was truncated to 30 June 2011 due to extraordinary high observations). The curve dynamics are modeled in dependence of one regressor: the maturity time. As described in Section 2 the members of the yield curve are fixed across time. Thus, we specify the knots as the time to maturity grid and the order of tensor B-splines is set to 1. The results of the selection of factors  $L$  are reported in Table 1. The higher the number of factors, the better is the general fit, however at the cost of parsimony and robustness of the model. In order to choose the optimal  $L$  one proceeds similarly to principal component analysis by selecting the number of factors according to their contribution to the total variation. For domestically modeled curves a two-factor DSFM specification is sufficient (Figures 2 and 3).

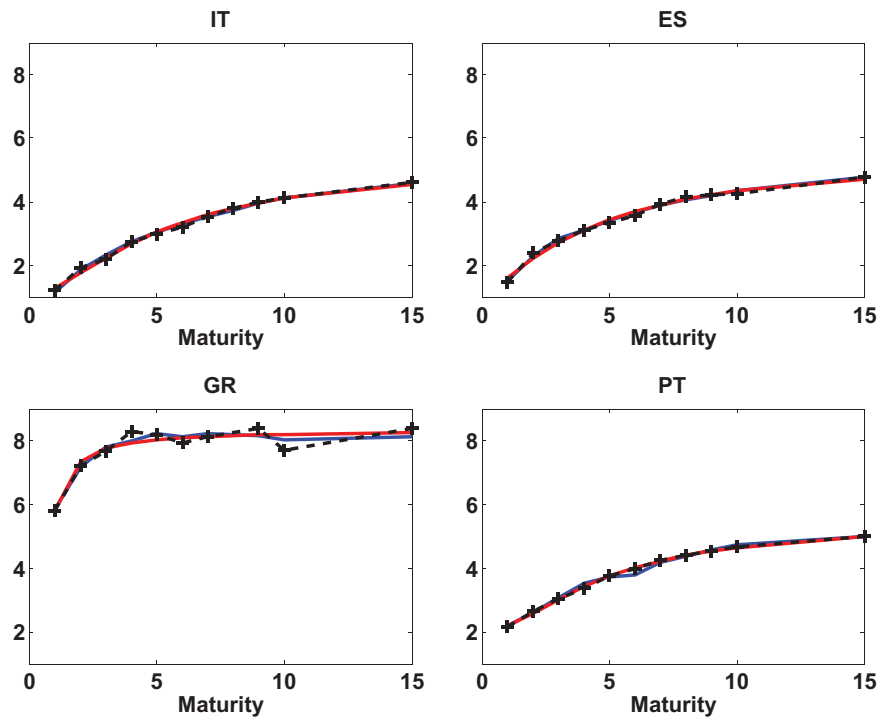


Figure 2. The term structure of interest rates (dotted black) observed on 20100331, DSFM (blue) and the NS fitted data. We use a DSFM specification with two factors.

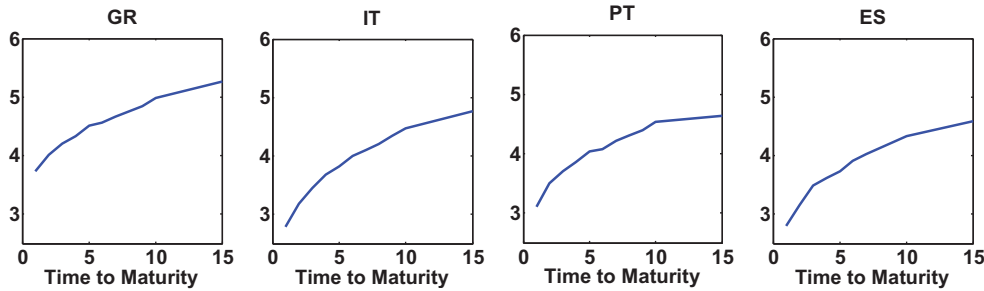


Figure 3. Estimated constant factors  $\hat{M}_0$  of the yield curve depending on time to maturity (years) using the domestic DSFM approach with two factors.

4.1.1 Estimated factors

Figure 4 depicts the estimated first factor. The first factor represents the slope similar to Diebold and Li (2006). We find out that the corresponding NS slope factor is strikingly different. The shape of the DSFM slope is remarkably similar across countries. The slope is steeper though for short maturities (especially for Greece), more weight is attributed to shorter maturities (1–3 years). We attribute it to the economic stagnation that depressed the short rates relative to the benchmark 10 year rate (although overall rates are high). The first factor for Greece is convex and increases slightly for the long maturities. For the remaining countries the slopes are almost identical. The second factor  $\hat{m}_2$  across countries is shown in Figure 5. We observe that they are different from the NS factors, decrease with the maturity, and exhibit a country-specific peak. The

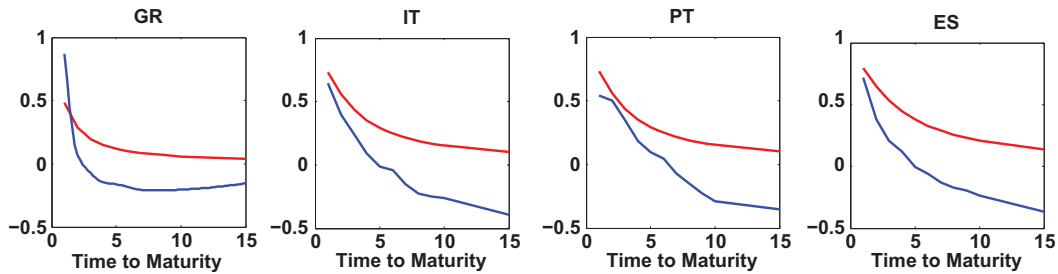


Figure 4. Estimated first factor of the yield curve depending on time to maturity (years) using the domestic DSFM approach (blue line) with two factors and Nelson–Sigel slope factor (red line) with  $\lambda_{GR} = 0.049$ ,  $\lambda_{IT} = 0.127$ ,  $\lambda_{PT} = 0.109$  and  $\lambda_{ES} = 0.174$ , respectively.

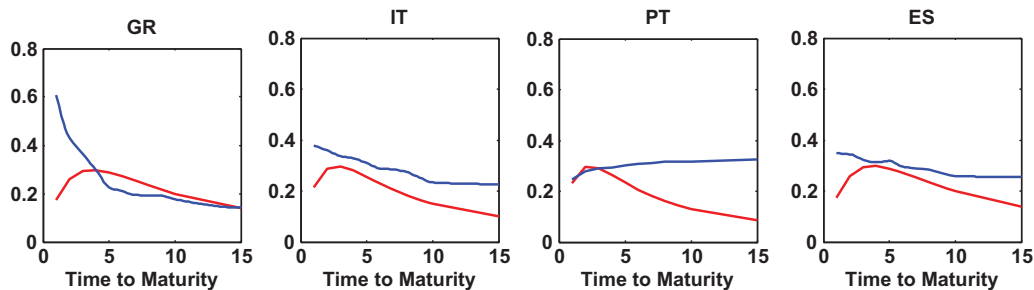


Figure 5. Estimated second factor of the yield curve depending on time to maturity (years) using the domestic DSFM approach (blue line) with two factors and Nelson–Sigel curvature factor (red line) with  $\lambda_{GR} = 0.049$ ,  $\lambda_{IT} = 0.127$ ,  $\lambda_{PT} = 0.109$  and  $\lambda_{ES} = 0.174$ .



DSFM second factor decreases, but for Portugal it increases with the time to maturity. We also attribute the second factor to the slope of the yield curve.

These findings can be summarized as follows. The nonparametric estimates are similar to the NS slope factor. There is no curvature factor present for the southern European yield curve dynamics. Their term structure of interest rates and extracted model factors are similar, just as characteristics of their economies are. The impact of the crisis is reflected by the steepness of the first DSFM factor, especially for severely struck Greece.

#### 4.1.2 Factor loadings and yield curve dynamics

Figure 6 displays the extracted time series  $\hat{Z}_t$  for the entire calibration period. The series shows high persistency and unit root  $I(1)$  behavior. This observation is in line with the general dynamics of the yield curve which does not change substantially over a small (monthly) time period. In Table 2 we report the stationarity and unit root tests on the first differences of extracted yield curve factor loadings.  $\Delta\hat{Z}_t \stackrel{\text{def}}{=} \hat{Z}_t - \hat{Z}_{t-1}$  are (weak) stationary processes ( $H_0$  is not rejected at significance level  $\alpha = 5\%$ ) for all analyzed countries. Based on those diagnostics we consider VAR as a suitable model for dynamics of the extracted  $\Delta\hat{Z}_t$ . The order  $p$  of VAR( $p$ ) is determined by Schwarz (SC) and Hannan-Quinn (HQ) information criteria (see Table A4). The selected specification will be kept for the remainder of the analysis.

#### 4.1.3 Yield curve modeling in dependence of further explanatory variables

Dynamic term structure models assume that the time evolution of the yield curve is driven by a (finite) number of latent state variables. A large body of literature studies the economic cause of yield curve factors, see Diebold, Piazzesi, and Rudebusch (2005) and Hautsch and Ou (2012).

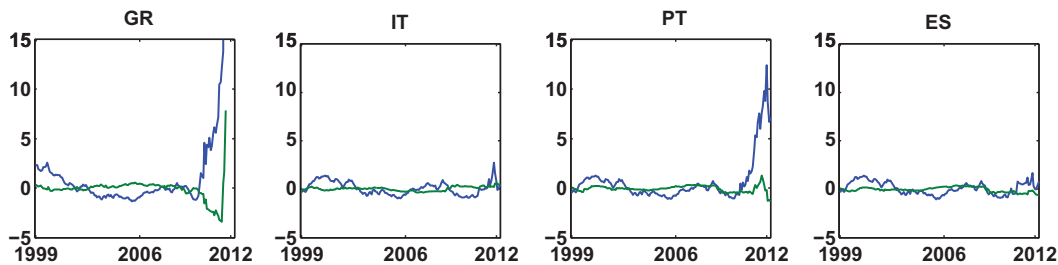


Figure 6. Estimated factor loadings  $\hat{Z}_t$  of the yield curve over whole sample using the domestic DSFM with two factors; blue line corresponds to  $\hat{Z}_{t,1}$ , green to  $\hat{Z}_{t,2}$ .

Table 2. KPSS, ADF for estimated first differences of factor loadings  $\Delta\hat{Z}_{t,1}$  (upper panel) and  $\Delta\hat{Z}_{t,2}$  (lower panel).

	GR	IT	PT	ES
KPSS	0.427	0.060	0.075	0.062
ADF	-2.492	-10.901	-15.454	-12.334
KPSS	0.209	0.068	0.068	0.072
ADF	-3.6425	-13.282	-11.502	-12.323

Note: Kwiatkowski–Phillips–Schmidt–Shin (KPSS):  $H_0$ : weak stationarity, critical values at 0.10, 0.05, 0.01 are 0.119, 0.146 and 0.216; augmented Dickey–Fuller (ADF):  $H_0$ : unit root, critical values at 0.01, 0.05, 0.10 are -1.61, -1.94 and -2.58.

The explicit relation between term structure and fundamental macroeconomic variables led to the Taylor rule (Taylor 1992; Ang and Piazzesi 2003). This approach provides a convenient way to relate yield curve dynamics with macro data. There are however residual variations in the term structure that are not captured and explained via the inclusion of macro variables. To this end we exploit the DSFM and implement additional regressors. The B-splines knots are an equally spaced grid (six knots). The lowest (highest) knot equals a minimum (maximum) of the explanatory variable, corrected by 2% and the quadratic B-splines basis is used. The results show stable behavior regarding the choice of the knots, see Table A5. In Figure 7 we show the estimated first factor  $\hat{m}_1(\Xi)$ ,  $\Xi_t \stackrel{\text{def}}{=} (X_{t,j}, \text{INF}_t)$  for domestic DSFM with harmonized consumer price index (INF) as a regressor. First, the structure of the factor does not differ much across countries, the impact of the inflation rate is similar for all states. Second, the highest impact is observed on the short rather than on the long rates with a peak at inflation rate around 2%. This central peak may be attributed to the target inflation rate of the central bank. For all countries we observe the decaying impact of the inflation rate for higher maturities; what is in line with expectations. Though the term structure and the harmonized consumer price index are interconnected, it does not improve the model's goodness-of-fit (see Table 3) due to complicity and computational limitations.

Figure 8 depicts the first factor  $\hat{m}_1(\Xi)$ ,  $\Xi_t \stackrel{\text{def}}{=} (X_{t,j}, \text{IP}_t)$  for domestic DSFM as a function of time to maturity and IP. We can summarize that the estimates does not differ much across countries and the changes in the production levels mainly affect the shorter maturities. As expected the impact of the current IP is decaying with time to maturity, slowly for Italy, Portugal and Spain and more rapidly for Greece. We observe that for Greece this dependence on long-term rates is almost negligible. Similarly to the results on the inflation as the additional regressor, the model's goodness-of-fit is not improved, see Table 3.

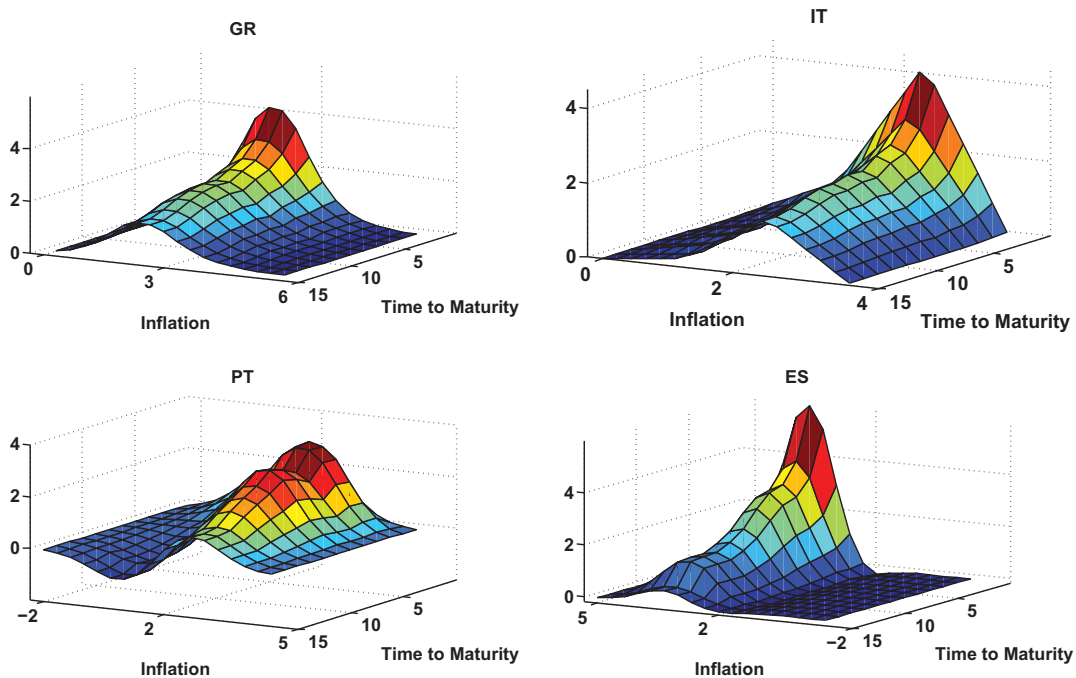


Figure 7. Estimated factors with respect to maturity and inflation rate using domestic DSFM approach with two factors.

Table 3. RMSE derived by NS model (NS), domestic DSFM, DSFM with inflation rate in dependence on time to maturity.

	GR	IT	PT	ES
NS	0.5600	0.0685	0.2009	0.0636
DSFM	0.2886	0.0872	0.4195	0.0695
DSFM(INF)	0.6813	0.1550	0.5520	0.2110
DSFM(IP)	0.6361	0.5141	0.7630	0.2423

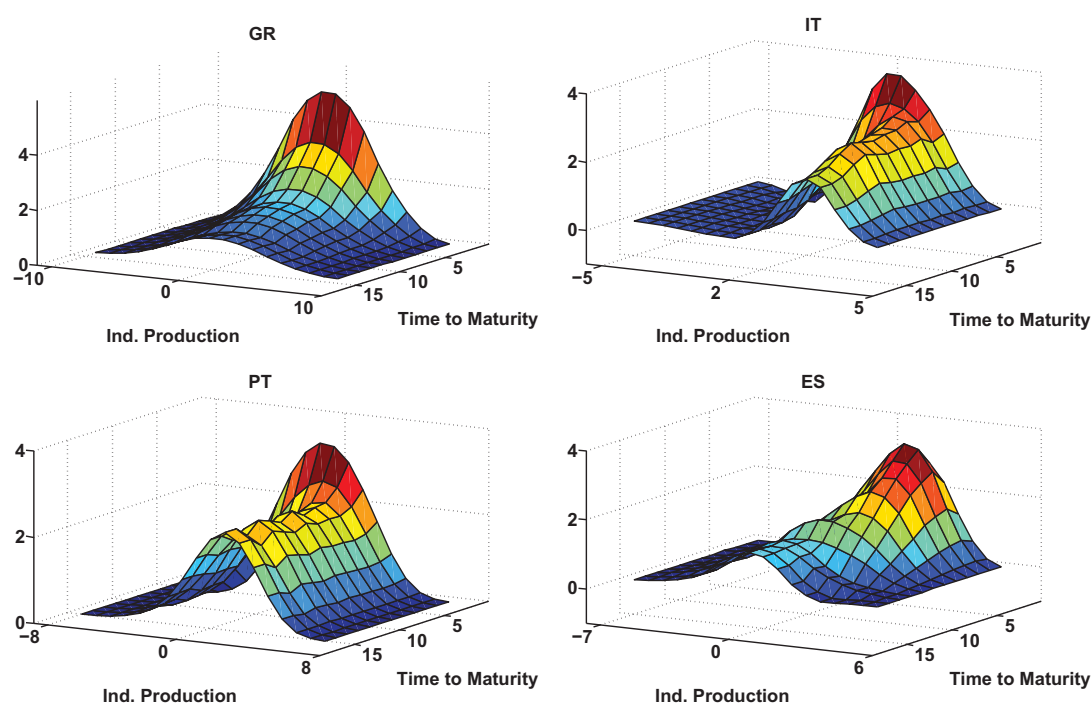


Figure 8. Estimated factors with respect to maturity and IP using domestic DSFM approach with two factors.

Table 3 presents the RMSE calculated for domestic, PDSFM approach and the NS model. One observes that the in-sample fit of the domestic DSFM and the dynamic NS model are remarkably similar. This stays in favor of the domestic DSFM approach, which captures the yield curve dynamics with just two dynamic factors. Second, the PDSFM (presented in Appendix A.1) fit is weaker. Thus, we concentrate on the domestic DSFM technique.

## 5. Factors and macroeconomic fundamentals

In this section we examine the relationship between the factor loadings and the macroeconomic environment. For simplicity of presentation we focus on Italy, which is the third largest bond market in the world and the largest economy among the countries considered. Our analysis is based on five macroeconomic variables: the harmonized consumer price index (INF), the manufacturing CU, the unemployment rate (EMP), IP and the real GDP ( $\Delta$ GDP). The variable selection is motivated by [Diebold, Rudebusch, and Aruoba \(2006\)](#) and [Hautsch and Ou \(2012\)](#). The analysis of the contemporaneous correlation between extracted yield curve factor loadings and macroeconomic

variables (observed monthly) is done by the regression

$$\Delta\hat{Z}_t = C + \beta_1\text{INF}_t + \beta_2\text{CU}_t + \beta_3\text{EMP}_t + \beta_4\text{IP}_t + \beta_5\Delta\text{GDP}_t + \varepsilon_t. \quad (11)$$

The results reported in Table 4 show that the differentiated estimated yield curve first factor loading is driven by the macroeconomic. The explanatory power of macroeconomic variables on the second factor reaches just 6%. We have to note here that both factor loadings and macroeconomic variables are relatively persistent what might cause spurious correlation effects. Thus, before analysis, the time series are detrended. Moreover, as expected, the Chow test (Chow 1960) for the regression models (for  $\Delta\hat{Z}_{t,1}$  and  $\Delta\hat{Z}_{t,2}$ ) before and after the bankruptcy of Lehman Brothers confirmed the structural break in the data at significance level  $\alpha = 0.05$ . The first factor is mainly driven by the inflation rate, real GDP and the IP (at a significance level  $\alpha = 0.05$ ). The positive signs of the INF and  $\Delta\text{GDP}$  coefficients are economically plausible and in line with the theory. For the second factor, due to the obvious structural break within the analyzed period, the shape of the yield curve cannot be explained by the macroeconomic fundamentals.

To investigate the predictability of the DSFM yield factors and their dynamic interdependencies between macroeconomic activity measures, we estimate a VAR(1) model of the yield factors and macroeconomic fundamentals

$$F_t = \mu + \text{A}F_{t-1} + \varepsilon_t, \quad (12)$$

where  $F_t \stackrel{\text{def}}{=} (\Delta\hat{Z}_{t,1}, \Delta\hat{Z}_{t,2}, \text{INF}_t, \text{CU}_t, \text{EMP}_t, \text{IP}_t, \Delta\text{GDP}_t)$ . The estimation results are shown in Table 5. We can summarize that the factor loadings primarily depend on their own lags and on those of other factors. Second, it is shown that factor loadings are not predictable, based on macroeconomic variables. The coefficients in the estimated VAR(1) matrices are significantly different than 0 for diagonal elements. We analyze the long-term relations between the yield curve factor loadings and macroeconomic variables by prediction error variance decomposition implied by the VAR estimates. We can summarize the following results. First, in the long perspective, prediction error variances of factor loadings  $\hat{Z}_t$  are not explainable by the macroeconomic fundamentals. The contribution is only up to 10%, see Figure A4. Hence, in line with Diebold, Rudebusch, and Aruoba (2006) we report that yield curve factor loadings are not predicable by the given macroeconomic data set.

## 6. Forecasting

### 6.1 Setup

The aim of this section of the paper is to analyze the model's forecasting performance, especially in comparison to the dynamic NS model as a natural competitor. We focus our analysis on the

Table 4. Linear regressions of monthly changes factor loadings  $\hat{Z}_t$  (separate approach) on (normalized) changes of the harmonized consumer price index (INF), log changes of the CU, changes of unemployment rate (EMP), changes of IP and the monthly changes in real log GDP ( $\Delta\text{GDP}$ ).

	CONST	INF	CU	EMP	IP	$\Delta\text{GDP}$	$R^2$
$\Delta\hat{Z}_{t,1}^{\text{IT}}$	-0.018	0.8353*	-0.018	-0.147	-0.621*	0.732*	0.16
$\Delta\hat{Z}_{t,2}^{\text{IT}}$	0.002	0.035	0.854	0.004	-0.202*	0.029	0.06

\*Significant at  $\alpha = 0.05$ .

Table 5. VAR(1) estimates of monthly IT data set: factor loadings  $\hat{Z}_t$  (domestic approach), (normalized) changes of the harmonized consumer price index (INF), log changes of the CU, changes of unemployment rate (EMP), changes of IP and the monthly changes in real GDP ( $\Delta$ GDP).

	$Z_{t,1}$	$Z_{t,2}$	$INF_t$	$CU_t$	$EMP_t$	$IP_t$	$\Delta GDP_t$
$Z_{t-1,1}$	0.158	-0.171	0.243	-0.021	-0.014	-0.154	0.179
$Z_{t-1,2}$	0.161	-0.076	-0.049	-0.137	-0.002	-0.242	0.034
$INF_{t-1}$	0.029	-0.107	0.306	-0.034	0.072	0.130	0.170
$CU_{t-1}$	0.068	-0.039	0.105	0.774	-0.001	0.036	0.020
$EMP_{t-1}$	-0.023	0.099	0.011	-0.029	-0.205	0.087	-0.315
$IP_{t-1}$	-0.070	0.030	0.019	0.036	-0.067	-0.383	0.297
$\Delta GDP_{t-1}$	-0.0323	0.123	-0.003	0.0927	-0.009	0.037	0.810

Note: Sample period January 1999–March 2012.

Italian and Spanish term structure data, the countries that kept access to the financial markets and were not bailed out. We undertake a short-term forecasting exercise in deriving term structure of interest rates monthly, in times of financial distress July 2007–March 2012 and January 2003–June 2007. The models are re-estimated every month exploiting the past information over a whole analyzed period. In accordance with our in-sample analysis reported in the previous section, the domestic DSFM approach with two factors without additional explanatory variables is applied. Second, the specified VAR( $p$ ) model for domestic term structure is used to forecast. As reported in Table A4 the order  $p = 1$  is chosen for both, Italy and Spain. A natural benchmark is the dynamic NS model, where the factor loadings are modeled by AR(1) processes (addressing high persistency and random-walk behavior, see Diebold and Li (2006); Diebold, Rudebusch, and Aruoba (2006)). The forecasting horizon is up to 12 months (observations) ahead. The prediction quality is measured using the RMSPE given by

$$\text{RMSPE} = \sqrt{\frac{1}{hJ} \sum_{t=1}^h \sum_{j=1}^J \left\{ Y_{t,j} - \sum_{l=0}^L \hat{Z}_{t,j} \hat{m}_l(X_{t,j}) \right\}^2}. \quad (13)$$

The prediction performance regarding particular maturities  $j$  is compared using the following formula:

$$\text{RMSPE}(j) = \sqrt{\frac{1}{h} \sum_{t=1}^h \left\{ Y_{t,j} - \sum_{l=0}^L \hat{Z}_{t,j} \hat{m}_l(X_{t,j}) \right\}^2}. \quad (14)$$

## 6.2 Forecasting results

The forecasting measures are displayed in Figures 9 and A5 for both the domestic DSFM and the dynamic NS model and show that the domestic DSFM does better than the dynamic NS model in a short-term forecasting exercise in times of financial distress. In the long horizon though, the dynamic NS model is a serious competitor. As expected, the term structure of interest rates cannot be well predicted based on its past observations in the long horizon. Second, the forecasting performance is better for short and long maturities in the crisis period. The non-parametrically estimated factors and parsimony of the model pay off, especially in times of financial distress. We refer here to the famous rule introduced by Zellner, Keuzenkamp, and McAleer (2002): ‘Keep

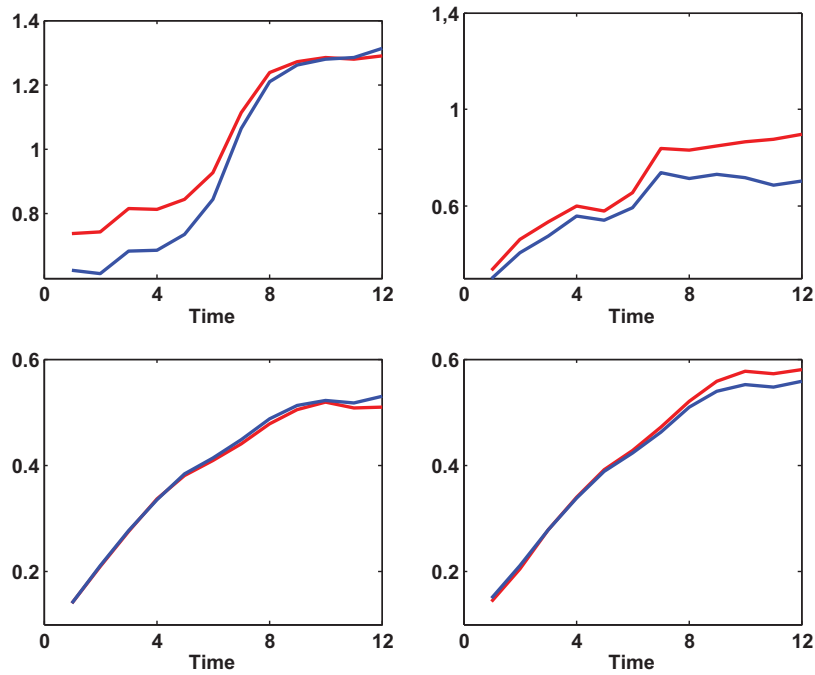


Figure 9. Root mean squared prediction error (RMSPE) derived by the domestic DSFM approach with two factors (blue) and by the dynamic NS for all forecasting horizons (in months) for Italy (left) and Spain (right); forecasting periods: 2007–2012 (upper panel) and 2003–2007 (lower panel).

it Sophisticatedly Simple’ (KISS). The inferior forecasting performance of dynamic NS model for long maturities might be explained by its general difficulty to fit for longer maturities. In the non-crisis period 2003–2007 we report comparable performance in the short and long forecasting horizon. The DSFM approach does not improve prediction power and both models can be used equivalently.

Table 6 shows the RMSPE averaged over short-term forecasting periods for the domestic DSFM approach and the dynamic NS model in the financial distress. Summarizing one concludes that the overall prediction performance of the DSFM approach is improved compared to the market benchmark for the crisis period 2007–2012. Nevertheless, both model perform comparably in low-volatile market conditions, as reported in Table 7 (Figure 10).

Table 6. Averaged RMSPE over six month forecasting horizon for the domestic DSFM approach and the dynamic NS model for 1, 5, 8, 10 year maturities and for the entire yield curve for Italy and Spain; forecasting period 2007–2012.

	1-year	5-year	8-year	10-year	Overall
<i>Italy</i>					
DSFM	0.8600	0.6893	0.5778	0.6575	0.6309
NS	1.2682	0.6379	0.6564	0.7191	0.7052
<i>Spain</i>					
DSFM	0.5142	0.7011	0.5940	0.5635	0.6123
NS	0.5595	0.7174	0.6250	0.6169	0.6569

Table 7. Averaged RMSPE over six month forecasting horizon for the domestic DSFM approach and the dynamic NS model for 1, 5, 8, 10 year maturities and for the entire yield curve for Italy and Spain; forecasting period 2003–2007.

	1-year	5-year	8-year	10-year	Overall
<i>Italy</i>					
DSFM	0.3161	0.3383	0.3505	0.3286	0.3355
NS	0.3098	0.3455	0.3274	0.3189	0.3312
<i>Spain</i>					
DSFM	0.3385	0.3472	0.3256	0.3212	0.3385
NS	0.3134	0.3622	0.3426	0.3311	0.3434

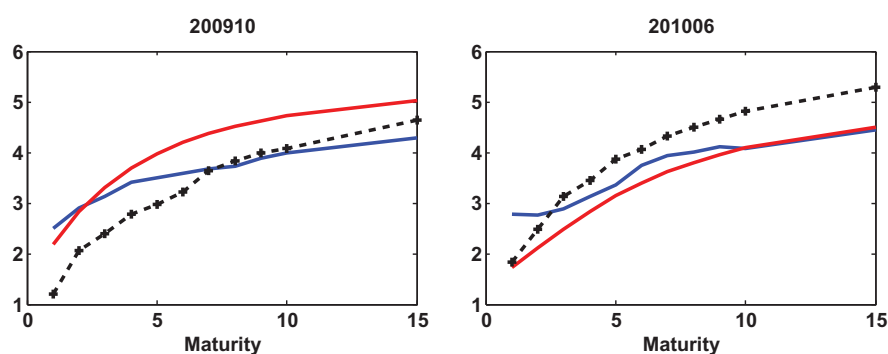


Figure 10. Term structure of interest rates (dotted black) observed on 30 October 2009 (left) and 30 June 2010 (right) for Italy with the DSFM (blue) and the dynamic NS (red) forecasts.

## 7. Conclusion

We propose a DSFM to model the term structure of interest rates. The DSFM approach was encouraged by the success of factor models. The assumption of parametric, exponential form of the NS factors is relaxed, they are estimated nonparametrically. Our framework is flexible and parsimonious. That makes it a useful tool, when standard models fail. The time evolution of south European zero-curves is described by two dynamic factor loadings and one constant function that corresponds to the ‘averaged’ yield curve.

The model is applied to four southern European bond markets over the period January 1999–March 2012. The focus is on the recent European sovereign-debt crisis. It is shown that two underlying factors can explain more than 95% of in-sample variations of the domestic zero-curve dynamics. Both factors (ordered in terms of explained variance) correspond to the yield curve’s slope. The proposed model achieves an explanatory power of 98%, where the inclusion of the third factor does not lead to a significantly better in-sample fit. The extracted factor loadings are unit root processes and reveal high persistency, similar to the zero-curves. The contemporaneous relation with macroeconomic fundamentals is not clearly revealed by the regression analysis due to a structural break in the data. We reported the  $R^2$  criterion 16% for the first factor and 6% for the second one. Though it is known that yield curves are driven by explanatory variables, i.e. the inflation rate, those variables do not improve the model’s goodness-of-fit.

## Acknowledgment

The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through CRC 649 'Economic Risk'.

## References

- Ang, A., and M. Piazzesi. 2003. "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables." *Journal of Monetary Economics* 50 (4): 745–787.
- Borak, S., and R. Weron. 2008. "A Semiparametric Factor Model for Electricity Forward Curve Dynamics." *Journal of Energy Markets* 1 (3): 3–16.
- Bowsher, C. G., and R. Meeks. 2008. "The Dynamics of Economic Functions: Modeling and Forecasting the Yield Curve." *Journal of the American Statistical Association* 103 (484): 1419–1437.
- Brüggemann, R., W. Härdle, J. Mungo, and C. Trenkler. 2008. "VAR Modelling for Dynamic Semiparametric Factors of Volatility Strings." *Journal of Financial Econometrics* 6 (3): 361–381.
- Chow, G. C. 1960. "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." *Econometrica* 28 (3): 591–605.
- Christensen, J., F. Diebold, and G. Rudebusch. 2009. "An Arbitrage-Free Generalized Nelson–Siegel Term Structure Model." *The Econometrics Journal* 12 (3): 33–64.
- Coroneo, L., K. Nyhlon, and R. Vidova-Koleva. 2008. "How Arbitrage Free Is Nelson–Siegel Model?" Working Paper 874, European Central Bank.
- Diebold, F., M. Piazzesi, and G. Rudebusch. 2005. "Modeling Bond Yields in Finance and Macroeconomics." *American Economic Review* 95 (2): 415–420.
- Diebold, F., G. Rudebusch, and S. Aruoba. 2006. "The Macroeconomy and Yield Curve: A Dynamic Latent Factor Approach." *Journal of Econometrics* 131 (1–2): 309–338.
- Diebold, F. X., and C. Li. 2006. "Forecasting the Term Structure of Government Bond Yields." *Journal of Econometrics* 130 (2): 337–364.
- Fengler, M. R., W. Härdle, and E. Mammen. 2007. "A Dynamic Semiparametric Factor Model for Implied Volatility String Dynamics." *Journal of Financial Econometrics* 5 (2): 189–218.
- Gürkaynak, R. S., B. Sack, and J. H. Wright. 2010. "The TIPS Yield Curve and Inflation Compensation." *American Economic Association* 2 (1): 70–92.
- Härdle, W., and S. Trück. 2010. "The Dynamics of Hourly Electricity Prices." SFB 649 DP 2009-013.
- Hautsch, N., and Y. Ou. 2012. "Yield Curve Factors, Term Structure Volatility, and Bond Risk Premia." SFB 649 DP 2008-53.
- Heath, D., R. Jarrow, and A. Morton. 1992. "Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation." *Econometrica* 60 (1): 77–105.
- Hull, J., and A. White. 1990. "Pricing Interest Rate Derivative Securities." *Review of Financial Studies* 3 (4): 573–592.
- Krivobokova, T., G. Kauermann, and T. Archontakis. 2006. "Estimating the Term Structure of Interest Rates Using Penalized Splines." *Statistical Papers* 47 (3): 443–459.
- Lin, B. H. 2002. "Fitting Term Structure of Interest Rates Using B-splines: The Case of Taiwanese Government Bonds." *Applied Financial Economics* 12 (1): 57–75.
- Longstaff, F. A., and E. S. Schwartz. 1992. "Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model." *Journal of Finance* 47 (4): 1259–1282.
- Park, B., E. Mammen, W. Härdle, and S. Borak. 2009. "Time Series Modelling with Semiparametric Factor Dynamics." *Journal of the American Statistical Association* 104 (485): 284–298.
- Ramsay, J., and B. Silverman. 1997. *Functional Data Analysis*. Springer series in statistics. New York: Springer.
- Svensson, L. E. O. 1995. "Estimating Forward Interest Rates with the Extended Nelson–Siegel Method." *Quarterly Review Sveriges Riksbank* 3: 13–26.
- Taylor, J. B. 1992. "Discretion Versus Policy Rules in Practice." *Carnegie-Rochester Conference Series on Public Policy* 39: 195–214.
- van Bömmel, A., S. Song, P. Majer, P. N. C. Mohr, H. R. Heekeren, and W. K. Härdle. 2013. "Risk Patterns and Correlated Brain Activities. Multidimensional Statistical Analysis of fMRI Data in Economic Decision Making Study." *Psychometrika*. doi:10.1007/s11336-013-9352-2
- Vasicek, O., and H. G. Fong. 1982. "Term Structure Modeling Using Exponential Splines." *Journal of Finance* 37 (2): 339–356.



Zellner, A., H. A. Keuzenkamp, and M. McAleer. 2002. *Simplicity, Inference and Modelling*. Cambridge: Cambridge University Press.

## Appendix

### A.1 Panel DSFM

Dynamics of the term structure of interest rates can be modeled separately for each country, similarly to other DSFM applications (Borak and Weron 2008; Härdle and Trück 2010). However, following the spirit of NS model we wish to have common factors for all the analyzed data, and the monetary-specific behavior captured by factor loadings  $Z_t^i$ , where  $i$  is the country index. Therefore, to analyze all investigated yield curves  $i$  simultaneously, we extend Equation (7) to a panel dynamic semiparametric factor model (PDSFM) van Bömmel et al. 2013:

$$Y_{t,j}^i = m_0(X_{t,j}) + \sum_{l=1}^L Z_{t,j}^i \bar{m}_l(X_{t,j}) + \varepsilon_{t,j}^i, \quad 1 \leq j \leq J, 1 \leq t \leq T, 1 \leq i \leq I. \quad (A1)$$

$Z_{t,j}^i$  is the fixed individual effect for country  $i$  on function  $\bar{m}_l$  at time point  $t$ .

The PDSFM (A1) ensures exactly the same spatial structure of factors among all investigated bond markets. The joint spatial factors are denoted as  $\bar{m}_l$ ,  $l = 1, \dots, L$ . The term structure differences between the countries and time evolution are captured by their loading time series  $Z_{t,j}^i$ . The model estimation procedure is similar to DSFM estimation, however instead of Equation (8), similarly to common panel data models the sum of squared residuals is minimized

$$S(Z^1, \dots, Z^I, A) \stackrel{\text{def}}{=} \sum_{i=1}^I \sum_{t=1}^T \sum_{j=1}^J \{Y_{t,j}^i - Z_t^{i\top} A \psi(X_{t,j})\}^2. \quad (A2)$$

It is worth noting that given  $(Z^1, \dots, Z^I)$  or  $A$ , function  $S$  in Equation (A2) is quadratic with respect to other variables and therefore the solution can be found by OLS method. To find the solution  $(\hat{Z}_t^1, \dots, \hat{Z}_t^I, \hat{A}) = \arg \min_{Z^1, \dots, Z^I, A} S(Z^1, \dots, Z^I, A)$  we adopt the following iterative algorithm, similarly to Fenger, Härdle, and Mammen (2007). (i) Given the initial choice of  $(Z^{1,(0)}, \dots, Z^{I,(0)})$  minimize  $S(Z^{1,(0)}, \dots, Z^{I,(0)}, A)$  with respect to  $A$ , the explicit solution is given by OLS estimate  $A^{(1)}$ . (ii) given the  $A^{(1)}$  minimize  $S(Z^1, \dots, Z^I, A^{(1)})$  with respect to  $(Z^1, \dots, Z^I)$ . (iii) iterate (i) and (ii) until convergence. The algorithm runs until only minor changes occur.

### A.2 Panel yield curve modeling

The domestic interest rate data are demeaned by the country-specific constant factor  $\hat{m}_0$ . For the PDSFM model selection the one-factor model achieves an explanatory power of 78%, while the inclusion of the second and third factors improves the fit to 94% and 98%, respectively. The marginal contribution of the fourth factor is relatively small, thus from now on we only consider results for PDSFM specification with  $L = 3$ . The sample period was truncated to 30 June 2011 due to extraordinary high observations.

The estimated three factors of PDSFM are depicted in Figure A1. The first factor is almost constant over all different maturities, thus one can attribute it to the overall level of the yield curve. The slope structure of the second PDSFM factor

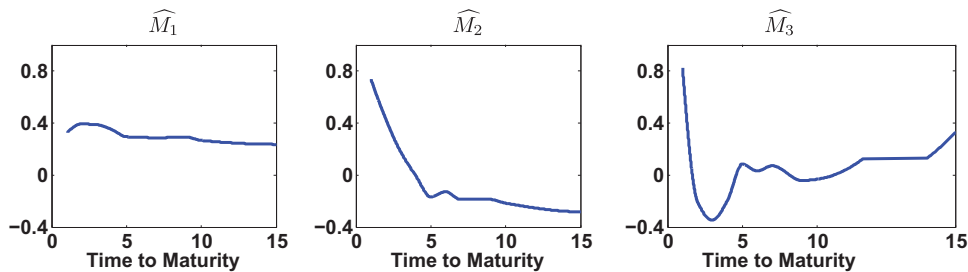


Figure A1. Estimated factors of the yield curve depending on time to maturity (years) using PDSFM with three factors.

is noticeably similar to the NS framework. The third function though does not have a counterpart in the NS model. It is decreasing for the short maturities and has a bump around the six year rate. It is worth noting that the overall performance of the PDSFM is worse than the domestic DSFM approach. One has to include additional factor to explain the same proportion of variation in the data. As expected, the analyzed countries, while sharing some common characteristics, are remarkably different with respect to the bond market (volume, liquidity) and economic policy. Those differences are reflected by the higher order of the model.

### A.3 Descriptive statistics and sensitivity analysis

Table A1. Explained variation in percent of the model with different numbers of factors  $L$  for the PDSFM.

	$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$	$L = 6$
<i>PDSFM</i>						
GR	0.9347	0.9782	0.9970	0.9984	0.9998	0.9999
IT	0.8529	0.9088	0.9857	0.9946	0.9967	0.9982
PT	0.9108	0.9507	0.9883	0.9957	0.9973	0.9973
ES	0.8529	0.9088	0.9857	0.9946	0.9968	0.9976
$\overline{EV}$	0.8999	0.9431	0.9896	0.9963	0.9906	0.9983

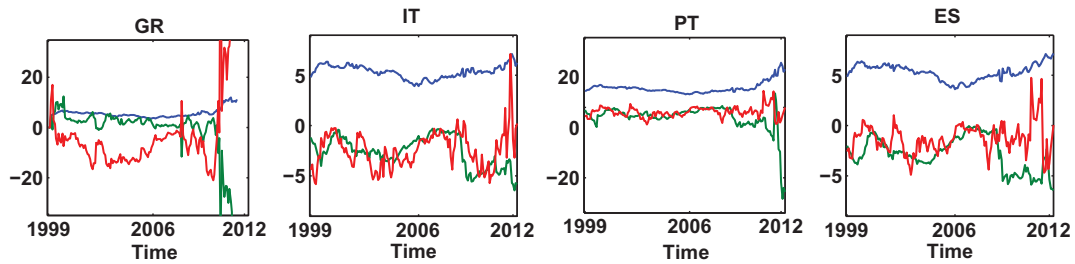


Figure A2. Estimated NS factors:  $L_t$  level (blue),  $S_t$  slope (green) and  $C_t$  curvature (red) for Greece, Italy, Portugal and Spain with  $\lambda_{GR} = 0.049$ ,  $\lambda_{IT} = 0.127$ ,  $\lambda_{PT} = 0.109$  and  $\lambda_{ES} = 0.174$ , respectively.

Table A2. Statistical summary of the level and change series of 1, 3, 5, 10-year zero-coupon bond yields.

	Mean	Median	SD	Skewness	Kurtosis
<i>Greece</i>					
<i>Levels</i>					
1-year	4.6969	3.6245	4.1007	4.7708	37.7046
3-year	5.3384	4.1219	4.2326	3.1321	13.9549
5-year	5.3026	4.3569	2.9831	2.4804	8.7310
10-year	5.7074	5.1632	2.0136	2.0951	7.4392
<i>Changes in</i>					
1-year	-0.0009	0.0155	1.9473	-8.4143	89.4413
3-year	-0.0008	0.0265	1.0061	-2.9892	24.7871

(Continued).

Table A2. Continued.

	Mean	Median	SD	Skewness	Kurtosis
5-year	-0.0002	0.0175	0.6759	-2.3611	17.9979
10-year	-0.0001	0.0012	0.3936	-2.3989	13.6143
<i>Italy</i>					
Levels					
1-year	2.8736	2.7843	1.1719	0.0523	2.1771
3-year	3.5394	3.4493	0.9836	0.4591	3.0203
5-year	3.9361	3.8228	0.8649	0.6655	3.5370
10-year	4.6039	4.4694	0.6801	0.5987	3.4534
Changes in					
1-year	0.0066	-0.0164	0.3646	1.3801	12.7428
3-year	-0.0048	-0.0056	0.3826	0.1485	9.5065
5-year	-0.0085	0.0062	0.3453	-0.0403	8.8564
10-year	-0.0101	0.0151	0.2476	-0.0234	7.2980

Note: The sample of Greek data is from January 1999 to June 2011; the sample for Italy is from January 1999 to March 2012; SD denotes standard deviation.

Table A3. Statistical summary of the level and change series of 1, 3, 5, 10-year zero-coupon bond yields.

	Mean	Median	SD	Skewness	Kurtosis
<i>Portugal</i>					
Levels					
1-year	3.5353	3.1029	2.1676	2.3915	10.3401
3-year	4.5377	3.7105	3.3085	2.9769	11.9225
5-year	4.8341	4.0366	2.9935	2.9600	11.7316
10-year	5.2021	4.5398	2.1242	2.6980	10.3645
Changes in					
1-year	-0.0039	-0.0136	0.7707	3.5002	28.6326
3-year	-0.0679	-0.0120	0.9918	0.2136	22.5985
5-year	-0.0669	0.0075	0.8773	-0.4221	20.3009
10-year	-0.0512	0.0002	0.4856	-0.6836	15.7143
<i>Spain</i>					
Levels					
1-year	2.8556	2.8062	1.1540	0.0056	2.0978
3-year	3.4844	3.4875	0.8802	0.1938	2.0900
5-year	3.8672	3.7370	0.7923	0.2639	2.0215
10-year	4.4929	4.3304	0.6960	0.2219	2.1219
Changes in					
1-year	0.0049	-0.0300	0.3226	1.6994	12.8158
3-year	-0.0050	0.0048	0.3662	0.6379	12.3061
5-year	-0.0087	0.0150	0.3364	0.0259	11.7166
10-year	-0.0124	0.0074	0.2611	0.0387	9.4403

Note: The sample is from January 1999 to March 2012.

Table A4. HQ and SC information criteria for the VAR( $p$ ) model for Italy and Spain.

IT	$p$	1	2	3	4
SC		-8.04	-7.98	-7.90	-7.83
HQ		-8.11	-8.10	-8.07	-8.04
ES	$p$	1	2	3	4
SC		-7.65	-7.53	-7.48	-7.40
HQ		-7.73	-7.68	-7.63	-7.61

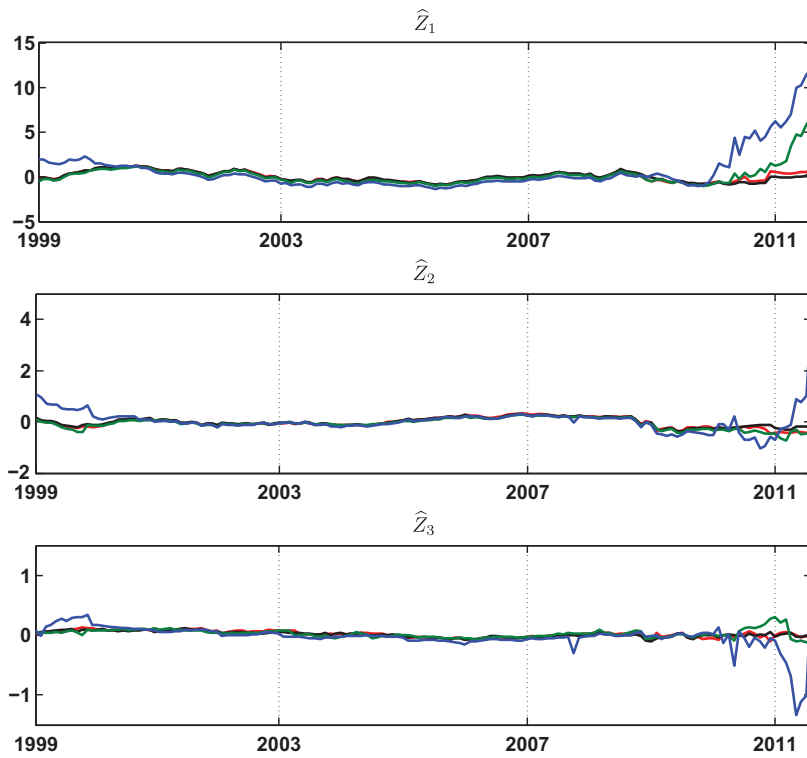


Figure A3. Estimated factor loadings  $\hat{Z}_{t,1}$  (top),  $\hat{Z}_{t,2}$  (middle) and  $\hat{Z}_{t,3}$  (bottom) of the yield curve over the whole sample using PDSFM with three factors; blue lines corresponds to  $\hat{Z}_t^{GR}$ , red to  $\hat{Z}_t^{ES}$ , green to  $\hat{Z}_t^{PT}$  and black to  $\hat{Z}_t^{IT}$ .

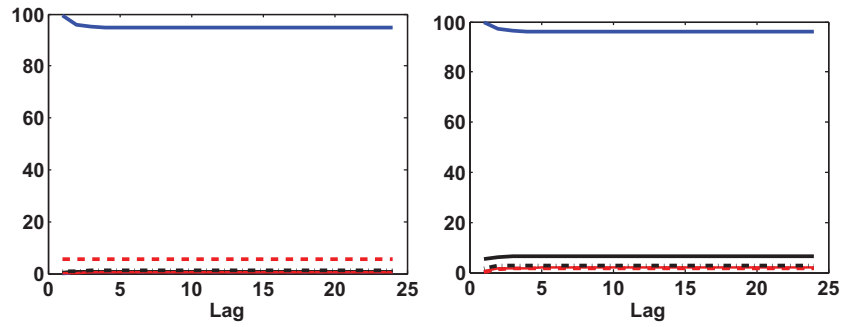


Figure A4. Prediction error decomposition of the first factor loadings  $\hat{Z}_{t,1}$  (left panel) and  $\hat{Z}_{t,2}$  (right panel). Based on a VAR(1) model of yield factors and macro factors using a Cholesky decomposition of the covariance. Extracted factor loadings and macroeconomic fundamentals for Italy.

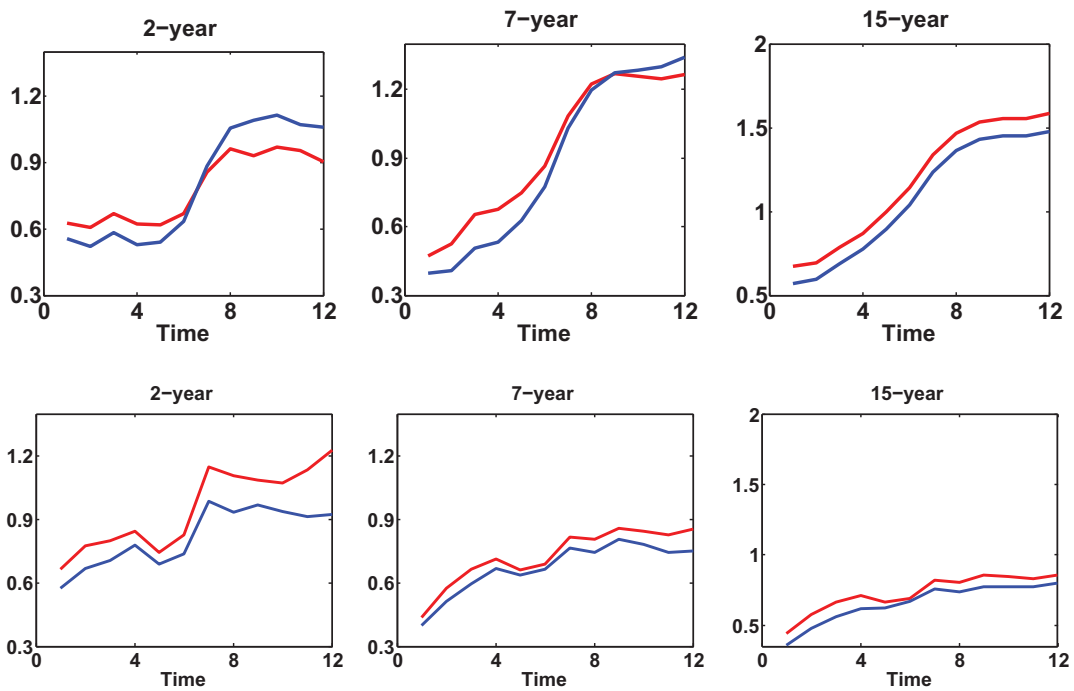


Figure A5. Root mean squared prediction errors (RMSPE( $j$ )) derived by the domestic DSFM approach with two factors (blue) and by the dynamic NS for all forecasting horizons (in months) for Italy (upper panel) and Spain (lower panel) for maturities: 2 years (1st column), 7 years (2nd column) and 15 years (3rd column); forecasting period: 2007–2012.

Table A5. Explained variation in percent of the model with different numbers of knots for the DSFM with inflation rate and IP.

	GR	IT	PT	ES
<i>INF</i>				
4	0.9466	0.9765	0.9597	0.9781
6	0.9523	0.9804	0.9611	0.9785
8	0.9597	0.9823	0.9645	0.9789
<i>IP</i>				
4	0.9355	0.9698	0.9422	0.9645
6	0.9408	0.9780	0.9451	0.9657
8	0.9437	0.9796	0.9477	0.9676

# Dynamic activity analysis model-based win-win development forecasting under environment regulations in China

Shiyi Chen · Wolfgang K. Härdle

Received: 31 October 2013 / Accepted: 14 May 2014 / Published online: 1 June 2014  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** Porter hypothesis states that environmental regulation may lead to win-win opportunities, that is, improve the productivity and reduce the undesirable output simultaneously. Based on directional distance function, this paper proposes a novel dynamic activity analysis model to forecast the possibilities of win-win development in Chinese industry between 2011 and 2050. The consistent bootstrap estimation procedures are also developed for statistical inference of the point forecasts. The evidence reveals that the appropriate energy-saving and emission-abating regulation will significantly result in both the net growth of potential output and the increasing growth of total factor productivity for most industrial sectors in a statistical sense. This favors Porter hypothesis.

**Keywords** Dynamic activity analysis model · Win win development · Environmental regulations · China industry

## 1 Introduction

In the recent 20 years, the relationship among energy, environment and economy (3E) has always been a focal topic of scholars and policy makers. The traditional

---

S. Chen (✉)

China Center for Economic Studies, School of Economics, Fudan University,  
Handan Road 220, Shanghai 200433, China  
e-mail: shiyichen@fudan.edu.cn

W. K. Härdle

Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin,  
Spandauer Straße 1, 10178 Berlin, Germany

W. K. Härdle

Lee Kong Chian School of Business, Singapore Management University,  
Singapore, Singapore

established notion on environmental protection is that the extra costs government imposes on the firms can jeopardize their international competitiveness. Porter, however, first challenged this argument in his one-page paper published in 1991 (Porter 1991). He regarded large energy consumption and pollutant emission as a form of economic waste and a sign of incompleteness and inefficiency of resources using. In his opinion, the amelioration of this inefficiency will provide firms with the win-win opportunity of improving both the productivity and environment. And the efforts of environmental protection can help firms to identify and eliminate the production inefficiency and regulatory disincentives that prevent the simultaneous improvements in both productivity and environmental quality. Thus, whether these types of environmental policy initiatives are successful depends on the extent to which such inefficiencies are widespread in the sub-industries, particularly in the energy/pollution intensive industries. However, due to deficient management systems, firms are not aware of certain opportunities and that environmental policy might open the eyes. Porter and Linde (1995) further emphasized that properly designed environmental protection policy in the form of economic incentives can trigger innovation that may partially or fully offset the costs of complying with them. Such innovation offsets occur mainly because pollution regulation is often coincident with improved efficiency of resource usage and the inference is that stiffer environmental regulation results in greater productivity and competence. These arguments are titled as Porter hypothesis (Ambec and Barla 2002). Admittedly, many scholars criticize Porter hypothesis, arguing that it is a fundamental challenge to efficient market hypothesis and neoclassical theory, and if it does exist it will be unnecessary for the government to impose extra environmental protective costs on the firms. They question why firms do not see these win-win opportunities by themselves, which at least implies that the argument does not have a general validity (Palmer et al. 1995; Jaffe et al. 1995; Faucheux and Nicolai 1998).

Empirical researches have provided arguments for both positions and have not been conclusive so far.<sup>1</sup> There are very rare studies to investigate the validity of Porter hypothesis in China, though it is critically important, too. Now China is the largest energy consumer and CO<sub>2</sub> emitter in the world, which brings China much abatement pressure from the outside world. The limited energy resources and serious pollution emissions have also made the traditional growth model in China unsustainable. To transform the economic growth model and challenge the climate change, in 2009 China decided to abate the CO<sub>2</sub> intensity by 40–45 % till 2020 as opposed to the benchmark level in 2005. Though it is only the relative carbon abatement, rather the absolute reduction employed by most countries, it is still challenging for China to realize it due to its coal oriented energy consumption structure and extensive factor-driving growth model. In particular, environment regulations will use up the limited resources which may be put into other productions and very likely influence the economic growth. Hence, an in-depth analysis is needed on both the positive and negative influence of

---

<sup>1</sup> Many empirical researches support Porter hypothesis, such as Karvonen (2001); Mohr (2002); Murty and Kumar (2003); Beaumont and Tinch (2004); Cerin (2006); Greaker (2006); Kuosmanen et al. (2009); Groom et al. (2010); Zhang and Choi (2013) There are also a few papers whose conclusion is neutral or against Porter hypothesis, see Boyd and McClelland (1999); Xepapadeas and Zeeuw (1999); Feichtinger et al. (2005).



environment regulations on China's economy, including the output and productivity growth. It is also a quite practical and edging issue to search for an optimal energy-saving and emission-abating path that can induce a win-win development for China in the following decades. Both motivate the research in this paper. As is known to all, on average, the industry counts for near 70 % of total energy consumption and over 80 % of the total CO<sub>2</sub> emission in China, which makes it the primal target to save energy and abate emissions. However, China is currently in the middle stage of industrialization, in which energy and emission intensive sectors such as iron and steel, cement and chemistry industries will continue to play pivotal roles in future economic growth. Thus, we can foresee there will be more negative impact brought by energy-saving and emission-abating activities on China's industry. Therefore, this paper focuses on the win-win forecasting in China's industrial economy.

As denoted above, the empirical results on win-win development possibilities are conflicting, which may be due to different dataset for analysis, the regulatory regime in a country, different cultural setting, the customer behavior, the type of industries or size of companies to be analyzed, and the time span and so on. However, the main reason may be the lack of a reasonable theoretical framework within which to investigate the links between environmental regulation and economic performance (Schaltegger and Synnestevedt 2002). For example, the commonly used CGE model fits static analysis well but its dynamic extension in empirical study is still rather scarce and too simple. Parametric econometric model is restricted to its priori functional form and distribution assumption. Traditional data envelopment analysis (DEA) and Shepherd distance function cannot distinguish the different characteristics between desirable output like GDP and undesirable output such as pollutions. Not until the presence of directional distance function (DDF) do we find a reasonable framework to capture the difference between desirable and undesirable outputs, and to model the behavior of increasing desirable output while decreasing undesirable output simultaneously. DDF allows for the type of inefficiency that is typified by Porter hypothesis, providing the most appropriate approach to examine Porter hypothesis. By following Boyd et al. (2002), this paper makes use of two types of DDF based on the strong and weak disposability assumption of pollution emissions to measure the potential output gain and loss, and uses the standard DDF based Malmquist–Luenberger Productivity Index (MLPI) to forecast the change of total factor productivity (TFP) and its components. In order to forecast the win-win development possibility from now on to the year of 2050 and find the optimal environment regulatory path, this paper designs different energy-saving and emission-abating paths and proposes a new dynamic activity analysis model (AAM) in which the different paths are introduced into the direction vector of DDF to examine the influence of different regulation paths on the win-win development possibilities in China in the following 40 years. However, there clearly exists the uncertainty surrounding these forecasts due to sampling variation. It is not enough to know whether the forecasts indicate increases or decreases in efficiency and productivity, but whether the indicated changes are significant in a statistical sense. This paper develops a consistent bootstrap estimation procedure to obtain the confidence intervals for potential net output gain and the index of productivity and its decompositions. The bootstrap methodology is an extension of earlier work by Simar and Wilson (1998, 1999).

The rest of this paper is organized as below: Sect. 2 introduce the dynamic activity analysis model firstly proposed in this paper. How to measure the potential output gain and loss and the specification of the Malmquist–Luenberger productivity index are also illustrated in the section. Section 3 firstly designs different energy-saving and emission-abating paths, which will be added into the direction vector of DDF so as to extend the AAM into dynamic version. The section also designs the bootstrap procedure that allows us to make the distinctions between a real change in potential output and productivity and an artifact of sampling noise. Section 4 selects an optimal environment regulatory path for China’s industrial win-win development during 2011 and 2050, and discusses the corresponding forecasts of potential output gain and loss, the evolution of productivity, technique and efficiency change among a set of sectors, and their statistical significance. Section 5 concludes this paper.

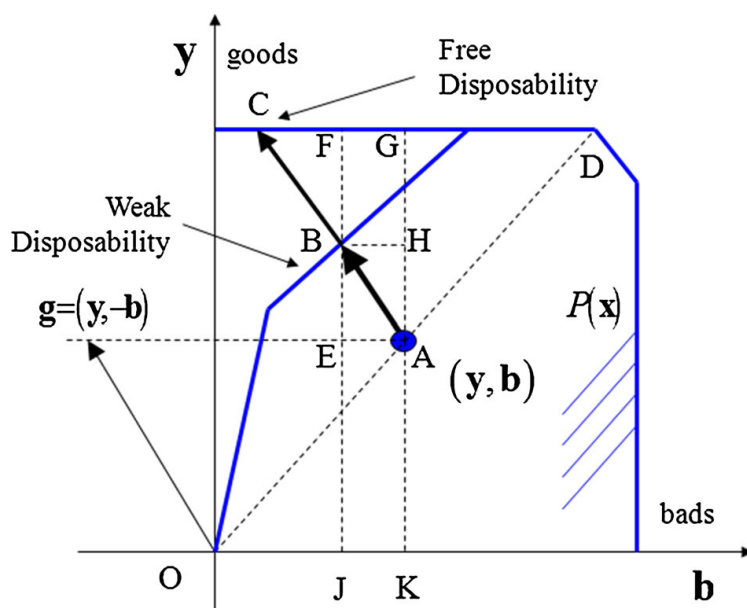
## 2 Dynamic activity analysis model

In the section, a novel dynamic activity analysis model (DAAM for short), not addressed so far, is proposed to forecast the effect of energy-saving and emission-abating regulations on economy in the long run, which is extended from the standard DDF-based AAM provided by Chambers et al. (1996) and Chung et al. (1997) and applied by Färe et al. (2001); Jeon and Sickles (2004) etc. How to simulate the potential output gain, output loss and the change of productivity, technique and efficiency by using the newly proposed DAAM approach is also introduced in the section. In the study, the decision-making units (DMU) are 38 two-digit industrial sectors ( $i = 1, 2, \dots, 38$ ). The forecasting time span is from 2011 to 2050 ( $t = 2011, \dots, 2050$ ). For each sector, there are three types of input ( $j = 1, 2, 3$ , corresponding to capital, labor and energy), one type of desirable output (gross industrial output value, GIOV), and one type of undesirable output (carbon dioxide emission, CO<sub>2</sub>). The historical dataset between 1980 and 2010 used for simulation is from Chen (2013). The panel data for 38 industrial sectors, rather than aggregate data, significantly enhances the information that could be obtained to analyze microeconomic performance, particularly when examining the efficiency of each unit.

For  $i$ th industrial sector, the column vectors of  $\mathbf{x}^i$ ,  $\mathbf{y}^i$  and  $\mathbf{b}^i$  represent the inputs, desirable output and undesirable output, respectively. Then the production technology for  $i$ th sector at time point  $t$  can be described by its output set:

$$P(\mathbf{x}^i) = \left\{ (\mathbf{y}^i, \mathbf{b}^i, -\mathbf{x}^i) : \mathbf{x}^i \text{ can produce } (\mathbf{y}^i, \mathbf{b}^i) \right\} \quad (1)$$

Same as Shephard distance function, DDF is also the representative function to describe such a production technology. The principle of DDF is illustrated in Fig. 1. The technology is represented by the output set  $P(\mathbf{x})$  to which the output vector of A point  $(\mathbf{y}, \mathbf{b})$  belongs. Shephard’s output distance function radially scales the original vector from point A proportionally to point D to describe the simultaneous increase of desirable and undesirable output. In contrast to this, the DDF starts at A and scales in the direction along ABC to capture the increase of desirable outputs (or goods) and decrease of undesirable outputs (or bads) simultaneously, which make it possible to



**Fig. 1** Principle of directional output distance function

investigate Porter hypothesis that allow for the possibility of crediting units for the reduction of pollutions. Formally, DDF is defined as

$$\bar{D}_o(x^i, y^i, b^i; g^i) = \sup \left\{ \beta : (y^i, b^i) + \beta g^i \in P(x^i) \right\} \tag{2}$$

where  $g$  is the direction vector in which outputs are scaled. In standard case,  $g = (y, -b)$ , as shown in Fig. 1.  $\beta$  is the maximum feasible expansion of the desirable outputs and contraction of the undesirable outputs when the expansion and contraction are identical proportions for a given level of inputs, which amounts to the value of DDF to be measured.

### 2.1 Production inefficiency and loss due to environmental regulation

As shown in Fig. 1, because the point A remains within the efficient production frontier, the inefficiencies resulted from such factors as wasteful energy consumption and serious pollution emissions give the producer the potential room to increase the output, given the inputs and current output, by saving energy and abating emission.<sup>2</sup> But whether the observation vector projects from the point A to point B or C depends on the weak or free disposal assumption of undesirable output. If assume that the undesirable output is strongly or freely disposal, that is, the disposability costs nothing, the producers will voluntarily get rid of the unwanted by-products, then the growth of potential output based on current desirable output will be maximized which amounts to the distance function value  $\beta_s$  (i.e., the ratio of AC/Og). In this case, energy and environment do not impose any restriction on output, then the production in point C is the most efficient. However, it's impossible to cost nothing to reduce undesirable output in reality. The producers therefore are not willing to reduce the undesirable

<sup>2</sup> In this case, the value of  $\beta$  is greater than zero which tell us the sizes of inefficiencies for the unit.

outputs because it makes use of the important inputs and then translates into the loss of desirable outputs given inputs. The reduction of undesirable outputs only can be achieved by environment regulations. Accordingly, the more appropriate assumption is weak disposability of undesirable output, the point A projecting into B on the frontier, which is the standard DDF, or referred to as environment regulatory AAM. Its value equals  $\beta_w$  or the ratio of AB/Og in the figure. In this case, the potential goods growth is a tradeoff between more goods and less bads, bound to below the maximized  $\beta_s$  under the strong disposability of bads.

The difference between  $\beta_w$  and  $\beta_s$  reflects the potential output loss caused by the observable lack of free disposability (more vividly, due to enforced environment regulations), i.e.,  $l = \beta_w - \beta_s < 0$  (Boyd et al. 2002). The value of  $l$  is analogous to the hyperbolic output loss measure introduced by Färe et al. (1989) and used by Boyd and McClelland (1999). The potential output loss  $l$  and potential output growth  $\beta_w$  reveal the extent of the win-win potential for each industrial sector, given current output at some time point. If potential  $\beta_w$  exceeds or equals the absolute value of  $l$ ,  $|l|$ , from the perspective of output, the win-win opportunity due to environment regulations that is described by Porter hypothesis happens, to some extent suggesting that improved production efficiency can make up for the losses imposed by regulations. If  $\beta_w < |l|$ , it indicates that environmental regulations will not lead to the win-win development. This paper will make use of this method to find the best energy-saving and emission-abating path that leads to the win-win development potentials in China.

### 2.2 Dynamic activity analysis model (DAAM)

As stated previously, the direction vector in DDF is  $\mathbf{g} = (\mathbf{y}, -\mathbf{b})$ , and the value of DDF,  $\beta$ , captures the maximum feasible proportion that the goods  $\mathbf{y}$  expand while the bads  $\mathbf{b}$  contract based on current output level  $(\mathbf{y}, \mathbf{b})$ , the negative sign of  $\mathbf{b}$  indicating the reduction of bads. To simulate the dynamic process of energy-saving and emission-reducing activity, in this paper, we introduce the time factor into the direction vector and redefine the output direction vector as  $\mathbf{g}^t = (\mathbf{y}^t, -\mathbf{b}^t) = [(1 + u)\mathbf{y}^{t-1}, -(1 + v)\mathbf{b}^{t-1}]$ , where  $u$  and  $v$  represent the changing rate of current industrial output (goods) and CO2 emissions (bads) relative to previous time point during the forecasting period, correspond to the different energy-saving and emission-abating paths to be designed in Sect. 3.1. Similarly, the dynamic changing path for the  $j$ th input vector is defined as  $\mathbf{x}_j^t = (1 + \sigma_j)\mathbf{x}_j^{t-1}$ , where  $\sigma_j$  is the changing rate of the  $j$ th input to be discussed also in Sect. 3.1. In terms of the defined dynamic direction vector, the technology in  $t$  period and observation also in  $t$  period, the linear programming of two types of DDF, the assumption of weak and strong disposability of undesirable output, is specified respectively for  $i$ th sector as below.

Directional distance function (weakly disposable bads)

$$\begin{aligned} \vec{D}_o^t(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t}) &= \underset{\lambda, \beta}{Max} \beta_w \\ s.t. \quad &\sum_{i=1}^{38} \lambda^i \mathbf{y}^{i,t} \geq (1 + \beta_w)(1 + u)\mathbf{y}^{i,t-1} \end{aligned}$$

$$\begin{aligned}
 \sum_{i=1}^{38} \lambda^i \mathbf{b}^{i,t} &= (1 - \beta_w) (1 + v) \mathbf{b}^{i,t-1} \\
 \sum_{i=1}^{38} \lambda^i \mathbf{x}_j^{i,t} &\leq (1 + \sigma_j) \mathbf{x}_j^{i,t-1} \quad (j = 1, 2, 3) \\
 \beta, \lambda^i &\geq 0 \quad (i = 1, 2, \dots, 38)
 \end{aligned}
 \tag{3}$$

In linear programming (3),  $\beta = 0$  means that the industrial sector lies on the possibility frontier and its production is efficient; while  $\beta > 0$  implies that the sector is inefficient in production. The proportion of the sectors with  $\beta > 0$  to all sectors shows us how widespread the inefficiencies are in the industry, which is related to the win-win opportunities by environmental regulation. The inequality for goods in (3) makes it freely disposable which means that the goods can be disposed of without the use of any inputs and then without the decrease of bads. The bads is modelled with equality that makes it weakly disposable. The inequality specification of inputs illustrates also that the inputs are strongly disposable; that is, the increase of inputs will not cause the decrease of output. The intensity variable  $\lambda^i$  is the weight assigned to each sector when constructing the production frontier. As shown in linear programming (3), novel definition of dynamic output and input direction vector not only introduces many possible energy-saving and emission-abating paths into DDF in order to capture the regulatory behavior but also makes it possible to forecast the dynamic impact of energy-saving and emission-abating activity on economy in the following decades. Therefore, we abuse terminology and refer to the extended DDF as dynamic (environmental regulatory) activity analysis model (DAAM), which distinguishes itself from the standard DDF and AAM in that it has introduced the time lag operator into the direction vector.

Directional distance function (strongly disposable bads)

$$\begin{aligned}
 \vec{D}_o^t \left( \mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t} \right) &= \underset{\lambda, \beta}{Max} \beta_s \\
 s.t. \quad \sum_{i=1}^{38} \lambda^i \mathbf{y}^{i,t} &\geq (1 + \beta_s) (1 + u) \mathbf{y}^{i,t-1} \\
 \sum_{i=1}^{38} \lambda^i \mathbf{b}^{i,t} &\geq (1 - \beta_s) (1 + v) \mathbf{b}^{i,t-1} \\
 \sum_{i=1}^{38} \lambda^i \mathbf{x}_j^{i,t} &\leq (1 + \sigma_j) \mathbf{x}_j^{i,t-1} \quad (j = 1, 2, 3) \\
 \beta, \lambda^i &\geq 0 \quad (i = 1, 2, \dots, 38)
 \end{aligned}
 \tag{4}$$

From the mathematical perspective, the equality constraint of undesirable output in linear programming (3) is changed into the same inequality constraint as on the desirable output to reveal the strong disposability of undesirable output in linear programming

(4). As mentioned above, the difference of solutions between (3) and (4) measures the potential production loss due to energy-saving and emission-reducing activity.

### 2.3 Malmquist–Luenberger Productivity Index (MLPI)

The DAAM approach summarized in linear programming (3) with the weak disposal assumption of undesirable output models the energy-saving and emission-abating activity; therefore, it can be used to measure the change of total factor productivity (TFP) and its decomposition under environmental regulations by calculating the Malmquist–Luenberger Productivity Index (MLPI). To the end, four different types of DDF must be solved for each sector: two use observations and technology at time period  $t$  and  $t + 1$ ,  $\bar{D}_o^t(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})$  and  $\bar{D}_o^{t+1}(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})$ ; and two use adjacent period, for example,  $\bar{D}_o^t(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})$  calculated from  $t$  period technology with the  $t + 1$  period observation, and  $\bar{D}_o^{t+1}(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})$  calculated from  $t + 1$  period technology with the  $t$  period observation. Then the Malmquist–Luenberger Productivity Index (MLPI) defined by Chung et al. (1997) can be computed using the following formulas:

$$MLPI^{t,t+1} = \left[ \frac{1 + \bar{D}_o^t(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})}{1 + \bar{D}_o^{t+1}(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})} \times \frac{1 + \bar{D}_o^{t+1}(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})}{1 + \bar{D}_o^t(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})} \right]^{1/2} \quad (5)$$

The Malmquist–Luenberger index is the most widely used productivity index and is particularly attractive when constructing it since it does not rely on prices, particularly the price of CO<sub>2</sub> appeared in this study. The MLPI can be decomposed as the product of two terms: the index of Malmquist–Luenberger technical change (MLTCH) and Malmquist–Luenberger efficiency change (MLECH); that is

$$MLPI^{t,t+1} = MLTCH^{t,t+1} \cdot MLECH^{t,t+1} \quad (6)$$

where,

$$MLTCH^{t,t+1} = \left( \frac{1 + \bar{D}_o^{t+1}(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})}{1 + \bar{D}_o^t(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})} \cdot \frac{1 + \bar{D}_o^{t+1}(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})}{1 + \bar{D}_o^t(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})} \right)^{1/2} \quad (7)$$

$$MLECH^{t,t+1} = \frac{1 + \bar{D}_o^t(\mathbf{x}^{i,t}, \mathbf{y}^{i,t}, \mathbf{b}^{i,t}; \mathbf{y}^{i,t}, -\mathbf{b}^{i,t})}{1 + \bar{D}_o^{t+1}(\mathbf{x}^{i,t+1}, \mathbf{y}^{i,t+1}, \mathbf{b}^{i,t+1}; \mathbf{y}^{i,t+1}, -\mathbf{b}^{i,t+1})} \quad (8)$$

If  $MLPI > 1$ , it means that TFP grows over the adjacent period; while  $MLPI < 1$  indicates that TFP declines.

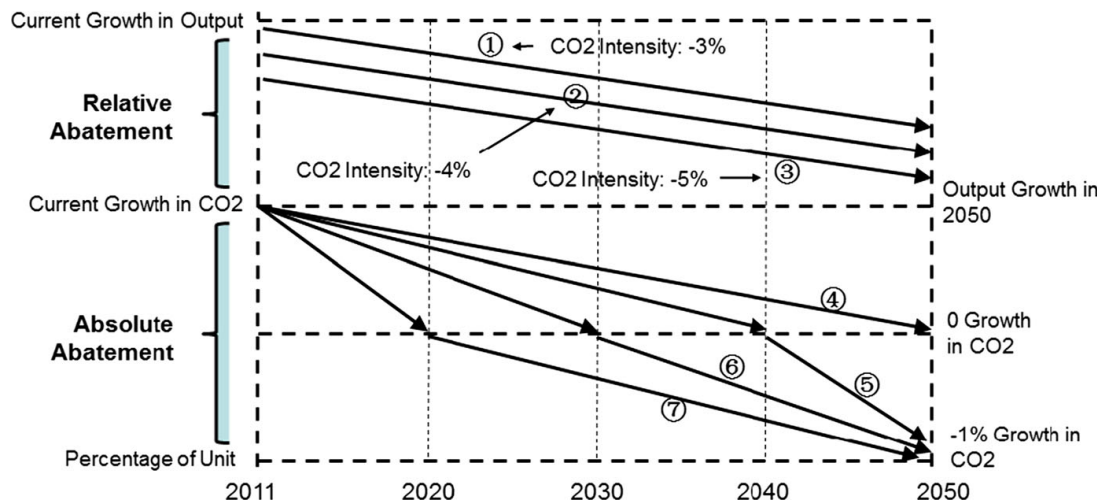
### 3 Forecasting scheme

#### 3.1 Design energy-saving and emission-abating paths

Different energy-saving and emission-abating paths will have obviously different impact on economy (Lee et al. 2007; Kuosmanen et al. 2009). This paper designs three energy-saving scenarios and seven emission-reducing scenarios, totally twenty one combinations of environment regulatory paths. By introducing different regulatory paths into the DAAM approach proposed in Sect. 2.2, this paper will forecast their effect on the potential growth of output and productivity in the following four decades so as to look for the best regulatory path leading to a win-win development possibility for Chinese industry.

Before the design of changing paths of energy and emission, we firstly specify the changing patterns for industrial output and other inputs such as capital stock and labor in the future. According to Chen and Golley (2014), between 1981 and 2010 the historical average growth rate of total industry is about 12.6, 9.3 and 2.6% for output, capital and labor, respectively. Their growth rates in 2010 are 20.4, 14.1 and 5.3%, higher than their historical average growth. However, after 30 years of rapid economic growth since the reform in 1978, China is facing a long-term decline in its economic growth rate, given its latest records. Thus, we assume that the growth rate of output, capital and labor will decrease from their respective growth rate in 2010 evenly to one third of their historical average growth in 2050 for each industrial sector and the aggregate industry. The design of energy saving scheme is based on the promissory targets to save energy stipulated by China government. Specifically, China central government promise to decrease the energy consumption per unit of output (i.e., energy intensity) by 20 and 16%, respectively, during the period of 11st and 12nd five-year-plan, translating into 4.36 and 3.43% annual reducing rate of energy intensity. In fact, during the 11st five-year-plan period, China decreased the energy intensity by 19%, 1% below the target rate. Therefore, this paper designs three scenarios for energy save; that is, the energy intensity will decrease by 3, 4, 5% per year in the following 40 years. According to the annual growth rate of industrial output specified already, this can be translated into three paths of energy save between 2011 and 2050.

This paper designs the emission abatement scheme according to two specification of relative and absolute abatement, the former of which caters to the state condition that China is a developing country whose major task is to develop. If output experiences a rapid growth, CO<sub>2</sub> emission may have a not low growth, too. As mentioned in introduction part, China officially announced that it will abate the CO<sub>2</sub> intensity by 40–45% in 2020 as opposed to the intensity in 2005. That means China should decrease the CO<sub>2</sub> intensity by 3.4–3.9% per year during that period. During the period of 12nd five-year-plan, China plans to reduce the CO<sub>2</sub> intensity by 17%, i.e. 3.66% per year. Based on this, we will design three scenarios for CO<sub>2</sub> relative abatement;



**Fig. 2** Carbon dioxide abatement paths (1–7) for Chinese industry (2011–2050)

that is, CO<sub>2</sub> intensity will decrease annually by 3, 4 and 5 %, respectively. As depicted in Fig. 2, given specified declining growth of industrial output, this will also lead to three decreasing emission paths (1–3 path). Most countries adopt the strategy of absolute abatement of CO<sub>2</sub> emissions. As illustrated in Fig. 2, this paper designs four scenarios for absolute abatement, which is attributable to the generalized understanding of emission abatement concept that emission reduction does not necessarily refer to the absolute decline in emission level; a declining emission growth rate or declining relative to BaU is also a type of emission abatement. Specifically, four scenarios include: (1) the growth rate of CO<sub>2</sub> for different sectors decreases from their respective growth in 2010 evenly to zero growth in 2050 (that is, the emission peak will appear in mid of this century); (2) the emission growth of all sectors reduces from the growth rate in 2010 to zero growth in 2039 and, after the emission peak, continuously decreases to –1 % growth rate in 2050; the 3rd and 4th path are similar to the 2nd path but the emission peak is moved on to the year of 2030 and 2020, respectively. The twenty one energy-saving and emission-abating paths, together with the varying paths of industrial output, capital and labor, will be introduced into the dynamic activity analytical model (DAAM) mainly through direction vector so as to forecast the effect of environment regulations on output and productivity in the following decades.

### 3.2 Bootstrapping potential output gain, productivity change and its components

Lovell (1993) have labeled the nonparametric DEA and its variants as the deterministic approaches, which seems to suggest that they do not have statistical underpinnings and are sensitive to the sampling variations. Since the pioneering work by Efron (1979) and its extensions in the frontier framework by Hall et al. (1995), the bootstrap methodology is often used to undertake the statistical inferences on distance function of DEA approach. The key to obtaining consistent bootstrap estimates of distance function lies in consistent replication of the data generating process. Simar and Wilson (1998) argued that resampling from the empirical distribution of the data (i.e., drawing with replacement from the set of original distance function estimates) will lead to



inconsistent bootstrap estimation. They proposed the smooth bootstrap to overcome this problem and yield consistent estimates of distance function. Simar and Wilson (1999) applied the principle to bootstrap the Malmquist productivity index. Following this, this paper extends the ideas to directional distance function and the Malmquist–Luenberger productivity index and its components.

Bootstrapping the distance function specified in the linear programming (3) is firstly exemplified. Its complete bootstrap algorithm could be summarized by the following steps:

- 1) By using the linear programming (3), compute  $\{\beta_w^k, k = 1, 2, \dots, n\}$ ;
- 2) Define the empirical distribution function for efficiency scores by putting mass  $\frac{1}{n}$  on  $\beta_w^i, i = 1, 2, \dots, n$ ;
- 3) By using the univariate kernel density estimator and the reflection method described in Simar and Wilson (1998), generate a random sample  $\{\beta_{w,b}^{i*}, i = 1, 2, \dots, n\}$ <sup>3</sup> from the empirical distribution function defined in 2);
- 4) Compute the pseudo-sample  $\{(\mathbf{x}^i, \mathbf{y}_b^{i*}, \mathbf{b}_b^{i*}), i = 1, 2, \dots, n\}$ , where  $\mathbf{y}_b^{i*} = \mathbf{y}^i (1 + \beta_w^i) / (1 + \beta_{w,b}^{i*})$  and  $\mathbf{b}_b^{i*} = \mathbf{b}^i (1 - \beta_w^i) / (1 - \beta_{w,b}^{i*})$ ;
- 5) By using the pseudo-sample produced in 4) and the linear programming (3), compute the bootstrap estimate of  $\beta_w^k : \{\beta_{w,b}^{k*}, k = 1, 2, \dots, n\}$
- 6) Repeat 3)-5) B times to obtain a set of estimates

$$\{\beta_{w,b}^{k*}, k = 1, 2, \dots, n, b = 1, \dots, B\}$$

In this study,  $n = 38, B = 2,000,$  and  $h = 0.02$ .<sup>4</sup>

Bootstrapping for the distance function  $\beta_f$  specified in the linear programming (4) in this study largely involves a straightforward translation of the notation in above steps. Once the bootstrap values have been computed, we can construct the confidence intervals of the distance function and its linear combinations at the desired level of significance.

The methodology for bootstrapping distance function in linear programming (3) and (4) could be easily adapted to the productivity index, except that the time-dependence structure of the panel data must be taken into account. According to formulas (5)–(8), we firstly obtain the point forecasts of Malmquist–

<sup>3</sup> The random sample is generated according to

$$\beta_{w,b}^{i*} = \begin{cases} \beta_{w,b}^{i,0*} + h\varepsilon_b^{i*} & \text{if } \beta_{w,b}^{i,0*} + h\varepsilon_b^{i*} \leq 1 \\ 2 - \beta_{w,b}^{i,0*} - h\varepsilon_b^{i*} & \text{otherwise} \end{cases}$$

where  $\{\beta_{w,b}^{i,0*}, i = 1, 2, \dots, n\}$  is a simple bootstrap sample from  $\{\beta_w^i, i = 1, 2, \dots, n\}$ , that is, obtained by drawing with replacement from  $\{\beta_w^i, i = 1, 2, \dots, n\}$ ,  $\varepsilon_b^{i*}$  is a random drawn from a standard normal, and  $h$  is the smoothing parameter of bandwidth.

<sup>4</sup> As Daraio and Simar (2007) denoted, B should be greater than 2,000. The choice of kernel bandwidth controls the smoothness of the probability density curve. Following Simar and Wilson (1998), we choose  $h = 0.02$  in this paper which provides a reasonably smooth estimate of the distribution function of efficiency scores.

Luenberger productivity index and its components in adjacent period of  $t_1$  and  $t_2$ ,  $\{(MLPI^{k,t_1,t_2}, MLTCH^{k,t_1,t_2}, MLECH^{k,t_1,t_2}), k = 1, 2, \dots, n\}$ . To bootstrap the productivity index and its components, we need the data in adjacent time periods to consider the possibility of temporal correlation. To preserve any temporal correlation present in the data, following [Simar and Wilson \(1999\)](#), we make use of bivariate kernel density estimator and reflection method to generate two joint random samples of  $\{\beta_{w,b}^{i,t_1,*}\}$  and  $\{\beta_{w,b}^{i,t_2,*}\}$   $i = 1, \dots, n$ , and then compute two adjacent pseudo-samples of  $\{\mathbf{x}^{i,t_1}, \mathbf{y}_b^{i,t_1,*}, \mathbf{b}_b^{i,t_1,*}\}$  and  $\{\mathbf{x}^{i,t_2}, \mathbf{y}_b^{i,t_2,*}, \mathbf{b}_b^{i,t_2,*}\}$   $i = 1, 2, \dots, n$ . Based on two pseudo-samples and formulas (5)–(8), we could compute one bootstrap estimate of Malmquist–Luenberger productivity index and its components of technical and efficiency change. This step will be repeated for  $B$  times to provide a set of estimate of  $\{(MLPI_b^{k,t_1,t_2,*}, MLTCH_b^{k,t_1,t_2,*}, MLECH_b^{k,t_1,t_2,*}), k = 1, 2, \dots, n, b=1, \dots, B\}$ . Likely, in this paper,  $n = 38$ ,  $B = 2,000$ , and the smoothing parameter for bivariate bivariate normal kernel  $h = (4/5n)^{1/6}$ .<sup>5</sup> The bootstrapping values of *MLPI*, *MLTCH* and *MLECH* could be used to test if there is a real change in productivity, technique and efficiency in the following 40 years from a statistical perspective.

## 4 Forecasting analysis

### 4.1 Simulate the win-win prospect under different environment regulatory paths

Table 1 reports the potential industrial output growth  $\beta_w$ , output loss  $l$  and corresponding net output gain averaged over the entire forecasting period under twenty one environmental regulatory paths combined by three energy-saving scenarios and seven emission-reducing scenarios.

As shown in Table 1, the former three emission abating paths are designed in terms of CO2 intensity reduction targets and classified into the relative abatement group and the latter four paths into the absolute abating group. On the whole, considering the fact of priority in development for China, seven emission abating paths specified in both groups are modest. On average, the abating path 1 in relative abatement group will not lead to the emission inflexion during entire forecasting period, while the emission peak appear in 2050 and 2048 for abating path 2 and 3, similar to the case in path 4 in absolute abating group, indicating that the emission abatement specified in relative abatement group is more modest than that in absolute group. Mostly, the distribution of the values of potential output growth, output loss and net gain display a quite regular varying pattern as shown in the table. For three energy saving paths, the potential output growth increases as the abating rate of emission intensity increases from 3 to 5 % (path 1–path 3); for first two energy saving paths, the potential output growth increases first and then turns to fall from abatement path 4 to path 7, and, corresponding to third energy saving path, the potential output growth always increases in the absolute abatement group. For three energy saving paths, the potential output loss exhibit a consistently deterioration

<sup>5</sup> [Silverman \(1978, 1986\)](#) and [Härdle \(1990\)](#) discuss considerations relevant to the choice of  $h$ . In the paper, we use [Silverman \(1986\)](#) suggestion for  $h$  setting since we are using a bivariate normal kernel.

**Table 1** Potential output gain-loss analysis corresponding to 21 energy-saving and emission-abating paths (%)

Energy saving and emission abating paths	Relative abatement			Absolute abatement			
	Path 1 Emission intensity, 3 %	Path 2 Emission intensity, 4 %	Path 3 Emission intensity, 5 %	Path 4 Emission peak in 2050	Path 5 Emission peak in 2040	Path 6 Emission peak in 2030	Path 7 Emission peak in 2020
<i>Energy intensity, 3 %</i>							
$\beta W$	17.43	18.27	18.98	17.73	18.40	17.74	16.60
$l = \beta W - \beta f$	-23.64	-23.74	-24.35	-24.76	-25.41	-26.04	-26.09
Net gain	-5.50	-5.47	-5.60	-5.80	-5.94	-6.16	-6.25
<i>Energy intensity, 4 %</i>							
$\beta W$	19.21	19.48	20.23	17.18	18.91	18.70	17.88
$l = \beta W - \beta f$	-26.50	-26.67	-26.89	-26.67	-27.20	-29.70	-30.29
Net gain	-6.19	-6.22	-6.23	-6.37	-6.41	-7.13	-7.35
<i>Energy intensity, 5 %</i>							
$\beta W$	17.62	18.99	20.19	17.60	17.81	18.57	18.73
$l = \beta W - \beta f$	-30.44	-32.43	-32.98	-33.43	-33.72	-34.71	-35.04
Net gain	-7.41	-7.88	-7.95	-8.26	-8.33	-8.56	-8.64

from the emission abating path 1 to path 7, except the value of  $-26.89$  and  $-26.67\%$  in path 3 and path 4 in the second energy path. Accordingly, the averaging net output gain also consistently increases from emission abating path 1 to path 7 no matter what kind of scenario for the energy save (with one exception of  $-5.50\%$  in first path of both energy save and emission abatement), implying that the optimal energy-saving and emission-abating path must be in the relative abating group. From the dimension of energy save, with the increasing of intensity of energy save the potential output growth does not exhibit a regular changing pattern but the potential output loss does experience the deteriorating process for all the seven emission abating paths, leading to a similarly deteriorating net output gain for all the abating scenarios. It is thus clear that appropriately decreasing the intensity of energy save will reduce the widespread extent of production inefficiencies, leading to the shrinking of improving space for potential output growth. Taken together, on average, the lowest potential net output gain is  $-5.47\%$ , appearing in the combination of first energy saving path and second emission abating path in relative abatement group. This is the optimal energy-saving and emission abating path we select for further investigation next; that is, according the scenarios simulation, the optimal environment regulatory path is to decrease the energy intensity by  $3\%$  per year and decrease the CO<sub>2</sub> emission intensity by  $4\%$  per year in the following 40 years. Since all the potential net gain shown in Table 1 are negative, it seems that all paths cannot lead to the win-win development suggested by Porter hypothesis, even though the best energy-saving and emission-abating path chosen above.

The findings in Table 1 are consistent with most other researches. [Schaltegger and Synnestvedt \(2002\)](#) argue that not merely the level of environmental performance, but mainly the kind of environmental management approach with which a certain level is achieved, influences the economic outcome, thus, the economic success resulted from the environmental protection finally depends on the chosen kind of regulatory approach rather the level. It's suggestion that research and business practice should focus more on the effect of different environmental management approaches on economic performance is consistent with the methodology used in our studies. [Roughgarden and Schneider \(1999\)](#) use a dynamic integrated climate-economy model to calculate an optimal rate of carbon tax and suggest that an efficient policy for slowing global warming would incorporate only a relatively modest amount of abatement of greenhouse gas emissions, via the mechanism of a small carbon tax. [Chen et al. \(2004\)](#) find that the earlier the emission reducing policy is implemented the greater the GDP loss will be. If the start of the emission reductions is the year of 2030, 2020 or 2010 instead of 2040, then the undiscounted total GDP losses in the whole planning horizon would be 0.58–0.74, 1.00–1.32, or 1.10–1.83 times higher. [Kuusmanen et al. \(2009\)](#) suggest that if one is only interested in greenhouse gases abatement at the lowest economic cost, then equal reduction of emissions over time is preferred. These researches all support the strategy of gradual and modest emission abatement. Similar to the idea of our paper that there is a close relationship between emission reduction and development, [Reddy and Assenza \(2009\)](#) also suggest that the integration of climate policies with those of development priorities that are vitally important for developing countries and stress the need for using sustainable development as a framework for climate change policies.

#### 4.2 The influence of environment regulation on future potential output

[Murty and Kumar \(2003\)](#) pointed out that the win-win opportunities under the environmental regulations could be found more in some industries and less in others, and the studies for specific industries could help us to identify the industries with no such opportunities so that the monitoring and enforcement could be directed to those industries in which incentives are absent. As a matter of fact, it is also the reason why we focus on the analysis of China's industrial sectors instead of merely the aggregated industry. Therefore, under the optimal path of energy save and emission reduction chosen in previous subsection, this subsection further simulates the potential output growth, output loss and net output gain for all sectors in the following 40 years. Table 2 illustrates the forecasting prospects for each sector in the first forecasting year 2011, the win-win turning year and the last forecasting year of 2050 with the bootstrapping confidence interval for the net output gain. Specifically, the second and third column contains the original estimate of  $\beta_w$  and output loss of  $l = \beta_w - \beta_f$  in 2011; the following three columns show the win-win turning year in which the potential output growth exceeds the output loss firstly in the forecasting period; the potential output gain, output loss, net gain and its confidence interval in 2050 are reported after the win-win information.

Table 1 has shown that the averaged net output gains brought by different regulatory paths are all negative, even though by the best energy-saving and emission-abating

**Table 2** Sectoral output gain and loss in 2011, win-win turning year and 2050

Sectors	Forecasting period			Win-win turning point			Last year(2050)			Confidence Interval	
	First year (2011)		Year	Win-win turning point		$l = \beta w - \beta f$	Last year(2050)		Net Gain	Lower limit	Upper limit
	$\beta w$	$l = \beta w - \beta f$		$\beta w$	$l = \beta w - \beta f$		$\beta w$	$l = \beta w - \beta f$			
Coal	19.87	-69.29	2025	19.50	-18.11	19.02	-0.57	18.45	16.48	20.41	
Petroleum Ext.	9.96	-595.28	<i>not exist</i>			8.62	-30.56	-21.94	-23.94	-19.93	
Ferrous Mi.	9.90	-590.76	<i>not exist</i>			9.27	-71.50	-62.23	-64.19	-60.27	
Non-ferrous Mi.	33.89	-79.31	2038	9.66	-8.88	9.45	-7.50	1.96	0.01	3.90	
Nonmetal Mi.	44.82	-198.94	2040	18.04	-16.95	14.79	-1.58	13.21	11.26	15.16	
Wood Exp	69.98	-476.20	2045	22.92	-22.56	22.91	-7.81	15.10	13.16	17.03	
Food Prod.	99.87	-179.99	2039	53.61	-50.75	53.41	-27.36	26.05	24.08	28.02	
Food Ma.	129.92	-224.26	2030	35.88	-33.80	15.71	-5.03	10.68	8.70	12.66	
Beverage	79.92	-166.50	2017	75.87	-70.61	17.65	-3.97	13.68	11.68	15.68	
Tobacco	9.67	-68.80	2028	9.12	-7.95	2.15	0.00	2.14	0.19	4.10	
Textile	109.93	-244.45	2031	47.64	-45.37	28.56	-18.36	10.20	8.24	12.15	
Apparel	26.70	-80.93	<i>not exist</i>			18.59	-25.22	-6.63	-8.55	-4.72	
Leather	69.62	-91.67	<i>not exist</i>			28.27	-30.15	-1.88	-3.84	0.07	
Wood Prod.	39.85	-123.07	2048	29.09	-29.09	29.12	-28.71	0.41	-1.56	2.37	
Furniture	9.18	-59.79	2035	7.07	-7.02	5.53	-3.51	2.01	0.02	4.01	
Paper	39.96	-112.10	2035	29.58	-27.57	28.89	-1.74	27.15	25.25	29.05	
Printing	68.89	-146.57	<i>not exist</i>			32.55	-41.85	-9.30	-11.20	-7.40	
Cultural articles	39.01	-76.96	2043	36.78	-35.52	35.94	-25.96	9.97	7.98	11.96	
Petroleum Prod.	140.00	-303.60	2016	100.00	-98.27	50.97	-0.43	50.53	48.58	52.49	
Chemical products	73.60	-99.15	2019	58.93	-55.06	37.07	-11.69	25.38	23.39	27.36	
Medicine	99.24	-49.84	From2011			20.84	-9.30	11.54	9.56	13.53	

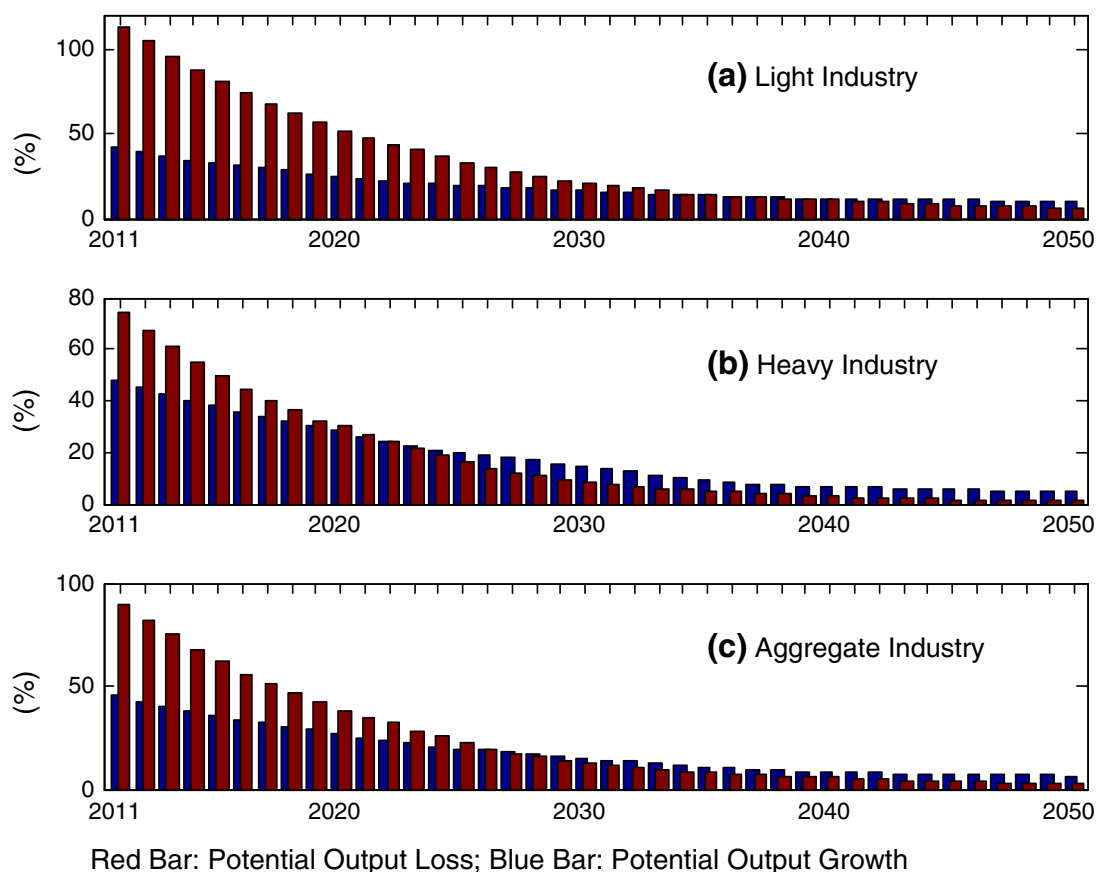
Table 2 continued

Forecasting period	First year (2011)		Win-win turning point		Last year(2050)		Confidence Interval			
	$\beta w$	$l = \beta w - \beta f$	Year	$\beta w$	$l = \beta w - \beta f$	$\beta w$	$l = \beta w - \beta f$	Net Gain		
								Lower limit	Upper limit	
Fibers	149.84	-212.87	2037	66.69	-66.60	66.18	-43.15	23.04	21.07	25.00
Rubber	141.84	-232.07	2031	98.44	-97.87	65.68	-39.09	26.59	24.68	28.51
Plastic	39.07	-139.72	<i>not exist</i>			26.36	-31.50	-5.13	-7.13	-3.13
Nonmetal Ma.	29.34	-73.62	2019	28.58	-27.36	26.97	-0.31	26.66	24.69	28.62
Ferrous press	120.11	-131.42	2034	109.09	-106.67	102.78	-4.83	97.95	96.03	99.86
Non-ferrous press	139.71	-176.34	2030	128.71	-128.55	117.44	-31.85	85.59	83.62	87.56
Metal products	96.89	-164.72	2043	32.58	-32.43	32.56	-16.18	16.38	14.42	18.35
General machinery	9.57	-105.30	2041	1.10	-0.90	1.04	-0.01	1.04	-0.92	2.99
Special machinery	9.81	-94.50	2029	9.23	-6.61	1.40	-0.02	1.38	-0.58	3.34
Transport equipment	9.63	-63.28	2025	8.57	-7.85	0.33	-0.04	0.29	-1.72	2.29
Electrical equipment	29.51	-40.09	2028	8.61	-7.96	6.12	-0.33	5.80	3.81	7.78
Computer	16.62	-11.58	From2011			3.30	-0.01	3.30	1.34	5.25
Measuring instrument	8.24	-36.78	2035	2.06	-1.49	1.74	-0.12	1.63	-0.30	3.55
Electric power	79.92	-141.96	2018	56.87	-55.86	9.99	-0.98	9.00	7.04	10.96
Gas Prod.	19.99	-178.93	2037	9.99	-7.43	9.83	-1.75	8.07	6.10	10.04
Water Prod.	49.87	-70.49	2029	49.52	-49.37	45.36	-26.05	19.31	17.42	21.19
Others	49.87	-142.61	2037	8.57	-8.45	5.36	-4.18	1.18	-0.78	3.13

path. However, if we look at the simulation results for 38 industrial sectors reported in Table 2 rather the aggregated industry only, the situation will be totally another story. Overall, the potential output loss exhibits an obviously declining trend for all sectors and the potential output growth of most sectors has a modest decline or does not change much. Except for six sectors such as extraction of petroleum and natural gas, mining and processing of ferrous metal ores, apparel manufacturing, leather manufacturing, printing, and plastic manufacturing, the potential output loss for all the other sectors appears to be smaller than potential output growth at some time point before 2050. Table 2 has listed the respective turning year for the remaining sectors in which the potential output growth exceeds the output loss firstly in the entire forecasting period. Note that two sectors such as medicine manufacturing and manufacture of computers, communication equipment and other electronic equipment have higher output growth than output loss even from the first forecasting year of 2011. This indicates that for most sectors, the energy-saving and emission-abating activity can bring the win-win development opportunity in the forecasting period. Even to the above exceptional six sectors, their potential output losses tend to decline, too, and are bound to be lower than the potential growth at certain time after the year of 2050, leading to an expected win-win development.

The reason why the averaged net gain for all paths, even the optimal path, is negative in Table 1 is that most sectors have large potential loss in the nearer future, as shown in Table 2. It is thus clear that the aggregation analysis is undependable and even leads you to the opposite conclusion. Particularly, the potential output loss of those energy and emission intensive sectors such as extraction of petroleum and natural gas, mining and processing of ferrous metal ores, exploiting of wood and bamboo, processing of petroleum and coking are extremely large, which should be one of the causes of the negative weighted potential net gain for aggregated industry. Moreover, what we care about the energy save and emission reduction is its final influential level instead of accumulative effect; hence, the high potential output loss in the nearer future is just meaningful for that period and useless for the analysis on the future opportunity of win-win development. The last three columns in Table 2 report the net output gain  $\beta_w - l = 2\beta_w - \beta_f$  in the last forecasting year of 2050 and its confidence interval at 5% significance level, estimated according to two independently bootstrap estimate of both  $\beta_w$  and  $\beta_f$ . This allows us to appreciate the sensitivity of the simulated win-win development possibility with respect to the sampling variations. Specifically, for the net output gain, we say it is significantly greater than zero (which would indicate the win-win development) if the confidence interval does not include zero and values below zero. As reported in Table 2, except six sectors denoted above that do not approach the turning point in the forecasting period and six sectors with confidence interval including negative values and zero,<sup>6</sup> the remaining twenty six industrial sectors, 68.4% of all sectors, enjoy a significant potential net output gain, a certain win-win development prospect without sample noise, in the last forecasting year

<sup>6</sup> They are wood processing, general machinery manufacturing, special machinery manufacturing, transport equipment manufacturing, manufacturing of measuring instruments and machinery, and others.



**Fig. 3** Averaged win-win development forecasting under the best energy-saving and emission-abating path for light, heavy industry and the industry as a whole (2011–2050)

of 2050. All in all, the sectoral simulation results shown in Table 2 manifests that, from the perspective of potential output, environment regulations can bring costs on output which means that Porter hypothesis will not be satisfied in the very nearer future, but when the time moves on, it will lead to the win-win development prospect for most industrial sectors, finally supporting the Porter hypothesis.

According to the theory in [Chenery et al. \(1986\)](#) and current empirical work in [Chen et al. \(2011\)](#), the standard perception of industrialization is a general shift in relative importance from light to heavy industry. Light industry is of great importance normally at the early stage of industrialization and labor-intensive in nature with relatively low ratios of capital to labor; while heavy industry is at the middle or late stage and capital-intensive with relatively high ratios of capital to labor. Therefore, we divide all industrial sectors into light and heavy industrial groups according to the ranking of capital to labor ratio ( $K/L$ ) in 2008. That is, the light industrial group corresponds to the top half of sectors with the lower  $K/L$  ratio, and the heavy industry to the last half of sectors with the larger  $K/L$  ratio. We refer to them as light industry and heavy industry in brief from now on in this paper. This is because 38 sectoral patterns of potential output growth and loss are too complicated to see clearly all at once, and sometimes we want to observe the difference just between the light and heavy industry instead. Figure 3 depicts the weighted average potential output loss (red bar) and output growth (blue bar) for light and heavy industry and aggregated



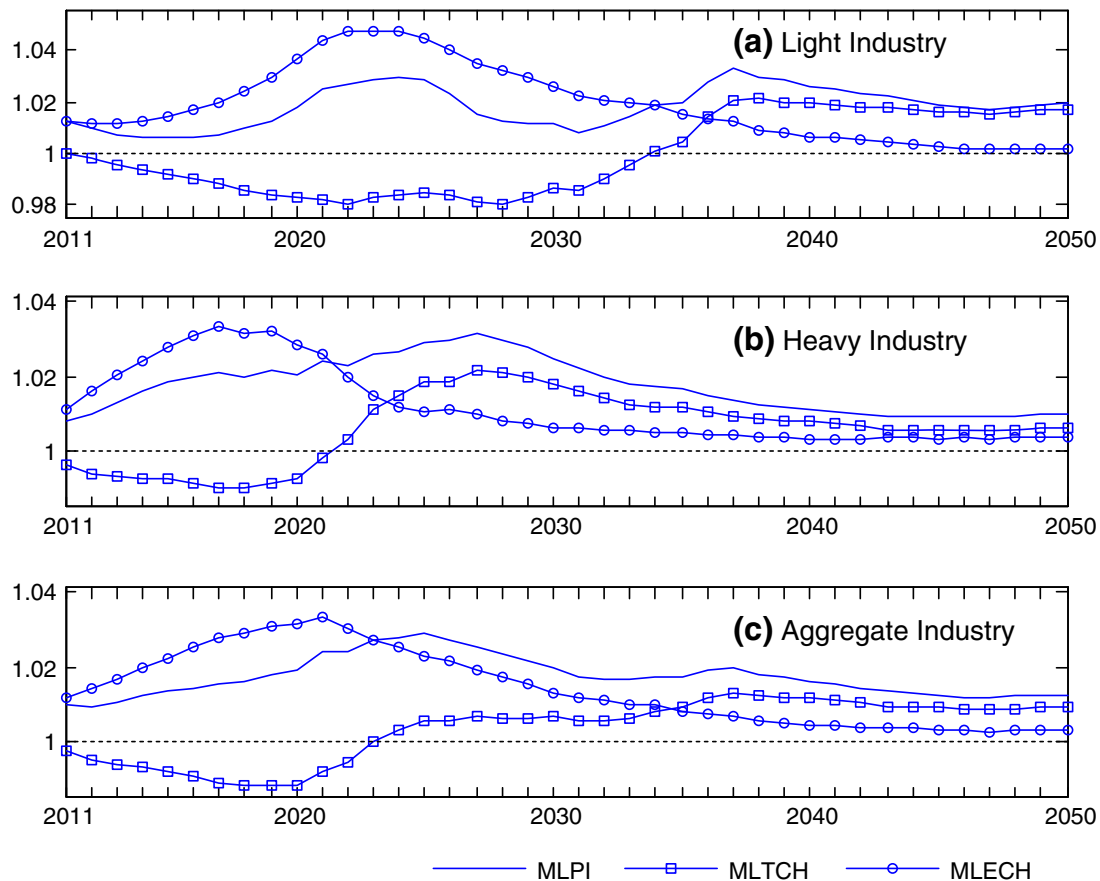
industry, under the best environmental regulatory path, in which the sectoral weight is its respective share of gross industrial output value.

Seen from Fig. 3, in light industry, the averaged potential output loss declines prominently from  $-112.76\%$  in 2011 to  $-6.35\%$  in 2050 while the potential output growth decreases less evidently from  $41.78\%$  in 2011 to  $10.26\%$  in 2050; in heavy industry, the corresponding varying range is  $(-74.03\%, -0.97\%)$  for averaged potential output loss and  $(48.16\%, 4.88\%)$  for output growth. Basically, the potential output loss in light industry is higher than that in heavy industry over the entire forecasting time span while the output growth in light industry is lower than that in heavy industry before 2025 and exchanges the position since then. The light industry does not reach a comparable level for potential output loss and output growth until 2035 and keeps the similar situation to the beginning of 2040s, just right meeting the win-win development condition. But for the heavy industry, the win-win situation is reached even since the earlier year of 2023 and the potential output growth holds a relatively large advantage over the output loss since that year. Therefore, heavy industry is obviously the beneficiaries of energy save and emission reduction, but light industry is also not the losers. For the aggregated industry, the potential output loss declines from  $-89.69\%$  in 2011 to  $-2.62\%$  in 2050, the potential output growth decreases from  $45.58\%$  in 2011 to  $6.53\%$  at the end of the forecasting period—being between that of light and heavy industry. Since heavy industrial sectors have larger weights, the varying pattern of the potential output in aggregated industry looks more similar to that in heavy industry—realizing the win-win development in the year of 2027.

#### 4.3 The influence of environment regulation on future industrial productivity

Sickles and Streitwieser (1998) have once investigated the impact of regulatory environment such as partial and gradual decontrol of natural gas prices on output change, technology and productivity in the interstate natural gas pipeline industry. Following this, this subsection also addresses the impact of optimal energy-saving and emission-abating activity on the foreseeable change of productivity, technique and efficiency in Chinese industry. Adopting the same group classification and weights as in Figs. 3, 4 exhibits the averaged changing trends of total factor productivity (TFP, i.e. MLPI) and its decompositions of MLTCH and MLECH under the optimal path of environment regulation for light, heavy and aggregated industry. Three subfigures show a similar pattern. That is, China's industrial TFP is firstly influenced by the efficiency change in which the catching-up effect of adoption of the frontier technologies due to the environment regulation is very obvious. When the efficiency attaches its utmost limits and the catching-up energy is almost released, the technical progress begins to serve as the major propelling force of industrial TFP through gradual accumulation and assimilation. The improvement of overall TFP index also reveals that the industrial development has generally shifted in a win-win fashion.

More specifically, at the early stage, energy-saving and emission-abating policy mainly negatively affects the industrial technical progress, and a little more on light industry than on heavy industry. For instance, for light industry, the level of techni-



**Fig. 4** Averaged productivity forecasting and its decomposition under the optimal energy-saving and emission-abating path (2011–2050)

cal progress in 2022 and 2028 is similarly 98.03 % of previous year, attaching the largest backward magnitude of production frontier,  $-1.97\%$ , over the whole forecasting period; the largest backward extent of technical progress for heavy industry is  $-0.98\%$  in 2018 and the largest one for the aggregated industry is  $-1.16\%$  in 2019. However, due to the obvious catching-up effect and the improvement of production efficiency (at the efficiency peak, the value of MLECH is 1.048 in 2022 for light industry, 1.033 in 2017 for heavy industry, and 1.033 in 2021 for aggregated industry), the TFP growth will keep an increasing trend at the earlier forecasting phase. The negative effect of environmental regulation on technical progress fades gradually and turns to be positive in the year of 2034, 2022 and 2023 for light, heavy and aggregated industry, respectively; at the same time, the catching-up effect turns to decline and begins to be lower than the effect of technical progress in 2036, 2024 and 2035 for light, heavy and aggregated industry. The technical progress then gradually reaches its peak due to the long-term introduction, absorption, adoption and innovation of the advanced technologies—say, the technical progress attains the highest value of 1.021 in 2038, 1.021 in 2027, and 1.013 in 2037 for light, heavy and aggregated industry. After that, the efficiency continues to decrease while the industrial technique and its dominated productivity keep a steady growth till the end of forecasting period.

In a word, the optimal energy-saving and emission-abating activity plays a positive role in improving industrial productivity, though different role in technical progress and efficiency change in different period. For example, the TFP of light industry grows steadily to the first peak in 2024 (1.030) and attains the second peak in 2037 (1.033); while the TFP of heavy and aggregated industry increases first and turns to decline after reaching its peak in 2027 (1.032) and 2025 (1.029); in 2050 the TFP growth is 1.94, 0.97 and 1.27 % for light, heavy and aggregated industry, respectively. During the entire forecasting period from 2011 to 2050, on average the TFP growth of light, heavy and aggregated industry attains 1.80, 1.73 and 1.74 %, and the aggregated industrial technical progress and efficiency change reach to 0.34 and 1.42 %. This is a win-win development prospect since the productivity, technique and efficiency are growing and the targets to save energy and reduce emission are also achieved. As [Chen and Golley \(2014\)](#) denoted, the traditionally estimated TFP that does not take the energy and environment into account often overestimates the real TFP. In this paper, we also choose another model, named basic DEA approach to forecast the change of productivity, technique and efficiency in the same forecasting period, in which the CO<sub>2</sub> emission will not be considered. The averaged change of productivity, technique and efficiency estimated by DEA approach over the entire forecasting time span is 2.34, 0.66 and 1.66 %, higher than their counterparts estimated by DAAM approach. To check the difference between basic DEA and DAAM measurements, we run the non-parametric Kolmogorov-Smirnov Z test in which the null hypothesis is that the DEA estimates are the same as the DAAM estimates. The test rejects the null hypothesis at the 0.000, 0.0108 and 0.005 significance level for series of productivity, technique and efficiency, respectively.

To investigate the heterogeneity and sensitivity of the estimates, we applied the bootstrap methods specified in Sect. 3.2 to test for significant differences from unity of sectoral Malmquist–Luenberger productivity index and its decomposition of technique and efficiency index, referring to Tables 3, 4 and 5, in which values greater than unity denote progress while values less than unity denote regress. Five adjacent time periods are exemplified. In 2011/2010, the original estimates tell us that the numbers of sectors that progress in productivity, technique and efficiency are 29, 24 and 33; while the bootstrapping test reveals that among them only 16, 5 and 15 sectors have a significant progress. Nine sectors regress in productivity in which only two of nonmetal products manufacturing and ferrous metals pressing are significant; 14 sectors decrease in technique and only 3 are significant (nonmetal ores mining, nonmetal products manufacturing, ferrous metals pressing); five sectors decrease in efficiency while only the sector of ferrous metals pressing is significant. In 2020/2019, twenty two sectors regress in technique and eighteen of them are significant, while sixteen sectors seem to progress in which there are only four to be significant, indicating a negative influence resulted from environment regulations. For change in efficiency, the original estimates tell us that thirty four sectors progress and the bootstrapping test denotes twenty seven of them are significant; four sectors that regress are all insignificant. Driven more by the efficiency, the performance of productivity looks not bad, in which twenty three sectors progress and only five are insignificant; fifteen sectors regress but only three are significant (textile manufacturing, leather manufacturing, and printing).

**Table 3** Sectoral changes in productivity in selected years

Sectors	2011/2010	2020/2019	2030/2029	2040/2039	2050/2049
Coal	1.0051	0.9960	0.9971	1.0304*	1.0806**
Petroleum Ext.	1.0034*	1.0023*	0.9893	0.8551**	1.0000
Ferrous Mi.	0.9985	0.9930	1.0179*	0.9522*	0.9484**
Non-ferrous Mi.	1.0034	1.0093*	0.9982	0.9941	1.0053*
Nonmetal Mi.	0.9936	0.9866	0.9867	1.0002*	1.0017**
Wood Exp.	0.9920	0.9857	0.9865	0.9980	0.9990
Food Prod.	1.0031	1.0049	1.0031	1.0157**	1.0076*
Food Ma.	1.0021	1.0062	0.9989	1.0070*	0.9989
Beverage	1.0096	1.0097	1.0027	1.0038*	1.0024
Tobacco	1.0584**	1.0433**	1.0692***	1.0583**	1.0639***
Textile	1.0069*	0.9927*	0.9792*	0.9902	1.0076**
Apparel	1.0100*	0.9956	0.9886*	0.9887	0.9952
Leather	1.0037*	0.9933*	0.9923	1.0036*	1.0030
Wood Prod.	1.0012	0.9904	0.9859**	0.9907	0.9721
Furniture	1.0047	0.9953	0.9943*	0.9892	0.9723
Paper	1.0027	1.0063	0.9995	1.0086*	0.9909
Printing	1.0033	0.9935*	0.9907	1.0049**	1.0055*
Cultural articles	1.0069*	0.9971	0.9937*	1.0078**	1.0107**
Petroleum Prod.	0.9939	1.0049**	1.0038**	0.9973	1.0057*
Chemical products	1.0142**	1.0550***	1.0518***	1.0066*	1.0368**
Medicine	1.0025*	1.0059*	1.0030*	1.0066*	1.0075**
Fibers	1.2570***	1.0097**	1.0025*	1.0025*	1.0011*
Rubber	0.9999	0.9897	0.9866	1.0003*	1.0010
Plastic	1.0020	0.9916	0.9949*	1.0031*	0.9934
Nonmetal Ma.	0.9654**	1.0117**	0.9944	1.0355***	1.0348**
Ferrous press	0.8613***	1.0016*	1.1040**	1.0001	0.9799
Non-ferrous press	1.0030	1.0074*	1.0071*	0.9535*	0.9310*
Metal products	1.0037	0.9927	0.9909	1.0270**	1.0618***
General machinery	1.0081	1.0264**	1.0314**	1.0676**	1.0501**
Special machinery	1.0082*	1.0074*	1.0385**	1.0270**	1.0135*
Transport equipment	1.0358***	1.0308***	1.0314***	1.0202*	1.0133*
Electrical equipment	1.0508***	1.0532***	1.0098*	1.0092*	1.0029**
Computer	1.0095*	1.0157**	1.0142**	1.0017	1.0023**
Measuring instrument	1.0145**	1.0057*	1.0063*	1.0166**	1.0076*
Electric power	0.9983	0.9953	1.1336***	1.0079*	1.0866**
Gas Prod.	1.0030*	1.0056*	1.0026*	1.0038	1.0023*
Water Prod.	0.9994	1.0031	1.0609**	1.0775**	1.3265**
Others	1.0439**	1.0482**	1.0216**	1.0219**	1.0236*

Single, double and triple asterisks (\*, \*\*, \*\*\*) indicate significant differences from unity at 0.10, 0.05 and 0.01 level, respectively

**Table 4** Sectoral changes in technique in selected years

Sectors	2011/2010	2020/2019	2030/2029	2040/2039	2050/2049
Coal	1.0049	1.0053	0.9958	1.0278*	1.0555**
Petroleum Ext.	1.0024	1.0044	0.9925	0.8612**	0.9656*
Ferrous Mi.	0.9978	0.9924	1.0005*	0.9791*	0.9482**
Non-ferrous Mi.	1.0023	1.0082	0.9993	0.9953	1.0003
Nonmetal Mi.	0.9934*	0.9864	0.9861	1.0003	1.0000
Wood Exp.	0.9918	0.9856*	0.9863	1.0002	1.0002*
Food Prod.	1.0027	1.0055	1.0006	1.0118**	1.0181**
Food Ma.	1.0016	1.0052*	0.9969*	1.0024*	0.9988
Beverage	1.0028	1.0054	1.0003	1.0007	0.9996
Tobacco	1.0029	1.0096	1.0776**	1.0526**	1.0544**
Textile	1.0059*	0.9917*	0.9844**	1.0004	1.0053*
Apparel	1.0072*	0.9928*	0.9858**	0.9882	0.9950
Leather	1.0001	0.9894**	0.9896*	1.0032*	1.0028*
Wood Prod.	1.0000	0.9893**	0.9856**	0.9880	0.9770
Furniture	0.9975	0.9883**	0.9864**	0.9835	0.9699*
Paper	1.0032	1.0055	1.0009	1.0094*	1.0134**
Printing	0.9983	0.9886*	0.9858*	1.0049*	1.0058*
Cultural articles	0.9986	0.9888*	0.9859**	1.0088**	1.0090*
Petroleum Prod.	0.9998	1.0035	0.9991	0.9606*	0.9531*
Chemical products	0.9994	1.0477**	1.0300**	1.0134**	1.0318**
Medicine	1.0012	1.0046**	1.0011	1.0037*	1.0048*
Fibers	1.0642**	1.0068	1.0000	0.9997	0.9997
Rubber	0.9991	0.9890*	0.9858	1.0004	1.0008
Plastic	1.0000	0.9897	0.9888	1.0041*	1.0058*
Nonmetal Ma.	0.9783**	0.9753**	0.9791**	1.0332**	1.0367**
Ferrous press	0.9147**	0.9583**	1.0623**	1.0010	0.9629*
Non-ferrous press	1.0029	1.0059	0.9846*	0.9675*	1.0226**
Metal products	1.0008	0.9898*	0.9879*	1.0321**	1.0605***
General machinery	0.9998	0.9852**	0.9877**	1.0396**	1.0400**
Special machinery	1.0071	0.9506***	0.9833**	1.0300**	1.0121**
Transport equipment	1.0013	0.9547**	1.0262**	1.0143**	1.0075*
Electrical equipment	1.0014	0.9960*	0.9875**	1.0053*	1.0022*
Computer	0.9995	1.0056	1.0096*	1.0007	0.9997
Measuring instrument	1.0129**	0.9492**	0.9865*	1.0111**	1.0065*
Electric power	1.0037	1.0015	1.0872***	1.0127**	1.0766***
Gas Prod.	1.0030*	1.0056**	1.0006*	0.9998	0.9996
Water Prod.	0.9988	0.9973	1.0497**	1.0744**	1.1519**
Others	1.0001	0.9693*	0.9902	1.0171*	1.0236*

Single, double and triple asterisks (\*, \*\*, \*\*\*) indicate significant differences from unity at 0.10, 0.05 and 0.01 level, respectively

**Table 5** Sectoral changes in efficiency in selected years

Sectors	2011/2010	2020/2019	2030/2029	2040/2039	2050/2049
Coal	1.0002	0.9908	1.0013	1.0026**	1.0238**
Petroleum Ext.	1.0010	0.9980	0.9968	0.9929	1.0357**
Ferrous Mi.	1.0008	1.0006	1.0173	0.9725	1.0002
Non-ferrous Mi.	1.0011*	1.0011*	0.9989	0.9988	1.0050*
Nonmetal Mi.	1.0002	1.0002	1.0006	0.9999	1.0017
Wood Exp.	1.0002	1.0002*	1.0002	0.9978	0.9987
Food Prod.	1.0004	0.9995	1.0025*	1.0039	0.9897
Food Ma.	1.0004	1.0010*	1.0020*	1.0045*	1.0001
Beverage	1.0067*	1.0042*	1.0024*	1.0031*	1.0027
Tobacco	1.0554**	1.0334**	0.9922	1.0054*	1.0090
Textile	1.0010	1.0010	0.9947	0.9898	1.0023
Apparel	1.0028	1.0028*	1.0028*	1.0005	1.0002
Leather	1.0036**	1.0039**	1.0027**	1.0004	1.0002
Wood Prod.	1.0012	1.0012	1.0004	1.0027	0.9949
Furniture	1.0071**	1.0071**	1.0081**	1.0058*	1.0025
Paper	0.9995	1.0008*	0.9986	0.9992	0.9777
Printing	1.0050*	1.0050**	1.0050**	1.0000	0.9997
Cultural articles	1.0083**	1.0085**	1.0079**	0.9990	1.0017
Petroleum Prod.	0.9941	1.0013*	1.0047	1.0381**	1.0553**
Chemical products	1.0147**	1.0070*	1.0212**	0.9933	1.0048*
Medicine	1.0013	1.0013	1.0020	1.0029	1.0028
Fibers	1.1812**	1.0029*	1.0025	1.0028	1.0013
Rubber	1.0008*	1.0008*	1.0008	0.9999	1.0002
Plastic	1.0020*	1.0020*	1.0061	0.9990	0.9877
Nonmetal Ma.	0.9868	1.0373**	1.0157**	1.0022	0.9982
Ferrous press	0.9416**	1.0452***	1.0393**	0.9991	1.0177**
Non-ferrous press	1.0000	1.0015*	1.0229**	0.9855	0.9104**
Metal products	1.0029	1.0029	1.0031	0.9950	1.0012
General machinery	1.0083	1.0418**	1.0442**	1.0269**	1.0097
Special machinery	1.0011	1.0597***	1.0562***	0.9971	1.0015
Transport equipment	1.0344**	1.0797***	1.0051	1.0059*	1.0057
Electrical equipment	1.0493**	1.0574**	1.0226**	1.0039*	1.0007
Computer	1.0101*	1.0100*	1.0045	1.0010	1.0026
Measuring instrument	1.0016	1.0595**	1.0201**	1.0055	1.0011
Electric power	0.9947	0.9938	1.0427*	0.9953	1.0093*
Gas Prod.	1.0000	1.0000	1.0020	1.0041	1.0026
Water Prod.	1.0006	1.0058**	1.0106	1.0030*	1.1516*
Others	1.0438**	1.0814*	1.0318	1.0047	1.0000

Single, double and triple asterisks (\*, \*\*, \*\*\*) indicate significant differences from unity at 0.10, 0.05 and 0.01 level, respectively

In 2030/2029, twenty four sectors regress in technique and half of them are significant in which ten sectors belong to light industry except for nonmetal products manufacturing and non-ferrous metals pressing; fourteen sectors progress in technique and nine of them are significant (six sectors belong to heavy industry such as ferrous ores mining, chemical products manufacturing, ferrous metals pressing, electric power producing, gas producing, and water producing). Efficiency performs not bad; say, thirty three sectors increase in efficiency and seventeen of them are significant; while five sectors that regress are all insignificant. Thus, twenty sectors progress in productivity in which only two of them are insignificant, and eighteen sectors decrease in productivity with six being significant—such as textile manufacturing, apparel manufacturing, wood processing, furniture manufacturing, cultural articles manufacturing and plastic manufacturing, most of them belonging to light industry. In 2040/2039, twenty eight sectors progress in productivity, twenty six of which being significant; ten sector decrease in productivity with only three heavy industrial sectors being significant (i.e., petroleum extraction, ferrous ores mining, non-ferrous metals). For technique change, twenty eight sectors also progress in which only seven sectors are insignificant, ten sectors regress with only four sectors being significant. There are only ten sectors that have significant change in efficiency, and all the sectors that regress in efficiency are not significant. In 2050/2049, the sectors with significant efficiency change are very rare. Specifically, only nine sectors are significant in the change of efficiency, one of which regresses. There are twenty seven sectors that progress in technique and only three of which are insignificant; eleven sectors decrease in technique with only five sectors are significant. As for productivity change, twenty eight progress with only four sectors being insignificant; ten sectors regress in which only two of ferrous ores mining and non-ferrous metals pressing are significant. Obviously, the bootstrapping estimates reveal more accurate forecasting of sectoral change of productivity, technique and efficiency in the following 40 years than original point prediction.

## 5 Conclusion

To challenge the climate change and boost the transformation of development model, developing the low carbon economy under the appropriate environment regulations have become the necessary approach for most countries to achieve the sustainable economic development (Chen 2011). However, both energy save and environment protection will seize the important materials originally planned to normal production, causing the declination of the desirable output and competitiveness. The conflicting views are also reflected in academic area, i.e., if in favor or against the Porter hypothesis. This paper makes use of the directional distance function that precisely embodies the spirit of Porter hypothesis in which the goods increase and bads decrease simultaneously and proposes a novel dynamic activity analysis model (DAAM) to forecast the win-win development possibilities for Chinese industrial sectors between 2011 and 2050, to investigate the existence of Porter hypothesis in China. To overcome the sample variation, the consistent bootstrapping estimates are developed for forecasting both potential output and change of productivity, technique and efficiency in the following decades.

From the perspective of potential output, the empirical results show that, on average, energy save and emission reduction will cause relatively large potential output loss in an early stage; but in long run, the loss will decline gradually and become lower than potential output growth finally, achieving the win-win development prospect stated in Porter hypothesis. Specifically, the bootstrapping estimates reveal that twenty six industrial sectors, 68.4 % of all sectors, enjoy a statistically significant potential net output gain, a certain win-win development prospect without sample noise, in the last forecasting year of 2050. From the viewpoint of productivity, the prediction analysis manifests that energy-saving and emission-reducing policy will have a larger negative impact on industrial technical progress at an early stage, especially for light industry; however, due to the obvious catching-up effect and increasing production efficiency in the early forecasting period and the rising technical progress dominated in the latter period, the industrial TFP is not negatively influenced by the environment regulation and always maintains an increasing trend. During the entire forecasting period from 2011 to 2050, on average the TFP growth of light, heavy and aggregated industry attains 1.80, 1.73 and 1.74 %, respectively. The bootstrapping estimates also support that most sectors experience a progress in productivity, technique and efficiency. Overall, although energy-saving and emission-abating regulation will cause certain loss at an early stage, in the long run, it will not only reach the target of improving environment quality but also increase the output and productivity, finally leading to the win-win development in the following 40 years. Our forecasting analysis in this paper favors the Porter hypothesis.

**Acknowledgments** The work is sponsored by Deutsche Forschungsgemeinschaft through SFB 649 “Economic Risk”. The supports from National Natural Science Foundation (71173048), National Social Science Foundation (12AZD047), Ministry of Education (11JJD790007), Shanghai Leading Talent Project and Fudan Zhuo-Shi Talent Plan are also acknowledged.

## References

- Ambec S, Barla P (2002) A theoretical foundation of the Porter hypothesis. *Econ Lett* 75(3):355–360
- Beaumont NJ, Tinch R (2004) Abatement cost curves: a viable management tool for enabling the achievement of win-win waste reduction strategies? *J Environ Manag* 71(3):207–215
- Boyd GA, McClelland JD (1999) The impact of environmental constraints on productivity improvement in integrated paper plants. *J Environ Econ Manag* 38:121–142
- Boyd GA, Tolley G, Pang J (2002) Plant level productivity, efficiency, and environmental performance of the container glass industry. *Environ Resour Econ* 23:29–43
- Cerin P (2006) Bringing economic opportunity into line with environmental influence: a discussion on the Coase theorem and the Porter and van der Linde hypothesis. *Ecol Econ* 56(2):209–225
- Chambers R, Chung YH, Färe R (1996) Benefit and distance function. *J Econ Theory* 70:407–419
- Chen S (2011) The abatement of carbon intensity in China: factor decomposition and policy implications. *World Econ* 34(7):1148–1167
- Chen S (2013) *Energy, environment and economic transformation in China*. Routledge Taylor & Francis Group, London
- Chen S, Golley J (2014) ‘Green’ productivity growth in China’s industrial economy. *Energy Econ* 44:89–98
- Chen S, Jefferson GH, Zhang J (2011) Structural change, productivity growth and industrial transformation in China. *China Econ Rev* 22(1):133–150
- Chen W, Gao P, He J (2004) Impacts of future carbon reductions on the Chinese GDP growth. *J Tsinghua Univ (Sci Technol)* 44(6):744–747



- Chenery HB, Robinson S, Syrquin M (1986) *Industrialization and growth: a comparative study*. Oxford University Press, New York
- Chung YH, Färe R, Grosskopf S (1997) Productivity and undesirable outputs: a directional distance function approach. *J Environ Manag* 51:229–240
- Daraio C, Simar L (2007) *Advanced robust and nonparametric methods in efficiency analysis: methodology and applications*. Springer, Berlin
- Efron B (1979) Bootstrap methods: another look at the jackknife. *Ann Stat* 7:1–26
- Färe R, Grosskopf S, Lovell K, Pasurka C (1989) Multilateral productivity comparisons when some outputs are undesirable: a nonparametric approach. *Rev Econ Stat* 71:90–98
- Färe R, Grosskopf S, Pasurka CA Jr (2001) Accounting for air pollution emissions in measures of state manufacturing productivity growth. *J Reg Sci* 41(3):381–409
- Faucheux S, Nicolai I (1998) Environmental technological change and governance in sustainable development policy. *Ecol Econ* 27:243–256
- Feichtinger G, Hartl RF, Kort PM, Veliov VM (2005) Environmental policy, the porter hypothesis and the composition of capital. *J Environ Econ Manag* 50(2):434–446
- Greker M (2006) Spillovers in the development of new pollution abatement technology: a new look at the Porter-hypothesis. *J Environ Econ Manag* 52(1):411–420
- Groom B, Grosjean P, Kontoleon A, Swanson T, Zhang S (2010) Relaxing rural constraints: a ‘win-win’ policy for poverty and environment in China? *Oxford Econ Pap* 62(1):132–156
- Hall P, Härdle W, Simar L (1995) Iterated bootstrap with application to frontier models. *J Product Anal* 6(1):63–76
- Härdle W (1990) *Applied nonparametric regression*. Cambridge University Press, Cambridge
- Jaffe A, Peterson S, Portney P, Stavins R (1995) Environmental regulation and the competitiveness of US manufacturing: what does the evidence tell us? *J Econ Lit* 33(1):132–163
- Jeon BM, Sickles RC (2004) The role of environmental factors in growth accounting. *J Appl Econ* 19(5):567–591
- Karvonen M (2001) Natural versus manufactured capital: win-lose or win-win? A case study of the Finnish pulp and paper industry. *Ecol Econ* 37(1):71–85
- Kuosmanen T, Bijsterbosch N, Dellink R (2009) Environmental cost-benefit analysis of alternative timing strategies in greenhouse gas abatement. *Ecol Econ* 68(6):1633–1642
- Lee CF, Lin SJ, Lewis C, Chang YF (2007) Effects of carbon taxes on different industries by fuzzy goal programming: a case study of the petrochemical-related industries, Taiwan. *Energy Policy* 35(8):4051–4058
- Lovell CAK (1993) Production frontiers and productive efficiency. In: Fried H, Lovell CAK, Schmidt SS (eds) *The measurement of productive efficiency: techniques and applications*. Oxford University Press, Oxford, pp 3–67
- Mohr RD (2002) Technical change, external economies, and the Porter hypothesis. *J Environ Econ Manag* 43(1):158–168
- Murty MN, Kumar S (2003) Win-win opportunities and environmental regulation: testing of porter hypothesis for Indian manufacturing industries. *J Environ Manag* 67(2):139–144
- Palmer K, Oates WE, Portney PR (1995) Tightening environmental standards: the benefit-cost or the no-cost paradigm. *J Econ Perspect* 9(4):97–118
- Porter ME (1991) America’s Green strategy. *Sci Am* 264(4):168
- Porter ME, van der Linde C (1995) Toward a new conception of the environment: competitiveness relationship. *J Econ Perspect* 9(4):97–118
- Reddy BS, Assenza GB (2009) The great climate debate. *Energy Policy* 37(8):2997–3008
- Roughgarden T, Schneider SH (1999) Climate change policy: quantifying uncertainties for damages and optimal carbon taxes. *Energy Policy* 27(7):415–429
- Schaltegger S, Synnestvedt T (2002) The link between green and economic success: environmental management as the crucial trigger between environmental and economic performance. *J Environ Manag* 65(4):339–346
- Sickles RC, Streitwieser ML (1998) An analysis of technology, productivity, and regulatory distortion in the interstate natural gas transmission industry: 1977–1985. *J Appl Econ* 13(4):377–395
- Silverman BW (1978) Choosing the window width when estimating a density. *Biometrika* 65:1–11
- Silverman BW (1986) *Density estimation for statistics and data analysis*. Chapman and Hall, London
- Simar L, Wilson PW (1998) Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. *Manag Sci* 44:49–61

- 
- Simar L, Wilson PW (1999) Estimating and bootstrapping Malmquist indices. *Eur J Oper Res* 115:459–471
- Xepapadeas A, De Zeeuw A (1999) Environmental policy and competitiveness: the Porter hypothesis and the composition of Capital. *J Environ Econ Manag* 37(2):165–182
- Zhang N, Choi Y (2013) Total-factor carbon emission performance of fossil fuel power plants in China: a metafrontier non-radial Malmquist index analysis. *Energy Econ* 40:549–559

## Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series

SONG SONG<sup>†</sup>, WOLFGANG K. HÄRDLE<sup>‡,§</sup> AND YA'ACOV RITOV<sup>‡,§</sup>

<sup>†</sup>*Department of Mathematics, University of Alabama, 318B Gordon Palmer Hall, Tuscaloosa, AL 35487, USA.*

E-mail: ssoonngg123@gmail.com

<sup>‡</sup>*School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099, Berlin, Germany.*

E-mail: haerdle@wiwi.hu-berlin.de

<sup>§</sup>*Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel.*

E-mail: yaacov@mscc.huji.ac.il

First version received: March 2012; final version accepted: November 2013

**Summary** High-dimensional non-stationary time series, which reveal both complex trends and stochastic behaviour, occur in many scientific fields, e.g. macroeconomics, finance, neuroeconomics, etc. To model these, we propose a generalized dynamic semi-parametric factor model with a two-step estimation procedure. After choosing smoothed functional principal components as space functions (factor loadings), we extract various temporal trends by employing variable selection techniques for the time basis (common factors). Then, we establish this estimator's non-asymptotic statistical properties under the dependent scenario ( $\beta$ -mixing and  $m$ -dependent) with the weakly cross-correlated error term. At the second step, we obtain a detrended low-dimensional stochastic process that exhibits the dynamics of the original high-dimensional (stochastic) objects and we further justify statistical inference based on this. We present an analysis of temperature dynamics in China, which is crucial for pricing weather derivatives, in order to illustrate the performance of our method. We also present a simulation study designed to mimic it.

**Keywords:** *Asymptotic inference, Factor model, Group Lasso, Periodic, Seasonality, Semi-parametric model, Spectral analysis, Weather.*

### 1. INTRODUCTION

Over the past few decades, high-dimensional data analysis has attracted increasing attention in various fields. We often face a high-dimensional vector of observations evolving in time (a very large interrelated time process), which is also possibly controlled by an exogenous covariate. For example, in macroeconomic forecasting, people use very large dimensional economic and financial time series (Stock and Watson, 2005b). In meteorology and agricultural economics, one of the primary interests is to study the fluctuations of temperatures at different nearby locations; for a recent summary, see Gleick et al. (2010). Such an analysis is also essential for pricing weather derivatives and hedging weather risks in finance (Odening et al., 2008). In

neuroeconomics, high-dimensional functional magnetic resonance imaging (fMRI) data are used to analyse the brain's response to certain risk-related stimuli, as well as to identify its activation area (Worsley et al., 2002). In financial engineering, the dynamics of the implied volatility surface (IVS) are studied for risk management, calibration and pricing purposes (Fengler et al., 2007). Other examples include mortality analysis (Lee and Carter, 1992), bond portfolio risk management or derivative pricing (Nelson and Siegel, 1987, and Diebold and Li, 2006), limit order book dynamics (Hall and Hautsch, 2006), yield curves (Bowsher and Meeks, 2006), and so on.

Empirical studies in economics and finance often involve non-stationary variables, such as real consumer price index, individual consumption, exchange rates, real gross domestic product, etc. For example, the large panel macroeconomic data, provided by Stock and Watson (2005a), contain some complex non-stationary behaviour, such as normal seasonality, large economic cycle and upward trend representing economic growth, etc. However, some studies have produced counterintuitive and contradictory results; see Campbell and Yogo (2006), Cai et al. (2009), Xiao (2009) and Wang and Phillips (2009a,b). This might partly be attributed to the use of methods that cannot capture non-stationarity or non-linear structural relations. In fact, in the econometrics literature, the study of such non-stationary time series is dominated by linear or, at most, parametric models, restricting non-stationarity to the unit root or long-memory autoregressive fractionally integrated moving average (ARFIMA) types of non-stationarity and restricting structural relations to linear or parametric types of cointegration models. General processes can be characterized by certain recurrence properties. These processes contain stationary, long-memory and unit-root type or nearly integrated processes as subclasses, and are more general than the class of locally stationary processes. As pointed out in the recent econometrics literature, when some covariates are non-stationary, conventional statistical tests are invalid, even though the predictive power in a non-parametric regression model can be improved if some covariates are non-stationary. While some asymptotic results for general non-parametric estimation methods for low-dimensional non-stationary time series have been obtained, semi-parametric modelling has hardly been investigated so far, especially for high-dimensional non-stationary time series. For the i.i.d. case, there have been many studies in the literature, including but not limited to Horowitz and Lee (2005), Horowitz et al. (2006) and Horowitz (2006) for the moderate-dimension case and Horowitz and Huang (2012) and Huang et al. (2010) for the high-dimension case.

In such situations, if we still use either high-dimensional static methods, which are initially designed for independent data or low-dimensional multivariate time series techniques (on a few concentrated series through naïve aggregation), we might lose potentially relevant information, such as the time dynamics or the space dependence structure. This might produce suboptimal forecasts and would be extremely inefficient. In macroeconomics studies, this potentially creates an omitted variable bias with adverse consequences for both structural analysis and forecasting. Christiano et al. (1999) has pointed out that the positive reaction of prices in response to a monetary tightening, the so-called price puzzle, is an artefact resulting from the omission of forward-looking variables, such as the commodity price index. The more scattered and dynamic the information is, the more severe this loss will be. To this end, an integrated solution addressing both issues is appealing. We need to analyse jointly time and space dynamics by simultaneously fitting a time series evolution and by fine tuning the factors involved. The solution we are seeking helps us to understand the spatial pattern, to gain strength from the different time points and, at the same time, to analyse the non-stationary temporal behaviour of the value at each spatial point. In this paper, we present and investigate the so-called generalized dynamic semi-parametric

factor model (GDSFM), together with its corresponding panel version, in order to address this problem.

Panel data have attracted much attention in econometrics; see, e.g. Baltagi (2005), Frees (2004) and Hsiao (1986). To address the above challenges in a large panel of economic and financial time series, some recent studies have proposed ways to impose restrictions on the covariance structure in order to limit the number of parameters to be estimated. Dynamic factor models introduced by Forni et al. (2000) and Stock and Watson (2002a,b), also discussed by Forni et al. (2005) and Giannone et al. (2005), have drawn upon the idea that the intertemporal dynamics can be explained and represented by a few common factors (low-dimensional time series). Another approach in this field has been presented by Park et al. (2009), where a latent  $L$ -dimensional process,  $Z_1, \dots, Z_T$ , is introduced, and the  $J$ -dimensional random process  $Y_t = (Y_{t,1}, \dots, Y_{t,J})^\top$ ,  $t = 1, \dots, T$ , is represented as

$$Y_{t,j} = Z_{t,1}m_{1,j} + \dots + Z_{t,L}m_{L,j} + \varepsilon_{t,j}, \quad j = 1, \dots, J, \quad t = 1, \dots, T. \quad (1.1)$$

Here,  $Z_{t,l}$  are the common factors depending on time,  $\varepsilon_{t,j}$  are errors or specific factors, and the coefficients  $m_{l,j}$  are factor loadings. The index  $t = 1, \dots, T$  reflects the time evolution,  $\{Z_t\}_{t=1}^T$  ( $Z_t = (Z_{t,1}, \dots, Z_{t,L})^\top$ ) is assumed to be a stationary random process, and  $m_l = (m_{l,1}, \dots, m_{l,J})^\top$  captures the spatial dependency structure. The study of the time behaviour of the high-dimensional  $Y_t$  is then simplified to the modelling of  $Z_t$ , which is more feasible when  $L \ll J$ . Model (1.1) reduces to a special case of the generalized dynamic factor model (approximate factor model) considered by Forni et al. (2000, 2005) and Hallin and Liska (2007), when  $Z_{t,l} = a_{l,1}(B)U_{t,1} + \dots + a_{l,q}(B)U_{t,q}$ . Here, the  $q$ -dimensional vector process  $U_t = (U_{t,1}, \dots, U_{t,q})^\top$  is an orthonormal white noise and  $B$  denotes the lag operator. In this case, model (1.1) is expressed as  $Y_{t,j} = m_{0,j} + \sum_{k=1}^q b_{k,j}(B)U_{t,k} + \varepsilon_{t,j}$ , where  $b_{k,j}(B) = \sum_{l=1}^L a_{l,k}(B)m_{l,j}$ . Less general models in the literature include the static factor models proposed by Stock and Watson (2002a,b) and the exact factor models suggested by Sargent and Sims (1977) and Geweke (1977).

Our goal of modelling high-dimensional non-stationary time series is achieved by using a sparse representation approach to regression. In fact, we combine spatio-temporal modelling with group Lasso (Yuan and Lin, 2006). We approximate both the temporal common factors and spatial factor loadings by a linear combination of series terms. Because the temporal non-stationarity behaviour might result from different sources, the choice of basis functions is important. We start by introducing an overparametrized model, which can capture (almost) any type of temporal behaviour, such as cyclic behaviour plus linear or quadratic trends, by utilizing series basis, such as powers, trigonometrics, local polynomials, periodic functions and B-splines. Then, we select a sparse submodel, using penalizing-Lasso and group-Lasso techniques.

In practice, there might be multiple subjects, each of which by itself corresponds to a set of high-dimensional time series. For example, in international economies, industrial organizations or financial studies, there are data for many countries, firms or assets, all of which are high-dimensional. Thus, we also need to provide a panel version of the high-dimensional time series model to address this issue. Compared with previous studies in the literature, the novelty of this paper lies in the following aspects.

1. When the time process is not stationary (i.e. the process has a non-linear, non-parametric temporal structure in time), using a skilful selection of time basis, we can handle such complex time series. To achieve a successful selection, the key assumption is that the initially proposed time basis should not be too dependent, even though the number can

be large (i.e. we should include as many orthogonal time basis functions as possible for the automatic selection). From the point of view of large panel time series modelling, we incorporate non-stationarity and non-linearity (complex trends) into time dynamics. We deviate from most of the current body of literature that still requires  $Z_t$  to be stationary and still needs a large number of observations (relative to dimensionality) to establish asymptotic properties.

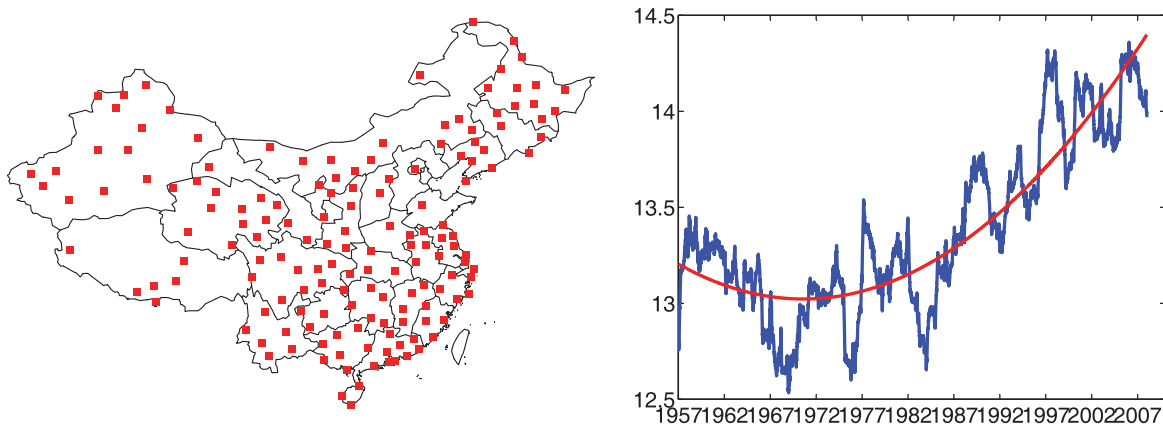
2. The contribution lies in the way the time dynamics is introduced for variable selection and regularization methods. Under the assumption that the product of the time basis, space basis and error term has a bounded second moment, and the error term  $\varepsilon_t$  is only weakly cross-correlated, the non-asymptotic theoretical properties of existing methods are established under the scenario of independence. We extend it to a dependent scenario ( $\beta$ -mixing and  $m$ -dependent process) with the weakly cross-correlated error term (the details are specified in Assumption 3.2), and we derive oracle sparsity inequalities (non-asymptotic risk bounds). The key assumption is that the temporal dependence level of the error term is controlled within some level. Also, this result is not built upon any specific forms of time and space basis.
3. When the space structure of  $m_l$  is complex, the low-dimensional parametrizations do not capture it properly. We employ a data-driven semi-parametric method, introduced by Hall et al. (2006), to capture the spatial dependence structure.
4. For the case that there might be multiple subjects, each of which corresponds to a set of high-dimensional time series, we provide a panel version of the model with a corresponding estimation method.

In a variety of applications, we have explanatory variables  $X_{t,j} \in \mathbb{R}^d$  at hand, e.g. the geo coordinates of weather stations, the voxels (volume elements, representing values on regular grids) of fMRI, or the moneyness and time-to-maturity variables for implied volatility modelling, which can influence the factor loadings  $m_l$ . An important refinement of model (1.1) is to incorporate the existence of observable covariates  $X_{t,j}$  from Park et al. (2009). The factor loadings are then generalized to functions of  $X_{t,j}$ . In the following, we write  $X_t = (X_{t,1}, \dots, X_{t,J})^\top$  and consider the generalization of (1.1),

$$Y_{t,j} = Z_t^\top m(X_{t,j}) + \varepsilon_{t,j}, \quad t = 1, \dots, T, \quad (1.2)$$

where  $Y_{t,j}, \varepsilon_{t,j} \in \mathbb{R}$ ,  $X_{t,j} \in \mathbb{R}^d$ ,  $m : \mathbb{R}^d \rightarrow \mathbb{R}^L$  and  $Z_t \in \mathbb{R}^{1 \times L}$ .

Our motivating example is from temperature analysis for pricing weather derivatives. The data set is taken from the Climatic Data Center (CDC) of the China Meteorological Administration (CMA). It contains daily observations from 159 weather stations across China from 1 January 1957 to 31 December 2009. We would not only like to address the question of whether there is a change in time, but also to permit a different trend in time, in different climate types, as shown by Figure 1 (left), which shows a map of the network of China's weather stations. Besides the well-known seasonality effect, we can expect a trend related to climate change. In Figure 1 (right), we show the moving average (of 730 nearby days) of temperatures in China from 1 January 1957 to 31 December 2009, which is  $(159 \times 730)^{-1} \sum_{s=-354}^{+365} \sum_{j=1}^{159} Y_{t+s,j}$ , where  $Y_{t,j}$  is the temperature of the  $j$ th weather station at time  $t$ . From this figure, we can see that there is a large period (around 10 years) between peaks and an upward trend for China's temperatures. Besides these trends, there is also stochasticity inherent in the remaining time dynamics, which is essential for pricing weather derivatives and hedging weather risks. By simultaneously studying the dynamics of temperatures in various places w.r.t.  $X_{t,j} = X_j$



**Figure 1.** Map of China’s weather stations and moving averages of temperature.

(the three-dimensional geographical information of the  $j$ th weather station), we will be able to estimate, forecast and price temperatures in time and space.

The rest of the paper is organized as follows. In the next section, we present details of the GDSFM, together with the corresponding basis selection and panel model. We present the estimator’s properties under various scenarios in Section 3. In Section 4, we apply the method to the motivating problem: the dynamic behaviour of temperatures. In Section 5, we present the results of simulation studies that mimic the previous empirical example. Section 6 contains concluding remarks. The estimation procedure and all technical proofs are sketched in Appendices A and B, respectively.

## 2. GENERALIZED DYNAMIC SEMI-PARAMETRIC FACTOR MODELS

We observe  $(X_{t,j}, Y_{t,j})$  for  $j = 1, \dots, J$  and  $t = 1, \dots, T, Y_{tj} \in \mathbb{R}, X_{tj} \in \mathbb{R}^d, \varepsilon_{tj} \in \mathbb{R}$  generated by

$$Y_{tj}^\top = Z_t^\top A^* \Psi(X_{tj}) + \varepsilon'_{tj} = (U_t^\top \Gamma^* + Z_{0,t}^\top) A^* \Psi(X_{tj}) + \varepsilon'_{tj},$$

where  $A^*$  and  $\Gamma^*$  are the  $L \times K$  and  $R \times L$  (unknown) underlying coefficient matrices and  $Z_t$  has two components  $\Gamma^{*\top} U_t$  and  $Z_{0,t}$ . Let  $Y_t = (Y_{t,1}, \dots, Y_{t,J})^\top, X_t = (X_{t,1}, \dots, X_{t,J})^\top, \varepsilon'_t = (\varepsilon'_{t,1}, \dots, \varepsilon'_{t,J})^\top$  and  $\Psi(X_t) = (\Psi(X_{t1}), \dots, \Psi(X_{tJ}))$  (abbreviated as  $\Psi_t$ ). We rewrite this in compact form as

$$\begin{aligned} Y_t^\top &= (U_t^\top \Gamma^* + Z_{0,t}^\top) A^* \Psi(X_t) + \varepsilon_t'^\top \\ &= U_t^\top \Gamma^* A^* \Psi(X_t) + Z_{0,t}^\top A^* \Psi_t + \varepsilon_t'^\top. \end{aligned} \tag{2.1}$$

Again, by introducing  $\beta^{*\top} = \Gamma^* A^*$  (the  $R \times K$  unknown underlying coefficient matrices consisting of  $\beta_{rk}$ ) and  $\varepsilon_t = Z_{0,t}^\top A^* \Psi_t + \varepsilon'_t$ , we could further simplify this as

$$Y_t^\top \stackrel{\text{def}}{=} U_t^\top \beta^{*\top} \Psi_t + \varepsilon_t^\top. \tag{2.2}$$

Note the following.

1. Time evolution/common factors.  $Z_t = (Z_{t,1}, \dots, Z_{t,L})^\top$  is an unobservable  $L$ -dimensional process consisting of both a deterministic portion,  $\Gamma^{*\top} U_t$ , and a stochastic portion,  $Z_{0,t}$ . Here,  $\{Z_{0,t}\}_{t=1}^T$  is a stationary process (to be detailed later). A key difference between our method and that of Park et al. (2009) is this additional non-stationary component  $\Gamma^{*\top} U_t$ .
2. Factor loading functions and error terms.  $m(X_{tj}) = A^* \Psi(X_{tj})$  is an  $L$ -tuple  $(m_1, \dots, m_L)$  of unknown real-valued functions  $m_l$  defined on a subset of  $\mathbb{R}^d$  and  $\varepsilon'_t = (\varepsilon'_{t,1}, \dots, \varepsilon'_{t,J})^\top$  are the errors. Throughout the paper, we assume that the covariates  $X_{t,j}$  have support  $[0, 1]^d$ . The error terms  $\varepsilon_t$  and  $\varepsilon'_t$  only need to satisfy some mild condition (details specified in Assumptions 3.2 and 3.3(c)), which allows them to be weakly dependent (over time) and cross-correlated (over space).
3. Time and space basis. We use a series expansion to capture the time trend and the space-dependent structure. Let  $U_t^\top = (u_1(t), \dots, u_R(t))$  be the  $1 \times R$  vector of time basis functions (polynomial and trigonometric functions, etc.), which are selected and weighted by the matrix  $\Gamma^*$ . For the space basis, we take  $\Psi_t = (\psi_1(X_t), \dots, \psi_K(X_t))^\top$  ( $K \times J$  matrix). For every  $\beta$  matrix, we introduce  $\beta_r = (\beta_{rk}, 1 \leq k \leq K)$ , which is the column vector formed by the coefficients corresponding to the  $r$ th time basis. Additionally, we define the mixed  $(2, 1)$  norm  $\|\beta\|_{2,1} = \sum_{r=1}^R \sqrt{\sum_{k=1}^K \beta_{rk}^2}$ . Finally, we set  $\mathcal{R}(\beta) = \{r : \beta_r \neq 0\}$  and  $M(\beta) = |\mathcal{R}(\beta)|$ , where  $|\mathcal{R}(\beta)|$  denotes the cardinality of set  $\mathcal{R}(\beta)$ . For the sake of simplicity and convenience, we use  $|\cdot|$  to denote the  $L_1$  norm for vectors and  $\|\cdot\|$  to denote the  $L_2$  norm for vectors or the mixed  $(2, 1)$  norm for matrices.

Because the non-stationary behaviour might be very complex, to ensure that all the trends causing the non-stationarity are considered, the dimension  $R$  of the initially included time basis might be large. For example, in the temperature analysis, because we never know the exact frequency (frequencies) of the period(s), at the beginning, we include all the basis functions. We think that this might be useful for capturing the non-stationary behaviour, e.g. 16 trigonometric functions w.r.t. different frequencies and  $53 \times 3$  (year by year) cubic polynomial basis. Consequently, we end up with  $R = 175$ . However, to avoid overfitting, variable selection with regularization techniques is necessary. A popular variable selection method is the Lasso (Tibshirani, 1996). An extension for factor-structured models is the group Lasso (Yuan and Lin, 2006), in which the penalty term is a mixed  $(2, 1)$  norm of the coefficient matrix. Here, we assume that the vectors  $\beta_r$  are not only sparse, but also have the same sparsity pattern across different factors. We study the estimator's theoretical sparsity properties related to the time basis selection, and we take (2.1) to be the true model. Because group LASSO permits overparametrization, this is a mild assumption. We would also like to emphasize that our non-asymptotic sparse oracle inequality results are independent of specifications of time and space basis. They apply equally to local polynomials, periodic functions, such as sin and cos, and B-splines, etc., while we just assume that there is no additional approximation error for obtaining the space basis at this non-asymptotic analysis step.

### 2.1. A panel version with multiple individuals

Here, we just present a panel version of (2.1) based on assumptions closely related to the fMRI neuroeconomics study (Mohr et al., 2010). It is reasonable to assume that different subjects have



different patterns of brain activation (to the external stimuli) represented by the time series  $Z_t$ , but they (and all human beings) share essentially the same spatial structure of the brain represented by the space function  $A^* \Psi_t$ . With a panel of  $I$  subjects, we formulate the following generalization of (2.1) and (2.2),

$$Y_{t,j}^i = \sum_{l=1}^L (Z_{0,t,l}^i + U_t^\top \Gamma_l^i) m_l(X_{t,j}) + \varepsilon_{t,j}^i, \quad 1 \leq j \leq J_t, \quad 1 \leq t \leq T, \quad 1 \leq i \leq I, \quad (2.3)$$

where the fixed effects  $Z_{0,t,l}^i$  and  $\Gamma_l^i$  are the individual effects on functions  $m_l$  for subject  $i$  at time point  $t$ . For identification purposes, assume

$$E\left[\sum_{i=1}^I \sum_{l=1}^L Z_{0,t,l}^i m_l(X_{t,j}) | X_{t,j}\right] = 0.$$

For this data structure, we use  $\bar{Y}_{t,j}$  to denote the average of  $Y_{t,j}^i$  across different subjects  $i$ . Thus, from (2.3), we have

$$\bar{Y}_{t,j} = \sum_{l=1}^L (U_t^\top \bar{\Gamma}_l) m_l(X_{t,j}) + \varepsilon_{t,j} \quad 1 \leq j \leq J.$$

The two-step estimation procedure for the panel version model is as follows.

STEP 1. Take the average of  $Y_{t,j}^i$  across different subjects  $i$ , and estimate the common basis function in space  $\hat{m}_l$  as in the original approach; see Appendix A for more details.

STEP 2. Given the common  $\hat{m}_l$ , estimate subject-specific time factors  $Z_{t,l}^i$ :

$$Y_{t,j}^i = \sum_{l=1}^L (Z_{0,t,l}^i + U_t^\top \Gamma_l^i) \hat{m}_l(X_{t,j}) + \varepsilon_{t,j}^i.$$

Next, we discuss the choice of time basis  $U_t$ , space basis  $\Psi_t$  and the estimation procedure for (2.2).

## 2.2. Choice of time basis

To capture the global trend in time, we can use any orthogonal polynomial basis, e.g.  $u_1(t) = 1/C_1$ ,  $u_2(t) = t/C_2$ ,  $u_3(t) = (3t^2 - 1)/C_3$ , ... (where  $C_i$  are generic constants with  $T^{-1} \sum_{t=1}^T u_r^2(t)/C_r^2 = 1$ ). We can also use the fact that there are natural frequencies in the data, and thus start with a few trigonometric functions. In the temperature example, the yearly cycle and a large period are two clear phenomena. To capture these periodic variations, we can use trigonometric functions,  $u_4(t) = \sin(2\pi t/p)/C_4$ ,  $u_5(t) = \cos(2\pi t/p)/C_5$ ,  $u_6(t) = \sin(2\pi t/(p/2))/C_6$ ,  $u_7(t) = \cos(2\pi t/(p/2))/C_7$ , ..., with the given period  $p$ : 365 and 10 for the yearly cycle and large period, respectively. In the fMRI application of Myšičková et al. (2013), the basic experiment is repeated every 29.5 seconds, and we have the period  $p = 11.8$  (there is a fMRI scan every 2.5 seconds). In general, to adopt various types of non-linearities, various basis functions could be employed, such as powers, trigonometrics, local polynomials, periodic functions, B-splines, etc. The theory to be presented later for selecting the significant time basis selection is actually independent of their specific forms, and thus is very useful in practice.

### 2.3. Choice of space basis

There are various choices for space basis. Park et al. (2009) have proposed a multidimensional B-spline basis. Alternatively, functional principal component analysis (PCA; Hall et al., 2006) can be employed, which combines smoothing techniques with ideas related to functional PCA. The basic steps are as follows.

STEP 1. Calculate the covariance operator (in a functional sense). Denote  $X_{tj} = (X_{tj}^1, \dots, X_{tj}^d)$ ,  $u = (u^1, \dots, u^d)$  and  $v = (v^1, \dots, v^d)$  (as for  $b, \hat{b}, b_1, \hat{b}_1, b_2$  and  $\hat{b}_2$ ). Given  $u \in [0, 1]^d$ , and bandwidths  $h_\mu$  and  $h_\phi$ , define  $(\hat{a}, \hat{b})$  to minimize

$$\min_{a,b} \sum_{t=1}^T \sum_{j=1}^{J_t} (Y_{tj} - a - b^\top(u - X_{tj}))^2 K\left(\frac{X_{tj} - u}{h_\mu}\right),$$

and take  $\hat{\mu}(u) = \hat{a}$ . Then, given  $u, v \in [0, 1]^d$ , choose  $(\hat{a}_0, \hat{b}_1, \hat{b}_2)$  to minimize

$$\sum_{t=1}^T \sum_{1 \leq j \neq k \leq J_t} (Y_{tj} Y_{tk} - a_0 - b_1^\top(u - X_{tj}) - b_2^\top(v - X_{tk}))^2 K\left(\frac{X_{tj} - u}{h_\phi}\right) K\left(\frac{X_{tk} - v}{h_\phi}\right).$$

Denote  $\hat{a}_0$  by  $\hat{\phi}(u, v)$  and construct  $\hat{\mu}(v)$  similarly to  $\hat{\mu}(u)$ . The estimate of the covariance operator is then

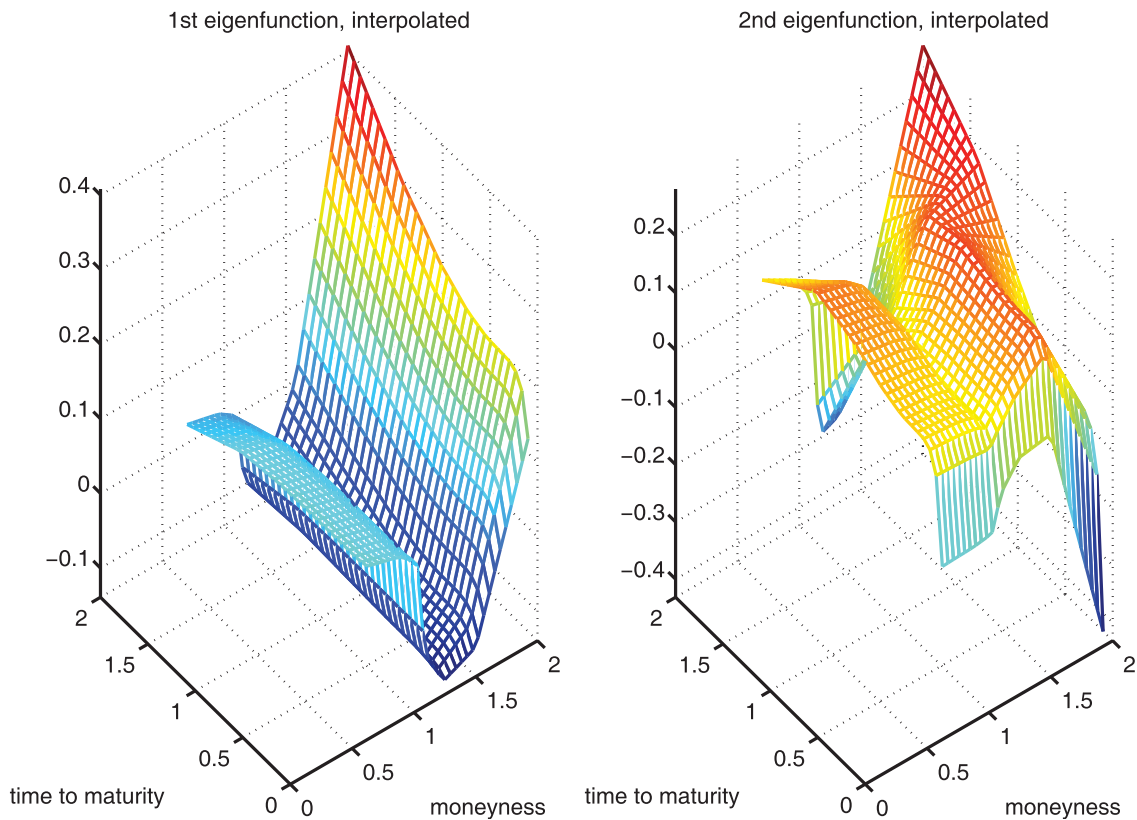
$$\hat{\psi}(u, v) = \hat{\phi}(u, v) - \hat{\mu}(u)\hat{\mu}(v). \quad (2.4)$$

STEP 2. Compute the principal space basis. Obtain from (2.4) the largest  $K$  eigenvalues and corresponding orthonormal eigenfunctions as the basis  $\hat{\psi}_1(x), \dots, \hat{\psi}_K(x)$ . For computational methods and practical considerations, we refer to Section 8.4 of Ramsay and Silverman (2005).

As remarked by Hall et al. (2006), the operator defined by (2.4) is not necessarily positive semi-definite, but it is assured to have real eigenvalues. Theorem 1 of Hall et al. (2006) provides theoretical foundations that the bandwidths  $h_\mu$  and  $h_\phi$  should be chosen as  $\mathcal{O}(T^{-1/5})$  to minimize the distance between the estimates  $\hat{\psi}$  and the corresponding true  $\hat{\psi}$ . In Section 4 (details presented later), we find that the performance of  $\hat{\beta}$  is very robust to the choice of the smoothing parameter.

We would like to emphasize that the space basis function  $\hat{\Psi}_t$  is only an estimator of the true (unobservable)  $\Psi_t$ . However, in proving the properties of the time basis selection, as in Theorem 3.2 and Corollary 3.1, we assume that this space basis estimation does not affect the study of selecting the temporal basis, because, otherwise, the non-asymptotic theoretical deviation will be too complex. If we still stick to the B-spline basis as in Park et al. (2009), all the proofs afterwards do not need to be modified. For simplicity of notation, we continue to use  $\Psi_t$  to denote this estimate of space basis from now.

We apply this method to the implied volatility modelling problem, which has been discussed in detail by Park et al. (2009). Figure 2 displays the space basis modelling using the functional PCA approach, which could capture the special ‘smiling’ effect well, while the spline basis modelling cannot.



**Figure 2.** Space basis using the functional PCA approach for IVS modelling.

### 3. PROPERTIES OF ESTIMATES

In this section, we study sparse oracle inequalities for the estimate  $\hat{\beta}$  defined in (A.1), assuming that the errors  $\varepsilon_t$  are dependent ( $\beta$ -mixing in Theorem 3.2 and  $m$ -dependent in Corollary 3.1). This work extends those of Lounici et al. (2009), Bickel et al. (2009) and Lounici (2008) concerning upper bounds on the prediction error and the distance between the estimator and the true matrix  $\beta^*$ .

For the second step of the estimation procedure, an important question arises: is it justified, from an inferential point of view, to base further statistical inference on the detrended stochastic time series? Theorem 3.4 shows that the difference between the inference based on the estimated time series and true unobserved time series is asymptotically negligible.

Before stating the first theorem, we make the following assumption.

ASSUMPTION 3.1. *There exists a positive number  $\kappa = \kappa(s)$  such that*

$$\min\left(\frac{\sqrt{\sum_t \|\Psi_t^\top \Delta U_t\|^2}}{\sqrt{T} \|\Delta_{\mathcal{R}}\|} : |\mathcal{R}| \leq s, \Delta \in \mathbb{R}^{K \times R} \setminus \{0\}, \|\Delta_{\mathcal{R}^c}\|_{2,1} \leq 3 \|\Delta_{\mathcal{R}}\|_{2,1}\right) \geq \kappa,$$

where  $\mathcal{R}^c$  denotes the complement of the set of indices  $\mathcal{R}$  and  $\Delta_{\mathcal{R}}$  denotes the matrix formed by stacking the rows of matrix  $\Delta$  w.r.t. row index set  $\mathcal{R}$ .

Assumption 3.1 is essentially a restriction on the eigenvalues of  $\sum_{t=1}^T U_t U_t^\top$  as a function of sparsity  $s$ . In fact, it requires that the initially involved time basis is not too dependent, which is naturally satisfied by orthogonal polynomials and trigonometric functions. Low sparsity means that  $s$  is big and therefore  $\kappa$  is small. Thus,  $\kappa(s)$  is a decreasing function of  $s$ ; see also Lemma 4.1 of Bickel et al. (2009) for more details and related discussions.

**THEOREM 3.1 (DETERMINISTIC PART).** *Consider the model (2.2). Assume that  $\Psi_t \Psi_t^\top = I_K$  (orthonormalized space basis),  $T^{-1} \sum_{t=1}^T U_t^\top U_t / R = 1$ , and the number of true non-zero time basis  $M(\beta^*) \leq s$ . If the random event*

$$\mathcal{A} = (2T^{-1} \max_{1 \leq r \leq R} \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}^\top \varepsilon_{tj} U_{tr} \leq \lambda) \quad (3.1)$$

holds for some  $\lambda > 0$  and Assumption 3.1 is satisfied, then, for any solution  $\hat{\beta}$  of (A.1), we have

$$T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\hat{\beta} - \beta^*) U_t\|^2 \leq 16s\lambda^2 \kappa^{-2}, \quad (3.2)$$

$$K^{-1/2} \|\hat{\beta} - \beta^*\|_{2,1} \leq 16s\lambda K^{-1/2} \kappa^{-2}, \quad (3.3)$$

$$M(\hat{\beta}) \leq 64\phi_{\max}^2 s \kappa^{-2}. \quad (3.4)$$

Note that Theorem 3.1 is valid for any  $J, R, T$  and any type of distribution of  $\varepsilon_t$ , and yields non-asymptotic bounds.

Because the standard assumption that  $\varepsilon_t$  is independent is often unsatisfied in practice, it is important to understand how the estimator behaves in a more general situation (i.e. with dependent error terms). As far as we know, our result is one of the first attempts to deal with dependent error terms for (group) Lasso variable selection techniques. We build it w.r.t.  $\beta$ -mixing, which is an important measure of dependence between  $\sigma$ -fields (for time series). A detailed definition can be found in Appendix B (before Proof of Theorem 3.2). A very natural question to ask is, to what extent the degree of dependence (in terms of  $\beta$ -mixing coefficients) is allowed, while we can still obtain certain sparse oracle inequalities (i.e. to study the relationship among high dimensionality  $R$ , moderate sample size  $T$  and  $\beta$ -mixing coefficients  $\beta$ ).

We use the following mild technical assumption similar to the typical bounded second-moment requirement for i.i.d. data.

**ASSUMPTION 3.2.** *The matrices  $\Psi_t$  and  $U_t$  and random variables  $\varepsilon_t$  are such that for  $V_t \stackrel{\text{def}}{=} K^{-1/2} \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj} \varepsilon_{tj} U_{tr}$ ,  $\exists \sigma^2$  such that  $\forall n, m, m^{-1} E[V_n + \dots + V_{n+m}]^2 \leq \sigma^2$  and  $\forall t, |V_t| \leq C'', \forall r$  and some constants  $\sigma^2, C'' > 0, t = 1, \dots, T$ .*

Note that because  $V_t$  (as a function of  $\varepsilon_{tj}$ ) is defined as a sum over  $j$ , it also indicates that the error term  $\varepsilon_t$  could be weakly cross-correlated. We can now state our main result.

**THEOREM 3.2 ( $\beta$ -MIXING).** *Consider the model (2.2). Assume the sequence  $\{V_t\}_{t=1}^T$  satisfies Assumption 3.2 and the  $\beta$ -mixing condition with the  $\beta$ -mixing coefficients*

$$\begin{aligned} & \beta(((3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2} C''^{-1} \log R^{-(1+\delta')/2}) - 1) \\ & \leq (24\sigma(1-\varepsilon)^{1/2} (R^{1+\delta'} \sqrt{\log R^{1+\delta'} T C''})^{-1}), \end{aligned}$$

for any  $\varepsilon > 0$ , some  $\delta' > 0$  and  $\lambda$  defined below.  $\Psi_t \Psi_t^\top = I_K$ ,  $T^{-1} \sum_{t=1}^T U_t^\top U_t / R = 1$ , and  $M(\beta^*) \leq s$ . Furthermore, let  $\kappa$  be defined as in Assumption 3.1 and let  $\phi_{\max}$  be the maximum eigenvalue of the matrix  $\sum_{t=1}^T U_t U_t^\top / T$ . Let

$$\lambda = \sqrt{\frac{16 \log R^{1+\delta'} K \sigma^2}{T(1-\varepsilon)}}.$$

Then, with a probability of at least  $1 - 3R^{-\delta'}$ , for any solution  $\hat{\beta}$  of (A.1), we have

$$T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\hat{\beta} - \beta^*) U_t\|^2 \leq 256s \left( \frac{\log R^{1+\delta'} K \sigma^2}{T(1-\varepsilon)} \right) \kappa^{-2}, \tag{3.5}$$

$$K^{-1/2} \|\hat{\beta} - \beta^*\|_{2,1} \leq 96s \sqrt{\frac{\log R^{1+\delta'} \sigma^2}{T(1-\varepsilon)}} \kappa^{-2}, \tag{3.6}$$

$$M(\hat{\beta}) \leq 64\phi_{\max}^2 s \kappa^{-2}. \tag{3.7}$$

REMARK 3.1. Before explaining the results, as also mentioned in Song and Bickel (2011), we would first like to discuss some related results. For technical simplicities, we consider the following simplest linear regression model with  $R \rightarrow \infty$ :

$$e_t = x_{t1}\theta_1 + \dots, x_{tR}\theta_R + \epsilon_t = x_t^\top \theta + \epsilon_t, \tag{3.8}$$

with the regressors  $(x_{t1}, \dots, x_{tR}) = x_t^\top$ , the coefficients  $(\theta_1, \dots, \theta_R) = \theta^\top$  and the error term  $\epsilon_t$ . Suppose  $x$  in (3.8) has full rank  $R$  and  $\epsilon_t$  is  $N(0, \sigma^2)$ . Consider the least-squares estimate ( $R \leq T$ )  $\hat{\theta}_{OLS} = (x x^\top)^{-1} x e$ . Then, from standard least-squares theory, we know that the prediction error  $\|x^\top (\hat{\theta}_{OLS} - \theta^*)\|_2^2 / \sigma^2$  is  $\chi_R^2$ -distributed, i.e.

$$E\left[\frac{\|x^\top (\hat{\theta}_{OLS} - \theta^*)\|_2^2}{T}\right] = \frac{\sigma^2}{T} R. \tag{3.9}$$

In the sparse situation, if  $\epsilon_t$  is  $N(0, \sigma^2)$  (different from our case), Corollary 6.2 of Bühlmann and van de Geer (2011) shows that the Lasso estimate obeys the following oracle inequality:

$$\frac{\|x^\top (\hat{\theta}_{Lasso} - \theta^*)\|_2^2}{T} \leq C_0 \frac{\sigma^2 \log R}{T} M(\theta^*), \tag{3.10}$$

with a large probability and some constant  $C_0$ . The additional  $\log R$  factor here could be seen as the price to pay for not knowing the set  $\{\theta_p^*, \theta_p^* \neq 0\}$  (Donoho and Johnstone, 1994). Similar to the i.i.d. Gaussian situation discussed above, the term  $(\log R)^{1+\delta'}$  in (3.5) could be interpreted as the price to pay for not knowing the set  $\{\theta_r^*, \theta_r^* \neq 0\}$ . Here, we have  $(\log R)^{1+\delta'}$  instead of  $\log R$  because we deviate from the typical i.i.d. Gaussian situation and establish the results under the more general Assumption 3.2, which can be thought of as the finite second-moment condition. Also, the  $\delta'$  term is the price to pay for this deviation.

REMARK 3.2. Because

$$\begin{aligned} & \beta(((3/8)\sigma\varepsilon^2T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}\log R^{-(1+\delta')/2}) - 1) \\ & \leq (24\sigma(1-\varepsilon)^{1/2}(R^{1+\delta'}\sqrt{\log R^{1+\delta'}TC''})^{-1}) \end{aligned}$$

is required, when dimensionality  $R$  increases, the allowed dependence level reflected by the  $\beta$ -mixing coefficients must decrease fast enough so that we still achieve similar risk bounds as in the independent case. Intuitively, this makes sense because if the dependence level inherent in  $Z_{0,t}$  (or  $\varepsilon_t$  equivalently) is too strong (i.e.  $\beta$  exceeds some level), then the amount of information provided by these observations is less, and therefore the estimate does not perform well. However, strong dependence in  $Z_{0,t}$  might be caused by some trend, which should be included in  $U_t^\top \Gamma$ , but is not, which results in the increased dependence. This tells us that at the beginning, we should include a large enough number  $R$  of pre-specified time basis functions such that it could include most of the deterministic (even though it could be segment by segment) time evolution and the remaining dependence level in  $Z_{0,t}$  is controlled.

COROLLARY 3.1 (*m*-DEPENDENT). Consider the model (2.2). Assume that the sequence  $\{V_t\}_{t=1}^T$  is an *m*-dependent process with order  $k$  ( $k \geq 1$ ) and satisfies the following conditions for some constants  $\sigma_0^2, C'' > 0, t = 1, \dots, T$ : (a)  $\forall t, E[V_t^2] \leq \sigma_0^2, |V_t| \leq C''$ ; (b)  $((3/8)\sigma\varepsilon^2T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}\log R^{-(1+\delta')/2}) - 1 \geq k + 1$  for any  $\varepsilon > 0$  and some  $\delta' > 0$ . Also,  $\Psi_t\Psi_t^\top = I_K, T^{-1}\sum_{t=1}^T U_t^\top U_t/R = 1$ , and  $M(\beta^*) \leq s$ . Furthermore, let  $\kappa$  be defined as in Assumption 3.1, let  $\phi_{\max}$  be the maximum eigenvalue of the matrix  $\sum_{t=1}^T U_t U_t^\top / T$  and let  $\lambda$  be defined as in Theorem 3.2. Then, with a probability of at least  $1 - 3R^{-\delta'}$ , for any solution  $\hat{\beta}$  of (A.1), we have

$$T^{-1} \sum_{t=1}^T \|\Psi_t^\top(\hat{\beta} - \beta^*)U_t\|^2 \leq 512s \left( \frac{\log R^{1+\delta'} K k \sigma_0^2}{T(1-\varepsilon)} \right) \kappa^{-2}, \tag{3.11}$$

$$K^{-1/2} \|\hat{\beta} - \beta^*\|_{2,1} \leq 96\sqrt{2}s \sqrt{\frac{\log R^{1+\delta'} k \sigma_0^2}{T(1-\varepsilon)}} \kappa^{-2}, \tag{3.12}$$

$$M(\hat{\beta}) \leq 64\phi_{\max}^2 s \kappa^{-2}. \tag{3.13}$$

REMARK 3.3. We can see that when  $k$  increases (i.e. the dependence in  $\{V_t\}_{t=1}^T$  becomes stronger and stronger), the risk bounds become larger and larger. To ensure  $((3/8)\sigma\varepsilon^2T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}) - 1 \geq k + 1$ , approximately we need  $T^{1/2}\log R^{-(1+\delta')/2} \geq ((3/4)\sigma_0\varepsilon^2\sqrt{(1-\varepsilon)})^{-1}C''\sqrt{k}$ , which gives the requirement on the sample size  $T$  (relative to the high dimensionality) and the amount of information from the data. Similar results could also be separately obtained for the generalized *m*-dependent process based on fractional cover theory and the (extended) McDiarmid inequality; see Theorem 2.1 of Janson (2004). At the second step,  $Z_{0,t}$  is estimated based on  $\hat{\beta}$  instead of  $\beta^*$ , so we need to show that the influence of this plug-in estimate is negligible. Our result relies on the following assumptions, which are similar to Assumptions (A1)–(A8) in Park et al. (2009).

ASSUMPTION 3.3. (a) The sets of variables  $(X_{1,1}, \dots, X_{T,J}), (\varepsilon'_{1,1}, \dots, \varepsilon'_{T,J})$  and  $(Z_{0,1}, \dots, Z_{0,T})$  are independent of each other; (b) for  $t = 1, \dots, T$ , the variables  $X_{t,1}, \dots, X_{t,J}$

are identically distributed, have support  $[0, 1]^d$  and a density  $f_t$  that is bounded from below and above on  $[0, 1]^d$ , uniformly over  $t = 1, \dots, T$ ; (c) we assume that  $E[\varepsilon'_{t,j}] = 0$  for  $1 \leq t \leq T, 1 \leq j \leq J$ , and for  $c > 0$  small enough,  $\sup_{1 \leq t \leq T, 1 \leq j \leq J} E[\exp(c(\varepsilon'_{t,j})^2)] < \infty$ ; (d) the vector of functions  $m = (m_1, \dots, m_L)^\top$  can be approximated by  $\Psi_k$ , i.e.

$$\delta_K \stackrel{\text{def}}{=} \sup_{x \in [0,1]^d} \inf_{A \in \mathbb{R}^{L \times K}} \|m(x) - A\Psi(x)\| \rightarrow 0$$

as  $K \rightarrow \infty$ , and we denote  $A$  that fulfils  $\sup_{x \in [0,1]^d} \|m(x) - A\Psi(x)\| \leq 2\delta_K$  by  $A^*$ ; (e) there exist constants  $0 < C_L < C_U < \infty$  such that all eigenvalues of the matrix  $T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^\top$  lie in the interval  $[C_L, C_U]$  with probability tending to one; (f) for all  $\beta$  and  $A$  ( $\beta^\top = \Gamma A$ ) in (A.1), with probability tending to one, we have

$$\sup_{x \in [0,1]^d} \max_{1 \leq t \leq T} \|Z_{0,t}^\top A\Psi(x)\| \leq M_T,$$

where the constant  $M_T$  satisfies  $\max_{1 \leq t \leq T} \|Z_{0,t}\| \leq M_T/C_m$  for a constant  $C_m$  such that  $\sup_{x \in [0,1]^d} \|m(x)\| < C_m$ ; (g) it holds that  $\rho^2 = (K + T)M_T^2 \log(JTM_T)/(JT) \rightarrow 0$ , and the dimension  $L$  is fixed.

Assumption 3.3(f) and the additional bound  $M_T$  in the minimization are introduced purely for technical reasons. They are similar to the assumption that  $V_t$  is upper bounded in Assumption 3.2 by noticing  $V_t = K^{-1/2} \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj} \varepsilon_{tj} U_{tr}$  and  $\varepsilon_t = Z_{0,t}^\top A^* \Psi_t + \varepsilon'_t$ . Recall that given  $\beta$ , the number of parameters still needing to be estimated equals  $KT$  ( $\{Z_{0,t}\}_{t=1}^T$ ) and  $KL$  ( $A$ ) (given  $\beta$ , if  $A$  is fixed,  $\Gamma$  is also fixed). Because  $L$  is fixed, Assumption 3.3(g) basically requires that, neglecting the factors  $M_T^2 \log(JTM_T)$ , the number of parameters grows slower than the number of observations  $JT$ .

**THEOREM 3.3.** *Suppose that model (2.1), all assumptions in Theorem 3.2 and Assumption 3.3 hold. Then, we have*

$$\frac{1}{T} \sum_{1 \leq t \leq T} \|\widehat{Z}_{0,t}^\top \widehat{A} - Z_{0,t}^\top A^*\|^2 = \mathcal{O}_P(\rho^2 + \delta_K^2). \tag{3.14}$$

In the following, we discuss how statistical analysis differs if the inference of stochasticity on  $Z_{0,t}$  is based on  $\widehat{Z}_{0,t}$  instead of using the (unobserved) process  $Z_{0,t}$ . We establish theoretical properties under a strong mixing condition, which is more general than the  $\beta$ -mixing considered in Theorem 3.2. For the statement of the theorem, we need the following assumptions, which are similar to Assumptions (A9)–(A11) in Park et al. (2009).

**ASSUMPTION 3.4.** (a) (i)  $Z_{0,t}$  is a strictly stationary sequence with  $E[Z_{0,t}] = 0$ ,  $E[\|Z_{0,t}\|^\gamma] < \infty$  for some  $\gamma > 2$ ; (ii) it is  $\alpha$ -mixing with  $\sum_{i=1}^\infty \alpha(i)^{(\gamma-2)/\gamma} < \infty$ ; (iii) the matrix  $E[Z_{0,t} Z_{0,t}^\top]^\top$  has full rank; (iv) the process  $Z_{0,t}$  is independent of  $X_{11}, \dots, X_{TJ}, \varepsilon'_{11}, \dots, \varepsilon'_{TJ}$ . (b) It holds that  $(\log(KT)^2((KM_T/J)^{1/2} + T^{1/2}M_T^4J^{-2} + K^{3/2}J^{-1} + K^{4/3}J^{-2/3}T^{-1/6}) + 1)T^{1/2}(\rho^2 + \delta_K^2) = \mathcal{O}(1)$ .

Assumption 3.4(b) imposes a very weak condition on the growth of  $J, K$  and  $T$ . Suppose, for example, that  $M_T$  is of logarithmic order and that  $K$  is of order  $(JT)^{1/5}$ , then the condition requires that  $T/J^2$  times a logarithmic factor converges to zero. As remarked by Doukhan (1994), if a stochastic process is  $\beta$ -mixing, then it is also  $\alpha$ -mixing with  $2\alpha(\mathcal{A}, \mathcal{B}) \leq \beta(\mathcal{A}, \mathcal{B})$ . If the

requirement on the  $\beta$ -mixing coefficient in Theorem 3.2 is satisfied, then the requirement on the  $\alpha$ -mixing coefficient in Assumption 3.4(a) is usually satisfied.

Furthermore, note that the minimization problem (A.1) only has a unique solution in  $\beta$ , but not in  $\Gamma$  and  $A$ . If  $(\widehat{Z}_{0,t}, \widehat{A})$  is a minimizer, then so is  $(B^\top \widehat{Z}_{0,t}, B^{-1}A)$ , where  $B$  is an arbitrary invertible matrix. With the choice  $B = (\sum_{t=1}^T Z_{0,t} \widehat{Z}_{0,t}^\top)^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^\top$ , we obtain  $\sum_{t=1}^T Z_{0,t} (\widetilde{Z}_{0,t} - Z_{0,t})^\top = 0$ , where  $\widetilde{Z}_{0,t} \stackrel{\text{def}}{=} B^\top \widehat{Z}_{0,t}$  and  $\widetilde{A} \stackrel{\text{def}}{=} B^{-1}A$ . Without loss of generality, we can assume  $T^{-1} \sum_{s=1}^T \widehat{Z}_{0,s} = T^{-1} \sum_{s=1}^T Z_{0,s} = 0$ . Additionally, we define

$$\begin{aligned}\widetilde{Z}_{n,t} &= (T^{-1} \sum_{s=1}^T \widetilde{Z}_{0,s} \widetilde{Z}_{0,s}^\top)^{-1/2} \widetilde{Z}_{0,t}, \\ Z_{n,t} &= (T^{-1} \sum_{s=1}^T Z_{0,s} Z_{0,s}^\top)^{-1/2} Z_{0,t}.\end{aligned}$$

**THEOREM 3.4.** *Suppose that model (2.1) holds. Besides all assumptions in Theorem 3.2, also let Assumptions 3.3 and 3.4 be satisfied. Then, there exists a random matrix  $B$  specified above such that, for  $h \geq 0$ ,*

$$T^{-1} \sum_{t=1}^{T-h} \widetilde{Z}_{0,t} (\widetilde{Z}_{0,t+h} - \widetilde{Z}_{0,t})^\top - Z_{0,t} (Z_{0,t+h} - Z_{0,t})^\top = o_P(T^{-1/2})$$

and

$$T^{-1} \sum_{t=1}^{T-h} \widetilde{Z}_{n,t} \widetilde{Z}_{n,t+h}^\top - Z_{n,t} Z_{n,t+h}^\top = o_P(T^{-1/2}).$$

In Theorem 3.4, we consider the autocovariances of the estimated stochastic process  $\widehat{Z}_{0,t}$  and the (unobserved) process  $Z_{0,t}$ , and we show that these estimators differ only by second-order terms. Thus, the statistical analysis based on  $\widehat{Z}_{0,t}$  is equivalent to that based on the (unobserved) process  $Z_{0,t}$ .

#### 4. DYNAMICS OF TEMPERATURE ANALYSIS

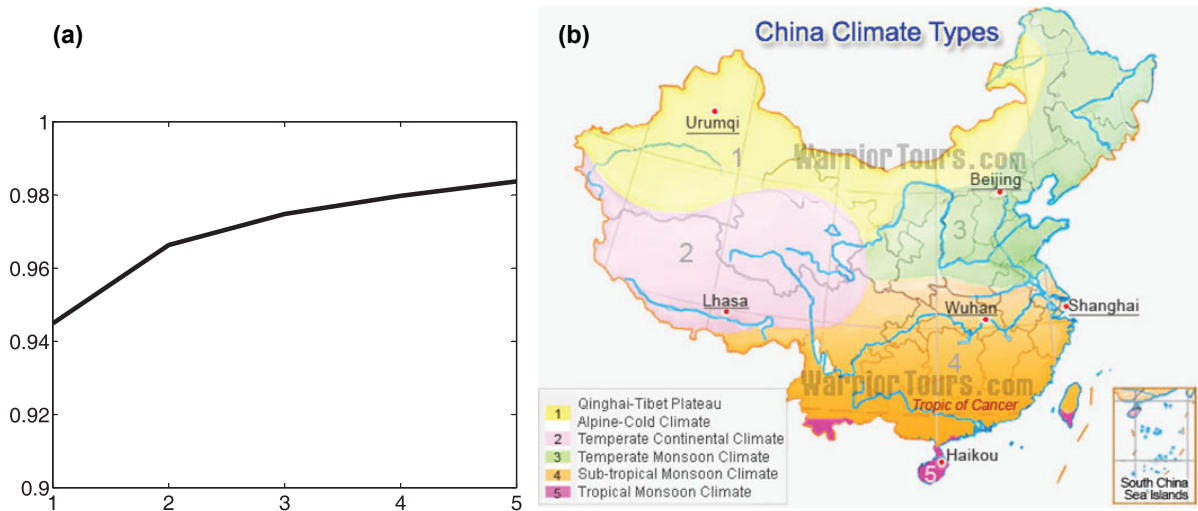
Since the first transaction in the weather derivatives market in 1971, the market has expanded rapidly. Many companies, who faced the possibility of significant declines in earnings because of abnormal weather fluctuations, decided to hedge their seasonal weather risk. Thus, weather derivative contracts have become particularly attractive. One essential task is to model the fluctuations of temperatures at many different weather stations. Thus, in this section, we present the application to the analysis of temperature dynamics by fitting the daily temperature observations provided by the CDC of the CMA; see Figure 1. To capture the upward trend, seasonal and large-period effects, similar to Racsco et al. (1991), Parton and Logan (1981) and Hedin (1991), we propose the following initial choice of time basis (rescaling factors omitted) in Table 1.

For the space basis, when we consider the relative proportion of variance explained by the first  $K$  basis (eigenvalues of the smoothed covariance operator) and the five climate types of China, as shown in Figure 3, the number of space basis  $K = 5$  is appealing. As we discuss in Appendix



**Table 1.** Initial choice of  $53 \times 3 + 16 = 175$  time basis.

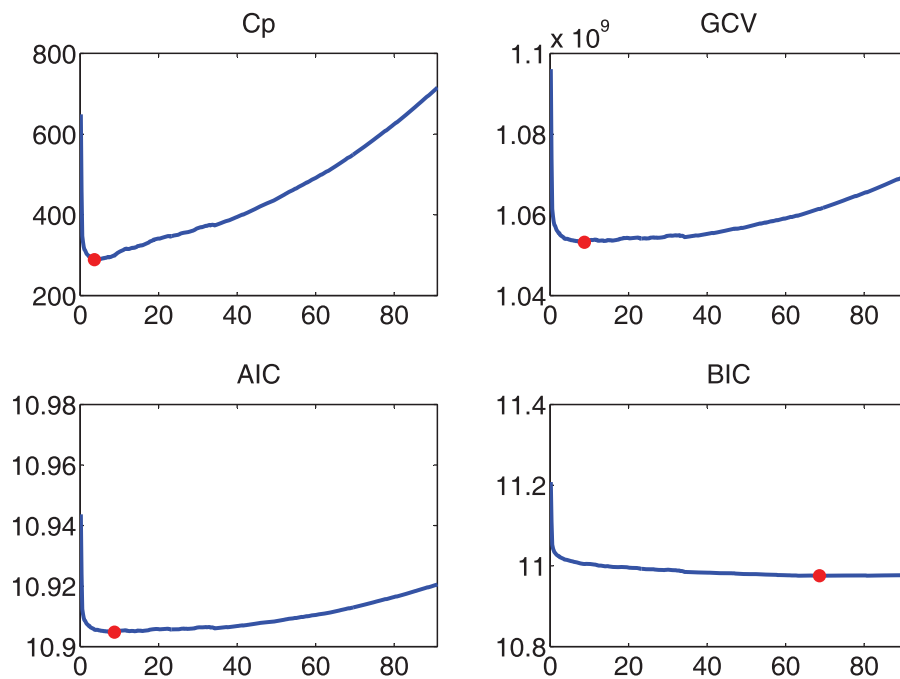
	Factors		Factors
Trend (year by year)	1	Large period	$\sin 2\pi t / (365 \times 15)$
	$t$		$\cos 2\pi t / (365 \times 15)$
	$3t^2 - 1$		$\sin 2\pi t / (365 \times 10)$
Seasonal effect	$\sin 2\pi t / 365$		$\cos 2\pi t / (365 \times 10)$
	$\cos 2\pi t / 365$		$\sin 2\pi t / (365 \times 5)$
	...		$\cos 2\pi t / (365 \times 5)$
	$\cos 10\pi t / 365$		



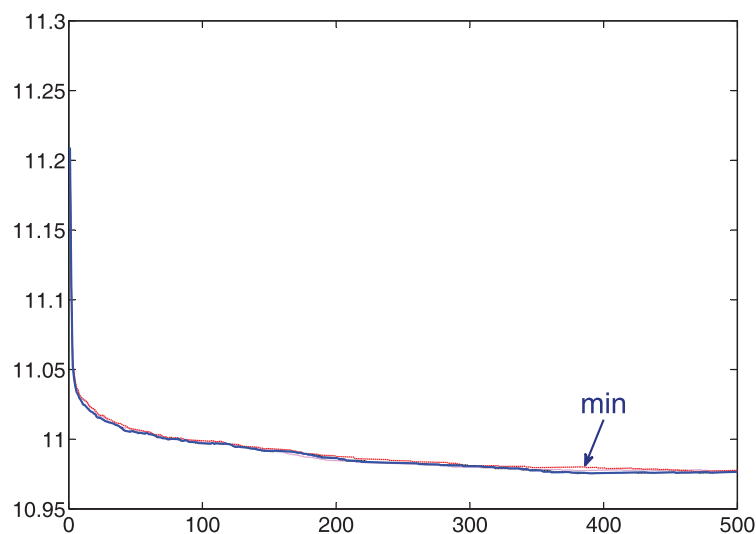
**Figure 3.** Relative proportion of variance and China’s climate types.

A, the choice of tuning parameter  $\lambda$  is crucial here. Figure 4 presents the solution path of four different selection criteria,  $C_p$ , GCV, AIC and BIC, evaluated on 500 equally spaced values of  $\lambda$ , where the minimizer is marked as the red dot. As we can see, the minimizers of  $C_p$ , GCV and AIC are significantly smaller than that of BIC, which confirms previous discussions in the literature that the AIC-type criterion (including GCV and  $C_p$ ) tends to overestimate the model size and thus overfits. Our estimate also involves the smoothing bandwidth in the smoothed functional PCA step, which, by Theorem 1 of Hall et al. (2006), should be chosen as  $\mathcal{O}(T^{-1/5})$  in order to minimize the distance between the estimates of the  $\hat{\psi}$  eigenfunctions and the corresponding true ones. Figure 5 presents the BIC solution path w.r.t. four different (by a constant factor) values of the smoothing parameter for the same 500 values of  $\lambda$  as above. As we can see, the solution path is very stable w.r.t. the choice of the smoothing parameter.

Figure 6 displays the estimated coefficients of the first factor with respect to the  $54 \times 3$  yearly polynomial time basis w.r.t.  $k = 1$  under the optimal choice of  $\lambda$  selected by the BIC criterion. The coefficients of constant, linear and quadratic terms are displayed as solid, dashed and dotted lines, respectively, and they are also coupled with the corresponding 90% confidence intervals (based on year-by-year ordinary least-squares (OLS) estimates) represented

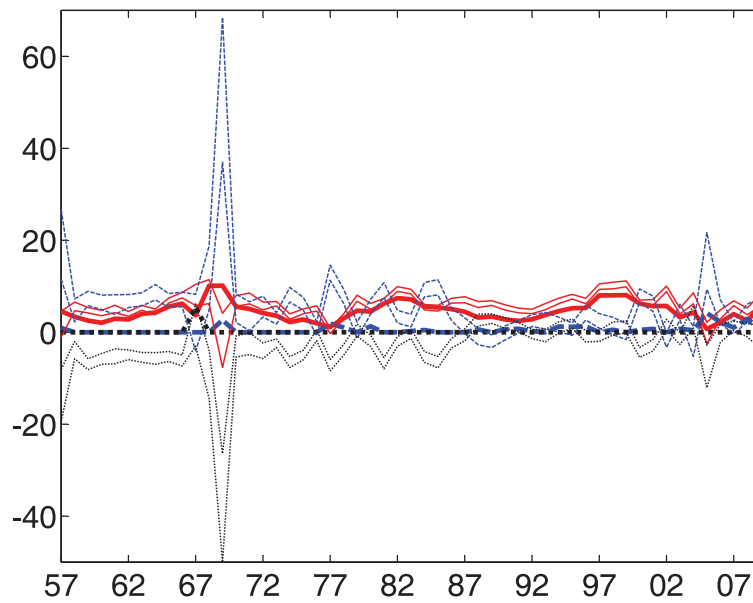


**Figure 4.** Comparison of  $C_p$ , GCV, AIC and BIC.



**Figure 5.** BIC solution path.

by the thin lines (with the same colour and style). The fact that all these coefficients are non-negative indicates that over the past 50 years, there might have been a warming effect across China. The confidence intervals are computed using OLS polynomial fitting to the year-by-year time series after removing the normal seasonality and large-period effects. We observe an unusual large positive (w.r.t. the linear term) and negative (w.r.t. the quadratic term) variation for the OLS estimates at the end of the 1960s, caused by the extreme temperatures



**Figure 6.** Estimated coefficients of the  $54 \times 3$  yearly polynomial time basis.

**Table 2.** Estimated coefficients of the five factors.

Basis	Estimates				
$\sin 2\pi t/365$	-25.4922	1.1059	2.4129	-2.6985	1.2320
$\cos 2\pi t/365$	-87.3303	1.8228	5.3358	-5.0823	1.6284
$\sin 4\pi t/365$	0.0000	0.0000	0.0000	0.0000	0.0000
$\cos 4\pi t/365$	-4.5532	0.8761	0.6752	-0.6709	0.9163
...	0.0000	...			
$\cos 10\pi t/365$	0.0000	...			
$\sin 2\pi t/(365 \times 15)$	11.7818	-0.0053	-1.4026	0.4743	-0.0214
$\cos 2\pi t/(365 \times 15)$	0.0000	...			
...	0.0000	...			
$\cos 2\pi t/(365 \times 5)$	0.0000	...			

in China at that time. By employing shrinkage techniques, we can remove this disadvantage and produce stabler estimates. The estimated coefficients of the five factors w.r.t. the 16 trigonometric functions time basis corresponding to the optimal  $\lambda$  are displayed in Table 2. It clearly indicates that the 15-year period effect, as some meteorologists claim, is related to solar activity.

Because the eigenvalues of  $\widehat{\beta}\widehat{\beta}^\top$  are (10140, 208, 118, 44, 14, 0, 0, ...) (with the first five being non-zero and the rest being zero), we choose  $L = 5$  and obtain the remaining five-dimensional random process  $\widehat{Z}_{0,t}$ , which could be further modelled by using multivariate time

series techniques. For example, if we use a VAR(1) process,  $\widehat{Z}_{0,t} = S\widehat{Z}_{0,t-1} + \varepsilon_{0,t}$ , where  $\varepsilon_{0,t}$  is a random vector, then the estimated coefficient matrix is

$$\begin{pmatrix} 0.7703 & 0.0103 & 0.0007 & 0.0015 & 0.0005 \\ -0.0552 & 0.1449 & -0.1841 & -0.0285 & 0.0003 \\ -0.3047 & -0.3419 & 0.3877 & -0.0436 & -0.0020 \\ 0.2078 & -0.1717 & -0.1337 & 0.8431 & 0.0071 \\ 0.6345 & -0.0484 & -0.0447 & 0.0184 & 0.8338 \end{pmatrix}.$$

Compared with the existing temperature modelling (pricing weather derivatives) techniques (e.g. Benth and Benth, 2005), our approach possesses the following advantages. First, based on high-dimensional time series data, it offers integrated analysis considering space (high dimensionality) and time (dynamics) parts simultaneously, while forecasting at different places other than the existing weather stations is also possible because the space basis is actually a function of the geographical location information. Second, it extracts the trend more clearly. Third, it provides theoretical justification for further inferential analysis of  $\widehat{Z}_{0,t}$  instead of  $Z_{0,t}$ .

## 5. SIMULATION STUDY

Because the simulation results about the performance of the group-Lasso estimator have been well illustrated in the literature, to evaluate the overall fitting performance of the GDSFM, we conduct a Monte Carlo experiment designed to mimic the previous empirical example.

We generate random variables  $\beta_1, \dots, \beta_{175} \in \mathbb{R}^4$  such that all coordinates are i.i.d. standard normal random variables. We randomly pick 80% of the  $\beta_r$  coefficients from  $\beta_1, \dots, \beta_{175}$  and assign them to be  $0 \in \mathbb{R}^4$ . We choose the same time basis as in Table 1 with  $p = 365$  and  $T = 19345$ . For the space part, inspired by Park et al. (2009), we consider  $d = 2$  and the following tuples of two-dimensional functions:

$$\begin{aligned} m_1(x_1, x_2) &= 1, & m_2(x_1, x_2) &= 3.46(x_1 - 0.5), \\ m_3(x_1, x_2) &= 9.45((x_1 - 0.5)^2 + (x_2 - 0.5)^2) - 1.6, \\ m_4(x_1, x_2) &= 1.41 \sin(2\pi x_2). \end{aligned}$$

These functions are chosen to be close to orthogonal. The design points  $X_{t,j}$  are independently generated from a uniform distribution on the unit square. We generate  $Y_t^\top = U_t^\top \beta^\top \Psi_t + \varepsilon_t$ ,  $t = 1, \dots, T$  with the following three types of error distributions:

1. all coordinates of  $\varepsilon_1, \dots, \varepsilon_T$  are i.i.d.  $N(0, 0.05)$  random variables;
2.  $\varepsilon_t$  are generated from a centred VAR(1) process  $\varepsilon_t = S\varepsilon_{t-1} + \eta_t$ , where  $S$  is a diagonal matrix with all diagonal entries equal to 0.4 and all entries of  $\eta_t$  are  $N(0, 0.84 \times 0.05)$  random variables (such that  $\text{Var}(\varepsilon_t)$  is still the same as that of the independent case);
3. the same as above except that all diagonal entries of  $S$  equal 0.8 (i.e. a stronger dependence level and  $\eta_t$  are  $N(0, 0.36 \times 0.05)$  random variables).

The algorithm presented in (B.3) converges fast (with a tolerance of  $10^{-3}$ ). The values of  $\beta$  are estimated by the group-Lasso technique as in (A.1) with tuning parameter  $\lambda$  selected by the BIC-type criterion, as in (A.2). After obtaining  $\hat{\beta}$ , we further estimate the stochastic process  $Z_{0,t}$  by a VAR(1) model. We take the remaining variation  $(1 - R^2)$  as a measure of the fitting

**Table 3.** Average values of  $1 - R^2$ .

	Independent	Weakly dependent	Strongly dependent
$1 - R^2$	5.30%	5.32%	5.40%

performance, where

$$1 - R^2 = \frac{\sum_{t=1}^T \|Y_t^\top - (U_t^\top \hat{\Gamma} + \hat{Z}_{0,t}^\top) \hat{A} \hat{\Psi}_t\|_2^2}{\sum_{t=1}^T \|Y_t^\top - \sum_{t=1}^T \sum_{j=1}^J Y_{t,j} / JT\|_2^2} \quad (5.1)$$

is the proportion of the remaining variation not explained by the model among total variation. We repeat this experiment 100 times and present the average values of  $1 - R^2$  in Table 3 for the independent, weakly dependent and strongly dependent cases. As we can see, when the dependence level (in  $\varepsilon_t$ ) increases, even though the remaining variation slightly increases because of the worse estimates of  $\beta$ , overall it is still relatively good.

## 6. CONCLUDING REMARKS

In this paper, we provide an integrated and yet flexible model for high-dimensional non-stationary time series that reveals both complex trends and stochastic components. When applying GDSFMs, we employ a non-parametric series expansion for both temporal and spatial components. After choosing smoothed (non-parametric) functional principal components as a space basis and extracting temporal trends utilizing time basis function selection techniques, the estimate's properties are investigated under the dependent scenario, together with the weakly cross-correlated error term. This is not built upon any specific forms of time and space basis. This enables us to explore the interplay among the degree of time dependence, high dimensionality and moderate sample size (relative to dimensionality). The presented theory is an extension to the current regularization techniques. We further justify statistical inference, e.g. estimation and classification based on the detrended low-dimensional stochastic process. Applications to the dynamic behaviour analysis of temperatures confirm its power.

## REFERENCES

- Baltagi, B. H. (2005). *Econometric Analysis of Panel Data* (3rd ed.). New York, NY: Wiley.
- Benth, F. and J. Benth (2005). Stochastic modelling of temperature variations with a view towards weather derivatives. *Applied Mathematical Finance* 12, 53–85.
- Bickel, P. J., Y. Ritov and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* 37, 1705–32.
- Bowsher, C. G. and R. Meeks (2006). High-dimensional yield curves: models and forecasting. Working Paper 2006-W12, Nuffield College, University of Oxford.
- Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.
- Cai, Z., Q. Li and J. Y. Park (2009). Functional-coefficient models for non-stationary time series data. *Journal of Econometrics* 148, 101–13.

- Campbell, J. Y. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics* 81, 27–60.
- Christiano, L., M. Eichenbaum and C. Evans (1999). Monetary policy shocks: what have we learned and to what end? In J. B. Taylor and M. Woodford (Eds.), *Handbook of Macroeconomics, Volume 1*, 65–148. Amsterdam: Elsevier.
- Diebold, F. X. and C. Li (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics* 130, 337–64.
- Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425–55.
- Doukhan, P. (1994). *Mixing: Properties and Examples*. Heidelberg: Springer.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–60.
- Fengler, M. R., W. Härdle and E. Mammen (2007). A semi-parametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics* 5, 189–218.
- Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000). The generalized dynamic-factor model: identification and estimation. *Review of Economics and Statistics* 82, 540–54.
- Forni, M., M. Hallin, M. Lippi and L. Reichlin (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association* 100, 830–40.
- Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.
- Fu, W. J. (1998). Penalized regressions: the bridge versus the Lasso. *Journal of Computational and Graphical Statistics* 7, 397–416.
- Geweke, J. (1977). The dynamic factor analysis of economic time series. In D. J. Aigner and A. S. Goldberg (Eds.), *Latent Variables in Socio-Economic Models*, 365–83. Amsterdam: North-Holland.
- Giannone, D., L. Reichlin and L. Sala (2005). Monetary policy in real time. In M. Gertler and K. Rogoff (Eds.), *NBER Macroeconomics Annual 2004, Volume 19*, 161–224. Cambridge, MA: National Bureau of Economic Research.
- Gleick, P. H. et al. (2010). Climate change and the integrity of science. *Science* 328, 689–90.
- Hall, A. and N. Hautsch (2006). Order aggressiveness and order book dynamics. *Empirical Economics* 30, 973–1005.
- Hall, P., H. G. Müller and J. L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics* 34, 1493–517.
- Hallin, M. and R. Liska (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association* 102, 603–17.
- Hedin, A. E. (1991). Extension of the msis thermosphere model into the middle and lower atmosphere. *Journal of Geophysical Research* 96, 1159–72.
- Horowitz, J. L. (2006). Testing a parametric model against a non-parametric alternative with identification through instrumental variables. *Econometrica* 74, 521–38.
- Horowitz, J. and J. Huang (2012). Penalized estimation of high-dimensional models under a generalized sparsity condition. CWP 17/12, Centre for Microdata Methods and Practice, Institute for Fiscal Studies and University College London.
- Horowitz, J. L. and S. Lee (2005). Non-parametric estimation of an additive quantile regression model. *Journal of the American Statistical Association* 100, 1238–49.
- Horowitz, J., J. Klemelä and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli* 12, 271–98.
- Hsiao, C. (1986). *Analysis of Panel Data*. Econometric Society Monographs No. 11. Cambridge: Cambridge University Press.

- Huang, J., J. L. Horowitz and F. Wei (2010). Variable selection in non-parametric additive models. *Annals of Statistics* 38, 2282–313.
- Janson, S. (2004). Large deviations for sums of partly dependent random variables. *Random Structures Algorithms* 24, 234–48.
- Lee, R. D. and L. Carter (1992). Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association* 87, 659–71.
- Leng, C., Y. Lin and G. Wahba (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica* 16, 1273–84.
- Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* 2, 90–102.
- Lounici, K., M. Pontil, A. B. Tsybakov and S. van de Geer (2009). Taking advantage of sparsity in multi-task learning. In S. Dasgupta and A. Klivans (Eds.), *Proceedings of the 22nd Conference on Learning Theory (COLT) 2009*, 73–82. Madison, WI: Omnipress.
- Mohr, P. N. C., G. Biele, L. K. Krugel, S-C. Li and H. R. Heekeren (2010). Neural foundations of risk-return trade-off in investment decisions. *NeuroImage* 49, 2556–63.
- Myšičková, A., S. Song, P. N. Mohr, H. R. Heekeren and W. K. Härdle (2013). Risk patterns and correlated brain activities. Forthcoming in *Psychometrika* (doi:10.1007/s11336-013-9352-2).
- Nelson, C. R. and A. F. Siegel (1987). Parsimonious modeling of yield curves. *Journal of Business* 60, 473–89.
- Odening, M., E. Berg and C. Turvey (2008). Management of climate risk in agriculture. *Special Issue of the Agricultural Finance Review* 68, 83–97.
- Park, B. U., E. Mammen, W. Härdle and S. Borak (2009). Time series modelling with semi-parametric factor dynamics. *Journal of the American Statistical Association* 104, 284–98.
- Parton, W. J. and J. A. Logan (1981). A model for diurnal variation in soil and air temperature. *Agricultural Meteorology* 23, 205–16.
- Racsko, P., L. Szeidl and M. Semenov (1991). A serial approach to local stochastic weather models. *Ecological Modelling* 57, 27–41.
- Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2nd ed.). Berlin: Springer.
- Sargent, T. J. and C. A. Sims (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Paper 55, Federal Reserve Bank of Minneapolis.
- Song, S. and P. Bickel (2011). Large vector auto regressions. Working Paper, University of California at Berkeley (arXiv:1106.3915).
- Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–79.
- Stock, J. H., and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* 20, 147–62.
- Stock, J. H. and M. W. Watson (2005a). An empirical comparison of methods for forecasting using many predictors. Working paper, Princeton University.
- Stock, J. H. and M. W. Watson (2005b, July). Implications of dynamic factor models for VAR analysis. Working Paper 11467, National Bureau of Economic Research. Available at <http://ideas.repec.org/p/nbr/nberwo/11467.html>.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–88.
- Wang, H. and C. Leng (2008). A note on adaptive group Lasso. *Computational Statistics and Data Analysis* 52, 5277–86.
- Wang, H., G. Li and C-L. Tsai (2007a). Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 69, 63–78.

- Wang, H., R. Li and C-L. Tsai (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94, 553–68.
- Wang, Q. and P. C. B. Phillips (2009a). Asymptotic theory for local time density estimation and non-parametric cointegrating regression. *Econometric Theory* 25, 710–38.
- Wang, Q. and P. C. B. Phillips (2009b). Structural non-parametric cointegrating regression. *Econometrica* 77, 1901–48.
- Worsley, K. C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales and A. Evans (2002). A general statistical analysis for fMRI data. *NeuroImage* 15, 1–15.
- Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics* 152, 81–92.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68, 49–67.

## APPENDIX A: ESTIMATION PROCEDURE

First, we present the estimation method.

- STEP 1. Given the pre-specified time and space basis, find significantly loaded time basis functions (i.e. coefficients  $\beta$ ) utilizing the group-Lasso technique by minimizing

$$\min_{\beta} T^{-1} \sum_{t=1}^T (Y_t^{\top} - U_t^{\top} \beta^{\top} \Psi_t)(Y_t^{\top} - U_t^{\top} \beta^{\top} \Psi_t)^{\top} + 2\lambda \|\beta\|_{2,1}. \quad (\text{A.1})$$

Here, we use  $T^{-1}$  instead of  $(JT)^{-1}$  because the space basis has been orthonormalized ( $\widehat{\Psi}_t \widehat{\Psi}_t^{\top} = I_K$ ).

- STEP 2. Split the joint matrix  $\widehat{\beta}$  into two separate coefficient matrices  $\widehat{\Gamma}$  and  $\widehat{A}$  by taking  $\widehat{\Gamma}$  as the  $L$  eigenvectors of  $\widehat{\beta} \widehat{\beta}^{\top}$  (w.r.t. the  $L$  largest eigenvalues) and  $\widehat{A} = \widehat{\Gamma}^{\top} \widehat{\beta}$ . Given  $Y_t^{\top} - U_t^{\top} \widehat{\beta}^{\top} \Psi_t$  and  $\widehat{A}$ ,  $\Psi_t$ , estimate  $Z_{0,t}$  by the OLS method.

It is worth noting that both  $\Gamma$  (and  $Z_{0,t}$ , respectively) and  $A$  are unidentifiable in model (2.1), because trivially  $\Gamma^* A^* = (\Gamma^* B)(B^{-1} A^*)$ . However, if we concentrate on prediction, the identification of  $\beta$  (as a product of  $\Gamma$  and  $A$ , as in (A.1)) is enough. Additionally, we show that for any version of  $\{Z_{0,t}\}$ , there exists a version of  $\{\widehat{Z}_{0,t}\}$  whose lagged covariances are asymptotically the same as those of  $\{Z_{0,t}\}$ .

The group-Lasso estimates depend on the tuning parameter  $\lambda$ . We implement an easily computable BIC-type criterion. The solution path is computed by evaluating some criteria on equally spaced  $\lambda$ 's between 0 and  $\lambda_{\max} = \max_r \|\sum_t \Psi_t Y_t U_{tr}\|$ . We select the  $\lambda$  that minimizes

$$\text{BIC}(\lambda) = \log \left( \sum_t \|Y_t^{\top} - U_t^{\top} \widehat{\beta}^{\top} \Psi_t\|^2 / T \right) + \log T \cdot df / T, \quad (\text{A.2})$$

$$df = \sum_r \mathbf{1}(\|\widehat{\beta}_r\| > 0) + \sum_r \frac{\|\widehat{\beta}_r\|}{\|\widehat{\beta}_{\text{OLS}}\|} (K - 1).$$

For reference purposes, we also list the formulae of the  $C_p$ , GCV and AIC criteria:

$$C_p(\lambda) = \sum_t \|Y_t^{\top} - U_t^{\top} \widehat{\beta}^{\top} \Psi_t\|^2 / \widehat{\sigma}^2 - T + 2df;$$

$$\widehat{\sigma}^2 = \sum_t \|Y_t^{\top} - U_t^{\top} \widehat{\beta}_{\text{OLS}}^{\top} \Psi_t\|^2 / (T - df);$$



$$\text{GCV}(\lambda) = \sum_t \| Y_t^\top - U_t^\top \widehat{\beta}^\top \Psi_t \|^2 / (1 - df/T)^2;$$

$$\text{AIC}(\lambda) = \log(\sum_t \| Y_t^\top - U_t^\top \widehat{\beta}^\top \Psi_t \|^2 / T) + 2df/T.$$

As pointed out by Yuan and Lin (2006) (for i.i.d. data), the performance of this approximate information criterion is generally comparable with that of the computationally much more expensive (especially for the massive data) fivefold cross-validation. More importantly, because the data here are observed in time, the order of observations is significant, and hence a simple cross-validation procedure is inappropriate in a time series context. Besides BIC, there are other parameter selection criteria, such as  $C_p$ , GCV and AIC. In terms of variable selection, Wang and Leng (2008) have found that BIC is superior to  $C_p$ . The reason for this is that when there exists a true model, AIC-type criteria (including GCV and  $C_p$ ) tend to overestimate the model size; see, e.g. Leng et al. (2006), Wang et al. (2007a) and Wang et al. (2007b). Subsequently, estimation accuracy using  $C_p$  can suffer. Wang et al. (2007b) have given a theoretical justification showing that GCV overfits the smoothly clipped absolute deviation (SCAD) method (Fan and Li, 2001). Analogous arguments also apply to the  $C_p$  methods.

### APPENDIX B: TECHNICAL PROOFS

In order to study the statistical properties of this estimator, it is useful to derive some optimality conditions for a solution of (A.1). Our implementation of group-Lasso-type estimator comes from Yuan and Lin (2006), which is an extension of the shooting algorithm of Fu (1998). As a direct consequence of the Karush–Kuhn–Tucker conditions, we have a necessary and sufficient condition for  $\widehat{\beta}$  to be a solution of (A.1):

$$T^{-1} \sum_{t=1}^T (\Psi_t(Y_t - \Psi_t^\top \widehat{\beta} U_t) U_t^\top)_r = \lambda \frac{\widehat{\beta}_r}{\|\widehat{\beta}_r\|}, \quad \text{if } \widehat{\beta}_r \neq 0; \tag{B.1}$$

$$T^{-1} \left\| \sum_{t=1}^T (\Psi_t(Y_t - \Psi_t^\top \widehat{\beta} U_t) U_t^\top)_r \right\| \leq \lambda, \quad \text{if } \widehat{\beta}_r = 0. \tag{B.2}$$

Recall that  $\Psi_t \Psi_t^\top = I_K$ . It can be easily verified that the solution to (B.1) and (B.2) is

$$\widehat{\beta}_r = (1 - \lambda / \|S_r\|)_+ S_r, \tag{B.3}$$

where  $S_r = \sum_{t=1}^T (\Psi_t(Y_t - \Psi_t^\top \widehat{\beta}_{-r} U_t) U_t^\top)_r$  with  $\widehat{\beta}_{-r} = (\widehat{\beta}_1, \dots, \widehat{\beta}_{r-1}, 0, \widehat{\beta}_{r+1}, \dots, \widehat{\beta}_R)$ . The solution to expression (A.1) can therefore be obtained by applying (B.3) to  $r = 1, \dots, R$  iteratively.

LEMMA B.1. Consider model (2.2). Assume that  $\Psi_t \Psi_t^\top = I_K$ ,  $T^{-1} \sum_{t=1}^T U_t^\top U_t / R = 1$ , and  $M(\beta^*) \leq s$ . If the random event

$$\mathcal{A} = (2T^{-1} \max_{1 \leq r \leq R} \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}^\top \varepsilon_{tj} U_{tr} \leq \lambda) \tag{B.4}$$

holds with high probability for some  $\lambda > 0$ . Then, for any solution  $\widehat{\beta}$  of problem (A.1) and  $\forall \beta$ , we have

$$T^{-1} \sum_{t=1}^T \| \Psi_t^\top (\widehat{\beta} - \beta^*) U_t \|^2 + \lambda \|\widehat{\beta} - \beta\|_{2,1}$$

$$\leq T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\beta - \beta^*) U_t\|^2 + 4\lambda \sum_{r \in \mathcal{R}(\beta)} \|\widehat{\beta}_r - \beta_r\|, \quad (\text{B.5})$$

$$T^{-1} \max_{1 \leq r \leq R} \left\| \sum_{t=1}^T (\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top)_r \right\| \leq 3\lambda/2, \quad (\text{B.6})$$

$$M(\widehat{\beta}) \leq \frac{4\phi_{\max}^2}{\lambda^{-2} T^{-2}} \sum_{t=1}^T \|(\widehat{\beta} - \beta^*) U_t\|_2^2, \quad (\text{B.7})$$

where  $\phi_{\max}$  is the maximum eigenvalue of the matrix  $\sum_{t=1}^T U_t U_t^\top / T$ .

**Proof:** The proof involves similar thoughts given in Lemma 3.1 of Lounici et al. (2009). By the definition of  $\widehat{\beta}$  as a minimizer of (A.1),  $\forall \beta$  we have

$$\begin{aligned} T^{-1} \sum_{t=1}^T \|\Psi_t^\top \widehat{\beta} U_t - Y_t\|^2 + 2\lambda \sum_{r=1}^R \|\widehat{\beta}_r\| \\ \leq T^{-1} \sum_{t=1}^T \|\Psi_t^\top \beta U_t - Y_t\|^2 + 2\lambda \sum_{r=1}^R \|\beta_r\|, \end{aligned} \quad (\text{B.8})$$

which, using  $Y_t = \Psi_t^\top \beta^* U_t + \varepsilon_t$ , is equivalent to

$$\begin{aligned} T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\widehat{\beta} - \beta^*) U_t\|^2 &\leq T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\beta - \beta^*) U_t\|^2 \\ &+ 2T^{-1} \sum_{t=1}^T \varepsilon_t^\top \Psi_t^\top (\widehat{\beta} - \beta) U_t + 2\lambda \sum_{r=1}^R (\|\beta_r\| - \|\widehat{\beta}_r\|). \end{aligned} \quad (\text{B.9})$$

Using the Hölder inequality, we have

$$2T^{-1} \sum_{t=1}^T \varepsilon_t^\top \Psi_t^\top (\widehat{\beta} - \beta) U_t \leq 2T^{-1} \sum_{t=1}^T \|\Psi_t \varepsilon_t U_t^\top\|_{2,\infty} \|\widehat{\beta} - \beta\|_{2,1}, \quad (\text{B.10})$$

where  $\|\sum_{t=1}^T \Psi_t \varepsilon_t U_t^\top\|_{2,\infty} \leq \max_{1 \leq r \leq R} \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}^\top \varepsilon_{tj} U_{tr}$ .

If the random event

$$\mathcal{A} = (2T^{-1} \max_{1 \leq r \leq R} \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}^\top \varepsilon_{tj} U_{tr} \leq \lambda) \quad (\text{B.11})$$

holds with high probability for some  $\lambda > 0$ , which we specify afterwards, then it follows from (B.9) and (B.10), on the event  $\mathcal{A}$ , that

$$\begin{aligned} T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\widehat{\beta} - \beta^*) U_t\|^2 + \lambda \sum_{r=1}^R \|\widehat{\beta}_r - \beta_r\| \\ \leq T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\beta - \beta^*) U_t\|^2 + 2\lambda \sum_{r=1}^R (\|\widehat{\beta}_r - \beta_r\| + \|\beta_r\| - \|\widehat{\beta}_r\|) \end{aligned}$$

$$\begin{aligned}
&\leq T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\beta - \beta^*) U_t\|^2 + 2\lambda \sum_{r \in \mathcal{R}(\beta)} (\|\widehat{\beta}_r - \beta_r\| + \|\beta_r\| - \|\widehat{\beta}_r\|) \\
&\quad + 2\lambda \sum_{r \in \mathcal{R}^c(\beta)} (\|\widehat{\beta}_r - \beta_r\| + \|\beta_r\| - \|\widehat{\beta}_r\|) \\
&\leq T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\beta - \beta^*) U_t\|^2 + 4\lambda \sum_{r \in \mathcal{R}(\beta)} \|\widehat{\beta}_r - \beta_r\|. \tag{B.12}
\end{aligned}$$

This proves (B.5).

To prove (B.4), we use (B.1) and (B.2), which yield the inequality

$$T^{-1} \max_{1 \leq r \leq R} \left\| \sum_{t=1}^T (\Psi_t (Y_t - \Psi_t^\top \widehat{\beta} U_t) U_t^\top)_r \right\| \leq \lambda. \tag{B.13}$$

Then

$$\begin{aligned}
&T^{-1} \left\| \sum_{t=1}^T (\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top)_r \right\| \\
&\leq T^{-1} \left\| \sum_{t=1}^T (\Psi_t (\Psi_t^\top \widehat{\beta} U_t - Y_t) U_t^\top)_r \right\| + T^{-1} \left\| \sum_{t=1}^T (\Psi_t \varepsilon_t U_t^\top)_r \right\|, \tag{B.14}
\end{aligned}$$

where we use  $Y_t = \Psi_t^\top \beta^* U_t + \varepsilon_t$  and the triangle inequality. Then, the bound (B.4) follows by combining (B.14) with (B.13) and using the definition of the event  $\mathcal{A}$ .

Finally, we show (B.7). First, observe that

$$\sum_{t=1}^T \Psi_t (Y_t - \Psi_t^\top \beta^* U_t) U_t^\top = \sum_{t=1}^T \Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top + \sum_{t=1}^T \Psi_t \varepsilon_t U_t^\top.$$

On the event  $\mathcal{A}$ , utilizing (B.1) and the triangle inequality, we have

$$T^{-1} \left\| \sum_{t=1}^T (\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top)_r \right\| \geq \lambda/2, \quad \text{if } \widehat{\beta}_r \neq 0.$$

The following arguments yield the bound (B.7) on the number of non-zero rows of  $\widehat{\beta}^\top$ :

$$\begin{aligned}
M(\widehat{\beta}) &\leq \frac{4}{\lambda^2 T^2} \sum_{r \in \mathcal{R}(\widehat{\beta})} \left\| \sum_{t=1}^T (\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top)_r \right\|^2 \\
&\leq \frac{4}{\lambda^2 T^2} \sum_{r=1}^R \left\| \sum_{t=1}^T (\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top)_r \right\|^2 \\
&= \frac{4}{\lambda^2 T^2} \left\| \sum_{t=1}^T (\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top) \right\|_2^2 \\
&\leq \frac{4\phi_{\max}^2}{\lambda^2 T} \sum_{t=1}^T \|(\widehat{\beta} - \beta^*) U_t\|_2^2.
\end{aligned}$$

Here, we use the fact that  $\Psi_t \Psi_t^\top = I_K$  and  $\phi_{\max}$  is the maximum eigenvalue of the matrix  $\sum_{t=1}^T U_t U_t^\top / T$ .  $\square$

**Proof of Theorem 3.1:** We proceed along the lines of Theorem 6.2 of Bickel et al. (2009) and Theorem 3.1 of Lounici et al. (2009). Let  $\mathcal{R} = \mathcal{R}(\beta^*) = \{r : \beta_r^* \neq 0\}$ .

Using inequality (B.5) in Lemma B.1 with  $\beta = \beta^*$ , on the event  $\mathcal{A}$  defined in (3.1), we have

$$T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\hat{\beta} - \beta^*) U_t\|^2 \leq 4\lambda \sum_{r \in \mathcal{R}} \|\hat{\beta}_r - \beta_r^*\| \leq 4\lambda \sqrt{s} \|(\hat{\beta} - \beta^*)_{\mathcal{R}}\|. \quad (\text{B.15})$$

Moreover, by the same inequality, on the event  $\mathcal{A}$ , we have  $\sum_{r=1}^R \|\hat{\beta}_r - \beta_r^*\| \leq 4 \sum_{r \in \mathcal{R}} \|\hat{\beta}_r - \beta_r^*\|$ , which implies that  $\sum_{r \in \mathcal{R}^c} \|\hat{\beta}_r - \beta_r^*\| \leq 3 \sum_{r \in \mathcal{R}} \|\hat{\beta}_r - \beta_r^*\|$ . Thus, by Assumption 3.1 with  $\Delta = (\hat{\beta} - \beta^*)$ ,

$$\|(\hat{\beta} - \beta^*)_{\mathcal{R}}\|^2 \leq \sum_{t=1}^T \|\Psi_t^\top (\hat{\beta} - \beta^*) U_t\|^2 / (\kappa^2 T). \quad (\text{B.16})$$

Now,  $T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\hat{\beta} - \beta^*) U_t\|^2 \leq 16s\lambda^2\kappa^{-2}$  (3.2) follows from (B.15) and (B.16).

Inequality (3.3) follows by noting that

$$K^{-1/2} \sum_{r=1}^R \|\hat{\beta}_r - \beta_r^*\| \leq 4K^{-1/2} \sum_{r \in \mathcal{R}} \|\hat{\beta}_r - \beta_r^*\| \leq 4K^{-1/2} \sqrt{s} \|(\hat{\beta} - \beta^*)_{\mathcal{R}}\| \leq 16s\lambda\kappa^{-2} K^{-1/2}, \quad (\text{B.17})$$

and then using (3.2). Inequality (3.4) follows from (B.7) and (3.2).  $\square$

**Definition of  $\beta$ -mixing:** Following Doukhan (1994), let  $(\Omega, \mathcal{F}, P)$  be a probability space and let  $\mathcal{A}$  and  $\mathcal{B}$  be two sub- $\sigma$  algebras of  $\mathcal{F}$ . Various measures of dependence between  $\mathcal{A}$  and  $\mathcal{B}$  have been defined as

$$\beta(\mathcal{A}, \mathcal{B}) = \sup \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J |P(A_i \cap B_j) - P(A_i)P(B_j)|, \quad (\text{B.18})$$

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup |P(A \cap B) - P(A)P(B)|, \quad A \in \mathcal{A}, B \in \mathcal{B}, \quad (\text{B.19})$$

where the supremum is taken over all pairs of (finite) partitions  $\{A_1, \dots, A_I\}$  and  $\{B_1, \dots, B_J\}$  of  $\Omega$  such that  $A_i \in \mathcal{A}$  for each  $i$  and  $B_j \in \mathcal{B}$  for each  $j$ . Now suppose  $\{V_t\}_{t \in \mathcal{T}}$  is a (not necessarily stationary) sequence of random variables. For  $-\infty \leq i \leq j \leq \infty$ , define the  $\sigma$ -field  $\sigma_i^j = \sigma(V_t, i \leq t \leq j, t \in \mathcal{T})$ . For each  $a \geq 1$ , define the following dependence coefficients:

$$\beta(a) = \sup_{t \in \mathcal{T}} \beta(\sigma_{-\infty}^t, \sigma_{t+a}^\infty), \quad \alpha(a) = \sup_{t \in \mathcal{T}} \alpha(\sigma_{-\infty}^t, \sigma_{t+a}^\infty).$$

In the special case where the sequence  $\{V_t\}_{t \in \mathcal{T}}$  is strictly stationary, they simply become

$$\beta(a) = \beta(\sigma_{-\infty}^t, \sigma_{t+a}^\infty), \quad \alpha(a) = \alpha(\sigma_{-\infty}^t, \sigma_{t+a}^\infty).$$

A stochastic process is said to be  $\beta$ -mixing (or  $\alpha$ -mixing) if  $\beta(a) \rightarrow 0$  (or  $\alpha(a) \rightarrow 0$ ) as  $a \rightarrow \infty$ . By definition, when  $\sigma_{-\infty}^t$  and  $\sigma_{t+a}^\infty$  are independent of each other,  $\beta(a) = 0$ ; the closer  $\beta(a)$  gets to 0, the more independent the time series is.

**Proof of Theorem 3.2:** The proofs of this theorem are similar to those of Theorem 3.1 up to a specification of the bound on  $P(\mathcal{A}^c)$  in Lemma B.1. Consider the event

$$\mathcal{A} = (2T^{-1} \max_{1 \leq r \leq R} \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}^\top \varepsilon_{tj} U_{tr} \leq \lambda).$$

Observe that

$$\begin{aligned}
 P(\mathcal{A}^c) &\leq RP\left(\sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj} \varepsilon_{tj} U_{tr} K^{-1/2} > 2^{-1} \lambda T K^{-1/2}\right) \\
 &\stackrel{\text{def}}{=} RP\left(\sum_{t=1}^T V_t > 2^{-1} \lambda T K^{-1/2}\right).
 \end{aligned}$$

Because Assumption 3.2 holds, applying the Bernstein-type inequality for  $\beta$ -mixing random variables  $\{V_t\}_{t=1}^T$  (Theorem 4 of Doukhan, 1994, p. 36) yields that  $\forall \varepsilon > 0$  and  $\forall 0 < q \leq 1$ ,

$$\begin{aligned}
 P\left(\sum_{t=1}^T V_t \geq 2^{-1} \lambda T K^{-1/2}\right) &\leq 2 \exp\left(-\underbrace{\frac{(1-\varepsilon)3(1+\varepsilon^2/4)\lambda^2 T K^{-1}}{4(6(1+\varepsilon^2/4)\sigma^2 + qC''\lambda T K^{-1/2})}}_{\stackrel{\text{def}}{=} T_1}\right) \\
 &\quad + \underbrace{\frac{(1+\varepsilon^2/4)\beta((qT\varepsilon^2/(4+\varepsilon^2)) - 1)}{q}}_{\stackrel{\text{def}}{=} T_2}.
 \end{aligned}$$

To make  $T_1 \leq R^{-(1+\delta')}$ ,  $\delta' > 0$  and  $T_2 \leq R^{-(1+\delta')}$ , we choose

$$\lambda = \sqrt{\frac{16 \log R^{1+\delta'} K \sigma^2}{T(1-\varepsilon)}}, \quad qC''\lambda T K^{-1/2} = 6(1+\varepsilon^2/4)\sigma^2$$

and

$$\beta((qT\varepsilon^2/(4+\varepsilon^2)) - 1) \leq qR^{-(1+\delta')}/(1+\varepsilon^2/4) = (24\sigma(1-\varepsilon)^{1/2}(R^{1+\delta'} \sqrt{\log R^{1+\delta'} T C''})^{-1}),$$

with  $qT\varepsilon^2/(4+\varepsilon^2) = (3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2} C''^{-1} \log R^{-(1+\delta')/2}$ . Then, we have

$$P(\mathcal{A}^c) \leq RP\left(\sum_{t=1}^T V_t > \lambda T / K\right) \leq 3R^{-\delta'}. \quad \square$$

**Proof of Corollary 3.1:** To prove this corollary, we need to show that Assumption 3.2 is satisfied, i.e. for an  $m$ -dependent process with order  $k$ ,  $\sigma^2$  in Assumption 3.2 is equal to  $2k\sigma_0^2$ . For simplicity, we assume that  $n = 1$  and that  $m$  is divisible by  $2k$ . Then,

$$\begin{aligned}
 E\left[\sum_{i=1}^m V_i\right]^2 &= E\left[\sum_{i=1}^k V_i + \sum_{i=k+1}^{2k} V_i + \dots + \sum_{i=m-k}^m V_i\right]^2 \\
 &= E\left[\underbrace{\sum_{j=0}^{m/2k-1} \sum_{i=2jk+1}^{2jk+k} V_i}_{\stackrel{\text{def}}{=} C} + \underbrace{\sum_{j=0}^{m/2k-1} \sum_{i=2jk+k+1}^{2(j+1)k+k} V_i}_{\stackrel{\text{def}}{=} D}\right]^2 \\
 &\leq 2E[C^2] + 2E[D^2].
 \end{aligned}$$

Because for  $j = 0, \dots, m/2k - 1$ ,  $\sum_{i=2jk+1}^{2jk+k} V_i$  are independent of each other by the definition of  $V_i$  and the same argument holds for  $\sum_{i=2jk+k+1}^{2(j+1)k+k} V_i$ , we have

$$\begin{aligned} 2E[C^2] + 2E[D^2] &= 2 \sum_{j=0}^{m/2k-1} E\left[\sum_{i=2jk+1}^{2jk+k} V_i\right]^2 + 2 \sum_{j=0}^{m/2k-1} E\left[\sum_{i=2jk+k+1}^{2(j+1)k+k} V_i\right]^2 \\ &\leq m/kk^2\sigma_0^2 + m/kk^2\sigma_0^2 = 2mk\sigma_0^2. \end{aligned} \quad \square$$

**Proof of Theorem 3.3:** Similar to  $\widehat{Y}_t^\top \stackrel{\text{def}}{=} Y_t^\top - U_t^\top \widehat{\beta} \Psi_t$ , define  $\widetilde{Y}_t^\top \stackrel{\text{def}}{=} Y_t^\top - U_t^\top \beta^* \Psi_t$  with the corresponding estimate  $\widetilde{Z}_{0,t}$ . Thus,

$$\frac{1}{T} \sum_{1 \leq t \leq T} \|\widehat{Z}_{0,t}^\top \widehat{A} - Z_{0,t}^\top A^*\|^2 \leq \frac{1}{T} \sum_{1 \leq t \leq T} \|\widehat{Z}_{0,t}^\top \widehat{A} - \widetilde{Z}_{0,t}^\top \widehat{A}\|^2 + \frac{1}{T} \sum_{1 \leq t \leq T} \|\widetilde{Z}_{0,t}^\top \widehat{A} - Z_{0,t}^\top A^*\|^2,$$

where the second term is bounded by  $\mathcal{O}_P(\rho^2 + \delta_K^2)$  by Theorem 2 of Park et al. (2009). For the first term, because

$$\begin{aligned} \widehat{Z}_{0,t} &= (\widehat{A} \Psi_t \Psi_t^\top \widehat{A}^\top)^{-1} \widehat{A} \Psi_t \widehat{Y}_t, \\ \widetilde{Z}_{0,t} &= (\widehat{A} \Psi_t \Psi_t^\top \widehat{A}^\top)^{-1} \widehat{A} \Psi_t \widetilde{Y}_t, \\ \widetilde{Z}_{0,t} - \widehat{Z}_{0,t} &= (\widehat{A} \Psi_t \Psi_t^\top \widehat{A}^\top)^{-1} \widehat{A} \Psi_t (\Psi_t^\top (\widehat{\beta} - \beta^*) U_t), \end{aligned}$$

Theorem 3.2 tells us that  $T^{-1} \sum_{t=1}^T \|\Psi_t^\top (\widehat{\beta} - \beta^*) U_t\|^2$  is bounded by  $\mathcal{O}(T^{-1})$ . From the definitions of  $\rho^2$  and  $\delta_K$ , we know that the first term is dominated by the second term.  $\square$

**Proof of Theorem 3.4:** The proof shares ideas with Park et al. (2009). We prove the first equation of the theorem for  $h \neq 0$ . The second equation follows from the first. We start by proving that the matrix  $T^{-1} \sum_{t=1}^T Z_{0,t} \widehat{Z}_{0,t}^\top$  is invertible. Suppose that the assertion is not true, then we can choose a random vector  $e$  such that  $\|e\| = 1$  and  $e^\top \sum_{t=1}^T Z_{0,t} \widehat{Z}_{0,t}^\top = 0$ . Note that

$$\begin{aligned} &\|T^{-1} \sum_{t=1}^T Z_{0,t} \widehat{Z}_{0,t}^\top \widehat{A} - T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^\top A^*\| \\ &\leq T^{-1} \sum_{t=1}^T \|Z_{0,t} (\widehat{Z}_{0,t}^\top \widehat{A} - Z_{0,t}^\top A^*)\| \\ &\leq (T^{-1} \sum_{t=1}^T \|Z_{0,t}\|^2)^{1/2} (T^{-1} \sum_{t=1}^T \|\widehat{Z}_{0,t}^\top \widehat{A} - Z_{0,t}^\top A^*\|^2)^{1/2} \\ &= \mathcal{O}_P(\rho + \delta_K), \end{aligned} \quad (\text{B.20})$$

because of Assumption 3.3(e) and Theorem 3.3. Thus, with  $f = T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^\top e$ , we obtain

$$\begin{aligned} \|f^\top m\| &= \|f^\top (A^* \Psi)\| + \mathcal{O}_P(\delta_K) \\ &= \|e^\top T^{-1} \sum_{t=1}^T Z_{0,t} Z_t^\top \widehat{A} \Psi\| + \mathcal{O}_P(\rho + \delta_K) \\ &= \mathcal{O}_P(\rho + \delta_K). \end{aligned}$$

This implies that  $m_1, \dots, m_L$  are linearly dependent, contradicting the construction that all space basis are independent.

Note that  $\tilde{Z}_{0,t} = B^\top \widehat{Z}_{0,t}$  and  $\tilde{A} = B^{-1}A$ . With (B.20) this gives

$$\begin{aligned} \|\tilde{A} - A^*\| &= \left\| T^{-1} \sum_{t=1}^T Z_{0,t} Z_t^\top (\tilde{A} - A^*) \right\|_{\mathcal{O}_P(1)} \\ &= \left\| T^{-1} \sum_{t=1}^T Z_{0,t} \tilde{Z}_{0,t}^\top \tilde{A} - T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^\top A^* \right\|_{\mathcal{O}_P(1)} \\ &= \mathcal{O}_P(\rho + \delta_K). \end{aligned} \tag{B.21}$$

From Assumptions 3.3(d), (B.21) and Theorem 3.3, we obtain

$$\begin{aligned} T^{-1} \sum_{t=1}^T \|\tilde{Z}_t^\top - Z_{0,t}\|^2 &= T^{-1} \sum_{t=1}^T \|\tilde{Z}_t^\top (m_1, \dots, m_L)^\top - Z_{0,t}^\top (m_1, \dots, m_L)^\top\|^2_{\mathcal{O}_P(1)} \\ &= T^{-1} \sum_{t=1}^T \|\tilde{Z}_t^\top A^* - \tilde{Z}_t^\top \tilde{A}\|^2_{\mathcal{O}_P(1)} \\ &\quad + T^{-1} \sum_{t=1}^T \|\tilde{Z}_t^\top \tilde{A} - Z_{0,t}^\top A^*\|^2_{\mathcal{O}_P(1)} + \mathcal{O}_P(\delta_K^2) \\ &\leq T^{-1} \sum_{t=1}^T \|\tilde{Z}_{0,t} - Z_{0,t}\|^2 \|\tilde{A} - A^*\|^2_{\mathcal{O}_P(1)} \\ &\quad + T^{-1} \sum_{t=1}^T \|Z_{0,t}\|^2 \|\tilde{A} - A^*\|^2_{\mathcal{O}_P(1)} \\ &\quad + T^{-1} \sum_{t=1}^T \|\tilde{Z}_t^\top \tilde{A} - Z_{0,t}^\top A^*\|^2_{\mathcal{O}_P(1)} + \mathcal{O}_P(\delta_K^2) \\ &= \mathcal{O}_P(\rho^2 + \delta_K^2). \end{aligned} \tag{B.22}$$

We show that for  $h \neq 0$ ,

$$T^{-1} \sum_{t=h+1}^T \left( (\tilde{Z}_{0,t+h} - Z_{0,t+h}) - (\tilde{Z}_{0,t} - Z_{0,t}) \right) Z_{0,t}^\top = \mathcal{O}_P(T^{-1/2}). \tag{B.23}$$

This implies the first statement of Theorem 3.4 because by (B.22),

$$T^{-1} \sum_{t=-h+1}^T (\tilde{Z}_{0,t} - Z_{0,t})(\tilde{Z}_{0,t+h} - Z_{0,t+h}) = \mathcal{O}_P(b^2) = \mathcal{O}_P(T^{-1/2}).$$

To prove (B.23), define

$$\tilde{\mathcal{S}}_{t,Z} = J^{-1} \sum_{j=1}^J \tilde{A} \Psi(X_{t,j}) \Psi(X_{t,j})^\top \tilde{A}^\top,$$

$$\begin{aligned}
S_{t,Z} &= A^* E[\Psi(X_{t,j})\Psi(X_{t,j})^\top] A^{*\top}, \\
\tilde{S}_\alpha &= (JT)^{-1} \sum_{t=1}^T \sum_{j=1}^J (\Psi(X_{t,j}) \otimes \tilde{Z}_{0,t})(\Psi(X_{t,j}) \otimes \tilde{Z}_{0,t})^\top, \\
S_\alpha &= T^{-1} \sum_{t=1}^T E[(\Psi(X_{t,j}) \otimes Z_{0,t})(\Psi(X_{t,j}) \otimes Z_{0,t})^\top | Z_{0,t}], \\
S &= J^{-1} A^* (\Psi(X_{t,j})\Psi(X_{t,j})^\top e - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top e]),
\end{aligned}$$

where  $e \in \mathbb{R}^K$  with  $\|e\| = 1$ . Let  $\tilde{a}$  be the stack form of  $\tilde{A}$ . It can be verified that

$$\tilde{Z}_{0,t} = \tilde{S}_{t,Z}^{-1} J^{-1} \sum_{j=1}^J (Y_{t,j} A \Psi(X_{t,j})), \quad (\text{B.24})$$

$$\tilde{a} = \tilde{S}_\alpha^{-1} (JT)^{-1} \sum_{t=1}^T \sum_{j=1}^J (\Psi(X_{t,j}) \otimes \tilde{Z}_{0,t}) Y_{t,j}. \quad (\text{B.25})$$

Let  $\gamma = T^{-1/2}/b$ . We argue that

$$\sup_{1 \leq t \leq T} \|\tilde{S}_{t,Z} - S_{t,Z}\| = o_P(\gamma), \quad \|\tilde{S}_\alpha - S_\alpha\| = o_P(\gamma). \quad (\text{B.26})$$

We show the first part of (B.26), and the second part can be obtained analogously. Because

$$\tilde{A} \Psi_t \Psi_t^\top \tilde{A}^\top = (\tilde{A} - A^* + A^*) (\Psi_t \Psi_t^\top - E[\Psi_t \Psi_t^\top] + E[\Psi_t \Psi_t^\top]) (\tilde{A} - A^* + A^*)^\top,$$

in order to prove the first part, it suffices to show that, uniformly for  $1 \leq t \leq T$ ,

$$J^{-1} \sum_{j=1}^J A^* (\Psi(X_{t,j})\Psi(X_{t,j})^\top - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top]) (\tilde{A} - A^*)^\top = o_P(\gamma), \quad (\text{B.27})$$

$$J^{-1} \sum_{j=1}^J (\tilde{A} - A^*) (\Psi(X_{t,j})\Psi(X_{t,j})^\top - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top]) (\tilde{A} - A^*)^\top = o_P(\gamma), \quad (\text{B.28})$$

$$J^{-1} \sum_{j=1}^J A^* (\Psi(X_{t,j})\Psi(X_{t,j})^\top - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top]) A^{*\top} = o_P(\gamma), \quad (\text{B.29})$$

$$J^{-1} \sum_{j=1}^J A^* E[\Psi(X_{t,j})\Psi(X_{t,j})^\top] (\tilde{A} - A^*)^\top = o_P(\gamma), \quad (\text{B.30})$$

$$J^{-1} \sum_{j=1}^J (\tilde{A} - A^*) E[\Psi(X_{t,j})\Psi(X_{t,j})^\top] (\tilde{A} - A^*)^\top = o_P(\gamma). \quad (\text{B.31})$$



The proof of (B.27)–(B.29) follows by simple arguments. We now show (B.30). Claim (B.31) can be shown similarly. To prove (B.30), we use the Bernstein inequality for the following sum:

$$P\left(\left|\sum_{j=1}^J W_j\right| > x\right) \leq 2 \exp\left(-\frac{1}{2} \frac{x^2}{V + Mx/3}\right). \tag{B.32}$$

For  $t$  with  $1 \leq t \leq T$ , the random variable  $W_j$  is an element of the  $L \times 1$ -matrix  $S = J^{-1}A^*\left(\Psi(X_{t,j})\Psi(X_{t,j})^\top e - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top e]\right)$ , where  $e \in \mathbb{R}^K$  with  $\|e\| = 1$ . In (B.32),  $V$  is an upper bound for the variance of  $\sum_{j=1}^J W_j$ , and  $M$  is a bound for the absolute values of  $W_j$  (i.e.  $|W_j| \leq M$  for  $1 \leq j \leq J$ , a.s.). With some constants  $C_1$  and  $C_2$  that do not depend on  $t$  and the row number, we obtain  $V \leq C_1 J^{-1}$  and  $M \leq C_2 K^{1/2} J^{-1}$ . The application of the Bernstein inequality gives that, uniformly for  $1 \leq t \leq T$  and  $e \in \mathbb{R}^K$  with  $\|e\| = 1$ , all  $L$  elements of  $S$  are of order  $\mathcal{O}_P(\gamma)$ . This completes the proof of claim (B.27).

From (B.21), (B.22) and (B.24)–(B.26), it follows that uniformly for  $1 \leq t \leq T$ ,

$$\begin{aligned} \tilde{Z}_{0,t} - Z_{0,t} &= S_{t,Z}^{-1} J^{-1} \sum_{j=1}^J \varepsilon'_{t,j} A^* \Psi(X_{t,j}) \\ &\quad + S_{t,Z}^{-1} J^{-1} \sum_{j=1}^J \varepsilon'_{t,j} (\tilde{A} - A^*) \Psi(X_{t,j}) + \mathcal{O}_P(T^{-1/2}) \\ &\stackrel{\text{def}}{=} \Delta_{t,1,Z} + \Delta_{t,2,Z} + \mathcal{O}_P(T^{-1/2}). \end{aligned} \tag{B.33}$$

To prove the theorem, it remains to show that for  $1 \leq j \leq 2$ ,

$$T^{-1} \sum_{t=-h+1}^T (\Delta_{t+h,j,Z} - \Delta_{t,j,Z}) Z_{0,t}^\top = \mathcal{O}_P(T^{-1/2}). \tag{B.34}$$

This can be checked easily for  $j = 1$ . For  $j = 2$ , it follows from  $\|\tilde{A} - A^*\| = \mathcal{O}_P(\rho + \delta_K)$  and

$$E\left[\|(JT)^{-1} \sum_{t=1}^T \sum_{j=1}^J \varepsilon'_{t,j} S_{t,Z}^{-1} \mathcal{M} \Psi(X_{t,j})\|^2\right] = \mathcal{O}(K(JT)^{-1}),$$

for any  $L \times K$  matrix  $\mathcal{M}$  with  $\|\mathcal{M}\| = 1$ . □



## Comment

Wolfgang Karl Härdle & Weining Wang

To cite this article: Wolfgang Karl Härdle & Weining Wang (2014) Comment, Journal of Business & Economic Statistics, 32:2, 173-174, DOI: [10.1080/07350015.2014.898585](https://doi.org/10.1080/07350015.2014.898585)

To link to this article: <http://dx.doi.org/10.1080/07350015.2014.898585>



Published online: 16 May 2014.



Submit your article to this journal [↗](#)



Article views: 121



View related articles [↗](#)



View Crossmark data [↗](#)

# Comment

## Wolfgang Karl HÄRDLE

Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany; Singapore Management University, 50 Stamford Road, Singapore 178899 ([haerdle@wiwi.hu-berlin.de](mailto:haerdle@wiwi.hu-berlin.de))

## Weining WANG

Center for Applied Statistics & Economics School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany ([wangwein@cms.hu-berlin.de](mailto:wangwein@cms.hu-berlin.de))

The authors are to be congratulated for a timely and important contribution. The article proposes a novel principal volatility component (PVC) technique based on a generalized kurtosis matrix in a time series context. The proposed test statistics allow deep insight into higher moments and tail behavior of multivariate time series. The article considers a weak stationary multivariate time series  $\mathbf{y}_t (k \times 1)$  with finite fourth moments, the lag  $l$  generalized kurtosis matrix is defined as

$$\gamma_l \stackrel{\text{def}}{=} \sum_i \sum_j \text{cov}^2(\mathbf{y}_t \mathbf{y}_t^\top, x_{ij,t-l}) \stackrel{\text{def}}{=} \sum_i \sum_j \gamma_{l,ij} \gamma_{l,ij}^\top, \quad (1)$$

where

$$\gamma_{l,ij} \stackrel{\text{def}}{=} \text{cov}(\mathbf{y}_t \mathbf{y}_t^\top, x_{ij,t-l}). \quad (2)$$

The PVCs are then defined as linear combinations  $m_v^\top \mathbf{y}_t$ , where the  $m_v$ 's are the  $v$ th eigenvectors of the cumulative generalized kurtosis matrix  $\Gamma_\infty$  for general multivariate GARCH-type models ( $\Gamma_m$  for ARCH( $m$ ) effects in  $\mathbf{y}_t$ ) with  $\Gamma_\infty \stackrel{\text{def}}{=} \sum_{l=1}^{\infty} \gamma_l$  ( $\Gamma_m \stackrel{\text{def}}{=} \sum_{l=1}^m \gamma_l$ ). Note that  $x_{ij,t-l}$  is a function of  $y_{i,t-l} y_{j,t-l}$ .

The kurtosis matrix indicates the correlations and cross-correlations between the current variance–covariance matrix and its lagged one, and thus would be a four-dimensional object ( $k \times k \times k \times k$ ). Nevertheless, the authors consider a  $k \times k$  generalized kurtosis matrix, which sums up all the effects of a lagged variance–covariance matrix. In some cases, one might like to look at the componentwise effects, which requires alternatives of defining a generalized kurtosis matrix. For example, one can analyze the variance–covariance matrix between  $\text{vec}(\mathbf{y}_t \mathbf{y}_t^\top)$  and  $\text{vec}(\mathbf{y}_{t-l} \mathbf{y}_{t-l}^\top)$ , whose dimension is  $k^2 \times k^2$  matrix. Moreover, to generalize the idea of impulse response functions, one can look at the matrix  $\sum_j \text{cov}^2(\mathbf{y}_t \mathbf{y}_t^\top, x_{i_0 j, t-l})$  to isolate the lagged variable  $i_0$ 's ( $i_0 = 1, \dots, k$ ) contribution.

The article employs Huber's function which is symmetric. One might by an asymmetric clip function address the well-known leverage effect, which means that negative returns increase future volatility by a larger amount than positive returns of the same magnitude. In particular, to model asymmetry in the ARCH process, for example, as in GJRARCH models introduced by Glosten, Jagannathan, and Runkle (1993). For instance, setting  $x_{ij,t-l} = y_{i,t-l}^- y_{j,t-l}^-$  may serve this propose, where  $y_{j,t-l}^- = y_{j,t-l}$  only when  $y_{j,t-l} < 0$  (negative part of  $y_{j,t-l}$ ).

The idea of PVC is a decomposition of a (mixed) moment matrix. In PCA, one considers the variance–covariance matrix,

which falls short on modeling a nonlinear and asymmetric multivariate distribution. This fact reminds us of a strand of literature on independent component analysis (ICA); see, for example, Chen, Härdle, and Spokoiny (2007); Chen et al. (2014). ICA looks for a projection that maximizes a non-Gaussianity measure. Similarly, a generalized kurtosis matrix in PVC is connected to measuring non-Gaussianity. However, kurtosis does not provide the whole picture of a distribution function, and therefore other perspectives of the conditional distribution (e.g., conditional skewness and conditional quantile) may also be of interest, see, for example, Lanne and Pentti (2007).

Another issue is possible nonstationarity in  $\mathbf{y}_t$ . In this situation, the nonstationarity can be modeled via switching parameters of a stationary model, see Härdle, Herwartz, and Spokoiny (2003). The eigenvectors  $m_v$ 's would then be time varying  $m_{vt}$ . Accordingly, at time  $t$  one can adopt local adaptive techniques (see Spokoiny, Wang, and Härdle 2013) to identify a local homogeneous interval  $[t - t_0, t]$ , in which one may apply PVC.

Once more we would like to congratulate the authors for this great advance. We are sure that this work will create a new strand of literature with implications on asset allocation: portfolio choice and factor models. If one is interested in the factors that have no ARCH effects, one can certainly employ the presented technique. The factors isolated can be used as factors in asset pricing model, taking into account of rare events as in Martin (2013).

## ACKNOWLEDGMENTS

The financial support From the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko," Humboldt-Universität zu Berlin and IRTG 1972 "High Dimensional Non Stationary Time Series" is gratefully acknowledged.

## REFERENCES

- Chen, Y., Chen, R.-B., and Härdle, W. K. (2014), "TVICA—Time Varying Independent Component Analysis and Its Application to Financial Data," *Journal of Computational Statistics and Data Analysis*, forthcoming, DOI: <http://dx.doi.org/10.1016/j.csda.2014.01.002>. [173]
- Chen, Y., Härdle, W., and Spokoiny, V. (2007), "Portfolio Value at Risk Based on Independent Component Analysis," *Journal of Computational and Applied Mathematics*, 205, 594–607. [173]
- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993), "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *The Journal of Finance*, 48, 1779–1801. [173]
- Härdle, W., Herwartz, H., and Spokoiny, V. (2003), "Time Inhomogeneous Multiple Volatility Modeling," *Journal of Financial Econometrics*, 1, 55–95. [173]
- Lanne, M., and Pentti, S. (2007), "Modeling Conditional Skewness in Stock Returns," *The European Journal of Finance*, 13, 691–704. [173]
- Martin, I. W. R. (2013), "Consumption-Based Asset Pricing With Higher Cumulants," *Review of Economic Studies*, 80, 745–773. [173]
- Spokoiny, V., Wang, W., and Härdle, W. K. (2013), "Local Quantile Regression" (with discussion), *Journal of Statistical Planning and Inference*, 143, 1109–1129. [173]

# Comment

**Michael McALEER**

Department of Quantitative Finance, College of Technology Management, National Tsing Hua University, Taiwan;  
 Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands;  
 Tinbergen Institute, The Netherlands;  
 Department of Quantitative Economics, Complutense University of Madrid, Spain

## DISCUSSION

It is a pleasure to provide some comments on the excellent and topical article by Hu and Tsay (2014).

The article extends principal component analysis (PCA) to principal volatility component analysis (PVCA), and should prove to be an invaluable addition to the existing multivariate models for dynamic covariances and correlations that are essential for sensible risk and portfolio management, including dynamic hedging.

One of the key obstacles to developing multivariate covariance and correlation models is the "curse of dimensionality," namely the number of underlying parameters to be estimated, the article is concerned with dimension reduction through the use of PCA, which is possible if there are some common volatility components in the time series.

In particular, the method searches for linear combinations of a vector time series for which there are no time-varying conditional variances or covariances, and hence no time-varying conditional correlations.

The authors extend PCA to PCVA in a clear, appealing, and practical manner. Specifically, they use a spectral analysis of a cumulative generalized kurtosis matrix to summarize the volatility dependence of multivariate time series and define the principal volatility components for dimension reduction.

The technical part of the article starts in Section 2 with a vectorization of the volatility matrix, and a connection to the BEKK model of Engle and Kroner (1995).

However, because a primary purpose of PCVA is to search for the absence of multivariate time-varying conditional heteroskedasticity in vector time series, it would have been helpful to see how PCVA might be connected to the conditional covariances arising from various specializations of BEKK (for further details, see below).

Theorem 1 assumes the existence of fourth moments of a weakly stationary vector time series, but Theorems 2 and

3 assume the existence of sixth moments. The latter two theorems beg the question as to whether the assumption is testable.

Interestingly, in the simulation study, the four simple ARCH models considered are understood to "not satisfy the moment condition of Theorems 2 and 3," with a reference to Box and Tiao (1977) that traditional PCA works well in finite samples for nonstationary time series.

However, the purported connection between PCA for nonstationary time series, on the one hand, and time-varying conditional covariances and correlations for a weakly stationary vector time series, on the other, is not entirely clear.

The empirical analysis considers a dataset of weekly log returns of seven exchange rates against the U.S. dollar from March 2000 to October 2011, giving 605 observations, which would be considered a relatively small number of observations for purposes of estimating dynamic vector covariance and correlation matrices.

The simple GARCH(1,1) model is used to estimate the conditional volatility models for the first six principal volatility components. It would have been useful to compare the GARCH estimates with the univariate asymmetric GJR and (possibly) leverage-based EGARCH alternatives.

A simple comparison is made of the PVCA results with the varying conditional correlation (VCC) model of Tse and Tsui (2002), though VCC is referred to as a "dynamic conditional correlation (DCC) model" (see Engle 2002).

As the effect of "news" in the VCC model has an estimated coefficient of 0.013 and a standard error of 0.004, it is stated

## Testing monotonicity of pricing kernels

Yuri Golubev · Wolfgang K. Härdle ·  
Roman Timofeev

Received: 6 March 2011 / Accepted: 9 January 2014 / Published online: 12 March 2014  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** The behaviour of market agents has been extensively covered in the literature. Risk averse behaviour, described by Von Neumann and Morgenstern (Theory of games and economic behavior. Princeton University Press, Princeton, 1944) via a concave utility function, is considered to be a cornerstone of classical economics. Agents prefer a fixed profit over an uncertain choice with the same expected value, however, lately there has been a lot of discussion about the empirical evidence of such risk averse behaviour. Some authors have shown that there are regions where market utility functions are locally convex. In this paper we construct a test to verify uncertainty about the concavity of agents' utility function by testing the monotonicity of empirical pricing kernels (EPKs). A monotonically decreasing EPK corresponds to a concave utility function while a not monotonically decreasing EPK means non-averse pattern on one or more intervals of the utility function. We investigate the EPKs for German DAX data for the years 2000, 2002 and 2004 and find evidence of non-concave utility functions: the null hypothesis of a monotonically decreasing pricing kernel is rejected for the data under consideration. The test is based on approximations

---

We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft and the Sonderforschungsbereich 649 "Ökonomisches Risiko". Roman Timofeev's research was supported by Deka Bank scholarship program.

---

Y. Golubev

CMI, Universite de Provence, 39, rue F. Joliot-Curie, 13453 Marseille Cedex 13, France  
e-mail: golubev@cmi.univ-mrs.fr

W. K. Härdle · R. Timofeev (✉)

Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin,  
Unter den Linden 6, 10099 Berlin, Germany  
e-mail: romant\_2000@mail.ru

W. K. Härdle

e-mail: haerdle@wiwi.hu-berlin.de

of spacings through exponential random variables. In a simulation we investigate its performance and calculate the critical values (surface).

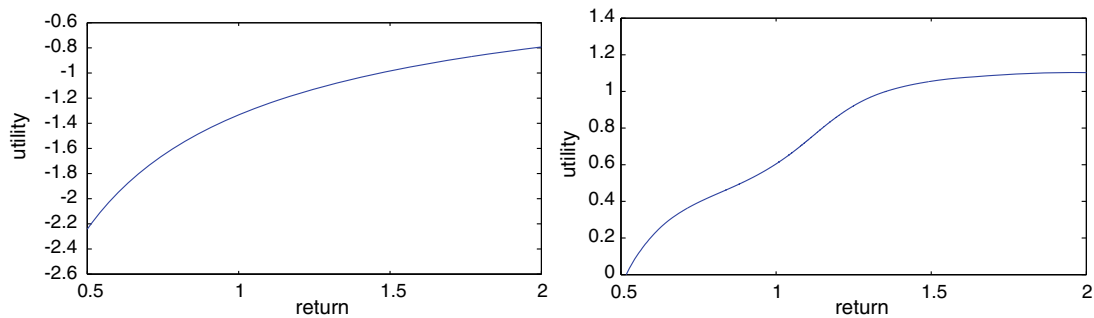
**Keywords** Monotonicity · Pricing kernel · Risk aversion

**JEL Classification** C12 · G12

## 1 Introduction

The behaviour of market agents has always been in focus in economic literature. [Von Neumann and Morgenstern \(1944\)](#) describe risk averse behaviour using concave utility functions. Agents prefer a fixed profit over an uncertain choice with the same expected value, however, lately there has been a lot of discussion about the empirical evidence of such risk averse behaviour. Recent studies by [Jackwerth \(2000\)](#) showed that there is a reference point near the initial wealth where market utility functions are convex. [Rosenberg and Engle \(2002\)](#) also observed a region of negative absolute risk aversion for the pricing kernel constructed using an orthogonal polynomial. A formal test procedure has not been given though. We want to fill this gap by testing the concavity of the utility function and thus checking the monotonicity of the corresponding empirical pricing kernel (EPK). A strictly decreasing EPK corresponds to a concave utility function which is consistent with the classic theory of risk averse behaviour, while rejection of a monotonically decreasing EPK would indicate non-riskaverse pattern of the utility function. By analysing empirical pricing kernels we can also identify on which interval or intervals the monotonicity of the EPK was rejected. Non-monotonicity of the pricing kernel for the S&P 500 was also shown in more recent research by [Constantinides et al. \(2009\)](#), [Bakshi et al. \(2010\)](#) and [Chaudhuri and Schroder \(2010\)](#).

The construction and estimation of empirical pricing kernels has been well described by [Ait-Sahalia and Lo \(2000\)](#). They analyze the concept of economic risk containing investors' preferences and statistical risk which provides information on the dynamics of the data generating process (DGP). Both these risk measures can be identified via distributions (risk neutral ( $Q$ ), physical ( $P$ )). The pricing kernel  $K$  is the Radon Nikodym derivative  $dQ/dP$  of these two measures. Economic risk is well approximated by Arrow-Debreu prices and can be estimated by the risk neutral density  $q$  obtained from the derivative market. By looking at option prices we can find out what stock prices or returns investors expect at time to maturity. Several accurate estimators of  $q$  using, for example, the [Black and Scholes \(1973\)](#) model or nonparametric estimators exist. In this paper the risk neutral density  $q$  is derived from the Heston model. Stochastic volatility models provide better results by fitting the observed volatility smile. Due to the large number of observations in the derivative option market, the risk neutral density  $q$  can be precisely estimated. Statistical risk is related to the properties of the DGP and is given by the pdf  $p$  of future prices conditional on current prices. The main difficulty for the estimation of  $p$  is, of course, that an assumption about the model for the underlying process  $S_t$  has to be made (e.g. geometric Brownian motion under the Black and Scholes model). The density  $p$  can



**Fig. 1** Classical utility function produced from Black Scholes model (*left*) and market utility function estimated from empirical pricing kernel on 06/30/2000 (*right*)

be estimated in several ways, for example, using a nonparametric diffusion model as in [Ait-Sahalia and Lo \(2000\)](#) or a GARCH model as in [Rosenberg and Engle \(2002\)](#). The historical density  $p$  can only be estimated using the past of the time series  $S_t$  and hence is influenced by model specification and data scarcity. The differences in the form of the EPK by various authors might occur due to uncertainty in the estimation of  $p$ . Therefore, we would like to test monotonicity of a pricing kernel constructed as a *ratio of estimated  $q$  and unknown  $p$* .

In [Fig. 1](#) we compare the market utility function obtained from the DAX index in the year 2000 and the utility function derived from the Black and Scholes model. In both cases the risk neutral density  $q$  was obtained via the option market: the state price density that replicates observed option prices is derived to fit the option pricing model (Black and Scholes). This setup provides us with the lognormal density. The historical density  $p$  was assumed to be lognormal for the Black and Scholes model, and nonparametric density estimation over historical time series of the DAX index was used to obtain  $q$  in case of the market utility function. The Black and Scholes model produces an increasing and concave utility function, while the market utility function has a slight bump over the region of zero returns. The aim of this paper is to find out whether observed non-concavity is significant. Obviously, the form of the utility function depends on choice of the DGP for  $S_t$ . As mentioned before, we would like to test monotonicity of the EPK for a general class of DGPs and, therefore, consider  $p$  unknown.

[Ait-Sahalia and Lo \(2000\)](#) in their paper offer another test for risk neutrality and specific preferences. Depending on the form of preferences they define  $H_0$  hypothesis as a relationship between the estimated neutral density  $q$  and the historical density  $p$ . We do not make any assumptions about the form of preferences and also consider the historical density  $p$  unknown. In our test the  $H_0$  hypothesis of a monotonically decreasing EPK is compared to a general class of functions under  $H_1$ . The test is constructed as follows: first the spacing method is used to reduce the problem to an exponential model. On the basis of this model a likelihood ratio test is applied for a fixed interval, then using intersection of tests for different intervals it is expanded to a test independent of intervals. Finally, the test statistics calculated on observed data are compared to simulated critical values, and a final decision about monotonicity is taken.

The paper is organized as follows. In Sect. 2 we introduce important notations and problem setup which is then reduced to an exponential model using the spacing method. In Sect. 3 we formulate the hypotheses, construct a likelihood test for a fixed interval  $[I, J]$  and then expand it to an independent test using the multiple testing technique. We also describe how to simulate critical values using the Monte-Carlo method. Section 4 provides empirical results on DAX data for 2000, 2002 and 2004.

## 2 Conceptual thoughts

### 2.1 Problem setup

Let  $[0, T]$  be the interval of investment in the financial market, where  $t = 0$  denotes the present time and  $t = T \in ]0, \infty[$  the time of maturity. Furthermore, it is assumed that a riskless bond and a risky asset are traded in the financial market as basic underlyings. The price process  $(B_t)_{t \in [0, T]}$  of the riskless bond is defined by

$$\frac{dB_t}{B_t} = r_t dt,$$

via a deterministic Riemannian-integrable interest process  $(r_t)_{t \in [0, T]}$ . The price process  $(S_t)_{t \in [0, T]}$  of the risky asset is assumed to be a nonnegative semimartingale with a constant  $S_0$  and continuously distributed marginals  $S_t$ ,  $t \in [0, T]$ . Furthermore, let us suppose that the financial market is arbitrage free in the sense that there exists at least one equivalent martingale measure. Throughout the paper we assume that the **risk valuation principle** is valid for a nonnegative payoff  $\psi(S_T)$ . That means that there is a Radon-Nikodym density  $\pi$  of a martingale measure such that the price of any  $\psi(S_T)$  is characterized by

$$\mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \pi \right\} = \mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \mathbf{E}_P(\pi | S_T) \right\}.$$

By factorization we may find some Borel-measurable  $K_\pi$  with  $\mathbf{E}(\pi | S_T) = K_\pi$ , so that

$$\mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \pi \right\} = \int_0^\infty e^{-\int_0^T r_t dt} \psi(S_T) K_\pi(x) p_{S_T}(x) dx,$$

where  $p_{S_T}$  denotes the density of the distribution of  $S_T$ . The last formula allows us to call  $K_\pi$  the **pricing kernel** (w.r.t.  $\pi$ ). Here the distribution  $Q_{S_T} \stackrel{\text{def}}{=} \int_{-\infty}^{S_T} K_\pi(z) p_{S_T}(z) dz$ , plays an important role. It is a continuous distribution with pdf  $q_{S_T}$  and is called the **risk neutral distribution** of  $S_T$  (w.r.t.  $\pi$ ). Since

$$\mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \pi \right\} = \mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(x) q_{S_T}(x) dx \right\}$$



holds for any nonnegative payoff  $\psi(S_T)$ , the pricing kernel  $K_\pi = \frac{q_{S_T}}{p_{S_T}}$  a.s. (w.r.t.  $P$ ).

Let us further assume that investors of the financial market are consumers whose consumptions depend on the price  $S_T$  of the stock at maturity only. Within the classical framework, where investors' preferences may be represented by expected utilities, there exists a link between the risk attitude of the investors and the pricing rule in the financial markets. It relies on the assumption of a representative agent whose indirect utility  $U\{\bar{e}(S_T)\}$  which is dependent on the aggregated market endowment  $\bar{e}(S_T)$  has expected utility representation  $U\{\bar{e}(S_T)\} = \mathbb{E}\{u(S_T)\}$  with concave Von Neumann-Morgenstern utility index  $u$ . Under further technical conditions on the investors preferences, see [Härdle et al. \(2012\)](#), [Grith et al. \(2013\)](#), and  $\bar{e}(S_T) = S_T$  there is a positive  $\beta$  such that

$$\frac{du}{dx}|_{x=S_T} = \beta K_\pi(S_T)$$

for almost any realization  $s_T$  of  $S_T$ . For a rigorous derivation we refer to [Karatzas and Shreve \(1998\)](#), sections 4.4 and 4.5. Without loss of generality consider  $q$  and  $K = K_\pi$  on a scale of regular returns  $X = \frac{S_T - S_0}{S_0}$ , where  $S_0$  is the known current price.

The concavity of utility  $U$  can be, therefore, tested by checking monotonicity of  $K$ : a strictly decreasing  $K$  corresponds to a concave utility function, while a non-monotone  $K$  would indicate a non-concave pattern. Our test idea is based on intervals  $[a, b]$ , where  $K$  is not monotonically decreasing.

Denote by  $X_{(1)}, \dots, X_{(n)}$  the order statistics related to a sequence of  $X_1, \dots, X_n$  of returns  $X$  i.e.

$$X_{(1)} \leq X_{(2)}, \dots, \leq X_{(n)}.$$

With these notations we can rephrase the monotonicity testing problem: find (if possible) integers  $I, J$  such that the sequence

$$K_k = K(X_{(k)}) = \frac{q(X_{(k)})}{p(X_{(k)})}, \quad I \leq k \leq J$$

is not monotonically decreasing.

The principal difficulty in this testing procedure is related to the fact that  $p$  is unknown and that violation of monotonicity may occur at different sub-intervals  $[a, b]$ . To solve this challenge we will use three basic ingredients:

1. the spacing method to reduce the stochastics to a simpler exponential model
2. the maximum likelihood test to check monotonicity of  $K_k$  for given  $I$  and  $J$
3. the multiple-testing procedure to find  $I$  and  $J$  on the basis of the data at hand.

## 2.2 The spacing method

Our method is based on Pyke's lemma about the distribution of order statistics, see [Pyke \(1965\)](#). It describes various ways of constructing the spacings, the differences between consecutive observations, in the context of distribution-free tests of fit. The distribution-free assumption is vital for our monotonicity test. Not assuming any form for  $p, q$  makes this test very general and allows to imply strong conclusions on the economic risk of market participants. Pyke's Lemma is based on the following thoughts.

Let  $U_1, \dots, U_n$  be i.i.d random variables with the uniform distribution on  $[0, 1]$  and  $U_0 = 0, U_{n+1} = 1$ . Then the uniform spacings associated with these random variables are defined as

$$S_k = U_{(k)} - U_{(k-1)}, \quad k = 1, \dots, n + 1,$$

where  $U_{(k)}$  are the order statistics  $0 \leq U_{(1)} \leq U_{(2)}, \dots, \leq U_{(n)} \leq 1$ .

The uniform spacings can be represented as exponential random variables proportional to their sum, [Pyke \(1965\)](#).

**Lemma 2.1** *Let  $e_1, \dots, e_{n+1}$  be i.i.d. standard exponentially distributed random variables and  $D = e_1 + e_2 + \dots + e_{n+1}$  be the sum of them. Then the joint distribution of  $\{e_k/D\}_{k=1}^{n+1}$  coincides with the distribution of the set of  $n + 1$  uniform spacings.*

*Using the fact that  $E(e_k) = 1$ , with the law of large numbers for  $D$ , i.e.  $D = n + \mathcal{O}_p(n^{-1/2})$ , we obtain the following approximation:*

$$\begin{aligned} n \{U_{(k)} - U_{(k-1)}\} &= n \cdot S_k \stackrel{\mathcal{L}}{=} n \cdot e_k/D = n \cdot e_k/n + \mathcal{O}_p(n^{-1/2}) \\ &= e_k + \mathcal{O}_p(n^{-1/2}) \approx e_k, k = 1, \dots, n + 1. \end{aligned} \quad (1)$$

We now apply (1), showing the approximation of spacings by a standard exponential random variable, to the problem of the pricing kernel. Let  $X_1, X_2, \dots, X_{n+1}$  be i.i.d. random variables (returns) with a historical density  $p(x)$ ,  $x \in \mathbb{R}^1$  and  $X_{(1)} \leq X_{(2)}, \dots, \leq X_{(n+1)}$  are the corresponding order statistics. By  $P(x)$  we denote the cdf associated with  $p(x)$ . The i.i.d. assumption might be seen as a too strong one, since log returns show volatility clustering effects. These occur though more frequently in highly sampled financial time series. In our case the frequency is low and therefore the identical marginal distribution appears to be justifiable.

The first order Taylor approximation  $P(x)$  at point  $X_{(k)}$  can be calculated using the value of the function at point  $X_{(k-1)}$ ;

$$P(X_{(k)}) \approx P(X_{(k-1)}) + P'(X_{(k-1)})\{X_{(k)} - X_{(k-1)}\}$$

Note that the spacings are of order  $\mathcal{O}_p(n^{-1})$  by Lemma 2.1.

Using the probability integral transformation we see that the random variables  $P(X_i)$  are uniformly distributed over  $(0, 1)$ . Combining first order Taylor approxima-

tion with (1) we obtain

$$e_k \approx n\{U_{(k)} - U_{(k-1)}\} = n\{P(X_{(k)}) - P(X_{(k-1)})\} \approx n \cdot p(X_{(k-1)}) \cdot \{X_{(k)} - X_{(k-1)}\}. \tag{2}$$

Equation (2) is the representation of the spacing of the historical density  $p$  in a form of exponential variables using ordered returns  $X_{(k)}$ . This way we do not make any assumptions about the distribution of  $X$ . Yet, in order to apply Pyke’s Lemma 2.1 the returns are assumed to be i.i.d. implying that in this case we deal with the unconditional density  $p$ . The test of monotonicity of the pricing kernel can now be constructed as the ratio of the risk-neutral density  $q$  and the unconditional historical density  $p$ .

Replacing  $p(x) = K^{-1}(x)q(x)$  in (2) allows to complete the test with respect to the pricing kernel  $K(x)$ :

$$n \cdot \{X_{(k)} - X_{(k-1)}\} \cdot K^{-1}(X_{(k-1)}) \cdot q(X_{(k-1)}) \approx e_{k-1} \quad k = 1, \dots, n + 1 \tag{3}$$

Let us denote for simplicity  $K_{(k-1)} = K(X_{(k-1)})$  and

$$Z_{k-1} = n \{X_{(k)} - X_{(k-1)}\} q(X_{(k-1)}), \quad k = 1, \dots, n + 1. \tag{4}$$

Thus the test problem based on (3) is to check monotonicity of  $K_{k-1}$  using:

$$Z_{k-1} \approx K_{k-1} \cdot e_{k-1}, \quad k = 1, \dots, n + 1. \tag{5}$$

Here again the approximation (5) is of order  $\mathcal{O}_p(n^{-1/2})$ .

### 3 Construction of the test

#### 3.1 Local test

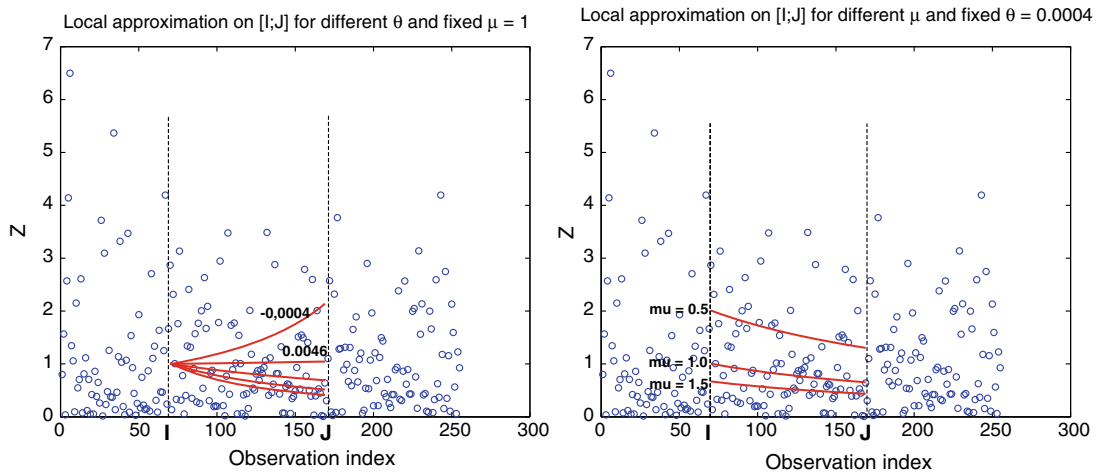
The approximation (2), (5) have been made in order to specify the stochastic fluctuation of  $Z_k$  as being approximately exponential. This will allow us to continue within a quasi likelihood framework.

For simplicity, let us first consider a fixed interval  $[I, J]$  of the sequence

$$Z_s \approx K_s e_s, \quad I \leq s \leq J. \tag{6}$$

where  $I$  and  $J$  are beginning and ending observation indexes of a selected interval. The test alternative on interval  $[I, J]$  implies that if  $K(x)$  is not decreasing, then one can find an index  $s$  that the subsequence  $K_s, I \leq s \leq J$  is increasing.

The local test is based on an inverse linear approximation of  $K(s)$ . The motivation behind this approach is rather simple: in contrast to the standard linear approximation, the inverse linear approximation results in quasi-concavity of the log-likelihood and, thus, permitting to reduce significantly the numerical complexity of the test.



**Fig. 2** Inverse linear approximation for different parameters  $\mu$  and  $\theta$

If  $K_s$ ,  $I \leq s \leq J$  is increasing, then the statistical model with the observations

$$\tilde{Z}_s = \frac{e_s}{\mu\{1 + \theta(s - I)\}}, \quad I \leq s \leq J \tag{7}$$

with parameters  $\mu$  and  $\theta$  and i.i.d. standard exponentially distributed random  $e_s$ , approximates the model (6) better with some negative  $\theta$  than with a positive one. It is important to notice that since  $Z_s$  can have only positive values,  $\theta$  and  $\mu$  are also limited.

Excluding the randomness generated by  $e_s$  by substituting it with  $E(e) = 1$  the approximation (7) takes the form presented in Fig. 2. The plots show different scenarios depending on parameters  $\mu$  and  $\theta$ , where  $\mu$  is responsible for the starting level and  $\theta$  controls the degree of the slope.

Therefore, two composite hypotheses can be formulated. Based on the observed sequence of  $Z_s$  from (6) and approximation (7) we have:

$$H_0 : \theta > 0$$

and  $K_s$ ,  $I \leq s \leq J$  is monotonically decreasing

$$H_1 : \theta \leq 0$$

and  $K_s$ ,  $I \leq s \leq J$  is not-monotonically decreasing.

The test is constructed using the maximum likelihood principle. Let  $P_{\mu,\theta}(\cdot)$  be the joint cdf and  $p_{\mu,\theta}(\cdot)$  be the joint pdf of the observations in (7). Using the fact that the  $e_s$  are i.i.d. standard exponential distributed, the corresponding log-likelihood function takes the form:

$$\log\{p_{\mu,\theta}(\tilde{Z})\} = -\mu \sum_{s=I}^J \tilde{Z}_s \{1 + \theta(s - I)\} + (J - I + 1) \log(\mu) + \sum_{s=I}^J \log\{1 + \theta(s - I)\} \tag{8}$$

Therefore, we can re-formulate the test hypotheses: accept  $H_0$  if

$$\max_{\mu, \theta > 0} \log p_{\mu, \theta}(\tilde{Z}) - \max_{\mu, \theta \leq 0} \log p_{\mu, \theta}(\tilde{Z}) \geq h_\alpha(I, J),$$

otherwise  $H_0$  is rejected.

Here the critical value  $h_\alpha(I, J)$  is computed as a root of equation

$$P \left\{ \max_{\mu, \theta > 0} \log p_{\mu, \theta}(e) - \max_{\mu, \theta \leq 0} \log p_{\mu, \theta}(e) < h_\alpha(I, J) \right\} = \alpha, \tag{9}$$

where  $\alpha$  is the type I error probability.

Now the problem is reduced to calculate the MLE's  $\hat{\mu}$  and  $\hat{\theta}$  for the observed data sequence  $\{Z_s\}$ . Fortunately, the numerical complexity of this test is not very high. First of all, the maximum in  $\mu$  of  $p_{\mu, \theta}(\cdot)$  may be computed very easily. By calculating  $\partial \log p_{\mu, \theta}(Z) / \partial \mu = 0$  we obtain the optimal value of  $\hat{\mu}$

$$\hat{\mu} = \frac{J - I + 1}{\sum_{s=I}^J Z_s \{1 + \theta(s - I)\}}$$

which results in the maximum of the log-likelihood function in  $\mu$

$$\begin{aligned} \max_{\mu} p_{\mu, \theta}(Z) &= \sum_{s=I}^J \log\{1 + \theta(s - I)\} - (J - I + 1) \log \left[ \sum_{s=I}^J Z_s \{1 + \theta(s - I)\} \right] \\ &+ (J - I + 1) \log \frac{J - I + 1}{\exp(1)} \end{aligned} \tag{10}$$

Due to quasi-concavity property, the function  $\max_{\mu} p_{\mu, \theta}$  has a maximum in  $\theta$ . In order to find the optimal value  $\hat{\theta}$  the part which contains  $\theta$  and the rest of the equation should be separated. Denote for brevity

$$L_{\theta}^{I, J}(Z) = \sum_{s=I}^J \log\{1 + \theta(s - I)\} - (J - I + 1) \log\{1 + \theta R^{I, J}(Z_s)\}, \tag{11}$$

where

$$R^{I, J} = R^{I, J}(Z) = \frac{\sum_{s=I}^J Z_s (s - I)}{\sum_{s=I}^J Z_s}. \tag{12}$$

is a random field.

By (10), it is easy to see that

$$\max_{\mu} p_{\mu, \theta}(Z) = L_{\theta}^{I, J}(Z_s) - (J - I + 1) \log \frac{J - I + 1}{\exp(1)} + (J - I + 1) \log \sum_{s=I}^J Z_s. \tag{13}$$

Since only  $L_{\theta}^{I,J}(Z)$  depends on  $\theta$ , the optimal value can be found as:

$$\hat{\theta} = \arg \max_{\theta} L_{\theta}^{I,J}(Z_s)$$

The simplest way to find the maximum of the function is to use the Newton-Raphson algorithm:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \frac{dL_{\hat{\theta}_k}^{I,J}(Z_s)/d\hat{\theta}_k}{d^2L_{\hat{\theta}_k}^{I,J}(Z_s)/d\hat{\theta}_k^2} \quad (14)$$

So, the final decision about monotonicity on interval  $[I, J]$  is based on:

$$\begin{aligned} \max_{\theta > 0} L_{\theta}^{I,J}(Z) - \max_{\theta \leq 0} L_{\theta}^{I,J}(Z) &= L_{\hat{\theta}}^{I,J}(Z) \mathbf{1}\{\hat{\theta} > 0\} - L_{\hat{\theta}}^{I,J}(Z) \mathbf{1}\{\hat{\theta} \leq 0\} \\ &= L_{\hat{\theta}}^{I,J}(Z) \operatorname{sign}(\hat{\theta}). \end{aligned}$$

With the above argument in mind, we propose the following local test on  $[I, J]$  for checking monotonicity of  $K_s$   $I \leq s \leq J$  in (6):

1. compute

$$\hat{\theta}(Z) = \arg \max_{\theta} L_{\theta}(Z)$$

with the help of the Newton-Raphson method (14),

2. accept the hypothesis that  $K_s$ ,  $I \leq s \leq J$  is decreasing if

$$L_{\hat{\theta}(Z)}^{I,J}(Z) \operatorname{sign}\{\hat{\theta}(Z)\} - h_{\alpha}(I, J) \geq 0 \quad (15)$$

otherwise reject the hypothesis.

Notice that the critical value  $h_{\alpha}(I, J)$  may be computed with the help of the Monte-Carlo method as a root of the equation

$$\mathbb{P}\left[L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\} - h_{\alpha}(I, J) < 0\right] = \alpha, \quad (16)$$

where  $e = (e_1, \dots, e_{J-I})$  is the sequence of i.i.d. standard exponential random variables.

### 3.2 Global test

The previously described approach considers each interval  $[I, J]$  separately, whereas the decision about monotonicity should be taken for all possible combinations of  $I$  and  $J$ . Therefore, the next step is to join the local tests described above in a global setup. The approach is related to a natural modification of the Bonferroni method which is also used in adaptive estimation in computing nearly optimal penalties for the empirical risk minimization method, see e.g. [Cavalier and Golubev \(2006\)](#). In

order to join the local tests, notice that if the underlying sequence is decreasing, then (15) must hold true for any  $I, J$  or equivalently

$$\min_{I,J} \left[ L_{\hat{\theta}(Z)}^{I,J}(Z) \operatorname{sign}\{\hat{\theta}(Z)\} - t_\alpha(I, J) \right] \geq 0. \tag{17}$$

Therefore we may use this relation as a test prototype. To construct the final test, it remains to redefine the critical values  $t_\alpha(I, J)$ . Obviously, we cannot stick with  $h_\alpha(I, J)$  defined by (16) because it does not control anymore the type I error probability. In fact, the critical values describe a surface  $t_\alpha(I, J)$  that must satisfy the following equation:

$$P\left(\min_{I,J} \left[ L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\} - t_\alpha(I, J) \right] < 0 \right) = \alpha. \tag{18}$$

In contrast to (15), this equation has no unique solution. Intuitively, to maximize the power of the test, i.e. the type II error probability, we should chose  $t_\alpha(I, J)$  as a “maximal” function satisfying (18). Unfortunately, the problem of computation of such a “critical surface” is extremely difficult from theoretical and numerical viewpoints. Therefore, we provide only an approximate solution of this problem. The main step in computing a nearly optimal  $t_\alpha(I, J)$  is to find out the probabilistic structure of  $L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\}$ . Notice that the stochastic part in this field is completely determined by the random field  $R^{I,J}$  given in (12). Therefore, we first focus on probabilistic properties of this field. Using Taylor expansion and the central limit theorem,  $R^{I,J}$  can be approximated as:

$$R^{I,J} \approx \frac{J - I}{2} + \sqrt{\frac{J - I}{12}} \xi \tag{19}$$

where  $\xi \sim N(0, 1)$ . Let us show the approximation for  $L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\}$  in more details. Recall the definition of  $R^{I,J}$  in (12).

Using Taylor expansion and the central limit theorem  $R^{I,J}$ :

$$\begin{aligned} R^{I,J} &= \frac{\sum_{s=I}^J e_s(s - I)}{\sum_{s=I}^J e_s} \\ &= \frac{\sum_{s=I}^J \left\{ (e_s - 1)(s - I) + (S - I) \right\}}{\sum_{s=I}^J \left\{ (e_s - 1) + 1 \right\}} \\ &= \left\{ \frac{(J - I)(J - I + 1)}{2} + \sum_{s=I}^J (s - I)(e_s - 1) \right\} \left\{ (J - I + 1) + \sum_{s=I}^J (e_s - 1) \right\}^{-1} \\ &= \left\{ \frac{(J - I)}{2} + \frac{1}{J - I + 1} \sum_{s=I}^J (s - I)(e_s - 1) \right\} \left\{ 1 + \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\}^{-1} \end{aligned}$$

Assuming that  $J - I$  is sufficiently large:

$$\begin{aligned} & \left\{ 1 + \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\}^{-1} \\ &= 1 - \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \Big/ \left[ 1 - \left\{ \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\}^2 \right] \\ &\approx 1 - \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \end{aligned}$$

which then results in:

$$\begin{aligned} R^{I,J} &\approx \left\{ \frac{(J - I)}{2} + \frac{1}{J - I + 1} \sum_{s=I}^J (s - I)(e_s - 1) \right\} \left\{ 1 - \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\} \\ &= \frac{(J - I)}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{J - I}{2} \right) (e_s - 1) \\ &\quad - \frac{1}{(J - I + 1)^2} \sum_{s=I}^J (e_s - 1) \sum_{s=I}^J (s - I)(e_s - 1) \\ &\approx \frac{J - I}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right) (e_s - 1). \end{aligned}$$

Using the CLT  $R^{I,J}$  is approximated:

$$R^{I,J} = \mu^{I,J} + \sigma^{I,J} \xi,$$

where  $\mu^{I,J}$  and  $\sigma^{I,J}$  are the mean and variance of  $R^{I,J}$  and  $\xi \sim N(0, 1)$ .

Note that:

$$\begin{aligned} \mu^{I,J} &= \mathbb{E} \left\{ \frac{J - I}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right) (e_s - 1) \right\} = \frac{J - I}{2} \\ \sigma^{2 I,J} &= \text{Var} \left\{ \frac{J - I}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right) (e_s - 1) \right\} \\ &= \frac{\text{Var}(e_s - 1)}{(J - I + 1)^2} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right)^2 \\ &= \frac{1}{(J - I + 1)^2} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right)^2 \end{aligned}$$



Using the fact that  $\sum_{i=k}^n i^2 = \sum_{i=1}^{n-k+1} (i+k-1)^2$ , we can derive:

$$\begin{aligned} \sum_{s=I}^J \left( s - \frac{J-I}{2} \right)^2 &= \sum_{s=1}^{J-I+1} \left( s - \frac{I+J}{2} + I - 1 \right)^2 \\ &= \sum_{s=1}^{J-I+1} \left( s - \frac{I+J}{2} + I - 1 \right)^2 \\ &= \sum_{s=1}^{J-I+1} \left( s - \frac{J-I}{2} - 1 \right)^2 \end{aligned}$$

Furthermore, as  $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$  the variance  $\sigma^{2 I, J}$  converges as follows:

$$\begin{aligned} \sigma^{2 I, J} &= \frac{1}{(J-I+1)^2} \sum_{s=1}^{J-I+1} \left( s - \frac{J-I}{2} - 1 \right)^2 \\ &= \frac{1}{(J-I+1)^2} \left\{ \sum_{s=1}^{J-I+1} s^2 - 2 \sum_{s=1}^{J-I+1} s \left( \frac{J-I-2}{2} \right) + \sum_{s=1}^{J-I+1} \left( \frac{J-I-2}{2} \right)^2 \right\} \\ &= \frac{(J-I+1)(J-I+2)(2(J-I)+3)}{6(J-I+1)^2} - \frac{(J-I+2)(J-I+1)(J-I-2)}{2(J-I+1)^2} + \\ &\quad + \frac{(J-I+1)(J-I-2)^2}{4(J-I+1)^2} \\ &\approx \frac{J-I}{3} - \frac{J-I}{2} + \frac{J-I}{4} = \frac{J-I}{12} \end{aligned}$$

Therefore,  $R^{I, J}$  can be approximated as:

$$R^{I, J}(e_s) \approx \frac{J-I}{2} + \sqrt{\frac{J-I}{12}} \xi$$

Next, combining (20) with the Taylor expansion for  $L_\theta^{I, J}(e)$ , we obtain

$$L(e) \approx -\theta \sqrt{\frac{(J-I)^3}{12}} \xi - \theta^2 \frac{(J-I)^3}{24}$$

Again all these approximations are of order  $\mathcal{O}_p(n^{-1/2})$ .

Thus, with simple algebra we arrive at the limit distribution of the test statistics

$$L_{\hat{\theta}}(e) \text{ sign}\{\hat{\theta}(e)\} \approx -\frac{1}{2} \xi^2 \text{ sign}(\xi).$$

The equation for the critical surface (18) therefore takes the following form

$$\mathbb{P}\left[\max_{I,J}\left\{\frac{1}{2}\xi^2 \operatorname{sign}(\xi) + t_\alpha(I, J)\right\} > 0\right] = \alpha. \quad (20)$$

In order to find a solution, we assume for a moment that the maximum in the above display is computed over couples  $I_k, J_k, k = 1, \dots, (n-1)/d$ , where  $I_k = 1 + d(k-1)$ ,  $J_k = I_k + d$  and  $d$  is a given integer. Thus, we are looking for a minimal  $t_\alpha(I_k, J_k)$  satisfying

$$\mathbb{P}\left[\max_{k \leq n/d}\left\{\frac{1}{2}(\xi^{I_k, J_k})^2 \operatorname{sign}(\xi^{I_k, J_k}) + t_\alpha(I_k, J_k)\right\} > 0\right] = \alpha.$$

Since the random variables  $(\xi^{I_k, J_k})^2 \operatorname{sign}(\xi^{I_k, J_k}), k = 1, \dots, n/d$  are i.i.d., it is clear that  $t_\alpha(I_k, J_k)$  is a constant depending only on  $\alpha, n$ , and  $d$ . Finally notice that

$$\max_{k \leq n/d} (\xi^{I_k, J_k})^2 \operatorname{sign}(\xi^{I_k, J_k}) \approx 2 \log \frac{n}{d}$$

because  $\xi^{I_k, J_k}$  are i.i.d. and nearly Gaussian  $N(0, 1)$ . Therefore it is clear that

$$t_\alpha(I_k, J_k) = -\tilde{t}_\alpha \log \frac{n}{d},$$

where  $\tilde{t}_\alpha$  is a constant close to 1. This argument prompts the following form of the critical surface (18):

$$t_\alpha(I, J) = -\tilde{t}_\alpha \log \frac{n}{J-I}. \quad (21)$$

The exact constant  $\tilde{t}_\alpha$  is finally computed with the help of the Monte-Carlo as a root of the equation:

$$\mathbb{P}\left(\min_{|I-J| \geq M} \left[ L_{\hat{\theta}(e)}^{I, J}(e) \operatorname{sign}\{\hat{\theta}(e)\} + \tilde{t}_\alpha \log \frac{n}{J-I} \right] < 0\right) = \alpha. \quad (22)$$

Hence the critical surface  $t_\alpha(I, J)$  in (21) is approximated as a function of a scalar critical value  $\tilde{t}_\alpha$ , sample size  $n$  and significance level  $\alpha$ , which definitely reduces the complexity of the computation.

Here  $M > 2$  is an integer which is needed to guarantee that the asymptotic approximation (22) holds true. Typically,  $M \approx 10$ . The inaccuracies due to small  $M$  and other approximations applied to derive the final results are compensated by  $\tilde{t}_\alpha$  critical value.

More precisely the calculation of the critical value  $\tilde{t}_\alpha$  is done in the following steps:

1. Generation of  $Z_{\text{gen}}$  as  $\exp(1)$  for a given sample size  $n$ .
2. Calculation of optimal parameters  $\hat{\theta}(I, J)$  and resulting  $L_{\hat{\theta}}(I, J)$  over generated sequences  $Z_{\text{gen}}$  for all possible intervals  $[I, J], 1 \leq I < J \leq n$

**Table 1** Simulated critical values for different sample sizes and  $\tilde{t}_\alpha$ 

$\alpha$ (%)	$n = 50$	$n = 100$	$n = 255$
20	2.5010	1.8003	1.2934
10	2.5789	1.8257	1.3065
5	2.6163	1.8358	1.3087
4	2.6229	1.8381	1.3093
3	2.6363	1.8414	1.3102
2	2.6425	1.8437	1.3111
1	2.6530	1.8453	1.3117

### 3. Calculation of the corresponding $\tilde{t}_\alpha$ as a root of equation

$$P\left(\min_{|I-J|\geq M} \left[ L_{\hat{\theta}(Z_{\text{gen}})}^{I,J}(Z_{\text{gen}}) \text{sign}\{\hat{\theta}(Z_{\text{gen}})\} + \tilde{t}_\alpha \log \frac{n}{J-I} \right] < 0\right) = \alpha \quad (23)$$

by repeating steps 1 and 2 using simulated data.

For reader's convenience Table 1 provides the critical values  $\tilde{t}_\alpha$  for  $\alpha = 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2$  and sample sizes  $n = 50, 100, 255$ . As can be seen for smaller size  $n$  the critical values  $\tilde{t}_\alpha$  are larger to counterbalance the inaccuracies in the estimation of  $L_{\hat{\theta}(Z_{\text{gen}})}$ .

With the given  $\tilde{t}_\alpha$  the monotonicity test on the observed data  $Z$  takes the following form: *we accept the hypothesis that  $K_s$  is a decreasing sequence if*

$$\min_{|I-J|\geq M} \left[ L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I} \right] > 0. \quad (24)$$

## 4 Empirical results

### 4.1 Data and estimation of risk neutral density

For the analysis we take the data used in [Detlefsen et al. \(2007\)](#) where the pricing kernels and the risk aversion are analysed in years 2000, 2002 and 2004 in order to consider different market regimes (30th of June, 28th of June and 25th of June correspondingly). These dates were selected in such a way that the DAX index was rising, remained stable and was falling during one year period prior to these dates. According to our test design the decision about monotonicity is made on the basis of (4):  $Z_k = n \cdot (X_{(k+1)} - X_{(k)}) \cdot q(X_{(k)})$  where  $X_{(k)}$  are the order statistics of DAX returns and  $q$  is an estimate of the risk neutral density.

The DAX returns  $\frac{S_t - S_{t-126}}{S_{t-126}}$  are calculated on half a year basis, where  $S_t$  are daily index observations. They are ordered into  $X_{(k)}$ . We started 1.5 year back from the dates mentioned above which resulted in exactly  $n = 255$  observations. The corresponding

ordered returns differences  $X_{(k+1)} - X_{(k)}$  for 2000, 2002 and 2004 are displayed in Fig. 3.

The risk neutral density  $q$  aggregates economic information about the prices by replicating observed option prices. An estimate of  $q$  can be found as the second derivative of the call price with respect to the strike. The estimation of  $q$  is then reduced to the problem of a proper option-pricing formula. Under the hypothesis of [Black and Scholes \(1973\)](#) we obtain a log-normal density  $q$ . A closed form solution can be also obtained under more general class of models. Here we use the [Heston \(1993\)](#) model calibrated to fit the observed smile in implied volatility surfaces (IVS) using the absolute error between observed and modeled quantities:

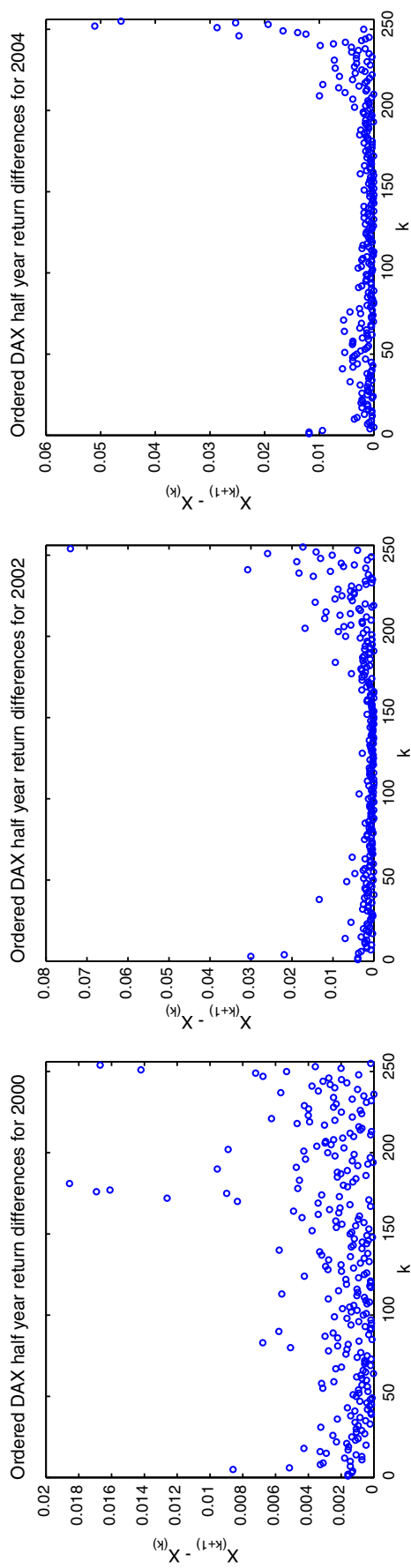
$$\text{ASE}_t = \sqrt{\sum_{i=1}^n n^{-1} \{\text{IV}_i^{\text{mod}}(t) - \text{IV}_i^{\text{mar}}(t)\}^2}$$

where mod refers to a model quantity, mar to a quantity observed on the market and  $\text{IV}(t)$  to an implied volatility on day  $t$ . The index  $i$  runs over all  $n$  observations of the surface on day  $t$ . Daily EUREX-settlement prices of European options on the DAX index are used to obtain observed option prices and corresponding implied volatilities. The model parameters are calibrated for each of three dates using the whole surface of implied volatilities, but we exclude observations that are deep out of the money because of illiquidity of these products. More precisely, we consider for the calibration only options with more than 1 month time to maturity and restrict ourselves to strikes 50% above or below the spot in the moneyness direction. For each trading day there are about 250 points in the volatility surface available for the calibration. Having obtained the model parameters we can estimate the risk neutral density for any time to maturity  $\tau$ . In this paper we analyse semiannual returns, therefore, we obtain the density  $q$  by fixing  $\tau = 0.5$  years. The corresponding densities for 2000, 2002 and 2004 can be seen in Fig. 4. The risk free interest rates are approximated by the EURIBOR. On each trading day we use the yields corresponding to the maturities of the implied volatility surface. As the DAX is a performance index it was adjusted to dividend payments. Thus, we do not have to consider dividend payments explicitly. For more details on the estimation of the risk neutral density refer to [Detlefsen and Härdle \(2007\)](#). Similar density  $q$  can be obtained using the minimization procedure mentioned in [Jackwerth \(2000\)](#). Alternatively, the density  $q(x)$  can be estimated semiparametrically or even nonparametrically, see [Ait-Sahalia and Lo \(2000\)](#).

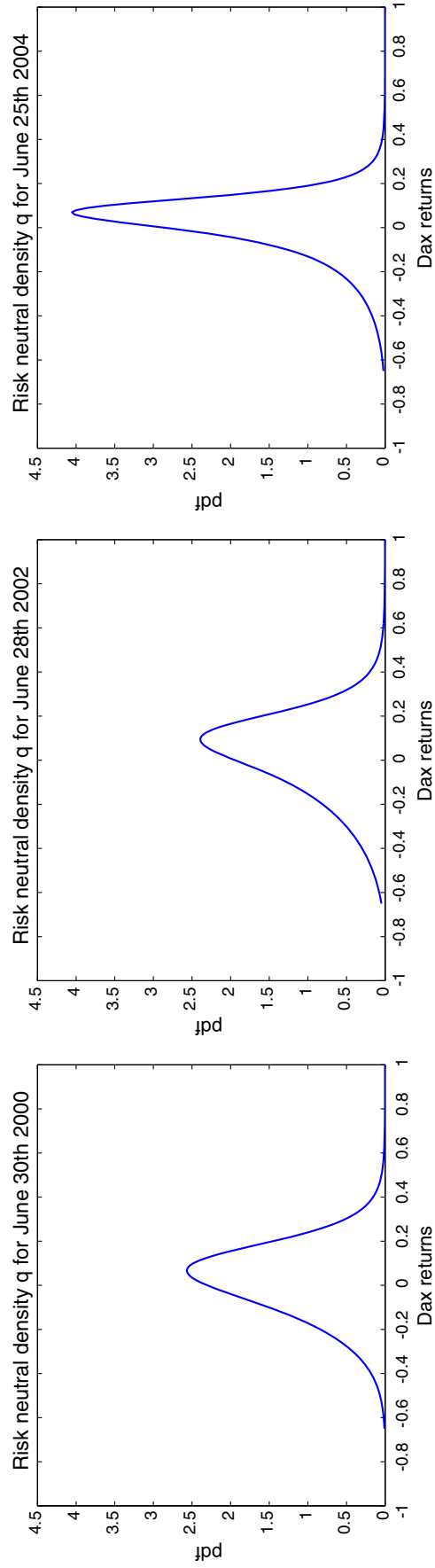
#### 4.2 Monotonicity of DAX EPKs

The final goal is to test an empirical pricing kernel obtained from observed data. Having obtained  $q$  and  $X_{(k)}$ ,  $Z_k$  can be calculated and the monotonicity testing becomes a technical exercise. Resulting values of  $Z_k$  are displayed in Fig. 5.

The calculated  $Z_k$  correspond to one year risk neutral density  $q$  and can be tested with the corresponding critical values for  $n = 255$  from Table 1. Similarly to the graphs showing the test ideas a minimum distance of 10 observations between  $I$  and  $J$  was set.



**Fig. 3** Half-year ordered returns differences  $X_{(k+1)} - X_{(k)}$  for years 2000, 2002 and 2004



**Fig. 4** Risk neutral densities  $q$  estimated on 30/06/2000, 28/06/2002 and 25/06/2004

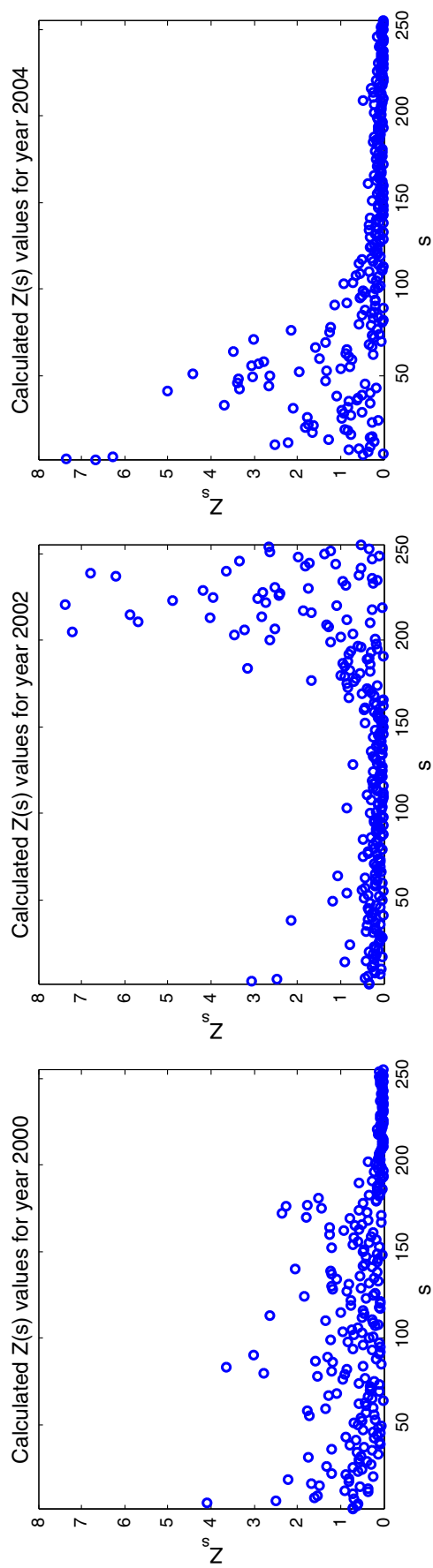


Fig. 5 Calculated  $Z_k$  for years 2000, 2002 and 2004

**Table 2**  $\min_{|I-J| \geq M} [L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I}]$  for  $\alpha = 10, 5$  and  $1\%$ 

$\alpha$ (%)	$\min[L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I}]$		
	2000	2002	2004
10	0.5038	-0.005	0.2114
5	0.5046	0.0021	0.2017
1	0.5058	0.0118	0.1946

The results are summarized in Table 2, the surfaces  $L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I}$  are given in Fig. 6.

The analysis of the DAX data in years 2000, 2002 and 2004 showed that  $H_0$  (monotonic pattern of the pricing kernel) is rejected under 90% significance level in year 2002.

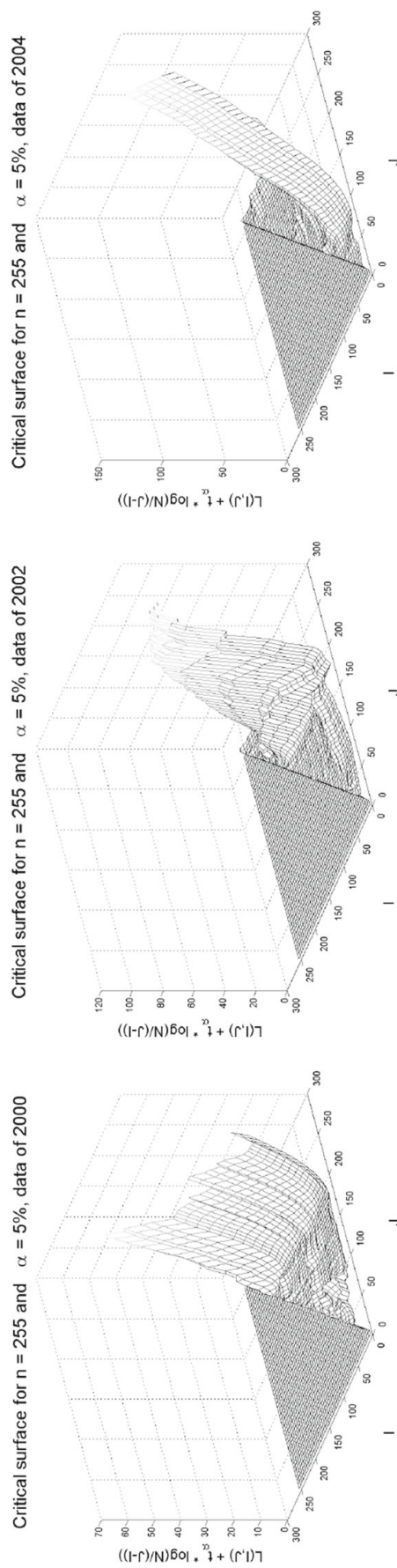
## 5 Conclusion

We describe a test that checks monotonicity of pricing kernels. By testing monotonicity of a pricing kernel we can determine whether the corresponding utility function is concave or not. A strictly decreasing pricing kernel corresponds to a concave utility function, while a non-decreasing EPK means that the utility function contains non-concave regions.

Pricing kernels are constructed as a ratio of the risk neutral density  $q$  and the historical density  $p$ . Investors' assessment of the future distribution of asset prices under risk neutral measure (density  $q$ ) can be estimated via the derivative market. By looking at option prices we can find out what stock prices or returns investors expect at the time of maturity. Due to the large number of observations  $q$  can be precisely estimated. The actual movement of  $S_t$  is described by the historical density  $p$  which is estimated using the time series of  $S_t$ . The main difficulty for the estimation of  $p$  is, of course, that an assumption about the model for the underlying process  $S_t$  has to be made. Due to scarcity of data and specification difficulties  $p$  is considered to be unknown. We, therefore, test the monotonicity via the ratio  $q/p$  of two densities, where  $q$  is given and  $p$  is unknown.

The test is constructed as follows: first the spacing method is used to reduce the problem to an exponential model. Using Pyke's lemma of order statistics, a pricing kernel  $K$  is represented as a sequence of observed values  $Z_k$  and standard exponential variables  $e_k$ . Based on this simple exponential model we construct the likelihood ratio test for a fixed interval  $[I, J]$ . A global test is built by the simultaneous testing on all possible intervals  $[I, J]$ , where the main difficulty is to calculate the corresponding critical surfaces for given  $I, J$ , sample size  $n$  and confidence level  $\alpha$ . The critical surfaces can be nearly approximated with a scalar critical value  $\tilde{t}_\alpha$  dependent only on sample size  $n$  and significance level  $\alpha$ , which significantly reduces the complexity of the test. The problem is then reduced to the simulation of the critical value  $\tilde{t}_\alpha$  for  $n$  and  $\alpha$  using the Monte-Carlo technique.





**Fig. 6** Resulting test statistic surfaces of  $L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}(\hat{\theta}(Z)) + \tilde{t}_{\alpha} \log \frac{\mu}{J-1}$  for years 2000, 2002 and 2004 with  $\alpha = 5\%$

We investigated the EPKs for German DAX data for the years 2000, 2002 and 2004 and found evidence of non-concave utility behaviour for the data under consideration.

## References

- Ait-Sahalia, Y., Lo, A.: Nonparametric risk management and implied risk aversion. *J. Econom.* **94**(12), 9–51 (2000)
- Bakshi, G., Madan, D., Panayotov, G.: Returns of claims on the upside and the viability of U-shaped pricing kernels. *J. Financ. Econ.* **97**, 130–154 (2010)
- Black, F., Scholes, M.: The pricing of options and corporate liabilities. *J. Polit. Econ.* **102**(3), 637–659 (1973)
- Cavalier, L., Golubev, Y.: Monotonicity of the stochastic discount factor and expected option returns. *Ann. Stat.* **34**(4), 1653–1677 (2006)
- Chaudhuri, R., Schroder, M.: Monotonicity of the stochastic discount factor and expected option returns. Working paper, School of Business Administration, Oakland University (2010)
- Constantinides, G., Jackwerth, J., Perrakis, S.: Mispricing of S&P 500 index options. *Rev. Financ. Stud.* **22**, 1247–1277 (2009)
- Detlefsen, K., Härdle, W.: Calibration risk for exotic options. *J. Deriv.* **14**(4), 47–63 (2007)
- Detlefsen, K., Härdle, W., Moro, R.: Empirical pricing kernels and investor preferences. *Math. Methods Econ. Financ.* **3**(1), 19–48 (2007)
- Grith, M., Härdle, W.K., Krätschmer, V.: An axiomatic and data driven view on the EPK paradox. *Rev. Financ.*, revise and resubmit (2013)
- Härdle, W., Okhrin, Y., Wang, W.: Uniform confidence bands for empirical pricing kernel. *J. Financ. Econom.*, revise and resubmit (2012)
- Heston, S.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financ. Stud.* **6**(2), 327–343 (1993)
- Jackwerth, J.: Recovering risk aversion from option prices and realized returns. *Rev. Financ. Stud.* **13**(2), 433–451 (2000)
- Karatzas, I., Shreve, S.: *Methods of Mathematical Finance*. Springer, NY (1998)
- Pyke, R.: Spacings. *J. R. Stat. Soc. B* **27**, 395–436 (1965)
- Rosenberg, J., Engle, R.: Empirical pricing kernels. *J. Financ. Econ.* **64**(3), 341–372 (2002)
- Von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press, Princeton (1944)

# Uniform Confidence Bands for Pricing Kernels

WOLFGANG KARL HÄRDLE

*Humboldt-Universität zu Berlin, CASE - Center for Applied Statistics and Economics*

YAREMA OKHRIN

*University of Augsburg*

WEINING WANG

*Humboldt-Universität zu Berlin*

## ABSTRACT

Pricing kernels implicit in option prices play a key role in assessing the risk aversion over equity returns. We deal with nonparametric estimation of the pricing kernel (PK) given by the ratio of the risk-neutral density estimator and the historical density (HD). The former density can be represented as the second derivative w.r.t. the European call option price function, which we estimate by nonparametric regression. HD is estimated nonparametrically too. In this framework, we develop the asymptotic distribution theory of the Empirical Pricing Kernel (EPK) in the  $L^\infty$  sense. Particularly, to evaluate the overall variation of the pricing kernel, we develop a uniform confidence band of the EPK. Furthermore, as an alternative to the asymptotic approach, we propose a bootstrap confidence band. The developed theory is helpful for testing parametric specifications of pricing kernels and has a direct extension to estimating risk aversion patterns. The established results are assessed and compared in a Monte-Carlo study. As a real application, we test risk aversion over time induced by the EPK. (JEL: C14, J01, J31)

**KEYWORDS:** empirical pricing kernel, confidence band, bootstrap, kernel smoothing, nonparametric fitting

---

The financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin is gratefully acknowledged. Address correspondence to Yarema Okhrin, Department of Statistics, University of Augsburg, D-86159 Augsburg, Germany, or e-mail: yarema.okhrin@wiwi.uni-augsburg.de

doi:10.1093/jfinc/mbu002 Advanced Access publication February 19, 2014

© The Author, 2014. Published by Oxford University Press. All rights reserved.

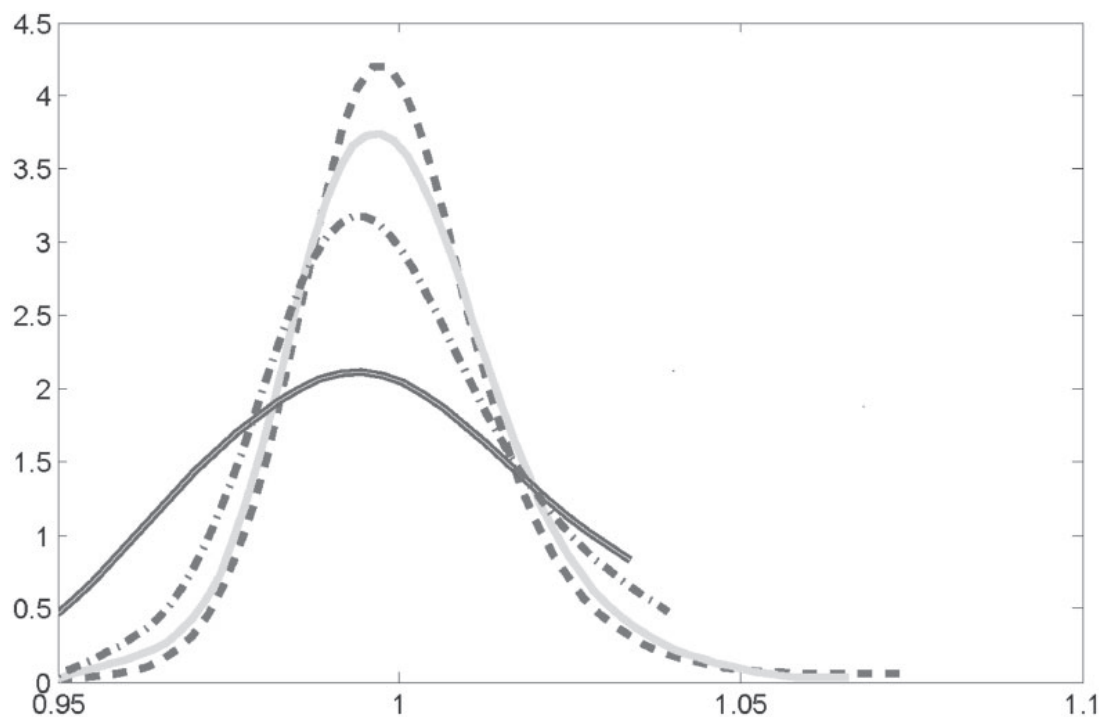
For Permissions, please email: journals.permissions@oup.com

A challenging task in financial econometrics is to understand investors' attitudes toward market risk in its evolution over time. Such a study naturally involves stochastic discount factors, empirical pricing kernels (EPK), and state price densities, see Cochrane (2001). Asset pricing kernels (PKs) summarize investors' risk preferences and the so called "EPK paradox" exhibits when PKs are estimated from data, as several studies including Aït-Sahalia and Lo (2000), Rosenberg and Engle (2002), Brown and Jackwerth (2004) have shown. Although in all these studies the EPK paradox (nonmonotonicity) became evident, a test for the nonmonotone behavior of the pricing kernel has not been devised yet. A uniform confidence band is a very simple tool for such shape inspection. Confidence bands drawn around an EPK based on asymptotic theory and bootstrap is the subject of our study. In addition, we relate critical values of our test to changing market conditions given by exogenous time series.

The common difficulty is that the investors' preference is implicit in the goods traded in the market and thus can not be directly observed from the path of returns. A profound martingale-based pricing theory provides us one approach to tackle the problem from a probabilistic perspective. An important concept involved is the State Price Density (SPD) or Arrow-Debreu prices reflecting fair prices of one unit gain or loss for the whole market. Under no arbitrage assumption, there exists at least one SPD, and when a market is complete, this SPD is unique. Assuming a market is complete, pricing is done by taking expected payoff under the risk neutral measure, which is related to the pdf of the historical measure by multiplying with a stochastic discount factor, see Section 1 for a detailed illustration. From an economic perspective, the price is formed according to the utility maximization theory, which admits that the risk preference of consumers is connected to the shape of utility functions. Specifically, a concave, convex, or linear utility function describes the risk averse, risk seeking, or risk neutral behavior. Importantly, a stochastic discount factor can be expressed via a utility function (Marginal Rate of Substitution), which links the shape of pricing kernel to the risk patterns of investors, see Kahneman and Tversky (1979), Jackwerth (2000), Rosenberg and Engle (2002) and others.

The above mentioned theory allows us to relate price processes of assets to risk preference of investors. This amounts to fitting a flexible model and making inference on the dynamics of EPKs over time. A well-known but restrictive approach is to assume the underlying following a geometric Brownian motion. In this setting, SPDs and HDs are log normal distributions with different drifts, and the parametrization of PK coincides with the conditional expectation of marginal utilities when assuming a power utility function. Thus it is decreasing in return and implies overall risk-averse behavior. However, across different markets, one observes quite often a nondecreasing pattern for EPKs, a phenomenon called the EPK paradox, see Chabi-Yo, Garcia, and Renault (2008), Christoffersen, Heston, and Jacobs (2011).

Two plots of pricing kernels are shown in Figures 1 and 2. Figure 1 depicts inter-temporal pricing kernels with fixed maturity, while Figure 2 depicts pricing

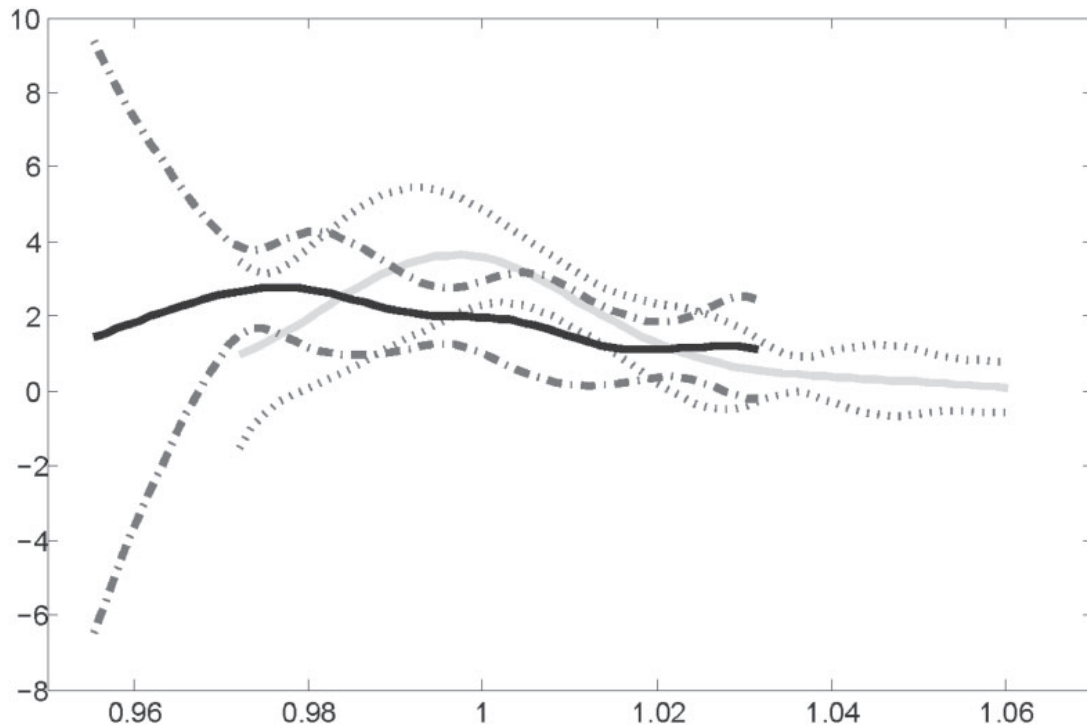


**Figure 1** Examples of estimated inter-temporal pricing kernels (as functions of moneyness) with fixed maturity 0.00833 (3 days) in years respectively on January 17, 2006 (dashed), April 18, 2006 (light gray), May 16, 2006 (dark gray), June 13, 2006 (dash-dotted), see Grith, Härdle, and Park (2013).

kernels with two different maturities and their confidence bands. The figures are shown on return scale. The curves present a bump in the middle and a switch from convexity to concavity in all cases. Especially, this shows that very unlikely the bands contain a monotone decreasing curve.

Besides the shape of the confidence bands, the time varying coverage probability of a uniform confidence band has implications on risk attitudes of investors. At a fixed point in time, it helps us to test against alternatives for a PK and thus yields insights into time varying risk patterns. The extracted time varying parameter, realized either from a low-dimensional model for PKs or given by the coverage probability, may thus be economically analyzed in connection with exogenous macroeconomic business cycle indicators, e.g., credit spread, yield curve, etc., see also Grith, Härdle, and Park (2013).

To our knowledge, there are no comparable approaches developed for uniform testing of the shape of EPK or of any continuous transformations of it. Golubev, Härdle, and Timofeev (2014) suggest a test of monotonicity of the PK, while other literature on testing PKs, for example, Jagannathan and Wang (1996), Wang (2002), Wang (2003), serves different purposes like verifying the significance of pricing errors. In contrary to these papers we do not address the issue of mispricing, but provide a solid statistical tool to testing the validity of any parametric shape of the PK.



**Figure 2** Examples of estimated inter-temporal pricing kernels with various maturities in years: 0.02222 (8 days, gray) 0.1 (36 days, black) on January 12, 2006 and their confidence bands.

Several econometric studies are concerned with estimating PKs by estimating a SPD and HD separately. See Section 1 for details. It is stressed in Aït-Sahalia and Lo (1998) that nonparametric inference from pricing kernels gives unbiased insights into the properties of asset markets. The stochastic fluctuation of EPK as measured by the maximum deviation has not been studied yet. Nevertheless, the asymptotic distribution of the maximum deviation and the uniform confidence band linked to it are very useful for model check.

Uniform confidence bands for smooth curves have first been developed for kernel density estimators by Bickel and Rosenblatt (1973), extension to regression smoothing can be found in Liero (1982) and Härdle (1989). But only recently, the results have been carried over to derivative smoothing by Claeskens and Van Keilegom (2003). Our theoretical path follows largely their results, but our results are applied to a ratio estimator instead of a local polynomial estimator. Also we have a realistic data situation that relates coverage to economic indicators. In addition we perform the smoothing in an implied volatility space which brings by itself an interesting modification of the results of that paper.

This article is organized as follows: In Section 1, we describe the theoretical connection between utility functions and pricing kernels. In Section 2, we present a nonparametric framework for the estimation of both the HD and the SPD and derive the asymptotic distribution of the maximum deviation. In Section 3, we simulate the asymptotic behavior of the uniform confidence band and compare it with the

bootstrap method. Moreover, we also compare the results with other parametric estimation procedures. In Section 4, we conclude and discuss our results.

## 1 EMPIRICAL PRICING KERNEL ESTIMATION

Consider an arbitrary risky financial security with the price process  $\{S_t\}_{t \in [0, T]}$ . The interest rate process  $r$  is deterministic. We assume that the market is complete, so there exists a unique risk neutral measure. By the change-of-measure argument the price at time  $t$  for the nonnegative payoff  $\psi(S_T)$  is

$$P_t \stackrel{\text{def}}{=} E^Q[e^{-r\tau}\psi(S_T)] = E[e^{-r\tau}\psi(S_T)\mathcal{K}(S_T)], \quad (1)$$

where  $\mathcal{K}(S_T)$  is defined as *the pricing kernel* or *stochastic discount factor* at time  $t$ ,  $E$  is the expectation under the historical measure  $\mathbb{P}_{S_T|S_t}(x)$  and  $E^Q$  is the expectation under the risk neutral measure  $\mathbb{Q}_{S_T|S_t}(x)$ ,  $\tau$  is the time to maturity. Thus the price of the security at time point  $t$  equals the expected net present value of its future payoffs, computed with respect to the risk-neutral measure. More explicitly,

$$P_t = \int_0^{+\infty} e^{-r\tau}\psi(x)q(x)dx = \int_0^{+\infty} e^{-r\tau}\psi(x)\mathcal{K}(x)p(x)dx, \quad (2)$$

where  $p(x)$  and  $q(x)$  are the pdf of the historical measure and the risk neutral density or state price density (SPD) of  $S_T$ , respectively. Note that  $p(x)$  and  $q(x)$  are conditional on the current price  $S_t$  and potentially may depend on other parameters as discussed below. We skip the indication of conditioning to keep the notation simple. Thus all expectations hereafter are conditional on  $S_t$  if not stated otherwise.

It follows from (2), that  $\mathcal{K} = \frac{q}{p}$  and both the pdf of the future payoff and the SPD are required to compute the pricing kernel. Several approaches are available to determine the EPK explicitly. First, we can impose strict parametric restrictions on the dynamics of the asset prices and on the distribution of the future payoff. Mixture normal distributions are an example, see Jackwerth (2000). In the case of more complex stochastic processes, usually no explicit solution is available. A possible technique though is to use the Brownian motion setup as a prior model. Subsequently the SPD is estimated by minimizing the distance to the prior SPD subject to the constraints characterizing the underlying securities, see Rubinstein (1994) and Jackwerth and Rubinstein (1996).

Another important perspective of specifying PK is done via the utility function in the consumption based pricing model, see Heaton and Lucas (1992).

Let the aim of the investor be to solve the problem:

$$\max_{W_t} \{u(W_t) + E[\beta u(W_T)]\},$$

where  $u(\cdot)$  denotes the utility function,  $W_t$  the wealth and  $\beta$  the subjective discount factor. The current price of an asset is

$$P_t = E[\beta_t \psi(S_T)], \quad (3)$$

where  $\beta_t$  is the stochastic discount factor and it equals the inter-temporal marginal rate of substitution. If both ways of pricing in (1) and (3) are admissible, then they lead to the same price in a complete market. Then the stochastic discount factor is given by  $\beta \frac{u'(s)}{u'(S_t)}$  and is proportional to the PK. This implies that by fixing the utility of the investor we can determine the PK, which is related to the standard risk aversion measures. In practice, however, usually the opposite procedure is applied. The PK is estimated via a ratio of  $\hat{q}$  and  $\hat{p}$  and used to determine the utility function or the risk aversion coefficient of the investor. Assessment of the temporal dynamics of the latter allows for inferences on the market risk behavior.

### 1.1 EPK and Option Pricing

EPK is calibrated from the data via an estimation of the ratio of the SPD  $q$  and the HD  $p$  respectively. In this section we describe the details for this calibration. The latter can easily be estimated either parametrically or nonparametrically from the time series of payoffs. On the contrary, the SPD depends on risk preferences and therefore the past observed stock price time series do not contain enough information. Option prices do reflect preferences and, therefore, can be used to estimate the SPD  $q$ . Let  $C(S_t, X, \tau, r, \sigma^2)$  denote the European call-option price as a function of the strike price  $X$ , price  $S_t$ , maturity  $\tau$ , interest rate  $r$  and volatility  $\sigma$ . Following Breeden and Litzenberger (1978) the SPD can be determined from the pricing equation by

$$q(S_T) = e^{r\tau} \frac{\partial^2 C}{\partial X^2} \Big|_{X=S_T}. \quad (4)$$

This result is very general and is valid for all European call options with the payoff function  $(S_T - X)_+$  and with the single assumption that the price is twice differentiable. No additional restrictions on the stochastic process for the underlying or on the preferences of market participants are needed. In a Black-Scholes (BS) framework, where the underlying asset price  $S_t$  follows a geometric Brownian motion, the European options are priced via:

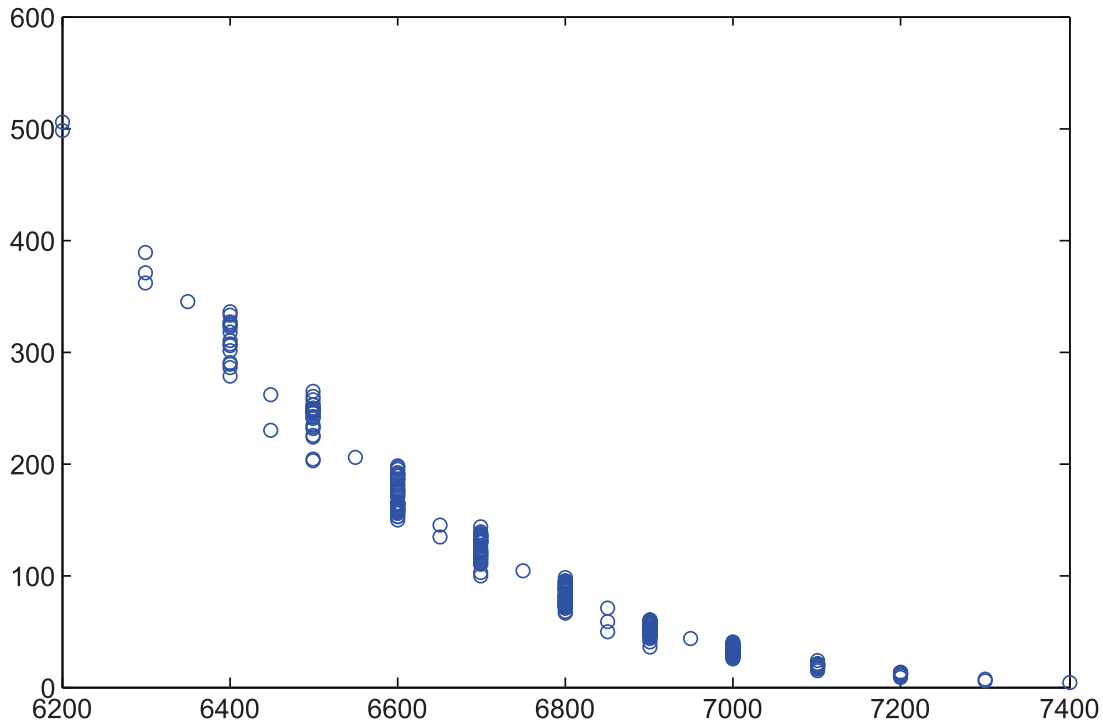
$$C(S_t, X, \tau, r, \sigma^2) = S_t \Phi(d_1) - X e^{r\tau} \Phi(d_2),$$

where  $d_1$  and  $d_2$  are known functions of  $\sigma^2$ ,  $\tau$ ,  $X$ , and  $S_t$ . This implies that both  $q(S_T)$  and  $p(S_T)$  are the densities of lognormal distributions:

$$q(S_T) = \frac{1}{S_T \sqrt{2\pi\sigma^2\tau}} \exp \left[ -\frac{\{\log(S_T/S_t) - (r - \sigma^2/2)\tau\}^2}{2\sigma^2\tau} \right] \quad (5)$$

and  $p(S_T)$  with  $\mu$  replacing  $r$  in (5).





**Figure 3** Plot of call option prices against strike prices on January 17, 2001

The restrictive parametric form of BS model may often induce modeling bias when fitted to the data, especially it is not possible to reflect the implicit volatility smile (surface) as a function of  $X$  and  $\tau$  via (5), see Renault (1997). The latter may be derived in a stochastic volatility model, such as those of Heston or Bates type. In order to study unbiased risk patterns, we need to guarantee models for the pricing kernel that are rich enough to reflect local risk aversion in time and space. This leads naturally to a smoothing approach.

We now describe how to estimate  $q(\cdot)$  nonparametrically. Consider call options with maturity  $\tau$ . We consider the following heteroscedastic nonparametric model for the observed option prices  $Y_i$

$$Y_i = C_\tau(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n_q, \quad (6)$$

where  $Y_i$  denotes the observed option price,  $X_i$  the strike price and  $C_\tau(\cdot)$  is a smooth function of the strike price. For simplicity of notation, we write  $C(\cdot)$  for  $C_\tau(\cdot)$ . The informational content of the model is similar to that of Yuan (2009). We argue that the perceived errors are due to neglected heterogeneity factors, rather than mispricings exploited by arbitrage strategies, see Renault (1997). Thus the pricing errors  $\varepsilon_i$  are assumed to be i.i.d. in the cross section. Figure 3 depicts the call option prices used to calculate a SPD. The observations are distributed, with different variances, at discrete grid points of strike prices.

As from (4), estimation of  $q(\cdot)$  boils down to the estimation of the second derivative of  $C(\cdot)$ . The following local polynomial approach allows us to estimate

$C(\cdot)$  and the derivatives of  $C(\cdot)$  simultaneously. Assuming that  $C(X)$  is continuously differentiable of order  $d=3$ , it can be locally approximated by

$$C(X, X_0) = \sum_{j=0}^d C_j(X_0)(X_0 - X)^j, \quad (7)$$

where  $C_j(X) = C^{(j)}(X)/j!$ ,  $j=0, \dots, d$ . See Cleveland (1979), Fan (1992), Fan (1993), Ruppert and Wand (1994) for more details. By assuming a local Gaussian quasi-likelihood model with  $\mathcal{L}\{Y, C(X)\} = \{Y - C(X)\}^2 / \{2\sigma^2(X)\}$  and maximizing the local likelihood, the function  $C(\cdot)$  can be approximated by  $\hat{C}(X, X_0)$  with:

$$\hat{C}(X) = \arg \max_{C(X)} \frac{1}{n_q} \sum_{i=1}^{n_q} w_i K_{h_{n_q}}(X_i - X) \mathcal{L}\{Y_i, C(X, X_i)\}, \quad (8)$$

where  $C(X) = \{C_0(X), C_1(X), \dots, C_d(X)\}^\top$ ,  $K_{h_{n_q}}(X_i - X) = K\{(X_i - X)/h_{n_q}\}/h_{n_q}$  is a kernel function with bounded support and a bandwidth sequence  $h_{n_q}$ . Following Ait-Sahalia and Duarte (2003) and Yuan (2009) the weights  $w_i$  may reflect the relative liquidity of different options, putting more weight on more heavily traded options. To simplify the exposition we assume  $w_i = 1$  for all  $i$ . The above maximum likelihood approach is equivalent to a minimum weighted square loss approach, since maximizing the local Gaussian likelihood function leads numerically to the same solution as minimizing the weighted least squares. The advantage of the likelihood framework is that it can be easily adapted to non-Gaussian distributions and to heteroscedastic pricing equations.

Solving the above optimization problem in (8) is equivalent to solving:

$$\mathbf{A}_{n_q}(X) \stackrel{\text{def}}{=} \frac{1}{n_q} \sum_{i=1}^{n_q} K_{h_{n_q}}(X_i - X) \frac{\partial Q\{Y_i, C(X, X_i)\}}{\partial C} \mathbf{X}_i = \mathbf{0}, \quad (9)$$

with  $\mathbf{X}_i \stackrel{\text{def}}{=} (1, X_i - X, (X_i - X)^2, (X_i - X)^3)^\top$ . We are concerned with  $2! \hat{C}_2(x) = \frac{\partial^2 C(X)}{\partial X^2} \Big|_{X=x}$ , which is shown by Breeden and Litzenberger (1978) to be proportional to  $q(x)$ . Note that the described procedure does not guarantee the feasibility of the estimator as a density. The constrained estimator of Ait-Sahalia and Duarte (2003) makes the large sample results below invalid. Therefore, we rely on the consistency and asymptotic validity of  $\hat{q}$  as a density estimator. This approach is justified by large samples available in financial applications. A multiplicative renormalizing of the estimator will shift the EPK curve and the corresponding confidence bands, while keeping the results of the monotonicity test unchanged. Furthermore, the renormalization introduces a bias, which is difficult to tackle analytically. Additional improvement of the estimator is elaborated in Section 2.1.

Note that we assume the parameter  $C(\cdot)$  and  $\sigma(\cdot)$  to be orthogonal to each other. Thus we can estimate them separately as in a single parameter case. The following lemma states the results on the existence of the solution and its consistency.

**Lemma 1:** Under conditions (A1)–(A7), there exists a sequence of solutions to the equations

$$\mathbf{A}_{n_q}(x) = \mathbf{0},$$

with  $x$  being an element of a compact set  $E$ , such that

$$\sup_{x \in E} |\hat{q}(x) - q(x)| = \mathcal{O}[h_{n_q}^{-2} \{\log n_q / (n_q h_{n_q})\}^{1/2} + h_{n_q}^2] \quad a.s.$$

*Proof.* The statement follows from Theorem 2.1 of Claeskens and Van Keilegom (2003). ■

The HD  $p(x)$  can be estimated separately from the SPD using historical prices  $S_t, \dots, S_{t+n_p+\tau-1}$  ( $t+n_p+\tau-1 < T$ ) of the underlying asset. The nonparametric kernel estimator of  $p(x)$  is given similarly to Aït-Sahalia and Lo (2000) by

$$\hat{p}_{return}(x) = n_p^{-1} \sum_{j=0}^{n_p-1} K_{h_{n_p}} \{x - \log(S_{t+j+\tau}/S_{t+j})\},$$

where  $K_{h_{n_p}}$  is a kernel function with bounded support and the bandwidth  $h_{n_p}$ . This kernel should necessarily coincide with the kernel for estimating SPD  $q(\cdot)$ . Also as in Aït-Sahalia and Lo (2000), the density of log-returns can be estimated as:

$$\hat{p}_{return}(x) = S_t \exp(x) \hat{p}\{S_t \exp(x)\}.$$

Alternatively, to eliminate the impact of serial dependence of overlapping returns over  $\tau$  periods, we can simulate independent pathes of the price process and use it to estimate the density of  $S_T$ , then  $n_p$  will become the number of paths simulated for the time  $T$ . Under assumption (A5), we know that

$$\sup_{x \in E} |\hat{p}(x) - p(x)| = \mathcal{O}\{(n_p h_{n_p} / \log n_p)^{-1/2} + h_{n_p}^2\} \quad a.s. \quad (10)$$

**Remark** The uniform convergence results for estimation of HD in the i.i.d case follows from Bickel and Rosenblatt (1973), and recently extended by Liu and Wu (2010) (Theorem 2.3) to the serial dependent data case.

The EPK is then given by the ratio of the estimated SPD and the HD  $p(x)$  i.e.  $\hat{\mathcal{K}}(x) = \hat{q}(x) / \hat{p}(x)$ . The next lemma provides the linearization of the ratio, which is important for further statements about the uniform confidence band of the EPK.

**Lemma 2:** Under conditions (A1)–(A7) it holds

$$\begin{aligned} & \sup_{x \in E} |\hat{\mathcal{K}}(x) - \mathcal{K}(x)| \\ &= \sup_{x \in E} \left| \frac{\hat{q}(x) - q(x)}{p(x)} - \frac{\hat{p}(x) - p(x)}{p(x)} \cdot \frac{q(x)}{p(x)} - \frac{\{\hat{q}(x) - q(x)\} \{\hat{p}(x) - p(x)\}}{p^2(x)} \right| \\ & \quad + \mathcal{O}[\max\{(n_p h_{n_p} / \log n_p)^{-1/2} + h_{n_p}^2, h_{n_q}^{-2} \{n_q h_{n_q} / \log n_q\}^{-1/2} + h_{n_q}^2\}] \quad a.s. \quad (11) \end{aligned}$$

This lemma implies that the stochastic deviation of  $\hat{\mathcal{K}}$  from  $\mathcal{K}$  can be linearized into a stochastic part containing the estimator of the SPD and a deterministic part containing  $E[\hat{p}(x)]$ . The uniform convergence can be proved by dealing separately with the two parts. The convergence of the deterministic part is shown by imposing mild smoothness conditions, while the convergence of the stochastic part is proved by following the approach of Claeskens and Van Keilegom (2003). Theorem 1 formalizes this uniform convergence of the EPK.

**Theorem 1:** *Under conditions (A1)–(A7), it holds*

$$\sup_{x \in E} |\hat{\mathcal{K}}(x) - \mathcal{K}(x)| = \mathcal{O}[\max\{(n_p h_{n_p} / \log n_p)^{-1/2} + h_{n_p}^2, h_{n_q}^{-2} \{n_q h_{n_q} / \log n_q\}^{-1/2} + h_{n_q}^2\}] \quad a.s.$$

The proof is given in the Appendix. The theoretical optimal rate of bandwidth follows by minimizing the bias and variance term together in Theorem 1 leading to  $(n_q \log n_q)^{-1/9}$ . In our simulation and applications, we use the bandwidth sequence which minimizes coverage error by performing a grid search, see Claeskens and Van Keilegom (2003).

## 2 CONFIDENCE INTERVALS AND CONFIDENCE BANDS

Confidence intervals characterize the precision of the EPK for a given fixed value of the payoff. This allows to make inference about the PK at each particular strike price, but does not allow conclusions about the global shape. The confidence bands, however, characterize the whole EPK curve and offer therefore the possibility to test for shape characteristics. In particular, it is a way to check the persistence of the bump as observed. Given a certain shape, one may verify the restriction imposed by the power utility and obtain insights on the market risk aversion. In addition, the confidence bands can be used to measure the global variability of the EPK. Also, the proportion of the BS-based EPK covered in nonparametric bands can be used as a measure of global risk aversion. The global variability is measured by the variance function of EPK and the BS-based EPK means the parametric fitting achieved by assuming that the underlying follows the geometric Brownian motion.

A confidence interval for the EPK at a fixed value  $x$  requires the asymptotic distribution of  $\hat{p}(x)$  and  $\hat{q}(x)$ . Hereafter, we use  $\mathcal{L}$  to denote the convergence in law. Under (A1)–(A7):

$$\sqrt{n_p h_{n_p}} \{\hat{p}(x) - p(x)\} \xrightarrow{\mathcal{L}} N\{0, p(x) \int K^2(u) du\} \quad (12)$$

and

$$\sqrt{n_q h_{n_q}^5} \{\hat{q}(x) - q(x)\} \xrightarrow{\mathcal{L}} N\{0, \sigma_q^2(x)\}, \quad (13)$$

where  $\sigma_q^2(x) = [B(x)^{-1} L^{-1} T L^{-1}]_{(3,3)}$ , with  $B(x)$  equal to the product of the density  $f_X(x)$  of the strike price and the local Fisher information matrix  $I\{C(x)\}$ . The matrices

$\mathbf{L}$  and  $\mathbf{T}$  are given by  $\mathbf{L} \stackrel{\text{def}}{=} [\int u^{i+j}K(u)du]_{i,j}$  and  $\mathbf{T} \stackrel{\text{def}}{=} [\int u^{i+j}K^2(u)du]_{i,j}$  with  $i, j = 0, \dots, 3$ . This implies the asymptotic normality of the EPK at a fixed payoff  $x$ . More precisely

$$\sqrt{n_q h_q^5} \{ \hat{\mathcal{K}}(x) - \mathcal{K}(x) \} \xrightarrow{\mathcal{L}} N\{0, \sigma_q^2(x)/p^2(x)\}. \tag{14}$$

The variance of  $\hat{\mathcal{K}} = \hat{\mathcal{K}}(x)$  is given by

$$\text{Var}\{\hat{\mathcal{K}}(x)\} \approx \{p(x)\}^{-2} B^{-1}(x) \mathbf{L}^{-1} \mathbf{T} \mathbf{L}^{-1}. \tag{15}$$

The above results on the limiting distribution of  $\hat{p}$ ,  $\hat{q}$ , and  $\hat{\mathcal{K}}$  can directly used to establish confidence intervals and the Bonferroni-type confidence bands for the considered densities. This approach, as argued by Eubank and Speckman (1993), are asymptotically conservative even though they do not explicitly account for potential bias. These bands are taken as benchmarks for comparison purposes in the simulation study.

Let  $\hat{\mathcal{D}}_{n_q}(x)$  be the standardized process with the estimated variance of the EPK:

$$\hat{\mathcal{D}}_{n_q}(x) \stackrel{\text{def}}{=} n_q^{1/2} h_{n_q}^{5/2} \{ \hat{\mathcal{K}}(x) - \mathcal{K}(x) \} / [\widehat{\text{Var}}\{\hat{\mathcal{K}}(x)\}]^{1/2}.$$

Relying on the linearization in Lemma 2, we derive the confidence band for  $\mathcal{K}$ .

**Theorem 2:** *Under assumptions (A1)-(A5) it follows*

$$P \left[ (-2\log h_{n_q})^{1/2} \left\{ \sup_{x \in E} |\hat{\mathcal{D}}_{n_q}(x)| - c_{n_q} \right\} < z \right] \rightarrow \exp\{-2\exp(-z)\},$$

where  $c_{n_q} = (-2\log h_{n_q})^{1/2} + (-2\log h_{n_q})^{-1/2} \{x_\alpha + \log(R/2\pi)\}$ .

The  $(1 - \alpha)100\%$  confidence band for the pricing kernel  $\mathcal{K}$  is thus:

$$[f : \sup_{x \in E} \{ |\hat{\mathcal{K}}(x) - f(x)| / \widehat{\text{Var}}(\hat{\mathcal{K}})^{1/2} \} \leq L_\alpha],$$

where  $L_\alpha \stackrel{\text{def}}{=} 2!(n_q h_{n_q}^5)^{-1/2} c_{n_q}$ ,  $x_\alpha = -\log\{-1/2\log(1 - \alpha)\}$  and

$$R \stackrel{\text{def}}{=} (\mathbf{L}^{-1} \mathbf{P} \mathbf{L}^{-1})_{3,3} / (\mathbf{L}^{-1} \mathbf{T} \mathbf{L}^{-1})_{3,3}$$

with  $\mathbf{P} \stackrel{\text{def}}{=} [\int u^{i+j}K'(u)]^2 du - \frac{1}{2} \{i(i-1) + j(j-1)\} \int u^{i+j-2}K^2(u)du]_{i,j=0,\dots,3}$ .

For the implementation with real data we need a consistent estimator of  $\text{Var}(\hat{\mathcal{K}})$ . For fixed  $\tau$ , we rely on the delta method and use the empirical sandwich estimator, see Carroll, Ruppert, and Welsh (1998). The latter method provides the variance estimator for the parameters obtained from estimating equations given by (9).

To estimate the variance function of the EPK we consider time series of the option prices and the corresponding strike prices,  $(X_{it}, Y_{it}), i = 1, \dots, n_q$ ;

$t = t + 1, \dots, t + \tau$ , we have

$$\widehat{\text{Var}}\{\hat{\mathcal{K}}(x)\} = \{\hat{p}(x)\}^{-2} \mathbf{V}(x)^{-1} \mathbf{U}(x) \mathbf{V}(x)^{-1}, \quad (16)$$

where

$$\mathbf{V}(x) \stackrel{\text{def}}{=} \frac{1}{n_q \tau} \sum_{i=1}^{n_q} \sum_{j=t+1}^{t+\tau} K_{h_{n_q}}^2(X_{ij} - x) \left[ \frac{\partial}{\partial C} Q\{Y_{ij}; \hat{C}(x, X_{ij})\} \right]^2 (\mathbf{H}_{n_q}^{-1} \mathbf{X}_{ij})(\mathbf{H}_{n_q}^{-1} \mathbf{X}_{ij})^\top, \quad (17)$$

$$\mathbf{U}(x) \stackrel{\text{def}}{=} \frac{1}{n_q \tau} \sum_{i=1}^{n_q} \sum_{j=t+1}^{t+\tau} K_{h_{n_q}}^2(X_{ij} - x) \left[ \frac{\partial^2}{\partial^2 C} Q\{Y_{ij}; \hat{C}(x, X_{ij})\} \right] (\mathbf{H}_{n_q}^{-1} \mathbf{X}_{ij})(\mathbf{H}_{n_q}^{-1} \mathbf{X}_{ij})^\top, \quad (18)$$

where  $\mathbf{X}_{ij} \stackrel{\text{def}}{=} (1, \dots, (X_{ij} - x)^3)^\top$  and  $\mathbf{H}_{n_q} \stackrel{\text{def}}{=} \text{diag}\{1, \dots, h_{n_q}^3\}$ . The estimator is consistent in our setup as motivated in Appendix A.2 of Carroll, Ruppert, and Welsh (1998).

It is important to note that the nonparametric estimators are biased, which leads to potentially wrongly centered confidence bands and misleading coverages.

To overcome this problem we deploy the bias-correcting technique of Xia (1998), which is based on the local polynomial estimation. It is used to correct the bias in estimated SPD, while the bias in the HD is corrected using the additive bias correction method mentioned in Jones, Linton, and Nielson (1995). In the next step we correct the bias in the EPK using the linearization in Lemma 2. The estimated leading term bias for EPK consists of the estimated bias of  $\hat{q}(x)$  and of  $\hat{p}(x)$  with a bigger bandwidth than what used in estimation. This is the oversmoothing idea proposed by Eubank and Speckman (1993).

## 2.1 Bootstrap Confidence Bands

In this subsection, we discuss a bootstrap version of the confidence band to obtain possibly better finite sample performance. The slow rate of convergence is known to us by Hall (1991), who showed that for density estimators, the supremum of  $\{\hat{q}(x) - q(x)\}$  converges at the slow rate  $(\log n_q)^{-1}$  to the Gumbel extreme value distribution. Therefore the confidence band may exhibit poor performance in finite samples. An alternative approach is to use the bootstrap method. Claeskens and Van Keilegom (2003) used smooth bootstrap for the numerical approximation to the critical value. Here we consider the bootstrap technique of the leading term in Lemma 2

$$\sup_{x \in E} \left| \frac{\hat{q}(x) - q(x)}{p(x)} \right|.$$

We resample data from the smoothed bivariate distribution of  $(X, Y)$  with the density estimator given by estimator is:

$$\hat{f}(x, y) = \frac{\hat{\sigma}_X}{n_q h_{n_q} h_{n_q} \hat{\sigma}_Y} \sum_{i=1}^{n_q} K \left\{ \frac{X_i - x}{h_{n_q}}, \frac{(Y_i - y) \hat{\sigma}_X}{h_{n_q} \hat{\sigma}_Y} \right\},$$

where  $\hat{\sigma}_X$  and  $\hat{\sigma}_Y$  are the estimated standard deviations of the distributions of  $X$  and  $Y$ . The motivation of using the smooth bootstrap procedure is that the Rosenblatt transformation requires the resampled data  $(X^*, Y^*)$  to be continuously distributed.

From the re-sampled data sets, we calculate the bootstrap analogue of the leading term in Lemma 2:

$$\sup_{x \in E} \left| \frac{\hat{q}^*(x) - \hat{q}(x)}{\hat{p}(x)} \right|.$$

One may argue that this resampling technique does not correctly reflect the bias arising in estimating  $q$ . Therefore, Härdle and Marron (1991) use therefore a resampling procedure based on a larger bandwidth. This refined bias correcting bootstrap method does not need to be applied in our case, since our bandwidth conditions ensure a negligible bias.

Correspondingly, we define the one-step estimator for the stochastic deviation:

$$h_{n_q}^2 \{\hat{\mathcal{K}}^*(x) - \hat{\mathcal{K}}(x)\} = -\{p(x)\}^{-2} \{\mathbf{U}^*(x)^{-1} \mathbf{H}_{n_q}^{-1} \mathbf{A}_{n_q}^*(x)\}_{3,3},$$

with  $\mathbf{U}^*(x)$  and  $\mathbf{A}_{n_q}^*(x)$  as  $\mathbf{U}(x)$  and  $\mathbf{A}_{n_q}(x)$  defined previously with bootstrap data  $(X_i^*, Y_i^*)$  and the variance given by:

$$\text{Var}\{\hat{\mathcal{K}}(x)\} \approx \{p(x)\}^{-2} B(x)^{-1} \mathbf{L}^{-1} \mathbf{P} \mathbf{L}^{-1}, \quad (19)$$

where  $B(x)$  is defined after equation (13).

**Corollary 1:** Assume conditions (A1)-(A7), a  $(1-\alpha)100\%$  bootstrap confidence band for the EPK  $\mathcal{K}(x)$  is:

$$[f(x) : \sup_{x \in E} \{|\hat{\mathcal{K}}(x) - f(x)| \widehat{\text{Var}}(\hat{\mathcal{K}})^{-1/2}\} \leq L_\alpha^*],$$

where the bound  $L_\alpha^*$  satisfies

$$P^*[-\{\mathbf{U}(x)^{-1} \mathbf{H}_{n_q}^{-1} \mathbf{A}_{n_q}^*(x)\}_{3,3} / \{B(x)^{-1} \mathbf{L}^{-1} \mathbf{P} \mathbf{L}^{-1}\}_{3,3} \leq L_\alpha^*] = 1 - \alpha.$$

The estimator  $\widehat{\text{Var}}(\hat{\mathcal{K}})$  is computed in a similar fashion as in the previous section.

## 2.2 Confidence Bands based on Smoothing Implied Volatility

Although the nonparametric estimator of the PK is reasonable in theoretical sense, it often fails to provide stable and economically treatable estimators with real data. One way to stabilizing the empirical SPD is the use of data-driven local bandwidths (see Vieu 1993) or a multiple-testing-type adaptive technique of Lepski and Spokoiny (1997). These alternative methods are tools of general purpose and address the bias-variance trade-off locally. They are known to be either asymptotically optimal or to have a near oracle property. Although the adaptive

bandwidth provides us with optimal estimators, it is still possible that the noise is too large and the algorithm fails to provide a curve with a small bias and easy interpretation. This point is stressed in Rookley (1997): “implied volatilities on the other hand tend to be less volatile and differences in implied volatilities convey much more economic information than option prices alone, as implied volatilities already embed much of the fundamental information available.”

We follow, however, an alternative approach and stabilize the empirical SPD by a two-step procedure as in Rookley (1997) and Fengler (2005). At the first step, we estimate the implied volatility (IV) function by a local polynomial regression. At the second step, we plug the smoothed IV into the BS formula to obtain a semiparametric estimator of the option price. This approach relies on a bijective transformation of the call prices to the IV space and reflects the tendency of investors to quote the options in terms of IV. Aït-Sahalia and Lo (1998) used a similar semiparametric technique for dimension reduction purposes. Note that the procedure does not require the BS model to hold, but leads to finite sample improvements, while being asymptotically equivalent to the original estimator (see Theorem 3). Thus we impose more assumptions on the functional form of the call price function and focus only on the nonparametric structure of the volatility surface. Thus the noise is relevant only in the estimation of the volatility surface. However, it is well recognized that the volatility is less noisy and its shape is more tractable and easy to interpret economically. Moreover, we can improve our two-step procedure further by adopting adaptive techniques for the volatility surface.

Formally we smooth the IV using a local polynomial regression in moneyness  $M$ , with the implicit assumption on the pricing formula is homogenous of degree 1 w.r.t. the asset price and the strike price as proved in Renault (1997). In the absence of dividends, the moneyness is defined at time  $t$  as  $M_{it} = S_t/X_i$ . The heteroscedastic model for the IV is given by:

$$\sigma_i = \sigma(M_{it}) + \sqrt{\eta(M_{it})}v_i, \quad i = 1, \dots, n_q, \quad (20)$$

where  $v_i$  are the i.i.d. errors with zero mean, unit variance and  $\eta(\cdot)$  is the volatility function. We make the same assumptions about the implied volatility  $\sigma(\cdot)$  as we did for the option prices  $C(\cdot)$  in Section 1.1.

Defining the rescaled call option price  $c(M_{it}) = C(X_i)/S_t$ , we obtain from the BS formula

$$c(M_{it}) = c\{M_{it}; \sigma(M_{it})\} = \Phi\{d_1(M_{it})\} - \frac{e^{-r\tau} \Phi\{d_2(M_{it})\}}{M_{it}},$$

where

$$d_1(M_{it}) = \frac{\log(M_{it}) + \left\{r + \frac{1}{2}\sigma(M_{it})^2\right\}\tau}{\sigma(M_{it})\sqrt{\tau}}, \quad d_2(M_{it}) = d_1(M_{it}) - \sigma(M_{it})\sqrt{\tau}.$$



Combining the result of Breeden and Litzenberger (1978) with the expression for  $c(M_{it})$  leads to the SPD

$$q(x) = e^{r\tau} \frac{\partial^2 C}{\partial X^2} \Big|_{X=x} = e^{r\tau} S_t \frac{\partial^2 c}{\partial X^2} \Big|_{X=x} \quad (21)$$

with

$$\frac{\partial^2 c}{\partial X^2} = \frac{d^2 c}{dM^2} \left( \frac{M}{X} \right)^2 + 2 \frac{dc}{dM} \frac{M}{X^2}. \quad (22)$$

As it is shown in the Appendix the derivatives in the last expression can be determined explicitly and are functions of  $V \stackrel{\text{def}}{=} \sigma(M)$ ,  $V' \stackrel{\text{def}}{=} \partial \sigma(M) / \partial M$  and  $V'' \stackrel{\text{def}}{=} \partial^2 \sigma(M) / \partial M^2$ . We estimate the latter quantities by the nonparametric local polynomial regression for the IV of the form

$$\sigma(M_{it}) \approx V(M) + V'(M)(M_{it} - M) + \frac{1}{2} V''(M)(M_{it} - M)^2,$$

for  $M$  near  $M_{it}$ . The respective estimators are denoted by  $\hat{V}$ ,  $\hat{V}'$  and  $\hat{V}''$ . Plugging the results into (21)-(22) we obtain the estimator of SPD in the smoothed IV space. Assuming that the IV process fulfills the (A1)–(A7) in the appendix instead of C(·), we conclude that Theorem 2.1 of Claeskens and Van Keilegom (2003) holds also for  $\hat{V}$ ,  $\hat{V}'$  and  $\hat{V}''$ . Note that the convergence rate of  $\hat{V}$  and  $\hat{V}'$  is lower than of  $\hat{V}''$ . Relying on this fact, we state the asymptotic behavior of  $\hat{q}(x) - q(x)$  in the next theorem.

**Theorem 3:** *Let  $\sigma(\cdot)$  satisfy the assumptions (A1)–(A7). Then with  $M = S_t/x$  it holds*

$$\sqrt{n_q h_{n_q}^5} \{ \hat{q}(x) - q(x) \} \xrightarrow{\mathcal{L}} N\{0, r(M)^2 \sigma_V^2(M)\}, \quad (23)$$

where

$$r(M) \stackrel{\text{def}}{=} e^{r\tau} S_t \frac{M^2}{x^2} \left[ \varphi\{d_1(M)\} \left\{ \sqrt{\tau}/2 - \frac{\log(M) + r\tau}{V(M)^2 \sqrt{\tau}} \right\} - e^{-r\tau} \varphi\{d_2(M)\} \left\{ -\sqrt{\tau}/2 - \frac{\log(M) + r\tau}{V(M)^2 \sqrt{\tau}} \right\} / M \right]$$

and  $\sigma_V^2(M) \stackrel{\text{def}}{=} [B_V(M)^{-1} \mathbf{L}^{-1} \mathbf{T} \mathbf{L}^{-1}]_{(3,3)}$ , with  $\sigma_V^2(M)$  defined as in (15).

Proof. The proof is given in the Appendix. ■

Theorem 3 allows us to construct the confidence bands of the SPD estimated semiparametrically using the confidence bands for the IV. The variance of the estimator is obtained by the delta method in the following way

$$\text{Var}\{\hat{q}(x) - q(x)\} = \left( \frac{\partial q}{\partial V''} \right)^2 \text{Var}\{\hat{V}''(M) - V''(M)\}.$$

The variance  $\text{Var}\{\hat{V}''(M) - V''(M)\}$  is estimated using a sandwich estimator similarly to (16), and  $\frac{\partial q}{\partial V''} = e^{r\tau} S_t \frac{M^2}{x^2} \left[ \varphi\{d_1(M)\} \left\{ \sqrt{\tau}/2 - \frac{\log(M)+r\tau}{V(M)^2\sqrt{\tau}} \right\} - e^{-r\tau} \varphi\{d_2(M)\} \left\{ -\sqrt{\tau}/2 - \frac{\log(M)+r\tau}{V(M)^2\sqrt{\tau}} \right\} / M \right]$ . Here we have proved that it is sufficient to consider only the variance of the second derivative of  $V$ , as the other terms involved are of higher order.

### 2.3 Extension to Dependent Data

In the previous sections we have assumed independent data in the estimation of both the historical density and the SPD. Violation of this assumption may lead to misspecified asymptotic results and wrong confidence bands. The assumption is, however, feasible in our study for both densities. The confidence bands for a simple density estimator of time series data are analysed by Liu and Wu 2010, see Section 2.1 and can be directly transferred to the historical density in our setup. Note however, that the impact of these results on the confidence bands for the EPK is flattened by the higher convergence rate of the historical density. Regarding the SPD note that assuming sufficient liquidity we can use options only with a given maturity traded on a single day. This implies that the data used to estimate SPD is not a time series data and there is no need to take the serial correlation into account.

Nevertheless, to serve a general purpose estimation, it is still interesting to generalize our theoretical results to dependent data. Liu and Wu (2010) developed uniform confidence bands for kernel density estimators and Nadaraya–Watson estimation for a general class of time series models. In this section we adopt their approach to our problem. We extend the setup to time dependence and consider the model as in (6)

$$Y_i = C(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n_q,$$

with the strike prices being a causal stationary process  $X_i = G(\dots, \eta_{i-1}, \eta_i)$ ,  $\eta_i$  are i.i.d. and independent with  $\varepsilon_i$ .

We focus on the estimation of  $C(\cdot)$  as in (8) and keep  $\sigma(\cdot)$  known for the derivations in this subsection. The physical dependence measure  $\theta_{(i,\gamma)} \stackrel{\text{def}}{=} \|X_i - X'_i\|_\gamma = (E|X_i - X'_i|^\gamma)^{1/\gamma}$ , where  $X'_i$  is a coupled process of  $X_i$  with  $\eta_0$  is replaced by an i.i.d. copy of  $\eta'_0$ , i.e.  $X'_i = G(\eta'_0, \dots, \eta_{i-1}, \eta_i)$ . Additionally define the dependence measure with coupled whole past as  $\Psi_{i,\gamma} \stackrel{\text{def}}{=} \|G(\eta_0, \dots, \eta_{i-1}, \eta_i) - G(\eta'_0, \dots, \eta'_{i-1}, \eta'_i)\|_\gamma$ . Suppose that  $\|X_i\|_\kappa \leq \infty$  for some  $\kappa > 0$ . Let  $\kappa' = \min(\kappa, 2)$  such that  $\Theta_{n_q} = \sum_{i=1}^{n_q} \theta_{(i,\kappa')}^{1/2}$ . Define  $Z_{n_q} \stackrel{\text{def}}{=} \sum_{k=-n_q}^{\infty} (\Theta_{n_q+k} - \Theta_k)^2$  and  $\tilde{\xi}_i \stackrel{\text{def}}{=} (\dots, \varepsilon_{i-1}, \varepsilon_i, \dots, \eta_{i-1}, \eta_i)$ .

(A8) Assume ( $\|X_i\|_\kappa \leq \infty$  for  $\kappa > 0$ ). The density of  $\eta_i$  is positive and uniformly bounded over its whole support up to the third derivative. There exists a constant  $M < \infty$  such that  $\sup[|f_{X_{n_q}|\tilde{\xi}_{n_q-1}}(x)| + |f'_{X_{n_q}|\tilde{\xi}_{n_q-1}}(x)| + |f''_{X_{n_q}|\tilde{\xi}_{n_q-1}}(x)|] \leq M$  almost surely.  $\varepsilon_i$  has bounded fourth moments.  $\Psi_{n_q,\gamma} = \mathcal{O}(n_q^{-r})$  for some  $\gamma$  and  $r > \delta_1/(1 - \delta_1)$ ,

$0 < \delta_1 < 1/4$ . There exists a constant  $\delta$  such that  $0 < \delta \leq \delta_1 < 1$  and  $h_{n_q} = \mathcal{O}(n_q^{-\delta})$ ,  $n_q^{-\delta} = \mathcal{O}(h_{n_q})$ . Furthermore  $\theta_{(n_q, \kappa)} = \mathcal{O}(\rho^{n_q})$  for some  $\kappa > 0$  and  $0 < \rho < 1$ .

Let  $\hat{\mathcal{F}}_{n_q}(x)$  be the standardized process:

$$\hat{\mathcal{F}}_{n_q}(x) \stackrel{\text{def}}{=} n_q^{1/2} h_{n_q}^{5/2} \{\hat{q}(x) - q(x)\} / [\hat{\sigma}_q(x)]^{1/2}.$$

**Theorem 4:** Under assumptions (A1)-(A6), (A8),  $h_{n_q} = \mathcal{O}\{(n_q \log n_q)^{-1/9}\}$ ,  $\mathcal{Z}_{n_q} h_{n_q}^3 = \mathcal{O}(n_q \log n_q)$ , it follows

$$\mathbb{P} \left[ (-2 \log h_{n_q})^{1/2} \left\{ \sup_{x \in E} |\hat{\mathcal{F}}_{n_q}(x)| - c_{n_q} \right\} < z \right] \rightarrow \exp\{-2 \exp(-z)\},$$

where  $c_{n_q} = (-2 \log h_{n_q})^{1/2} + (-2 \log h_{n_q})^{-1/2} \{x_\alpha + \log(R/2\pi)\}$ .

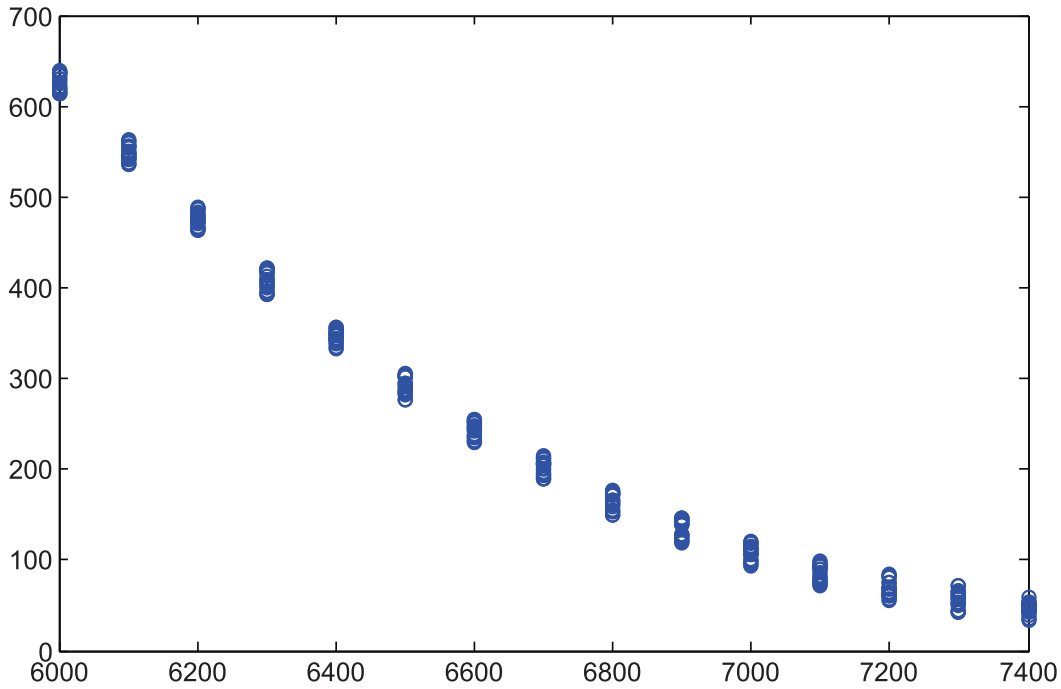
Liu and Wu (2010) note an interesting dichotomy phenomenon, where the rate of convergence is the consequence of an interplay between the strength of dependency and the bandwidth  $h_{n_q}$ . Accordingly, we suggest to undersmooth as a smaller bandwidth would both reduce the bias and the effect of dependency. The rate of  $h_{n_q}$  is set to  $\mathcal{O}\{(n_q \log n_q)^{-1/9}\}$ .

### 3 MONTE-CARLO STUDY

The practical performance of the above theoretical considerations is investigated via two Monte-Carlo studies. The first simulation aims at evaluating the performance under the BS hypothesis, while the second simulation setup does the same under a realistically calibrated surface. The confidence bands are applied to DAX index options. We first study the confidence bands under a BS null model (Section 3.1). Naturally, without volatility smile, both the BS estimator and nonparametric estimator are expected to be covered by the bands. While in the presence of volatility smile (Section 3.2), we expect our tests to reject the BS hypothesis in most cases.

#### 3.1 How Well is the BS Model Covered?

In the first setting, we calibrate a BS model on day 20010117 with the interest rate set equal to the short rate  $r = 0.0481$ ,  $S_0 = 6500$ , strike prices in the interval  $[6000, 7400]$ . We refer to Ait-Sahalia and Duarte (2003) on the sources of the noise and use an identical simulation setting, with the noise being uniformly distributed in the interval  $[0, 6]$ . Figure 4 is a scatter plot of generated observations of European call option prices against strikes, the data is clustered in discrete values of the strike price. Recall that bandwidths in the following context are all selected to



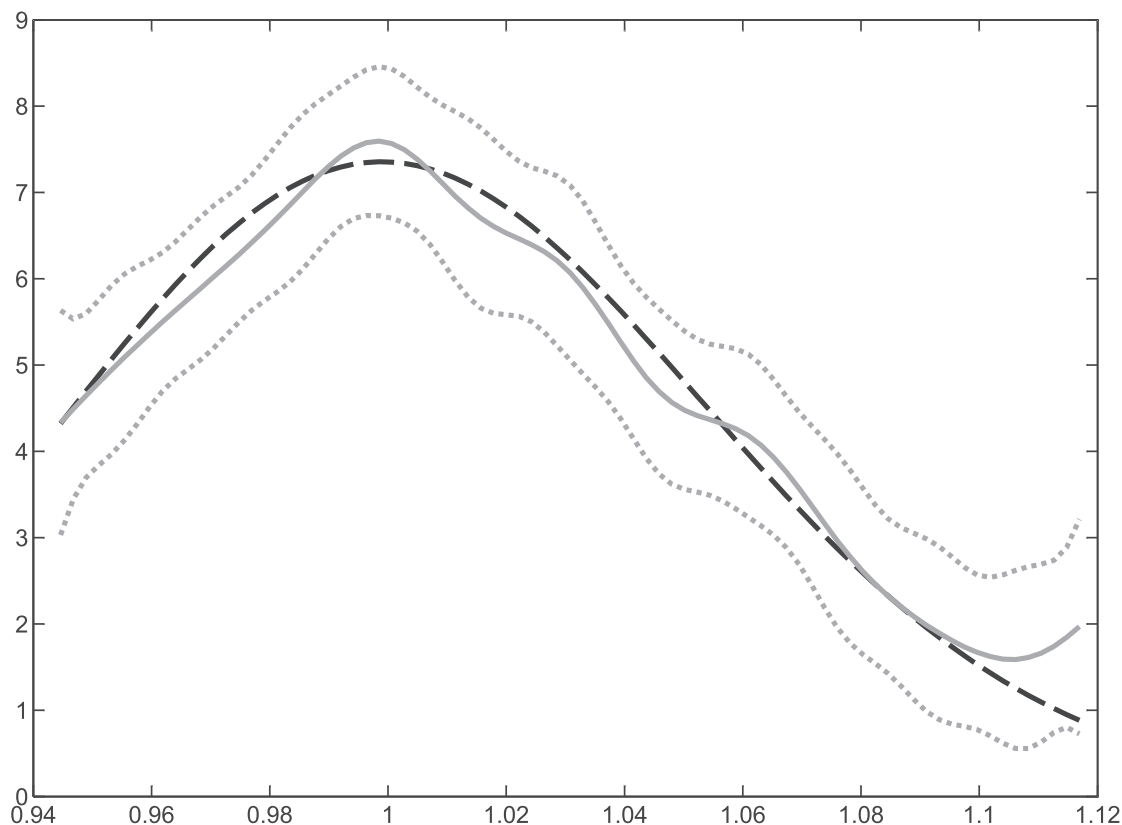
**Figure 4** Generated noisy BS call option prices against strike prices.

minimize pricing error using a leave-one-out approach on the bivariate grid  $[1/n_p; 1] \times [1/n_q; 1]$ .

Figure 5 shows a nonparametric estimator for the SPD and a parametric BS estimator. The two estimators roughly coincide except for a small wiggle, thus the bands drawn around the nonparametric curve also fully cover the parametric one. The accuracy is evaluated by calculating the coverage probabilities and average area within the bands, see Table 1 (see the rows labeled “null”). The coverage probabilities is determined via 500 simulations, whenever the hypothesized curve calculated on a grid of 100. The coverage probability approaches its nominal level with increasing sample size, but never reaches it. This may well be attributed to the above mentioned poor convergence of Gaussian maxima to the Gumbel distribution. The area within the bands reflects the stability of the estimation procedure. The bands get narrower with increasing sample size.

The bias correction for the SPD follows the approach of Xia (1998). The HD is corrected as in Jones, Linton, and Nielson (1995). The bias correction for EPK relies on the linear term from Lemma 2. The correction of the bands mimics the Bonferroni correction in Eubank and Speckman (1993) and is based on the asymptotic confidence intervals in (13) and (14). We conclude that the bias correction approach and the Bonferroni correction are not better than the proposed method for all sample sizes.

HDs are estimated from simulated stock prices following geometric Brownian motion with  $\mu = 0.23$ . A BS EPK estimator could be tested using the above procedure. Due to boundary effects, we concentrate on moneyness ( $M_t = S_t/X$ ) in



**Figure 5** Estimation of SPD (gray), bands (dotted) and the BS SPD (dashed), with  $h_{n_q} = 0.085$ ,  $\alpha = 0.05$ ,  $n_q = 300$ .

[0.95, 1.1]. Figure 6 displays the nonparametric EPK with confidence band and the BS EPK covered in the band. We observe that the BS EPK is strictly monotonically decreasing. The summary statistics are given in Table 1, due to the additional source of randomness introduced through the estimation of  $p(x)$ , the coverage probabilities are less precise than the corresponding coverage probabilities for SPD. Nevertheless, the probabilities are getting closer to their nominal values and the bands get narrower when the sample size increases.

### 3.2 How Well is the Band in Reality?

Section 3.1 studied the performance of the bands under the null hypothesis with BS assumption, while this section is designed to investigate the performance of the bands when the null hypothesis is violated by a realistic volatility smile observed in the market. Keeping the parameters identical to the setup of the first study, we generated the data with a smoothed volatility function based on data for or options traded on 20010117 with  $\tau = 3M, 6M$  to maturity.

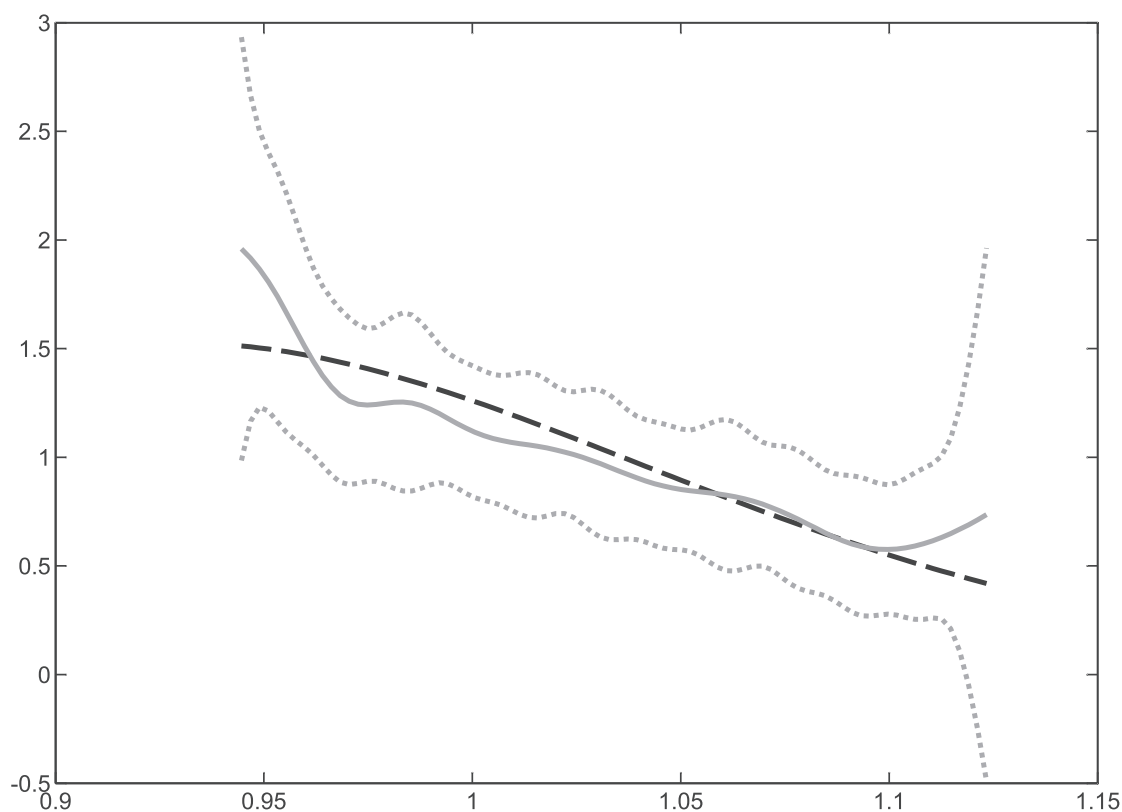
Figures 7 and 8 report the estimators for SPD and EPK. The bands do not cover the BS estimator. Correspondingly, Table 1 (see the rows labeled "alter.") show the

**Table 1** Averaged, coverage probability (area) of the uniform confidence band over 500 simulations in different cases

Level	Maturity	Method	$n_q=300$	450	600
5% (null)	3M	EPK	0.782 (2.54)	0.798 (2.49)	0.802 (2.38)
		EPK (bias)	0.798 (2.51)	0.800 (2.43)	0.802 (2.31)
		EPK (Bonfer.)	0.673 (2.98)	0.697 (2.88)	0.754 (2.76)
		SPD	0.906 (2.40)	0.914 (2.20)	0.923 (1.99)
		SPD (Bonfer.)	0.873 (2.68)	0.924 (2.57)	0.929 (2.44)
	6M	EPK	0.860 (2.50)	0.875 (2.43)	0.890 (2.41)
		EPK (bias)	0.862 (2.53)	0.883 (2.41)	0.899 (2.42)
		EPK (Bonfer.)	0.785 (2.90)	0.801 (2.67)	0.824 (2.74)
		SPD	0.896 (2.44)	0.906 (2.13)	0.920 (2.07)
		SPD (Bonfer.)	0.883 (2.73)	0.894 (2.69)	0.903 (2.52)
10%(null)	3M	EPK	0.706 (2.47)	0.736 (2.34)	0.762 (2.23)
		EPK (bias)	0.712 (2.45)	0.737 (2.33)	0.771 (2.23)
		EPK (Bonfer.)	0.673 (2.33)	0.686 (2.12)	0.734 (2.01)
		SPD	0.795 (2.17)	0.812 (2.06)	0.853 (1.88)
		SPD (Bonfer.)	0.764 (2.12)	0.801 (2.00)	0.833 (1.98)
	6M	EPK	0.729 (2.50)	0.774 (2.23)	0.829 (2.31)
		EPK (bias)	0.713 (2.47)	0.753 (2.26)	0.835 (2.30)
		EPK (Bonfer.)	0.671 (2.86)	0.745 (2.88)	0.798 (2.72)
		SPD	0.800 (2.34)	0.814 (2.08)	0.860 (1.94)
		SPD (Bonfer.)	0.763 (2.55)	0.800 (2.46)	0.847 (2.36)
5% (alter.)	3M	EPK	0.512 (2.43)	0.178 (2.23)	0.050 (2.02)
		EPK (bias)	0.543 (2.42)	0.235 (2.27)	0.145 (1.99)
		EPK (Bonfer.)	0.372 (2.51)	0.239 (2.37)	0.099 (2.12)
	6M	EPK	0.592 (2.53)	0.410 (2.17)	0.178 (2.02)
		EPK (bias)	0.541 (2.49)	0.349 (2.12)	0.251 (2.01)
		EPK (Bonfer.)	0.331 (2.34)	0.136 (2.16)	0.150 (2.15)
10% (alter.)	3M	EPK	0.258 (2.12)	0.050 (2.04)	0.030 (2.01)
		EPK (bias)	0.268 (2.13)	0.043 (2.01)	0.001 (2.00)
		EPK (Bonfer.)	0.148 (2.78)	0.030 (2.61)	0.001 (2.54)
	6M	EPK	0.375 (2.22)	0.410 (2.13)	0.178 (2.00)
		EPK (bias)	0.362 (2.21)	0.432 (2.13)	0.176 (2.01)
		EPK (Bonfer.)	0.231 (2.46)	0.221 (2.35)	0.110 (2.25)

(bias) means bias correction, (Bonfer.) means Bonferoni correction.

coverage probabilities, which rapidly decrease when sample sizes are increasing. However, the area within the bands does not change significantly when compared with the results of Section 3.1. We conclude that the confidence bands are useful for detecting the deviation from the BS model.



**Figure 6** Estimation of EPK (gray), bands (dotted), and the BS EPK (dashed), with  $h_{n_q} = 0.085$ ,  $h_{n_p} = 0.060$  and  $\alpha = 0.05$ ,  $n_p = 2000$ ,  $n_q = 300$ .

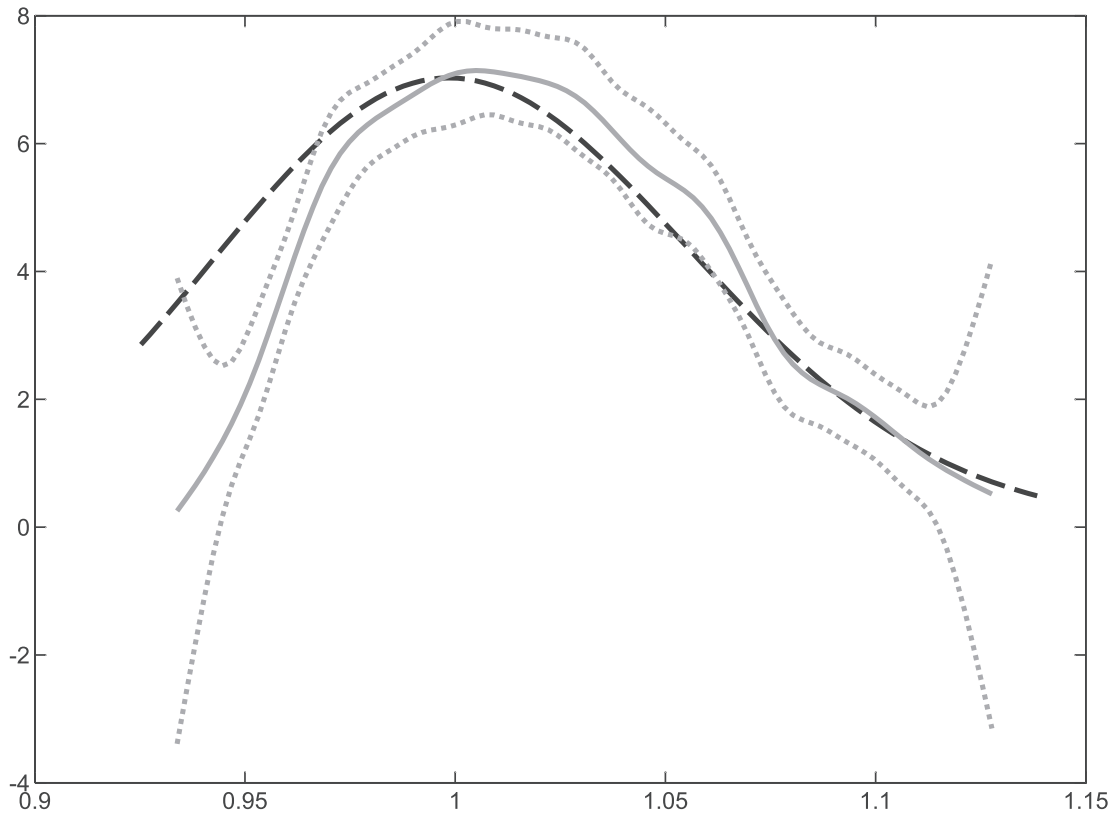
Note that the suggested method works for other processes, for example, the Heston model, which allows for flexible forms of the pricing kernels. The simulation study confirms the good performance of the confidence bands. The results are not reported here for the brevity of presentation.

#### 4 AN ILLUSTRATION WITH DAX DATA

This section aims at illustrating the functionality of our bands by checking the coverage of BS EPK, which indicates how much the market risk behavior deviates from the BS model. The procedure can be seen as a test of monotonicity of pricing kernels. The available tests for monotonicity (Ghosal, Sen, and van der Vaart (2000), Lee, Linton, and Whang (2009), Chetverikov (2012)) work for (regression) functions and not for derivative estimation as required here. We take a dynamic point of view by considering the EPK estimated at different dates.

##### 4.1 Data

In contrast to previous studies that are mainly based on S&P500 data, we focus on intraday European options on the DAX options. The source is the



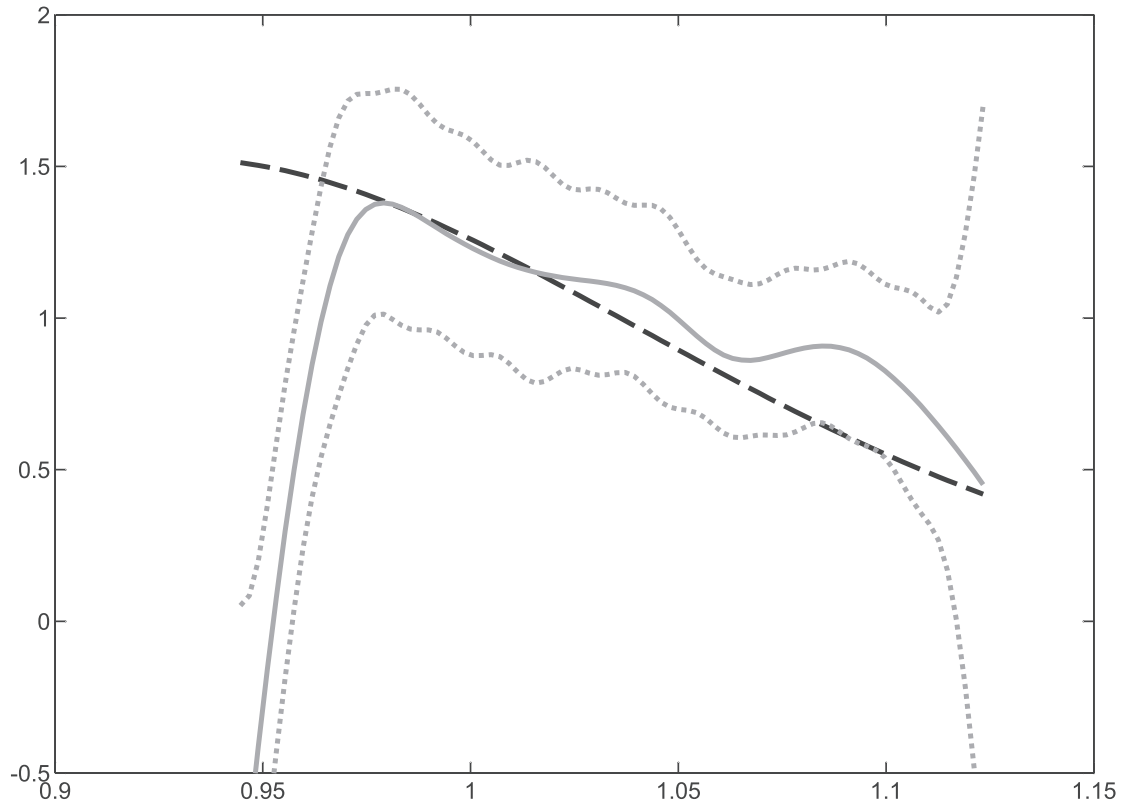
**Figure 7** Plot of confidence bands (dotted), nonparametrically estimated SPD (gray), the fitted BS (dashed) SPD with simulated volatility smile,  $n_q = 300$ ,  $h_{n_q} = 0.066$ ,  $\alpha = 0.05$ .

European Exchange EUREX and data available by C.A.S.E., RDC SFB 649 (<http://sfb649.wiwi.hu-berlin.de>) in Berlin. The extracted observations for our analysis cover the period between 1998 and 2008. The smoothing in volatility approach described in Section 2.2 is applied to estimate the EPK (denoted as Rookley method). As we cannot find traded options with the same maturity on each day, we consider options with maturity 15 days (10 trading days) across several years. Specifically, we extract a time series of options for every month from January 2001 to December 2006; this adds up to 63 days.

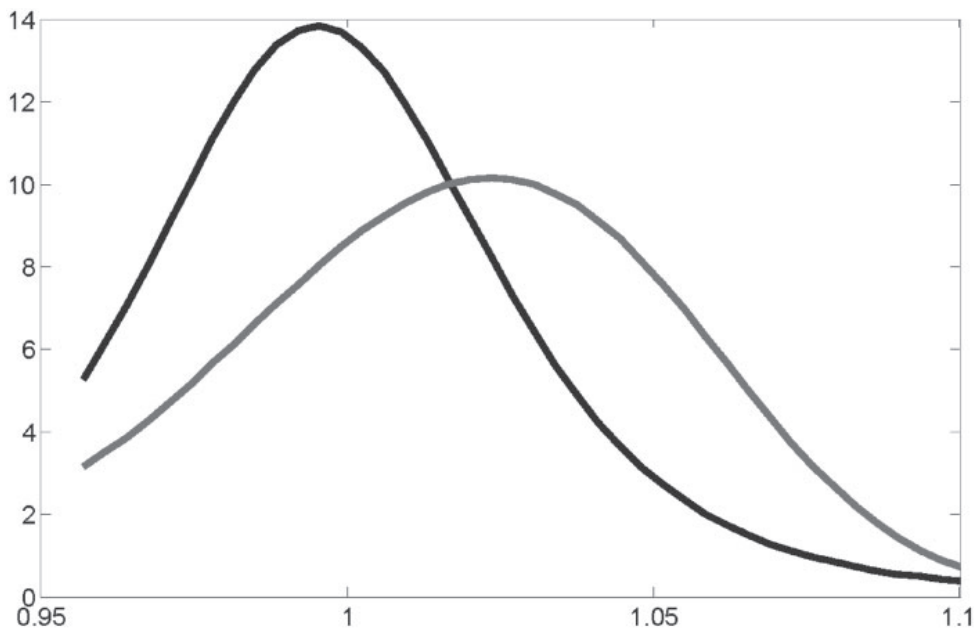
To make sure that the data correctly represents the market conditions, we use several cleaning criteria. In our sample, we eliminate the observations with  $\tau < 1D$  and  $IV > 0.7$ . Also, we skip the option quotes violating general no-arbitrage condition i.e.,  $S > C > \max\{0, S - Xe^{-r\tau}\}$ . Due to the put-call parity, both out-of-the-money call options and in-the-money puts are used to compute the smoothed volatility surface. The median of intra day stock prices is used to compute the SPD. We use a window of 500 returns for nonparametric kernel density estimators of HD.

Figure 9 describes the relative position of the HD and SPD on a specific day, the EPK peak is apparently created through the different probability mass contributions at different moneyness states.

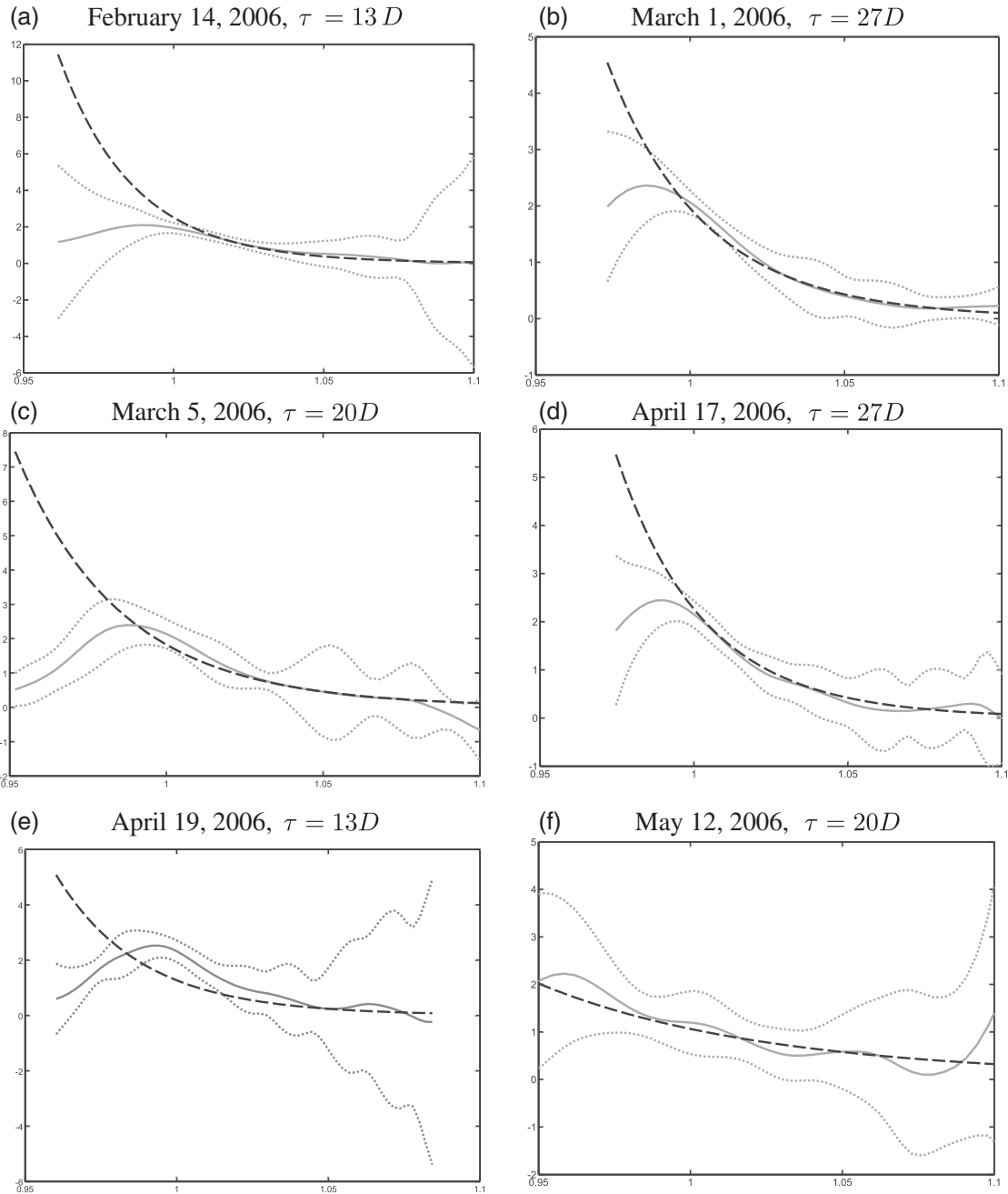




**Figure 8** Plot of confidence bands (dotted), nonparametrically estimated EPK (gray), the fitted BS (dashed) EPK with simulated volatility smile  $n_p=2000$ ,  $n_q=300$ ,  $h_{n_q}=0.063$ ,  $h_{n_p}=0.011$ ,  $\alpha=0.05$ .



**Figure 9** Plot of estimated SPD on February 28, 2006 (Rookley,  $h_{n_q}=0.063$ , black) and HD ( $h_{n_p}=0.0106$ , gray).

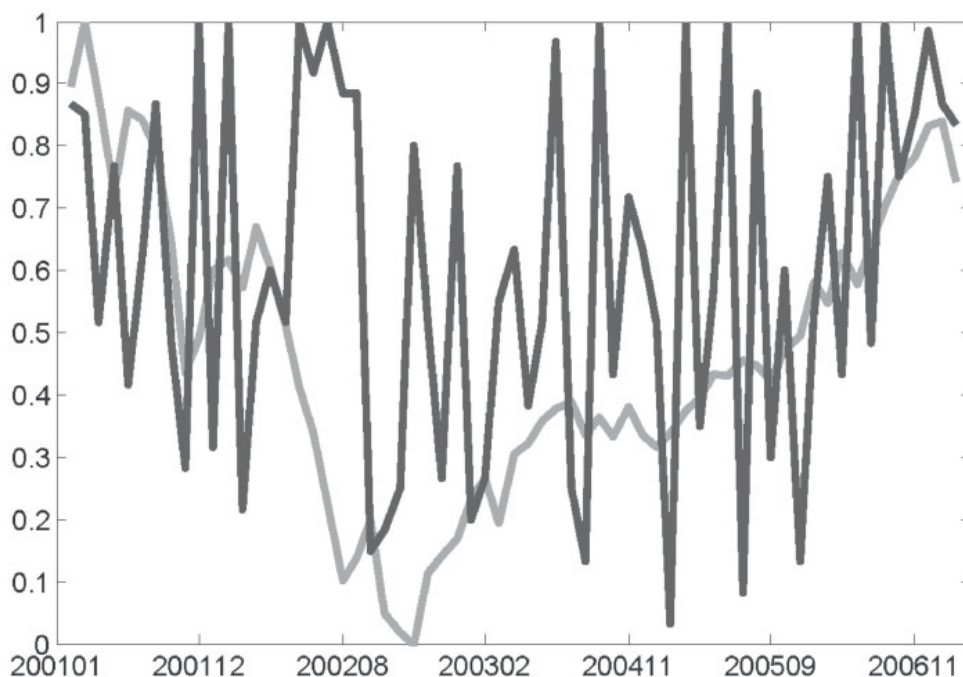


**Figure 10** Estimated BS EPK (dashed), Rookley EPK (gray), uniform confidence band (dotted),  $\alpha = 0.05$ .

## 4.2 Estimation of DAX EPK and its Uniform Confidence Band

We consider two specifications for the pricing kernels. In the first specification, the BS pricing kernels have a marginal rate of substitution with power utility function:

$$\mathcal{K}(M) = \beta_0 M^{-\beta_1}, \quad (24)$$



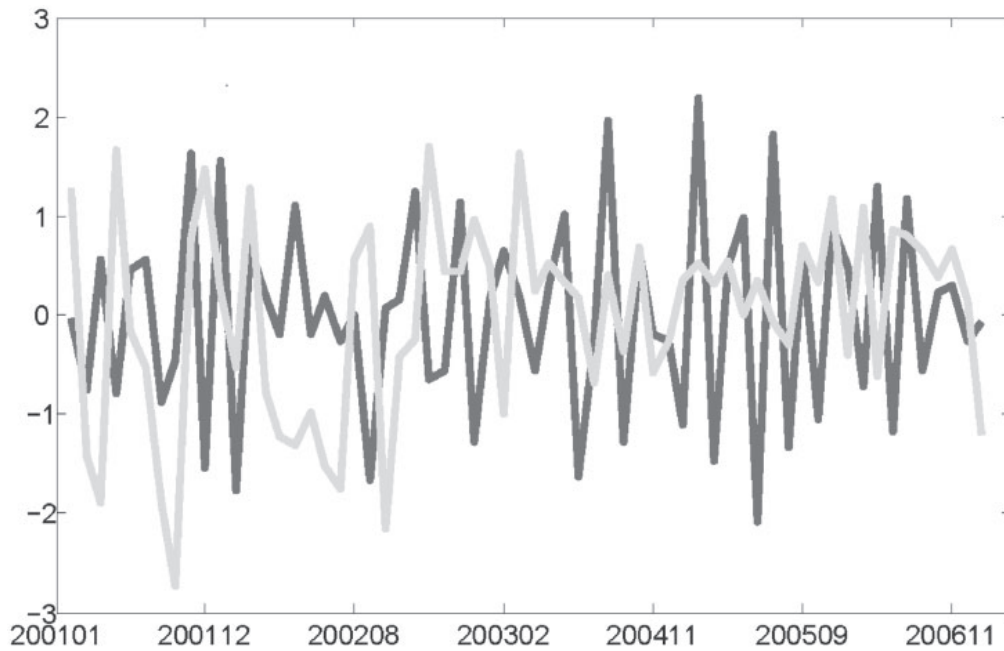
**Figure 11** Coverage probability ( $\alpha=0.05$ ) estimated at 63 trading days and the DAX index (gray, rescaled to  $[0, 1]$ ),  $\tau=3M$ .

where  $\beta_0$  is a scaling factor and  $\beta_1$  determines the slope of pricing kernel. Thus the BS calibration is realized by linearly regressing the (ordered) log-EPK on log-moneyness. In the second specification, we construct the nonparametric confidence bands as described in Section 2.2. A sequence of EPKs and corresponding bands are shown in Figure 10. In most of the cases, the BS EPKs are rejected via the confidence bands. The amount of deviation from the hypothesized BS specification though provides us valuable information about how risk hungry investors are. Besides, the area of between the bands varies over time, which gives us insights into the variabilities of the prevailing risk patterns. In sum, the bands do not only provide a simple test for hypothesized EPKs, but also help us to study the dynamics of risk patterns over time.

### 4.3 Linking Economic Conditions to EPK Dynamics

We use two different indicators for the deviation from a simple BS model. As an approximation to the coverage probability, we calculate the proportion of grid points of the band which covers the BS EPK. As a second measure, we introduce the average width of the confidence bands over the moneyness interval  $[0.95, 1.1]$  as a proxy for the area between the confidence bands. This provides us with a measure of variability, see also Theorem 2.

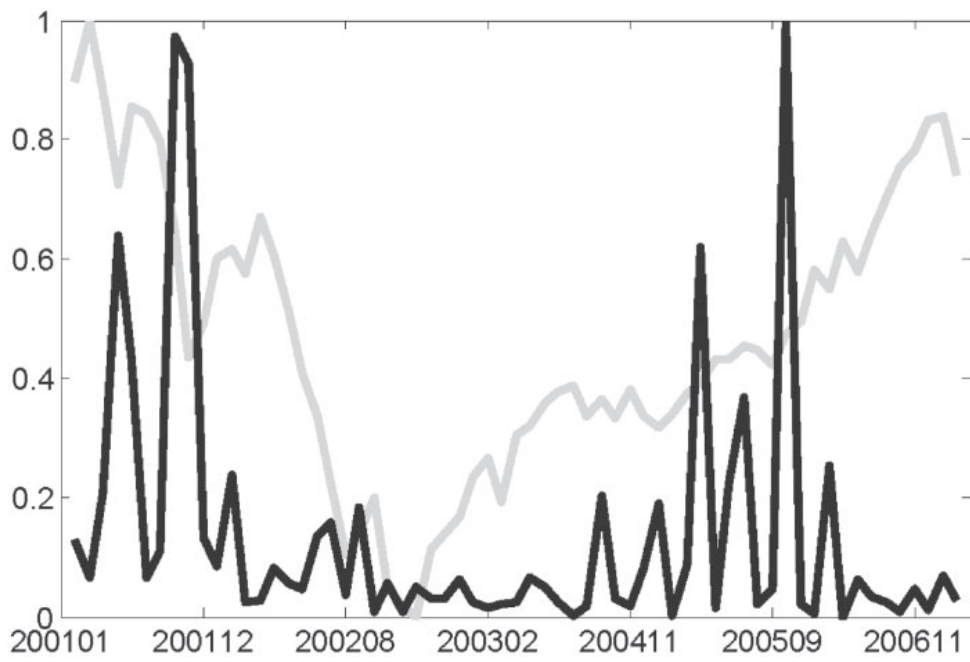
The first risk pattern time series is given in Figure 11, where we display the DAX index (scaled to  $[0, 1]$ ) together with the coverage probability. We discover that



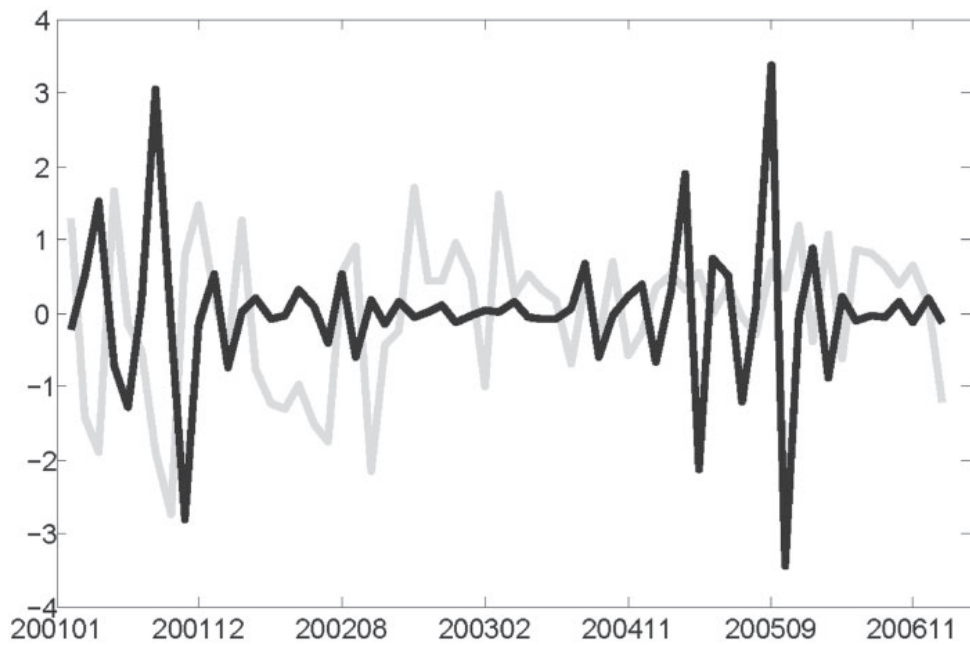
**Figure 12** First difference of coverage probability and the DAX index return (gray, standardized).

the coverage probability becomes less volatile when the DAX index level is high. Figure 12 shows the differenced time series. From a simple correlation analysis, we argue that the change in coverage probability and DAX return (with a lag of 3M) are highly negatively correlated (correlation  $-0.3543$ ) when the DAX index goes down (200101-200302). On the contrary, in the period when the DAX goes up, one observes a large positive correlation ( $0.3151$ ). What does this mean economically? This implies in a period of worsening economic condition, a positive part of the monthly DAX returns induces a greater hunger for risk in a delay horizon of 3 months. Positive returns have just the opposite effects. With boosting and bullish markets, the positive correlation indicates a 3-month horizon of decreasing risk aversion. The exercise we have done so far support these economic reasoning. Risk aversion seems to be higher in recessions and lower in boom times. This corresponds to the findings in the economics literature. Economic agents (e.g., corporate firms, banks, households) with higher risk aversion tend to hold more liquid assets, driving down the interest rate. At the same time, the higher risk aversion calls for a higher rate of return on risky assets. A lower interest rate and a higher rate on risky assets generate a higher risk premium, Gilchrist and Zakrajšek (2012).

As far as the average width of the bands is concerned, we may conclude from Figure 13 and Figure 14 that in periods of clearly bullish or bearish momentum, the volatility of the width of the confidence band is higher. This may be caused by the uncertainty of the market participants about the long-term persistence of the trend. The lag effect on risk hunger is also detectable for this constructed indicator. Over the whole observation interval, the correlation between the monthly DAX return and the change in the average width is  $-0.3230$  for a 1M lag and  $-0.2717$  for 3M.



**Figure 13** Area of the confidence bands ( $\alpha=0.05$ ) estimated at 63 trading days and the DAX index (gray, rescaled to  $[0, 1]$ ).



**Figure 14** First difference of area and the DAX index return (gray, standardized).

## 5 CONCLUSIONS

Pricing Kernels are important elements in understanding investment behavior since they reflect the relative weights given by investors' states of nature (Arrow–Debreu securities). Pricing kernels may be deduced in either parametric or nonparametric approaches. Parametric approaches like a simple BS model are too restrictive to account for the dynamics of the risk patterns, which induces the well-known EPK paradox. Nonparametric approaches allow more flexibility and reduce the modeling bias. Simple tools like uniform confidence bands help us to conduct tests against any parametric assumption of the EPKs i.e., shape inspection. Considering the numerical stability, we smooth the IV surface via the Rookley's method, and obtain SPD estimator.

We have studied systematically the methodology of constructing the uniform bands for both semiparametric or nonparametric estimators. Based on the confidence bands, we explored two indicators to measure risk aversion over time and linked it with DAX index; the first one is the coverage probability measuring the proportion of the BS curve covered in bands, while the second one is the area indicator measuring the variability of the estimator. We found out that there are strong correlations between DAX index and our indicators with lag effects. The smooth bootstrap is also studied without a significant improvement in finite sample performance. One interest further extension is employing robust smoothers to improve the bootstrap performance.

## 6 APPENDIX

Assumptions:

(A1)  $h_{n_q} \rightarrow 0$  in such a way that  $\{\log n_q / (n_q h_{n_q})\}^{1/2} \cdot h_{n_q}^3 \rightarrow 0$ , and the optimal rate bandwidth, to guarantee undersmoothing, would be  $\mathcal{O}\{(\log n_q \cdot n_q)^{-1/9}\}$ .

(A2) The kernel functions  $K \in C^{(1)}[-1, 1]$  (adopted for estimating both HD  $p(x)$  and SPD  $q(x)$ ) are symmetric and takes value 0 on the boundary.

(A3) For the likelihood function  $\mathcal{L} \in C^{(1)}(E)$  it holds that  $\inf_{x \in E} \mathcal{L}(x) > 0$ .  $C(x) \in C^{(4)}(E)$ . Additionally the third partial derivatives of  $\mathcal{L}(Y, C)$  with respect to  $C$  exists and is continuous in  $C$  for every  $y$ . The Fisher information  $I(C(x))$  has a continuous derivative and  $\inf_{x \in E} I\{C(x)\} > 0$ .

(A4) There exists a neighborhood  $N(C(x))$  such that

$$\max_{k=1,2} \sup_{x \in E} \left\| \sup_{C \in N\{C(x)\}} \frac{\partial^k}{\partial C^k} \mathcal{L}(y; C) \right\|_\lambda < \infty$$

for some  $\lambda \in (2, \infty]$ . Furthermore

$$\sup_{x \in E} E \left[ \sup_{C \in N\{C(x)\}} \left| \frac{\partial^3}{\partial C^3} \mathcal{L}(y; C) \right| \right] < \infty.$$

(A5) The HD of underlying  $p(x)$  is three times continuously differentiable and is bounded by a positive constant from below on the compact set  $E$ .

(A6) Let  $a_{n_p} = (n_p h_{n_p} / \log n_p)^{-1/2} + h_{n_p}^2$  from (10) and  $b_{n_q} = h_{n_q}^{-2} (n_q h_{n_q} / \log n_q)^{-1/2} + h_{n_q}^2$  from Lemma 1. We assume that  $n_q / n_p = \mathcal{O}(1)$ ,  $h_{n_q}^5 / h_{n_p} = \mathcal{O}(1)$ .

(A7) The pricing errors  $\varepsilon_i$  are independent and identically distributed random variables.

(A1) is a bandwidth assumption for estimating SPD. We can undersmooth to reduce the bias. (A2) is the assumption on kernel function which facilitates the derivation of results. In a typical setting,  $h_{n_q}$  is chosen to be as in (A1), while  $h_{n_p}$  is chosen to be  $\mathcal{O}(n_p^{-1/5})$ , and  $h_{n_q}$  is larger than  $h_{n_p}$ . (A4) contains moment conditions defined via likelihood functions. (A5) is an assumption imposed on the smoothness of  $p(x)$ . In our empirical setting,  $n_q = 715$ ,  $n_p = 200$ , thus for a typical data situations, (A6) is reasonable. The assumptions (A1) and (A2) ensure  $a_{n_p} / b_{n_q} = \mathcal{O}(1)$ .

## 6.1 Proof of Lemma 2

Recall from Lemma 1 and (10) that

$$\begin{aligned} \sup_{x \in E} |\hat{p}(x) - p(x)| &= \mathcal{O}\{\{\log n_q / (n_q h_{n_q})\}^{1/2} + h_{n_p}^2\} = \mathcal{O}(a_{n_p}), \\ \sup_{x \in E} |\hat{q}(x) - q(x)| &= \mathcal{O}[h_{n_q}^{-2} \{\log n_q / (n_q h_{n_q})\}^{1/2} + h_{n_q}^2] = \mathcal{O}(b_{n_q}). \end{aligned}$$

To determine the order of the EPK we linearize the ratio  $q(x)/p(x)$ .

$$\hat{\mathcal{K}}(x) - \mathcal{K}(x) = \frac{\hat{q}(x)}{\hat{p}(x)} - \frac{q(x)}{p(x)} = \frac{\hat{q}(x)p(x) - \hat{p}(x)q(x)}{p^2(x)} \cdot \frac{1}{1 + \frac{\hat{p}(x) - p(x)}{p(x)}}. \quad (25)$$

We decompose the first factor as  $\hat{q}(x)p(x) - \hat{p}(x)q(x) = \{\hat{q}(x) - q(x)\}p(x) - \{\hat{p}(x) - p(x)\}q(x)$ , while for the second factor we use the first order Taylor expansion. Putting together we obtain

$$\begin{aligned} \sup_{x \in E} |\hat{\mathcal{K}}(x) - \mathcal{K}(x)| &= \sup_{x \in E} \left| \frac{\hat{q}(x) - q(x)}{p(x)} - \frac{\hat{p}(x) - p(x)}{p(x)} \cdot \frac{q(x)}{p(x)} \right. \\ &\quad \left. - \frac{\{\hat{q}(x) - q(x)\}\{\hat{p}(x) - p(x)\}}{p^2(x)} + \frac{\{\hat{p}(x) - p(x)\}^2}{p^2(x)} \cdot \frac{q(x)}{p(x)} \right|. \end{aligned}$$

The first two elements are of order  $\mathcal{O}(b_{n_q})$  and  $\mathcal{O}(a_{n_p})$  respectively, while the last element is of order  $\mathcal{O}(a_{n_p})$ . Summarizing we conclude that

$$\sup_{x \in E} |\hat{\mathcal{K}}(x) - \mathcal{K}(x)| = \mathcal{O}[\max\{a_{n_p}, b_{n_q}\}].$$

## 6.2 Proof of Theorem 2

The basic idea of the proof is to approximate the process

$$D_{n_q}(x) = n_q^{1/2} h_{n_q}^{5/2} \{\hat{\mathcal{K}}(x) - \mathcal{K}(x)\} / [\widehat{\text{Var}}\{\hat{\mathcal{K}}(x)\}]^{1/2}$$

by a process with nonstochastic variance term, which will then be further approximated by a process that can be treated with the tools of Claeskens and Van Keilegom (2003). Here we have dropped for the simplicity of notation the  $q$  in  $n_q$  and the  $n_q$  in  $h_{n_q}$ . More precisely, we define, as first approximation,

$$D_n^{(1)}(x) \stackrel{\text{def}}{=} n^{1/2} h^{5/2} \{\hat{\mathcal{K}}(x) - \mathcal{K}(x)\} / \{\text{Var}\{\hat{\mathcal{K}}(x)\}\}^{1/2},$$

where  $\text{Var}\{\hat{\mathcal{K}}(x)\}$  is given in (15). Lemma 2 ensures that the approximation by

$$n^{1/2} h^{5/2} \{\hat{q}(x) - q(x)\} / \{p(x) \text{Var}\{\hat{\mathcal{K}}(x)\}\}^{1/2} \quad (26)$$

is uniformly of order  $\mathcal{O}_p\{(\log n)^{-1/2}\}$ . The process in equation (26) can be approximated as in Claeskens and Van Keilegom (2003) by

$$2! \exp(r\tau) h^2 f_X(x)^{-1/2} \text{Var}\{\hat{\mathcal{K}}(x)\}^{-1/2} I\{C(x)\}^{-1/2} \sum_{i=0}^3 f_X(x)^{-1/2} I\{C(x)\}^{-1/2} \{\mathbf{L}^{-1}\}_{3,i+1} A_{n,i}(x) \quad (27)$$

For the definition of the local Fisher information,  $I\{C(x)\}$ , the matrix  $\mathbf{L}$  and the process  $A_{ni}(x)$ , we refer to Section 2, and Section 6. Define

$$Z_{ni}(x) \stackrel{\text{def}}{=} (nh)^{1/2} h^{-i} [I\{C(x)\} f_X(x)]^{-1/2} A_{ni}(x).$$

Then equation (27) can be written as

$$F_n(x) \stackrel{\text{def}}{=} 2! \exp(r\tau) h^2 \{f_X(x)\}^{-1/2} \text{Var}\{\hat{\mathcal{K}}(x)\}^{-1/2} I\{C(x)\}^{-1/2} \sum_{i=0}^3 h^i \{\mathbf{L}^{-1}\}_{3,i+1} Z_{ni}(x)$$

Please note that  $\mathbf{L}$  is not a function of  $x$  as Claeskens and Van Keilegom (2003) erroneously write. Following their line of thoughts, we replace  $Z_{ni}(x)$  (uniformly) by

$$Z'_{ni}(x) = h^{1/2} \int K_h(z-x) \left(\frac{z-x}{h}\right) dz$$

In order to apply corollary A1 of Bickel and Rosenblatt (1973), define the covariance function  $r(x)$  of the Gaussian process  $F_{n_q}(x)$ , and we know that

$$\begin{aligned} r(x) &= \text{Cov}(Z'_{nj}(x), Z'_{nj}(0)) \\ &= C_1 - C_2|x|^2 + \mathcal{O}(|x|^2), \end{aligned}$$

for  $x \in E$ , where  $C_1$  and  $C_2$  are two constants, so the regularity conditions satisfies, the result follows.



Finally, we have to show that  $\sup_{x \in E} |\widehat{\text{Var}}\{\hat{\mathcal{K}}(x)\} - \text{Var}\{\hat{\mathcal{K}}(x)\}| = o_p(1)$ .

$$\begin{aligned} & \sup_{x \in E} |\widehat{\text{Var}}\{\hat{\mathcal{K}}(x)\} - \text{Var}\{\hat{\mathcal{K}}(x)\}| \\ &= \sup_{x \in E} |\widehat{\text{Var}}\left\{\frac{\hat{q}(x) - q(x)}{p(x)}\right\} - \text{Var}\left\{\frac{\hat{q}(x) - q(x)}{p(x)}\right\}| + o_p\{(nh)^{-(1/2+\alpha)}(\log n)^{1+\alpha}\}, \end{aligned}$$

where  $0 < \alpha < 1$ .

According to corollary 2.1 in Claeskens and Van Keilegom (2003), for  $j=3, k=3$ ,

$$\sup_{x \in E} |\widehat{\text{Var}}\{\hat{q}(x)\} - \text{Var}\{\hat{q}(x)\}| = o_p\{(nh \log n)^{-1/2}\}.$$

So we have,

$$\sup_{x \in E} |\widehat{\text{Var}}\{\hat{\mathcal{K}}(x)\} - \text{Var}\{\hat{\mathcal{K}}(x)\}| = o_p(1).$$

### 6.3 Expressions for the Semiparametric Estimator of SPD and Proof of Theorem 3

To prove the statement we show that  $\sqrt{n_q h_{n_q}}(\hat{q}(x) - q(x))$  has asymptotically the same distribution as  $\sqrt{n_q h_{n_q}}\{\hat{V}''(M) - V''(M)\}$  with proper scaling. Thus we derive the following equation.

$$\hat{q}(x) - q(x) = e^{r\tau} S_t \frac{M^2}{x^2} \left[ \varphi\{\hat{d}_1(M)\} \left\{ \sqrt{\tau}/2 - \frac{\log(M) + r\tau}{\hat{V}(M)^2 \sqrt{\tau}} \right\} \right. \quad (28)$$

$$\left. - e^{-r\tau} \varphi\{\hat{d}_2(M)\} \left\{ -\sqrt{\tau}/2 - \frac{\log(M) + r\tau}{\hat{V}(M)^2 \sqrt{\tau}} \right\} / M \right] \{\hat{V}''(M) - V''(M)\}$$

$$+ o\{\hat{V}''(M) - V''(M)\}, \quad (29)$$

where  $\hat{d}_i$  and  $\hat{c}$  are the terms defined in Section 2.2 with  $\hat{V}(M)$  replaced by the true function. We now describe how to derive (29) Taking the derivatives of  $c(M_{it})$  with respect to moneyness ( $M$ ) and noting that both  $d_1(M_{it})$  and  $d_2(M_{it})$  depend on  $M_{it}$  we obtain

$$\begin{aligned} \frac{dc}{dM} &= \varphi(d_1) \frac{dd_1}{dM} - e^{-r\tau} \frac{\varphi(d_2)}{M} \frac{dd_2}{dM} + e^{-r\tau} \frac{\Phi(d_2)}{M^2}, \\ \frac{d^2c}{dM^2} &= \varphi(d_1) \left\{ \frac{d^2d_1}{dM^2} - d_1 \left( \frac{dd_1}{dM} \right)^2 \right\} \\ &\quad - \frac{e^{-r\tau} \varphi(d_2)}{M} \left\{ \frac{d^2d_2}{dM^2} - \frac{2}{M} \frac{dd_2}{dM} - d_2 \left( \frac{dd_2}{dM} \right)^2 \right\} - \frac{2e^{-r\tau} \Phi(d_2)}{M^3} \end{aligned}$$

Computing the first and second order differentials for  $d_1$  and  $d_2$  using the notation  $V = \sigma(M)$ ,  $V' = \partial\sigma(M)/\partial M$  and  $V'' = \partial^2\sigma(M)/\partial M^2$ , we obtain

$$\begin{aligned} \frac{dd_1}{dM} &= \frac{1}{MV\sqrt{\tau}} + \left\{ -\frac{\log(M)+r\tau}{V^2\sqrt{\tau}} + \sqrt{\tau}/2 \right\} V', \\ \frac{dd_2}{dM} &= \frac{1}{MV\sqrt{\tau}} + \left\{ -\frac{\log(M)+r\tau}{V^2\sqrt{\tau}} - \sqrt{\tau}/2 \right\} V', \\ \frac{d^2d_1}{dM^2} &= -\frac{1}{MV\sqrt{\tau}} \left\{ \frac{1}{M} + \frac{V'}{V} \right\} + V'' \left\{ \frac{\sqrt{\tau}}{2} - \frac{\log(M)+r\tau}{V^2\sqrt{\tau}} \right\} \\ &\quad + V' \left\{ 2V' \frac{\log(M)+r\tau}{V^3\sqrt{\tau}} - \frac{1}{MV^2\sqrt{\tau}} \right\}, \\ \frac{d^2d_2}{dM^2} &= -\frac{1}{MV\sqrt{\tau}} \left\{ \frac{1}{M} + \frac{V'}{V} \right\} + V'' \left\{ -\frac{\sqrt{\tau}}{2} - \frac{\log(M)+r\tau}{V^2\sqrt{\tau}} \right\} \\ &\quad + V' \left\{ 2V' \frac{\log(M)+r\tau}{V^3\sqrt{\tau}} - \frac{1}{MV^2\sqrt{\tau}} \right\}. \end{aligned}$$

To prove (29), we know from (22) and (21) that

$$q(x) - \hat{q}(x) = \left( \frac{d^2c}{dM^2} - \frac{d^2\hat{c}}{dM^2} \right) \left( \frac{M}{X} \right)^2 + 2 \left( \frac{dc}{dM} \frac{M}{X^2} - \frac{d\hat{c}}{dM} \frac{M}{X^2} \right). \tag{30}$$

The stochastic terms involved are

$$\begin{aligned} \frac{d^2c}{dM^2} - \frac{d^2\hat{c}}{dM^2} &= \left\{ \varphi(d_1) \frac{d^2d_1}{dM^2} - \varphi(\hat{d}_1) \frac{d^2\hat{d}_1}{dM^2} \right\} - \left\{ \varphi(d_1) d_1 \left( \frac{dd_1}{dM} \right)^2 - \varphi(\hat{d}_1) \hat{d}_1 \left( \frac{d\hat{d}_1}{dM} \right)^2 \right\} \\ &\quad - \left\{ \frac{e^{-r\tau} \varphi(d_2)}{M} \frac{d^2d_2}{dM^2} - \frac{e^{-r\tau} \varphi(\hat{d}_2)}{M} \frac{d^2\hat{d}_2}{dM^2} \right\} + \left\{ \frac{e^{-r\tau} \varphi(d_2)}{M} \frac{2}{M} \frac{dd_2}{dM} - \frac{e^{-r\tau} \varphi(\hat{d}_2)}{M} \frac{2}{M} \frac{d\hat{d}_2}{dM} \right\} \\ &\quad + \left\{ \frac{e^{-r\tau} \varphi(d_2)}{M} d_2 \left( \frac{dd_2}{dM} \right)^2 - \frac{e^{-r\tau} \varphi(\hat{d}_2)}{M} \hat{d}_2 \left( \frac{d\hat{d}_2}{dM} \right)^2 \right\} - \left\{ \frac{2e^{-r\tau} \Phi(d_2)}{M^3} - \frac{2e^{-r\tau} \Phi(\hat{d}_2)}{M^3} \right\} \\ &\stackrel{\text{def}}{=} g_{t1} - g_{t2} - g_{t3} + g_{t4} + g_{t5} - g_{t6} \end{aligned}$$

and

$$\begin{aligned} \frac{dc}{dM} - \frac{d\hat{c}}{dM} &= \left\{ \varphi(d_1) \frac{dd_1}{dM} - \varphi(\hat{d}_1) \frac{d\hat{d}_1}{dM} \right\} - \left\{ e^{-r\tau} \frac{\varphi(d_2)}{M} \frac{dd_2}{dM} - e^{-r\tau} \frac{\varphi(\hat{d}_2)}{M} \frac{d\hat{d}_2}{dM} \right\} \\ &\quad + \left\{ e^{-r\tau} \frac{\Phi(d_2)}{M^2} - e^{-r\tau} \frac{\Phi(\hat{d}_2)}{M^2} \right\} \\ &\stackrel{\text{def}}{=} g'_{t1} - g'_{t2} + g'_{t3}. \end{aligned}$$

Now we analyze each term obtaining,

$$g_{t1} = \frac{d^2 d_1}{dM^2} \left\{ \varphi(d_1) - \varphi(\hat{d}_1) \right\} - \varphi(\hat{d}_1) \left\{ \frac{d^2 \hat{d}_1}{dM^2} - \frac{d^2 d_1}{dM^2} \right\}$$

$$g_{t2} = d_1 \left( \frac{dd_1}{dM} \right)^2 \left\{ \varphi(d_1) - \varphi(\hat{d}_1) \right\} + \left( \frac{dd_1}{dM} \right)^2 \varphi(\hat{d}_1) \left\{ d_1 - \hat{d}_1 \right\}$$

$$+ \hat{d}_1 \varphi(\hat{d}_1) \left\{ \left( \frac{dd_1}{dM} \right)^2 - \left( \frac{d\hat{d}_1}{dM} \right)^2 \right\},$$

and similarly for  $g_{t3}, g_{t4}, g_{t5}, g_{t6}, g'_{t1}, g'_{t2}, g'_{t3}$ . The further analysis of the rate boils down to the analysis of the rates for  $\frac{d\hat{d}_1}{dM} - \frac{dd_1}{dM}, \frac{d^2 d_1}{dM^2} - \frac{d^2 \hat{d}_1}{dM^2}, d_1 - \hat{d}_1, \Phi(d_1) - \Phi(\hat{d}_1)$  and similar terms for  $d_2$ , as  $d_2(M) = (d_1(M) - \sigma(M))\sqrt{\tau}$ .

By the mean value theorem

$$\Phi(d_1) - \Phi(\hat{d}_1) = \varphi(d_0)(d_1 - \hat{d}_1)$$

$$d_1 - \hat{d}_1 = \frac{\{\log(M) + r\tau\}(\hat{V} - V)}{V\hat{V}\sqrt{\tau}} + \sqrt{\tau}(V - \hat{V})/2$$

$$\frac{dd_1}{dM} - \frac{d\hat{d}_1}{dM} = \frac{\hat{V} - V}{MV\hat{V}\sqrt{\tau}} + \left\{ \frac{(\log(M) + r\tau)(V^2 - \hat{V}^2)}{V^2\hat{V}^2\sqrt{\tau}} + \sqrt{\tau}/2 \right\} V'$$

$$+ \left\{ \frac{(\log(M) + r\tau)}{\hat{V}^2\sqrt{\tau}} + \sqrt{\tau}/2 \right\} (\hat{V}' - V'),$$

$$= \mathcal{O}\left( \left\{ \frac{(\log(M) + r\tau)}{\hat{V}^2\sqrt{\tau}} + \sqrt{\tau}/2 \right\} (\hat{V}' - V') \right)$$

$$\frac{d^2 d_1}{dM^2} - \frac{d^2 \hat{d}_1}{dM^2} = - \left\{ \frac{1}{M\sqrt{\tau}} \frac{\hat{V} - V}{V\hat{V}} \frac{1}{M} + \frac{1}{M} \frac{\hat{V} - V}{V\hat{V}\sqrt{\tau}} \frac{V'}{V} + \frac{1}{M\hat{V}\sqrt{\tau}} \frac{V'(\hat{V} - V) + V(V' - \hat{V}')}{V\hat{V}M} \right\}$$

$$+ \left\{ V'' - \hat{V}'' \right\} \left\{ \frac{\sqrt{\tau}}{2} - \frac{\log(M) + r\tau}{V^2\sqrt{\tau}} \right\} + \hat{V}'' \left\{ - \frac{(\log(M) + r\tau)(\hat{V}^2 - V^2)}{V^2\hat{V}^2\sqrt{\tau}} \right\}$$

$$+ 2\hat{V}^3 \{V^2 - \hat{V}^2\} \frac{\log(M) + r\tau}{V^3\hat{V}^3\sqrt{\tau}} + 2\hat{V}^2 \{\hat{V}^3 - V^3\} \frac{\log(M) + r\tau}{V^3\hat{V}^3\sqrt{\tau}}$$

$$- \left\{ V^2(V' - \hat{V}') \frac{1}{MV^2\hat{V}^2\sqrt{\tau}} + \hat{V}'(V^2 - \hat{V}^2) \frac{1}{MV^2\hat{V}^2\sqrt{\tau}} \right\}$$

$$= \mathcal{O}\left( \left\{ V'' - \hat{V}'' \right\} \left\{ \frac{\sqrt{\tau}}{2} - \frac{\log(M) + r\tau}{V^2\sqrt{\tau}} \right\} \right).$$

So the dominant term in the equation (30) is  $S_t e^{r\tau} \left( \frac{d^2 c}{dM^2} - \frac{d^2 \hat{c}}{dM^2} \right) \left( \frac{M}{X} \right)^2$  and this term is dominated by  $g_{t1}$  and  $g_{t3}$ .

$$\begin{aligned} q(x) - \hat{q}(x) &= \mathcal{O} \left( S_t e^{r\tau} \left( \frac{M}{X} \right)^2 \left\{ \varphi(d_1) \frac{d^2 d_1}{dM^2} - \varphi(\hat{d}_1) \frac{d^2 \hat{d}_1}{dM^2} \right\} \right. \\ &\quad \left. - \left( \frac{M}{X} \right)^2 S_t e^{r\tau} \left\{ \frac{e^{-r\tau} \varphi(d_2)}{M} \frac{d^2 d_2}{dM^2} - \frac{e^{-r\tau} \varphi(\hat{d}_2)}{M} \frac{d^2 \hat{d}_2}{dM^2} \right\} \right) \\ &= \mathcal{O} \left( -S_t e^{r\tau} \left( \frac{M}{X} \right)^2 \varphi(\hat{d}_1) \left\{ \frac{d^2 \hat{d}_1}{dM^2} - \frac{d^2 d_1}{dM^2} \right\} \right. \\ &\quad \left. + S_t e^{r\tau} e^{-r\tau} \left( \frac{M}{X} \right)^2 \varphi(\hat{d}_2) / M \left\{ \frac{d^2 \hat{d}_2}{dM^2} - \frac{d^2 d_2}{dM^2} \right\} \right) \\ &= \mathcal{O} \left( -S_t e^{r\tau} \left( \frac{M}{X} \right)^2 \varphi(\hat{d}_1) \left\{ \hat{V}'' - V'' \right\} \left\{ \frac{\sqrt{\tau}}{2} - \frac{\log(M) + r\tau}{V^2 \sqrt{\tau}} \right\} \right. \\ &\quad \left. + S_t \left( \frac{M}{X} \right)^2 \varphi(\hat{d}_2) / M \left\{ \hat{V}'' - V'' \right\} \left\{ \frac{-\sqrt{\tau}}{2} - \frac{\log(M) + r\tau}{V^2 \sqrt{\tau}} \right\} \right). \end{aligned}$$

### 6.4 Proof of Theorem 4

Firstly, the expansion under (A8) in Corollary 2.1 in Claeskens and Van Keilegom (2003) is still valid with remainder term,

$$\mathbf{H}_{n_q} \{ \hat{\mathbf{C}}(x) - \mathbf{C}(x) \} = \mathbf{J}(x)^{-1} \mathbf{H}_{n_q}^{-1} \mathbf{A}_{n_q}(x) + \mathbf{R}_{n_q}(x), \tag{31}$$

where using  $I\{C(x)\}$ ,  $\mathbf{L}$  defined in Section 2,

$$\mathbf{J}(x) \stackrel{\text{def}}{=} f_X(x) I\{C(x)\} \mathbf{L} \tag{32}$$

and

$$\mathbf{R}_{n_q}(x) = -\mathbf{B}_{n_q}^{-1}(x) \mathbf{J}^{-1}(x) \{ \mathbf{J}(x) + \mathbf{B}_{n_q}(x) \} \mathbf{H}_{n_q}^{-1} \mathbf{A}_{n_q}(x) \tag{33}$$

$$+ \{ \mathbf{B}_{n_q}^{-1}(x) \mathbf{J}^{-1}(x) \mathbf{B}_{n_q}(x) - \mathbf{J}^{-1}(x) \} \mathbf{H}_{n_q}^{-1} \mathbf{A}_{n_q}(x) - \mathbf{B}_{n_q}^{-1}(x) \mathbf{D}_{n_q}(x). \tag{34}$$

Define

$$\begin{aligned} \mathbf{D}_{n_q}(x) &\stackrel{\text{def}}{=} \frac{1}{2n_q} \sum_{i=1}^{n_q} K_{h_{n_q}}(X_i - x) \frac{\partial^3}{\partial C^3} \log f\{Y_i; \xi(x, X_i)\} (\hat{C}(x) - C(x)) \\ &\quad \mathbf{X}_i^\top \mathbf{X}_i (\hat{C}(x) - C(x)) \mathbf{H}_{n_q}^{-1} \mathbf{X}_i \\ \mathbf{B}_{n_q}(x) &\stackrel{\text{def}}{=} \frac{1}{n_q} \sum_{i=1}^{n_q} K_{h_{n_q}}(X_i - x) \frac{\partial^2}{\partial C^2} \log f\{Y_i; C(x, X_i)\} \mathbf{H}_{n_q}^{-1} \mathbf{X}_i (\mathbf{H}_{n_q}^{-1} \mathbf{X}_i)^\top. \end{aligned}$$

where  $\xi(x, X_i)$  is in between  $C(x, X_i)$  and  $\hat{C}(x, X_i)$ . Masry (1996) proved that the remainder term is theoretically ignorable uniformly under strong mixing conditions and under (A8) it can be shown that

$$\sup_{x \in E} \mathbf{R}_{n_{qj}}(x) = \mathcal{O}_p((h_{n_q} \log n_q / n_q)^{1/2}). \quad (35)$$

Thus we concentrate on the scaled first-order term

$$S_{nj} \stackrel{\text{def}}{=} (n_q h_{n_q})^{1/2} h_{n_q}^{-j} \{g(x)\}^{-1/2} \mathbf{A}_{n_{qj}}(x)$$

with  $g(x) \stackrel{\text{def}}{=} I(C(x)) f_X(x)$ .

Recall that with a Gaussian likelihood, the components involved in  $\mathbf{J}(x)^{-1} \mathbf{H}_n^{-1} \mathbf{A}_n(x)$  are

$$\begin{aligned} S_{n_{qj}} &= -(n_q)^{-1/2} h_{n_q}^{1/2} \{g(x)\}^{-1/2} \sum_{i=1}^{n_q} K_h(X_i - x) \{(X_i - x)^j / h_{n_q}^j\} \{Y_i - C(x, X_i)\} / \sigma^2(x) \\ &= T_{n_{q1}}(x) + T_{n_{q2}}(x), \quad j = 0, \dots, d, \end{aligned}$$

where

$$\begin{aligned} T_{n_{q1}} &\stackrel{\text{def}}{=} -(n_q)^{-1/2} h_{n_q}^{1/2} g(x)^{-1/2} \sum_{i=1}^{n_q} K_{h_{n_q}}(X_i - x) \{(X_i - x)^j / h_{n_q}^j\} (C(x) - C(x, X_i)) / \sigma^2(x), \\ T_{n_{q2}} &\stackrel{\text{def}}{=} -(n_q)^{-1/2} h_{n_q}^{1/2} g(x)^{-1/2} \sum_{i=1}^{n_q} K_{h_{n_q}}(X_i - x) \{(X_i - x)^j / h_{n_q}^j\} (\sigma(X_i) \varepsilon_i) / \sigma^2(x). \end{aligned}$$

Then we rely on an easy modification of Theorem 2.4 in Liu and Wu (2010), which implies

$$\sup_{x \in E} |T_{n_{q1}}(x)| = \mathcal{O}_p \left( h_{n_q} \sqrt{\log n_q} + \frac{\mathcal{Z}_{n_q}^{1/2} h_{n_q}^{3/2}}{n_q^{1/2}} \right), \quad (36)$$

$$\sup_{x \in E} |T_{n_{q2}}(x) - \frac{h_{n_q}^{1/2}}{(n_q f_X(x))^{1/2}} \sum_{i=1}^{n_q} K_{h_{n_q}}(X_i - x) (X_i - x)^j \varepsilon_i / h_{n_q}^j| = \mathcal{O}_p(h_{n_q} \sqrt{\log n_q}). \quad (37)$$

We need a linear combination of the component  $S_{nqj}$  to analyze  $\hat{q}(x) - q(x)$ , in particular  $\sum_{j=0}^p \mathbf{L}_{3,j+1}^{-1} S_{nqj}$ .

The rest of the proof then follows from the modification of the proof of Proposition 2.1 in Liu and Wu (2010) and is similar to Theorems 4 and 5 in Zhao and Wu (2008).

*Received April 15, 2011; revised December 4, 2013; accepted January 8, 2014.*

## REFERENCES

- Aït-Sahalia, Y., and J. Duarte. 2003. Nonparametric Option Pricing under Shape Restrictions. *Journal of Econometrics* 116: 9–47.
- Aït-Sahalia, Y., and A. W. Lo. 1998. Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices. *The Journal of Finance* 53: 499–547.
- Aït-Sahalia, Y., and A. W. Lo. 2000. Nonparametric Risk Management and Implied Risk Aversion. *Journal of Econometrics* 94: 9–51.
- Bickel, P. J., and M. Rosenblatt. 1973. On Some Global Measures of the Deviations of Density Function Estimates. *The Annals of Statistics* 1: 1071–1095.
- Breeden, D., and R. Litzenberger. 1978. Prices of State-Contingent Claims Implicit in Options Prices. *The Journal of Business* 51: 621–651.
- Brown, D. P., and J. C. Jackwerth. 2004. The Pricing Kernel Puzzle: Reconciling Index Option Data and Economic Theory. *Manuscript*.
- Carroll, R., D. Ruppert, and A. Welsh. 1998. Local Estimating Equations. *Journal of American Statistical Association* 93: 214–227.
- Chabi-Yo, Y., R. Garcia, and E. Renault. 2008. State Dependence can Explain the Risk Aversion Puzzle. *Review of Financial Studies* 21: 973–1011.
- Chetverikov, D. (2012). Testing Regression Monotonicity in Econometric Models. *Technical report*, MIT.
- Christoffersen, P., S. Heston, and K. Jacobs. 2011. A GARCH Option Model with Variance-Dependent Pricing Kernel. *Technical report*, University of Maryland.
- Claeskens, G., and I. Van Keilegom. 2003. Bootstrap Confidence Bands for Regression Curves and their Derivatives. *The Annals of Statistics* 31: 1852–1884.
- Cleveland, W. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 74: 829–836.
- Cochrane, J. H. 2001. *Asset Pricing*. Princeton and Oxford: Princeton University Press.
- Eubank, R. L., and P. L. Speckman. 1993. Confidence Bands in Nonparametric Regression. *Journal of the American Statistical Association* 88: 1287–1301.
- Fan, J. 1992. Design-Adaptive Nonparametric Regression. *Journal of the American Statistical Association* 87: 998–1004.
- Fan, J. 1993. Local Linear Regression Smoothers and their Minimax Efficiencies. *The Annals of Statistics* 21: 196–216.
- Fengler, M. 2005. *Semiparametric Modeling of Implied Volatility*. Berlin: Springer Verlag.

- Ghosal, S., A. Sen, and A. van der Vaart. 2000. Testing Monotonicity of a Regression Function. *The Annals of Statistics* 28: 1054–1082.
- Gilchrist, S., and E. Zakrajšek. 2012. Credit Spreads and Business Cycle Fluctuations. *American Economic Review* 102: 1692–1720.
- Golubev, Y., W. Härdle, and R. Timofeev. 2014. Testing Monotonicity of Pricing Kernels. *Advances in Statistical Analysis* 98: 305–326.
- Grith, M., W. Härdle, and J. Park. 2013. Shape Invariant Modelling Pricing Kernels and Risk Aversion. *Journal of Financial Econometrics* 11: 370–399.
- Hall, P. 1991. *Edgeworth Expansions for Nonparametric Density Estimators, with Applications*, Vol. 22. New York: Academic Press.
- Härdle, W. 1989. Asymptotic Maximal Deviation of  $M$ -smoothers. *Journal of Multivariate Analysis* 29: 163–179.
- Härdle, W., and J. Marron. 1991. Bootstrap Simultaneous Error Bars for Nonparametric Regression. *The Annals of Statistics* 19: 778–796.
- Heaton, J., and D. Lucas. 1992. The Effects of Incomplete Insurance Markets and Trading Costs in a Consumption-based Asset Pricing Model. *Journal of Economic Dynamics and Control* 16: 601–620.
- Jackwerth, J., and M. Rubinstein. 1996. Recovering Probability Distributions from Option Prices. *Journal of Finance* 51: 1611–1631.
- Jackwerth, J. 2000. Recovering Risk Aversion from Option Prices and Realized Returns. *Review of Financial Studies* 13: 433–451.
- Jagannathan, R., and Z. Wang. 1996. The Conditional CAPM and the Cross-Section of Expected Returns. *Journal of Finance* 51: 3–53.
- Jones, M. C., O. Linton, and J. P. Nielson. 1995. A Simple Bias Reduction Method for Density Estimations. *Biometrika* 82: 327–38.
- Kahneman, D., and A. Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* 47: 263–291.
- Lee, S., O. Linton, and Y. Whang. 2009. Testing for Stochastic Monotonicity. *Econometrica* 77: 585–602.
- Lepski, O. V., and V. G. Spokoiny. 1997. Optimal Pointwise Adaptive Methods in Nonparametric Estimation. *The Annals of Statistics* 25: 2512–2546.
- Liero, H. 1982. On the Maximal Deviation of the Kernel Regression Function Estimate. *Mathematische Operationsforschung und Statistik, Serie Statistics* 13: 171–182.
- Liu, W., and W. Wu. 2010. Simultaneous Nonparametric Inference of Time Series. *The Annals of Statistics* 38: 2388–2421.
- Masry, E. 1996. Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* 17: 571–599.
- Renault, E. (1997). “Econometric Models of Option Pricing Errors.” In D. M. Kreps and K. F. Wallis (eds.), *Proceedings the Seventh World Congress of the Econometric Society*, Econometric Society Monographs. London: Cambridge University Press, pp. 223–278.
- Rookley, C. 1997. Fully Exploiting the Information Content of Intra Day Option Quotes: Applications in Option Pricing and Risk Management.

- Rosenberg, J., and R. F. Engle. 2002. Empirical Pricing Kernels. *Journal of Financial Economics* 64: 341–372.
- Rubinstein, M. 1994. Implied Binomial Trees. *Journal of Finance* 49: 771–818.
- Ruppert, D. and M. Wand. 1994. Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics* 22: 1346–1370.
- Vieu, P. 1993. Nonparametric Regression: Optimal Local Bandwidth Choice. *Journal of the Royal Statistical Society, Series B* 53: 453–464.
- Wang, K. 2002. Nonparametric Tests of Conditional Mean-Variance Efficiency of a Benchmark Portfolio. *Journal of Empirical Finance* 9: 133–169.
- Wang, K. 2003. Asset Pricing with Conditioning Information: a New Test. *Journal of Finance* 58: 161–196.
- Xia, Y. 1998. Bias-Corrected Confidence Bands in Nonparametric Regression. *Journal of the Royal Statistical Society: Series B* 4: 797–811.
- Yuan, M. 2009. State Price Density Estimation via Nonparametric Mixtures. *Journal of Applied Statistics* 3: 963–984.
- Zhao, Z., and W. Wu. 2008. Confidence Bands in Nonparametric Time Series Regression. *The Annals of Statistics* 36: 1854–1878.



## Testing monotonicity of pricing kernels

Yuri Golubev · Wolfgang K. Härdle ·  
Roman Timofeev

Received: 6 March 2011 / Accepted: 9 January 2014 / Published online: 12 March 2014  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** The behaviour of market agents has been extensively covered in the literature. Risk averse behaviour, described by Von Neumann and Morgenstern (Theory of games and economic behavior. Princeton University Press, Princeton, 1944) via a concave utility function, is considered to be a cornerstone of classical economics. Agents prefer a fixed profit over an uncertain choice with the same expected value, however, lately there has been a lot of discussion about the empirical evidence of such risk averse behaviour. Some authors have shown that there are regions where market utility functions are locally convex. In this paper we construct a test to verify uncertainty about the concavity of agents' utility function by testing the monotonicity of empirical pricing kernels (EPKs). A monotonically decreasing EPK corresponds to a concave utility function while a not monotonically decreasing EPK means non-averse pattern on one or more intervals of the utility function. We investigate the EPKs for German DAX data for the years 2000, 2002 and 2004 and find evidence of non-concave utility functions: the null hypothesis of a monotonically decreasing pricing kernel is rejected for the data under consideration. The test is based on approximations

---

We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft and the Sonderforschungsbereich 649 "Ökonomisches Risiko". Roman Timofeev's research was supported by Deka Bank scholarship program.

---

Y. Golubev

CMI, Universite de Provence, 39, rue F. Joliot-Curie, 13453 Marseille Cedex 13, France  
e-mail: golubev@cmi.univ-mrs.fr

W. K. Härdle · R. Timofeev (✉)

Center for Applied Statistics and Economics (CASE), Humboldt-Universität zu Berlin,  
Unter den Linden 6, 10099 Berlin, Germany  
e-mail: romant\_2000@mail.ru

W. K. Härdle

e-mail: haerdle@wiwi.hu-berlin.de

of spacings through exponential random variables. In a simulation we investigate its performance and calculate the critical values (surface).

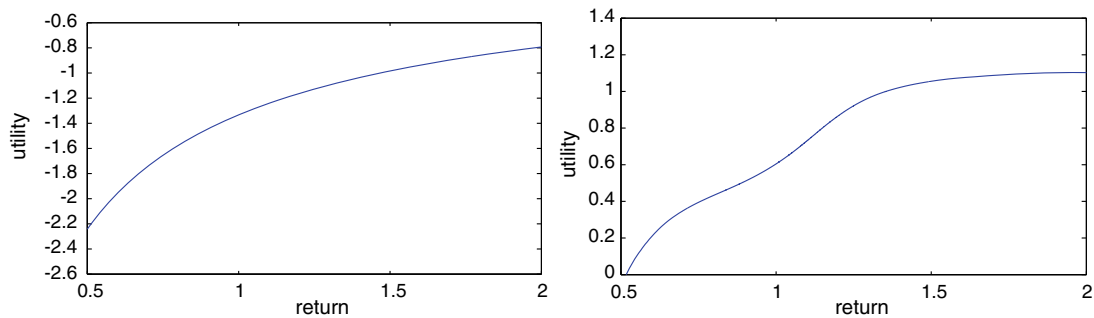
**Keywords** Monotonicity · Pricing kernel · Risk aversion

**JEL Classification** C12 · G12

## 1 Introduction

The behaviour of market agents has always been in focus in economic literature. [Von Neumann and Morgenstern \(1944\)](#) describe risk averse behaviour using concave utility functions. Agents prefer a fixed profit over an uncertain choice with the same expected value, however, lately there has been a lot of discussion about the empirical evidence of such risk averse behaviour. Recent studies by [Jackwerth \(2000\)](#) showed that there is a reference point near the initial wealth where market utility functions are convex. [Rosenberg and Engle \(2002\)](#) also observed a region of negative absolute risk aversion for the pricing kernel constructed using an orthogonal polynomial. A formal test procedure has not been given though. We want to fill this gap by testing the concavity of the utility function and thus checking the monotonicity of the corresponding empirical pricing kernel (EPK). A strictly decreasing EPK corresponds to a concave utility function which is consistent with the classic theory of risk averse behaviour, while rejection of a monotonically decreasing EPK would indicate non-riskaverse pattern of the utility function. By analysing empirical pricing kernels we can also identify on which interval or intervals the monotonicity of the EPK was rejected. Non-monotonicity of the pricing kernel for the S&P 500 was also shown in more recent research by [Constantinides et al. \(2009\)](#), [Bakshi et al. \(2010\)](#) and [Chaudhuri and Schroder \(2010\)](#).

The construction and estimation of empirical pricing kernels has been well described by [Ait-Sahalia and Lo \(2000\)](#). They analyze the concept of economic risk containing investors' preferences and statistical risk which provides information on the dynamics of the data generating process (DGP). Both these risk measures can be identified via distributions (risk neutral ( $Q$ ), physical ( $P$ )). The pricing kernel  $K$  is the Radon Nikodym derivative  $dQ/dP$  of these two measures. Economic risk is well approximated by Arrow-Debreu prices and can be estimated by the risk neutral density  $q$  obtained from the derivative market. By looking at option prices we can find out what stock prices or returns investors expect at time to maturity. Several accurate estimators of  $q$  using, for example, the [Black and Scholes \(1973\)](#) model or nonparametric estimators exist. In this paper the risk neutral density  $q$  is derived from the Heston model. Stochastic volatility models provide better results by fitting the observed volatility smile. Due to the large number of observations in the derivative option market, the risk neutral density  $q$  can be precisely estimated. Statistical risk is related to the properties of the DGP and is given by the pdf  $p$  of future prices conditional on current prices. The main difficulty for the estimation of  $p$  is, of course, that an assumption about the model for the underlying process  $S_t$  has to be made (e.g. geometric Brownian motion under the Black and Scholes model). The density  $p$  can



**Fig. 1** Classical utility function produced from Black Scholes model (*left*) and market utility function estimated from empirical pricing kernel on 06/30/2000 (*right*)

be estimated in several ways, for example, using a nonparametric diffusion model as in [Ait-Sahalia and Lo \(2000\)](#) or a GARCH model as in [Rosenberg and Engle \(2002\)](#). The historical density  $p$  can only be estimated using the past of the time series  $S_t$  and hence is influenced by model specification and data scarcity. The differences in the form of the EPK by various authors might occur due to uncertainty in the estimation of  $p$ . Therefore, we would like to test monotonicity of a pricing kernel constructed as a *ratio of estimated  $q$  and unknown  $p$* .

In [Fig. 1](#) we compare the market utility function obtained from the DAX index in the year 2000 and the utility function derived from the Black and Scholes model. In both cases the risk neutral density  $q$  was obtained via the option market: the state price density that replicates observed option prices is derived to fit the option pricing model (Black and Scholes). This setup provides us with the lognormal density. The historical density  $p$  was assumed to be lognormal for the Black and Scholes model, and nonparametric density estimation over historical time series of the DAX index was used to obtain  $q$  in case of the market utility function. The Black and Scholes model produces an increasing and concave utility function, while the market utility function has a slight bump over the region of zero returns. The aim of this paper is to find out whether observed non-concavity is significant. Obviously, the form of the utility function depends on choice of the DGP for  $S_t$ . As mentioned before, we would like to test monotonicity of the EPK for a general class of DGPs and, therefore, consider  $p$  unknown.

[Ait-Sahalia and Lo \(2000\)](#) in their paper offer another test for risk neutrality and specific preferences. Depending on the form of preferences they define  $H_0$  hypothesis as a relationship between the estimated neutral density  $q$  and the historical density  $p$ . We do not make any assumptions about the form of preferences and also consider the historical density  $p$  unknown. In our test the  $H_0$  hypothesis of a monotonically decreasing EPK is compared to a general class of functions under  $H_1$ . The test is constructed as follows: first the spacing method is used to reduce the problem to an exponential model. On the basis of this model a likelihood ratio test is applied for a fixed interval, then using intersection of tests for different intervals it is expanded to a test independent of intervals. Finally, the test statistics calculated on observed data are compared to simulated critical values, and a final decision about monotonicity is taken.

The paper is organized as follows. In Sect. 2 we introduce important notations and problem setup which is then reduced to an exponential model using the spacing method. In Sect. 3 we formulate the hypotheses, construct a likelihood test for a fixed interval  $[I, J]$  and then expand it to an independent test using the multiple testing technique. We also describe how to simulate critical values using the Monte-Carlo method. Section 4 provides empirical results on DAX data for 2000, 2002 and 2004.

## 2 Conceptual thoughts

### 2.1 Problem setup

Let  $[0, T]$  be the interval of investment in the financial market, where  $t = 0$  denotes the present time and  $t = T \in ]0, \infty[$  the time of maturity. Furthermore, it is assumed that a riskless bond and a risky asset are traded in the financial market as basic underlyings. The price process  $(B_t)_{t \in [0, T]}$  of the riskless bond is defined by

$$\frac{dB_t}{B_t} = r_t dt,$$

via a deterministic Riemannian-integrable interest process  $(r_t)_{t \in [0, T]}$ . The price process  $(S_t)_{t \in [0, T]}$  of the risky asset is assumed to be a nonnegative semimartingale with a constant  $S_0$  and continuously distributed marginals  $S_t$ ,  $t \in [0, T]$ . Furthermore, let us suppose that the financial market is arbitrage free in the sense that there exists at least one equivalent martingale measure. Throughout the paper we assume that the **risk valuation principle** is valid for a nonnegative payoff  $\psi(S_T)$ . That means that there is a Radon-Nikodym density  $\pi$  of a martingale measure such that the price of any  $\psi(S_T)$  is characterized by

$$\mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \pi \right\} = \mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \mathbf{E}_P(\pi | S_T) \right\}.$$

By factorization we may find some Borel-measurable  $K_\pi$  with  $\mathbf{E}(\pi | S_T) = K_\pi$ , so that

$$\mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \pi \right\} = \int_0^\infty e^{-\int_0^T r_t dt} \psi(S_T) K_\pi(x) p_{S_T}(x) dx,$$

where  $p_{S_T}$  denotes the density of the distribution of  $S_T$ . The last formula allows us to call  $K_\pi$  the **pricing kernel** (w.r.t.  $\pi$ ). Here the distribution  $Q_{S_T} \stackrel{\text{def}}{=} \int_{-\infty}^{S_T} K_\pi(z) p_{S_T}(z) dz$ , plays an important role. It is a continuous distribution with pdf  $q_{S_T}$  and is called the **risk neutral distribution** of  $S_T$  (w.r.t.  $\pi$ ). Since

$$\mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(S_T) \pi \right\} = \mathbf{E}_P \left\{ e^{-\int_0^T r_t dt} \psi(x) q_{S_T}(x) dx \right\}$$

holds for any nonnegative payoff  $\psi(S_T)$ , the pricing kernel  $K_\pi = \frac{q_{S_T}}{p_{S_T}}$  a.s. (w.r.t.  $P$ ).

Let us further assume that investors of the financial market are consumers whose consumptions depend on the price  $S_T$  of the stock at maturity only. Within the classical framework, where investors' preferences may be represented by expected utilities, there exists a link between the risk attitude of the investors and the pricing rule in the financial markets. It relies on the assumption of a representative agent whose indirect utility  $U\{\bar{e}(S_T)\}$  which is dependent on the aggregated market endowment  $\bar{e}(S_T)$  has expected utility representation  $U\{\bar{e}(S_T)\} = \mathbb{E}\{u(S_T)\}$  with concave Von Neumann-Morgenstern utility index  $u$ . Under further technical conditions on the investors preferences, see [Härdle et al. \(2012\)](#), [Grith et al. \(2013\)](#), and  $\bar{e}(S_T) = S_T$  there is a positive  $\beta$  such that

$$\frac{du}{dx}|_{x=S_T} = \beta K_\pi(S_T)$$

for almost any realization  $s_T$  of  $S_T$ . For a rigorous derivation we refer to [Karatzas and Shreve \(1998\)](#), sections 4.4 and 4.5. Without loss of generality consider  $q$  and  $K = K_\pi$  on a scale of regular returns  $X = \frac{S_T - S_0}{S_0}$ , where  $S_0$  is the known current price.

The concavity of utility  $U$  can be, therefore, tested by checking monotonicity of  $K$ : a strictly decreasing  $K$  corresponds to a concave utility function, while a non-monotone  $K$  would indicate a non-concave pattern. Our test idea is based on intervals  $[a, b]$ , where  $K$  is not monotonically decreasing.

Denote by  $X_{(1)}, \dots, X_{(n)}$  the order statistics related to a sequence of  $X_1, \dots, X_n$  of returns  $X$  i.e.

$$X_{(1)} \leq X_{(2)}, \dots, \leq X_{(n)}.$$

With these notations we can rephrase the monotonicity testing problem: find (if possible) integers  $I, J$  such that the sequence

$$K_k = K(X_{(k)}) = \frac{q(X_{(k)})}{p(X_{(k)})}, \quad I \leq k \leq J$$

is not monotonically decreasing.

The principal difficulty in this testing procedure is related to the fact that  $p$  is unknown and that violation of monotonicity may occur at different sub-intervals  $[a, b]$ . To solve this challenge we will use three basic ingredients:

1. the spacing method to reduce the stochastics to a simpler exponential model
2. the maximum likelihood test to check monotonicity of  $K_k$  for given  $I$  and  $J$
3. the multiple-testing procedure to find  $I$  and  $J$  on the basis of the data at hand.

## 2.2 The spacing method

Our method is based on Pyke's lemma about the distribution of order statistics, see [Pyke \(1965\)](#). It describes various ways of constructing the spacings, the differences between consecutive observations, in the context of distribution-free tests of fit. The distribution-free assumption is vital for our monotonicity test. Not assuming any form for  $p, q$  makes this test very general and allows to imply strong conclusions on the economic risk of market participants. Pyke's Lemma is based on the following thoughts.

Let  $U_1, \dots, U_n$  be i.i.d random variables with the uniform distribution on  $[0, 1]$  and  $U_0 = 0, U_{n+1} = 1$ . Then the uniform spacings associated with these random variables are defined as

$$S_k = U_{(k)} - U_{(k-1)}, \quad k = 1, \dots, n + 1,$$

where  $U_{(k)}$  are the order statistics  $0 \leq U_{(1)} \leq U_{(2)}, \dots, \leq U_{(n)} \leq 1$ .

The uniform spacings can be represented as exponential random variables proportional to their sum, [Pyke \(1965\)](#).

**Lemma 2.1** *Let  $e_1, \dots, e_{n+1}$  be i.i.d. standard exponentially distributed random variables and  $D = e_1 + e_2 + \dots + e_{n+1}$  be the sum of them. Then the joint distribution of  $\{e_k/D\}_{k=1}^{n+1}$  coincides with the distribution of the set of  $n + 1$  uniform spacings.*

*Using the fact that  $E(e_k) = 1$ , with the law of large numbers for  $D$ , i.e.  $D = n + \mathcal{O}_p(n^{-1/2})$ , we obtain the following approximation:*

$$\begin{aligned} n \{U_{(k)} - U_{(k-1)}\} &= n \cdot S_k \stackrel{\mathcal{L}}{=} n \cdot e_k/D = n \cdot e_k/n + \mathcal{O}_p(n^{-1/2}) \\ &= e_k + \mathcal{O}_p(n^{-1/2}) \approx e_k, k = 1, \dots, n + 1. \end{aligned} \quad (1)$$

We now apply (1), showing the approximation of spacings by a standard exponential random variable, to the problem of the pricing kernel. Let  $X_1, X_2, \dots, X_{n+1}$  be i.i.d. random variables (returns) with a historical density  $p(x)$ ,  $x \in \mathbb{R}^1$  and  $X_{(1)} \leq X_{(2)}, \dots, \leq X_{(n+1)}$  are the corresponding order statistics. By  $P(x)$  we denote the cdf associated with  $p(x)$ . The i.i.d. assumption might be seen as a too strong one, since log returns show volatility clustering effects. These occur though more frequently in highly sampled financial time series. In our case the frequency is low and therefore the identical marginal distribution appears to be justifiable.

The first order Taylor approximation  $P(x)$  at point  $X_{(k)}$  can be calculated using the value of the function at point  $X_{(k-1)}$ ;

$$P(X_{(k)}) \approx P(X_{(k-1)}) + P'(X_{(k-1)})\{X_{(k)} - X_{(k-1)}\}$$

Note that the spacings are of order  $\mathcal{O}_p(n^{-1})$  by Lemma 2.1.

Using the probability integral transformation we see that the random variables  $P(X_i)$  are uniformly distributed over  $(0, 1)$ . Combining first order Taylor approxima-

tion with (1) we obtain

$$e_k \approx n\{U_{(k)} - U_{(k-1)}\} = n\{P(X_{(k)}) - P(X_{(k-1)})\} \approx n \cdot p(X_{(k-1)}) \cdot \{X_{(k)} - X_{(k-1)}\}. \tag{2}$$

Equation (2) is the representation of the spacing of the historical density  $p$  in a form of exponential variables using ordered returns  $X_{(k)}$ . This way we do not make any assumptions about the distribution of  $X$ . Yet, in order to apply Pyke’s Lemma 2.1 the returns are assumed to be i.i.d. implying that in this case we deal with the unconditional density  $p$ . The test of monotonicity of the pricing kernel can now be constructed as the ratio of the risk-neutral density  $q$  and the unconditional historical density  $p$ .

Replacing  $p(x) = K^{-1}(x)q(x)$  in (2) allows to complete the test with respect to the pricing kernel  $K(x)$ :

$$n \cdot \{X_{(k)} - X_{(k-1)}\} \cdot K^{-1}(X_{(k-1)}) \cdot q(X_{(k-1)}) \approx e_{k-1} \quad k = 1, \dots, n + 1 \tag{3}$$

Let us denote for simplicity  $K_{(k-1)} = K(X_{(k-1)})$  and

$$Z_{k-1} = n \{X_{(k)} - X_{(k-1)}\} q(X_{(k-1)}), \quad k = 1, \dots, n + 1. \tag{4}$$

Thus the test problem based on (3) is to check monotonicity of  $K_{k-1}$  using:

$$Z_{k-1} \approx K_{k-1} \cdot e_{k-1}, \quad k = 1, \dots, n + 1. \tag{5}$$

Here again the approximation (5) is of order  $\mathcal{O}_p(n^{-1/2})$ .

### 3 Construction of the test

#### 3.1 Local test

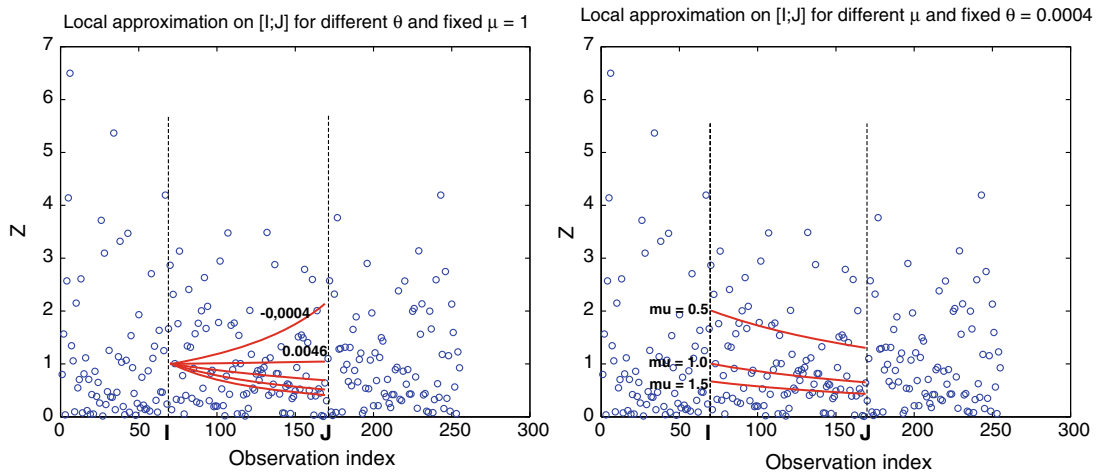
The approximation (2), (5) have been made in order to specify the stochastic fluctuation of  $Z_k$  as being approximately exponential. This will allow us to continue within a quasi likelihood framework.

For simplicity, let us first consider a fixed interval  $[I, J]$  of the sequence

$$Z_s \approx K_s e_s, \quad I \leq s \leq J. \tag{6}$$

where  $I$  and  $J$  are beginning and ending observation indexes of a selected interval. The test alternative on interval  $[I, J]$  implies that if  $K(x)$  is not decreasing, then one can find an index  $s$  that the subsequence  $K_s, I \leq s \leq J$  is increasing.

The local test is based on an inverse linear approximation of  $K(s)$ . The motivation behind this approach is rather simple: in contrast to the standard linear approximation, the inverse linear approximation results in quasi-concavity of the log-likelihood and, thus, permitting to reduce significantly the numerical complexity of the test.



**Fig. 2** Inverse linear approximation for different parameters  $\mu$  and  $\theta$

If  $K_s$ ,  $I \leq s \leq J$  is increasing, then the statistical model with the observations

$$\tilde{Z}_s = \frac{e_s}{\mu\{1 + \theta(s - I)\}}, \quad I \leq s \leq J \tag{7}$$

with parameters  $\mu$  and  $\theta$  and i.i.d. standard exponentially distributed random  $e_s$ , approximates the model (6) better with some negative  $\theta$  than with a positive one. It is important to notice that since  $Z_s$  can have only positive values,  $\theta$  and  $\mu$  are also limited.

Excluding the randomness generated by  $e_s$  by substituting it with  $E(e) = 1$  the approximation (7) takes the form presented in Fig. 2. The plots show different scenarios depending on parameters  $\mu$  and  $\theta$ , where  $\mu$  is responsible for the starting level and  $\theta$  controls the degree of the slope.

Therefore, two composite hypotheses can be formulated. Based on the observed sequence of  $Z_s$  from (6) and approximation (7) we have:

$$H_0 : \theta > 0$$

and  $K_s$ ,  $I \leq s \leq J$  is monotonically decreasing

$$H_1 : \theta \leq 0$$

and  $K_s$ ,  $I \leq s \leq J$  is not-monotonically decreasing.

The test is constructed using the maximum likelihood principle. Let  $P_{\mu,\theta}(\cdot)$  be the joint cdf and  $p_{\mu,\theta}(\cdot)$  be the joint pdf of the observations in (7). Using the fact that the  $e_s$  are i.i.d. standard exponential distributed, the corresponding log-likelihood function takes the form:

$$\log\{p_{\mu,\theta}(\tilde{Z})\} = -\mu \sum_{s=I}^J \tilde{Z}_s \{1 + \theta(s - I)\} + (J - I + 1) \log(\mu) + \sum_{s=I}^J \log\{1 + \theta(s - I)\} \tag{8}$$



Therefore, we can re-formulate the test hypotheses: accept  $H_0$  if

$$\max_{\mu, \theta > 0} \log p_{\mu, \theta}(\tilde{Z}) - \max_{\mu, \theta \leq 0} \log p_{\mu, \theta}(\tilde{Z}) \geq h_\alpha(I, J),$$

otherwise  $H_0$  is rejected.

Here the critical value  $h_\alpha(I, J)$  is computed as a root of equation

$$P \left\{ \max_{\mu, \theta > 0} \log p_{\mu, \theta}(e) - \max_{\mu, \theta \leq 0} \log p_{\mu, \theta}(e) < h_\alpha(I, J) \right\} = \alpha, \tag{9}$$

where  $\alpha$  is the type I error probability.

Now the problem is reduced to calculate the MLE's  $\hat{\mu}$  and  $\hat{\theta}$  for the observed data sequence  $\{Z_s\}$ . Fortunately, the numerical complexity of this test is not very high. First of all, the maximum in  $\mu$  of  $p_{\mu, \theta}(\cdot)$  may be computed very easily. By calculating  $\partial \log p_{\mu, \theta}(Z) / \partial \mu = 0$  we obtain the optimal value of  $\hat{\mu}$

$$\hat{\mu} = \frac{J - I + 1}{\sum_{s=I}^J Z_s \{1 + \theta(s - I)\}}$$

which results in the maximum of the log-likelihood function in  $\mu$

$$\begin{aligned} \max_{\mu} p_{\mu, \theta}(Z) &= \sum_{s=I}^J \log\{1 + \theta(s - I)\} - (J - I + 1) \log \left[ \sum_{s=I}^J Z_s \{1 + \theta(s - I)\} \right] \\ &+ (J - I + 1) \log \frac{J - I + 1}{\exp(1)} \end{aligned} \tag{10}$$

Due to quasi-concavity property, the function  $\max_{\mu} p_{\mu, \theta}$  has a maximum in  $\theta$ . In order to find the optimal value  $\hat{\theta}$  the part which contains  $\theta$  and the rest of the equation should be separated. Denote for brevity

$$L_{\theta}^{I, J}(Z) = \sum_{s=I}^J \log\{1 + \theta(s - I)\} - (J - I + 1) \log\{1 + \theta R^{I, J}(Z_s)\}, \tag{11}$$

where

$$R^{I, J} = R^{I, J}(Z) = \frac{\sum_{s=I}^J Z_s (s - I)}{\sum_{s=I}^J Z_s}. \tag{12}$$

is a random field.

By (10), it is easy to see that

$$\max_{\mu} p_{\mu, \theta}(Z) = L_{\theta}^{I, J}(Z_s) - (J - I + 1) \log \frac{J - I + 1}{\exp(1)} + (J - I + 1) \log \sum_{s=I}^J Z_s. \tag{13}$$

Since only  $L_\theta^{I,J}(Z)$  depends on  $\theta$ , the optimal value can be found as:

$$\hat{\theta} = \arg \max_{\theta} L_\theta^{I,J}(Z_s)$$

The simplest way to find the maximum of the function is to use the Newton-Raphson algorithm:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \frac{dL_\theta^{I,J}(Z_s)/d\hat{\theta}_k}{d^2L_\theta^{I,J}(Z_s)/d\hat{\theta}_k^2} \quad (14)$$

So, the final decision about monotonicity on interval  $[I, J]$  is based on:

$$\begin{aligned} \max_{\theta > 0} L_\theta^{I,J}(Z) - \max_{\theta \leq 0} L_\theta^{I,J}(Z) &= L_{\hat{\theta}}^{I,J}(Z) \mathbf{1}\{\hat{\theta} > 0\} - L_{\hat{\theta}}^{I,J}(Z) \mathbf{1}\{\hat{\theta} \leq 0\} \\ &= L_{\hat{\theta}}^{I,J}(Z) \operatorname{sign}(\hat{\theta}). \end{aligned}$$

With the above argument in mind, we propose the following local test on  $[I, J]$  for checking monotonicity of  $K_s$   $I \leq s \leq J$  in (6):

1. compute

$$\hat{\theta}(Z) = \arg \max_{\theta} L_\theta(Z)$$

with the help of the Newton-Raphson method (14),

2. accept the hypothesis that  $K_s$ ,  $I \leq s \leq J$  is decreasing if

$$L_{\hat{\theta}(Z)}^{I,J}(Z) \operatorname{sign}\{\hat{\theta}(Z)\} - h_\alpha(I, J) \geq 0 \quad (15)$$

otherwise reject the hypothesis.

Notice that the critical value  $h_\alpha(I, J)$  may be computed with the help of the Monte-Carlo method as a root of the equation

$$\mathbb{P}\left[L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\} - h_\alpha(I, J) < 0\right] = \alpha, \quad (16)$$

where  $e = (e_1, \dots, e_{J-I})$  is the sequence of i.i.d. standard exponential random variables.

### 3.2 Global test

The previously described approach considers each interval  $[I, J]$  separately, whereas the decision about monotonicity should be taken for all possible combinations of  $I$  and  $J$ . Therefore, the next step is to join the local tests described above in a global setup. The approach is related to a natural modification of the Bonferroni method which is also used in adaptive estimation in computing nearly optimal penalties for the empirical risk minimization method, see e.g. [Cavalier and Golubev \(2006\)](#). In

order to join the local tests, notice that if the underlying sequence is decreasing, then (15) must hold true for any  $I, J$  or equivalently

$$\min_{I,J} \left[ L_{\hat{\theta}(Z)}^{I,J}(Z) \operatorname{sign}\{\hat{\theta}(Z)\} - t_\alpha(I, J) \right] \geq 0. \tag{17}$$

Therefore we may use this relation as a test prototype. To construct the final test, it remains to redefine the critical values  $t_\alpha(I, J)$ . Obviously, we cannot stick with  $h_\alpha(I, J)$  defined by (16) because it does not control anymore the type I error probability. In fact, the critical values describe a surface  $t_\alpha(I, J)$  that must satisfy the following equation:

$$P\left(\min_{I,J} \left[ L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\} - t_\alpha(I, J) \right] < 0 \right) = \alpha. \tag{18}$$

In contrast to (15), this equation has no unique solution. Intuitively, to maximize the power of the test, i.e. the type II error probability, we should chose  $t_\alpha(I, J)$  as a “maximal” function satisfying (18). Unfortunately, the problem of computation of such a “critical surface” is extremely difficult from theoretical and numerical viewpoints. Therefore, we provide only an approximate solution of this problem. The main step in computing a nearly optimal  $t_\alpha(I, J)$  is to find out the probabilistic structure of  $L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\}$ . Notice that the stochastic part in this field is completely determined by the random field  $R^{I,J}$  given in (12). Therefore, we first focus on probabilistic properties of this field. Using Taylor expansion and the central limit theorem,  $R^{I,J}$  can be approximated as:

$$R^{I,J} \approx \frac{J - I}{2} + \sqrt{\frac{J - I}{12}} \xi \tag{19}$$

where  $\xi \sim N(0, 1)$ . Let us show the approximation for  $L_{\hat{\theta}(e)}^{I,J}(e) \operatorname{sign}\{\hat{\theta}(e)\}$  in more details. Recall the definition of  $R^{I,J}$  in (12).

Using Taylor expansion and the central limit theorem  $R^{I,J}$ :

$$\begin{aligned} R^{I,J} &= \frac{\sum_{s=I}^J e_s(s - I)}{\sum_{s=I}^J e_s} \\ &= \frac{\sum_{s=I}^J \left\{ (e_s - 1)(s - I) + (S - I) \right\}}{\sum_{s=I}^J \left\{ (e_s - 1) + 1 \right\}} \\ &= \left\{ \frac{(J - I)(J - I + 1)}{2} + \sum_{s=I}^J (s - I)(e_s - 1) \right\} \left\{ (J - I + 1) + \sum_{s=I}^J (e_s - 1) \right\}^{-1} \\ &= \left\{ \frac{(J - I)}{2} + \frac{1}{J - I + 1} \sum_{s=I}^J (s - I)(e_s - 1) \right\} \left\{ 1 + \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\}^{-1} \end{aligned}$$

Assuming that  $J - I$  is sufficiently large:

$$\begin{aligned} & \left\{ 1 + \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\}^{-1} \\ &= 1 - \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) / \left[ 1 - \left\{ \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\}^2 \right] \\ &\approx 1 - \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \end{aligned}$$

which then results in:

$$\begin{aligned} R^{I,J} &\approx \left\{ \frac{(J - I)}{2} + \frac{1}{J - I + 1} \sum_{s=I}^J (s - I)(e_s - 1) \right\} \left\{ 1 - \frac{1}{J - I + 1} \sum_{s=I}^J (e_s - 1) \right\} \\ &= \frac{(J - I)}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{J - I}{2} \right) (e_s - 1) \\ &\quad - \frac{1}{(J - I + 1)^2} \sum_{s=I}^J (e_s - 1) \sum_{s=I}^J (s - I)(e_s - 1) \\ &\approx \frac{J - I}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right) (e_s - 1). \end{aligned}$$

Using the CLT  $R^{I,J}$  is approximated:

$$R^{I,J} = \mu^{I,J} + \sigma^{I,J} \xi,$$

where  $\mu^{I,J}$  and  $\sigma^{I,J}$  are the mean and variance of  $R^{I,J}$  and  $\xi \sim N(0, 1)$ .

Note that:

$$\begin{aligned} \mu^{I,J} &= \mathbb{E} \left\{ \frac{J - I}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right) (e_s - 1) \right\} = \frac{J - I}{2} \\ \sigma^{2 I,J} &= \text{Var} \left\{ \frac{J - I}{2} + \frac{1}{(J - I + 1)} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right) (e_s - 1) \right\} \\ &= \frac{\text{Var}(e_s - 1)}{(J - I + 1)^2} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right)^2 \\ &= \frac{1}{(J - I + 1)^2} \sum_{s=I}^J \left( s - \frac{I + J}{2} \right)^2 \end{aligned}$$

Using the fact that  $\sum_{i=k}^n i^2 = \sum_{i=1}^{n-k+1} (i+k-1)^2$ , we can derive:

$$\begin{aligned} \sum_{s=I}^J \left( s - \frac{J-I}{2} \right)^2 &= \sum_{s=1}^{J-I+1} \left( s - \frac{I+J}{2} + I - 1 \right)^2 \\ &= \sum_{s=1}^{J-I+1} \left( s - \frac{I+J}{2} + I - 1 \right)^2 \\ &= \sum_{s=1}^{J-I+1} \left( s - \frac{J-I}{2} - 1 \right)^2 \end{aligned}$$

Furthermore, as  $\sum_{i=1}^n i^2 = n(n+1)(2n+1)/6$  the variance  $\sigma^{2 I, J}$  converges as follows:

$$\begin{aligned} \sigma^{2 I, J} &= \frac{1}{(J-I+1)^2} \sum_{s=1}^{J-I+1} \left( s - \frac{J-I}{2} - 1 \right)^2 \\ &= \frac{1}{(J-I+1)^2} \left\{ \sum_{s=1}^{J-I+1} s^2 - 2 \sum_{s=1}^{J-I+1} s \left( \frac{J-I-2}{2} \right) + \sum_{s=1}^{J-I+1} \left( \frac{J-I-2}{2} \right)^2 \right\} \\ &= \frac{(J-I+1)(J-I+2)(2(J-I)+3)}{6(J-I+1)^2} - \frac{(J-I+2)(J-I+1)(J-I-2)}{2(J-I+1)^2} + \\ &\quad + \frac{(J-I+1)(J-I-2)^2}{4(J-I+1)^2} \\ &\approx \frac{J-I}{3} - \frac{J-I}{2} + \frac{J-I}{4} = \frac{J-I}{12} \end{aligned}$$

Therefore,  $R^{I, J}$  can be approximated as:

$$R^{I, J}(e_s) \approx \frac{J-I}{2} + \sqrt{\frac{J-I}{12}} \xi$$

Next, combining (20) with the Taylor expansion for  $L_{\theta}^{I, J}(e)$ , we obtain

$$L(e) \approx -\theta \sqrt{\frac{(J-I)^3}{12}} \xi - \theta^2 \frac{(J-I)^3}{24}$$

Again all these approximations are of order  $\mathcal{O}_p(n^{-1/2})$ .

Thus, with simple algebra we arrive at the limit distribution of the test statistics

$$L_{\hat{\theta}}(e) \text{ sign}\{\hat{\theta}(e)\} \approx -\frac{1}{2} \xi^2 \text{ sign}(\xi).$$

The equation for the critical surface (18) therefore takes the following form

$$\mathbb{P}\left[\max_{I,J}\left\{\frac{1}{2}\xi^2 \operatorname{sign}(\xi) + t_\alpha(I, J)\right\} > 0\right] = \alpha. \quad (20)$$

In order to find a solution, we assume for a moment that the maximum in the above display is computed over couples  $I_k, J_k, k = 1, \dots, (n-1)/d$ , where  $I_k = 1 + d(k-1)$ ,  $J_k = I_k + d$  and  $d$  is a given integer. Thus, we are looking for a minimal  $t_\alpha(I_k, J_k)$  satisfying

$$\mathbb{P}\left[\max_{k \leq n/d}\left\{\frac{1}{2}(\xi^{I_k, J_k})^2 \operatorname{sign}(\xi^{I_k, J_k}) + t_\alpha(I_k, J_k)\right\} > 0\right] = \alpha.$$

Since the random variables  $(\xi^{I_k, J_k})^2 \operatorname{sign}(\xi^{I_k, J_k}), k = 1, \dots, n/d$  are i.i.d., it is clear that  $t_\alpha(I_k, J_k)$  is a constant depending only on  $\alpha, n$ , and  $d$ . Finally notice that

$$\max_{k \leq n/d} (\xi^{I_k, J_k})^2 \operatorname{sign}(\xi^{I_k, J_k}) \approx 2 \log \frac{n}{d}$$

because  $\xi^{I_k, J_k}$  are i.i.d. and nearly Gaussian  $N(0, 1)$ . Therefore it is clear that

$$t_\alpha(I_k, J_k) = -\tilde{t}_\alpha \log \frac{n}{d},$$

where  $\tilde{t}_\alpha$  is a constant close to 1. This argument prompts the following form of the critical surface (18):

$$t_\alpha(I, J) = -\tilde{t}_\alpha \log \frac{n}{J-I}. \quad (21)$$

The exact constant  $\tilde{t}_\alpha$  is finally computed with the help of the Monte-Carlo as a root of the equation:

$$\mathbb{P}\left(\min_{|I-J| \geq M} \left[ L_{\hat{\theta}(e)}^{I, J}(e) \operatorname{sign}\{\hat{\theta}(e)\} + \tilde{t}_\alpha \log \frac{n}{J-I} \right] < 0\right) = \alpha. \quad (22)$$

Hence the critical surface  $t_\alpha(I, J)$  in (21) is approximated as a function of a scalar critical value  $\tilde{t}_\alpha$ , sample size  $n$  and significance level  $\alpha$ , which definitely reduces the complexity of the computation.

Here  $M > 2$  is an integer which is needed to guarantee that the asymptotic approximation (22) holds true. Typically,  $M \approx 10$ . The inaccuracies due to small  $M$  and other approximations applied to derive the final results are compensated by  $\tilde{t}_\alpha$  critical value.

More precisely the calculation of the critical value  $\tilde{t}_\alpha$  is done in the following steps:

1. Generation of  $Z_{\text{gen}}$  as  $\exp(1)$  for a given sample size  $n$ .
2. Calculation of optimal parameters  $\hat{\theta}(I, J)$  and resulting  $L_{\hat{\theta}}(I, J)$  over generated sequences  $Z_{\text{gen}}$  for all possible intervals  $[I, J], 1 \leq I < J \leq n$

**Table 1** Simulated critical values for different sample sizes and  $\tilde{t}_\alpha$ 

$\alpha$ (%)	$n = 50$	$n = 100$	$n = 255$
20	2.5010	1.8003	1.2934
10	2.5789	1.8257	1.3065
5	2.6163	1.8358	1.3087
4	2.6229	1.8381	1.3093
3	2.6363	1.8414	1.3102
2	2.6425	1.8437	1.3111
1	2.6530	1.8453	1.3117

### 3. Calculation of the corresponding $\tilde{t}_\alpha$ as a root of equation

$$P\left(\min_{|I-J|\geq M} \left[ L_{\hat{\theta}(Z_{\text{gen}})}^{I,J}(Z_{\text{gen}}) \text{sign}\{\hat{\theta}(Z_{\text{gen}})\} + \tilde{t}_\alpha \log \frac{n}{J-I} \right] < 0\right) = \alpha \quad (23)$$

by repeating steps 1 and 2 using simulated data.

For reader's convenience Table 1 provides the critical values  $\tilde{t}_\alpha$  for  $\alpha = 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2$  and sample sizes  $n = 50, 100, 255$ . As can be seen for smaller size  $n$  the critical values  $\tilde{t}_\alpha$  are larger to counterbalance the inaccuracies in the estimation of  $L_{\hat{\theta}(Z_{\text{gen}})}$ .

With the given  $\tilde{t}_\alpha$  the monotonicity test on the observed data  $Z$  takes the following form: *we accept the hypothesis that  $K_s$  is a decreasing sequence if*

$$\min_{|I-J|\geq M} \left[ L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I} \right] > 0. \quad (24)$$

## 4 Empirical results

### 4.1 Data and estimation of risk neutral density

For the analysis we take the data used in [Detlefsen et al. \(2007\)](#) where the pricing kernels and the risk aversion are analysed in years 2000, 2002 and 2004 in order to consider different market regimes (30th of June, 28th of June and 25th of June correspondingly). These dates were selected in such a way that the DAX index was rising, remained stable and was falling during one year period prior to these dates. According to our test design the decision about monotonicity is made on the basis of (4):  $Z_k = n \cdot (X_{(k+1)} - X_{(k)}) \cdot q(X_{(k)})$  where  $X_{(k)}$  are the order statistics of DAX returns and  $q$  is an estimate of the risk neutral density.

The DAX returns  $\frac{S_t - S_{t-126}}{S_{t-126}}$  are calculated on half a year basis, where  $S_t$  are daily index observations. They are ordered into  $X_{(k)}$ . We started 1.5 year back from the dates mentioned above which resulted in exactly  $n = 255$  observations. The corresponding

ordered returns differences  $X_{(k+1)} - X_{(k)}$  for 2000, 2002 and 2004 are displayed in Fig. 3.

The risk neutral density  $q$  aggregates economic information about the prices by replicating observed option prices. An estimate of  $q$  can be found as the second derivative of the call price with respect to the strike. The estimation of  $q$  is then reduced to the problem of a proper option-pricing formula. Under the hypothesis of [Black and Scholes \(1973\)](#) we obtain a log-normal density  $q$ . A closed form solution can be also obtained under more general class of models. Here we use the [Heston \(1993\)](#) model calibrated to fit the observed smile in implied volatility surfaces (IVS) using the absolute error between observed and modeled quantities:

$$\text{ASE}_t = \sqrt{\sum_{i=1}^n n^{-1} \{\text{IV}_i^{\text{mod}}(t) - \text{IV}_i^{\text{mar}}(t)\}^2}$$

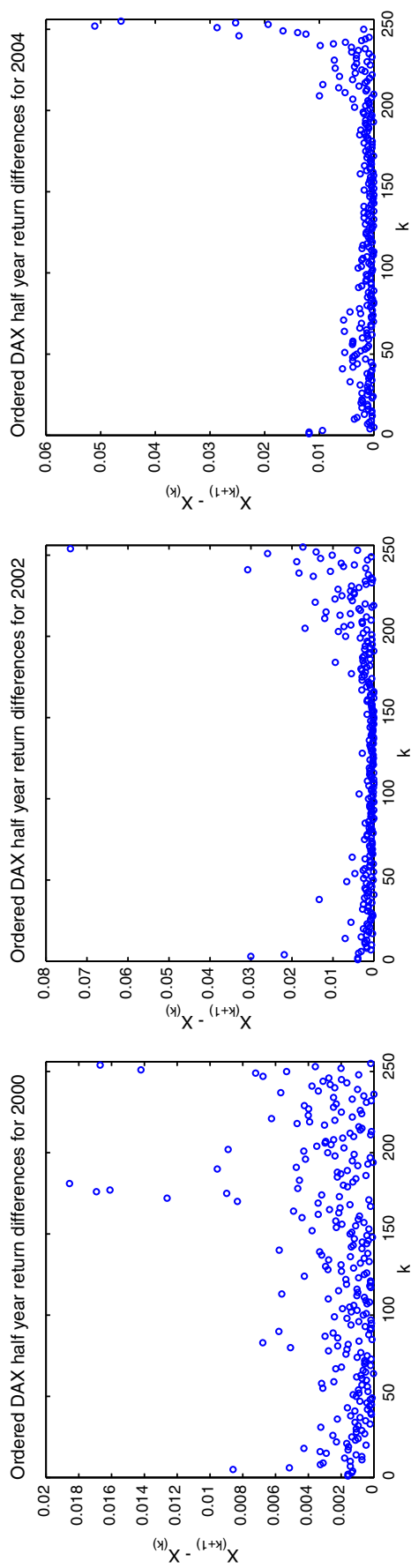
where mod refers to a model quantity, mar to a quantity observed on the market and  $\text{IV}(t)$  to an implied volatility on day  $t$ . The index  $i$  runs over all  $n$  observations of the surface on day  $t$ . Daily EUREX-settlement prices of European options on the DAX index are used to obtain observed option prices and corresponding implied volatilities. The model parameters are calibrated for each of three dates using the whole surface of implied volatilities, but we exclude observations that are deep out of the money because of illiquidity of these products. More precisely, we consider for the calibration only options with more than 1 month time to maturity and restrict ourselves to strikes 50% above or below the spot in the moneyness direction. For each trading day there are about 250 points in the volatility surface available for the calibration. Having obtained the model parameters we can estimate the risk neutral density for any time to maturity  $\tau$ . In this paper we analyse semiannual returns, therefore, we obtain the density  $q$  by fixing  $\tau = 0.5$  years. The corresponding densities for 2000, 2002 and 2004 can be seen in Fig. 4. The risk free interest rates are approximated by the EURIBOR. On each trading day we use the yields corresponding to the maturities of the implied volatility surface. As the DAX is a performance index it was adjusted to dividend payments. Thus, we do not have to consider dividend payments explicitly. For more details on the estimation of the risk neutral density refer to [Detlefsen and Härdle \(2007\)](#). Similar density  $q$  can be obtained using the minimization procedure mentioned in [Jackwerth \(2000\)](#). Alternatively, the density  $q(x)$  can be estimated semiparametrically or even nonparametrically, see [Ait-Sahalia and Lo \(2000\)](#).

#### 4.2 Monotonicity of DAX EPKs

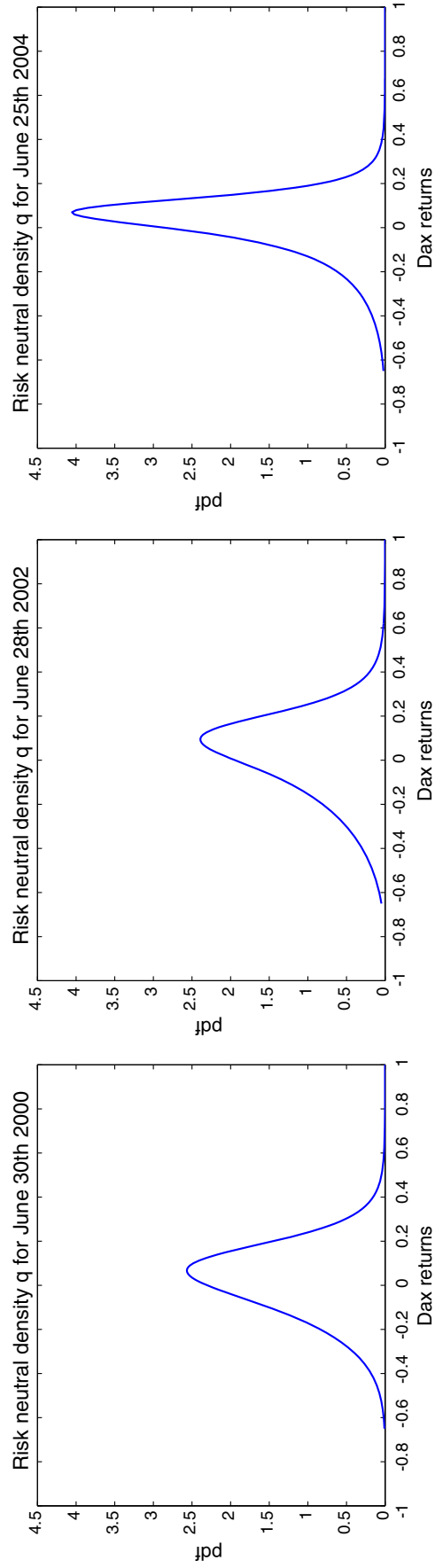
The final goal is to test an empirical pricing kernel obtained from observed data. Having obtained  $q$  and  $X_{(k)}$ ,  $Z_k$  can be calculated and the monotonicity testing becomes a technical exercise. Resulting values of  $Z_k$  are displayed in Fig. 5.

The calculated  $Z_k$  correspond to one year risk neutral density  $q$  and can be tested with the corresponding critical values for  $n = 255$  from Table 1. Similarly to the graphs showing the test ideas a minimum distance of 10 observations between  $I$  and  $J$  was set.





**Fig. 3** Half-year ordered returns differences  $X_{(k+1)} - X_{(k)}$  for years 2000, 2002 and 2004



**Fig. 4** Risk neutral densities  $q$  estimated on 30/06/2000, 28/06/2002 and 25/06/2004

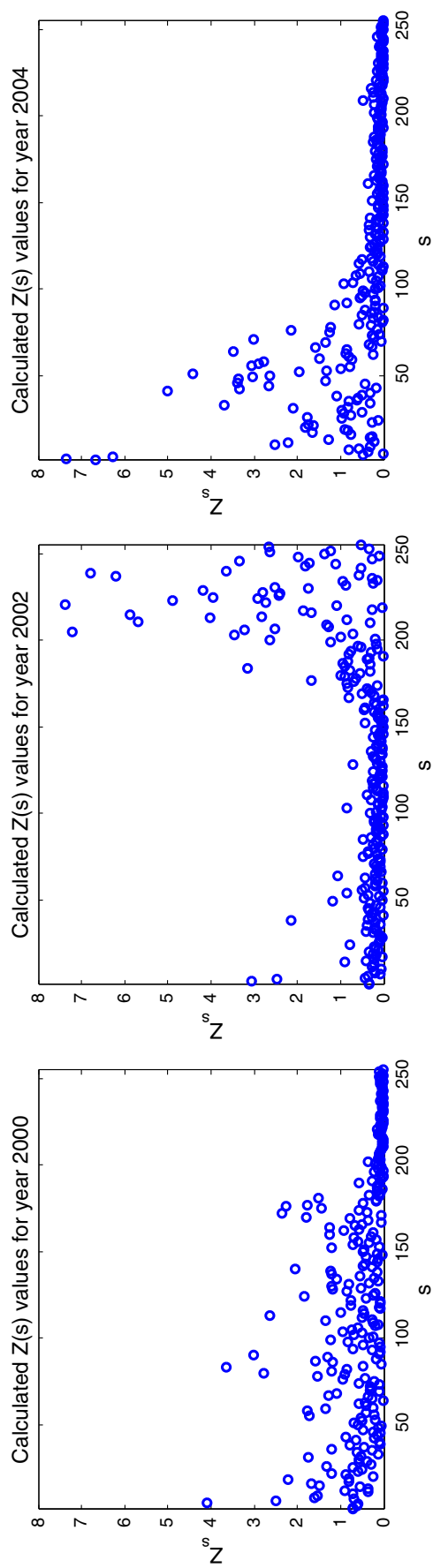


Fig. 5 Calculated  $Z_k$  for years 2000, 2002 and 2004

**Table 2**  $\min_{|I-J| \geq M} [L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I}]$  for  $\alpha = 10, 5$  and  $1\%$ 

$\alpha$ (%)	$\min[L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I}]$		
	2000	2002	2004
10	0.5038	-0.005	0.2114
5	0.5046	0.0021	0.2017
1	0.5058	0.0118	0.1946

The results are summarized in Table 2, the surfaces  $L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}\{\hat{\theta}(Z)\} + \tilde{t}_\alpha \log \frac{n}{J-I}$  are given in Fig. 6.

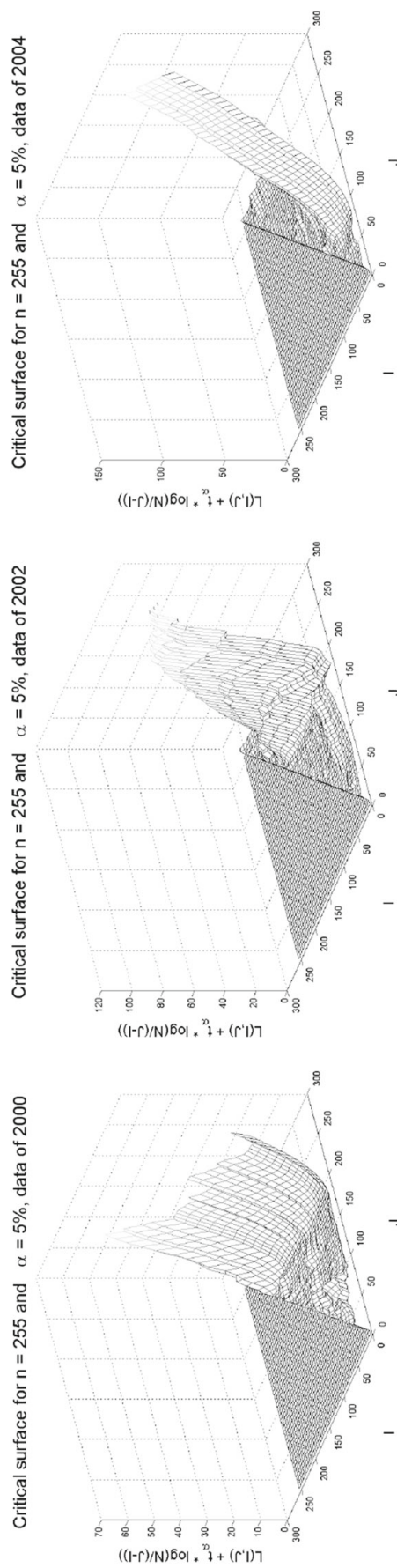
The analysis of the DAX data in years 2000, 2002 and 2004 showed that  $H_0$  (monotonic pattern of the pricing kernel) is rejected under 90% significance level in year 2002.

## 5 Conclusion

We describe a test that checks monotonicity of pricing kernels. By testing monotonicity of a pricing kernel we can determine whether the corresponding utility function is concave or not. A strictly decreasing pricing kernel corresponds to a concave utility function, while a non-decreasing EPK means that the utility function contains non-concave regions.

Pricing kernels are constructed as a ratio of the risk neutral density  $q$  and the historical density  $p$ . Investors' assessment of the future distribution of asset prices under risk neutral measure (density  $q$ ) can be estimated via the derivative market. By looking at option prices we can find out what stock prices or returns investors expect at the time of maturity. Due to the large number of observations  $q$  can be precisely estimated. The actual movement of  $S_t$  is described by the historical density  $p$  which is estimated using the time series of  $S_t$ . The main difficulty for the estimation of  $p$  is, of course, that an assumption about the model for the underlying process  $S_t$  has to be made. Due to scarcity of data and specification difficulties  $p$  is considered to be unknown. We, therefore, test the monotonicity via the ratio  $q/p$  of two densities, where  $q$  is given and  $p$  is unknown.

The test is constructed as follows: first the spacing method is used to reduce the problem to an exponential model. Using Pyke's lemma of order statistics, a pricing kernel  $K$  is represented as a sequence of observed values  $Z_k$  and standard exponential variables  $e_k$ . Based on this simple exponential model we construct the likelihood ratio test for a fixed interval  $[I, J]$ . A global test is built by the simultaneous testing on all possible intervals  $[I, J]$ , where the main difficulty is to calculate the corresponding critical surfaces for given  $I, J$ , sample size  $n$  and confidence level  $\alpha$ . The critical surfaces can be nearly approximated with a scalar critical value  $\tilde{t}_\alpha$  dependent only on sample size  $n$  and significance level  $\alpha$ , which significantly reduces the complexity of the test. The problem is then reduced to the simulation of the critical value  $\tilde{t}_\alpha$  for  $n$  and  $\alpha$  using the Monte-Carlo technique.



**Fig. 6** Resulting test statistic surfaces of  $L_{\hat{\theta}(Z)}^{I,J}(Z) \text{sign}(\hat{\theta}(Z)) + \tilde{t}_{\alpha} \log \frac{n}{J-I}$  for years 2000, 2002 and 2004 with  $\alpha = 5\%$

We investigated the EPKs for German DAX data for the years 2000, 2002 and 2004 and found evidence of non-concave utility behaviour for the data under consideration.

## References

- Ait-Sahalia, Y., Lo, A.: Nonparametric risk management and implied risk aversion. *J. Econom.* **94**(12), 9–51 (2000)
- Bakshi, G., Madan, D., Panayotov, G.: Returns of claims on the upside and the viability of U-shaped pricing kernels. *J. Financ. Econ.* **97**, 130–154 (2010)
- Black, F., Scholes, M.: The pricing of options and corporate liabilities. *J. Polit. Econ.* **102**(3), 637–659 (1973)
- Cavalier, L., Golubev, Y.: Monotonicity of the stochastic discount factor and expected option returns. *Ann. Stat.* **34**(4), 1653–1677 (2006)
- Chaudhuri, R., Schroder, M.: Monotonicity of the stochastic discount factor and expected option returns. Working paper, School of Business Administration, Oakland University (2010)
- Constantinides, G., Jackwerth, J., Perrakis, S.: Mispricing of S&P 500 index options. *Rev. Financ. Stud.* **22**, 1247–1277 (2009)
- Detlefsen, K., Härdle, W.: Calibration risk for exotic options. *J. Deriv.* **14**(4), 47–63 (2007)
- Detlefsen, K., Härdle, W., Moro, R.: Empirical pricing kernels and investor preferences. *Math. Methods Econ. Financ.* **3**(1), 19–48 (2007)
- Grith, M., Härdle, W.K., Krätschmer, V.: An axiomatic and data driven view on the EPK paradox. *Rev. Financ.*, revise and resubmit (2013)
- Härdle, W., Okhrin, Y., Wang, W.: Uniform confidence bands for empirical pricing kernel. *J. Financ. Econom.*, revise and resubmit (2012)
- Heston, S.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev. Financ. Stud.* **6**(2), 327–343 (1993)
- Jackwerth, J.: Recovering risk aversion from option prices and realized returns. *Rev. Financ. Stud.* **13**(2), 433–451 (2000)
- Karatzas, I., Shreve, S.: *Methods of Mathematical Finance*. Springer, NY (1998)
- Pyke, R.: Spacings. *J. R. Stat. Soc. B* **27**, 395–436 (1965)
- Rosenberg, J., Engle, R.: Empirical pricing kernels. *J. Financ. Econ.* **64**(3), 341–372 (2002)
- Von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press, Princeton (1944)