IRTG 1792 Discussion Paper 2020-021

Improved Estimation of Dynamic Models of Conditional Means and Variances

Weining Wang [*]
Jeffrey M. Wooldridge [*2]
Mengshan Xu [*3]

[*] University of York, UK
[*2] Michigan State University, USA
[*3] London School of Economics, UK

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

Improved Estimation of Dynamic Models of Conditional Means and Variances

Weining Wang*, Jeffrey M. Wooldridge† and Mengshan Xu‡

September 11, 2020

Abstract

Modelling dynamic conditional heteroscedasticity is the daily routine in time series econometrics. We propose a weighted conditional moment estimation to potentially improve the efficiency of the QMLE (quasi maximum likelihood estimation). The weights of conditional moments are selected based on the analytical form of optimal instruments, and we nominally decide the optimal instrument based on the third and fourth moments of the underlying error term. This approach is motivated by the idea of general estimation equations (GEE). We also provide an analysis of the efficiency of QMLE for the location and variance parameters. Simulations and applications are conducted to show the better performance of our estimators.

1 Introduction

Nonlinear, dynamic models of means, variances, and covariances are routinely estimated in financial economics, macroeconomics, and other disciplines. A leading example is the set of models coming from the GARCH (generalized autoregressive conditional heteroskedasticity) family, see Bollerslev [1986]. Bollerslev and Wooldridge [1992] show that the Gaussian (normal) quasi maximum likelihood estimator (QMLE) has a critical robustness property in the general multivariate case: provided that the first two conditional moments are correctly specified, the Gaussian QMLE consistently estimates the parameters indexing the means, variances, and covariances under weak regularity conditions. Moreover, inferences that are robust to arbitrary departures from normality (subject to enough finite moments) are readily available.

*Department of Economics and Related Studies, University of York, YO10 5DD, United Kingdom. email: weining.wang@york.ac.uk

†Department of Economics, Michigan State University, East Lansing, MI 48824 USA; email: wooldri1@msu.edu

‡Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK. Email: m.xu8@lse.ac.uk.

Newey and Steigerwald [1997] suggest that for dynamic models with conditional means and variance components to be estimated, identification will be related to the symmetry of the density, excluding some special cases. Of course, if the normality assumption fails, then it is possible to obtain more efficient estimators. One possibility is to model the deviations from normality: assume that the new distribution is correctly specified and then use MLE to estimate the mean and variance parameters along with any new parameters. A leading example is Bollerslev [1987], who proposed replacing the normal distribution with a t-distribution with unknown degrees of freedom (df). The degrees-of-freedom parameter is estimated along with the mean and variance parameters. For specific parametric models such as GARCH, Fan et al. [2014] consider Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. Hafner and Rombouts [2007] looks at a nonparametric estimation for innovation distributions of the multivariate Garch model to improve the efficiency.

It is difficult to extend to multivariate settings with a general likelihood function tailored to heavy tail distributions. People can use copulas, see, for example, Chen and Fan [2006], who consider a copulae model for the temporal dependency. But estimation and theory might be involved. Semiparametric efficient properties for dynamic models have been studied in the literature, see, for example, Drost et al. [1997] with assumptions on the independent and identically distributed (iid) innovations. To obtain the most efficient estimator, which can be easily computed under very general model conditions, might be too demanding for practitioners.

In general, a QMLE framework is to construct an efficient estimator, and an efficiency lower bound is available under certain regularity conditions. However, it might be too much to ask for hitting the efficiency lower bound in practice. Instead, we can focus on a class of simple and useful estimators obtained from solving a set of general estimation equations (GEE). The GEE is set up in a way that the estimators would be efficient at the elements of certain "ideal submodel", and we would in turn have more robustness and better finite sample properties. For estimation in dynamic models, a few conditional moment equations are convenient to obtain, and optimal instruments are used to achieve estimation efficiency in a general GMM framework. Optimal weights are theoretically feasible but difficult to be estimated as it implicitly depends on the data generating process, in particular on moments of the underlying data generating processes.

To mimic the GEE idea in our moment estimation and optimal instrument framework, we can restrict the data generating process to a class of submodels under which the optimal instrument matrix is easy to be estimated. In particular, connected to the GEE approach as in Liang and Zeger [1986], we specify "a working" optimal instrument matrix. This involves a 'working" conditional variance-covariance matrix for the residual function that defines the first two conditional moments. Namely, we consider an estimator that, like the QMLE, requires only the first two moments to be correctly specified for consistency. Within this

class, we would like to find the optimal IV estimator under weaker assumptions than are commonly imposed. We then show that the estimator is likely to be more efficient than the QMLE quite generally when only the first two moments are correctly specified. The simulation results in many univariate and multivariate models strongly support this point: If the first two moment conditions are correctly specified, for example, the underline distribution of innovation term follows the standard normal, QMLE and our methods have similar performances. If we change the underline data generating process to a skewed normal distribution, our method outperforms QMLE.

We contribute to the literature in three aspects. First, we propose an easy to use estimator for dynamic models with dynamic moments. Second, we illustrate its efficient properties for many commonly used dynamic models. Third, we show its theoretical properties. The paper is organized as follows. In Section 2, we show the basic univariate framework and our proposed estimator. In Section 3, we extend the estimator to multivariate models. We study in Section 4 some model scenarios under which QMLE is indeed efficient. Simulations and applications are set in Section 5 and Section 6. We show the theoretical properties of our estimator in Section 7.

2 Univariate models

We start with the univariate case, where y_t is a scalar response. Let $\mathbf{x}_t$ be a vector of conditioning variables, which would generally include lagged values of y_t. It can also include contemporaneous values of some other series, say $\mathbf{z}_t$, as well as lags of $\mathbf{z}_t$. Let $\mathbf{w}_t = (y_t, \mathbf{x}_t)$. We assume models of the conditional mean and variance:

$$m_t(\mathbf{x}_t, \theta),\ v_t(\mathbf{x}_t, \theta),\ \theta \in \mathbf{\Theta}.$$

where $v_t(\mathbf{x}_t, \theta) > 0$ for all $\mathbf{x}_t$ and $\theta \in \mathbf{\Theta}$. The assumption that these models are correctly specified is that for some $\theta_o \in \mathbf{\Theta}$,

$$E(y_t|\mathbf{x}_t) = m_t(\mathbf{x}_t, \theta_o) \tag{1}$$

$$Var(y_t|\mathbf{x}_t) = v_t(\mathbf{x}_t, \theta_o),\ t = 1, 2, ... \tag{2}$$

We will assume that we have T observations from the stochastic process, and, as an indexing convention, we also assume that $t = 1$ is the first period we observe all elements of $(y_t, \mathbf{x}_t)$.

The setup is general enough to allow the variance parameters and mean parameters to overlap, or to be completely separate. This allows traditional models where the variance is modeled separately from the

variance, as well as ARCH-in-mean type models. See, for example, Engle et al. [1987].

Assumptions (1) and (2) are generally sufficient for consistency and $\sqrt{T}$-asymptotic normality under suitable regularity conditions and weak dependence requirements. However, it is traditional in the settings of interest to assume that the models are dynamically complete in mean and variance:

$$E(y_t|\mathbf{x}_t, I_{t-1}) = E(y_t|\mathbf{x}_t) \quad (3)$$

$$Var(y_t|\mathbf{x}_t, I_{t-1}) = Var(y_t|\mathbf{x}_t), \quad (4)$$

where

$$I_{t-1} = (y_{t-1}, \mathbf{x}_{t-1}, y_{t-2}, \mathbf{x}_{t-2}, ..., y_1, \mathbf{x}_1)$$

is the information observed through time $t-1$. Dynamic completeness in the first two moments is always essentially assumed in ARCH and GARCH models, and so are their numerous variations and extensions. A more realistic setting would allow misspecification of all kinds. However, this is rarely done in practice. We follow convention and assume the first two conditional moments are correctly specified and dynamically complete. This is the framework assumed in Bollerslev and Wooldridge [1992].

It is also important that we do not try to exploit all of the moment conditions implied by correct dynamics beyond the lags of variables included in $\mathbf{x}_t$. In other words, the optimal instrumental variables we derive depend only on $\mathbf{x}_t$ and not further lags. We take this approach for a few reasons. First, we want our estimator to be a direct extension of the Gaussian QMLE – henceforth, QMLE for brevity – which can be viewed as a particular IV estimator whose instruments depend only on $\mathbf{x}_t$. Second, the efficiency gains in going beyond functions of $\mathbf{x}_t$ are unlikely to be impressive, given that we are assuming the dynamics in the first two moments are completely captured by $\mathbf{x}_t$. Third, it would not be clear how to add additional moment conditions to ensure nontrivial efficiency gains. These points will become clear as we derive the proposed estimator. Along the way, we will derive the conditions under which the QMLE is the asymptotically efficient estimator based only on the first two moments and dynamic completeness. Henceforth, the dynamic completeness assumption is taken as given.

To see how to potentially improve over the QMLE, define the error term as well as the standardized error as

$$u_t = y_t - m_t(\mathbf{x}_t, \theta_o)$$

$$e_t = \frac{u_t}{\sqrt{v_t(\mathbf{x}_t, \theta_o)}}.$$

By construction,

$$\begin{aligned} E\left(u_t|\mathbf{x}_t\right) &= 0,\, Var\left(u_t|\mathbf{x}_t\right) = v_t(\mathbf{x}_t,\theta_o) \\ E\left(e_t|\mathbf{x}_t\right) &= 0,\, Var\left(e_t|\mathbf{x}_t\right) = 1, \end{aligned}$$

and these conditional moments continue to hold conditional on $(\mathbf{x}_t, I_{t-1})$. It is important to observe that e_t is not guaranteed to be even independent of $\mathbf{x}_t$, let alone $(\mathbf{x}_t, I_{t-1})$. Treatments such as Newey and Steigerwald [1997] make the assumption

$$e_t \text{ is independent of } \left(\mathbf{x}_t, I_{t-1}\right),\, t = 1, 2, ... \tag{5}$$

assume e_t is independent of $(\mathbf{x}_t, I_{t-1})$, an extra assumption that is not implied by the specification of the first two moments. As discussed in Bollerslev and Wooldridge [1992], the correct specification of the first two conditional moments implies that the score of the quasi-log likelihood is a vector martingale difference sequence (MDS). Along with weak dependence requirements, the MDS result ensures that the QMLE is $\sqrt{T}$-asymptotically normal. Assumption (5) can be used to simplify the verification of regularity conditions, but it has no substantive effect on the asymptotic properties of the QMLE. See also Wooldridge [1994] for a more general discussion.

To obtain a simple estimator that is asymptotically more efficient than the QMLE, we *nominally* assume

$$E\left(e_t^3|\mathbf{x}_t\right) = \kappa_3^o \tag{6}$$

$$E\left(e_t^4|\mathbf{x}_t\right) = \kappa_4^o, \tag{7}$$

which is implied by the independence assumption (5). Written in terms of the errors u_t,

$$\begin{aligned} E\left(u_t^3|\mathbf{x}_t\right) &= \kappa_3^o \left[v_t(\mathbf{x}_t,\theta_o)\right]^{3/2}, \\ E\left(u_t^4|\mathbf{x}_t\right) &= \kappa_4^o \left[v_t(\mathbf{x}_t,\theta_o)\right]^2. \end{aligned}$$

Under the assumption of normality, we have $\kappa_3^o = 0$ and $\kappa_4^o = 3$. Bollerslev and Wooldridge [1992] show that neither of these restrictions is necessary for the consistency of the QMLE. In fact, neither is the assumption that these $E\left(e_t^3|\mathbf{x}_t\right)$ and $E\left(e_t^4|\mathbf{x}_t\right)$ are constant. For the estimators here, we use these assumptions to derive optimal instruments for estimating θ_o, but (6) and (7) are not required for the consistency of our estimator. Later, we will need to consistently estimate κ_3^o and κ_4^o, but this is easily done given an initial preliminary

estimator of θ_o, which would typically be the QMLE. In deriving the asymptotic properties, we will only assume that the estimators converge to some constant without invoking (6) or (7).

Our estimator is motivated by finding the optimal instruments under the assumption that the model as dynamically complete in the first two moments and the auxiliary assumptions (6) and (7). For each t, define the 2×1 residual function ($\mathbf{w}_t$ not defined here)

$$\mathbf{r}_t(\mathbf{w}_t, \theta) = \begin{pmatrix} y_t - m_t(\mathbf{x}_t, \theta) \\ [y_t - m_t(\mathbf{x}_t, \theta)]^2 - v_t(\mathbf{x}_t, \theta) \end{pmatrix}.$$

If the model is dynamically complete then,

$$E\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right] = E\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t, I_{t-1}\right] = \mathbf{0}.$$

As discussed in Wooldridge [1994], the optimal instrumental variables based on these moment conditions depend on

$$\mathrm{E}\left[\nabla_\theta \mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right],$$
$$\mathbf{Var}\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right].$$

The first matrix is easily obtained under correct specification of the first two moments as the $2 \times P$ matrix

$$\mathbf{R}_t(\mathbf{x}_t, \theta_o) \equiv E\left[\nabla_\theta \mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right] = -\begin{pmatrix} \nabla_\theta m_t(\mathbf{x}_t, \theta_o) \\ \nabla_\theta v_t(\mathbf{x}_t, \theta_o) \end{pmatrix}.$$

In general, $\mathbf{Var}\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right]$ can be any positive semi-definite matrix function of $\mathbf{x}_t$, making it difficult to implement an always efficient IV estimator. Our key innovation is to impose a "working" version of $Var\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right]$, where we borrow the term "working" from the generalized estimating equations (GEE) literature (for example, Zeger and Liang [1986]). In particular, if we impose (6) and (7), then

$$\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o) \equiv Var\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right] = \begin{pmatrix} v_t(\mathbf{x}_t, \theta_o) & \kappa_3^o \left[v_t(\mathbf{x}_t, \theta_o)\right]^{3/2} \\ \kappa_3^o \left[v_t(\mathbf{x}_t, \theta_o)\right]^{3/2} & \left(\kappa_4^o - 1\right)\left[v_t(\mathbf{x}_t, \theta_o)\right]^2 \end{pmatrix}.$$

Rather than being unrestricted, $\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)$ has a relatively simple form and depends on only two additional parameters. Under normality and other symmetric distributions, this structure holds with $\kappa_3^o = 0$. Specific distributions also imply a value for κ_4^o, or in some cases – such as the skewed normal distribution or t distribution – treat it as a parameter to be estimated using an MLE approach. Here, we use this structure

to obtain an estimator more efficient than the QMLE, if (6) and (7) hold. As in the GEE literature, we expect efficiency will carry over even if we drop (6) and (7).

As discussed in Wooldridge [1994], the optimal instruments, obtained only from the moments conditional on $\mathbf{x}_t$, are

$$[\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)]^{-1} \mathbf{R}_t(\mathbf{x}_t, \theta_o).$$

In order to implement the optimal IV estimator, we need consistent estimators of θ_o and κ_o. For θ_o, the obvious choice is the QMLE, say $\check{\theta}$. For κ_3^o, a natural method of moments estimator is to obtain the standardized residuals,

$$\check{e}_t = \frac{\left[y_t - m_t(\mathbf{x}_t, \check{\theta})\right]}{\sqrt{v_t(\mathbf{x}_t, \check{\theta})}},$$

and then

$$\check{\kappa}_3 = T^{-1} \sum_{t=1}^{T} \check{e}_t^3 = T^{-1} \sum_{t=1}^{T} \left[\frac{\check{u}_t}{\sqrt{v_t(\mathbf{x}_t, \check{\theta})}} \right]^3.$$

Next, define $\eta_4^o = \kappa_4^o - 1$, so that

$$\eta_4^o = E\left[\left(e_t^2 - 1\right)^2\right].$$

Therefore, a method of moments estimator that ensures nonnegativity is

$$\check{\eta}_4 = T^{-1} \sum_{t=1}^{T} \left(\check{e}_t^2 - 1\right)^2 = T^{-1} \sum_{t=1}^{T} \left[\frac{\check{u}_t^2}{v_t(\check{\theta})} - 1 \right]^2,$$

and an unbiased version can be switched with the denominator T to be $T - 1$.

Given the preliminary estimators, we now obtain what would be estimates of the optimal instruments under correct model specification, including complete dynamics, (6) and (7):

$$\check{\mathbf{Z}}_t = \check{\mathbf{D}}_t^{-1} \check{\mathbf{R}}_t. \tag{8}$$

These can be used to obtain an optimal IV estimator. Namely, $\hat{\theta}$ solves

$$\sum_{t=1}^{T} \check{\mathbf{Z}}_t' \mathbf{r}_t(\hat{\theta}) = \mathbf{0}, \tag{9}$$

which is a set of P nonlinear equations in the P unknowns $\hat{\theta}$.

It is true that, under correct specification (dynamic completeness is not needed for this), the preliminary estimation of θ_o and κ_o does not affect the asymptotic distribution of $\sqrt{T}\left(\hat{\theta}-\theta_o\right)$, see conditions in Theorem 3. In fact, in the formal statement of our result, we will drop (6) and (7), and simply assume that $\check{\kappa}_3$ converges in probability to some constant, say κ_3^*, and $\check{\eta}_4$ converges to a positive constant η_4^*. The estimator will not be the optimal IV estimator without (6) and (7), but it could still be asymptotically more efficient.

To avoid solving the nonlinear first-order condition, we can consider a one-step estimation. The idea of a one-step estimator is to consider a linear approximation to the moment condition in (9). Provided that we have a well-chosen initial estimator $\check{\theta}$, for example, a consistent QMLE. Then we can improve upon the estimator $\check{\theta}$ using our one-step procedure. The one-step estimation can be defined as:

$$\overline{\theta}=\check{\theta}-\{\sum_{t=1}^{T}\check{\mathbf{Z}}_t'\nabla\mathbf{r}_t(\check{\theta})\}^{-1}\sum_{t=1}^{T}\check{\mathbf{Z}}_t'\mathbf{r}_t(\check{\theta}). \tag{10}$$

Suppose that $\check{\theta}$ is a consistent and asymptotic normal estimator. $\hat{\theta}$ is an "ideal" consistent estimator for θ^0. Then we can prove that $\tilde{\theta}$ is asymptotic equivalent to the estimator $\hat{\theta}$.

3 The multivariate case

We now formulate our estimation in the multivariate case. We can similarly look at the vector $d\times 1$ vector $\mathbf{y}_t$, and the conditional mean specification $\mathrm{E}(\mathbf{y}_t|\mathbf{x}_t)=\mathbf{m}(\mathbf{x}_t,\theta_o)$, $\mathbf{Var}(\mathbf{y}_t|\mathbf{x}_t)=\Sigma_t(\mathbf{x}_t,\theta_o)$, and $\mathbf{u}_t(\theta)=\mathbf{y}_t-\mathbf{m}(\mathbf{x}_t,\theta_o)$. We suppress the dependence of the function $\Sigma_t(\mathbf{x}_t,\theta_o)$ on $\mathbf{x}_t$. We can define the standardized vector as $\mathbf{e}_t=\Sigma_t(\theta_o)^{-1/2}(\mathbf{y}_t-\mathbf{m}(\mathbf{x}_t,\theta_o))=\Sigma_t(\theta_o)^{-1/2}\mathbf{u}_t(\theta_o)$. Note that $\mathbf{e}_t$ is a $d\times 1$ dimensional random vector, and θ is a P-dimensional parameter. As a generalization in the multivariate case, we look at the moment vector $\underset{(d+d(d+1)/2)\times 1}{\mathbf{r}_t(\theta)}=(\mathbf{u}_t(\theta)',\mathrm{Vech}(\mathbf{u}_t(\theta)\mathbf{u}_t'(\theta)-\Sigma_t(\theta))')'$. It should be noted that Vec is the vectorization of a matrix, and Vech is denoted as a half vectorization. The switching between Vec and Vech is via a duplication and elimination matrix, i.e., $D_n\mathrm{Vech}A=\mathrm{Vec}(A)$, and $L_n\mathrm{Vec}A=\mathrm{Vech}A$ for an $n\times n$ symmetric matrix A.

$$\begin{aligned}\mathrm{E}\left(\mathbf{u}_t(\theta_o)|\mathbf{x}_t\right) &= \mathbf{0},\ \mathbf{Var}\left(\mathbf{u}_t(\theta_o)|\mathbf{x}_t\right)=\Sigma_t(\mathbf{x}_t,\theta_o)\\ \mathrm{E}\left(\mathbf{e}_t|\mathbf{x}_t\right) &= \mathbf{0},\ \mathrm{Vec}\{\mathbf{Var}\left(\mathbf{e}_t|\mathbf{x}_t\right)\}=\mathrm{E}(\mathbf{e}_t\otimes\mathbf{e}_t|\mathbf{x}_t)=\mathrm{Vec}(I_d)\end{aligned}$$

We define the third and fourth condition moments of $\mathbf{e}_t$ as $\mathrm{E}(\mathbf{e}_t\otimes\mathbf{e}_t\mathbf{e}_t'|\mathbf{x}_t)=\mathbf{K}_3$, and $\mathrm{E}(\mathbf{e}_t\otimes\mathbf{e}_t\mathbf{e}_t'\otimes\mathbf{e}_t'|\mathbf{x}_t)=\mathbf{K}_4$, which in the case of multivariate Gaussian has constant matrix forms. In particular for independent

standard normal random vector, $\mathbf{K}_4 = I_{d^2} + \tilde{\mathbf{K}}_d + \text{Vec}(I_d)\text{Vec}(I_d)'$, where $\tilde{\mathbf{K}}_d$ is a commutation matrix defined for example in Magnus and Neudecker [1979], and $\mathbf{K}_3 = \mathbf{0}$.

We can see that $\mathbf{K}_3$ and $\mathbf{K}_4$ can be similarly estimated from the sample. In particular, let j_1, j_2, j_3, i_1, and i_2 be integers taking values from 1 to d, and we have

$$\begin{aligned} \check{\mathbf{K}}_3 \text{ is } d^2 \times d \qquad & \text{with } [\check{\mathbf{K}}_3]_{(j_1-1)d+j_2,j_3} = \check{e}_{tj_1}\check{e}_{tj_2}\check{e}_{tj_3} \\ \check{\mathbf{K}}_4 \text{ is } d^2 \times d^2 \qquad & \text{with } [\check{\mathbf{K}}_4]_{(j_1-1)d+j_2,(j_3-1)d+j_4} = \check{e}_{tj_1}\check{e}_{tj_2}\check{e}_{tj_3}\check{e}_{tj_4} \end{aligned}$$

, where

$$\check{\mathbf{e}}_t = \Sigma_t(\check{\theta})^{-1/2}(\mathbf{y}_t - \mathbf{m}(\mathbf{x}_t, \check{\theta}))$$

The estimate of $((j_1-1)d+j_2) \times j_3$ element of $\check{\mathbf{K}}_3$ is $T^{-1}\sum_{t=1}^{T}\check{e}_{tj}^3$ for $j_1 = j_2 = j_3 = j$, and otherwise 0, due to the fact that $\text{E}(e_{j_1}e_{j_2}e_{j_3}) = 0$, when two of the elements in j_1, j_2, j_3 are unequal. Moreover, the nonzero elements $(j_1-1)d+j_2, (j_3-1)d+j_4$ in $\check{K}_4$ is when $j_1 = j_2 = i_1, j_3 = j_4 = i_2$, $j_1 = j_4 = i_1, j_2 = j_3 = i_2$, $j_1 = j_3 = i_1, j_2 = j_4 = i_2$ and $j_1 = j_2 = j_3 = j_4 = j$. We can then estimate the nonzero element by $(T^{-1}\sum_{t=1}^{T}\check{e}_{ti_1}^2)(T^{-1}\sum_{t=1}^{T}\check{e}_{ti_2}^2)$ for the first three cases, and $T^{-1}\sum_{t=1}^{T}\check{e}_{tj}^4$ for $j_1 = j_2 = j_3 = j_4 = j$.
The matrix $\mathbf{D}$ in Section 2 becomes

$$\mathbf{D}(\mathbf{x}_t, \theta_o) = \begin{pmatrix} \Sigma_t & \Sigma_{12,t} \\ \Sigma_{12,t}' & \Sigma_{22,t} \end{pmatrix},$$

where

$$\begin{aligned} \Sigma_{12t}' &= L_d\Sigma_t(\theta_o)^{1/2} \otimes \Sigma_t(\theta_o)^{1/2}\mathbf{K}_3\Sigma_t(\theta_o)^{1/2}, \text{ and} \\ \Sigma_{22t} &= L_d(\Sigma_t(\theta_o)^{1/2} \otimes \Sigma_t(\theta_o)^{1/2}\mathbf{K}_4\Sigma_t(\theta_o)^{1/2} \otimes \Sigma_t(\theta_o)^{1/2} - \text{Vec}\Sigma_t(\theta_o)\text{Vec}\Sigma_t(\theta_o)')L_d'. \end{aligned}$$

The partial derivative of the moment functions $(d + d(d+1)/2) \times P$ is

$$\mathbf{R}_t = \nabla\mathbf{r}_t(\mathbf{x}_t, \theta_o) = (\nabla\mathbf{m}_t(\mathbf{x}_t, \theta_o)', (L_d(\mathbf{u}_t \otimes I + I \otimes \mathbf{u}_t)\nabla\mathbf{m}_t(\mathbf{x}_t, \theta_o))')'.$$

The optimal instrument matrix is $\mathbf{D}(\mathbf{x}_t, \theta_o)^{-1}\mathbf{R}_t(\mathbf{x}_t, \theta_o)$, and define $\check{\mathbf{Z}}_t = \mathbf{D}(\mathbf{x}_t, \check{\theta})^{-1}\mathbf{R}_t(\mathbf{x}_t, \check{\theta})$ we solve the estimation as in equation (9).

4 When is the Gaussian QMLE asymptotically efficient?

To understand better that the circumstances under which we are going to improve upon QMLE, we study in this section the efficiency of QMLE for some specific dynamic models. In particular, we study the effect of third and fourth moments on the asymptotic variance-covariance matrix. It is not difficult to obtain a sufficient condition for the Gaussian QMLE to be the asymptotically efficient estimator. We follow Bollerslev and Wooldridge [1992] to introduce the setup in the Gaussian QMLE briefly. The Gaussian QMLE is defined to be

$$\breve{\theta} \stackrel{\text{def}}{=} \text{argmax}_{\theta\in\Theta} T^{-1}\sum_{t=1}^{T}\ell_t(\theta;\mathbf{y}_t,\mathbf{x}_t). \tag{11}$$

The log likelihood function is denoted as

$$\ell_t(\theta;\mathbf{y}_t,\mathbf{x}_t) = -1/2\log|\Sigma_t(\mathbf{x}_t,\theta)| - 1/2(\mathbf{u}_t(\mathbf{w}_t,\theta)')\Sigma_t^{-1}(\mathbf{x}_t,\theta)(\mathbf{u}_t(\mathbf{w}_t,\theta)). \tag{12}$$

We breviate $\mathbf{u}_t(\mathbf{w}_t,\theta)$ to $\mathbf{u}_t(\theta)$. The score function is

$$\mathbf{s}_t(\theta) = \nabla_\theta\mathbf{m}_t(\theta)'\Sigma_t(\theta)^{-1}\mathbf{u}_t(\theta) + 1/2\nabla_\theta\Sigma_t(\theta)'(\Sigma_t^{-1}(\theta)\otimes\Sigma_t^{-1}(\theta))\text{Vec}(\mathbf{u}_t(\theta)\mathbf{u}_t(\theta)-\Sigma_t(\theta)). \tag{13}$$

where $\nabla_\theta\Sigma_t(\theta) = \nabla_{\theta'}\text{Vec}(\Sigma_t(\theta))$ is $d^2\times P$ matrix, and $\nabla_\theta\mathbf{m}_t(\theta)' = -\nabla_\theta\mathbf{u}_t(\theta)'$ is a $P\times d$ dimension matrix.

The negative Hessian evaluated by expectation at the θ_o, i.e., $\text{E}(\nabla_\theta(s_t(\theta_o)'))$ is

$$I_t(\theta_o) = \nabla_\theta\mathbf{m}_t(\theta_o)'\Sigma_t(\theta_o)^{-1}\nabla_\theta\mathbf{u}_t(\theta_o) + 1/2\nabla_\theta\Sigma_t(\theta_o)'(\Sigma_t^{-1}(\theta_o)\otimes\Sigma_t^{-1}(\theta_o))\nabla_\theta\Sigma_t(\theta_o). \tag{14}$$

It is not hard to see that

$$\begin{aligned}
J_t(\theta_o) \stackrel{\text{def}}{=}\ & \text{E}(\mathbf{s}_t(\theta_o)\mathbf{s}_t'(\theta_o)) = \nabla_\theta\mathbf{u}_t(\theta_o)'\Sigma_t^{-1}(\theta_o)\nabla_\theta\mathbf{u}_t(\theta_o) \\
& +1/2\nabla_\theta\Sigma(\theta_o)'\Sigma_t(\theta_o)^{-1/2}\otimes\Sigma_t(\theta_o)^{-1/2}\mathbf{K}_3\Sigma_t(\theta_o)^{-1/2}\nabla\mathbf{u}_t(\theta) \\
& +1/2\{\nabla_\theta\Sigma(\theta_o)'\Sigma_t(\theta_o)^{-1/2}\otimes\Sigma_t(\theta_o)^{-1/2}\mathbf{K}_3\Sigma_t(\theta_o)^{-1/2}\nabla\mathbf{u}_t(\theta)\}' \\
& +1/4\nabla_\theta\Sigma_t(\theta_o)'\Sigma_t(\theta_o)^{-1/2}\otimes\Sigma_t(\theta_o)^{-1/2}\mathbf{K}_4\Sigma_t(\theta_o)^{-1/2}\otimes\Sigma_t(\theta_o)^{-1/2}\nabla_\theta\Sigma_t(\theta_o)) \\
& -1/4\nabla_\theta\Sigma_t(\theta_o)'\text{Vec}(\Sigma_t(\theta_o)^{-1})\text{Vec}'(\Sigma_t(\theta_o)^{-1})\nabla_\theta\Sigma_t(\theta_o) \\
\stackrel{\text{def}}{=}\ & J_{t1}+J_{t2}+J_{t2}'+J_{t3}+J_{t4}
\end{aligned}$$

Under regularity conditions, the asymptotic variance of the QMLE is $(I_o^{-1}J_oI_o^{-1})$, see, for example, Theorem 7, where $J_o = \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}J_t(\theta_o)$ corresponds to the information matrix, and $I_o = \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}I_t(\theta_o)$.

If the conditional distribution of e_t is indeed a standardized multivariate Gaussian, we have $I_o = J_o$, and the asymptotic variance would hit the lower bound J_o^{-1}.

The efficiency of a Gaussian QMLE would depend on the analytical form of $J_t(\theta_o)$ and $I_t(\theta_o)$. In the one dimensional case, for a GARCH(p,q) model, $\mathbf{m}_t(\theta)' = \mathbf{0}$, and thus $J_{t1} = 0$ and $J_{t2} = 0$, we can see that $J_t(\theta_o) = (\kappa_4^o - 1)/2I_t(\theta_o)$. If $\kappa_4^o = 3$ corresponding to the fourth moment of a Gaussian random variable $J_t(\theta_o) = I_t(\theta_o)$, the asymptotic variance would hit the lower bound. This corresponds to the finding of Francq and Zakoïan [2004].

Now we let

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} (J_{t2} + J_{t2}' + J_{t3} + J_{t4}) := J_{2o} + J_{2o}' + J_{3o} + J_{4o},$$

and

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} 1/2 \nabla_\theta \Sigma_t(\theta_o)' (\Sigma_t^{-1}(\theta_o) \otimes \Sigma_t^{-1}(\theta_o)) \nabla_\theta \Sigma_t(\theta_o) := I_o.$$

Proposition 1. *If (i) the model is dynamic complete, (ii) satisfying 1, 2, 3, 4, (iii) has consistent QMLE, (iv) and the innovations* $\mathbf{e}_t$ *have a finite fourth-order moment, then the efficiency loss of the Gaussian QMLE by misspecifying the moments are* $V_o \overset{\text{def}}{=} I_o^{-1}(J_{2o} + J_{2o}' + J_{3o} + J_{4o} - I_{2o})I_o^{-1}$*, with* $V_{o1} = I_o^{-1}(J_{2o} + J_{2o}')I_o^{-1}$ *coming from skewness part, and* $V_{o2} = I_o^{-1}(J_{3o} + J_{4o} - I_{2o})I_o^{-1}$ *coming from the fourth moment part. In particular, when the mean parameter and the variance parameter overlap, the* V_{o1}*,* V_{o2} *would be of specific structure so that the skewness matrix would affect the variance of the mean and the variance of the QMLE, and the fourth-moment matrix would only affect the variance of the QMLE.*

To investigate the role of third and fourth moments of the underlying distribution of $\mathbf{e}_t$ on the efficiency of estimation, we look at J_{t2} concerning the third moment matrix $\mathbf{K}_3$ and J_{t3} concerning the fourth moment $\mathbf{K}_4$. Ideally, for innovation distribution that is symmetric $\mathbf{K}_3 = \mathbf{0}$, which is true for a Gaussian vector, any deviation from $\mathbf{K}_3 = \mathbf{0}$ would contribute to the deviation of J_t to I_t.

If we were to look at the case when the mean parameter and the variance parameter do not overlap, which is to say $\theta = (\beta', \gamma')'$, where β' is $1 \times P_1$, and γ' is $1 \times P_2$, then $\nabla_\theta \mathbf{u}_t(\theta_o)$ consists of $\underset{d\times P_1}{\nabla_\beta \mathbf{u}_t(\theta_o)}$ and $\underset{d\times P_2}{\nabla_\gamma \mathbf{u}_t(\theta_o)} = \mathbf{0}$, and $\nabla_\theta \Sigma_t(\theta_o)$ contains $\nabla_\beta \Sigma_t(\theta_o) = \mathbf{0}$ and $\nabla_\gamma \Sigma_t(\theta_o)$.

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} I_t$$
$$= \begin{pmatrix} \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \nabla_\beta \mathbf{u}_t(\theta_o)' \Sigma_t(\theta_o)^{-1} \nabla_\beta \mathbf{u}_t(\theta_o) & 0 \\ 0 & \lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \nabla_\gamma \Sigma_t(\theta_o)' O_{2t} \nabla_\gamma \Sigma_t(\theta_o) \end{pmatrix},$$

with $O_{2t} \stackrel{\text{def}}{=} \Sigma_t(\theta_o)'(\Sigma_t^{-1}(\theta_o) \otimes \Sigma_t^{-1}(\theta_o))$, which is defined to be

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} I_t \stackrel{\text{def}}{=} \begin{pmatrix} I_1 & 0 \\ 0 & I_2 \end{pmatrix},$$

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} (J_{t2} + J_{t2}') = \begin{pmatrix} 0 & D_{\kappa_3^o} \\ D_{\kappa_3^o}' & 0 \end{pmatrix},$$

where $D_{\kappa_3^o}$ is a $P_1 \times P_2$ matrix corresponding to the formulae of J_t, and $\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} J_{t3} = \begin{pmatrix} 0 & 0 \\ 0 & E_{\kappa_4^o} \end{pmatrix}$ where $E_{\kappa_4^o}$ is a $P_2 \times P_2$ matrix corresponding to the formulae of J_t.

We can then find that V_o would be consistent of the following part $(I_o^{-1}(J_{2o} + J_{2o}' + J_{3o})I_o^{-1})$, which would take the following form,

$$\begin{pmatrix} 0 & 0 \\ 0 & I_2^{-1} E_{\kappa_4^o} I_2^{-1} \end{pmatrix} + \begin{pmatrix} 0 & I_1^{-1} D_{\kappa_3^o} I_2^{-1} \\ I_1^{-1} D_{\kappa_3^o}' I_2^{-1} & 0 \end{pmatrix}.$$

In sum, we can see that if the parameter does not overlap, the third moment would not affect the asymptotic variance of the estimated mean parameter $\hat{\beta}$ and estimated variance parameter $\hat{\gamma}$, but rather play a role in their covariance. And the fourth moment would play a role only on the variance of the estimated variance parameter $\hat{\gamma}$. However, if they overlap, both the third and fourth moments of error would play a role in efficiency. Now, if the mean and variance parameters vary separately, the Gaussian QMLE is efficient for any value of κ_3^o. And we need $\kappa_4^o = 3$ only for the efficiency of $\hat{\gamma}$. Generally, if the mean and variance depend on the same parameters, $\kappa_4^o = 0$ and $\kappa_4^o = 3$ are needed.

It is worth noting that one can also implement a one-step QMLE, which is indicated for example in Newey and McFadden [1994]. For any consistent initial estimator $\check{\theta}$ of θ_o, we denote the estimated Hessian of the likelihood function to be

$$I_T(\theta) \stackrel{\text{def}}{=} -T^{-1} \sum_{t=1}^{T} \nabla_{\theta\theta'} \ell_t(\mathbf{y}_t, \mathbf{x}_t, \theta). \tag{15}$$

Then the one-step estimation of the Gaussian QMLE is defined to be

$$\overline{\theta} \stackrel{\text{def}}{=} \check{\theta} + I_T(\check{\theta})^{-1} T^{-1} \sum_{t=1}^{T} \mathbf{s}_t(\check{\theta}). \tag{16}$$

The theoretical properties of the one-step QMLE can be found in Theorem 8 in the Appendix.

5 Simulations

To evaluate the performance of our method, we run four Monte Carlo simulation experiments. We compare the finite sample performances of our method and QMLE. We show that for both univariate models and multivariate models, our method significantly outperforms QMLE in the case that the assumption on the normality is violated, while they perform similarly in the case that the innovation terms ε_t are normally distributed. The sample sizes of our experiments are $n = 500$, 1000, and 2000. The number of Monte Carlo replications is 500. We use the function *rsnorm* of the R package *fGarch* to generate series of skewed normal distribution. The resulting series has a skewness $\kappa_3^0 \approx 0.78$ and a kurtosis $\kappa_4^0 \approx 3.49$[1], compared with the standard normal distribution where $\kappa_3^0 = 0$ and a kurtosis $\kappa_4^0 = 3$.

5.1 Univariate models

We first show our results for a univariate GARCH/ARCH model.

5.1.1 Data generating process

We let the conditional variance $\sigma_t^2 = v_t(\mathbf{x}_t, \theta_0)$. Based on our discussion in Section 2, we have

$$\begin{aligned}\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o) \equiv Var\left[\mathbf{r}_t(\mathbf{w}_t, \theta_o)|\mathbf{x}_t\right] &= \begin{pmatrix} v_t(\mathbf{x}_t, \theta_o) & \kappa_3^o \left[v_t(\mathbf{x}_t, \theta_o)\right]^{3/2} \\ \kappa_3^o \left[v_t(\mathbf{x}_t, \theta_o)\right]^{3/2} & (\kappa_4^o - 1)\left[v_t(\mathbf{x}_t, \theta_o)\right]^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_t^2 & \kappa_3^0 \sigma_t^3 \\ \kappa_3^0 \sigma_t^3 & (\kappa_4^0 - 1)\sigma_t^4 \end{pmatrix}\end{aligned}$$

Let $\mathbf{D}_t(x_t, \theta_0, \kappa_0) := \mathbf{D}_t$, we have

$$\begin{aligned}\mathbf{D}_t^{-1} &= \frac{1}{(\kappa_4^0 - 1)\sigma_t^6 - (\kappa_3^0)^2 \sigma_t^6} \begin{pmatrix} (\kappa_4^0 - 1)\sigma_t^4 & -\kappa_3^0 \sigma_t^3 \\ -\kappa_3^0 \sigma_t^3 & \sigma_t^2 \end{pmatrix} \\ &=: \frac{1}{c^k} \begin{pmatrix} (\kappa_4^0 - 1)\sigma_t^{-2} & -\kappa_3^0 \sigma_t^{-3} \\ -\kappa_3^0 \sigma_t^{-3} & \sigma_t^{-4} \end{pmatrix}. \end{aligned} \tag{17}$$

[1]The arguments of the function *rsnorm* are chosen as: $mean = 0$, $sd = 1$, $xi = 2$

Case 1: ARCH(1)

We apply the score function (9) to ARCH(1) model,

$$\varepsilon_t = \sigma_t \eta_t,$$
$$\sigma_t^2 = \omega_0 + \alpha_0 \varepsilon_{t-1}^2,$$

where the noise term $\eta_t \overset{i.i.d}{\sim} (0,1)$. Recall that σ_t is the conditional volatility. The parameter of interest is $\theta = (\omega, \alpha)$. We recall from (8) that our estimator solves the following equation,

$$\sum_{t=1}^{T} (\check{\mathbf{D}}_t^{-1} \check{\mathbf{R}}_t)' \mathbf{r}_t(\mathbf{w}_t, \hat{\theta}) = 0. \tag{18}$$

For ARCH(1), we have

$$\mathbf{r}_t(\theta) = \begin{pmatrix} \varepsilon_t \\ \varepsilon_t^2 - \sigma_t^2 \end{pmatrix} = \begin{pmatrix} \varepsilon_t \\ \varepsilon_t^2 - \omega - \alpha \varepsilon_{t-1}^2 \end{pmatrix}, \tag{19}$$

$$\mathbf{R}_t = -\begin{pmatrix} 0 & 0 \\ 1 & \varepsilon_{t-1}^2 \end{pmatrix}. \tag{20}$$

Combining (18), (19), and (20), we can solve our optimal IV estimator from the following score function.

$$\begin{aligned} 0 &= \sum_{t=1}^{T} \check{\mathbf{R}}_t \check{\mathbf{D}}_t^{-1} \mathbf{r}_t(\hat{\theta}) \\ &= -\sum_{t=1}^{T} \frac{1}{\hat{c}^k} \begin{pmatrix} -\hat{\kappa}_3 \hat{\sigma}_t^{-3} & \hat{\sigma}_t^{-4} \\ -\hat{\kappa}_3 \hat{\sigma}_t^{-3} \varepsilon_{t-1}^2 & \hat{\sigma}_t^{-4} \varepsilon_{t-1}^2 \end{pmatrix} \begin{pmatrix} \varepsilon_t \\ \varepsilon_t^2 - \hat{\omega} - \hat{\alpha} \varepsilon_{t-1}^2 \end{pmatrix} \\ &= -\frac{1}{\hat{c}^k} \sum_{t=1}^{T} \begin{pmatrix} -\hat{\kappa}_3 \hat{\sigma}_t^{-3} \varepsilon_t + \frac{\varepsilon_t^2 - \hat{\omega} - \hat{\alpha} \varepsilon_{t-1}^2}{\hat{\sigma}_t^4} \\ -\hat{\kappa}_3 \hat{\sigma}_t^{-3} \varepsilon_{t-1}^2 \varepsilon_t + \frac{\varepsilon_t^2 - \hat{\omega} - \hat{\alpha} \varepsilon_{t-1}^2}{\hat{\sigma}_t^4} \varepsilon_{t-1}^2 \end{pmatrix}. \end{aligned} \tag{21}$$

The QMLE score function of ARCH(1) is (See, e.g., p.147 of Francq and Zakoian (2010))

$$\sum_{t=1}^{T} \begin{pmatrix} \frac{\varepsilon_t^2 - \hat{\omega} - \hat{\alpha} \varepsilon_{t-1}^2}{\hat{\sigma}_t^{-4}} \\ \frac{\varepsilon_t^2 - \hat{\omega} - \hat{\alpha} \varepsilon_{t-1}^2}{\hat{\sigma}_t^{-4}} \varepsilon_{t-1}^2 \end{pmatrix} = 0. \tag{22}$$

Case 2: GARCH (1,1) with mean

For a GARCH(1,1) model we have

$$
\begin{aligned}
y_t &= \mu + \lambda\sigma_t + \varepsilon_t, \\
\varepsilon_t &= \sigma_t \eta_t, \\
\sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \\
\mathbf{r}_t(\mathbf{w}_t, \theta_0) &= \begin{bmatrix} y_t - (\mu + \lambda\sigma_t), \\ \varepsilon_t^2 - (\omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2) \end{bmatrix}.
\end{aligned}
$$

The parameter of interest is $\theta = (\mu, \lambda, \omega, \alpha, \beta)$. And we have

$$
\mathbf{R}_t = -\begin{pmatrix} 1 & \sigma_t & \frac{\lambda}{2\sigma_t} & \frac{\lambda}{2\sigma_t}\varepsilon_{t-1}^2 & \frac{\lambda}{2\sigma_t}\sigma_{t-1}^2 \\ 0 & 0 & 1 & \varepsilon_{t-1}^2 & \sigma_{t-1}^2 \end{pmatrix}.
$$

Correspondingly, we have $\mathbf{D}_t^{-1}$ as in (17)

$$
\begin{aligned}
\mathbf{D}_t^{-1} &= \frac{1}{(\kappa_4^0 - 1)\sigma_t^6 - (\kappa_3^0)^2\sigma_t^6} \begin{pmatrix} (\kappa_4^0 - 1)\sigma_t^4 & -\kappa_3^0\sigma_t^3 \\ -\kappa_3^0\sigma_t^3 & \sigma_t^2 \end{pmatrix}, \\
&=: \frac{1}{c^k} \begin{pmatrix} (\kappa_4^0 - 1)\sigma_t^{-2} & -\kappa_3^0\sigma_t^{-3} \\ -\kappa_3^0\sigma_t^{-3} & \sigma_t^{-4} \end{pmatrix}.
\end{aligned}
$$

Also the optimal weight matrix is obtained as

$$
\check{\mathbf{R}}_t\check{\mathbf{D}}_t^{-1} = -\frac{1}{c^k} \begin{pmatrix} (\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-2} & -\hat{\kappa}_3\hat{\sigma}_t^{-3} \\ (\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-1} & -\hat{\kappa}_3\hat{\sigma}_t^{-2} \\ (\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-3}\frac{\lambda}{2} - \hat{\kappa}_3\hat{\sigma}_t^{-3} & \hat{\sigma}_t^{-4} - \hat{\kappa}_3\hat{\sigma}_t^{-4}\frac{\lambda}{2} \\ [(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-3}\frac{\lambda}{2} - \hat{\kappa}_3\hat{\sigma}_t^{-3}]\varepsilon_{t-1}^2 & [\hat{\sigma}_t^{-4} - \hat{\kappa}_3\hat{\sigma}_t^{-4}\frac{\lambda}{2}]\varepsilon_{t-1}^2 \\ [(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-3}\frac{\lambda}{2} - \hat{\kappa}_3\hat{\sigma}_t^{-3}]\hat{\sigma}_{t-1}^2 & [\hat{\sigma}_t^{-4} - \hat{\kappa}_3\hat{\sigma}_t^{-4}\frac{\lambda}{2}]\hat{\sigma}_{t-1}^2 \end{pmatrix}'.
$$

Let $\hat{\varepsilon}_t := y_t - \hat{\mu} - \lambda\hat{\sigma}_t$ and $\hat{\sigma}_t^2 = \omega + \alpha\hat{\varepsilon}_{t-1}^2 + \beta\hat{\sigma}_{t-1}^2$, the estimator solves the following equation.

$$
\begin{aligned}
0 &= \sum_{t=1}^{T} \check{\mathbf{R}}_t \check{\mathbf{D}}_t^{-1} \mathbf{r}_t(\hat{\theta}) \\
&= -\frac{1}{\hat{c}^k} \sum_{t=1}^{T} \begin{pmatrix}
(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-2}\hat{\varepsilon}_t - \hat{\kappa}_3\hat{\sigma}_t^{-3}(\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2) \\
(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-1}\hat{\varepsilon}_t - \hat{\kappa}_3\hat{\sigma}_t^{-2}(\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2) \\
\frac{\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2}{\hat{\sigma}_t^4}(1 - \hat{\kappa}_3\frac{\lambda}{2}) + [(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-3}\frac{\lambda}{2} - \hat{\kappa}_3\hat{\sigma}_t^{-3}]\hat{\varepsilon}_t \\
\frac{\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2}{\hat{\sigma}_t^4}(1 - \hat{\kappa}_3\frac{\lambda}{2})\hat{\varepsilon}_{t-1}^2 + [(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-3}\frac{\lambda}{2} - \hat{\kappa}_3\hat{\sigma}_t^{-3}]\hat{\varepsilon}_t\hat{\varepsilon}_{t-1}^2 \\
\frac{\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2}{\hat{\sigma}_t^4}(1 - \hat{\kappa}_3\frac{\lambda}{2})\hat{\sigma}_{t-1}^2 + [(\hat{\kappa}_4 - 1)\hat{\sigma}_t^{-3}\frac{\lambda}{2} - \hat{\kappa}_3\hat{\sigma}_t^{-3}]\hat{\varepsilon}_t\hat{\sigma}_{t-1}^2
\end{pmatrix}. \qquad (23)
\end{aligned}
$$

The score function of QMLE (multiplied by 2) is

$$
\sum_{t=1}^{T} \begin{pmatrix}
2\hat{\sigma}_t^{-2}\hat{\varepsilon}_t \\
2\hat{\sigma}_t^{-1}\hat{\varepsilon}_t \\
\lambda\hat{\sigma}_t^{-3}\hat{\varepsilon}_t + \frac{\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2}{\hat{\sigma}_t^4} \\
\lambda\hat{\sigma}_t^{-3}\hat{\varepsilon}_t\varepsilon_{t-1}^2 + \frac{\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2}{\hat{\sigma}_t^4}\varepsilon_{t-1}^2 \\
\lambda\hat{\sigma}_t^{-3}\hat{\varepsilon}_t\hat{\sigma}_{t-1}^2 + \frac{\hat{\varepsilon}_t^2 - \hat{\sigma}_t^2}{\hat{\sigma}_t^4}\hat{\sigma}_{t-1}^2
\end{pmatrix} = 0. \qquad (24)
$$

5.1.2 Simulation results

Table 1 and Table 2 present the Monte Carlo averages $\hat{\mu}$ and variances $\hat{\sigma}^2$ (multiplied by n) of each 500 estimates in the above two cases. The true values are chosen as $(\omega_0, \alpha_0) = (1, 0.1)$ for ARCH(1), and are $(\mu_0, \lambda_0, \omega_0, \alpha_0, \beta_0) = (2, 1.5, 1, 0.3, 0.3)$ for the GARCH (1,1) with mean model.

Table 1: Simulation results for ARCH(1), with $\omega_0 = 1$ and $\alpha_0 = 0.1$. OLS denotes the regular ordinary least square estimators; OPIV is our estimator.

		η_t is standard normal.				η_t is skewed normal.			
n	Methods	$\hat{\mu}_\omega$	$\hat{\mu}_\alpha$	$\hat{\sigma}^2_\omega$	$\hat{\sigma}^2_\alpha$	$\hat{\mu}_\omega$	$\hat{\mu}_\alpha$	$\hat{\sigma}^2_\omega$	$\hat{\sigma}^2_\alpha$
	OLS	1.0047	0.0951	3.8726	1.8665	1.0051	0.0911	4.4729	2.3845
500	QMLE	1.0015	0.0982	3.5846	1.6287	1.0001	0.0958	3.7918	1.9290
	OPIV	0.9962	0.0967	3.7452	1.7384	0.9960	0.0954	**3.2009**	**1.7298**
	OLS	1.0000	0.0987	4.3567	2.2480	0.9988	0.0982	5.0749	2.7023
1000	QMLE	1.0003	0.0984	3.8703	1.8652	0.9967	0.1000	4.1823	2.1477
	OPIV	0.9973	0.0979	3.9271	1.8916	0.9970	0.0991	**3.3964**	**1.6315**
	OLS	0.9997	0.0987	4.1172	2.0477	1.0016	0.0965	4.6833	2.7404
2000	QMLE	0.9993	0.0991	3.6429	1.7455	1.0001	0.0979	4.1662	2.1456
	OPIV	0.9978	0.0990	3.6565	1.7301	0.9994	0.0978	**3.6539**	**1.8497**

In Case 1, we can first compare (21) to (22). Note that the term $-\frac{1}{\hat{c}^k}$ does not affect the solution of (21) and can be ignored. The score function of the optimal IV estimator (21) has an additional item in each row. These terms are products of estimated skewness κ_3^0 and some zero-mean term. We notice that κ_4 does not play a role in the estimation efficiency. If the original distribution is symmetric, such as normal distribution, we have $\kappa_3^0 = 0$, so the optimal IV estimator should be approximately equivalent to QMLE. If the skewness of η_t is not zero, the optimal IV estimator should capture this information and beat QMLE in terms of efficiency. Table 1 confirms this conjecture.

In Table 1, we also include OLS estimators into our comparison since the ARCH model can be estimated by OLS. By comparing $\hat{\mu}$'s and the true parameters, we see that three estimators seem to be all asymptotically unbiased in general. Therefore, to compare the efficiency, we can focus on the variances. The left panel shows the simulation results with $\eta_t \overset{i.i.d}{\sim} N(0,1)$, which implies $\kappa_3^0 = 0$. OLS has the largest variances in every sample size since it does not take heteroskedasticity into account. If the sample size is small, the QMLE performs slightly better than our method since our method uses an estimated version of κ_3, which might be considerably different from the true $\kappa_3^0 = 0$ in small samples. If we increase the sample size to 2000, variances of QMLE and our method become very close to each other. The right panel shows the case where η_t follows the skewed normal distribution and $\kappa_3^0 \approx 0.78$. OLS still has the worst performance. With a non-zero κ_3^0, QMLE is no longer efficient, and we see that our method has achieved better performance in terms of variances, in every sample size and every parameter (as marked in bold). For example, in the sample

size $n = 2000$, the Monte Carlo variance of ω has dropped from 4.16 to 3.65, and the Monte Carlo variance of α has dropped from 2.15 to 1.85.

In Case 2, the results are similar. First, we can compare (23) and (24). We see that they are equivalent if $k_3^0 = 0$ and $k_4^0 = 3$, which are satisfied with $\eta_t \sim N(0,1)$. In this case, QMLE and our method should perform similarly. This is confirmed in the upper panel of Table 2. If η_t follows the skewed normal distribution, our estimator should capture the departure from the standard normal distribution and beat QMLE in terms of efficiency. This phenomenon is illustrated in the lower panel of Table 2. As marked in bold, our method outperforms QMLE in the case $\kappa_3^0 \approx 0.78$ and $\kappa_4^0 \approx 3.49$, for all different sample sizes and almost every coefficient.

Table 2: Simulation results for GARCH (1,1), with mean $\mu_0 = 2$, $\lambda_0 = 1.5$, $\omega_0 = 1$, $\alpha_0 = 0.3$, and $\beta_0 = 0.3$.

n	Methods	$\hat{\mu}_\mu$	$\hat{\mu}_\lambda$	$\hat{\mu}_\omega$	$\hat{\mu}_\alpha$	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\mu$	$\hat{\sigma}^2_\lambda$	$\hat{\sigma}^2_\omega$	$\hat{\sigma}^2_\alpha$	$\hat{\sigma}^2_\beta$
					η_t is standard normal.						
500	QMLE	1.9676	1.5313	1.0053	0.2939	0.2993	87.75	43.36	26.92	1.81	6.43
	OPIV	1.9465	1.5454	1.0059	0.2920	0.3003	94.50	46.63	25.73	1.80	6.32
1000	QMLE	1.9923	1.5101	0.9940	0.2969	0.3035	96.05	45.21	25.93	2.12	5.88
	OPIV	1.9958	1.5074	0.9940	0.2969	0.3036	84.62	40.59	26.02	2.07	5.95
2000	QMLE	1.9990	1.5031	0.9946	0.2990	0.3017	89.83	42.63	25.40	2.10	6.04
	OPIV	1.9856	1.5124	0.9950	0.2981	0.3022	108.48	51.69	25.63	2.22	6.18
					η_t is skewed normal.						
500	QMLE	1.9798	1.5183	1.0445	0.2903	0.2861	94.17	43.90	46.33	3.08	9.94
	OPIV	1.9211	1.5588	1.0436	0.2891	0.2874	**84.87**	**40.85**	**36.70**	3.33	**8.91**
1000	QMLE	1.9520	1.5369	1.0177	0.2966	0.2925	175.74	85.56	37.75	3.59	10.29
	OPIV	1.9849	1.5138	1.0134	0.2992	0.2923	**79.75**	**39.56**	**37.75**	**3.34**	10.30
2000	QMLE	2.0033	1.4988	1.0106	0.2995	0.2956	103.21	45.30	39.25	3.02	9.37
	OPIV	1.9908	1.5080	1.0102	0.2983	0.2965	**70.38**	**34.64**	**35.41**	3.04	**8.72**

5.2 Multivariate models

5.2.1 Data generating process

In this subsection, we show our methods can perform well in multivariate models. For simplicity, we consider two dimensional cases. Now we have $\eta_t \overset{i.i.d}{\sim} (0,1)^2$. Let $\kappa_3^0 = E(\eta_{1,t}^3)$, and $\kappa_4^0 = E(\eta_{1,t}^4)$.

In Section 3 we define

$$\mathbf{D} \equiv \text{Var}[\mathbf{r}_t(\mathbf{w}_t, \theta_0)|\mathbf{x}_t] = \begin{pmatrix} \Sigma_t & \Sigma_{12,t} \\ \Sigma'_{12,t} & \Sigma_{22,t} \end{pmatrix}, \tag{25}$$

where Σ_t is the abbreviation of $\Sigma_t(\mathbf{x}_t, \theta_o)$ defined in Section 3, and $\Sigma'_{12,t} = L_2\Sigma_t^{1/2} \otimes \Sigma_t^{1/2}\mathbf{K}_3^0\Sigma_t^{1/2}$, $\Sigma_{22,t} = L_2(\Sigma_t^{1/2} \otimes \Sigma_t^{1/2}\mathbf{K}_4^0\Sigma_t^{1/2} \otimes \Sigma_t^{1/2} - \text{Vec}\Sigma_t\text{Vec}\Sigma'_t\,)L'_2$,

$$\text{with } L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{K}_3^0 = E(\eta_t \otimes \eta_t\eta'_t) = \begin{pmatrix} \kappa_3^0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & \kappa_3^0 \end{pmatrix} \text{ and } \mathbf{K}_4^0 = E(\eta_t \otimes \eta_t\eta'_t \otimes \eta'_t) = \begin{pmatrix} \kappa_4^0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & \kappa_4^0 \end{pmatrix}.$$

Case 3: CCC-GARCH

Now let ε_t be two dimensional. We have the Constant Conditional Correlations model

$$\begin{cases} \varepsilon_t = \Sigma_t^{1/2}\eta_t \\ \Sigma_t = \Lambda_t\Gamma_t\Lambda_t \\ \sigma_{1,t}^2 = \omega_1 + \alpha_1\varepsilon_{1,t-1}^2 + \beta_1\sigma_{k,t-1}^2 \\ \sigma_{2,t}^2 = \omega_2 + \alpha_2\varepsilon_{2,t-1}^2 + \beta_2\sigma_{2,t-1}^2 \end{cases},$$

with

$$\Lambda_t = \begin{pmatrix} \sigma_{1,t} & 0 \\ 0 & \sigma_{2,t} \end{pmatrix} \Gamma_t = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We have to ensure that Σ_t is positive definite. Thus we need $-1 < \rho < 1$. We reparameterize $\rho = \sin\delta$, to avoid adding parameter restriction in the estimation. Note $\delta = \arcsin\rho$.

Then we have the score function written as,

$$\mathbf{r}_t(\mathbf{w}_t,\theta)=\begin{pmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\\ \varepsilon_{1,t}^2-(\omega_1+\alpha_1\varepsilon_{1,t-1}^2+\beta_1\sigma_{1,t-1}^2)\\ \varepsilon_{1,t}\varepsilon_{2,t}-\sin\delta\sigma_{1,t}\sigma_{2,t}\\ \varepsilon_{2,t}^2-(\omega_2+\alpha_2\varepsilon_{2,t-1}^2+\beta_2\sigma_{2,t-1}^2)\end{pmatrix},$$

where the parameter of interest are $\theta=(\omega_1,\alpha_1,\beta_1,\omega_2,\alpha_2,\beta_2,\delta)$.

$$\mathbf{R}_t=-\begin{pmatrix}0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 1&\varepsilon_{1,t-1}^2&\sigma_{1,t-1}^2&0&0&0&0\\ \frac{\sin\delta\sigma_{2,t}}{2\sigma_{1,t}}&\frac{\sin\delta\sigma_{2,t}}{2\sigma_{1,t}}\varepsilon_{1,t-1}^2&\frac{\sin\delta\sigma_{2,t}}{2\sigma_{1,t}}\sigma_{1,t-1}^2&\frac{\sin\delta\sigma_{1,t}}{2\sigma_{2,t}}&\frac{\sin\delta\sigma_{1,t}}{2\sigma_{2,t}}\varepsilon_{2,t-1}^2&\frac{\sin\delta\sigma_{1,t}}{2\sigma_{2,t}}\sigma_{2,t-1}^2&\sigma_{1,t}\sigma_{2,t}\cos\delta\\ 0&0&0&1&\varepsilon_{2,t-1}^2&\sigma_{2,t-1}^2&0\end{pmatrix}.$$

With D_t defined in (25), we can construct our sample moment condition $\sum_{t=1}^{T}\check{\mathbf{R}}_t\check{\mathbf{D}}_t^{-1}\mathbf{r}_t(\hat{\theta})=0$ accordingly.

Case 4: BEKK-GARCH

We consider the BEKK-GARCH model

$$\begin{cases}\varepsilon_t=\Sigma_t^{1/2}\eta_t\\ \Sigma_t=C+A\varepsilon_{t-1}\varepsilon_{t-1}'A'+B\Sigma_{t-1}B'\end{cases}.$$

In the simulation, we set

$$C=\begin{pmatrix}c_{11}&c_{12}\\ c_{12}&c_{22}\end{pmatrix}\ A=\begin{pmatrix}a_{11}&0\\ 0&a_{22}\end{pmatrix}\ B=0.$$

For this model we have

$$\mathbf{r}_t(\mathbf{w}_t,\theta)=\begin{pmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\\ \varepsilon_{1,t}^2-(c_{11}+a_{11}^2\varepsilon_{1,t-1}^2)\\ \varepsilon_{1,t}\varepsilon_{2,t}-(c_{12}+\varepsilon_{1,t-1}\varepsilon_{2,t-1}a_{11}a_{22})\\ \varepsilon_{2,t}^2-(c_{22}+a_{22}^2\varepsilon_{2,t-1}^2)\end{pmatrix}.$$

where the parameter of interest are $\theta = (c_{11}, c_{12}, c_{22}, a_{11}, a_{22})$. And we have

$$\mathbf{R}_t = -\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2a_{11}\varepsilon_{1,t-1}^2 & 0 \\ 0 & 1 & 0 & a_{22}\varepsilon_{1,t-1}\varepsilon_{2,t-1} & a_{11}\varepsilon_{1,t-1}\varepsilon_{2,t-1} \\ 0 & 0 & 1 & 0 & 2a_{22}\varepsilon_{2,t-1}^2 \end{pmatrix},$$

with D defined in (25). We can solve the sample moment condition $\sum_{t=1}^{T} \check{\mathbf{R}}_t \check{\mathbf{D}}_t^{-1} \mathbf{r}_t(\hat{\theta}) = 0$ accordingly.

5.2.2 Simulation results

Table 3 and Table 4 present the Monte Carlo averages $\hat{\mu}$ and variances $\hat{\sigma}^2$ (multiplied by n) of each 500 estimates for Case 3 and Case 4. The true values are $(\omega_{0,1}, \alpha_{0,1}, \beta_{0,1}, \omega_{0,2}, \alpha_{0,2}, \beta_{0,2}, \rho_0) = (0.4, 0.8, 0.15, 0.2, 0.7, 0.2, 0.7)$ for the CCC-GARCH(1,1) model, and $(c_{0,11}, c_{0,12}, c_{0,22}, a_{0,11}, a_{0,22}) = (0.8, 0.5, 0.7, 0.6, 0.5)$ in the BEKK-GARCH model.

Table 3: Simulation results of CCC-GARCH (1,1), with $(\omega_{0,1}, \alpha_{0,1}, \beta_{0,1}, \omega_{0,2}, \alpha_{0,2}, \beta_{0,2}, \rho_0) = (0.4, 0.8, 0.15, 0.2, 0.7, 0.2, 0.7)$

η_t is standard normal.								
mean								
n	Methods	$\hat{\mu}_{\omega_1}$	$\hat{\mu}_{\alpha_1}$	$\hat{\mu}_{\beta_1}$	$\hat{\mu}_{\omega_2}$	$\hat{\mu}_{\alpha_2}$	$\hat{\mu}_{\beta_2}$	$\hat{\mu}_{\rho}$
500	QMLE	0.4097	0.7982	0.1466	0.2043	0.6968	0.1978	0.7028
	OPIV	0.4071	0.8051	0.1452	0.2034	0.7015	0.1962	0.7042
1000	QMLE	0.4048	0.7982	0.1474	0.2012	0.6985	0.1990	0.7002
	OPIV	0.4042	0.8015	0.1458	0.2005	0.7020	0.1979	0.7008
2000	QMLE	0.4011	0.7986	0.1499	0.2009	0.6973	0.1996	0.7004
	OPIV	0.4015	0.8027	0.1485	0.2009	0.7003	0.1985	0.7009
variance								
n	Methods	$\hat{\sigma}^2_{\omega_1}$	$\hat{\sigma}^2_{\alpha_1}$	$\hat{\sigma}^2_{\beta_1}$	$\hat{\sigma}^2_{\omega_2}$	$\hat{\sigma}^2_{\alpha_2}$	$\hat{\sigma}^2_{\beta_2}$	$\hat{\sigma}^2_{\rho}$
500	QMLE	1.8855	4.2735	0.9867	0.5283	3.5460	1.3148	0.2551
	OPIV	1.8424	4.2181	1.0391	0.5024	3.5502	1.2925	0.2301
1000	QMLE	1.7578	4.0821	0.9992	0.4972	4.0567	1.2688	0.2668
	OPIV	1.7521	4.0926	1.0886	0.4450	4.3099	1.3089	0.2515

2000	QMLE	1.6287	4.7403	0.9895	0.491	3.6837	1.3723	0.2554
	OPIV	1.6286	4.7654	1.0249	0.4387	3.7833	1.4295	0.2440

η_t is skewed normal.

mean

n	Methods	$\hat{\mu}_{\omega_1}$	$\hat{\mu}_{\alpha_1}$	$\hat{\mu}_{\beta_1}$	$\hat{\mu}_{\omega_2}$	$\hat{\mu}_{\alpha_2}$	$\hat{\mu}_{\beta_2}$	$\hat{\mu}_{\rho}$
500	QMLE	0.4057	0.7971	0.1466	0.2054	0.6942	0.1967	0.7011
	OPIV	0.4025	0.8048	0.1433	0.2031	0.7017	0.1953	0.7020
1000	QMLE	0.4022	0.7993	0.1486	0.2022	0.7008	0.1969	0.7013
	OPIV	0.4035	0.8051	0.1461	0.2012	0.7036	0.1965	0.7021
2000	QMLE	0.4024	0.7961	0.1495	0.2004	0.6977	0.2000	0.7008
	OPIV	0.4033	0.7985	0.1487	0.2006	0.7005	0.1989	0.7010

variance

n	Methods	$\hat{\sigma}^2_{\omega_1}$	$\hat{\sigma}^2_{\alpha_1}$	$\hat{\sigma}^2_{\beta_1}$	$\hat{\sigma}^2_{\omega_2}$	$\hat{\sigma}^2_{\alpha_2}$	$\hat{\sigma}^2_{\beta_2}$	$\hat{\sigma}^2_{\rho}$
500	QMLE	2.3803	4.6442	1.3126	0.7193	4.4313	1.7120	0.2642
	OPIV	**1.6862**	**3.6131**	**1.0543**	**0.5342**	**3.4227**	**1.4295**	**0.2509**
1000	QMLE	2.1661	5.1899	1.1421	0.6381	3.9831	1.6058	0.2399
	OPIV	**1.6571**	**3.9749**	**0.8762**	**0.4389**	**2.9682**	**1.2012**	**0.2278**
2000	QMLE	2.2955	5.2833	1.3333	0.6570	4.5339	1.7018	0.2270
	OPIV	**1.8981**	**4.3921**	**1.0868**	**0.4812**	**3.6749**	**1.3623**	**0.2119**

In Case 3, both QMLE and our method seem to be consistent and asymptotically unbiased. We can see that in both upper panel and lower panel of Table 3, the Monte Carlo mean of estimates $\hat{\mu}$'s are converging to the true values as sample sizes grow. Therefore, we can focus on the comparison of variances. For multivariate cases, as illustrated in Section 3, both κ_3 and κ_4 matter. The upper panel shows the simulation results with $\eta_t \overset{i.i.d}{\sim} N(0,1)$, which implies $\kappa_3^0 = 0$ and $\kappa_4^0 = 3$. We see that across all the sample sizes, QMLE and our method have similar performances. The Monte Carlo variances of these two methods are very close across the whole upper panel. The differences in variances are mostly smaller than 0.1. The lower panel shows the case where η_t follows the skewed normal distribution with $\kappa_3^0 \approx 0.78$ and $\kappa_4^0 \approx 3.49$. As predicted, QMLE is no longer efficient due to the model misspecification. We see that our method has achieved smaller variances in every sample size and every coefficient, as marked in bold. For example, in the sample size $n = 2000$, our method decreases the Monte Carlo variances by approx. 21%, 20%, 23%, 37%, 23%, 25%, and 7%.

In Case 4 for the BEKK-GARCH model, we see the same pattern: our methods and QMLE have similar performances if the underline η_t is correctly specified as $N(0,1)$. If η_t is drawn from a skewed normal distribution, our method outperforms QMLE in every sample size and every parameter, as marked in bold.

To summarize, in both univariate models and multivariate models, the simulation results show that our method has good performances. While QMLE and our methods perform similarly in the case of normally distributed η_t, our methods outperform QMLE in the case of skewed normally distributed η_t. Overall, the simulation results are encouraging to support our estimation strategy.

Table 4: Simulation results of BEKK-GARCH, with $c_{0,11} = 0.5$, $c_{0,12} = 0.5$, $c_{0,22} = 0.7$, $a_{0,11} = 0.6$, and $a_{0,22} = 0.5$.

					η_t is standard normal.						
n	Methods	$\hat{\mu}_{c_{11}}$	$\hat{\mu}_{c_{12}}$	$\hat{\mu}_{c_{22}}$	$\hat{\mu}_{a_{11}}$	$\hat{\mu}_{a_{22}}$	$\hat{\sigma}^2_{c_{11}}$	$\hat{\sigma}^2_{c_{12}}$	$\hat{\sigma}^2_{c_{22}}$	$\hat{\sigma}^2_{a_{11}}$	$\hat{\sigma}^2_{a_{22}}$
500	QMLE	0.8040	0.5054	0.7035	0.5958	0.4938	2.2722	1.1816	1.3283	1.3559	1.2791
	OPIV	0.7991	0.5024	0.699	0.5938	0.4919	2.1995	1.1181	1.2595	1.3311	1.2721
1000	QMLE	0.7998	0.4995	0.6992	0.5972	0.4968	2.4484	1.2653	1.3792	1.3667	1.2274
	OPIV	0.7973	0.4981	0.6973	0.5962	0.4959	2.4135	1.2464	1.3583	1.3577	1.2146
2000	QMLE	0.7992	0.4997	0.6991	0.5993	0.4988	2.3611	1.2553	1.5122	1.4003	1.1858
	OPIV	0.7982	0.4991	0.6982	0.5988	0.4984	2.3353	1.2221	1.4894	1.398	1.1806
					η_t is skewed normal.						
n	Methods	$\hat{\mu}_{c_{11}}$	$\hat{\mu}_{c_{12}}$	$\hat{\mu}_{c_{22}}$	$\hat{\mu}_{a_{11}}$	$\hat{\mu}_{a_{22}}$	$\hat{\sigma}^2_{c_{11}}$	$\hat{\sigma}^2_{c_{12}}$	$\hat{\sigma}^2_{c_{22}}$	$\hat{\sigma}^2_{a_{11}}$	$\hat{\sigma}^2_{a_{22}}$
500	QMLE	0.8005	0.5028	0.7024	0.5943	0.4933	2.5267	1.3466	1.6874	1.5667	1.4354
	OPIV	0.794	0.4996	0.6992	0.5949	0.493	**1.9389**	**1.127**	**1.2894**	**1.2861**	**1.1772**
1000	QMLE	0.7978	0.5002	0.6994	0.5987	0.4984	2.7234	1.3224	1.7553	1.6537	1.4653
	OPIV	0.7968	0.4988	0.6965	0.5982	0.4979	**2.3056**	**1.1948**	**1.3877**	**1.4348**	**1.1959**
2000	QMLE	0.8013	0.501	0.6996	0.5959	0.4965	2.7604	1.3792	2.0095	1.6591	1.5410
	OPIV	0.8018	0.5011	0.6998	0.5952	0.4966	**2.2133**	**1.1999**	**1.7316**	**1.3656**	**1.2623**

6 Application

In this section, we illustrate the use of our methods by modeling the processes of two stock price series, i.e. Apple (APPL) and Amazon (AMZN). The observations are dated from Feb. 2013 to April. 2014. The data is downloaded from Yahoo Finance, *finance.yahoo.com*.

Figure 1 shows the original series of these two stock prices. The solid red line represents the stock price

of APPL, and the dashed blue line represents the stock price of AMZN. Compared to the price of APPL, the price of AMZN is on a higher level and seems to show a larger range of variation. Therefore, we work with their logarithmic first-difference series. Figure 2 shows these series (re-scaled by 100), and Table 5 summarizes their descriptive statistics.

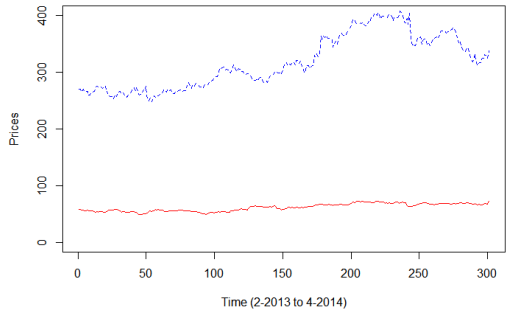

Figure 1: Plot of stock price. The solid red line represents that of APPL, and the dashed blue line represents that of AMZN.

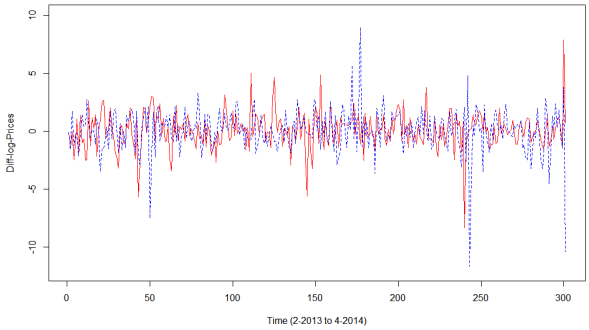

Figure 2: Plot of the first difference of log price. The solid red line represents that of APPL, and the dashed blue line represents that of AMZN.

Table 5: Summary statistics of the differenced logarithmic series of stock prices

Stock	Mean	Std. dev.	Min.	Max	Std.skewness	Std.kurtosis
APPL	0.0756	1.5812	-8.3302	7.8795	-0.1685	4.8011
AMZN	0.0398	1.8743	-11.6503	8.9709	-1.1588	9.2150

In the first-difference logarithmic series, the price of AMZN still shows a larger variance and data range. The sample skewness and kurtosis of both series suggest clear deviations from the normal distribution. In this application, we use AR(1) to model the conditional mean function of the first-difference logarithmic series of these two stock prices and apply QMLE and our methods to their AR(1) residuals. For both QMLE and our methods, we fit the CCC-GARCH (1,1) model described in Section 5.2.1. The parameters of interest are $\theta = (\omega_1, \alpha_1, \beta_1, \omega_2, \alpha_2, \beta_2, \rho)$, and the model is specified as

$$\mathbf{r}_t(\mathbf{w}_t, \theta) = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{1,t}^2 - (\omega_1 + \alpha_1 \varepsilon_{1,t-1}^2 + \beta_1 \sigma_{1,t-1}^2) \\ \varepsilon_{1,t}\varepsilon_{2,t} - \rho\sigma_{1,t}\sigma_{2,t} \\ \varepsilon_{2,t}^2 - (\omega_2 + \alpha_2 \varepsilon_{2,t-1}^2 + \beta_2 \sigma_{2,t-1}^2) \end{bmatrix}.$$

Table 6 summarizes the estimation results of QMLE and our method. $(\omega_{ap}, \alpha_{ap}, \beta_{ap})$ are coefficients for APPL, and $(\omega_{am}, \alpha_{am}, \beta_{am})$ are coefficients for AMZN. From the table, we see that the estimates from QMLE and our methods are close to each other. All the coefficients of the constant terms are significant. It seems that in this dataset, the constant term and the ARCH effects dominate since all the estimates of GARCH parameters are close to zero and most of them are insignificant. It suggests that during this period, the main factor affecting the variation of stock prices is the realized historical innovation. Among the parameters for ARCH effects, the coefficients in our method are all significant, and the coefficients of the second stock price in QMLE are insignificant. We see a weak correlation between the prices of two stocks. The estimated ρ of QMLE and our method are significant at 5% and 10% level. This is in line with the pattern in Figure 2. During this period, the stock prices of these two companies are rarely moving in the same direction together.

Table 6: Estimation results of CC-MGARCH(1,1) for QMLE and our method. Subscript $_{ap}$ denotes parameters for Apple, and $_{am}$ denotes parameters for Amazon.

Methods	ω_{ap}	α_{ap}	β_{ap}	ω_{am}	α_{am}	β_{am}	ρ
QMLE	1.5212	0.3606	0.1361	1.9509	0.5280	0.0000	0.1195
	(0.3809)***	(0.1373)***	(0.0792)*	(0.1154)***	(0.4088)	(0.1294)	(0.0577)**
OPIV	1.4953	0.4400	0.1092	1.9950	0.5723	0.0296	0.1127
	(0.3733)***	(0.1648)***	(0.1504)	(0.3970)***	(0.2786)**	(0.0532)	(0.0667)*

7 Proof

We show in this section a few theoretical properties of our proposed estimators.

7.1 Consistency

Assume that $\{\mathbf{y}_t, \mathbf{x}_t\}_t$ are strictly stationary and $\mathcal{F}_t = \sigma(\mathbf{y}_t, \mathbf{x}_t)$ is the sigma field generated by $\mathbf{y}_t, \mathbf{x}_t$. We assume that the score function $\{\mathbf{r}_t(\theta)\}_{t\geq 1}$ is constructed as stationary, ergodic, and square-integrable

martingale difference sequence with respect to $\mathcal{F}_t$ under the dynamic completeness assumption (3) and (4). The weights $\mathbf{R}_t(\theta_o)\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1}$ are measurable with respect to $\mathcal{F}_{t-1}$. Moreover, to ensure that our nuisance parameter κ_3, κ_4 exists, we assume that for the process $\{\mathbf{y}_t, \mathbf{x}_t\}_t$, $T^{-1}\sum_{t=1}^{T} e_{tj}^4 \to_p c_{4o}$ and $T^{-1}\sum_{t=1}^{T} e_{tj}^3 \to_p c_{3o}$, where $\kappa_o = (\kappa_{3o}, \kappa_{4o})$ are two constants.

Define for our general estimation equation on the parameter space $\Theta \times \Gamma$. θ is known to be the parameter of interest, and $\kappa = (\kappa_3, \kappa_4) \in \Gamma$ is denoted as the nuisance parameter. Define the score function

$$Q_T(\check{\theta}, \check{\kappa}, \theta) = T^{-1}\sum_{t}^{T} \check{\mathbf{R}}_t' \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})^{-1}\mathbf{r}_t(\theta).$$

And we can see that

$$(\theta_o, \kappa_o) = \text{argzero}_{\theta\in\Theta, \kappa\in\Gamma} Q_\infty(\theta, \kappa),$$

where $Q_\infty(\theta, \kappa) = \lim_{T\to\infty} T^{-1}\sum_t^T \mathrm{E}[\mathbf{R}_t(\theta)'\mathbf{D}_t(\mathbf{x}_t, \theta, \kappa)^{-1}\mathbf{r}_t(\theta)]$. Also, the theoretical estimator is defined to be

$$\theta_o = \text{argzero}_{\theta\in\Theta} Q_\infty(\theta_o, \kappa_o, \theta),$$

where $Q_\infty(\theta_o, \kappa_o, \theta) = \lim_{T\to\infty} T^{-1}\sum_t^T \mathrm{E}[\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1}\mathbf{r}_t(\theta)]$. Our estimator is

$$\hat{\theta} = \text{argzero}_{\theta\in\Theta} Q_T(\check{\theta}, \check{\kappa}, \theta),$$

with plugged-in estimators $\check{\theta}$, $\check{\kappa}$.

We shall show the consistency of $\hat{\theta} \to_p \theta_o$. We need to set a few high order conditions. Define the $|v|_2$ as the l_2 norm or a vector v, and $|A|_2$ as the 2-norm of a matrix A.

A.1 $\{\mathbf{y}_t, \mathbf{x}_t\}_t$ are strictly stationary. The score function $\{\mathbf{r}_t(\theta)\}_{t\geq 1}$ is constructed as stationary, ergodic, and square-integrable martingale difference sequence with respect to $\mathcal{F}_t$ under the dynamic completeness assumption (3) and (4).

A.2 (Uniform consistency) $\sup_{\theta\in\Theta}|Q_T(\theta_o, \kappa_o, \theta) - Q_\infty(\theta_o, \kappa_o, \theta)|_2 \to_p 0$

A.3 (Identifiability) For any constant $\varepsilon > 0$, we have $\sup_{|\theta-\theta_o|_2\geq\varepsilon} |Q_\infty(\theta_o, \kappa_o, \theta)|_2 > 0 = |Q_\infty(\theta_o, \kappa_o, \theta_o)|_2$.

A.4 The empirical estimator is almost the point of our first-order function $Q_T(\check{\theta}, \check{\kappa}, \hat{\theta}) = \mathcal{O}_p(1)$, and the plugged-in estimator is not affecting our estimation, i.e., $\sup_{\theta\in\Theta}|Q_T(\check{\theta}, \check{\kappa}, \theta) - Q_T(\theta_o, \kappa_o, \theta)|_2 = \mathcal{O}_p(1)$.

Theorem 2. *Under A.1-A.4, we have $\hat{\theta} \to_p \theta_o$.*

Proof. See Theorem 5.9 in Van der Vaart [2000]. By [A.3] for any positive constant ε, exits a positive constant

δ such that $\mathrm{P}(|\hat{\theta}-\theta_o)|_2>\varepsilon)\leq \mathrm{P}(|Q_\infty(\theta_o,\kappa,\hat{\theta}|_2>\delta))$. $\mathrm{P}(|Q_\infty(\theta_o,\kappa_o,\hat{\theta})|_2>\delta)\to 0$ as $|Q_\infty(\theta_o,\kappa_o,\hat{\theta})|_2\leq |Q_\infty(\theta_o,\kappa_o,\hat{\theta})|_2-|Q_T(\theta_o,\kappa_o,\hat{\theta})|_2+\mathcal{o}_p(1)\leq \sup_{\theta\in\Theta}|Q_\infty(\theta_o,\kappa_o,\theta)-Q_T(\theta_o,\kappa_o,\theta)|_2+\mathcal{o}_p(1)=\mathcal{o}_p(1)$, which is implied by [A.2] and [A.4]. □

Uniform convergence of the criterion functions is not hard to obtain in normal cases.

In principle, [A.2] is a relatively very strong assumption, and we can switch to a different set of assumptions such as lower semi-continuous and compactness parameter set from the Wald consistency proof.

A.1' For every $\theta\in\Theta$, $\liminf_{T\to\infty}|Q_T(\theta_o,\kappa_o,\theta)|_2\geq \lim_{T\to\infty}|\mathrm{E}Q_T(\theta_o,\kappa_o,\theta)|_2$, a.s.

A.2' Θ is compact. The moment is bounded $|\mathrm{E}\sup_{\theta\in\Theta}Q_T(\theta_o,\kappa_o,\theta)|_2<\infty$.

A.4' We have pointwise weak law of large numbers. $Q_T(\theta_o,\kappa_o,\theta)\to_p Q_\infty(\theta_o,\kappa_o,\theta)$ for every θ.

Remark. *The nuisance parameters need not to converge to the true parameters θ_o,κ_o. Therefore, the above condition can be changed to $Q_T(\theta^*,\kappa^*,\theta)\to_p Q_\infty(\theta^*,\kappa^*,\theta)$ for any θ^*,κ^* in the parameter space.*

Theorem 3. *Under A.1, A.1', A.2', A.3, A.4', we have $\hat{\theta}\to_p\theta_o$.*

Proof. $V_k(\theta_i)\overset{\text{def}}{=}\{\theta:|\theta-\theta_i|_2<1/k\}$ is a sequence of open balls centered around $\theta_i\in\Theta$. For any $\theta\in\Theta\neq\theta_o$, with a sequence of increase k, $V_k(\theta)$ is shrinking in size and thus

$$\liminf_{\theta\to V_k(\theta)}|\lim_{T\to\infty}\mathrm{E}Q_T(\theta_o,\kappa_o,\theta)|_2\uparrow|Q_\infty(\theta_o,\kappa_o,\theta)|_2>0=|Q_\infty(\theta_o,\kappa_o,\theta_o)|_2 \tag{26}$$

by monotone convergence theorem and the notion of low semi-continuous as in [A.1'].

Thus, for any $\theta_i\neq\theta_o\in\Theta$ we can find a $k(\theta_i)$ such that

$$\liminf_{\theta\in V_{k(\theta_i)}(\theta_i)}|\lim_{T\to\infty}\mathrm{E}|Q_T(\theta_o,\kappa_o,\theta)|_2>0=|Q_\infty(\theta_o,\kappa_o,\theta_o)|_2.$$

Define $\delta=\min\{\inf_{i\in1,\cdots,l}\liminf_{\theta\in V_{k(\theta_i)}(\theta_i)}\lim_{T\to\infty}\mathrm{E}|Q_T(\theta_o,\kappa_o,\theta)|_2,1\}$.

Consider the compact set $\Theta_\varepsilon\overset{\text{def}}{=}\{\theta:|\theta-\theta_o|_2\geq\varepsilon\}$ for a positive constant ε, and we can find a finite subcover of $\Theta_\varepsilon\subset\cup_{i=1}^l V_{k(\theta_i)}(\theta_i)$.

Notice that

$$\mathrm{P}(|\hat{\theta}-\theta_o|\geq\varepsilon)\leq\mathrm{P}(\inf_{i=1,\cdots,l}\inf_{\theta'\in V_{k(\theta_i)}(\theta_i)}|Q_T(\check{\theta},\check{\kappa},\theta')|_2-|Q_T(\theta_o,\kappa_o,\theta_o)|_2\leq 0).$$

We now prove that the right-hand side of the equation is $\mathcal{o}(1)$.

By [A.4'] and (26), $\inf_{i=1,\cdots,l}\inf_{\theta'\in V_{k(\theta_i)}(\theta_i)}|Q_T(\check{\theta},\check{\kappa},\theta')|_2$ can be made big enough, for example, δ, with probability approaching 1. By [A.4'], $|Q_T(\theta_o,\kappa_o,\theta_o)|_2$ can be made close to $|Q_\infty(\theta_o,\kappa_o,\theta_o)|_2 = 0$ with probability 1, say $\mathrm{P}(|Q_T(\theta_o,\kappa_o,\theta_o)|_2 > \delta/2) = \mathcal{o}(1)$, for a positive constant δ.

Thus, $\mathrm{P}(|\hat{\theta}-\theta_o|\geq\varepsilon) \leq \mathrm{P}(\delta/2\leq 0) + \mathcal{o}(1) = \mathcal{o}(1)$. □

7.2 Normality

In this section we show the asymptotic normality of our estimator. We first introduce a few definitions, recall that

$$
\begin{aligned}
&\mathbf{R}_t(\tilde{\theta}) = \nabla\mathbf{r}_t(\tilde{\theta}) = \partial\mathbf{r}_t(\theta))/\partial\theta'|_{\theta=\tilde{\theta}},\\
&\check{Q}_T(\theta_o) = Q_T(\check{\theta},\check{\kappa},\theta_o) \overset{\text{def}}{=} T^{-1}\sum_{t=1}^{T}\check{\mathbf{R}}_t'\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}\mathbf{r}_t(\theta_o),\\
&\check{A}_o(\theta_o) = Q_\infty(\check{\theta},\check{\kappa},\theta_o) \overset{\text{def}}{=} \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{E}(\check{\mathbf{R}}_t'(\check{\theta})\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}\mathbf{r}_t(\theta_o)),\\
&\check{B}_T(\tilde{\theta}) \overset{\text{def}}{=} T^{-1}\sum_{t=1}^{T}\check{\mathbf{R}}_t'\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}\nabla\mathbf{r}_t(\tilde{\theta}),\\
&B_T(\tilde{\theta}) \overset{\text{def}}{=} T^{-1}\sum_{t=1}^{T}\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}(\tilde{\theta}),\\
&\check{B}_o(\tilde{\theta}) \overset{\text{def}}{=} \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{E}(\check{\mathbf{R}}_t'\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}\nabla\mathbf{r}_t(\tilde{\theta})).
\end{aligned}
\tag{27}
$$

In addition we assume that

$$
\begin{aligned}
&A_o \overset{\text{def}}{=} \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o)),\\
&B_o \overset{\text{def}}{=} \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)).
\end{aligned}
\tag{28}
$$

Define $V_{\theta_o}(c)$ as an $1/\sqrt{T}$-ball around θ_o, i.e., $\{\theta : |\theta-\theta_o|_2 \leq c/\sqrt{T}\}$. C_o is defined to be

$$
C_o \overset{\text{def}}{=} \lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)).
$$

We denote $\|X\|_2$ as $(\mathrm{E}(X^2))^{1/2}$.

B.1 $\check{\theta}$ and $\check{\kappa}$ are pre-estimators lies in the interior of Θ and Γ, and are consistent.

B.2 $\hat{\theta}$ is $\sqrt{T}$-consistent.

B.3 Each element of $\mathbf{r}_t(\theta)$ is measurable and twice continuously differentiable. Each element of $\mathbf{D}_t(\theta, \kappa)$ is first continuously differentiable.

B.4 $\| \sup_{\theta \in \Theta} |\mathbf{r}_t(\theta)|_2^2 \|_2 \leq M_1$, $\| \sup_{\theta \in \Theta} |\mathbf{R}_t(\theta)|_2^2 \|_2 \leq M_2$, and $\mathrm{E}(tr(\mathbf{R}_t' \mathbf{D}_t^{-1} \mathbf{R}_t)) \leq M_3$, where M_1, M_2, and M_3 are constants.

B.5 C_o and B_o are positive definite.

B.6 $\sup_{\theta \in V_{\theta_o}(c)} |B_T(\theta) - B_o(\theta)|_2 = \mathcal{O}_p(1)$.

B.7 The map $\theta \mapsto B_o(\theta)$ is element-wise first order differentiable in $V_{\theta_o}(c)$.

Theorem 4. *Under conditions B.1- B.7, the estimator $\hat{\theta}$ can be linearized through the following form,*

$$\sqrt{T}(\hat{\theta} - \theta_o) = -\sqrt{T} B_o^{-1} T^{-1} \sum_{t=1}^{T} \mathbf{R}_t(\theta_o)' \mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1} \mathbf{r}_t(\theta_o) + \mathcal{O}_p(1). \tag{29}$$

Furthermore, we have the asymptotic normality of our estimator $\hat{\theta}$,

$$\sqrt{T}(\hat{\theta} - \theta_o) \to_d \mathbb{N}(0, B_o^{-1} C_o B_o^{-1}). \tag{30}$$

Moreover, if $E(\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)|\mathcal{F}_{t-1}) = \mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)$, then we have the minimum variance of the form B_o^{-1}.

Remark. *The consistency in B.1 is needed for the minimum variance, but not necessary for asymptotic normality. If $\check{\theta}$ and $\check{\kappa}$ are converging to θ^* and κ^*, we have*

$$\sqrt{T}(\hat{\theta} - \theta_o) = -\sqrt{T} B_o^{*-1} T^{-1} \sum_{t=1}^{T} \mathbf{R}_t(\theta^*)' \mathbf{D}_t(\mathbf{x}_t, \theta^*, \kappa^*)^{-1} \mathbf{r}_t(\theta_o) + \mathcal{O}_p(1), \tag{31}$$

where $B_o^* \stackrel{\text{def}}{=} lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} E(\mathbf{R}_t(\theta^*)' \mathbf{D}_t(\mathbf{x}_t, \theta^*, \kappa^*)^{-1} \mathbf{R}_t(\theta_o))$.

Proof. In our procedure, (9) becomes

$$T^{-1} \sum_{t}^{T} \check{\mathbf{R}}_t' \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})^{-1} \mathbf{r}_t(\theta) = 0,$$

and we are interested in $\hat{\theta}$, the solution of θ in this score function.

A Taylor expansion would lead to

$$\begin{aligned} & T^{-1}\sum_{t=1}^{T}\check{\mathbf{R}}_t'\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}\mathbf{r}_t(\hat{\theta}) \\ = \quad & T^{-1}\sum_{t=1}^{T}\check{\mathbf{R}}_t'\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}(\mathbf{r}_t(\theta_o)+\nabla\mathbf{r}_t(\tilde{\theta})(\hat{\theta}-\theta_o)), \end{aligned}$$

where $\tilde{\theta}$ is a point lies in the line segment between $\hat{\theta}$ and θ_o.

It is known that under proper assumptions and by the uniform weak law of large numbers, we would have the closeness of the partial sums $\check{Q}_T(\hat{\theta})$ and $\check{B}_T(\tilde{\theta})$ to $\check{A}_o(\theta_o)$ and $\check{B}_o(\tilde{\theta})$ respectively. In B.5 we assume that B_o is positive definite, so the minimum eigenvalue $\lambda_{min}(B_o) > c$, for a positive constant c.

We would attain the following expansion

$$\sqrt{T}(\hat{\theta}-\theta_o) = -\sqrt{T}B_o^{-1}T^{-1}\sum_{t=1}^{T}\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o) - \sqrt{T}B_o^{-1}(\check{A}_o(\theta_o)-A(\theta_o)) + I_T \tag{32}$$

I_T is a residual term that is needed to be proved to be of small order. Let us rewrite it into

$$\begin{aligned} I_T \quad = \quad & -\sqrt{T}B_o^{-1}(\check{B}_T(\tilde{\theta})-B_o)(\hat{\theta}-\theta_o) \\ - \quad & \{\sqrt{T}B_o^{-1}T^{-1}\sum_{t=1}^{T}(\check{\mathbf{R}}_t'\mathbf{D}_t(\mathbf{x}_t,\check{\theta},\check{\kappa})^{-1}-\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1})\mathbf{r}_t(\theta_o) - \sqrt{T}B_o^{-1}(\check{A}(\theta_o)-A(\theta_o))\} \\ =: \quad & I_{T1}+I_{T2}. \end{aligned}$$

The drift term $\sqrt{T}B_o^{-1}(\check{A}_o(\theta_o)-A(\theta_o))$ characterizes the dependency of our pre-estimated nuisance parameter $\check{\theta},\check{\kappa}$. $\sqrt{T}B_o^{-1}(\check{A}_o(\theta_o)-A(\theta_o)) = \mathbf{0}$ by construction since the conditional moment property of $\mathrm{E}[\mathbf{r}_t(\theta_o)|\mathbf{x}_t] = 0$.

We need to prove that $I_T = \mathcal{O}_p(1)$. Recall that $V_{\theta_o}(c)$ as a $1/\sqrt{T}$-ball around θ_o, i.e., $\{\theta : |\theta-\theta_o|_2 \le c\frac{1}{\sqrt{T}}\}$. Define

$$\begin{aligned} B_T(\tilde{\theta}) &= T^{-1}\textstyle\sum_{t=1}^{T}\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}(\tilde{\theta}), \\ B_T(\theta) &= T^{-1}\textstyle\sum_{t=1}^{T}\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}(\theta), \\ \text{and } B_o(\theta) &= \lim_{T\to\infty}\mathrm{E}B_T(\theta). \end{aligned}$$

Looking at I_{T1} first. Note

$$|I_{T1}|_2 \le \sqrt{T}|B_o^{-1}|_2(|\check{B}_T(\tilde{\theta})-B_T(\tilde{\theta})|_2 + |B_T(\tilde{\theta})-B_o(\tilde{\theta})|_2 + |B_o(\tilde{\theta})-B_o|_2)|\hat{\theta}-\theta_o|_2.$$

If we have $|\check{B}_T(\tilde{\theta}) - B_T(\tilde{\theta})|_2 \leq \sup_{\theta \in V_{\theta_o}(c)} |B_T(\theta) - \check{B}_T(\theta)|_2$, we would have the right-hand side $= \mathcal{O}_p(1)$. To check this point, note

$$\begin{aligned}
& \sup_{\theta \in V_{\theta_o}(c)} |B_T(\theta) - \check{B}_T(\theta)|_2 \\
\leq & \sup_{\theta \in V_{\theta_o}(c)} T|T^{-1} \sum_{t=1}^{T} (\check{\mathbf{R}}_t - \mathbf{R}_t(\theta_o)) \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})^{-1} \\
- & \mathbf{R}_t(\theta) \mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1} (\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o) - \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})) \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})^{-1}|_2 \cdot |T^{-1} \sum_{t=1}^{T} \mathbf{R}_t(\theta)|_2
\end{aligned}$$

by Cauchy Schwartz inequality. As $|T^{-1} \sum_{t=1}^{T} \mathbf{R}_t(\theta)|_2$ consists of a partial sum of martingale differences, we can apply Burkholder inequality and show that $|T^{-1} \sum_{t=1}^{T} \mathbf{R}_t(\theta)|_2 = \mathcal{O}_p(\frac{1}{\sqrt{T}})$ by assuming that the second moment of $\mathrm{E}(\sup_{\theta \in V_{\theta_o}(c)} |\mathbf{R}_t(\theta)|_2^2)$ is bounded (B.4). And we can show

$$\begin{aligned}
& |T^{-1} \sum_{t=1}^{T} (\check{\mathbf{R}}_t - \mathbf{R}_t(\theta_o)) \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})^{-1} \\
- & \mathbf{R}_t(\theta) \mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1} (\mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o) - \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})) \mathbf{D}_t(\mathbf{x}_t, \check{\theta}, \check{\kappa})^{-1}|_2 \\
= & \mathcal{o}(\frac{1}{\sqrt{T}}),
\end{aligned}$$

due to the differentiability and consistency of $\check{\theta}$ and $\check{\kappa}$ (B.1 and B.3). Also by the uniform law of large number $|B_T(\tilde{\theta}) - B_o(\tilde{\theta})|_2 \leq sup_{\theta \in V_{\theta_o}(c)} |B_T(\theta) - B_o(\theta)|_2$. Finally, we need $|B_o - B_o(\tilde{\theta})|_2 = \mathcal{o}(1)$, this is implied by the differentiability of the map $\theta \longmapsto \lim_{T \to \infty} \mathrm{E} B_T(\theta)$ at the neighborhood around θ_o (B.7). $I_{T2} = \mathcal{o}(1)$ is implied by the uniform law of large numbers (B.6).

Finally, we apply a central limit theorem to the leading term

$$-\sqrt{T} B_o^{-1} T^{-1} \sum_{t=1}^{T} \mathbf{R}_t(\theta_o)' \mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1} \mathbf{r}_t(\theta_o) - \sqrt{T} B_o^{-1} \check{A}_o(\theta_o)$$

as in El Machkouri et al. [2013] as follows.

Lemma 5 (Theorem 1 of El Machkouri et al. [2013])**.** *Denote $Y_t = f(\mathcal{F}_t)$, where f is some measurable function. Let $S_n = \sum_{t=1}^{n} Y_t$, and $\delta_{\varsigma,t} = \|Y_t - Y_t^*\|_\varsigma$. ($Y_t^*$ is Y_t replaced by a i.i.d. copy of the underlying innovations at time point zero.) If $E(Y_i) = 0$, $\sum_{t=0}^{\infty} \delta_{\varsigma,t} < \infty$, some $\varsigma \geq 2$, and $\sigma_n^2 \stackrel{\text{def}}{=} E(S_n^2) \to \infty$, then*

$$\sigma_n^{-1} S_n \to_L \mathbb{N}(0,1).$$

$-\sqrt{T} B_o^{-1} T^{-1} \sum_{t=1}^{T} \nabla \mathbf{r}_t(\theta_o) \mathbf{D}_t(\mathbf{x}_t, \theta_o, \kappa_o)^{-1} \mathbf{r}_t(\theta_o)$ would give us the asymptotic normality of the estima-

tion.

We are applying Lemma 5 to the object

$$-\sqrt{T}B_o^{-1}T^{-1}\sum_{t=1}^{T}\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o) \stackrel{\text{def}}{=} -\sqrt{T}T^{-1}\sum_{t=1}^{T}\tilde{d}_t.$$

Let c to be any vector in $\mathbf{R}^P$, and take ς to be 2, then we can verify that $c'\tilde{d}_t$ is MDS, and we then have $\sum_{t=0}^{\infty}\delta_{\varsigma,t} \leq C\|c'\tilde{d}_t\|_2 < M_3$, by imposing [B.4]. In addition, we have $\lim_{T\to\infty}\mathrm{E}(S_T^2) = T^{-1}\sum_{t=1}^{T}\mathrm{E}c'B_o^{-1}C_oB_o^{-1}c$.

We have then following the above lemma 5, the Cramár Wold device, and the expansion as in (32).

$$\sqrt{T}(\hat{\theta}-\theta_o) \to_L \mathbb{N}(0, B_o^{-1}C_oB_o^{-1}), \tag{33}$$

where C_o is defined to be

$$\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)).$$

It is not hard to see that if we correctly specify the variance-covariance matrix of the conditional moment, namely $\mathrm{E}(\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)|\mathcal{F}_{t-1}) = \mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)$, then

$$\begin{aligned}
& \mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)) \\
= \;& \mathrm{E}(\mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)|\mathcal{F}_{t-1})) \\
= \;& \mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathrm{E}(\mathbf{r}_t(\theta_o)\mathbf{r}_t'(\theta_o)|\mathcal{F}_{t-1})\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)) \\
= \;& \mathrm{E}(\mathbf{R}_t(\theta_o)'\mathbf{D}_t(\mathbf{x}_t,\theta_o,\kappa_o)^{-1}\mathbf{R}_t(\theta_o)).
\end{aligned}$$

Thus $C_o = B_o$ and the asymptotic variance would be B_o^{-1}. □

7.3 One-step estimation

In this subsection, we show the asymptotic equivalence of the one-step estimator $\overline{\theta}$ and the ideal estimator $\hat{\theta}$, by updating any $\sqrt{T}$ -consistent estimator $\check{\theta}$.

B_o, B_T, and $\check{B}_T$ in the following assumptions are defined in (27) and (28).

C.1 For each random sequence $\theta_T = \theta_o + \mathcal{O}(\frac{1}{\sqrt{T}})$, $|B_T(\theta_T) - B_o(\theta_o)|_2 = \mathcal{o}_p(1)$. B_o is invertible.

C.2 $|\check{B}_T(\theta) - B_T(\theta_T)|_2 \to_p 0$.

C.3 Both $\hat{\theta}$ and $\check{\theta}$ are $\sqrt{T}$-consistent.

C.4 $B_T(\theta)$ converges in probability to $B_o(\theta)$ within $V_{\theta_o}(c)$.

Theorem 6. *Under assumption C.1- C.4, we have the asymptotic equivalence of the one-step estimator $\overline{\theta}$ and the ideal estimator $\hat{\theta}$, namely, $\overline{\theta} - \hat{\theta} = \mathcal{o}_p(\frac{1}{\sqrt{T}})$.*

Proof. Define $\check{Q}_T(\theta) = T^{-1}\sum_{t=1}^{T} \check{\mathbf{Z}}_t' \mathbf{r}_t(\theta)$, the score function leading to the ideal estimator $\hat{\theta}$. Recall that

$$\check{B}_T(\check{\theta}) = T^{-1}\partial \check{Q}_T(\theta)/\partial\theta'|_{\theta=\check{\theta}},$$
$$\overline{\theta} = \check{\theta} - (\check{B}_T(\check{\theta}))^{-1}\check{Q}_T(\check{\theta}).$$

Rearrange the last equation, we have

$$\check{B}_T(\check{\theta})(\overline{\theta} - \check{\theta}) = -\check{Q}_T(\check{\theta}) \tag{34}$$

By the property $\check{Q}_T(\hat{\theta}) = 0$ and Assumption C.3, we have

$$0 = \check{Q}_T(\hat{\theta}) = \check{Q}_T(\check{\theta}) + \check{B}_T(\check{\theta})(\hat{\theta} - \check{\theta}) + \mathcal{o}_p(\frac{1}{\sqrt{T}}) \tag{35}$$

Rearrange (35), we have

$$-\check{Q}_T(\check{\theta}) = \check{B}_T(\check{\theta})(\hat{\theta} - \check{\theta}) + \mathcal{o}_p(\frac{1}{\sqrt{T}}) \tag{36}$$

Combining (34) and (36) and rescaling them by $\sqrt{T}$, we have

$$\begin{aligned}
\mathcal{o}_p(1) &= \sqrt{T}\check{B}_T(\check{\theta})(\overline{\theta} - \hat{\theta}) \\
&= \sqrt{T}(B_o(\theta) + \mathcal{o}_p(1))(\overline{\theta} - \hat{\theta})) \\
&= \sqrt{T}B_o(\theta)(\overline{\theta} - \hat{\theta})) + \mathcal{o}_p(1)
\end{aligned}$$

where the second equality follows from C.2 and C.4. The last equality follows from C.3.

By the invertibility of the matrix B_o, we have

$$|\sqrt{T}(\overline{\theta} - \hat{\theta})|_2 \leq |B_o|_2^{-1}|\mathcal{o}_p(1)|_2 = \mathcal{o}_p(1), \tag{37}$$

This shows the asymptotic equivalence of $\hat{\theta}$ and $\tilde{\theta}$. □

7.4 Asymptotic consistent of the QMLE and the one-step estimation

The asymptotic normality and consistency are well understood in the literature.

M.1 $\theta_o \in \text{int}(\Theta)$.

M.2 $\ell_t(\mathbf{y}_t, \mathbf{x}_t, \theta)$ are measurable and second-order differentiable.

M.3 I_o is positive definite.

M.4 $\sup_{\theta\in\Theta} |T^{-1}\sum_{t=1}^{T} I_t(\theta) - I_o(\theta)|_2 \to 0$.

M.5 $\frac{1}{\sqrt{T}}\mathbf{s}_t(\theta_o) \to_L \mathbb{N}(0, J_o)$.

Theorem 7 (Wooldridge [1994])**.** *Under assumption M.1 to M.5, the QMLE is asymptotically normal,* $\sqrt{T}(\check{\theta} - \theta_o) \to \mathbb{N}(0, I_o^{-1}J_oI_o^{-1})$.

Theorem 8. *Under assumption M.1 to M.5, the one-step estimation of QMLE is asymptotically normal, and* $\overline{\theta} - \check{\theta} = \mathcal{O}(\frac{1}{\sqrt{T}})$, $\sqrt{T}(\overline{\theta} - \theta_o) \to \mathbb{N}(0, I_o^{-1}J_oI_o^{-1})$.

Proof. The proof is similar to the proof in Theorem 6 and therefore omitted. □

References

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3): 307–327, 1986.

Tim Bollerslev and Jeffrey M Wooldridge. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric reviews*, 11(2):143–172, 1992.

Whitney K Newey and Douglas G Steigerwald. Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. *Econometrica: Journal of the Econometric Society*, pages 587–599, 1997.

Tim Bollerslev. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The review of economics and statistics*, pages 542–547, 1987.

Jianqing Fan, Lei Qi, and Dacheng Xiu. Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics*, 32(2):178–191, 2014.

Christian M Hafner and Jeroen VK Rombouts. Semiparametric multivariate volatility models. *Econometric Theory*, 23(2):251–280, 2007.

Xiaohong Chen and Yanqin Fan. Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335, 2006.

Feike C Drost, Chris AJ Klaassen, Bas JM Werker, et al. Adaptive estimation in time-series models. *The Annals of Statistics*, 25(2):786–817, 1997.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Robert F Engle, David M Lilien, and Russell P Robins. Estimating time varying risk premia in the term structure: The arch-m model. *Econometrica: journal of the Econometric Society*, pages 391–407, 1987.

Jeffrey M Wooldridge. Estimation and inference for dependent processes. *Handbook of econometrics*, 4: 2639–2738, 1994.

Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

Jan R Magnus and Heinz Neudecker. The commutation matrix: some properties and applications. *The Annals of Statistics*, pages 381–394, 1979.

Christian Francq and Jean-Michel Zakoïan. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, 10(4):605–637, 2004. ISSN 1350-7265. doi: 10.3150/bj/1093265632. URL `https://doi.org/10.3150/bj/1093265632`.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Mohamed El Machkouri, Dalibor Volnỳ, and Wei Biao Wu. A central limit theorem for stationary random fields. *Stochastic Processes and their Applications*, 123(1):1–14, 2013.

IRTG 1792 Discussion Paper Series 2020

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Estimation and Determinants of Chinese Banks' Total Factor Efficiency: A New Vision Based on Unbalanced Development of Chinese Banks and Their Overall Risk" by Shiyi Chen, Wolfgang K. Härdle, Li Wang, January 2020.

002 "Service Data Analytics and Business Intelligence" by Desheng Dang Wu, Wolfgang Karl Härdle, January 2020.

003 "Structured climate financing: valuation of CDOs on inhomogeneous asset pools" by Natalie Packham, February 2020.

004 "Factorisable Multitask Quantile Regression" by Shih-Kang Chao, Wolfgang K. Härdle, Ming Yuan, February 2020.

005 "Targeting Cutsomers Under Response-Dependent Costs" by Johannes Haupt, Stefan Lessmann, March 2020.

006 "Forex exchange rate forecasting using deep recurrent neural networks" by Alexander Jakob Dautel, Wolfgang Karl Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2020.

007 "Deep Learning application for fraud detection in financial statements" by Patricia Craja, Alisa Kim, Stefan Lessmann, May 2020.

008 "Simultaneous Inference of the Partially Linear Model with a Multivariate Unknown Function" by Kun Ho Kim, Shih-Kang Chao, Wolfgang K. Härdle, May 2020.

009 "CRIX an Index for cryptocurrencies" by Simon Trimborn, Wolfgang Karl Härdle, May 2020.

010 "Kernel Estimation: the Equivalent Spline Smoothing Method" by Wolfgang K. Härdle, Michael Nussbaum, May 2020.

011 "The Effect of Control Measures on COVID-19 Transmission and Work Resumption: International Evidence" by Lina Meng, Yinggang Zhou, Ruige Zhang, Zhen Ye, Senmao Xia, Giovanni Cerulli, Carter Casady, Wolfgang K. Härdle, May 2020.

012 "On Cointegration and Cryptocurrency Dynamics" by Georg Keilbar, Yanfen Zhang, May 2020.

013 "A Machine Learning Based Regulatory Risk Index for Cryptocurrencies" by Xinwen Ni, Wolfgang Karl Härdle, Taojun Xie, August 2020.

014 "Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects" by Daniel Jacob, August 2020.

015 "Tail-risk protection: Machine Learning meets modern Econometrics" by Bruno Spilak, Wolfgang Karl Härdle, October 2020.

016 "A data-driven P-spline smoother and the P-Spline-GARCH models" by Yuanhua Feng, Wolfgang Karl Härdle, October 2020.

017 "Using generalized estimating equations to estimate nonlinear models with spatial data" by Cuicui Lu, Weining Wang, Jeffrey M. Wooldridge, October 2020.

018 "A supreme test for periodic explosive GARCH" by Stefan Richter, Weining Wang, Wei Biao Wu, October 2020.

019 "Inference of breakpoints in high-dimensional time series" by Likai Chen, Weining Wang, Wei Biao Wu, October 2020.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2020

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

020 "Long- and Short-Run Components of Factor Betas: Implications for Stock Pricing" by Hossein Asgharian, Charlotte Christiansen, Ai Jun Hou, Weining Wang, October 2020.

021 "Improved Estimation of Dynamic Models of Conditional Means and Variances" by Weining Wang, Jeffrey M. Wooldridge, Mengshan Xu, October 2020.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.