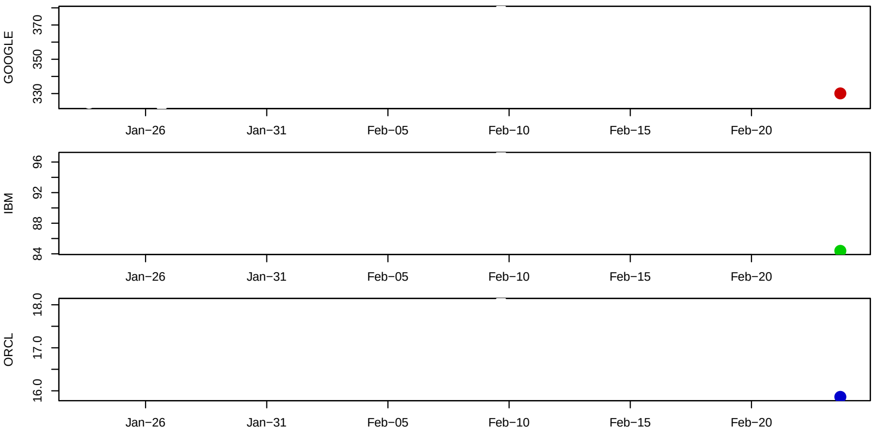
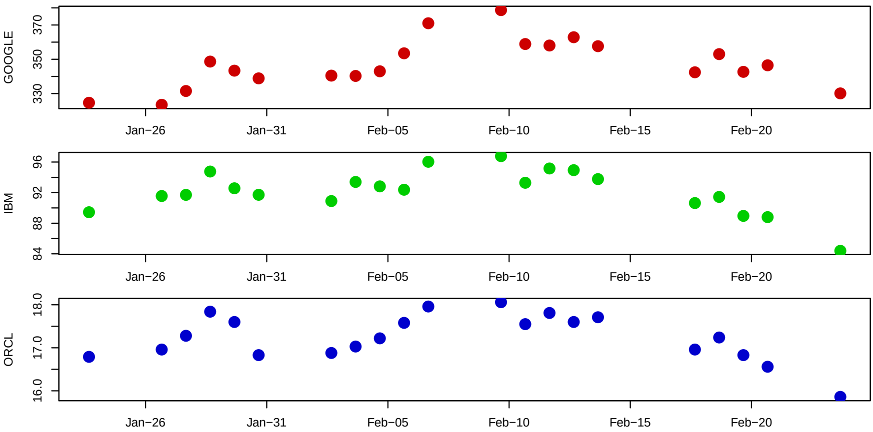
Realized Copula

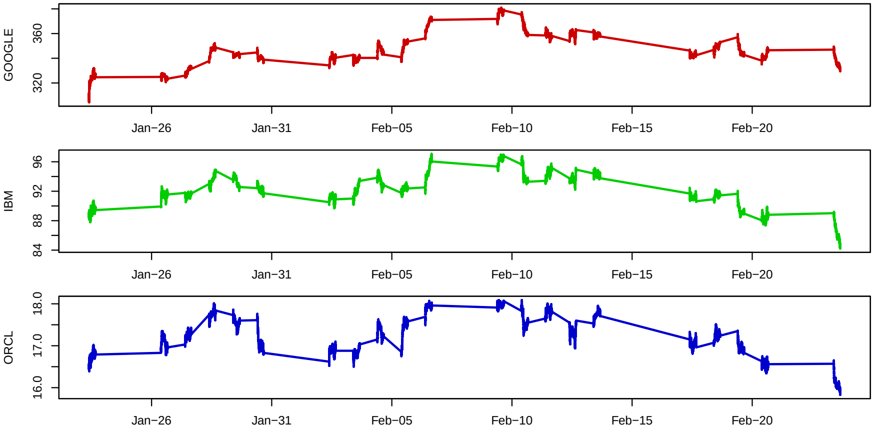
Matthias R. Fengler
Ostap Okhrin

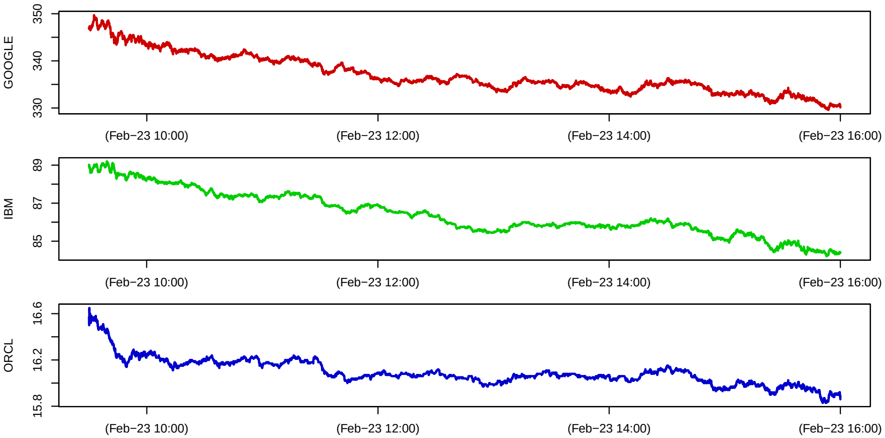
Ladislaus von Bortkiewicz
Chair of Statistics
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin

Chair of Financial Econometrics
Universität St. Gallen


GOOGLE
330
350
370
IBM
84
88
92
96
ORCL
16.0
17.0
18.0
Jan−26
Jan−31
Feb−05
Feb−10
Feb−15
Feb−20


GOOGLE
330
350
370
IBM
84
88
92
96
ORCL
16.0
17.0
18.0
Jan−26
Jan−31
Feb−05
Feb−10
Feb−15
Feb−20


GOOGLE
320
360
IBM
84
88
92
96
ORCL
16.0
17.0
18.0
Jan−26
Jan−31
Feb−05
Feb−10
Feb−15
Feb−20



GOOGLE
350
340
330
(Feb−23 10:00)
(Feb−23 12:00)
(Feb−23 14:00)
(Feb−23 16:00)
IBM
89
87
85
(Feb−23 10:00)
(Feb−23 12:00)
(Feb−23 14:00)
(Feb−23 16:00)
ORCL
16.6
16.2
15.8
(Feb−23 10:00)
(Feb−23 12:00)
(Feb−23 14:00)
(Feb−23 16:00)

Realized Variance of Google-IBM-Oracle

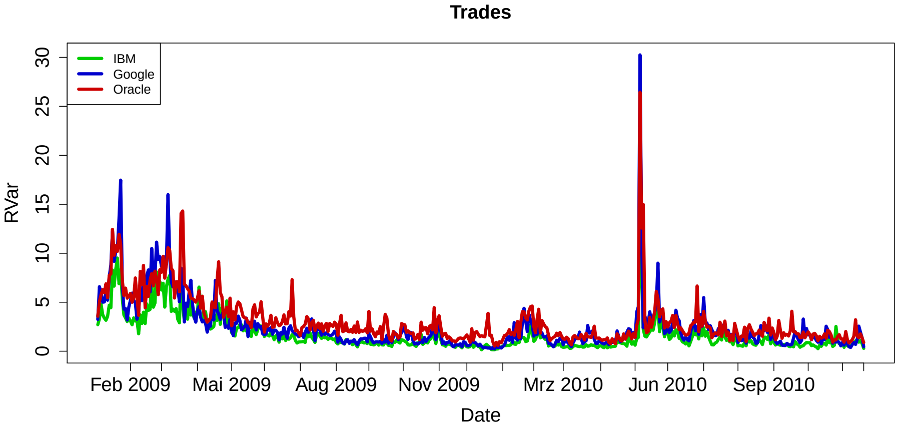


Figure 1: Realized kernel (variance) of Google-IBM-Oracle.

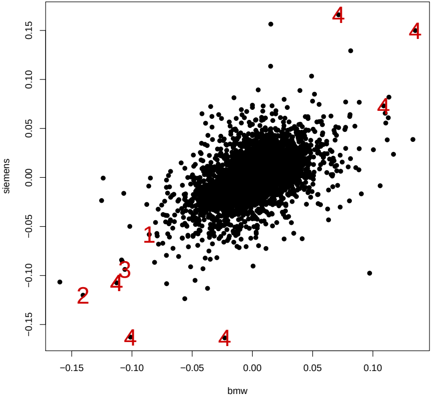
RV: Exploiting intra-day information

Literature of the past 10yrs on high-frequency data shows:

- daily realized (co)variance (RV, RCov) computed from intra-day data serves as an accurate measures of conditional (co)variance of daily returns;
- no specific model is needed (like GARCH);
- can treat an inherently latent variable like an observed one;
- shows excellent forecasting performance.

Heavily discussed in *derivatives pricing, portfolio optimization, risk-management, and volatility forecasting.*

Dependency



1. 19.10.1987
   Black Monday
2. 16.10.1989
   Berlin Wall
3. 19.08.1991
   Kremlin
4. 17.03.2008, 19.09.2008, 10.10.2008, .10.2008, 15.10.2008, 29.10.2008
   Crisis

Copulae

Copulae is a convenient tool to capture nonlinear dependence.

Copulae

Copulae is a convenient tool to capture nonlinear dependence.

Multivariate RCov models have an underlying *Gaussian* structure.

Copulae

Copulae is a convenient tool to capture nonlinear dependence.

Multivariate RCov models have an underlying *Gaussian* structure.

How can we suitably combine intra-day RCov information into a *non-Gaussian* model framework?

Copulae

Copulae is a convenient tool to capture nonlinear dependence.

Multivariate RCov models have an underlying *Gaussian* structure.

How can we suitably combine intra-day RCov information into a *non-Gaussian* model framework?

realized copula (RCop)

Outline

1. Motivation ✓
2. Copula and realized copula
3. Benchmark models
4. Empirical Part
5. References

Copulae

A **copula** is a multivariate distribution with all univariate margins being $U(0,1)$.

Theorem (Sklar, 1959)

Let $X_1,\ldots,X_d$ *be random variables with marginal distribution functions* $F_1,\ldots,F_d$ *and joint distribution function* F. *Then there exists a* d*-dimensional copula* $C:[0,1]^d \to [0,1]$ *such that* $\forall x_1,\ldots,x_d \in \overline{\mathbb{R}} = [-\infty,\infty]$

$$F(x_1,\ldots,x_d) = C\{F_1(x_1),\ldots,F_d(x_d)\}$$

Archimedean copula $C : [0,1]^d \to [0,1]$ defined as

$$C(u_1, \dots, u_d) = \phi\{\phi^{-1}(u_1) + \cdots + \phi^{-1}(u_d)\}, \qquad (1)$$

where $\phi : [0,\infty) \to [0,1]$ is strictly decreasing with $\phi(0) = 1$, $\phi(\infty) = 0$ and ϕ^{-1} its (pseudo)inverse.

Example

$$\begin{aligned} \phi_{Gumbel}(u,\theta) &= \exp\{-u^{1/\theta}\}, \text{ where } 1 \le \theta < \infty \\ \phi_{Clayton}(u,\theta) &= (\theta u + 1)^{-1/\theta}, \text{ where } \theta \in [-1,\infty)\backslash\{0\} \end{aligned}$$

Rotated copula as an example of a non-Archimedean copula:

$$C_{rot}(u_1, u_2) = C(1-u_1, 1-u_2) + u_1 + u_2 - 1,$$

which in term of copula density is given through
$c_{rot}(u_1, \dots, u_d) = c(1-u_1, \dots, 1-u_d)$

Realized Copula, I

Lemma (Hoeffding)

Suppose there are two random variables X_i and X_j with marginal distributions F_i and F_j and joint distribution F_{ij} and finite second moments

$$
\begin{aligned}
\sigma_{ij}(\theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big\{F_{i,j}(x_i, x_j, \theta) - F_i(x_i)F_j(x_j)\Big\} dx_i dx_j \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Big[C_\theta\{F_i(x_i), F_j(x_j)\} - F_i(x_i)F_j(x_j)\Big] dx_i dx_j \ .
\end{aligned}
$$

Realized Copula, II

For the notion of *realized copula*, we define θ implicitly through

$$\begin{aligned} h_{ij,t} &= \mathrm{f}_{ij}(\theta_t) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[C_{\theta_t}\{F_{i,t}(x_i), F_{j,t}(x_j)\} - F_{i,t}(x_i)F_{j,t}(x_j)\right] dx_i dx_j \end{aligned}$$

where $h_{ij,t}$ denotes an element of the RCov matrix measured at day t.

This moment condition, together with the assumptions on the copula and the marginal distributions, identifies the ex-post daily distribution as materialized in RCov.

Method-of-moments estimator, I

Let $d = 2$, with one off-diagonal element $h_{12,t}$ in the RCov. An estimate of θ_t is given by

$$\widehat{\theta}_t^{\mathsf{MM}} = \mathrm{f}_{12}^{-1}(h_{12,t}).$$

Similar to method-of-moments approaches where the copula parameter of an Archimedean copula is estimated from Kendall's tau (Genest and Rivest, 1993).

Method-of-moments estimator, II

For $d > 2$, define

$$\mathrm{g}_{ij}(\theta) = h_{ij,t} - \mathrm{f}_{ij}(\theta),$$

where $i < j$ and $i, j = 1, \ldots, d$.

Stacking all g_{ij} into a vector $\mathbf{g}$ of size $d(d-1)/2$, we define the estimator as

$$\widehat{\theta}_t^{\mathrm{MM}} = \arg\min_{\theta} \ \mathbf{g}^\top(\theta)\boldsymbol{\Omega}\mathbf{g}(\theta) \ ,$$

with $\boldsymbol{\Omega}$ denoting a $d(d-1)/2$-dimensional pd weight matrix. A conventional choice would be the unit matrix $\mathbf{I}_{d(d-1)/2}$.

Ad hoc estimator

Under Gaussianity, Kendall's τ is $\tau^{G}_{ij,t} = \frac{2}{\pi}\arcsin \rho_{ij,t}$, and generally, for general Archimedean copulae (Genest and Rivest, 1993):

$$\tau \equiv \mathrm{f}_\tau(\theta) = 4\int_0^1 {\phi_\theta}^{-1}(v)/({\phi_\theta}^{-1})'(v)\,dv + 1\ .$$

family	ϕ_θ	f_τ
Gumbel	$\exp\{-x^{1/\theta}\}$	$1 - 1/\theta$
Clayton	$(\theta x + 1)^{-1/\theta}$	$\theta/(2+\theta)$

We define an ad-hoc estimator by

$$\widehat{\theta}^{\text{ad hoc}}_t = \frac{2}{d(d-1)}\sum_{i<j} \mathrm{f}_\tau^{-1}(\widehat{\tau}^{G}_{ij,t})\ .$$

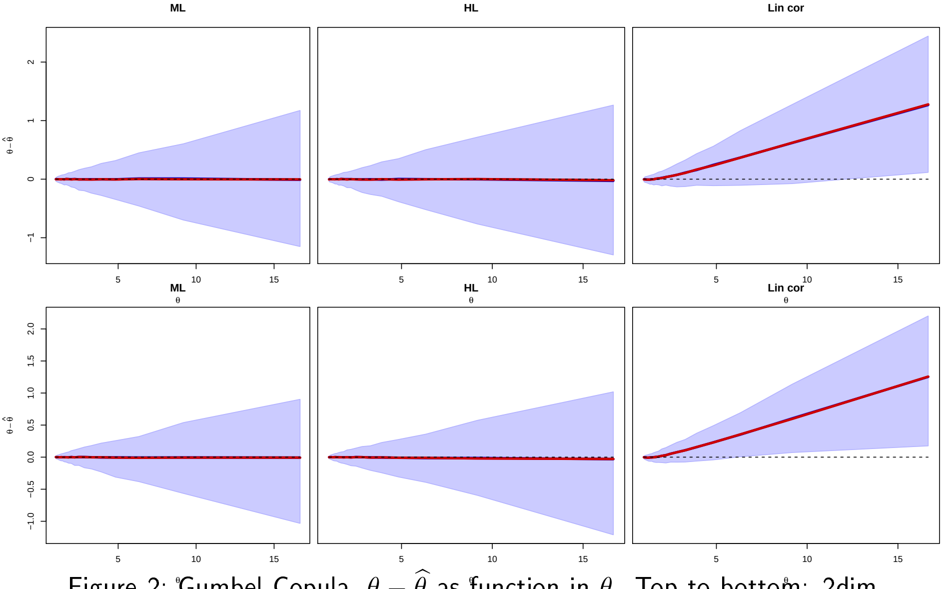


Figure 2: Gumbel Copula, $\theta - \widehat{\theta}$ as function in θ. Top to bottom: 2dim, 3dim. $n = 1000$, $N = 1000$. Shaded area is the simulation based 95% interval.

Forecasting framework for RCop

Let $P_t = (P_{1t}, \ldots, P_{dt})^\top$ and $r_t = P_t - P_{t-1},\ t = 1, \ldots, T$ be daily log-prices and their log-returns with

$$r_{t+1} \sim F_{r_{t+1}|\mathcal{F}_t}(\widehat{H}_{t+1|t})$$

where $\widehat{H}_{t+1|t}$ is an $\mathcal{F}_t$-measurable forecast of the RC matrix of r_t and

$$F_{r_{t+1}|\mathcal{F}_t}(\widehat{H}_{t+1|t}) = C_{\widehat{\theta}_{t+1|t}}\{F_{1,t}(\widehat{h}_{1,t+1|t}), \ldots, F_{d,t}(\widehat{h}_{d,t+1|t})\}$$

As reported in Andersen et al. (2001) returns standardized by ex post RV are close to standard normal, we thus assume that

$$F_{j,t}(\widehat{h}_{j,t+1|t}) = \mathbb{N}(0, \widehat{h}_{j,t+1|t})$$

Forecasting framework

Consider the following multivariate forecasting rule:

$$\begin{pmatrix} \log \widehat{h}_{1,t+1|t} \\ \vdots \\ \log \widehat{h}_{d,t+1|t} \\ \widehat{\theta}_{t+1|t} \end{pmatrix} = \boldsymbol{E}_t \begin{pmatrix} \log h_{1,t+1} \\ \vdots \\ \log h_{d,t+1} \\ \theta_{t+1} \end{pmatrix} = \begin{pmatrix} \beta_0^1 + \beta_{\mathrm{D}}^1 \log h_t^{\mathrm{D}} + \beta_{\mathrm{W}}^1 \log h_t^{\mathrm{W}} + \beta_{\mathrm{M}}^1 \log h_t^{\mathrm{M}} \\ \vdots \\ \beta_0^d + \beta_{\mathrm{D}}^d \log h_t^{\mathrm{D}} + \beta_{\mathrm{W}}^d \log h_t^{\mathrm{W}} + \beta_{\mathrm{M}}^d \log h_t^{\mathrm{M}} \\ \alpha_0 + \alpha_{\mathrm{D}} \theta_t^{\mathrm{D}} + \alpha_{\mathrm{W}} \theta_t^{\mathrm{W}} + \alpha_{\mathrm{M}} \theta_t^{\mathrm{M}} \end{pmatrix},$$

where $x_t^{\mathrm{D}} = x_t$ are daily, $x_t^{\mathrm{W}} = \frac{1}{5}\sum_{i=0}^{4} x_{t-i}$ weekly, and $x_t^{\mathrm{M}} = \frac{1}{21}\sum_{i=0}^{20} x_{t-i}$ monthly averages of past realizations of x_t.

Borrowed from the heterogenous autoregressive model (HAR) of Corsi (2009); extended here to the copula parameter.

Empirical application

Compare one day ahead VaR forecasting performance of RCop against a number of standard benchmark models:

- models based on daily data
  - naive rolling window
  - local adaptive estimation
- models based on intra-day data (RV models)
  - Logm-model
  - Cholesky factorization

Rolling window and adaptive estimation

Naive approach:

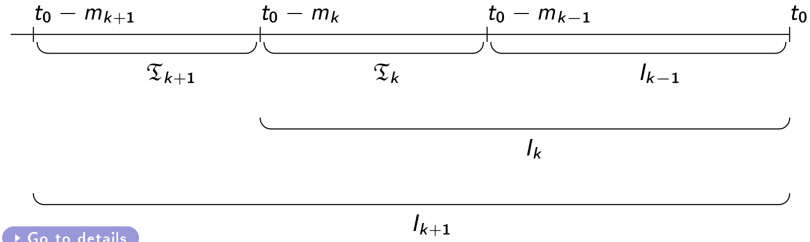
- estimate copula parameter on a fixed rolling window

LCP:

- adaptively estimate largest interval where homogeneity hypothesis is accepted
- *Local Change Point* detection (LCP): sequentially test whether θ_t is constant (i.e. $\theta_t = \theta$) within some interval I (local parametric assumption).

Local Change Point Detection

1. define the family of nested intervals
   $I_0 \subset I_1 \subset I_2 \subset \ldots \subset I_K = I_{K+1}$ with length m_k as
   $I_k = [t_0 - m_k, t_0]$
2. define $\mathfrak{T}_k = [t_0 - m_k, t_0 - m_{k-1}]$



Go to details

Data used in this study

- $d = 3$
- daily (Yahoo Finance) and tick trades (LOBSTER) prices for the two portfolios
  - IBM, Google, Oracle;
  - IBM, Pfizer, Exxon
- timespan = [02.01.2009 till 31.12.2010] ($n = 470$ days) for tick data and $n = 800$ days for daily data
- cleaning high-frequency data as in BNHLS (2008): *9:45-16:00, one stock exchange, multiple quotes or trades with same time stamp, negative spread, etc.*
- Rotated Gumbel and Clayton copulae.

Basis

Let $Y = (Y_1, \ldots, Y_d)^\top$ be a d-dim efficient *(log)price* process

$$dY_t = \mu_t dt + \sigma_t dW_t$$

The market microstructure effect is modeled through an additive component

$$\begin{aligned} P_{jt} &= Y_{jt} + U_{jt}, \text{ with } \boldsymbol{E}(U_{jt}) = 0 \\ \sum_h |h\Omega_h| &< \infty, \text{ where } \Omega_{jh} = Cov(U_{jt}, U_{j,t-h}). \end{aligned}$$

Usual aim: Estimate the *quadratic variation* of Y, i.e. $[Y] = \int_0^1 \Sigma_u du$, with $\Sigma = \sigma\sigma^\top$.

Naive Estimator (realized co/variance)

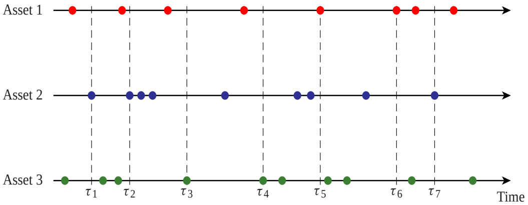
Synchronization – *last traded*: for time t, the log-price for asset j is given by P_{j,t^*} with $t^* = \max\{t_{j,i} | t_{j,i} \leq t,\ \forall i = 1, \dots, N_j\}$.
$M = M(m)$ number of subintervals of length m (in seconds)

$$
\begin{aligned}
\mathrm{RC}_{t_1,m,j_1,j_2}(P) &= \sum_{i=1}^{M} (P_{j_1,t_i} - P_{j_1,t_{i-1}})(P_{j_2,t_i} - P_{j_2,t_{i-1}}), \\
\mathrm{RC}_{t_1,m}(P) &= \{\mathrm{RC}_{m,j_1,j_2}\}_{j_1,j_2}, \text{ for } j_1, j_2 = 1, \dots, d
\end{aligned}
$$

Realized Kernels, BNHLS (2011, JoE)

Synchronization – *refresh time sampling*

$$\tau_1 = \max\{t_{1,1}, \ldots, t_{d,1}\}$$
$$\tau_{i+1} = \arg\min\{t_{j,k_j} | t_{j,k_j} > \tau_i, \ \forall j \in 1 \ldots d\}$$



Leads to new high-frequency vector of returns $p_i = P_{\tau_i} - P_{\tau_{i-1}}$, where $i = 1, \ldots, n$ and n is the of refresh time observations.

Realized Variance of Google-IBM-Oracle

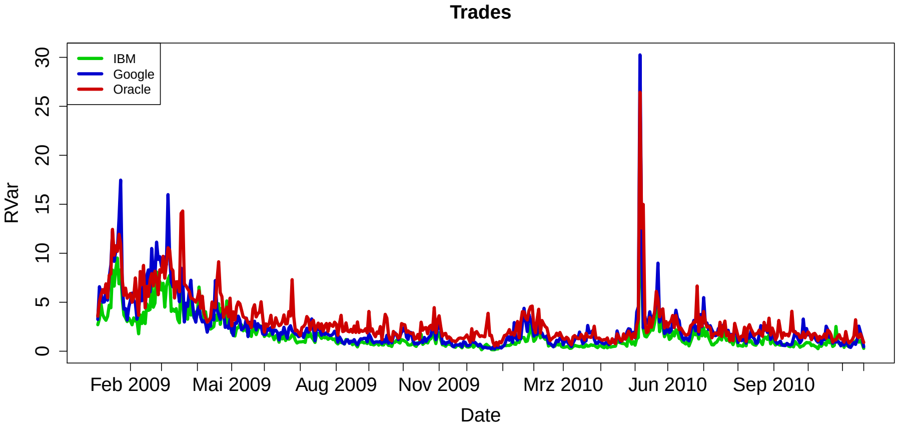


Figure 3: Realized kernel (variance) of Google-IBM-Oracle.

Realized Covariance of Google-IBM-Oracle

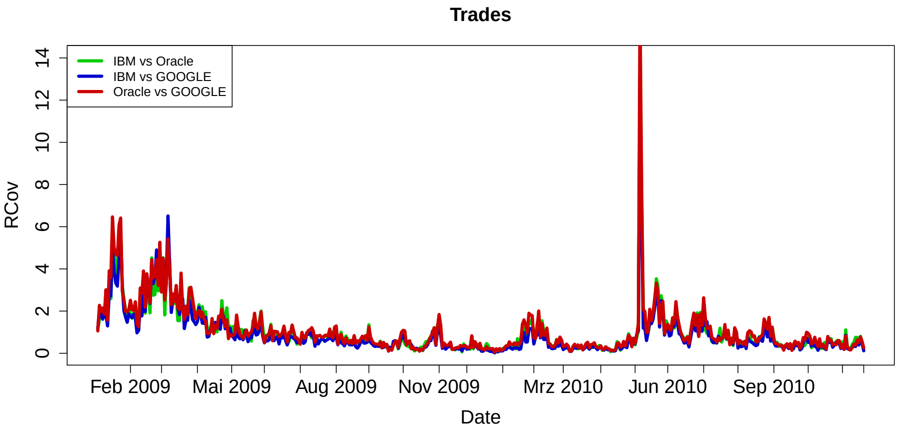


Figure 4: Realized Kernel (covariance) of Google-IBM-Oracle.

Realized Correlation of Google-IBM-Oracle

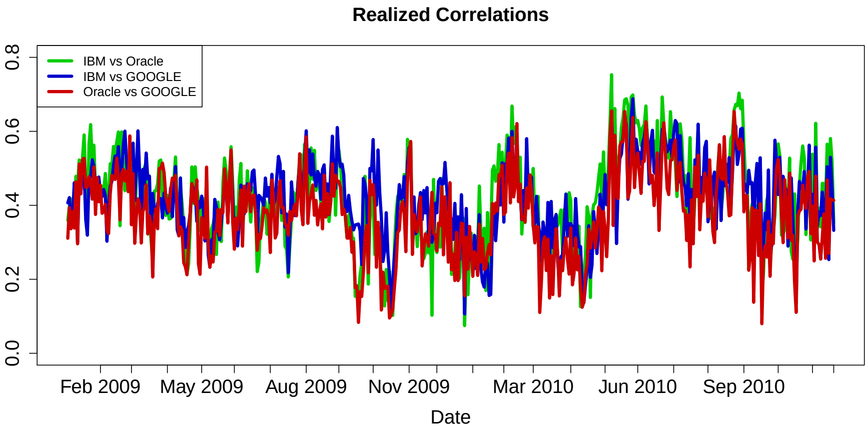


Figure 5: Realized Kernel (correlation) of Google-IBM-Oracle.

Realized Correlation of IBM-Pfizer-Exxon

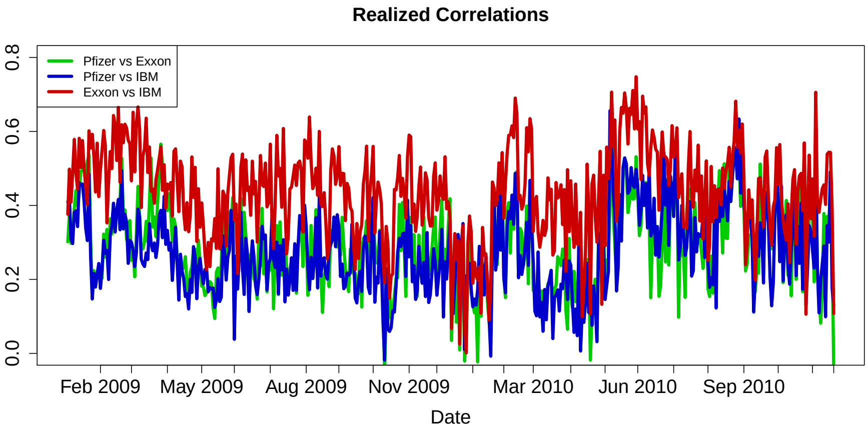


Figure 6: Realized Kernel (correlation) of IBM-Pfizer-Exxon.

Descriptive Statistics

	min.	median	mean	max.	std.
RV(Google)	2.277e-5	1.714e-4	2.503e-4	0.003	0.269e-3
RV(IBM)	*1.431e-5*	*1.048e-4*	*1.704e-4*	*0.001*	*0.180e-3*
RV(Oracle)	5.220e-5	2.208e-4	3.082e-4	0.002	0.253e-3
RC(Google,IBM)	1.978e-6	5.758e-5	9.112e-5	0.001	0.110e-3
RC(Google,Oracle)	5.359e-6	7.628e-5	1.112e-4	0.001	0.128e-3
RC(IBM,Oracle)	2.106e-6	6.749e-5	1.015e-4	0.001	0.113e-3
RV(IBM)	*1.474e-5*	*1.014e-4*	*1.704e-4*	*0.194e-4*	*1.820e-4*
RV(Pfizer)	2.819e-5	2.067e-4	2.837e-4	0.311e-4	2.467e-4
RV(Exxon)	2.455e-5	1.281e-4	1.810e-4	0.229e-4	1.786e-4
RC(IBM,Pfizer)	-1.550e-6	4.069e-5	6.553e-5	0.161e-4	9.599e-5
RC(IBM,Exxon)	4.231e-8	5.198e-5	8.442e-5	0.111e-4	1.010e-4
RC(Pfizer,Exxon)	-3.858e-6	4.691e-5	7.187e-5	0.112e-4	8.744e-5

Table 1: Descriptive statistics of the realized kernels (Var and Cov).

LCP for Google-IBM-Oracle

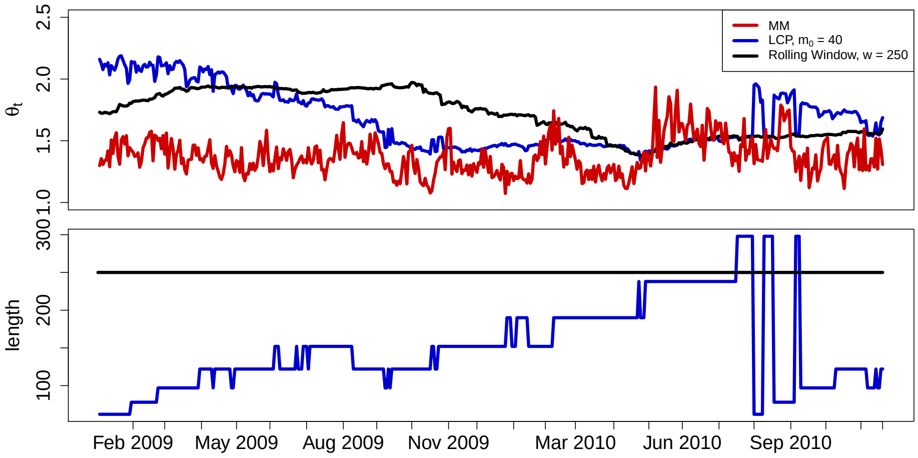


Figure 7: All copulae for Google-IBM-Oracle portfolio.

LCP for IBM-Pfizer-Exxon

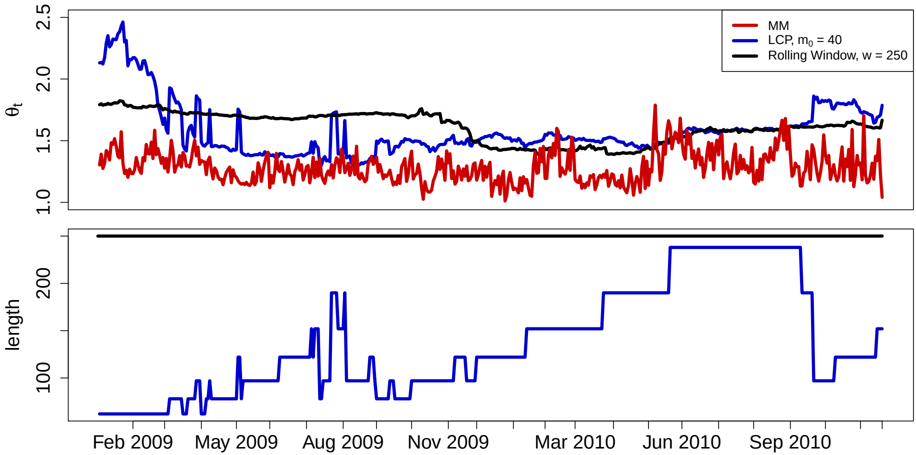


Figure 8: All copulae for IBM-Pfizer-Exxon portfolio.

Gaussian models

Recent suggestions in the multivariate RV literature: the matrix-log model (Bauer and Vorkink, 2010) and the Cholesky factorization (Chiriac and Voev, 2011).

For the logm-model, apply the logm to the RV matrix

$$A_t = \text{logm}(H_t)$$

and apply the vech-operator

$$a_t = \text{vech}(A_t)$$

which yields a $d(d+1)/2$ vector a_t.

To this vector the same HAR-forecasting rule is applied. Predictions $\widehat{a}_{t+1|t}$ are converted to positive-definite predicted covariance matrices by applying the reverse vech-operator and the matrix exponential:

$$\widehat{H}_{t+1|t} = \text{expm}(\widehat{A}_{t+1|t})\,.$$

Likewise, for the Cholesky decomposition, find a matrix A such that

$$H = AA^\top\,.$$

For predicitons, use a HAR model on the vector obtained from the vech-operation, and convert predicted Cholesky factors back:

$$\widehat{H}_{t+1|t} = \widehat{A}_{t+1|t}\widehat{A}_{t+1|t}^\top\,.$$

Overview on models

- daily models: LCP ($m_0 = 40$) and rolling window ($w = 250$)
- 2 methods of copula estimation (MM, ad hoc)
- 2 copula functions (rotated Gumbel, Clayton)
- 2 RV Gaussian Models (Chiriac and Voev (2011); Bauer and Vorkink (2010))

Value at Risk (VaR), I

Let $a = \{a_1, \ldots, a_d\},\ a_i \in \mathbb{Z}$ be the portfolio. The value V_t of a is given by

$$V_t = \sum_{j=1}^{d} a_j S_{j,t}$$

and the *profit and loss (P&L) function* of the portfolio

$$L_{t+1} = (V_{t+1} - V_t) = \sum_{j=1}^{d} a_j S_{j,t} \left\{\exp(X_{j,t+1}) - 1\right\},$$

where $w_j = a_{j,t} S_{j,t} / \sum_{i=1}^{d} (a_{i,t} S_{i,t})$ and $w_i = 1/d,\ 1, \ldots, d$.

VaR, II

The distribution function of L is given by

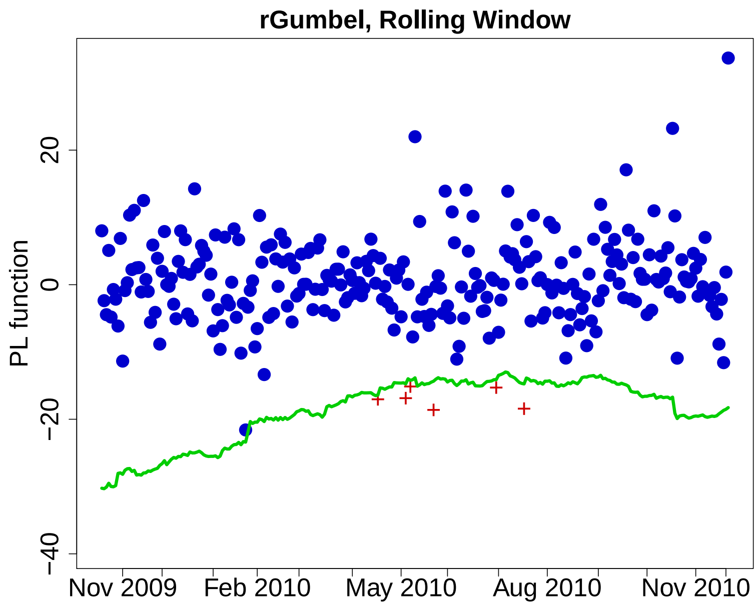
$$F_L(x) = P(L \le x).$$

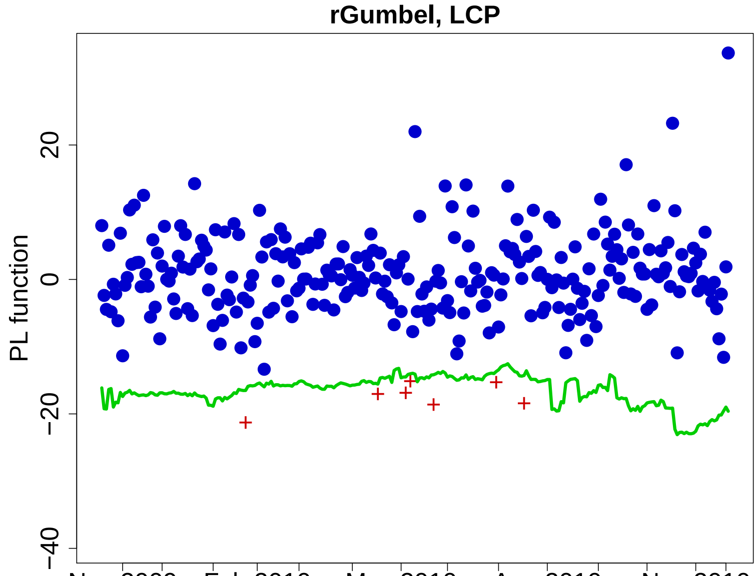
The *Value-at-Risk* at level α from w is defined as the α-quantile from F_L:

$$\mathrm{VaR}(\alpha) = F_L^{-1}(\alpha).$$

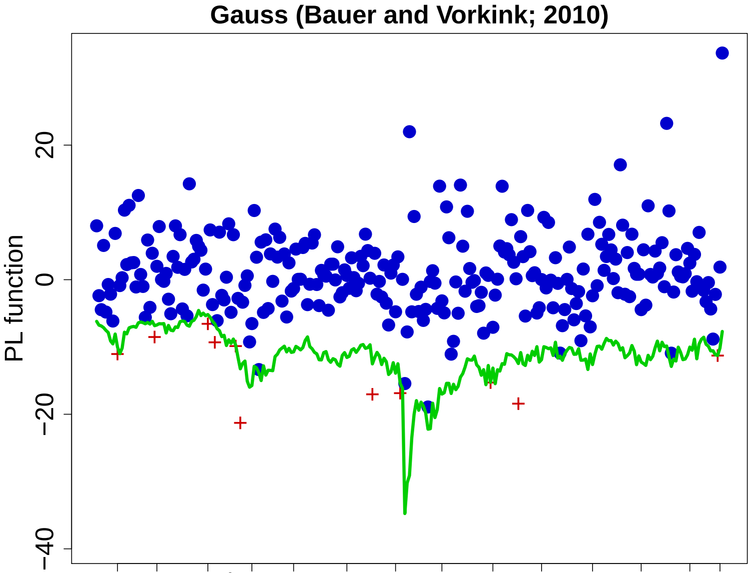
Backtesting: estimated values of the VaR are compared with the true $\{l_t\}$ of the function L_t, an *exceedance* occurring for each l_t smaller than $\widehat{VaR}_t(\alpha)$. The *exceedances ratio* $\widehat{\alpha}$ is given by:

$$\widehat{\alpha} = \frac{1}{T} \sum_{t=r}^{T} \mathbf{I}\{l_t < \widehat{VaR}_t(\alpha)\}.$$

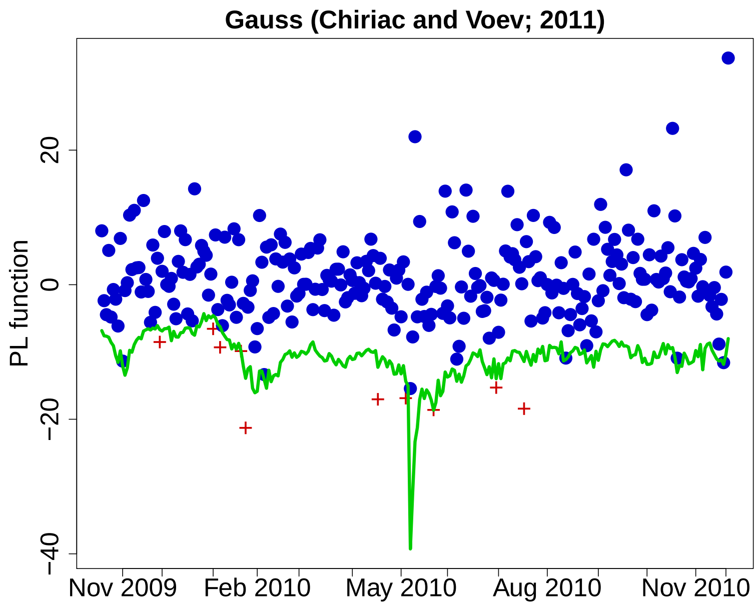

rGumbel, Rolling Window
PL function
20
0
−20
−40
Nov 2009
Feb 2010
May 2010
Aug 2010
Nov 2010

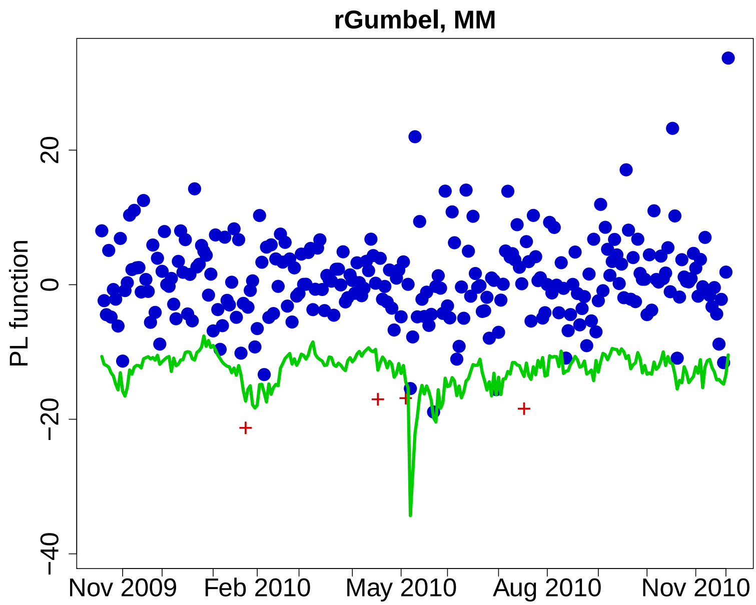


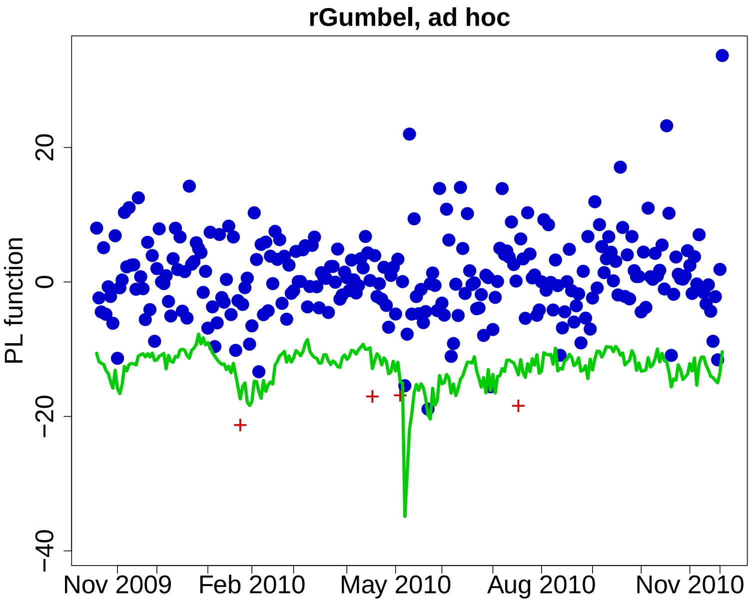
rGumbel, LCP
PL function
20
0
−20
−40
Nov 2009
Feb 2010
May 2010
Aug 2010
Nov 2010



Gauss (Bauer and Vorkink; 2010)
PL function
20
0
−20
−40
Nov 2009
Feb 2010
May 2010
Aug 2010
Nov 2010


Gauss (Chiriac and Voev; 2011)
PL function
20
0
−20
−40
Nov 2009
Feb 2010
May 2010
Aug 2010
Nov 2010


rGumbel, MM
PL function
20
0
−20
−40
Nov 2009
Feb 2010
May 2010
Aug 2010
Nov 2010


rGumbel, ad hoc
PL function
20
0
−20
−40
Nov 2009
Feb 2010
May 2010
Aug 2010
Nov 2010

VaR Performance for Google-IBM-Oracle

model \ α	0.01	0.05	0.1
LCP $m_0 = 40$ (rGumbel)	0.0258 (0.028)	0.0369 (0.300)	0.0775 (0.200)
ROL $w = 250$ (rGumbel)	0.0221 (0.083)	0.0332 (0.177)	0.0664 (0.051)
MM (rGumbel)	**0.0148** (0.462)	0.0590 (0.506)	**0.0996** (0.983)
ad hoc (rGumbel)	**0.0148** (0.462)	0.0590 (0.506)	**0.0996** (0.983)
LCP $m_0 = 40$ (Clayton)	0.0258 (0.028)	**0.0517** (0.900)	0.0849 (0.395)
ROL $w = 250$ (Clayton)	0.0221 (0.083)	0.0443 (0.659)	0.0738 (0.133)
MM (Clayton)	**0.0148** (0.462)	0.0554 (0.690)	0.0959 (0.822)
ad hoc (Clayton)	**0.0148** (0.462)	0.0554 (0.690)	0.0886 (0.522)
Gauss (Bauer and Vorkink; 2010)	0.0406 (1e-04)	0.0738 (0.092)	0.1218 (0.246)
Gauss (Chiriac and Voev; 2011)	0.0369 (6e-04)	0.0812 (0.030)	0.1255 (0.177)

Table 2: VaR performance ($\widehat{\alpha}$) for the Google-IBM-Oracle portfolio. p-values of the Kupiec test in brackets.

VaR Performance for IBM-Pfizer-Exxon

model \ α	0.01	0.05	0.1
LCP $m_0 = 40$ (rGumbel)	**0.0111** (0.861)	0.0443 (0.659)	0.0701 (0.084)
ROL $w = 250$ (rGumbel)	**0.0111** (0.861)	0.0332 (0.177)	0.0517 (0.003)
MM (rGumbel)	0.0074 (0.649)	0.0554 (0.691)	**0.1033** (0.856)
ad hoc (rGumbel)	0.0074 (0.649)	**0.0517** (0.900)	**0.1033** (0.856)
LCP $m_0 = 40$ (Clayton)	0.0185 (0.211)	0.0554 (0.690)	0.0923 (0.667)
ROL $w = 250$ (Clayton)	**0.0111** (0.861)	0.0369 (0.300)	0.0590 (0.015)
MM (Clayton)	0.0074 (0.649)	0.0554 (0.690)	**0.1033** (0.856)
ad hoc (Clayton)	0.0074 (0.649)	0.0554 (0.690)	**0.1033** (0.856)
Gauss (Bauer and Vorkink; 2010)	0.0369 (0.000)	0.0738 (0.092)	0.1107 (0.563)
Gauss (Chiriac and Voev; 2011)	0.0406 (0.000)	0.0738 (0.092)	0.1144 (0.439)

Table 3: VaR performance ($\widehat{\alpha}$) for the IBM-Pfizer-Exxon portfolio. p-values of the Kupiec test in brackets.

Conclusions

- We introduce the notion of realized copula.
- We suggest a forecasting framework for RCop and thus extend the literature on multivariate RCov models.
- Empirically, we find that model relying on daily data are too inert for good forecasts.
- Standard RCov model are more adaptive, but are dominated by copula models.
- RCop unites both advantages and shows nice forecasting performance.

References

H. Joe
Multivariate Models and Concept Dependence
Chapman & Hall, 1997

V. Spokoiny
Local Parametric Methods in Nonparametric Estimation
Springer Verlag, 2009

E. Giacomini, W. Härdle and V. Spokoiny
Inhomogeneous Dependence Modeling with Time-Varying Copulae
Journal of Business and Economic Statistics, 27(2), 2009

O. Okhrin and Y. Okhrin and W. Schmid
On the Structure and Estimation of Hierarchical Archimedean Copulas
under revision in *Journal of Econometrics*, 2009

Appendix

- Realized kernels
- ML estimation
- Details on LCP
- Kupiec (1995) test

Basis

Let $Y = (Y_1, \ldots, Y_d)^\top$ be a d-dim efficient *(log)price* process

$$dY_t = \mu_t dt + \sigma_t dW_t$$

The market microstructure effect is modeled through an additive component

$$\begin{aligned} P_{jt} &= Y_{jt} + U_{jt}, \text{ with } \boldsymbol{E}(U_{jt}) = 0 \\ \sum_h |h\Omega_h| &< \infty, \text{ where } \Omega_{jh} = Cov(U_{jt}, U_{j,t-h}). \end{aligned}$$

Usual aim: Estimate the *quadratic variation* of Y, i.e. $[Y] = \int_0^1 \Sigma_u du$, with $\Sigma = \sigma\sigma^\top$.

Naive Estimator (realized co/variance)

Synchronization – *last traded*: for time t, the log-price for asset j is given by P_{j,t^*} with $t^* = \max\{t_{j,i} | t_{j,i} \leq t, \ \forall i = 1, \ldots, N_j\}$.
$M = M(m)$ number of subintervals of length m (in seconds)

$$
\begin{aligned}
\mathrm{RC}_{t_1,m,j_1,j_2}(P) &= \sum_{i=1}^{M} (P_{j_1,t_i} - P_{j_1,t_{i-1}})(P_{j_2,t_i} - P_{j_2,t_{i-1}}), \\
\mathrm{RC}_{t_1,m}(P) &= \{\mathrm{RC}_{m,j_1,j_2}\}_{j_1,j_2}, \text{ for } j_1, j_2 = 1, \ldots, d
\end{aligned}
$$

Realized Kernels, BNHLS (2011, JoE)

Synchronization – *refresh time sampling*

$$\begin{aligned}\tau_1 &= \max\{t_{1,1}, \ldots, t_{d,1}\} \\ \tau_{i+1} &= \arg\min\{t_{j,k_j} | t_{j,k_j} > \tau_i,\ \forall j \in 1 \ldots d\}\end{aligned}$$

Leads to new high-frequency vector of returns $p_i = P_{\tau_i} - P_{\tau_{i-1}}$, where $i = 1, \ldots, n$ and n is the of refresh time observations.

Realized Kernels

The multivariate realized kernel is defined as

$$K(P) = \sum_{h=-H}^{H} k\left(\frac{|h|}{H+1}\right)\Gamma_h,$$

with Γ_h being a matrix of autocovariances given by

$$\Gamma_h = \begin{cases} \sum_{j=|h|+1}^{n} p_j p_{j-h}^{\top}, & h \geq 0 \\ \sum_{j=|h|+1}^{n} p_{j-h} p_j^{\top}, & h < 0 \end{cases},$$

and $k(x)$ being a weight function of the *Parzen kernel*, defined through

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3 & 0 \leq x \leq 1/2 \\ 2(1-x)^3 & 1/2 \leq x \leq 1 \\ 0 & x > 1 \end{cases}.$$

Realized Kernels

The multivariate bandwidth parameter

$$H = \left[d^{-1} \sum_{j=1}^{d} H_j \right]$$

where $H_j,\ j = 1, \ldots, d$ is chosen by *mean squared error* criteria as

$$H_j = c^* \xi_j^{4/5} n^{3/5}$$

with $c^* = \left\{ k''(0)^2 / \int_0^1 k(x)^2 dx \right\}^{1/5}$, which is equal to $c^* = 3.511678$ for Parzen kernel.
$\xi^2 = \omega / \sqrt{IQ}$ denotes the *noise-to-signal ratio*, where ω^2 is the *measure of microstructural noise variance* and IQ is *the integrated quarticity* as defined in Barndorff-Nielsen and Shephard (2002).

ML estimation of copula parameters

For a sample of observations $\{x_t\}'_{t=1}$ and $\vartheta = (\delta_1, \ldots, \delta_d; \theta) \in \mathbb{R}^{d+1}$ the likelihood function is

$$L(\vartheta; x_1, \ldots, x_T) = \prod_{t=1}^{T} f(x_{1,t}, \ldots, x_{d,t}; \delta_1, \ldots, \delta_d; \theta)$$

and the corresponding log-likelihood function

$$\begin{aligned} \ell(\vartheta; x_1, \ldots, x_T) &= \sum_{t=1}^{T} \log c\{F_{X_1}(x_{1,t}, \delta_1), \ldots, F_{X_d}(x_{d,t}, \delta_d); \theta\} \\ &+ \sum_{t=1}^{T} \sum_{j=1}^{d} \log f_j(x_{j,t}, \delta_j) \end{aligned}$$

"*Oracle*" choice: largest interval $I = [t_0 - m_{k^*}, t_0]$ where the small modelling bias condition (SMB)

$$\triangle_I(\theta) = \sum_{t \in I} \mathcal{K}\{C(\cdot;\theta_t), C(\cdot;\theta)\} \leq \triangle.$$

holds for some $\triangle \geq 0$. m_{k^*} is the ideal scale, θ is ideally estimated from $I = [t_0 - m_{k^*}, t_0]$ and $\mathcal{K}(\cdot,\cdot)$ is the *Kullback-Leibler* divergence

$$\mathcal{K}\{C(\cdot;\theta_t), C(\cdot;\theta)\} = \boldsymbol{E}_{\theta_t} \log \frac{c(\cdot;\theta_t)}{c(\cdot;\theta)}$$

Under the SMB condition on I_{k^*} and assuming that $\max_{k \le k^*} \boldsymbol{E}_{\theta_t} |\mathcal{L}(\widetilde{\theta}_k) - \mathcal{L}(\theta)|^r \le \mathcal{R}_r(\theta_t)$, we obtain

$$\boldsymbol{E}_{\theta_t} \log \left\{ 1 + \frac{|\mathcal{L}(\widetilde{\theta}_{\widehat{k}}) - \mathcal{L}(\theta)|^r}{\mathcal{R}_r(\theta)} \right\} \le 1 + \triangle,$$

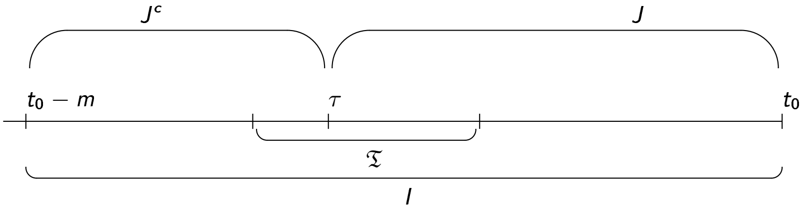
$$\boldsymbol{E}_{\theta_t} \log \left\{ 1 + \frac{|\mathcal{L}(\widetilde{\theta}_{\widehat{k}}) - \mathcal{L}(\widehat{\theta}_{\widehat{k}})|^r}{\mathcal{R}_r(\theta)} \right\} \le 1 + \triangle,$$

where $\widehat{a}_I$ is an adaptive estimator on I and $\widetilde{a}_I$ is any other parametric estimator on I.

Test of homogeneity

Interval $I = [t_0 - m, t_0], \mathfrak{T} \subset I$

$$H_0 \quad : \quad \forall \tau \in \mathfrak{T},\ \theta_t = \theta, \forall t \in J = [\tau, t_0], \forall t \in J^c = I \setminus J$$
$$H_1 \quad : \quad \exists \tau \in \mathfrak{T},\ \theta_t = \theta_1;\ \forall t \in J,\ \theta_t = \theta_2 \neq \theta_1;\ \forall t \in J^c$$



Test of homogeneity

Likelihood ratio test statistic for fixed change point location:

$$
\begin{aligned}
T_{I,\tau} &= \max_{\theta_1,\theta_2}\{L_J(\theta_1) + L_{J^c}(\theta_2)\} - \max_{\theta} L_I(\theta) \\
&= ML_J + ML_{J^c} - ML_I
\end{aligned}
$$

Test statistic for unknown change point location:

$$
T_I = \max_{\tau \in \mathfrak{T}_I} T_{I,\tau}
$$

Reject H_0 if for a critical value ζ_I

$$
T_I > \zeta_I
$$

Selection of I_k and ζ_k

- set of numbers m_k defining the length of I_k and $\mathfrak{T}_k$ are in the form of a geometric grid
- $m_k = [m_0 c^k]$ for $k = 1, 2, \ldots, K$, $m_0 \in \{20,\ 40\}$, $c = 1.25$ and $K = 10$, where $[x]$ means the integer part of x
- $I_k = [t_0 - m_k, t_0]$ and $\mathfrak{T}_k = [t_0 - m_k, t_0 - m_{k-1}]$ for $k = 1, 2, \ldots, K$

(Mystery Parameters)

Sequential choice of ζ_k

- after k steps there are two cases: change point at some step $\ell \le k$ or no change points.
- let $\mathcal{B}_\ell$ be the event meaning the rejection at step ℓ

$$\mathcal{B}_\ell = \{ T_1 \le \zeta_1, \ldots, T_{\ell-1} \le \zeta_{\ell-1}, T_\ell > \zeta_\ell \},$$

and $(\widehat{\theta}_k) = (\widetilde{\theta}_{\ell-1})$ on $\mathcal{B}_\ell$ for $\ell = 1, \ldots, k$.

- we find sequentially such a minimal value of ζ_ℓ that ensures the inequality

$$\max_{k=l,\ldots,K} \boldsymbol{E}_{\theta^*}[|\mathcal{L}(\widetilde{\theta}_k) - \mathcal{L}(\widetilde{\theta}_{\ell-1})|^r \mathbf{I}(\mathcal{B}_\ell)] \le \rho \mathcal{R}_r(\theta^*) k/(K-1)$$

return to LCP

Kupiec (1995) test

LR test based on the binomial model.
$H_0 : \widehat{\alpha} = \alpha$ with test statistic

$$LR_{uc} = 2 \log \frac{\widehat{\alpha}^N (1 - \widehat{\alpha})^{T-N}}{\alpha^N (1 - \alpha)^{T-N}}$$