Robust Multivariate Scale Estimation

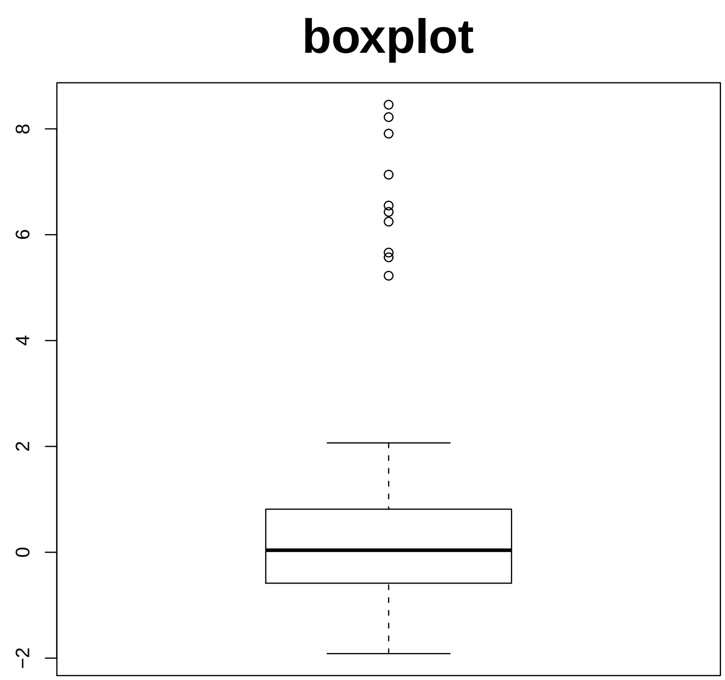
Application to Intra-day Volatility estimation of financial time series

Christophe Croux - Kris Boudt - Sebastien Laurent

CFE09 and ERCIM 2009

Univariate Scale estimation

Univariate Data $y_1, y_2, \ldots, y_n$.



SD =2.16

IQR=1.03

MAD=1.10

Breakdown point

The breakdown point of a scale estimator S is the smallest fraction of observations that you need to replace to other values such that $S \uparrow \infty$ or $S \downarrow 0$.

The breakdown point of a scale estimator S is the smallest fraction of observations that you need to replace to other values such that $S \uparrow \infty$ or $S \downarrow 0$.

- Standard deviation: breakdown point= 1/n ≈ 0.

Breakdown point

The breakdown point of a scale estimator S is the smallest fraction of observations that you need to replace to other values such that $S \uparrow \infty$ or $S \downarrow 0$.

- Standard deviation: breakdown point= 1/n ≈ 0.
- IQR $= 0.74 * \left|y_{(\lfloor 0.75*n \rfloor)} - y_{(\lfloor 0.25*n \rfloor)}\right|$

InterQuartileRange: breakdown point $\approx 25\%$

Breakdown point

The breakdown point of a scale estimator S is the smallest fraction of observations that you need to replace to other values such that $S \uparrow \infty$ or $S \downarrow 0$.

- Standard deviation: breakdown point= 1/n ≈ 0.
- IQR $= 0.74 * |y_{(\lfloor 0.75*n \rfloor)} - y_{(\lfloor 0.25*n \rfloor)}|$

InterQuartileRange: breakdown point $\approx 25\%$

- MAD $= 1.48 * \text{med}_i \, |y_i - \text{med}_j \, y_j|$

Median Absolute deviation: breakdown point $\approx 50\%$

Influence Function

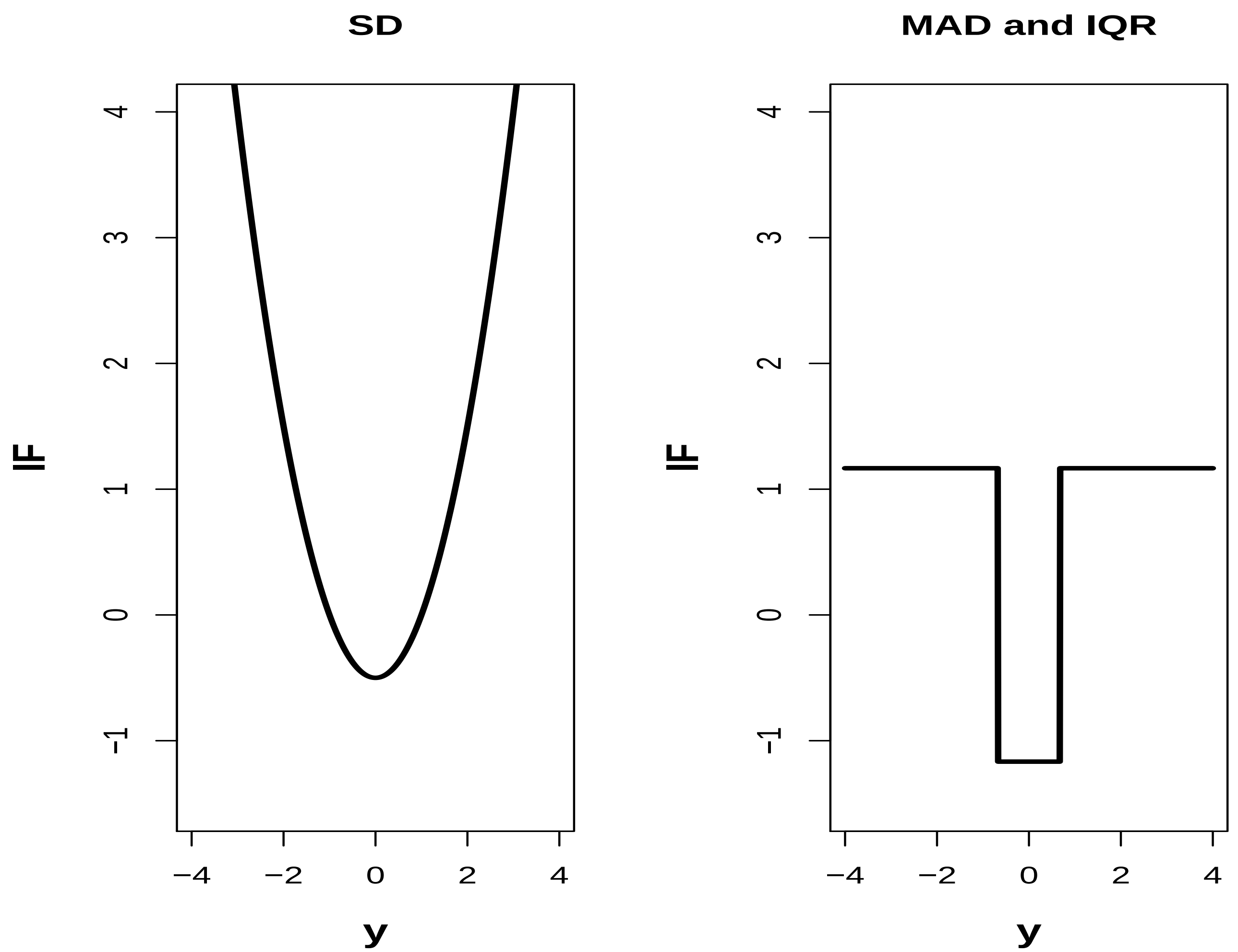
The influence that an observation at position y has on the scale S when sampling from a distribution F:

$$IF(y; S, F) = \lim_{\varepsilon \downarrow 0} \frac{S((1-\varepsilon)F + \varepsilon\Delta_y) - S(F)}{\varepsilon},$$

with Δ_y a Dirac measure at y (see Hampel, 1986).

- The influence function is a local measure of robustness, the breakdown point a global measure.
- The influence function should be bounded.

Influence Function at F=N(0,1)



Multivariate Scale estimation

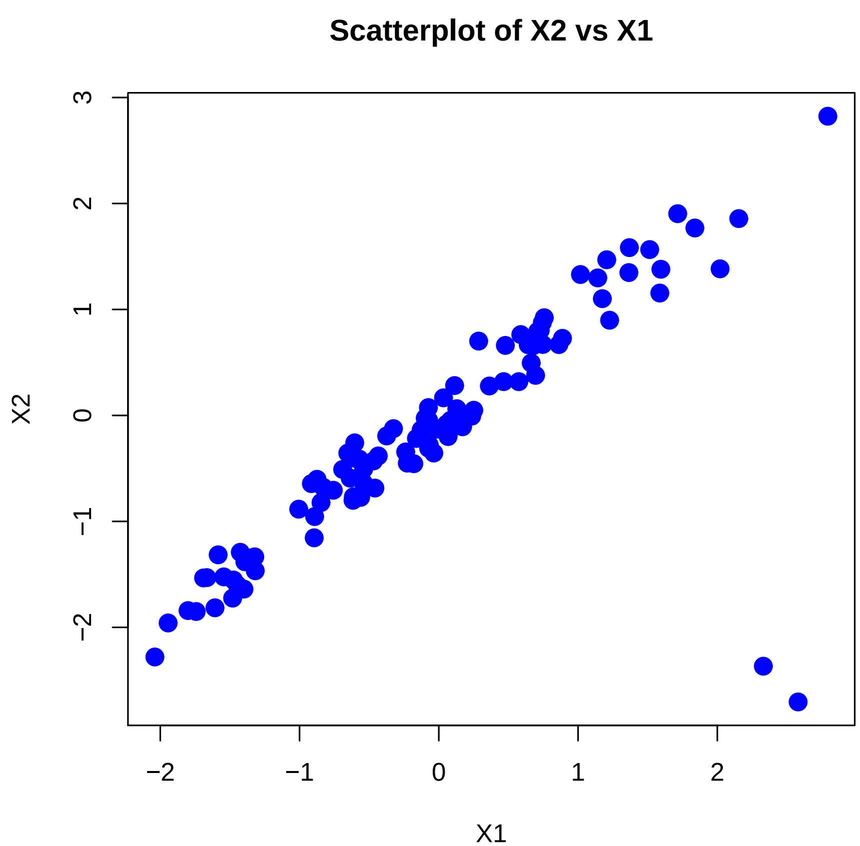
Multivariate Scale estimation

In higher dimensions:

- Construction of robust estimators more challenging.

  Ranking of the observations from smallest to largest is not possible.

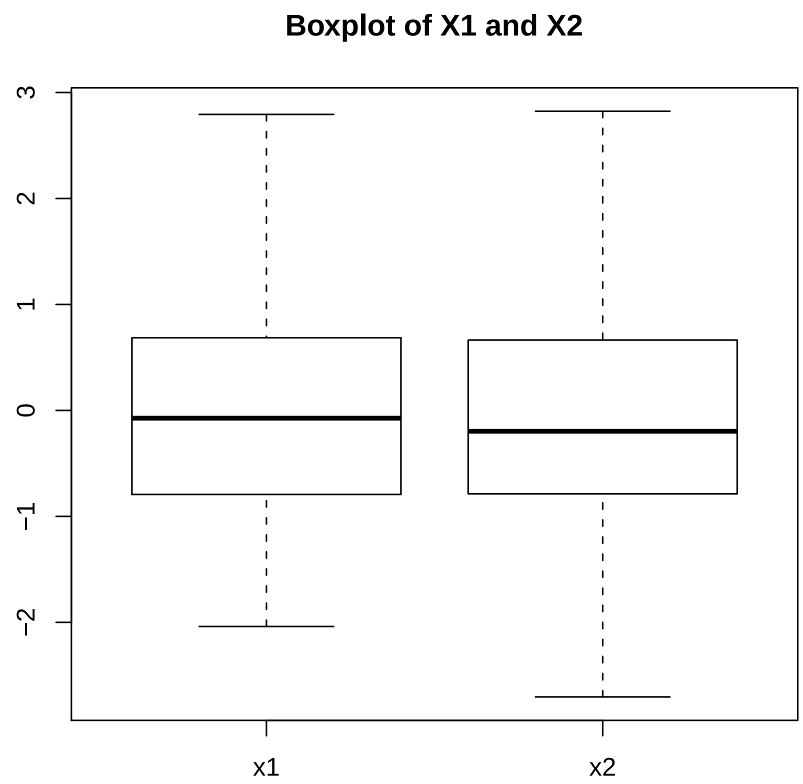
- Outlier detection becomes more difficult.



Outliers for the correlation structure.

$\rho = 0.980$, $\hat{\rho} = 0.774$, $\hat{\rho}_{\text{Robust}} = 0.978$.

Boxplots of the marginals



The correlation outliers are not detected in the marginals !

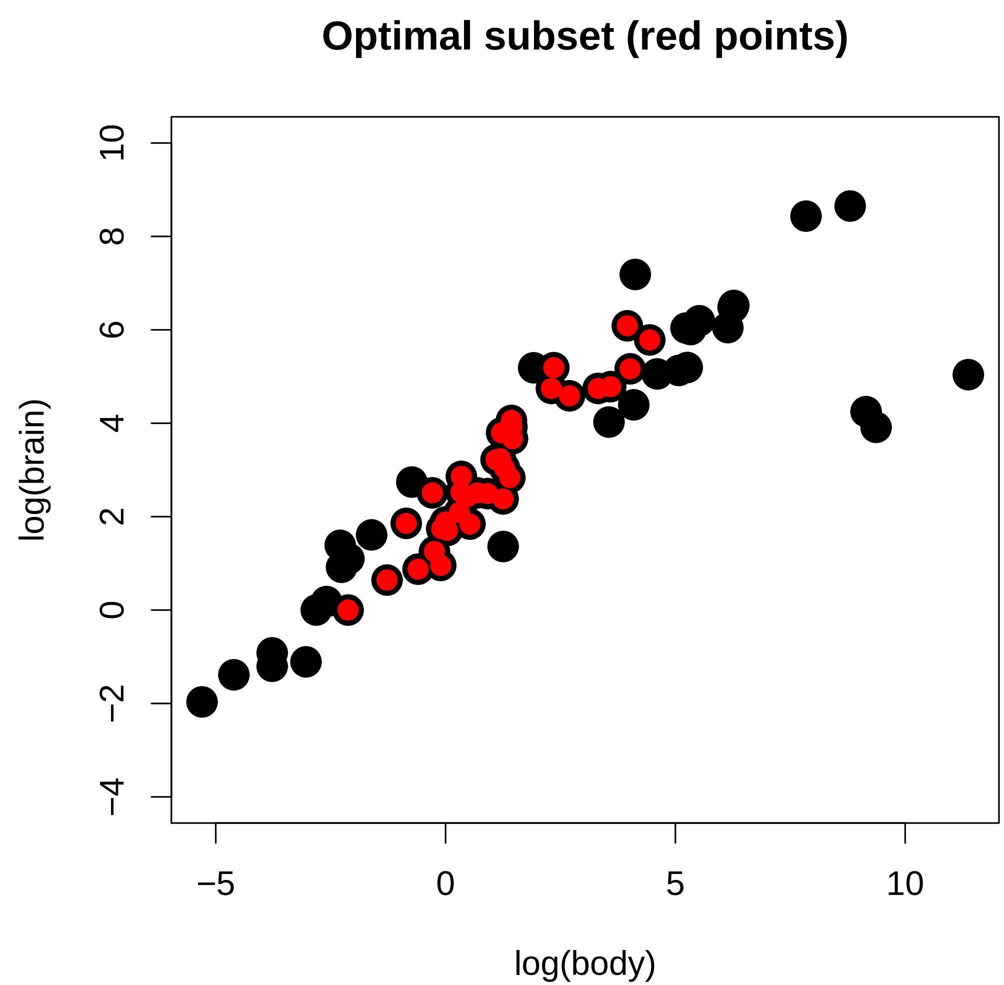
The Minimum Covariance Determinant (MCD)

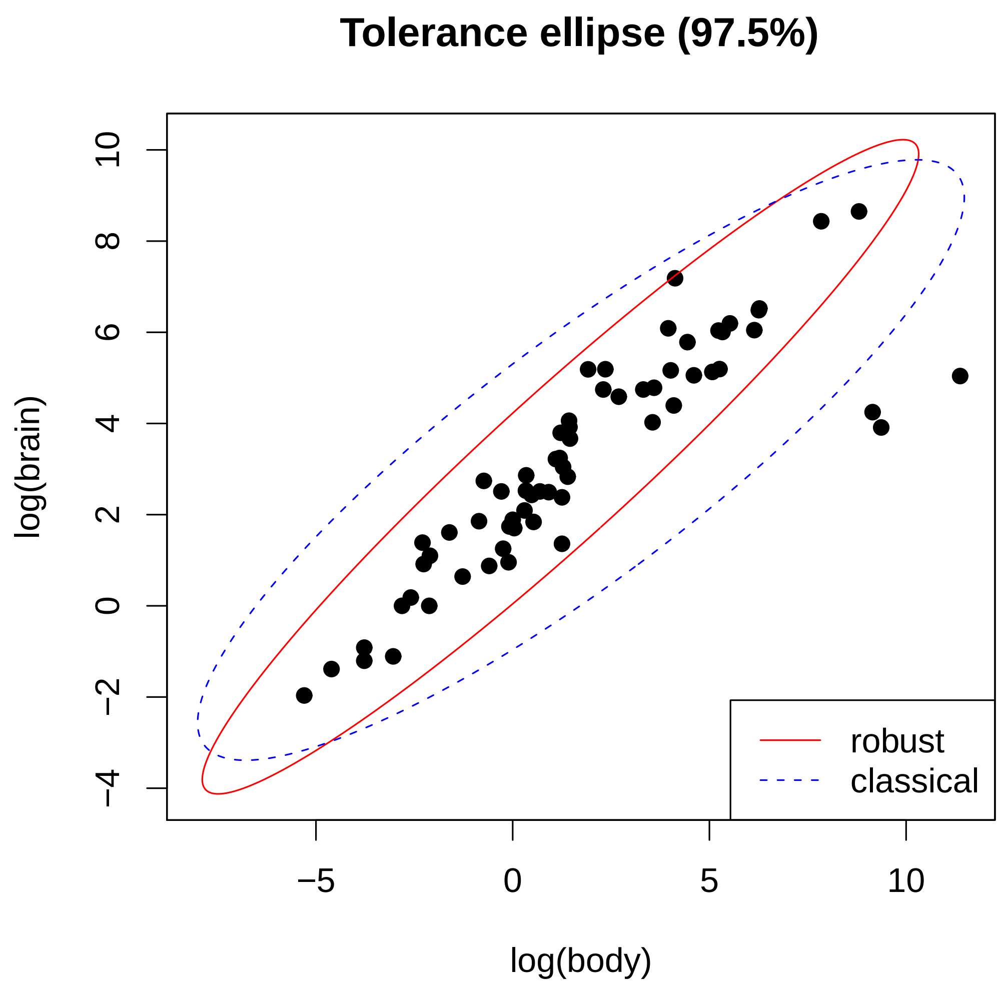
For a p-variate multivariate sample $y_1, \ldots, y_n$:

- Let $h \leq n$ be the size of the *optimal subsample*. Typically $h \approx n/2$.
- For every subsample H of size h, compute $\det(Cov\{y_i | i \in H\})$.
- The optimal subsample is given by $\hat{H} = \text{argmin}_H \det(Cov\{y_i | i \in H\})$.
- The MCD scale estimator is MCD $= Cov\{y_i | i \in \hat{H}\}$.

(Rousseeuw 1985)

Brain and Body Weights for 65 Species of Land Animals




Tolerance ellipse (97.5%)
log(brain)
log(body)
−4
−2
0
2
4
6
8
10
−5
0
5
10
robust
classical

The MCD estimator is a multivariate scale estimator S that is

- robust with respect to outliers, including correlation outliers.
- positive definite
- affine equivariant, meaning that

$$S(Ay_1, \ldots, Ay_n) = A\, S(y_1, \ldots, y_n)\, A^t.$$

Moreover, the MCD is

1. Asymptotically normal (Lopuhaä 09, C and Haesbroek 99)
2. Fast to compute (Rousseeuw and Van Driessen 99)
   - FAST-MCD algorithm, based on concentration steps.
   - Aims at finding the “most concentrated” subsample of size h.
   - R-packages: robustbase (covMcd) or rrcov (CovMcd)

(assume location equal to zero, for simplicity)

Let H_0 be a starting subsample of size h. Perform Concentration steps

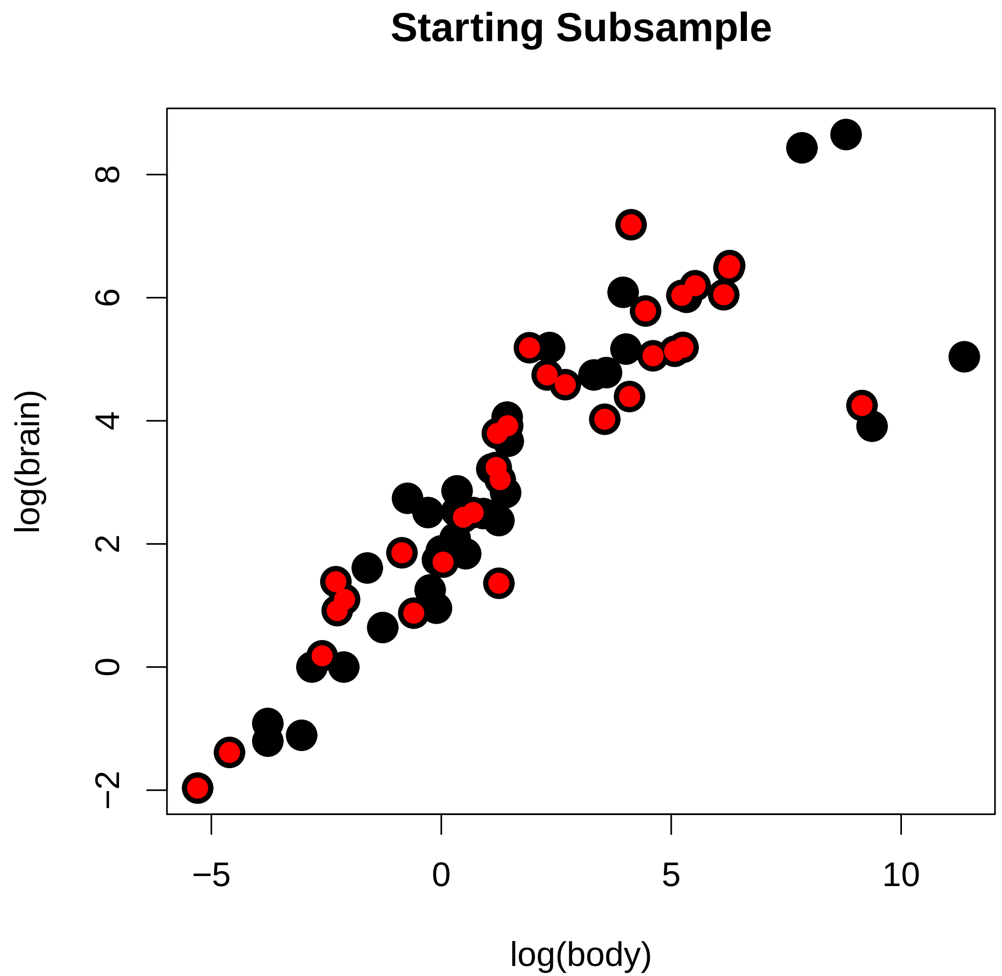
1. (a) $S_0 = Cov\{y_i | i \in H_0\}$.

   (b) H_1 collects the observations with the h smallest values of
   $$y_i^t S_0^{-1} y_i$$

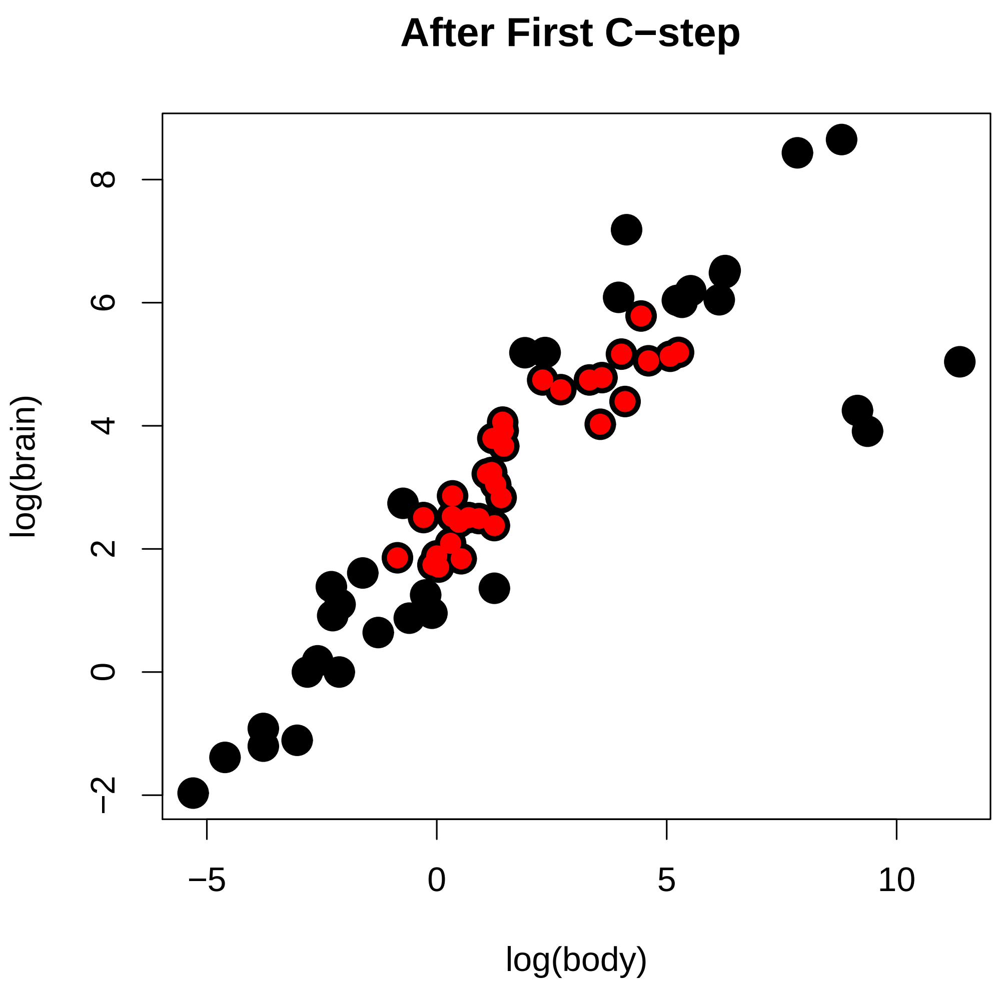
2. (a) $S_1 = Cov\{y_i | i \in H_1\}$.

   (b) H_2 collects the observations with the h smallest values of
   $$y_i^t S_1^{-1} y_i$$

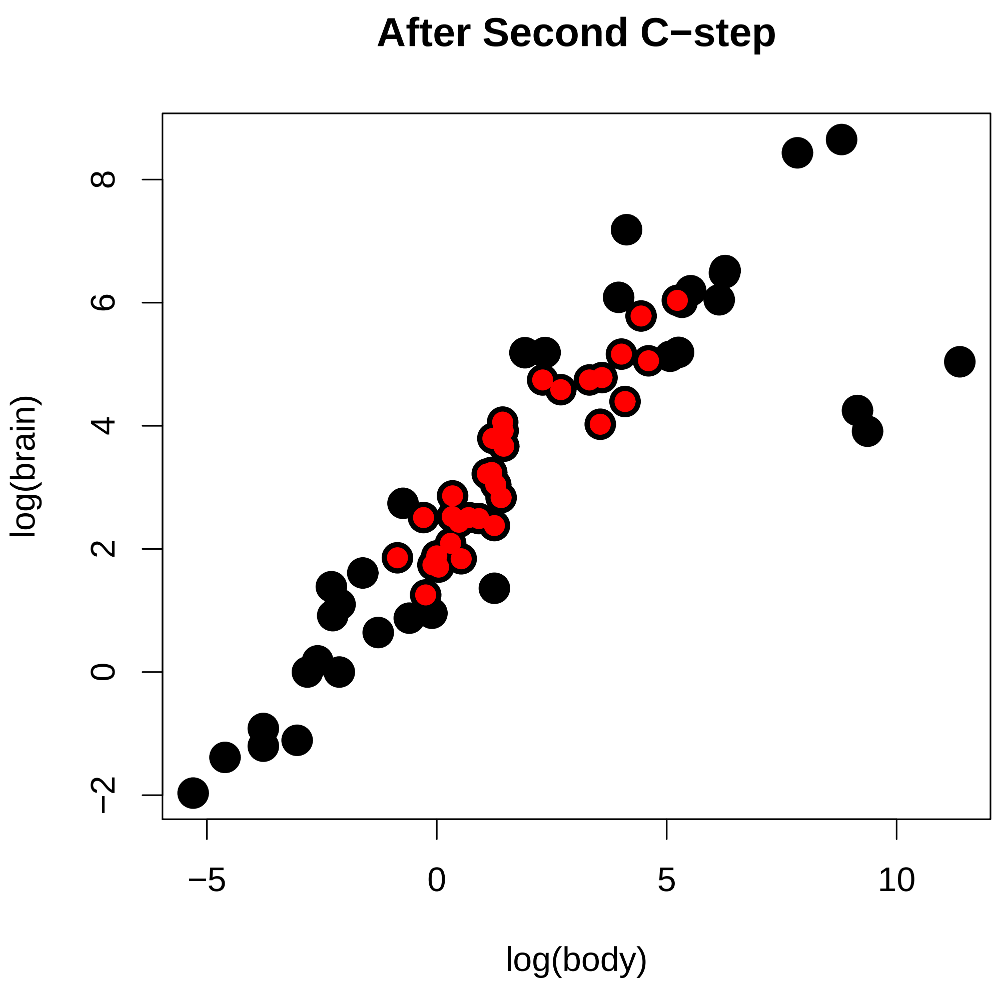
3. ...

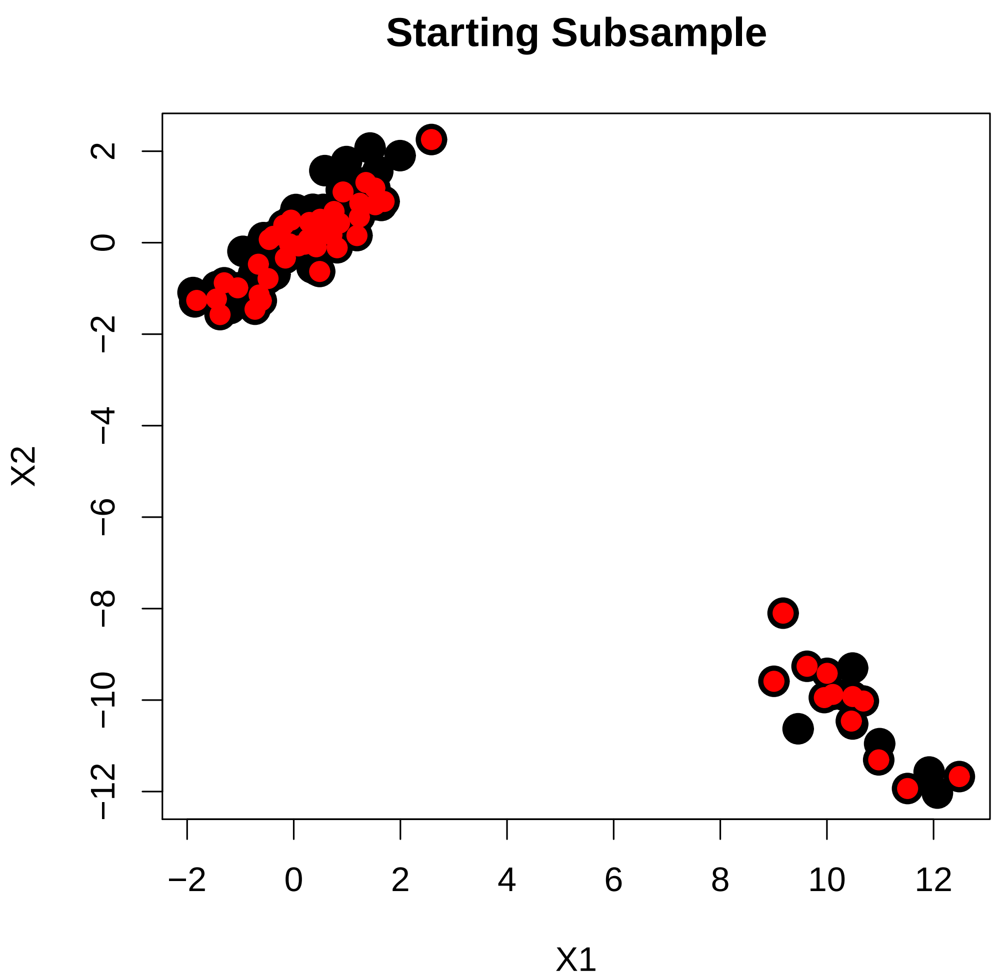

Starting Subsample
log(brain)
log(body)
−2
0
2
4
6
8
−5
0
5
10

Best subsample after step 1



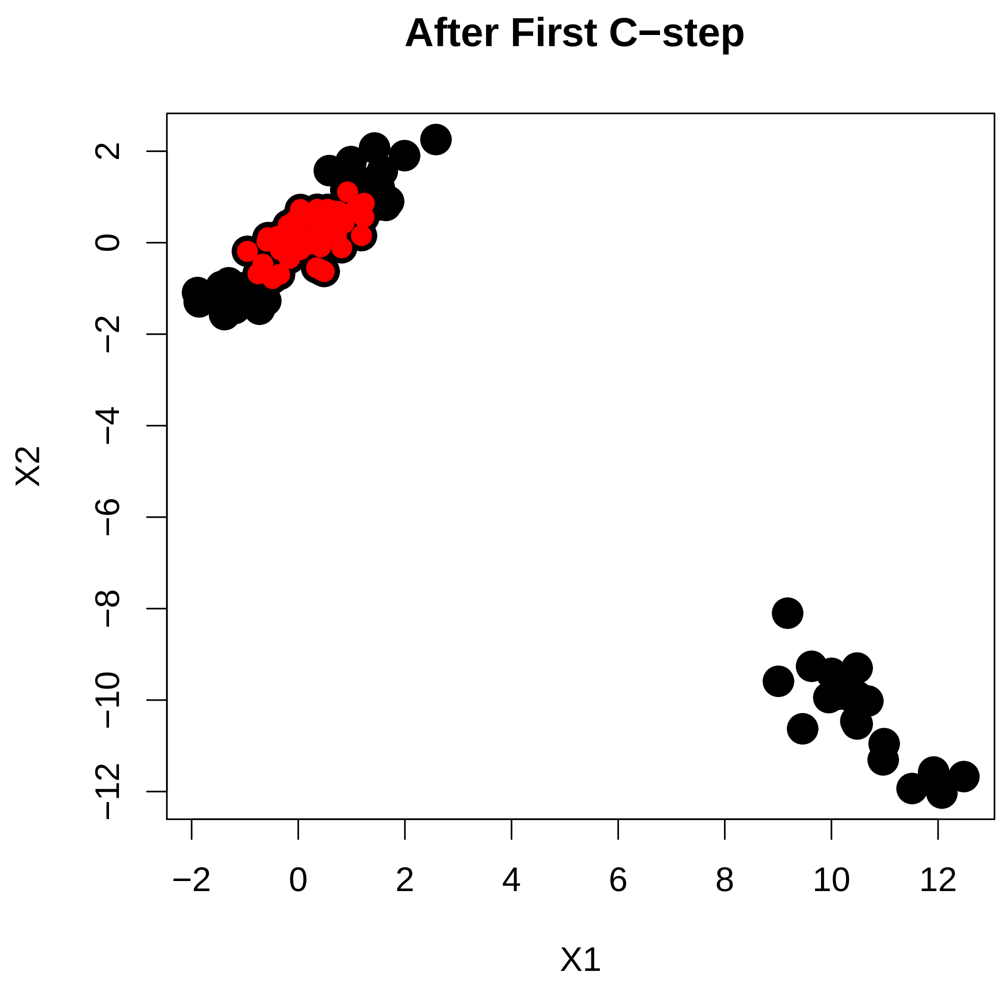
Best subsample after step 2



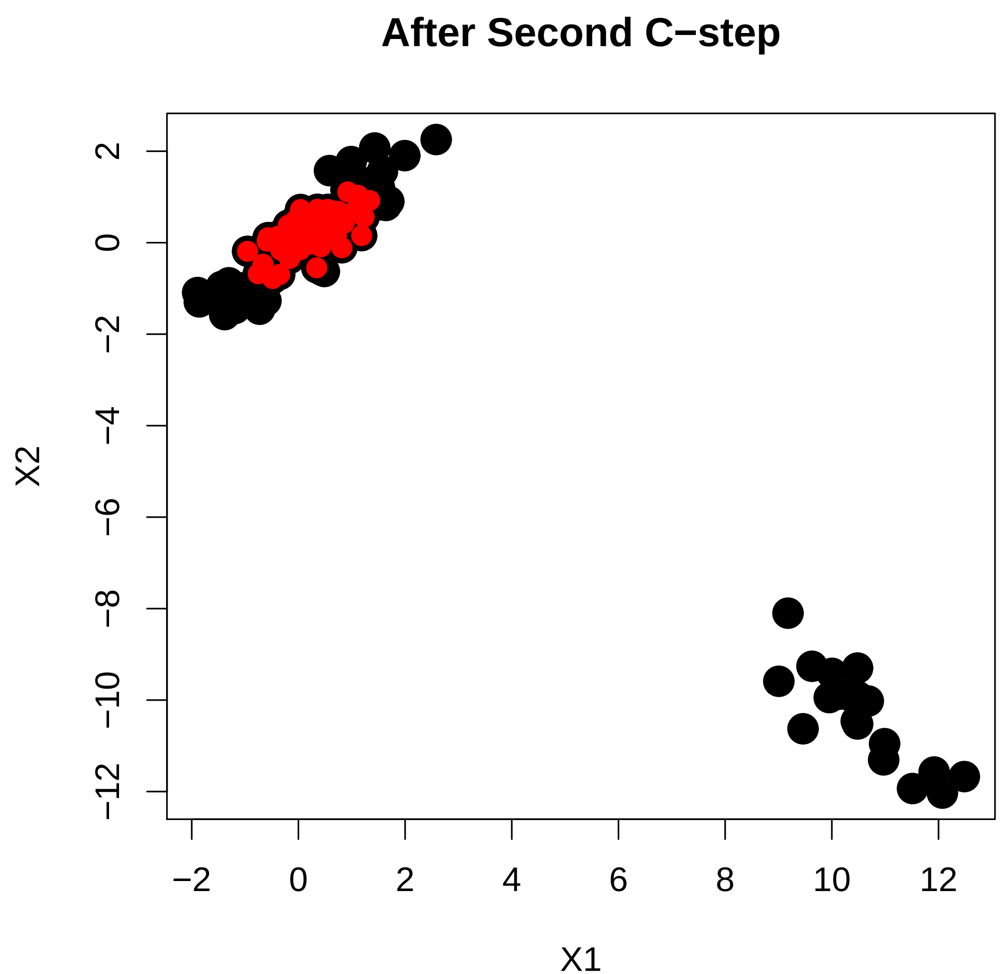


Starting Subsample
X2
X1
2
0
−2
−4
−6
−8
−10
−12
−2
0
2
4
6
8
10
12

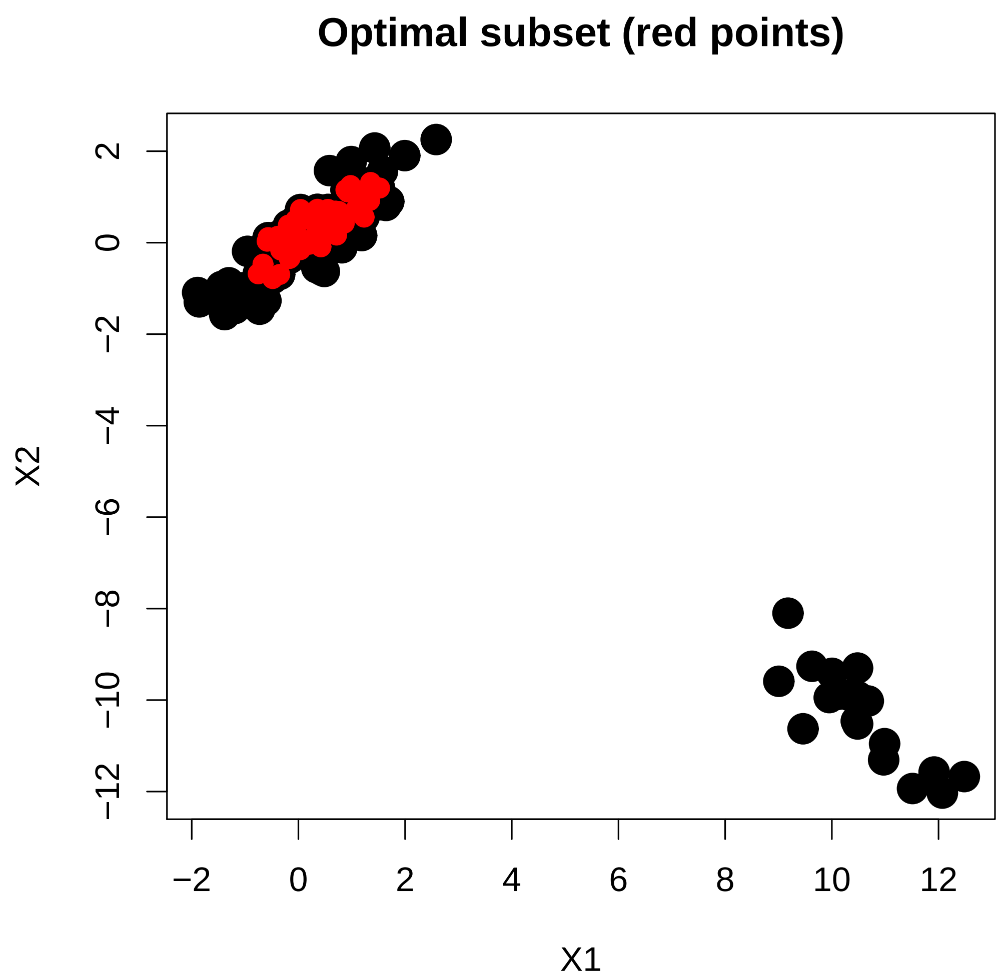
Best subsample after step 1

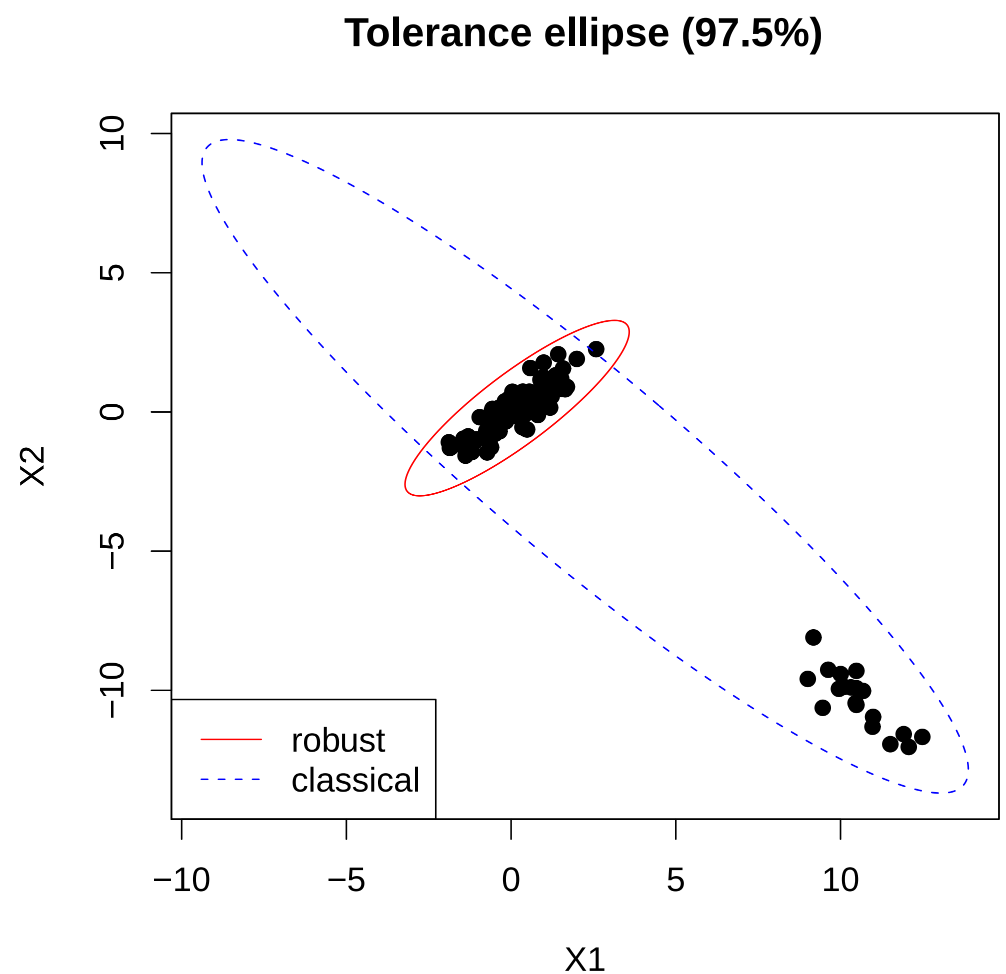


Best subsample after step 2



Final Solution




Tolerance ellipse (97.5%)
10
5
0
−5
−10
X2
robust
classical
−10
−5
0
5
10
X1

1. Compute Spatial Sign Covariance matrix (Oja et al)

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{||y_i|||} \frac{y_i^t}{||y_i|||}$$

How to find a starting value?

1. Compute Spatial Sign Covariance matrix (Oja et al)

   $$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{\|y_i\|} \frac{y_i^t}{\|y_i\|}$$

2. SVD-decomposition $\hat{\Sigma} = UDU^t$. Denote u_j, for $j = 1, \ldots, p$, the eigenvectors.

3. $\hat{\sigma}_j = \text{MAD}(u_j^t y_1, \ldots, u_j^t y_n)$, for $1 \leq j \leq p$
   $\tilde{D} = diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$ and $\hat{\Sigma}_1 = U\tilde{D}U^t$.

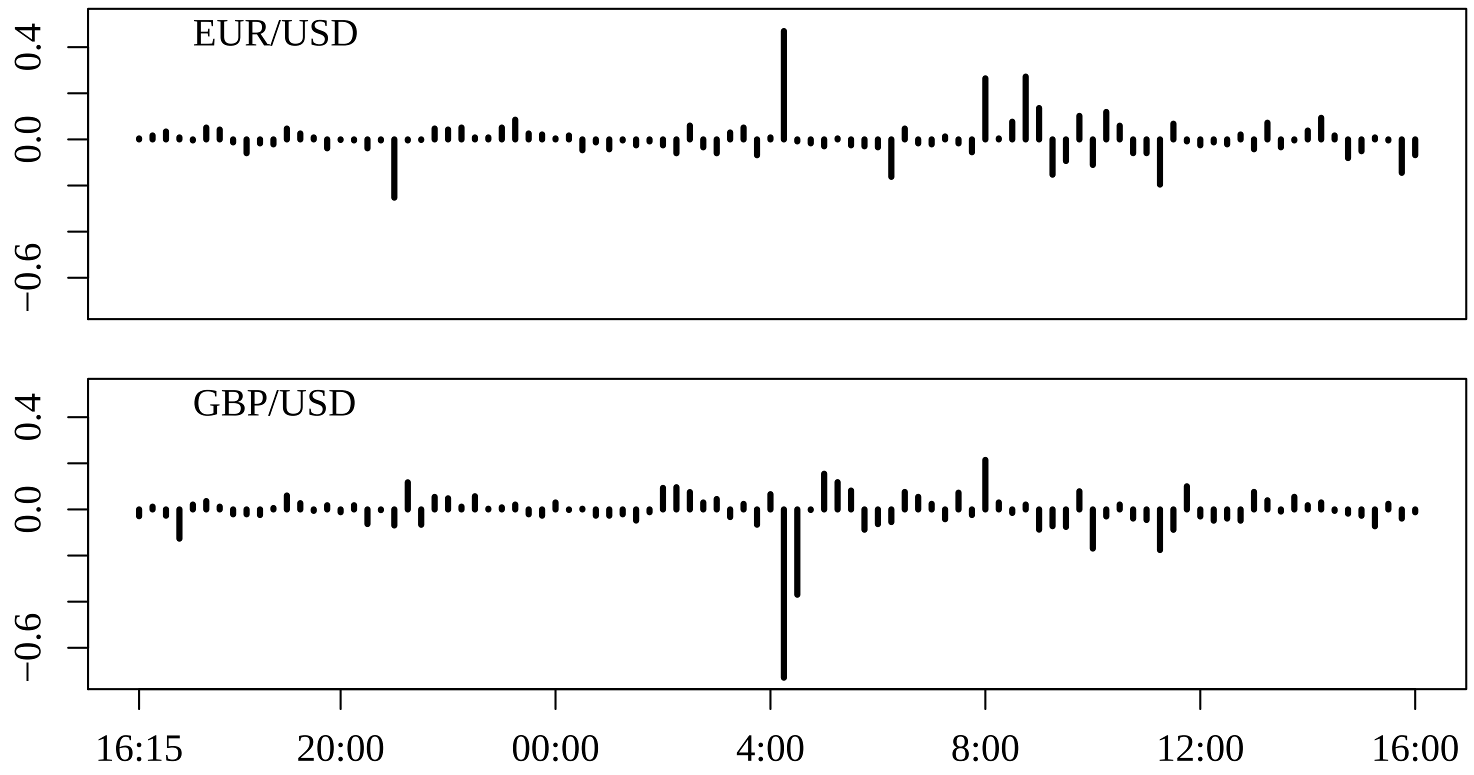
4. H_0 collects the h observations with smallest values of $y_i^t(\hat{\Sigma}_1)^{-1}y_i$.

(see also Verdonck, Hubert, Rousseeuw, Ercim 09)

Intra-day Volatility of Multivariate return series

High frequency data

EUR/USD and GBP/USD exchange rates returns on June 9, 2003.



Aim: measure daily variance + daily correlation.

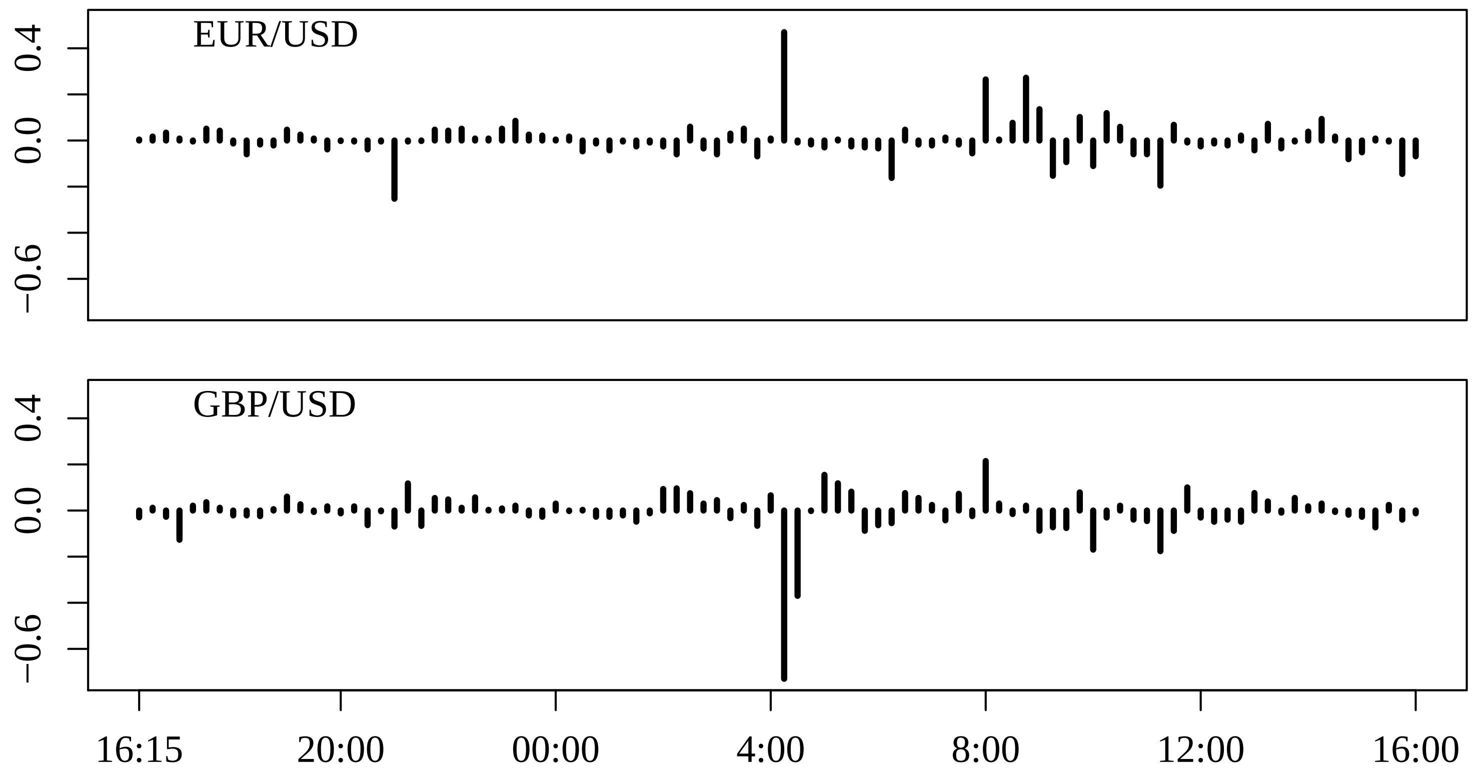
Notation

- Denote $p(s)$ the log of the price vector observed at time s
- Normalize the length of one day to 1
- We observe the price at every Δ units of time (Δ small)
- The i-th intraday return vector at day t is

$$r_i \equiv r_{t,i,\Delta} = p(t + i\Delta) - p(t + (i-1)\Delta),$$

with $i = 1, \ldots, \lfloor 1/\Delta \rfloor$.

Daily Realized Quadratic Covariation



$$\text{RQCOV}_{t,\Delta} \quad = \sum_i r_i r_i'$$

Brownian Semi-Martingale model (BSM)

Log-price process:

$$dp(s) = \mu(s)ds + \Omega(s)dw(s),$$

with $w(s)$ a Brownian motion, $\mu(s)$ is the drift process,

$\Sigma(s) = \Omega(s)\Omega'(s)$ is the spot volatility/covariance process.

Log-price process:

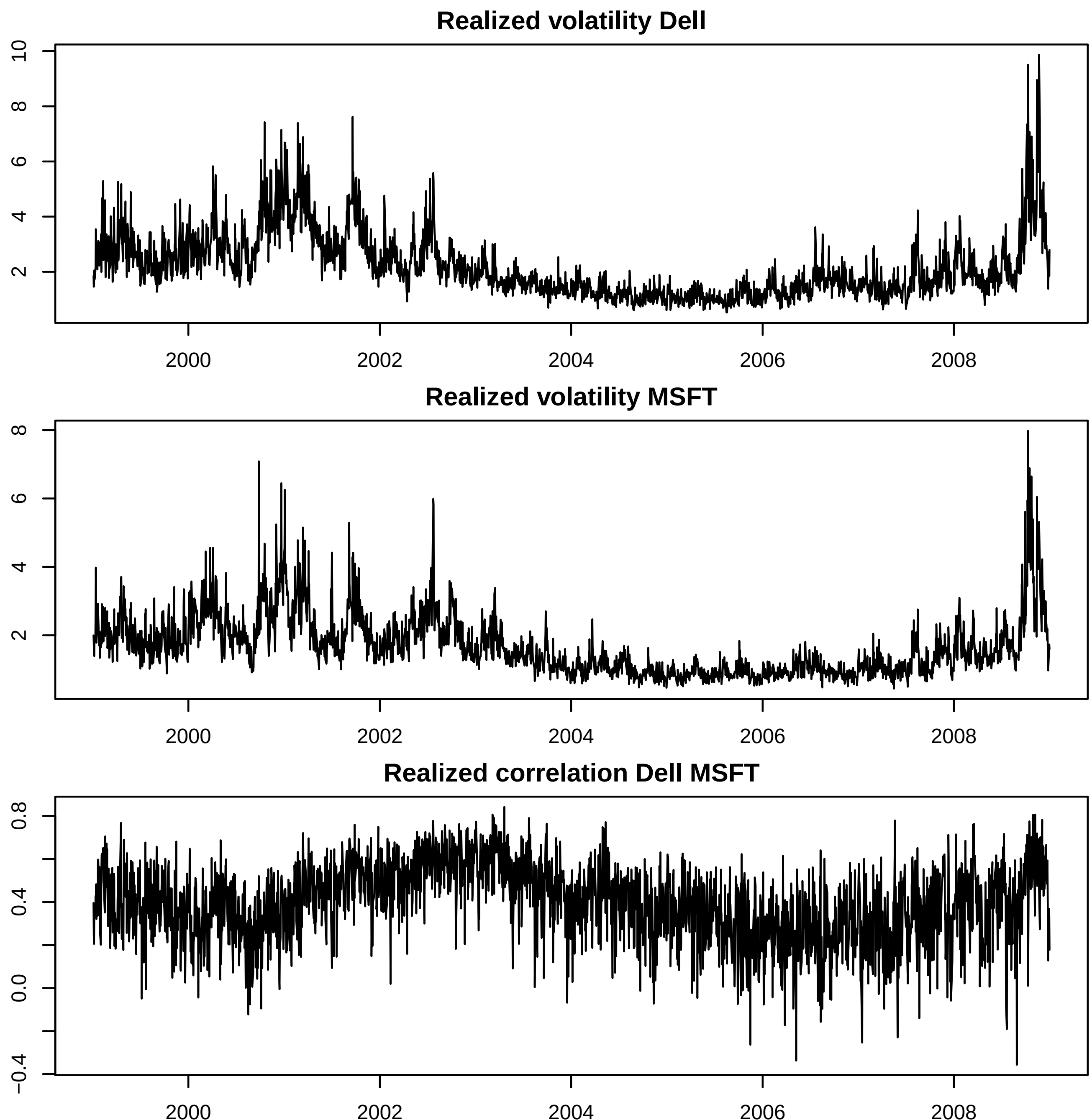
$$dp(s) = \mu(s)ds + \Omega(s)dw(s),$$

with $w(s)$ a Brownian motion, $\mu(s)$ is the drift process, $\Sigma(s) = \Omega(s)\Omega'(s)$ is the spot volatility/covariance process.

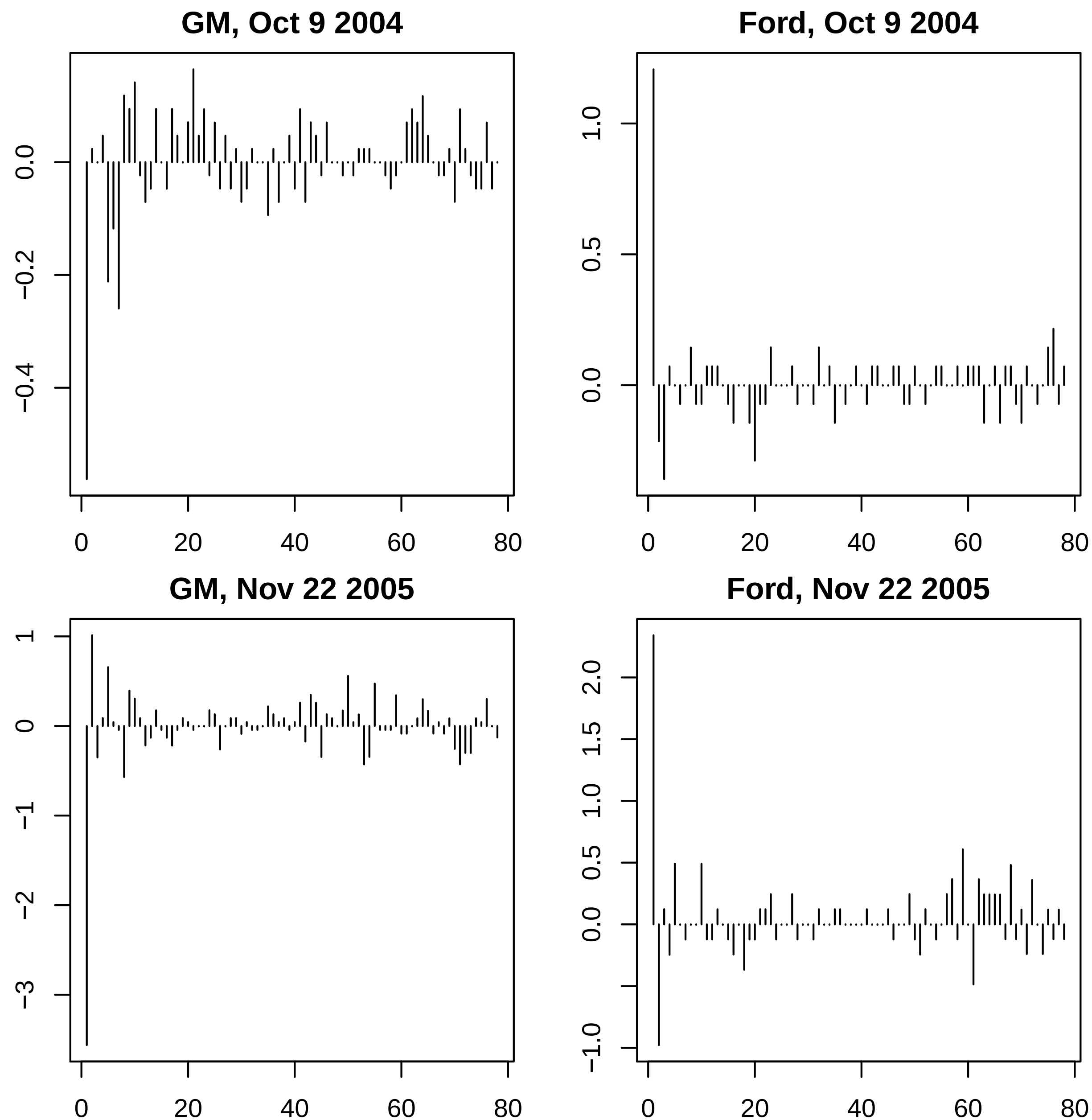
$$\text{RQCov}_{t,\Delta} = \sum_i r_i r_i' \xrightarrow{\Delta \downarrow 0} \int_{t-1}^{t} \Sigma(s)ds = \text{ICov}_t$$

with ICov_t the Integrated CoVariance at day t.

Example: Dell and MSFT



But price jumps may exist



Price process:

$$dp(s) = \mu(s)ds + \Omega(s)dw(s) + \kappa(s)dq(s)$$

Jump process $\kappa(s)dq(s)$ has two components:

- A count process $q(s)$ governing jump occurrences
- A process generating the size of the jumps $\kappa(s)$

Price process:

$$dp(s) = \mu(s)ds + \sigma(s)dw(s) + \kappa(s)dq(s)$$

One has

$$\text{RQCov}_{t,\Delta} = \sum_i r_i r_i' \xrightarrow{\Delta \downarrow 0} \int_{t-1}^{t} \Sigma(s)ds + \sum_{j=1}^{j_t} \kappa_j \kappa_j'$$

$\int_{t-1}^{t} \Sigma(s)ds$ = ICov_t = Integrated Covariance at day t

$\sum_{j=1}^{j_t} \kappa_j \kappa_j'$ = Jump contribution to Quadratic Covariation.

AIM: estimate ICov_t

Motivation

Aim: disentangle the continuous component and the jump component of the volatility

Aim: disentangle the continuous component and the jump component of the volatility

Why? Better volatility forecasts.

Some references:

Andersen, Bollerslev, Diebold, and Lapys, P. (2001, 2003)

Barndorff-Nielsen and Shephard (2004)

Andersen, Bollerslev, and Diebold (2007)

Aim: disentangle the continuous component and the jump component of the volatility

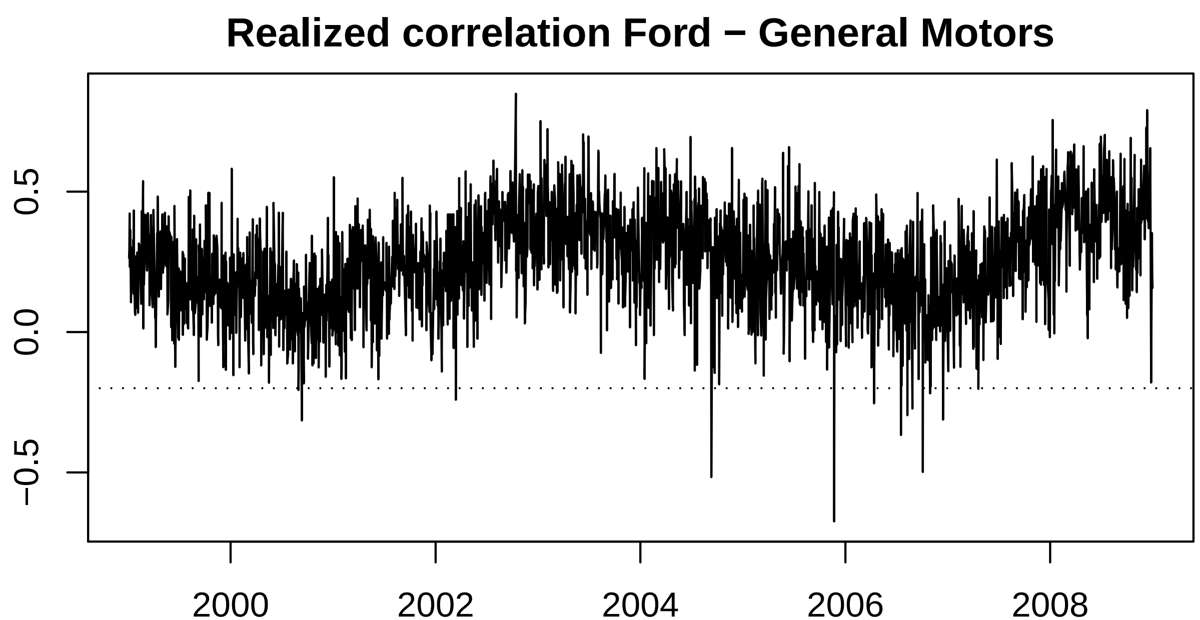
- Univariate: several proposals, starting with the Realized Bipower Variation (Barndorff-Nielsen and Shephard, 2004).

Aim: disentangle the continuous component and the jump component of the volatility

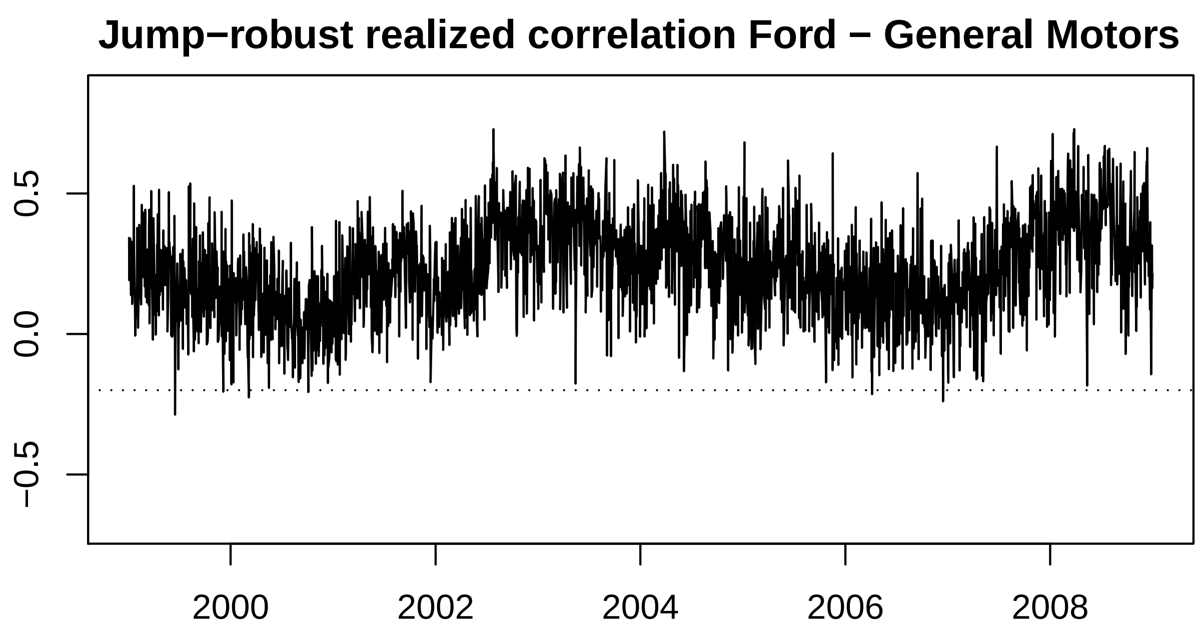
- Univariate: several proposals, starting with the Realized Bipower Variation (Barndorff-Nielsen and Shephard, 2004).
- Multivariate: we propose the
  Realized Outlyingness Weighted Quadratic Covariation
  (ROWQCov)

  Positive Definite + Affine Equivariant + Jump Robust

Example 1



Realized correlation Ford − General Motors
0.5
0.0
−0.5
2000
2002
2004
2006
2008



Jump−robust realized correlation Ford − General Motors
0.5
0.0
−0.5
2000
2002
2004
2006
2008

- As an alternative to $\text{RQCov}_t = \sum_i r_i r_i'$
- the Realized Outlyingness Weighted Quadratic Covariation matrix is defined as

$$\text{ROWQCov}_t = c \frac{\sum_i w_i r_i r_i'}{\sum_i w_i}.$$

ROWQCov

- As an alternative to $\text{RQCov}_t = \sum_i r_i r_i'$
- the Realized Outlyingness Weighted Quadratic Covariation matrix is defined as

$$\text{ROWQCov}_t = c \frac{\sum_i w_i r_i r_i'}{\sum_i w_i}.$$

The weight $w_i = w(r_i' \hat{\Sigma}_i^{-1} r_i)$, with $w(\cdot)$ is a descending weight function.

$\hat{\Sigma}_i$ is a robust estimate of the covariance of r_i.

How to compute $\hat{\Sigma}_i$?

Divide the day in local windows of length λ.

- Compute $\hat{\Sigma}_i$ as the MCD of the returns belonging to the same window as r_i.

How to compute $\hat{\Sigma}_i$?

Divide the day in local windows of length λ.

- Compute $\hat{\Sigma}_i$ as the MCD of the returns belonging to the same window as r_i.

Select λ

(i) small enough to have locally constant scale

(ii) large enough to have enough observations in the window.

$\lambda \to 0, \Delta \to 0$ and $\lambda/\Delta \to \infty$.

Weight function

Jumps are outliers with respect to the regular returns.

Weight for return i is $w_i = w(r_i'\hat{\Sigma}_i^{-1}r_i) = w(d_i)$.

For $\Delta \downarrow 0$:

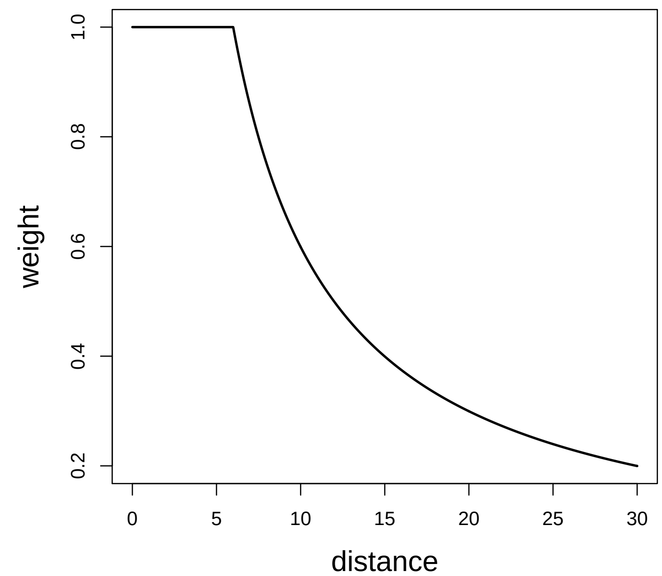
- Return not affected by a jump:

$$d_i = r_i'\hat{\Sigma}_i^{-1}r_i \sim \chi_p^2$$

- Return affected by a jump:

$$d_i = r_i'\hat{\Sigma}_i^{-1}r_i \rightarrow \infty$$

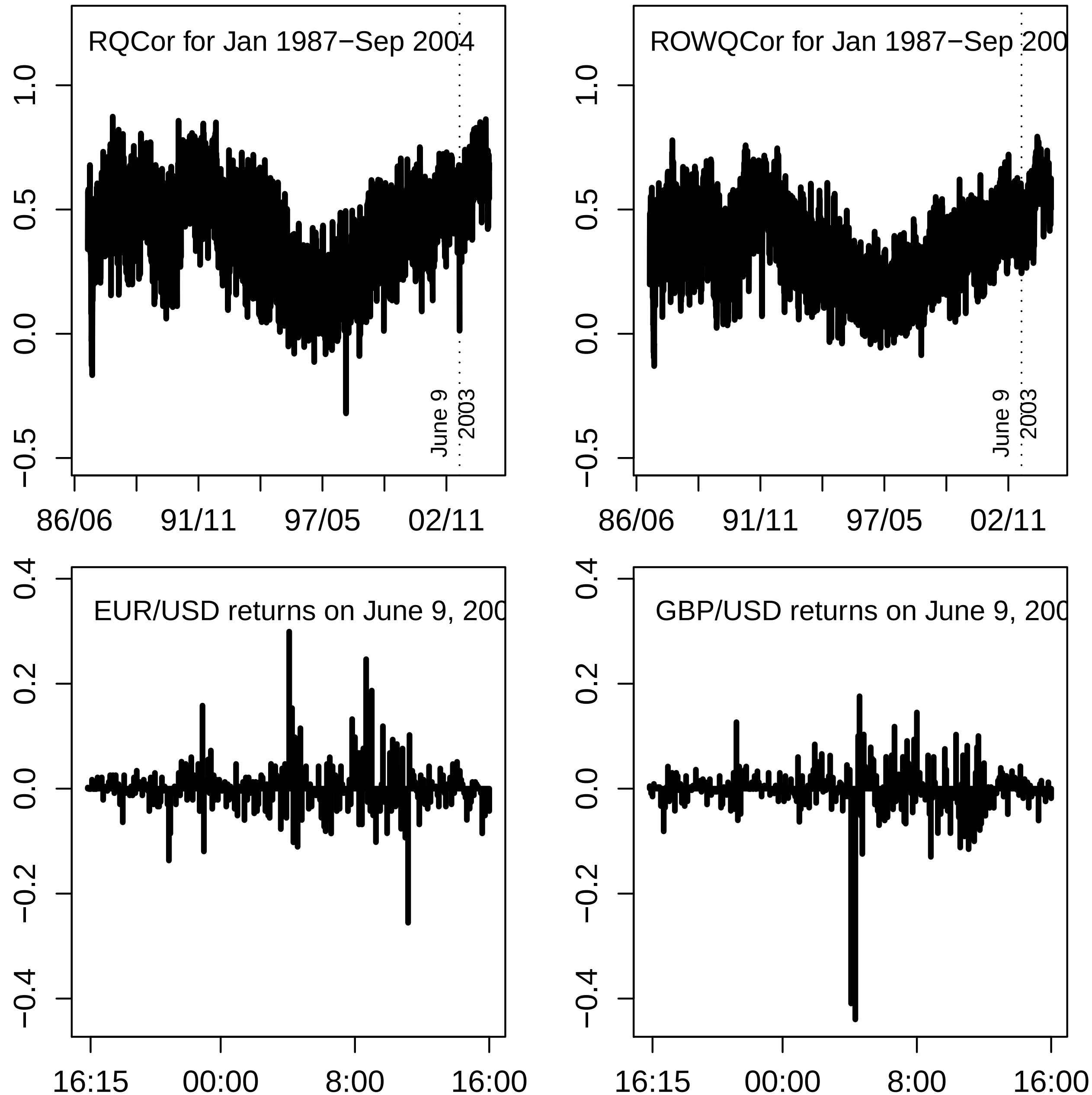
Redescending Weight function



$w(d) = \min(1, \chi^2_{p,1-\alpha}/d)$; we take $\alpha = 0.05$.

! By letting α tend to zero, the efficiency gets arbitrarily close to 1.

Example 2: EUR/USD and GBP/USD exchange



Other proposals for jump-robust intra-day volatility measures

Let $r_1, \dots, r_n$ be the return series within a day. Recall that

$$\text{RVar} = \frac{1}{n} \sum_{i=1}^{n} r_i^2$$

- Realized Bipower variation (BN, Shephard 2004):

$$\text{RBPVar} = \frac{\pi}{2} \sum_{i=2}^{n} |r_i||r_{i-1}|$$

Let $r_1, \ldots, r_n$ be the return series within a day. Recall that

$$\text{RVar} = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

- Realized Bipower variation (BN, Shephard 2004):

$$\text{RBPVar} = \frac{\pi}{2}\sum_{i=2}^{n} |r_i||r_{i-1}|$$

- MinRV estimator of Andersen, Dobrev, Schaumburg (2008).

$$\text{MinRV} = \frac{\pi}{\pi - 2}\sum_{i=2}^{n} \min(|r_i|, |r_{i-1}|)^2$$

- ...

RBPVar and MinRV are consistent estimators of daily integrated variance in presence of jumps ($\Delta \downarrow 0$).

They do not require estimates of local scale, but

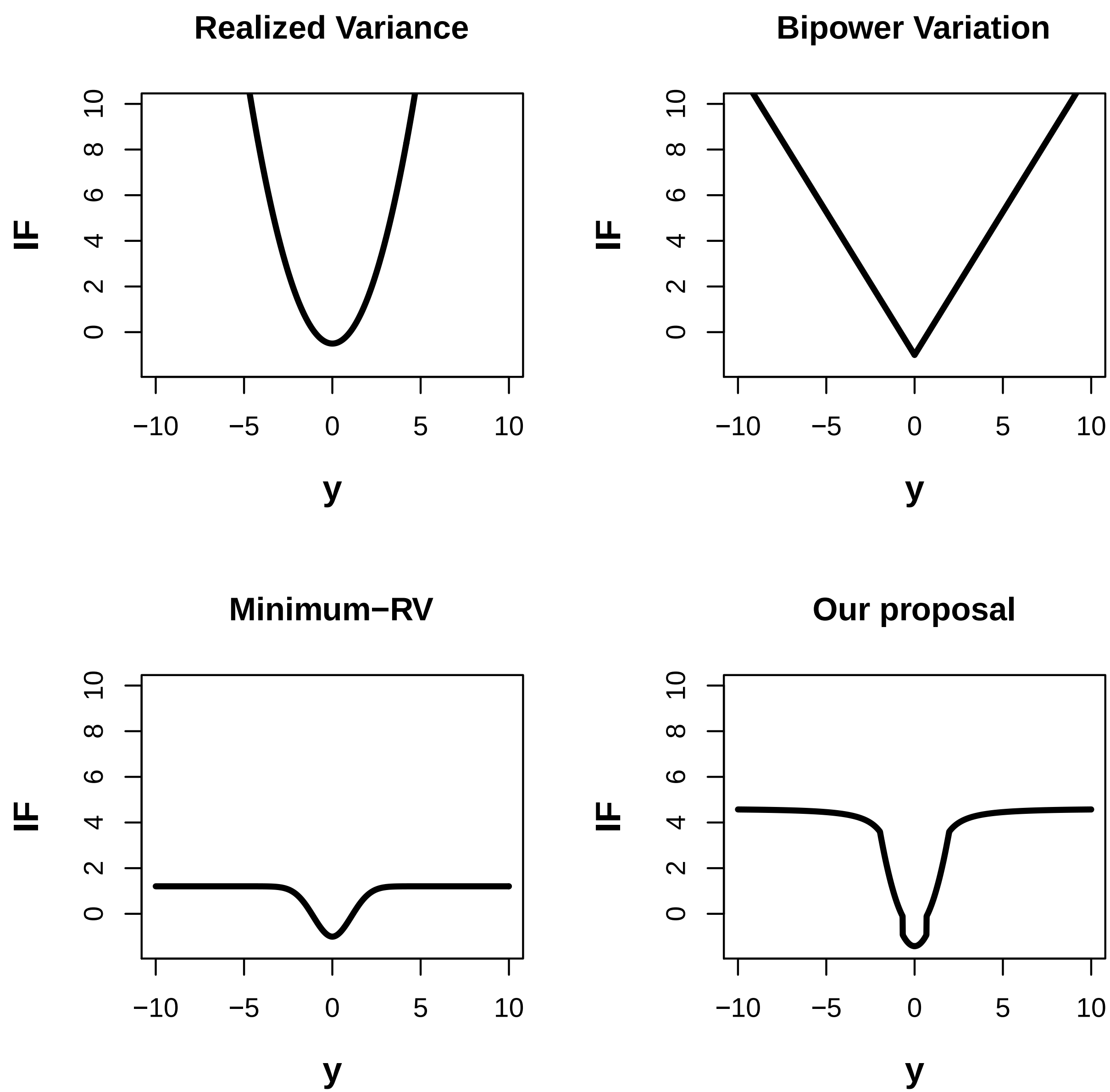
- RBPVar $= \frac{\pi}{2} \sum_i |r_i||r_{i-1}|$

  Breakdown point = 1/n

- MinRV $= \frac{\pi}{\pi-2} \sum_i \min(|r_i|, |r_{i-1}|)^2$

  Breakdown point = 2/n

Influence Functions



Bivariate case

If S^2 is an estimator of the variance, then

$$Cov(X,Y) = \frac{1}{4}\{S^2(X+Y) - S^2(X-Y)\}$$

provides an estimator of covariance.

Bivariate case

If S^2 is an estimator of the variance, then

$$Cov(X,Y) = \frac{1}{4}\{S^2(X+Y) - S^2(X-Y)\}$$

provides an estimator of covariance.

S^2 =Realized Bipower Variation $\longrightarrow$ Realized Bipower Covariation

If S^2 is an estimator of the variance, then

$$Cov(X,Y) = \frac{1}{4}\{S^2(X+Y) - S^2(X-Y)\}$$

provides an estimator of covariance.

S^2 =Realized Bipower Variation $\longrightarrow$ Realized Bipower Covariation

Matrix of pairwise covariances can be constructed from S, but this will not result in a positive definite matrix.

Let c be a threshold value.

If $|r_{i,1}| < c$ and $|r_{i,2}| < c$, then r_i is below the treshold.

Compute $\sum_i r_i r_i'$ but only over the r_i below the threshold.

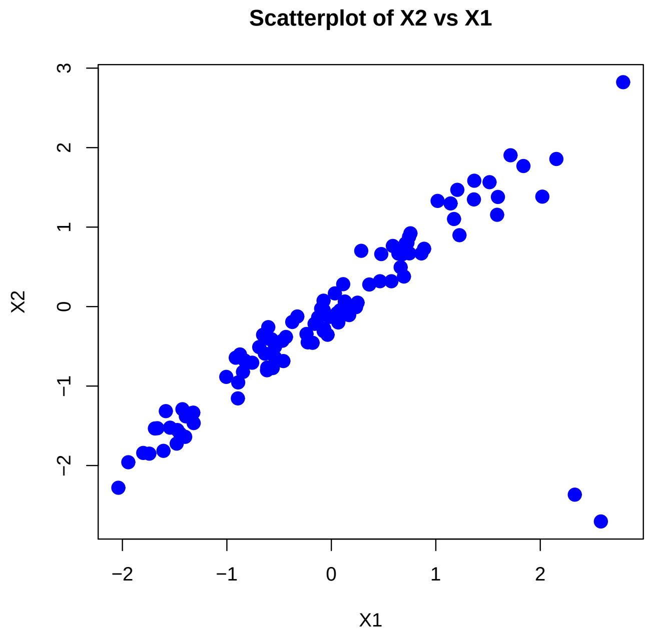
Gobbi and Mancini (2008)

Let c be a threshold value.

If $|r_{i,1}| < c$ and $|r_{i,2}| < c$, then r_i is below the treshold.

Compute $\sum_i r_i r_i'$ but only over the r_i below the threshold.

Not affine equivariant. Cannot detect correlation outliers

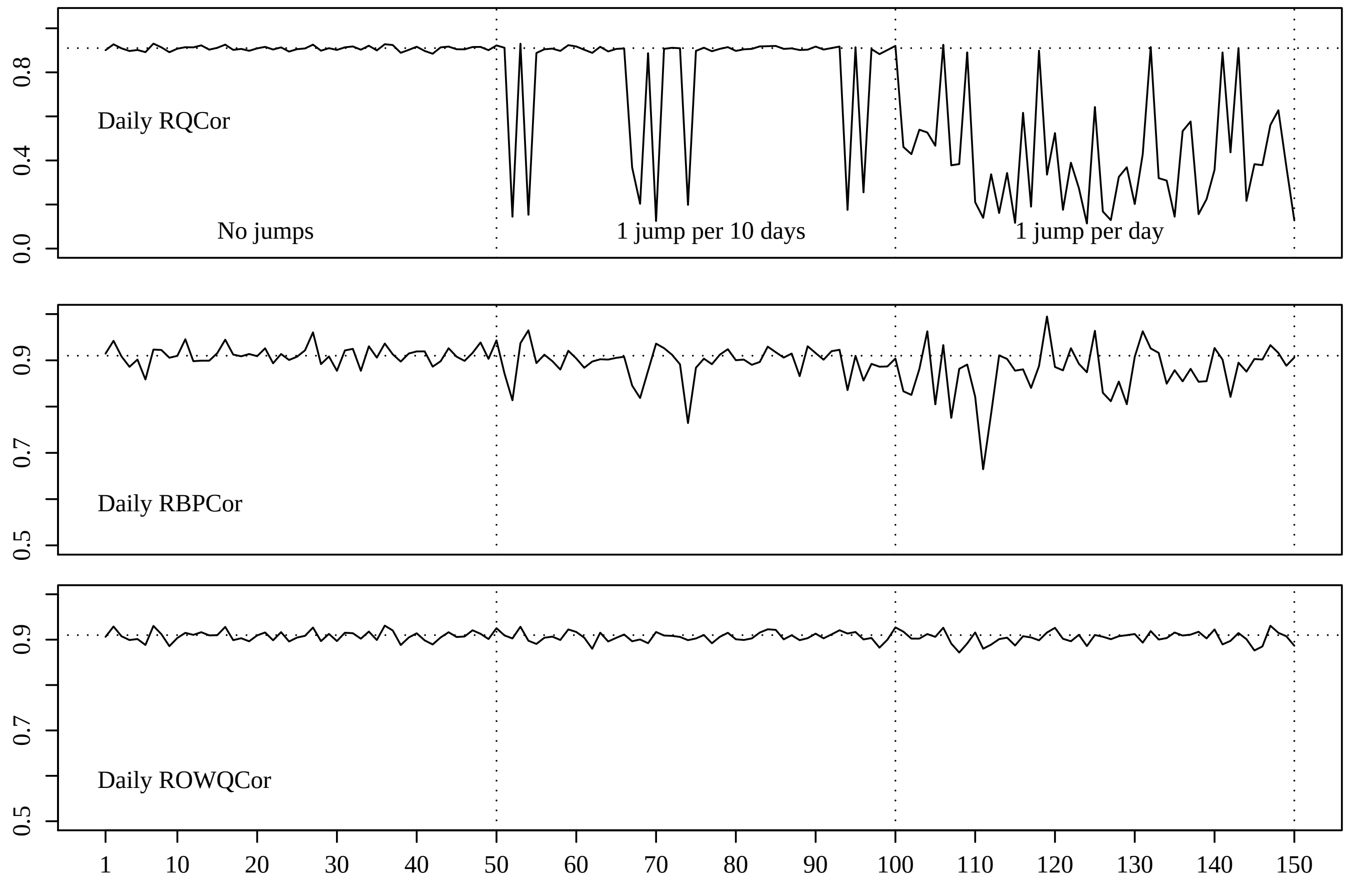


Simulation

Bivariate process $(p^{(1)}(s), p^{(2)}(s))$ where the log-returns have a constant correlation equal to 0.91. We compare

- Realized Quadratic Covariation (RQCov)
- Realized Bipower Covariation (RBPCov)
- Realized Outyingness Weighted Quadratic Covariation (ROWQCOV)

Bivariate process $(p^{(1)}(s), p^{(2)}(s))$ where the log-returns have a constant correlation equal to 0.91. Daily correlations over 150 days.



Root Mean Squared Errors

Jumps per day κ^*:	0	1	1
Magnitude of jumps m:		0.5	1
5-minute returns $(\Delta = 1/288)$			
RQCov	**0.084**	1.413	2.804
RBPCov	0.096	0.242	0.339
ROWQCov	0.089	**0.088**	**0.091**

Conclusions

Limitations

Robustness to market micro-structure noise. Furthermore:

- Intraday periodicity in volatility
- Infrequent trading (zero returns)
- Nonsynchronous trading
- Large dimensions, small sample size
- Forecasting models

Boudt et al, Janus et al, Schulz-Mosler, Haysashi-Yoshida, Cornelissen et al, Corsi et al. *All at CFE 09.*

Conclusions

We propose an estimator of the continuous component of the volatility of a multivariate price process in presence of jumps.

- Consistent and highly efficient
- Robust to jumps
- Affine equivariant and positive definite
- Implemented in the Ox package GARCH (www.garch.org). R code is also available.

www.econ.kuleuven.be/christophe.croux/public