Estimation of Extreme Risk Regions Under Multivariate Regular Variation

Juan-Juan Cai[1] John H. J. Einmahl[2] Laurens de Haan[3]

[1]Technology University of Delft

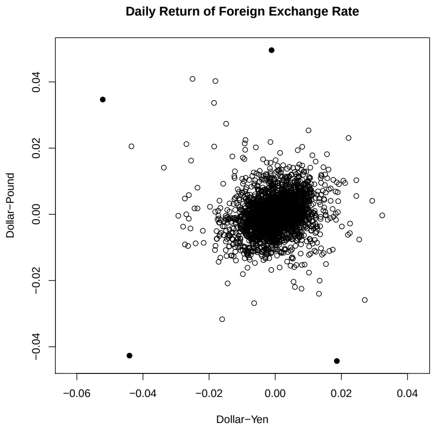
[2]Tilburg University

[3]University of Lisbon and Erasmus University Rotterdam

The Hague, May 22, 2013

An Example : which event is rarest?

- Returns of daily exchange rates of yen-dollar and pound-dollar from 1999 to 2009.
- Which one is the most extreme among those marked with solid circles?



- Let $\mathbf{Z}$ be a random vector on $\mathbb{R}^d$ $(d \geq 2)$.
- A risk region is a set Q such that $\mathbb{P}(\mathbf{Z} \in Q) = p$, *extremely* small.

- Let $\mathbf{Z}$ be a random vector on $\mathbb{R}^d$ $(d \geq 2)$.
- A risk region is a set Q such that $\mathbb{P}(\mathbf{Z} \in Q) = p$, *extremely* small.
- Events in Q hardly happen. The interest of these events originates from their potential large consequences.

- Let $\mathbf{Z}$ be a random vector on $\mathbb{R}^d$ $(d \geq 2)$.
- A risk region is a set Q such that $\mathbb{P}(\mathbf{Z} \in Q) = p$, *extremely* small.
- Events in Q hardly happen. The interest of these events originates from their potential large consequences.

- Suppose $\mathbf{Z}$ has probability density f. Denote the corresponding probability measure with P.
- The risk regions of interest are defined in this form:

$$Q = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta\},$$

  where β is an unknown number such that $PQ = p$.
- $Q^c = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) > \beta\}$.
- Q is the set of less likely points.

- The goal is to estimate Q based on a random sample from $\mathbf{Z}$. The sample size is n.
- For asymptotics, we consider $p = p(n) \to 0$, as $n \to \infty$.
- We write:

$$Q_n = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta_n\}.$$

Multivariate Regular Variation

There exist a positive number α and a positive function q, such that

$$\lim_{t\to\infty} \frac{\mathbb{P}(\| \mathbf{Z} \| > tx)}{\mathbb{P}(\| \mathbf{Z} \| > t)} = x^{-\alpha}, \qquad \text{for all } x > 0,$$

and

$$\lim_{t\to\infty} \frac{f(t\mathbf{z})}{t^{-d}\mathbb{P}(\| \mathbf{Z} \| > t)} = q(\mathbf{z}), \qquad \text{for all } \mathbf{z} \neq 0,$$

where $\| \cdot \|$ denotes the L_2 norm.

Multivariate Regular Variation

There exist a positive number α and a positive function q, such that

$$\lim_{t\to\infty} \frac{\mathbb{P}(\| \mathbf{Z} \| > tx)}{\mathbb{P}(\| \mathbf{Z} \| > t)} = x^{-\alpha}, \qquad \text{for all } x > 0,$$

and

$$\lim_{t\to\infty} \frac{f(t\mathbf{z})}{t^{-d}\mathbb{P}(\| \mathbf{Z} \| > t)} = q(\mathbf{z}), \qquad \text{for all } \mathbf{z} \neq 0,$$

where $\| \cdot \|$ denotes the L_2 norm.

1. Examples: Cauchy distributions and all elliptical distributions with a heavy tailed radius.

Some results from the assumption

- The distribution of the radius has a right heavy tail. α is the tail index.
- q is homogenous: $q(a\mathbf{z}) = a^{-d-\alpha} q(\mathbf{z})$.
- Define $\nu(B) = \int_B q(\mathbf{z}) d\mathbf{z}$. Then, for a Borel set B with positive distance from the origin,

$$\lim_{t \to \infty} \frac{\mathbb{P}(\mathbf{Z} \in tB)}{\mathbb{P}(\| \mathbf{Z} \| \geq t)} = \nu(B).$$

- Recall that we try to estimate

$$Q_n = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta_n\},$$

such that $\mathbb{P}(\mathbf{Z} \in Q_n) = p$.

- Recall that we try to estimate

$$Q_n = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta_n\},$$

such that $\mathbb{P}(\mathbf{Z} \in Q_n) = p$.

- Link Q_n to $S = \{\mathbf{z} \in \mathbb{R}^d : q(\mathbf{z}) \leq 1\}$.

- Recall that we try to estimate

$$Q_n = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta_n\},$$

  such that $\mathbb{P}(\mathbf{Z} \in Q_n) = p$.
- Link Q_n to $S = \{\mathbf{z} \in \mathbb{R}^d : q(\mathbf{z}) \leq 1\}$.
- Inflate S with the factor u_n: $\bar{Q}_n := u_n S$, where u_n is such that $\mathbb{P}(||\mathbf{Z}|| > u_n) = \frac{\nu(S)}{p}$.

- Recall that we try to estimate

$$Q_n = \{\mathbf{z} \in \mathbb{R}^d : f(\mathbf{z}) \leq \beta_n\},$$

  such that $\mathbb{P}(\mathbf{Z} \in Q_n) = p$.
- Link Q_n to $S = \{\mathbf{z} \in \mathbb{R}^d : q(\mathbf{z}) \leq 1\}$.
- Inflate S with the factor u_n: $\bar{Q}_n := u_n S$, where u_n is such that $\mathbb{P}(||\mathbf{Z}|| > u_n) = \frac{\nu(S)}{p}$.
- $\bar{Q}_n$ is a good approximation of Q_n.

- Suppose we have $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ i.i.d copies of $\mathbf{Z}$.
- Write $R_i = \| \mathbf{Z}_i \|$ and $\mathbf{W}_i = \frac{\mathbf{Z}_i}{R_i}$, $i = 1, 2, \ldots, n$.
- Put $\Theta := \{\mathbf{z} : \| \mathbf{z} \| = 1\}$. Then $\mathbf{W}_i \in \Theta, i = 1, 2, \ldots, n$.

Estimation of u_n

- Note that u_n is the tail quantile of R_1: $\mathbb{P}(R_1 > u_n) = \frac{\nu(S)}{p}$.
- Suppose that we know $\nu(S)$. Applying the univariate extreme value technique, we define the estimator given by

$$\hat{u}_n = R_{n-k,n} \left(\frac{k\nu(S)}{np} \right)^{1/\hat{\alpha}},$$

where $k = k(n)$ such that $k \to \infty$ and $k/n \to 0$, as $n \to \infty$ and $R_{n-k,n}$ is the $(n-k)$-th order statistics of $\{R_i, i = 1, \ldots, n\}$.

Estimation of u_n

- Note that u_n is the tail quantile of R_1: $\mathbb{P}(R_1 > u_n) = \frac{\nu(S)}{p}$.
- Suppose that we know $\nu(S)$. Applying the univariate extreme value technique, we define the estimator given by

$$\hat{u}_n = R_{n-k,n} \left(\frac{k\nu(S)}{np} \right)^{1/\hat{\alpha}},$$

  where $k = k(n)$ such that $k \to \infty$ and $k/n \to 0$, as $n \to \infty$ and $R_{n-k,n}$ is the $(n-k)$-th order statistics of $\{R_i, i = 1, \ldots, n\}$.
- We need to estimate $\nu(S)$. It is sufficient to estimate q, the density of ν.

Estimation of q

- For a Borel set $A \in \Theta$, $\lim_{t\to\infty} \mathbb{P}(W_1 \in A | R_1 > t) =: \Psi(A)$ exists.
- The density of Ψ, $\psi(\mathbf{w}) = \frac{1}{\alpha} q(\mathbf{w})$, $\mathbf{w} \in \Theta$.
- The estimation of ψ is based on $W_{(i)}$, where the corresponding radius $R_{(i)} > R_{n-k,n}$.
- We propose a kernel density estimator $\hat{\psi}$.
- Then $\hat{q} = \hat{\alpha}\hat{\psi}$. The estimation of S and $\nu(S)$ follow directly.

We obtain our estimator:

$$\widehat{Q}_n = \hat{u}_n \widehat{S} = R_{n-k,n} \left(\frac{k\widehat{\nu(s)}}{np} \right)^{1/\hat{\alpha}} \{\mathbf{z} : \hat{q}(\mathbf{z}) < 1\}.$$

We obtain our estimator:

$$\widehat{Q}_n = \hat{u}_n \widehat{S} = R_{n-k,n} \left(\frac{k\widehat{\nu(s)}}{np} \right)^{1/\hat{\alpha}} \{\mathbf{z} : \hat{q}(\mathbf{z}) < 1\}.$$

Theorem

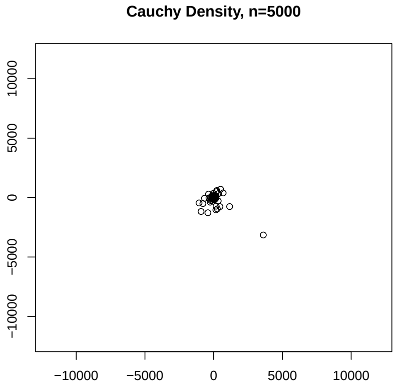
Under some regular conditions, we have, as $n \to \infty$,

$$\frac{P\left(\hat{Q}_n \triangle Q_n\right)}{p} \xrightarrow{\mathbb{P}} 0,$$

Here $\triangle$ denotes the symmetric difference. $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

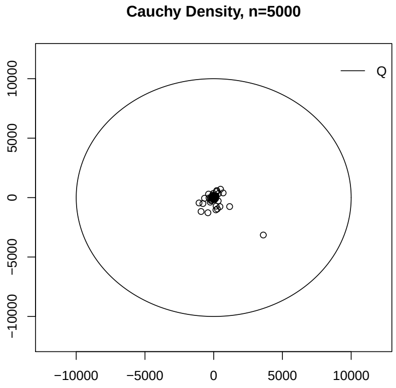
Bivariate Cauchy Distribution

Bivariate Cauchy Distribution



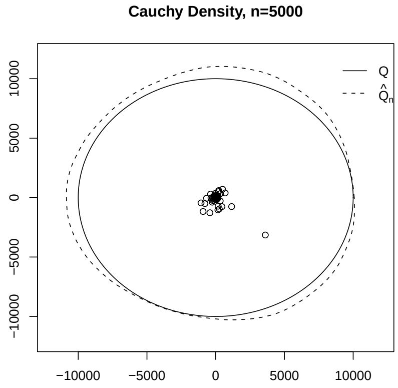
- Data are simulated from the bivariate Cauchy distribution. $n = 5000$.

Bivariate Cauchy Distribution



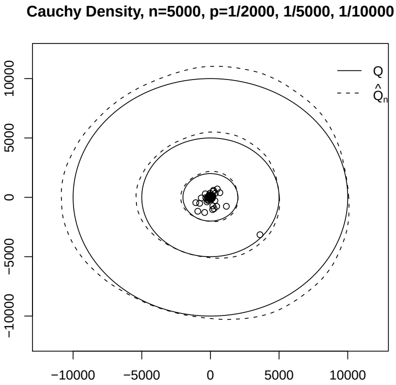
- Data are simulated from the bivariate Cauchy distribution. $n = 5000$.
- The area outside the solid line is the true risk region. $PQ = 10^{-4}$.

Bivariate Cauchy Distribution



- Data are simulated from the bivariate Cauchy distribution. $n = 5000$.
- The area outside the solid line is the true risk region. $PQ = 10^{-4}$.
- The area outside the dotted curve corresponds to the estimated risk region.

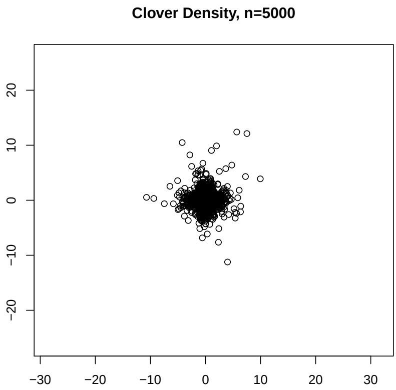
Bivariate Cauchy Distribution



- Data are simulated from the bivariate Cauchy distribution. $n = 5000$.
- The area outside the solid line is the true risk region. $PQ = 10^{-4}$.
- The area outside the dotted curve corresponds to the estimated risk region.

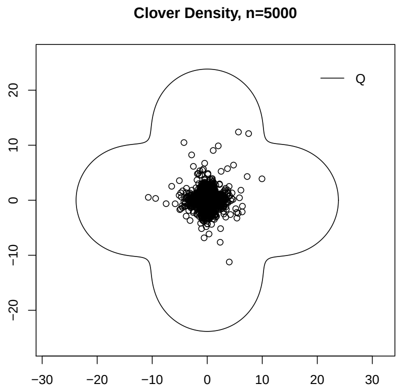
Clover Density

Clover Density



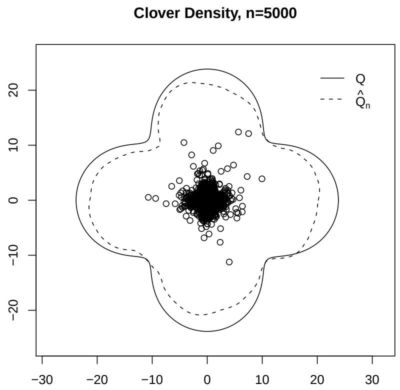
- $n = 5000.$

Clover Density

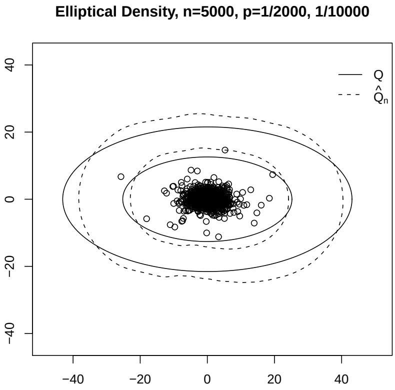


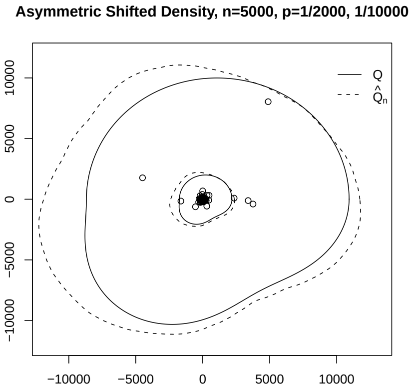
- $n = 5000$.
- The area outside the solid line is the true risk region, Q. $PQ = 10^{-4}$.

Clover Density



- $n = 5000$.
- The area outside the solid line is the true risk region, Q. $PQ = 10^{-4}$.
- The area outside the dotted curve corresponds to the estimated risk region.


Elliptical Density, n=5000, p=1/2000, 1/10000
Q
$\hat{Q}_n$
40
20
0
−20
−40
−40
−20
0
20
40


Asymmetric Shifted Density, n=5000, p=1/2000, 1/10000
Q
$\hat{Q}_n$
−10000
−5000
0
5000
10000
−10000
−5000
0
5000
10000

Competitor I

A "Parametric" estimator

- The method works for bivariate distributions only.
- Estimate $\nu(S)$ and S by assuming a parametric model to q:
$q(\mathbf{w}) = q(\cos\theta, \sin\theta) = \alpha(4\pi)^{-1}(2 + \sin(2(\theta - \rho)))$, $\theta \in [0, 2\pi]$.

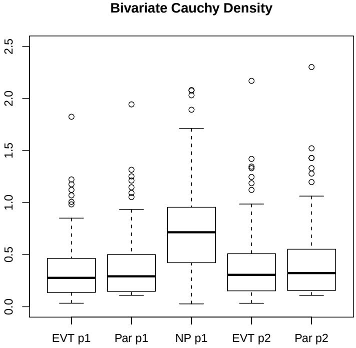
Competitor II

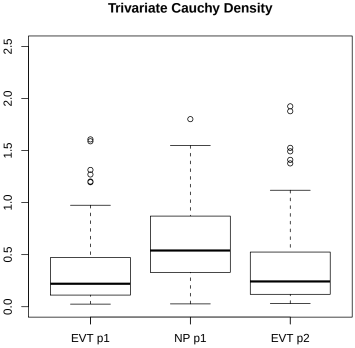
A non-parametric estimator

- Compute the smallest ellipsoid containing half of the data, the so-called MVE.
- Inflate this ellipsoid such that largest observation lies on its boundary.
- It works for $p = 1/n$ only.

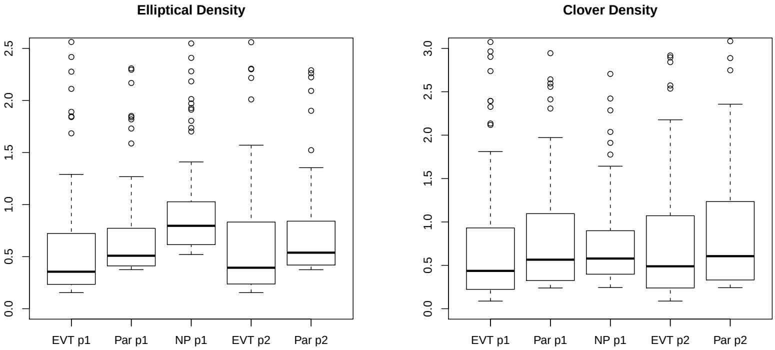
- We simulate 100 data sets from four bivariate distributions and the trivariate Cauchy distribution. Each data set is of size 5000.
- The main theorem states that $\frac{P(\hat{Q}_n \triangle Q_n)}{p} \xrightarrow{\mathbb{P}} 0$.

- $e_{evt} = \frac{P(\hat{Q}_n \triangle Q_n)}{p}$, $p_1 = 1/5000$ and $p_2 = 1/10000$.
- $e_{np} = \frac{P(\widehat{Q}_{np} \triangle Q_n)}{p}$, $p_1 = 1/5000$.
- $e_{par} = \frac{P(\widehat{Q}_{par} \triangle Q_n)}{p}$, $p_1 = 1/5000$ and $p_2 = 1/10000$.

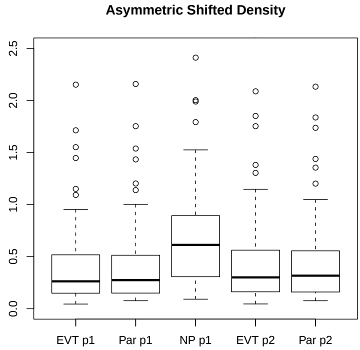




- $e_{evt} = \frac{P(\hat{Q}_n \triangle Q_n)}{p}$, $p_1 = 1/5000$ and $p_2 = 1/10000$.
- $e_{np} = \frac{P(\widehat{Q}_{np} \triangle Q_n)}{p}$, $p_1 = 1/5000$.
- $e_{par} = \frac{P(\widehat{Q}_{par} \triangle Q_n)}{p}$, $p_1 = 1/5000$ and $p_2 = 1/10000$.



- $e_{evt} = \dfrac{P(\hat{Q}_n \triangle Q_n)}{p}$, $p_1 = 1/5000$ and $p_2 = 1/10000$.
- $e_{np} = \dfrac{P(\widehat{Q}_{np} \triangle Q_n)}{p}$, $p_1 = 1/5000$.
- $e_{par} = \dfrac{P(\widehat{Q}_{par} \triangle Q_n)}{p}$, $p_1 = 1/5000$ and $p_2 = 1/10000$.

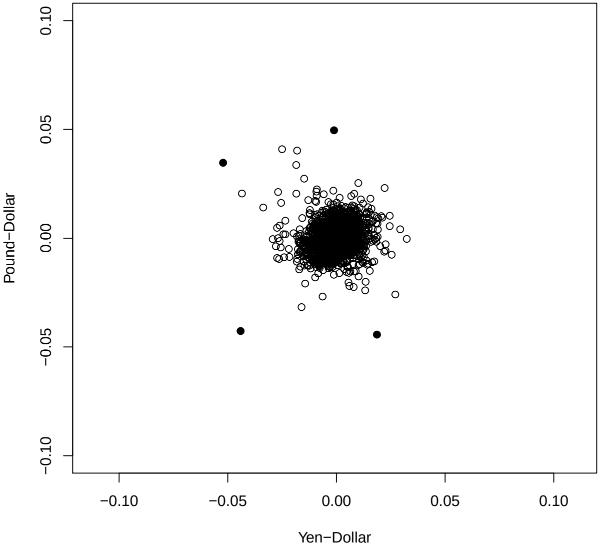


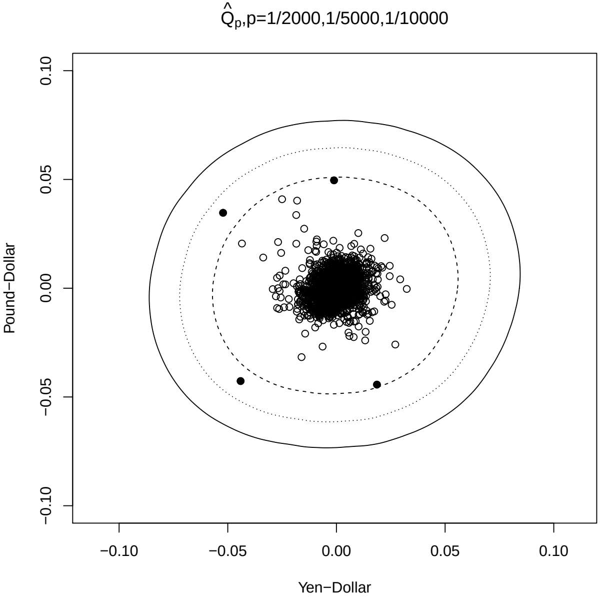
We apply our method to foreign exchange rate data.

- Data: daily exchange rates of yen-dollar and pound-dollar, dating from 4 Jan 1999 to 31 July 2009. $n = 2665$.
- We consider the log-return.

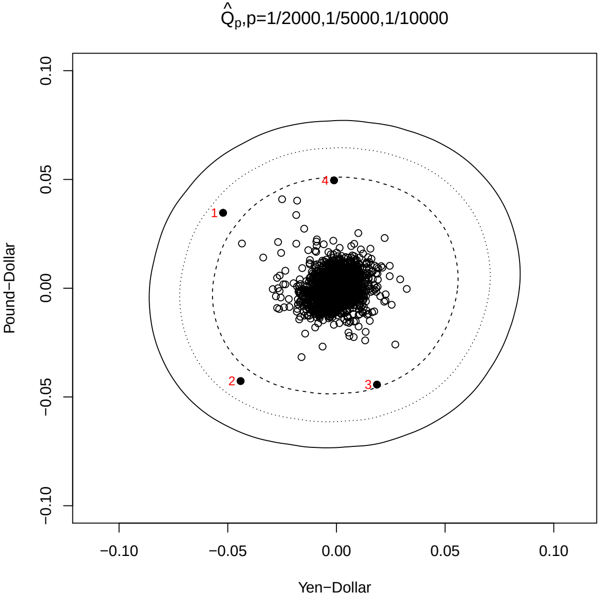
$$X_{t,i} = \log \frac{Y_{t,i}}{Y_{t-1,i}},$$

where $t = 1, \ldots, 2664$, $i = 1, 2$ and $Y_{t,1}$ is the daily exchange rate of yen-dollar and $Y_{t,2}$ pound-dollar.


0.10
0.05
0.00
−0.05
−0.10
Pound−Dollar
−0.10
−0.05
0.00
0.05
0.10
Yen−Dollar



$\hat{Q}_p$,p=1/2000,1/5000,1/10000
Pound−Dollar
0.10
0.05
0.00
−0.05
−0.10
Yen−Dollar
−0.10
−0.05
0.00
0.05
0.10


$\hat{Q}_p$,p=1/2000,1/5000,1/10000
Pound−Dollar
Yen−Dollar
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
1
2
3
4

Thank you very much for your attention!