Estimation of Extreme Risk Regions Under Multivariate Regular Variation

Juan-Juan Cai\textsuperscript{1} John H. J. Einmahl\textsuperscript{2} Laurens de Haan\textsuperscript{3}

\textsuperscript{1}Technology University of Delft
\textsuperscript{2}Tilburg University
\textsuperscript{3}University of Lisbon and Erasmus University Rotterdam

University of Strasbourg, December 2014
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 \( Z \) be a random vector on \( \mathbb{R}^d \) \((d \geq 2)\).

A risk region is a set \( Q \) such that \( \mathbb{P}(Z \in Q) = p \), extremely small.

Events in \( Q \) hardly happen. The interest of these events originates from their potential large consequences.
Let $Z$ be a random vector on $\mathbb{R}^d$ ($d \geq 2$).

A risk region is a set $Q$ such that $\mathbb{P}(Z \in Q) = p$, extremely small.

Events in $Q$ hardly happen. The interest of these events originates from their potential large consequences.
Suppose \( Z \) has probability density \( f \). Denote the corresponding probability measure with \( P \).

The risk regions of interest are defined in this form:

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

where \( \beta \) is an unknown number such that \( PQ = p \).
Suppose $Z$ has probability density $f$. Denote the corresponding probability measure with $P$.

The risk regions of interest are defined in this form:

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

where $\beta$ is an unknown number such that $PQ = p$.

$$Q^c = \{z \in \mathbb{R}^d : f(z) > \beta\}.$$

$Q$ is the set of less likely points.
The goal is to estimate $Q$ based on a random sample from $Z$. The sample size is $n$.

For asymptotics, we consider $p = p(n) \rightarrow 0$, as $n \rightarrow \infty$.

We write:

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

Multivariate Regular Variation

There exist a positive number $\alpha$ and a positive function $q$, such that

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

and

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

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

Multivariate Regular Variation

There exist a positive number $\alpha$ and a positive function $q$, such that

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

and

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

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

1. Examples: Cauchy distributions and all elliptical distributions with a heavy tailed radius.
The distribution of the radius has a right heavy tail. $\alpha$ is the tail index.

$q$ is homogenous: $q(az) = a^{-d-\alpha}q(z)$.

Define $\nu(B) = \int_B q(z)dz$. Then, for a Borel set $B$ with positive distance from the origin,

$$\lim_{t \to \infty} \frac{\mathbb{P}(Z \in tB)}{\mathbb{P}(\|Z\| \geq t)} = \nu(B).$$
Recall that we try to estimate 

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

such that $$\mathbb{P}(Z \in Q_n) = p.$$
Recall that we try to estimate

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

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

Link \( Q_n \) to \( S = \{ z \in \mathbb{R}^d : q(z) \leq 1 \} \).
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 \( u_n \): \( \tilde{Q}_n := u_n S \), 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 \( u_n : \tilde{Q}_n := u_n S \), where \( u_n \) is such that \( \mathbb{P}(\|\mathbf{Z}\| > u_n) = \frac{\nu(S)}{p} \).
Main Result

$\tilde{Q}_n$ is a good approximation of $Q_n$. We show that as $n \to \infty$, 

$$\frac{P(Q_n \Delta \tilde{Q}_n)}{p} \to 0,$$

where $\Delta$ denotes the symmetric difference. $A \Delta B = (A \setminus B) \cup (B \setminus A)$.

To estimate $Q_n$ is now to estimate $\tilde{Q}_n$. 

Estimation
Main Result

Estimation

Estimation of $\tilde{Q}_n = u_n S$

- Suppose we have $Z_1, \ldots, Z_n$ i.i.d copies of $Z$.
- Write $R_i = \| Z_i \|$ and $W_i = \frac{Z_i}{R_i}$, $i = 1, 2, \ldots, n$.
- Put $\Theta := \{ z : \| z \| = 1 \}$. Then $W_i \in \Theta$, $i = 1, 2, \ldots, n$. 
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 $\{R_i, i = 1, \ldots, n\}$. 

Cai, Einmahl, de Haan
Multivariate Extreme Risk Regions

10 / 25
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 $q$. 

Estimation of $q$

- For a Borel set $A \in \Theta$, $\lim_{t \to \infty} P(W_1 \in A | R_1 > t) =: \Psi(A)$ exists.
- The density of $\Psi$ exists: $\psi(w) = \frac{1}{\alpha} q(w), w \in \Theta$.
- We propose a kernel density estimator of $\psi$ making use of observations with big radius.
Estimation of $q$

- For a Borel set $A \in \Theta$, $\lim_{t \to \infty} P(W_1 \in A | R_1 > t) =: \Psi(A)$ exists.
- The density of $\Psi$ exists: $\psi(w) = \frac{1}{\alpha} q(w)$, $w \in \Theta$.
- We propose a kernel density estimator of $\psi$ making use of observations with big radius.
- As $n \to \infty$,

$$\sup_{w \in \Theta} \left| \hat{\psi}(w) - \psi(w) \right| \overset{P}{\to} 0.$$
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 $\Psi$ exists: $\psi(w) = \frac{1}{\alpha} q(w)$, $w \in \Theta$.
- We propose a kernel density estimator of $\psi$ making use of observations with big radius.
- As $n \to \infty$,
  \[
  \sup_{w \in \Theta} \left| \hat{\psi}(w) - \psi(w) \right| \overset{\mathbb{P}}{\to} 0.
  \]
- Then $\hat{q} = \hat{\alpha} \hat{\psi}$. The estimations of $S$ and $\nu(S)$ follow directly.
Main Result

Estimation

We obtain our estimator:

\[ \hat{Q}_n = \hat{u}_n \hat{S} = R_{n-k,n} \left( \frac{k \nu(S)}{np} \right)^{1/\alpha} \{ z : \hat{q}(z) < 1 \} . \]
We obtain our estimator:

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

**Theorem**

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

\[ P \left( \frac{\hat{Q}_n \triangle Q_n}{p} \right) \overset{\mathbb{P}}{\to} 0, \]

Here \( \triangle \) denotes the symmetric difference.
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. \( n = 5000 \).
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

Clover Density, n=5000

$n = 5000$. 
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
Asymmetric Shifted Density, n=5000, p=1/2000, 1/10000
Two competitors

A “Parametric” estimator

- Estimate $\nu(S)$ and $S$ by assuming
  $\psi(w_1, w_2) = \psi(\cos \theta, \sin \theta) = (4\pi)^{-1}(2 + \sin(2(\theta - \rho))))$, $\theta \in [0, 2\pi]$.
- The method works for bivariate distributions only.

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 \( \frac{P(\hat{Q}_n \triangle Q_n)}{p} \to 0 \).
\[ \tilde{e}_{evt} = \frac{P(\hat{Q}_n \triangle Q_n)}{p}, \quad p_1 = \frac{1}{5000} \text{ and } p_2 = \frac{1}{10000}. \]
\[ \tilde{e}_{np} = \frac{P(\hat{Q}_{np} \triangle Q_n)}{p}, \quad p_1 = \frac{1}{5000}. \]
\[ \tilde{e}_{par} = \frac{P(\hat{Q}_{par} \triangle Q_n)}{p}, \quad p_1 = \frac{1}{5000} \text{ and } p_2 = \frac{1}{10000}. \]
\[ e_{evt} = \frac{P(\hat{Q}_n \triangle Q_n)}{p}, \quad p_1 = 1/5000 \text{ and } p_2 = 1/10000. \]
\[ e_{np} = \frac{P(\hat{Q}_{np} \triangle Q_n)}{p}, \quad p_1 = 1/5000. \]
\[ e_{par} = \frac{P(\hat{Q}_{par} \triangle Q_n)}{p}, \quad p_1 = 1/5000 \text{ and } p_2 = 1/10000. \]
\[ e_{evt} = \frac{P(\hat{Q}_n \triangle Q_n)}{p}, \quad p_1 = \frac{1}{5000} \text{ and } p_2 = \frac{1}{10000}. \]

\[ e_{np} = \frac{P(\hat{Q}_{np} \triangle Q_n)}{p}, \quad p_1 = \frac{1}{5000}. \]

\[ e_{par} = \frac{P(\hat{Q}_{par} \triangle Q_n)}{p}, \quad p_1 = \frac{1}{5000} \text{ and } p_2 = \frac{1}{10000}. \]
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.
\[ \hat{\alpha} = 3.9 \]
\( \hat{\alpha} = 3.9 \)
\( \hat{\alpha} = 3.9 \)
\( \hat{\alpha} = 3.9 \)
Ordering multivariate extreme observations

- For an extreme observation $Z_i$, we associate a $p$-value given by

$$H(Z_i) = P(z : f(z) \leq f(Z_i)).$$

- Estimation of $H(Z_i)$ follows easily from the current procedure.
- Order multivariate extremes accordingly.
Outlier detection

An issue: It is not clear how to formulate an alternative hypothesis.