Aalto University

On extreme quantile region estimation under heavy-tailed elliptical distributions

To be published in the *Journal of Multivariate Analysis*

Jaakko Pere
Co-authors: Pauliina Ilmonen and Lauri Viitasaari

March 25, 2024

Agenda of the presentation

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

Table of Contents

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

Multivariate extreme quantile region estimation

Consider an m-dimensional random variable $X : \Omega \to \mathbb{R}^m$ with density f. Estimate multivariate quantile region of the form

$$Q_p = \{x \in \mathbb{R}^m : f(x) \leq \beta\},$$

where β is chosen such that $\mathbb{P}(X \in Q_p) = p$ and p is very small. See (Cai, Einmahl, and Laurens De Haan 2011).

Extreme value theory (EVT)

- Provides rigorous mathematical framework to infer about behavior of extremes (extrapolate outside the sample).

Extreme value theory (EVT)

- Provides rigorous mathematical framework to infer about behavior of extremes (extrapolate outside the sample).
- Extreme quantile estimation, estimating the probability of extreme event, endpoint estimation of a given distribution, etc.

Problems posed by multidimensionality

- How to measure centrality of an observation?

Table of Contents

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

Left-continuous inverse

Left-continuous inverse of a nondecreasing function g is given by

$$g^{\leftarrow}(y) = \inf\{x : g(x) \geq y\}.$$

Left-continuous inverse

Left-continuous inverse of a nondecreasing function g is given by

$$g^{\leftarrow}(y) = \inf\{x : g(x) \geq y\}.$$

Quantile function of a distribution F is defined as $F^{\leftarrow}(p)$.

Problem formulation in one-dimensional case

Let $X : \Omega \to \mathbb{R}$ be a univariate random variable with cumulative distribution function F. We wish to estimate $(1-p)$-quantile

$$x_p = F^{\leftarrow}(1-p),$$

when p is very small.

Tail quantile function U

Define tail quantile function $U : (1, \infty) \mapsto \mathbb{R}$ corresponding to distribution F by

$$U(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right).$$

Tail quantile function U

Define tail quantile function $U : (1, \infty) \mapsto \mathbb{R}$ corresponding to distribution F by

$$U(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right).$$

Notice that $U(1/p)$ is the $(1-p)$-quantile of the distribution F. That is, $x_p = U(1/p)$.

Heavy-tailed distributions

Definition

Distribution F is heavy-tailed with an extreme value index $\gamma > 0$ if the corresponding tail quantile function U is regularly varying with the index γ, i.e.,

$$\lim_{t \to \infty} \frac{U(tx)}{U(t)} = x^{\gamma},$$

for all $x > 0$. Notation: $U \in RV_{\gamma}$, $F \in RV_{\gamma}$, $X \in RV_{\gamma}$.

Extreme quantile estimator

Definition

Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from a heavy-tailed distribution F. Denote the order statistics by $X_{1,n} \leq \cdots \leq X_{n,n}$. Then we can define an estimator for the extreme $(1 - p_n)$-quantile $x_{p_n} = F^{\leftarrow}(1 - p_n)$ as

$$\hat{x}_{p_n} = X_{n-k_n,n} \left(\frac{k_n}{np_n} \right)^{\hat{\gamma}_n},$$

where $k_n \in \{1, 2, \ldots, n-1\}$ and $\hat{\gamma}_n$ is an estimator for the extreme value index γ.

See (L. De Haan and Ferreira 2006) for a review about extreme quantile estimation, estimation of extreme value index etc.

Table of Contents

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

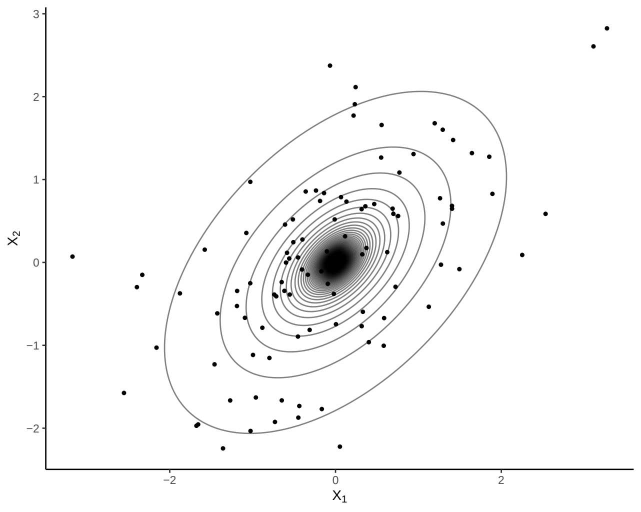

Figure: Sample of size 100 from a bivariate t-distribution with 10 degrees of freedom. Figure also includes some contours from the probability density function.

Elliptical distributions

Definition (Cambanis, Huang, and Simons 1981)

We say that a random variable $X : \Omega \to \mathbb{R}^m$ is elliptically distributed if

$$X \stackrel{d}{=} \mu + \mathcal{R}\Lambda\mathcal{U},$$

where

- $\mu \in \mathbb{R}^m$ is called the *location vector*;
- $\mathcal{R} : \Omega \to \mathbb{R}_{\geq 0}$ is a nonnegative random variable called the *generating variate*;
- $\Lambda \in \mathbb{R}^{m \times m}$ is a matrix such that $\Sigma = \Lambda\Lambda^\intercal$ is a symmetric positive definite matrix (matrix Σ is called the *scatter matrix*);
- $\mathcal{U}$ is an m-dimensional random vector uniformly distributed over the unit-sphere $\{x \in \mathbb{R}^m : x^\intercal x = 1\}$;
- Random variables $\mathcal{R}$ and $\mathcal{U}$ are independent.

Role of the generating variate $\mathcal{R}$

- $\mathcal{R}$ heavy-tailed $\iff$ X is heavy-tailed (Hult and Lindskog 2002, Theorem 4.3).

Role of the generating variate $\mathcal{R}$

- $\mathcal{R}$ heavy-tailed $\iff$ X is heavy-tailed (Hult and Lindskog 2002, Theorem 4.3).
- $\mathcal{R} \stackrel{d}{=} \sqrt{(X-\mu)^T \Sigma^{-1} (X-\mu)}$

Table of Contents

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

From now on assume that

- Random variable X is m-dimensional and elliptically distributed with the representation $X \stackrel{d}{=} \mu + \mathcal{R}\Lambda\mathcal{U}$.
- $\mathcal{R}$ is absolutely continuous and heavy-tailed.
- $X_1, \ldots, X_n$ are i.i.d. copies of X.

Mahalanobis distance

For $x \in \mathbb{R}^m$ define a norm induced by positive definite matrix $H \in \mathbb{R}^{m \times m}$,

$$\|x\|_H = \sqrt{x^\top H^{-1} x}.$$

Mahalanobis distance

For $x \in \mathbb{R}^m$ define a norm induced by positive definite matrix $H \in \mathbb{R}^{m \times m}$,

$$\|x\|_H = \sqrt{x^\top H^{-1} x}.$$

- Remember that $\|X - \mu\|_\Sigma \stackrel{d}{=} \mathcal{R}$.

Quantile regions

Under ellipticity, and under certain natural assumptions about density f, quantile regions Q_{p_n} can be expressed as

$$Q_{p_n} = \{x \in \mathbb{R}^m : \|x - \mu\|_\Sigma \geq r_{p_n}\},$$

where r_{p_n} is the $(1 - p_n)$-quantile of the generating variate $\mathcal{R}$.

Estimated Mahalanobis distance

Let $\hat{\mu}_n$ and $\hat{\Sigma}_n$ be estimators of location and scatter, respectively, such that

$$\sqrt{n}(\hat{\mu}_n - \mu) = O_{\mathbb{P}}(1) \quad \text{and} \quad \sqrt{n}(\hat{\Sigma}_n - \Sigma) = O_{\mathbb{P}}(1).$$

Define

$$\hat{R}_i = \|X_i - \hat{\mu}_n\|_{\hat{\Sigma}_n}.$$

Estimator

We wish to estimate Q_{p_n} for a very small p_n. Define estimator $\hat{Q}_{p_n}$ as

$$\hat{Q}_{p_n} = \{x \in \mathbb{R}^m : \|x - \hat{\mu}_n\|_{\hat{\Sigma}_n} \geq \hat{r}_{p_n}\},$$

where

$$\hat{r}_{p_n} = \hat{R}_{n-k_n,n} \left(\frac{k_n}{np_n} \right)^{\hat{\gamma}_n},$$

and $\hat{\gamma}_n$ is the separating Hill estimator (Heikkilä, Dominicy, and Ilmonen 2019).

Table of Contents

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

Formulation

- For example, in multivariate density estimation it is common to define consistency through probability mass in symmetric difference (Tsybakov 1997).

Formulation

- For example, in multivariate density estimation it is common to define consistency through probability mass in symmetric difference (Tsybakov 1997).
- Let $S, T \subset \mathbb{R}^m$. Then $S \triangle T = (S \setminus T) \cup (T \setminus S)$.

Formulation

- For example, in multivariate density estimation it is common to define consistency through probability mass in symmetric difference (Tsybakov 1997).
- Let $S, T \subset \mathbb{R}^m$. Then $S \triangle T = (S \setminus T) \cup (T \setminus S)$.
- Consider probability space $(\mathbb{R}^m, \mathcal{B}, P)$, where $\mathcal{B}$ is the Borel σ-algebra and P is the law of X. Then $P(S \triangle T) : \mathcal{B} \times \mathcal{B} \mapsto [0, 1]$ is a pseudometric in $\mathcal{B}$.

Formulation

- For example, in multivariate density estimation it is common to define consistency through probability mass in symmetric difference (Tsybakov 1997).
- Let $S, T \subset \mathbb{R}^m$. Then $S \triangle T = (S \setminus T) \cup (T \setminus S)$.
- Consider probability space $(\mathbb{R}^m, \mathcal{B}, P)$, where $\mathcal{B}$ is the Borel σ-algebra and P is the law of X. Then $P(S \triangle T) : \mathcal{B} \times \mathcal{B} \mapsto [0, 1]$ is a pseudometric in $\mathcal{B}$.
- So can we say that $\hat{Q}_{p_n}$ is consistent estimator of Q_{p_n}, if

$$\mathbb{P}(X \in \hat{Q}_{p_n} \triangle Q_{p_n}) \overset{\mathbb{P}}{\to} 0, \quad n \to \infty?$$

Formulation

- For example, in multivariate density estimation it is common to define consistency through probability mass in symmetric difference (Tsybakov 1997).
- Let $S, T \subset \mathbb{R}^m$. Then $S \triangle T = (S \setminus T) \cup (T \setminus S)$.
- Consider probability space $(\mathbb{R}^m, \mathcal{B}, P)$, where $\mathcal{B}$ is the Borel σ-algebra and P is the law of X. Then $P(S \triangle T) : \mathcal{B} \times \mathcal{B} \mapsto [0, 1]$ is a pseudometric in $\mathcal{B}$.
- So can we say that $\hat{Q}_{p_n}$ is consistent estimator of Q_{p_n}, if

$$\mathbb{P}(X \in \hat{Q}_{p_n} \triangle Q_{p_n}) \xrightarrow{\mathbb{P}} 0, \quad n \to \infty?$$

- Well, no... Consider setting $\hat{Q}_{p_n} = \emptyset$.

Consistency

Theorem

Let X be an m-variate elliptically distributed random variable with absolutely continuous generating variate $\mathcal{R}$, location vector μ and scatter matrix Σ. Let $X_1, X_2, \ldots, X_n$ be i.i.d. copies of X. Let $\mathcal{R} \in RV_\gamma$, $\gamma > 0$, and let $\hat{\gamma}_n$ denote the corresponding separating Hill estimator. Assume that the following conditions hold:

1. $k_n \to \infty$, $k_n/n \to 0$, *as* $n \to \infty$.
2. $np_n = o(k_n)$, $\log(np_n) = o\left(\sqrt{k_n}\right)$ *and* $1/p_n = O\left(n^{1/(2\gamma)}\right)$, *as* $n \to \infty$.
3. $\sqrt{n}(\hat{\mu}_n - \mu) = O_{\mathbb{P}}(1)$ *and* $\sqrt{n}(\hat{\Sigma}_n - \Sigma) = O_{\mathbb{P}}(1)$.
4. *Some other technical conditions...*

Then as $n \to \infty$,

$$\frac{\mathbb{P}(X \in \hat{Q}_{p_n} \triangle Q_{p_n})}{p_n} \xrightarrow{\mathbb{P}} 0.$$

Table of Contents

Problem statement

Univariate extreme quantile estimation

Elliptical distributions

Multivariate extreme quantile estimator

Consistency

Application

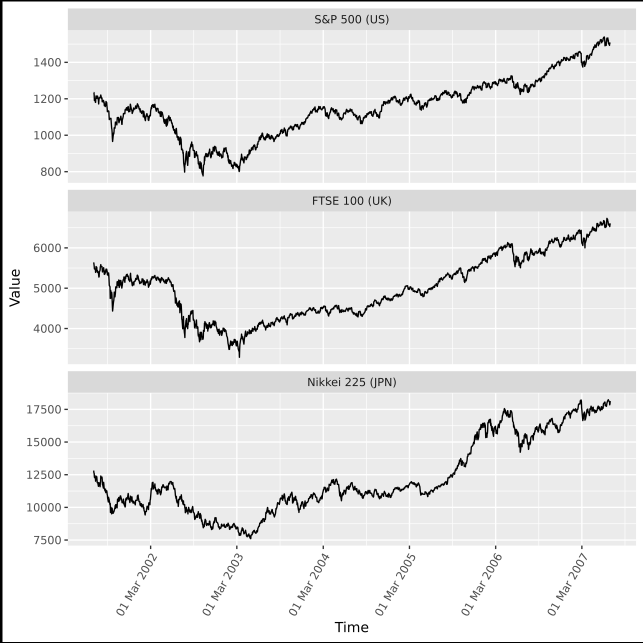

Figure: Three different stock indices from the time period 2.7.2001 - 29.6.2007 (1565 observations)

Preprocessing of the data

1. Compute log returns $Y_t^{(i)} = \log\left(X_t^{(i)}/X_{t-1}^{(i)}\right)$ for each time series $i \in \{1, 2, 3\}$.
2. Fit $\mathrm{EGARCH}(1, 1)$ model to each time series and save the (estimated) residuals.
3. The following assumptions hold for the residuals:
   - Independence (Ljung-Box test);
   - Ellipticity (Huffer and Park 2007);
   - Heavy-tailedness of the generating variate $\mathcal{R}$ (Hüsler and Li 2006).

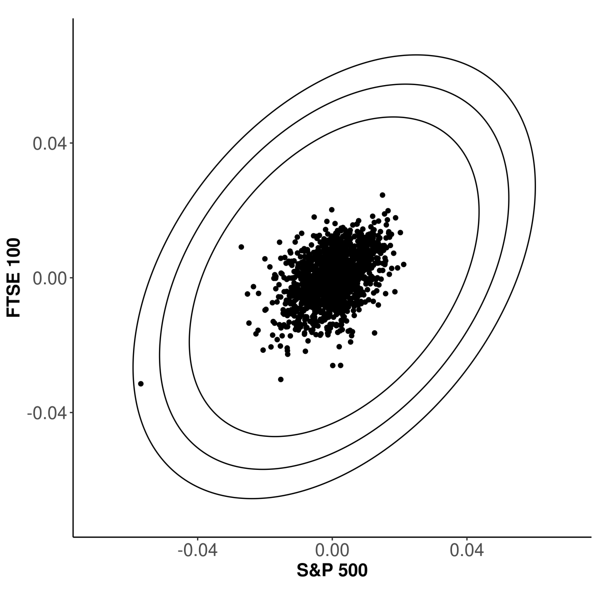

Figure: Predicted extreme quantile regions corresponding the log returns of 2.7.2007 for the pair (USA, UK), $p \in \{1/2000, 1/5000, 1/10000\}$. In the estimation we chose $k = 160$ and location and scatter were estimated with MCD with $\alpha = 0.5$.

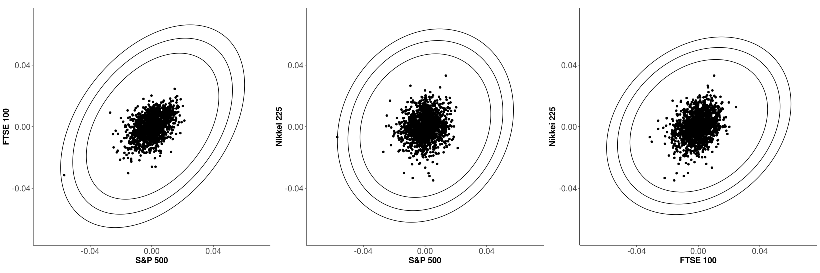

Figure: Same as above, but for all the pairs (USA, UK), (USA, JPN) and (UK, JPN).

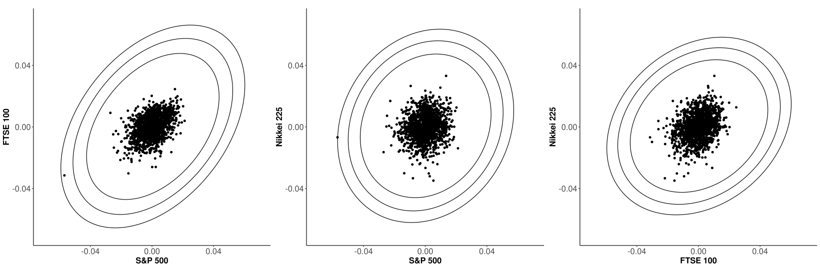

Figure: Same as above, but for all the pairs (USA, UK), (USA, JPN) and (UK, JPN).

- Day 27.2.2007 hits the three-dimensional estimated extreme quantile region for $p = 1/5000$ and $k = 160$.

Pre-proof of the article:

- https://www.sciencedirect.com/science/article/pii/S0047259X24000216

Simulations:

- https://github.com/perej1/elliptical-sim

Empirical example:

- https://github.com/perej1/elliptical-empirical

References

Cai, Juan-Juan, John HJ Einmahl, and Laurens De Haan (2011). "Estimation of extreme risk regions under multivariate regular variation". In: *The Annals of Statistics* 39.3, pp. 1803–1826.

Cambanis, Stamatis, Steel Huang, and Gordon Simons (1981). "On the theory of elliptically contoured distributions". In: *Journal of Multivariate Analysis* 11.3, pp. 368–385.

De Haan, L. and A. Ferreira (2006). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

Heikkilä, Matias, Yves Dominicy, and Pauliina Ilmonen (2019). "On multivariate separating Hill estimator under estimated location and scatter". In: *Statistics* 53.2, pp. 301–320.

Huffer, Fred W and Cheolyong Park (2007). "A test for elliptical symmetry". In: *Journal of Multivariate Analysis* 98.2, pp. 256–281.

References II

Hult, Henrik and Filip Lindskog (2002). “Multivariate extremes, aggregation and dependence in elliptical distributions”. In: *Advances in Applied probability* 34.3, pp. 587–608.

Hüsler, Jürg and Deyuan Li (2006). “On testing extreme value conditions”. In: *Extremes* 9, pp. 69–86.

Tsybakov, Alexandre B (1997). “On nonparametric estimation of density level sets”. In: *The Annals of Statistics* 25.3, pp. 948–969.