ON THE IMPACT OF APPROXIMATION ERRORS ON EXTREME QUANTILE ESTIMATION WITH APPLICATIONS TO FUNCTIONAL DATA ANALYSIS

Based on collaboration with Pauliina Ilmonen, Lauri Viitasaari, Valentin Garino and Benny Avelin

https://doi.org/10.48550/arXiv.2307.03581 (submitted to a journal)

Jaakko Pere

7th of May, 2025

Dep. of Mathematics and Statistics, University of Helsinki

Agenda of the presentation

Univariate Extreme Value Theory

Multidimensional Extremes and Impact of Approximation Errors

Extreme Quantile Estimation for L^p-Norms

Table of Contents

What is Extreme Value Theory?

Extreme value theory is concerned about inference of rare events.

What is Extreme Value Theory?

Extreme value theory is concerned about inference of rare events.

- Extreme quantile estimation
- Tail probability estimation
- Estimation of the endpoint of a given distribution

What is Extreme Value Theory?

Extreme value theory is concerned about inference of rare events.

- Extreme quantile estimation
- Tail probability estimation
- Estimation of the endpoint of a given distribution

See de Haan and Ferreira, 2006 for a review.

Maximum Domain of Attraction

Definition

Let $Y_1, \ldots, Y_n$ be i.i.d. observations of a random variable Y. If there exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$, and a random variable G with a nondegenerate distribution such that

$$\frac{\max(Y_1, \ldots, Y_n) - b_n}{a_n} \xrightarrow{\mathcal{D}} G, \quad n \to \infty,$$

we say that Y belongs to the maximum domain of attraction of G, and denote $Y \in \mathrm{MDA}(G)$.

Extreme Value Index

Theorem (Fisher and Tippett, 1928; Gnedenko, 1943)

Up to location and scale, the distribution of $G = G_\gamma$ is characterized by the parameter γ, called the extreme value index. That is, the distribution of G_γ is of the type

$$F_{G_\gamma}(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \quad 1+\gamma x > 0 & \text{if} \quad \gamma \neq 0, \\ \exp\left(-e^{-x}\right), \quad x \in \mathbb{R} & \text{if} \quad \gamma = 0. \end{cases}$$

Extreme Value Index

Theorem (Fisher and Tippett, 1928; Gnedenko, 1943)

Up to location and scale, the distribution of $G = G_\gamma$ is characterized by the parameter γ, called the extreme value index. That is, the distribution of G_γ is of the type

$$F_{G_\gamma}(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \quad 1+\gamma x > 0 & \text{if} \quad \gamma \neq 0, \\ \exp\left(-e^{-x}\right), \quad x \in \mathbb{R} & \text{if} \quad \gamma = 0. \end{cases}$$

In the case $\gamma > 0$ the type of G_γ is Fréchet,

$$\Phi_\gamma(x) = \begin{cases} 0, & x \leq 0 \\ \exp\left(-x^{-1/\gamma}\right), & x > 0. \end{cases}$$

Tail Quantile Function

Define the tail quantile function corresponding to a distribution F by

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

where we denote the left-continuous inverse of a nondecreasing function by $f^{\leftarrow}(y) = \inf\{x \in \mathbb{R} : f(x) \geq y\}$.

Tail Quantile Function

Define the tail quantile function corresponding to a distribution F by

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

where we denote the left-continuous inverse of a nondecreasing function by $f^{\leftarrow}(y) = \inf\{x \in \mathbb{R} : f(x) \geq y\}$.

That is, $U(1/p)$ is the $(1 - p)$-quantile.

Definition (Regular variation)

A Lebesgue measurable function $f : \mathbb{R}^+ \to \mathbb{R}$ that is eventually positive is regularly varying with index $\alpha \in \mathbb{R}$ if for all $x > 0$,

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^\alpha.$$

Then we denote $f \in \mathrm{RV}_\alpha$. Furthermore, we say that a function f is slowly varying if $f \in \mathrm{RV}_0$.

Definition (Regular variation)

A Lebesgue measurable function $f : \mathbb{R}^+ \to \mathbb{R}$ that is eventually positive is regularly varying with index $\alpha \in \mathbb{R}$ if for all $x > 0$,

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\alpha.$$

Then we denote $f \in \mathrm{RV}_\alpha$. Furthermore, we say that a function f is slowly varying if $f \in \mathrm{RV}_0$.

Intuition:

$$f \in \mathrm{RV}_\alpha \iff f(x) = L(x)x^\alpha, \quad L \in \mathrm{RV}_0 \,.$$

We also have

$$\lim_{x\to\infty} x^{-\varepsilon} L(x) = 0, \quad \forall\, \varepsilon > 0.$$

Construction of an Extreme Quantile Estimator

Theorem (de Haan, 1970; Gnedenko, 1943)

Let $\gamma > 0$. *We have*

$$Y \in \mathrm{MDA}\,(G_\gamma) \iff 1 - F \in \mathrm{RV}_{-1/\gamma} \iff U \in \mathrm{RV}_\gamma\,.$$

Construction of an Extreme Quantile Estimator

Theorem (de Haan, 1970; Gnedenko, 1943)

Let $\gamma > 0$. *We have*

$$Y \in \mathrm{MDA}\left(G_\gamma\right) \iff 1 - F \in \mathrm{RV}_{-1/\gamma} \iff U \in \mathrm{RV}_\gamma \, .$$

Choose $t = n/k$ and $x = k/(np)$ to get the approximation

$$U\left(\frac{1}{p}\right) \approx U\left(\frac{n}{k}\right)\left(\frac{k}{np}\right)^\gamma .$$

Extreme Quantile Estimation

Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is an i.i.d. sample of $Y \in \mathrm{MDA}\,(G_\gamma)$, $\gamma > 0$. Denote order statistics corresponding to the sample $\boldsymbol{Y}$ by $\boldsymbol{Y}_{1,n} \leq \cdots \leq \boldsymbol{Y}_{n,n}$. Then an estimator for the extreme $(1-p)$-quantile $x_p = U\,(1/p)$ can be given as

$$\hat{x}_p\,(\boldsymbol{Y}) = \boldsymbol{Y}_{n-k,n} \left(\frac{k}{np} \right)^{\hat{\gamma}(\boldsymbol{Y})},$$

where $\hat{\gamma}$ is an estimator for the extreme value index γ.

The Hill Estimator (Hill, 1975; Mason, 1982)

Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is an i.i.d. sample of $Y \in \mathrm{MDA}\,(G_\gamma)$, $\gamma > 0$. The Hill estimator is defined as

$$\hat{\gamma}_H\,(\boldsymbol{Y}) = \frac{1}{k}\sum_{i=0}^{k-1} \ln\left(\frac{\boldsymbol{Y}_{n-i,n}}{\boldsymbol{Y}_{n-k,n}}\right).$$

The Hill Estimator (Hill, 1975; Mason, 1982)

Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is an i.i.d. sample of $Y \in \mathrm{MDA}\,(G_\gamma)$, $\gamma > 0$. The Hill estimator is defined as

$$\hat{\gamma}_H\,(\boldsymbol{Y}) = \frac{1}{k}\sum_{i=0}^{k-1} \ln\left(\frac{\boldsymbol{Y}_{n-i,n}}{\boldsymbol{Y}_{n-k,n}}\right).$$

If additionally as $n \to \infty$, $k = k_n \to \infty$, $k/n \to 0$, then

$$\hat{\gamma}_H\,(\boldsymbol{Y}) \xrightarrow{\mathbb{P}} \gamma, \quad n \to \infty.$$

Table of Contents

A Traditional Approach to Extremes

Definition (Multivariate regular variation)

Let Θ be a probability measure on the unit sphere $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. A d-dimensional random vector X is multivariate regularly varying with the extreme value index $\gamma > 0$ and the probability measure Θ if

$$\lim_{t \to \infty} \frac{\mathbb{P}\left(\|X\|_2 \geq tx, X/\|X\|_2 \in A\right)}{\mathbb{P}\left(\|X\|_2 \geq t\right)} = x^{-1/\gamma}\Theta\left(A\right),$$

for every $x > 0$ and for every Borel set A in $\mathbb{S}^{d-1}$ with $\Theta\left(\partial A\right) = 0$.

A Traditional Approach to Extremes

Definition (Multivariate regular variation)

Let Θ be a probability measure on the unit sphere $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. A d-dimensional random vector X is multivariate regularly varying with the extreme value index $\gamma > 0$ and the probability measure Θ if

$$\lim_{t\to\infty} \frac{\mathbb{P}\left(\|X\|_2 \geq tx, X/\|X\|_2 \in A\right)}{\mathbb{P}\left(\|X\|_2 \geq t\right)} = x^{-1/\gamma}\Theta\left(A\right),$$

for every $x > 0$ and for every Borel set A in $\mathbb{S}^{d-1}$ with $\Theta\left(\partial A\right) = 0$.

- Estimation of multivariate extreme quantile regions under multivariate regular variation based on
  - density (Cai et al., 2011), and
  - half-space depth (He & Einmahl, 2016).

An Alternative Framework

1. Approach in multidimensional extremes:
    - Let $X \in \mathbb{S}$ be a random object, where, e.g., $\mathbb{S} = \mathbb{R}^d$ or $\mathbb{S} = \mathsf{L}^p([0, 1]^d)$.

An Alternative Framework

1. Approach in multidimensional extremes:
   - Let $X \in \mathbb{S}$ be a random object, where, e.g., $\mathbb{S} = \mathbb{R}^d$ or $\mathbb{S} = \mathsf{L}^p([0,1]^d)$.
   - Apply extreme value theory to $g(X)$, where g is a suitable map depending on the context.

An Alternative Framework

1. Approach in multidimensional extremes:
   - Let $X \in \mathbb{S}$ be a random object, where, e.g., $\mathbb{S} = \mathbb{R}^d$ or $\mathbb{S} = \mathrm{L}^p([0,1]^d)$.
   - Apply extreme value theory to $g(X)$, where g is a suitable map depending on the context.
2. Often instead of the sample $\boldsymbol{Y}$, only approximations $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ are available.

An Alternative Framework

1. Approach in multidimensional extremes:
   - Let $X \in \mathbb{S}$ be a random object, where, e.g., $\mathbb{S} = \mathbb{R}^d$ or $\mathbb{S} = \mathrm{L}^p([0,1]^d)$.
   - Apply extreme value theory to $g(X)$, where g is a suitable map depending on the context.
2. Often instead of the sample $\boldsymbol{Y}$, only approximations $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ are available.
   - How the approximation error affects the asymptotics?

Applications

- Elliptical extreme quantile region estimation (Pere et al., 2024).
- Extreme value index estimation for latent model (Virta et al., 2024).
- Estimation of the extreme value index corresponding to functional PCA scores (Kim & Kokoszka, 2019).

Approximated L^p-Norms

- Let $X \in \mathrm{L}^p\left([0,1]^d\right)$, and let $X_1, \ldots, X_n$ be i.i.d. copies of X.
- We wish to estimate extreme value index and extreme quantiles corresponding to $\|X\|_p \in \mathrm{MDA}\left(G_\gamma\right)$, $\gamma > 0$.

Approximated L^p-Norms

- Let $X \in \mathrm{L}^p\left([0,1]^d\right)$, and let $X_1, \ldots, X_n$ be i.i.d. copies of X.
- We wish to estimate extreme value index and extreme quantiles corresponding to $\|X\|_p \in \mathrm{MDA}\left(G_\gamma\right)$, $\gamma > 0$.
- In practice we never observe $X_1, \ldots, X_n$.

Approximated L^p-Norms

- Let $X \in \mathrm{L}^p\left([0,1]^d\right)$, and let $X_1, \ldots, X_n$ be i.i.d. copies of X.
- We wish to estimate extreme value index and extreme quantiles corresponding to $\|X\|_p \in \mathrm{MDA}\left(G_\gamma\right)$, $\gamma > 0$.
- In practice we never observe $X_1, \ldots, X_n$.
- Approximate norms with Riemann sums or Monte Carlo integration.
- Use approximated norms $\hat{Y}_i$ in the estimation.

Approximated L^p-Norms

- Let $X \in \mathrm{L}^p\left([0,1]^d\right)$, and let $X_1, \ldots, X_n$ be i.i.d. copies of X.
- We wish to estimate extreme value index and extreme quantiles corresponding to $\|X\|_p \in \mathrm{MDA}\left(G_\gamma\right)$, $\gamma > 0$.
- In practice we never observe $X_1, \ldots, X_n$.
- Approximate norms with Riemann sums or Monte Carlo integration.
- Use approximated norms $\hat{Y}_i$ in the estimation.
- As the estimator of the extreme value index we choose the Hill estimator

$$\hat{\gamma}(\hat{\boldsymbol{Y}}) = \frac{1}{k} \sum_{i=0}^{k-1} \ln\left(\frac{\hat{\boldsymbol{Y}}_{n-i,n}}{\hat{\boldsymbol{Y}}_{n-k,n}}\right).$$

Draft of the Main Result

Let $\gamma > 0$. Let $Y_1, \ldots, Y_n$ be i.i.d. copies of $Y \in \mathrm{MDA}\left(G_\gamma\right)$ and $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ the corresponding approximations. Denote errors by $E_i = \left|\hat{Y}_i - Y_i\right|$. If

$$\sqrt{k}\frac{\boldsymbol{E}_{n,n}}{U_Y\left(n/k\right)} \xrightarrow{\mathbb{P}} 0, \quad n \to \infty,$$

then

$$\sqrt{k}\left(\hat{\gamma}(\hat{\boldsymbol{Y}}) - \gamma\right) \quad \text{and} \quad \frac{\sqrt{k}}{\ln\left(k/(np)\right)}\left(\frac{\hat{x}_p(\hat{\boldsymbol{Y}})}{U(1/p)} - 1\right)$$

are asymptotically normally distributed under the standard assumptions (second-order condition, rate for $p = p_n$, $k = k_n \to \infty$, $k/n \to 0$, as $n \to \infty$).

Table of Contents

Riemann Sum Approximated Norms

Let $\gamma > 0$. Let X_i be i.i.d. copies of $X \in L^p([0,1])$, $p \in [1,\infty]$, s.t. $Y = \|X\|_p \in \mathrm{MDA}(G_\gamma)$. Let $\hat{Y}_i$ be the Riemann sum approximated norms (based on discretizations with m equidistant observed points). Suppose for all $s, t \in [0,1]$, X satisfies

$$|X(t) - X(s)| \leq V\phi(|t-s|) \quad a.s.,$$

for some random variable $V \in \mathrm{MDA}\left(G_{\gamma'}\right)$, $\gamma' > 0$, and for some continuous decreasing function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ with $\phi(0) = 0$. Then the condition

$$\sqrt{k}\frac{\boldsymbol{E}_{n,n}}{U_Y(n/k)} \xrightarrow{\mathbb{P}} 0, \quad n \to \infty,$$

translates into

$$\sqrt{k}\phi\left(\frac{1}{m}\right)k^{\gamma}n^{\gamma'-\gamma} \to 0, \quad n \to \infty.$$

Concentration for $\hat{\gamma}(\hat{\boldsymbol{Y}})$

In order to give concentration inequality for $\mathbb{P}\left(\left|\hat{\gamma}(\hat{\boldsymbol{Y}}) - \hat{\gamma}(\boldsymbol{Y})\right| > x\right)$ one needs to control the errors

$$\mathbb{P}\left(\frac{\boldsymbol{E}_{n,n}}{U_Y\,(n/k)} > x\right)$$

and the convergence

$$\mathbb{P}\left(\left|\frac{\boldsymbol{Y}_{n-k,n}}{U(n/k)} - 1\right| > x\right).$$

Chernoff-Type Bound for Intermediate Order Statistics

Let $\gamma > 0$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be an i.i.d. sample of $Y \in \mathrm{MDA}\,(G_\gamma)$ and assume that, as $n \to \infty$, $k = k_n \to \infty$, and $k/n \to 0$. Then for sufficiently large n

$$\mathbb{P}\left(\left|\frac{\boldsymbol{Y}_{n-k,n}}{U(n/k)} - 1\right| > x\right) \leq C_1 e^{-C_2 k},$$

where the constants $C_1 > 0$ and $C_2 > 0$ depend on x and γ.

Thank you for your attention!

- Link to the manuscript (arXiv):
  https://doi.org/10.48550/arXiv.2307.03581
- Link to slides (Github):
  https://github.com/perej1/ics-and-related

References I

Cai, J.-J., Einmahl, J. H. J., & de Haan, L. (2011).Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*, *39*(3), 1803–1826. https://doi.org/10.1214/11-AOS891

de Haan, L., & Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer. https://doi.org/10.1007/0-387-34471-3

de Haan, L. (1970). *On Regular Variation and Its Application to Weak Convergence of Sample Extremes* [Doctoral dissertation].

Fisher, R. A., & Tippett, L. H. C. (1928).Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(2), 180–190. https://doi.org/https://doi.org/10.1017/S0305004100015681

References II

Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire. *Annals of Mathematics*, *44*(3), 423–453. https://doi.org/https://doi.org/10.2307/1968974

He, Y., & Einmahl, J. H. J. (2016). Estimation of extreme depth-based quantile regions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(2), 449–461. https://doi.org/10.1111/rssb.12163

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, *3*(5), 1163–1174. https://doi.org/https://doi.org/10.1214/aos/1176343247

Kim, M., & Kokoszka, P. (2019). Hill estimator of projections of functional data on principal components. *Statistics*, *53*(4), 699–720. https://doi.org/10.1080/02331888.2019.1609476

References III

Mason, D. M. (1982). Laws of Large Numbers for Sums of Extreme Values. *The Annals of Probability*, *10*(3), 754–764. https://doi.org/https://doi.org/10.1214/aop/1176993783

Pere, J., Ilmonen, P., & Viitasaari, L. (2024). On extreme quantile region estimation under heavy-tailed elliptical distributions. *Journal of Multivariate Analysis*, *202*, 105314. https://doi.org/https://doi.org/10.1016/j.jmva.2024.105314

Virta, J., Lietzén, N., Viitasaari, L., & Ilmonen, P. (2024). Latent model extreme value index estimation. *Journal of Multivariate Analysis*, *202*, 105300. https://doi.org/https://doi.org/10.1016/j.jmva.2024.105300