한양대학교 HANYANG UNIVERSITY

Introduction to Extreme Value Statistics

The AIGS X IE graduate seminar for Fall semester, 2025

2025-09-10

Seoncheol Park
Department of Mathematics
Hanyang University
pscstat@hanyang.ac.kr

Contents

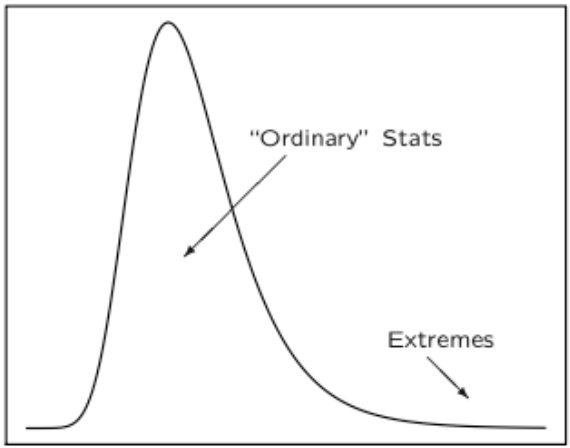
Topics

- Motivation
- Block maxima approach (using GEV)
- Peak-over threshold approach (using GPD)
- Extremal quantile regression
- Recent advances in incorporating machine learning techniques in extremes

Motivation

Conventional vs Extreme Value Statistics

- Ordinary statistics: Describes bulk of distribution.
- **Extremes**: Characterizes the tail of the distribution.



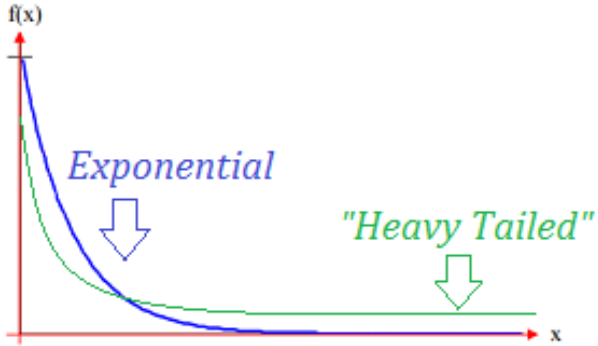
Why Study Extremes?

i Note

네덜란드 영토의 40%는 해수면보다도 낮은데 이는 제방 둑으로 보호되고 있다. 그러나 겨울철 불어오는 폭풍우는 해수면을 밀어 올리고 해변가에 위치한 제방 둑은 이를 견뎌 내야만 한다.

이를 위해 네덜란드 정부는 경비와 안정성을 모두 고려하여 연중 최대 해수면이 제방 둑을 넘칠 확률이 0.0001이 되도록 제방 둑의 높이를 정하고자 한다. 이때 사용할 수 있는 해수면 자료는 100년 남짓이라고 한다. 이 자료를 이용하여 해수면이 10,000년에 한번 정도 넘어설 정도의 제방 둑의 높이를 추정할 수 있을까?

Heavy-Tailed Distribution



Heavy-Tailed Distribution (cont.)

Definition 1:

- A distribution function F is said to be **heavy-tailed** if and only if, $\forall \mu > 0$,

$$\limsup_{x \to \infty} \frac{1 - F(x)}{e^{-\mu x}} = \limsup_{x \to \infty} \frac{\overline{F}(x)}{e^{-\mu x}} = \infty. \tag{1}$$

- A random variable X is said to be heavy-tailed if its distribution function is heavy-tailed. (Nair et al., 2022)

- Consider a random variable X. Then the following statements are equivalent:
  - X is heavy-tailed.
  - The MGF $M_X(t) := E[e^{tX}] = \infty, \quad \forall t > 0.$

Concentration Inequality

Q. Given a random variable Y, how **concentrated** is Y (e.g., around its mean or median)?

- Chernoff bound: When $\lambda \geq 0$, we have

$$P[Z \geq t] \leq e^{-\lambda t} E[e^{\lambda Z}]. \tag{2}$$

- If we define the log-moment generating function $\psi_Z(\lambda)$ of a random variable Z is defined as

$$\psi_Z(\lambda) = \log E[e^{\lambda Z}], \quad \lambda \geq 0. \tag{3}$$

- If Z is sufficiently heavy-tailed, it could even be that $E[e^{\lambda Z}] = \infty$ for all $\lambda > 0$, in which case, the Chernoff bound **cannot** be used.
- In randomized algorithms and data structures, Chernoff bounds are used to bound error probabilities and runtimes.
- With heavy tails, performance can be dominated by rare but extreme cases.

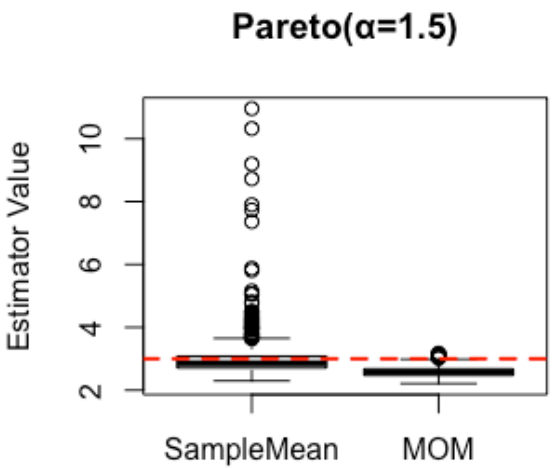
Example: Pareto Distribution

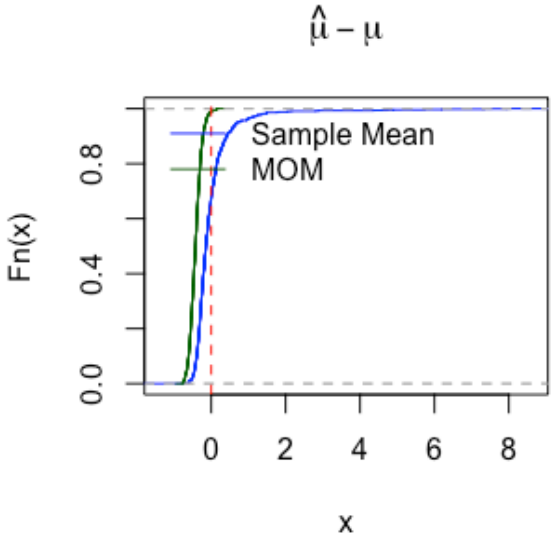
- Suppose that $X \sim \text{Pareto}(\alpha, x_m)$. Then the tail probability and the pdf of X are

$$P(X \geq x) = \overline{F}(x) = \left(\frac{x}{x_m}\right)^{-\alpha}, \quad \text{for } \alpha > 0, x \geq x_m > 0$$
$$f(x) = \frac{\alpha x_m}{x^{\alpha+1}}, \quad x \geq x_m \tag{4}$$

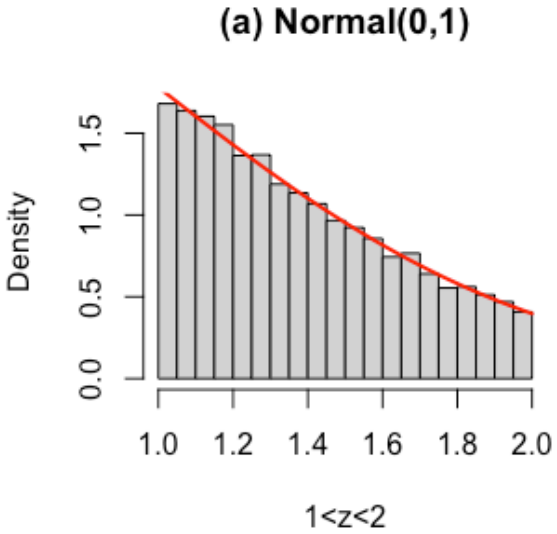
- Since the MGF $M_X(\lambda) = E[e^{\lambda X}] = \infty$ for all $\lambda \geq 0$, we cannot use Chernoff bound.
  - When $X \sim \text{Pareto}(\alpha, x_m)$, sample mean will be failed.
  - Instead, we can use **median-of-means**: Divide the data into k blocks, compute the mean of each block, then take the median of these block means as the final estimator.
- The Pareto distribution is characterized by a special **scale-invariance** property: Intuitively this means that the shape of the distribution does not change by changing the scale of measurements.

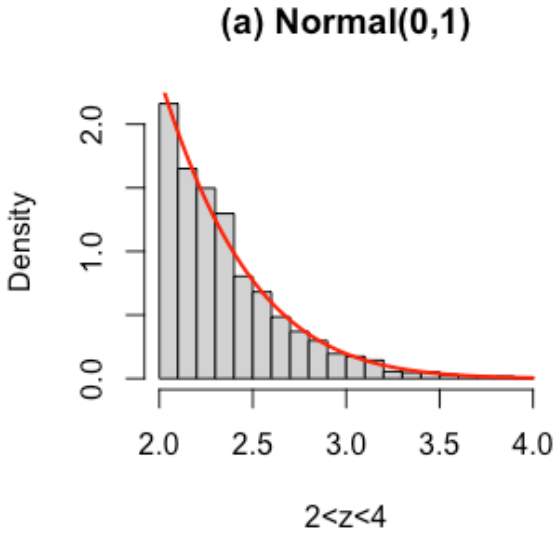
Example: Pareto Distribution (cont.)



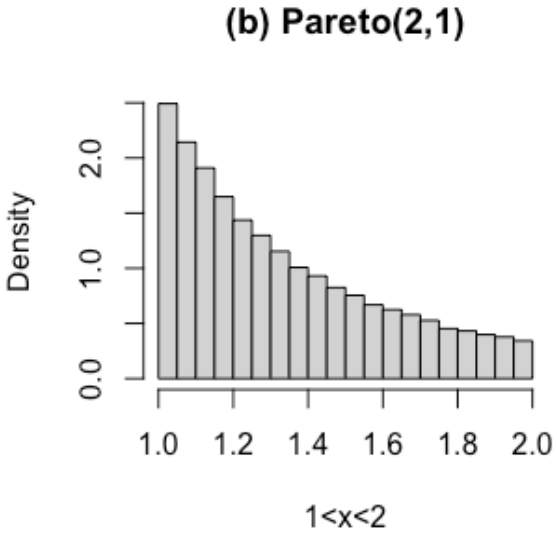


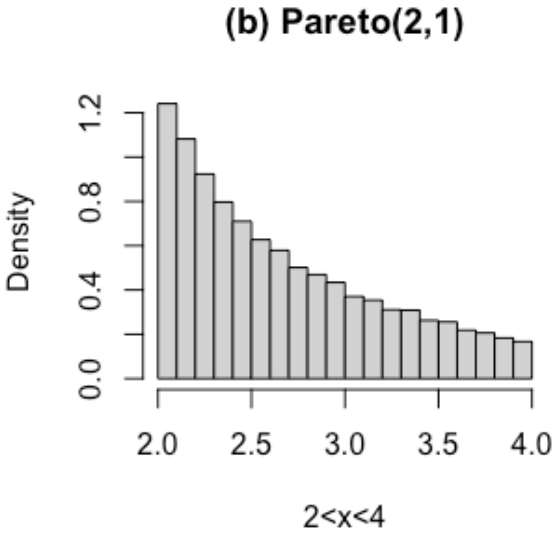
Example: Pareto Distribution (cont.)





Example: Pareto Distribution (cont.)





Scale Invariance

Definition 2 (Scale invariance):

- A distribution function F is **scale invariant** (or F has a **power law tail**) if $\exists x_0 > 0$ and a continuous positive function g such that

$$F(\lambda x) = g(\lambda)\overline{F}(x) \tag{5}$$

for all x, λ, satisfying $x, \lambda x \geq x_0$.

- Pareto is scale invariant,

$$\overline{F}(\lambda x) = \left(\lambda \frac{x}{x_m}\right)^{-\alpha} = \overline{F}(x)\lambda^{-\alpha}, \text{whenever } x, \lambda x > x_m. \tag{6}$$

- In practice, however, it is very difficult to find distributions that follow an **exact** power-law tail.

Regular Variation

Definition 3 (Regular variation):

- A funtion $f : \mathbb{R}_+ \to \mathbb{R}_+$ is **regularly varying** (asymptotically scale invariant) of index $\rho \in \mathbb{R}$, denoted $f \in \mathcal{RV}(\rho)$, if $\forall y > 0$,

$$\lim_{x \to \infty} \frac{f(xy)}{f(x)} = y^{\rho}. \tag{7}$$

- Furthermore, for $\rho \leq 0$, a distribution F is regularly varying of index ρ, denoted as $F \in \mathcal{RV}(\rho)$, if $\overline{F}(x) = 1 - F(x)$ is a regularly varying function of index ρ.

- Asymptotic scale invariance focuses only on the tail of the distribution, the body of such a distribution may behave in an arbitrary manner as long as the tail is approximately scale invariant.
- Intuitively, a heavy-tailed distribution satisfies scale invariance only in its **tail** region.

Regular Variation (cont.)

- Relationship between heavy-tailed and regularly varying (Nair et al., 2022)

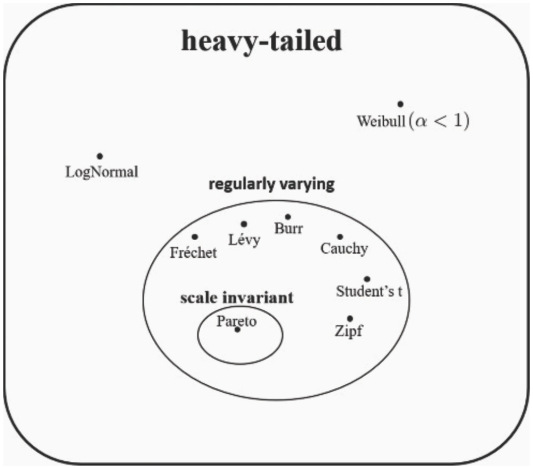


Figure 2.5 Scale invariant and regularly varying distributions.

Block Maxima Approach (GEV)

Review: Central Limit Theorem (CLT)

- Let $Y_1, Y_2, ..., \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and consider the **partial sum** $S_n = \sum_{i=1}^{n} Y_i$.
- We can easily check that $\frac{S_n - n\mu}{\sqrt{n}\sigma} \overset{D}{=} \mathcal{N}(0, 1)$.
- **CLT**: For any $Y, Y_1, Y_2, ..., \overset{\text{i.i.d.}}{\sim} F$ such that $\text{var}\ (Y) < \infty$, there exist sequences $a_n > 0$ and b_n such that, as $n \to \infty$,

$$\frac{S_n - b_n}{a_n} \xrightarrow{D} \mathcal{N}(0, 1). \tag{8}$$

  These sequences are $a_n = \sqrt{n\ \text{var}\ (Y)}$ and $b_n = nE(Y)$.

Review: Central Limit Theorem (CLT) (cont.)

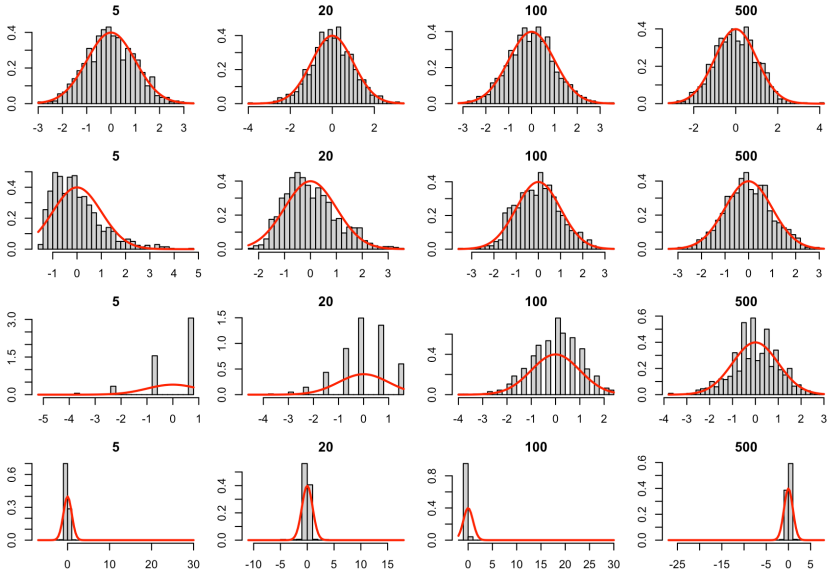


Figure 1: 정규, 감마, 베르누이, 코시 분포와 CLT.

Extremal Types Theorem

- Let $Y_1, Y_2, ..., \overset{\text{i.i.d.}}{\sim}$ with right endpoint $y_F = \sup\{y : F(y) < 1\}$, and consider the **partial maximum**

$$M_n = \max(Y_1, ..., Y_n). \tag{9}$$

- **Q**. Can we find sequences $a_n > 0$ and b_n such that the renormalized maximum $(M_n - b_n)/a_n$ has a **non-degenerate** limiting distribution? If so, what is the limit?
  - Exponential: $F(y) = 1 - \exp(-y/\lambda), \quad y > 0, \lambda > 0$
  - Fréchet: $F(y) = \exp(-y^{-\alpha}), \quad y > 0, \alpha > 0$
  - Uniform: $F(y) = y, \quad 0 < y < 1$
- Similarly to the CLT, the **extremal types theorem** (Fisher & Tippett, 1928) gives a partial answer to the above questions.

Extremal Types Theorem (cont.)

Theorem 1 (Extremal types theorem): If there exist sequences of constants $a_n > 0$ and b_n such that, as $n \to \infty$, $\frac{M_n - b_n}{a_n} \xrightarrow{D} Z \sim G$, where G is non-degenerate, then G must be a generalized extreme-value (GEV) distribution, i.e.,

$$G(y) = \begin{cases} \exp\left\{-\left(1 + \xi\frac{y-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}, & \xi = 0, \end{cases} \tag{10}$$

where $a_+ = \max(0, a)$ and $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are location, scale, and shape parameters.

The GEV Distribution

- $Y \sim \text{GEV}\ (\mu, \sigma, \xi)$ distribution, where

$$G(y) = \begin{cases} \exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, & \xi \neq 0 \\ \exp\left\{-\exp\left(-\frac{y-\mu}{\sigma}\right)\right\}, & \xi = 0, \end{cases} \tag{11}$$

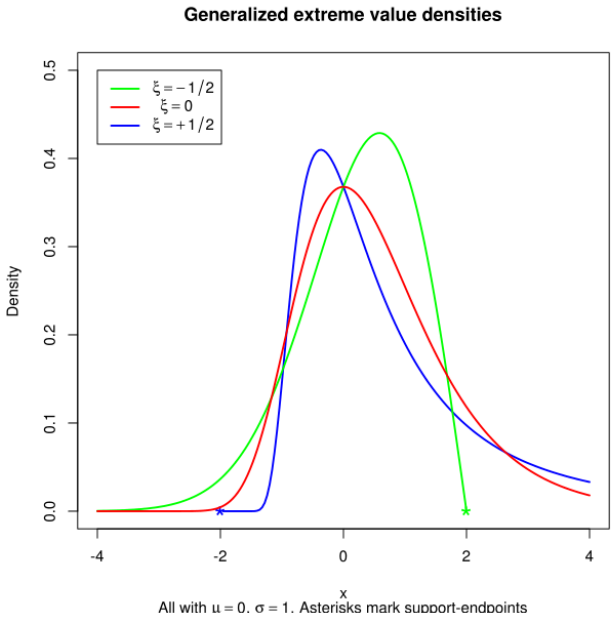
  where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$, defined on $\left\{y : 1+\xi\frac{y-\mu}{\sigma} > 0\right\}$.

- The crucial parameter is ξ, which determines the heaviness of the tail:
  - $\xi < 0$ (Bounded-tailed): **Reversed Weibull**
  - $\xi = 0$ (Light-tailed): **Gumbel**
  - $\xi > 0$ (Heavy-tailed): **Fréchet**

- Moments of $Y \sim \text{GEV}\ (\mu, \sigma, \xi)$:

$$E(Y^r) < \infty \Leftrightarrow \xi r < 1. \tag{12}$$

The GEV Distribution (cont.)

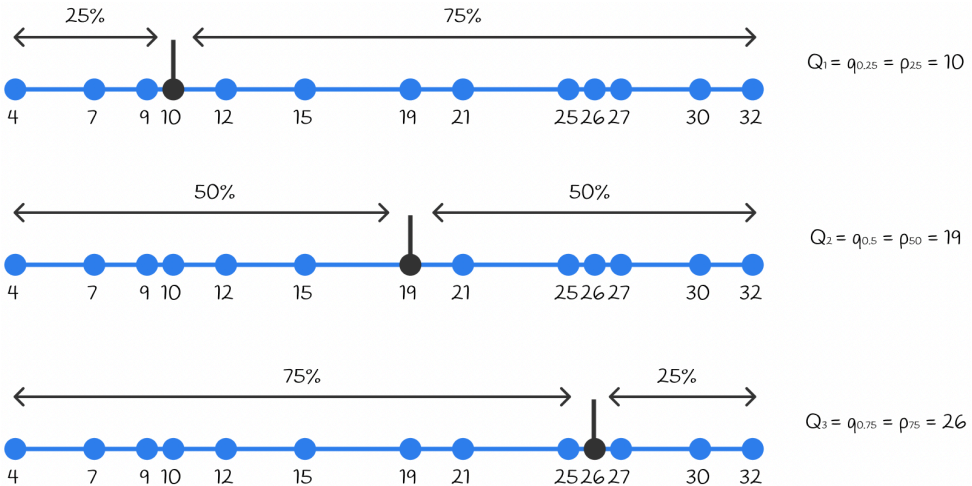


Max-Domains of Attraction (MDA)

- A distribution F is siad to be in the **max-domain of attraction** of G, denoted by $F \in \mathrm{MDA}\ (G)$, if there exist sequences $a_n > 0$ and b_n such that for any y, $F^n(a_n y + b_n) \to G(y)$, as $n \to \infty$.
- MDA of the **reversed Weibull** distribution
  - Beta, Uniform
- MDA of the **Gumbel** distribution
  - Normal, Gamma, Log-normal, Weibull, Exponential
- MDA of the **Fréchet** distribution
  - Fréchet, Pareto, Cauchy, Log-gamma, Student-t
- No domain of attraction
  - Geometric, Poisson
- If the tail is regularly varying with index $-1/\xi$, the maximum converges to the Fréchet type GEV with shape parameter $\xi > 0$.

Quantile

- Image from towardsdatascience.com



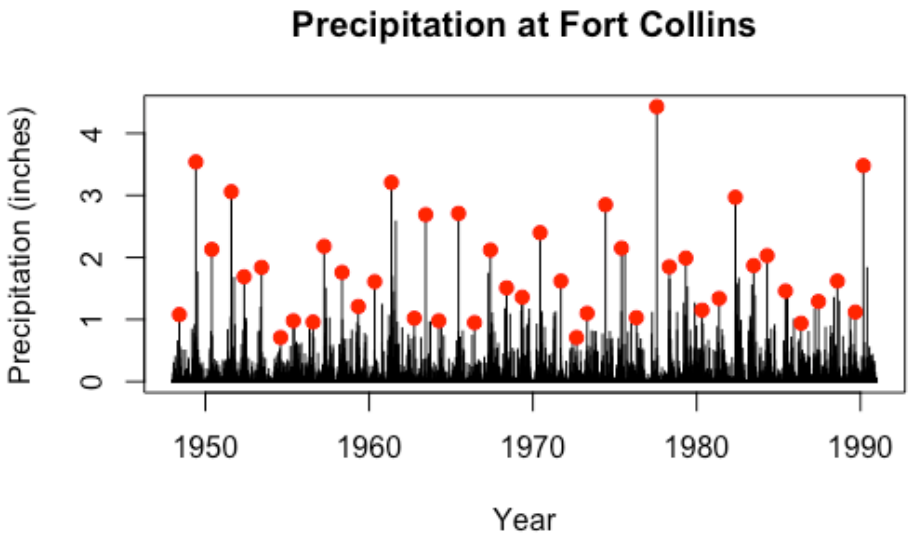
Return Levels (Quantiles)

- Practitioners usually consider **return levels (quantiles)**, since parameters μ, σ, ξ are often difficult to interpret.
- The value that is exceeded on average once every N years is called the **N year-return level**. The corresponding **return period** is N years.
- If $G \sim \text{GEV}\ (\mu, \sigma, \xi)$ is used to model annual maxima of a quantity of interest Y, the N year-return level y_N satisfies the equation

$$y_N = G^{-1}\left(1 - \frac{1}{N}\right) = \begin{cases} \mu + \frac{\sigma}{\xi}\left[\left\{-\log\left(1 - \frac{1}{N}\right)\right\}^{-\xi}\right], \xi \neq 0, \\ \mu - \sigma \log\left\{-\log\left(1 - \frac{1}{N}\right)\right\}, \quad \xi = 0 \end{cases}. \tag{13}$$

- This formula allows **extrapolation** beyond the range of the data when we have the **true** parameter values.

Example: Fort Collins Precipitation



Example: Fort Collins Precipitation (cont.)

- Measured value for 1997 event: 4.63 inches
- Let's analyze data preceding the event (1948-1990) and try to estimate the **return period** associated with an event of 4.63 inches.
- That means, we need to answer the question: "What is the probability the annual maximum event is larger than 4.63 inches?"
- Note that the largest observation (1948-1990) is 4.43 inches (**extrapolation**).
- Modeling process:

1. Model using non-extreme value distribution.
2. Model using GEV distribution.

(1) Gamma distribution

- Let X_t be the daily **summer** (April-October) precipitation amount for Fort Collins.
- To model precipitation, we need to account for zeroes.

$$X_t > 0 \text{ with prob. } p, \quad X_t = 0 \text{ with prob. } 1-p, \hat{p} = 0.263. \tag{14}$$

- Further, assume that $[X_t \mid X_t > 0] \sim \text{Gamma } (\alpha, \beta)$. ML estimates are $\hat{\alpha} = 0.656, \hat{\beta} = 3.20$.

$$\begin{aligned} P(X_t > 4.63) &= P(X_t > 4.63 \mid X_t > 0) P(X_t > 0) \\ &= (1 - F_X(4.63)) \times 0.263 \\ &= (1 - 0.9999999) \times 0.263 = 3.01 \times 10^{-8} \end{aligned} \tag{15}$$

$$\begin{aligned} P(\text{ann max } > 4.63) &= 1 - P(\text{entire year's obs } < 4.63) \\ &= 1 - (1 - P(\text{indiv obs } > 4.63))^{214} \\ &= 1 - \left(1 - 3.01 \times 10^{-8}\right)^{214} = 6.441 \times 10^{-6}. \end{aligned} \tag{16}$$

Modelling (cont.)

(Assume independence of daily observations, 214 *summer days* in a year.)

- **Return period**=$\left(6.441 \times 10^{-6}\right)^{-1} = 155,255$ years.

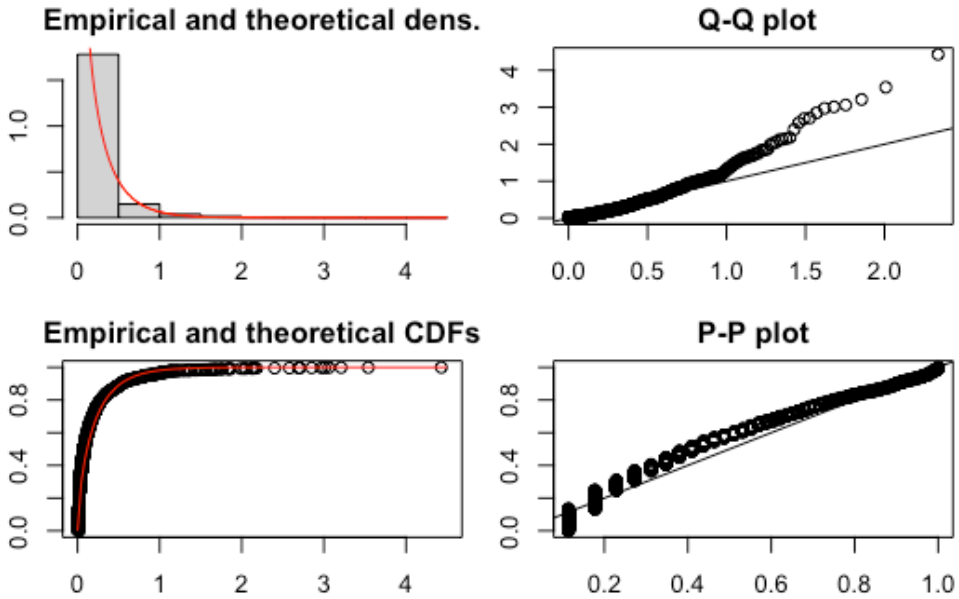
GEV distribution

- Let $M_n = \max_{t=1,\ldots,n}(X_t)$. Assume $M_n \sim \text{GEV}\ (\mu, \sigma, \xi)$.

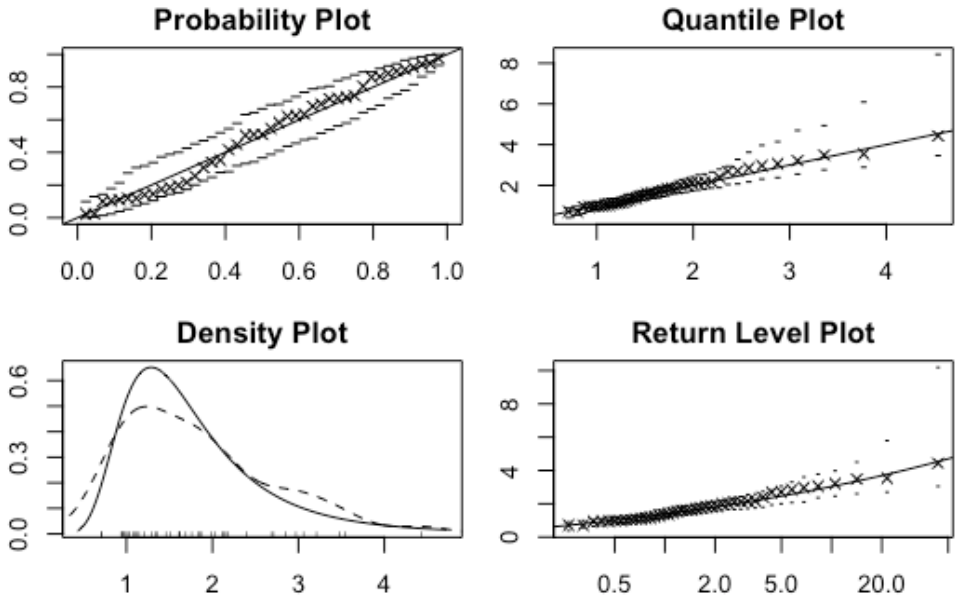
$$F_{M_n}(x) = P(M_n \leq x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{17}$$

- ML estimates: $\hat{\mu} = 1.384$, $\hat{\sigma} = 0.574$, $\hat{\xi} = 0.188$
- $P(\text{ann max} > 4.63) = 1 - F_{M_n}(4.63) = 0.021$
- Return period point estimate: $(0.021)^{-1} = 47.6$ years.

Results: (1) Gamma

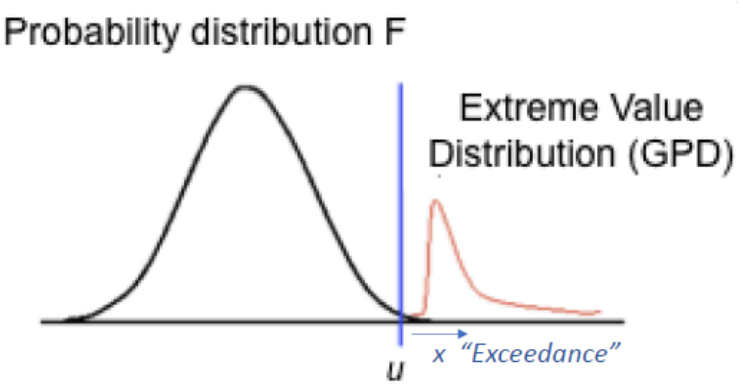


Results: (2) GEV



Peak-over Threshold Approach (GPD)

Basic Idea

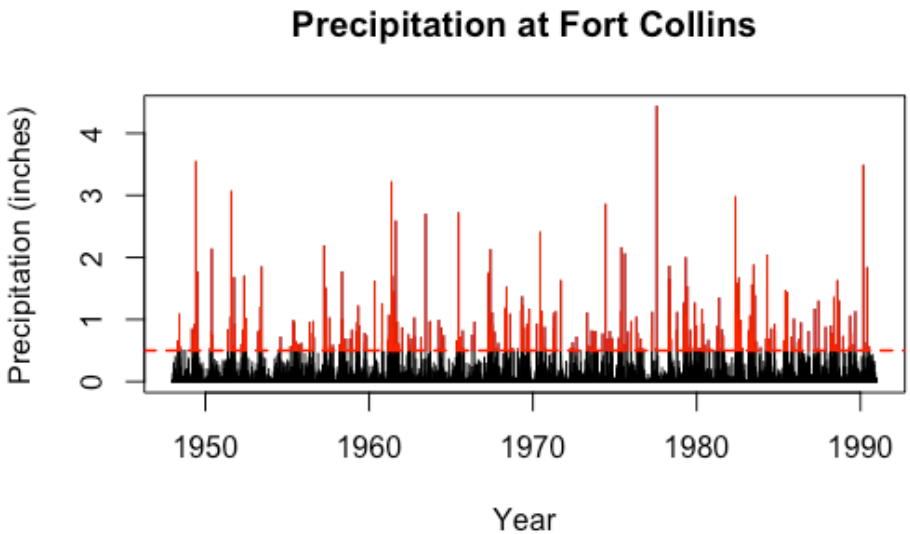


Generalized
Pareto Distribution

$\xi = shape,$
$\beta = scale$

$$G_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-1/\xi}, & if\ \xi \neq 0 \\ 1 - e^{\left(-\frac{x}{\beta}\right)}, & if \xi = 0 \end{cases}$$

Fort Collins Data with Threshold



Generalized Pareto Distribution (GPD)

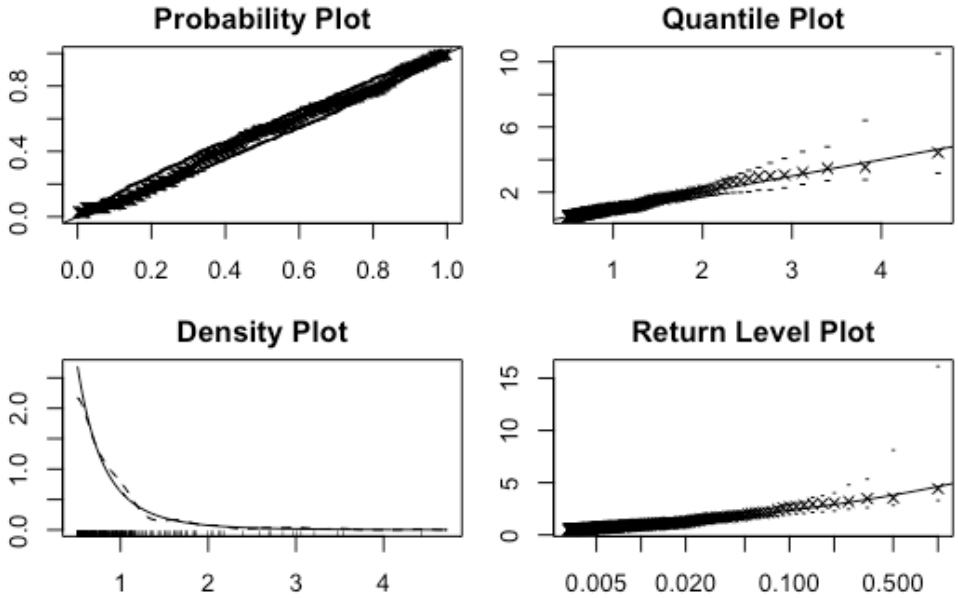
- The result is consistent with the Extremal Types Theorem.
- High threshold exceedances may be approximated by the **generalized Pareto distribution**, since for $y > 0$ (Coles, 2001),

$$P\left(\frac{Y-b_n}{a_n} > u+y \mid \frac{Y-b_n}{a_n} > u\right) = \frac{P\{(Y-b_n)/a_n > u+y\}}{P\{(Y-b_n)/a_n > u\}} \approx \frac{\{1+\xi(y+u-\mu)/\sigma\}_+^{-1/\xi}}{\{1+\xi(u-\mu)/\sigma\}_+^{-1/\xi}} = (1+\xi y/\tilde{\sigma})_+^{-1/\xi}, \tag{18}$$

where $\tilde{\sigma} = \sigma + \xi(u-\mu) > 0$.

- The right-hand side of the equation above corresponds to $1 - H(y)$, where $H(y)$ is the **GPD**$(\tilde{\sigma}, \xi)$ distribution.
- If the tail is regularly varying with index $-1/\xi$, the exceedances converge to the GPD with shape parameter ξ.

Results: (3) GPD



GPD vs GEV

GEV	GPD
$n = 43$	$n = 309$
$\hat{\mu} = 1.38(0.106)$	$u = 0.5$
$\hat{\sigma} = 0.57(0.086)$	$\hat{\beta} = 0.36(0.033)$
$\hat{\xi} = 0.19(0.171)$	$\hat{\xi} = 0.22(0.072)$
$\hat{r}_{100} = 5.6$ inches	$\hat{r}_{100} = 5.8$ inches
95% CI:$(2.1, 9.0)$	95% CI:$(3.3, 8.3)$

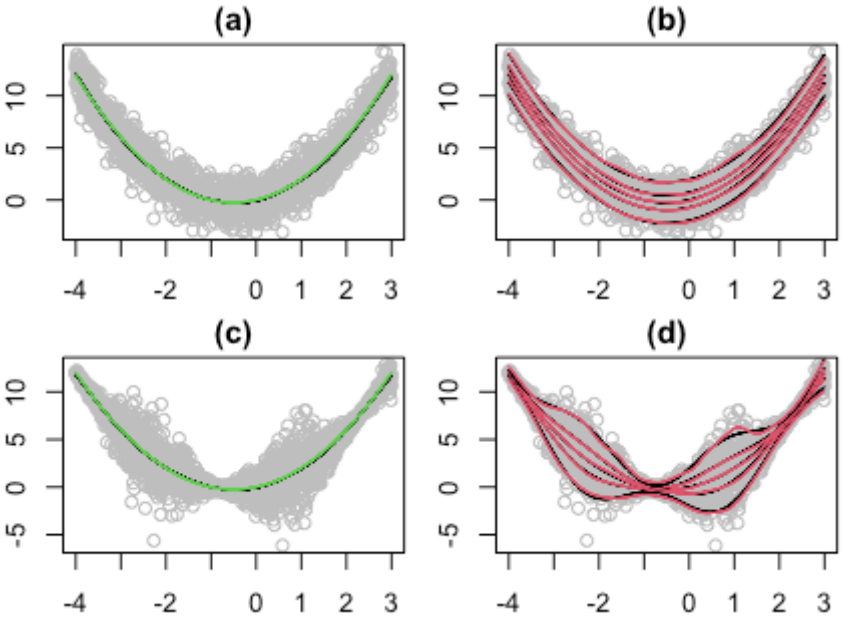
- Lower standard error for ξ
- Narrower confidence interval

Machine Learning and Extreme Value Statistics

Area of Extreme Value Statistics

- Univariate / Multivariate modeling
- Time series of extremes
- Spatial extremes and max-stable process
- Bayesian inference for extreme value modeling
- Extreme conditional quantiles
- Extreme dependence models

Conditional Quantile Estimation



Conditional Quantile Estimation (cont.)

- Sometimes, it is useful to compute conditional quantiles to understand the underlying data structure.

Q. How do we estimate 0.9995 conditional quantiles?

Research topics on conditional quantiles

- Regression + Quantiles $\rightarrow$ Quantile regression (Koenker & Bassett, 1978)
- Smoothing splines + Quantiles $\rightarrow$ Quantile smoothing splines (Koenker et al., 1994)
- Random forests + Quantiles $\rightarrow$ Quantile regression forests (Meinshausen, 2006)

Research topics on conditional extreme quantiles

- Regression + Extreme quantiles $\rightarrow$ Extreme quantile regression (Chernozhukov, 2005)
- Random forests + Extreme quantiles $\rightarrow$ Extremal random forests (Gnecco et al., 2024)

Quantile Regression

- Let $0 < \tau_L < 1$ be a fixed constant that is close to one. Consider the following linear quantile regression model:

$$Q_Y(\tau \mid \mathbf{x}) = \alpha(\tau) + \mathbf{x}^T\boldsymbol{\beta}, \quad \tau \in [\tau_L, 1], \tag{19}$$

  where $\alpha(\tau) \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the unknown quantile coefficients.

- Given the random sample $\{(y_i, \mathbf{x}_i), i = 1, ..., n\}$, the quantile coefficients can be estimated by

$$\left(\hat{\alpha}(\tau), \hat{\boldsymbol{\beta}}\right) = \text{argmin}_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \mathbf{x}_i^T\boldsymbol{\beta}), \tag{20}$$

  where $\rho_\tau(u) = \{\tau - I(u < 0)\}u$ is the **quantile loss function**.

Quantile Regression (cont.)

- Image from towardsdatascience.com

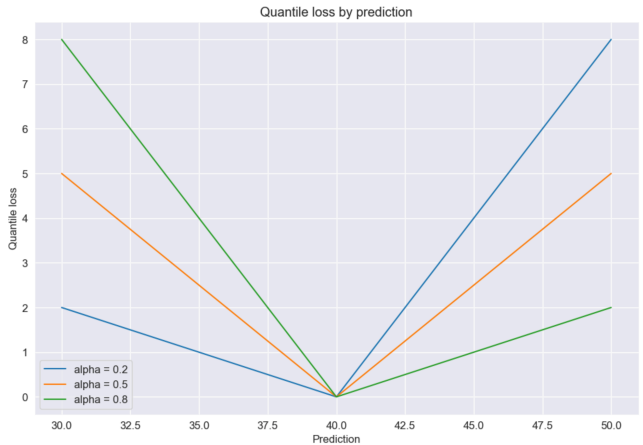


Figure 2: An example quantile loss functions.

Extreme Quantile Regression

- At the extreme quantiles $\tau_n \to 1$ as $n \to \infty$, the conventional quantile regression estimators $\hat{\alpha}(\tau)$ and $\hat{\beta}(\tau)$ are ofen not precise due to **data sparsity**.
- (Wang & Li, 2013) suggested a Weismann-type extrapolation estimator for $Q_Y(\tau_n \mid \mathbf{x})$,

$$\hat{Q}_Y(\tau_n \mid \mathbf{x}) = \hat{Q}_Y(\tau_{n-k} \mid \mathbf{x}) \left(\frac{1-\tau_{n-k}}{1-\tau_n} \right)^{\hat{\xi}(\mathbf{x})}, \tag{21}$$

where $\hat{\xi}(\mathbf{x})$ is a **Hill estimator**, which is an average of k log-excesses above random threshold $u = \hat{Q}_Y(\tau_j \mid \mathbf{x})$ ($k+1$th largest observation),

$$\begin{aligned} \hat{\xi} &= \frac{1}{k-[n^\eta]} \sum_{j=[n^\eta]}^{k} \log \frac{\hat{Q}_Y(\tau_{n-j} \mid \mathbf{x})}{\hat{Q}_Y(\tau_{n-k} \mid \mathbf{x})} \\ &= \frac{1}{k-[n^\eta]} \sum_{j=[n^\eta]}^{k} \Big(\log\big(\hat{Q}_Y(\tau_{n-j} \mid \mathbf{x})\big) - \log\big(\hat{Q}_Y(\tau_{n-k} \mid \mathbf{x})\big) \Big) \end{aligned} \tag{22}$$

with $\hat{Q}_Y(\tau_{n-k} \mid \mathbf{x}) \le \hat{Q}_Y(\tau_{n-k+1} \mid \mathbf{x}) \le \cdots \le \hat{Q}_Y\big(\tau_{n-[n^\eta]} \mid \mathbf{x}\big)$

Random Forests

- **Random forests**: Grow $t = 1, ..., k$ single trees via

1. Generate a boostrap sample from training data $\mathbf{Z}^*$
2. Also randomly select m variables from set of parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$
3. Fit a new regression tree $T(\boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_t \subset \boldsymbol{\theta}$

- When we have new data $\mathbf{X} = \mathbf{x}$, the predicted value of a single regression tree $T(\boldsymbol{\theta})$ is a mean of all observations within a terminal node (leaf) $\mathcal{R}_{l(\mathbf{x},\boldsymbol{\theta})}$ where $\mathbf{X}$ lives:

$$\text{single tree:} \quad \hat{\mu}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}, \boldsymbol{\theta}) Y_i, \quad w_i(\mathbf{x}, \boldsymbol{\theta}) = \frac{1_{\{X_i \in \mathcal{R}_{l(\mathbf{x},\boldsymbol{\theta})}\}}}{\#\{j : X_j \in \mathcal{R}_{l(\mathbf{x},\boldsymbol{\theta})}\}}. \tag{23}$$

Generalized Regression Forests

- (Athey et al., 2019) pointed out that problems arise in QRF when splitting parent nodes into child nodes, as it tends to split based on regions where the difference in means is large rather than conditional quantiles.
- Expressing quantile regression analysis with an estimating equation, the parameter function we need to estimate is the τ-th quantile of Y given $X = x$, which can be represented as $\theta_\tau(\mathbf{x})$.
- From the perspective of GRF, $\hat{\theta}_{\tau, P(X_i)}$ will partition in a way that the τ-th quantile of the parent node $P(X_i)$ separates the lower observations as much as possible.
- Random forests approximate $E(Y \mid \mathbf{X} = \mathbf{x})$ as an ensemble mean of k regression trees

$$\hat{E}(Y \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) Y_i, \quad w_i(\mathbf{x}) = \frac{\sum_{t=1}^{k} w_i(\mathbf{x}, \boldsymbol{\theta}_t)}{k}. \tag{24}$$

Generalized Regression Forests (cont.)

- In quantile regression forests, computes conditional cumulative distribution function (CDF) of Y, given $\mathbf{X} = \mathbf{x}$.

$$F(y \mid \mathbf{X} = \mathbf{x}) = P(Y \leq y \mid \mathbf{X} = \mathbf{x}) = E\Big(1_{\{Y \leq y\}} \mid \mathbf{X} = \mathbf{x}\Big). \tag{25}$$

- (Meinshausen, 2006) suggested that we can use w_i to estimate the conditional CDF:

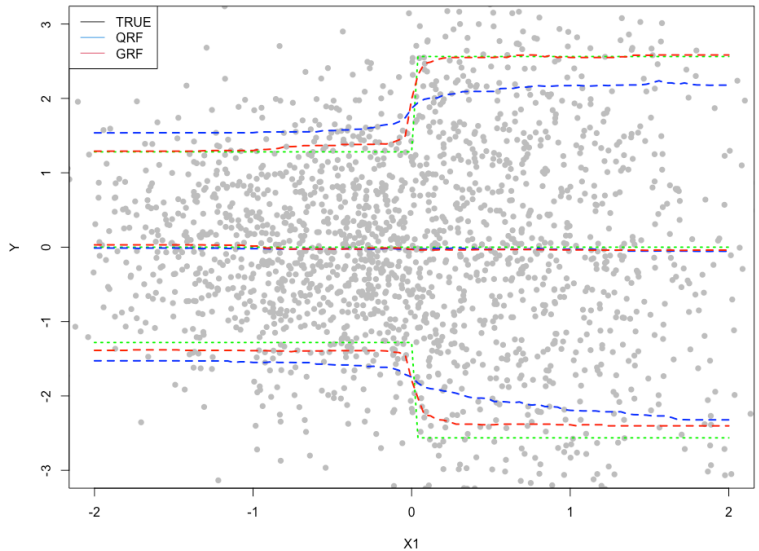
$$\hat{f}(y \mid \mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) I_{\{Y_i \leq y\}}. \tag{26}$$

- Based on the conditional CDF, we can also estimate the conditional τ-quantile:

$$\hat{Q}_\tau(y \mid \mathbf{x}) = \inf\Big\{y : \hat{F}_{Y \mid \mathbf{X}}(y \mid \mathbf{x}) \geq \tau\Big\}. \tag{27}$$

- Application: (Park et al., 2018)

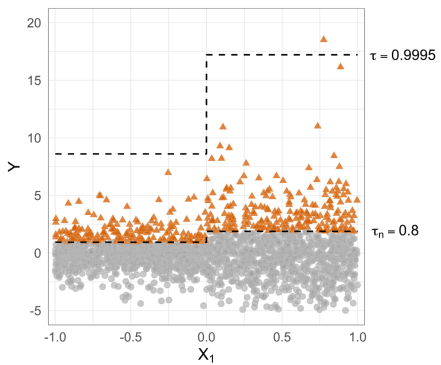
Generalized Regression Forests (cont.)



Extremal Random Forests

- Extremal random forests (Gnecco et al., 2024) : intermediate quantile function estimation + Estimate extreme quantile by using the GPD extrapolation formula

$$\hat{Q}(\tau) \approx \hat{Q}(\tau_n) + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(\frac{1-\tau}{1-\tau_n} \right)^{-\hat{\xi}} - 1 \right]. \tag{28}$$



VAE Approach to Multivariate Extremes

- Image from (Lafon et al., 2023)

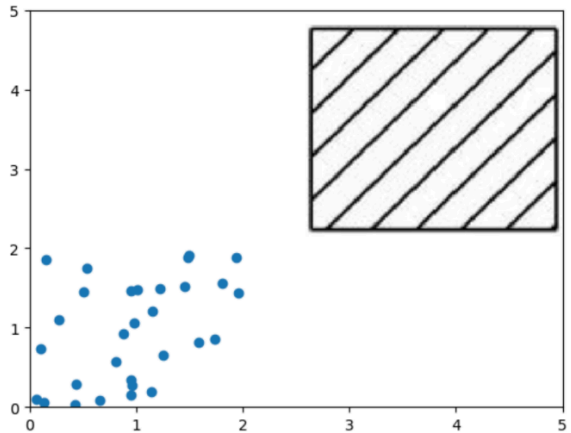


Figure 3: How to sample from observations (blue dots) in extreme regions (black square) to estimate probability of rare events?

Standardized Data

- Random vectors $\mathbf{X} = (X_1, ..., X_d), \quad X_j \geq 0$
- Margins: $X_j \sim F_j, \quad 1 \leq j \leq d$ (continuous)
- Standardization with unit Pareto margins

$$Y_j = \frac{1}{1 - F_j(X_j)}, \quad P(Y_j > v) = \frac{1}{v} \tag{29}$$

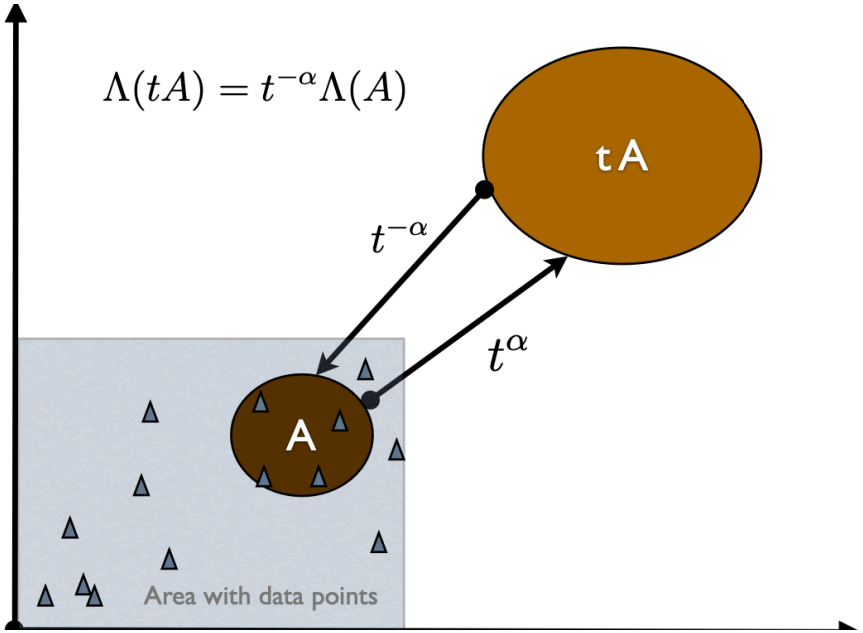
- (Radial homogeneity assumption) $P(Y \in A) \simeq tP(Y \in tA)$
- If we assume that $\mathbf{Y}$ has **multivariate regular variation** (Lafon et al., 2023), for $\mathbf{0} \notin \text{clousure } (A)$ (extreme region),

$$tP(\mathbf{Y} \in tA) \overset{t \to \infty}{\to} \Lambda(A), \tag{30}$$

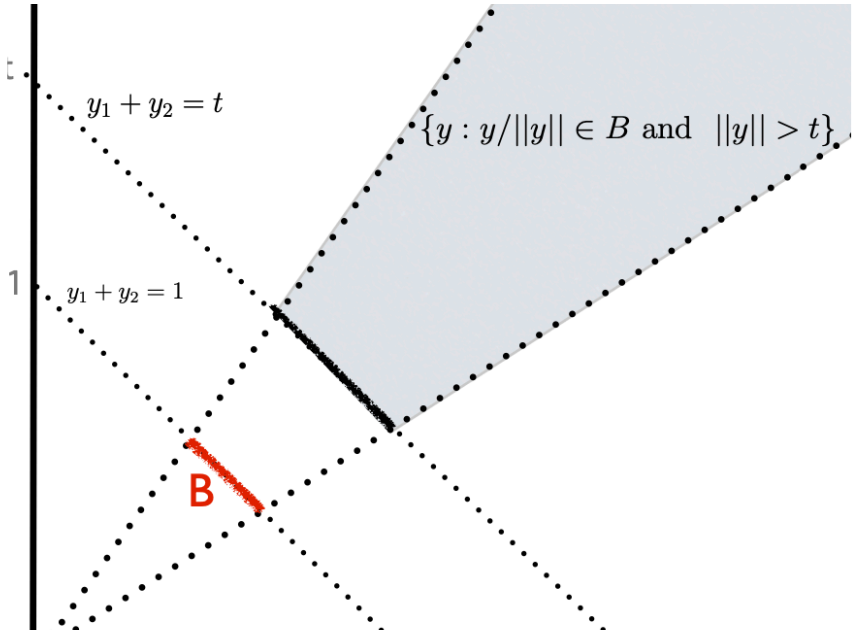
where Λ is called an **exponent measure**.

- Necessarily Λ is homogeneous: $\Lambda(tA) = \frac{1}{t}\Lambda(A)$

Scaling Property



Scaling Property (cont.)



Scaling Property (cont.)

- The above Figure interpretes the scaling property $\Lambda(tA) = \frac{1}{t}\Lambda(A)$ with $\| \mathbf{y} \| = y_1 + y_2$.
- Consider a special case,

$$A = \{\mathbf{y} = (y_1, y_2) : \mathbf{y}/ \| \mathbf{y} \| \in B \ \text{ and } \ \| \mathbf{y} \| > 1\}, \tag{31}$$

  where $\| \mathbf{y} \| = y_1 + y_2$ and B any set belonging to the unit simplex.
- Then

$$\begin{aligned} tA &= \{t\mathbf{y} : \mathbf{y}/ \| \mathbf{y} \| \in B \ \text{ and } \ \| \mathbf{y} \| > 1\} \\ &= \{\mathbf{u} : \mathbf{u}/ \| \mathbf{u} \| \in B \ \text{ and } \| \mathbf{u} \| > t\}, \quad \text{with } \ \mathbf{u} = t\mathbf{y}. \end{aligned} \tag{32}$$

- Let the radial component $R = Y_1 + Y_2 + \cdots + Y_m = \| \mathbf{Y} \|$ and an angular component of the $(m-1)$-dimensional simplex $\mathbf{\Theta} = \frac{\mathbf{Y}}{\|\mathbf{Y}\|}$, and an **angular measure** $\mathbf{S}$ on the corresponding simplex is $\mathbf{S}(B) = \Lambda\{R > 1, \mathbf{\Theta} \in B\}$. Then

$$\Lambda(\{\mathbf{u} : \mathbf{u}/ \| \mathbf{u} \| \in B \ \text{ and } \| \mathbf{u} \| > t\}) = \frac{1}{t}\mathbf{S}(B). \tag{33}$$

Multivariate Regular Variation

- Let $\boldsymbol{X}$ be a random vector in $(\mathbb{R}^+)^m$. We decompose $\boldsymbol{X}$ into
  - A **radial component** $R = X_1 + \cdots + X_m = \|\boldsymbol{X}\|$, and
  - An **angular component** of the $(m-1)$-dimensional simplex $\boldsymbol{\Theta} = \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}$.

Definition 4 (Multivariate Regular Variation): $\boldsymbol{X}$ has **multivariate regular variation** if the two following properties are fulfilled:

1. The **radius** R is regularly varying with tail index α, i.e.,

$$\lim_{t \to +\infty} P(R > \mathrm{tr} |\ R > t) = r^{-\alpha}, \quad r > 0. \tag{34}$$

2. There exist a probability mreasure $\mathbf{S}$ defined on the $(m-1)$-dimensional simplex such that $(R, \boldsymbol{\Theta})$ verifies

$$P(\boldsymbol{\Theta} \in \bullet \mid R > r) \xrightarrow{w} \mathbf{S}(\bullet), \tag{35}$$

where $\xrightarrow{w}$ denotes week convergence. $\mathbf{S}$ is called **angular measure**.

Multivariate Regular Variation (cont.)

- The exponent measure represents the entire joint tail behavior over an unbounded domain, which makes it challenging to work with in practice.
- Therefore, we often turn to the **angular (spectral) measure**, which arises from the polar decomposition of the exponent measure.
- Multivariate regular variation indicates that if the radius is above a sufficiently high threshold, the respective distributions of the radius and the angle can be considered **independent**. (Lafon et al., 2023)
- **Estimation**:
  - Radial component R: Since R has a Pareto tail, we estimate tail index α using Hill-estimator.
  - Angular component $\boldsymbol{\Theta}$: We can use either (1) empirical measure

$$\hat{\mathbf{S}}(A) = \frac{1}{n_{\text{ext}}} \sum_{i=1}^{n_{\text{ext}}} \mathbb{1}\{\Theta_i \in A\} \tag{36}$$

  or (2) kernel smoothing to approximate the angular density.

VAE to Sample Multivariate Extremes

- Image from (Lafon et al., 2023)

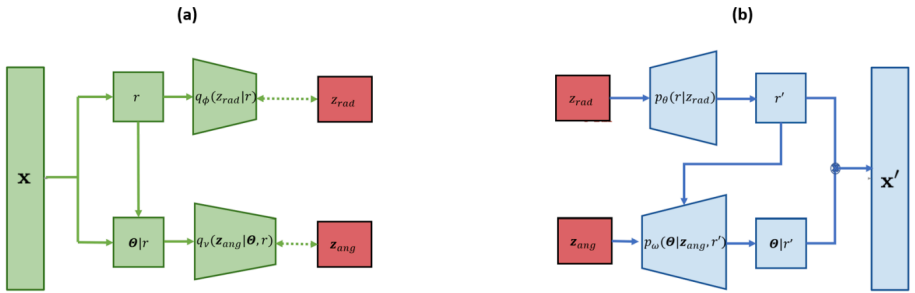


Figure 2: Global architecture of our approach with **(a)** the probabilistic encoders and **(b)** the probabilistic decoders. Ideally, distributions of $\mathbf{x}$ and $\mathbf{x}'$ are similar. Solid arrows show a causal link between the different blocks. Dashed double arrows in **(a)** indicate that the distributions in the pointed blocks are compared using a Kullback-Leibler divergence criterion (Equation 2).

VAE to Sample Multivariate Extremes (cont.)

- Although the angular measure $\mathbf{S}$ should be estimated, Lafon et al. (2023) uses transformed variable approach based on the conditional distribution of $\mathbf{\Theta} \mid R$.
- To generate a sample $\mathbf{x}^{(i)}$ of a multivariate regularly varying random vector, Lafon et al. (2023) proposed three-step VAE scheme:

1. Using a VAE, a radius $r^{(i)}$ is drawn from a univariate heavy-tail distribution R.
2. Conditionally on the drawn radius $r^{(i)}$, sample $\mathbf{\Theta}^{(i)}$ an element of the $(m-1)$-dimensional simplex from the conditional distribution $\mathbf{\Theta} \mid \left[R = r^{(i)}\right]$ while forcing the independence between radius R and angle $\mathbf{\Theta}$ for larger value of the radius.
3. Multiply component-wise the angle vector by the radius to obtain the desired sample, i.e., $\mathbf{x}^{(i)} = r^{(i)}\mathbf{\Theta}^{(i)}$.

Smpling heavy-tailed R

- Goal: model R through a latent variable Z_{rad}

VAE to Sample Multivariate Extremes (cont.)

Conditions in (Lafon, Naveau, and Fablet 2023)

1. Z_{rad} follows the inverse-gamma distribution $f_{\text{Inv }\Gamma}(z_{\text{rad}}; \alpha, \beta)$,

$$f_{\text{Inv }\Gamma}(z_{\text{rad}}; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z_{\text{rad}}^{-\alpha-1} \exp(-\beta/z_{\text{rad}}), \quad \alpha, \beta, z_{\text{rad}} > 0. \tag{37}$$

2. R is linked to Z_{rad} throughout a multiplicative model with a positive random coefficient W,

$$R \overset{d}{=} W \times Z_{\text{rad}}, \tag{38}$$

where $\overset{d}{=}$ corresponds to a equality in distribution, and the random variable W is absolutely continuous and independent of Z_{rad}. We also assume that $0 < E[W^{\alpha+\epsilon}] < \infty$ for some positive ϵ.

- The inverse-gamma distribution is **heavy-tailed** with tail index α and has a positive support.
- The moment condition $0 < E[W^{\alpha+\epsilon}] < \infty$ means that W has a significantly lighter tail than Z_{rad}.

VAE to Sample Multivariate Extremes (cont.)

! Important

- If two conditions hold, by Breiman's lemma (Kulik, 2020) R is heavy-tailed with tail index α.
- In particular, since GPD can be represented as an exponential-gamma mixture, if W follows an exponential distribution with scale parameter c, then R follows a generalized Pareto distribution with $\xi = \frac{1}{\alpha}$ and $\tilde{\sigma} = \frac{\beta c}{\alpha}$.

Sampling from heavy-tailed radius distributions

$$\begin{aligned} p_\alpha(z_{\text{rad}}) &= f_{\text{Inv }\mathbf{\Gamma}}(z_{\text{rad}}; \alpha, 1), \\ p_\theta(r \mid z_{\text{rad}}) &= f_{\mathbf{\Gamma}}(r; \alpha_\theta(z_{\text{rad}}), \beta_\theta(z_{\text{rad}})), \\ q_\phi(z_{\text{rad}} \mid r) &= f_{\text{Inv }\mathbf{\Gamma}}\big(z_{\text{rad}}; \alpha_\phi(r), \beta_\phi(r)\big), \end{aligned} \tag{39}$$

where $\alpha_\theta, \beta_\theta, \alpha_\phi, \beta_\phi$ are ReLU neural network functions with parameters θ and ϕ.

VAE to Sample Multivariate Extremes (cont.)

Sampling on the multivariate simplex

$$
\begin{aligned}
p(\mathbf{z}_{\text{ang}}) &= \mathcal{N}(\mathbf{z}_{\text{ang}}; \mathbf{0}, \mathbf{I}_n), \\
p_\nu(\boldsymbol{\Theta} \mid \mathbf{z}_{\text{ang}}, r) &= \int_{\mathbf{S}(\boldsymbol{\Theta})} \mathcal{N}\left(\mathbf{s}; \mu_\nu(\mathbf{z}_{\text{ang}}, r), \text{diag}\ \left(\sigma_\nu(\mathbf{z}_{\text{ang}}, r)\right)^2\right), \\
q_\omega(\mathbf{z}_{\text{ang}} \mid \boldsymbol{\Theta}, r) &= \mathcal{N}\left(\mathbf{z}_{\text{ang}}; \mu_\omega(\boldsymbol{\Theta}, r), \text{diag}\ \left(\sigma_\omega(\boldsymbol{\Theta}, r)\right)^2\right),
\end{aligned}
\tag{40}
$$

where n is the dimension of the latent space, $\mu_\nu, \sigma_\nu, \mu_\omega, \sigma_\omega$ are neural network functions with parameters ν and ω.

- To generate $p_\nu(\boldsymbol{\Theta} \mid \mathbf{z}_{\text{ang}}, r)$, Lafon et al. (2023) sample from $\mathcal{N}\left(\mu_\nu(\mathbf{z}_{\text{ang},r}), \text{diag}\ \left(\sigma_\nu(\mathbf{z}_{\text{ang}}, r)\right)^2\right)$ and then projectin on the sphere through a projection

$$
\boldsymbol{\Pi}(\mathbf{s}) = \frac{\mathbf{s}}{\| \mathbf{s} \|}, \tag{41}
$$

  where the norm is the $\mathcal{L}_1$ −norm.

Angular Measure Estimation Result

- Image from (Lafon et al., 2023)

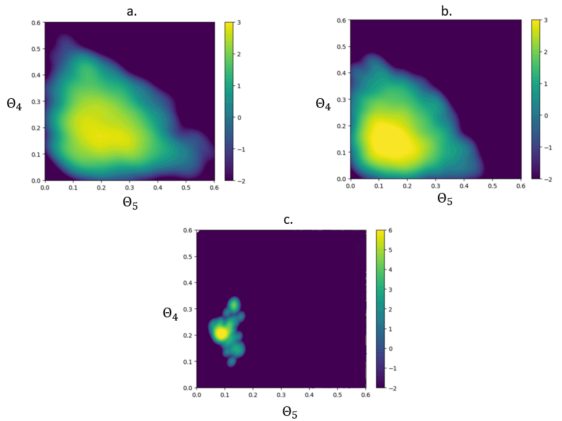
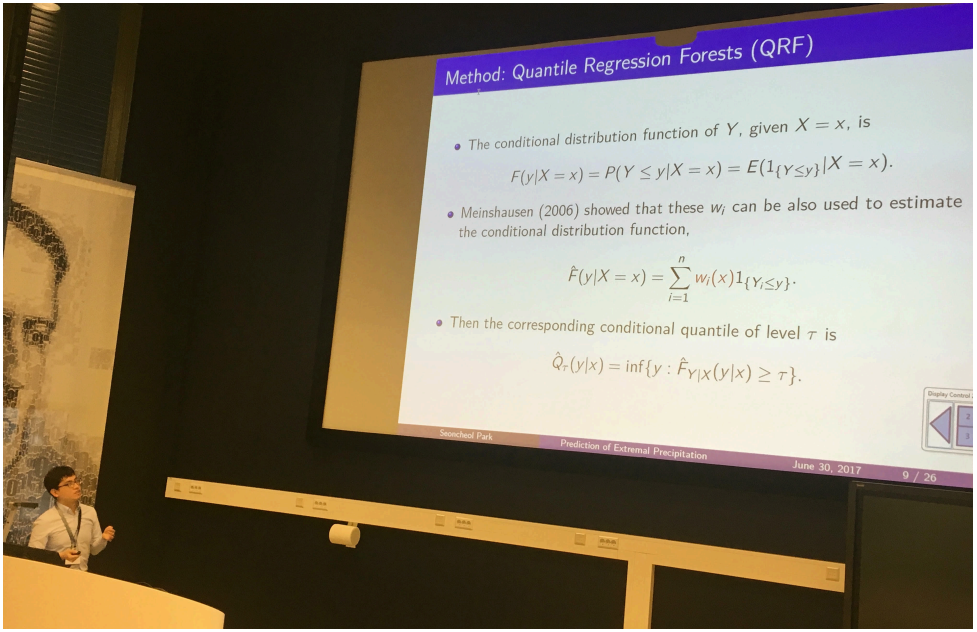


Figure 6: Log probability of the angular measure obtained with a. ExtVAE, b. true distribution, c. ParetoGAN, projected on axes 4 and 5 (named θ_4 and θ_5). For ParetoGAN, the estimation is based on 10000 samples at a high value of radius, typically above 10, which corresponds to the upper percentile of R_1 distribution.

Extreme Value Analysis Conference



Bibliography

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-aos1709

Chernozhukov, V. (2005). Extremal quantile regression. *The Annals of Statistics*, *33*(2), 806–839.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London.

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.

Gnecco, N., Terefe, E. M., & Engelke, S. (2024). Extremal Random Forests. *Journal of the American Statistical Association*, 1–14. https://doi.org/10.1080/01621459.2023.2300522

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, *46*(1), 33–50. https://www.jstor.org/stable/1913643

Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*(4), 673–680.

Bibliography (cont.)

Kulik, R. (2020). *Heavy-tailed time series* (P. Soulier, Ed.). Springer Nature. https://doi.org/10.1007/978-1-0716-0737-4

Lafon, N., Naveau, P., & Fablet, R. (2023,). *A VAE Approach to Sample Multivariate Extremes*. arXiv. https://doi.org/10.48550/ARXIV.2306.10987

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, *7*, 983–999.

Nair, J., Wierman, A., & Zwart, B. (2022). *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. Cambridge University Press. https://doi.org/10.1017/9781009053730

Park, S., Kwon, J., Kim, J., & Oh, H.-S. (2018). Prediction of extremal precipitation by quantile regression forests: from SNU Multiscale Team. *Extremes*, *21*(3), 463–476. https://doi.org/10.1007/s10687-018-0323-y

Wang, H. J., & Li, D. (2013). Estimation of Extreme Conditional Quantiles Through Power Transformation. *Journal of the American Statistical Association*, *108*(503), 1062–1074. https://doi.org/10.1080/01621459.2013.820134