

Machine Learning goes to extremes

Baptiste Leroux

October 14, 2026

Table of contents

Introduction

Extreme Value Theory

- Univariate Extreme Value Theory

- Multivariate extremes

Learning with extremes

- An introduction to statistical learning

- Machine learning with regular variation

Introduction

Extreme Values: Motivation

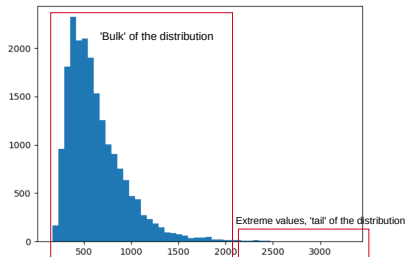


Extreme Values: Motivation



Extreme Values: Motivations

Extreme value Theory is the study of "the tails" rather than "the bulk" of a distribution



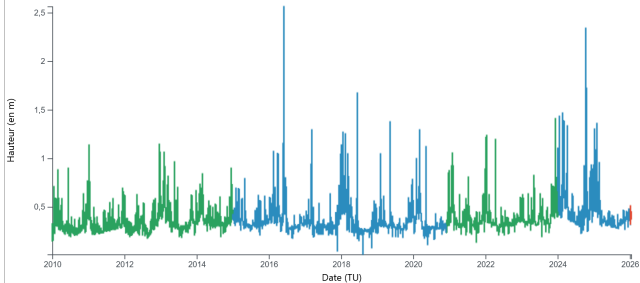
- ▶ Study events with potentially devastating impacts
- ▶ Applications in earth science, finance, insurance, telecommunications...
- ▶ Extrapolation, Out-of-domain generalization

Yvette river: An example where EVT is relevant



**Hauteur instantanée minimale n journalière (n=1, non glissant) - Données les plus valides de l'entité -
F466 0001 01 - L'Yvette à Villebon-sur-Yvette - du 01/01/2010 00:00 au 14/01/2026 23:59 (TU)**

Généré le 15/01/2026 09:12 (TU)



Extreme Value Theory

A CLT for maxima?

- ▶ Extremes $\longrightarrow$ Maxima of random variables: Given a sequence $X_1, \dots, X_n$ of *i.i.d.* random variables, we consider

$$\bigvee_{i=1}^n X_i = \max(X_1, \dots, X_n).$$

A CLT for maxima?

- ▶ Extremes $\longrightarrow$ Maxima of random variables: Given a sequence $X_1, \dots, X_n$ of *i.i.d.* random variables, we consider

$$\bigvee_{i=1}^n X_i = \max(X_1, \dots, X_n).$$

- ▶ CLT: If $(X_i)_{i \in \mathbb{N}^*}$ is a sequence of *i.i.d.* random variables, with $\mu = \mathbb{E}[X_1], \sigma = \sqrt{\mathbb{E}[X_1^2]} < \infty$, then

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

- ▶ Can we find a_n, b_n and some non degenerate Z such that

$$\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} Z?$$

Maximum Domain of Attraction

The convergence

$$\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} Z$$

is called the **maximum domain of attraction** (MDA) condition.

Maximum Domain of Attraction

The convergence

$$\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} Z$$

is called the **maximum domain of attraction** (MDA) condition.

Denote $F(x) = \mathbb{P}(X_1 \leq x)$ the cdf of X_1 . We have:

$$\begin{aligned} \mathbb{P}\left(\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \leq x\right) &= \mathbb{P}\left(\bigcap_{i=1}^n X_i \leq a_n x + b_n\right) \\ &= F^n(a_n x + b_n) \end{aligned}$$

Maximum Domain of Attraction

The convergence

$$\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} Z$$

is called the **maximum domain of attraction** (MDA) condition. Denote $F(x) = \mathbb{P}(X_1 \leq x)$ the cdf of X_1 . We have:

$$\begin{aligned} \mathbb{P}\left(\frac{\bigvee_{i=1}^n X_i - b_n}{a_n} \leq x\right) &= \mathbb{P}\left(\bigcap_{i=1}^n X_i \leq a_n x + b_n\right) \\ &= F^n(a_n x + b_n) \end{aligned}$$

The MDA condition can be rephrased as:

$$F^n(a_n x + b_n) \xrightarrow[n \rightarrow \infty]{} G(x)$$

To obtain this convergence, the distribution G needs to be **max-stable**:

$$G^n(a_n x + b_n) = G(x) \tag{1}$$

Max stable distributions

Theorem

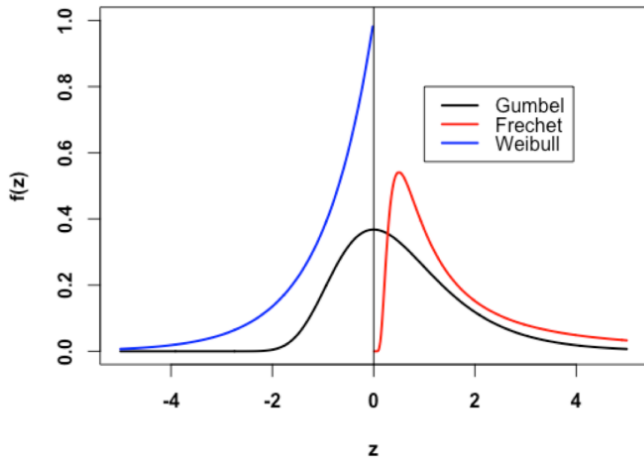
If G is max-stable, then it is one of the three types:

1. *Fréchet: $G(x) = \exp(-x^{-\alpha})\mathbb{1}_{x \geq 0}$, $\alpha > 0$*
2. *Weibull: $G(x) = \exp(-(-x)^{-\alpha})$ if $x < 0$, $G(x) = 1$ if $x \geq 0$, with $\alpha < 0$.*
3. *Gumbel: $G(x) = \exp(-\exp(-x))$*

Another formulation: If the Maximum Domain of attraction condition is satisfied, then there exists μ, ξ, σ such that

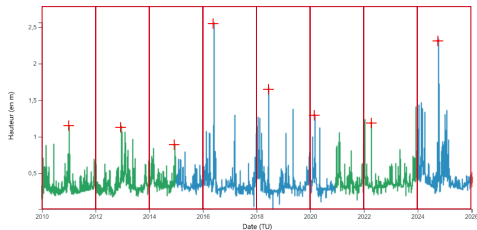
$$G(x) = \exp\left(-\left(1 + \xi \frac{x - \mu}{\sigma}\right)_+^{-1/\xi}\right)$$

G follows a **Generalized Pareto Distribution (GPD)**.



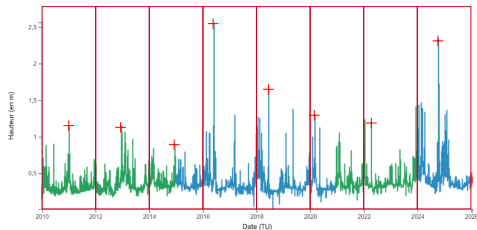
Block maxima and Peak-over-threshold

Block maxima approach

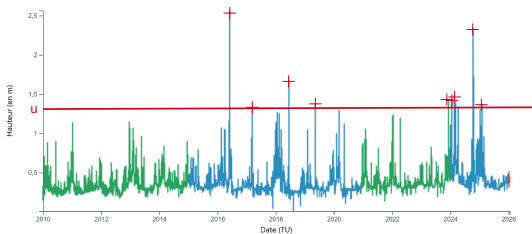


Block maxima and Peak-over-threshold

Block maxima approach



Alternatively: Peak-over-threshold (pot)



Excesses above thresholds (informal)

One alternative way to deal with extremes: Look at the behaviour of X in the tails: For some threshold u , we consider

$$\mathbb{P}\left(\frac{X - u}{\sigma(u)} > t \mid X > u\right)$$

The maximum domain of attraction condition is equivalent to:

$$\mathbb{P}\left(\frac{X - u}{\sigma(u)} > t \mid X > u\right) \xrightarrow{u \rightarrow \infty} (1 + \xi t)_+^{-\frac{1}{\xi}}$$

Regularly varying functions

Definition

b is regularly varying with tail index α if

$$\frac{b(tx)}{b(t)} \xrightarrow{t \rightarrow \infty} x^\alpha$$

if $\alpha = 0$: b is slowly varying.

Regularly varying functions

Definition

b is regularly varying with tail index α if

$$\frac{b(tx)}{b(t)} \xrightarrow[t \rightarrow \infty]{} x^\alpha$$

if $\alpha = 0$: b is slowly varying.

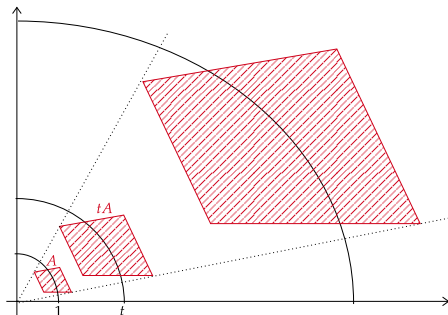
If the distribution F belongs to the Fréchet maximum domain of attraction ($\alpha > 0$): $1 - F$ is regularly varying with the same tail index $\alpha > 0$.

In this case: for any $x > 0, t > 0$:

$$\frac{\mathbb{P}(X > tx)}{\mathbb{P}(X > t)} \xrightarrow[t \rightarrow \infty]{} x^{-\alpha}$$

Towards regular variation

Goal: We see $\mu_t(A) = \frac{\mathbb{P}(X \in tA)}{\mathbb{P}(X > t)}$ as a measure.



Notion of convergence for the measure μ_t as $t \rightarrow \infty$?

Vague convergence

$$\mathbb{M}_0 = \{\mu \text{ Borel measure on } \mathbb{R}^d \setminus \{0\} : \forall r > 0, \mu(\mathbb{R}^d \setminus B_r) < \infty\}$$

(Borel measures finite on sets bounded away from zero)

$$\mathcal{C}_0 = \{f \text{ continuous bounded, } f|_{B_r} = 0 \text{ for some } r > 0\}$$

Vague convergence

$$\mathbb{M}_0 = \{\mu \text{ Borel measure on } \mathbb{R}^d \setminus \{0\} : \forall r > 0, \mu(\mathbb{R}^d \setminus B_r) < \infty\}$$

(Borel measures finite on sets bounded away from zero)

$$\mathcal{C}_0 = \{f \text{ continuous bounded, } f|_{B_r} = 0 \text{ for some } r > 0\}$$

Definition

The sequence of measures μ_t converges weakly to μ if, for any $f \in \mathcal{C}_0$,

$$\int f d\mu_t \xrightarrow[t \rightarrow \infty]{} \int f d\mu$$

Theorem (Portmanteau's theorem)

μ_t converges weakly to μ if and only if, for all A such that $0 \notin \overline{A}$ and $\mu(\partial A) = 0$, $\mu_t(A) \xrightarrow[t \rightarrow \infty]{} \mu(A)$.

Theorem

The following are equivalent:

- 1. F belongs to the maximum domain of attraction of the Fréchet distribution (for some $\alpha > 0$).*
- 2. $1 - F$ is regularly varying with tail index α*
- 3. Let $\mu_n(\cdot) = \mathbb{P}(X \in a_n \cdot)$. μ_n converges vaguely to ν_α defined by $\nu_\alpha([x, \infty)) = x^{-\alpha}$*

Multivariate extremes

- ▶ The same work as in the univariate case can be done (CLT for Maxima, Max-stable distributions, ...)
- ▶ Here we only consider regular variation in the multivariate case

Multivariate extremes

- ▶ The same work as in the univariate case can be done (CLT for Maxima, Max-stable distributions, ...)
- ▶ Here we only consider regular variation in the multivariate case

Definition

We say that $X \in \mathbb{R}^d$ is regularly varying if there exists some regularly varying function $b(t)$ with tail index $\alpha > 0$, and a non-zero Borel measure μ such that, for any A bounded away from 0 with $\mu(\partial A) = 0$

$$b(t)\mathbb{P}(X \in tA) \xrightarrow[t \rightarrow \infty]{} \mu(A)$$

Multivariate extremes: Regular Variation II

Theorem

If X is regularly varying with limit measure μ , then:

- 1. The measure μ is homogeneous:*

$$\mu(tA) = t^{-\alpha} \mu(A)$$

- 2. We can always assume that the measure μ is such that $\mu\{X: \|X\| \geq 1\} = 1$, and assume $b(t) = \frac{1}{\mathbb{P}(\|X\| \geq t)}$*

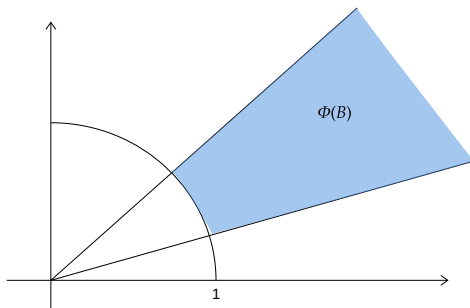
Homogeneity of μ : polar coordinates are useful!

Angular measure

Definition

Let $\theta(x) = \frac{x}{\|x\|}$. The angular measure of $B \subset \mathbb{S}$ is

$$\Phi(B) = \mu\{\|X\| \geq 1, \theta(X) \in B\}.$$



Regular variation: Implications

Theorem

If X is multivariate, regularly varying with limit measure μ , then:

- ▶ *For any $B \subset \mathbb{S}$, $\mu\{\|x\| \geq t, \theta(x) \in B\} = t^{-\alpha} \Phi(B)$*
- ▶ *The convergence in distribution holds:*

$$(X/t \mid \|X\| \geq t) \xrightarrow[t \rightarrow \infty]{} X_\infty$$

with $\|X_\infty\| \perp \theta(X_\infty)$ and $\|X_\infty\|$ follows a Pareto distribution with parameter α .

Pareto distribution (with scale parameter = 1): $\mathbb{P}(X > t) = t^{-\alpha}$ for $t > 1$.

Unit Pareto: $\mathbb{P}(X > t) = t^{-1}$; case where $\alpha = 1$.

Standardization to Unit-Pareto Margins

It is often necessary to standardize each margin X_j of the vector X to a unit Pareto. Consider

$$V_j = \frac{1}{1 - F_j(X_j)}$$

Standardization to Unit-Pareto Margins

It is often necessary to standardize each margin X_j of the vector X to a unit Pareto. Consider

$$V_j = \frac{1}{1 - F_j(X_j)}$$

For any $t > 0$:

$$\begin{aligned}\mathbb{P}(V_j > t) &= \mathbb{P}(1 - F_j(X_j) \leq \frac{1}{t}) \\ &= \mathbb{P}(U \leq \frac{1}{t}) \text{ where } U \sim \mathcal{U}([0, 1]) \\ &= \frac{1}{t}\end{aligned}$$

and V_j is a standard Pareto.

Standardization: How to do it in practice

In practice: F_j is unknown. Standardization relies on an estimation of the distribution of each margin X_j .

- In the bulk of the distribution:

$$\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_j \leq x\}$$

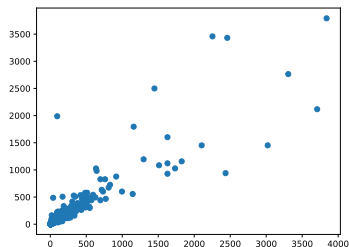
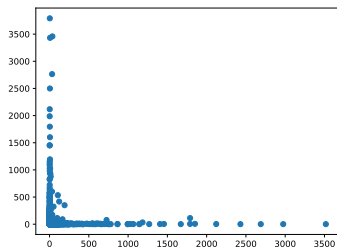
- In the tail of the distribution (above a threshold u): Leverage univariate Extreme Value Theory and fit a generalized Pareto distribution:

$$\hat{F}_j(x) = \exp\left(-\left(1 + \hat{\xi}_j \frac{x - u}{\hat{\sigma}_j}\right)_+^{-1/\hat{\xi}_j}\right) \quad (2)$$

for some estimation of ξ_j, σ_j .

Asymptotic dependence, asymptotic independence

Take $d = 2$ for simplicity, $X = (X_1, X_2)$ regularly varying. Several situations may arise...



- Extreme values for X_1, X_2 may or may not happen simultaneously. Notion of **asymptotic dependence** is introduced to model this phenomenon.

Learning with extremes

Machine learning: what are we trying to achieve?

$(X, Y) \sim \mathbb{P}$. The goal is to predict Y by $h(X)$.

How accurate is the prediction? Given a loss function ℓ , the mean error is measured by the risk

$$\mathcal{R}(h) = \mathbb{E}[\ell(Y, h(X))]$$

Minimum achievable risk: Bayes risk

$$\mathcal{R}^* = \inf\{\mathcal{R}(h): h \text{ measurable}\}$$

Machine learning: what are we trying to achieve?

$(X, Y) \sim \mathbb{P}$. The goal is to predict Y by $h(X)$.

How accurate is the prediction? Given a loss function ℓ , the mean error is measured by the risk

$$\mathcal{R}(h) = \mathbb{E}[\ell(Y, h(X))]$$

Minimum achievable risk: Bayes risk

$$\mathcal{R}^* = \inf\{\mathcal{R}(h): h \text{ measurable}\}$$

Goal: Get a predictor h which achieves the Bayes risk (the Bayes predictor h^*).

Problem: The probability distribution $\mathbb{P}$ is unknown. We only have access to a sample $((X_1, Y_1), \dots, (X_n, Y_n)) \sim \mathbb{P}$.

→ Need another strategy to obtain h .

Empirical Risk Minimization

Idea: Replace the true risk by an empirical estimate based on the sample $((X_1, Y_1), \dots, (X_n, Y_n))$

$$\mathcal{R}(h) = \mathbb{E}[\ell(Y, h(X))] \longleftrightarrow \hat{\mathcal{R}}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i)$$

Empirical Risk Minimization

Idea: Replace the true risk by an empirical estimate based on the sample $((X_1, Y_1), \dots, (X_n, Y_n))$

$$\mathcal{R}(h) = \mathbb{E}[\ell(Y, h(X))] \longleftrightarrow \hat{\mathcal{R}}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i)$$

Restrict the search of $\hat{h}$ to some set $\mathcal{F}$: new predictor $\hat{h}_{\mathcal{F}}$

$$\hat{h}_{\mathcal{F}} \in \operatorname{argmin}\{\hat{\mathcal{R}}(h) : h \in \mathcal{F}\}$$

Empirical Risk Minimization

Idea: Replace the true risk by an empirical estimate based on the sample $((X_1, Y_1), \dots, (X_n, Y_n))$

$$\mathcal{R}(h) = \mathbb{E}[\ell(Y, h(X))] \longleftrightarrow \hat{\mathcal{R}}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i)$$

Restrict the search of $\hat{h}$ to some set $\mathcal{F}$: new predictor $\hat{h}_{\mathcal{F}}$

$$\hat{h}_{\mathcal{F}} \in \operatorname{argmin}\{\hat{\mathcal{R}}(h) : h \in \mathcal{F}\}$$

Example: $\mathcal{F} = \{x \mapsto \langle w, x \rangle, w \in \mathbb{R}^d\} \longrightarrow$ Linear regression.

Statistical guarantees

Once we have our predictor $\hat{h}$, how do we certify its good performance?

Statistical guarantees

Once we have our predictor $\hat{h}$, how do we certify its good performance?

- Performance is measured by the **excess risk**:

$$\mathcal{R}(\hat{h}) - \mathcal{R}^*$$

- Provide a **high probability bound** for this excess risk: with probability $1 - \delta$,

$$\mathcal{R}(\hat{h}) - \mathcal{R}^* \leq f(\delta, n)$$

Statistical guarantees

Once we have our predictor $\hat{h}$, how do we certify its good performance?

- Performance is measured by the **excess risk**:

$$\mathcal{R}(\hat{h}) - \mathcal{R}^*$$

- Provide a **high probability bound** for this excess risk: with probability $1 - \delta$,

$$\mathcal{R}(\hat{h}) - \mathcal{R}^* \leq f(\delta, n)$$

Relies on a **decomposition** of the excess risk

Decomposition of the excess risk in the ERM framework

- Our estimator is

$$\hat{h}_{\mathcal{F}} \in \operatorname{argmin}_{h \in \mathcal{F}} \hat{\mathcal{R}}(h).$$

Denote

$$h_{\mathcal{F}}^* \in \operatorname{argmin}_{h \in \mathcal{F}} \mathcal{R}(h)$$

Decomposition of the excess risk in the ERM framework

- Our estimator is

$$\hat{h}_{\mathcal{F}} \in \operatorname{argmin}_{h \in \mathcal{F}} \hat{\mathcal{R}}(h).$$

Denote

$$h_{\mathcal{F}}^* \in \operatorname{argmin}_{h \in \mathcal{F}} \mathcal{R}(h)$$

- Decomposition of the excess risk:

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}^* = \underbrace{\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*)}_{\text{Variance term}} + \underbrace{\mathcal{R}(h_{\mathcal{F}}^*) - \mathcal{R}^*}_{\text{Approximation error}}$$

Control of the excess risk

Decomposition we obtained:

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}^* = \underbrace{\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*)}_{\text{Variance term}} + \underbrace{\mathcal{R}(h_{\mathcal{F}}^*) - \mathcal{R}^*}_{\text{Approximation error}}$$

In the ERM framework: the variance term can be bounded with

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*) \leq 2 \sup_{h \in \mathcal{F}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)|$$

Control of the excess risk

Decomposition we obtained:

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}^* = \underbrace{\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*)}_{\text{Variance term}} + \underbrace{\mathcal{R}(h_{\mathcal{F}}^*) - \mathcal{R}^*}_{\text{Approximation error}}$$

In the ERM framework: the variance term can be bounded with

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*) \leq 2 \sup_{h \in \mathcal{F}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)|$$

Bias-variance tradeoff for the choice of $\mathcal{F}$: Broader $\mathcal{F}$ reduces the approximation error but inflates the variance.

Control of the excess risk

Decomposition we obtained:

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}^* = \underbrace{\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*)}_{\text{Variance term}} + \underbrace{\mathcal{R}(h_{\mathcal{F}}^*) - \mathcal{R}^*}_{\text{Approximation error}}$$

In the ERM framework: the variance term can be bounded with

$$\mathcal{R}(\hat{h}_{\mathcal{F}}) - \mathcal{R}(h_{\mathcal{F}}^*) \leq 2 \sup_{h \in \mathcal{F}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)|$$

Bias-variance tradeoff for the choice of $\mathcal{F}$: Broader $\mathcal{F}$ reduces the approximation error but inflates the variance.

Under reasonable assumptions (VC theory) on $\mathcal{F}$: There is a high probability bound on the variance term, typically of order $\frac{1}{\sqrt{n}}$.

Learning with extremes: Framework

X, Y random variables, $Y \in \mathbb{R}, X \in \mathbb{R}^d$. $(X, Y) \sim \mathbb{P}$.

Goal: estimate $m(x)$ the conditional expectation of Y given $X = x$:

$$m(x) = \mathbb{E}[Y \mid X = x]$$

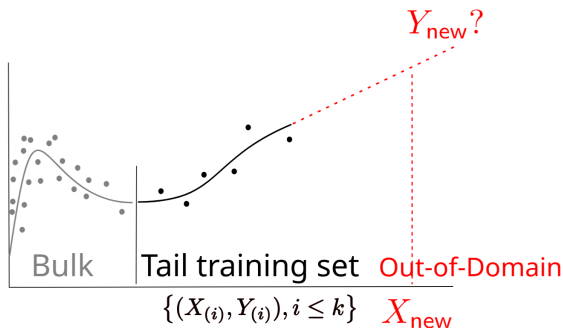
Learning with extremes: Framework

X, Y random variables, $Y \in \mathbb{R}, X \in \mathbb{R}^d$. $(X, Y) \sim \mathbb{P}$.

Goal: estimate $m(x)$ the conditional expectation of Y given $X = x$:

$$m(x) = \mathbb{E}[Y \mid X = x]$$

with a twist: we want a reliable estimation **when the covariate X is large**



Usual strategy

The conditional expectation minimizes the quadratic risk

$$\mathcal{R}(h) = \mathbb{E} \left[(Y - h(X))^2 \right]$$

Usual strategy

The conditional expectation minimizes the quadratic risk

$$\mathcal{R}(h) = \mathbb{E} \left[(Y - h(X))^2 \right]$$

Strategy: Empirical Risk minimization (ERM) The predictor is

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, h(X_i))$$

Usual strategy

The conditional expectation minimizes the quadratic risk

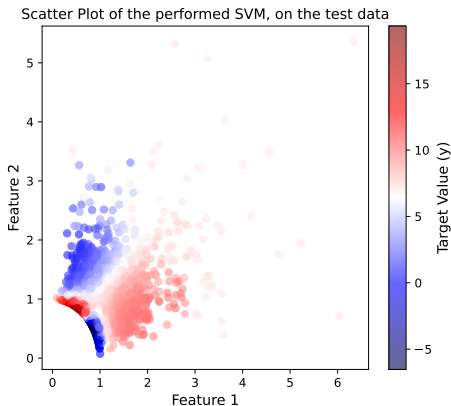
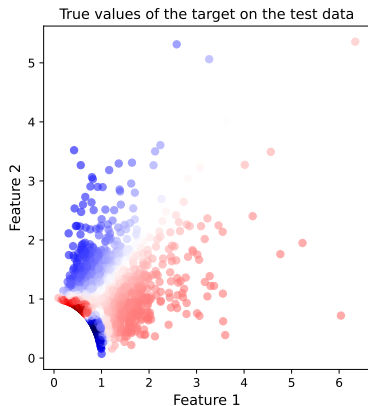
$$\mathcal{R}(h) = \mathbb{E} \left[(Y - h(X))^2 \right]$$

Strategy: Empirical Risk minimization (ERM) The predictor is

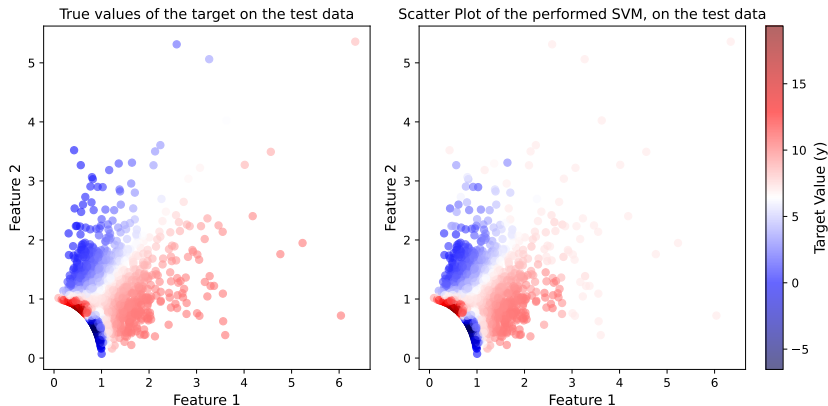
$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, h(X_i))$$

Question: Is this estimator reliable when n is large?

X, Y random variables, $Y \in \mathbb{R}, X \in \mathbb{R}^2$. $(X, Y) \sim \mathbb{P}$. training dataset $D_n = ((X_1, Y_1) \dots, (X_n, Y_n))$.



X, Y random variables, $Y \in \mathbb{R}, X \in \mathbb{R}^2$. $(X, Y) \sim \mathbb{P}$. training dataset $D_n = ((X_1, Y_1) \dots, (X_n, Y_n))$.



→ Poor performance in the 'extreme region' of the dataset.

Why does this happen?

The algorithm aims at minimizing:

$$\mathbb{E} [\ell(h(X), Y)]$$

With the following decomposition:

$$\begin{aligned} \mathbb{E} [\ell(h(X), Y)] &= \mathbb{P}(\|X\| \leq t) \mathbb{E} [\ell(h(X), Y) \mid \|X\| \leq t] \\ &\quad + \underbrace{\mathbb{P}(\|X\| \geq t) \mathbb{E} [\ell(h(X), Y) \mid \|X\| \geq t]}_{\text{negligible for large } t} \end{aligned}$$

Why does this happen?

The algorithm aims at minimizing:

$$\mathbb{E} [\ell(h(X), Y)]$$

With the following decomposition:

$$\begin{aligned} \mathbb{E} [\ell(h(X), Y)] &= \mathbb{P}(\|X\| \leq t) \mathbb{E} [\ell(h(X), Y) \mid \|X\| \leq t] \\ &\quad + \underbrace{\mathbb{P}(\|X\| \geq t) \mathbb{E} [\ell(h(X), Y) \mid \|X\| \geq t]}_{\text{negligible for large } t} \end{aligned}$$

The algorithm may ignore the behavior for $\|X\|$ large!

Looking for a new strategy

- ▶ Conditional quadratic risk: Given a threshold $t > 0$,

$$\mathcal{R}_t(f) = \mathbb{E}[\ell_\tau(Y, f(X)) \mid \|X\| \geq t]$$

- ▶ Asymptotic quadratic risk:

$$\mathcal{R}_\infty(f) = \limsup_{t \rightarrow \infty} \mathcal{R}_t(f)$$

- ▶ Predictor: Minimizer of the asymptotic quadratic risk $\mathcal{R}_\infty$

First-component regular variation

Assumption

The pair (X, Y) is first-component regularly varying:

$$(X/t, Y \mid \|X\| \geq t) \xrightarrow[t \rightarrow \infty]{\mathcal{L}} (X_\infty, Y_\infty)$$

First-component regular variation

Assumption

The pair (X, Y) is first-component regularly varying:

$$(X/t, Y \mid \|X\| \geq t) \xrightarrow[t \rightarrow \infty]{\mathcal{L}} (X_\infty, Y_\infty)$$

- ▶ **Extreme risk** can be defined:

$$\mathcal{R}_{P_\infty}(f) = \mathbb{E}[\ell(Y_\infty, f(X_\infty))]$$

- ▶ $m_{P_\infty}(x)$ the conditional expectation of Y_∞ given $X_\infty = x$:

$$m_{P_\infty}(x) = \mathbb{E}[Y_\infty \mid X_\infty = x]$$

- ▶ m_{P_∞} minimizes the extreme quadratic risk.

Properties of m_{P_∞}

Theorem

1. *Under the first-component regular variation assumption, the extreme conditional expectation is of angular type, that is, for all $x \in \mathbb{R}^d$,*

$$m_{P_\infty}(x) = m_{P_\infty}(\theta(x))$$

2. *If the target Y is bounded, if m_{P_∞} is continuous and satisfies the regularity assumption*

$$\mathbb{E}[|m_{P_\infty}(X) - m(X)| \mid \|X\| \geq t]$$

*then m_{P_∞} **also minimizes the asymptotic risk** $\mathcal{R}_\infty$*

New strategy

The previous theorem suggests to only learn from extreme **angles**:

- ▶ Approximate a sample $((\theta_{\infty,1}, Y_{\infty,1}), \dots, (\theta_{\infty,k}, Y_{\infty,k}))$ with the most extreme observations: Sort the observations by decreasing magnitude of the norm for X :

$$\|X_{(1)}\| \geq \dots \geq \|X_{(n)}\|$$

and retain only the first k most extreme observations.

- ▶ New empirical risk minimization strategy:

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{F}} \frac{1}{k} \sum_{i=1}^k \ell(Y_{(i)}, h(\theta(X_{(i)})))$$

- ▶ Estimated conditional expectation function: $\hat{m}(x) = \hat{h} \circ \theta(x)$

Statistical guarantees

Theorem

Consider the estimated quantile function $\hat{h}$ defined previously. Then,

$$\mathcal{R}_\infty(\hat{h} \circ \theta) - \mathcal{R}_\infty^* \leq B_1(n, k) + B_2(\mathcal{F}) + D(n, k)$$

where:

- 1. $B_1(n, k)$ is a bias term due to the choice of k to approximate the law of $(\Theta_\infty, Y_\infty)$.*
- 2. $B_2(\mathcal{F})$ is a second bias term (approximation error) due to the choice of the family $\mathcal{F}$*
- 3. $D(n, k)$ is a variance term, and can be bounded with high probability by a term of order $\frac{\log(1/\delta)}{\sqrt{k}}$ (under some assumptions...).*

Statistical guarantees

Theorem

Consider the estimated quantile function $\hat{h}$ defined previously. Then,

$$\mathcal{R}_\infty(\hat{h} \circ \theta) - \mathcal{R}_\infty^* \leq B_1(n, k) + B_2(\mathcal{F}) + D(n, k)$$

where:

- 1. $B_1(n, k)$ is a bias term due to the choice of k to approximate the law of $(\Theta_\infty, Y_\infty)$.*
- 2. $B_2(\mathcal{F})$ is a second bias term (approximation error) due to the choice of the family $\mathcal{F}$*
- 3. $D(n, k)$ is a variance term, and can be bounded with high probability by a term of order $\frac{\log(1/\delta)}{\sqrt{k}}$ (under some assumptions...).*

We recover a variance term of order $\frac{1}{\sqrt{k}}$, but with an additional bias-variance tradeoff for k .

My contributions

The work presented so far was done in [?].

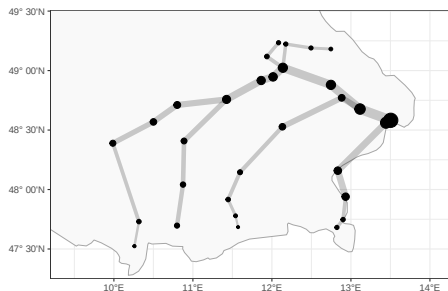
- Extension of the framework to quantile regression with a **penalized** procedure

$$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, h(X_i)) + \underbrace{\lambda \|h\|_{\mathcal{H}}^2}_{\text{penalization}}$$

- Allows to relax some assumptions (boundedness of the target, assumptions on $\mathcal{F}$) and obtain more tractable guarantees.
- Study case on a real multivariate dataset.

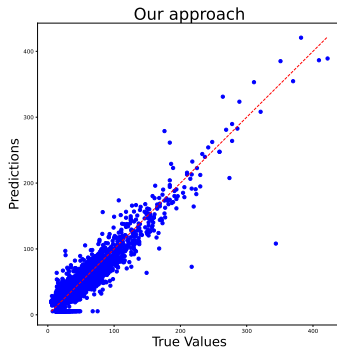
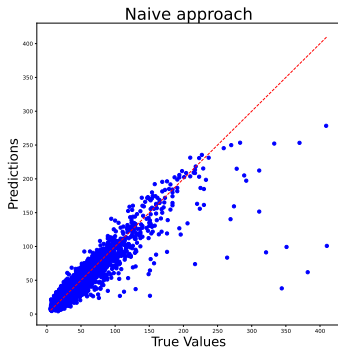
Application to a real dataset

- Data from 31 gauging stations measuring the river flow in the Danube Basin.

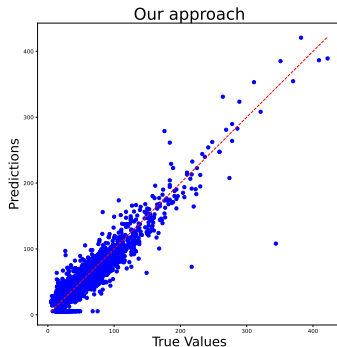
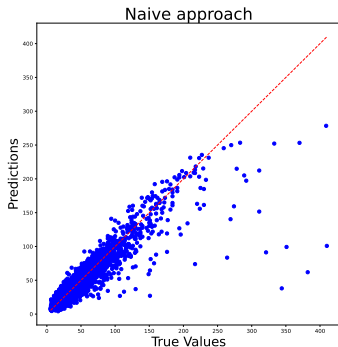


- Prediction of the conditional quantile of the flow measured in one of the station given the flow measured in the other stations.
- Prediction needs to be reliable when a flood occurs in one of the stations (= crisis scenario).

Application to a real dataset



Application to a real dataset



Thank you for listening!