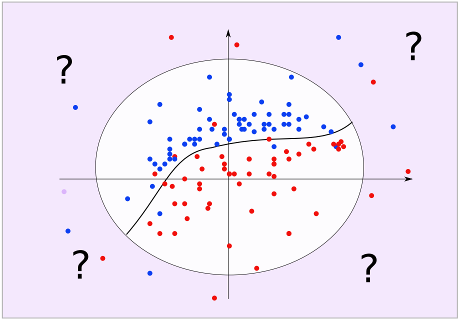
Statistical learning viewpoints on extreme value analysis

Anne Sabourin

Laboratoire MAP5, Université Paris Cité, CNRS, Paris, France

JDS 2025, Marseille



Outline

Multivariate Extreme Values, Heavy-Tails: Why, What, How?

Tail processes, Non-asymptotic deviation bounds

Learning on extreme covariates for out-of-domain generalization
- Classification and Regression on Extremes
- Extension: High dimensional extreme covariates (XLASSO)
- Applications
- Cross-Validation

Why?

- Earth sciences, Finance, Insurance, Telecommunications: unusually large values of (Rain - Temperature - Wind - Sea levels - Streamflow - Traffic - Negative log-returns - Insurance Claims), devastating impacts.



- Such events hard to “predict” (proba. of occurrence hard to estimate) due to
  - Small sample sizes
  - Potentially heavy tails, not satisfying convenient 'Boundedness - subgaussianity - subsomething' assumptions.
- Anomaly detection (all sectors): Anomalies often in the tails. Distinguish 'normal' extreme values from 'abnormal' ones?

Extreme Value Theory: textbook story

Probability Theory: Under minimal assumptions, distributions of maxima/excesses converge to a certain class. Early works Fréchet (1927), Fisher, Tipett (1928), Karamata (1930), Gumbel (1935), Gnedenko (1943), ...

Modelling: Use those limits to model maxima/excesses above large thresholds.

$\mathbf{X}$: random object (variable / vector/ process) $\quad \mathbf{X}_i \overset{i.i.d.}{\sim} \mathbf{X}$.

$$\max_{i=1}^{n} \mathbf{X}_i \quad \overset{d}{\approx} \quad \text{Max-stable} \qquad (n \text{ large})$$

$$\left[\, \mathbf{X} \;\middle|\; \|\mathbf{X}\| \geq r \,\right] \quad \overset{d}{\approx} \quad \text{Generalized Pareto} \qquad (r \text{ large})$$

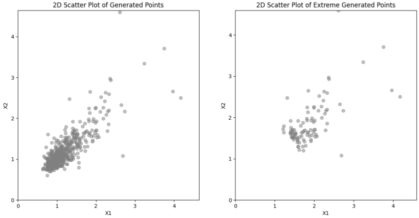
$$\sum_{i=1}^{n} \delta_{(i, X_i)} \quad \overset{d}{\approx} \quad \text{Poisson point process} \qquad (n \text{ large, above large } r)$$

Peaks-Over-Threshold and out-of-domain generalization

- Goal: learn μ/Φ.
- Use $\widehat{P}_k$: empirical distribution of k largest observations $(1 \ll k \ll n)$ (w.r.t. their norm) as a proxi for

$$P_{t_{1-k/n}} = \text{Law}\big(\mathbf{X} \mid \|\mathbf{X}\| > t(1-k/n)\big)$$

where t_{1-p} true $(1-p)$-quantile of the "radial variable" $\|\mathbf{X}\|$



- Hope that $P_{t_{1-k/n}}$ is close to P_∞

Generic strategy for statistical analysis

- Error analysis (in spirit: "k-NN at infinity" / local method)

$$\text{Error}(\widehat{P}_k, \mu) \leq \underbrace{\text{Error}(\widehat{P}_k, P_{t(1-k/n)})}_{\text{Variance(k)}} + \underbrace{\text{Error}(P_{t(1-k/n)}, \mu)}_{Bias(k/n)}$$

- Obvious Bottlenecks:

Bias ($k/n < \infty$) or Variance ($k \ll n$)

Heavy-tails

$X_{(1)}, \ldots, X_{(k)}$ **are not** i.i.d. data

Machine Learning / AI / High dimensions + Extremes since 2015

- (Many environmental) applications with Deep Learning involved for parameter fitting, generative modelling, auto-encoding, Neural Bayes . . . Lafon et al. (2023); Dahal et al. (2024); De Monte et al. (2025); Richards et al. (2024), ...

- Graphical models and causality Velthoen et al. (2023); Gnecco et al. (2024, 2021), some finite sample error bounds (Engelke et al., 2021)

- Sparse support identification Goix et al. (2016, 2017); Meyer and Wintenberger (2021, 2024), feature clustering Chiapino and Sabourin (2016); Chiapino et al. (2019, 2020), Dimension selection Butsch and Fasen-Hartmann (2024, 2025) Supervised dimension reduction: for high dimensional tail index estimation (Chen and Zhou, 2024), identification of tail conditional independence (extreme targets/covariates) (Gardes, 2018; Aghbalou et al., 2024b; Gardes and Podgorny, 2024; Girard and Pakzad, 2024)

Generic research goals and bottlenecks

- Develop **non-asymptotic** guarantees for Extreme Value estimators/learning algorithms, in a **non-parametric** framework, with minimal assumptions, robust to ill-behaved bias

  How to avoid "second order" assumptions that traditionally control bias decrease in CLT's ?
  Until $\approx$ 2015, literature exclusively asymptotic.

- Bridge the gap (Extremes| |**High dimensional** statistics)

  Back in 2015: multivariate modeling envisioned for $d \leq 5$ or 10, except for spatial extremes with parametric spatial structure or parametric models wih fixed, low number of parameters

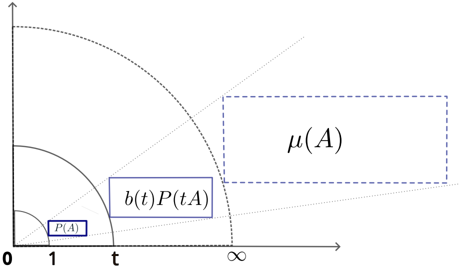
Ingredients for this talk

- Survey paper (preprint) Clémençon and Sabourin (2025)
- Joint works with several colleagues: Patrice Bertail, Chloé Clavel, Eric Gaussier, Philippe Naveau, François Portier, Johan Segers; and students: Nicolas Goix, Hamid Jalalzai, Anass Aghbalou, Nathan Huet (chron. order)+ Pierre Colombo, Stéphane Lhaut

Multivariate Regular Variation in two slides I

$X : \Omega \to \mathbb{R}^d$ is **regularly varying** if $\exists$ scaling $b(t) \to \infty$, and $\exists \mu$ a non-zero **limit measure** on $\mathbb{R}^d \setminus \{0\}$, finite on sets **bounded away from** 0, s.t. as $t \to \infty$, (Resnick, 2008; Hult and Lindskog, 2006)

$$b(t)\mathbb{P}(X \in tA) \to \mu(A), \qquad A \text{ measurable}, 0 \notin \partial A. \tag{1}$$



Then for some $\alpha > 0$, for all $x > 0$,

$$\frac{b(tx)}{b(t)} \to x^{-\alpha} \text{ (regularly varying scaling) and}$$

$$\mu(tA) = t^{-\alpha}\mu(A) \text{ (homogeneous limit measure).}$$

Multivariate regular variation in two slides II

- μ **rules the (probabilistic) behaviour of extremes**: if A is far from the origin, then

$$\mathbb{P}(X \in A) \approx \mu(A) .$$

Namely

$$\mathbb{P}\left(X \in tA\right) = L(t)\mu(tA),$$

with L a slowly varying function.

- **Examples:** Max stable vectors with standardized margins, multivariate Student, ...

- Preliminary **componentwise standardization** is often necessary: then (1) concerns the standard version V of X,

$$V_j := 1/(1 - F_j(X_j)), \quad V = (V_1, \dots, V_d).$$

In practice: empirical $\hat{F}_j$. In spirit $\approx$ empirical copula, but non-linear (unstable propagation of $|\hat{F}_j - F_j|$)

Angular Measure (a third slide was needed)

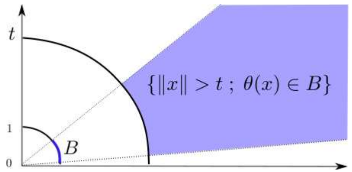
- Homogeneity of $\mu \Rightarrow$ polar coordinates are convenient

$$r(x) = \|x\| \quad ; \quad \theta(x) = r(x)^{-1}x.$$

- **Angular measure** Φ on the $\|\cdot\|$-sphere: $\Phi(B) = \mu\{r > 1, \theta \in B\}$.

- Then μ decomposes as a **product measure**

$$\mu \circ \textit{Polar-transform}^{-1}\{r > t, \theta \in B\} = t^{-\alpha}\Phi(B)$$



$$\text{Multiv. reg. var.} \iff \text{Law}(\theta(X) \mid r(X) > t) \xrightarrow{w} \Phi(\,\cdot\,)$$

$$\big(+ \mathbb{P}\,(r(X) > t) = t^{-\alpha}L(t)\big)$$

Outline

Supremum deviations on low probability classes

$\mathbb{A}$: a VC-class of sets with VC-dimension $\mathcal{V}_{\mathbb{A}}$, with $\mathbb{P}\left(\bigcup_{A\in\mathbb{A}} A\right) \leq p$.

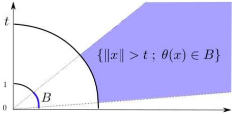
- In Goix et al. (2015) (with universal constant) and Lhaut et al. (2022) (variants, explicit constants), we show

$$\sup_{A\in\mathbb{A}} |P_n(A) - P(A)| \leq \sqrt{\frac{2p}{n}}\Big(\sqrt{2\log(1/\delta)}+ \ldots \sqrt{\log 2 + \mathcal{V}_{\mathbb{A}} \log(2np+1)} + \sqrt{2}/2\Big) \cdots + \frac{2}{3n}\log(1/\delta)$$

- Existing normalized VC inequalities had an extra $\sqrt{\log n}$ factor, see Vapnik and Chervonenkis (2015); Anthony and Shawe-Taylor (1993).
- Tools: McDiarmid (1998)'s Bernstein type concentration inequality + conditioning trick to control Rademacher average
- Possible improvement (factor $\sqrt{2}$) using Bousquet-Talagrand inequality (in preparation with B. Leroux, A. Marchina)

Empirical Angular Measure of extremes

$X_i \overset{i.i.d.}{\sim} F$ in $\mathbb{R}^d$, $1 \ll k \ll n$ to be 'chosen by the user' (choice of k ...)



Rank-transformed variables:

$$\widehat{V}_{i,j} = \frac{1}{1 - \frac{n}{n+1}\widehat{F}_j(X_{i,j})} \qquad (j \leq d, i \leq n)$$

"Radial" order statistics:

$$\widehat{V}_{(1)}, \ldots, \widehat{V}_{(n)} \text{ such that } \|\widehat{V}_{(1)}\| \geq \|\widehat{V}_{(2)} \geq \cdots \geq \|\widehat{V}_{(n)}\|$$

Empirical Angular measure:

$$\widehat{\Phi}(A) = \frac{1}{k}\sum_{i \leq k} \mathbb{1}_A(\|\widehat{V}_{(i)}\|^{-1}\widehat{V}_{(i)})$$

Existing guarantees < 2023**:** Asymptotic, 2nd order assumptions, $d = 2$ **only**. (Einmahl et al., 2001; Einmahl and Segers, 2009)

Concentration of the empirical angular measure

In Clémençon et al. (2023) we assume:

- $\mathcal{A}$ a class of sets on $\mathbb{S}_+$ (positive orthant of the sphere) with some regularity assumptions
- $\mathcal{A}$ is uniformly bounded away from the boundary of the positive orthant
- One can construct "framing " classes of sets accounting for the propagation of uncertainty due to marginal standardization,

and we show:

$$\sup_{A \in \mathcal{A}} |\widehat{\Phi}(A) - \Phi(A)| \leq \frac{C_1(\delta, d, \mathcal{V}_\Gamma, k)}{\sqrt{k}} + \frac{C_2(\delta, d, \mathcal{V}_\Gamma, k)}{k} + \mathrm{Bias}(k, n),$$

where $\mathrm{Bias}(k, n) \to 0$ as $k/n \to 0$ under RV assumptions; $\mathcal{V}_\Gamma$ is the VC dimension of the framing sets; $C_1(\delta, d, \mathcal{V}_\Gamma, k)$, $C_2(\delta, d, \mathcal{V}_\Gamma, k)$ are explicit and have logarithmic dependence on $(k, 1/\delta)$, and polynomial dependence on $d, \mathcal{V}_\Gamma$.

Outline

Outline

Multivariate Extreme Values, Heavy-Tails: Why, What, How?

Tail processes, Non-asymptotic deviation bounds

Learning on extreme covariates for out-of-domain generalization
- Classification and Regression on Extremes
- Extension: High dimensional extreme covariates (XLASSO)
- Applications
- Cross-Validation

Learning on extreme covariates

- X: **Heavy tailed random covariates,** Y: **bounded target** to be predicted, $Y \in I = \{-1, 1\}$ (Jalalzai et al., 2018, Binary classification) or $I = [-M, M]$ (Huet et al., 2023, Regression)

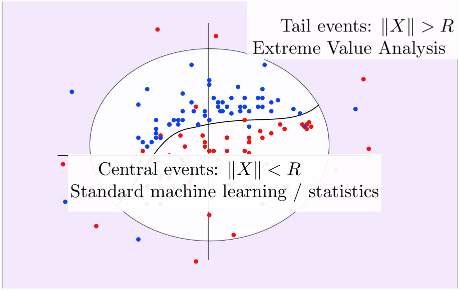
- **Goal:** make acurate prediction in **'crisis scenarios'** where new observed covariable are (unusally) large

- Example 1: $X =$ (temperature, air quality), $Y =$ daily proportion of admissions to the pneumology department in a hospital.
  Covariate shifts with climate change

- Example 2: Prediction of an unobserved component in $(\tilde{X}_1, \ldots, \tilde{X}_{d+1})$ a heavy-tailed r.v.

$$X = (\tilde{X}_1, \ldots, \tilde{X}_d)\ ;\ Y = \tilde{X}_{d+1}/\|\tilde{X}\| \text{ or } Y = \mathbb{1}\{\tilde{X}_{d+1} \geq c\|X\|\}$$

Learning on extremes: Meta-algorithm



1. Pick your favorite predictor (random forest, SVM, logistic regression, deep neural network, . . .)

2. Train it on a fraction of your data (those with the largest norm)

3. For a new (unlabelled) point x_{new}:
    - If $\|x_{new}\|$ is small, use an of-the-shelf ML predictor
    - If $\|x_{new}\|$ is large, use the predictor dedicated to extremes.

Conditional risk minimization, obvious issues

- Learning task (first naive attempt): minimize over $f \in \mathcal{F}$ for "large t"

$$R_t(f) = \mathbb{E}\left(c(f(X), Y) \mid \|X\| > t\right).$$

- Since $\mathbb{P}\left(\|X\| > t\right)$ is small, even though $R(\hat{f})$ is $\approx$ optimal, $R_t(\hat{f})$ may not be so (negligible weight for R_t in the law of total expectations).

- Even though $R_t(\hat{f}_t)$ is $\approx$ optimal for some t, no guarantee for $t' \gg t$.

- For fixed, arbitrary predictor f, the conditional risk $R_t(f)$ may not converge as $t \to \infty$

Asymptotic risk and learning problem

- Issues in previous slide $\rightarrow$ change of focus

$$R_\infty(f) = \limsup_{t\to\infty} R_t(f).$$

- Learning problem:

Minimize $R_\infty(f)$ over $f \in \mathcal{F}$ a class of prediction functions, based on i.i.d. data $(X_i, Y_i)_{i\leq n} \sim (X, Y)$

- Done (and shown today): Stylized settings. $\mathcal{F}$ a VC class, 0-1 loss and squared error loss, no penalization term (except for XLASSO), no convexification. ...

- TODO: quantile regression, unbounded targets, more realistic algorithm: work in progress, (With C. Dombry, B. Leroux's intenship).

Conditional/One component Regular Variation

- Some stability assumptions regarding dependence $Y \sim X$ necessary for extrapolation

- Classification: in Jalalzai et al. (2018) and Clémençon et al. (2023) with standardization step, we assume:

$$b(t)\mathbb{P}(t^{-1}X \in (\,\cdot\,) \mid Y = \pm 1) \xrightarrow[t\to\infty]{} \mu(\,\cdot\,)$$

(same tail index: no class becomes a minority as $\|X\| \to \infty$)

- Regression (Huet et al., 2023): simplification with "one-component regular variation":

$$b(t)\mathbb{P}((t^{-1}X, Y) \in (\,\cdot\,)\,) \xrightarrow[t\to\infty]{} \mu(\,\cdot\,)$$

Consequences: extreme pair (X_∞, Y_∞)

- Scaling function b may be chosen as a quantile function of $\|X\|$, so that $\mu\{(x,y) : \|x\| \geq 1\} = 1$ (probability measure).

- Define
$(X_\infty, Y_\infty) \sim \mu|_{\{\|x\|\geq 1, y \in I\}} = \lim \mathbb{P}\left((X/t, Y) \in (\,\cdot\,) \mid \|X\| \geq t\right).$

- Let $\Theta_\infty = \theta(X_\infty)$. Then (by homogeneity again)

$$(Y_\infty, \Theta_\infty) \perp\!\!\!\perp \|X_\infty\|.$$

- Consequence on the extreme Bayes regression function

$$\begin{aligned} f^*_\infty(x) &:= \mathbb{E}\left(Y_\infty \mid X_\infty = x\right) \quad a.s. \\ &= \mathbb{E}\left(Y_\infty \mid \Theta_\infty = \theta(x), \|X_\infty\| = r(x)\right) \\ &= f^*_\infty(\theta(x)). \end{aligned}$$

The Bayes regression function for the extreme pair is 'angular', *i.e.* it depends only on $\theta(x)$.

Meta-Algorithm

Prediction based on **angles** of observations with **largest radii**

Input: Training dataset $\mathcal{D}_n = \{(X_1, Y_1),\ \ldots,\ (X_n, Y_n)\}$ in $\mathbb{R}^d \times \mathbb{R}$; Class $\mathcal{H}$ of predictive functions $\mathbb{S} \to \mathbb{R}$; number $k \leq n$ of 'extremes'; Norm $\| \cdot \|$ on $\mathbb{R}^d$.

Selection of extremes: Sort the training data by decreasing radial order, $||X_{(1)}|| \geq \ldots \geq ||X_{(n)}||$ and form a set of k *extreme training observations*

$$\left\{ \left(X_{(1)}, Y_{(1)}\right),\ \ldots,\ \left(X_{(k)}, Y_{(k)}\right) \right\}.$$

Empirical risk minimization: Solve

$$\min_{h \in \mathcal{H}} \frac{1}{k} \sum_{i=1}^{k} \left(Y_{(i)} - h\left(\theta\left(X_{(i)}\right)\right)\right)^2, \tag{2}$$

where $\theta(x) = \|x\|^{-1}x$. producing the solution $\hat{h}$.

Output: Predictive function $(\hat{h} \circ \theta)(x)$, to be used for predicting Y based on new examples X such that $\|X\| \geq \|X_{(k)}\|$.

Stability of solutions: additional assumptions

Additional working assumptions

$$\text{classification} \quad \sup_{\{x \in \mathbb{R}_+^d : \|x\| \geq t\}} |f^*(x) - f_\infty^*(x)| \xrightarrow[t \to \infty]{} 0.$$

$$\text{regression} \quad \mathbb{E}\left(|f^*(X) - f_\infty^*(X)| \;\middle|\; \|X\| > t \right) \to 0.$$

Satisfied under (classical) assumptions of regular variation of densities, similar to De Haan and Resnick (1987); Cai et al. (2011)

Main structural results (classification/regression)

(i) Under one-component RV assumption, for any angular function $f(x) = h \circ \theta(x)$, where h is continuous on $\mathbb{S}$, the conditional risk converges

$$R_t(f) \xrightarrow[t \to \infty]{} R_{P_\infty}(f),$$

so that $R_\infty(f) = \lim_{t \to +\infty} R_t(f) = R_{P_\infty}(f)$.

If the above additional assumption (convergence of regression function) holds, then also

(ii) As $t \to +\infty$, the minimum value of R_t converges to that of R_{P_∞}, *i.e.* $R_t^* \xrightarrow[t \to +\infty]{} R_{P_\infty}^*$.

(iii) The minimum values of R_∞ and R_{P_∞} coincide, *i.e.* $R_\infty^* = R_{P_\infty}^*$.

(iv) The regression function $f_{P_\infty}^*$ minimizes the asymptotic conditional risk:

$$R_\infty^* = R_\infty(f_{P_\infty}^*).$$

Statistical guarantees: classification

- Preliminary covariate rank transformation is performed (to Pareto margins)
- Leveraging concentration of empirical angular measure, in Clémençon et al. (2023) we show: with proba. $1-\delta$,

$$\sup_{h\in\mathcal{H}} |\widehat{R}^{>\tau}(h) - R_\infty^\tau(h)| \leq \frac{C_1(\delta/2, d, \mathcal{V}_{\bar{\mathcal{A}}}, k)}{\sqrt{k}} + \frac{C_2(\delta/2, d, \mathcal{V}_{\bar{\mathcal{A}}}, k)}{k} + \text{Bias}(k, n),$$

  $\widehat{R}^{>\tau}$, R_∞^τ restrictions of risks to x's such that $\min\theta(\widehat{v}(x)) > \tau$, resp. $\min\theta(v(x)) > \tau$

- τ is not an artifact from the proof, see simulations in Clémençon et al. (2023)

- Stylized setting in Jalalzai et al. (2018) with marginal distribution known: same rate $1/\sqrt{k}$, τ restriction not required

Statistical guarantees: Regression

Huet et al. (2023); Aghbalou et al. (2024a)

- Same spirit, different proof techniques and bottlenecks (e.g. How to control error due to rank transformation: open question). Standard assumption that $\mathcal{H}$ is "VC subgraph" $\rightarrow$ Localization arguments (conditioning) leveraging Giné and Guillou (2001)'s control of expected sup deviations
- Under standard pointwise measurability assumptions, with proba $1-\delta$,

$$\sup_{h\in\mathcal{H}} \left| \widehat{R}_k(h\circ\theta) - R_{t(n,k)}(h\circ\theta) \right| \leq \frac{8M^2\sqrt{2\log(3/\delta)} + C\sqrt{V_{\mathcal{H}}}}{\sqrt{k}} + \frac{16M^2\log(3/\delta)/3 + 4M^2 V_{\mathcal{H}}}{k},$$

Outline

XLASSO: LASSO on extreme covariates

Section 5 in Clémençon and Sabourin (2025):

$$\min_{\beta \in \mathbb{R}^d,} \frac{1}{2k} \sum_{i=1}^{k} (Y_{(i)} - h_\beta \circ \theta(X_{(i)}))^2 + \lambda \|\beta\|_1.$$

- Design matrix of extreme angles

$$\mathbf{Z} = (\theta(X_{(1)})^\top, \ldots, \theta(X_{(k)})^\top)^\top \in \mathbb{R}^{k \times p};$$

Target $\mathbf{y} = (Y_{(1)}, \ldots, Y_{(k)}) \in \mathbb{R}^k$, residual vector $\mathbf{w} = \mathbf{y} - \mathbf{Z}\beta^*$.

Asymptotic linear Model

- Assumption: For some $\beta^* \in \mathbb{R}^d$,

$$Y = \theta(X)^\top \beta^* + b(X) + \varepsilon,$$

where ε is a bounded noise, $|\varepsilon| \leq M_\epsilon$ almost surely, and $b : \mathbb{R}^d \to \mathbb{R}$ is a bounded function that vanishes at infinity,

$$\bar{b}(t) := \sup_{x: r(x) > t} |b(x)| \xrightarrow[t \to \infty]{} 0.$$

- Ensures required assumptions for regression on extremes Huet et al. (2023) are met.
- Example: regularly varying pair (X, Z) such that X is regul. varying. and

$$Z = X^\top \beta^* + B(X) + \underbrace{\epsilon \|X\|}_{\text{todo: simplify}}, \quad Y = Z/\|X\|,$$

where perturbation function B s.t. $\sup_{x \in \mathbb{R}^d} |B(x)|/\|x\| = M_B < \infty$ and $\sup_{\|x\| > t} |B(x)|/\|x\| \to 0$; ϵ: centered noise s.t. $|\epsilon| \leq M_\epsilon$

XLASSO: Main result: minimal prediction guarantees

Theorem (XLASSO: prediction error guarantees)

Let

$$\lambda \geq M_\varepsilon \sqrt{\frac{\log(4d/\delta)}{2k}} + \bar{b}(t_{1-\tilde{k}(\delta/2)/n}),$$

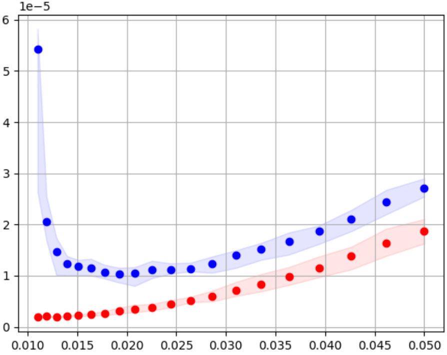
where $\tilde{k}(\delta) \approx k$, $\tilde{k}(\delta) = k\left(1 + \sqrt{\frac{3\log(1/\delta)}{k}} + \frac{3\log(1/\delta)}{k}\right)$, *and* $\bar{b}(t) = \sup_{\|x\|>t} b(x)$.

Then w.p. at least $1-\delta$,

$$k^{-1}\|\mathbf{Z}^\top(\widehat{\beta} - \beta^*)\|_2^2 \leq 12\|\beta^*\|_1\lambda.$$

Experiments: Simulated data

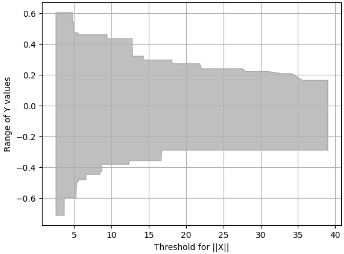
- $Y = \langle \theta(X), \beta_0 \rangle + \frac{1}{\log(1+\|X\|)} \langle \theta(X), \beta_1 \rangle + \epsilon$, with $\beta_1 \equiv 1$ and β_0 5-sparse, $d = 100$, $n = 10^4$, $d = 100$. $k \in [0.011n, 0.05n]$. Test set radial quantile: $1 - 0.01$. 20 replications

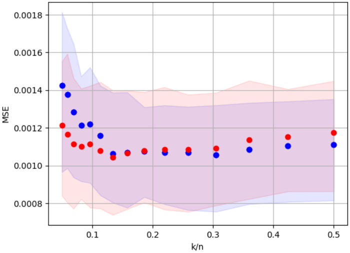


Red dots: XLASSO; Blue dots: linear regression

Experiments: Real data

- **Industry Portfolio Dataset** (Meyer and Wintenberger, 2024; Huet et al., 2023). Target: $Z =$ "Transportation sector", $d = 49$, $n = 13577$.
- Target rescaling: $Y = Z/\|X\|$, X: other variables.
- Threshold for $\|X\|$: $1 - 0.005$ quantile for test, $1 - [0.05, 0.5]$ for train.
- left panel: boundedness of Y?



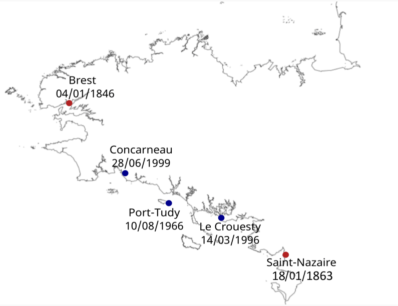


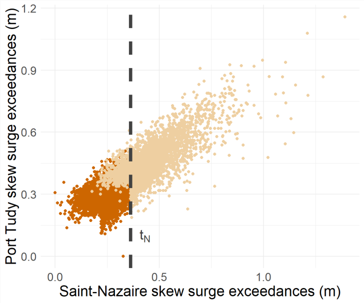
Outline

Extreme sea levels reconsruction

Huet et al. (2025)

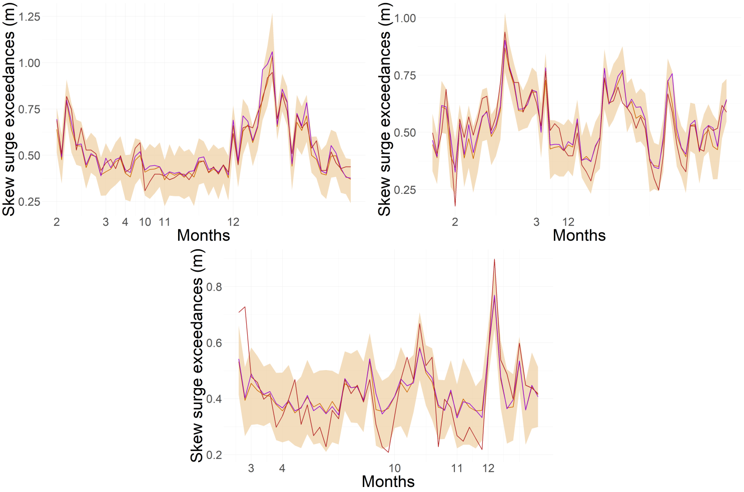
- Extreme surges (tidal component removed)
- Goal: reconstruct missing coastal gauges records from nearby stations with longer historical records
- Input stations: Brest, St Nazaire; output: Port Tudy, Concarneau, and Le Crouesty.





Methodology

- Implementation of the 'learning on extreme covariates' meta-algorithm (instances: Random Forest, OLS)
- Sanity check: Comparison with a parametric plug-in method (Multivariate Generalized Pareto families, similar working assumptions, different marginal standardization and methodology), “distributional regression” of the conditional distribution at one gauge given an extreme value at another gauge.
- Comparable performance in terms of mean square errors and qualitative behavior from visual inspection



Predicted skew surge exceedances at Port Tudy station for the years 1989 (left), 1978 (middle), 1977 (right). Red curves represent the true values; purple curves represent the predicted values by the ROXANE procedure with OLS algorithm; orange curves represent the predicted values by MGPRED with bootstrap 0.95 confidence bands (lightorange).

Application to Natural Language Processing

Jalalzai et al. (2020)

- Extension of the previous framework to datasets who are **NOT** regularly varying.
- Dataset: text embeddings (BERT). $X =$ vector in $\mathbb{R}^d$, d large (768).
- label $Y =$ positive/negative sentiment.
- Two goals:

  (i) improved classification in low probability regions of $\mathcal{X}$

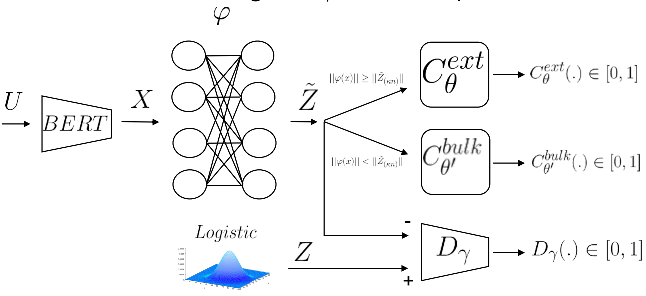
  (ii) label preserving data augmentation

Learning a regularly varying representation for NLP

- Key step: adversarial strategy, (Goodfellow et al. 2014) mixed loss function involving

  - $0-1$ loss in extreme/ non-extreme regions

  - Jensen-Shannon divergence between the learnt representation and a Max-stable multivariate Logistic, $\neq$ common practice Gaussian



- Output: a transformed vector $\tilde{Z} = \varphi(X)$ which is (experimentally) regularly varying (low correlations $\theta(\tilde{Z}) \leftrightarrow \|\tilde{Z}\|$).

Outline

Starting point of Aghbalou et al. (2024a): facts and wishes

- Cross-Validation (CV): widely used (even in extremes) for

  1. **Model or Hyper-parameter selection**
  2. **Estimating the generalisation risk** (= expected error on new examples) of a learning algorithm/estimator.

  Arlot&Celisse, 2010; Wager, 2020; Bates et al., 2023.

  Theoretically analysed in a variety of settings: density estimation Arlot, 2008; Arlot&Lerasle, 2016 or least-squares regression Homrighausen&McDonald, 2013; Xu et al., 2020, . . .

- **Our wish:** Make a first step towards theoretical guarantees for CV **in an EVT framework (existing works:$\varnothing$)**.

- **Main challenge**: statistical properties of CV are hard to analyze in general (dependence between folds / bias).

Motivating examples for using CV in EVA

- **Unsupervised**:
  - Parametric modeling of multivariate tail dependence Einmahl et al. (2012, 2018, 2016); Kiriliouk et al. (2019), . . . : CV for goodness-of-fit assessment on model selection?
  - Dimension reduction in Multivariate Extremes: **Support identification** Goix et al. 2017: Choosing the number of subcones in $\mathbb{R}^d$ supporting the tail measure? **PCA** Cooley & Thibaud, 2019; Jiang et al., 2020; Drees and S.,2021: Estimating the reconstruction error for dimensionality selection? **Clustering** Janssen&Wan, 2020; Jalalzai&Leluc, 2021: Number of clusters?
- **Supervised**:
  - Extreme Quantile Regression Chernozhukov et al. 2017: Choice of Kernel bandwidth? **Trees** Farkas et al., 2021: number of splits? **Gradient boosting** Velthoen et al., 2023, Random Forests Gnecco et al.,2023: number of trees and minimum node size?
  - **Classification/Regression on extreme covariates** Jalalzai et al. 2018, Jalalzai et al. 2020, Clémençon et al. 2022, Huet et al. 2022: Penalty level?

Considered framework (l'art de la répétition)

- Leading example: Classification setup, constrained Logistic-LASSO regression

$$\min_{\beta \in \mathbb{R}^d} \sum_{i \leq k} c(g_\beta(X_{(i)}), Y_{(i)}) \quad \text{subject to} \quad \|\beta\|_1 \leq u,$$

  where $u > 0$ is a hyper-parameter to be selected by CV, $g_\beta(x) = \beta^\top \theta(x)$ following Jalalzai et al. (2018)
- Why not regression? Because it was not ready yet.
- Why not (unconstained) Lasso? Because —//—
- More generally: ERM machine learning algorithms minimizing empirical versions of the risk:

$$R(g, Z) = \mathbb{E}\left(c(g, Z) | \|Z\| > t_p\right),$$

  $\| \cdot \|$ is a semi-norm on $\mathcal{Z}$, and t_p is the $1 - p$ quantile of $\|Z\|$.

CV for ERM generalization risk on covariate tails

- Focus: learning rules Ψ that take a sample $\mathcal{S}$ as input and return the ERM solution $\Psi(\mathcal{S}) = \hat{g}(\mathcal{S}) = \arg\min_{g \in \mathcal{G}} \widehat{R}(g, \mathcal{S})$.
- Goal: estimate generalization risk $R(\hat{g}_n)$ of the ERM predictor $\hat{g}_n = \Psi(\{1, \dots, n\})$ trained on the full dataset.
- CV estimator

$$\widehat{R}_{\mathrm{CV},p}(\Psi, V_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} \widehat{R}(\Psi(T_j), V_j),$$

where $(V_j, j \leq K)$ are validation sets and $T_j = \{1, \dots, n\} \setminus V_j$ are training sets.

Main results

Working assumptions

- Loss class $\{z \mapsto c(g,z), g \in \mathcal{G}\}$ associated with the predictor class $\mathcal{G}$ is VC subgraph
- Bounded cost function
- Balance condition on the CV scheme (met by K fold, lpo, loo)

Exponential error bound, w.p. $1 - 15\delta$,

$$\left|\widehat{R}_{\mathrm{CV},p}(\Psi, V_{1:K}) - R\left(\hat{g}_n\right)\right| \leq E_{CV}(n_T, n_V, p) + \frac{20}{3np}\log(1/\delta) + 20\sqrt{\frac{2}{np}\log(1/\delta)},$$

where $E_{CV}(n_T, n_V, p) = C\sqrt{\mathcal{V}_{\mathcal{G}}}(1/\sqrt{n_V p} + 4/\sqrt{n_T p}) + 5/(n_T p)$.

Applicable to K-fold, not l.o.o because of $1/\sqrt{n_T}$
NB: also a polynomial bound, applicable to loo but not suitable for parameter selection guarantees via union bounds

Application to constrained LASSO problem

- grid search over a range U of plausible values for u, union bound: With proba $1 - 15\delta$

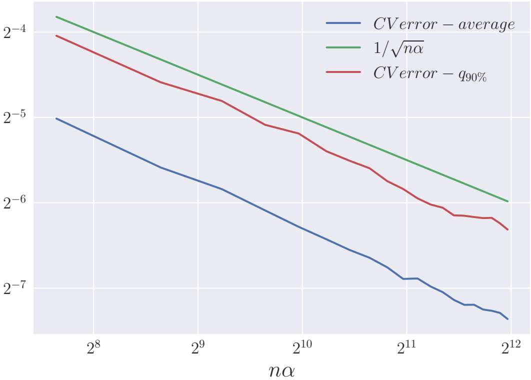
$$\left|\widehat{R}_{\mathrm{CV},p}(\Psi_{\widehat{u}}, V_{1:K}) - R\big(\hat{g}_n\big)\right| \leq \max(U)\Bigg[2E(n,K,p) + \frac{40}{3np}\log\left(|U|/\delta\right)\big) \cdots$$

$$\cdots + 40\sqrt{\frac{2}{np}\log\left(|U|/\delta\right)}\Bigg],$$

where $\widehat{u}$ is the minimizer of the CV risks $\widehat{R}_{\mathrm{CV},p}(\Psi_u V_{1:K}), u \in U$, and $E(n,K,p) = 5C\sqrt{(d+1)K/(np)} + 5K/((K-1)np)$.

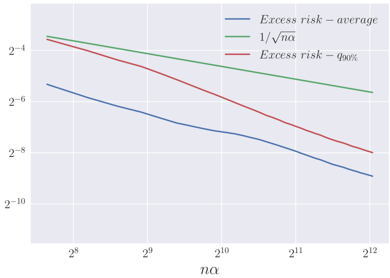
Risk estimation error $|\widehat{R}_{\mathrm{CV},p}(\Psi_\alpha, V_{1:K}) - R_\alpha(\{1,\ldots,n\})|$: rate $1/\sqrt{n\alpha}$?

- Toy example: simulated data, dimension 1,
  Class distributions: Student, threshold classifier, Hamming loss
- $n = 2.10^4$, $\alpha \in [1\%, 20\%]$
- Average absolute error of the K-fold ($K = 10$) and upper quantile at level 0.90, logarithmic scale, over 10^4 experiments.



Logistic-LASSO: excess risk $R_\alpha(\hat{g}_{\hat{\lambda}}) - R_\alpha(\hat{g}_{\lambda^*})$

- Penalized version of the LASSO: $R + \lambda\|\beta\|_1$: computationally (much) easier and strong connections with constrained version.
- data: $X \in \mathbb{R}^{50}$, $Y \sim$ Bernoulli(0.5), class distribution: multivariate student, same tail index + scale but different centers.
- $\alpha \in [0.01, 0.1]$, $n = 10^4$, with 2000 repetitions
- grid $\lambda_i = 10^{i/30} - 1$, $i \leq 30$.



CV on extremes: Discussion, perspectives

- Replacing ERM assumption with algorithmic stability $\rightarrow$ wider class of algorithms and improved bounds for the l-p-o.
- Extension to other rare events (imbalanced classification)?
- Beyond sanity check bounds? (even for $\alpha = 1$?)
- Extension to other EVA settings by relaxing the bounded loss assumption?

- This talks: contributions to setting up a theoretical grounding for ML approaches in EVA

- Field still emerging, more open questions than answers

- Not shown: Choice of k, without second order assumptions $\rightarrow$ Lederer et al. (2025) for tail index estimation. Applications of rare classes arguments to imbalanced classification Aghbalou et al. (2024c).
- Just released

SOFTWARE

MLExtreme Python Package

`https://github.com/hi-paris/MLExtreme/`

  - Unsupervised: anomaly scoring with MV sets, support identification (feature clustering), PCA
  - Supervised: Classification, Regression (compatible with any learner with a fit and predict method, à la scikit-learn)
  - Tutorial notebooks

References I

Aghbalou, A., Bertail, P., Portier, F., and Sabourin, A. (2024a). Cross-validation on extreme regions. *Extremes*, 27(4):505–555.

Aghbalou, A., Portier, F., Sabourin, A., and Zhou, C. (2024b). Tail inverse regression: dimension reduction for prediction of extremes. *Bernoulli*, 30(1):503–533.

Aghbalou, A., Sabourin, A., and Portier, F. (2024c). Sharp error bounds for imbalanced classification: how many examples in the minority class? In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR.

Anthony, M. and Shawe-Taylor, J. (1993). A result of Vapnik with applications. *Discrete Applied Mathematics*, 47(3):207 – 217.

Butsch, L. and Fasen-Hartmann, V. (2024). Information criteria for the number of directions of extremes in high-dimensional data. *arXiv preprint arXiv:2409.10174*.

Butsch, L. and Fasen-Hartmann, V. (2025). Estimation of the number of principal components in high-dimensional multivariate extremes. *arXiv preprint arXiv:2505.22437*.

Cai, J., Einmahl, J., and De Haan, L. (2011). Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*, pages 1803–1826.

Chen, L. and Zhou, C. (2024). High dimensional inference for extreme value indices. *arXiv preprint arXiv:2407.20491*.

References II

Chiapino, M., Clémençon, S., Feuillard, V., and Sabourin, A. (2020). A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, 35(2):607–628.

Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer.

Chiapino, M., Sabourin, A., and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222.

Clémençon, S., Jalalzai, H., Lhaut, S., Sabourin, A., and Segers, J. (2023). Concentration bounds for the empirical angular measure with statistical learning applications. *Bernoulli*, 29(4):2797–2827.

Clémençon, S. and Sabourin, A. (2025). Weak signals and heavy tails: Machine-learning meets extreme value theory. *arXiv preprint arXiv:2504.06984*.

Dahal, A., Huser, R., and Lombardo, L. (2024). At the junction between deep learning and statistics of extremes: formalizing the landslide hazard definition. *Journal of Geophysical Research: Machine Learning and Computation*, 1(3):e2024JH000164.

De Haan, L. and Resnick, S. (1987). On regular variation of probability densities. *Stochastic processes and their applications*, 25:83–93.

References III

De Monte, L., Huser, R., Papastathopoulos, I., and Richards, J. (2025). Generative modelling of multivariate geometric extremes using normalising flows. *arXiv preprint arXiv:2505.02957*.

Einmahl, J. H., de Haan, L., and Piterbarg, V. I. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Stat.*, 29:1401–1423.

Einmahl, J. H. J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Stat.*, 37:2953–2989.

Engelke, S., Lalancette, M., and Volgushev, S. (2021). Learning extremal graphical structures in high dimensions. *arXiv preprint arXiv:2111.00840*.

Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1):57–95.

Gardes, L. and Podgorny, A. (2024). Dimension reduction for the estimation of the conditional tail-index.

Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. de l'IHP Probab. et Stat.*, 37(4):503–522.

Girard, S. and Pakzad, C. (2024). Functional extreme-pls. *arXiv preprint arXiv:2410.05517*.

References IV

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. (2021). Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778.

Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal random forests. *Journal of the American Statistical Association*, 119(548):3059–3072.

Goix, N., Sabourin, A., and Clémençon, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory*, pages 843–860.

Goix, N., Sabourin, A., and Clémençon, S. (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pages 75–83.

Goix, N., Sabourin, A., and Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31.

Huet, N., Clémençon, S., and Sabourin, A. (2023). On regression in extreme regions. *arXiv preprint arXiv:2303.03084*.

Huet, N., Naveau, P., and Sabourin, A. (2025). Multi-site modelling and reconstruction of past extreme skew surges along the french atlantic coast. *arXiv preprint arXiv:2505.00835*.

References V

Hult, H. and Lindskog, F. (2006). Regular variation for measures on metric spaces. *Publications de l'Institut Mathematique*, 80(94):121–140.

Jalalzai, H., Clémençon, S., and Sabourin, A. (2018). On binary classification in extreme regions. In *Advances in Neural Information Processing Systems*, pages 3092–3100.

Jalalzai, H., Colombo, P., Clavel, C., Gaussier, E., Varni, G., Vignon, E., and Sabourin, A. (2020). Heavy-tailed representations, text polarity classification & data augmentation. *Advances in Neural Information Processing Systems*, 33:4295–4307.

Lafon, N., Naveau, P., and Fablet, R. (2023). A vae approach to sample multivariate extremes. *arXiv preprint arXiv:2306.10987*.

Lederer, J., Sabourin, A., and Taheri, M. (2025). Adaptive tail index estimation: minimal assumptions and non-asymptotic guarantees.

Lhaut, S., Sabourin, A., and Segers, J. (2022). Uniform concentration bounds for frequencies of rare events. *Statistics & Probability Letters*, 189:109610.

McDiarmid, C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, Berlin.

Meyer, N. and Wintenberger, O. (2021). Sparse regular variation. *Advances in Applied Probability*, 53(4):1115–1148.

References VI

Meyer, N. and Wintenberger, O. (2024). Multivariate sparse clustering for extremes. *Journal of the American Statistical Association*, 119(547):1911–1922.

Resnick, S. I. (2008). *Extreme values, regular variation, and point processes*, volume 4. Springer Science & Business Media, Berlin.

Richards, J., Sainsbury-Dale, M., Zammit-Mangion, A., and Huser, R. (2024). Neural bayes estimators for censored inference with peaks-over-threshold models. *Journal of Machine Learning Research*, 25(390):1–49.

Vapnik, V. N. and Chervonenkis, A. Y. (2015). *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. Springer International Publishing, Cham.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*, 26(4):639–667.