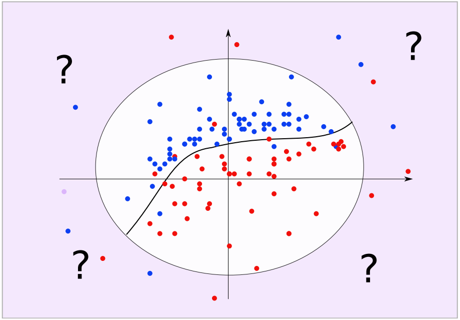
Statistical learning viewpoints on extreme value analysis

Anne Sabourin

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

EVA 2025, Chapel Hill



Outline

Outline

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) $\quad \mathbf{X}_i \overset{i.i.d.}{\sim} \mathbf{X}$.

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

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

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

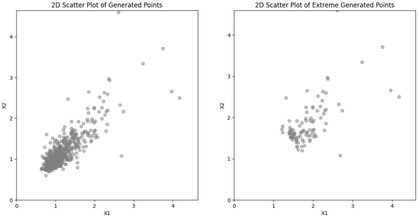
Peaks-Over-Threshold and out-of-domain generalization

- 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

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

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_∞

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), recent special issue in 'Extremes', ...

- 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 (This talk)

- 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

Generic strategy for statistical analysis (This talk)

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

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

- Obvious Bottlenecks:

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

Heavy-tails

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

Ingredients

- Survey paper (preprint) Clémençon and Sabourin (2025)

- Joint works with many colleagues (chronological order): Stephan Clémençon, Alexandre Gramfort, Chloé Clavel, Eric Gaussier, Johan Segers, François Portier, Patrice Bertail, Philippe Naveau;
  and students: Nicolas Goix, Maël Chiapino, Hamid Jalalzai, Anass Aghbalou, Nathan Huet + Pierre Colombo, Stéphane Lhaut
- 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)
  - Data generation + basic EVT tools
  - **Tutorial notebooks**

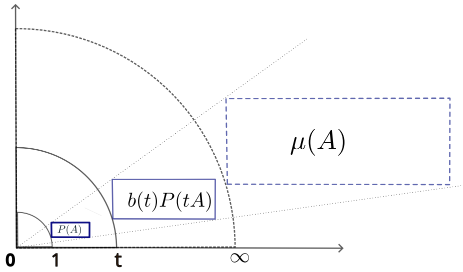
Outline

Multivariate Regular Variation I

$X : \Omega \to \mathbb{R}^d$ is **regularly varying** (Resnick (2008); Hult and Lindskog (2006), ...) if

- $\exists$ scaling $b(t) \to \infty$,
- $\exists \mu$ a non-zero **limit measure** on $\mathbb{R}^d \setminus \{0\}$, s.t. as $t \to \infty$, for any A **bounded away from** 0 with $\mu(\partial A) = 0$,

$$b(t)\mathbb{P}(X \in tA) \to \mu(A) < \infty \quad (1)$$

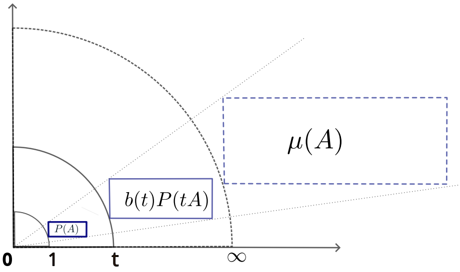


Multivariate Regular Variation II

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 III

- 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, ...

Angular Measure

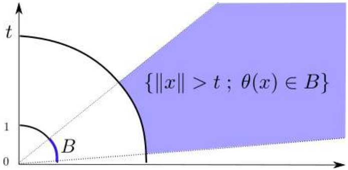
- Homogeneity of $\mu \Rightarrow$ polar coordinates are convenient

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

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

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

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



$$\text{MRV} \iff \left[\theta(X) \mid r(X) > t\right] \xrightarrow{w} \Phi(\,\cdot\,)$$

and $\mathbb{P}\left(r(X) > t\right) = t^{-\alpha}L(t)$

General domain of attraction, marginal standardization

- Different X_j's may have different 'tail indices' or even 'domains of attraction' (Weibull/Gumbel/Fréchet), while still, for some vectors (a_n, b_n), $a_n \succ 0$ there is convergence in distribution of

$$\frac{M_n - b_n}{a_n}, \quad \text{or equivalently} \quad \left[\frac{X - b_n}{a_n} \;\Big|\; X \not\preceq b_n\right].$$

- Luckily, the above 2 equivalent conditions are also equivalent to
  1. Marginal convergence of margins $X_j, j \le d$;
  2. Convergence on the standard scale of the conditional distribution

$$\left[t^{-1} V \;\middle|\; \|V\| > t\right]$$

  with $V = v(X)$, and $v_j(x_j) = \frac{1}{1 - F_j(x_j))}, j \le d$.

Equivalent statements, on standard scale

1.
$$\left[t^{-1}V \;\middle|\; \|V\| > t\right] \xrightarrow{w} Z_\infty,$$

where

$$\begin{aligned} Z_\infty &\overset{d}{=} R_\infty \Theta_\infty \\ R_\infty &\sim \text{Pareto}(1) \quad \perp\!\!\!\perp \quad \Theta_\infty \sim \Phi \end{aligned}$$

2.
$$\left[(t^{-1}\|V\|,\ \theta(V)) \;\middle|\; \|V\| > t\right] \xrightarrow{w} (R_\infty, \Theta_\infty)$$

3.
$$\left[t^{-1}V \mid V \in tA\right] \xrightarrow{w} \frac{\mu(\,\cdot\,)}{\mu(A^c)}$$

for all set A s.t. $0 \notin A, \mu(\partial A) = 0$, and

$$\mathrm{d}\mu \;\;\text{"}=\text{"}\;\; \frac{\mathrm{d}r}{r^2}\,\mathrm{d}\Phi \quad \text{in polar coodinates}$$

Related frameworks I

- Multivariate Generalized Pareto Rootzén and Tajvidi (2006); Rootzén et al. (2018a,b); Kiriliouk et al. (2019)
  - **Same working assumptions** (multivariate max-domain of attraction)
  - Different Standardization choice ("Standard" = Exponential)
  - Different affine transformations

$$(X - b)/a \text{ not just } X/a$$

  - Different typical conditioning events

$$\text{``}\exists j \leq d : X_j > b_j\text{''} \text{ not } \text{``}\|V\| > t\text{''}$$

  - Different representation of the limit

$$E + S \quad \text{not} \quad R \times \Theta$$

Related frameworks II

- Asymptotically independent components
  - Concomitant extremes have negligible probability compared with isolated ones.
  - Angular measure / limit measure concentrated on the axes: not informative about subasymptotic dependence
  - Long history of models allowing for asymptotic independence: Ledford and Tawn (1996) ... Heffernan and Resnick (2007)... Wadsworth et al. (2017) ... Huser and Wadsworth (2019)
  - (not today)

Outline

Complexity controls of classes of sets I

- $\mathcal{A}$: class of subsets of $\mathcal{X}(= \mathbb{R}^d$ here).
- $D_n = x_1, \ldots, x_n$ is "fully shattered" if all possible subsets of D_n can be selected by applying a mask from $\mathcal{A}$, *i.e.*

$$\left|\left\{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{A}\right\}\right| = 2^n.$$

- **Shattering coefficient** of the class: the maximum cardinality of the above family of subset, as D_n varies.

$$S_{\mathcal{A}}(n) = \max_{(x_1,\ldots,x_n)\in\mathcal{X}^n} |\{A \cap \{x_1, \ldots, x_n : A \in \mathcal{A}\}| \leq 2^n$$

Complexity controls of classes of sets II

- VC dimension: **complexity control** of $\mathcal{A}$. The maximum size n such that $\exists D_n$ that can be shattered by $\mathcal{A}$

$$\mathcal{V}_{\mathcal{A}} = \sup\{n : S_{\mathcal{A}}(n) \leq 2^n\}$$

- Also used in Asymptotic Statistics, see van der Vaart (1998); van der Vaart and Wellner (1996)

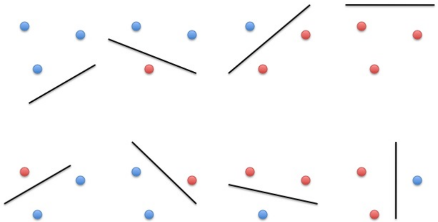
- "Standard assumption" in statistical ML, a good starting point, maybe not the endpoint.

VC-dimension of Hyperplanes I

- What is the VC-dimension of hyperplanes in $\mathbb{R}^2$ (denoted $\mathbb{H}_2$)?

- Obviously $\text{VCdim}(\mathbb{H}_2) \geq 2$

- Let us try with 3 points:



VC-dimension of Hyperplanes II

- Thus $\mathrm{VCdim}(\mathbb{H}_2) \geq 3$
- For any set of 4 points, either 3 of them (at least) are aligned or no triplet of points is aligned.

(a) (b)

It is not possible for $\mathbb{H}_2$ to shatter 4 points.

- Thus $\mathrm{VCdim}(\mathbb{H}_2) = 3$.

VC-dimension of Hyperplanes III

- More generally, one can prove :

$$VCdim(\mathbb{H}_d) = d + 1$$

Vapnik Chervonenkis inequality (71)

- $X, X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$; P_n: empirical measure.

- **Key historical result**: uniform bound on the deviations of the empirical measure P_n from the true law P of X, over a class of sets of controlled complexity

- Sauer's lemma: **polynomial growth of shattering coefficients** for VC classes.

$$S_{\mathcal{A}}(n) \leq (n+1)^{\mathcal{V}_{\mathcal{A}}} \ll_{n \text{ large }, \mathcal{V}_{\mathcal{A}} < \infty} 2^n$$

Vapnik and Chervonenkis (1971)'s inequality

With probability $\geq 1 - \delta$,

$$\sup_{A \in \mathcal{A}} |P_n - P|(A) \leq 2\sqrt{\frac{2}{n}[\log(4/\delta) + \log(S_{\mathcal{A}}(2n))]}$$

Classification, Empirical Risk Minimization (ERM)

- A classifier: a function $g : \mathcal{X} \mapsto \{-1, 1\}$, $\mathcal{X}$: feature space.

- Choose a family of such classifiers $\mathcal{G}$ ($\sim$ a 'model').

- $\mathcal{G}$ is 1-to-1 with $\mathcal{A} = \{A_g = \{(x, y) : g(x) \neq y\}, g \in \mathcal{G}\}$

- Empirical risk $R_n(g) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{g(X_i) \neq Y_i\} = P_n(A_g)$.

- Upper bound the generalization (excess) risk R via

$$R(\widehat{g}) \leq R(g^*) + 2 \underbrace{\sup_{g \in \mathcal{G}} |R(g) - R_n(g)|}_{= \sup_{A \in \mathcal{A}} |P(A) - P_n(A)| \lesssim \sqrt{\frac{\mathcal{V}_{\mathcal{A}}}{n}}}$$

Supremum deviations: Classical proofs (no EVT) I

Lugosi (2002); Bousquet et al. (2003); Boucheron et al. (2005), ...

- (random) supremum absolute deviations:

$$\boldsymbol{Z} = \sup_{A \in \mathcal{A}} \left| P_n(A) - P(A) \right|$$

- "Obvious" decomposition

$$\boldsymbol{Z} \leq \underbrace{\mathbb{E}(\boldsymbol{Z})}_{\text{I: Symmetrization} \to \text{Rademacher process}} + \underbrace{\boldsymbol{Z} - \mathbb{E}(\boldsymbol{Z})}_{\text{II: Concentration / Mc Diarmid}}$$

I: About $\mathbb{E}\boldsymbol{Z}$ (Boucheron et al., 2005; Lugosi, 2002; Bousquet et al., 2003)...

- Symmetrization and Rademacher complexity:
  - ghost sample $X'_{1,\ldots,n}$
  - $f_A(x) = 2\mathbb{1}_A(x) - 1 \in \{\pm 1\}$

$$
\begin{aligned}
\mathbb{E}\boldsymbol{Z} &\leq \frac{1}{2}\mathbb{E}\sup_f \left| P_n f - \mathbb{E}\left(P'_n f \mid X_{1,\ldots,n}\right)\right| \\
&\leq \frac{1}{2n}\mathbb{E}\sup_f \left| \sum_{i\leq n} \sigma_i \big(f(X_i) - f(X'_i)\big)\right| \\
&\leq \mathbb{E}\sup_f \left| \frac{1}{n}\sum_{i\leq n} \sigma_i f(X_i)\right| \qquad \sigma_i \in \{\pm 1\} \text{ white noise} \\
&:= \textbf{ Rademacher complexity, } \mathrm{RAD}(n)
\end{aligned}
$$

(How well can the class fit arbitrary random labels)

I: About $\mathbb{E}Z$ (Boucheron et al., 2005; Lugosi, 2002; Bousquet et al., 2003)

- Rademacher bound (projection on $X_{1:n}$)

$$\mathrm{RAD}(n) \leq \sqrt{\frac{2\log S_{\mathcal{A}}(n)}{n}} \leq_{\text{Sauer's Lemma}} \sqrt{\frac{2\mathcal{V}_{\mathcal{A}}\log(n+1)}{n}}$$

II: Concentration around $\mathbb{E}(Z)$

- Recall $\mathbf{Z} = \sup_A |P_n(A) - P(A)| = \varphi(X_1, \ldots, X_n)$ where φ has the stability property ('bounded differences')

$$\left| \varphi(x_1, \ldots, x_i, \ldots, x_n) - \varphi(x_1, \ldots, x_i', \ldots, x_n) \right| \quad \leq \quad \frac{1}{n}.$$

- McDiarmid's inequality (McDiarmid, 1998):

$$\mathbb{P}\left(\ \mathbf{Z} - \mathbb{E}\left(\mathbf{Z}\right) \ > \ \varepsilon \ \right) \quad \leq \quad e^{-2n\epsilon^2}.$$

Solving $\delta = 2e^{-2n\varepsilon}$: with probability at least $1 - \delta$,

$$\mathbf{Z} - \mathbb{E}\left(\mathbf{Z}\right) \leq \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The case of rare classes (I) relative deviations

- Anthony and Shawe-Taylor (1993): with probability $1 - 2\delta$,

$$\sup_{A \in \mathcal{A}} \frac{P(A) - P_n(A)}{\sqrt{P(A)}} \leq 2\sqrt{\frac{\log S_{\mathcal{A}}(2n) + \log \frac{4}{\delta}}{n}},$$

- Consequence, with $p = \sup_{A \in \mathcal{A}} P(A)$

$$\frac{1}{p} \sup_{A \in \mathcal{A}} P(A) - P_n(A) \leq 2\sqrt{\frac{\log S_{\mathcal{A}}(2n) + \log \frac{4}{\delta}}{np}},$$

How to replace the $S_{\mathcal{A}}(n)$ term with $S_{\mathcal{A}}(np)$ or $V \log(np)$ or C?

Other central topics

- Regression problems: $Y \in [-M, M]$, prediction function $f : \mathcal{X} \to [-M, M]$
$$\min_{f \in \mathcal{F}} \mathbb{E}\left((Y - f(X))^2\right)$$
Under (different but related) complexity controls of the class $\mathcal{F}$
- Regularization:
$$\min_{g} R_n(g) + \lambda \text{ Complexity } (g)$$
- Beyond the ERM paradigm:
    - Local methods (k-nn, trees)
    - Aggregation
    - Stable aglorithms
    - ...
- Neural Networks

Outline

Outline

Tail empirical process: asymptotic literature

- Convenient re-writing: (1) $\iff tP\big(U(t)\,\cdot\,\big) \to c\mu(\,\cdot\,)$ (vaguely) where $U(t) =$ quantile of order $1 - 1/t$ of the norm, $c > 0$ a constant

Estimating $\mu \approx$ Estimating $tP\big(U(t)\,\cdot\,\big)$ (up to a vanishing bias term)

- As $n \to \infty, k \to \infty, n/k \to \infty$: the **'tail empirical process'** converges weakly (1D case, $\alpha = 1$)

$$\sqrt{k}\,\frac{n}{k}(P_n - P)\Big[U(n/k)y, \infty\Big) \xrightarrow[w]{D(0,\infty)} W(y),$$

W: brownian motion, see Resnick 2007, thm. 9.1 + references

In practice: $U(n/k) \leftarrow X_{(k)}$ (the k^{th} largest order statistic)

- Multivariate extensions Einmahl et al. (2006); Einmahl and Segers (2009); Einmahl et al. (2012); Aghbalou et al. (2024b); Lhaut and Segers (2024)...

Low probability classes in EVT I

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

- $p \approx k/n$, where k is the (componentwise) number of extreme samples used for inference

- Motivating example: Stable tail dependence function in $\mathbb{R}^d$ (cdf-type characterization of μ), $\ell(x) = \lim_t t\mathbb{P}\left(\exists j : V_j \geq t/x_j\right), \; x \succeq 0, x \neq 0$.
  Empirical version: involves in particular

$$P_{\mathcal{U},n}\Big(\underbrace{\{y \in \mathbb{R}^d \;\; | \;\; \exists j \leq d : \; y_j < (k/n)x_j\}}_{A(x)}\Big), \qquad 0 \leq x_j \leq T$$

where $P_{\mathcal{U},n}$: empirical measure associated with $U_i = (F_j(X_{i,j}))_{j=1}^n$, with

$$P_{\mathcal{U}}\Big(\bigcup_{\|x\|_\infty \leq T} A(x)\Big) \leq dkT/n$$

Low probability classes in EVT II

- Follow-up applications:
  - limit support identification (Goix et al., 2016, 2017; Chiapino and Sabourin, 2016; Simpson et al., 2020)
  - Anomaly detection in mutlivariate tails via mass-volume sets estimation (Thomas et al., 2017)
  - Empirical angular measure (Clémençon et al., 2023), out-of-domain classification Jalalzai et al. (2018), . . .

Supremum deviations on low probability classes

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

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

- Recall: existing normalized VC inequalities had an extra $\sqrt{\log n}$ factor (Vapnik and Chervonenkis, 2015; Anthony and Shawe-Taylor, 1993).

Key argument I: conditioning trick, control of $\mathbb{E}(\mathbf{Z})$

- $K \sim \text{Binomial}(n, p)$: random number of extreme points $X_i \in \mathbb{A}$

- Conditionally on K:

$$\Big[P_n(A), A \in \mathcal{A} \ \ |K = k\Big] \stackrel{d}{=} \Big(\frac{k}{n} P'_{\mathbb{A},k}(A), A \in \mathcal{A}\Big)$$

  where $P'_{\mathbb{A},k}$: empirical sample of an independant sample $(Y_i, i \leq k)$ following $P(\cdot \ |Y \in \mathbb{A})$

- Consequence

$$\underbrace{\mathbb{E}\left(\sup_A |P_n - P|(A)\right)}_{\mathbb{E}(\mathbf{Z})} \leq \mathbb{E}\left(\mathbb{E}\left(\frac{K}{n} \sup_A |P'_{\mathbb{A},K} - P_{\mathbb{A}}(A)| \ \Big| \ K\right)\right) + \sqrt{p/n}$$

...VC inequality conditional on K + concavity

$$\leq \sqrt{\frac{2p}{n}(\log 2 + \mathcal{V}_{\mathcal{A}} \log(2np + 1))} + \sqrt{p/n}$$

Key argument II: Concentration with small variance

$$\mathbf{Z} - \mathbb{E}(\mathbf{Z})?$$

- use $\mathbb{V}\mathrm{ar}(\mathbb{1}_A(X_i)) \leq p \ll 1$

- Replace the bounded difference inequality with a Bernstein-like uniform bound, also proved in McDiarmid (1998) incorporating $\mathbb{V}\mathrm{ar}$, by martingale arguments

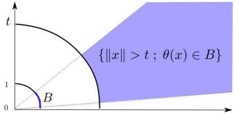
- Result: with proba $1 - \delta$,

$$\mathbf{Z} - \mathbb{E}(\mathbf{Z}) \leq 2\sqrt{\frac{p}{n}\log(1/\delta)} + \frac{2\log(1/\delta)}{3n}$$

- 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:

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

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)

Outline

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,

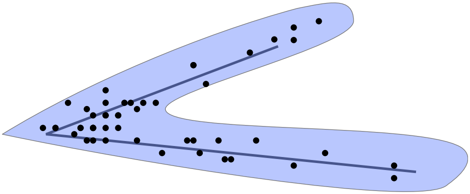
and we show:

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

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 $d, \mathcal{V}_\Gamma$.

Anomaly detection

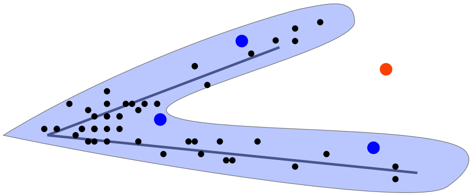
(Goix et al., 2016; Thomas et al., 2017)



- **Training step:**
  Learn a **'normal region'** (*e.g.* approximate support)

Anomaly detection

(Goix et al., 2016; Thomas et al., 2017)



- **Training step:**
  Learn a **'normal region'** (*e.g.* approximate support)
- **Prediction step:** (with new data)
  Anomalies = points outside the 'normal region'

If 'normal' data are heavy tailed, **Abnormal** $\not\Leftrightarrow$ **Extreme** .
There may be **extreme** 'normal data'.

How to distinguish between large anomalies and normal extremes?

Anomaly detection and Minimum Volume sets (no EVT)

- Multivariate generalizations of quantiles, Scott and Nowak (2006); Polonik (1997); Cai et al. (2011)
- At fixed 'level' α (=1-false positive), $\mathcal{G}$ a class of subsets of $\mathcal{X}$:

$$\Omega^*_\alpha \in \arg\min_{\Omega \in \mathcal{G}} \lambda(\Omega) \text{ s. t. } P(\Omega) \geq \alpha.$$

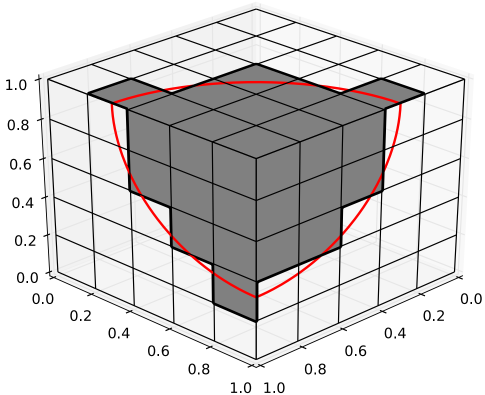
- new X_{test} flagged as abnormal if $X_{test} \notin \Omega^*_\alpha$
- Ω^*_α is the 'best' normality set (Neyman Pearson) if anomalies are uniformly distributed according to reference measure λ.
- Ω^*_α is a level set of the density if $\mathcal{G} =$ all measurable sets
- non asymptotic bounds for $\lambda(\widehat{\Omega})$ (false negative) and $P[\widehat{\Omega}]$ (true negative) in Scott and Nowak (2006).

Angular MV sets for extremes Thomas et al. (2017)

- $\mathcal{G}$ class of subsets of the sphere
- Construct angular MV sets for extremes, based on $\widehat{\Phi}$

$$\widehat{\Omega}_\alpha = \arg\min_{\mathcal{G}} \lambda(\Omega) \text{ s.t. } \widehat{\Phi}(\Omega) \geq \alpha - \psi(\delta)$$

where ψ: tolerance $\geq$ (supremum) error bound for $\widehat{\Phi}$.



Angular Guarantees and heuristic strategy

With probability $1-\delta$, (radially integrated) error rates are controlled,

$$\Phi(\widehat{\Omega}_\alpha) \geq \alpha - 2\psi\,, \qquad \text{and}$$
$$\lambda(\widehat{\Omega}_\alpha) \;\leq\; \inf\left\{\lambda(A):\; A \in \mathcal{A},\, \Phi(A) \geq \alpha\right\}.$$

- Heuristic tail scoring function

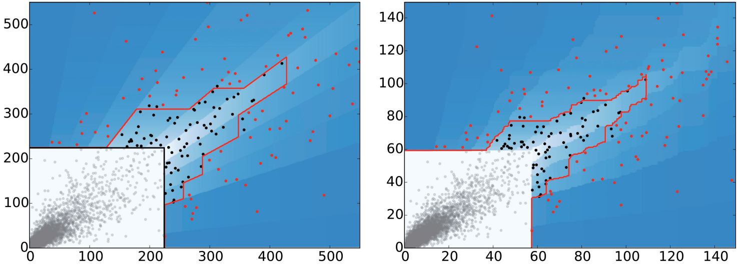
$$\widehat{s}(v) = \widehat{s}_\theta(\theta(v)) * r(v)^{-2}$$

  where $\widehat{s}_\theta(\theta(v))$ is constructed from a family of nested angular volume sets $(\widehat{\Omega}_{\alpha(1)}, \ldots, \widehat{\Omega}_{\alpha(J)})$

- (Justified when $\widehat{s}_\theta$ estimates in fact the density of V)

Anomaly detection in the tail

- $X_{new} \rightarrow$ rank-transform $\hat{V}$
- If $\|\hat{V}\|_\infty \leq$ training threshold for $\widehat{\Phi}$, use an AD algorithm for the bulk
- Otherwise score abnormality of X_{new} among extremes using $\widehat{s}(V)$
- Further guarantees in terms of Neyman Pearson as in Clémençon and Jakubowicz (2013); Clémençon and Thomas (2018)?



score on the transformed scale/original scale

DEMO 1: MVsets

Outline

Outline

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)

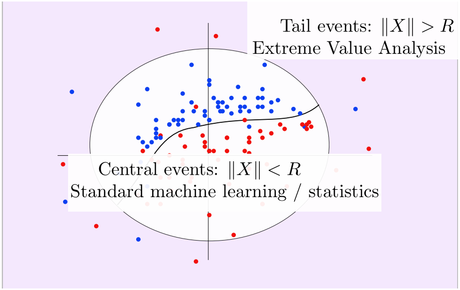
- **Goal:** make acurate prediction in **'crisis scenarios'** where new observed covariable are (unusally) large

- 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 $\to$ 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, talk later this week), no convexification. ...

- Work in progress: quantile regression, unbounded targets, more realistic algorithm (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 $\mathbb{P}(\|X\| > t)^{-1}$, 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 $k \le n$ of 'extremes'; Norm $\| \cdot \|$ on $\mathbb{R}^d$.

Selection of extremes: Sort the training data by decreasing radial order, $||X_{(1)}|| \ge \ldots \ge ||X_{(n)}||$ 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 $\|X\| \ge \|X_{(k)}\|$.

DEMO 2: Classification/Regression

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. $1-\delta$,

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

  $\widehat{R}^{>\tau}$, 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 $1-\delta$,

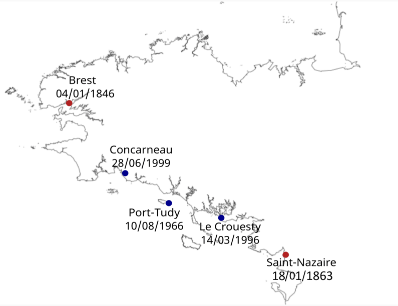
$$\sup_{h\in\mathcal{H}} \left| \widehat{R}_k(h\circ\theta) - R_{t(n,k)}(h\circ\theta) \right| \leq \frac{8M^2\sqrt{2\log(3/\delta)} + C\sqrt{V_{\mathcal{H}}}}{\sqrt{k}} + \frac{16M^2\log(3/\delta)/3 + 4M^2 V_{\mathcal{H}}}{k},$$

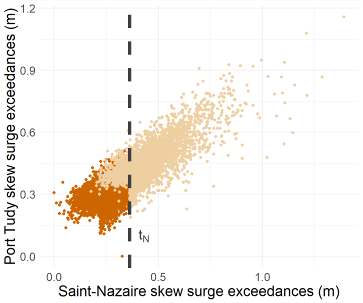
Outline

Extreme sea levels reconsruction

Huet et al. (2025)

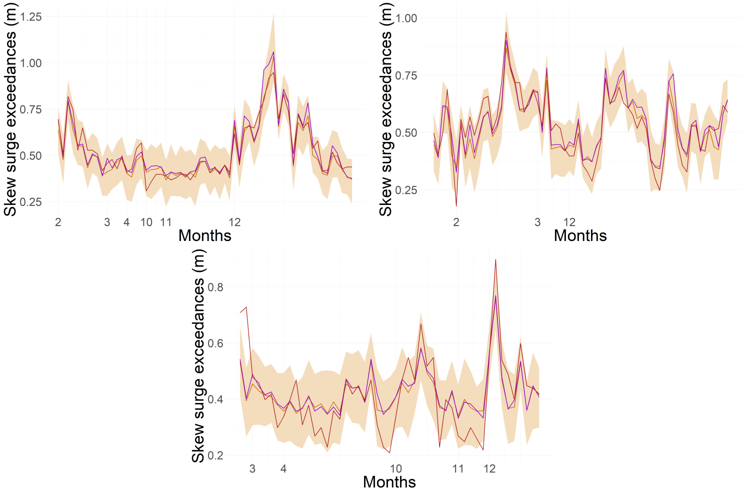
- Extreme surges (tidal component removed)
- Goal: reconstruct missing coastal gauges records from nearby stations with longer historical records
- Input stations: Brest, St Nazaire; output: Port Tudy, Concarneau, and Le Crouesty.





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 with GAN's

Jalalzai et al. (2020)

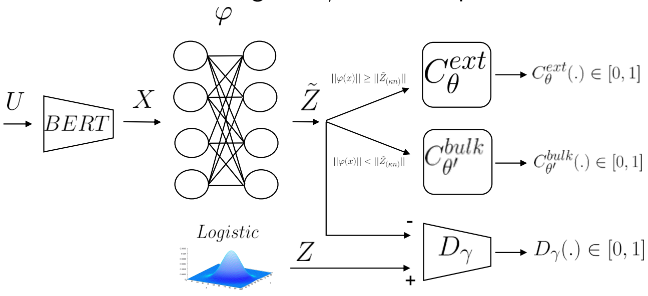
- Extension of the previous framework to datasets who are **NOT** regularly varying.
- Dataset: text embeddings (BERT). $X =$ vector in $\mathbb{R}^d$, d large (768).
- label $Y =$ positive/negative sentiment.
- Two goals:

  (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, GAN-like 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 (again)

- 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 $1 - 15\delta$

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

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

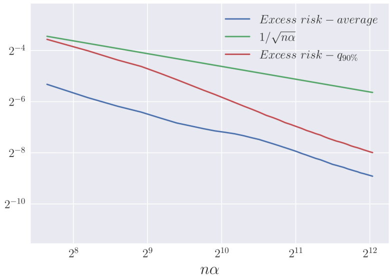
where $\widehat{u}$ is the minimizer of the CV risks $\widehat{R}_{\mathrm{CV},p}(\Psi_u V_{1:K}), u \in U$, and $E(n,K,p) = 5C\sqrt{(d+1)K/(np)} + 5K/((K-1)np)$.

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?

Outline

Dimension reduction in EVA - overview

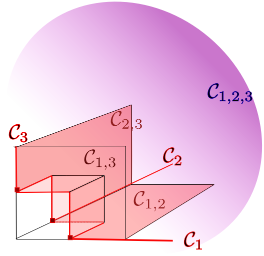
- First work (tmbk) Chautru et al. (2015): Assuming some mixture model for the angular measure, where each component is 'low dimensional'.
- Goix et al. (2016, 2017): notion of 'sparse' angular measure, determining which subgroups of components 'may' be simultaneously large. Finite sample error bounds. Applications to Anomaly Detection. Variants with alternative definitions of sparsity Meyer and Wintenberger (2021, 2024)
- In Janßen and Wan (2020); Jalalzai and Leluc (2021): clustering of extremes; with temporal dependence in Boulin et al. (2025b,a)

Outline

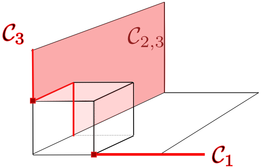
Sparse support recovery (Goix, S. Clémençon, 2016, 2017)

- Reasonable hope: the groups $\{X_j, j \in J\}$'s wich may be simultaneously large are (i) few (ii) small. $\rightarrow$ **sparse angular measure**

Our goal: Estimate the (sparse) support of the angular measure (*i.e.* the dependence structure).

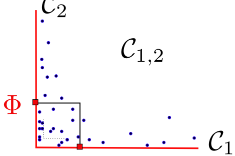


Full support:
anything may happen

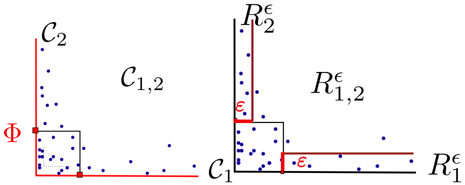


Sparse support
(X_1 not large if X_2 or X_3 large)

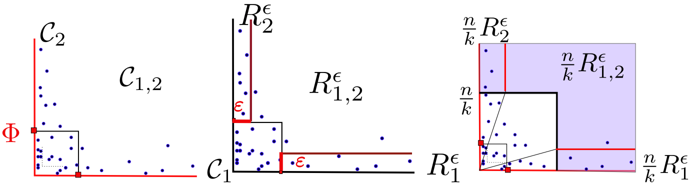
Identifying non empty subspaces (Goix et al. 2016)



Identifying non empty subspaces (Goix et al. 2016)

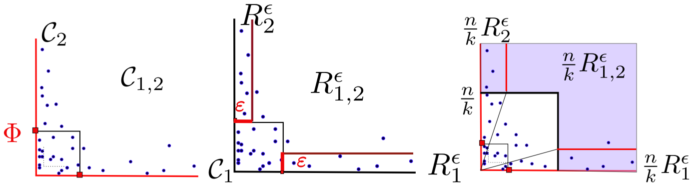


Identifying non empty subspaces (Goix et al. 2016)



Two parameters: (i) Tolerance parameter ε (otherwise empirical measure of subspaces $= 0$) ; (ii) k (number of observations considered as extremes)

Identifying non empty subspaces (Goix et al. 2016)



Two parameters: (i) Tolerance parameter ε (otherwise empirical measure of subspaces $= 0$) ; (ii) k (number of observations considered as extremes)

Theorem (Goix, S., Clémençon, 2016)

If the margins F_j are continuous and if the density of the angular measure is bounded by $M > 0$ on each subface (infinity norm),
There is a constant C s.t. for any n, d, k, $\delta \geq e^{-k}$, $\varepsilon \leq 1/4$, w.p. $\geq 1 - \delta$,

$$\max_{J \subset \{1,\dots,d\}} |\hat{\mu}_n(\mathcal{C}_J) - \mu(\mathcal{C}_J)| \leq Cd \left(\sqrt{\frac{1}{k\varepsilon} \log \frac{d}{\delta}} + Md\varepsilon \right) + \mathrm{Bias}_{\frac{n}{k},\varepsilon}(F, \mu).$$

$$\text{Regular variation} \iff \mathrm{Bias}_{t,\varepsilon} \xrightarrow[t\to\infty]{} 0$$

Related works

- Relaxed problem and feature clustering (Chiapino and Sabourin, 2016), statistical tests, Chiapino et al. (2019)
- link with hidden regular variation (Simpson et al., 2020)
- Modeling on multiple subspaces (Mourahib et al., 2024), (brand new) penalized method, mixture model (Mourahib et al., 2025)
- Clustering (Janßen and Wan, 2020), Spatial clustering and time series (Boulin et al., 2025a)

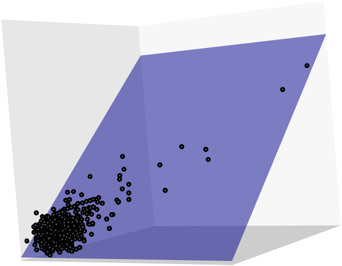
DEMO 3: DAMEX / CLEF

Outline

Principal Component Analysis for Extremes

- Motivation: assume that μ concentrates on $S^* \subset \mathbb{R}^d$ of dimension p.

- Consequence: Extreme shocks 'mostly' happen along directions $u \in S$, whyle for $u \notin S$, $\mathbb{P}(\langle u, X\rangle \gg 1)$ is comparatively negligible



- How can PCA recover the 'tail support' ($= S^*$)?

- Error control requires in theory 4th moments $\rightarrow$ what with heavy tails?

Angular PCA of extremes

- Main idea (Drees and S., 2021; Cooley and Thibaud, 2019): eigen decomposition of
$$\Sigma_t = \mathbb{E}\left(\Theta\Theta^\top \mid \|X\| > t\right)$$
with $\Theta = \theta(X) = \|X\|^{-1}X$, Euclidean norm.

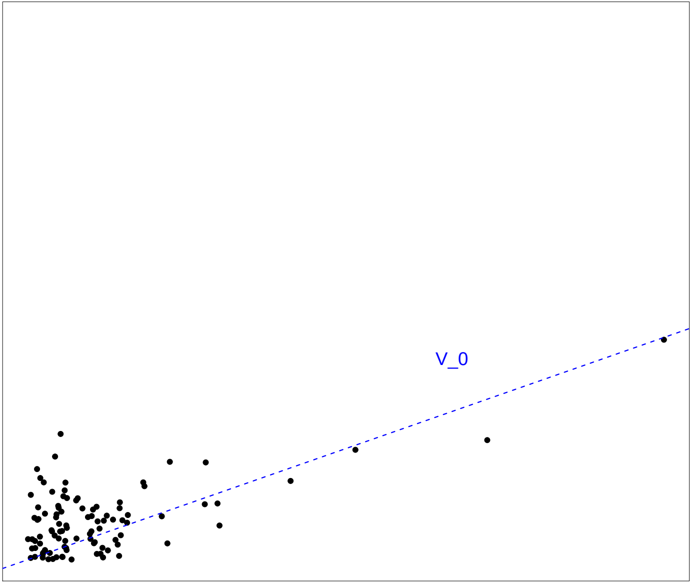
- Upcoming chapter (Drees & S.) in upcoming 'Handbook of Statistics of Extremes': a posteriori merging of independent similar ideas

- If X is regularly varying, then $\Sigma = \lim_{t\to\infty} \Sigma_t$ exists and
$$\Sigma_\infty = \mathbb{E}\left(\Theta_\infty\Theta_\infty^\top\right),$$
where $\Theta_\infty \sim \Phi$, with $\Phi = \lim \operatorname{Law}(\Theta \mid \|X\| > t)$.

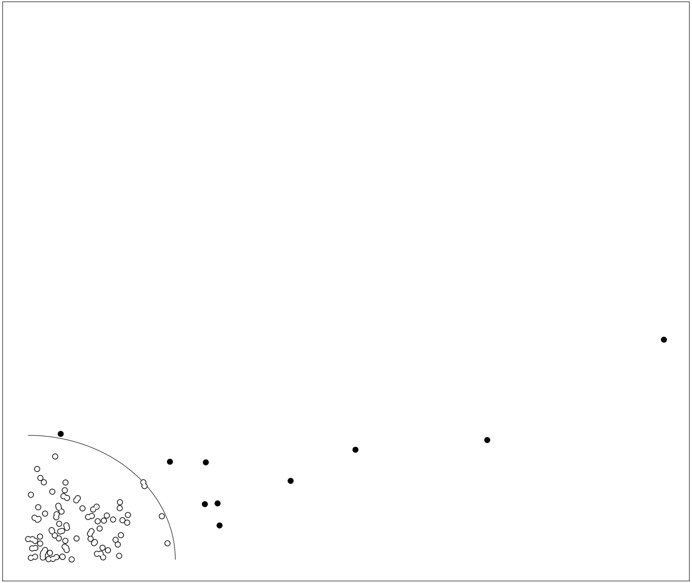
- Applications to rainfall data (Jiang et al., 2020), flood simulation (Rohrbeck and Cooley, 2023), follow-up work (faster rates) (Drees, 2025)

Angular PCA of extremes: in practice

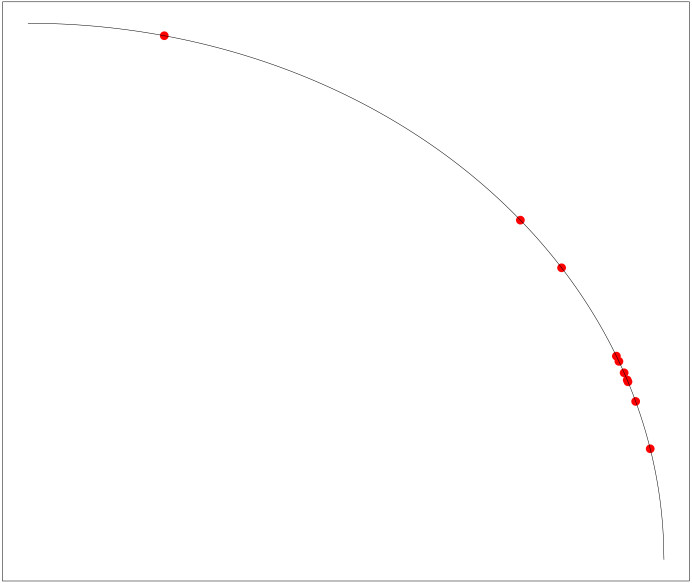
Angular PCA of extremes: in practice



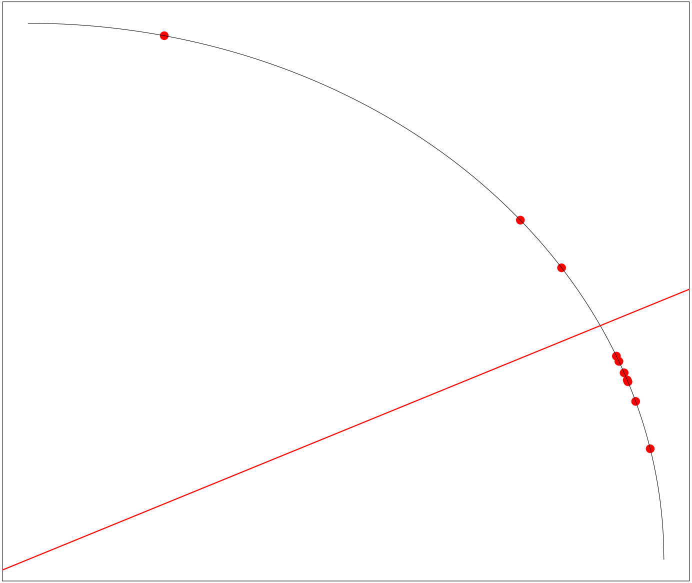
Angular PCA of extremes: in practice



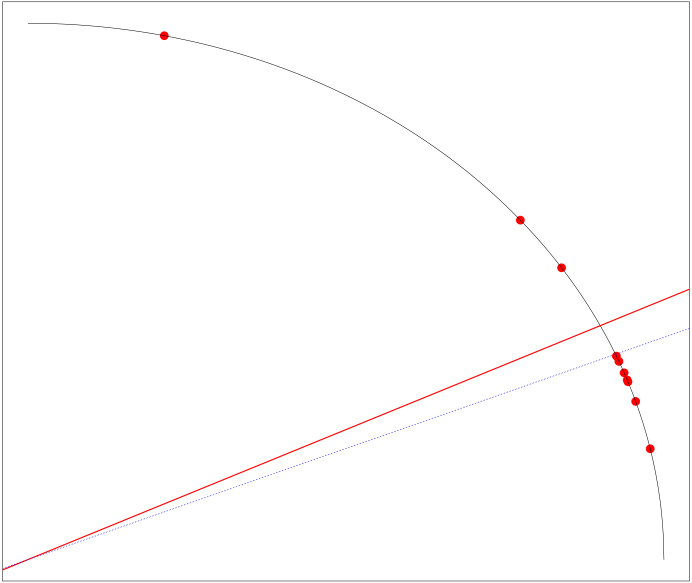
Angular PCA of extremes: in practice



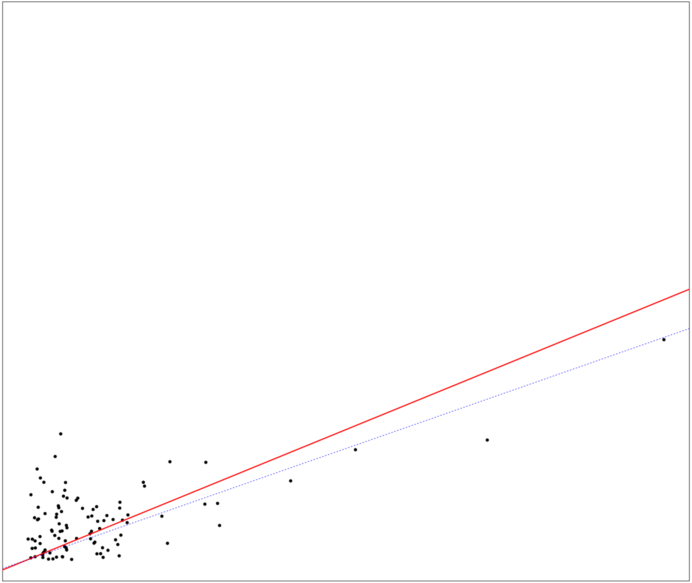
Angular PCA of extremes: in practice



Angular PCA of extremes: in practice



Angular PCA of extremes: in practice



Reconstruction risk analysis Drees and S. (2021)

- S: p-dim subspace, $\Pi_S^\perp(\Theta)$: orth. projection on $S^\perp$ (angular residual).
- Reconstruction Risk $R_t(S) = \mathbb{E}\left(\|\Pi_S^\perp \Theta\|^2 \mid \|X\| > t\right)$
- Risk Minimization over $\mathcal{S}_p :=$ p-dim subspaces $\to$ solution $S^*_{t,p}$ generated by first p eigenvectors of Σ_t (with distinct eignevalues).

Limit support recovery (Lemma 2.5)

If μ is concentrated on a p-dim subspace S^*, then S^* minimizes R_∞ over $\mathcal{S}_p$

Eigenspaces stability (Theorem 2.4)

$$\text{operator.norm.distance}(S^*_{t,p}, S^*_{\infty,p}) \to 0$$

Guarantees for empirical versions Drees and S. (2021)

- Empirical angular 2*nd* moments matrix $\widehat{\Sigma}_k$ with k largest observations, eigenspaces $\widehat{S}_{k,p}$. We show

Consistency

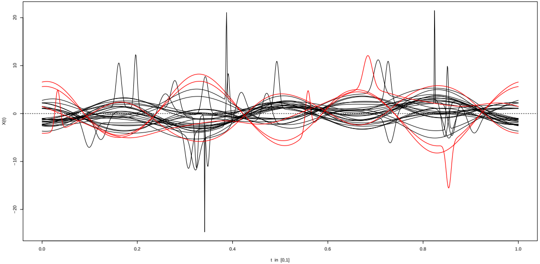
$\rho(\widehat{S}_{k,p}, S_p) \to 0$ as $k, n \to \infty, k/n \to 0$ in probability

Finite sample bounds on the reconstruction error

$$\sup_{S \in \mathcal{S}_p} |\widehat{R}_k(S) - R_{t(n,k)}(S)| \le \sqrt{\frac{\min(p, d-p)(1-k/n)\,\mathrm{tr}(\Sigma^2_{t(n,k)})}{k}} + \ldots$$

$$\sqrt{\frac{8(1+k/n)\log(4/\delta)}{k}} + \frac{4\log(1/\delta)}{3k}.$$

Functional extension (Clémençon et al., 2024)



- Focus: 'High energy' functional data (measured with L_2 norm)

- How to **adapt functional PCA to heavy-tailed functions** and obtain a finite-dimensional representation for the tail angular process $\lim \mathcal{L}(X/\|X\| \mid \|X\| > t)$ with $\|X\| = (\int X^2(s)\,\mathrm{d}s)^{1/2}$?

- Statistical **guarantees** under **regular variation in** $\mathbb{H} = L_2([0,1])$

From finite dim Drees and S. (2021) to L_2: bottlenecks

- Proofs in Drees and S. (2021) use **compactness** of the sphere: for stability and consistency of estimators $\hat{V}_k$.

- Excess risk bounds **depend on the dimension** (factor $\min(p, d-p)$)

Solution: *convergence of covariance operators* (in the HS norm)

- Useful reference for FDA Hsing and Eubank (2015)

Covariance operator in $\mathbb{H}$

- Generalization of rank-1 matrix $gf^\top$ in $\mathbb{R}^d$: For $f, g \in \mathbb{H}$,

$$\begin{aligned} f \otimes g : \mathbb{H} &\to \mathbb{H} \\ h &\mapsto \langle h, f \rangle g \end{aligned}$$

- Covariance operator of Z a random element in $\mathbb{H}$, with $\mathbb{E}(Z) = 0$:

$$\begin{aligned} \mathrm{C} = \mathbb{E}(Z \otimes Z) : \mathbb{H} &\to \mathbb{H} \\ h &\mapsto \mathrm{C}\, h = \mathbb{E}(\langle X, h \rangle X) \end{aligned}$$

  N.B. $\mathbb{E}$ is in the Bochner sense.

- C is a Hilbert-Schmidt, self adjoint operator $\Rightarrow$ **spectral theorem**

$$\mathrm{C} = \sum_{i \in \mathbb{N}} \lambda_j v_j \otimes v_j$$

  $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$: eigenvalues, $(v_j)_j$ eigenfunctions, form a CONS.

Conditional angular covariance operator in $\mathbb{H}$

- For extreme value analysis,

$$\mathrm{C}_t = \mathbb{E}\left(\Theta \otimes \Theta \mid \|X\| > t\right) \; ; \; \mathrm{C}_\infty = \mathbb{E}\left(\Theta_\infty \otimes \Theta_\infty\right).$$

Theorem Clémençon et al. (2024)

As $t \to \infty$, $\|C_t - C_\infty\|_{\mathrm{HS}(\mathbb{H})} \to 0$

Proof Short!
weak convergence of $\Theta_t \otimes \Theta_t$ + Skorohod's representation theorem (separability) + Jensen inequality (Bochner sense) + Dominated convergence (boundedness of Θ_t).

Corollary

If $\delta_p = \lambda_p - \lambda_{p+1} > 0$, then $\rho(S_{p,t}, S_{p,\infty}) \to 0$,
where $\rho(E, F) = \|\Pi_E - \Pi_F\|_{\mathrm{HS}(\mathbb{H})}$.

Proof: perturbation theory: deviations of eigen spaces/values of $C + \delta$ are controlled by $\|\delta\|_{\mathrm{HS}(\mathbb{H})}$.

Covariance Estimation and support recovery in $\mathbb{H}$

Theorem: Concentration of the empirical covariance

Let $\delta \in (0,1)$. With probability at least $1-\delta$, we have

$$\|\widehat{C}_k - C_{t_{n,k}}\|_{HS} \leq \frac{1}{\sqrt{k}} + 6\sqrt{\frac{\log(2/\delta)}{k}} + O\Big(\frac{\log(2/\delta)}{k}\Big).$$

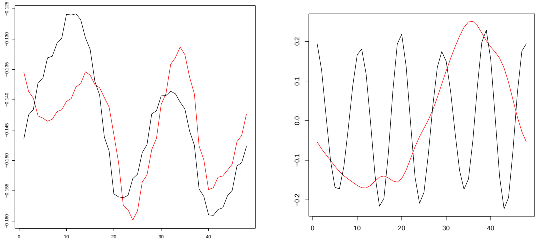
Corollary: $\rho(\hat{S}^p_{\hat{t}_{n,k}}, S^p_{t_{n,k}}) \leq \frac{1}{\text{spectral.gap}(p, t_{n,k})} \times \Big($ latter bound $\Big)$
(with multiplicative factor: spectral gap)

Corollary: Consistent estimation of S_p if $\gamma_p > 0$.
(Consequence of Weyl's inequality)

Tools for the proof
(i) 'Luckily' (boundedness) $\mathbb{E}\|\widehat{C}_k - C_{t_{n,k}}\| \leq 1/\sqrt{k}$
(ii) Mc Diarmid's Bernstein -type inequality applied to
$\varphi(X_1, \dots, X_n) = \|\bar{C}_k - C_{t_{n,k}}\|$

First eigenvector: real and simulated data

Black: Using all angles   Red: Using extreme angles.



real (left) and simulated (right) data.

real data `pm_10_gr_sqrt` Half-hour(squared-root) measurements of concentration in particulate matter, over 24h. Package `ftsa`. $n = 182$, $d = 48$.

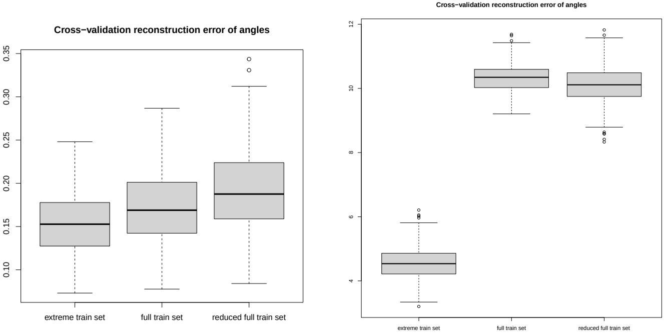
simulated data $X = \sum_1^4 A_j e_j$, with

e_j: trigonometric polynomials ; A_j: Pareto variables.

A_1, A_2 have common (heavier) tail index than (A_3, A_4), all have comparable variance. $n = 500$.

Reconstruction Error above large t, Cross-Validation

- Reconstruction of extreme angles above radial quantile $1 - k/n = 0.78$ (real data) or 0.9 (simulated data)
- Comparison: train on
  (i) extreme angles, (ii) all angles, (iii) subsample of size k among all angles.



real data (left) ; simulated data (right)

The End

- This talks: contributions to setting up a theoretical grounding for ML approaches in EVA, proof techniques, implementation with illustrative purpose

- More open (theoretical and applied) 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).

  Supervised dimension reduction Gardes (2018); Aghbalou et al. (2024b); Bousebata et al. (2023); Girard and Pakzad (2024)

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.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375.

Boulin, A., Di Bernardino, E., Laloë, T., and Toulemonde, G. (2025a). High-dimensional variable clustering based on maxima of a weakly dependent random process. *Journal of the American Statistical Association*, (just-accepted):1–21.

Boulin, A., Di Bernardino, E., Laloë, T., and Toulemonde, G. (2025b). Identifying regions of concomitant compound precipitation and wind speed extremes over europe. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlaf014.

References II

Bousebata, M., Enjolras, G., and Girard, S. (2023). Extreme partial least-squares. *Journal of Multivariate Analysis*, 194:105101.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer.

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.

Chautru, E. et al. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic journal of statistics*, 9(1):383–418.

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

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.

References III

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., Huet, N., and Sabourin, A. (2024). Regular variation in Hilbert spaces and principal component analysis for functional extremes. *Stochastic Processes and their Applications*, 174:104375.

Clémençon, S. and Jakubowicz, J. (2013). Scoring anomalies: a m-estimation formulation. In *Artificial Intelligence and Statistics*, pages 659–667. PMLR.

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*.

Clémençon, S. and Thomas, A. (2018). Mass volume curves and anomaly ranking. *Electron. J. Statist.*, 12(2):2806 – 2872.

References IV

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.

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.

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*.

Drees, H. (2025). Asymptotic behavior of principal component projections for multivariate extremes. *arXiv preprint arXiv:2503.22296*.

Drees, H. and S., A. (2021). Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943.

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.

References V

Einmahl, J. H. J., de Haan, L., and Li, D. (2006). Weighted approximations of tail copula processes with application to testing the bivariate extreme value condition. *Ann. Statist.*, 34(4):1987–2014.

Einmahl, J. H. J., Krajina, A., and Segers, J. (2012). An m-estimator for tail dependence in arbitrary dimensions. *Ann. Statist.*, 40(3):1764–1793.

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 VI

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.

Heffernan, J. E. and Resnick, S. I. (2007). Limit laws for random vectors with an extreme component. *The Annals of Applied Probability*, 17(2):537–571.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

References VII

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*.

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

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American statistical association*, 114(525):434–444.

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.

Jalalzai, H. and Leluc, R. (2021). Feature clustering for support identification in extreme regions. In *International Conference on Machine Learning*, pages 4733–4743. PMLR.

References VIII

Janßen, A. and Wan, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233.

Jiang, Y., Cooley, D., and Wehner, M. F. (2020). Principal component analysis for extremes and application to us precipitation. *Journal of Climate*, 33(15):6441–6451.

Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized pareto distributions. *Technometrics*, 61(1):123–135.

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.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

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

Lhaut, S. and Segers, J. (2024). An asymptotic expansion of the empirical angular measure for bivariate extremal dependence. In *Recent Advances in Econometrics and Statistics: Festschrift in Honour of Marc Hallin*, pages 187–208. Springer.

References IX

Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, Berlin.

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.

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

Mourahib, A., Kiriliouk, A., and Segers, J. (2024). Multivariate generalized pareto distributions along extreme directions. *Extremes*, pages 1–34.

Mourahib, A., Kiriliouk, A., and Segers, J. (2025). A penalized least squares estimator for extreme-value mixture models. *arXiv preprint arXiv:2506.15272*.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24.

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

References X

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.

Rohrbeck, C. and Cooley, D. (2023). Simulating flood event sets using extremal principal components. *The Annals of Applied Statistics*, 17(2):1333–1352.

Rootzén, H., Segers, J., and L. Wadsworth, J. (2018a). Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145.

Rootzén, H., Segers, J., and Wadsworth, J. L. (2018b). Multivariate generalized pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131.

Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized pareto distributions. *Bernoulli*, 12(5):917–930.

Scott, C. and Nowak, R. (2006). Learning minimum volume sets. *JMLR*, 7:665–704.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Thomas, A., Clemencon, S., Gramfort, A., and Sabourin, A. (2017). Anomaly detection in extreme regions via empirical mv-sets on the sphere. In *AISTATS*, pages 1011–1019.

References XI

van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

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.

Wadsworth, J. L., Tawn, J. A., Davison, A. C., and Elton, D. M. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(1):149–175.