Learning the Dependence Structure of Rare Events: a Non-Asymptotic Study

Nicolas Goix   Anne Sabourin   Stéphan Clémençon

Motivation

Learning the Dependence Structure of **Rare** Events

- **Extreme Value Theory (EVT)** for ML
  - Learning the unusual
  - $\neq$ averaging effect / mean behaviour
  - $\rightarrow$ application to Anomaly Detection
- **EVT** by Statistical Learning
  - VC-type bounds for estimating the **Asymptotic Dependence Structure.**

Framework and Extreme Dependence Structure

- **Context**
  - Random vector $X = (X_1, \dots, X_d)$
  - Margins: $X_j \sim F_j$ (F_j continuous)
- **Preliminary step: Standardization of each marginal**
  - Standard Pareto: $V_j = \frac{1}{1-F_j(X_j)}$ $\mathbb{P}(V_j \geq x) = \frac{1}{x}$, $x \geq 1$

Goal: $\mathbb{P}(V \in A)$? (A 'far from the origin').

Fundamental hypothesis and consequences

Standard assumption: let A extreme region,

$$\mathbb{P}[V \in t\, A] \simeq t^{-1}\mathbb{P}[V \in A] \quad \text{(radial homogeneity)}$$

Formally,

regular variation (after standardization):

$$0 \notin \overline{A}$$

$$t\mathbb{P}[V \in t\ A] \rightarrow \mu(A), \quad t \rightarrow \infty. \qquad \mu: \text{ exponent measure}$$

Necessarily: $\mu(tA) = t^{-1}\mu(A)$

$\Rightarrow$ **angular measure** on sphere S_{d-1}: $\Phi(B) = \mu\{tB, t \geq 1\}$

General model in multivariate EVT

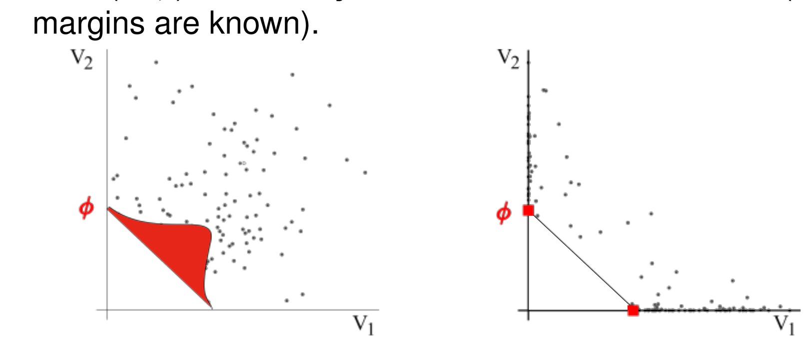
Model for excesses

For an extreme region A: $\mathbb{P}[\mathbf{V} \in A] \simeq \mu(A)$

$\Leftrightarrow$ For a large $r > 0$ and a region on the unit sphere B:

$$\mathbb{P}\left[\|\mathbf{V}\| > r, \ \frac{\mathbf{V}}{\|\mathbf{V}\|} \in B\right] \simeq \frac{1}{r}\Phi(B)$$

$\Rightarrow$ Φ (or μ) **rules the joint distribution of extremes** (if margins are known).



$\Rightarrow$ **Anomaly Detection**:
- μ or φ = "normal behaviour" in extreme regions
- $\rightarrow$ precision in extreme regions - better false alarm rate

The Standard Tail Dependence Function (STDF)

Why considering the STDF ?

Problem: Hard to study deviation of empirical $\hat{\mu}_n$ (or $\hat{\varphi}_n$) (existing work: $d = 2$)

Idea: Consider the **restriction** of μ **to a convenient VC-class**:

stable tail dependence function (STDF)

$$l(x) = \mu([0, x^{-1}]^c)$$

The STDF l is an analytic tool:
- knowledge of l $\Rightarrow$ knowledge of μ $\Rightarrow$ structure of extremes
- 'trick': allows to work on rectangles

$$\mu(A) = \lim_{t\to\infty} t\mathbb{P}(V \in tA)$$

spectral measure μ

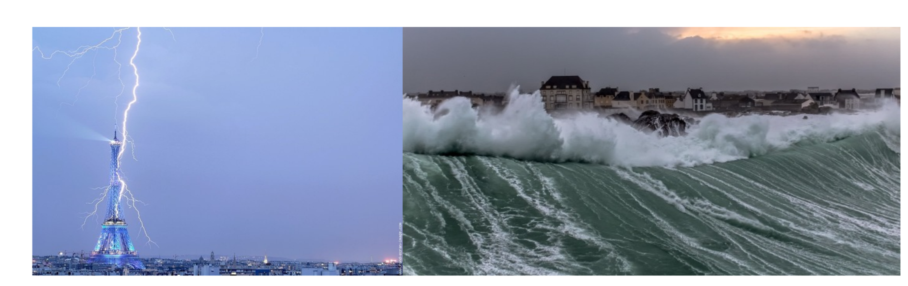
$\Updownarrow$

stdf $l = \mu\,\big|_{\text{VC-class}}$

$$l(x) = \mu([0, x^{-1}]^c) = \lim_{t\to\infty} \mathbb{P}(V_1 \geq tx_1^{-1} \text{ or } V_2 \geq tx_2^{-1})$$

Intuition behind the STDF

Two rare events $X_1 \geq x_1$, $X_2 \geq x_2$ with small proba p_1, p_2



Suppose that we know p_1 and p_2. Investigate:

$$p_{12} = \mathbb{P}(X_1 \geq x_1 \text{ or } X_2 \geq x_2)$$

STDF l verif es:

$p_{12} \simeq l(p_1, p_2)$ (if p_1 and p_2 small enough)

Estimation of the STDF

Related work and goal

- Results on l: asymptotic normality, under smoothness assumption.
- **Goal:** Derive non-asymptotic bounds with no assumption other than existence (Leftrightarrow regular variation Assumption).

Standard estimator of l

$$l(x_1, x_2) = \lim_{t\to\infty} t\ \mathbb{P}\left(V_1 \geq tx_1^{-1} \text{ or } V_2 \geq tx_2^{-1}\right)$$

$t \rightarrow \frac{k}{n}$
$V \rightarrow \hat{V}$

$$l_n(x_1, x_2) := \frac{n}{k}\ \hat{\mathbb{P}}_n\left(\hat{V}_1 \geq \frac{n}{k}x_1^{-1} \text{ or } \hat{V}_2 \geq \frac{n}{k}x_2^{-1}\right)$$

with
- $k \rightarrow \infty$, $\frac{k}{n} \rightarrow 0$
- $V_j = (1 - F_j(X_j))^{-1}$ and $\hat{V}_j = (1 - \hat{F}_j(X_j))^{-1}$

$\hat{F}(X_j) = rank(X_j)/n$

Main Issue

Would like to use concentration inequality...

Usually: $\sup_{A\in\mathcal{A}} |(\mathcal{P} - \mathcal{P}_n)(A)|$

In our case: $\sup_{A\in\mathcal{A}} \frac{n}{k}\left|(\mathcal{P} - \mathcal{P}_n)\left(\frac{k}{n}A\right)\right|$

- scaling $\frac{n}{k}$: to compense the decreasing proba of $\frac{k}{n}A$.
- classical VC-inequality: $\frac{k}{n}$ nice but not used !
  $\rightarrow$ high proba bound in

$$\frac{n}{k} \times \sqrt{\frac{1}{n}\log\frac{1}{\delta}} \quad \rightarrow \quad \infty \quad !!$$

$\Rightarrow$ Needs to take into account that the proba of $\frac{k}{n}A$ is small.

Solution

Key: VC-inequality adapted to rare regions $\rightarrow$ bound in

$$\sqrt{\mathbf{p}}\ \frac{n}{k}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}$$

with p the probability to be in the union class $\cup_{A\in\mathcal{A}} A$.

$$\mathbf{p} \lesssim \frac{k}{n}$$

$\Rightarrow$ bound in

$$\sqrt{\frac{1}{k}\log\frac{1}{\delta}}$$

interpretation of k:
- $k \simeq$ to the 'number of data considered as extreme'
- $k \simeq$ number of data used for estimation

Final result

Theorem

With proba. $\geq 1 - \delta$:

$$\sup_{0\preceq x\preceq T} \left|l_n(x) - l(x)\right| \leq Cd\sqrt{\frac{T}{k}\log\frac{d+3}{\delta}} + \text{bias}(k, n, T)$$

- T: to bound $\sqrt{\mathbf{p}}$ ($x \preceq T \Leftrightarrow \mathbf{p} \leq T\frac{k}{n}$)
- bias: to avoid assumptions, 'how far are we in the tail ?'

References

- J. H. J. Einmahl, Andrea Krajina, J. Segers. An M-estimator for tail dependence in arbitrary dimensions, 2012.
- P. Embrechts, L. de Haan, X. Huang. Modelling multivariate extremes, 2000.
- L. de Haan, A. Ferreira. Extreme value theory, 2006
- V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization (with discussion), 2006.
- Colin McDiarmid. Concentration, 1998
- Y. Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics, 1997
- S. Resnick. Extreme Values, Regular Variation, Point Processes, 1987
- V. Vapnik and A. Chervonenkis. Theory of Pattern Recognition, 1974.

Conclusion

- Learning theory adapted to multivariate EVT
- Tools for the study of low probability regions
- Pave the way to the use of multivariate EVT in machine learning and anomaly detection (ongoing work)