Learning the Dependence Structure of Rare Events: a non-asymptotic study

Nicolas Goix, Anne Sabourin, Stéphan Clémencon
Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI

COLT, July 2015, Paris

Learning the Dependence Structure of **Rare** Events

- **multivariate Extreme Value Theory (EVT)** for ML
  - Learning the *unusual*
  - $\neq$ averaging effect / mean behavior
  - $\rightarrow$ application to Anomaly Detection

- Statistical Learning for **multivariate EVT**
  - VC-type bounds for estimating the **Asymptotic Dependence Structure.**

1. Multivariate EVT & Extreme Dependence

2. Estimation of the STDF

Framework

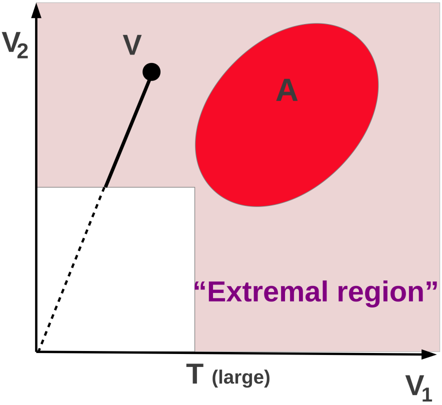
- **Context**
  - Random vector $\mathbf{X} = (X_1, \ldots, X_d)$
  - Margins: $X_j \sim F_j$ (F_j continuous)

- **Preliminary step: Standardization of each marginal**
  - Standard Pareto: $V_j = \frac{1}{1-F_j(X_j)}$ $\left(\mathbb{P}(V_j \geq x) = \frac{1}{x}, \quad x \geq 1\right)$

Problematic

Joint extremes: $\mathbf{V}$'s distribution above large thresholds?

$$\mathbb{P}(\mathbf{V} \in A)? \quad (A \text{ 'far from the origin'}).$$



Fundamental hypothesis and consequences

- Standard assumption: let A extreme region,

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

- Formally,

regular variation (after standardization):

$0 \notin \overline{A}$

$$t\mathbb{P}[\mathbf{V} \in t\, A] \xrightarrow[t\to\infty]{} \mu(A), \qquad \mu: \text{ exponent measure}$$

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

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

General model in multivariate EVT

Model for excesses

For an extreme region A:

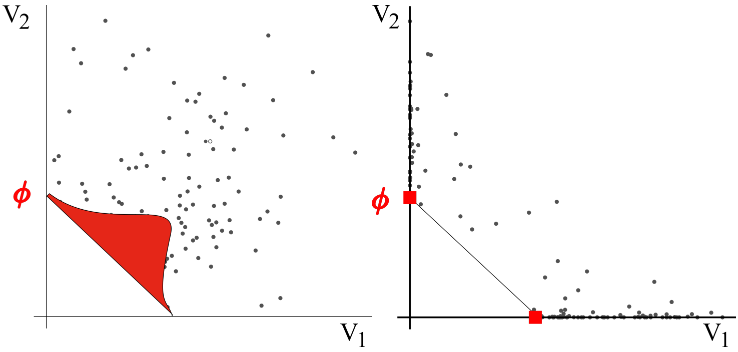
$$\mathbb{P}[\mathbf{V} \in A] \simeq \mu(A)$$

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

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

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

ϕ rules the joint distribution of extremes



$\Rightarrow$ **Anomaly Detection**:

- μ or ϕ = ”normal behavior” in extreme regions
- $\rightarrow$ precision in extreme regions - better false alarm rate

Why considering the STDF ?

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

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

stable tail dependence function (STDF)

$\mathbf{x} = (x_1, \dots, x_d), \quad \mathbf{x}^{-1} = (x_1^{-1}, \dots, x_d^{-1})$

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

The STDF l is an analytic tool:

- knowledge of $l \Rightarrow$ knowledge of $\mu \Rightarrow$ structure of extremes
- 'trick': allows to work on rectangles

Intuition behind the STDF

Two rare events $\left\{ \begin{array}{lll} X_1 & \geq & x_1 \\ X_2 & \geq & x_2 \end{array} \right.$ with small proba $\left\{ \begin{array}{l} p_1 \\ p_2 \end{array} \right.$

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 verifies:

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

Alternative definition of STDF

spectral measure μ $\Longleftrightarrow$ stdf $l = \mu\big|_{\text{VC-class}}$

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

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

- Equivalent definition of l:

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

- $\text{bias}(t, T) = \sup_{0\leq x_1, x_2\leq T} \left| l(x_1, x_2) - t\ \mathbb{P}\left(V_1 \geq tx_1^{-1} \text{ or } V_2 \geq tx_2^{-1}\right)\right|$

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 non parametric estimator of l

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

$t \to \frac{n}{k}$
$V \to \hat{V}$ yields the estimate of l:

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

with

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

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

Main Issue

Would like to use concentration inequality...

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

But usually: $\sup_{A \in \mathcal{A}} \left| (\mathcal{P} - \mathcal{P}_n)(A) \right|$

- scaling $\frac{n}{k}$
- classical VC-inequality: $\frac{k}{n}$ nice but not used !
  $\longrightarrow$ 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

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

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

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

$\Rightarrow$ bound in

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

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. $\geq 1 - \delta$:

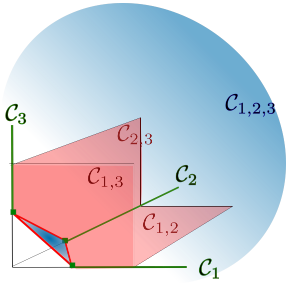
$$\sup_{0 \preceq x \preceq T} \left| l_n(x) - l(x) \right| \leq Cd\sqrt{\frac{T}{k}\log\frac{d+3}{\delta}} + \mathrm{bias}(\frac{n}{k}, T)$$

- T: to bound $\sqrt{\mathbf{p}}$ $\quad (x \preceq T \;\Leftrightarrow\; \mathbf{p} \leq dT\frac{k}{n})$
- bias $\to 0$ by existence of l. No assumptions needed about 'how far is k in the tail'.

$$\mathrm{bias}(\frac{n}{k}, T) = \sup_{0 \leq x_1, x_2 \leq T} \Big| l(x_1, x_2) - \frac{n}{k}\, \mathbb{P}\Big(V_1 \geq \frac{n}{k}x_1^{-1} \text{ or } V_2 \geq \frac{n}{k}x_2^{-1}\Big)\Big|$$

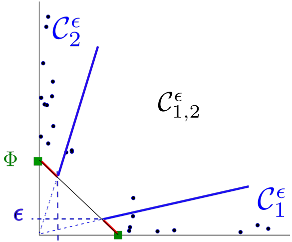
Idea for applications to Anomaly Detection

Structure of Φ in dimension 3 $\longrightarrow$



$2^d - 1$ faces on simplex

Hope: Sparse structure representing 'normal behavior'



Data are non-asymptotic
$\longrightarrow$ tolerance parameter ϵ
$\longrightarrow$ VC-class close to the one defining the STDF

Conclusion

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

Some 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.