Sparse Representation of Multivariate Extremes with Application to Anomaly Detection

Nicolas Goix **Anne Sabourin** **Stéphan Clémençon**

TELECOM ParisTech

INSTITUT Mines-Télécom

Goal & Motivations

Learning the Dependence Structure of **Rare** Events

Extreme Value Theory (EVT) + statistical learning

- Learning the unusual
- $\neq$ averaging effect / mean behaviour
- VC-type bounds for estimating the **Asymptotic Dependence Structure.**

For Anomaly Detection

- Extreme observations play a special role when dealing with outlying data.
- But no algorithm has **specific treatment for such multivariate extreme observations**.
- Our goal: Provide a method which can improve performance of standard AD algorithms by combining them with a **multivariate extreme analysis** of the **dependence structure**.

Framework and Extreme Dependence Structure

- **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: $V_j = \frac{1}{1-F_j(X_j)}$ $\mathbb{P}(V_j \geq x) = \frac{1}{x}$, $x \geq 1$

Goal: $\mathbb{P}(\mathbf{V} \in A)$? (A '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):

$$0 \notin \bar{A}$$

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

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

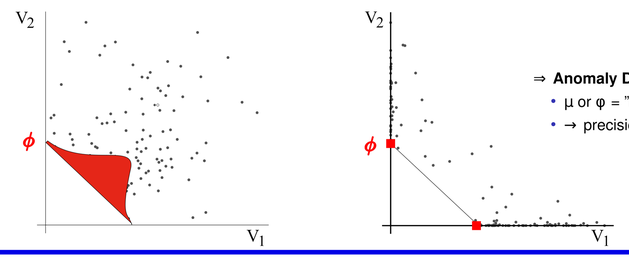
$\Rightarrow$ **angular measure** on sphere S_{d-1}: $\Phi(B) = \mu\{tB, t \geq 1\}$

General model in multivariate EVT

Intuitively: $\mathbb{P}[\mathbf{V} \in A] \simeq \mu(A)$ For a large $r > 0$ and a region B on the unit sphere:

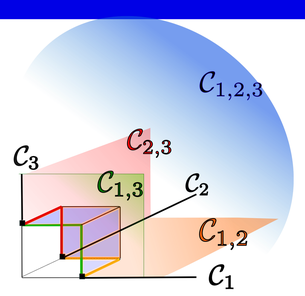
$$\mathbb{P}\left[\|\mathbf{V}\| > r, \ \frac{\mathbf{V}}{\|\mathbf{V}\|} \in B\right] \sim \frac{1}{r}\Phi(B) = \mu(\{tB, t \geq r\}) \ , r \to \infty$$

$\Rightarrow$ 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

Representation of extreme data



- Sub-cones:
  $\mathcal{C}_\alpha = \{v \geq 0, \ \|v\|_\infty \geq 1, \ v_j > 0 \ \text{ for } j \in \alpha, \ v_j = 0 \ \text{ for } j \notin \alpha\}$
- Corresponding sub-spheres:
  $\Omega_\alpha = \mathcal{C}_\alpha \cap \mathbf{S}_{d-1}$

- Natural decomposition of the angular measure :

$$\Phi = \sum_{\alpha \subset \{1,\ldots,d\}} \Phi_\alpha \qquad \text{with } \Phi_\alpha = \Phi_{|\Omega_\alpha} \leftrightarrow \mu_{|\mathcal{C}_\alpha}$$

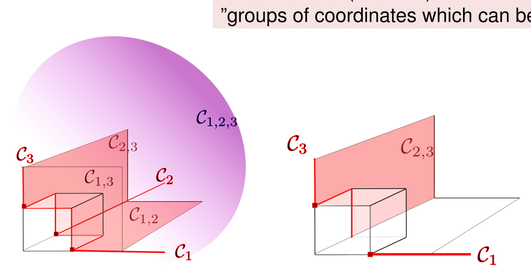
- $\Rightarrow$ yields a representation

$$\mathcal{M} = \left\{ \Phi(\Omega_\alpha): \quad \emptyset \neq \alpha \subset \{1, \ldots, d\} \right\}$$
$$= \left\{ \mu(\mathcal{C}_\alpha): \quad \emptyset \neq \alpha \subset \{1, \ldots, d\} \right\}$$

- Assumption: $\frac{d\mu_{|\mathcal{C}_\alpha}}{dv_\alpha} = O(1)$.
- Remark: Representation $\mathcal{M}$ is linear (after non-linear transform of the data $\mathbf{X} \to \mathbf{V}$).

Sparse dependence structure

Small number ($<<< 2^d$) of "groups of coordinates which can be large together"



Full pattern : anything may happen — $\mathcal{M}$ non-sparse

Sparse pattern (V_1 not large if V_2 or V_3 large) — $\mathcal{M}$ sparse

Estimation

Related work and goal

- asymptotic normality, under smoothness assumption
- **Goal:** Derive non-asymptotic bounds with no assumption other than existence ($\Longleftrightarrow$ regular variation)

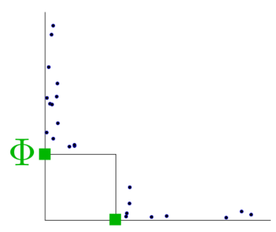
Problem: $\mathcal{M}$ is an **asymptotic** representation

$$\mathcal{M} = \{ \Phi(\Omega_\alpha), \ \alpha \} = \{ \mu(\mathcal{C}_\alpha), \ \alpha \}$$

is the restriction of an asymptotic measure

$$\mu(A) = \lim_{t \to \infty} t\mathbb{P}[\mathbf{V} \in t\,A]$$

to a representative class of set $\{\mathcal{C}_\alpha, \ \alpha\}$, but only the central sub-cone has positive Lebesgue measure!

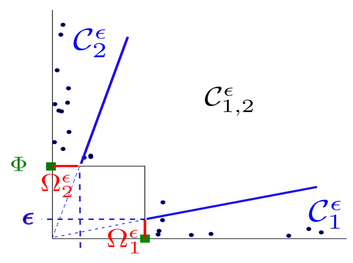


$\Rightarrow$ Cannot just do, for large t:

$$\Phi(\Omega_\alpha) = \mu(\mathcal{C}_\alpha) \simeq t\widehat{\mathbb{P}}(t\mathcal{C}_\alpha)$$

Solution

Fix $\epsilon > 0$. Affect data ϵ-close to an edge, to that edge.



$$\Omega_\alpha \to \Omega_\alpha^\epsilon = \{v \in \mathbf{S}_{d-1} : v_i > \epsilon \ (j \in \alpha), \ v_j \leq \epsilon \ (j \notin \alpha)\}.$$
$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\epsilon = \{t\,\Omega_\alpha^\epsilon, t \geq 1\}$$

New partition of $\mathbf{S}_{d-1}$, compatible with non asymptotic data.

$\hat{V}_i^j = \frac{1}{1-\hat{F}_j(X_i^j)}$ with $\hat{F}_j(X_i^j) = \frac{rank(X_i^j)-1}{n}$

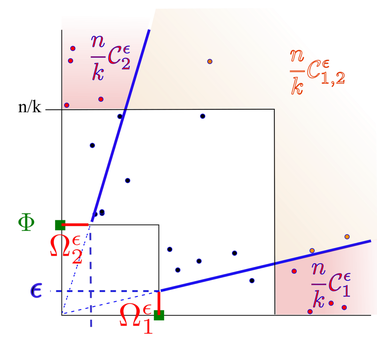
$\Rightarrow$ get an natural estimate of $\Phi(\Omega_\alpha)$

$$\widehat{\Phi}(\Omega_\alpha) := \frac{n}{k}\mathbb{P}_n(\hat{V} \in \frac{n}{k}\mathcal{C}_\alpha^\epsilon)$$

($\frac{n}{k}$ large, ϵ small)

$\Rightarrow$ we obtain

$$\widehat{\mathcal{M}} := \{ \widehat{\Phi}(\Omega_\alpha), \ \alpha \}$$



Final result

Theorem

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

$$\sup_{\emptyset \neq \alpha \subset \{1, \ldots, d\}} |\widehat{\mathcal{M}}(\alpha) - \mathcal{M}(\alpha)| \leq Cd\left(\sqrt{\frac{1}{\epsilon k}\log\frac{d}{\delta}} + Md\epsilon\right) + \text{bias}(\epsilon, k, n)$$

Comments:
- C: depends on $M = \sup$(density on subfaces)
- Existing litterature (for spectral measure) **Einmahl Segers 09, Einmahl *et.al.* 01**

$$d = 2.$$

asymptotic behaviour, rates in $1/\sqrt{k}$.
Here: $1/\sqrt{k} \to 1/\sqrt{\epsilon k} + \epsilon$. Price to pay for biasing our estimator with ϵ.

Algorithm

Algorithm

DAMEX in $O(dn\log n)$
Input: *parameters* $\epsilon > 0$, $k = k(n)$, $\Phi_{\min} \geq 0$.

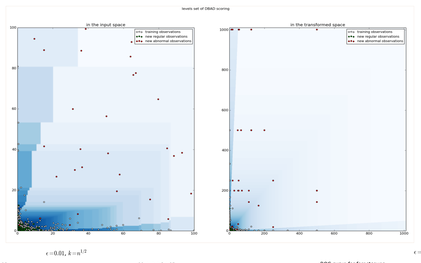
1. *Standardize via marginal rank-transformation:* $\hat{V}_i := \left(1/(1 - \hat{F}_j(X_i^j))\right)_{j=1,\ldots,d}$.
2. *Assign to each $\hat{V}_i$ the cone $\mathcal{C}_\alpha^\epsilon$ it belongs to.*
3. *Compute $\Phi_n^{\alpha,\epsilon} := \widehat{\Phi}(\Omega_\alpha) = \frac{n}{k}\mathbb{P}_n(\hat{V} \in \frac{n}{k}\mathcal{C}_\alpha^\epsilon)$ the estimate of the α-mass of Φ.*
4. *Set to 0 the $\Phi_n^{\alpha,\epsilon}$ below some small threshold $\Phi_{\min} \geq 0$ to eliminate cones with negligible mass*

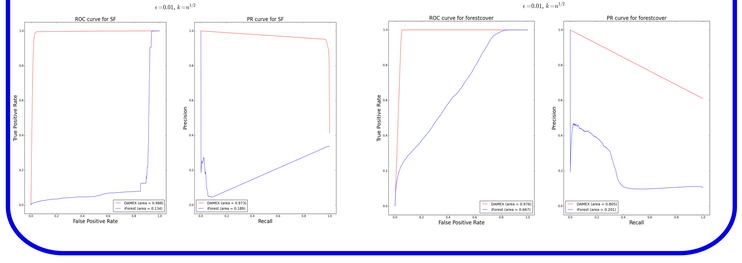
Output: *(sparse) representation of the dependence structure*

$$\widehat{\mathcal{M}} := (\Phi_n^{\alpha,\epsilon})_{\alpha \subset \{1,\ldots,d\}, \Phi_n^{\alpha,\epsilon} > \Phi_{\min}}$$

Anomaly Detection

Detecting Anomalies with Multivariate Extremes (DAMEX): anomalies = points violating the sparsity pattern.





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.