

Scoring anomalies among multivariate extreme observations

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

Plan

- 1 What is Anomaly Detection?
- 2 Multivariate EVT & Extreme Dependence
- 3 Sparse dependence structure
- 4 Estimation
- 5 Experiments

Anomaly Detection (AD)

What is Anomaly Detection ?

"Finding patterns in the data that do not conform to expected behavior"

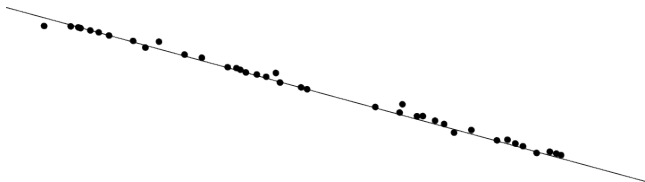


Machine Learning context

Different kind of Anomaly Detection

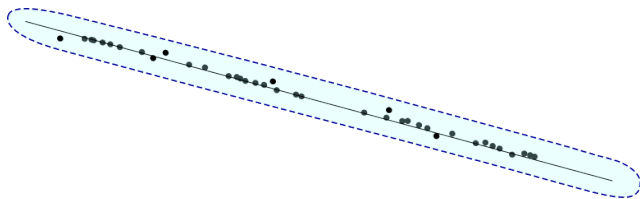
- **Supervised AD**
 - Labels available for both normal data and anomalies
 - Similar to rare class mining
- **Semi-supervised AD**
 - Only normal data available to train
 - The algorithm learns on normal data only
- **Unsupervised AD**
 - no labels, training set = normal + abnormal data
 - Assumption: anomalies are very rare

Anomaly Detection Schemes



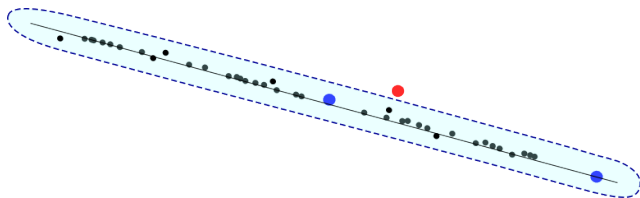
- **Step 1: Learn a profile** of the "normal" behavior
Profile can be patterns, summary statistics,...

Anomaly Detection Schemes



- **Step 1: Learn a profile** of the "normal" behavior
Profile can be patterns, summary statistics,...
- **Step 2:** Use the "normal" profile to build a **decision function**.

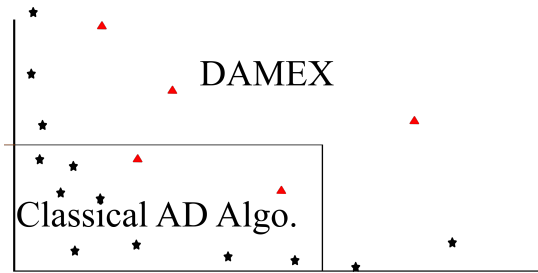
Anomaly Detection Schemes



- **Step 1: Learn a profile** of the "normal" behavior
Profile can be patterns, summary statistics,...
- **Step 2:** Use the "normal" profile to build a **decision function**.
- **Step 3: Detect anomalies** among new observations.
Anomalies are observations whose characteristics differ significantly from the normal profile

General idea of our work

- 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**.



Towards high dimension

Main idea :

Sparsity Assumption on the dependence structure:

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

Our Goal:

Learn the **sparsity pattern** of the dependence structure

Applications:

- **Detecting Anomalies with Multivariate Extremes** (DAMEX): anomalies = points violating the sparsity pattern.
- **Dimension reduction** If furthermore only a moderate number of X_j 's may be large together. Preliminary step before inference of the joint distribution of extremes.

Plan

- 1 What is Anomaly Detection?
- 2 Multivariate EVT & Extreme Dependence**
- 3 Sparse dependence structure
- 4 Estimation
- 5 Experiments

Framework

- **Context**

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

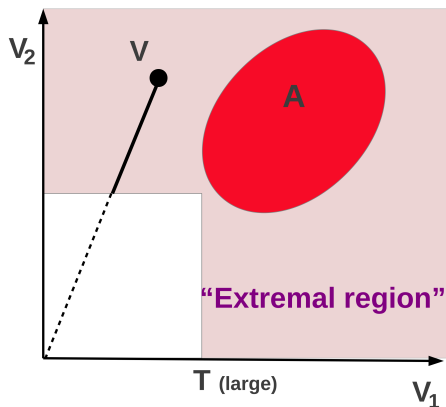
- **Preliminary step: Standardization of each marginal**

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

Problematic

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

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



Fundamental hypothesis and consequences

- Standard assumption: let A extreme region,

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

- Formally,

regular variation (after standardization):

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

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

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 :

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

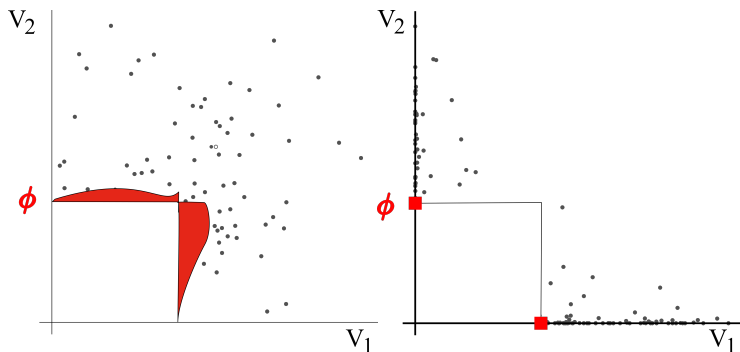
\Leftrightarrow For a large $r > 0$ and a region B on the unit sphere:

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

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

Angular distribution

- Φ rules the joint distribution of extremes



- ▶ Asymptotic dependence: (V_1, V_2) may be large together.

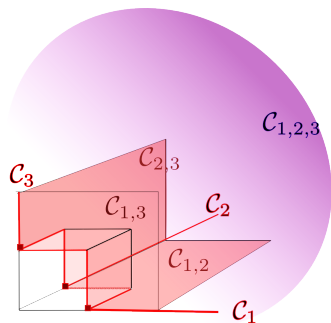
vs

- ▶ Asymptotic independence: only V_1 or V_2 may be large.

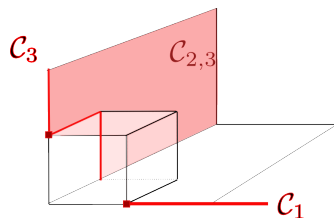
Plan

- 1 What is Anomaly Detection?
- 2 Multivariate EVT & Extreme Dependence
- 3 Sparse dependence structure**
- 4 Estimation
- 5 Experiments

Dependence structure



Full pattern :
anything may happen

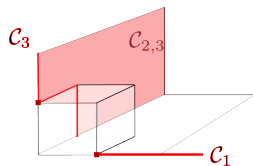


Sparse pattern
(V_1 not large if V_2 or V_3 large)

- Sub-cones:
$$\mathcal{C}_\alpha = \{ \|v\| \geq 1, v_i > 0 (i \in \alpha), v_j = 0 (j \notin \alpha) \}$$
- Corresponding sub-spheres: $\{ \Omega_\alpha, \alpha \subset \{1, \dots, d\} \}$
($\Omega_\alpha = \mathcal{C}_\alpha \cap \mathbf{S}_{d-1}$)

Representation of extreme data

$$\begin{aligned}\mathcal{M} &= \left\{ \Phi(\Omega_\alpha) : \emptyset \neq \alpha \subset \{1, \dots, d\} \right\} \\ &= \left\{ \mu(\mathcal{C}_\alpha) : \emptyset \neq \alpha \subset \{1, \dots, d\} \right\}\end{aligned}$$



Assumption: $\frac{d\Phi|_{\Omega_\alpha}}{dV_\alpha} = O(1)$.

Representation \mathcal{M} is natural and linear (after non-linear transform of the data $\mathbf{X} \rightarrow \mathbf{V}$).

Plan

- 1 What is Anomaly Detection?
- 2 Multivariate EVT & Extreme Dependence
- 3 Sparse dependence structure
- 4 Estimation**
- 5 Experiments

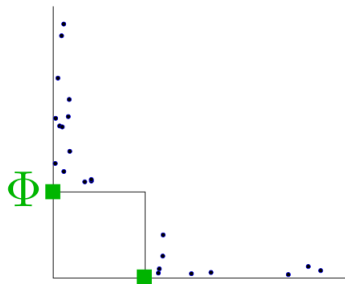
Problem: \mathcal{M} is an **asymptotic** representation

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

is the restriction of an asymptotic measure

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

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

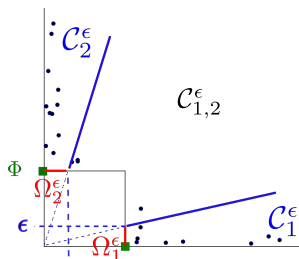


⇒ Cannot just do, for large t :

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

Solution

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



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

$$C_\alpha \rightarrow C_\alpha^\epsilon = \{t \Omega_\alpha^\epsilon, t \geq 1\}$$

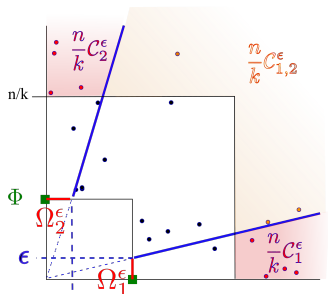
New partition of \mathbf{S}_{d-1} , compatible with non asymptotic data.

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

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

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

($\frac{n}{k}$ large, ϵ small)



⇒ we obtain

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

Theorem:

$$\|\hat{\mathcal{M}} - \mathcal{M}\|_\infty \leq Cd \left(\sqrt{\frac{1}{\epsilon k} \log \frac{d}{\delta}} + Md\epsilon \right) + \text{bias}(\epsilon, k, n),$$

Algorithm

DAMEX in $O(dn \log n)$

Input: parameters $\epsilon > 0$, $k = k(n)$, $\Phi_{\min} \geq 0$.

- 1 *Standardize via marginal rank-transformation:*
 $\hat{V}_i := (1/(1 - \hat{F}_j(X_i^j)))_{j=1, \dots, d}$.
- 2 *Assign to each \hat{V}_i the cone $\mathcal{C}_\alpha^\epsilon$ it belongs to.*
- 3 *Compute $\Phi_n^{\alpha, \epsilon} := \hat{\Phi}(\Omega_\alpha) = \frac{n}{k} \mathbb{P}_n(\hat{V} \in \frac{n}{k} \mathcal{C}_\alpha^\epsilon)$ 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*

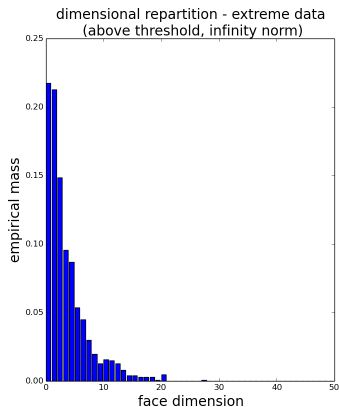
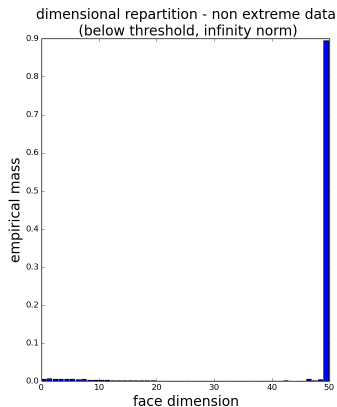
Output: (sparse) representation of the dependence structure

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

Sparse low-dimensional representation?

(Shell Research, and thanks J. Wadsworth)

data=50 wave direction from buoys in North sea.



nb of faces with positive mass
nb of faces with positive mass after thresholding
nb of faces with positive mass after 2nd thresholding

Non-extreme data	Extreme Data
2761	782
21	76
1	26

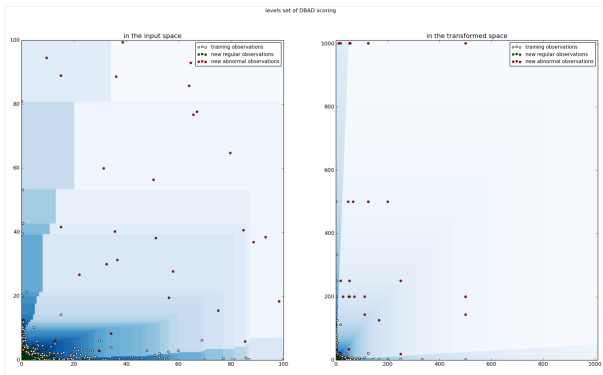
Application to Anomaly Detection

After standardization of marginals: $\mathbb{P}[R > r, \mathbf{W} \in B] \simeq \frac{1}{r} \Phi(B)$

→ scoring function = $\Phi_n^\epsilon \times 1/r$:

$$s_n(\mathbf{x}) := (1/\|\hat{T}(\mathbf{x})\|_\infty) \sum_{\alpha} \Phi_n^{\alpha, \epsilon} \mathbb{1}_{\hat{T}(\mathbf{x}) \in \mathcal{C}_\alpha^\epsilon}$$

where $T : \mathbf{X} \mapsto \mathbf{V}$ $(V_j = \frac{1}{1-F_j(X_j)})$



Plan

- 1 What is Anomaly Detection?
- 2 Multivariate EVT & Extreme Dependence
- 3 Sparse dependence structure
- 4 Estimation
- 5 Experiments**

	number of samples	number of features
shuttle	85849	9
forestcover	286048	54
SA	976158	41
SF	699691	4
http	619052	3
smtp	95373	3

Table: Datasets characteristics

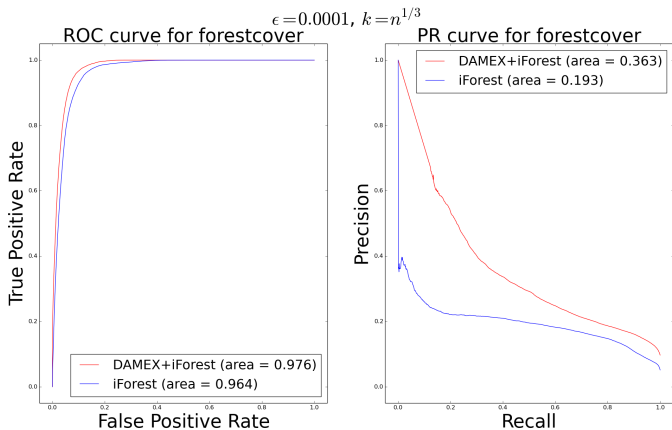


Figure: ROC and PR curve on forestcover dataset

$$\epsilon = 0.01, k = n^{1/2}$$

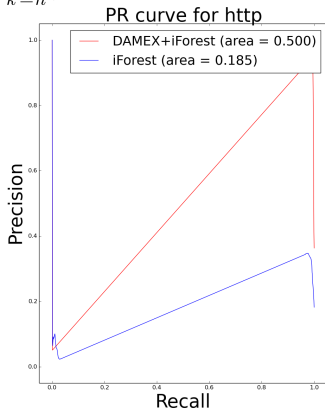
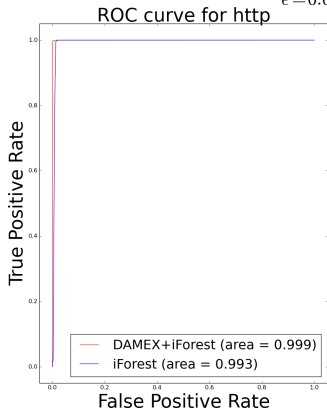


Figure: ROC and PR curve on http dataset

$$\epsilon = 0.01, k = n^{1/2}$$

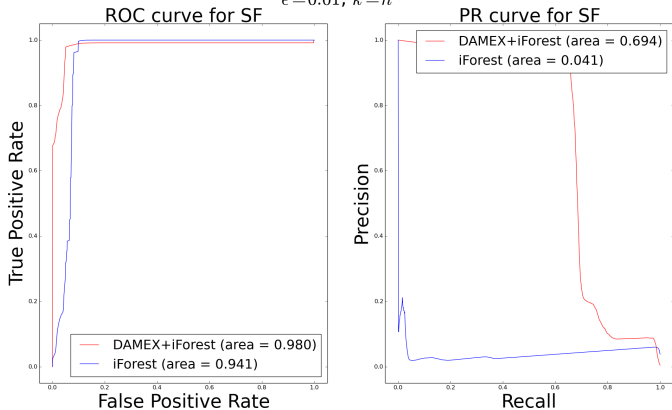


Figure: ROC and PR curve on SF dataset

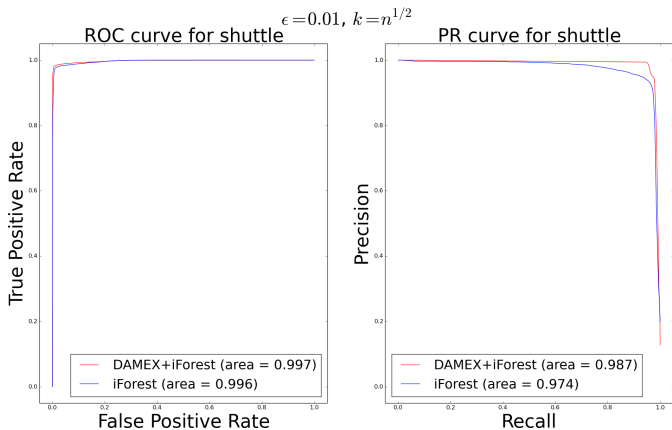


Figure: ROC and PR curve on shuttle dataset

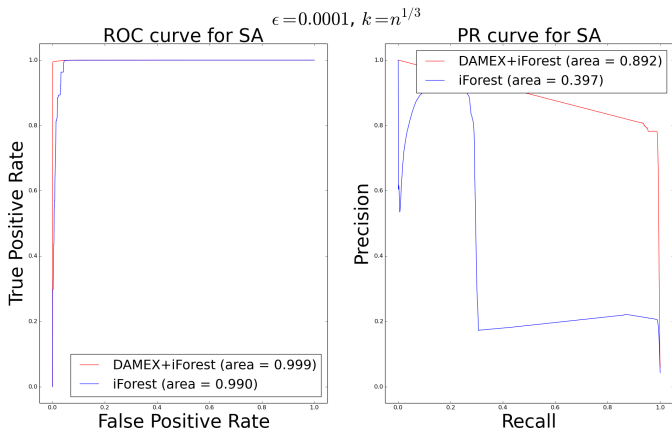


Figure: ROC and PR curve on SA dataset

- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey, 2009
- J. H. J. Einmahl , J. Segers Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution
- J. H. J. Einmahl, Andrea Krajina, J. Segers. An m-estimator for tail dependence in arbitrary dimensions, 2012.
- N. Goix, A. Sabourin, S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study.
- L. de Haan , A. Ferreira. Extreme value theory, 2006
- FT Liu, Kai Ming Ting, Zhi-Hua Zhou. Isolation forest, 2008
- 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
- S.J. Roberts. Novelty detection using extreme value statistics, Jun 1999
- J. Segers. Max-stable models for multivariate extremes, 2012