Scoring anomalies among multivariate extreme observations

Nicolas Goix, Anne Sabourin, Stéphan Clémencon

Institut Mines-Télécom, Télécom ParisTech, CNRS-LTCI

Séminaire de Statistique AgroParisTech, Mai 2015

Anomaly: ”an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism (Hawkins 1980)”

What is Anomaly Detection ?

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

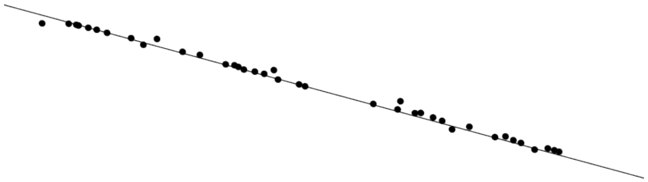
Synonyms of "Anomalies"

Outliers
Discordant observations
Exceptions
Aberrations
Peculiarities

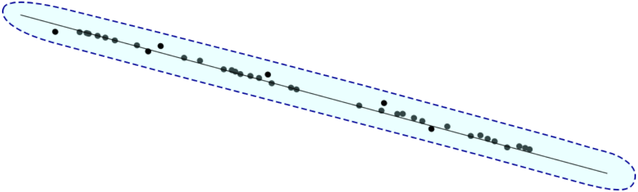
Huge number of applications: Network intrusions, credit card fraud detection, insurance, finance, military surveillance,...

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

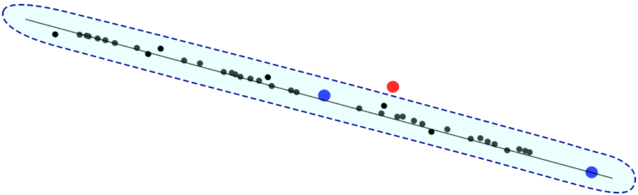


- **Step 1: Learn a profile** of the ”normal” behavior
  Profile can be patterns, summary statistics,...



- **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**.



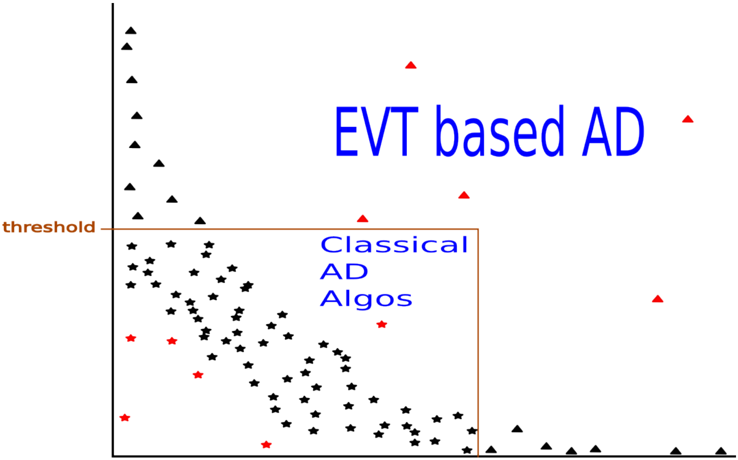
- **Step 1: Learn a profile** of the ”normal” behavior
  Profile can be patterns, summary statistics,...
- **Step 2:** Use the ”normal” profile to build a **decision function.**
- **Step 3: Detect anomalies** among new observations.
  Anomalies are observations whose characteristics differ significantly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

Different tools in Anomaly Detection:

- **statistical AD techniques**
  fit a statistical model for normal behavior
- **distance-based**
  - ex: Nearest Neighbors distance
  - Drawback: problem in high-dimensional spaces
- **density-based**
  - ex: Local Outlier Factor (LOF)
  - Drawback: Density estimation hard in high dimension
- **others**: spectral techniques (PCA), clustering-based, random forest,...

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**.


EVT based AD
threshold
Classical
AD
Algos

Main idea :

Sparsity Assumption on the dependence structure:

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

Our Goal:

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

Applications:

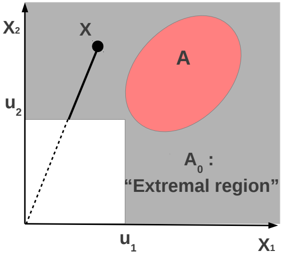
- **Dependence Based Anomaly Detection** (DBAD): anomalies = points violating the sparsity pattern.
- **Dimension reduction** If furthermore only a moderate number of X_j's may be large together. Preliminary step before inference of the joint distribution of extremes.

Program:

- Define adequate notion of ‘sparsity’ for multivariate extremes.
- Guidelines to learn this sparse asymptotic pattern from non-asymptotic data.
- Theoretical results: finite sample error bounds (tools from statistical learning theory)

- Random vectors $\mathbf{Y} = (Y_1, \ldots, Y_d) \quad Y_j \geq 0$
- Margins: $Y_j \sim F_j$, $1 \leq j \leq d$ (any).
- **Standardization** (standard Pareto): $X_j = (1 - F_j(Y_j))^{-1}$
- Joint extremes: $\mathbf{X}$'s distribution above large thresholds?

$$P(\mathbf{X} \in A)? \quad (A \subset A_0, \ \ A_0 \text{ `far from the origin'}).$$

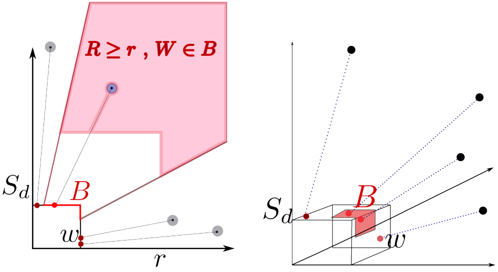


- Polar coordinates: $R = \|\mathbf{X}\|_\infty$; $\mathbf{W} = \frac{\mathbf{X}}{R}$.
- $\mathbf{W} \in$ sphere $\mathbf{S}_{d-1} = \{\mathbf{w} : \quad w_j \geq 0, \|w_j\|_\infty = 1\}$.

Characterize $\mathbb{P}(\mathbf{X} \in A), \; A \in A_0$

$\Leftrightarrow$

Characterize $\mathbb{P}(R > r, \mathbf{W} \in B), \; r > r_0$



- Regular variation: For $A \subset [0,\infty]^d \setminus \{0\}, \quad 0 \notin \bar{A}$,

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

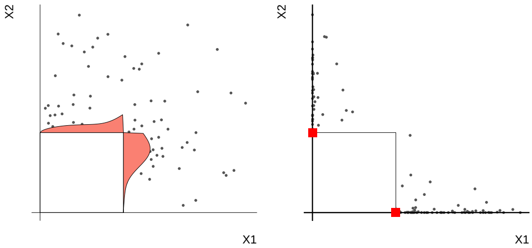
  Necessarily: $\mu(tA) = t^{-1}\mu(A)$, Radial homogeneity.
- **Angular measure** on $\mathbf{S}_{d-1}$: $\Phi(B) = \mu\{tB, t \geq 1\}$

Model for excesses above large radial threshold :

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

- Φ rules the joint distribution (if margins are known).

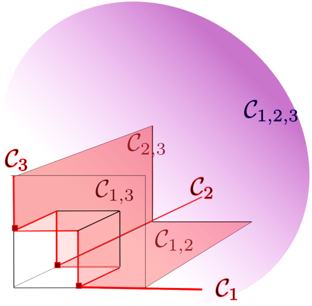
- Φ rules the joint distribution of extremes



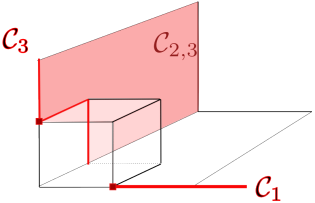
- Asymptotic dependence: (X_1, X_2) may be large together.

  vs

- Asymptotic independence: only X_1 *or* X_2 may be large.

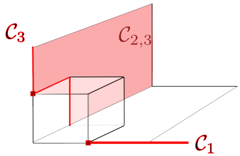


Full pattern :
anything may happen



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

- Subcones: $\emptyset \neq \alpha \subset \{1, \dots, d\}$,
  $\mathcal{C}_\alpha = \{\|x\| \geq 1, \;\; x_i > 0 \;(i \in \alpha), \;\;\; x_j = 0 \;(j \notin \alpha)\}$



- $\{\mathcal{C}_\alpha, \alpha \subset \{1,\dots,d\}$: partition of $\{x : \|x\| \geq 1\}$
- $\{\Omega_\alpha, \alpha \subset \{1,\dots,d\}$: corresponding partition of $\mathbf{S}_{d-1}$
  $\Phi_\alpha = \Phi_{|\Omega_\alpha} \leftrightarrow \mu_{|\mathcal{C}_\alpha}$
- Decomposing the angular measure : $\Phi_\alpha = \Phi_{|\Omega_\alpha}$

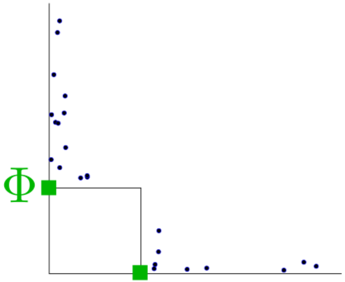
$$\Phi = \sum_{\alpha \subset \{1,\dots,d\}} \Phi_\alpha$$

- assumption: $\frac{d\phi_\alpha}{dx_\alpha} = O(1)$.

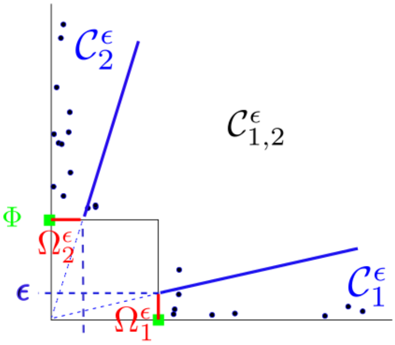
Identifying non empty edges

Issue : real data = non-asymptotic : $X_j > 0$

$\Rightarrow$ **Cannot just count data on each edge :**
Only the largest-dimensional sphere has empirical mass!



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



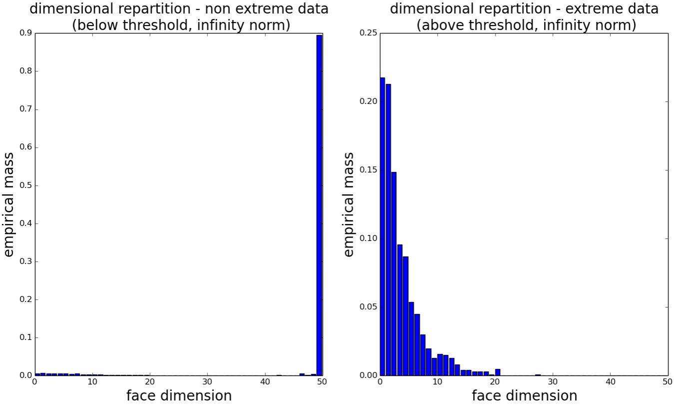
$$\Omega_\alpha \to \Omega_\alpha^\epsilon = \{x \in \mathbf{S}_{d-1} : x_i > \epsilon \,(i \in \alpha),\; x_j \leq \epsilon \,(j \notin \alpha)\}.$$

$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\epsilon = \{t\,\Omega_\alpha^\epsilon,\, t \geq 1\}$$

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

(Shell Research, and thanks J. Wadsworth)

data=50 wave direction from buoys in North sea.

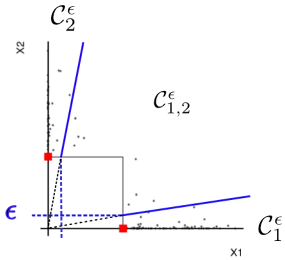


	Non-extreme data	Extreme Data
nb of faces with positive mass	2761	782
nb of faces with positive mass after thresholding	21	76
nb of faces with positive mass after 2[nd] thresholding	1	26

Goal:

Construct an empirical estimator of $\Phi_\alpha(\Omega_\alpha)$, based on the ϵ-cones.

Idea: Count the standardized points in the $\mathcal{C}^\epsilon_\alpha$.



Empirical estimator of $\Phi(\Omega_\alpha)$

data : $\mathbf{Y}_i, i = 1, \dots, n$, $\mathbf{Y}_i = (Y_{i,1}, \dots, Y_{i,d})$.

- standardization: $\hat{X}_i^j = \frac{1}{1-\hat{F}_j(Y_i^j)}$ with $\hat{F}_j(Y_i^j) = \frac{rank(Y_i^j)-1}{n}$
- $\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(\hat{X}_i) \quad \rightarrow \quad \mu_n(.) = \frac{n}{k}\mathbb{P}_n(\frac{n}{k}.)$
- $k = O(n^{2/3})$

- remind: $\phi(\Omega_\alpha^\epsilon) = \mu(\mathcal{C}_\alpha^\epsilon)$, $\alpha \subset \{1, \dots, d\}$: indices of 'large' components, while other components 'not large'.

Estimator: empirical measure of the cone

$$\Phi_n^\epsilon(\Omega_\alpha) := \mu_n(\mathcal{C}_\alpha^\epsilon) = \frac{n}{k}\mathbb{P}_n(\frac{n}{k}\mathcal{C}_\alpha^\epsilon)$$

Tool : stable tail dependence function, $l(x) = \mu[0, 1/x]^c$

(In fact, stdf generalized : $l_\alpha(x, z) = \mu[R(x, z, \alpha)]^c$ with $R(x, z, \alpha)$ VC-class.)

Regular variation of standardized marginal
$\iff \sup_{0 \preceq x \preceq T} |l(x) - t\tilde{F}(x/t)| \to 0$, where
$\tilde{F}(x) = \bar{F}(\bar{F}_j^{\leftarrow}(x_j)_{j=1..d})$

Theorem

If the density is bounded on each subface :
There is a constant C, s.t. for any n, d, k, with probability $\geq 1 - \delta$,

$$|\Phi_n^\epsilon(\Omega_\alpha) - \Phi(\Omega_\alpha)| \leq C\left(d\sqrt{\frac{1}{k}\log\frac{d}{\delta}} + 3\sqrt{\epsilon}\right) + bias_{n,k}(l, \tilde{F})$$

Comments:

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

$$d = 2.$$

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

Theorem's proof

1. **Maximal deviation on VC-class:**

$$\sup_{x \succeq \epsilon} |\mu_n - \mu|([0,x]^c \cap [0,1]^c) \leq Cd\sqrt{\frac{2}{k}\log\frac{d}{\delta}} + \text{bias}_{\tilde{F} \leftrightarrow l, x \leq 1/\epsilon}$$

Tools: Vapnik-Chervonenkis inequality adapted to small probability sets: bounds in $\sqrt{p}\sqrt{\frac{1}{n}\log\frac{1}{\delta}}$

On the VC class $[0, \frac{k}{n}x]^c$, $x \preceq 1/\epsilon$

Theorem's proof

1. **Maximal deviation on VC-class:**
2. **Decompose error:**

$$|\mu_n(\mathcal{C}_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)| \leq \underbrace{|\mu_n - \mu|(\mathcal{C}_\alpha^\epsilon)}_{A} + \underbrace{|\mu(\mathcal{C}_\alpha^\epsilon) - \mu(\mathcal{C}_\alpha)|}_{B}$$

- B : density on $\mathcal{C}_\alpha^\epsilon \times$ *Lebesgue* : small
- A : approximate $\mathcal{C}_\alpha^\epsilon$ by the VC-class.

Algorithm

1. *Let* $k = n^{2/3}$, $\epsilon = \frac{1}{k}$
2. *Class each observation in its corresponding cone* $\mathcal{C}_\alpha^\epsilon$
   $\rightarrow$ *yields: (small number of) set of cones with non-zero mass*
3. *Threshold to eliminate cones with too small mass*
   $\rightarrow$ *yields: (sparse) representation of the dependence structure* $\Phi_n^\epsilon := \sum_\alpha \mu_n(\mathcal{C}_\alpha^\epsilon) \mathbb{1}_{\{\Omega_\alpha\}}$

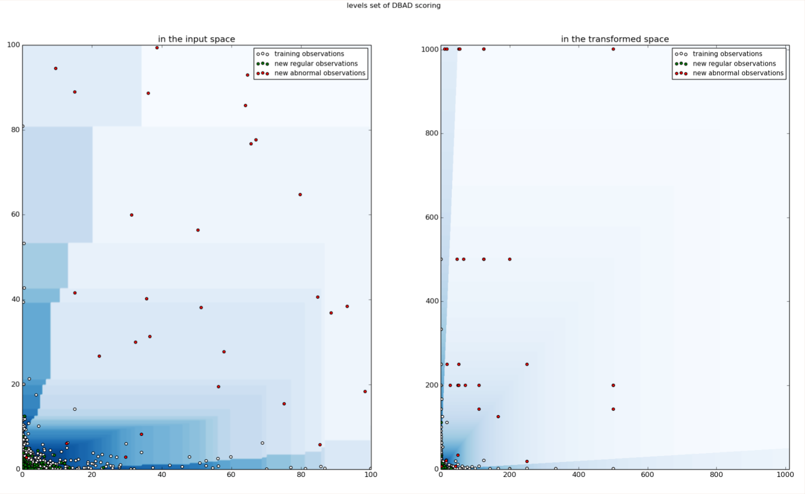
- **AD context**: Recall that after standardization of marginals:
  $\mathbb{P}[R > r, \mathbf{W} \in B] \simeq \frac{1}{r}\Phi(B)$
  $\rightarrow$ scoring function = $\Phi_n^\epsilon \times 1/r$

- **Dimension Reduction** context: Does the small edges ($|\alpha|$ small) recover 99% of the empirical mass?

Application to Anomaly Detection:

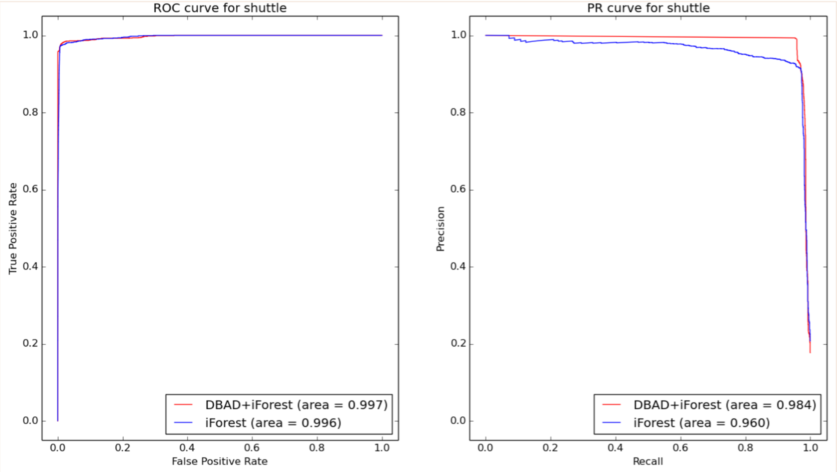
Scoring extremes according to their degree of abnormality

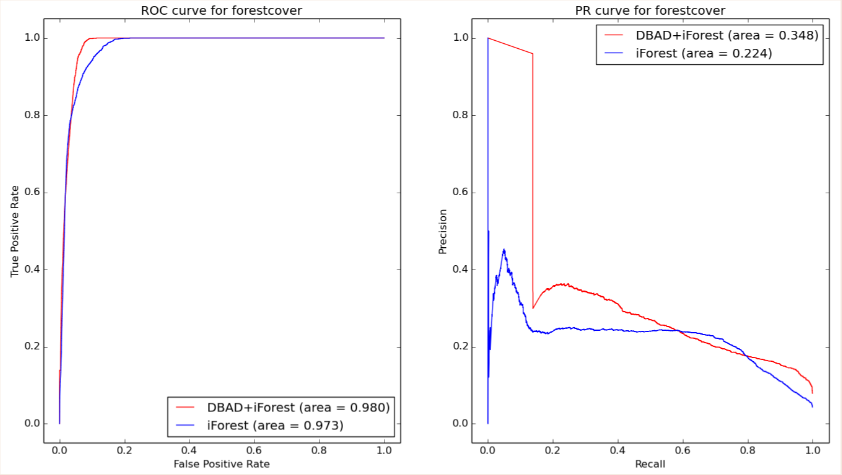
Angular part of the score = $\Phi_n^\epsilon(\Omega_\alpha)$ if $\hat{X}_i \in \Omega_\alpha^\epsilon$.

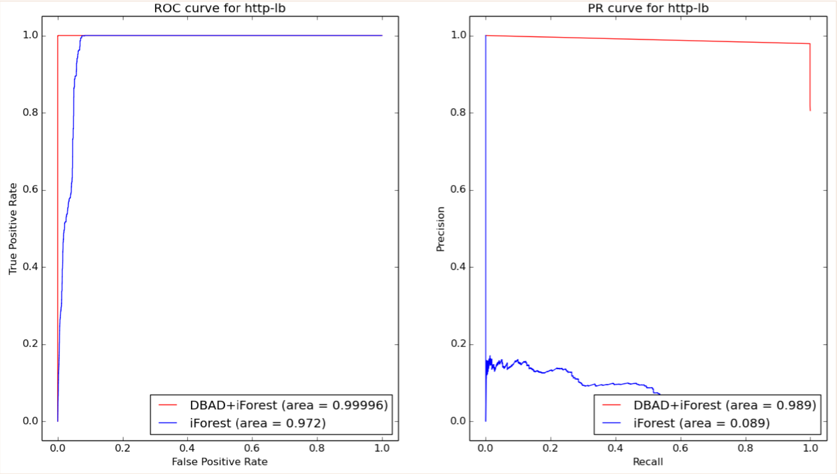


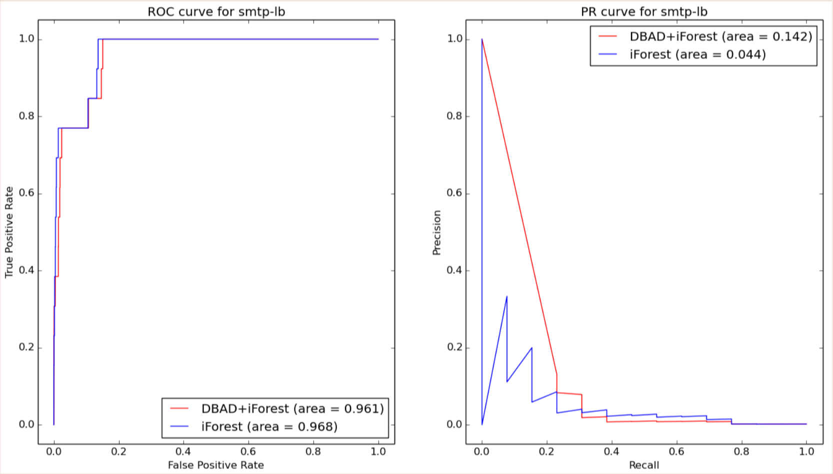
5 Experiments

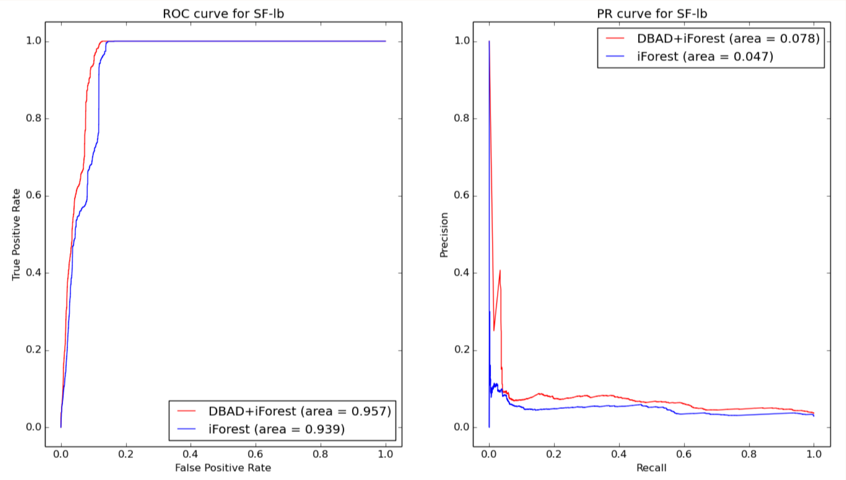
	IForest only		Iforest + DBAD	
	AUC-ROC	PR-ROC	AUC-ROC	PR-ROC
shuttle	0.996	0.96	0.997	0.984
forestcover	0.973	0.224	0.98	0.348
http	0.972	0.089	0.999	0.989
smtp	0.968	0.044	0.961	0.142
SF	0.939	0.047	0.957	0.078
SA	0.997	0.306	0.999	0.94

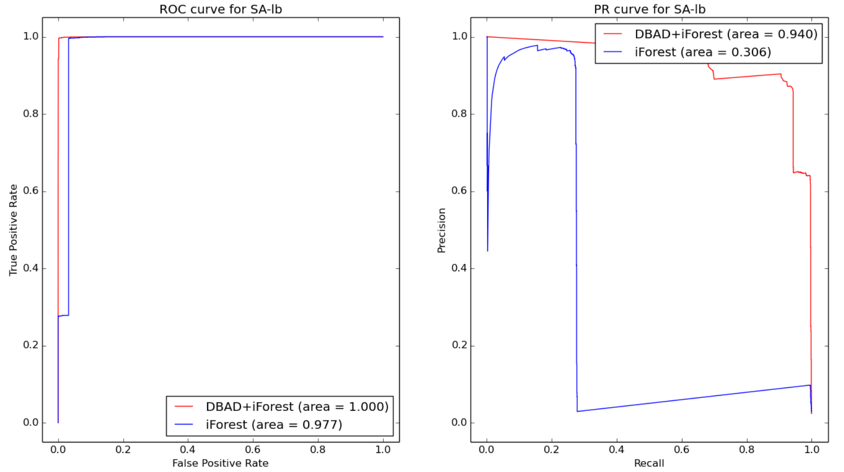

ROC curve for shuttle
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
DBAD+iForest (area = 0.997)
iForest (area = 0.996)
PR curve for shuttle
Precision
Recall
DBAD+iForest (area = 0.984)
iForest (area = 0.960)


ROC curve for forestcover
True Positive Rate
False Positive Rate
DBAD+iForest (area = 0.980)
iForest (area = 0.973)
PR curve for forestcover
Precision
Recall
DBAD+iForest (area = 0.348)
iForest (area = 0.224)


ROC curve for http-lb
True Positive Rate
False Positive Rate
DBAD+iForest (area = 0.99996)
iForest (area = 0.972)
PR curve for http-lb
Precision
Recall
DBAD+iForest (area = 0.989)
iForest (area = 0.089)


ROC curve for smtp-lb
True Positive Rate
False Positive Rate
DBAD+iForest (area = 0.961)
iForest (area = 0.968)
PR curve for smtp-lb
Precision
Recall
DBAD+iForest (area = 0.142)
iForest (area = 0.044)


ROC curve for SF-lb
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
DBAD+iForest (area = 0.957)
iForest (area = 0.939)
PR curve for SF-lb
Precision
Recall
DBAD+iForest (area = 0.078)
iForest (area = 0.047)


ROC curve for SA-lb
True Positive Rate
False Positive Rate
DBAD+iForest (area = 1.000)
iForest (area = 0.977)
PR curve for SA-lb
Precision
Recall
DBAD+iForest (area = 0.940)
iForest (area = 0.306)

Conclusion:

- A sparse, potentially low-dimensionnal representation of the dependence structure of extremes.
- Relatively coarse (angular part constant on each ϵ face)

- Futur work: refine this representation (dirichlet on each face?) while preserving the robustness with respect to outliers (avoid fitting anomalies).

- 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