Machine Learning and Extremes for Anomaly Detection

Nicolas Goix

LTCI, CNRS, Telecom ParisTech, Université Paris-Saclay, France

GT valeurs extremes, Jussieu, November 15th 2016

Anomaly Detection (AD)

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

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

Machine Learning context

Different kind of Anomaly Detection

- **Supervised** AD (not dealt with)
  Labels available for both normal data and anomalies
  (similar to rare class mining)

- **Novelty Detection** (our theoretical framework)
  The algorithm learns on normal data only

- **Outlier Detection** (extended application framework)
  Training set (unlabeled) = normal + abnormal data
  (assumption: anomalies are very rare)

Some litterature in Anomaly Detection

- **Statistical AD techniques**
  fit a statistical model for normal behavior (ex: gaussian, gaussian mixture)

- **K-nearest neighbors**
  - ex: Local Outlier Factor (LOF) [Breunig *et al.* 2000]

- **Support estimation**
  - One-Class-SVM [Schölkopf *et al.* 2000, Vert and Vert 2006]
  - Minimum Volume set estimate [Einmahl and Mason 92, Polonik 97, Scott and Nowak 2006]

- **High-dimensional techniques**
  - Dimensionality reduction [Aggarwal and Yu 2001, Shyu *et al.* 2003]
  - One-class Random Forests [Shi and Horvath 2012, Désir *et al.* 2012]
  - Isolation Forest [Liu *et al.* 2008]

Outline

An AD algorithm returns a scoring function $s : \mathbb{R}^d \to \mathbb{R}$.
It represents the 'degree of abnormality' of an observation $x \in \mathbb{R}^d$

- Part I: Performance criterion on s.
  (model selection)

- Part II: Building good s on extreme regions.
  (model design)

Part I: performance criterion

Definition

Learning a scoring function

Evaluating a scoring function

Part II: Learning accurate scoring functions on extreme regions

Multivariate EVT & Representation of Extremes

Estimation

Experiments

(unsupervised) performance criterion

Such a criterion allows:

- 1- To build good s by optimizing this criterion.
- 2- To evaluate any AD algorithm without using any labels.

Practical motivations:

Most of the time, data come without any label.
$\rightarrow$ no ROC or PR curves!

Idea:

How good is an anomaly detection algorithm?

$\downarrow$

How good is it estimating the level sets?

Context

Novelty Detection ('One-Class Classification', 'semi-supervised AD')

- **Data: inliers**.
  i.i.d. observations in $\mathbb{R}^d$ from the normal behavior, density f.

- **Output to evaluate: scoring function $s : \mathbb{R}^d \to \mathbb{R}$**
  - s defines a pre-order on $\mathbb{R}^d$ = 'degree of abnormality'.
  - s level sets are estimates of f level sets.
  - s can be interpreted as a box which contains an infinite number of level sets estimates (at different levels).

Remark. Ideal scoring functions: $s = f$ or $s = 2f + 3$ or $s = T \circ f$ any increasing transform of f.

Problem reformulation

We want a criterion $\mathcal{C}(s)$ which measures *how well the level sets of f are approximated by those of s*.

- **Fact:** For any strictly increasing transform T, level sets of $T \circ f$ are exactly those of f.
  $\Rightarrow$ Criterion $\mathcal{C}(s) = \|s - f\|$ is not relevant! ($\mathcal{C}(2f) > 0$)

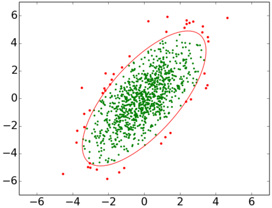
- **We are looking for a criterion s.t:**
  - $\mathcal{C}^{\Phi}(s) = \|\Phi(s) - \Phi(f)\|$ with Φ s.t. $\Phi(T \circ s) = \Phi(s)$.
  - $\{$level sets of optimal $s^*\} = \{$level sets of $f\}$.
  - $\mathcal{C}^{\Phi}(s)$ = 'distance' between level sets of s and those of f.

Question: How to choose $\Phi(s)$?

Existing criterion: Mass-Volume curve

Minimum volume set [Polonik, 1997]

$$\Gamma_\alpha^* = \underset{\Gamma \text{ borelian}}{\arg\min} \{\mathrm{Leb}(\Gamma), \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha\}$$



Under regularity assumptions, minimum-volume sets are **density level sets**:

$$\exists t_\alpha > 0, \quad \Gamma_\alpha^* = \{f > t_\alpha\} =: \Omega_{t_\alpha}$$

Mass Volume curve of a scoring function s [Clémençon and Jakubowicz, 2013]:

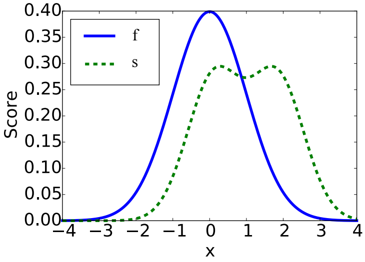
$$\begin{aligned} MV_s(\alpha) &:= \inf_{\Omega \text{ level-set of } s} \left\{ \mathrm{Leb}(\Omega) \quad s.t. \quad \mathbb{P}(\mathbf{X} \in \Omega) \geq \alpha \right\} \\ MV^*(\alpha) &:= MV_f(\alpha) \end{aligned}$$

Existing criterion: Mass-Volume curve

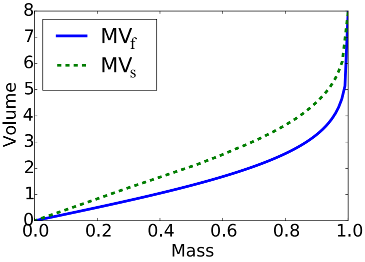
Properties:

- For any scoring function s, $MV^*(\alpha) \leq MV_s(\alpha)$
- For any increasing transform T, $MV^*(\alpha) = MV_{T \circ f}(\alpha)$
- $MV^*(\alpha) = \text{Leb}(\Gamma^*_\alpha) := \min_{\Omega \text{ borelian}} \{\text{Leb}(\Omega) \;\; s.t. \;\; \mathbb{P}(\mathbf{X} \in \Omega) \geq \alpha\}$

$$(MV_s(\alpha) \; := \; \inf_{\Omega \text{ level-set of } s} \;\; \{\text{Leb}(\Omega) \quad s.t. \quad \mathbb{P}(\mathbf{X} \in \Omega) \geq \alpha\})$$



(a) Scoring functions



(b) Mass Volume curves

Drawbacks and alternative criterion: Excess-Mass curve

Drawbacks of MV:

- When optimized *w.r.t.* different levels α, produces not necessarily nested empirical level sets.
- $\rightarrow$ low convergence rates – of order $O(n^{-1/4})$.
- MV diverges in 1 in case of unbounded support.

Solution: Excess-Mass curve

- **Definitions:**

$$EM_s(t) \;=\; \sup_{\Omega \text{ level-set of } s} \;\; \{\mathbb{P}(\mathbf{X} \in \Omega) \;-\; t\text{Leb}(\Omega)\}$$

- **Optimal curves:**

$$EM^*(t) \;:=\; EM_f(t) = EM_{T\circ f}(t) = \max_{\Omega \text{ borelian}} \;\{\mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega)\}$$

$$EM^*(t) \geq EM_s(t) \qquad\qquad \text{for all scoring function } \; s$$

Property: Previous drawbacks are fixed with EM.

MV and EM criteria

- **Interpretation:** $(EM_f - EM_s)(t) \simeq \ \inf_{u>0} \ \mathrm{Leb}(\{s > u\} \, \Delta \, \{f > t\})$

- How well t-level sets of f are approximated by level sets of s, $t \in I$?

$$\updownarrow$$

how small is $EM_f - EM_s$ on I ?

$$\updownarrow$$

how large is EM_s on I ? $\quad \rightarrow \mathcal{C}^{EM}(s) = \|EM_s\|_{1,I}$

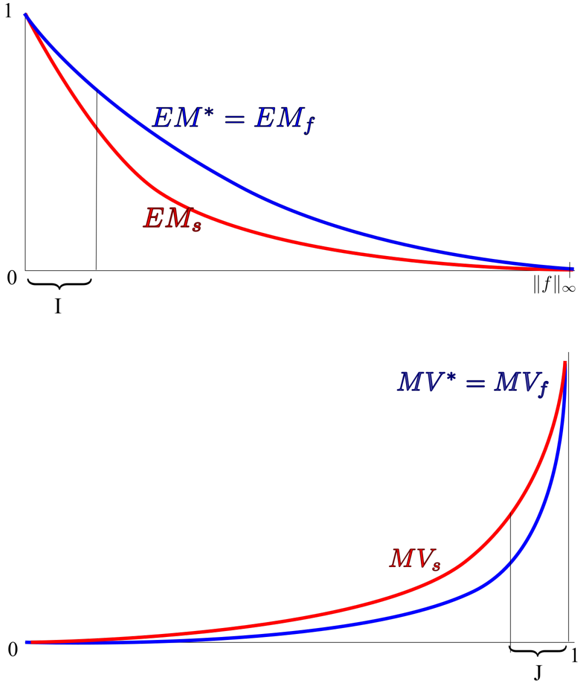
- How well α minimum-volume sets of f are approximated by level sets of s, $\alpha \in J$?

$$\updownarrow$$

how small is $MV_s - MV_f$ on J ?

$$\updownarrow$$

how small is MV_s on J ? $\quad \rightarrow \mathcal{C}^{MV}(s) = \|MV_s\|_{1,J}$


1
$EM^* = EM_f$
EM_s
0
$\|f\|_\infty$
I
$MV^* = MV_f$
MV_s
0
1
J

Part I: performance criterion

Definition

Learning a scoring function

Evaluating a scoring function

Part II: Learning accurate scoring functions on extreme regions

Multivariate EVT & Representation of Extremes

Estimation

Experiments

Learning a scoring function with M-estimation

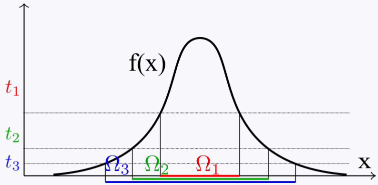
We are looking for nearly optimal scoring functions of the form $s = \sum_{j=1}^{N} a_j \mathbb{1}_{x \in \Omega_j}$, with $a_j \geq 0$, $\Omega_j \in \mathcal{G}$.

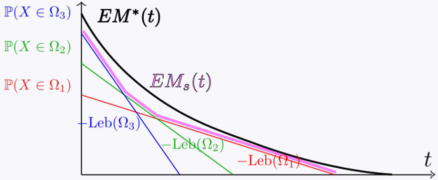
Procedure: Fixe $t_0 > 0$
For $k = 1, \ldots, N$,

$$t_{k+1} = \frac{t_k}{(1 + \frac{1}{\sqrt{n}})}$$

$$\widehat{\Omega}_{t_{k+1}} = \underset{\Omega \supset \widehat{\Omega}_{t_k}}{\arg\max} \quad \mathbb{P}_n(X \in \Omega) - t_{k+1}\mathrm{Leb}(\Omega)$$

$$s_N(x) := \sum_{j=1}^{N} (t_j - t_{j+1}) \mathbb{1}_{x \in \Omega_{t_j}}$$

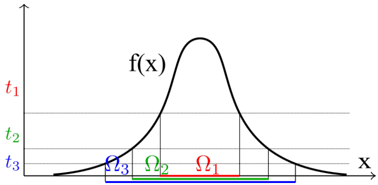


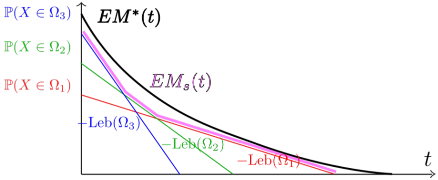


Learning a scoring function with M-estimation

Rates: (if density bounded and without flat parts, $\mathcal{G}$ VC-class)
With probability at least $1-\delta$,

$$\sup_{t \in]0,t_1]} |EM^*(t) - EM_{s_N}(t)| \leq \left[A + \sqrt{2\log(1/\delta)}\right] \frac{1}{\sqrt{n}} + bias(\mathcal{G}) + o_N(1).$$





Part I: performance criterion

- Definition
- Learning a scoring function
- Evaluating a scoring function

Part II: Learning accurate scoring functions on extreme regions

- Multivariate EVT & Representation of Extremes
- Estimation
- Experiments

Evaluation of scoring functions

Inputs: scoring function s

- **Estimation:**

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \quad \mathrm{Leb}(s \geq u) \quad \textit{s.t.} \quad \mathbb{P}_n(s \geq u) \geq \alpha$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \quad \mathbb{P}_n(s \geq u) \; - \; t\mathrm{Leb}(s \geq u)$$

- **Empirical criteria:**

$$\widehat{\mathcal{C}}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I)} \qquad\qquad I = [0, \widehat{EM}^{-1}(0.9)],$$

$$\widehat{\mathcal{C}}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(J)} \qquad\qquad J = [0.9, 1],$$

- **Issue:** The volume $\mathrm{Leb}(s \geq u)$ has to be estimated (Monte-Carlo). Challenging in high dimensions.

Evaluation: Heuristic solution

Feature sub-sampling (random projection) and averaging

Inputs: AD algorithm $\mathcal{A}$, data set X size $n \times d$, feature sub-sampling size d', number of draws m.

for $k = 1, \ldots, m$ **do**
-randomly select a sub-group F_k of d' features
-compute the associated scoring function $s_k = \mathcal{A}\big((x_i^j)_{1 \leq i \leq n,\ j \in F_k}\big)$
-compute $\widehat{\mathcal{C}}_k^{EM} = \|\widehat{EM}_{s_k}\|_{L^1(I)}$ or $\widehat{\mathcal{C}}_k^{MV} = \|\widehat{MV}_{s_k}\|_{L^1(J)}$
end for

Return performance criteria:

$$\widehat{\mathcal{C}}^{EM}_{high\_dim}(\mathcal{A}) = \frac{1}{m}\sum_{k=1}^{m} \widehat{\mathcal{C}}_k^{EM} \quad \text{or} \quad \widehat{\mathcal{C}}^{MV}_{high\_dim}(\mathcal{A}) = \frac{1}{m}\sum_{k=1}^{m} \widehat{\mathcal{C}}_k^{MV} \, .$$

Part I: performance criterion

Definition

Learning a scoring function

Evaluating a scoring function

Part II: Learning accurate scoring functions on extreme regions

Multivariate EVT & Representation of Extremes

Estimation

Experiments

Why dealing with extremes?

General ideas:

- Extreme observations play a special role when dealing with outlying data.
- But no anomaly detection algorithm has **specific treatment for such multivariate extreme observations**. Univariate EVT: [Roberts 99, Lee and Roberts 2008, Clifton *et al.* 2011]
- Our goal:
  - Define a notion of sparsity for extremes observations.
  - Provide a method which can improve performance of standard AD algorithms by combining them with a **multivariate extreme analysis** of the **dependence structure**, using this notion of sparsity.

Purpose

$$\mathbf{X} = (X_1, \dots, X_d)$$

Goal: find the groups of features which can be large together

ex: $\{X_1, X_2\}, \; \{X_1, X_3, X_4\}, \; \{X_5\}$

Namely: characterize the extreme dependence structure

$\rightarrow$ Anomalies = points which violate this structure

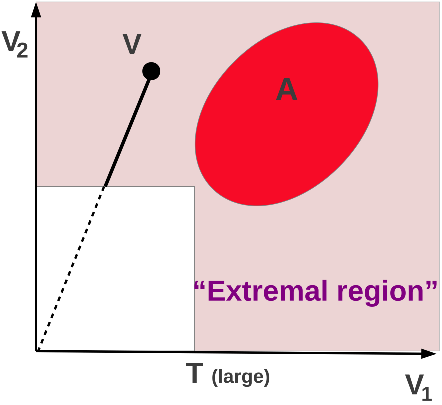
Theoretical framework

- **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 of Extreme Value Theory

Describe $\mathbf{V}$'s distribution, when $\mathbf{V}$ exceeds some large threshold.

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



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):

$$\text{If} \quad 0 \notin \overline{A}, \qquad t\mathbb{P}[\mathbf{V} \in t\,A] \xrightarrow[t\to\infty]{} \mu(A).$$

μ: 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 of multivariate EVT

$\mathbb{P}[\mathbf{V} \in A] \quad \simeq \quad \mu(A),$ if A extreme region.

Model for excesses

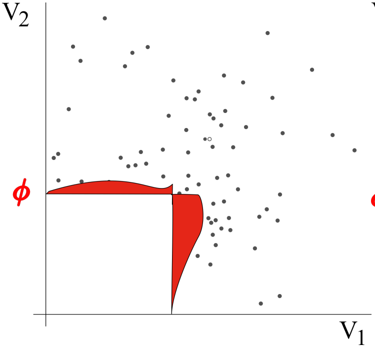
For a large $r > 0$ and a region B on the unit sphere:

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

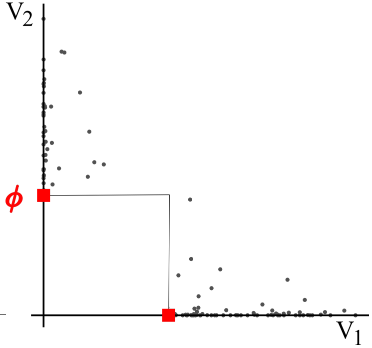
$\Rightarrow \Phi$ (or μ) **rules the joint distribution of extremes** (if margins are known).

Angular distribution

- Φ rules the joint distribution of extremes

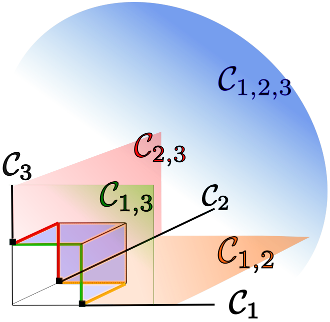


Asymptotic dependence:
(V_1, V_2) may be large together.



Asymptotic independence:
Only V_1 *or* V_2 may be large.

General Case



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

Representation of extreme data

- Natural decomposition of the angular measure :

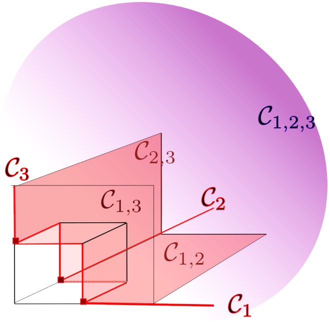
$$\Phi = \sum_{\alpha \subset \{1,\dots,d\}} \Phi_\alpha \qquad \text{with} \quad \Phi_\alpha = \Phi_{|\Omega_\alpha} \leftrightarrow \mu_{|\mathcal{C}_\alpha}$$

- $\Rightarrow$ yields a representation

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

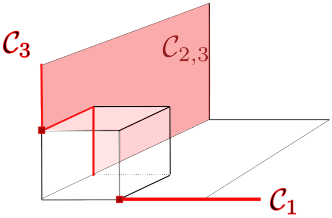
- Assumption: $\frac{\mathrm{d}\mu_{|\mathcal{C}_\alpha}}{\mathrm{d}v_\alpha} = O(1)$.

Sparse Representation ?



Full pattern :

anything may happen



Sparse pattern

(V_1 not large if V_2 or V_3 large)

Part I: performance criterion
- Definition
- Learning a scoring function
- Evaluating a scoring function

Part II: Learning accurate scoring functions on extreme regions
- Multivariate EVT & Representation of Extremes
- Estimation
- Experiments

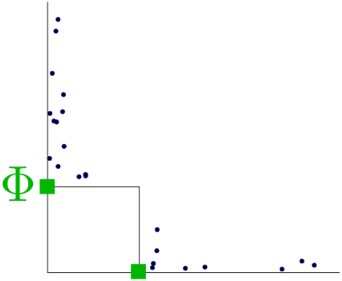
Problem: $\mathcal{M}$ is an **asymptotic** representation

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

is the restriction of an asymptotic measure

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

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

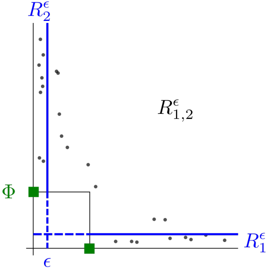


$\Rightarrow$ Cannot just do, for large t:

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

Solution

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



$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\epsilon = \{\|v\| \geq 1, \quad v_j > \epsilon \, (j \in \alpha), \; v_j \leq \epsilon \, (j \notin \alpha)\}.$$

$$\Omega_\alpha \to \Omega_\alpha^\epsilon = \mathcal{C}_\alpha^\epsilon \cap \mathbf{S}_{d-1}$$

New partition of $\mathbf{S}_{d-1}$

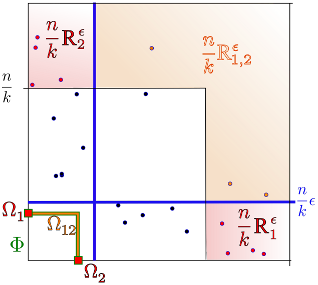
Resulting estimation procedure

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

$\Rightarrow$ 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

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

Statistical guaranties: 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

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

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

Statistical guaranties: 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

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

$k \propto$ number of data considered as extreme (data used for estimation)

Statistical guaranties

Theorem

There is an absolute constant $C > 0$ such that for any $n > 0,\ k > 0,\ 0 < \epsilon < 1,\ \delta > 0$ such that $0 < \delta < e^{-k}$, with probability at least $1 - \delta$,

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

Comments:

- $M = \mathrm{sup}(\text{density on sub-cones})$
- 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.

Theorem's proof

Decompose error:

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

- term A : Bounded with VC inequality adapted to small probability regions.
- term B : Leb$(\mathcal{C}_\alpha^\epsilon \setminus \mathcal{C}_\alpha)$ is small when ϵ is small.

Corresponding Algorithm

DAMEX in $O(dn \log n)$

Input: parameters $\epsilon > 0$, $k = k(n)$

for $i = 1, \ldots, n$ **do**
  # Standardize *via* marginal rank-transformation:
  $\hat{V}_i := \left(1/(1 - \hat{F}_j(X_i^j))\right)_{j=1,\ldots,d}$.
  if $\hat{V}_i > \frac{n}{k}$ **then**
    # Assign to each $\hat{V}_i$ the cone $\frac{n}{k}\mathcal{C}_\alpha^\epsilon$ it belongs to:
    $\alpha = \alpha(V_i)$
    c_α ++
  end if
end for
$\Phi_n^{\alpha,\epsilon} := \frac{n}{k} c_\alpha$

Output: (sparse) representation of the dependence structure:
$\Phi_n^{\alpha,\epsilon} = \widehat{\Phi}(\Omega_\alpha) = \frac{n}{k}\mathbb{P}_n(\hat{V} \in \frac{n}{k}\mathcal{C}_\alpha^\epsilon)$, estimate of the α-mass of Φ for every α.

Application to Anomaly Detection

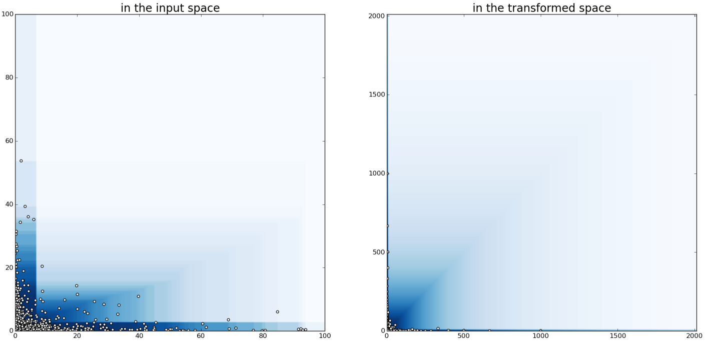
Recall that after standardization of marginals:
$\mathbb{P}[R > r, \mathbf{W} \in B] \simeq \frac{1}{r}\,\Phi(B)$

$\rightarrow$ scoring function = $\Phi_n^{\epsilon} \times 1/r$:

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

where $\hat{T} : \mathbf{X} \mapsto \mathbf{V}$ $\quad (\hat{V}_j = \frac{1}{1-\hat{F}_j(X_j)})$

levels set of DAMEX scoring function



Experiments

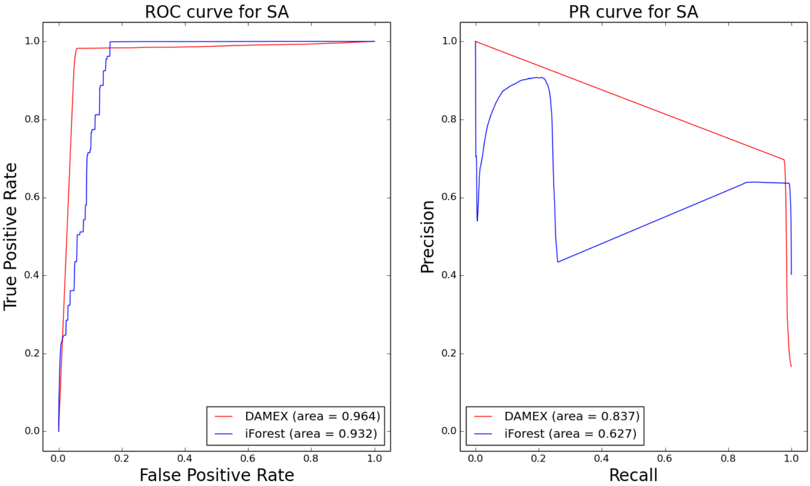


Figure: ROC and PR curve on SA dataset

Thank you!

Some references:

- Aggarwal and Yu 2001, Outlier detection for high dimensional data.
- Chandola, Banerjee, and Kumar 2009, Anomaly detection: A survey.
- Clifton, Hugueny and Tarassenko 2011, Novelty detection with multivariate extreme value statistics.
- Désir, Bernard, Petitjean and Heutte 2012, A New Random Forest Method for One-Class Classification.
- Einmahl and Mason 1992, Generalized quantile processes.
- Einmahl, Krajina and Segers 2012, An m-estimator for tail dependence in arbitrary dimensions.
- Goix, Sabourin and Clémençon 2015, Learning the dependence structure of rare events: a non-asymptotic study.
- Goix, Sabourin and Clémençon 2015, Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking.
- Goix, Sabourin and Clémençon 2015, On Anomaly Ranking and Excess-Mass Curves.
- Goix 2016, How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?
- Goix, Brault, Drougard and Chiapino 2016, One Class Splitting Criteria for Random Forests with Application to Anomaly Detection.

Some references:

- de Haan and Ferreira, Extreme value theory, 2006.
- Lee and Roberts 2008, On-line novelty detection using the Kalman filter and extreme value theory.
- Liu, Ting, Zhou 2008, Isolation forest.
- Polonik 1997, Minimum volume sets and generalized quantile processes.
- Qi 1997, Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics.
- Resnick 1987, Extreme Values, Regular Variation, Point Processes.
- Roberts 1999, Novelty detection using extreme value statistics.
- Schölkopf, Platt, Shawe-Taylor, Smola, and Williamson 2001, Estimating the support of a high-dimensional distribution.
- Scott and Nowak 2006, Learning minimum volume sets.
- J. Segers 2012, Max-stable models for multivariate extremes.
- Shi and Horvath 2012, Unsupervised learning with random forest predictors.
- Shyu, Chen, Sarinnapakorn and Chang 2003, A novel anomaly detection scheme based on principal component classifier
- Vert and Vert 2006, Consistency and Convergence Rates of One-Class SVMs and Related Algorithms

Part I: performance criterion
Definition
Learning a scoring function
Evaluating a scoring function

Part II: Learning accurate scoring functions on extreme regions
Multivariate EVT & Representation of Extremes
Estimation
Experiments

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

Table: Datasets characteristics

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

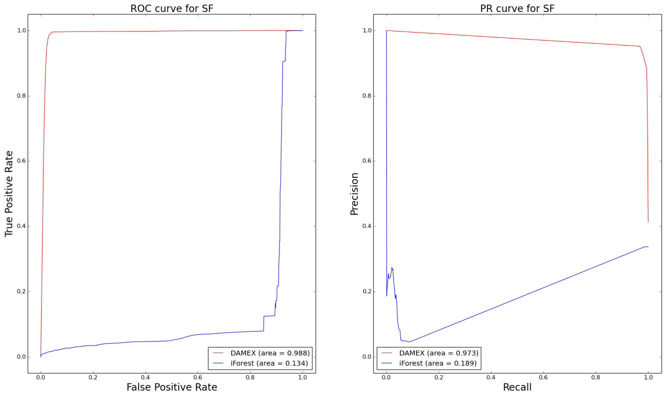


Figure: ROC and PR curve on SF dataset

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

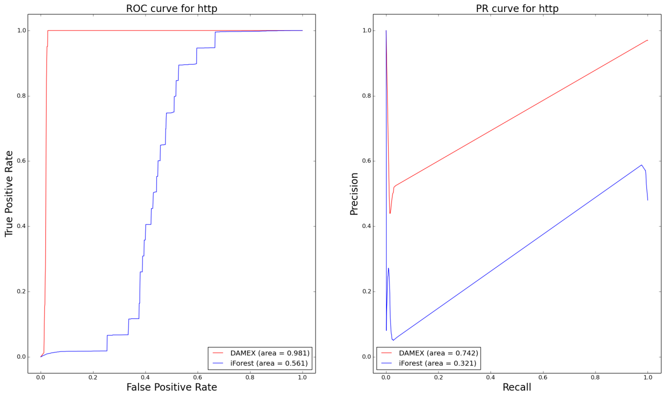


Figure: ROC and PR curve on http dataset

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

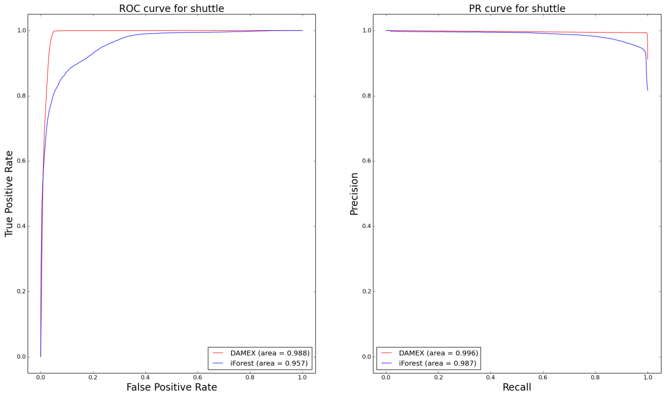


Figure: ROC and PR curve on shuttle dataset

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

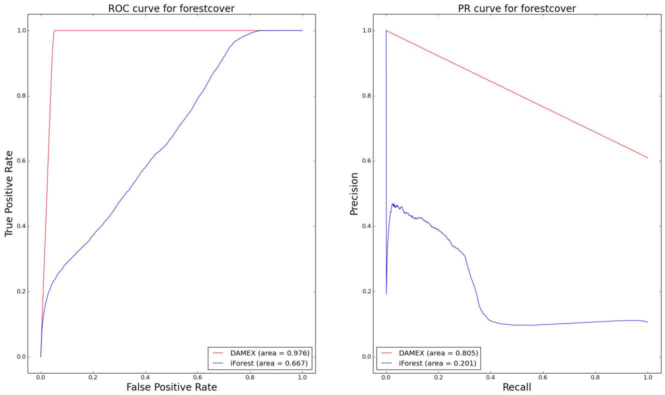


Figure: ROC and PR curve on forestcover dataset

Benchmarks for EM/MV

Does performance in term of EM/MV correspond to performance in term of ROC/PR?

- **Experiments:** 12 datasets, 3 AD algorithms (LOF, OCSVM, iForest) $\rightarrow$ 36 possible pairwise comparisons:

$$\left\{ \Big(A_1 \text{ on } \mathcal{D},\ A_2 \text{ on } \mathcal{D}\Big),\ \ A_1, A_2 \in \{\text{iForest, LOF, OCSVM}\}, \\ \mathcal{D} \in \{\text{adult, http}, \ldots, \text{spambase}\} \right\}.$$

- **Results:** If we only consider the pairs *s.t. ROC and PR agree on which algorithm is the best*, we are able (with EM and MV scores) to recover it in 80% of the cases.

Table: Original Datasets characteristics

	nb of samples	nb of features	anomaly class	
adult	48842	6	class '$> 50K$'	(23.9%)
http	567498	3	attack	(0.39%)
pima	768	8	pos (class 1)	(34.9%)
smtp	95156	3	attack	(0.03%)
wilt	4839	5	class 'w' (diseased trees)	(5.39%)
annthyroid	7200	6	classes $\neq 3$	(7.42%)
arrhythmia	452	164	classes $\neq 1$ (features 10-14 removed)	(45.8%)
forestcover	286048	10	class 4 (vs. class 2)	(0.96%)
ionosphere	351	32	bad	(35.9%)
pendigits	10992	16	class 4	(10.4%)
shuttle	85849	9	classes $\neq 1$ (class 4 removed)	(7.17%)
spambase	4601	57	spam	(39.4%)

Table: Results for the novelty detection setting. One can see that ROC, PR, EM, MV often do agree on which algorithm is the best (in bold), which algorithm is the worse (underlined) on some fixed datasets. When they do not agree, it is often because ROC and PR themselves do not, meaning that the ranking is not clear.

Dataset	iForest				OCSVM				LOF			
	ROC	PR	EM	MV	ROC	PR	EM	MV	ROC	PR	EM	MV
adult	**0.661**	**0.277**	**1.0e-04**	**7.5e01**	0.642	0.206	2.9e-05	4.3e02	<u>0.618</u>	<u>0.187</u>	<u>1.7e-05</u>	<u>9.0e02</u>
http	0.994	0.192	1.3e-03	9.0	**0.999**	**0.970**	**6.0e-03**	**2.6**	<u>0.946</u>	<u>0.035</u>	<u>8.0e-05</u>	<u>3.9e02</u>
pima	0.727	0.182	5.0e-07	**1.2e04**	**0.760**	**0.229**	**5.2e-07**	<u>1.3e04</u>	<u>0.705</u>	<u>0.155</u>	<u>3.2e-07</u>	2.1e04
smtp	0.907	<u>0.005</u>	<u>1.8e-04</u>	<u>9.4e01</u>	<u>0.852</u>	**0.522**	**1.2e-03**	8.2	**0.922**	0.189	1.1e-03	**5.8**
wilt	0.491	0.045	4.7e-05	<u>2.1e03</u>	<u>0.325</u>	<u>0.037</u>	**5.9e-05**	**4.5e02**	**0.698**	**0.088**	<u>2.1e-05</u>	1.6e03
annthyroid	**0.913**	**0.456**	**2.0e-04**	2.6e02	<u>0.699</u>	<u>0.237</u>	<u>6.3e-05</u>	**2.2e02**	0.823	0.432	6.3e-05	<u>1.5e03</u>
arrhythmia	**0.763**	**0.487**	**1.6e-04**	**9.4e01**	0.736	0.449	1.1e-04	1.0e02	<u>0.730</u>	<u>0.413</u>	<u>8.3e-05</u>	<u>1.6e02</u>
forestcov.	<u>0.863</u>	<u>0.046</u>	<u>3.9e-05</u>	<u>2.0e02</u>	0.958	0.110	5.2e-05	1.2e02	**0.990**	**0.792**	**3.5e-04**	**3.9e01**
ionosphere	<u>0.902</u>	<u>0.529</u>	<u>9.6e-05</u>	<u>7.5e01</u>	**0.977**	**0.898**	**1.3e-04**	**5.4e01**	0.971	0.895	1.0e-04	7.0e01
pendigits	0.811	0.197	2.8e-04	2.6e01	<u>0.606</u>	<u>0.112</u>	<u>2.7e-04</u>	<u>2.7e01</u>	**0.983**	**0.829**	**4.6e-04**	**1.7e01**
shuttle	0.996	0.973	1.8e-05	5.7e03	<u>0.992</u>	<u>0.924</u>	**3.2e-05**	**2.0e01**	**0.999**	**0.994**	<u>7.9e-06</u>	<u>2.0e06</u>
spambase	**0.824**	**0.371**	**9.5e-04**	**4.5e01**	<u>0.729</u>	0.230	4.9e-04	1.1e03	0.754	<u>0.173</u>	<u>2.2e-04</u>	<u>4.1e04</u>