

# Machine Learning and Extremes for Anomaly Detection

Nicolas Goix

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

PhD Defense, Telecom Paristech, November 28th 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,...

## Different kinds 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 literature in Anomaly Detection

- ▶ **Statistical AD techniques**

[Hawkins 1980, Liu and Weng 1991, Eskin 2000, Agarwal 2006]

- ▶ **K-nearest neighbors**

[Breunig *et al.* 2000, Tang *et al.* 2002, Papadimitriou *et al.* 2002, Hautamaki *et al.* 2004]

- ▶ **Support estimation**

[Einmahl and Mason 92, Polonik 97, Schölkopf *et al.* 2000, Vert and Vert 2006, Scott and Nowak 2006]

- ▶ **High-dimensional techniques**

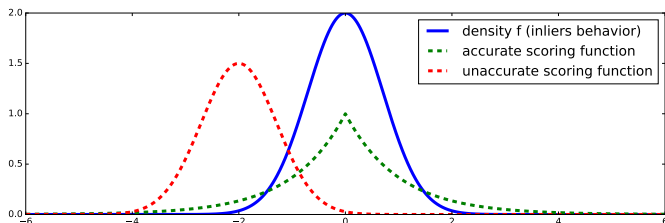
[Aggarwal and Yu 2001, Shyu *et al.* 2003, Shi and Horvath 2012, Liu *et al.* 2008]

# Background: scoring functions

AD algorithm returns a **scoring function**  $s : \mathbb{R}^d \rightarrow \mathbb{R}$

- ▶  $s$  defines a **pre-order** on  $\mathbb{R}^d$  = ‘degree of normality’.
- ▶  $s$  level sets are estimates of  $f$  level sets.
- ▶  $s$  can be interpreted as a **continuum of level sets estimates** (at different levels).

**Remark.** Ideal scoring functions:  $s = T \circ f$  any increasing transform of  $f$ .



# Outline

Part I: Performance Criterion for Scoring Functions

Part II: Learning accurate scoring functions on extreme regions

Part III: Heuristic Contributions and Perspectives

Part I: Performance Criterion for Scoring Functions

Part II: Learning accurate scoring functions on extreme regions

Part III: Heuristic Contributions and Perspectives

## 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  $s : \mathbb{R}^d \rightarrow \mathbb{R}$** 
  - $s$  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  $s$* .

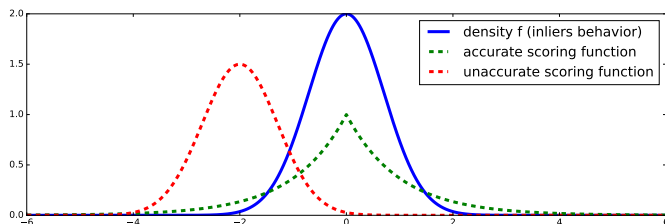
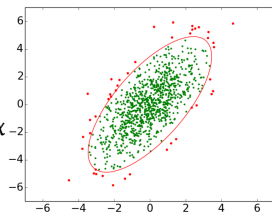
- ▶ **Fact:** For any strictly increasing transform  $T$ , level sets of  $T \circ f$  are exactly those of  $f$ .  
 $\Rightarrow$  Criterion  $\mathcal{C}(s) = \|s - f\|$  is not relevant! ( $\mathcal{C}(2f) > 0$ )
- ▶ **We are looking for a criterion s.t:**
  - $\mathcal{C}^\Phi(s) = \|\Phi(s) - \Phi(f)\|$  with  $\Phi$  s.t.  $\Phi(T \circ s) = \Phi(s)$ .
  - $\{\text{level sets of optimal } s^*\} = \{\text{level sets of } f\}$ .
  - $\mathcal{C}^\Phi(s)$  = '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]

$$\Gamma_{\alpha}^* = \arg \min_{\Gamma \text{ borelian}} \{\text{Leb}(\Gamma) \text{ s.t. } \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha\}$$



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

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

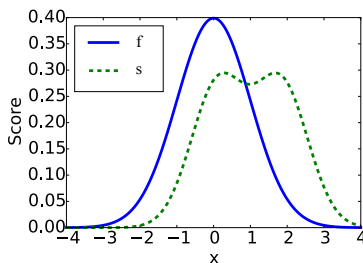
$$MV^*(\alpha) := MV_{\cdot}(\alpha)$$

# Existing criterion: Mass-Volume curve

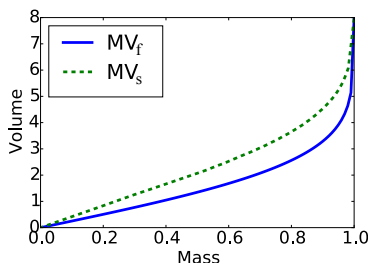
Properties:

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

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



(a) Scoring functions



(b) Mass Volume curves

# Drawbacks and alternative criterion: Excess-Mass curve

Drawbacks of MV:

- ▶ When optimized *w.r.t.* different levels  $\alpha$ , 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:

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

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

**Property: Previous drawbacks are fixed with EM.**

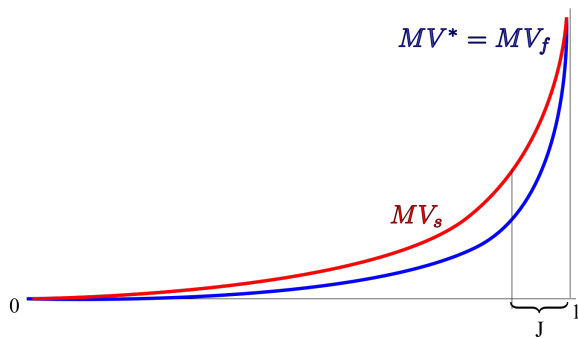
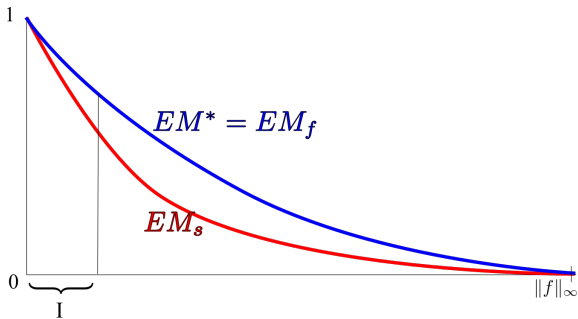
# MV and EM criteria

- **Interpretation:**  $(EM_f - EM_s)(t) \simeq \inf_{u>0} \text{Leb}(\{s > u\} \Delta \{f > t\})$
- 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  $I$ ?  $\rightarrow \mathcal{C}^{EM}(s) = \|EM_s\|_{1,I}$

- How well  $\alpha$  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  $J$ ?  $\rightarrow \mathcal{C}^{MV}(s) = \|MV_s\|_{1,J}$



# Learning a scoring function with M-estimation

We are looking for nearly optimal scoring functions of the form

$$s = \sum_{j=1}^N a_j \mathbb{1}_{x \in \Omega_j}, \text{ with } a_j \geq 0, \Omega_j \in \mathcal{G}.$$

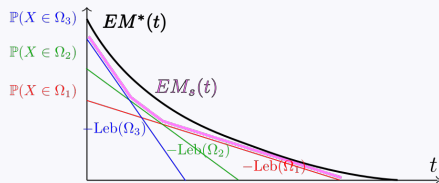
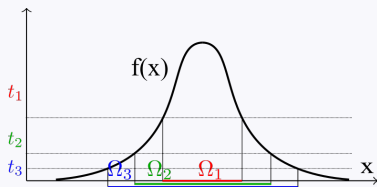
**Procedure:** Fixe  $t_0 > 0$

For  $k = 1, \dots, N$ ,

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

$$\hat{\Omega}_{t_{k+1}} = \arg \max_{\Omega \supset \hat{\Omega}_{t_k}} \mathbb{P}_n(X \in \Omega) - t_{k+1} \text{Leb}(\Omega)$$

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

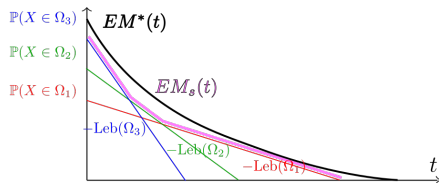
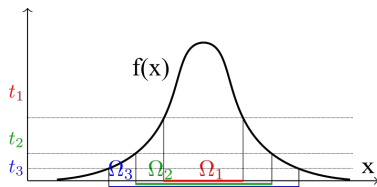


# Learning a scoring function with M-estimation

**Rates:** (if density bounded and without flat parts,  $\mathcal{G}$  VC-class)

With probability at least  $1 - \delta$ ,

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





Part I: Performance Criterion for Scoring Functions

Part II: Learning accurate scoring functions on extreme regions

Part III: Heuristic Contributions and Perspectives

# 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

→ Anomalies = points which violate this structure

# Theoretical framework

- ▶ **Context**

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

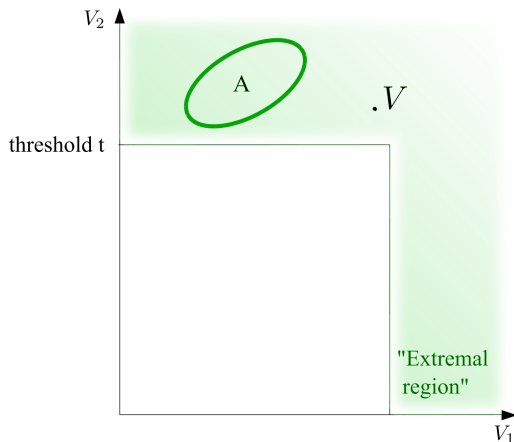
- ▶ **Preliminary step: Standardization of each marginal**

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

# Problematic of Extreme Value Theory

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

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



# 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):

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

$\mu$ : 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 of multivariate EVT

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

## Model for excesses

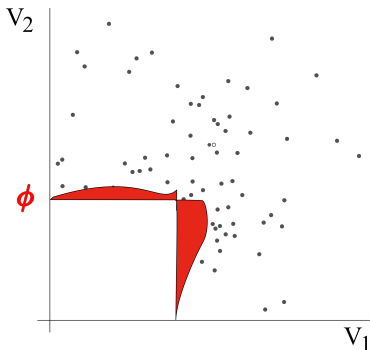
For a large  $r > 0$  and a region  $B$  on the unit sphere:

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

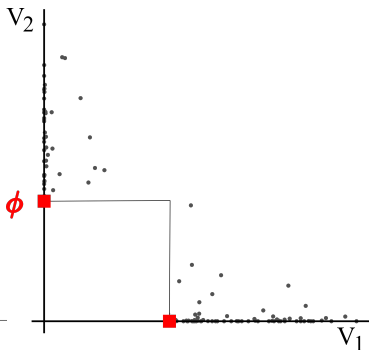
$\Rightarrow \Phi$  (or  $\mu$ ) **rules the joint distribution of extremes** (if margins are known).

# Angular distribution

- $\Phi$  rules the joint distribution of extremes



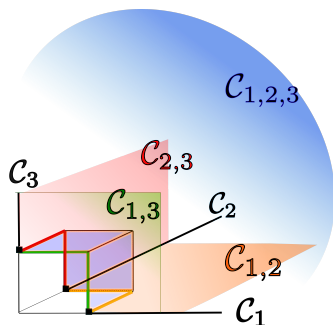
Asymptotic dependence:  
( $V_1$ ,  $V_2$ ) may be large together.



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



# General Case



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

# Representation of extreme data

- ▶ Natural decomposition of the angular measure :

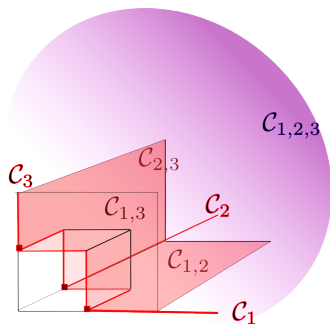
$$\Phi = \sum_{\alpha \subset \{1, \dots, d\}} \Phi_{\alpha} \quad \text{with } \Phi_{\alpha} = \Phi|_{\Omega_{\alpha}} \leftrightarrow \mu|_{\mathcal{C}_{\alpha}}$$

- ▶  $\Rightarrow$  yields a representation

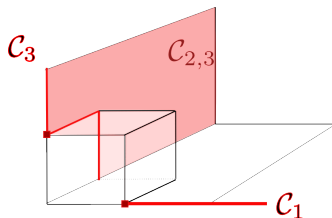
$$\begin{aligned} \mathcal{M} &= \left\{ \Phi(\Omega_{\alpha}) : \emptyset \neq \alpha \subset \{1, \dots, d\} \right\} \\ &= \left\{ \mu(\mathcal{C}_{\alpha}) : \emptyset \neq \alpha \subset \{1, \dots, d\} \right\} \end{aligned}$$

- ▶ Assumption:  $\frac{d\mu|_{\mathcal{C}_{\alpha}}}{dv_{\alpha}} = O(1)$ .

# Sparse Representation ?



Full pattern :  
anything may happen



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

# Estimation

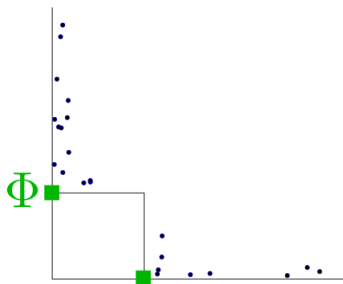
**Estimation Problem:**  $\mathcal{M}$  is an **asymptotic** representation

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

is the restriction of an asymptotic measure

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

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

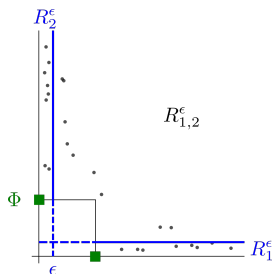


$\Rightarrow$  Cannot just do, for large  $t$ :

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

# Solution

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



$$\begin{aligned}\mathcal{C}_\alpha &\rightarrow \mathcal{C}_\alpha^\epsilon = \{\|v\| \geq 1, \ v_j > \epsilon \ (j \in \alpha), \ v_j \leq \epsilon \ (j \notin \alpha)\}. \\ \Omega_\alpha &\rightarrow \Omega_\alpha^\epsilon = \mathcal{C}_\alpha^\epsilon \cap \mathbf{S}_{d-1}\end{aligned}$$

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

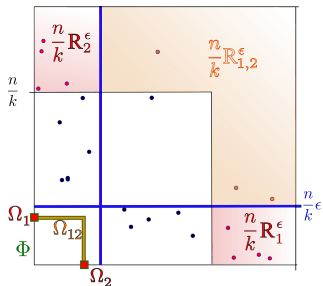
# Resulting estimation procedure

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

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

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

( $\frac{n}{k}$  large,  $\epsilon$  small)



$\Rightarrow$  we obtain

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

# Statistical guarantees: 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:  $\sup_{A \in \mathcal{A}} |(\mathcal{P} - \mathcal{P}_n)(A)|$

- ▶ scaling  $\frac{n}{k}$
- ▶ classical VC-inequality:  $\frac{k}{n}$  nice but not used !  
→ high proba bound in

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

⇒ 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

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

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

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

$\Rightarrow$  bound in

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

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



## 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  $1 - \delta$ ,*

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

## Comments:

- ▶  $M = \sup(\text{density on sub-cones})$
- ▶ Existing litterature (for spectral measure) Einmahl Segers 09, Einmahl *et al.* 01

$$d = 2.$$

asymptotic behaviour, rates in  $1/\sqrt{k}$ .

**Here:**  $1/\sqrt{k} \rightarrow 1/\sqrt{\epsilon k} + \epsilon$ . Price to pay for biasing our estimator with  $\epsilon$ .

# 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  $B$ :  $\text{Leb}(\mathcal{C}_\alpha^\epsilon \setminus \mathcal{C}_\alpha)$  is small when  $\epsilon$  is small.

# Corresponding Algorithm

DAMEX in  $O(dn \log n)$

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

**for**  $i = 1, \dots, n$  **do**

  # Standardize via marginal rank-transformation:

$$\hat{V}_i := (1/(1 - \hat{F}_j(X_i^j)))_{j=1, \dots, d}.$$

**if**  $\hat{V}_i > \frac{n}{k}$  **then**

    # Assign to each  $\hat{V}_i$  the cone  $\frac{n}{k}C_\alpha^\epsilon$  it belongs to:

$$\alpha = \alpha(V_i)$$

$$C_\alpha ++$$

**end if**

**end for**

$$\Phi_n^{\alpha, \epsilon} := \frac{n}{k} C_\alpha$$

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

$\Phi_n^{\alpha, \epsilon} = \hat{\Phi}(\Omega_\alpha) = \frac{n}{k} \mathbb{P}_n(\hat{V} \in \frac{n}{k} C_\alpha^\epsilon)$ , estimate of the  $\alpha$ -mass of  $\Phi$  for every  $\alpha$ .

# Application to Anomaly Detection

Recall that after standardization of marginals:

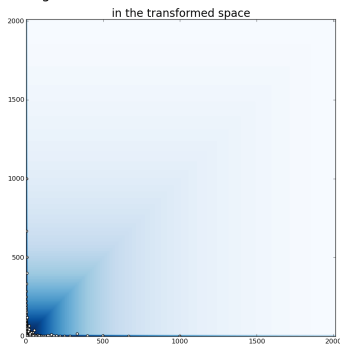
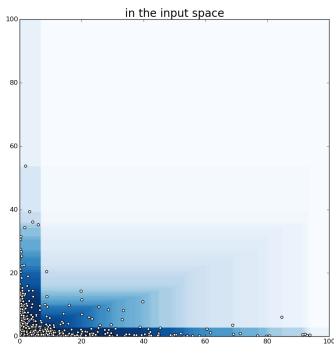
$$\mathbb{P}[R > r, \mathbf{W} \in B] \simeq \frac{1}{r} \Phi(B)$$

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

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

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

levels set of DAMEX scoring function



# Experiments

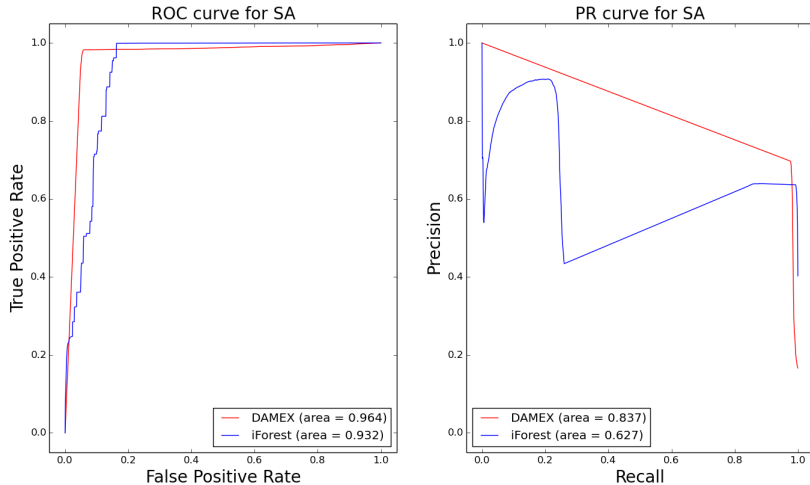


Figure: ROC and PR curve on SA dataset

Part I: Performance Criterion for Scoring Functions

Part II: Learning accurate scoring functions on extreme regions

Part III: Heuristic Contributions and Perspectives

# EM and MV curves for model selection?

In moderate dimension: [Thomas et al. 2015, 2016]

## Practical motivations:

Most of the time, data come without any label.

→ no ROC or PR curves!

## Implicit (natural) simplification of the problem:

How good is an anomaly detection algorithm?

→ How good is it estimating the level sets of the inliers distribution?

## Estimation:

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \text{Leb}(s \geq u) \quad \text{s.t.} \quad \mathbb{P}_n(s \geq u) \geq \alpha \quad \rightarrow \widehat{C}^{EM}(s)$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \mathbb{P}_n(s \geq u) - t \text{Leb}(s \geq u) \quad \rightarrow \widehat{C}^{MV}(s)$$

**Issue in large dimension:** The volume  $\text{Leb}(s \geq u)$  has to be estimated!

# Heuristic extension for large dimension

## 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**  $k = 1, \dots, m$  **do**

-randomly select a sub-group  $F_k$  of  $d'$  features

-compute the associated scoring function  $s_k = \mathcal{A}((x_i^j)_{1 \leq i \leq n, j \in F_k})$

-compute  $\hat{C}_k^{EM} = \|\widehat{EM}_{s_k}\|_{L^1(I)}$  or  $\hat{C}_k^{MV} = \|\widehat{MV}_{s_k}\|_{L^1(J)}$

**end for**

**Return** performance criteria:

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

Seems to work in practice but no statistical guarantees.



# One Class Random Forests

Existing litterature: [Désir et al., 2013, Liu et al., 2008, Shi and Horvath, 2012].

## One-Class splitting criterion

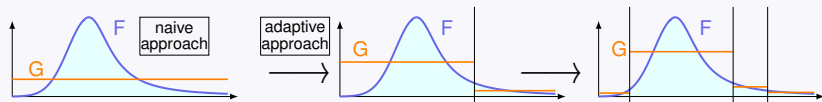


Figure:  $F$  is the inliers distribution,  $G$  is the assumed outliers distribution.

Does theorical guaranties on RFs apply ? [Biau et al. 2008, Biau and Scornet, 2016]

# AD and Extremes perspectives

A more accurate sparse representation

Thank you!

# Some references:

- ▶ Agarwal 2006, Detecting anomalies in cross-classified streams: a bayesian approach.
- ▶ 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.
- ▶ Eskin 2000, Anomaly detection over noisy data using learned probability distributions.
- ▶ Goix, Sabourin and Cléménçon 2015, Learning the dependence structure of rare events: a non-asymptotic study.
- ▶ Goix, Sabourin and Cléménçon 2015, Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking.
- ▶ Goix, Sabourin and Cléménçon 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.
- ▶ Hautamaki, Karkkainen and Franti 2004. Outlier detection using k-nearest neighbour graph.

# Some references:

- ▶ de Haan and Ferreira, Extreme value theory, 2006.
- ▶ Hawkins 1980, Identification of outliers.
- ▶ Lee and Roberts 2008, On-line novelty detection using the Kalman filter and extreme value theory.
- ▶ Liu, Ting, Zhou 2008, Isolation forest.
- ▶ Liu and Weng 1991, Detection of outlying data in bioavailability/bioequivalence studies.
- ▶ Papadimitriou, Kitagawa, Gibbons and Faloutsos 2002, Loci: Fast outlier detection using the local correlation integral.
- ▶ 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
- ▶ Tang, Chen, Fu, and W.Cheung 2002. Enhancing effectiveness of outlier detections for low density patterns.
- ▶ Thomas, Feuillard and Gramford 2015, Calibration of One-Class SVM for MV set estimation.
- ▶ Thomas, Clémenceau, Feuillard and Gramford 2016, Learning Hyperparameters

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

Table: Datasets characteristics

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

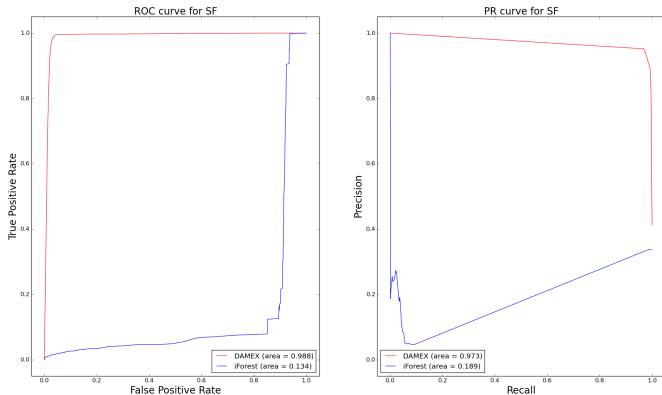


Figure: ROC and PR curve on SF dataset

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

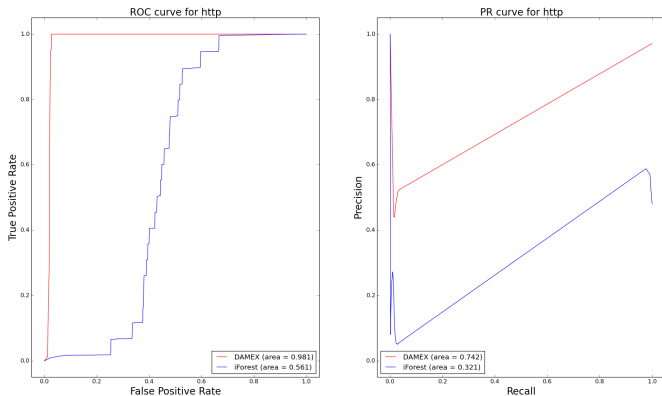


Figure: ROC and PR curve on http dataset



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

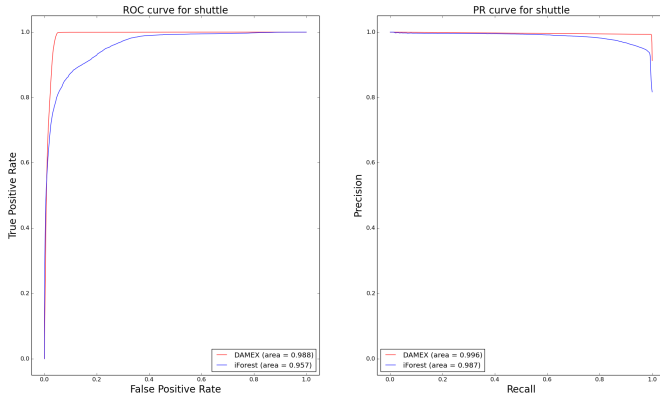


Figure: ROC and PR curve on shuttle dataset

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

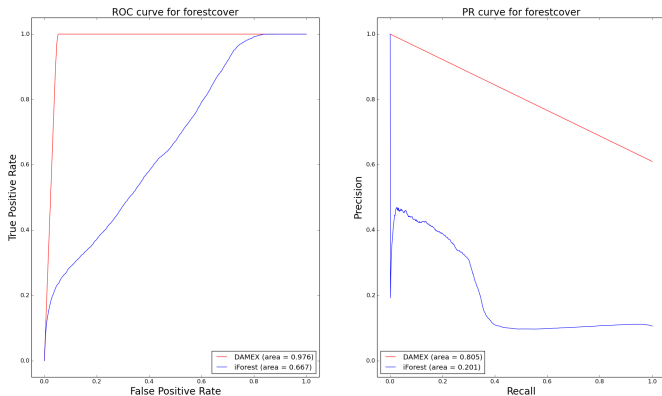


Figure: ROC and PR curve on forestcover dataset

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

- ▶ **Experiments:** 12 datasets, 3 AD algorithms (LOF, OCSVM, iForest)  
→ 36 possible pairwise comparisons:

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

- ▶ **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	<b>0.661</b>	<b>0.277</b>	<b>1.0e-04</b>	<b>7.5e01</b>	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	<b>0.999</b>	<b>0.970</b>	<b>6.0e-03</b>	<b>2.6</b>	<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	<b>1.2e04</b>	<b>0.760</b>	<b>0.229</b>	<b>5.2e-07</b>	<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>	<b>0.522</b>	<b>1.2e-03</b>	8.2	<b>0.922</b>	0.189	1.1e-03	<b>5.8</b>
wilt	0.491	<u>0.045</u>	4.7e-05	<u>2.1e03</u>	<u>0.325</u>	<u>0.037</u>	<b>5.9e-05</b>	<b>4.5e02</b>	<b>0.698</b>	<b>0.088</b>	<u>2.1e-05</u>	1.6e03
annthyroid	<b>0.913</b>	<b>0.456</b>	<b>2.0e-04</b>	2.6e02	<u>0.699</u>	<u>0.237</u>	<u>6.3e-05</u>	<b>2.2e02</b>	0.823	0.432	6.3e-05	<u>1.5e03</u>
arrhythmia	<b>0.763</b>	<b>0.487</b>	<b>1.6e-04</b>	<b>9.4e01</b>	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	<b>0.990</b>	<b>0.792</b>	<b>3.5e-04</b>	<b>3.9e01</b>
ionosphere	<u>0.902</u>	<u>0.529</u>	<u>9.6e-05</u>	<u>7.5e01</u>	<b>0.977</b>	<b>0.898</b>	<b>1.3e-04</b>	<b>5.4e01</b>	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>	<b>0.983</b>	<b>0.829</b>	<b>4.6e-04</b>	<b>1.7e01</b>
shuttle	0.996	0.973	1.8e-05	5.7e03	<u>0.992</u>	<u>0.924</u>	<b>3.2e-05</b>	<b>2.0e01</b>	<b>0.999</b>	<b>0.994</b>	<u>7.9e-06</u>	<u>2.0e06</u>
spambase	<b>0.824</b>	<b>0.371</b>	<b>9.5e-04</b>	<b>4.5e01</b>	<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>