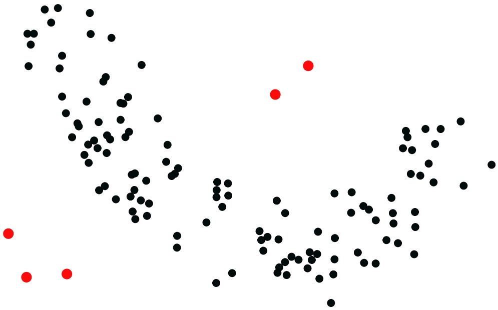
Machine Learning and Extremes for Anomaly Detection

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

PhD Defense, Telecom Paristech, November 28th 2016

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



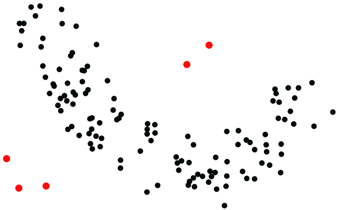
Applications: Network intrusions, credit card fraud detection, insurance, finance, military surveillance, predictive maintenance...

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)
  Synonym: one-class classification. 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]

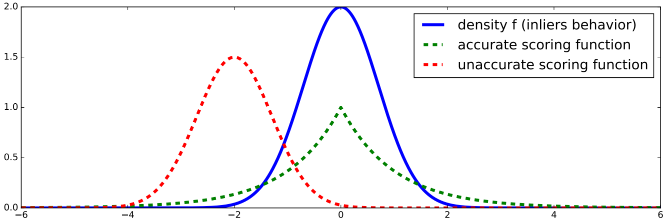
Background: scoring functions

Notation: Normal behavior $\rightarrow$ density f.

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

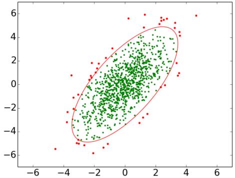
Part II: Learning accurate scoring functions on extreme regions

Part III: Heuristic contributions and perspectives

Existing criterion: Mass-Volume curve

Minimum volume set [Einmahl and Mason 1992, Polonik 1997]

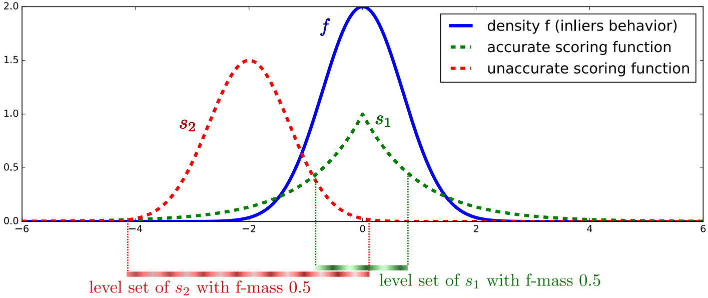
$$\Gamma^*_\alpha = \underset{\Gamma \text{ borelian}}{\arg\min} \quad \mathrm{Leb}(\Gamma) \quad s.t. \quad \mathbb{P}(\mathbf{X} \in \Gamma) \geq \alpha$$



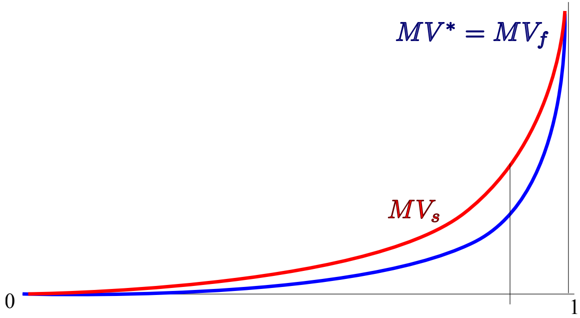
Mass Volume curve of a scoring function s [Clémençon and Jakubowicz, 2013]:

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

$$MV^*(\alpha) \; := \; MV_f(\alpha)$$



Existing criterion: Mass-Volume curve



Main drawbacks of MV:

- When optimized *w.r.t.* different levels α over a finite class, 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.

Our contribution: Excess-Mass curve [Goix, Sabourin, Clémençon 2015]

Excess-Mass [Hartigan 1987, Polonik 1995]

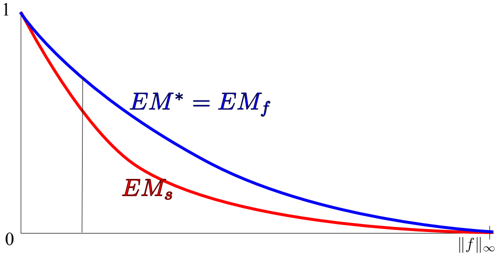
Excess-Mass curve

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

$$EM^*(t) := EM_f(t)$$

Property: Previous drawbacks are fixed with EM.

- Produces **nested** empirical level sets.
- $\rightarrow$ **convergence rates** of order $O(n^{-1/2})$.
- EM curve **finite** even in case of unbounded support.



Learning a scoring function with M-estimation

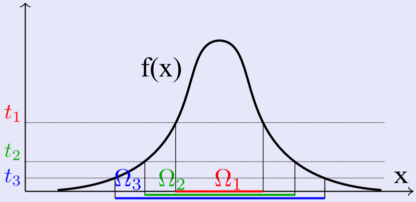
$\mathcal{G}$: VC-class of sets.

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

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

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

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

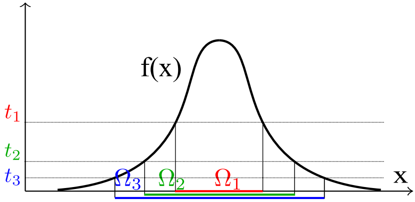


Learning a scoring function with M-estimation

Theorem

Assume the density f bounded, with compact support and without flat parts and $\mathcal{G}$ VC-class. Then if $t_N = \mathcal{O}(n^{-1/2})$, with probability at least $1-\delta$,

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



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

$$\mathbf{X} = (X_1, \ldots, X_5)$$

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

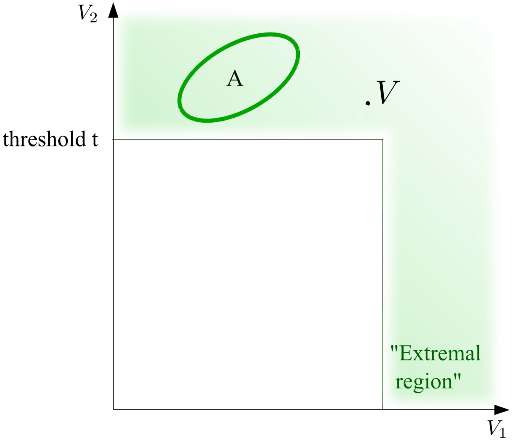
Theoretical framework

- **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.

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



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 $\mathbf{S}_{d-1}$: $\Phi(B) = \mu\{tB, t \geq 1\}$

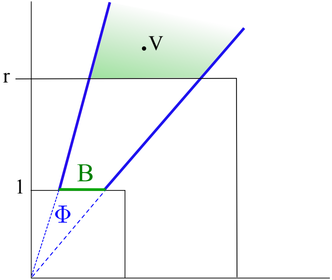
$\mathbb{P}[\mathbf{V} \in A] \quad \simeq \quad \mu(A)$, if A extreme region.

Model for excesses

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

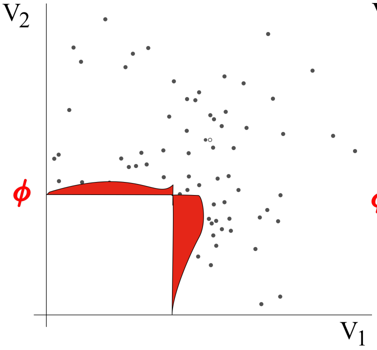
$$\mathbb{P}\left[\|\mathbf{V}\| > r, \ \frac{\mathbf{V}}{\|\mathbf{V}\|} \in B\right] \underset{r\to\infty}{\sim} \frac{1}{r}\Phi(B) = \mu(\{tB, t \geq r\})$$

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

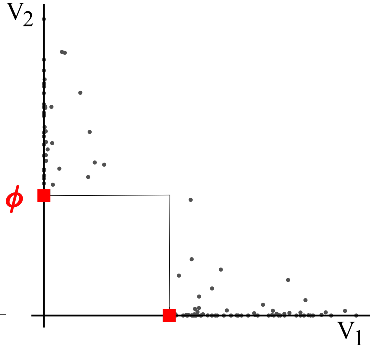


Angular distribution

Φ rules the joint distribution of extremes:

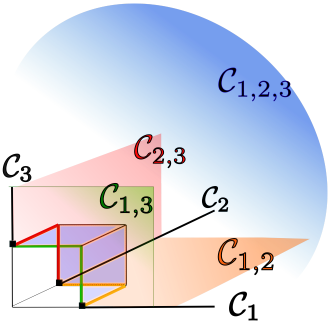


Asymptotic dependence:
(V_1, V_2) may be large together.



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

General Case



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

Representation of extreme data

- Natural decomposition of the angular measure :

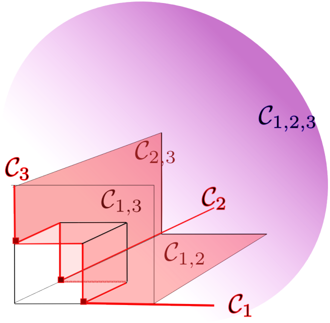
$$\Phi = \sum_{\alpha \subset \{1,\dots,d\}} \Phi_\alpha \qquad \text{with } \Phi_\alpha = \Phi_{|\Omega_\alpha} \leftrightarrow \mu_{|\mathcal{C}_\alpha}$$

- $\Rightarrow$ yields a representation

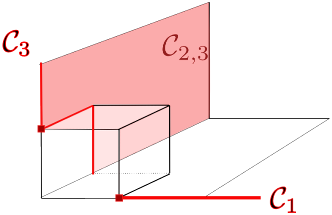
$$\begin{aligned} \mathcal{M} &= \left\{ \Phi(\Omega_\alpha) : \quad \emptyset \neq \alpha \subset \{1, \dots, d\} \right\} \\ &= \left\{ \mu(\mathcal{C}_\alpha) : \quad \emptyset \neq \alpha \subset \{1, \dots, d\} \right\} \end{aligned}$$

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

Sparse Representation ?



Full pattern :
anything may happen



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

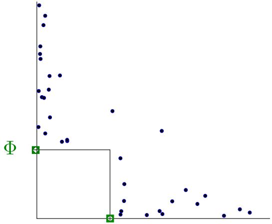
Estimation Problem: $\mathcal{M}$ is an **asymptotic** representation

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

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!

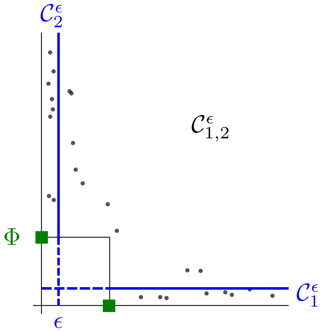


$\Rightarrow$ Cannot just do, for large t:

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

Solution

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



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

New partition.

Resulting estimation procedure

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

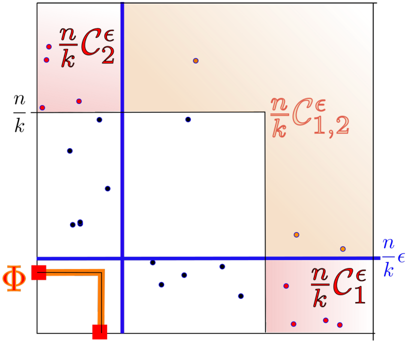
Recall that:

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

$\Rightarrow$ 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

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

Theorem [Goix, Sabourin, Clémençon 2016]

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 $1 - \delta$,

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

Comments:

- $M \simeq \sum_\alpha$ sup(density on cones α)
- Existing literature (for spectral measure) [Einmahl and Segers 09, Einmahl *et al.* 01]

$$d = 2, \text{ asymptotic behavior, rates in } 1/\sqrt{k}.$$

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

Theorem's proof: key ingredient

Would like to use concentration inequality...

In our case: $\sup_{A \in \mathcal{A}} \frac{n}{k} \left| (\mathcal{P} - \mathcal{P}_n) \left(\frac{n}{k} A \right) \right|$

But usually: $\sup_{A \in \mathcal{A}} \left| (\mathcal{P} - \mathcal{P}_n) (A) \right|$

- scaling $\frac{n}{k}$
- classical VC-inequality: $\frac{n}{k}$ 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 $\frac{n}{k} A$ is small.

Theorem's proof: key ingredient

Key: VC-inequality adapted to rare regions $\longrightarrow$ 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 $\cup_{A\in\mathcal{A}} A$.

$$\mathbf{p} \leq d\frac{k}{n}$$

$\Rightarrow$ bound in

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

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

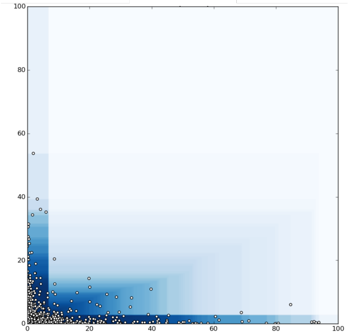
Application to Anomaly Detection

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



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?

Practical motivations:

Most of the time, data come without any label.

$\rightarrow$ no ROC or PR curves!

Estimation:

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

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

[Thomas et al. 2015]

Issue in large dimensions:

The volume $\mathrm{Leb}(s \geq u)$ has to be estimated!

EM and MV curves for model selection?

Heuristic extension for large dimension:

Random projection and averaging [Goix 2016]

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 $s_k = \mathcal{A}\big((x_i^j)_{1 \le i \le n,\ j \in F_k}\big)$
-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:

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

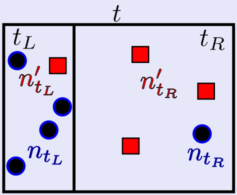
Seems to work in practice but no statistical guaranties.

Two-class Random Forests [Breiman, 2001]

Two-Class impurity decrease

$$I_G(t_L, t_R) = \frac{n_{t_L} n'_{t_L}}{n_{t_L} + n'_{t_L}} + \frac{n_{t_R} n'_{t_R}}{n_{t_R} + n'_{t_R}}.$$

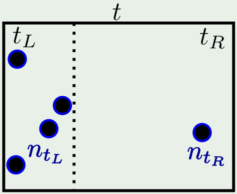
n_t : nb of observations with label 0 in node t.
n'_t: nb of observations with label 1 in node t.



Random Forests for one-class classification

Two-Class: $I_G(t_L, t_R) = \frac{n_{t_L} n'_{t_L}}{n_{t_L} + n'_{t_L}} + \frac{n_{t_R} n'_{t_R}}{n_{t_R} + n'_{t_R}}.$

One-Class: $n'_{t_L}, \; n'_{t_R} = ?$

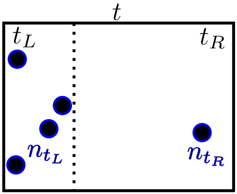


Existing literature: [Liu et al., 2008, Désir et al., 2013, Shi and Horvath, 2012].
Based on **second-class sampling**.

Random Forests for one-class classification

Two-Class: $I_G(t_L, t_R) = \frac{n_{t_L} n'_{t_L}}{n_{t_L} + n'_{t_L}} + \frac{n_{t_R} n'_{t_R}}{n_{t_R} + n'_{t_R}}.$

One-Class: $n'_{t_L}, \; n'_{t_R} = ?$

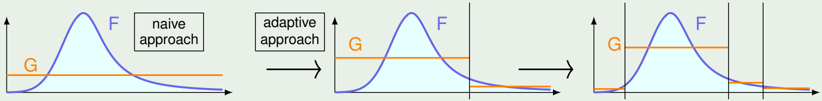


One-Class splitting criterion

[Goix, Brault, Drougard, Chiapino 2016]:

Naive approach: $n'_{t_L} \to \mathbf{n}\frac{\text{Leb}(t_L)}{\text{Leb}(\mathbf{t_0})}$ and $n'_{t_R} \to \mathbf{n}\frac{\text{Leb}(t_R)}{\text{Leb}(\mathbf{t_0})}$ ($\mathbf{t_0}$ root node)

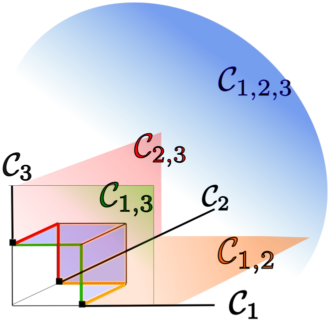
Adaptive approach: $n'_{t_L} \to \mathbf{n_t}\frac{\text{Leb}(t_L)}{\text{Leb}(\mathbf{t})}$ and $n'_{t_R} \to \mathbf{n_t}\frac{\text{Leb}(t_R)}{\text{Leb}(\mathbf{t})}$



F is the inliers distribution, G is the resulting outliers distribution.

Theoretical guaranties ? [Biau et al. 2008, Biau and Scornet, 2016]

Perspectives on AD with Extremes



- How to choose ϵ in practice?
- Refine our representation?
- An alternative notion of sparsity?

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 Clémençon 2015, Learning the dependence structure of rare events: a non-asymptotic study.
- Goix, Sabourin and Clémençon 2016, 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.
- Hautamaki, Karkkainen and Franti 2004. Outlier detection using k-nearest neighbour graph.
- 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.

Some references:

- 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 1995, Measuring mass concentrations and estimating density contour cluster-an excess mass approach.
- 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, Clémençon, Feuillard and Gramford 2016, Learning Hyperparameters for Unsupervised Anomaly Detection.
- Vert and Vert 2006, Consistency and Convergence Rates of One-Class SVMs and Related Algorithms

DAMEX in $O(dn \log n)$

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

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

Standardize *via* marginal rank-transformation:

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

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.

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

EM/MV random projection benchmark

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:

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

- **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.

EM/MV random projection benchmark

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>

OCRF Benchmark

Table: Results for the novelty detection setting

Datasets	OneClassRF		iForest		OCRFsampl.		OCSVM		LOF		Orca		LSAD		RFC	
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
adult	**0.665**	**0.278**	0.661	0.227	NA	NA	0.638	0.201	0.615	0.188	0.606	0.218	0.647	0.258	NA	NA
annthyroid	**0.936**	0.468	0.913	0.456	0.918	**0.532**	0.706	0.242	0.832	0.446	0.587	0.181	0.810	0.327	NA	NA
arrhythmia	0.684	0.510	0.763	0.492	0.639	0.249	**0.922**	**0.639**	0.761	0.473	0.720	0.466	0.778	0.514	0.716	0.299
forestcover	0.968	0.457	0.863	0.046	NA	NA	NA	NA	**0.990**	**0.795**	0.946	0.558	0.952	0.166	NA	NA
http	**0.999**	**0.838**	0.994	0.197	NA	NA	NA	NA	NA	NA	**0.999**	0.812	0.981	0.537	NA	NA
ionosphere	0.909	0.643	0.902	0.535	0.859	0.609	0.973	0.849	0.959	0.807	0.928	**0.910**	**0.978**	0.893	0.950	0.754
pendigits	0.960	0.559	0.810	0.197	0.968	0.694	0.603	0.110	0.983	0.827	**0.993**	**0.925**	0.983	0.752	NA	NA
pima	0.719	0.247	0.726	0.183	**0.759**	**0.266**	0.716	0.237	0.700	0.152	0.588	0.175	0.713	0.216	0.506	0.090
shuttle	**0.999**	**0.998**	0.996	0.973	NA	NA	0.992	0.924	**0.999**	0.995	0.890	0.782	0.996	0.956	NA	NA
smtp	0.922	0.499	0.907	0.005	NA	NA	0.881	**0.656**	**0.924**	0.149	0.782	0.142	0.877	0.381	NA	NA
spambase	**0.850**	0.373	0.824	0.372	0.797	**0.485**	0.737	0.208	0.746	0.160	0.631	0.252	0.806	0.330	0.723	0.151
wilt	0.593	0.070	0.491	0.045	0.442	0.038	0.323	0.036	0.697	0.092	0.441	0.030	0.677	0.074	**0.896**	**0.631**
average:	**0.850**	**0.495**	0.821	0.311	0.769	0.410	0.749	0.410	0.837	0.462	0.759	0.454	**0.850**	0.450	0.758	0.385
cum. train time:	**61s**		68s		NA		NA		NA		2232s		73s		NA	

Datasets

Datasets	nb of samples	nb of features	anomaly class	
adult	48842	6	class '$> 50K$'	(23.9%)
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%)
http	567498	3	attack	(0.39%)
ionosphere	351	32	bad	(35.9%)
pendigits	10992	16	class 4	(10.4%)
pima	768	8	pos (class 1)	(34.9%)
shuttle	85849	9	classes $\neq 1$ (class 4 removed)	(7.17%)
smtp	95156	3	attack	(0.03%)
spambase	4601	57	spam	(39.4%)
wilt	4839	5	class 'w' (diseased trees)	(5.39%)