Weak Signals: machine-learning meets extreme value theory

Stephan Clémençon
Télécom ParisTech, LTCI, Université Paris Saclay
machinelearningforbigdata.telecom-paristech.fr

2017-11-29, Workshop on Machine Learning and FinTech

Agenda

- Motivation - Health monitoring in aeronautics
- Anomaly detection in the Big Data era: a statistical learning view
- Anomalies and extremal dependence structure: a MV-set approach
- Theory and practice
- Conclusion - Lines of further research

Motivation - Context

- Era of Data - **Ubiquity of sensors**
  ex: an aircraft engine can equipped with more than 2000 sensors monitoring its functioning (pressure, temperature, vibrations, *etc.*)

- **Very high dimensional** setting: traditional survival analysis is inappropriate for **predictive maintenance**
- **Health monitoring**: avoid failures via early detection of abnormal behavior of a complex infrastructure

- The vast majority of the data are **unlabeled**
  Rarity should replace labels...

Anomalies correspond to **multivariate extreme** observations, but the reverse is not true in general

- False alarms are **very expensive** and should be **interpretable** by professional experts

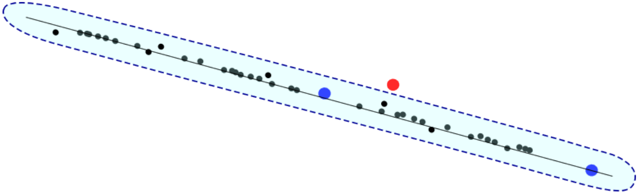
The many faces of Anomaly Detection

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"

Learning how to detect anomalies automatically



- <u>**Step 1:**</u> Based on **training data**, **learn a region** in the space of observations describing the "normal" behavior
- <u>**Step 2:**</u> **Detect anomalies** among new observations. Anomalies are observations lying outside the critical region

The many faces of Anomaly Detection

Different frameworks for 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

Supervised Learning Framework for Anomaly Detection

- (X, Y) random pair, valued in $\mathbb{R}^d \times \{-1, +1\}$ with $d >> 1$
  A positive label '$Y = +1$' is assigned to anomalies.
- **Observation:** sample $\mathcal{D}_n$ of i.i.d. copies of (X, Y)

$$(X_1, Y_1),\ \ldots,\ (X_n, Y_n)$$

- **Goal:** from labeled data $\mathcal{D}_n$, learn to **predict** labels assigned to new data $X'_1,\ \ldots,\ X'_{n'}$
- A typical binary classification problem...
  except that $p = \mathbb{P}\{Y = +1\}$ may be extremely small

The Flagship Machine-Learning Problem: Supervised Binary Classification

- $X \in$ observation with dist. $\mu(dx)$ and $Y \in \{-1, +1\}$ binary label
- *A posteriori* probability $\sim$ **regression function**

$$\forall x \in \mathbb{R}^d, \quad \eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$$

- $g : \mathbb{R}^d \to \{-1, +1\}$ prediction rule - **classifier**
- Performance measure = **classification error**

$$L(g) = \mathbb{P}\{g(X) \neq Y\} \quad \to \min_g L(g)$$

- Solution: **Bayes classifier** $g^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1$
- **Bayes error** $L^* = L(g^*) = 1/2 - \mathbb{E}[|2\eta(X) - 1|]/2$

Empirical Risk Minimization - Basics

- Sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ with i.i.d. copies of (X, Y)
- Class $\mathcal{G}$ of classifiers of a given **complexity**
- **Empirical Risk Minimization principle**

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} L_n(g)$$

with $L_n(g) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{g(X_i) \neq Y_i\}$

- Mimic the best classifier among the class

$$\bar{g} = \arg\min_{g \in \mathcal{G}} L(g)$$

Guarantees - Empirical processes in classification

- **Bias-variance decomposition**

$$\begin{aligned} L(\hat{g}_n) - L^* &\leq (L(\hat{g}_n) - L_n(\hat{g}_n)) + (L_n(\bar{g}) - L(\bar{g})) + (L(\bar{g}) - L^*) \\ &\leq 2\left(\sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid\right) + \left(\inf_{g\in\mathcal{G}} L(g) - L^*\right) \end{aligned}$$

- **Concentration results**

With probability $1-\delta$:

$$\sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid \leq \mathbb{E}\left[\sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid\right] + \sqrt{\frac{2\log(1/\delta)}{n}}$$

Main results in classification theory

1. Bayes risk consistency and rate of convergence

   Complexity control:

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \mid L_n(g) - L(g) \mid \right] \leq C\sqrt{\frac{V}{n}}$$

   if $\mathcal{G}$ is a VC class with VC dimension V.

2. Fast rates of convergence

   Under variance control: rate faster than $n^{-1/2}$

3. Convex risk minimization: Boosting, SVM, Neural Nets, *etc.*

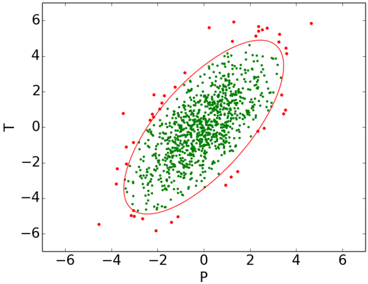
4. Oracle inequalities - Model selection

Unsupervised anomaly detection

$X_1, \ldots, X_n \in \mathbb{R}^d$ i.i.d. realizations of unknown probability measure $\mu(dx) = f(x)\lambda(dx)$

- Anomalies are supposed to be rare events, located in the tail of the distribution

  a critical region should be defined as the complementary of a **density sublevel set**
- Estimation of the region where the data are most concentrated: region of **minimum volume** for a given probability content α close to 1
- M-estimation formulation



Minimum Volume set, $\alpha = 0.95$

Minimum Volume set (MV set) - the Excess Mass approach

Definition [Einmahl & Mason, 1992]

- $\alpha \in [0, 1]$ (for anomaly detection α is close to 1)
- $\mathcal{C}$ class of measurable sets
- $\mu(dx)$ unknown probability measure of the observations
- λ Lebesgue measure

$$Q(\alpha) = \arg\min_{C \in \mathcal{C}} \{\lambda(C), \mathbb{P}(X \in C) \geqslant \alpha\}$$

- For small values of α, one recovers the **modes**.
- For large values:
  - Samples that belong to the MV set will be considered as **normal**
  - Samples that do not belong to the MV set will be considered as **anomalies**

Theoretical MV sets

Consider the following assumptions:

- The distribution μ has a density $f(x)$ w.r.t. λ such that $f(X)$ is bounded,
- The distribution of the r.v. $f(X)$ has no plateau, *i.e.* $\mathbb{P}(f(X) = c) = 0$ for any $c > 0$.

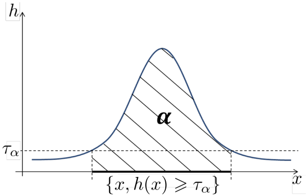
Under these hypotheses, there exists a unique MV set at level α:

$$G_\alpha^* = \{x \in \mathbb{R}^d : \ h(x) \geq t_\alpha\}$$

is a *density level set*, t_α is the quantile at level $1 - \alpha$ of the r.v. $h(X)$.

MV set estimation

Goal: learn a MV set $Q(\alpha)$ from $X_1, \ldots, X_n$



Empirical Risk Minimization paradigm: replace the unknown distribution μ by its statistical counterpart

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

and solve $\min_{G \in \mathcal{G}} \lambda(G)$ subject to $\widehat{\mu}_n(G) \geq \alpha - \phi_n$, where ϕ_n is some *tolerance level* and $\mathcal{G} \subset \mathcal{C}$ is a class of measurable subsets whose volume can be computed/estimated (*e.g.* Monte Carlo).

Connection with ERM, Scott & Nowak '06

- The approach is valid, provided $\mathcal{G}$ is **simple enough**, *i.e.* of controlled complexity (*e.g.* finite VC dimension)

$$\sup_{G\in\mathcal{G}} |\widehat{\mu}_n(G) - \mu(G)| \leq c\sqrt{\frac{V}{n}}$$

- The approach is accurate, provided that $\mathcal{G}$ is **rich enough**, *i.e.* contains a reasonable approximant of a MV set at level α

- The **tolerance level** should be chosen of the same order as $\sup_{G\in\mathcal{G}} |\widehat{\mu}_n(G) - \mu(G)|$

- **Model selection:** $\mathcal{G}_1, \ldots, \mathcal{G}_K \Rightarrow \widehat{G}_1, \ldots, \widehat{G}_K$

$$\widehat{k} = \arg\min_k \left\{ \lambda(\widehat{G}_k) + 2\phi_k :\ \widehat{\mu}_n(\widehat{G}_k) \geq \alpha - \phi_k \right\}$$

Statistical Methods

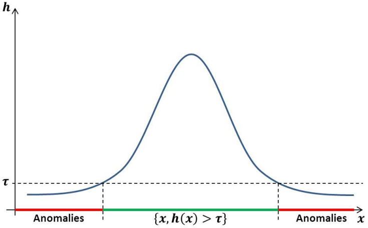
- Plug-in techniques (fit a model for $f(x)$)
- Turning unsupervised AD into binary classification
- Histograms
- Decision trees
- SVM
- Isolation Forest

Unsupervised anomaly detection - Mass Volume curves

- Anomalies are the rare events, located in the low density regions
- Most unsupervised anomaly detection algorithms learn a scoring function

$$s : x \in \mathbb{R}^d \mapsto \mathbb{R}$$

  such that the smaller $s(x)$ the more abnormal is the observation x.
- Ideal scoring functions: any increasing transform of the density $h(x)$



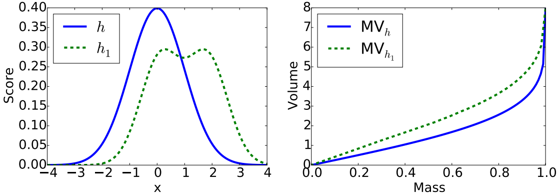
Mass Volume curve

$X \sim h$, scoring function s, t-level set of s: $\{x, s(x) \geq t\}$

- $\alpha_s(t) = \mathbb{P}(s(X) \geq t)$ **mass** of the t-level set
- $\lambda_s(t) = \lambda(\{x, s(x) \geq t\})$ **volume** of t-the level set.

Mass Volume curve MV_s of $s(x)$ [Clémençon and Jakubowicz, 2013]:

$$t \in \mathbb{R} \mapsto (\alpha_s(t), \lambda_s(t))$$



Mass Volume curve

MV_s also defined as the function

$$\text{MV}_s : \alpha \in (0,1) \mapsto \lambda_s(\alpha_s^{-1}(\alpha)) = \lambda(\{x, s(x) \geq \alpha_s^{-1}(\alpha)\})$$

where α_s^{-1} generalized inverse of α_s.

Property [Clémençon and Jakubowicz, 2013]

Let MV^* be the MV curve of the underlying density h and assume that h has no flat parts, then for all s with no flat parts,

$$\forall \alpha \in (0,1), \quad \text{MV}^*(\alpha) \leq \text{MV}_s(\alpha)$$

The closer is MV_s to MV^* the better is s

A MEVT Approach to Anomaly Detection

Main assumption:

Anomalies correspond to unusual simultaneous occurrence of extreme values for specific variables.

State of the Art: experts/practicioners set thresholds by hand

Anomaly detection in 'extreme' data

'Extremes' = points located in the tail of the distribution.
In Big Data samples, extremes can be observed with high probability
Learn statistically what 'normal' among extremes means?

Requirement: beyond interpretability and false alarm rate reduction, the method should be insensitive to unit choices

Multivariate EVT for Anomaly detection

- If ‘normal’ data are heavy tailed, there may be **extreme** normal data.

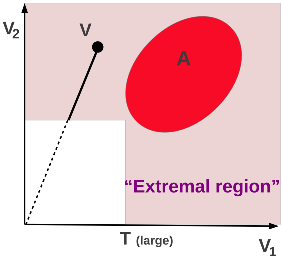
How to distinguish between large anomalies and normal extremes?

- Anomalies among extremes are those which direction $X/||X||_\infty$ is unusual.

  Our proposal: critical regions should be complementary sets of MV-sets of the **angular measure**, that describes the *dependence structure*

Multivariate extremes

- Random vectors $\mathbf{X} = (X_1, \ldots, X_{d,})$; $\quad X_j \geq 0$
- Margins: $X_j \sim F_j$, $1 \leq j \leq d$ (continuous).
- **Preliminary step: Standardization** $V_j = T(X_j) = \frac{1}{1-F_j(X_j))}$
  $\Rightarrow \mathbb{P}(V_j > v) = \frac{1}{v}$.
- Goal : $\mathbb{P}\{\mathbf{V} \in A\}$, A 'far from 0' ?



Fundamental assumption and consequences   de Haan, Resnick, 70's, 80's

Intuitively: $\mathbb{P}(\mathbf{V} \in tA) \simeq \frac{1}{t}\mathbb{P}(\mathbf{V} \in A)$

Multivariate regular variation

$$0 \notin \bar{A}: \qquad t\,\mathbb{P}\left(\frac{\mathbf{V}}{t} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \qquad \mu: \text{ Exponent measure}$$

necessarily: $\mu(tA) = t^{-1}\mu(A)$ (Radial homogeneity)
$\to$ **angular measure** on the sphere : $\Phi(B) = \mu\{tB, t \geq 1\}$

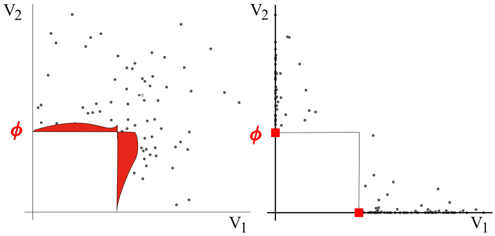
General model for extremes

$$\mathbb{P}\left(\|\mathbf{V}\| \geq r\,; \quad \frac{\mathbf{V}}{\|\mathbf{V}\|} \in B \right) \simeq r^{-1}\,\Phi(B)$$

Polar coordinates: $r(\mathbf{V}) = \|\mathbf{V}\|$, $\theta(\mathbf{V}) = \mathbf{V}/\|\mathbf{V}\|$

Angular measure

- Φ rules the joint distribution of extremes



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

  vs

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

No assumption on Φ: non-parametric framework.

MV-set estimation on the Sphere

Let λ_d be Lebesgue measure on $\mathbb{S}_{d-1}$ Fix $\alpha \in (0, \Phi(\mathbb{S}_{d-1}))$. Consider the 'asymptotic' problem:

$$\min_{\Omega \in \mathcal{B}(\mathbb{S}_{d-1})} \lambda_d(\Omega) \text{ subject to } \Phi(\Omega) \geq \alpha.$$

Replace the limit measure by the *sub-asymptotic* angular measure at finite level t:

$$\Phi_t(\Omega) = t\mathbb{P}\{r(\mathbf{V}) > t, \theta(\mathbf{V}) \in \Omega\}$$

We have $\Phi_t(\Omega) \to \Phi(\Omega)$ as $t \to \infty$. Replace the problem above by a non asymptotic version:

$$\min_{\Omega \in \mathcal{B}(\mathbb{S}_{d-1})} \lambda_d(\Omega) \text{ subject to } \Phi_t(\Omega) \geq \alpha.$$

The radius threshold t plays a role in the statistical method

Algorithm - Empirical estimation of an angular MV-set

Inputs: Training data $X_1, \ldots, X_n$, $k \in \{1, \ldots, n\}$, mass level α, confidence level $1-\delta$, tolerance $\psi_k(\delta)$, collection $\mathcal{G}$ of measurable subsets of $\mathbb{S}_{d-1}$

Standardization: Apply the rank-transformation, yielding

$$\widehat{V}_i = \widehat{T}(X_1) = \left(\frac{1}{1-\widehat{F}_1(X_i^{(1)})}, \ldots, \frac{1}{1-\widehat{F}_d(X_i^{(d)})} \right)$$

Thresholding: With $t = n/k$, extract the indexes

$$\mathcal{I} = \left\{ i : \; r(\widehat{V}_i) \geq n/k \right\} = \left\{ i : \; \exists j \leq d, \; \widehat{F}_i(X_i^{(j)}) \geq 1 - k/n \right\}$$

and consider the population of angles $\{\theta_i = \theta(\widehat{V}_i), \; i \in \mathcal{I}\}$

Empirical MV-set estimation: Form $\widehat{\Phi}_{n,k} = (1/k) \sum_{i \in \mathcal{I}} \delta_{\theta_i}$ and solve

$$\min_{\Omega \in \mathcal{G}} \lambda_d(\Omega) \text{ subject to } \widehat{\Phi}_{n,k}(\Omega) \geq \alpha - \psi_k(\delta)$$

Output: Empirical MV-set $\widehat{\Omega}_\alpha$

Theoretical guarantees - Assumptions

- For any $t > 1$, $\Phi_t(d\theta) = \phi_t(\theta) \cdot \lambda_d(d\theta)$ and $\forall c > 0$

$$\mathbb{P}\{\phi_t(\theta(\mathbf{V})) = c\} = 0$$

- $\sup_{t>1 \theta \in \mathbb{S}_{d-1}} \phi_t(\theta) < \infty$

Under these assumptions, the MV set problem at level α has a unique solution

$$B^*_{\alpha,t} = \{\theta \in \mathbb{S}_{d-1} : \ \phi_t(\theta) \geq K^{-1}_{\Phi_t}(\Phi(\mathbb{S}_{d-1}) - \alpha)\},$$

where $K_{\Phi_t}(y) = \Phi_t(\{\theta \in \mathbb{S}_{d-1} : \ \phi_t(\theta) \leq y\})$.

If the continuity assumption is not fulfilled?

Dimensionality reduction in the extremes

- Reasonable hope: only a moderate number of V_j's may be simultaneously large → **sparse angular measure**

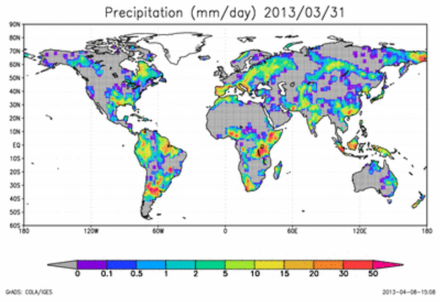
- In Clémençon, Goix and Sabourin (JMVA, 2017):

Estimation of the (sparse) support of the angular measure (*i.e.* the dependence structure).

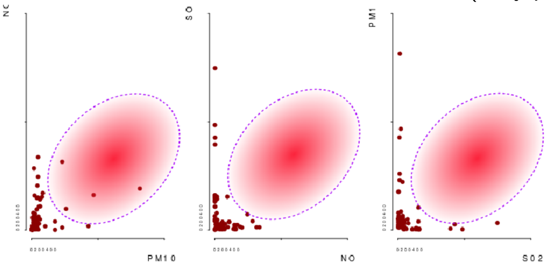
Which components may be large together, while the other are small?

  - Recover the asymptotically dependent groups of components → apply empirical MV-set estimation on the sphere to these groups/subvectors.

It cannot rain everywhere at the same time

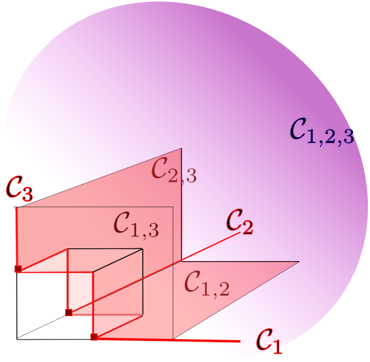


(daily precipitation)

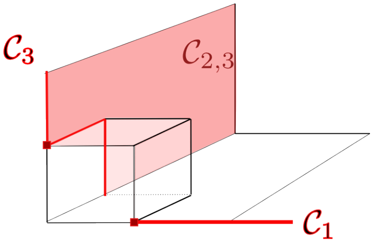


(air pollutants)

Recovering the (hopefully) sparse angular support



Full support:
anything may happen

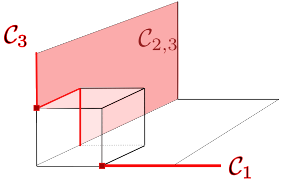


Sparse support
(V_1 not large if V_2 or V_3 large)

Where is the mass?

Subcones of $\mathbb{R}_+^d$: $\mathcal{C}_{\boldsymbol{\alpha}} = \{x \succeq 0, x_i \geq 0\ (i \in \alpha),\ x_j = 0\ (j \notin \alpha), \|x\| \geq 1\}$
$\alpha \subset \{1, \ldots, d\}$.

Support recovery + representation



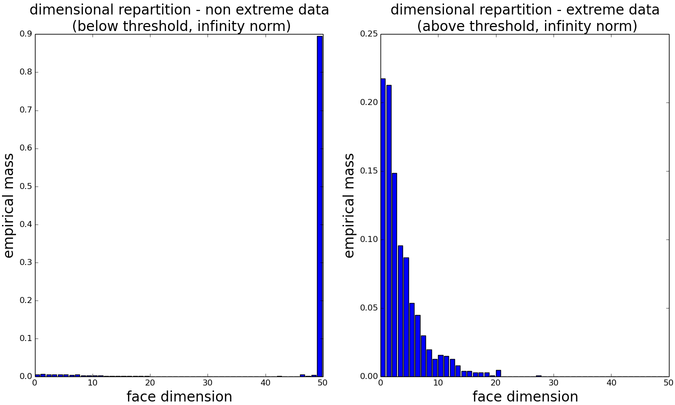
- $\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}$: partition of the unit sphere
- $\{\mathcal{C}_\alpha, \alpha \subset \{1, \ldots, d\}$: corresponding partition of $\{x : \|x\| \geq 1\}$
- μ-mass of subcone $\mathcal{C}_\alpha$: $\mathcal{M}(\alpha)$ (unknown)
- **Goal:** learn the $2^d - 1$-dimensional representation (potentially sparse)

$$\mathcal{M} = \Big(\mathcal{M}(\alpha)\Big)_{\alpha \subset \{1,\ldots,d\}, \alpha \neq \emptyset}$$

- $\mathcal{M}(\alpha) > 0 \iff$
  features $j \in \alpha$ may be large together while the others are small.

Sparsity in real datasets

Data=50 wave direction from buoys in North sea.
(Shell Research, thanks J. Wadsworth)



	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 2nd thresholding	1	26

Theoretical guarantees - Results

Theorem

Suppose $\mathcal{G}$ is of finite VC dimension $V_{\mathcal{G}}$ and set

$$\psi_k(\delta) = \sqrt{\frac{d}{k}}\left\{2\sqrt{V_{\mathcal{G}}\log(dk+1)} + 3\sqrt{\log(1/\delta)}\right\}.$$

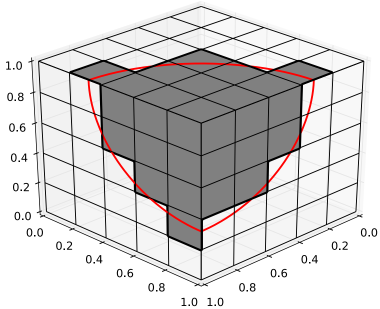
Then, with probability at least $1-\delta$, we have:

$$\Phi_{n/k}(\widehat{\Omega}_\alpha) \geq \alpha - 2\psi_k(\delta) \text{ and } \lambda_d(\widehat{\Omega}_\alpha) \leq \inf_{\Omega\in\mathcal{G},\Phi(\Omega)\geq\alpha} \lambda_d(\Omega)$$

- The learning rate is of order $O_{\mathbb{P}}(\sqrt{(\log k)/k})$
- Main tool: VC inequality for **small probability classes** (Goix, Sabourin & Clémençon 2015)
- The rank transformation **does not damage** the rate
- Oracle inequalities for **model selection** (choice of $\mathcal{G}$) by additive complexity penalization can be straightforwardly derived

Example: paving the sphere

- Let $J \geq 1$. Consider the partition of $\mathbb{S}_{d-1}$ made of $\mathcal{J} = dJ^{d-1}$ 'hypecubes' of same volume



- The class $\mathcal{G}_J$ is made of all possible unions of such hypercubes S_j, $|\mathcal{G}_J| = \exp(dJ^{d-1} \log 2)$

Example: paving the sphere

Algorithm

1. Sort the S_j's so that

$$\widehat{\Phi}_{n,k}(S_{(1)}) \geq \ldots \geq \widehat{\Phi}_{n,k}(S_{(\mathcal{J})})$$

2. Bind together the subsets with largest mass

$$\widehat{\Omega}_{J,\alpha} = \bigcup_{j=1}^{\mathcal{J}(\alpha)} S_{(j)},$$

where $\mathcal{J}(\alpha) = \min\{j \geq 1 : \sum_{l=1}^{j} \widehat{\Phi}_{n,k}(S_{(j)}) \geq \alpha - \psi_k(\delta)\}$

Application to Anomaly Detection

Anomalies correspond to observations

with directions lying in a region where the angular density takes low values

or

with very large sup norm

$\Rightarrow$ abnormal regions are of the form

$$\{(r,\theta):\ \phi(\theta)/r^2 \le s_0\}$$

Define $\widehat{s}((r(\mathbf{V}),\theta(\mathbf{V})) = (1/r(\mathbf{V})^2)\widehat{s}_\theta(\theta(\mathbf{V}))$, where

$$\widehat{s}_\theta(\theta) = \sum_{j=1}^{\mathcal{J}} \widehat{\Phi}_{n,k}(S_j)\mathbb{I}\{\theta \in S_j\}$$

Preliminary Numerical Experiments

UCI machine learning repository
First results on real datasets are encouraging

Table: ROC-AUC

Data set	**OCSVM**	**Isolation Forest**	**Score $\hat{s}$**
shuttle	0.981	0.963	**0.987**
SF	0.478	0.251	**0.660**
http	**0.997**	0.662	0.964
ann	0.372	**0.610**	0.518
forestcover	0.540	0.516	**0.646**

References

- N. Goix, A. Sabourin, S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study, COLT 2015
- N. Goix, A. Sabourin, S. Clémençon. Sparse representations of multivariate extremes with applications to anomaly detection, JMVA 2017
- S. Clémençon and A. Thomas. Mass Volume Curves and Anomaly Ranking. Preprint, https://arxiv.org/abs/1705.01305.
- A. Thomas, S. Clémençon, A. Gramfort, and A. Sabourin. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In AISTATS 2017
- A. Sabourin, S. Clémençon. Nonasymptotic bounds for empirical estimates of the angular measure of multivariate extremes. Preprint