Anomaly Detection:
A Machine-Learning View

Stephan Clémençon

Chaire ML4BGD - Télécom ParisTech

The Many Faces of Anomaly Detection

Anomaly Detection (AD)

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

What is Anomaly Detection ?

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

Synonyms of ”Anomalies”

Outliers
Discordant observations
Exceptions
Aberrations
Peculiarities

Huge number of applications: predictive maintenance, network intrusions, fraud detection, insurance, finance,...

Machine Learning Approaches

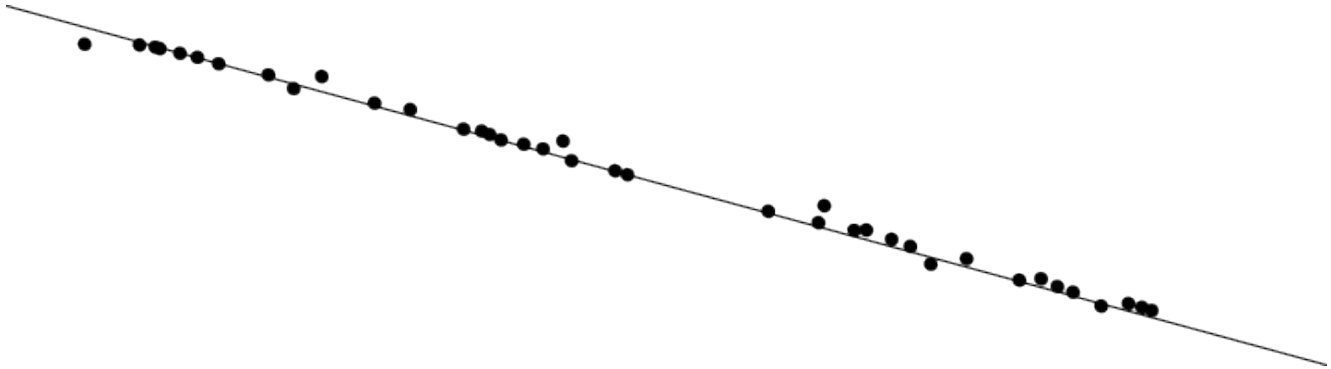
Different types of Anomaly Detection

- **Supervised** AD
  - Labels available for both normal data and anomalies
  - Similar to rare class mining

- **Semi-supervised** AD
  - Only normal data available to train
  - The algorithm learns on normal data only

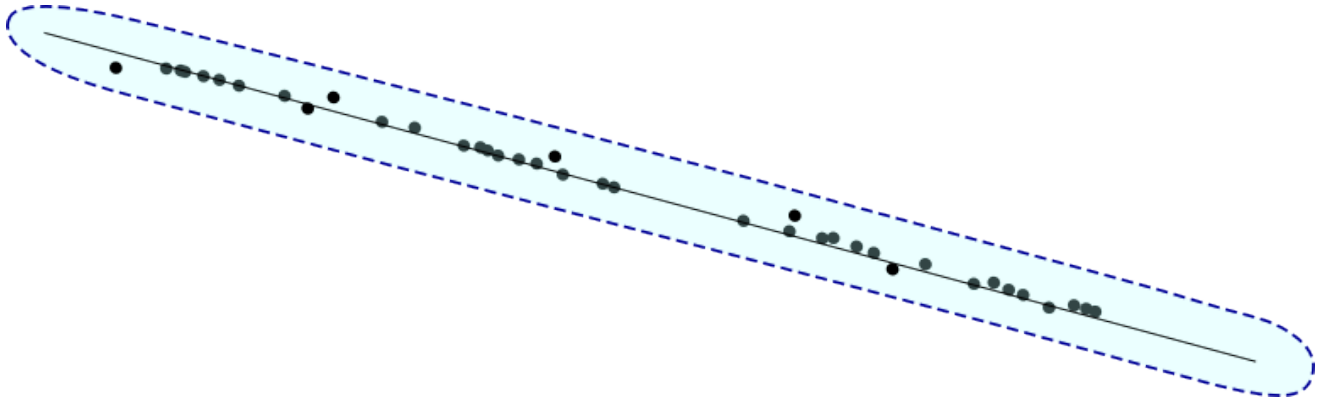
- **Unsupervised** AD
  - no labels, training set = normal + abnormal data
  - Assumption: anomalies are very rare

Anomaly Detection Schemes



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

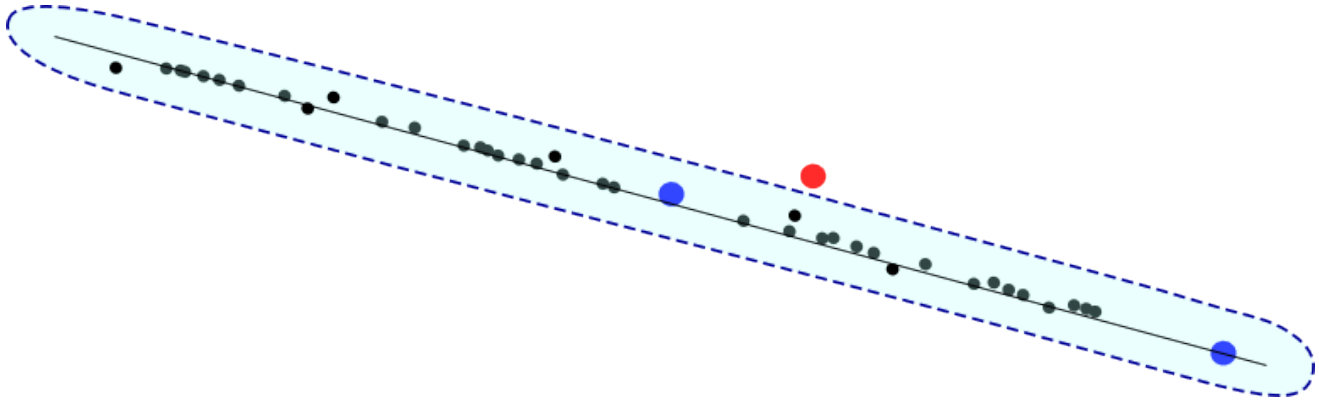
Anomaly Detection Schemes



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

- **Step 2:** Use the ”normal” profile to build a **decision function**.

Anomaly Detection Schemes



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

Different tools in Anomaly Detection:

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

A first go: supervised anomaly detection

Supervised setup

- (X, Y) random pair, valued in $\mathbb{R}^d \times \{-1, +1\}$ with $d >> 1$
  A positive label '$Y = +1$' is assigned to anomalies.

Supervised setup

- (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)

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

Supervised setup

- (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 **order** new data $X'_1, \ldots, X'_{n'}$

$$X'_7 \quad X'_{n'-2} \quad X'_3 \quad X'_6 \quad \ldots$$

Supervised setup

- (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 **order** new data $X'_1, \ldots, X'_{n'}$

$$\begin{array}{cccc} X'_7 & X'_{n'-2} & X'_3 & X'_6 & \ldots \\ + & + & - & + & \ldots \end{array}$$

in order to recover **anomalies on top of the list** with large probability

Supervised AD as Bipartite Ranking

- Exactly the same setup as **binary classification**...

Supervised AD as Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!

Supervised AD as Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** churn, credit-scoring, medical diagnosis, anomaly detection, information retrieval, target marketing, *etc.*

Supervised AD as Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** churn, credit-scoring, medical diagnosis, anomaly detection, information retrieval, target marketing, *etc.*
- **Goals**:
  - Quantitative functional performance criterion for the ranking problem

Supervised AD as Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** churn, credit-scoring, medical diagnosis, anomaly detection, information retrieval, target marketing, *etc.*
- **Goals**:
  - Quantitative functional performance criterion for the ranking problem
  - Design learning algorithms optimizing the ranking criterion
  - Develop dedicated **statistical learning theory** using approximation theory

Supervised AD as Bipartite Ranking

- Exactly the same setup as **binary classification**...
- ... except the nature of the problem is **global**!
- **Applications:** churn, credit-scoring, medical diagnosis, anomaly detection, information retrieval, target marketing, *etc.*
- **Goals**:
  - Quantitative functional performance criterion for the ranking problem
  - Design learning algorithms optimizing the ranking criterion
  - Develop dedicated **statistical learning theory** using approximation theory

Bipartite Ranking

- Same data, different questions:

  Classifying is a **local** task, while ranking is **global**!

- Ranking and scoring a set of instances

Bipartite Ranking

- Same data, different questions:

  Classifying is a **local** task, while ranking is **global**!

- Ranking and scoring a set of instances
  ... through a **scoring function** $s : \mathcal{X} \to \mathbb{R}$

- **Challenge:** develop theory and algorithms

- **Question:** are advances in classification of any use?

Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$

Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$
- **Want to:** rank $X_1, \ldots, X_n$ through a scoring function $s : \mathcal{X} \to \mathbb{R}$, so that a large number instances with $+1$ labels appear on the top with high probability

Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$
- **Want to:** rank $X_1, \ldots, X_n$ through a scoring function $s : \mathcal{X} \to \mathbb{R}$, so that a large number instances with $+1$ labels appear on the top with high probability
- **Class of solutions:**

$$\mathcal{S}^* = \{T \circ \eta \mid T : [0, 1] \to \mathbb{R} \text{ increasing}\}$$

Rigorous problem statement

- **Data:** $(X_1, Y_1), \ldots, (X_n, Y_n) \in \left(\mathcal{X} \times \{-1, +1\}\right)^{\otimes n}$
- **Want to:** rank $X_1, \ldots, X_n$ through a scoring function $s : \mathcal{X} \to \mathbb{R}$, so that a large number instances with $+1$ labels appear on the top with high probability
- **Class of solutions:**

$$\mathcal{S}^* = \{T \circ \eta \mid T : [0, 1] \to \mathbb{R} \text{ increasing}\}$$

- **Need to:** find an optimization criterion reflecting performance

ROC Curve and AUC

ROC Curve and AUC

- True positive rate:

$$\mathrm{TPR}_s(t) = \mathbb{P}(s(X) \geq t \mid Y = 1)$$

- False positive rate:

$$\mathrm{FPR}_s(t) = \mathbb{P}(s(X) \geq t \mid Y = -1)$$

ROC Curve and AUC

- True positive rate:

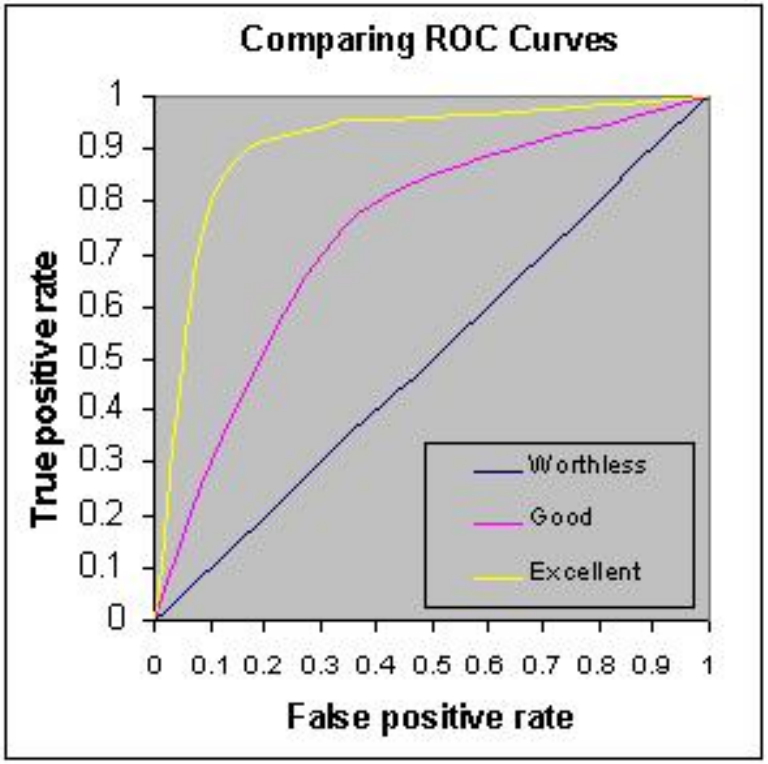
$$\mathrm{TPR}_s(t) = \mathbb{P}\left(s(X) \geq t \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(t) = \mathbb{P}\left(s(X) \geq t \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $t \mapsto \left(\mathrm{FPR}_s(t), \mathrm{TPR}_s(t)\right)$

ROC Curve and AUC



- True positive rate:

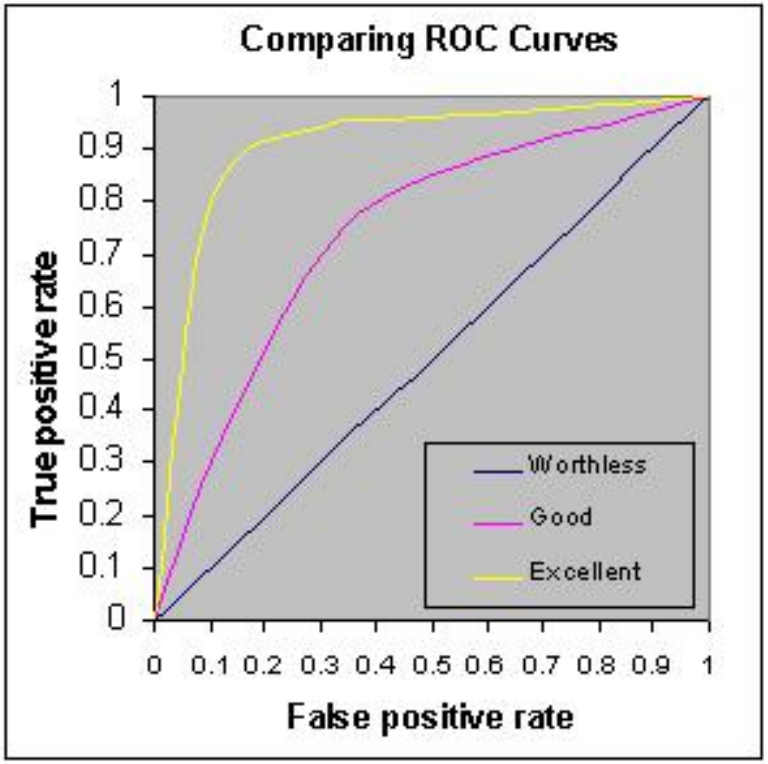
$$\mathrm{TPR}_s(t) = \mathbb{P}\left(s(X) \geq t \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(t) = \mathbb{P}\left(s(X) \geq t \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $t \mapsto \left(\mathrm{FPR}_s(t), \mathrm{TPR}_s(t)\right)$

ROC Curve and AUC



- True positive rate:

$$\mathrm{TPR}_s(t) = \mathbb{P}\left(s(X) \geq t \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(t) = \mathbb{P}\left(s(X) \geq t \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $t \mapsto \left(\mathrm{FPR}_s(t), \mathrm{TPR}_s(t)\right)$

AUC = Area Under an ROC Curve

Performance - The AUC summary criterion

- The L_1-**metric** is a convenient distance in the ROC space:

$$\min_s \int_{\alpha=0}^{1} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}\, d\alpha = \mathrm{AUC}^* - \max_s \mathrm{AUC}(s),$$

where the **area under the** ROC **curve** is defined by

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \mathrm{ROC}(s,\alpha) d\alpha$$

and $\mathrm{AUC}^* = \mathrm{AUC}(s^*)$ for $s \in \mathcal{S}^*$.

Performance - The AUC summary criterion

- The L_1-**metric** is a convenient distance in the ROC space:

$$\min_s \int_{\alpha=0}^{1} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s,\alpha)\}\, d\alpha = \mathrm{AUC}^* - \max_s \mathrm{AUC}(s),$$

where the **area under the** ROC **curve** is defined by

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \mathrm{ROC}(s,\alpha) d\alpha$$

and $\mathrm{AUC}^* = \mathrm{AUC}(s^*)$ for $s \in \mathcal{S}^*$.

- **Probabilistic interpretation:** If $s(X)$ is a continuous r.v., then

$$\begin{aligned} \mathrm{AUC}(s) &= \mathbb{P}\{s(X) > s(X') \mid Y = 1, Y' = -1\} \\ &= \frac{1}{2p(1-p)} \mathbb{P}\{(s(X) - s(X'))(Y - Y') > 0\} \ . \end{aligned}$$

AUC maximization and Pairwise Classification

- A natural estimate of the ranking performance

$$U(s) = \mathbb{P}\{(s(X) - s(X'))(Y - Y') > 0\}$$

is the U-**statistic** given by

$$\widehat{U}_n(s) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\{s(X_i) - s(X_j))(Y_i - Y_j) > 0\}$$

(rate of concording pairs = Kendall's association coefficient)

AUC maximization and Pairwise Classification

- A natural estimate of the ranking performance

$$U(s) = \mathbb{P}\{(s(X) - s(X'))(Y - Y') > 0\}$$

is the U-**statistic** given by

$$\widehat{U}_n(s) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \mathbb{I}\{s(X_i) - s(X_j))(Y_i - Y_j) > 0\}$$

(rate of concording pairs = Kendall's association coefficient)

- **A wide collection of algorithms**: SVMRank, LambdaRank, RankBoost, TreeRank, Ranking Forest, *etc.*

Scalability of pairwise classification

- Training sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
- Computing the empirical risk (gradient) requires to average over $O(n^2)$ pairs!
- Is there a smarter strategy than minimizing the risk using mini-batches?

How to apply the ERM paradigm to Massive Data?

- Suppose that n is too large to evaluate the empirical risk $L_n(g)$
- Common sense: run your preferred learning algorithm using a subsample of "reasonable" size $B << n$, *e.g.* by drawing with replacement in the original training data set...

How to apply the ERM paradigm to Massive Data?

- Suppose that n is too large to evaluate the empirical risk $L_n(g)$
- Common sense: run your preferred learning algorithm using a subsample of "reasonable" size $B << n$, *e.g.* by drawing with replacement in the original training data set...
- ... but of course, statistical performance is **downgraded**

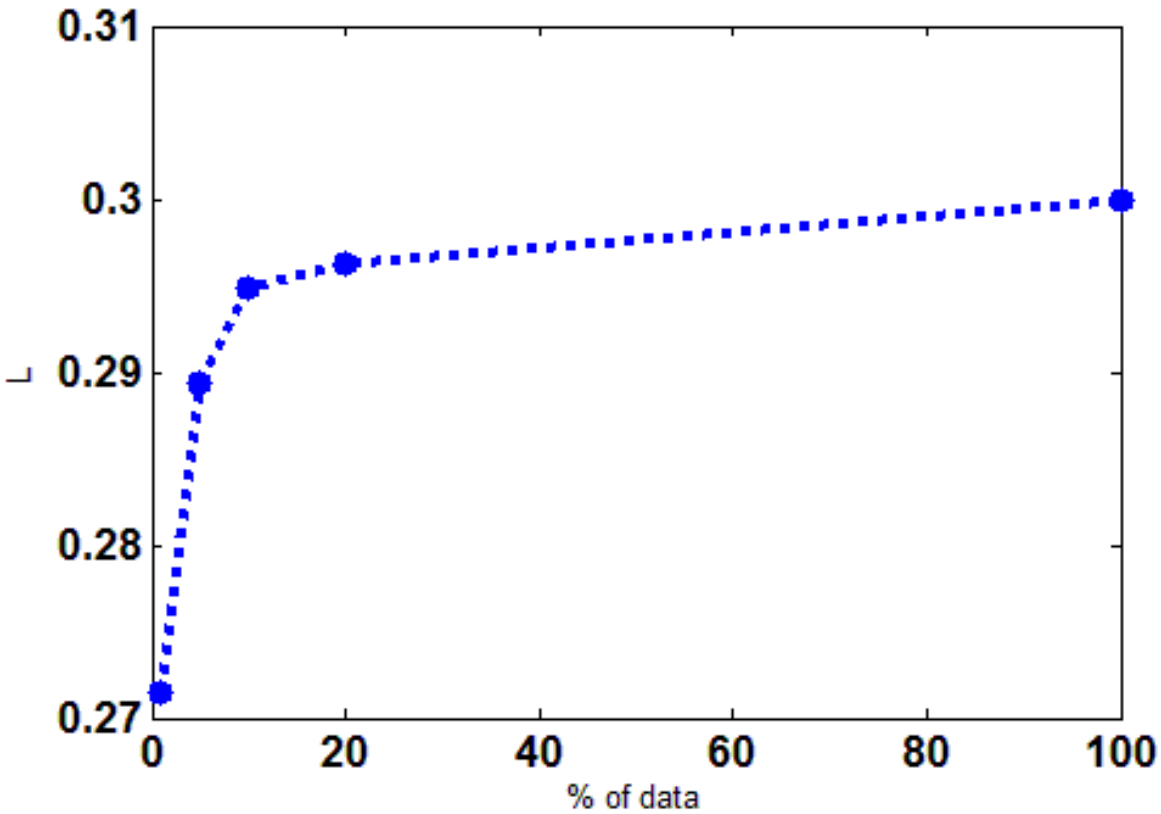
$$1/\sqrt{n} \; << \; 1/\sqrt{B}$$

Sampling pairs rather than individual observations

- Empirical risk sampling with $B = O(n)$ pairs yields a rate bound of the order $O(\sqrt{\log n / n})$
- One suffers **no loss** in terms of learning rate, while **drastically reducing computational cost** - Clémençon *et al.* (2016)

Example: AUC maximization

Empirical ranking performance for SVMRANK based on 1%, 5%, 10%, 20% and 100% of the "LETOR 2007" dataset.

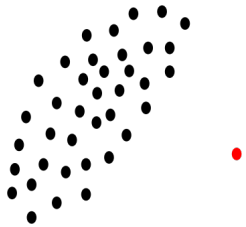


Unsupervised Anomaly Detection

Anomaly/Outlier/Novelty

Definition [Hawkins, 1980]

"An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"



From this definition:

- An anomaly does not neccesarily mean a failure of the system: many different outcomes possible
- An anomaly is likely to contain useful information

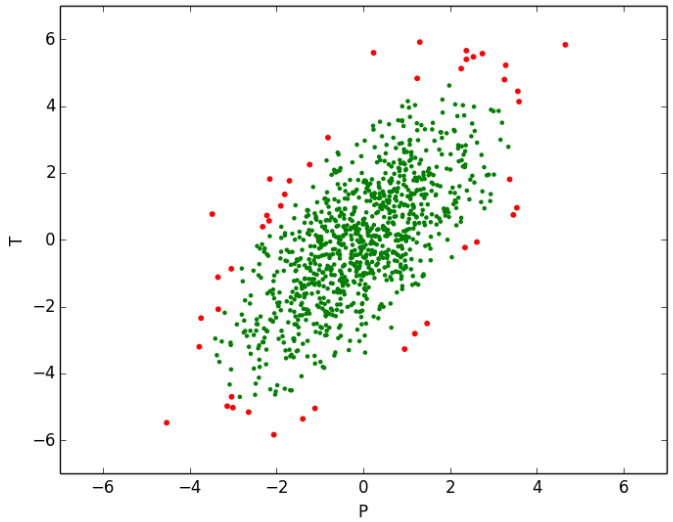
Our goal: develop models to detect abnormal observations

Anomaly detection

Goal

Detect abnormal observations in a given data set $X_1, \ldots, X_n \in \mathbb{R}^d$

- Labelled dataset (normal and abnormal observations)
  Supervised anomaly detection
  - Imbalanced dataset
  - Novelty detection?

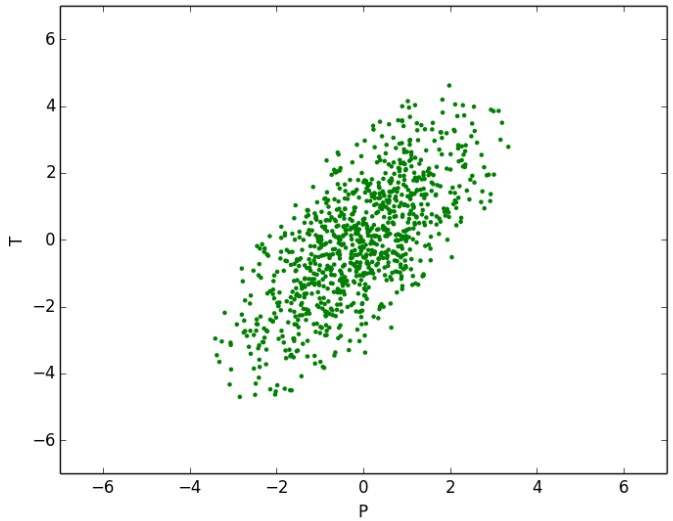


Anomaly detection

Goal

Detect abnormal observations in a given data set $X_1, \ldots, X_n \in \mathbb{R}^d$

- Labelled dataset (normal and abnormal observations)
  Supervised anomaly detection
  - Imbalanced dataset
  - Novelty detection?
- Normal data available
  Semi-supervised anomaly detection

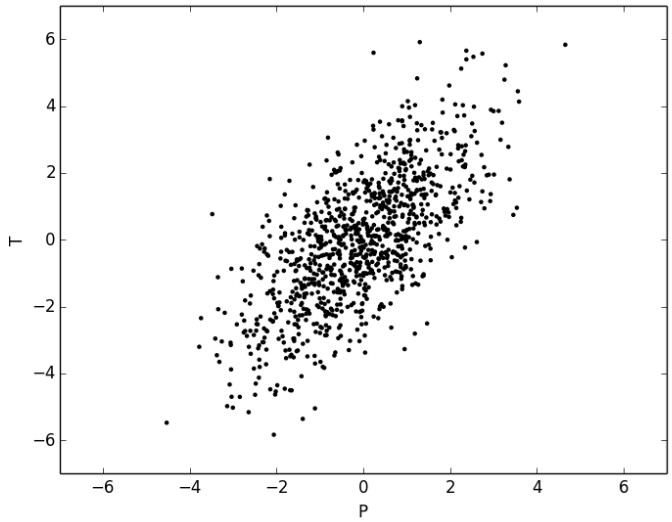


Anomaly detection

Goal

Detect abnormal observations in a given data set $X_1, \ldots, X_n \in \mathbb{R}^d$

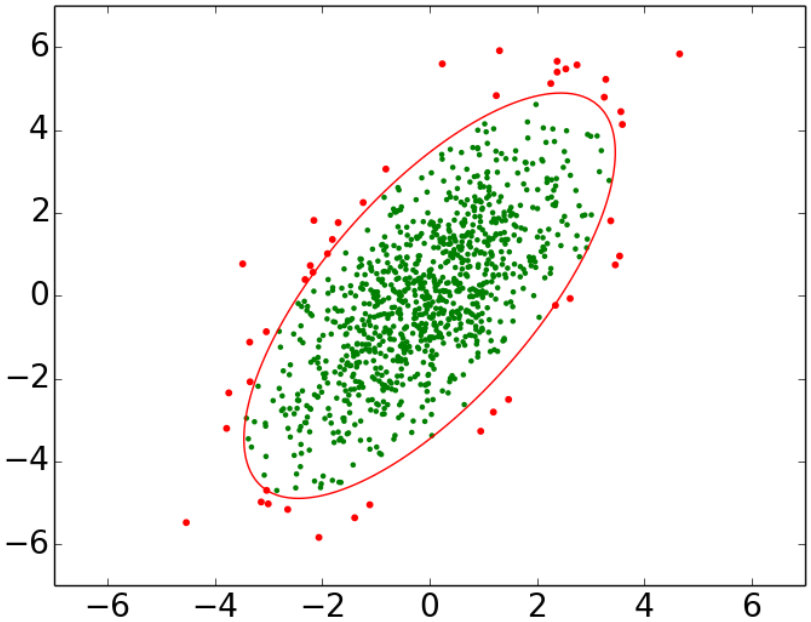
- Labelled dataset (normal and abnormal observations)
  Supervised anomaly detection
  - Imbalanced dataset
  - Novelty detection?
- Normal data available
  Semi-supervised anomaly detection
- Unlabelled dataset
  Unsupervised anomaly detection
  - Assumption: anomalies = rare events



Unsupervised anomaly detection

$X_1, \ldots, X_n \in \mathbb{R}^d$ i.i.d. realizations of unknown probability measure $\mathbb{P}$

- Anomalies are supposed to be rare events, located in the tail of the distribution
- Estimation of the region where the data are most concentrated: region of minimum volume regions for a given probability content α close to 1



Minimum Volume set, $\alpha = 0.95$

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

Definition [Polonik, 1997]

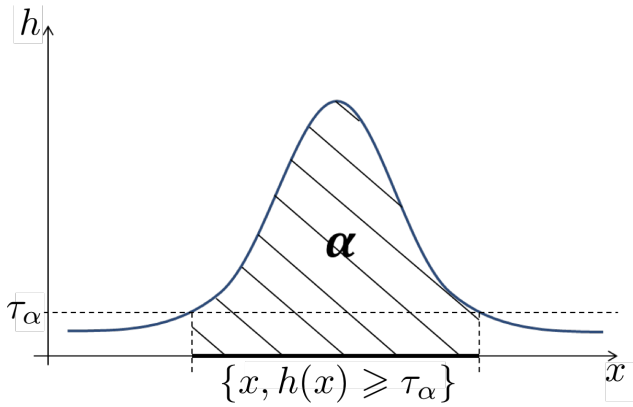
- $\alpha \in [0, 1]$ (for anomaly detection α is close to 1)
- $\mathcal{C}$ class of measurable sets
- $\mu(dx)$ unknown probability measure of the observations
- λ Lebesgue measure

$$Q(\alpha) = \underset{C \in \mathcal{C}}{\text{argmin}} \{\lambda(C), \mathbb{P}(X \in C) \geqslant \alpha\}$$

- 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**

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 $\widehat{\mu}_n(G) \geqslant \alpha - \phi_n$, 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 Empirical Risk Minimization

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

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

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

$$\widehat{k} = \operatorname*{argmin}_k \left\{ \lambda(\widehat{G}_k) + 2\phi_k : \ \widehat{\mu}_n(\widehat{G}_k) \geqslant \alpha - \phi_k \right\}$$

Theoretical MV sets

Consider the following assumptions:

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

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

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

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

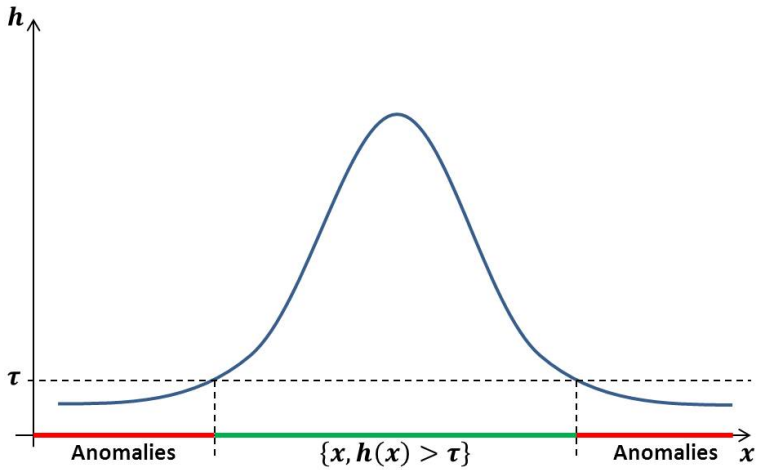
Unsupervised anomaly detection - Mass Volume curves

- 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 $h(x)$



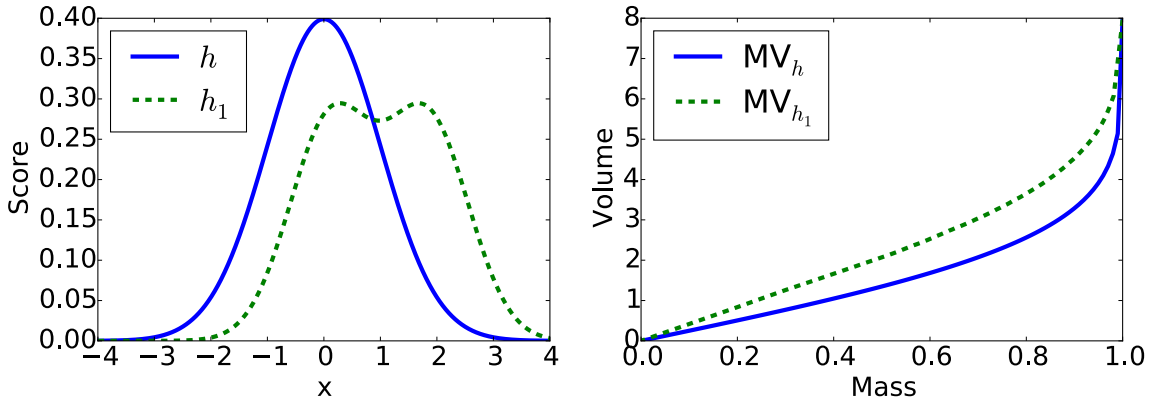
Mass Volume curve

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

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

$$MV_s : \alpha \in (0,1) \mapsto \lambda_s(\alpha_s^{-1}(\alpha)) = \lambda(\{x, s(x) \geqslant \alpha_s^{-1}(\alpha)\})$$

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,

$$\forall \alpha \in (0,1), \quad MV^*(\alpha) \leqslant MV_s(\alpha)$$

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

Algorithms for Anomaly Detection

Methods

- Plug-in techniques
- Turning unsupervised AD into binary classification
- Histograms
- Decision trees
- SVM
- Isolation Forest

Plug-in techniques

- MV sets are **density level sets**:

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

- **Naive approach**: *'2-split' trick*

  1. Compute a *density estimator* $\widehat{h}(x)$ based on $X_1, \ldots, X_n$
  2. Based on an extra sample $X'_1, \ldots, X'_{n'}$ (independent from $X_1, \ldots, X_n$), compute the *empirical quantile* $\widehat{h}_n(X'_{(\lfloor n(1-\alpha) \rfloor)})$, where

$$\widehat{h}(X'_{(1)}) \geqslant \widehat{h}(X'_{(2)}) \geqslant \cdots \geqslant \widehat{h}(X'_{(n')})$$

  3. Output: $\widehat{G}_\alpha^* = \{x : \ \widehat{h}_n(x) \geqslant \widehat{h}_n(X'_{(\lfloor n\alpha \rfloor)})\}$

Plug-in techniques

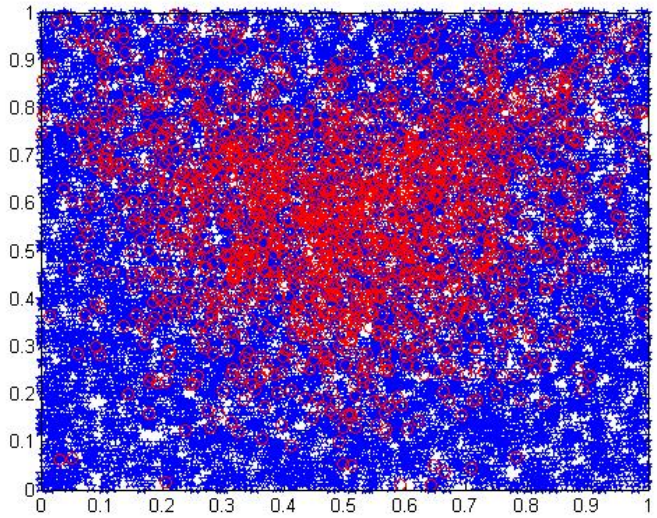
- Many methods are available for density estimation:
  - Parametric models (*e.g.* Neural Networks): $\{h_\theta(x);\ \theta \in \Theta\}$

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \log h_\theta(X_i)$$

  - Local averaging:
    - ⋆ Nearest neighbours
    - ⋆ Histograms
    - ⋆ Kernel smoothing
    - ⋆ etc.
- Major drawbacks: curse of dimensionality, overfitting

Turning unsupervised AD into binary classification

- Rather than rescaling the observations, suppose that X takes its values in $[0,1]^d$
- The restriction of λ to $[0,1]^d$ coincides with the **uniform distribution** $\mathcal{U}([0,1]^d)$
- Generate an i.i.d. sample $X_1^-, \ldots, X_m^-$ drawn from $\mathcal{U}([0,1]^d)$ s.t. $n/(n+m) \sim p$ and assign a **negative label** to these data, whereas a **positive label** is assigned to observations in the original sample



Turning unsupervised AD into binary classification

- Solve the binary classification problem based on the pooled labeled training set
- The solution mimics the **Bayes classifier**, which predicts $+1$ on the set

$$\{x \in [0,1]^d : \ h(x) \geqslant (1-p)/p\}$$

  and -1 everywhere else.
- It provides a MV set at level $\alpha = \mathbb{P}(h(X) \geqslant 1/p - 1)$
- In practice α is fixed in advance, p is picked via *gridsearch*

Histograms

- Consider a **compact** feature space, *e.g.* $\mathcal{X} = [0,1]^d$
- Consider a partition $\mathcal{P}$, $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of $\mathcal{X}$ formed of measurable subsets of same volume

$$\lambda(\mathcal{C}_1) = \cdots = \lambda(\mathcal{C}_K)$$

- The class $\mathcal{G}_\mathcal{P}$ is the ensemble composed of unions of $\mathcal{C}_k$
- How to get a solution of

$$\min_{G \in \mathcal{P}:\, \widehat{\mu}_n(G) \geqslant \alpha - \phi} \lambda(G) ?$$

Histograms

- A fast procedure:

  1. For $k = 1, \ldots, K$, compute $\widehat{\mu}_n(\mathcal{C}_k)$
  2. Sort the cells $\mathcal{C}_k$ of the partition by decreasing order of $\widehat{\mu}_n(\mathcal{C}_k)$:

$$\widehat{\mu}_n(\mathcal{C}_{(1)}) \geqslant \ldots \geqslant \widehat{\mu}_n(\mathcal{C}_{(K)})$$

  3. Find

$$\widehat{k} = \text{argmin} \left\{ k : \sum_{i=1}^{k} \widehat{\mu}_n(\mathcal{C}_{(i)}) \geqslant \alpha - \phi \right\}$$

  4. Output:

$$\widehat{G}_\alpha^* = \bigcup_{i=1}^{\widehat{k}} \mathcal{C}_{(i)}$$

Histograms

- Partition $\mathcal{P}$ formed of **hypercubes** of sidelength 2^{-j}

$$\Rightarrow \#\mathcal{G}_{\mathcal{P}} = 2^{2^j d} - 1$$

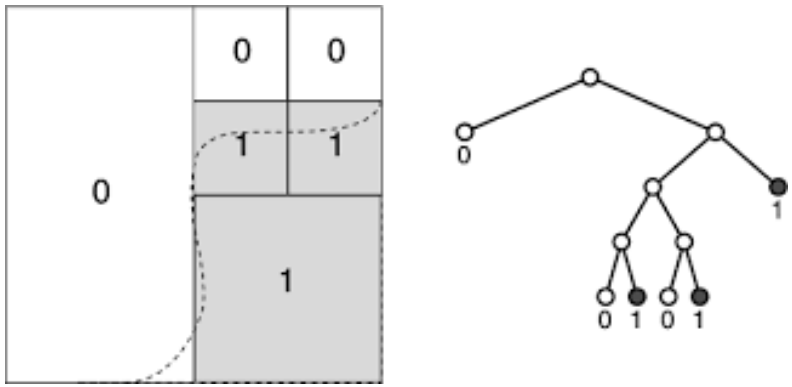
- The **volume** of elements of $\mathcal{G}_{\mathcal{P}}$ can be **explicitely computed**
- The **tolerance level** can be chosen as the upper confidence bound fat level $1 - \delta$ for $\sup_{G \in \mathcal{G}_{\mathcal{P}}} |$ given by (Hoeffding + union bounds)

$$\phi = \sqrt{\frac{\log(2^{2^{jd}} - 1) + \log(2/\delta)}{2n}}.$$

- **Tune** the complexity parameter j using **model selection techniques**
- Drawback: such partitions are **not flexible enough** in practice

Decision trees

- In practice the partition should be determined **based on the data**.
- For simplicity, consider $\mathcal{X} = [0,1]^d$ and **dyadic decision trees** (Donoho, '99): axis-orthogonal dyadic plits. Top-down strategy, start from the root node $(0,0)$: $C_{0,0} = [0,1]^d$. The volume of any cell $C_{j,k}$, $j \geqslant 0$ and $0 \leqslant k \leqslant 2^j - 1$ can be computed in a recursive manner.



Decision trees

- Hyperrectangles with label 1 are subsets of the MV set estimate (labels of two siblings are different). At each internal node $C_{j,k}$, the split leading to siblings $C_{j+1,2k}$ and $C_{j+1;2k+1}$ is characterized by the index $c_{j,k}$ of the coordinate used to minimize **recursively** the volume of the current decision set G under the constraint that

$$\widehat{\mu}_n(G) \geqslant \alpha \ (-\phi).$$

- In practice:
  - splits are not necessarily dyadic
  - in the growing stage, one fixes the maximal depth and/or the minimum of points in a cell
  - The growing step if followed by a **pruning stage**, in order to avoid overfitting

Isolation Forest (Liu, Ting & Zhou, '08)

- Build a **forest of isolation trees**
- **Heuristic:** *Anomalies are more susceptible to isolation under random partitioning*
- An isolation tree is a binary tree built recursively with a top-down strategy (starting from the root node), iterating the following steps:
  1. Choose randomly a *splitting variable* $X^{(i)}$ and a *split value* t
  2. Split the cell $C_{j,k}$ into $C_{j+1,2k} = \{X^{(i)} \leqslant t\}$ and $C_{j+1,2k+1} = \{X^{(i)} > t\}$
  3. Stop when:
     - ⋆ A depth limit is attained
     - ⋆ The observations X are constant on the cell
     - ⋆ The cell is of cardinality 1
- **Path length of an observation** x: number of edges that x traverses from the root node to its terminal node

Isolation Forest (Liu, Ting & Zhou, '08)

- **Path length of an observation x w.r.t. an iTree:** number of edges that x traverses from the root node to its terminal node
- **Anomaly score:** given $B \geqslant 1$ iTrees, $\mathcal{T}_1, \ldots, \mathcal{T}_B$

$$s(x) = 2^{-\frac{\frac{1}{B}\sum_{b=1}^{B} h_b(x)}{c_n}},$$

where $h_b(x)$ is the path length of observation x related to iTree $\mathcal{T}_b$ and c_n is a normalizing constant (full average path length)

- Anomalies are observations with a score close to 1

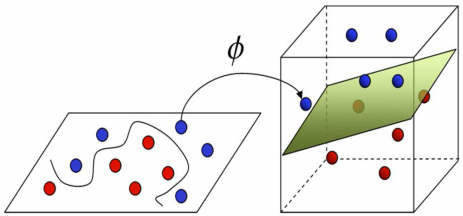
Isolation Forest (Liu, Ting & Zhou, '08)

Python code available at
`https://github.com/ngoix/scikit-learn/tree/iforest`

Beyond Greedy Methods: One-Class SVM

Observations $x_1, \ldots, x_n \in \mathbb{R}^d$

- Map the data in a high dimensional feature space



$$\langle \Phi(x), \Phi(x') \rangle = k(x, x')$$

- Find a separating hyperplane between the mapped data and the origin by solving the optimization problem: $\nu \in (0, 1]$

$$\min_{\mathbf{w}, \xi, \rho} \quad \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad \langle \mathbf{w}, \Phi(x_i) \rangle \geqslant \rho - \xi_i, \quad \xi_i \geqslant 0 \qquad 1 \leqslant i \leqslant n$$

Convex problem with tractable solution

We use the Gaussian kernel:

$$k_\sigma(x, x') = \exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right), \quad \sigma > 0$$

One-Class SVM

Observations $x_1, \ldots, x_n \in \mathbb{R}^d$

1. The user needs to choose $\nu \in (0, 1]$ and $\sigma > 0$
2. Optimization problem is solved
3. Solution

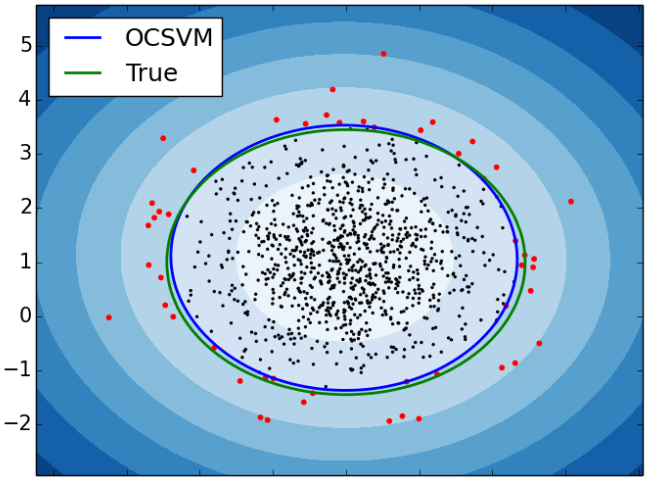
$$f_\sigma(x) = \sum_{i=1}^{n} \alpha_i k_\sigma(x, x_i) \qquad \text{Support Vectors (SV): } \alpha_i > 0$$

offset ρ_ν

$$\hat{C} = \{x, f_\sigma(x) \geqslant \rho_\nu\}$$

Outliers: $\{x_i, f_\sigma(x_i) < \rho_\nu\}$

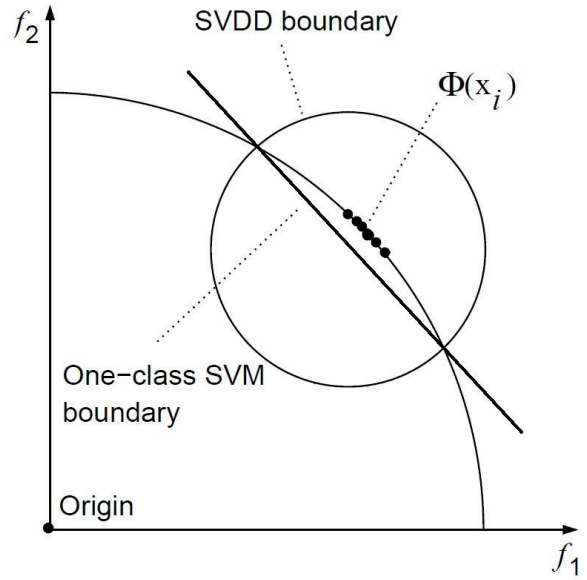
SV: $\{x_i, f_\sigma(x_i) \leqslant \rho_\nu\}$



Understanding the connection with MV sets

Support Vector Data Description (SVDD) [Tax et al., 2004]: find the smallest ball that contains an arbitrary proportion of data

- With a gaussian kernel, all the x_i have the same norm in H and lie in the same orthant

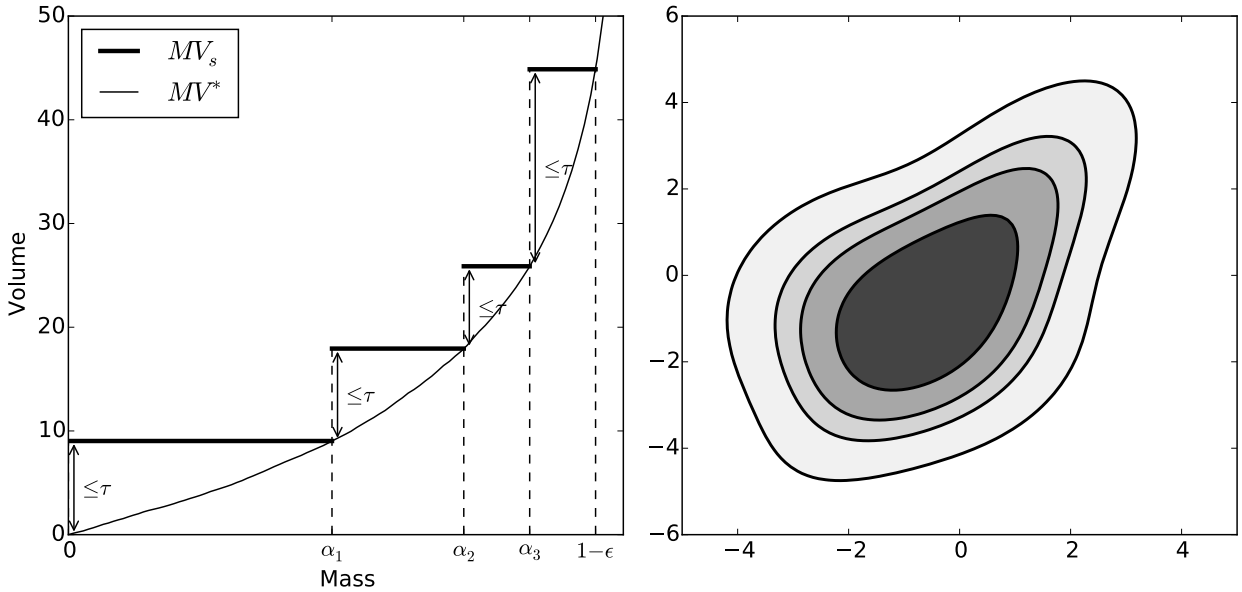


SVDD = OCSVM

One can show that the optimization problems are equivalent if the Gaussian kernel is used.

Mass Volume curve minimization: a continuum of MV-set problems

The A-Rank algorithm: an adaptive discretization
Clémençon & Thomas (2017)



Dimension reduction in multivariate extremes
Application to anomaly detection

Goix, Sabourin Clémençon (2017)

Dimension reduction in multivariate extremes

Exhibit sparsity?

Anomaly detection in 'extreme' data

'Extremes' = points located in the tail of the distribution.

What does 'normal' mean among extremes?

Goix, Sabourin Clémençon (2017)

Dimension reduction in multivariate extremes

Exhibit sparsity?

Error ('uncertainty') assessment

Bounds on the error?

Anomaly detection in 'extreme' data

'Extremes' = points located in the tail of the distribution.

What does 'normal' mean among extremes?

Multivariate EVT for Anomaly detection

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

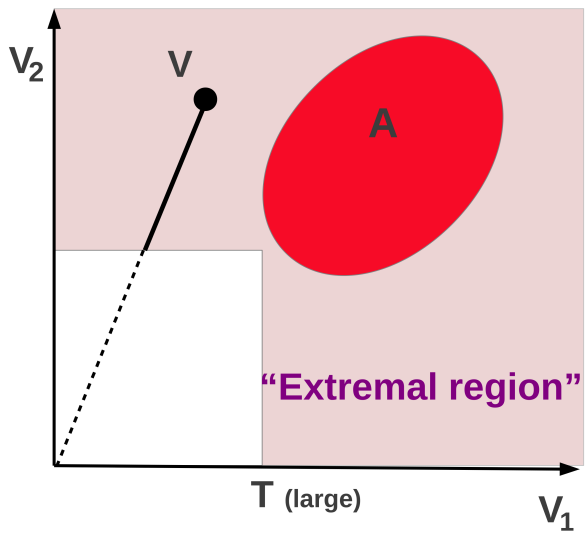
How to distinguish between large anomalies and normal extremes?

- Yet: no multivariate AD algorithm has a specific treatment for multivariate extreme data

- **Our goal** (from an AD point of view): Improve performance of standard AD algorithms on extremal regions using MEVT.
  $\rightarrow$ reduce # false positives

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** $V_j = \frac{1}{1-F_j(X_j))}$, $\mathbb{P}(V_j > v) = \frac{1}{v}$.
- 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

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

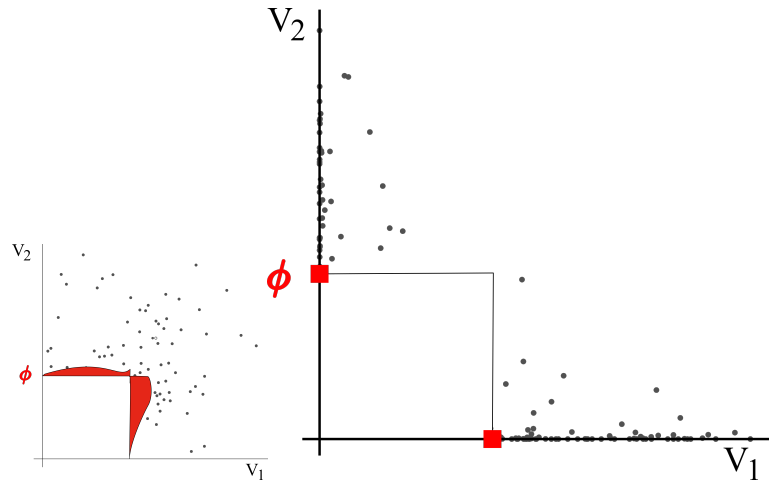
necessarily: $\mu(tA) = t^{-1}\mu(A)$ (Radial homogeneity)
$\rightarrow$ **angular measure** on the sphere : $\Phi(B) = \mu\{tB, t \geq 1\}$

General model for extremes

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

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.

Multivariate extremes in large dimension?

- Flexible multivariate models for **moderate dimension** ($d \simeq 5$)

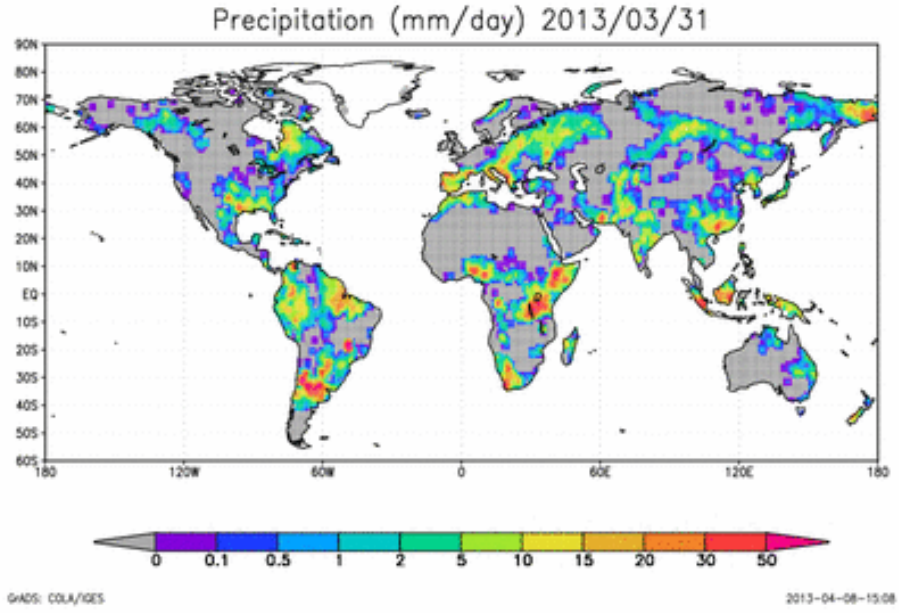
  Dirichlet Mixtures (Boldi,Davison 07; S., Naveau 12), Logistic family (Stephenson 09, Fougères *et.al*, 13), Pairwise Beta (Cooley *et.al*) ...

- Theory for angular measure (dependence) estimation: **asymptotic, $\mathbf{d = 2}$**, rates under **second order conditions**

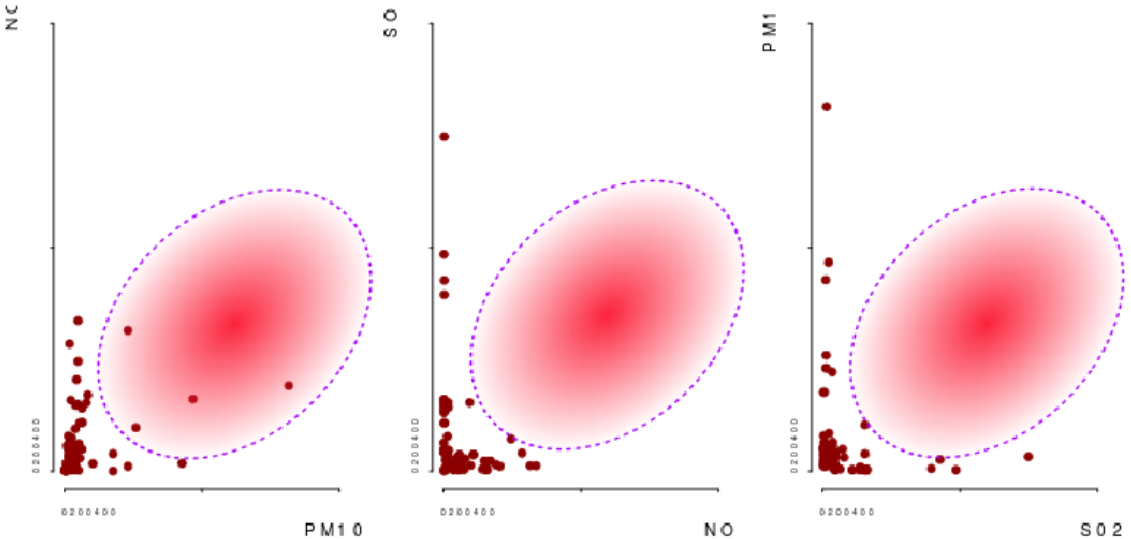
  (Einmahl, 01) Empirical likelihood (Einmahl, Segers 09)

- **High dimension?** ($d \gg 1$):
  - Spatial $\rightarrow$ max-stable models (parametric)
  - **Non spatial $\rightarrow$??**
    (multiple air pollutants, assets, features for AD ...)
  - Theory for integrated versions (tail dependence function)
    Asymptotic normality (Einmahl *et. al.*, 12, 15) (parametric case),
    Finite sample bounds (Goix *et. al*, 15)
    $\not\rightarrow$ structure of extremes (which components may be large together)

It cannot rain everywhere at the same time



(daily precipitation)



(air pollutants)

Towards high dimension

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

- **Our goal** from a MEVT point of view:

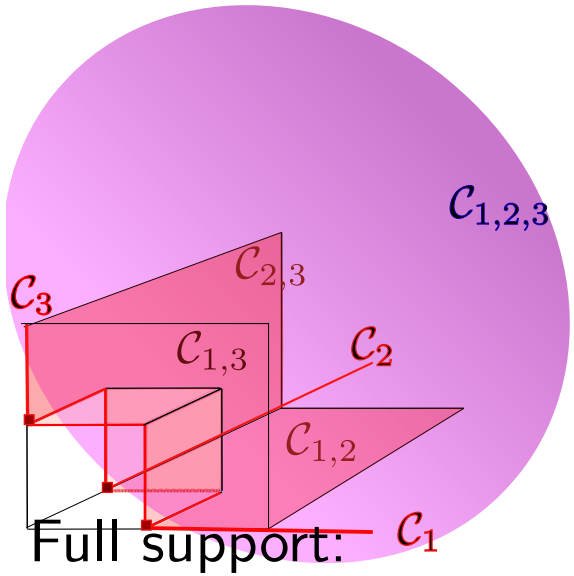
Estimate the (sparse) support of the angular measure (*i.e.* the dependence structure).

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

  - For MEVT modeling: recover the asymptotically dependent groups of components $\rightarrow$ use simplified model.

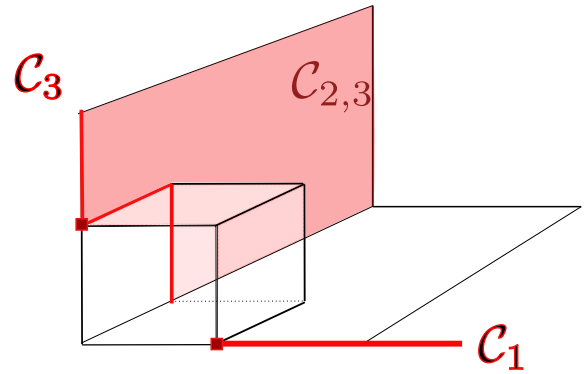
  - for AD: support = normal profile
    $\rightarrow$ anomalies = points 'far away' from the support.

Sparse angular support



Full support:

anything may happen



Sparse support

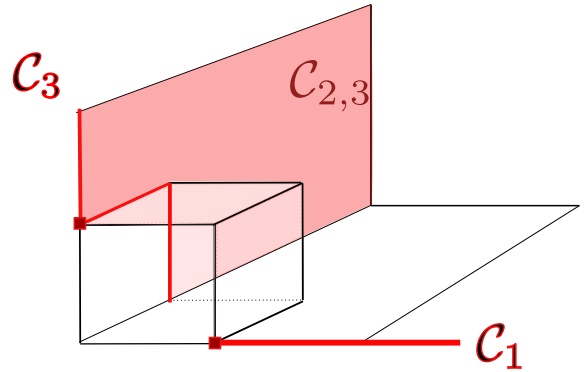
(V_1 not large if V_2 or V_3 large)

Where is the mass?

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

$\alpha \subset \{1, \ldots, d\}$.

Support recovery + representation



- $\{\Omega_\alpha, \alpha \subset \{1,\dots,d\}$: partition of the unit sphere
- $\{\mathcal{C}_\alpha, \alpha \subset \{1,\dots,d\}$: 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)

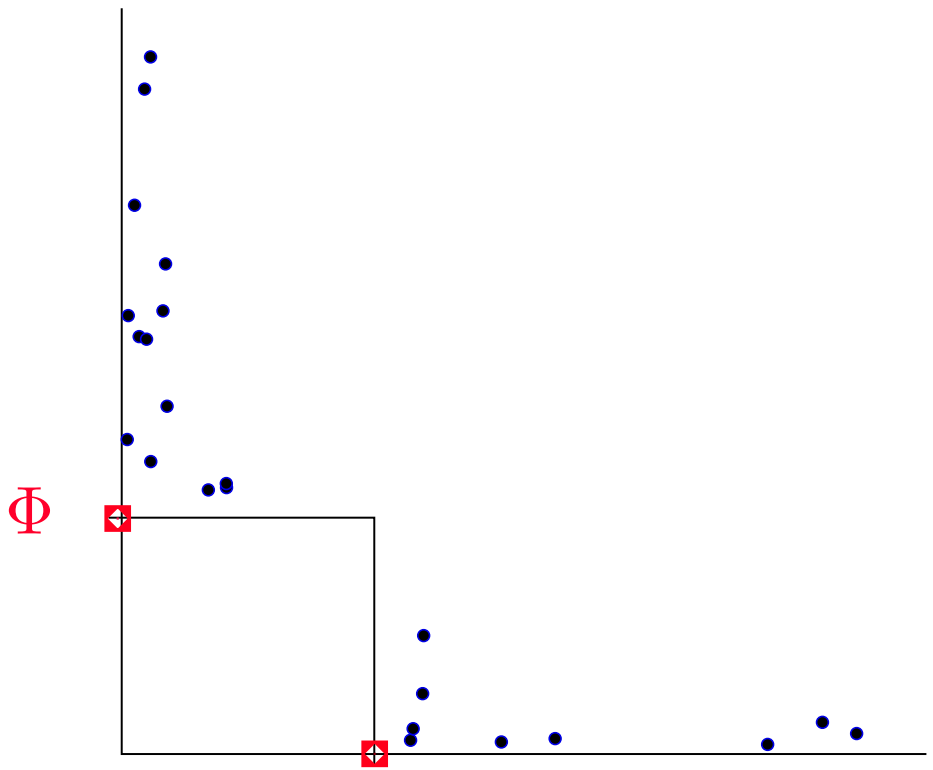
$$\mathcal{M} = \Big(\mathcal{M}(\alpha)\Big)_{\alpha \subset \{1,\dots,d\}, \alpha \neq \emptyset}$$

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

Identifying non empty edges

Issue: real data = non-asymptotic: $V_j > 0$.

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

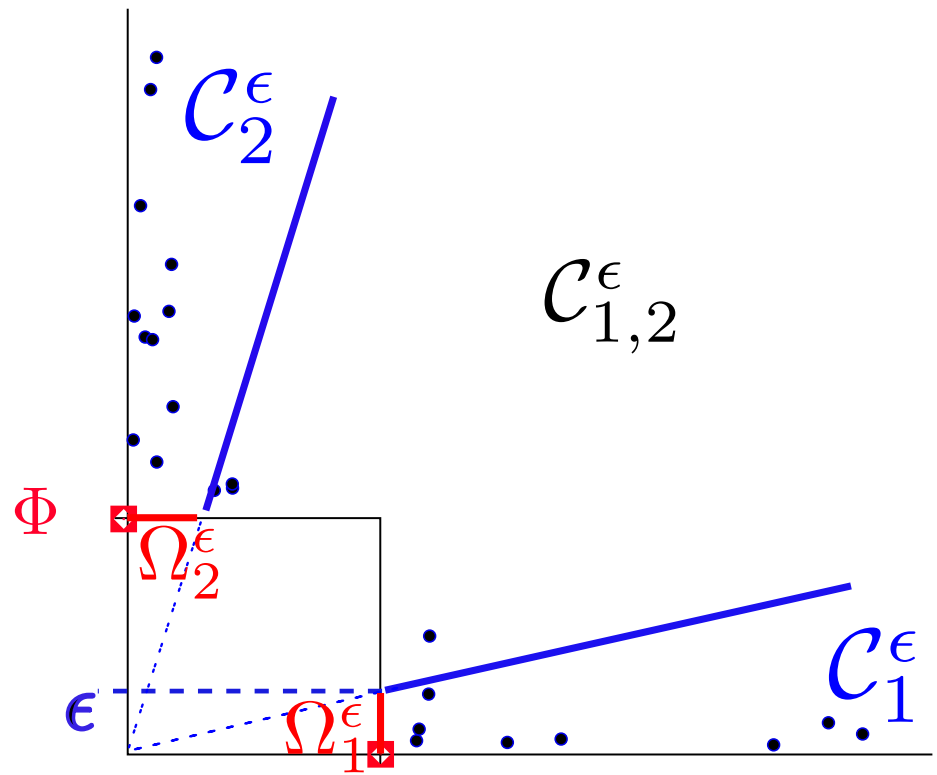


Identifying non empty edges

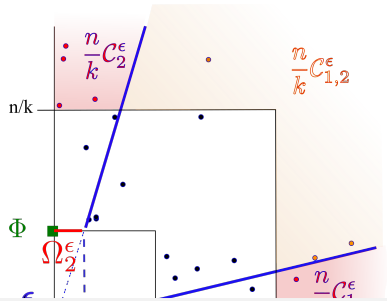
Fix $\varepsilon > 0$. Affect data ε-close to an edge, to that edge.

$$\Omega_\alpha \to \Omega_\alpha^\varepsilon = \{w : \quad w_i > \varepsilon\,(i \in \alpha), \quad w_j < \varepsilon\ (j \notin \alpha)\}.$$
$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\varepsilon = \{t\ \Omega_\alpha^\varepsilon, t \geq 1\}.$$

$\to$ New partition of the input space, compatible with non asymptotic data.



Empirical estimator: Counts the standardized points in $\mathcal{C}_\alpha^\varepsilon$, far from 0.



The DAMEX algorithm

data: $\mathbf{X}_i, i = 1, \ldots, n$, $\mathbf{X}_i = (V_{i,1}, \ldots, X_{i,d})$.

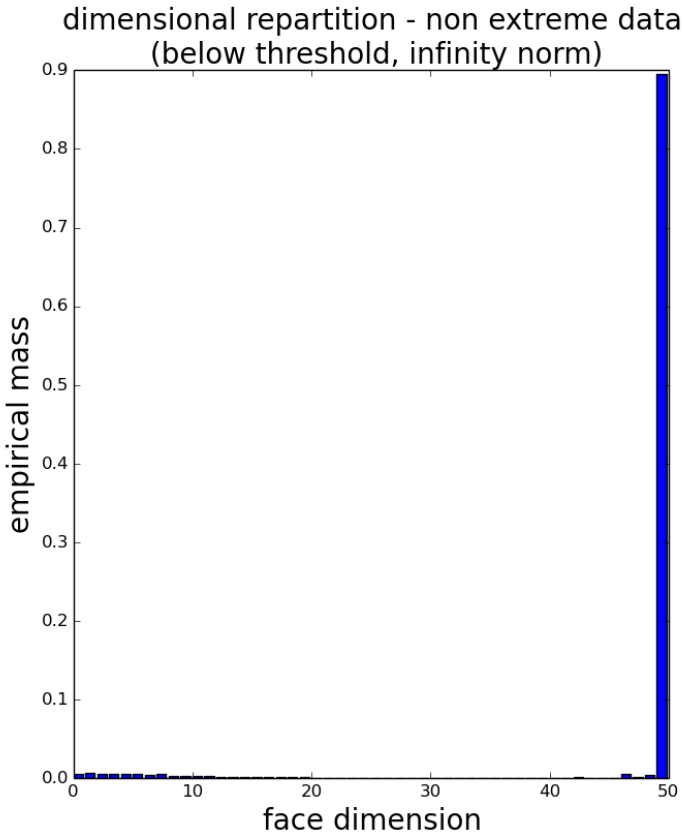
- Standardize: $\hat{V}_i = \frac{1}{1-\hat{F}_j(X_{i,j})}$, with $\hat{F}_j(X_{i,j}) = \frac{rank(X_{i,j})-1}{n}$
- Natural estimator

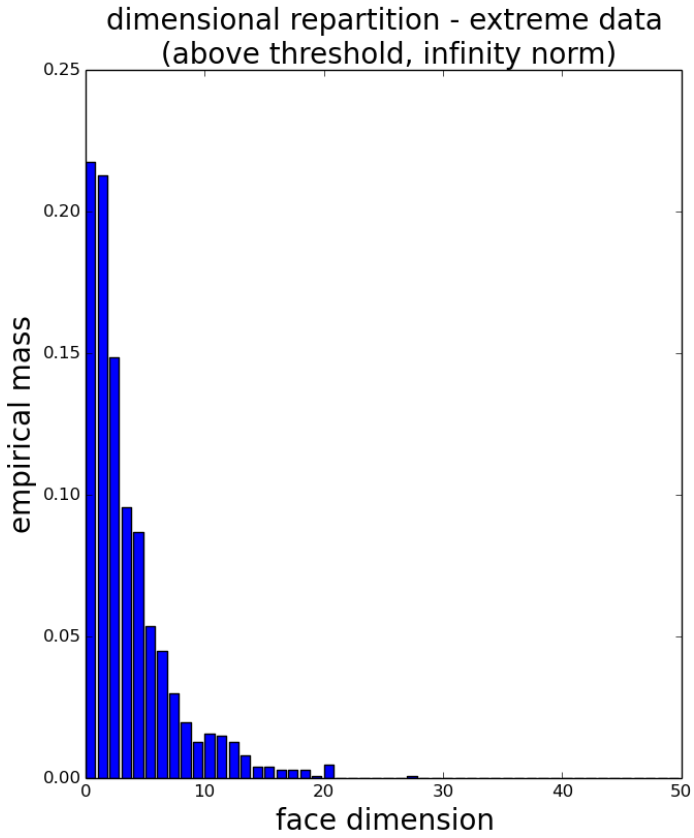
$$\hat{\Phi}_n(\Omega_\alpha) = \mu_n(\mathcal{C}_\alpha^\varepsilon) = \frac{n}{k}\mathbb{P}_n(\hat{\mathbf{V}} \in \frac{n}{k}\mathcal{C}_\alpha^\varepsilon).$$

$$\longrightarrow \hat{\mathcal{M}} = (\hat{\Phi}_n(\Omega_\alpha), \alpha \subset \{1, \ldots, d\})$$

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

Results: support recovery

- Asymmetric logistic, $d = 10$, dependence parameter $\alpha = 0.1$
  $\rightarrow$ Non asymptotic data (not exactly Generalized Pareto)
- K randomly chosen (asymptotically) non-empty faces.
- parameters: $k = \sqrt{n}$, $\epsilon = 0.1$
- Additional (heuristic) step: eliminate faces supporting less than 1% of total mass.

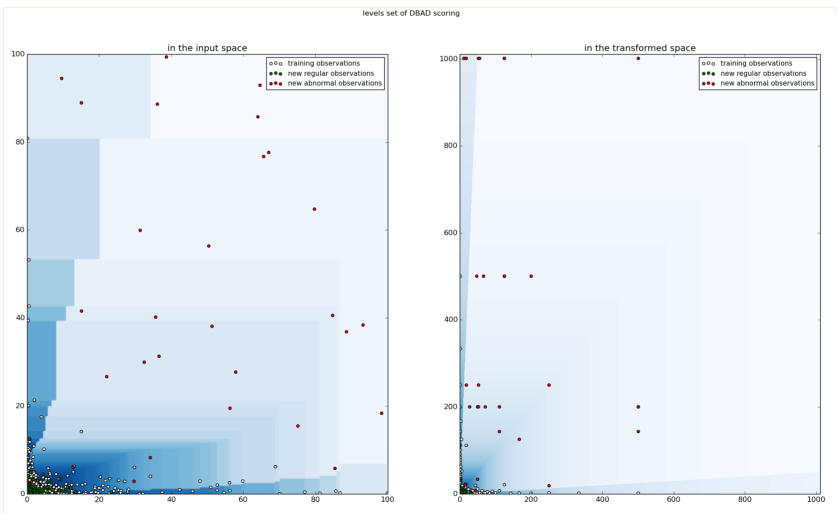
# sub-cones K	10	15	20	30	35	40	45	50
Aver. # errors (n=5e4)	0.01	0.09	0.39	1.82	3.59	6.59	8.06	11.21
Aver. # errors (n=15e4)	0.06	0.02	0.14	0.98	1.85	3.14	5.23	7.87

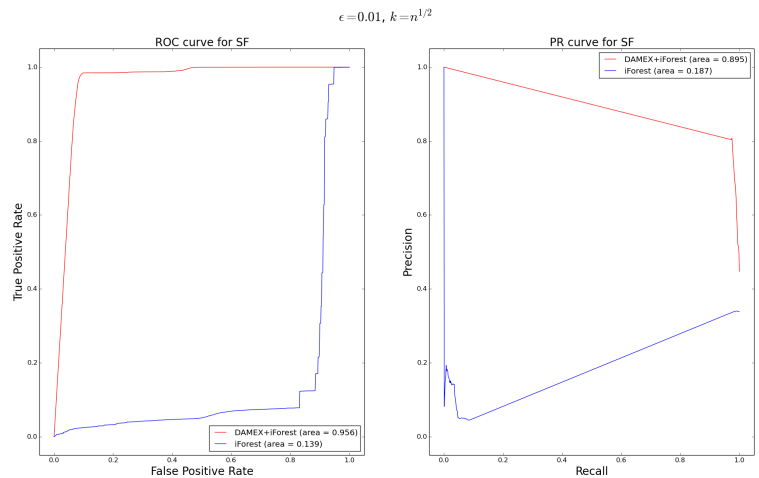
Algorithm DAMEX (Detecting Anomalies with Multivariate Extremes)

Anomaly = new observation 'violating the sparsity pattern': observed in empty or light subcone.

Scoring function: for x such that $\hat{v} \in \mathcal{C}_\alpha^\varepsilon$,

$$s_n(x) = \frac{1}{\|\hat{v}\|} \hat{\phi}_n(\Omega_\alpha^\epsilon) \quad \simeq_{x \text{ large}} \quad \mathbb{P}(V \in \mathcal{C}_\alpha, \|V\| > x)$$



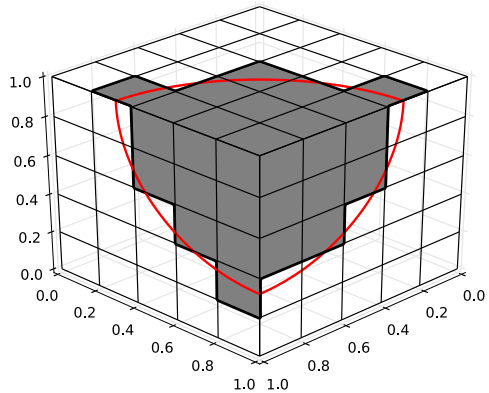


Conclusion

- Adequate notion of **'sparsity' for MEVT**: sparse **angular measure**
- **Empirical estimation** ($\rightarrow$ algorithm) to learn this sparse asymptotic support **from non-asymptotic, non sparse data**.
- **Finite sample error bounds** Goix et. al. , 2015
- **Applications:**
  - Immediate application to AD
  - View towards multivariate extreme (or spatial?) modeling:

    use sparsity information to build a simplified model
    exploit the underlying similarity to design a visualization tool
    (ongoing works)

Minimum Volume sets on the sphere
Thomas, Clémençon, Sabourin & Gramfort (2017)



- **Find anomalies w.r.t. the dependence structure in the extremes:** Apply the minimum volume set estimation methods to the angular probability measure with the **spherical measure** as volume
- Insensitive to scaling effects
- Reduction of the **false alarm** rate
- Rate bounds in $O(1/\sqrt{k})$ in spite of the **rank transformation**

Some references

- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey, 2009
- E. Chautru. Dimension reduction in multivariate extreme value analysis, 2015
- J. H. J. Einmahl , J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution, 2009.
- J. H. J. Einmahl, A. Krajina, J. Segers. An M-estimator for tail dependence in arbitrary dimensions, 2012.
- J.H.J Einmahl, A. Kiriliouk, A. Krajina, J. Segers. An M-estimator of spatial tail dependence, 2015
- 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 Statistical Representation of Multivariate Extremes with Applications to Anomaly Detection, JMVA 2017
- A. Thomas, S. Clémençon, A. Sabourin, A. Gramfort. Minimum Volume sets on the sphere (2017)
- S.J. Roberts. Novelty detection using extreme value statistics, 1999