Machine Learning Methods for Anomaly Detection

Nicolas Goix

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

PhD Defence, Telecom Paristech, October 2016

Anomaly Detection (AD)

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

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

Machine Learning context

Different kind of Anomaly Detection

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

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

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

Outlines

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

- Part I: Performance criterion on s.
  (model selection)
- Part II: Building good s on extreme regions.
  (model design)

(unsupervised) performance criterion

Such a criterion allows:

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

Practical motivations:

Most of the time, data come without any label.

$\rightarrow$ no ROC or PR curves!

Idea:

How good is an anomaly detection algorithm?

$\downarrow$

How good is it estimating the level sets?

Context

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

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

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

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

Problem reformulation

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

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

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

$\Rightarrow \Phi(s) := MV_s$ or EM_s, the Mass-Volume and Excess-Mass curves of s.

Criteria satisfying these requirements: MV and EM

Mass-volume and excess-mass curves

- **Definitions:**

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

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

- **Optimal curves:**

$$MV^*(\alpha) = \min_{\Omega \text{ borelian}} \left\{ \text{Leb}(\Omega) \quad s.t. \ \mathbb{P}(\mathbf{X} \in \Omega) \geq \alpha \right\}$$

$$= MV_f(\alpha) = MV_{T \circ f}(\alpha)$$

$$EM^*(t) = \max_{\Omega \text{ borelian}} \left\{ \mathbb{P}(\mathbf{X} \in \Omega) - t\text{Leb}(\Omega) \right\}$$

$$= EM_f(t) = EM_{T \circ f}(t)$$

MV and EM criteria

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

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

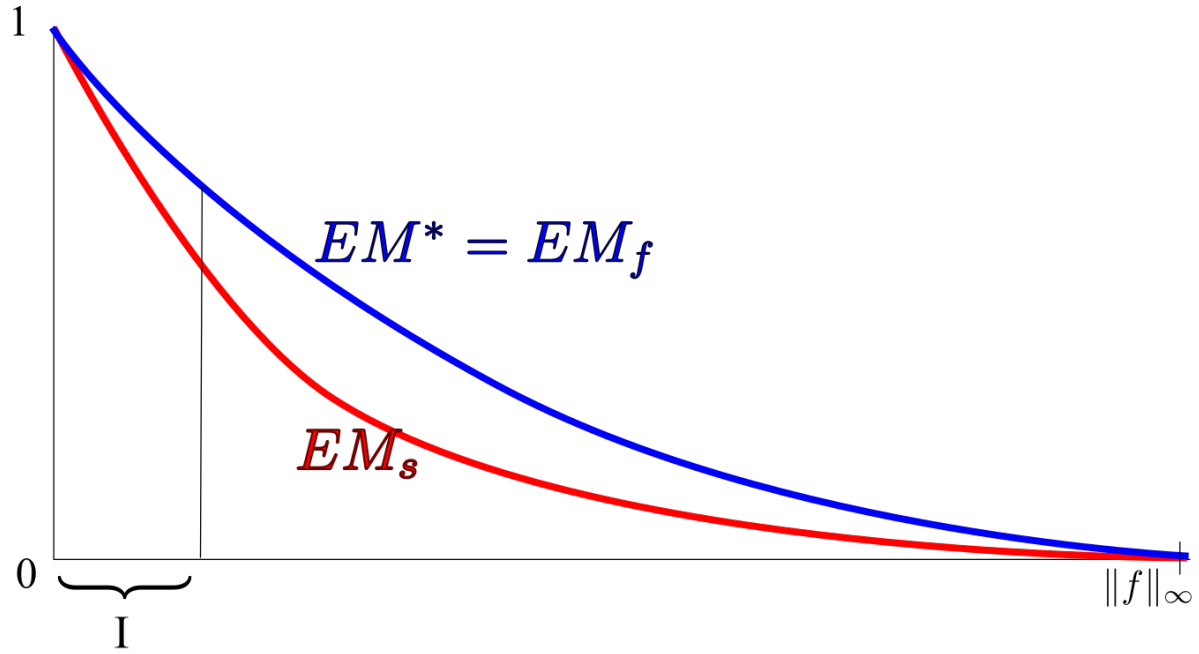
$$\downarrow$$

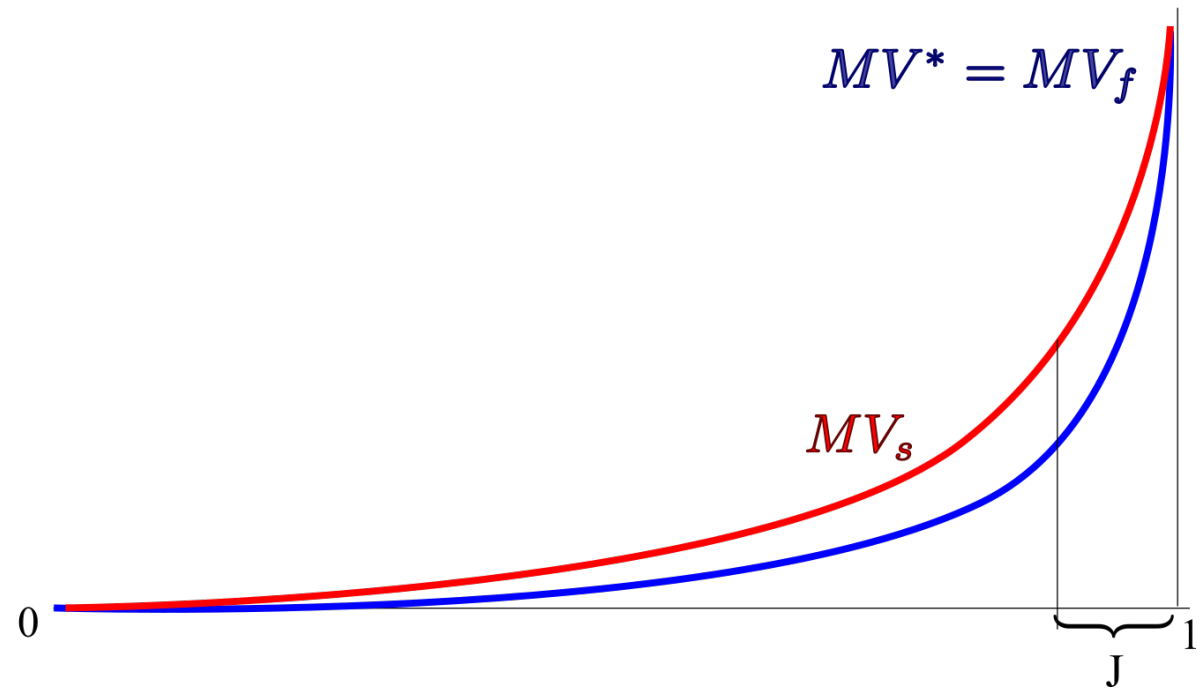
how small is $EM_s - EM_f$ on I ? $\Leftrightarrow$ how large is EM_s on I ?

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

$$\downarrow$$

how small is $MV_s - MV_f$ on J ? $\Leftrightarrow$ how small is MV_s on J ?


1
$EM^* = EM_f$
EM_s
0
I
$\|f\|_\infty$


$MV^* = MV_f$
MV_s
0
1
J

Learning a scoring function with M-estimation

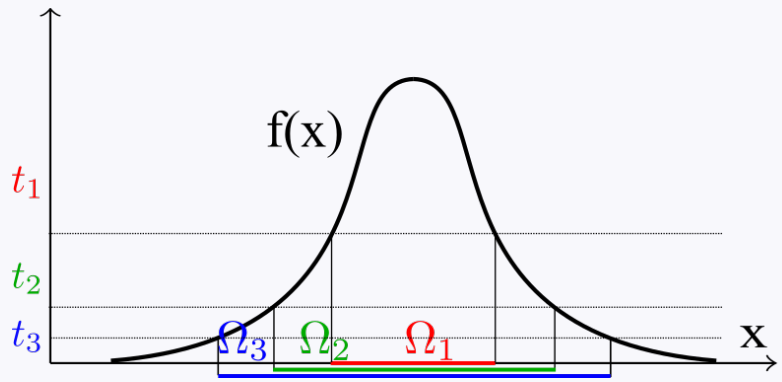
We are looking for nearly optimal scoring functions of the form $s = \sum_{j=1}^{N} a_j \mathbb{1}_{x \in \Omega_j}$, with $a_j \geq 0$, $\Omega_j \in \mathcal{G}$.

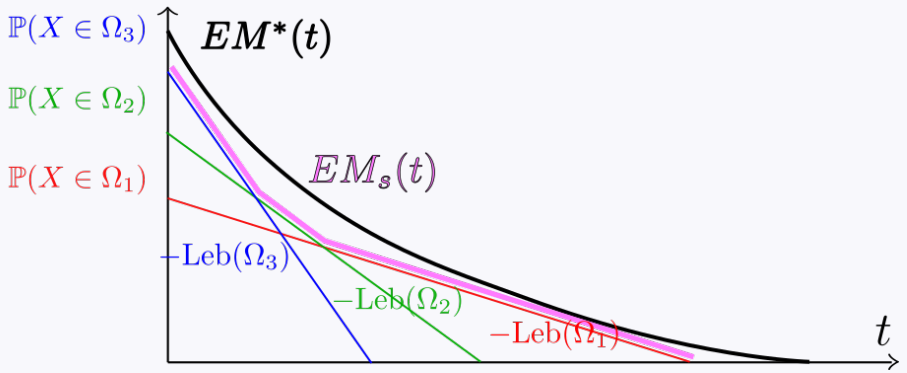
Procedure: For $k = 1, \ldots, N$,

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

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

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

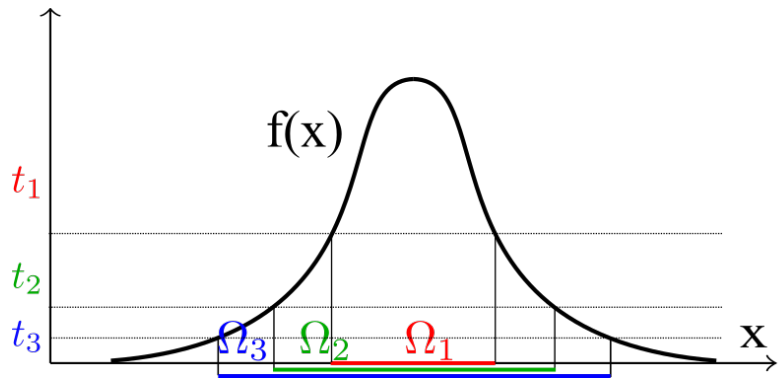


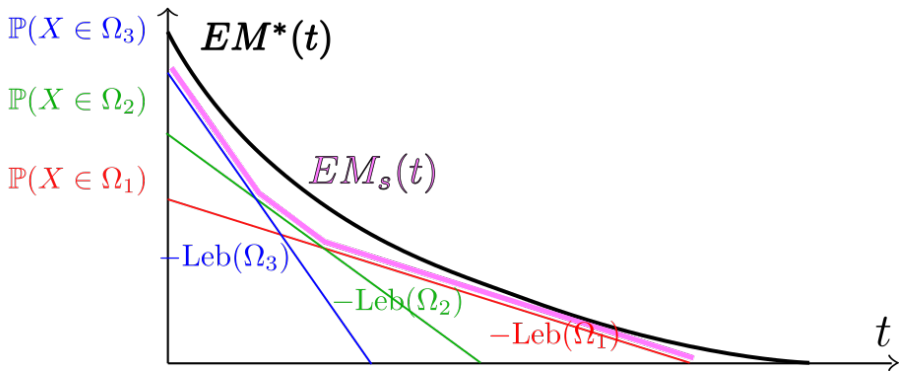


Learning a scoring function with M-estimation

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

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





Evaluation of scoring functions

- **Estimation:**

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

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

- **Empirical criteria:**

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

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

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

Evaluation: Heuristic solution

Feature sub-sampling (random projection) and averaging

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

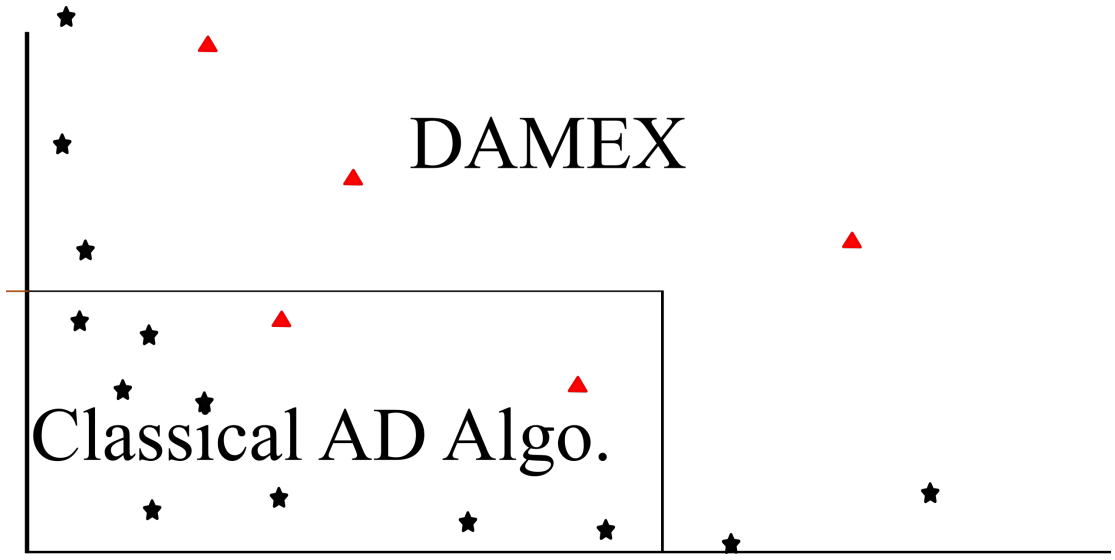
for $k = 1, \ldots, m$ **do**
-randomly select a sub-group F_k of d' features
-compute the associated scoring function $s_k = \mathcal{A}\left((x_i^j)_{1 \leq i \leq n,\ j \in F_k}\right)$
-compute $\widehat{C}_k^{EM} = \|\widehat{EM}_{s_k}\|_{L^1(I)}$ or $\widehat{C}_k^{MV} = \|\widehat{MV}_{s_k}\|_{L^1(J)}$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

Return performance criteria:

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

General idea of our work

- Extreme observations play a special role when dealing with outlying data.
- But no algorithm has **specific treatment for such multivariate extreme observations**.
- Our goal: Provide a method which can improve performance of standard AD algorithms by combining them with a **multivariate extreme analysis** of the **dependence structure**.


DAMEX
Classical AD Algo.

Goal:

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

Find the groups of features which can be large together

ex: $\{X_1, X_2\}$, $\{X_3, X_6, X_7\}$, $\{X_2, X_4, X_{10}, X_{11}\}$

$\Leftrightarrow$ Characterize the extreme dependence structure

Anomalies = points which violate this structure

Framework

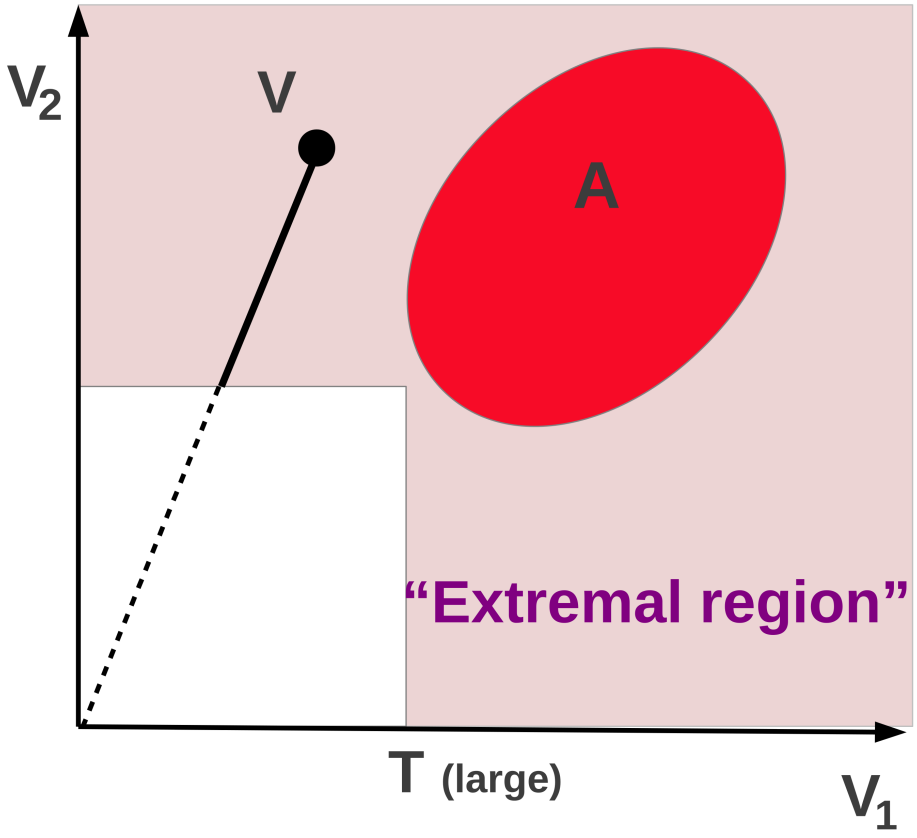
- **Context**
  - Random vector $\mathbf{X} = (X_1, \ldots, X_d)$
  - Margins: $X_j \sim F_j$ (F_j continuous)

- **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

Joint extremes: $\mathbf{V}$'s distribution above large thresholds?

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



Fundamental hypothesis and consequences

- Standard assumption: let A extreme region,

$$\mathbb{P}[\mathbf{V} \in t\ A] \simeq t^{-1}\mathbb{P}[\mathbf{V} \in A] \quad \text{(radial homogeneity)}$$

- Formally,

regular variation (after standardization):
$0 \notin \overline{A}$

$$t\mathbb{P}[\mathbf{V} \in t\ A] \xrightarrow[t\to\infty]{} \mu(A), \qquad \mu: \text{ exponent measure}$$

Necessarily: $\mu(tA) = t^{-1}\mu(A)$

- $\Rightarrow$ **angular measure** on sphere $\mathbf{S}_{d-1}$: $\Phi(B) = \mu\{tB, t \geq 1\}$

General model in multivariate EVT

Model for excesses

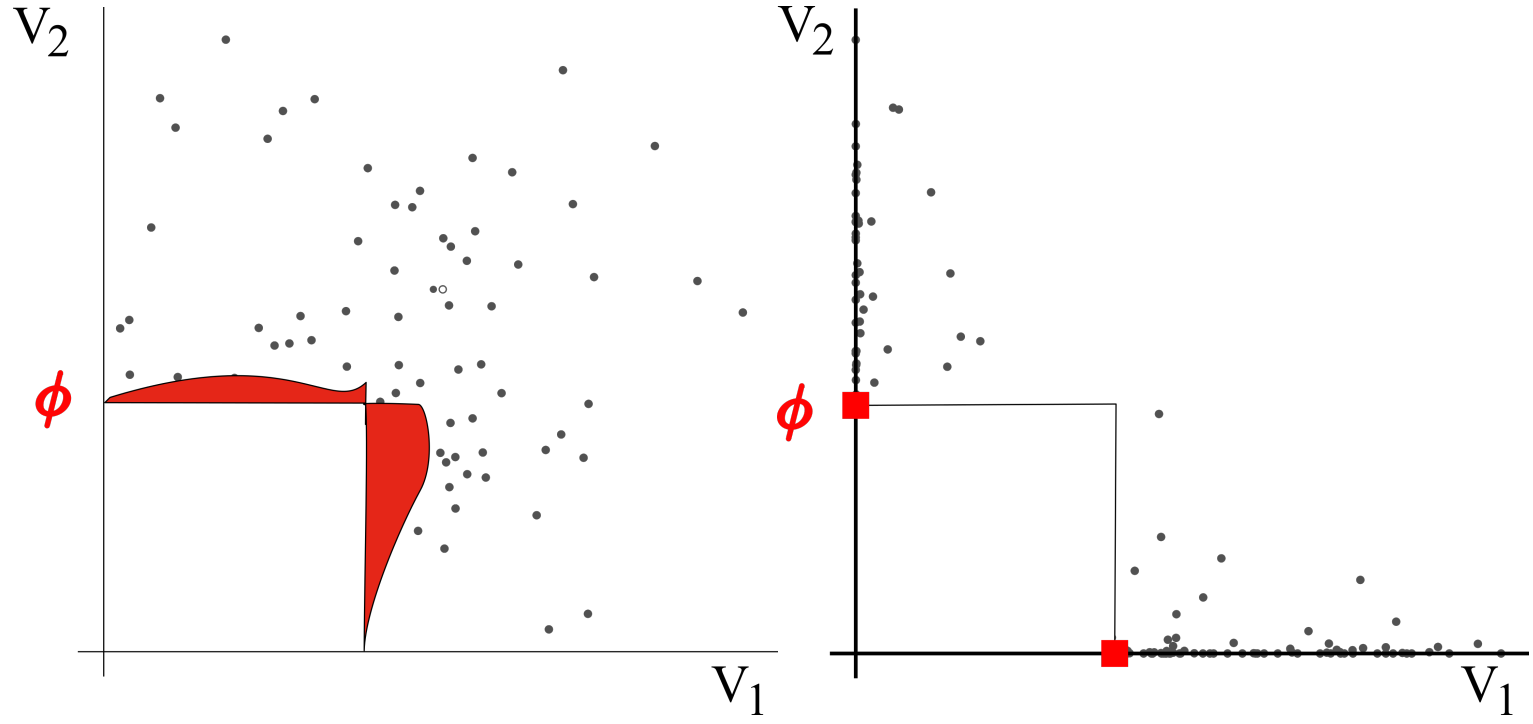
Intuitively: $\mathbb{P}[\mathbf{V} \in A] \simeq \mu(A)$ For a large $r > 0$ and a region B on the unit sphere:

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

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

Angular distribution

- Φ rules the joint distribution of extremes

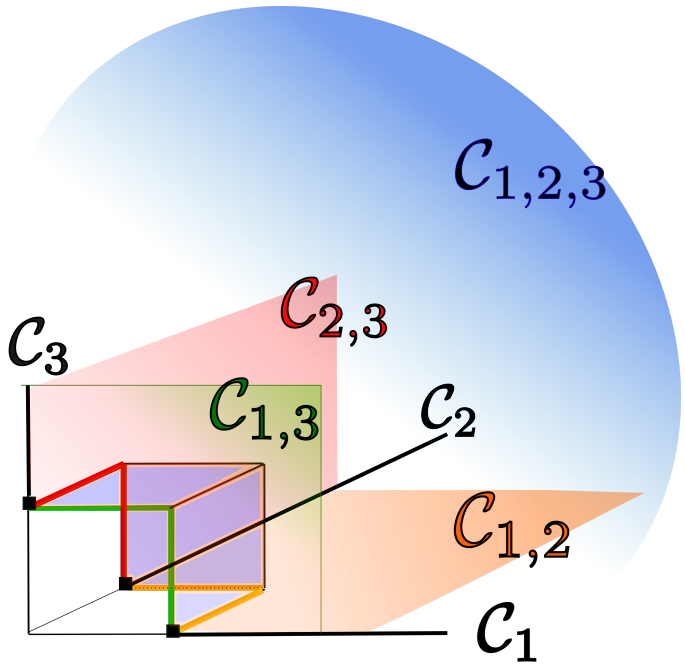


- Asymptotic dependence: (V_1, V_2) may be large together.

vs

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

General Case



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

Representation of extreme data

- Natural decomposition of the angular measure :

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

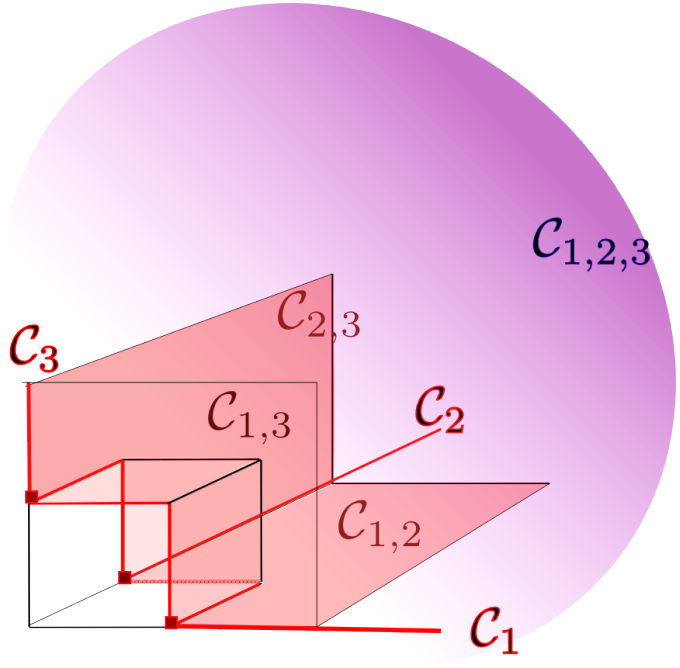
- $\Rightarrow$ yields a representation

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

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

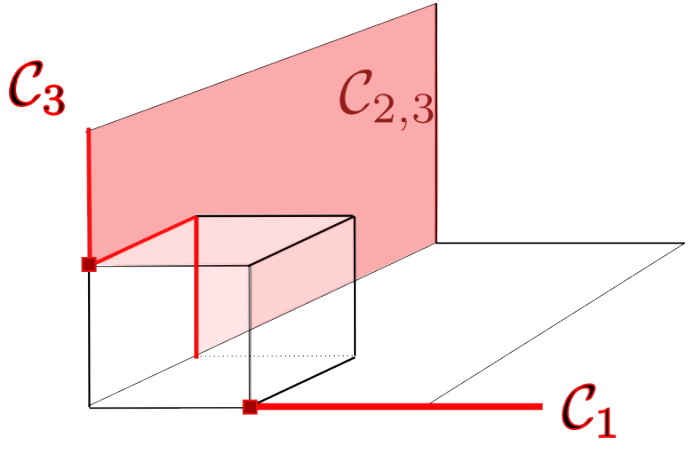
- Remark: Representation $\mathcal{M}$ is linear (after non-linear transform of the data $\mathbf{X} \to \mathbf{V}$).

Sparse Representation ?



Full pattern :

anything may happen



Sparse pattern

(V_1 not large if V_2 or V_3 large)

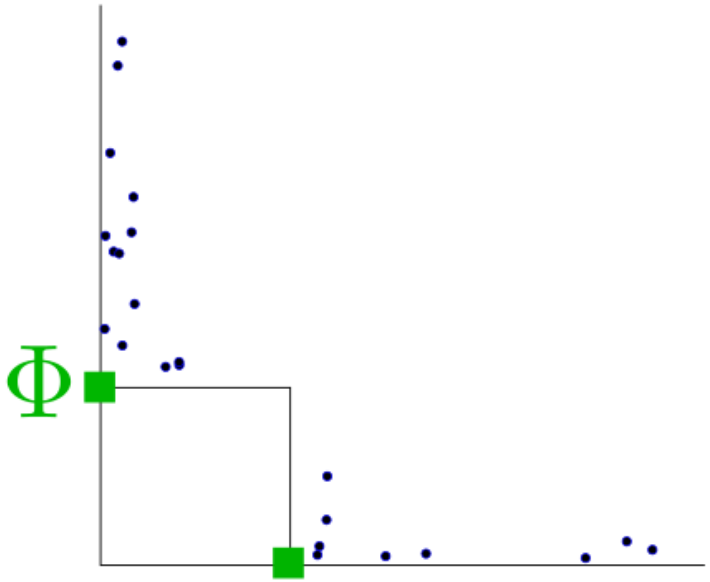
Problem: $\mathcal{M}$ is an **asymptotic** representation

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

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!

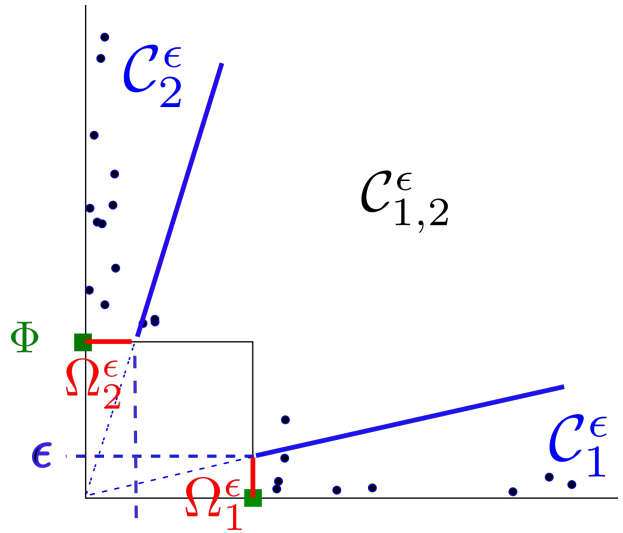


$\Rightarrow$ Cannot just do, for large t:

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

Solution

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



$$\Omega_\alpha \to \Omega_\alpha^\epsilon = \{v \in \mathbf{S}_{d-1} : v_i > \epsilon\,(j \in \alpha),\ v_j \leq \epsilon\,(j \notin \alpha)\}.$$
$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\epsilon = \{t\,\Omega_\alpha^\epsilon, t \geq 1\}$$

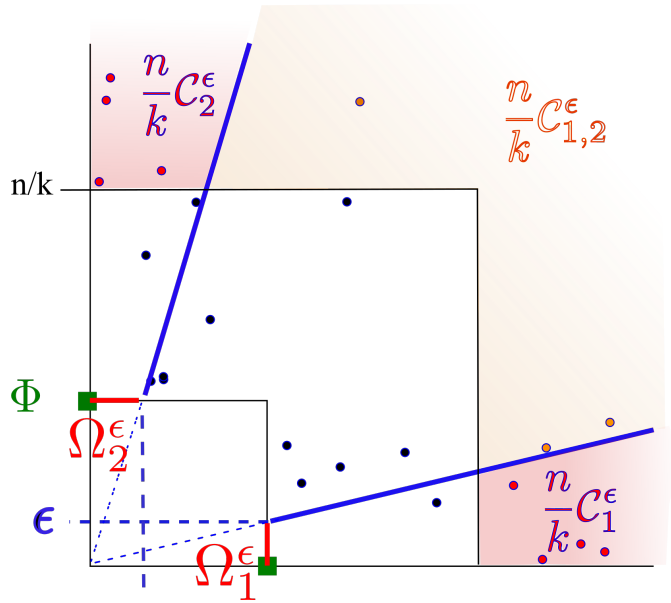
New partition of $\mathbf{S}_{d-1}$, compatible with non asymptotic data.

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

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

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

$(\frac{n}{k}$ large, ϵ small$)$



$\Rightarrow$ we obtain

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

Theorem

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

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

Comments:

- C: depends on $M = \sup$(density on subfaces)
- Existing litterature (for spectral measure) **Einmahl Segers 09, Einmahl *et.al.* 01**

  $$d = 2.$$

  asymptotic behaviour, rates in $1/\sqrt{k}$.
  Here: $1/\sqrt{k} \to 1/\sqrt{\epsilon k} + \epsilon$. Price to pay for biasing our estimator with ϵ.

Theorem's proof

1. **Maximal deviation on VC-class:**

$$\sup_{x \succeq \epsilon} |\mu_n - \mu|([x, \infty[) \leq Cd\sqrt{\frac{2}{k} \log \frac{d}{\delta}} + \text{bias}(\epsilon, k, n)$$

Tools: Vapnik-Chervonenkis inequality adapted to small probability sets: bounds in $\sqrt{p}\sqrt{\frac{1}{n} \log \frac{1}{\delta}}$

On the VC class $\{[\frac{n}{k}x, \infty], x \geq \epsilon\}$

Theorem's proof

1. **Maximal deviation on VC-class:**
2. **Decompose error:**

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

- A : First step.
- B : density on $\mathcal{C}_\alpha^\epsilon \times$ *Lebesgue* : small

DAMEX in $O(dn \log n)$

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

Standardize *via* marginal rank-transformation:

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

Assign to each $\hat{V}_i$ the cone $\frac{n}{k}\mathcal{C}_\alpha^\epsilon$ it belongs to.

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

Output: (sparse) representation of the dependence structure

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

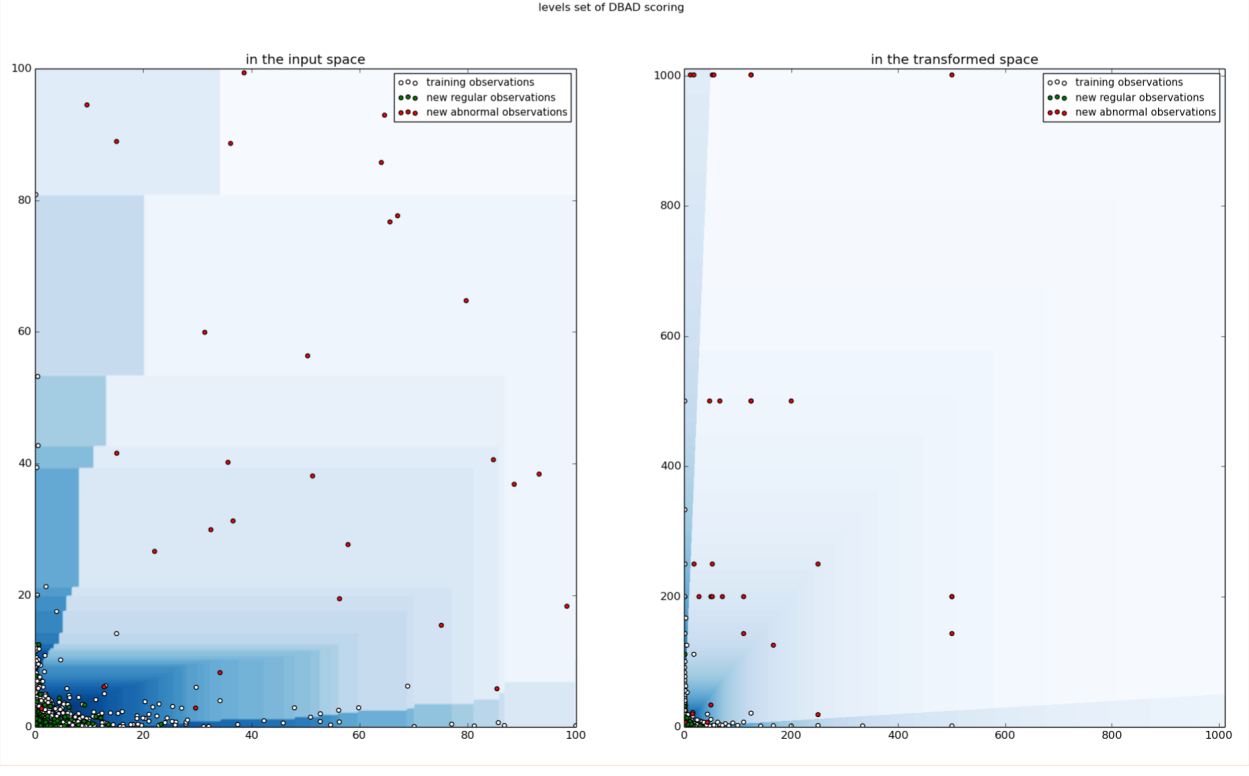
Application to Anomaly Detection

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 $T : \mathbf{X} \mapsto \mathbf{V}$ $\qquad (V_j = \frac{1}{1-F_j(X_j)})$



	number of samples	number of features
shuttle	85849	9
forestcover	286048	54
SA	976158	41
SF	699691	4
http	619052	3
smtp	95373	3

Table: Datasets characteristics

Thank you!

Benchmarks

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

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

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

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

Table: Original Datasets characteristics

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

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

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