Dimension reduction and feature clustering in multivariate extremes

Anne Sabourin[1]
Joint work with Maël Chiapino[1] , Johan Segers[2],
Nicolas Goix[1] , Stephan Clémençon [1]

[1] LTCI, Télécom ParisTech, Université Paris-Saclay
[2] Université catholique de Louvain, Louvain-La-Neuve

Statistics/Learning at Paris-Saclay,
January 19, 2017

This talk

Random vector $X = (X^1, \dots, X^d)$, $\quad d$ 'large'

- **Focus** on extremes : $\mathcal{L}\Big[X \;\Big|\; \|X\| \gg 1\Big] \approx \mu$
- **Dimension reduction:**

  Identify **supporting subspaces** of μ $\qquad (\star)$

Multivariate extreme value theory (MEVT) tells us:

$(\star) \iff$ Identify the **groups of features** $\alpha \subset \{1, \dots d\}$ which **may be large together** (while the others stay small), given that one of them is large.

1. Support recovery, finite sample error, concentration

   (Goix, S., Clémençon, 15, 16)

2. Subspaces/features clustering Chiapino, S., 2016

Outline

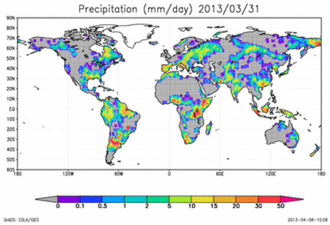
Motivation

Multivariate extremes

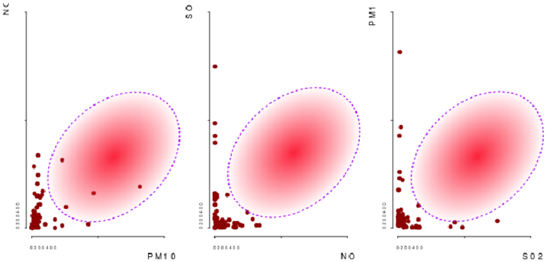
Estimating the (sparse) support of Φ / μ

Feature clustering

It cannot rain everywhere at the same time



(daily precipitation)



(air pollutants)

question (e.g. for risk management):
Which groups of sensors/components are likely to be jointly impacted ?

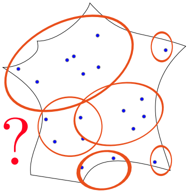
Applications to risk management

Sensors network (road traffic, river streamflow, temperature, internet traffic . . .):

→ extreme event = traffic jam, flood, heatwave, network congestion

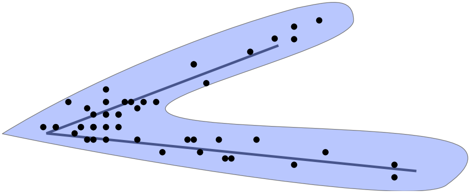
→ **question**: which groups of sensors are likely to be jointly impacted ?

→ how to define **alert regions** (alert groups of features)?



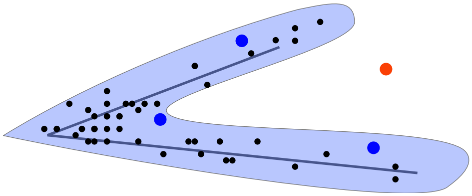
spatial case: one feature = one sensor

Applications to anomaly detection



- **Training step:**
  Learn a **'normal region'** (*e.g.* approximate support)

Applications to anomaly detection



- **Training step:**
  Learn a **'normal region'** (*e.g.* approximate support)
- **Prediction step:** (with new data)
  Anomalies = points outside the 'normal region'

If 'normal' data are heavy tailed, **Abnormal** $\not\Leftrightarrow$ **Extreme** .
There may be **extreme** 'normal data'.

How to distinguish between large anomalies and normal extremes?

Outline

Motivation

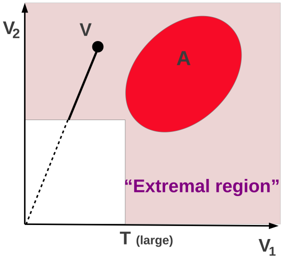
Multivariate extremes

Estimating the (sparse) support of Φ / μ

Feature clustering

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' ?



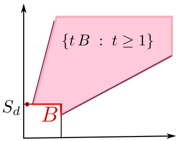
Regular variation assumption

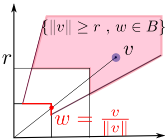
de Haan, Resnick, 70's, 80's

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

Polar coordinates ($R = \|\mathbf{V}\|, \mathbf{W} = \frac{\mathbf{V}}{\|\mathbf{V}\|}$): a product $\mathrm{d}\mu(r, \mathbf{w}) = \frac{\mathrm{d}r}{r^2}\,\mathrm{d}\Phi(\mathbf{w})$.

Φ: a finite **angular measure** on the sphere, $\Phi(B) = \mu\{tB, t \geq 1\}$.





$$\text{'model' for large } V\text{'s: } \mathbb{P}\left(\ \|\mathbf{V}\| \geq r\ ;\quad \frac{\mathbf{V}}{\|\mathbf{V}\|} \in B\ \right) \approx r^{-1}\,\Phi(B)$$

Estimation of the dependence structure: $\Phi(B)$ or $\mu[0, x]^c$

- 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*) ...

- **Asymptotic** theory: rates under **second order conditions**

  (Einmahl, 01) Empirical likelihood (Einmahl, Segers 09) Asymptotic normality (Einmahl *et. al.*, 12, 15) (parametric)

- **Finite sample** error bounds, non parametric, on

$$\sup_{x \succeq R} |\hat{\mu}_n[0, x]^c - \mu[0, x]^c| \qquad \text{(Goix, S., Clémençon, 15)}$$

Does **not** tell 'which components may be large together'

A bound on the stdf

$\mathbf{x} \in \mathbb{R}_d^+ \setminus \{0\}$, $\qquad l(\mathbf{x}) = \mu[0, 1/\mathbf{x}]^c$.
$k = o(n), k \to \infty$,
Rank transform: $\hat{F}_j(x) = \frac{1}{n}\sum \mathbf{1}_{X_i^j \le x} \qquad \hat{V}_i^j = \frac{1}{1-\hat{F}_j(X_i^j)}$

Empirical estimator of l

$$l_n(\mathbf{x}) = \frac{n}{k}\left(\frac{1}{n}\sum_1^n \delta_{\hat{V}_i^j}\left(\frac{n}{k}\Big[\mathbf{0}, 1/\mathbf{x}\Big]^c\right)\right)$$

Theorem (Goix, S. Clémençon, 15)

$$\text{for } T > \frac{7}{2}\left(\frac{\log d}{k} + 1\right),\ \delta > e^{-k},$$

$$\sup_{0 \preceq \mathbf{x} \preceq \mathbf{T}} |l_n - l|(\mathbf{x}) \le Cd\sqrt{\frac{T}{k}\log\frac{d+3}{\delta}} + \mathrm{Bias}_{\frac{n}{k}, T}(F, \mu)$$

Existing litterature ($\mathbf{d = 2}$): Einmahl Segers 09, Einmahl *et.al.* 01: asymptotic, $O(1/\sqrt{k})$.

Tools for the proof

N.B

$$\mathrm{Bias}_{\frac{n}{k},T}(F,\mu) = \sup_{0 \preceq x \preceq \mathbf{T}} \left| \frac{n}{k} \mathbb{P}\left(\exists j \leq d : 1 - F_j(X^j) \leq \frac{k}{n} x_j \right) - l(x) \right|$$

$$\xrightarrow{n\to\infty} 0 \qquad \text{(regular variation assumption)}$$

1. Mc Diarmid (98) 's Bernstein type concentration inequality involving the *variance* of martingale differences.
2. $\rightarrow$ VC inequality for small probability classes (Goix *et.al.*, 2015)

   $\rightarrow$ max deviations $\leq \sqrt{p} \times$ (usual bound)

3. Apply it on VC-class of rectangles $\{\frac{k}{n}[0,\mathbf{x}]^c\}$

$$\rightarrow p \leq d\frac{kT}{n} \quad \Rightarrow \quad \sup_{\alpha} |\hat{\mu}_n - \mu|(R_\alpha^\epsilon) \leq Cd\sqrt{\frac{T}{k}\log\frac{d}{\delta}}$$

Outline

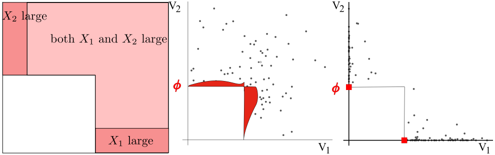
Motivation

Multivariate extremes

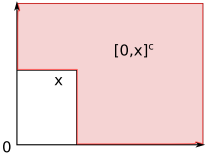
Estimating the (sparse) support of Φ / μ

Feature clustering

Back to problem: 'which components may be large together, while the others are small?'

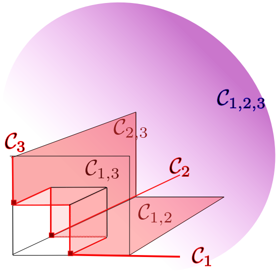


- Φ's support determines the answer.
- Unfortunately, above results concern $\mu[\mathbf{0}, \mathbf{x}]^c$, which is:



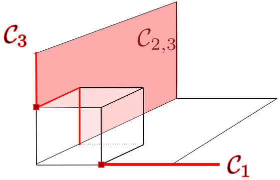
- Inclusion/exclusion: scary in high dimension (error terms pile up).

in higher dimensions: sparse angular support ?



Full support:

anything may happen



Sparse support

(V_1 not large if V_2 or V_3 large)

Cones: $\mathcal{C}_{\boldsymbol{\alpha}} = \{\mathbf{x} \succeq 0 : \ \|\mathbf{x}\| \geq 1, \ \ x_j = 0 \, (j \notin \alpha)\}, \quad \alpha \subset \{1, \ldots, d\}$

Subspheres: $\boldsymbol{\Omega}_{\boldsymbol{\alpha}} := \text{Projections on the sphere}$

Where is the mass?

$$\mu(\mathcal{C}_\alpha) > 0 \iff \Phi(\Omega_\alpha) > 0 \iff$$

features $j \in \alpha$ may be large together while the others are small.

Identifying non empty edges

Issue:

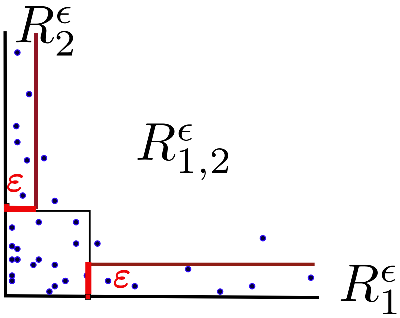
real data are non asymptotic.

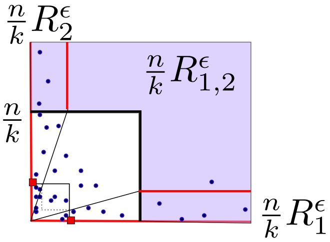
$\rightarrow$ all points belong to the interior cone $\mathcal{C}_{\{1,\dots,d\}}$.

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

$$\mathcal{C}_\alpha \rightarrow R_\alpha^\varepsilon$$

New partition of the sample space, compatible with non asymptotic data.





Empirical estimator of $\mu(\mathcal{C}_\alpha)$)
(Counts the standardized points in R_α^ε, far from 0.)

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

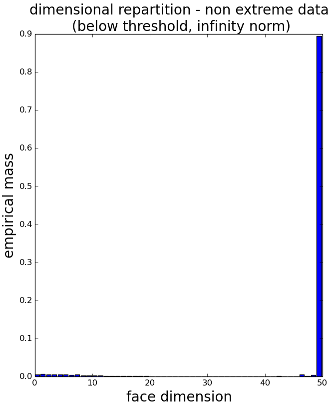
- Standardize: $\hat{V}_{i,j} = \frac{1}{1-\hat{F}_j(X_{i,j})}$, with $\hat{F}_j(X_{i,j}) = \frac{rank(X_{i,j})-1}{n}$
- Natural estimator

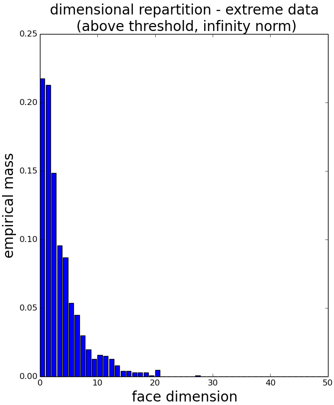
$$\hat{\mu}_n(\mathcal{C}_\alpha^\varepsilon) = \frac{n}{k}\mathbb{P}_n(\hat{\mathbf{V}} \in \frac{n}{k} R_\alpha^\varepsilon)$$

- Estimated support $\hat{\mathcal{S}} = \{\alpha : \hat{\mu}_n(\mathcal{C}_\alpha) > \mu_0\}$.

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

Finite sample error bound

VC-bound adapted to low probability regions (see Goix, S., Clémençon, 2015)

Theorem

If the margins F_j are continuous and if the density of the angular measure is bounded by $M > 0$ on each subface,
There is a constant C s.t. for any n, d, k, $\delta \geq e^{-k}$, $\varepsilon \leq 1/4$,
with probability $\geq 1 - \delta$,

$$\max_{\alpha} |\hat{\mu}_n(\mathcal{C}_\alpha) - \mu(\mathcal{C}_\alpha)| \leq Cd \left(\sqrt{\frac{1}{k\varepsilon} \log \frac{d}{\delta}} + Md\varepsilon \right) + \mathrm{Bias}_{\frac{n}{k},\varepsilon}(F, \mu).$$

Bias: using non asymptotic data to learn about an asymptotic quantity

$$\text{Regular variation} \iff \mathrm{Bias}_{t,\varepsilon} \xrightarrow[t\to\infty]{} 0$$

- relaxed bound: $1/\sqrt{k\varepsilon} + Md\varepsilon$. Price for biasing estimator with ε.
- Choice of ε: cross-validation or '$\varepsilon = 0.1$'

Tools for the proof

1. Apply the deviation bound for low-probability region on the VC-class of rectangles $\{\frac{k}{n} R(x, z, \alpha),\ x, z \succ \varepsilon\}$

$$\to p \le d \frac{k}{\varepsilon n} \quad \Rightarrow \quad \sup_{\alpha} |\hat{\mu}_n - \mu|(R_\alpha^\epsilon) \le Cd \sqrt{\frac{1}{\varepsilon k} \log \frac{d}{\delta}}$$

   ($1/\varepsilon$ plays the role of T in the previous bound for the stdf)

2. Approach $\mu(\mathcal{C}_\alpha)$ with $\mu(R_\alpha^\varepsilon) \to$ error $\le Md\varepsilon$ (bounded angular density).

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: $k = \sqrt{n}$, $\epsilon = 0.1$
- Heuristic for setting minimum mass μ_0: 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

Outline

Motivation

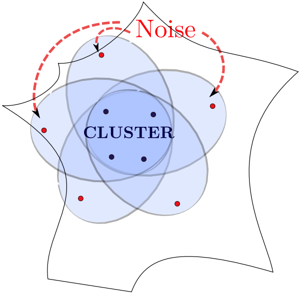
Multivariate extremes

Estimating the (sparse) support of Φ / μ

Feature clustering

Feature clustering (Chiapino, S, 2016)

Toy example: River stream-flow dataset, $d = 92$ gauging stations:
Typical groups jointly impacted by extreme records include noisy additional features !
$\rightarrow$ Empirical μ-mass **scattered** over many $\mathcal{C}_\alpha$'s



$$Z_{[i,j]} = \mathbb{I}_{V_i^j > t}$$

i \j	1	2	3	4	5	6	7	8	9	....
1	0	1	1	1	1	1	0	0	0	
2	1	1	1	1	1	0	0	0	0	
3	0	1	1	1	1	0	1	0	0	
4	0	1	1	1	1	0	0	1	0	
5	0	1	1	1	1	0	0	0	1	....
⋮									⋮	

$\rightarrow$ No apparent sparsity pattern.
How to **gather** 'closeby' α's into **feature clusters**? (= maximal groups of dependent features)

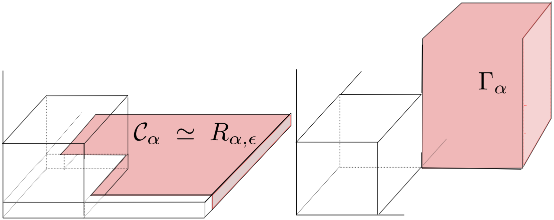
Relaxed constraints on the region of interest

Initial regions of interest:

$$\mathcal{C}_\alpha = \{\mathbf{v} \succeq 0 : v^j \text{ large for } j \in \alpha,\ v^j \text{ small for } j \notin \alpha\}$$

Modified regions (relaxed constraints, larger and nested regions)

$$\Gamma_\alpha = \{\mathbf{v} \succeq 0 : v^j \text{ large for } j \in \alpha\}$$



$$\alpha \text{ is maximal in } \{\alpha : \mu(\mathcal{C}_\alpha) > 0\}$$
$$\Longleftrightarrow$$
$$\alpha \text{ is maximal in in } \{\alpha : \mu(\Gamma_\alpha) > 0\}$$

Conditional criterion

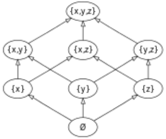
- One needs an empirical criterion for 'testing' dependence: $\mu(\Gamma_\alpha > 0)$. *e.g.* $\hat{\mu}_n(\Gamma_\alpha) > \mu_0$.
- Issue: $\mu(\Gamma_\alpha) \searrow$ as $|\alpha| \nearrow$ **set the threshold according to** $|\alpha|$ **?**
- Way around: **condition** upon excess of **all but one** components.

$$\begin{aligned}\kappa_\alpha &= \lim_{t\to\infty} \mathbb{P}(\forall j \in \alpha, V^j > t \mid V^j > t \text{ for all but at most one } j \in \alpha\} \\ &= \frac{\mu(\Gamma_\alpha)}{\mu\big(\bigcup_{\beta\subset\alpha, |\beta|\geq|\alpha|-1} \mu(\Gamma_\beta)\big)}\end{aligned}$$

Empirical criterion $$\hat{\kappa}_{\alpha,t} = \frac{\sum_{i=1}^n \mathbf{1}_{\widehat{V}_i^j > t \text{ for all } j\in\alpha}}{\sum_{i=1}^n \mathbf{1}_{\widehat{V}_i^j > t \text{ for all but at most one } j\in\alpha}}$$

Coping with combinatorial complexity

- $O(2^d)$ subsets $\alpha \subset \{1, \dots\}$ to be examined!
- Good news: $\mu(\Gamma_\alpha) = 0 \Rightarrow \forall \beta \supset \alpha, \mu(\Gamma_\beta) = 0$
  $\rightarrow$ the search should 'follow' the Hasse diagram



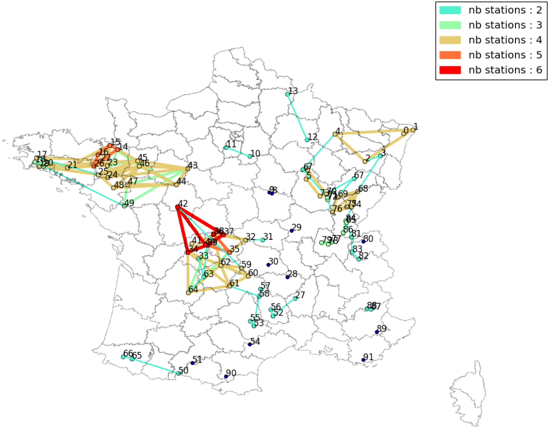
CLEF algorithm (CLustering Extreme Features):

- Start with pairs: $\hat{\mathbf{A}}_2 = \{\alpha : |\alpha| = 2, \quad \hat{\kappa}_t(\alpha) > \kappa_0\}$.
- Stage k: $\hat{\mathbf{A}}_k = \{\alpha : |\alpha| = k, \hat{\kappa}_t(\alpha) > \kappa_0\}$; $\rightarrow$ Candidates for $\mathbf{A}_{k+1}$:

$$\{\alpha : |\alpha| = k+1,\ \forall \beta \subset \alpha \text{ s.t. } |\beta| = k, \beta \in \hat{\mathbf{A}}_k.\} \qquad \text{Not too many !}$$

- Related data mining literature: 'frequent itemsets mining'
  Apriori algorithm (Agrawal et al., 94), feature clustering (Agrawal et al., 2005), fault-tolerant pattern discovery (Pei et al., 2001)

Toy example: output on stream-flow data



dependent groups are large in the North-West (oceanic climate), small in the south west (mediterranean climate, rain-storms)

(time: $\sim 1s$)

Simulated data

Generation:
20 datasets with $N = 100 \,.\, 10^3, d = 100$.
From *asymmetric logistic* extreme value model [4,5]. For each dataset, p subsets $\alpha_1, .., \alpha_p$ of $\{1, .., 100\}$ are randomly chosen.
Noise:
For each $i \leq N$, one additional noisy feature is added to each true α.

p	# errors CLEF	# errors with R_α^ϵ regions (Goix et. al., 16)
40	1.2	72.2
50	3.5	91.0
60	10.1	134.0

Average number of errors (non recovered and falsely discovered clusters).

(average computation time : $\sim 1s$ on a laptop)

Link with extremal coefficients joint work with Johan Segers

- Recall the extremal coefficient $\ell_\alpha (= \theta_\alpha) = \mu(\exists \mathbf{j} \in \alpha : x_j > 1)$.
  (Schlather & Tawn, 03, Einmahl Kiriliouk, Segers 16, ...)

- Define $\boldsymbol{\rho_\alpha} := \mu(\Gamma_\alpha) = \mu(\forall \mathbf{j} \in \alpha, x_j > 1)$
- Inclusion/exclusion $\to$ our incremental criterion κ_α re-writes

$$\kappa_\alpha = \frac{\rho_\alpha}{\sum_{j \in \alpha} \rho_{\alpha \setminus \{j\}} - (|\alpha| - 1)\rho_\alpha.}$$

- Inclusion/exclusion again $\to \rho_\alpha = \sum_{\beta \subset \alpha} (-1)^{|\beta|+1} \ell_\alpha$
  Nice! because the asymptotic joint distribution of $(\hat{\ell}_\alpha)_{\alpha \subset \{1,\dots,d\}}$ is known. (Einmahl, Kiriliouk, Segers, 16)

- Delta method $\to$ (work in progress)
  Gaussian asymptotics for $\sqrt{k}(\hat{\kappa}_\alpha - \kappa_\alpha)_{\emptyset \neq \alpha \subset \{1,\dots,d\}}$, statistical tests
  ... to be continued.

Conclusion

- Adequate notion of **'sparsity' for MEVT**: sparse **angular measure**
- **Empirical estimation** ($\rightarrow$ simple algorithms) to learn this sparse asymptotic support **from non-asymptotic, non sparse data**.
- **Finite sample error bounds** (tools from statistical learning theory)
- **When sparsity structure not apparent**: feature clustering may be necessary
- **Applications:**
  - Extreme values modeling: identification of dependent subgroups
  - Anomaly detection among extremes.

Some references

- R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan . Automatic subspace clustering of high dimensional data for data mining applications, 1998
- R. Agrawal, R. Srikant. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, 94
- E. Chautru. Dimension reduction in multivariate extreme value analysis, 2015
- M. Chiapino, A. Sabourin. Feature clustering for extreme events analysis, with application to extreme stream-flow data. In ECML-PKDD 2016, workshop NFmcp2016.
- J. H. J. Einmahl , J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution, 2009.
- 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 representation of multivariate extremes with applications to anomaly ranking, to appear in AISTATS (2016)
- N. Goix, A. Sabourin, S. Clémençon. Sparsity in multivariate extremes with applications to anomaly detection , arXiv preprint arXiv:1507.05899 (2015).
- S. Resnick. Extreme Values, Regular Variation, Point Processes, 1987