Extreme Value Theory and Machine Learning

Habilitation defense (HDR)

Anne Sabourin

LTCI, Télécom Paris, Institut polytechnique de Paris, France.

October 27, 2021

Jury:

Stéphane Boucheron (Reviewer), Richard Davis (Reviewer), Matthieu Lerasle (Internal Reviewer), Holger Drees, Gabor Lugosi, Johan Segers.

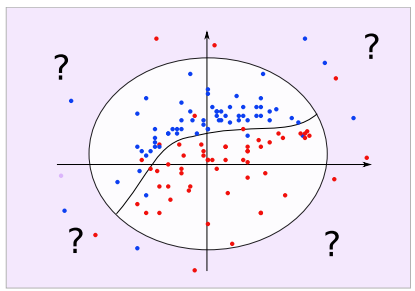
Acknowledgments

The works presented today are the result of collaborations with many people

- S2A team (Télécom Paris): Stephan Clémençon, Chloé Clavel, François Portier, Giovanna Varni.
- Others: Patrice Bertail, Holger Drees, Vincent Feuillard, Eric Gaussier, Alexandre Gramfort, Johan Segers, Chen Zhou.
- PhD students (past and present, co-advised by me or not, chronological order) Nicolas Goix, Maël Chiapino, Albert Thomas, Hamid Jalalzai, Pierre Colombo, Anass Aghbalou, Stéphane Lhaut, Nathan Huet.

Generic research goals

- Obtain statistical learning guarantees regarding estimators of standard quantities issued from extreme value theory (EVT)
- Develop dimensionality reduction tools in order to extend the range of application of multivariate EVT
- Propose EVT-based solutions to standard machine learning tasks (classification/anomaly detection)



Selected list of publications

Satistical learning guarantees: empirical measure and ERM classification

- Goix, S., Clémençon (2015). Learning the dependence structure of rare events: a non-asymptotic study. Conference on Learning Theory.
- Jalalzai, Clémençon, S. (2018). On Binary Classification in Extreme Regions. NeurIPS proceedings.
- Jalalzai, Colombo, Clavel, Gaussier, Varni, Vignon, S. (2020). Heavy-tailed Representations, Text Polarity Classification and Data Augmentation , NeurIPS proceedings.

Dimensionality reduction

- Goix, S., Clémençon (2017). Sparse representation of multivariate extremes with applications to anomaly detection. Journal of Multivariate Analysis
  and
  Goix, S., Clémençon (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. AISTATS proceedings.
- Chiapino, S. (2016) Feature clustering for extreme events analysis, with application to extreme stream-flow data. International Workshop on New Frontiers in Mining Complex Patterns
  and
  Chiapino, S., Segers (2019) Identifying groups of variables with the potential of being large simultaneously. Extremes.
- Drees, S. (2021) Principal component analysis for multivariate extremes. Electronic Journal of Statistics

Anomaly detection and clustering

- Thomas, Clémençon, Gramfort, S. (2017) Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. AISTATS proceedings.
- Chiapino, Clémençon, Feuillard, S. (2019) A multivariate extreme value theory approach to anomaly clustering and visualization. Computational Statistics.

preprints

- Aghbalou, Portier S., Zhou (2021). Tail inverse regression for dimension reduction with extreme response. arXiv:2108.01432.
- Clémençon, Jalalzai, S., Segers (2021). Concentration bounds for the empirical angular measure with statistical learning applications. arXiv:2104.03966.
- Lhaut, S., Segers, (2021). Uniform concentration bounds for frequencies of rare events. arXiv:2110.05826.

Misc.

- S. , Segers (2017). Marginal standardization of upper semicontinuous processes. with application to max-stable processes. Journal of Applied Probability.

Outline

Outline

Extreme Values: why bother?

'Il est impossible que l'improbable n'arrive jamais'

Emil Julius Gumbel, 1891-1966

1934 flood at Port Pirie, Australia

Ideas behind Extreme Value Theory

Theory: Under minimal assumptions, distributions of maxima/excesses converge to a certain class.

Modelling: Use those limits to model maxima/excesses above large thresholds.

X: random object (variable / vector/ process) $\quad \mathsf{X}_i \overset{i.i.d.}{\sim} \mathsf{X}$.

$$\max_{i=1}^{n} \mathsf{X}_i \quad \overset{d}{\approx} \quad \text{Max-stable} \qquad (n \text{ large})$$

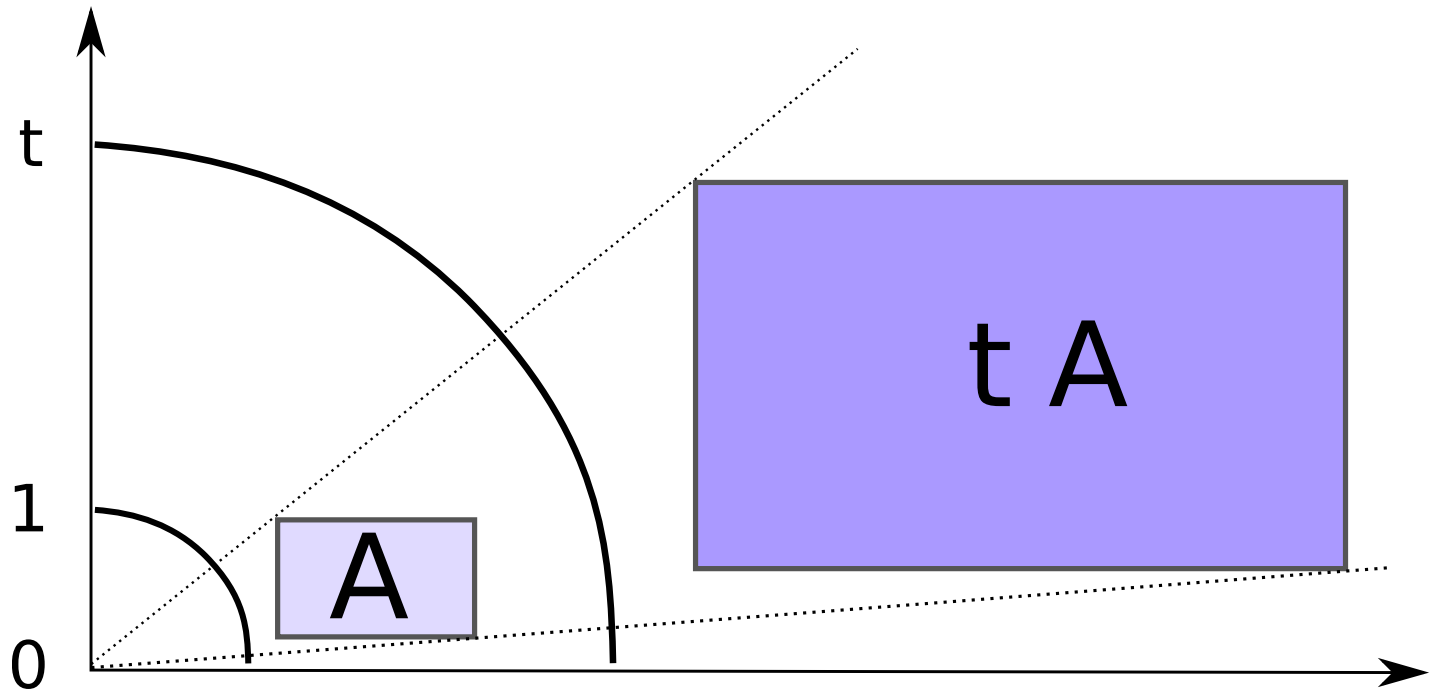
$$\left[\mathsf{X} \;\middle|\; \|\mathsf{X}\| \geq r \right] \quad \overset{d}{\approx} \quad \text{Generalized Pareto} \qquad (r \text{ large})$$

$$\sum_{i=1}^{n} \delta_{(i,\mathsf{X}_i)} \quad \overset{d}{\approx} \quad \text{Poisson point process} \quad (n \text{ large, above large } r)$$

Multivariate regular variation (MRV) (I)

- A random vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is regularly varying if there exists a **limit measure** μ such that for all set $A \subset \mathbb{R}^d$ such that $0 \notin \text{closure}(A)$ and $\mu(\partial A) \neq 0$,

$$\frac{\mathbb{P}(X \in tA)}{\mathbb{P}(\|X\| > t)} \xrightarrow[t\to\infty]{} \mu(A)\,. \tag{MRV}$$



- The limit measure is **homogeneous**: $\mu(rA) = r^{-\alpha}\mu(A)$ for some $\alpha > 0$ (tail index).

Multivariate regular variation (MRV) (II)

- μ **rules the (probabilistic) behaviour of extremes**: if A is far from the origin, then

$$\mathbb{P}(X \in A) \approx \mu(A) \, .$$

Namely

$$\mathbb{P}\left(X \in tA\right) = L(t)\mu(tA),$$

with L a slowly varying function, even $L(t) \xrightarrow[t\to\infty]{} C$ (Constant) after suitable marginal standardization.

- Examples: Max stable vectors with standardized margins, multivariate Student, ...

- In practice: preliminary **componentwise standardization** is often necessary: then (MRV) concerns the standard version V of X,

$$V_j := 1/(1 - F_j(X_j)), \quad V = (V_1, \ldots, V_d).$$

(using empirical $\hat{F}_j$ work well even in theory)

MRV (III): Angular Measure

- Homogeneity of $\mu \to$ polar coordinates are convenient

$$r(x) = \|x\| \quad ; \quad \theta(x) = r(x)^{-1}x.$$

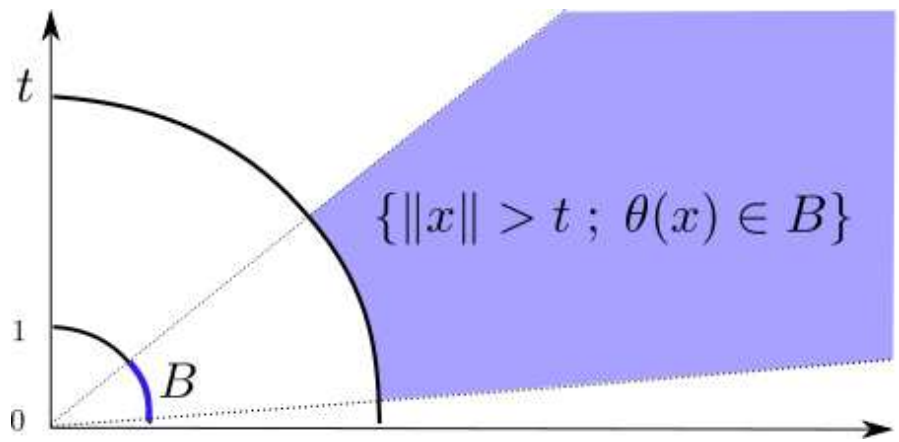
($\|\cdot\|$: any norm, typically $\|\cdot\|_\infty, \|\cdot\|_2$ or $\|\cdot\|_1$)

- **Angular measure** Φ on the $\|\cdot\|$-sphere:

$$\Phi(B) = \mu\{r > 1, \theta \in B\}.$$

- Then μ decomposes as a **product measure**

$$\mu \circ \textit{Polar-transform}^{-1}\{r > t, \theta \in B\} = t^{-\alpha}\Phi(B)$$



- The angular component Φ only is **non-parametric**.

Tail empirical process: asymptotic literature

- Convenient re-writing: (MRV) $\iff$ $tP\big[b(t)\cdot\,\big] \to c\mu(\cdot)$ (vaguely) where $b(t) =$ quantile of order $1-1/t$ of the norm, $c>0$ a constant

Estimating $\mu \approx$ Estimating $tP\big[(b(t)\cdot\,\big]$ (up to a vanishing bias term)

- $X, X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ satisfying (MRV), P_n: empirical measure.

- As $n\to\infty, k\to\infty, n/k\to\infty$: the **'tail empirical process'** converges weakly (1D case, $\alpha = 1$)

$$\sqrt{k}\,\frac{n}{k}(P_n - P)\Big[b(n/k)y,\infty\Big) \xrightarrow[w]{D(0,\infty)} W(y),$$

W: brownian motion, see Resnick 2007, thm. 9.1 + references.

In practice: $b(n/k) \leftarrow X_{(k)}$ (the k^{th} largest order statistic)

- Multivariate extensions are available
(Einmahl et al., 2006, Einmahl & Segers 2009, Einmahl et al. 2012, ...)

Outline

Bounding the deviations of empirical measures

Key historical result: uniform bound on the deviations of the empirical measure P_n from the true law P of X, over a class of sets of controlled complexity

- $\mathcal{A}$: class of subsets of $\mathcal{X}(= \mathbb{R}^d$ here).
- $S_{\mathcal{A}}(n) = \max_{(x_1,\ldots,x_n)\in\mathcal{X}^n} |\{A \cap \{x_1, \ldots, x_n : A \in \mathcal{A}\}|$
  $= \textbf{Shattering coefficient} \leq 2^n$.

Vapnik-Chervonenkis inequality (probability bound)

(Vapnik, Chervonenkis, 71) with probability $\geq 1 - \delta$,

$$\sup_{A\in\mathcal{A}} |P_n - P|(A) \leq 2\sqrt{\frac{2}{n}[log(4/\delta) + \log(S_{\mathcal{A}}(2n))]}$$

- Vapnik-Chervonenkis dimension $d_{\mathcal{A}}$: **complexity control** of $\mathcal{A}$
- Sauer 's lemma : $S_{\mathcal{A}}(n) \leq (n+1)^{d_{\mathcal{A}}} \ll 2^n$.

Application to classification

Control of the generalization risk in classification via ERM

- A classifier: a function $g : \mathcal{Z} \mapsto \{-1, 1\}$, $\mathcal{Z}$: feature space.
- Choose a family of such classifiers $\mathcal{G}$ ($\sim$ a 'model').
- $\mathcal{G}$ is 1-to-1 with $\mathcal{A} = \{A_g = \{(z, y) : g(z) \neq y\}, g \in \mathcal{G}\}$
- Empirical risk $R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(Z_i) \neq Y_i\} = P_n(A_g)$.
- $\sup_{g \in \mathcal{G}} |R - R_n|(g) = \sup_{A \in \mathcal{A}} |P_n - P|(A)$

$\rightarrow$ upper bounds on the generalization risk R.

Outline

Outline

Our wish list

- Recall the asymptotic convergence rate:

$$\sqrt{k}\frac{n}{k}(P_n - P)(b(n/k)A) \to \text{non degenerate Gaussian r.v.}$$

with $P(b(n/k)A) = O(k/n)P(A)$, $A \notin \bar{A}$, (b is a quantile function)

- in other terms, with $p = k/n$

$$\sqrt{pn}\frac{1}{p}(P_n - P)(A_p) \to \text{non degenerate Gaussian r.v.}$$

where
  - $P(A_p) = O(p)P(A)$
  - $np \approx$ number of points X_i in the extreme regions used for estimation.

- Reasonable hope: prove that with high probability,

$$\frac{1}{p}\sup_{A\in\mathcal{A}_p}|P_n - P|(A) \leq O\left(\sqrt{\frac{C \text{ or } d_{\mathcal{A}}\log(np)}{np}}\right)$$

where $\mathcal{A}_p$: a class of sets with low probability $O(p)$ and VC-dim $d_{\mathcal{A}}$

Existing literature: normalized VC inequalities

- Vapnik 2015, Bousquet et al. 2003:
  with probability $1 - 2\delta$,

$$\sup_{A \in \mathcal{A}} \left| \frac{P_n(A) - P(A)}{\sqrt{P(A)}} \right| \leq 2\sqrt{\frac{\log S_{\mathcal{A}}(2\boldsymbol{n}) + \log \frac{4}{\delta}}{n}},$$

- Consequence, with $p = \sup_{A \in \mathcal{A}} P(A)$

$$\frac{1}{p} \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \leq 2\sqrt{\frac{\log S_{\mathcal{A}}(2\boldsymbol{n}) + \log \frac{4}{\delta}}{np}},$$

How to replace the $S_{\mathcal{A}}(n)$ term with $S_{\mathcal{A}}(np)$ or $d_{\mathcal{A}} \log(np)$ or C?

A VC-type inequality for rare classes

(Goix, S. , Clémençon, 2015)

Theorem: Supremum deviation on low probability regions

Let $X_1, \ldots, X_n$ be *i.i.d.* realizations of a r.v. X with distribution P and let $\mathcal{A}$ be a VC-class of sets with VC-dimension $d_{\mathcal{A}}$. Consider the class union $\mathbb{A} = \cup_{A \in \mathcal{A}} A$, and let $p = P(\mathbb{A})$. Then there is a universal constant C such that for all $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\frac{1}{p} \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \quad \leq \quad C \left[\sqrt{\frac{d_{\mathcal{A}}}{np} \log \frac{1}{\delta}} + \frac{1}{np} \log \frac{1}{\delta} \right] . \qquad (1)$$

Tool #1: 'Bernstein-bounded difference' inequality

(Mc Diarmid, 98)

- Recall Bernstein inequality for $Z = \sum_1^n X_i$, $(X_i)_i$ *i.i.d.*,
  $v = \mathbb{V}\mathrm{ar}(Z), \qquad |X_i - \mathbb{E}(X_i)| < b,$

$$\mathbb{P}(Z - \mathbb{E}(Z) > t) \leq \exp\left(-\frac{t^2/2}{v + bt/3}\right)$$

Theorem (Mc Diarmid, 98)

The above inequality is also true for $Z = f(X_1, \ldots, X_n)$, with

- Variance term v maximum sum of variances:
  $v = \sup_{x_1,\ldots,x_n} \sum_{k=1}^n v_k(x_1, \ldots, x_{k-1})$, with

$$v_k(x_1, \ldots, x_{k-1}) = \mathbb{V}\mathrm{ar}\Big(\mathbb{E}(Z|X_k, x_{1:k-1})\Big)$$

- maximum positive deviation $b = \sup_{k,x_{1:k}} \mathrm{dev}_k^+(x_{1:k-1})$, with

$$\mathrm{dev}_k^+(x_{1:k-1}) = \sup_{x_k} \mathbb{E}(Z|X_k = x_k, x_{1:k-1}) - \mathbb{E}(Z|x_{1:k-1})$$

Mc Diarmid's Bernstein/Bounded difference inequality applied to rare classes

- Set $Z = \sup_{A \in \mathcal{A}} |P_n - P|(A)$, $p = P(\cup_{A \in \mathcal{A}} A)$.

Lemma **(Goix, S. Clémençon, 2015)**

The maximum sum of variances v and maximum positive deviation b involved in Mc Diarmid's bound (in the previous slide) satisfy

$$v \leq 2p/n \, ; \qquad b \leq 1/n$$

Corollary: concentration of the maximum deviations

$$\text{With proba } \geq 1-\delta, \quad p^{-1} \sup_{A \in \mathcal{A}} |P_n - P|(A)$$

$$\leq \mathbb{E}\left(p^{-1} \sup_{A \in \mathcal{A}} |P_n - P|(A) \right) + \sqrt{\frac{4 \log(1/\delta)}{np}} + \frac{2}{3np} \log(1/\delta)$$

Tool #2: 'Conditioning trick'

(Lhaut, S., Segers, 21+, Goix, S. Clémençon, 15, see also tail empirical process literature)

- Let $K = \#\{i : X_i \in \mathbb{A}\} \sim \text{Binomial}(n, p)$.

$$\left[\left(P_n(A) \right)_{A \in \mathcal{A}} \mid K = k \right] \stackrel{d}{=} \left(\frac{k}{n} P_k^Y(A) \right)_{A \in \mathcal{A}}$$

where $P_k^Y(A) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_A(Y_i)$, $Y_i \overset{i.i.d.}{\sim} P(\cdot \cap \mathbb{A})/P(\mathbb{A})$.

Bounding the expected maximal deviations

$$\mathbb{E}\left(\frac{\sup_A |P_n - P|(A)}{p} \right) \leq C \sqrt{\frac{d_{\mathcal{A}}}{np}}$$

(GSC15, using chaining results)

(C universal constant)

$$\mathbb{E}\left(\frac{\sup_A |P_n - P|(A)}{p} \right) \leq \sqrt{\frac{2 d_{\mathcal{A}} \log(2np+1)}{np}} + \frac{3}{\sqrt{np}}$$

(LSS21+, using VC inequality for expectation)

Outline

Classification in extreme regions: set up

(Jalalzai, Clémençon, S., 2018)

- Random pair (V, Y), $V \in \mathbb{R}_+^d$: observed input/features,
  $Y \in \{-1, 1\}$: label to be predicted.

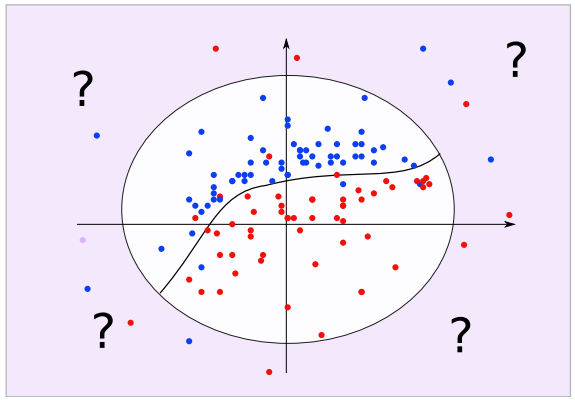
- Assumption (MRV), standard case, for each class:

$$t\mathbb{P}(t^{-1}V \in B \mid Y = \pm 1) \xrightarrow[t\to\infty]{} \mu_{\pm}(B)$$

  (no standardization required or margins known).

- Classification loss $L_t(g)$ for a classifier $g : \mathbb{R}^d \to \{-1, 1\}$ above level t:

$$L_t(g) = \mathbb{P}\{Y \neq g(V) | \ \|V\| > t\},$$

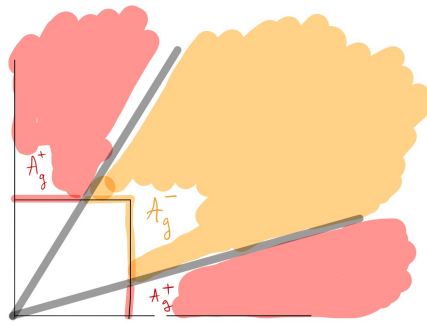


Main findings (Jalalzai, Clémençon, S., 2018)

1. For an **angular classifier** $g(v) = g\big(\theta(v)\big)$,

$$L_\infty(g) = \lim_{t\to\infty} L_t(g) \text{ exists.}$$

2. There **exists an angular classifier** g^* **minimizing** $\limsup_t L_t(g)$ over all possible classifiers.



3. The ERM strategy for $\mathcal{G} \subset$ **angular classifiers,**

$$\min_{g\in\mathcal{G}} \sum_{i=1}^{n} \mathbb{1}\{g(V_i) \neq Y_i, \|V_i\| > \hat{t}_{n/k}\}$$

with $\hat{t}_{n/k}$: $1 - k/n$ empirical quantile of $\|V\|$, yields a $\hat{g}$ s.t.

$$L_\infty(\hat{g}) - L_\infty(g^*) \leq 4C\sqrt{d_{\mathcal{G}} \ln(1/\delta)/k} + \text{bias}(n,k) + \text{bias}(\mathcal{G}) + O(1/k).$$

Application: Natural Language Processing (NLP)

Jalalzai, Colombo, Clavel, Gaussier, Varni, Vignon, S. (2020)

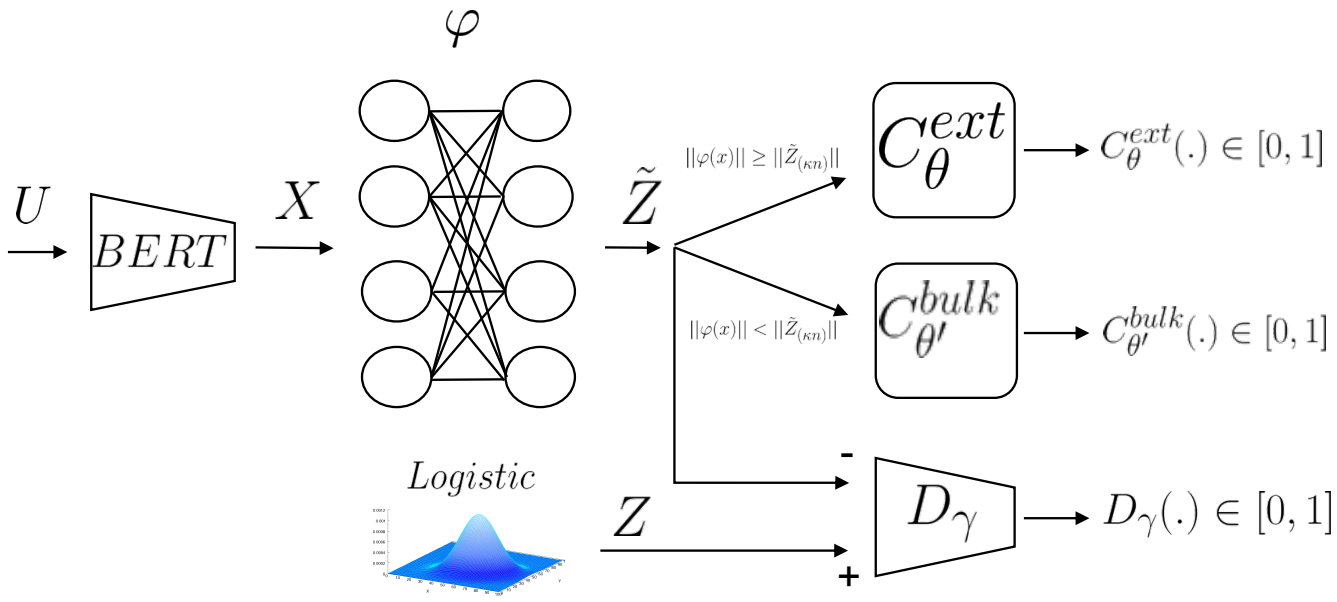
- Extension of the previous framework to datasets who are **NOT** regularly varying.
- Dataset: text embeddings (BERT). X = vector in $\mathbb{R}^d$, d large (768).
- label Y = positive/negative sentiment.
- Two goals:

  (i) improved classification in low probability regions of $\mathcal{X}$

  (ii) label preserving data augmentation

Learning a regularly varying representation for NLP

- Key step: adversarial strategy, (Goodfellow et al. 2014) mixed loss function involving
  - $0-1$ loss in extreme/ non-extreme regions
  - Jensen-Shannon divergence between the learnt representation and a Max-stable multivariate Logistic, $\neq$ common practice Gaussian



- Output: a transformed vector $\tilde{Z} = \varphi(X)$ which is (experimentally) regularly varying (low correlations $\theta(\tilde{Z}) \leftrightarrow \|\tilde{Z}\|$).

Outline

Other related results/applications

- Uniform control of the deviations of the empirical *c.d.f.* of the limit measure (=the standard tail dependence function):
  Goix, S. Clémençon, 15

- Mass-Volume set estimation of the angular component (angular measure) for anomaly detection far from the origin:
  Thomas, Clémençon, Gramfort, S. 2017

- Uniform deviations of the empirical angular measure of extremes (unknown marginal distributions): Clémençon, Jalalzai, S. Segers, 21+
  $\rightarrow$ Extension of the classification set-up to (X, Y) with marginal distribution X_j unknown

- Alternative bounds with explicit constants: Lhaut, Segers, S. 21+

Outline

Motivation

- Multivariate heavy-tailed random vector $X = (X_1, \ldots, X_d)$
  e.g. spatial field (temperature, precipitation), asset (negative) prices

- Focus on the tail: $\{x \mid \|x\| > t)$, $t \gg 1$ with $\mathbb{P}(\|X\| > t)$ small.

  Possible goals: modeling and simulation (stress test), anomaly detection/clustering among extreme values, ...

- $d \gg 1$: modeling $\mathrm{Law}(X \mid \|X\| > t)$ unfeasible.

- Dimension reduction problem(s) :

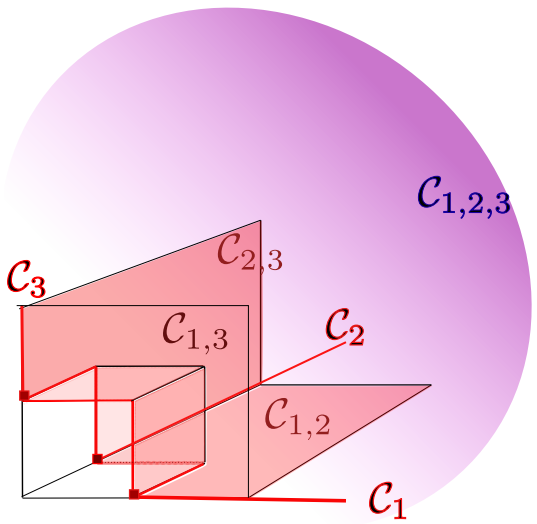
1. Identify the **groups of features** $J \subset \{1, \ldots d\}$ which **may be large together** (while the others stay small), given that one of them is large.
2. Identify **a single low dimensional projection subspace** S_0 such that $\mathrm{Law}(X \mid \|X\| > t) \approx$ concentrated on S_0.

Outline

Sparse support recovery (Goix, S. Clémençon, 2016, 2017)

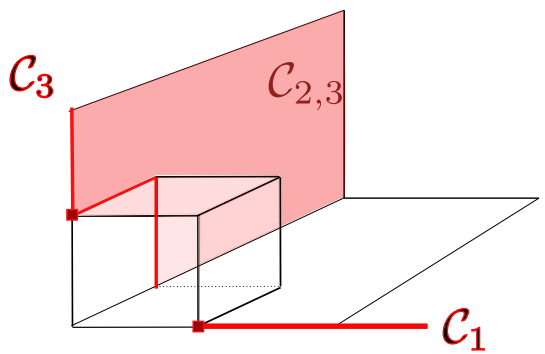
- Reasonable hope: the groups $\{X_j, j \in J\}$'s wich may be simultaneously large are (i) few (ii) small. $\rightarrow$ **sparse angular measure**

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



Full support:

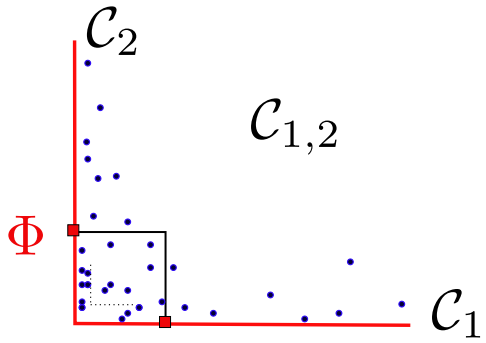
anything may happen



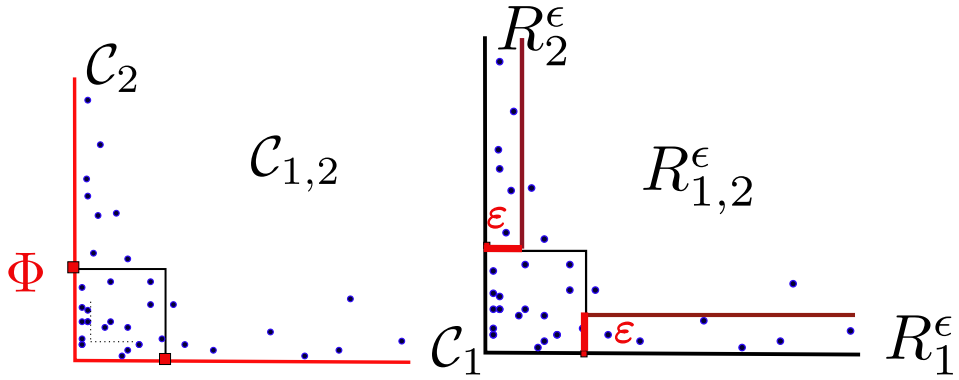
Sparse support

(X_1 not large if X_2 or X_3 large)

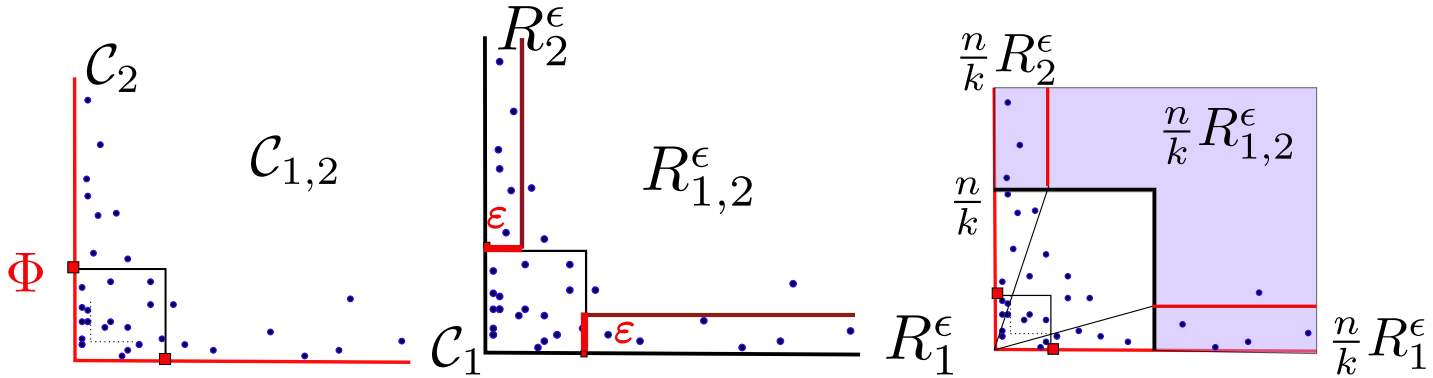
Identifying non empty subspaces (Goix et al. 2016)



Identifying non empty subspaces (Goix et al. 2016)

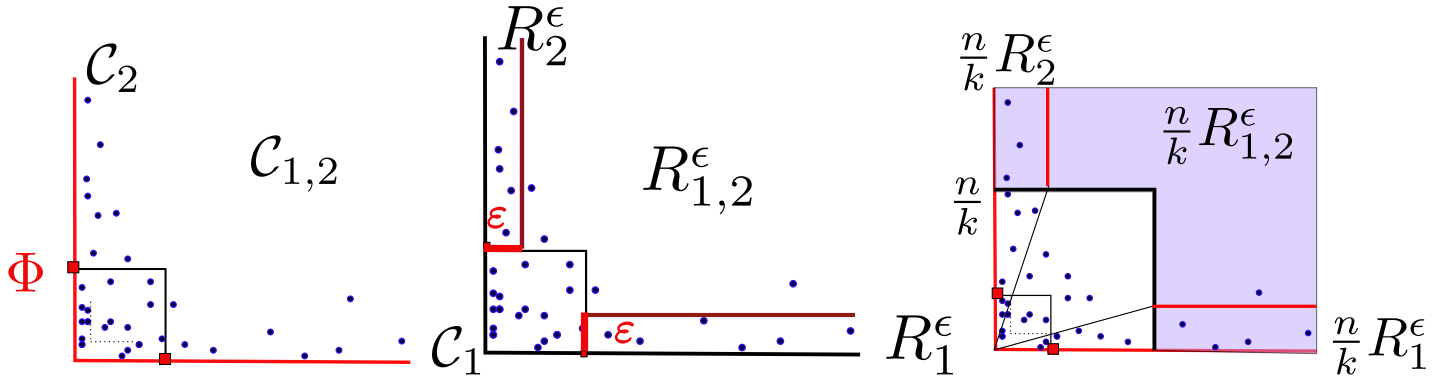


Identifying non empty subspaces (Goix et al. 2016)



Two parameters: (i) Tolerance parameter ε (otherwise empirical measure of subspaces $= 0$) ; (ii) k (number of observations considered as extremes)

Identifying non empty subspaces (Goix et al. 2016)



Two parameters: (i) Tolerance parameter ε (otherwise empirical measure of subspaces $= 0$) ; (ii) k (number of observations considered as extremes)

Theorem (Goix, S., Clémençon, 2016)

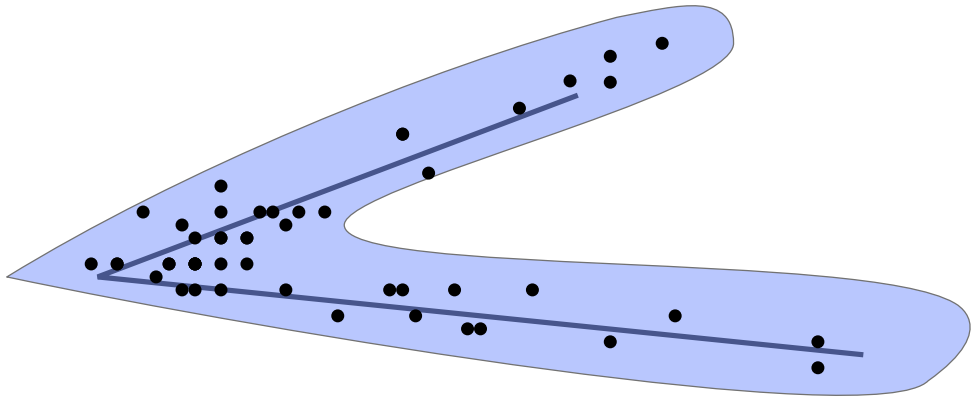
If the margins F_j are continuous and if the density of the angular measure is bounded by $M > 0$ on each subface (infinity norm),
There is a constant C s.t. for any n, d, k, $\delta \geq e^{-k}$, $\varepsilon \leq 1/4$, w.p. $\geq 1-\delta$,

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

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

Application (I): anomaly detection

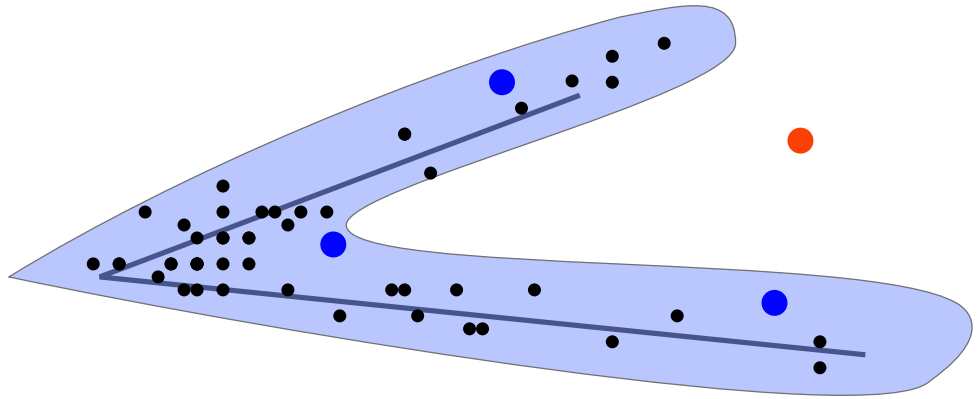
(Goix, S, Clémençon (2016), similar spirit without dimensionality reduction in Thomas, Clémençon, Gramfort, S. (2017))



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

Application (I): anomaly detection

(Goix, S, Clémençon (2016), similar spirit without dimensionality reduction in Thomas, Clémençon, Gramfort, S. (2017))



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

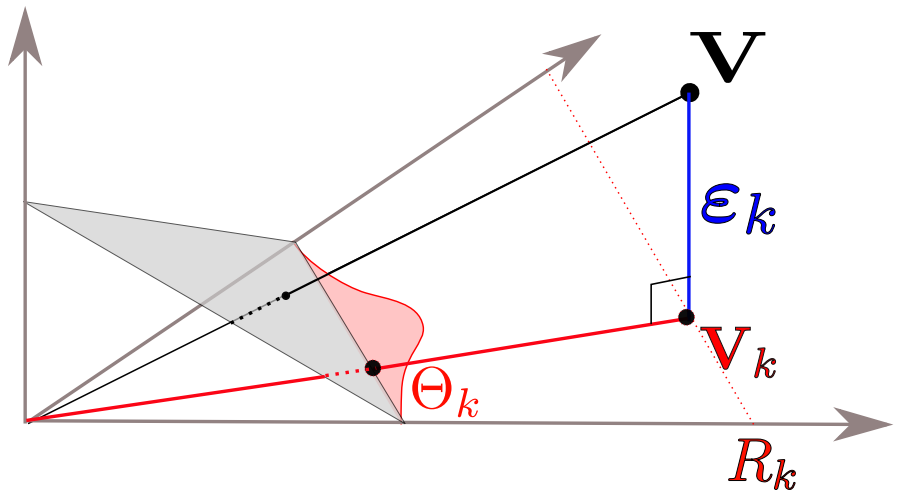
If 'normal' data are heavy tailed, **Abnormal** $\not\Leftrightarrow$ **Extreme** .
There may be **extreme** 'normal data'.

How to distinguish between large anomalies and normal extremes?

Application (II): Clustering of anomalies

(Chiapino, Clémençon, Feuillard, S. 2019)

- Motivation: Airbus flight data (times series/logs): dimension $= O(100)$. How to cluster together similar anomalies to help interpretation?
- Here: an anomaly = any X such that $\|X\|$ is large.
- Idea: Use the subspaces output by DAMEX (or variants) as the components of a pre-asymptotic mixture model



- Use posterior probabilities $\mathbb{P}(X_i \in \text{Component j})$ to construct a pairwise similarity matrix between extreme data.
- Use *e.g.* graph clustering techniques to issue final clusters

Dimension reduction: Alternative methods/extensions

- Simpson, Wadsworth, Tawn, 2020: hidden regular variation
- Chiapino, S. , 2018 ; Chiapino, S., Segers 2019: Gathering 'close' subspaces + asymptotic analysis and statistical tests (different null H_0 considered).
- Clustering: Chautru 2015, Janßen & Wan 2020, Jalalzai & Leluc 2021, Fomichov & Ivanovs 2021+
- Alternative definition of sparsity (Euclidean projections on the simplex) Meyer & Winterberger 2021+
- Graphical models for extremes (Hitz, Evans, 2016, Engelke, Hitz 2020, Engelke, Volgushev 2020+)
- PCA: Drees, S. 2021 (some theory) Cooley, Thibaud, 2019 (specific preliminary transformation, with applications), see also Jiang, Cooley, Wehner 2020, Rohrbeck, Cooley 2021+

Outline

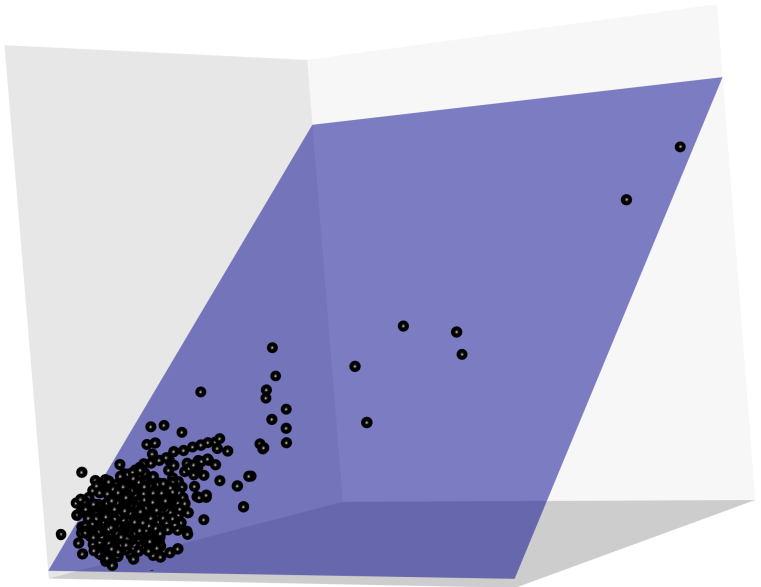
Drees, S. (2021)'s PCA: Context, Motivation

- $(X_1, \ldots, X_d)$ a multivariate r. vector with tail index $\alpha > 0$ and limit measure μ

Motivating Assumption (not necessary for our results)

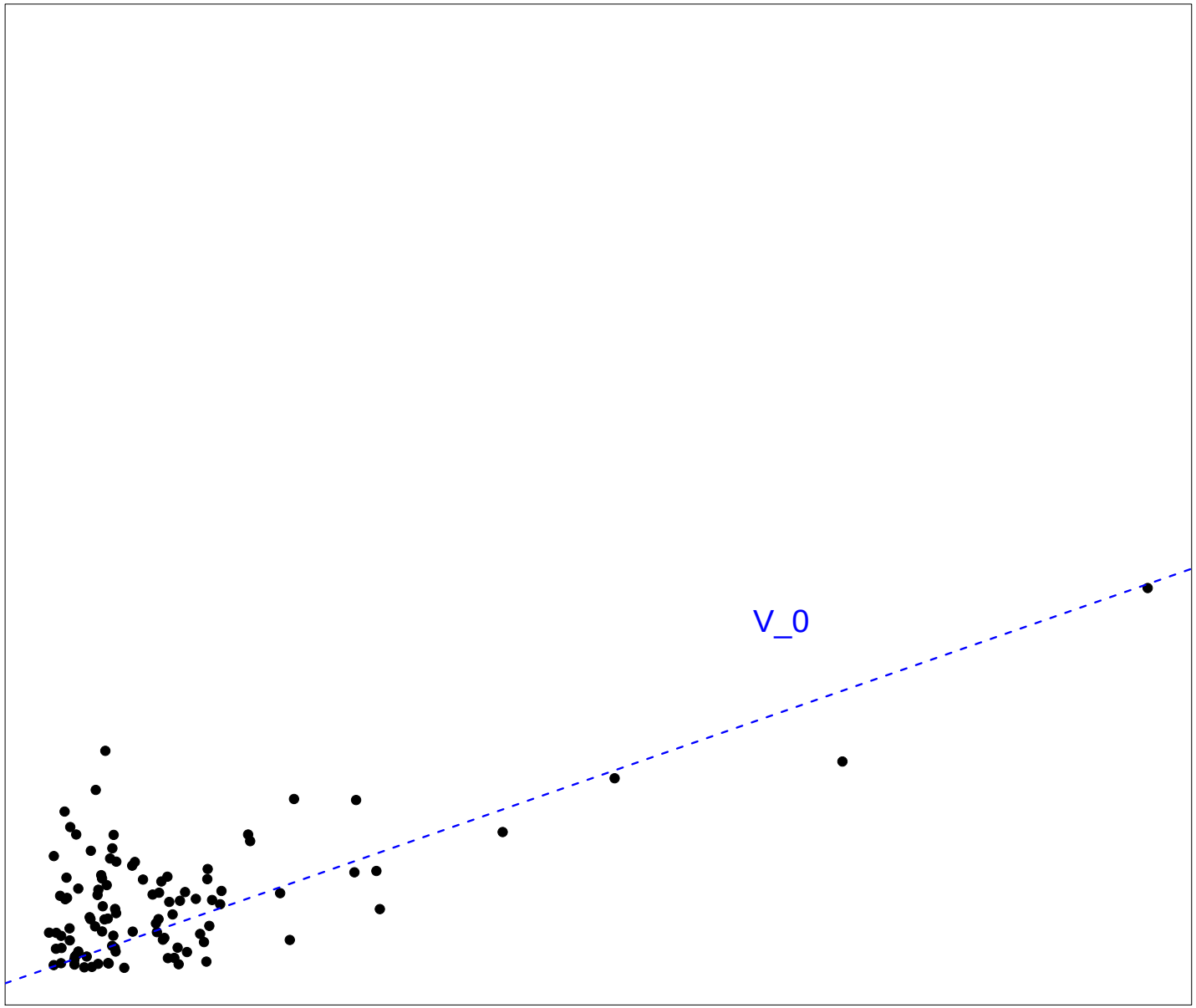
The vector space $S_0 = \text{span}(\text{supp}\,\mu)$ generated by the support of μ has dimension $p < d$.

- **Purpose of this work**: Recover S_0 from the data, with guarantees concerning the reconstruction error.

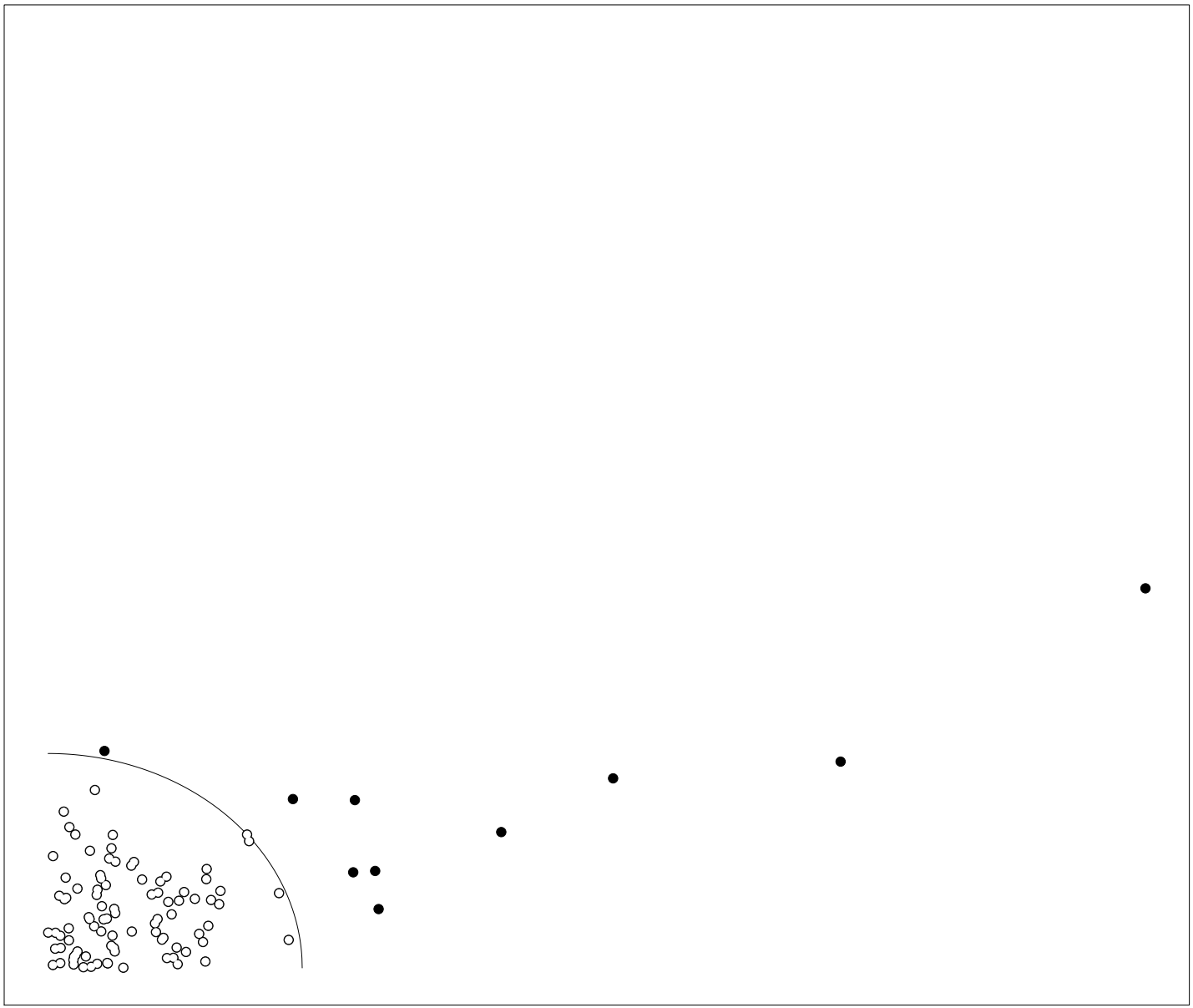


Toy example

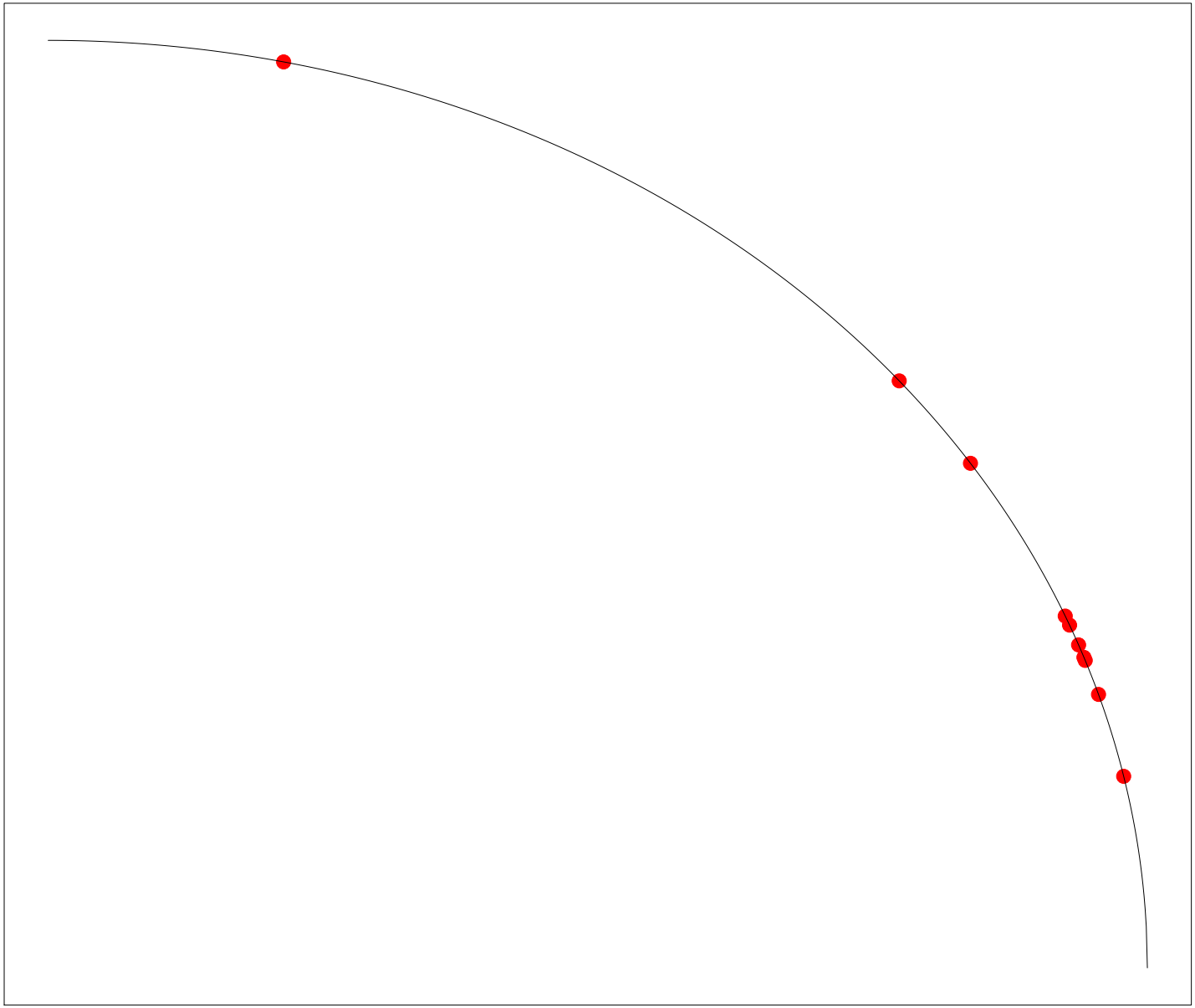
Toy example



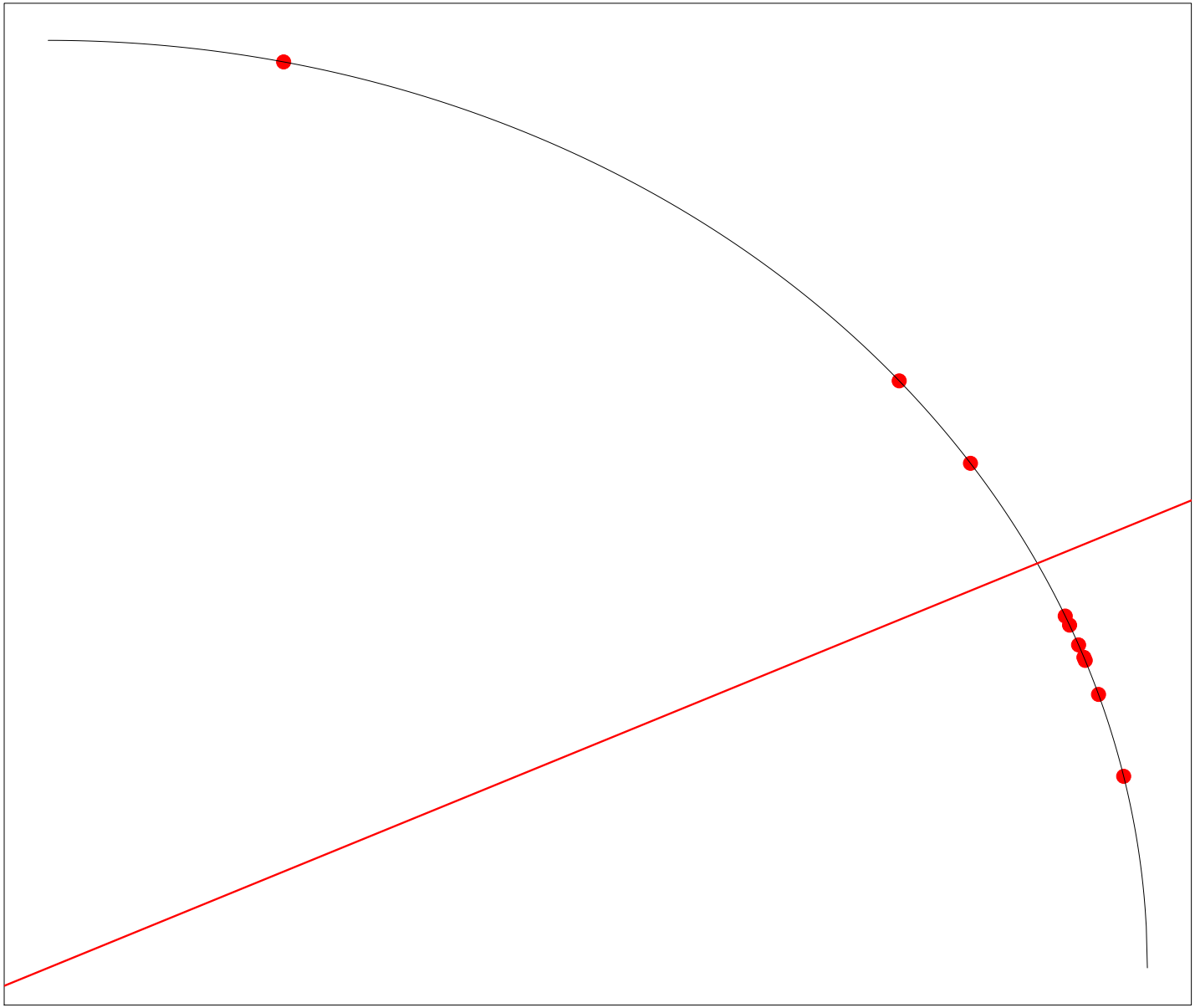
Toy example, proposed method



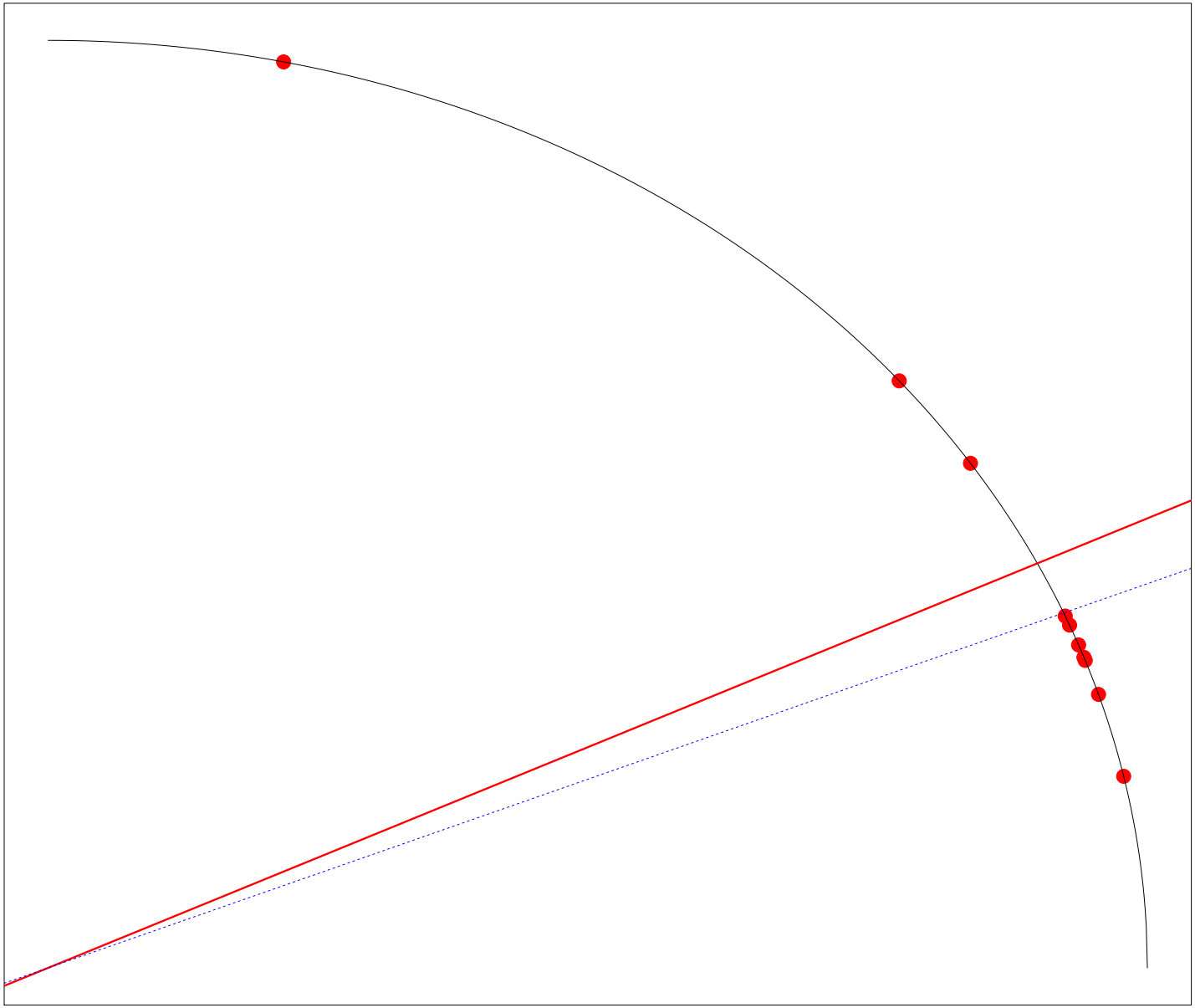
Toy example, proposed method



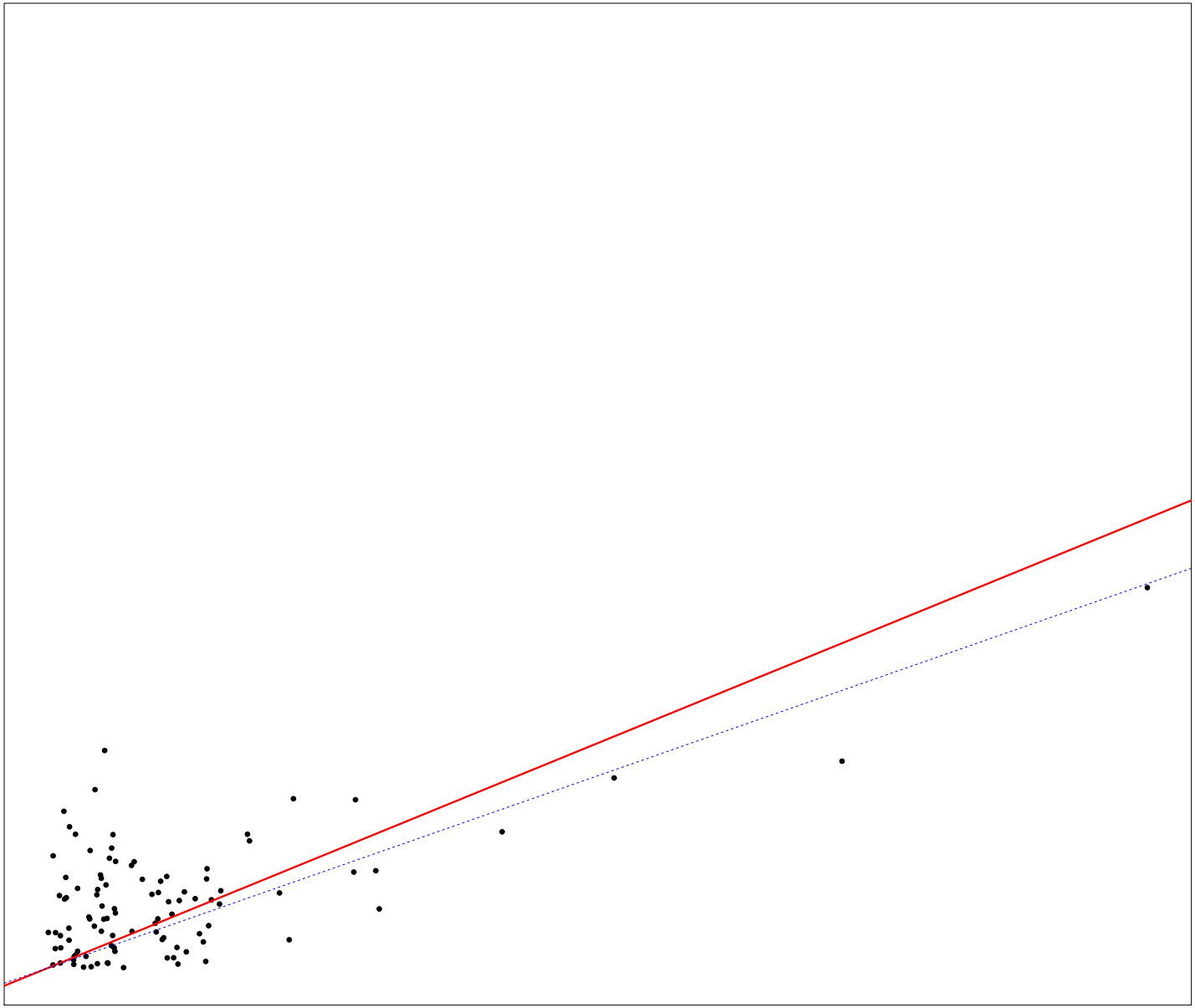
Toy example, proposed method



Toy example, proposed method



Toy example, proposed method



Uniform risk bound on the L_2 reconstruction error

- $t_{n,k}$: quantile of level $1 - k/n$ for $\|X\|$, $\Theta = \|X\|^{-1} X$
- $\Xi_t := \mathbb{E}\left(\|\Theta\|^4 - \pi_t \operatorname{tr}(\Sigma_t^2) \mid \|X\| > t\right)$; $\Sigma_t = \mathbb{E}\left(\Theta\Theta^\top \mid \|X\| > t\right)$

Theorem (Drees, S., 20++), simplified version

With probability at least $1 - \delta$,

$$\sup_{S \in \mathcal{E}_p} |R_{n,k}(S) - R_{t_{n,k}}(S)| \leq \left[\frac{p \wedge (d-p)}{k} \Xi_{t_{n,k}}\right]^{1/2} + \ldots$$
$$\ldots \left[\frac{8}{k}(1 + k/n)\log(4/\delta)\right]^{1/2} + \ldots$$
$$\ldots \frac{4\log(4/\delta)}{3k}.$$

(tools: McDiarmid, 98's Bernstein-type bound + arguments from Blanchard et al. 07)

- **NB**: unknown term Ξ_t: an alternative statement is proven with only empirical quantities in the upper bound.

Outline

Perspectives

- Unsupervised dimensionality reduction $\rightarrow$ supervised? (explain large values of Y given a low dimensional representation of the multivariate input X) (Aghbalou, Portier, S. , Zhou, 21+)

- Alternative concentration tools: Talagrand inequality /Bousquet inequality could replace McDiarmid's Bernstein-type one (they have a variance term)

- Tightness of the bounds? (Lower bounds?)

- Universal upper-bounds $\rightarrow$ data-dependent bounds? (e.g.Cross-Validation), algorithmic stability?
  ongoing PhD Anass Aghbalou.
  $\rightarrow$ model selection?

- From multivariate extremes to infinite dimensional one: dimensionality reduction tools for extremes of time series/ functional data? (ongoing PhD Nathan Huet)

Bibliography I

Concentration / statistical learning / machine learning

- Lugosi (2002). Pattern classification and learning theory. In Principles of nonparametric learning (pp. 1-56). Springer, Vienna.
- Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: A survey of some recent advances. ESAIM: probability and statistics, 9, 323-375.
- McDiarmid (1998). Concentration, Probabilistic methods for algorithmic discrete mathematics, 195–248. Algorithms Combin, 16.
- Blanchard, G., Bousquet, O., & Zwald, L. (2007). Statistical properties of kernel principal component analysis. Machine Learning, 66(2-3), 259-294.
- Goodfellow et al. (2014). Generative adversarial nets. NeurIPS proceedings, 27

Concentration of extreme order statistics, tail index estimation

- Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. Electronic Communications in Probability, 17.
- Boucheron, S. and Thomas, M. (2015). Tail index estimation, concentration and adaptivity. Electronic Journal of Statistics, 9(2):2751–2792.
- Carpentier, A. and Kim, A. K. (2015). Adaptive and minimax optimal estimation of the tail coefficient. Statistica Sinica, pages 1133–1144

Bibliography II

Extreme Value Theory

- Resnick (1987). Extreme Values, Regular Variation, and Point Processes. *Springer Series in Operations Research and Financial Engineering*.
- Resnick (2007). Heavy-tail phenomena: probabilistic and statistical modeling. *Springer Science & Business Media.*
- Einmahl, J. and Mason, D. (1988). Strong limit theorems for weighted quantile processes. Annals of Probability, 16(4):1623–1643.

Dimensionality reduction for extremes

- Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. Electronic journal of statistics, 9(1), 383-418.
- Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. Biometrika, 107(3):513–532.
- Cooley, D., & Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. Biometrika, 106(3), 587-604.
- Rohrbeck, C., & Cooley, D. (2021). Simulating flood event sets using extremal principal components. arXiv preprint arXiv:2106.00630.

Bibliography III

- Jiang, Y., Cooley, D., & Wehner, M. F. (2020). Principal Component Analysis for Extremes and Application to US Precipitation. Journal of Climate, 33(15), 6441-6451.
- Jalalzai, H., & Leluc, R. (2021). Feature Clustering for Support Identification in Extreme Regions. In International Conference on Machine Learning (pp. 4733-4743). PMLR.
- Janßen, A., & Wan, P. (2020). k-means clustering of extremes. Electronic Journal of Statistics, 14(1), 1211-1233.
- Engelke, S., & Hitz, A. S. Graphical models for extremes. arXiv:1812.01734.
- Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. Extremes, 21(1), 57-95.
- Hitz & Evans (2016). One-component regular variation and graphical modeling of extremes. *Journal of Applied Probability*
- Engelke, Hitz (2020). Graphical models for extremes. *JRSS-B*
- Engelke, Ivanovs (2020). Sparse structures for multivariate extremes. Annual Review of Statistics and its Application, 8.

Bibliography IV

Anomaly detection and Extremes (recent)

- Vignotto, E., & Engelke, S. (2020). Extreme value theory for anomaly detection–the GPD classifier. Extremes, 23(4), 501-520.
- Siffer, A., Fouque, P. A., Termier, A., & Largouet, C. (2017, August). Anomaly detection in streams with extreme value theory. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1067-1075).
- Suboh, S., & Aziz, I. A. (2020, November). Anomaly Detection with Machine Learning in the Presence of Extreme Value-A Review Paper. In 2020 IEEE Conference on Big Data and Analytics (ICBDA) (pp. 66-72). IEEE.