Feature Clustering for Support Identification in Extreme Regions

Hamid Jalalzai, Rémi Leluc

June 29th, 2021

Extreme Value Analysis

Today's Agenda

Introduction

Mathematical Background

ERM in Extreme Regions

Numerical Experiments

Conclusion

Introduction

Motivations

- Random vector $X = (X^1, \ldots, X^p) \in \mathbb{R}_+^p$, $p \geq 1$ with Pareto margins.

e.g. spatial fields, asset prices, in risk management: sensor networks (road/internet traffic) or financial assets

Motivations

- Random vector $X = (X^1, \ldots, X^p) \in \mathbb{R}_+^p$, $p \geq 1$ with Pareto margins.

e.g. spatial fields, asset prices, in risk management: sensor networks (road/internet traffic) or financial assets

- Extreme regions $\{x \in \mathbb{R}^p, \|x\| > t\}$, $t \gg 0$.

e.g. traffic jam, flood, network congestion, falling price

Motivations

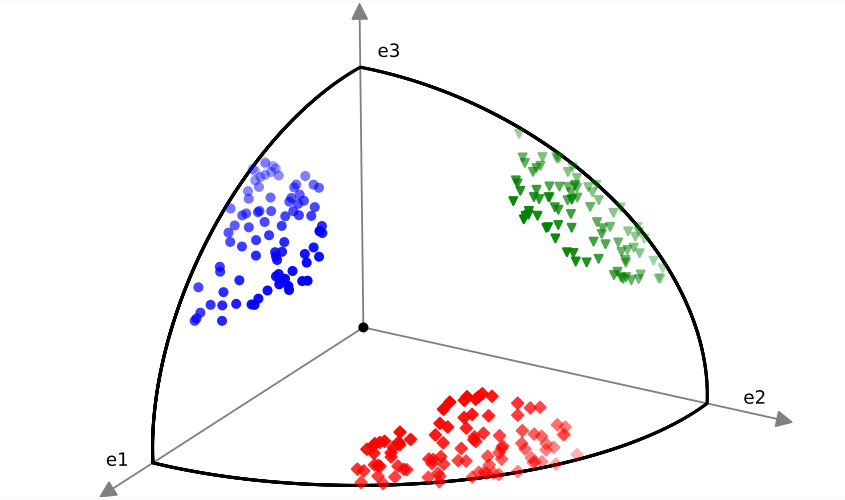
- Random vector $X = (X^1, \ldots, X^p) \in \mathbb{R}_+^p$, $p \geq 1$ with Pareto margins.

e.g. spatial fields, asset prices, in risk management: sensor networks (road/internet traffic) or financial assets

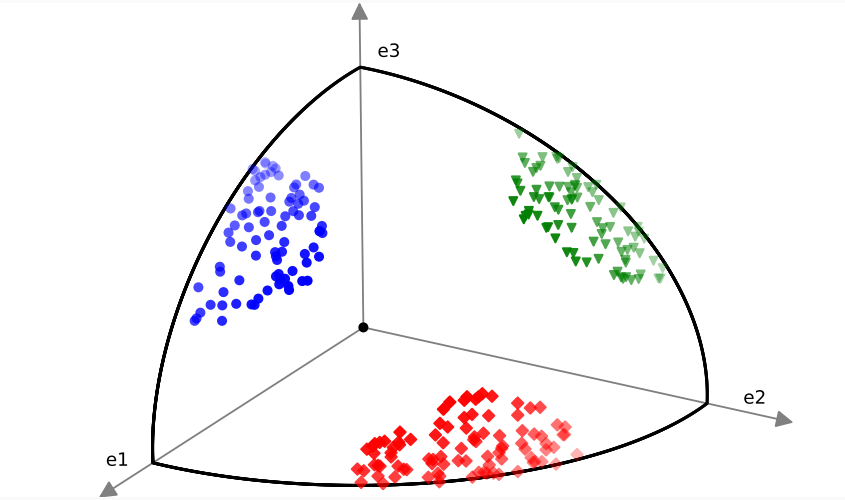
- Extreme regions $\{x \in \mathbb{R}^p, \|x\| > t\}$, $t \gg 0$.

e.g. traffic jam, flood, network congestion, falling price

- Our interest lies in the extreme dependence: Identifying the features X^j's contributing to X being extreme $\rightarrow$ feature clustering.



Goal: Identify Clusters of Features



Goal

Identify **clusters of features** $K \subset [\![1, p]\!]$ such that the variables $\{X^j : j \in K\}$ may be large while the other variables X^j for $j \notin K$ simultaneously remain small.

Assume that $K_i \cap K_j = \emptyset$ for $i \neq j$ (e.g. smart grids, portfolio diversity,...), $|K_i| > 1$ for $i \leq m$.

Our Intuition

Search a subset K of features such that the ℓ_1-norms of X and its restriction $X^{(K)}$ are almost equal *i.e.*

$$\|X\|_1 \approx \|X^{(K)}\|_1.$$

Example: $p = 7$ and $K = \{3, 4, 5\}$

$$X = (*, *, \mathbf{*}, \mathbf{*}, \mathbf{*}, *, *), \quad \|X\|_1 = 3\mathbf{*} + 3*$$
$$X^{(K)} = (0, 0, \mathbf{*}, \mathbf{*}, \mathbf{*}, 0, 0), \quad \|X^{(K)}\|_1 = 3\mathbf{*}$$

Related work

- **Analysis of the (Sparse) Dependence Structure**

Chautru (2015); Chiapino and Sabourin (2016); Goix et al. (2016); Engelke and Hitz (2018); Chiapino et al. (2019)

- **Dimension reduction techniques (PCA and derivatives)**

(Wold et al., 1987; Cutler and Breiman, 1994; Tipping and Bishop, 1999; Cooley and Thibaud, 2019; Drees and Sabourin, 2019)

- **Sparse support of multivariate extremes**

(De Haan and Ferreira, 2007; Chiapino and Sabourin, 2016; Meyer and Wintenberger, 2019; Engelke and Ivanovs, 2020)

Our Contributions

Problem

How to jointly find the extremes' structure dependence ?

• **Optimization** approach to perform subspace clustering of extreme regions: Empirical Risk Minimization (ERM) on the probability simplex with a non-asymptotic bound.

• **Algorithm**: find a sparse representation for the structure dependence.
Multivariate **EX**treme **I**nformative **C**lustering by **O**ptimization

• **Numerical Experiments** on both *feature clustering* and *anomaly detection* tasks in extreme regions.

Mathematical Background

Multivariate Regular Variation

$X = (X^1, \ldots, X^p)$ with continuous marginal cdf's $F^1, \ldots, F^p$

Definition: Multivariate regular variation (Resnick (1987))

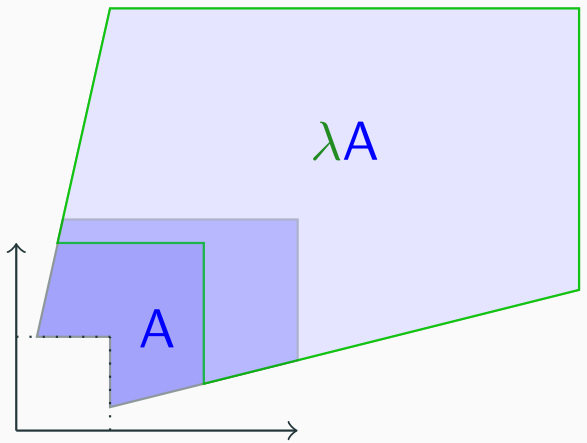
For subsets of $\mathbb{R}_+^p \setminus \{0\}$ bounded away from origin:

$$t\mathbb{P}\{t^{-1}X \in \cdot\} \xrightarrow[t\to\infty]{} \mu(\cdot),$$

The limit measure μ on $\mathbb{R}_+^d \setminus \{0\}$ is **homogeneous**:

$$\forall \lambda > 0, \qquad \mu(\lambda A) = \lambda^{-1}\mu(A)$$

with $0 \notin A, \mu(\partial A) = 0$.



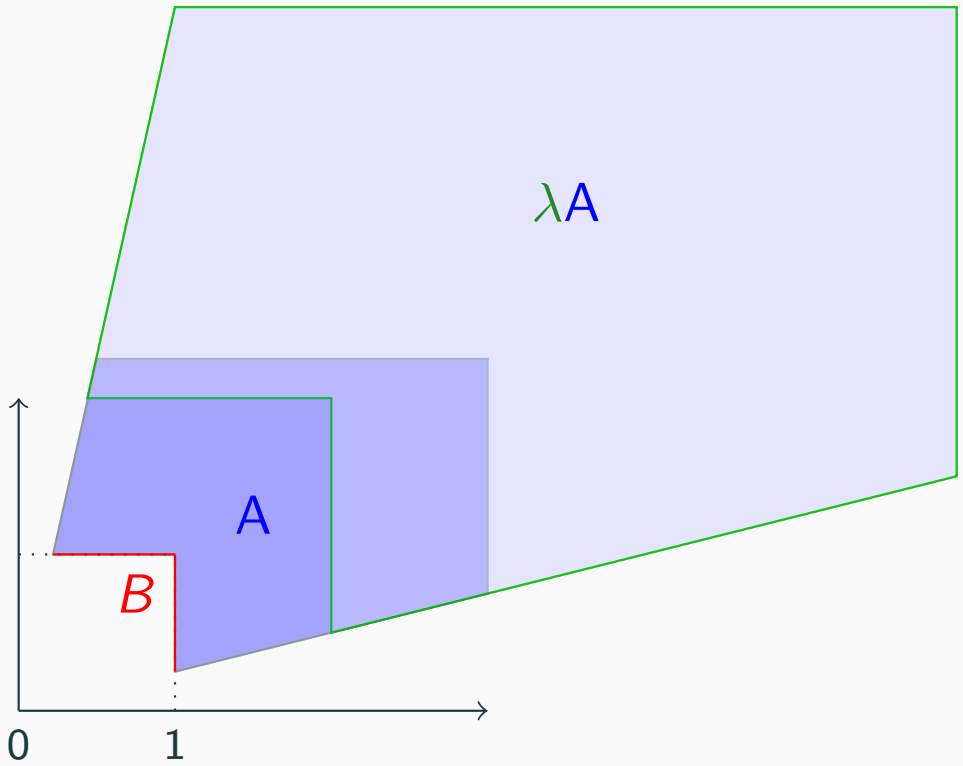
From Exponent Measure to Angular Measure

Angular measure Φ and directions of extremes

Φ is defined on $\mathbb{S} = \{x \in \mathbb{R}_+^d,\ ||x||_\infty = 1\}$,

$$\Phi(B) = \mu(\{x \in \mathbb{R}_+^d, ||x||_\infty \geq 1, \Theta(x) \in B\})$$

with $\Theta(x) = x/||x||_\infty$.



Angular Measure and Feature Clustering

The angular measure Φ characterizes the directions where extremes are more likely to occur.

- The support of $\Phi \rightarrow$ features that are more likely to jointly be large.
- We address the problem of finding different feature clusters $K_j \subset [\![1, p]\!]$ with $j = 1, \ldots, m$ and $m < p$ such that all features in a same subset may be large together.
- Relying on the m clusters of features $K_1, \ldots, K_m$, Φ can be approximated as

$$\Phi(\cdot) \approx \sum_{j=1}^{m} \Phi_{K_j}(\cdot).$$

Each component Φ_{K_j} is concentrated on the subregion given by the features of cluster K_j.

Empirical Risk Minimization (ERM)

- Observed *i.i.d.* copies $z_1, \ldots, z_n \in \mathcal{Z}$ of random variable z
- Loss function $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbb{R}$
- Goal is to minimize the *unknown* true risk $\mathcal{R}(g) = \mathbb{E}_z[\ell(g, z)]$
- Empirical counterpart, for all $g \in \mathcal{G}$,

$$\mathcal{R}(g) = \mathbb{E}_z[\ell(g, z)] \qquad \widehat{\mathcal{R}}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g, z_i).$$

Examples: $z = (x, y)$ with data $x \in \mathcal{X} \subset \mathbb{R}^p$ and label $y \in \mathcal{Y}$.

- (L_2 Regression) $\mathcal{Y} = \mathbb{R}$, $\ell(g, (x, y)) = (y - g(x))^2$
- (Classification) $\mathcal{Y} = \{-1, +1\}$, $\ell(g, (x, y)) = \mathbb{1}\{g(x) \neq y\}$

ERM in Extreme Regions

ERM in Extreme Regions

- $n \geq 1$ *i.i.d* copies $X_1, \ldots, X_n$ of X (Pareto margins)
- Loss function $\ell : \mathbb{R}_+^p \times \mathbb{R}_+ \to \mathbb{R}_+$ measuring the discrepancy between the true extreme dependence structure of X and its prediction $g(X)$.
- Find g to minimize the risk at level t_γ

$$\mathcal{R}_{t_\gamma}(g) = \mathbb{E}_X \left[\ell(X, g(X)) \middle| \|X\|_\infty > t_\gamma \right],$$

ERM in Extreme Regions

- $n \geq 1$ *i.i.d* copies $X_1, \ldots, X_n$ of X (Pareto margins)
- Loss function $\ell : \mathbb{R}_+^p \times \mathbb{R}_+ \to \mathbb{R}_+$ measuring the discrepancy between the true extreme dependence structure of X and its prediction $g(X)$.
- Find g to minimize the risk at level t_γ

$$\mathcal{R}_{t_\gamma}(g) = \mathbb{E}_X \left[\ell(X, g(X)) \,\middle|\, \|X\|_\infty > t_\gamma \right],$$

- Based on the extreme observations $X_{(1)}, \ldots, X_{(k)}$, the empirical risk is

$$\widehat{\mathcal{R}}_k(g) = \frac{1}{k} \sum_{i=1}^{k} \ell\big(X_{(i)}, g(X_{(i)})\big),$$

where $\|X_{(1)}\| \geq \ldots \geq \|X_{(k)}\| \geq \ldots \geq \|X_{(n)}\|$.

- Denote $\mathbf{X} \in \mathbb{R}_+^{k \times p}$ the data matrix of extreme observations.

Representation $g(X)$ and Loss function ℓ

- Approximate $\|X\|_1$ with mixtures of components of X
- Consider the probability simplex $\Delta_p = \{x \in \mathbb{R}_+^p, x_1 + \ldots + x_p = 1\}$ and let $\mathbf{W} \in \mathcal{A}_p^m$ with $m < p$ be a *mixture matrix* (columns belonging to Δ_p).
- Each column $\mathbf{W}^j$ for $j \in [\![1, m]\!]$ is modelling a mixture of components and represents a cluster K_j.

$$\ell(X, W) = \|X\|_1 - \vee_{j=1}^m XW^j$$

Representation $g(X)$ and Loss function ℓ

- Approximate $\|X\|_1$ with mixtures of components of X
- Consider the probability simplex $\Delta_p = \{x \in \mathbb{R}_+^p, x_1 + \ldots + x_p = 1\}$ and let $\mathbf{W} \in \mathcal{A}_p^m$ with $m < p$ be a *mixture matrix* (columns belonging to Δ_p).
- Each column $\mathbf{W}^j$ for $j \in [\![1, m]\!]$ is modelling a mixture of components and represents a cluster K_j.

$$\ell(X, W) = \|X\|_1 - \vee_{j=1}^m XW^j$$

Example: $p = 7, K_1 = \{1, 2\}, K_2 = \{3, 4, 5\}, K_3 = \{6, 7\}$

$$\mathbf{X} = \begin{pmatrix} * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \\ * & * & * & * & * & * & * \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1/2 & 0 & 0 \\ 1/2 & 0 & 0 \\ 0 & 0 & 1/3 \\ 0 & 0 & 1/3 \\ 0 & 0 & 1/3 \\ 0 & 1/2 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}$$

(Non-Convex) Optimization Problem

- For each row X_i, seek a column $j \in [\![1, m]\!]$ for which $\widetilde{X}^j = (X_i \mathbf{W})^j$ is the closest to $\|X_i\|_1$.
- Column index of a good mixture through the mapping

$$\varphi : [\![1, k]\!] \to [\![1, m]\!], \quad \varphi(i) = \underset{1 \leq j \leq m}{\arg\max}\, \widetilde{X}^j_{(i)}$$

- Learn the mixture matrix $\widehat{\mathbf{W}}_k$ such that

$$\widehat{\mathbf{W}}_k \in \underset{\mathbf{W} \in \mathcal{A}_p^m}{\arg\max} \left\{ \frac{1}{k} \sum_{i=1}^{k} (\mathbf{XW})_i^{\varphi(i)} = \frac{1}{k} \sum_{i=1}^{k} e_i(\mathbf{XW}) e^{\varphi(i)} \right\}.$$

Warning computationally intractable (all combinations)

$\rightarrow$ Relaxed version of the problem !

Relaxed Version and Regularization

$$(\widehat{\mathbf{W}}_k, \widehat{\mathbf{Z}}_k) \in \underset{(\mathbf{W},\mathbf{Z})\in\mathcal{A}_p^m\times\mathcal{A}_m^k}{\arg\max} f(\mathbf{W},\mathbf{Z}) = \frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i\mathbf{W}\mathbf{Z}^i = Tr(\mathbf{X}\mathbf{W}\mathbf{Z})/k.$$

Relaxed Version and Regularization

$$(\widehat{\mathbf{W}}_k, \widehat{\mathbf{Z}}_k) \in \underset{(\mathbf{W},\mathbf{Z})\in\mathcal{A}_p^m\times\mathcal{A}_m^k}{\arg\max} f(\mathbf{W},\mathbf{Z}) = \frac{1}{k}\sum_{i=1}^{k}\mathbf{X}_i\mathbf{W}\mathbf{Z}^i = Tr(\mathbf{XWZ})/k.$$

• Constraint of disjoint clusters by forcing the columns of the mixture matrix $\mathbf{W}$ to be orthogonal, *i.e.*, for all $i < j, \langle W^i, W^j\rangle = 0$.

• Penalized version of the objective function with a regularization parameter $\lambda > 0$:

$$f_\lambda(\mathbf{W},\mathbf{Z}) = Tr(\mathbf{XWZ})/k - \lambda\sum_{i<j}\langle W^i, W^j\rangle$$

with partial derivatives given by

$$\begin{cases} \nabla_{\mathbf{Z}} f_\lambda(\mathbf{W},\mathbf{Z}) &= (\mathbf{XW})^T/k \\ \nabla_{\mathbf{W}} f_\lambda(\mathbf{W},\mathbf{Z}) &= (\mathbf{ZX})^T/k - \lambda\widetilde{\mathbf{W}}, \qquad \widetilde{W}^j = \sum_{i<j} W^i. \end{cases}$$

Projection onto Simplex

- Recover clusters that are not unit sets $\rightarrow$ avoid the vertices.
- Projection step $\Pi_{\mathcal{S}}(\cdot)$ of each column of $\mathbf{W}$ onto a convex set $\mathcal{S}$.
- $\bar{x} = (1/p, \ldots, 1/p)$ the barycenter of the probability simplex Δ_p.
- To escape from the curse of dimensionality, we introduce the convex set where we cut off the vertices using a threshold τ of the distance $L = \|\bar{x} - e_j\|_2 = \sqrt{(p-1)/p}$ between the barycenter and a vertex.

$$\mathcal{S}_p^\tau = \left\{ x \in \Delta_p | \max_{1 \leq j \leq p} \langle x - \bar{x}, e_j - \bar{x} \rangle \leq \tau \|e_j - \bar{x}\|_2 \right\}.$$

Define the radius $r_\infty^p(\tau) = 1 - (1-\tau)(p-1)/p$ then

$$\mathcal{S}_p^\tau = \Delta_p \cap B_{\infty,p}(\bar{x}, \tau L) = \Delta_p \cap B_{\infty,p}(0, r_\infty^p(\tau)).$$

Region of interest: the $\mathbb{M}$-set

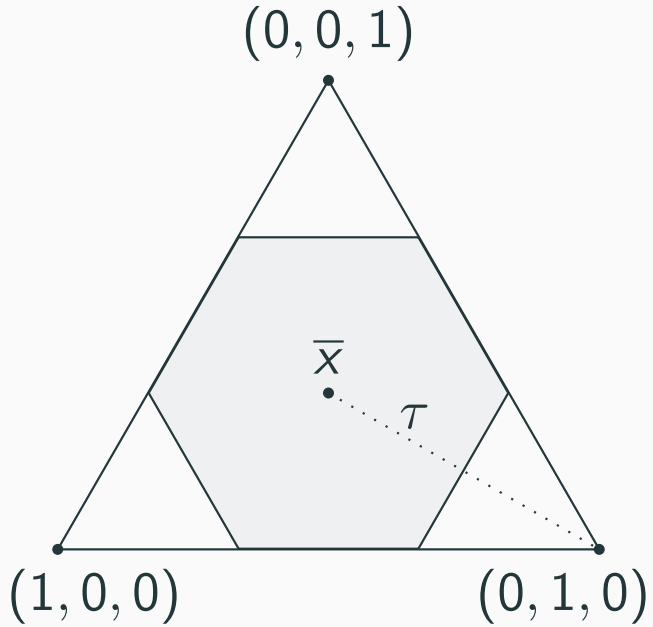


Figure 1: Simplex of $\mathbb{R}^3$ with our region of interest.

Bounding the Excess Risk

Non-asymptotic bound

Consider the risk $\mathcal{R}_{t_\gamma}, k = \lfloor n\gamma \rfloor$ and denote by $\mathbf{W}_{mex}$ the mixture matrix obtained by MEXICO. Then for $\delta \in (0,1)$, $n \geq 1$ and $\tau \leq 1$ we have with probability at least $1-\delta$,

$$\mathcal{R}_{t_\gamma}(\mathbf{W}_{mex}) - \mathcal{R}_{t_\gamma}(\mathbf{W}^{\star}_{t_\gamma}) \leq \frac{1}{\sqrt{k}} C(\gamma,\delta) \ + \frac{1}{k} C'(\gamma,\delta) + C''(\tau).$$

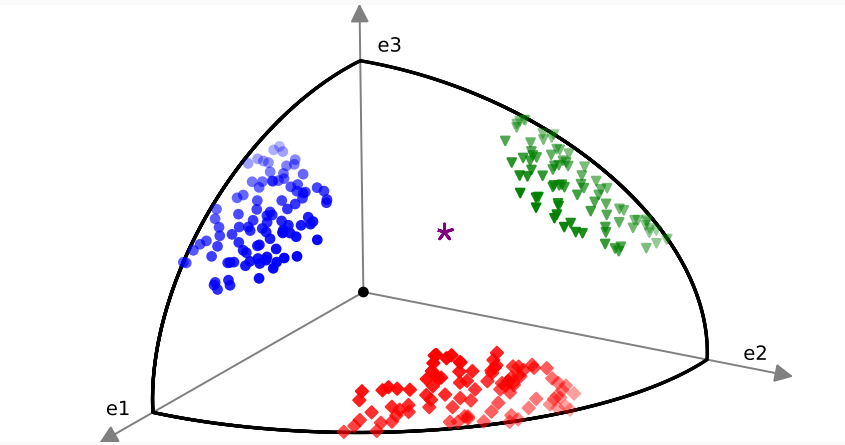
• Convergence rate of order $O_{\mathbb{P}}(1/\sqrt{k})$ where k is the actual size of the dataset required to estimate the support of extreme.

Numerical Experiments

Numerical Experiments: Anomaly Detection

Anomaly Detection

Predict if a new extreme sample $X_{\text{new}} \in \mathbb{R}_+^p$ is an anomaly, using the value of the loss function $\ell(X_{\text{new}}, \mathbf{W}_{\text{mex}})$ as an anomaly score.

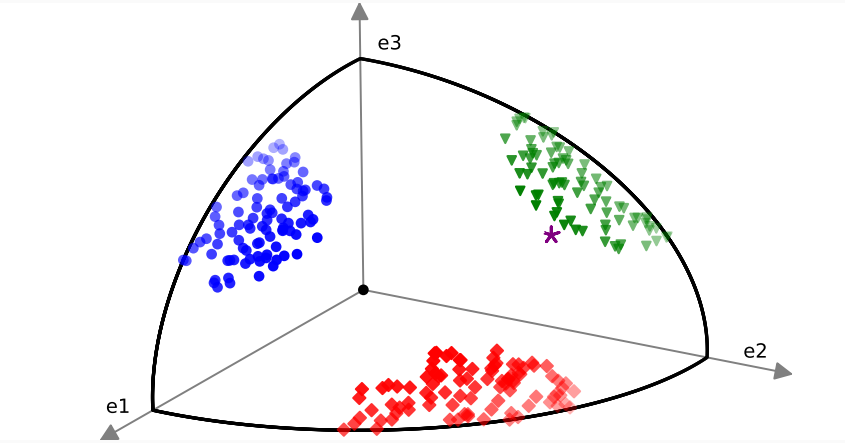


- small loss $\rightarrow$ X_{new} behavior is rather *normal*
- large loss $\rightarrow$ X_{new} more likely to be an *anomaly*.

Numerical Experiments: Feature Clustering

Feature Clustering

A new extreme sample $X_{\text{new}} \in \mathbb{R}_+^p$ is to be analyzed.



• Since X_{new} is extreme $\rightarrow$ predict the features that are large simultaneously based on the clusters given by MEXICO.

• Compute the transformed sample $\widetilde{X}_{\text{new}} = X_{\text{new}}\mathbf{W}_{mex}$ and assign the predicted cluster of features by $\text{Pred}(X_{\text{new}}) = \arg\max_{1 \le j \le m} \widetilde{X}^j_{\text{new}}$.

Numerical Experiments: Details

Feature Clustering

Since MEXICO is an inductive clustering method, compare with spectral clustering Ding et al. (2005) and spherical K-means Janßen et al. (2020).

- Simulated data from an (asymmetric) logistic distribution.
- Parameter setting: dimension $p \in \{75, 100, 150, 200\}$, number of train samples $n_{\text{train}} = 1000$ and test samples $n_{\text{test}} = 100$.

Anomaly Detection

Comparison of three algorithms for anomaly detection in extreme regions: Isolation Forest (Liu et al., 2008), DAMEX (Goix et al., 2017) and our method MEXICO.

- Five reference AD datasets are studied: shuttle, forestcover, http, SF and SA.

Conclusion

Conclusion

- Optimization framework (ERM) for clustering features in extreme regions
- Our approach does not scan all the multiple possible subsets and outperforms existing algorithms
- Future work will focus on the statistical properties of the developed algorithm by further exploring links with kernel methods

Questions and Answers

- Our paper: **https://arxiv.org/abs/2008.07365**

- Thank You!

References

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic journal of statistics 9*(1), 383–418.

Chiapino, M. and A. Sabourin (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pp. 132–147. Springer.

Chiapino, M., A. Sabourin, and J. Segers (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes 22*(2), 193–222.

Cooley, D. and E. Thibaud (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika 106*(3), 587–604.

Cutler, A. and L. Breiman (1994). Archetypal analysis. *Technometrics 36*(4), 338–347.

De Haan, L. and A. Ferreira (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.

Ding, C., X. He, and H. D. Simon (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pp. 606–610. SIAM.

Drees, H. and A. Sabourin (2019). Principal component analysis for multivariate extremes. *arXiv preprint arXiv:1906.11043*.

Engelke, S. and A. S. Hitz (2018). Graphical models for extremes. *arXiv preprint arXiv:1812.01734*.

Engelke, S. and J. Ivanovs (2020). Sparse structures for multivariate extremes. *arXiv preprint arXiv:2004.12182*.

Goix, N., A. Sabourin, and S. Clémençon (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pp. 75–83.

Goix, N., A. Sabourin, and S. Clémençon (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis 161*, 12–31.

Janßen, A., P. Wan, et al. (2020). k-means clustering of extremes. *Electronic Journal of Statistics 14*(1), 1211–1233.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE.

Meyer, N. and O. Wintenberger (2019). Sparse regular variation. *arXiv preprint arXiv:1907.00686*.

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Wold, S., K. Esbensen, and P. Geladi (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems 2*(1-3), 37–52.