Estimating Max-Stable Random Vectors with Discrete Spectral Measure using Model-Based Clustering

Alexis Boulin[1,2]

15 mai 2024

[1] Université Côte d'Azur, CNRS, LJAD, Nice, France.

[3] Inria, Lemon.

Introduction

Presentation of linear factor models

- We consider the linear factor model where $\mathbf{X}$ is an observable random vector in $\mathbb{R}^d$ which takes the following decomposition

$$\mathbf{X} = \mathrm{A}\mathbf{Z} + \mathbf{E}$$

where $A \in \mathbb{R}^{d \times K}$ is a loading matrix that parametrizes the factorization of $\mathbf{X}$ through $\mathbf{Z} \in \mathbb{R}^K$, an unobservable latent random vector, and $\mathbf{E}$ is a random vector serving as noise.

Linear factor models has a long story outside EVT

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Becoming popular as dimension reduction tools.

Linear factor models has a long story outside EVT

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Becoming popular as dimension reduction tools.
- Offering an efficient means of modeling dependencies in high dimensions, contigent a limited number of latent factors.

Linear factor models has a long story outside EVT

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Becoming popular as dimension reduction tools.
- Offering an efficient means of modeling dependencies in high dimensions, contigent a limited number of latent factors.
- Joint normality of the common factors is typically assumed and maximum likelihood estimation is employed.

Linear factor models inside EVT

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- An observable random vector $\mathbf{X} \in \mathbb{R}^d$;

Linear factor models inside EVT

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- An observable random vector $\mathbf{X} \in \mathbb{R}^d$;
- a latent random vector $\mathbf{Z} \in \mathbb{R}^K$ which is regularly varying with tail index $\alpha = 1$ and having the subsequent exponent measure

$$\Lambda_{\mathbf{Z}} = \sum_{k=1}^{K} \delta_0 \otimes \cdots \otimes \Lambda_{Z^{(k)}} \otimes \cdots \otimes \delta_0, \quad \Lambda_{Z^{(k)}}(dy) = y^{-2} dy.$$

Linear factor models inside EVT

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- An observable random vector $\mathbf{X} \in \mathbb{R}^d$;
- a latent random vector $\mathbf{Z} \in \mathbb{R}^K$ which is regularly varying with tail index $\alpha = 1$ and having the subsequent exponent measure

$$\Lambda_{\mathbf{Z}} = \sum_{k=1}^{K} \delta_0 \otimes \cdots \otimes \Lambda_{Z^{(k)}} \otimes \cdots \otimes \delta_0, \quad \Lambda_{Z^{(k)}}(dy) = y^{-2} dy.$$

- a light-tailed noise $\mathbf{E} \in \mathbb{R}^d$, independent of factors.

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- An observable random vector $\mathbf{X} \in \mathbb{R}^d$;
- a latent random vector $\mathbf{Z} \in \mathbb{R}^K$ which is regularly varying with tail index $\alpha = 1$ and having the subsequent exponent measure

$$\Lambda_{\mathbf{Z}} = \sum_{k=1}^{K} \delta_0 \otimes \cdots \otimes \Lambda_{Z^{(k)}} \otimes \cdots \otimes \delta_0, \quad \Lambda_{Z^{(k)}}(dy) = y^{-2} dy.$$

- a light-tailed noise $\mathbf{E} \in \mathbb{R}^d$, independent of factors.
- This model is also very interpretable :

$$X^{(1)} = \underbrace{0.5}_{\substack{\text{half of extremes} \\ \text{are due to } Z^{(1)}}} Z^{(1)} + \underbrace{0.5}_{\substack{\text{half of extremes} \\ \text{are due to } Z^{(2)}}} Z^{(2)}$$

$\mathbf{X}$ is also regularly varying

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Let $||\cdot||$ be a norm, $E = [0,\infty)^d \setminus \{\mathbf{0}\}$, $S_d = \{\mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}|| = 1\}$ and $\Theta = S_d \cap E$.

$\mathbf{X}$ is also regularly varying

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Let $||\cdot||$ be a norm, $E = [0, \infty)^d \setminus \{\mathbf{0}\}$, $S_d = \{\mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}|| = 1\}$ and $\Theta = S_d \cap E$.
- The following weak convergence holds true on Θ

$$\lim_{x \to \infty} \mathbb{P}\left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in \cdot \mid \|\mathbf{X}\| > x \right\} = \Phi(\cdot),$$

$\mathbf{X}$ is also regularly varying

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Let $||\cdot||$ be a norm, $E = [0, \infty)^d \setminus \{\mathbf{0}\}$, $S_d = \{\mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}|| = 1\}$ and $\Theta = S_d \cap E$.
- The following weak convergence holds true on Θ

$$\lim_{x \to \infty} \mathbb{P}\left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in \cdot \mid \|\mathbf{X}\| > x \right\} = \Phi(\cdot),$$

- Φ has the discrete representation

$$\Phi(\cdot) = w^{-1} \sum_{k=1}^{K} \|A_{\cdot k}\| \delta_{\frac{A_{\cdot k}}{\|A_{\cdot k}\|}}(\cdot), \quad w = \sum_{k=1}^{K} \|A_{\cdot k}\|,$$

X is also regularly varying

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- Let $||\cdot||$ be a norm, $E = [0,\infty)^d \setminus \{\mathbf{0}\}$, $S_d = \{\mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}|| = 1\}$ and $\Theta = S_d \cap E$.
- The following weak convergence holds true on Θ

$$\lim_{x\to\infty} \mathbb{P}\left\{ \frac{\mathbf{X}}{\|\mathbf{X}\|} \in \cdot \mid \|\mathbf{X}\| > x \right\} = \Phi(\cdot),$$

- Φ has the discrete representation

$$\Phi(\cdot) = w^{-1} \sum_{k=1}^{K} \|A_{\cdot k}\| \delta_{\frac{A_{\cdot k}}{\|A_{\cdot k}\|}}(\cdot), \quad w = \sum_{k=1}^{K} \|A_{\cdot k}\|,$$

- The linear factor model is a linear adaptation of the max-linear models, sharing the same angular measure Φ.

Max Linear Factor Model

$$\mathbf{X} = \left(\bigvee_{a=1}^{K} A_{1a} Z^{(a)}, \ldots, \bigvee_{a=1}^{K} A_{da} Z^{(a)} \right)$$

Main contributions

Outline of the litterature

- Since there is no Lebesgue density for the angular measure, estimating A in linear factor models is difficult.

- Since there is no Lebesgue density for the angular measure, estimating A in linear factor models is difficult.
- [Einmahl et al., 2012] and [Einmahl et al., 2018] opt for a least squares estimator based on the stable tail dependence function.

Outline of the litterature

- Since there is no Lebesgue density for the angular measure, estimating A in linear factor models is difficult.
- [Einmahl et al., 2012] and [Einmahl et al., 2018] opt for a least squares estimator based on the stable tail dependence function.
- [Janßen and Wan, 2020, Avella-Medina et al., 2021] propose spectral clustering designed for extremes employing its output to estimate $A_{\cdot 1}/||A_{\cdot 1}||, \dots, A_{\cdot K}/||A_{\cdot K}||$.

Outline of the litterature

- Since there is no Lebesgue density for the angular measure, estimating A in linear factor models is difficult.
- [Einmahl et al., 2012] and [Einmahl et al., 2018] opt for a least squares estimator based on the stable tail dependence function.
- [Janßen and Wan, 2020, Avella-Medina et al., 2021] propose spectral clustering designed for extremes employing its output to estimate $A_{\cdot 1}/||A_{\cdot 1}||, \dots, A_{\cdot K}/||A_{\cdot K}||$.
- [Avella-Medina et al., 2021, Avella-Medina et al., 2022], introduce a procedure coupled with screeplot to aid in the selection of K.

Outline of the litterature

- Since there is no Lebesgue density for the angular measure, estimating A in linear factor models is difficult.
- [Einmahl et al., 2012] and [Einmahl et al., 2018] opt for a least squares estimator based on the stable tail dependence function.
- [Janßen and Wan, 2020, Avella-Medina et al., 2021] propose spectral clustering designed for extremes employing its output to estimate $A_{\cdot 1}/||A_{\cdot 1}||, \dots, A_{\cdot K}/||A_{\cdot K}||$.
- [Avella-Medina et al., 2021, Avella-Medina et al., 2022], introduce a procedure coupled with screeplot to aid in the selection of K.
- Methods for estimating A in higher dimensions have emerged under the condition of a squared matrix $A \in \mathbb{R}^{d \times d}$ (see, e.g., [Klüppelberg and Krali, 2021, Kiriliouk and Zhou, 2022]).

Some (theoretical) limits

- No results have been obtained outside the independent and identically distributed (i.i.d.) assumption.

Some (theoretical) limits

- No results have been obtained outside the independent and identically distributed (i.i.d.) assumption.
- Theoretical results are derived within the framework of a fixed d and as n approaches infinity.

Our contribution

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- We propose a model-based clustering via A.

Our contribution

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- We propose a model-based clustering via A.
- We consider two components $X^{(i)}$ and $X^{(j)}$ as akin if they share a non-zero association.

Our contribution

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- We propose a model-based clustering via A.
- We consider two components $X^{(i)}$ and $X^{(j)}$ as akin if they share a non-zero association.
- Variables exhibiting this similarity are grouped together within the cluster denoted as G_a :

$$G_a = \{j \in \{1, \dots, d\} : A_{ja} \neq 0\}, \text{ for each } a \in \{1, \dots, K\}.$$

Our contribution

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- We propose a model-based clustering via A.
- We consider two components $X^{(i)}$ and $X^{(j)}$ as akin if they share a non-zero association.
- Variables exhibiting this similarity are grouped together within the cluster denoted as G_a :

$$G_a = \{j \in \{1, \ldots, d\} : A_{ja} \neq 0\}, \text{ for each } a \in \{1, \ldots, K\}.$$

- The definition of A within linear factor models, lacks uniqueness without imposing additional constraints :

Our contribution

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- We propose a model-based clustering via A.
- We consider two components $X^{(i)}$ and $X^{(j)}$ as akin if they share a non-zero association.
- Variables exhibiting this similarity are grouped together within the cluster denoted as G_a :

$$G_a = \{j \in \{1, \ldots, d\} : A_{ja} \neq 0\}, \text{ for each } a \in \{1, \ldots, K\}.$$

- The definition of A within linear factor models, lacks uniqueness without imposing additional constraints :
  - **Condition (i)** $\sum_{a=1}^{K} A_{ja} = 1$;

Our contribution

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- We propose a model-based clustering via A.
- We consider two components $X^{(i)}$ and $X^{(j)}$ as akin if they share a non-zero association.
- Variables exhibiting this similarity are grouped together within the cluster denoted as G_a :

$$G_a = \{j \in \{1, \ldots, d\} : A_{ja} \neq 0\}, \text{ for each } a \in \{1, \ldots, K\}.$$

- The definition of A within linear factor models, lacks uniqueness without imposing additional constraints :
  - **Condition (i)** $\sum_{a=1}^{K} A_{ja} = 1$;
  - **Condition (ii)** For every $a \in \{1, \ldots, K\}$, there exist at least one indice $j \in \{1, \ldots, d\}$ such that $A_{ja} = 1$ and $A_{jb} = 0$, $\forall b \neq a$.

Identifiability theorem

Factor with second moments

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- In this frame we suppose that $Z \sim \mathcal{N}_K(0, I_K)$.

Factor with second moments

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- In this frame we suppose that $Z \sim \mathcal{N}_K(0, I_K)$.
- Using independence between factors, we have $Cov(\mathbf{Z}) = I_K$, the identity matrix.

Factor with second moments

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- In this frame we suppose that $Z \sim \mathcal{N}_K(0, I_K)$.
- Using independence between factors, we have $Cov(\mathbf{Z}) = I_K$, the identity matrix.
- It is possible to show that under **Condition (i)** and **Condition (ii)** that the matrix A can be recovered solely using $Cov(\mathbf{X}) = AA^\top$.

Factor with second moments

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- In this frame we suppose that $Z \sim \mathcal{N}_K(0, I_K)$.
- Using independence between factors, we have $Cov(\mathbf{Z}) = I_K$, the identity matrix.
- It is possible to show that under **Condition (i)** and **Condition (ii)** that the matrix A can be recovered solely using $Cov(\mathbf{X}) = AA^\top$.
- In our framework, the covariance matrix of $\mathbf{Z}$ does not exists.

Factor with second moments

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

- In this frame we suppose that $Z \sim \mathcal{N}_K(0, I_K)$.
- Using independence between factors, we have $Cov(\mathbf{Z}) = I_K$, the identity matrix.
- It is possible to show that under **Condition (i)** and **Condition (ii)** that the matrix A can be recovered solely using $Cov(\mathbf{X}) = AA^\top$.
- In our framework, the covariance matrix of $\mathbf{Z}$ does not exists.
- Can we find a similar, but different bivariate measures having desirable properties ?

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

Theorem

Let $\mathbf{X}$ be a LFM and A satisfies Condition (i). Then $\mathbf{X}$ is regularly varying and its extremal correlation matrix $\mathcal{X}$ can be written as

$$\mathcal{X} = A \odot A^{\top},$$

with

$$\chi(i,j) = \sum_{k=1}^{K} A_{ik} \wedge A_{jk}.$$

Some properties about pure variables

- For any given matrix A, the pure variable set is outlined as follows

$$I = \cup_{a=1}^{K} I_a, \quad I_a := \{i \in [d] : A_{ia} = 1, A_{ib} = 0, \forall b \neq a\}.$$

- By Condition (ii), $\forall a \in [K]$, $\exists i_a \in \{1, \ldots, d\}$ such that $X^{(i_a)} = Z^{(a)}$.
- Per construction, the vector $(X^{(i_1)}, \ldots, X^{(i_K)})$ is the largest asymptotically independent vector.
- If $i, j \in I_a$, then $\chi(i, j) = 1$.

Pure variables are identifiable

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

Theorem

Let $\mathbf{X}$ be a LFM and Conditions **(i)**-**(ii)** hold. Then :

1. The set $[K]$ is a maximal clique of the undirected graph $G = (V, E)$ where $V = [d]$ and $(i, j) \in E$ if $\chi(i, j) = 0$.
2. The pure variable set I can be determined uniquely from $\mathcal{X}$. Moreover its partition $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$ is unique and can be determined from $\mathcal{X}$ up to label permutations.

Non-pure coefficients are identifiable

By designing $J := [d] \setminus I$, the set of impure variables, we show that A_J is identifiable.

- For each $i \in I_k$ for some $k \in [K]$ and any $j \in J$, the model imposes :

$$\chi(i,j) = \sum_{a=1}^{K} A_{ia} \wedge A_{ja} = A_{jk}$$

- After averaging over all $i \in I_k$,

$$A_{jk} = \frac{1}{|I_k|} \sum_{i \in I_k} \chi(i,j).$$

- Repeating this for every $k \in [K]$, we obtain the formula

$$A_{j\cdot} = \left(\frac{1}{|I_1|} \sum_{i \in I_1} \chi(i,j), \dots, \frac{1}{|I_K|} \sum_{i \in I_K} \chi(i,j) \right).$$

The matrix is identifiable

Linear Factor Model

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}$$

Theorem

Assume that $\mathbf{X}$ is a LFM and Conditions **(i)**-**(ii)** hold. Then, there exist a unique matrix A, up to a permutation, such that $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$. This implies that the associated soft clusters G_a, for $1 \leq a \leq K$, are identifiable, up to label switching.

Estimation

Data generative process

Let $(\mathbf{X}_t, t \in \mathbb{Z})$ be a multivariate strictly stationary random process and $(\mathbf{X}_t, t = 1, \dots, n)$ an excerpt. Consider $m \in \{1, \dots, n\}$ and C_m be the copula of the m-componentwise maxima of $(\mathbf{X}_t, t \in \mathbb{Z})$. We suppose that there exist a copula C_∞ such that

$$\lim_{m \to \infty} C_m(\mathbf{u}) = C_\infty(\mathbf{u}), \mathbf{u} \in [0,1]^d,$$

where

$$C_\infty(\mathbf{u}) = \exp\left\{-L\left(-\ln(u^{(1)}), \dots, -\ln(u^{(d)})\right)\right\},$$

and the stable tail dependence function $L : [0,\infty)^d \to [0,\infty)$ is described by

$$L(z^{(1)}, \dots, z^{(d)}) = \sum_{a=1}^{K} \vee_{j=1}^{d} A_{ja} z^{(j)}.$$

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
    - Construct the graph $G = (V, E)$ where $V = [d]$ and $(i, j) \in E$ if $\hat{\chi}_{n,m}(i, j) \leq \delta$.
    - Find a maximal clique, $\bar{\mathcal{G}}$, of G.
    - $\hat{I}^{(i)} = \{j \in [d] : 1 - \hat{\chi}_{n,m}(i, j) \leq \delta\}$, $\hat{I}^{(i)} = \hat{I}^{(i)} \cup \{i\}$, $i \in \bar{\mathcal{G}}$.

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
2. Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
2. Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;
   - $\hat{A}_{ka} = \hat{A}_{la} = 1, \quad \text{for } k, l \in \hat{I}_a, \quad a \in [\hat{K}].$

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
2. Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;
3. Estimate A_J, the submatrix of A with rows $A_{j\cdot}$ that correspond to $j \in J$;

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
2. Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;
3. Estimate A_J, the submatrix of A with rows $A_{j\cdot}$ that correspond to $j \in J$;
   - $\bar{\chi}^{(j)} = \left(\frac{1}{|\hat{I}_1|} \sum_{i \in \hat{I}_1} \hat{\chi}_{n,m}(i,j), \ldots, \frac{1}{|\hat{I}_{\hat{K}}|} \sum_{i \in \hat{I}_K} \hat{\chi}_{n,m}(i,j) \right).$
   - $\bar{\beta}_a^{(j)} = \bar{\chi}_a^{(j)} \mathbb{1}_{\{\bar{\chi}_a^{(j)} > \delta\}}$, $a \in [\hat{K}]$
   - By denoting $\hat{\mathcal{S}} = \operatorname{supp}(\bar{\beta}^{(j)})$, we obtain $\hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}} = \mathcal{P}_{\Delta_{\hat{K}-1}} \left(\bar{\beta}^{(j)}\big|_{\widehat{\mathcal{S}}} \right), \quad \hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}^c} = 0.$

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
2. Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;
3. Estimate A_J, the submatrix of A with rows $A_{j\cdot}$ that correspond to $j \in J$;
4. Estimate the overlapping clusters $\hat{\mathcal{G}} = \{\hat{G}_1, \ldots, \hat{G}_{\hat{K}}\}$.

The estimation procedure

1. Estimate the number of clusters K, the pure variable set I and its partition $\mathcal{I}$;
2. Estimate A_I, the submatrix of A with rows $A_{i\cdot}$ that correspond to $i \in I$;
3. Estimate A_J, the submatrix of A with rows $A_{j\cdot}$ that correspond to $j \in J$;
4. Estimate the overlapping clusters $\hat{\mathcal{G}} = \{\hat{G}_1, \ldots, \hat{G}_{\hat{K}}\}$.
    - $\hat{\mathcal{G}} = \{\hat{G}_1, \ldots, \hat{G}_{\hat{K}}\}$, $\hat{G}_a = \{j \in [d] : \hat{A}_{ja} \neq 0\}$, for each $a \in [\hat{K}]$.

Step 1

Find a maximal clique

Step 2

Estimate A_I

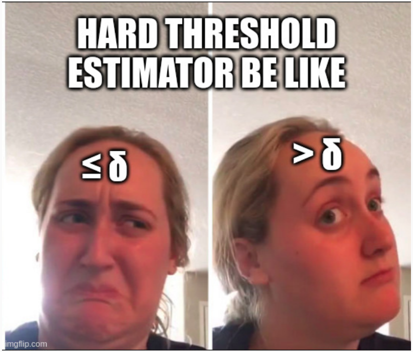
Step 3

Estimate A_J :

$$\bar{\chi}^{(j)} = \left(\frac{1}{|\hat{I}_a|} \sum_{i \in \hat{I}_a} \hat{\chi}_{n,m}(i,j) \right)_{a \in [\hat{K}]}$$

Step 3

Estimate the support : $\bar{\beta}_a^{(j)} = \bar{\chi}_a^{(j)} \mathbb{1}_{\{\bar{\chi}_a^{(j)} > \delta\}}$

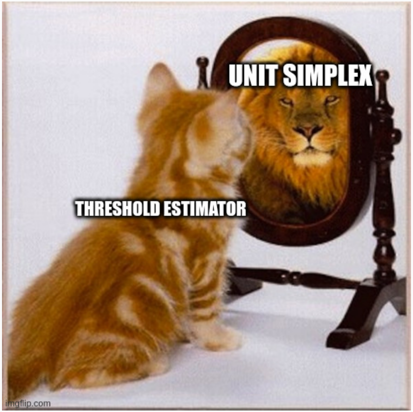


The procedure in memes

Step 3

Projection into the sparse simplex :

$$\hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}} = \mathcal{P}_{\Delta_{\hat{K}-1}}\left(\bar{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}}\right), \quad \hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}^c} = 0$$



The procedure in memes

Step 4

Estimate overlapping clusters :

$\hat{G}_a = \{j \in [d] \, : \, \hat{A}_{ja} \neq 0\}$, $a \in [\hat{K}]$

Statistical guarantees

A concentration inequality

1. Let $\hat{\chi}_{n,m}(i,j)$ the madogram-based estimator of the extremal correlation, n is the sample size and m the block's length, $i, j \in [d]$.

A concentration inequality

1. Let $\hat{\chi}_{n,m}(i,j)$ the madogram-based estimator of the extremal correlation, n is the sample size and m the block's length, $i,j \in [d]$.
2. Let
$$d_m = \sup_{1 \leq i < j \leq d} |\chi_m(i,j) - \chi(i,j)|,$$
where $\chi_m(i,j)$ is the *pre-asymptotic* extremal correlation.

A concentration inequality

1. Let $\hat{\chi}_{n,m}(i,j)$ the madogram-based estimator of the extremal correlation, n is the sample size and m the block's length, $i, j \in [d]$.
2. Let
$$d_m = \sup_{1 \leq i < j \leq d} |\chi_m(i,j) - \chi(i,j)|,$$
where $\chi_m(i,j)$ is the *pre-asymptotic* extremal correlation.
3. Define
$$\mathcal{E} = \mathcal{E}(\delta) := \left\{ \sup_{1 \leq i < j \leq d} |\hat{\chi}_{n,m}(i,j) - \chi(i,j)| \leq \delta \right\}.$$

A concentration inequality

1. Let $\hat{\chi}_{n,m}(i,j)$ the madogram-based estimator of the extremal correlation, n is the sample size and m the block's length, $i, j \in [d]$.
2. Let
$$d_m = \sup_{1 \leq i < j \leq d} |\chi_m(i,j) - \chi(i,j)|,$$
where $\chi_m(i,j)$ is the *pre-asymptotic* extremal correlation.
3. Define
$$\mathcal{E} = \mathcal{E}(\delta) := \left\{ \sup_{1 \leq i < j \leq d} |\hat{\chi}_{n,m}(i,j) - \chi(i,j)| \leq \delta \right\}.$$
4. If $(\mathbf{X}_t, t \in \mathbb{Z})$ has exponential decaying strong mixing coefficients, then there exists $c_0 > 0, c_1 > 0$ such that
$$\mathbb{P}(\mathcal{E}) \geq 1 - d^{-c_0},$$
where
$$\delta = d_m + c_1 \left(\sqrt{\frac{\ln(kd)}{k}} + \frac{\ln(k) \ln\ln(k) \ln(kd)}{k} \right),$$
and $k = \lfloor n/m \rfloor \geq 4$, the number of blocks.

The procedure is consistent with high probability

Theorem

Set $s = \max_{j \in [d]} ||A_{j\cdot}||_0$. Let $(\mathbf{X}_t, t \in \mathbb{Z})$ verifies the data generative process and some strong signal conditions. Then for the estimator $\hat{A}$ the following holds.

1. Recovery of latent factors :

$$\hat{K} = K,$$

with probability larger than $1 - d^{-c_0}$ for a positive constant c_0.

The procedure is consistent with high probability

Theorem

Set $s = \max_{j \in [d]} ||A_{j\cdot}||_0$. Let $(\mathbf{X}_t, t \in \mathbb{Z})$ verifies the data generative process and some strong signal conditions. Then for the estimator $\hat{A}$ the following holds.

2. An upper bound :

$$L_2(\hat{A}, A) \leq 4\sqrt{s}\delta,$$

where $L_2(A, A') := \min_{P \in S_K} ||AP - A'||_{\infty,2}$, and
$||A||_{\infty,2} := \max_{1 \leq j \leq d} ||A_{j\cdot}||_2$,

with probability larger than $1 - d^{-c_0}$ for a positive constant c_0.

The procedure is consistent with high probability

Theorem

Set $s = \max_{j \in [d]} ||A_{j\cdot}||_0$. Let $(\mathbf{X}_t, t \in \mathbb{Z})$ verifies the data generative process and some strong signal conditions. Then for the estimator $\hat{A}$ the following holds.

3. A guarantee for support recovery :

$$supp(A_{J_1}) \subseteq supp(\hat{A}) \subseteq supp(A),$$

where $J_1 = \{j \in J : \text{ for any } a \in [K] \text{ with } A_{ja} \neq 0, A_{ja} > 2\delta\}$,

with probability larger than $1 - d^{-c_0}$ for a positive constant c_0.

The procedure is consistent with high probability

Theorem

Set $s = \max_{j \in [d]} ||A_{j\cdot}||_0$. Let $(\mathbf{X}_t, t \in \mathbb{Z})$ verifies the data generative process and some strong signal conditions. Then for the estimator $\hat{A}$ the following holds.

4. Cluster recovery :

$$TFPP(\widehat{\mathcal{G}}) = \frac{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}=0, \hat{A}_{ja}>0\}}}{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}=0\}}} = 0,$$

$$TFNP(\widehat{\mathcal{G}}) = \frac{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0, \hat{A}_{ja}=0\}}}{\sum_{j \in [d], a \in [K]} \mathbb{1}_{\{A_{ja}>0\}}} \leq \frac{\sum_{j \in J \setminus J_1} t(j)}{|I| + \sum_{j \in J} s(j)},$$

where $s(j) = \sum_{a=1}^{K} \mathbb{1}_{\{A_{ja}>0\}}$ and $t(j) = \sum_{a=1}^{K} \mathbb{1}_{\{A_{ja} \leq 2\delta\}}$,

with probability larger than $1 - d^{-c_0}$ for a positive constant c_0.

Application(s)

Extreme precipitations in France

- We focus on weekly maxima of hourly precipitation recorded at 92 weather stations in France during the fall season (September-November, 1993-2011).

Extreme precipitations in France

- We focus on weekly maxima of hourly precipitation recorded at 92 weather stations in France during the fall season (September-November, 1993-2011).
- We thus have 228 block maxima.

Extreme precipitations in France

- We focus on weekly maxima of hourly precipitation recorded at 92 weather stations in France during the fall season (September-November, 1993-2011).
- We thus have 228 block maxima.
- This dataset was provided by Météo-France and has been previously used in [Bernard et al., 2013].

Extreme precipitations in France

- We focus on weekly maxima of hourly precipitation recorded at 92 weather stations in France during the fall season (September-November, 1993-2011).
- We thus have 228 block maxima.
- This dataset was provided by Météo-France and has been previously used in [Bernard et al., 2013].
- Using a data-driven selection method to choose δ, we unveil four latent factors.

Spatial representation

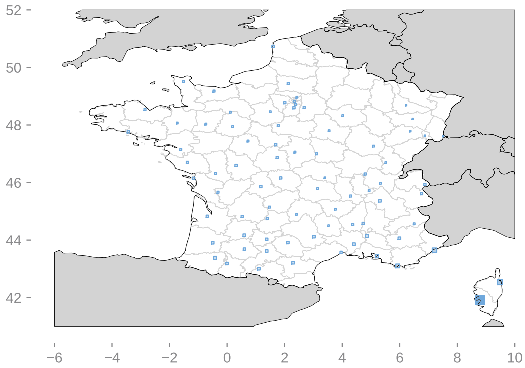


Figure 1 – Each location's strength of association with the respective latent variable is conveyed through the size and color intensity of the square.

Spatial representation

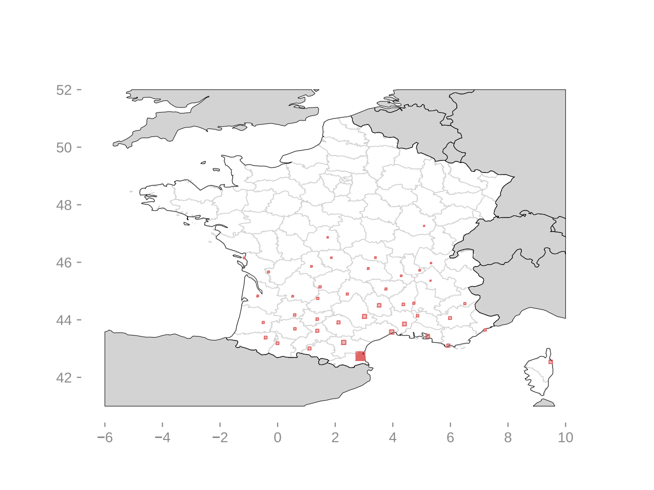


Figure 1 – Each location's strength of association with the respective latent variable is conveyed through the size and color intensity of the square.

Spatial representation

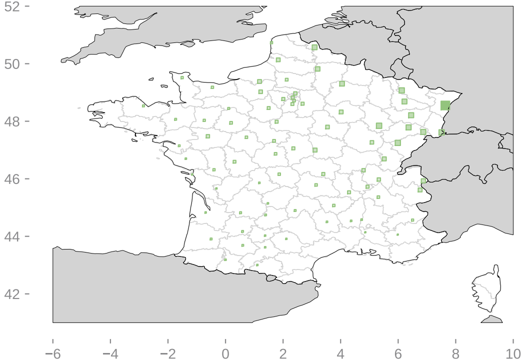


Figure 1 – Each location's strength of association with the respective latent variable is conveyed through the size and color intensity of the square.

Spatial representation

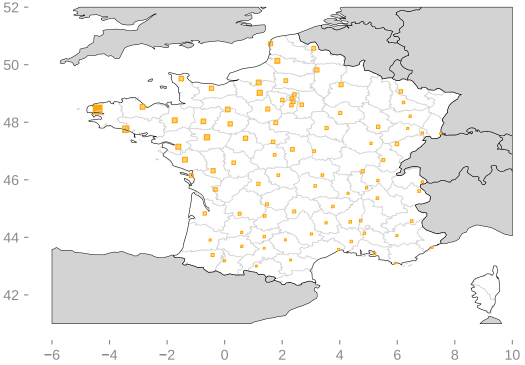


Figure 1 – Each location's strength of association with the respective latent variable is conveyed through the size and color intensity of the square.

Wildfires in French Mediterranean

- Our case study focuses on the southeastern part of France, covering an area of 80500 km2.

- Our case study focuses on the southeastern part of France, covering an area of 80500 km2.
- Gridded weather reanalysis data from the SAFRAN model of Météo-France, with an 8km resolution, is utilized for analysis.

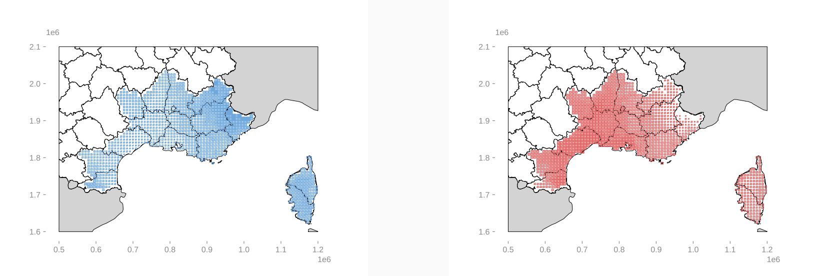
Wildfires in French Mediterranean

- Our case study focuses on the southeastern part of France, covering an area of 80500 km2.
- Gridded weather reanalysis data from the SAFRAN model of Météo-France, with an 8km resolution, is utilized for analysis.
- Various meteorological indices on fire activity patterns have been developed, including the widely used unitless Fire Weather Index (FWI).

- Our case study focuses on the southeastern part of France, covering an area of 80500 km2.
- Gridded weather reanalysis data from the SAFRAN model of Météo-France, with an 8km resolution, is utilized for analysis.
- Various meteorological indices on fire activity patterns have been developed, including the widely used unitless Fire Weather Index (FWI).
- In our methodology, we extract monthly maxima over the 1143 pixels, resulting in 100 observations.

- Our case study focuses on the southeastern part of France, covering an area of 80500 km2.
- Gridded weather reanalysis data from the SAFRAN model of Météo-France, with an 8km resolution, is utilized for analysis.
- Various meteorological indices on fire activity patterns have been developed, including the widely used unitless Fire Weather Index (FWI).
- In our methodology, we extract monthly maxima over the 1143 pixels, resulting in 100 observations.
- Using a data-driven approach to select the threshold δ, we obtain $\hat{K} = 2$ and $\hat{A}$.

Spatial variability of FWI



(a) (b)

Figure 2 – In panel a, we depict the spatial representation of cluster associated to the first latent variable. Panel b exhibits spatial association with the second latent variable. Each location's strength of association with the respective latent variable is conveyed through the proportionate size and color intensity of the square.

Conclusions

Extensions

- Minimax risk ? Very recently, [Zhang et al., 2023] were able to obtain a minimax risk for LFM with $K \geq d$.
- Despite making significant progress in understanding potential proofs by studying their methodologies, I am still encountering challenges in deriving the desired result.

Avella-Medina, M., Davis, R. A., and Samorodnitsky, G. (2021).
Spectral learning of multivariate extremes.
arXiv preprint arXiv :2111.07799.

Avella-Medina, M., Davis, R. A., and Samorodnitsky, G. (2022).
Kernel pca for multivariate extremes.
arXiv preprint arXiv :2211.13172.

Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013).
Clustering of maxima : Spatial dependencies among heavy rainfall in france.
Journal of climate, 26(20) :7929–7937.

Einmahl, J. H., Kiriliouk, A., and Segers, J. (2018).
A continuous updating weighted least squares estimator of tail dependence in high dimensions.
Extremes, 21 :205–233.

Einmahl, J. H. J., Krajina, A., and Segers, J. (2012).
An M-estimator for tail dependence in arbitrary dimensions.
The Annals of Statistics, 40(3) :1764 – 1793.

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

Kiriliouk, A. and Zhou, C. (2022).
Estimating probabilities of multivariate failure sets based on pairwise tail dependence coefficients.
arXiv preprint arXiv :2210.12618.

Klüppelberg, C. and Krali, M. (2021).
Estimating an extreme bayesian network via scalings.
Journal of Multivariate Analysis, 181 :104672.

Zhang, X., Blanchet, J., Marzouk, Y., Nguyen, V. A., and Wang, S. (2023).
Wasserstein-based minimax estimation of dependence in multivariate regularly varying extremes.
arXiv preprint arXiv :2312.09862.