Estimation de vecteurs aléatoires à variation régulière avec mesure spectrale discrète via le partitionnement de variables.

Alexis Boulin [1,2]

[1] *Université Côte d'Azur, CNRS, LJAD, France, aboulin@unice.fr*
[2] *Inria, Lemon*

Résumé. Cette étude présente une nouvelle méthode d'estimation pour les entrées et la structure d'une matrice A dans le modèle à facteurs linéaires $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$. Cela est appliqué à un vecteur observable $\mathbf{X} \in \mathbb{R}^d$ avec $\mathbf{Z} \in \mathbb{R}^K$, un vecteur composé de variables aléatoires à variation régulière indépendantes, et un bruit indépendant de $\mathbf{Z}$ à queue légère $\mathbf{E} \in \mathbb{R}^d$. $\mathbf{X}$ est à variation régulière et sa mesure spectrale est alors discrète et entièrement caractérisée par la matrice A. Nous supposerons que chaque ligne de la matrice A somme à 1 et est parcimonieuse. De plus, la valeur de K n'est pas connue a priori. Le problème d'identification de la matrice A à partir de sa matrice de corrélation extrémale est abordé. En présence de variables pures, qui sont des éléments de $\mathbf{X}$ liés, via A, à un unique facteur latent, la matrice A peut être reconstruite à partir de la matrice de corrélation extrémale. Nos preuves d'identifiabilité sont constructives et ouvrent la voie à notre estimation innovante pour déterminer le nombre de facteurs K et la matrice A à partir de n observations faiblement dépendantes sur $\mathbf{X}$.

Mots-clés. Théorie des valeurs extrêmes, Grande dimension, Modèles à facteurs latents, Partitionnement de variables souples.

Abstract. This study introduces a novel estimation method for the entries and structure of a matrix A in the linear factor model $\mathbf{X} = A\mathbf{Z} + \mathbf{E}$. This is applied to an observable vector $\mathbf{X} \in \mathbb{R}^d$ with $\mathbf{Z} \in \mathbb{R}^K$, a vector composed of independently regularly varying random variables, and light-tailed independent noise $\mathbf{E} \in \mathbb{R}^d$. $\mathbf{X}$ is hence regularly varying and its spectral measure is subsequently discrete and completely characterized by the matrix A. Each row of the matrix A is both scaled and sparse. Additionally, the value of K is not known a priori. The problem of identifying the matrix A from its matrix of pairwise extremal correlation is addressed. In the presence of pure variables, which are elements of $\mathbf{X}$ linked, through A, to a single latent factor, the matrix A can be reconstructed from the extremal correlation matrix. Our proofs of identifiability are constructive and pave the way for our innovative estimation for determining the number of factors K and the matrix A from n weakly dependent observations on $\mathbf{X}$.

Keywords. Extremes, High dimensional estimation, Latent model, Soft clustering, Variable clustering.

1 Introduction

Dans cette étude, nous souhaitons estimer la matrice d'association $d \times K$, A, qui peut présenter de la parcimonie et sert de paramètre pour la décomposition d'un vecteur aléatoire observable $\mathbf{X}$. Cela peut être exprimé comme

$$\mathbf{X} = A\mathbf{Z} + \mathbf{E}. \tag{1}$$

Dans cette équation, $\mathbf{Z}$ représente un vecteur aléatoire non observable de dimension K, servant de facteur latent sous-jacent. $E \in \mathbb{R}^d$ est un bruit aléatoire non observable, avec des entrées indépendantes. Le nombre précis de facteurs, K, reste non divulgué et à la fois d et K sont autorisés à augmenter et à être plus grands que n, le nombre d'observations. Pour établir les fondements de notre cadre dans la théorie des valeurs extrêmes, nous supposons que $\mathbf{Z}$ est composé de variables aléatoires asymptotiquement indépendantes (voir [4, page 192] caractérisées par un paramètre de forme α que nous fixerons à $\alpha = 1$. Selon cette construction, le vecteur $\mathbf{Z}$ est à variation régulière dont la mesure exponentielle (voir [3, Definition 2.1.2]) est

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

Le vecteur de pertubation $E \in \mathbb{R}^d$ est postulé posséder une distribution avec une queue plus légère que celle des facteurs associés. De plus, il présente une indépendance complète par rapport à ces facteurs. Par conséquent, $\mathbf{X}$ est également à variation régulière, ce qui peut être décrit de manière équivalente (voir, par exemple, [3, Section 2.2]) par l'existence d'une mesure angulaire $S_{\mathbf{X}}$ où la convergence faible suivante est vraie sur la sphère unité Δ_{d-1} de $\mathbb{R}^d_+$

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

où $S_{\mathbf{X}}$ a la représentation discrète

$$S_{\mathbf{X}}(\cdot) = w^{-1} \sum_{k=1}^{K} \|A_{\cdot k}\| \delta_{\frac{A_{\cdot k}}{\|A_{\cdot k}\|}}(\cdot), \quad w = \sum_{k=1}^{K} \|A_{\cdot k}\|, \tag{2}$$

avec $\delta_x(\cdot)$ étant la mesure de Dirac qui place une masse unitaire sur x et $A_{\cdot k}$ est la k-ème colonne de la matrice A.

Dans ce travail, nous proposons un partitionnement de variables basé sur ce modèle via A. Dans le cadre du modèle (1), nous considérons deux composantes, à savoir $X^{(i)}$ et $X^{(j)}$, appartenant au vecteur $\mathbf{X}$, comme similaires si elles partagent une association non nulle. Cette association est établie à travers la matrice A, les reliant à un facteur latent commun $Z^{(a)}$. Les variables présentant cette similitude sont regroupées dans le cluster désigné par G_a:

$$G_a = \{j \in \{1, \ldots, d\}, : A_{ja} \neq 0\}, \quad \text{pour chaque } a \in \{1, \ldots, K\}. \tag{3}$$

Étant donné que $X^{(j)}$ peut potentiellement être lié à plusieurs facteurs latents, un cluster peut déborder sur un autre.

Il convient de noter, cependant, que la définition de A dans le modèle (1) n'est pas systématiquement identifiable sans imposer de contraintes supplémentaires. Pour remédier à cela, nous envisageons une variante du modèle (1) où chaque ligne de A est mise à l'échelle. Pour être précis, nous posons l'hypothèse suivante:

Condition (i) $\sum_{a=1}^{K} A_{ja} = 1$.

Les poids $A_{i1}, \ldots, A_{iK}$ indiquent le degré auquel les composantes s'alignent avec chaque cluster. Cette condition divise notre modèle en partitionnements durs et souples. Néanmoins, s'appuyer uniquement sur la Condition (i) est insuffisant pour garantir l'unicité de A dans le modèle (1). Nous introduisons une hypothèse additionelle qui suppose la présence d'au moins une variable pure $X^{(j)}$, parmi les composantes de $\mathbf{X}$. Ces variables pures sont associées de manière unique à un seul facteur latent et à aucun autre.

Condition (ii) *Pour chaque* $a \in \{1, \ldots, K\}$*, il existe au moins un indice* $j \in \{1, \ldots, d\}$ *tel que* $A_{ja} = 1$ *et* $A_{jb} = 0$, $\forall b \neq a$.

2 Identifiabilité

Dans cette section, nous présentons une démonstration que la matrice d'association A, telle que définie par le modèle (1) et soumise aux conditions (i)-(ii), est identifiable, à l'exception d'une multiplication par une matrice de permutation.

Selon la construction, le vecteur $\mathbf{Z}$ est à variation régulière, il possède une matrice de corrélation extrémale représentée par I_K. Par conséquent, nous déduisons que le vecteur $\mathbf{X}$ est à variation régulière, conduisant à la présence d'une matrice de corrélation extrémale notée $\mathcal{X} = [\chi(i,j)]_{i=1,\ldots,d;j=1,\ldots,d}$, où

$$\chi(i,j) = \lim_{x\to\infty} \frac{\mathbb{P}\{X^{(i)} > x, X^{(j)} > x\}}{\mathbb{P}\{X^{(i)} > x\}}.$$

Le théorème suivant est destiné à démontrer que la matrice de corrélation extrémale peut être élégamment formulée en utilisant exclusivement la matrice d'association A. Cependant, avant d'approfondir, nous introduisons une nouvelle opération de matrice définie sur les matrices $A \in \mathcal{M}_{p,K}(\mathbb{R})$ et $B \in \mathcal{M}_{K,q}(\mathbb{R})$.

Définition 1 *Nous appelons* $\odot$ *l'application:*

$$\begin{aligned} \odot\colon \mathcal{M}_{p,K}(\mathbb{R}) \times \mathcal{M}_{K,q}(\mathbb{R}) &\longrightarrow \mathcal{M}_{p,q}(\mathbb{R}) \\ (a_{ik}, b_{mj}) &\mapsto c_{ij}, \end{aligned}$$

où, en notant par $a \wedge b := \min\{a, b\}$,

$$c_{ij} = a_{i1} \wedge b_{1j} + \cdots + c_{iK} \wedge b_{Kj}.$$

Avec tous les outils à notre disposition, nous sommes prêts à présenter le théorème fondamental suivant.

Théorème 1 *Soit* $\boldsymbol{X}$ *défini dans* (1) *et* A *satisfaisant la Condition (i). Alors* $\boldsymbol{X}$ *est à variation régulière et sa matrice de corrélation extrémale* $\mathcal{X}$ *peut être écrite comme*

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

avec

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

Pour toute matrice d'association A qui adhère au modèle (1), nous pouvons subdiviser l'ensemble $[d] = \{1, \ldots, d\}$ en deux ensembles distincts : I et son complémentaire, désigné par J. Nous noterons A_I (resp. A_J), la matrice $|I| \times K$ (resp. $|J| \times K$) extraite de A formant des lignes dans l'ensemble d'indice I (resp. J). Dans chaque ligne $A_{i\cdot}$ de A_I, il existe précisément au moins une valeur $a \in [K]$ pour laquelle $A_{ia} = 1$. Nous attribuons le terme "ensemble de variables pures" à I, tandis que J correspond à l' "ensemble de variables impures". Pour être plus précis, pour toute matrice donnée A, l'ensemble de variables pures est défini comme suit

$$I = \cup_{a=1}^{K} I_a, \quad I_a := \{i \in [d] : A_{ia} = 1, A_{ib} = 0 \; \forall b \neq a\}. \tag{4}$$

Il convient de mentionner que les ensembles $\{I_a\}_{1 \leq a \leq K}$ constituent une partition de l'ensemble de variables pures I.

Pour établir l'identifiabilité de la matrice A, notre tâche est simplifiée en se concentrant sur l'identifiabilité distincte de A_I et A_J. En ce qui concerne la définition de A_I, son identifiabilité est assurée tant que la partition de l'ensemble de variables pures I reste identifiable. Le cœur du défi réside dans l'identifiabilité de l'ensemble I et le problème inhérent de distinguer entre I et J, basé uniquement sur la matrice de corrélation extrémale du vecteur $\mathbf{X}$. Cela constitue l'obstacle central du problème. Le théorème 2 est le pinacle de notre méthodologie. Dans la première partie **(a)**, il offre à la fois une condition nécessaire et suffisante pour identifier $[K]$ en examinant la matrice de corrélation extrémale $\mathcal{X}$. Dans la deuxième partie **(b)**, il fournit une caractérisation nécessaire et suffisante pour identifier l'ensemble I lorsque la cardinalité de I_a est supérieure à un. Enfin, dans la troisième partie **(c)**, il illustre que I et sa partition en sous-ensembles $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$ peuvent être identifiés avec succès. Soit

$$M_i = \max_{j \in [d] \setminus \{i\}} \chi(i,j) \tag{5}$$

désignant la plus grande valeur parmi les entrées de la ligne i de la matrice $\mathcal{X}$ à l'exclusion de $\chi(i,i) = 1$. De plus, soit S_i l'ensemble d'indices pour lesquels M_i atteint son maximum

$$S_i = \{j \in [d] \setminus \{i\}, \; \chi(i,j) = M_i\}. \tag{6}$$

Théorème 2 *Supposons que le modèle* (1) *et les conditions (i)-(ii) sont satisfaites. Alors :*

(a) *L'ensemble* $[K]$ *est une clique maximale du graphe non orienté* $G = (V, E)$ *où* $V = [d]$ *et* $(i, j) \in E$ *si* $\chi(i, j) = 0$.

(b) *Soit* $i \in I_a$, $a \in [K]$ *et* $|I_a| \geq 2$, *alors*

$$j \in I \iff \chi(i, j) = 1 \text{ pour tout } j \in S_i.$$

(c) *L'ensemble de variables pures* I *peut être déterminé de manière unique à partir de* $\mathcal{X}$. *De plus, sa partition* $\mathcal{I} = \{I_a\}_{1 \leq a \leq K}$ *est unique et peut être déterminée à partir de* $\mathcal{X}$ *jusqu'à des permutations d'étiquettes.*

Théorème 3 *Supposons que le modèle* (1) *et les conditions (i)-(ii) sont satisfaites. Alors, il existe une unique matrice* A, *à une permutation près, telle que* $\boldsymbol{X} = A\boldsymbol{Z} + \boldsymbol{E}$ *dans* (1). *Cela implique que les clusters souples associés* G_a, *pour* $1 \leq a \leq K$, *sont identifiables, à une permutation près.*

3 Estimation

Supposons que $(\mathbf{X}_t, t \in \mathbb{Z}) = (X_t^{(1)}, \ldots, X_t^{(d)}, t \in \mathbb{Z})$ soit un processus strictement stationnaire multivarié, et que $(\mathbf{X}_t, t = 1 \ldots, n)$ soit des données observables. Soit $m \in \{1, \ldots, n\}$ un paramètre de taille de bloc et, pour $i = 1, \ldots, k$ et $j = 1, \ldots, d$, soit $M_{m,i}^{(j)} = \max\{X_t^{(j)}, : , t \in [(i-1)m, \ldots, im]\}$ le maximum des observations du $i$ème bloc dans la $j$ème coordonnée. Pour $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$, soit

$$\begin{aligned}
&\mathbf{M}_{m,i} = (M_{m,i}^{(1)}, \ldots, M_{m,i}^{(d)}),\\
&F_m^{(j)}(x) = \mathbb{P}\{M_{m,1}^{(j)} \leq x\},\\
&\mathbf{F}_m(\mathbf{x}) = (F_m^{(1)}(x^{(1)}), \ldots F_m^{(d)}(x^{(d)})),\\
&U_{m,i}^{(j)} = F_m^{(j)}(M_{m,i}^{(j)}),\\
&\mathbf{U}_{m,i} = (U_{m,i}^{(1)}, \ldots, U_{m,i}^{(d)}).
\end{aligned}$$

Ensuite, nous supposons que les marginales de $X_1^{(1)}, \ldots, X_1^{(d)}$ sont continues. Dans ce cas, les marginales de $\mathbf{M}_{m,1}$ sont également continues et

$$C_m(\mathbf{u}) = \mathbb{P}\{U_{m,1}^{(1)} \leq \mathbf{u}\}, \quad \mathbf{u} \in [0, 1]^d,$$

est la copule unique associée à $\mathbf{M}_{m,1}$. Soit $\Delta_{d-1} = \{(w^{(1)}, \ldots, w^{(d)}) \in [0, \infty)^d : \sum_{j=1}^d w^{(j)} = 1\}$ le simplexe unité dans $\mathbb{R}^d$. Tout au long, nous travaillerons sous le mécanisme suivante de génération de données.

Définition 2 (Mécanisme de génération de données) *Soit* $(\boldsymbol{X}_t, t \in \mathbb{Z})$ *un processus strictement stationnaire multivarié,* $(\boldsymbol{X}_t, t = 1, \ldots, n)$ *les données observables et* C_m *la copule*

associée aux maxima composante par composante pour $m \in \{1, \ldots, n\}$. *Il existe une copule* C_∞, *et une mesure Borélienne finie* $S_{\boldsymbol{X}}$ *sur* Δ_{d-1} *comme définie par* (2) *telle que*

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

où

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

et la fonction de dépendance caudale $L : [0,\infty)^d \to [0,\infty)$ *est décrite par*

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

Notre procédure d'estimation s'inspire de [1] et se compose des quatre étapes suivantes :

(a) Estimer le nombre de clusters K, l'ensemble de variables pures I et la partition $\mathcal{I}$;

(b) Estimer A_I, la sous-matrice de A avec les lignes $A_{i\cdot}$ correspondant à $i \in I$;

(c) Estimer A_J, la sous-matrice de A avec les lignes $A_{j\cdot}$ correspondant à $j \in J$;

(d) Estimer les clusters souples $\mathcal{G} = \{G_1, \ldots, G_K\}$.

Dans le contexte de notre analyse, nous devons estimer les sous-matrices, désignées par A_I et A_J, séparément. Pour commencer avec A_I, nous lançons la procédure d'estimation en déterminant $[K]$, ce qui nous permet ensuite d'identifier I et sa partition, désignée par $\mathcal{I} = \{I_1, \ldots, I_K\}$. Pour effectuer cette étape, nous utilisons la preuve constructive fournie par le Théorème 2, en substituant l'inconnue $\mathcal{X}$ par sa version échantillonnée $\widehat{\mathcal{X}} = [\hat{\chi}_{n,m}(i,j)]_{i,j=1,\ldots,d}$. Pour plus de détails sur cette étape, veuillez vous réferer à l'Algorithme PureVar.

Algorithm PureVar

procedure PUREVAR($\widehat{\mathcal{X}}$,δ)
  Initilisation : $\mathcal{I} = \emptyset$
  Construire le graphe $G = (V, E)$ où $V = [d]$ et $(i,j) \in E$ si $\hat{\chi}_{n,m}(i,j) \leq \delta$
  Trouver une clique maximal, $\mathcal{G}$, de G
  for $i \in \mathcal{G}$ **do**
    $\hat{I}^{(i)} = \{j \in [d] \setminus \{i\} : 1 - \hat{\chi}_{n,m}(i,j) \leq \delta\}$
    $\hat{I}^{(i)} = \hat{I}^{(i)} \cup \{i\}$
    $\hat{\mathcal{I}} = \mathbf{MERGE}(\hat{I}^{(i)}, \hat{\mathcal{I}})$
  Retourne $\hat{\mathcal{I}}$ et $\hat{K}$ comme le nombre d'ensembles dans $\hat{\mathcal{I}}$

Nous continuons en estimant la matrice A_J, ligne par ligne. Pour expliquer notre approche, nous décrivons d'abord la structure de chaque ligne, notée $A_{j\cdot}$, dans la matrice A_J,

pour $j \in J$. Nous devons noter que chaque $A_{j\cdot}$ satisfait des conditions de parcimonie et $\sum_{a=1}^{K} A_{ja} = 1$, tel que stipulé par la Condition (i). Ainsi, pour chaque $i \in I_a$ avec $a \in [K]$ et $j \in J$, nous avons

$$\chi(i,j) = A_{ja}.$$

En moyennant l'affichage ci-dessus sur tous les $i \in I_a$, nous obtenons

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

Par conséquent

$$\beta^{(j)} := A_{j\cdot} = \left(\frac{1}{|I_1|} \sum_{i \in I_1} \chi(i,j), \ldots, \frac{1}{|I_K|} \sum_{i \in I_K} \chi(i,j) \right),$$

qui peut être estimé à partir des données comme suit. Pour chaque $j \in \hat{J}$, nous désignons un estimateur pour le a-ième élément de $\beta^{(j)}$ en utilisant une approche simple. Cet estimateur est représenté comme suit

$$\bar{\chi}^{(j)} = \left(\frac{1}{|\hat{I}_1|} \sum_{i \in \hat{I}_1} \hat{\chi}_{n,m}(i,j), \ldots, \frac{1}{|\hat{I}_K|} \sum_{i \in \hat{I}_K} \hat{\chi}_{n,m}(i,j) \right).$$

Il est important de noter que cet estimateur n'est ni parcimonieux ni un élément du simplexe unitaire. Étant donné la valeur $\bar{\chi}^{(j)}$, notre objectif est de déterminer une projection euclidienne de $\bar{\chi}^{(j)}$ qui se situe dans l'espace $\mathbb{B}_0(s) = \{\beta \in \mathbb{R}^{\hat{K}}, \sum_{j=1}^{\hat{K}} \mathbb{1}_{\{\beta^{(j)} \neq 0\}} \leq s\}$, c'est-à-dire, les vecteurs ayant au plus s entrées non nulles, et le simplexe unitaire $\Delta^{\hat{K}-1} = \{\beta \in \mathbb{R}^{\hat{K}},\ \beta^{(j)} \geq 0, \sum_{j=1}^{\hat{K}} \beta^{(j)} = 1\}$:

$$\mathcal{P}(\hat{\beta}^{(j)}) \in \underset{\beta : \beta \in \mathbb{B}_0(s) \cap \Delta^{\hat{K}-1}}{\operatorname{argmin}} ||\beta - \bar{\chi}^{(j)}||_2. \tag{7}$$

Par conséquent, pour construire un estimateur du support, nous sélectionnons uniquement les coordonnées indexées par a où $\bar{\chi}_a^{(j)}$ dépasse un seuil δ. Cette sélection donne un estimateur parcimonieux pour $\beta_a^{(j)}$ comme suit

$$\bar{\beta}_a^{(j)} = \bar{\chi}_a^{(j)} \mathbb{1}_{\{\bar{\chi}_a^{(j)} > \delta\}},\ j = 1, \ldots, \hat{K}.$$

Cependant, il est essentiel de noter que l'estimateur $\bar{\beta}^{(j)}$ n'appartient pas intrinsèquement au simplexe unitaire. Pour remédier à cela, nous pouvons obtenir un estimateur alternatif, noté $\hat{\beta}^{(j)}$, en projetant $\bar{\beta}^{(j)}$ sur le simplexe unitaire dans l'espace $\hat{K}$-dimensionnel. L'opération de projection sur le simplexe unitaire est réalisée en utilisant un opérateur mathématique spécifique, défini comme suit

$$(\mathcal{P}_{\Delta_{\hat{K}-1}}(\beta))_j = [\beta^{(j)} - \tau]_+,\ \tau := \frac{1}{\rho} \left(\sum_{i=1}^{\rho} \beta^{(j)} - 1 \right),$$

pour $\rho := \max k, , \beta^{(j)} > \frac{1}{k}(\sum_{j=1}^{k} w^{(j)} - 1)$. Ainsi, en désignant $\widehat{\mathcal{S}} = \operatorname{supp}(\bar{\beta}^{(j)})$, nous obtenons

$$\hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}} = \mathcal{P}_{\Delta_{\hat{K}-1}}(\bar{\beta}^{(j)}\big|_{\widehat{\mathcal{S}}}), \quad \hat{\beta}^{(j)}\Big|_{\widehat{\mathcal{S}}^c} = 0. \tag{8}$$

Ensuite, nous construisons la matrice $\hat{A}_{\hat{J}}$ avec des lignes correspondant aux estimateurs $\hat{\beta}^{(j)}$ pour chaque $j \in \hat{J}$. Notre estimateur final, noté $\hat{A}$, pour la matrice A, est obtenu en concaténant $\hat{A}_{\hat{I}}$ et $\hat{A}_{\hat{J}}$. Les propriétés statistiques de l'estimateur final sont examinées dans la Section 4, où nous fournissons également des spécifications détaillées du paramètre de réglage nécessaire à son développement.

4 Garanties statistiques

Nous plongeons dans l'analyse des performances statistiques de notre estimateur, noté $\hat{A}$, qui vise à estimer A. En plus de cette estimation, nous considérons également sa partition associée. Dans le contexte de notre section, introduisons quelques notations et concepts. Considérons la quantité $\chi(i,j)$, qui représente la corrélation extrême entre $X^{(i)}$ et $X^{(j)}$, dans le domaine d'attraction tel que spécifié par la condition MDA. Nous définissons également un paramètre crucial noté :

$$d_m = \sup_{1 \le i < j \le d} |\chi_m(i,j) - \chi(i,j)|,$$

où $\chi_m(i,j)$ est la corrélation extrême sous-asymptotique entre $M_{m,1}^{(i)}$ et $M_{m,1}^{(j)}$. Ce paramètre caractérise le biais explicite entre le cadre sous-asymptotique et le domaine d'attraction maximal. Il quantifie essentiellement le taux de convergence du système vers son comportement asymptotique. De plus, nous introduisons le nouvel événement suivant :

$$\mathcal{E} = \mathcal{E}(\delta) := \left\{ \sup_{1 \le i < j \le d} |\hat{\chi}_{n,m}(i,j) - \chi(i,j)| \le \delta \right\}. \tag{9}$$

En prenant $c_1 > 0$ suffisamment grand et

$$\delta = d_m + c_1 \left(\sqrt{\frac{\ln(kd)}{k}} + \frac{\ln(k) \ln\ln(k) \ln(kd)}{k} \right),$$

[2, Theorem 4] garantit que $\mathcal{E}$ est vérifié avec grande probabilité :

$$\mathbb{P}(\mathcal{E}) \ge 1 - d^{-c_0},$$

pour une certaine constante positive $c_0 > 0$. Nous considérons la fonction de perte pour deux matrices $d \times K$ A, A' comme

$$L_2(A, A') := \min_{P \in S_K} ||AP - A'||_{\infty,2} \tag{10}$$

où S_K est le groupe de toutes les matrices de permutation $K \times K$ et

$$||A||_{\infty,2} := \max_{1 \le j \le d} ||A_{j\cdot}||_2 = \max_{1 \le j \le d} \left(\sum_{j=1}^{K} |A_{ij}|^2 \right)^{1/2};$$

pour une matrice générique $A \in \mathbb{R}^{d \times K}$. Enfin, étant donné δ, nous définissons

$$J_2 = \{ j \in J : \text{pour chaque } a \in [K] \text{ avec } A_{ja} \neq 0, A_{ja} > 2\delta \}. \tag{11}$$

Théorème 4 *Soit* $(\boldsymbol{X}_t, t \in \mathbb{Z})$ *une séquence de variables aléatoires décrit par la Définition 2 avec des coefficients de mélange fort décroissants exponentiellement. Fixons* $s = \max_{j\in[d]} ||A_{j\cdot}||_0$. *Alors, sous des conditions de signaux suffisaments forts, pour l'estimateur* $\hat{A}$, *les résultats suivants sont vérifiés.*

(a) *Récupération des facteurs latents :*

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

(b) *Une borne supérieure sur* $\hat{A}$ *:*

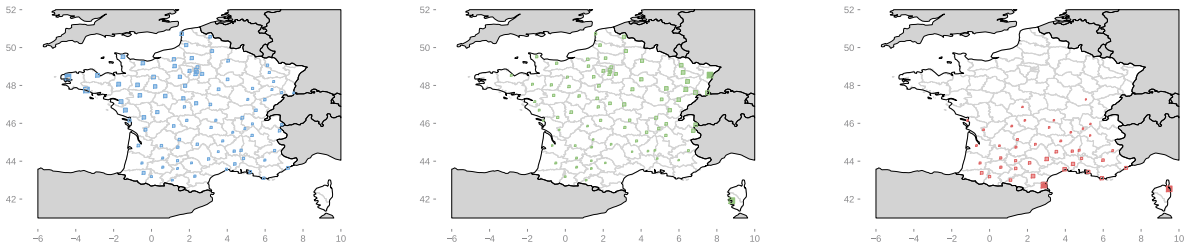
$$L_2(\hat{A}, A) \leq 4\sqrt{s}\delta,$$

(c) *Une garantie pour la récupération du support :*

$$supp(A_{J_2}) \subseteq supp(\hat{A}) \subseteq supp(A),$$

avec probabilité plus grande que $1 - d^{-c_0}$ *pour une constante positive* c_0.

5 Précipitations extrêmes en France



(a) (b) (c)

Figure 1: Dans le Panel (a), nous avons la représentation spatiale des clusters liés à la variable latente ouest. En passant au Panel (b), nous rencontrons les clusters associés à la variable latente est. Enfin, dans le Panel (c), la représentation spatiale se déploie pour les clusters liés à la variable latente sud. La force d'association de chaque emplacement avec la variable latente respective est transmise par la taille proportionnelle et l'intensité de couleur du carré.

Dans notre analyse, nous nous concentrons sur les maxima hebdomadaires des précipitations horaires enregistrées dans 92 stations météorologiques en France pendant la saison automnale, s'étendant de septembre à novembre, pour les années 1993 à 2011, ce qui donne un total de 228 maxima par bloc.

Nous proposons une approche basée sur les données pour sélectionner la valeur seuil de δ. En utilisant le seuil désigné, nous révélons trois variables latentes situées dans les

régions ouest, est et sud de la France. Il est crucial de souligner que notre processus fonctionne uniquement sur la base des enregistrements de précipitations, sans aucune information géographique. Par conséquent, discerner des structures spatiales cohérentes à partir de mesures de précipitations uniquement n'est pas un résultat évident. La représentation spatiale des clusters est illustrée dans la Figure 1. La zone ouest au-dessus de Bordeaux, indiquée dans la Figure 1 Panel (a), présente des dépendances robustes avec la région centrale autour de Paris. Cependant, au-delà de ces régions, les associations avec la variable latente diminuent rapidement. Symétriquement, la région est, s'étendant de Lyon et couvrant les Vosges, l'Alsace, la Franche-Comté et les régions du nord-est de la France, représentée dans la Figure 1 Panel (b), montre des dépendances avec les régions centrales tout en diminuant rapidement en dehors de cette zone. Il est à noter que plus un emplacement est éloigné de la variable pure, moins est l'affiliation correspondante. Le cluster sud, dans la Figure 1 Panel (c), met en évidence les dépendances spatiales sur la Corse et les villes méditerranéennes. Ces associations diminuent rapidement, ce qui entraîne la formation d'un cluster moins étendu.

References

[1] Xin Bing et al. "Adaptive estimation in structured factor models with applications to overlapping clustering". In: *The Annals of Statistics* 48.4 (2020), pp. 2055–2081. DOI: 10.1214/19-AOS1877. URL: https://doi.org/10.1214/19-AOS1877.

[2] Alexis Boulin. "Estimating Max-Stable Random Vectors with Discrete Spectral Measure using Model-Based Clustering". In: *arXiv preprint arXiv:2402.01609* (2024).

[3] Rafal Kulik and Philippe Soulier. *Heavy-tailed time series*. Springer, 2020.

[4] Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.