High-dimensional clustering of sub-asymptotic maxima of a weakly dependent process

Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde

High-dimensional variable clustering based on sub-asymptotic maxima of a weakly dependent random process

Alexis Boulin [*,1,3] & Elena Di Bernardino [1] & Thomas Laloë [1] & Gwladys Toulemonde [2,3]

[1] *Université Côte d'Azur, CNRS, LJAD, France*
[2] *Univ Montpellier, CNRS, Montpellier, France*
[3] *Inria, Lemon*
[*] *Corresponding author: aboulin@unice.fr*

Résumé. Nous présentons une classe de modèles novateurs pour le partitionnement de variables, nommée modèles Asymptotic Independent block (AI-block). Cette classe de modèles définit les clusters en se basant sur l'indépendance extrémale entre les composantes d'un processus stationnaire multivarié. Nous avons observé que cette classe de modèles est identifiable, dans le sens où il existe un élément maximal selon un certain ordre partiel entre les partitions, ce qui permet une inférence statistique. Nous avons développé un algorithme dédié pour identifier les clusters de variables, sans la nécessité de spécifier leur nombre préalablement. Nos travaux ont également fourni des éléments théoriques sur la consistance de l'algorithme, démontrant que dans certaines conditions, il identifie les clusters avec une complexité polynomiale en la dimension. Nous avons étudié l'inférence pour un processus stationnaire multivarié mélangeant, ce qui permet une large gamme d'applications, telles que les séries temporelles financières ou climatiques. Afin d'atteindre cet objectif, nous avons étendu certains des résultats théoriques concernant les dépendances extrémales pour des sous-vecteurs aléatoires dans un contexte de dépendance temporelle.

Mots-clés. Estimation consistante, Théorie des valeurs extrêmes, Modèles à grande dimension, Clustering de variables.

Abstract. We propose a new class of models for variable clustering called Asymptotic Independent block (AI-block) models, which defines population-levels clusters based on the independence of the maxima of a multivariate stationary random process among clusters. This class of models is identifiable, meaning that there exists a maximal element with a partial order between partitions, allowing for statistical inference. We also present a dedicated algorithm for recovering the clusters of variables without specifying the number of clusters a priori. Our work provides some theoretical insights into the consistency of our algorithm, demonstrating that under certain conditions, it can effectively identify clusters in the data with a computational complexity that is polynomial in the dimension. This inference is examined in the framework of a multivariate stationary mixing random process to stay within a wide range of statistical practices, where notable examples include financial and climate time-series data. For that purpose, some theoretical results for extremal independence are extended to the context of random subvectors and temporal dependence.

Keywords. Consistent estimation, Extreme value theory, High dimensional models, Variable clustering.

1 Introduction

Les résultats de la théorie des valeurs extrêmes ont montré que la structure de dépendance des extrêmes peut être décrite de diverses manières, telles que par la mesure exponentielle, la fonction de dépendance de Pickands, la fonction de dépendance caudale, le madogramme et la copule des valeurs extrêmes. Bien que la modélisation des extrêmes dans un cadre univarié et en petite dimension ait été largement étudiée, la modélisation des extrêmes multivariées pourrait être améliorée, en particulier dans le cas de la grande dimension. Des recherches récentes dans ce domaine se concentrent sur la connexion entre l'étude des extrêmes multivariés et les méthodes de statistiques modernes et les techniques de machine learning. Dans cette direction de recherche, nous proposons un algorithme de clustering pour apprendre la structure de dépendance des extrêmes multivariés.

Le clustering peut être appliqué sur $\mathbf{X}_1, \ldots, \mathbf{X}_n$, avec n étant le nombre d'observations d'un vecteur aléatoire $\mathbf{X} \in \mathbb{R}^d$, à travers deux approches distinctes : la partition de l'ensemble des indices des lignes $\{1, \ldots, n\}$ pour le clustering de données, et la partition de l'ensemble des indices des colonnes $\{1, \ldots, d\}$ pour le clustering de variables, qui est le sujet principal de cette présentation. Le clustering de variables a des applications pratiques importantes, notamment dans l'étude des événements climatiques extrêmes (Bador et al. (2015); Bernard et al. (2013); Saunders et al. (2021)), où un phénomène spatial est observé sur un nombre limité de sites. Un cas particulier consiste à partitionner ces sites en fonction de leurs dépendances extrêmes, ce qui peut être effectué à l'aide de techniques d'apprentissage non supervisé telles que les k-means ou le clustering hiérarchique avec une dissimilarité propre aux extrêmes. Toutefois, les propriétés statistiques de ces procédures n'ont pas encore été étudiées de manière exhaustive et il reste à déterminer les modèles probabilistes sur $\mathbf{X}$ susceptibles d'être estimés par ces techniques. Dans cet article, nous envisageons un model-based clustering où les clusters sont bien définis, offrant ainsi une interprétation des résultats et d'évaluer la performance d'un algorithme spécifique.

L'hypothèse d'indépendance et d'identité de distribution des observations est une hypothèse fondamentale dans la théorie statistique mais rarement vérifiée dans l'étude des séries temporelles, car les observations sont souvent dépendantes les unes aux autres. Les processus multivariés mélangeants sont une approche courante pour modéliser les séries temporelles dépendantes. Ces processus supposent que la dépendance va diminuer progressivement le long de la trajectoire.

Voici notre contribution : nous proposons un cadre probabiliste, les modèles AI-block, pour regrouper les composantes d'un vecteur selon les extrêmes. Ces modèles sont construits en supposant que les extrêmes d'un cluster de variables à l'autre d'un processus multivarié sont indépendants. Cette approche présente l'avantage d'être étudiable théoriquement et nous démontrons que cette classe de modèle est identifiable (Théorème 1). Nous développons un algorithme et nous montrons sa consistance (voir Théorème 2). Cette analyse est effectuée dans le contexte de l'estimation non paramétrique avec la méthode des maxima par blocs, dont la longueur de bloc est un paramètre de réglage.

Notations Les lettres en gras $\mathbf{x}$ représentent des vecteurs dans $\mathbb{R}^d$. En notant $B \subseteq \{1,\dots,d\}$, le $|B|$-sous-vecteur de $\mathbf{x}$ est noté $\mathbf{x}^{(B)} = (X^{(j)})_{j\in B}$. Nous définissons $\mathbf{X} \in \mathbb{R}^d$ comme un vecteur aléatoire avec une fonction de répartition H, et $\mathbf{X}^{(B)}$ comme un sous-vecteur aléatoire de $\mathbf{X}$ avec une fonction de répartition

$$H^{(B)}(\mathbf{x}^{(B)}) = H(\mathbf{1}, \mathbf{x}^{(B)}, \mathbf{1}), \quad (X^{(j)})_{j\in B} \in [0,1]^{|B|},$$

où $(\mathbf{1}, \mathbf{x}^{(B)}, \mathbf{1})$ a sa $j$ème composante égale à $x^{(j)}\mathbb{1}_{j\in B} + \mathbb{1}_{j\notin B}$. De même, nous notons $(\mathbf{0}, \mathbf{x}^{(B)}, \mathbf{0})$ un vecteur de $\mathbb{R}^d$ pour lequel la $j$ème composante est égale à $x^{(j)}$ si $j \in B$ et 0 sinon. Lorsque $B = \{1,\dots,d\}$, nous notons H au lieu de $H^{(\{1,\dots,d\})}$. Soit $O = \{O_g\}_{g=\{1,\dots,G\}}$ une partition de $\{1,\dots,d\}$ en G groupes, et $\mathbf{X}^{(O_g)}$, $g \in \{1,\dots,G\}$ des vecteurs aléatoires avec $\mathbf{X} = (\mathbf{X}^{(O_1)},\dots,\mathbf{X}^{(O_G)})$. Nous disons que les vecteurs $\mathbf{X}^{(O_1)},\dots,\mathbf{X}^{(O_G)}$ sont indépendants si et seulement si

$$H(\mathbf{x}) = \Pi_{g=1}^{G} H^{(O_g)}\left(\mathbf{x}^{(O_g)}\right), \quad \mathbf{x} \in \mathbb{R}^d.$$

Dans la Section 2, nous allons présenter les concepts nécessaires pour comprendre notre proposition, qui se concentrent sur les processus faiblement dépendants ainsi que la théorie des valeurs extrêmes. Nous présenterons ensuite notre classe de modèles pour le partitionement des extrêmes. Nous allons également présenter un algorithme dédié à cette classe de modèles dans la Section 3. Nous démontrerons que cet algorithme peut retrouver les clusters de variables avec une grande probabilité (voir Théorème 2). En outre, des résultats numériques dans la Section 4 confirmeront nos résultats théoriques et nous permettront de comparer notre algorithme avec d'autres méthodes de clustering bien connues dans la littérature.

2 Un modèle pour le clustering des variables

2.1 Cadre d'analyse

Considérons $\mathbf{Z} = (Z^{(1)},\dots,Z^{(d)})$ et $\mathbf{Z}_t = (Z_t^{(1)},\dots,Z_t^{(d)})$ où $t \in \mathbb{Z}$ comme étant respectivement un vecteur aléatoire de loi F, et un processus multivarié stationnaire distribué selon $\mathbf{Z}$. Pour le processus $(\mathbf{Z}_t, t \in \mathbb{Z})$, nous définissons

$$\mathcal{F}_k = \sigma(\mathbf{Z}_t, t \le k), \quad \mathcal{G}_k = \sigma(\mathbf{Z}_t, t \ge k),$$

qui sont respectivement la filtration naturelle et la filtration rétrograde de $(\mathbf{Z}_t, t \in \mathbb{Z})$. En considérant la quantité suivante

$$\varphi(\mathcal{A}_1, \mathcal{A}_2) = \sup_{A_1,A_2\in\mathcal{A}_1\times\mathcal{A}_2} \left|\mathbb{P}(A_2|A_1) - \mathbb{P}(A_1)\right|,$$

nous pouvons définir le coefficient de mélange φ comme suit

$$\varphi(\ell) = \sup_{t\in\mathbb{Z}} \varphi(\mathcal{F}_t, \mathcal{G}_{t+\ell}). \tag{1}$$

Soit $\mathbf{M}_m = (M_m^{(1)}, \ldots, M_m^{(d)})$ le vecteur des maxima composante par composante où $M_m^{(j)} = \bigvee_{j=1}^{d} Z_i^{(j)}$. Nous considérons un vecteur aléatoire $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ distribué selon H. Une fonction normalisante a sur $\mathbb{R}$ est une fonction croissante, continue à droite, qui converge vers $\pm\infty$. Dans la théorie des valeurs extrêmes, un problème fondamental est de caractériser la distribution limite H dans l'équation suivante

$$\lim_{m\to\infty} \mathbb{P}\{\mathbf{M}_m \leq \mathbf{a}_m(\mathbf{x})\} = H(\mathbf{x}),$$

où $\mathbf{a}_m = (a_m^{(1)}, \ldots, a_m^{(d)})$ avec $a_m^{(j)}, 1 \leq j \leq d$ sont des fonctions normalisantes, et H est une distribution non dégénérée. Typiquement, H est une distribution à valeurs extrêmes et $\mathbf{X}$ est un vecteur max-stable avec des marges distribuées selon des lois à valeurs extrêmes généralisées. Dans ce cas, nous pouvons écrire

$$\mathbb{P}\{\mathbf{X} \leq \mathbf{x}\} = \exp\{-\Lambda(E \setminus [0, \mathbf{x}])\},$$

où Λ est une mesure de Radon sur le cône $E = [0, \infty)^d \setminus \{\mathbf{0}\}$. Lorsque $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ sont des vecteurs aléatoires à valeurs extrêmes indépendants, i.e.,

$$H(\mathbf{x}) = \Pi_{g=1}^{G} H^{(O_g)}(\mathbf{x}^{(O_g)}), \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

nous dirons que $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ sont des vecteurs aléatoires asymptotiquements indépendants.

2.2 Modèles AI-block

Lorsqu'on parle de vecteurs aléatoires à valeurs extrêmes $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$, il est important de noter que, sous l'hypothèse d'indépendance, le vecteur aléatoire $\mathbf{X} = (\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)})$ est à valeurs extrêmes. Dans notre modèle, défini comme suit, nous considérons un processus stationnaire d-varié $(\mathbf{Z}_t, t \in \mathbb{Z})$ avec loi F, ainsi qu'un vecteur aléatoire à valeur extrême $\mathbf{X}$ de distribution H :

Définition 1. *Supposons $(\mathbf{Z}_t, t \in \mathbb{Z})$ être un processus d-varié stationnaire avec loi F, et $\mathbf{X}$ un vecteur aléatoire à valeur extrême de distribution H. On dit que le processus aléatoire $(\mathbf{Z}_t, t \in \mathbb{Z})$ suit un modèle AI-block si $F \in D(H)$ et si les vecteurs aléatoires $\mathbf{X}^{(O_g)}$ pour tout $g \in \{1, \ldots, G\}$ sont des vecteurs aléatoires à valeurs extrêmes asymptotiquement indépendants.*

Afin de pouvoir identifier notre modèle, nous introduisons la notation $\mathbf{Z} \sim O$, où O est une partition de $\{1, \ldots, d\}$, pour dire que $\mathbf{Z}$ suit un modèle AI-block avec la partition O. Nous notons également par $\mathcal{L}(\mathbf{Z})$ l'ensemble des partitions possibles de $\mathbf{Z}$, qui est non vide et fini, et qui admet des éléments maximums. Nous définissons un ordre partiel sur les partitions comme suit : pour deux partitions $O = \{O_g\}_g$ et $S = \{S_{g'}\}_{g'}$, nous disons que S est une sous-partition de O si pour tout g', il existe g tel que $S_{g'} \subseteq O_g$. Nous définissons alors $\leq$, l'ordre partiel entre deux partitions O et S de $\{1, \ldots, d\}$ comme suit :

$$O \leq S, \quad \text{si } S \text{ est une sous-partition de } O. \tag{2}$$

En utilisant cette notation, nous pouvons maintenant énoncer précisément l'identifiabilité de notre modèle dans le théorème ci-dessous.

Théorème 1. *L'ensemble $\mathcal{L}(\mathbf{Z})$ a un unique élément maximal $\bar{O}(\mathbf{Z})$ par rapport à l'ordre partiel $\leq$ défini en* (2).

La prochaine section abordera la forme particulière de la mesure exponentielle lorsque les vecteurs aléatoires à valeurs extrêmes $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ sont indépendants. Il est possible d'obtenir d'autres caractérisations, qui sont détaillées dans Boulin et al. (2023).

2.3 Structure de dépendance extrémale pour les modèles AI-block

Considérons le vecteur aléatoire $\mathbf{X} = (\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)})$. Le vecteur aléatoire $\mathbf{X}$ étant à valeur extrême, il existe une mesure exponentielle décrivant sa structure de dépendance. Plus particulièrement, sa mesure exponentielle est définie par la proposition suivante, qui décrit sa forme spécifique lorsque les composantes $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ sont indépendantes.

Proposition 1. *Supposons que le vecteur $\mathbf{X}$ est a valeurs extrêmes et que sa mesure exponentielle Λ est concentrée sur le cône $E = [0, \infty]^d \setminus \mathbf{0}$. Les trois propriétés suivantes sont équivalentes :*

1. *Les vecteurs $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ sont indépendants.*
2. *Les vecteurs sont indépendants par blocs : pour tout $1 \leq g < h \leq G$, les vecteurs $\mathbf{X}^{(O_g)}$ et $\mathbf{X}^{(O_h)}$ sont des vecteurs aléatoires indépendants.*
3. *La mesure exponentielle Λ se concentre sur l'ensemble*

$$\bigcup_{g=1}^{G} \{\mathbf{0}\}^{d_1} \times \cdots \times]0, \infty[^{d_g} \times \cdots \times \{\mathbf{0}\}^{d_G},$$

ce qui signifie que, pour toute valeur $\mathbf{y} > 0$,

$$\Lambda\left(\bigcup_{1 \leq g < h \leq G} \left\{ \mathbf{x} \in E, \exists a \in O_g, x^{(a)} > y^{(a)}, \exists b \in O_h, x^{(b)} > y^{(b)} \right\} \right) = 0.$$

Dans cette section, nous présentons une condition importante qui concerne la dépendance entre les variables d'un vecteur aléatoire à valeurs extrêmes dans chaque cluster. Cette condition nous permet d'introduire un algorithme simple mais puissant pour comparer les dépendances extrêmes bivariées entre les différentes variables. L'algorithme détient l'avantage d'identifier l'élement maximal de notre modèle sans avoir à spécifier le nombre de clusters a priori. De plus nous pouvons alors tirer des conclusions théoriques sur l'efficacité de notre méthode, ainsi que sur la force de la dépendance nécessaire entre les composantes de chaque cluster pour que la partition la plus fine puisse être retrouvée. Les détails de cette condition et de l'algorithme seront présentés dans la Section 3.

Condition $\mathcal{A}$. *Pour chaque $g \in \{1, \dots, G\}$, le vecteur aléatoire à valeur extrême $\mathbf{X}^{(\bar{O}_g)}$, où $\bar{O}_g$ est l'élément maximal de $\mathcal{L}(\mathbf{Z})$, présente de la dépendance entre chaque composante.*

Une condition suffisante pour satisfaire l'hypothèse $\mathcal{A}$ est de supposer que la mesure exponentielle de chaque vecteur aléatoire $\mathbf{X}^{\bar{O}_g}$ est dominée par la mesure de Lebesgue restreinte à $[0, \infty)^{d_g} \setminus \{\mathbf{0}\}^{d_g}$ pour chaque $g \in \{1, \dots, G\}$.

3 Inférence dans les modèles AI-block

Dans cette section, nous présentons une version adaptée de l'algorithme Bunea et al. (2020) pour partitionner des variables, basée sur une métrique entre covariance, nommée CORD. Notre adaptation implique l'utilisation de la corrélation extrémale qui mesure la dépendance extrémale entre deux variables. Cette grandeur est définie dans Coles et al. (1999)

$$\chi(a, b) = \lim_{q \to 0} \chi_q(a, b) = \mathbb{P}\{H^{(a)}(X^{(a)}) > 1 - q | H^{(b)}(X^{(b)}) > 1 - q\},$$

si la limite existe. En particulier, si $\mathbf{X}$ est une distribution multivariée à valeurs extrêmes, alors $\chi_q(a, b) = \chi(a, b)$ pour tout $q \in (0, 1)$ et $a, b \in \{1, \dots, d\}$. Dans un modèle AI-block, l'énoncé

$$\mathbf{X}^{(O_g)} \perp\!\!\!\perp \mathbf{X}^{(O_h)}, \quad g \neq h$$

est équivalent à

$$\chi(a, b) = \chi(b, a) = 0, \quad \forall a \in O_g, \forall b \in O_h, g \neq h.$$

En utilisant la Condition $\mathcal{A}$ et la Proposition 1, la corrélation extrémale est une quantité suffisante pour retrouver les clusters dans un modèle AI-block.

Nous définissons $\mathcal{X} = (\chi(a, b))_{a,b \in \{1,\dots,d\}}$ comme étant la matrice des corrélations extrêmes, et $\hat{\mathcal{X}} = (\hat{\chi}_{n,m}(a, b))_{a,b \in \{1,\dots,d\}}$ comme étant sa version estimée obtenue en utilisant le madogramme (voir, par exemple, Cooley et al. (2006); Naveau et al. (2009); Boulin et al. (2022)) où m désigne la longueur du bloc dans lequel sont évalués les maxima. Nous présentons notre algorithme, nommé (ECO) (Extremal COrrelation), pour estimer la partition $\bar{O}$.

Une considération importante est le choix du seuil τ. Si $\tau \approx 0$, l'algorithme va vraisemblablement retourner le cluster $\{1, \dots, d\}$, tandis que si $\tau \approx 1$, l'algorithme aura tendance à retourner le plus grand cluster $\{\{1\}, \dots, \{d\}\}$. Pour aider au choix de la valeur de ce paramètre, nous menons une étude non asymptotique de l'algorithme, en considérant d et n fixés, ce qui permet une étude dans le cas d'une dimension croissante en apportant une borne non asymptotique de la probabilité de retrouver les clusters via l'algorithme (ECO). Avant d'introduire ce résultat, nous présentons une métrique MECO (Minimal Extremal COrrelation), qui permet de mesurer la difficulté de clustering dans un modèle AI-block :

$$\mathrm{MECO}(\mathcal{X}) := \min_{a \overset{\bar{O}}{\sim} b}, \chi(a, b).$$

Dans les modèles AI-block, sous la Condition $\mathcal{A}$, nous avons $\mathrm{MECO}(\mathcal{X}) > \eta$ avec $\eta = 0$. Cependant, une valeur de η plus élevée sera nécessaire pour récupérer la partition $\bar{O}$.

Algorithm (ECO) Clustering procedure for AI-block models

procedure ECO(S, τ, $\hat{\mathcal{X}}$)
  Initialize: $S = \{1, \dots, d\}$, $\hat{\chi}_{n,m}(a, b)$ for $a, b \in \{1, \dots, d\}$ and $l = 0$
  while $S \neq \emptyset$ **do**
    $l = l + 1$
    if $|S| = 1$ **then**
      $\hat{O}_l = S$
    if $|S| > 1$ **then**
      $(a_l, b_l) = \arg\max_{a,b \in S} \hat{\chi}_{n,m}(a, b)$
      if $\hat{\chi}_{n,m}(a_l, b_l) \leq \tau$ **then**
        $\hat{O}_l = \{a_l\}$
      if $\hat{\chi}_{n,m}(a_l, b_l) > \tau$ **then**
        $\hat{O}_l = \{s \in S : \hat{\chi}_{n,m}(a_l, s) \wedge \hat{\chi}_{n,m}(b_l, s) \geq \tau\}$
    $S = S \setminus \hat{O}_l$
  return $\hat{O} = (\hat{O}_l)_l$

Théorème 2. *Nous considérons un modèle AI-block comme défini dans la Définition 1 sous la Condition $\mathcal{A}$ et $(\mathbf{Z}_t, t \in \mathbb{Z})$ un processus multivarié stationnaire avec des coefficients φ-mélangeant tel que $\varphi(n) \leq \lambda n^{-\zeta}$ où $\lambda > 0$ et $\zeta > 1$ et φ défini dans l'Equation (1). Définissons*

$$d_m = \max_{a \neq b} |\chi_m(a, b) - \chi(a, b)| .$$

Soit (τ, η) des paramètres satisfaisant

$$\tau = O\left(d_m + \sqrt{\frac{\ln(d)}{k}}\right), \quad \eta = \tau + O\left(d_m + \sqrt{\frac{\ln(d)}{k}}\right),$$

où k étant le nombre de maxima. Pour une certaine matrice $\mathcal{X}$ et son estimateur correspondant $\hat{\mathcal{X}}$, si $\mathrm{MECO}(\mathcal{X}) > \eta$*, alors la sortie de l'algorithme (ECO) est consistante, c'est-à-dire*

$$\mathbb{P}\left\{\hat{O} = \bar{O}\right\} \geq 1 - 2(1 + \sqrt{e})d^{-2\gamma}, \quad \gamma > 0.$$

En ignorant le biais, disons $d_m = 0$, pour chaque $m \in \mathbb{N}$, une conséquence immédiate du théorème est que notre algorithme est capable de retrouver les clusters dans un régime polynomial, i.e., $d = o(n^p)$, avec $p > 0$. En effet, dans ce régime, la métrique η diminue lorsque que le nombre d'observations augmente, ce qui rend possible l'inférence pour un nombre suffisant de données.

4 Résultats numériques

Au sein de cette section, nous examinons en détail les performances de notre algorithme dans sa capacité à identifier des clusters dans les modèles AI-block. Nous avons pu construire un processus stochastique vérifiant l'ensemble des hypothèses du Théorème 2, voir

Boulin et al. (2023). Pour comparer les performances de notre algorithme (ECO) avec d'autres méthodes déjà bien connues dans la littérature, nous avons opté pour le clustering hiérarchique (HC) utilisant le madogramme comme dissimilarité, ainsi que pour les k-means sphériques (SKmeans). Nos expériences numériques se concentrent sur la partition suivante dans le modèle limite :

E1 On considère $\mathbf{X}$ composé de $G = 10$ clusters dont 5 de tailles aléatoires $d_1, \ldots, d_5$. Chaque vecteur aléatoire est distribué selon une copule logistique de paramètre $10/7$. Les 5 clusters restants consistent en des singletons.

Pour l'expérience E1, nous avons étudié plusieurs cadres pour évaluer la performance de notre méthode. Voici les deux cadres que nous avons considérés :

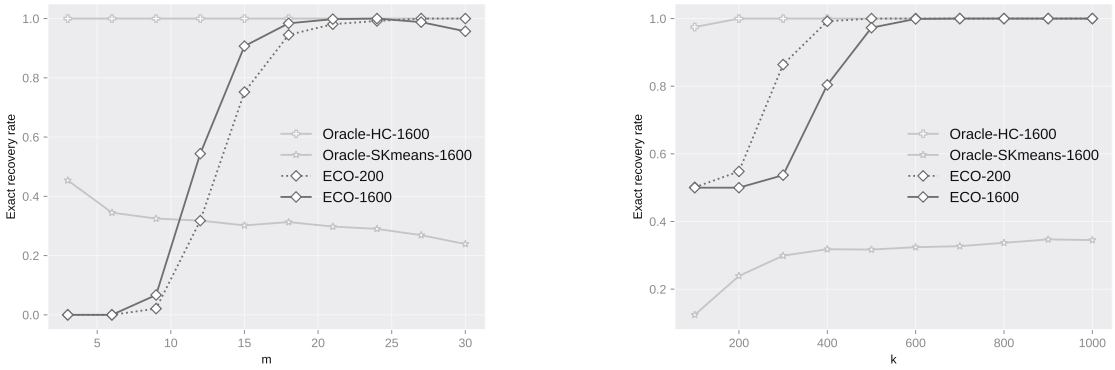
F1 Tout d'abord, nous avons examiné le choix de la suite intermédiaire m, qui représente la taille de bloc utilisée pour l'estimation. Pour ce faire, nous avons effectué des tests en prenant $m \in \{3, 6, \ldots, 30\}$, avec une taille d'échantillon $n = 10000$ et $k = \lceil n/m \rceil$.

F2 Ensuite, nous avons évalué la performance de notre méthode pour différentes tailles d'échantillon. Comme la valeur optimale de m pour une estimation consistante n'est pas connue en pratique, nous avons choisi $m = 20$ comme point de référence.

Dans le cadre de notre étude, nous avons utilisé le seuil $\tau = 2 \times (m^{-1} + \sqrt{\ln(d)k^{-1}})$, comme suggéré par l'ordre de grandeur dans le Théorème 2 (des détails concernant l'ordre de grandeur m^{-1} du biais sont donnés dans (Boulin et al., 2023)). Nous avons ensuite présenté les résultats dans la Figure 1, en affichant le taux de recouvrement exact de notre algorithme ((ECO)) pour deux valeurs de d, à savoir $d = 200$ et $d = 1600$. Dans le cas de la "grande dimension" avec $d = 1600$, nous avons également évalué les performances des algorithmes HC et SKmeans.

Il est intéressant de constater que dans le cadre F1, la performance de notre algorithme augmente avec la taille de bloc m jusqu'à un certain point, puis décroît. Cette tendance reflète le compromis entre le régime sous-asymptotique et la précision de la méthode d'inférence. Dans le cadre F2, nous observons que la performance de notre algorithme s'améliore à mesure que le nombre de maxima par bloc augmente.

Il est à noter que l'algorithme HC utilisant le madogramme comme dissimilarité présente de très bonnes performances dans chaque configuration, même lorsque l'inférence est biaisée. Cela peut s'expliquer par le fait que le madogramme est plus faible lorsque $a \overset{\bar{O}}{\sim} b$ et plus grand lorsque $a \overset{\bar{O}}{\not\sim} b$, ce qui est vrai tant dans le domaine d'attraction que dans le régime sous-asymptotique. Cependant, le SKmeans présente une performance fragile dans les deux cadres considérés.



F1 F2

Figure 1: Résultats pour les expériences dans les cadres F1 et F2. Le panel F1 présente la taux de recouvrement exact par rapport à la valeur de la taille du block m tandis que le panel F2 montre le recouvrement exact par rapport au nombre de blocs k. Les résultats sont présentés pour notre algorithme (diamants, noir) et les algorithmes concurrents sont le HC avec le madogramme (cercle plein, gris) et le SKmeans (étoiles, gris).

Conclusions

Nous avons concentré nos efforts sur le développement et l'analyse d'un algorithme capable de récupérer les partitions dans les modèles AI-block et de comprendre comment la structure de dépendance extrême affecte la difficulté de retrouver les groupes dans ces modèles. Notre objectif est particulièrement ambitieux car nous cherchons à travailler dans un cadre de grande dimension avec des observations faiblement dépendantes, qui sont sous asymptotiques. Nous supposons une indépendance asymptotique par blocs, qui est l'hypothèse fondamentale de notre modèle. Cette hypothèse nous permet de développer et d'analyser théoriquement un algorithme capable de récupérer les groupes dans ces modèles. Cette procédure peut récupérer les groupes avec une grande probabilité en fixant un seuil logarithmique en la dimension d. Toutefois, relâcher certaines hypothèses afin de couvrir un plus grand nombre de scénarios rencontrés en pratique est d'un grand intérêt. Nous pourrions atténuer la Condition $\mathcal{A}$ en cherchant une méthode tirant parti de la géométrie de la mesure exponentielle dans les modèles AI-block en inférant les groupes à l'aide de méthodes issues de l'analyse topologique des données.

References

Bador, M., P. Naveau, E. Gilleland, M. Castellà, and T. Arivelo (2015). Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over europe. *Weather and climate extremes 9*, 17–24.

Bernard, E., P. Naveau, M. Vrac, and O. Mestre (2013). Clustering of maxima: Spatial

dependencies among heavy rainfall in France. *Journal of climate 26*(20), 7929–7937.

Boulin, A., E. Di Bernardino, T. Laloë, and G. Toulemonde (2023). High-dimensional variable clustering based on sub-asymptotic maxima of a weakly dependent random process. *arXiv preprint arXiv:2302.00934*.

Boulin, A., E. Di Bernardino, T. Laloë, and G. Toulemonde (2022). Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework. *Journal of Multivariate Analysis 192*, 105059.

Bunea, F., C. Giraud, X. Luo, M. Royer, and N. Verzelen (2020). Model assisted variable clustering: Minimax-optimal recovery and algorithms. *The Annals of Statistics 48*(1), 111 – 137.

Coles, S., J. Heffernan, and J. Tawn (1999). Dependence measures for extreme value analyses. *Extremes 2*(4), 339–365.

Cooley, D., P. Naveau, and P. Poncet (2006). Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*, pp. 373–390. Springer.

Naveau, P., A. Guillou, D. Cooley, and J. Diebolt (2009). Modelling pairwise dependence of maxima in space. *Biometrika 96*(1), 1–17.

Saunders, K., A. Stephenson, and D. Karoly (2021). A regionalisation approach for rainfall based on extremal dependence. *Extremes 24*(2), 215–240.