High-dimensional variable clustering based on sub-asymptotic maxima of a weakly dependent random 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] and Gwladys Toulemonde[2,3]

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

Abstract.
We propose a new class of models for variable clustering called Asymptotic Independent block (AI-block) models, which defines population-level clusters based on the independence of the maxima of a multivariate stationary mixing 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 an 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 implies that groups can be learned nonparametrically in which block maxima of a dependent process are only sub-asymptotic. To further illustrate the significance of our work, we applied our method to neuroscience and environmental real-datasets. These applications highlight the potential and versatility of the proposed approach.

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

MSC Codes: 60G70; 62H05; 62M99

1. Introduction

Motivation Multivariate extremes arise when two or more extreme events occur simultaneously. These events are of prime interest to assess natural hazard, stemming from heavy rainfall, wind storms and earthquakes since they are driven by joint extremes of several of meteorological variables. Results from multivariate extreme value theory show that the possible dependence structure of extremes satisfy certain constraints. Indeed, the dependence structure may be described in various equivalent ways (Beirlant et al. 2004): by the exponent measure (Balkema and Resnick 1977), by the Pickands dependence function (Pickands 1981), by the stable tail dependence function (Huang 1992), by the madogram (Naveau et al. 2009; Boulin et al. 2022), and by the extreme value copula (Gudendorf and Segers 2010).

While the modeling of univariate and low-dimensional extreme events has been well-studied, it remains a challenge to model multivariate extremes, particularly when multiple rare events may occur simultaneously. Recent research in this area has focused on connecting the study of multivariate extremes to modern statistical and machine learning techniques. This has involved the development of new methods for characterizing complex dependence structures between extreme observations, such as sparsity-based approaches (Goix, Sabourin, and Stéphan Clémençon 2016; Meyer and Wintenberger 2021, 2023), conditional independence and graphical models (Engelke and Hitz 2020; Gissibl and Klüppelberg 2018; Segers 2020), dimensionality reduction (Chautru 2015; Drees and Sabourin 2021), and clustering methods (Cooley and Thibaud 2019; Janßen and Wan 2020). Our work is aligned with this direction of research as we propose a clustering algorithm for learning the dependence structure of multivariate extremes and, withal, to bridge important ideas from modern statistics and machine learning to the framework of extreme-value theory. Our

approach is remotely related to extremal graphical models. The probabilistic framework of this paper can effectively be seen as a disconnected extremal graph where the connected components are mutually independent of each other (see Engelke, Ivanovs, and Strokorb 2022, Section 8).

It is possible to perform clustering on $\mathbf{X}_1, \dots, \mathbf{X}_n$, where n is the number of observations of a random vector $\mathbf{X} \in \mathbb{R}^d$, through two different approaches: by partitioning the set of row indices $\{1, \dots, n\}$ or by partitioning the set of column indices $\{1, \dots, d\}$. The first problem is known as the data clustering problem, while the second is called the variable clustering problem, which is the focus of this paper. In data clustering, observations are drawn from a mixture distribution, and clusters correspond to different realizations of the mixing distribution, which is a distribution over all of $\mathbb{R}^d$.

The problem of variable clustering (see, e.g., Bunea et al. 2020) involves grouping similar components of a random vector $\mathbf{X} = (X^{(1)}, \dots, X^{(d)})$ into clusters. The goal is to recover these clusters from observations $\mathbf{X}_1, \dots, \mathbf{X}_n$. Instead of clustering similar observations based on a dissimilarity measure, the focus is on defining cluster models that correspond to subsets of the components $X^{(j)}$ of $\mathbf{X} \in \mathbb{R}^d$. The goal is to cluster similar variables such that variables within the same cluster are more similar to each other than they are to variables in other clusters. Variable clustering is of particular interest in the study of weather extremes, with examples in the literature on regionalization (Bador et al. 2015; Bernard et al. 2013; Saunders, Stephenson, and Karoly 2021), where spatial phenomena are observed at a limited number of sites. A specific case of interest is clustering these sites according to their extremal dependencies. This can be done using techniques such as k-means or hierarchical clustering with a dissimilarity measure designed for extremes. However, the statistical properties of these procedures have not been extensively studied, and it is not currently known which probabilistic models on $\mathbf{X}$ can be estimated using these techniques. In this paper, we consider model-based clustering, where the population-level clusters are well-defined, offering interpretability and a benchmark to evaluate the performance of a specific clustering algorithm.

The assumption that data are realizations of independent and identically distributed (i.i.d.) random variables is a fundamental assumption in statistical theory and modeling. However, this assumption is often unrealistic for modern datasets or the study of time series. Developing methods and theory to handle departures from this assumption is an important area of research in statistics. One common approach is to assume that the data are drawn from a multivariate stationary and mixing random process, which implies that the dependence between observations weakens over the trajectory. This assumption is widely used in the study of non-i.i.d. processes.

Our contribution is twofold. First, we develop a probabilistic setting for Asymptotic Independent block (AI-block) models to address the problem of clustering extreme values of the target vector. These models are based on the assumption that clusters of components of a multivariate random process are independent relative to their extremes. This approach has the added benefit of being amenable to theoretical analysis, and we show that these models are identifiable (see Theorem 1). Second, we motivate and derive an algorithm specifically designed for these models (see Algorithm (ECO)). We analyze its performance in terms of exact cluster recovery for minimally separated clusters, using a cluster separation metric (see Theorem 2). The issue is investigated in the context of nonparametric estimation over block maxima of a multivariate stationary mixing random process, where the block length is a tuning parameter.

Notations All bold letters $\mathbf{x}$ correspond to vector in $\mathbb{R}^d$. Let $O = \{O_g\}_{g=1,\dots,G}$ be a partition of $\{1, \dots, d\}$ into G groups and let $s : \{1, \dots, d\} \to \{1, \dots, G\}$ be a variable index assignment function, thus $O_g = \{a \in \{1, \dots, d\} : s(a) = g\} = \{i_{g,1}, \dots, i_{g,d_g}\}$ with $d_1 + \cdots + d_G = d$. Using these

notations, the variable $X^{(i_g,\ell)}$ should be read as the ℓth element from the gth cluster. By considering $B \subseteq \{1,\dots,d\}$, we denote the $|B|$-subvector of $\mathbf{x}$ by $\mathbf{x}^{(B)} = (x^{(j)})_{j\in B}$. When $B = \{1,\dots,d\}$, we will write H instead of $H^{(\{1,\dots,d\})}$. We define by $\mathbf{X} \in \mathbb{R}^d$ a random vector with law H and $\mathbf{X}^{(O_g)}$ a random subvector of $\mathbf{X}$ with marginal distribution $H^{(O_g)}$ with domain $\mathbb{R}^{d_g}$. Classical inequalities of vectors such as $\mathbf{x} > 0$ should be understand componentwise. The notation δ_x corresponds to the Dirac measure at x. Let $\mathbf{X}^{(O_g)}$, $g \in \{1,\dots,G\}$ be extreme value random vectors with $\mathbf{X} = (\mathbf{X}^{(O_1)},\dots,\mathbf{X}^{(O_G)})$, we say that $\mathbf{X}^{(O_1)},\dots,\mathbf{X}^{(O_G)}$ are independent if and only if $H(\mathbf{x}) = \Pi_{g=1}^{G} H^{(O_g)}\left(\mathbf{x}^{(O_g)}\right), \mathbf{x} \in \mathbb{R}^d$.

Structure of the paper In Section 2, we provide background on extreme-value theory and describe the probabilistic framework of AI-block models. We show that these models are identifiable and provide a series of equivalent characterizations. In Section 3, we develop a new clustering algorithm for AI-block models and prove that it can recover the target partition with high probability under mixing conditions over the random process. We provide a process that satisfies our probabilistic and statistical assumptions in Section 4. To exemplify further motivation for our research, we applied our method to real-data from neuroscience and environmental sciences, as discussed in Section 5. We illustrate the finite sample performance of our approach on simulated datasets in Appendix A. Mathematical details and proofs of our main, auxiliary, and supplementary results are provided in Appendix B, Appendix C, Appendix D, and Appendix E of the supplementary material, respectively. The access to all the codes and data are provided with the GitHub repository at the following link: https://github.com/Aleboul/ai_block_model.

2. A model for variable clustering

2.1 Background setting

Consider $\mathbf{Z}_t = (Z_t^{(1)},\dots,Z_t^{(d)})$, where $t \in \mathbb{Z}$ be a strictly stationary multivariate random process identically distributed as $\mathbf{Z}$, a d-dimensional random vector. Let $\mathbf{M}_m = (M_m^{(1)},\dots,M_m^{(d)})$ be the vector of component-wise maxima, where $M_{m,j} = \max_{i=1,\dots,m} Z_i^{(j)}$. Consider a random vector $\mathbf{X} = (X^{(1)},\dots,X^{(d)})$ with distribution H. A normalizing function a on $\mathbb{R}$ is a non-decreasing, right continuous function that goes to $\pm\infty$ as $x \to \pm\infty$. In extreme value theory (see, for example, the monograph of Beirlant et al. 2004), a fundamental problem is to characterize the limit distribution H in the following limit:

$$\lim_{m\to\infty} \mathbb{P}\left\{\mathbf{M}_m \leq \mathbf{a}_m(\mathbf{x})\right\} = H(\mathbf{x}), \tag{1}$$

where $\mathbf{a}_m = (a_m^{(1)},\dots,a_m^{(d)})$ with $a_m^{(j)}, 1 \leq j \leq d$ are normalizing functions and H is a non-degenerate distribution. Typically, H is an extreme value distribution, and $\mathbf{X}$ is a max-stable random vector with generalized extreme value margins. In this case, we can write:

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

where Λ is a Radon measure on the cone $E = [0,\infty)^d \setminus \{\mathbf{0}\}$. When (1) holds with H an extreme value distribution, the vector $\mathbf{Z}$ is said to be in the max-domain of attraction of the random vector $\mathbf{X}$ with law H, denoted as $F \in \mathcal{D}(H)$. In our context of a dependent process $(\mathbf{Z}_t, t \in \mathbb{Z})$, the limit in (1) will in general be different from a multivariate extremal types distribution and further conditions over the regularity (or mixing conditions) are thus needed to obtain an extremal distribution. In particular, if the random process $(\mathbf{Z}_t, t \in \mathbb{Z})$ is β-mixing, then a Fisher-Tippett-Gnedenko's type theorem holds for multivariate stationary random processes (see Hsing 1989, Theorem 4.2).

The max-domain of attraction can be translated into terms of copulae. Denote by C_m the unique copula associated with $\mathbf{M}_m$. Throughout, we will work under the following fundamental domain of attraction condition.

Condition $\mathcal{A}$. There exists a copula C_∞ such that

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

Typically, the limit C_∞ is an extreme value copula, that is, the copula C_∞ is max-stable $C_\infty(\mathbf{u}^{1/s})^s = C_\infty(\mathbf{u})$, for all $s > 0$ and it can be expressed as follows for $\mathbf{u} \in [0,1]^d$:

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

where $L : [0,\infty]^d \to [0,\infty]$ is the associated stable tail dependence function (see Gudendorf and Segers 2010 for an overview of extreme value copulae). However, C_∞ is in general different from the extreme value copula, denoted C_∞^{iid}, obtained when the process $(\mathbf{Z}_t, t \in \mathbb{Z})$ is serially independent (see, e.g., Bücher and Segers 2014, Section 4.1).

As L is an homogeneous function of order 1, i.e., $L(a\mathbf{z}) = aL(\mathbf{z})$ for all $a > 0$, we have, for all $\mathbf{z} \in [0,\infty)^d$,

$$L(\mathbf{z}) = (z^{(1)} + \cdots + z^{(d)})A(\mathbf{t}),$$

with $t^{(j)} = z^{(j)}/(z^{(1)} + \cdots + z^{(d)})$ for $j \in \{2, \ldots, d\}$, $t^{(1)} = 1 - (t^{(2)} + \cdots + t^{(d)})$, and A is the restriction of L into the d-dimensional unit simplex, viz.

$$\Delta_{d-1} = \{(v^{(1)}, \ldots, v^{(d)}) \in [0,1]^d : v^{(1)} + \cdots + v^{(d)} = 1\}.$$

The function A is known as the Pickands dependence function and is often used to quantify the extremal dependence among the elements of $\mathbf{X}$. Indeed, A satisfies the constraints $1/d \leq \max(t^{(1)}, \ldots, t^{(d)}) \leq A(\mathbf{t}) \leq 1$ for all $\mathbf{t} \in \Delta_{d-1}$, with lower and upper bounds corresponding to the complete dependence and independence among maxima. For the latter, it is commonly said that the stationary random process $(\mathbf{Z}_t, t \in \mathbb{Z})$ exhibits asymptotic independence, i.e., the multivariate extreme value distribution H in the max-domain of attraction is equal to the product of its marginal extreme value distributions.

2.2 Proposed AI-block models

In this paper, our main focus lies on the concept of asymptotic independence, which has been observed in various applications. Building upon these applications, we introduce a novel class of models called AI-block models for variable clustering. These models define population-level clusters as groups of variables that exhibit dependence within clusters but extremes are independent from variables in other clusters. Formally, these variables can be partitioned into an unknown number, denoted as G, of clusters represented by $O = \{O_1, \ldots, O_G\}$. Within each cluster, the variables display dependence, while the clusters themselves are asymptotically independent. In this section, our primary focus is on the identifiability of the model, specifically addressing the existence of a unique maximal element according to a specific partial order on the partition. We provide an explicit construction of this maximal element, which represents the thinnest partition where the desired property holds. This maximal element serves as a target for statistical inference within our framework.

Let us consider $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ to be extreme value random vectors with extreme value copulae $C_\infty^{(O_1)}, \ldots, C_\infty^{(O_G)}$ respectively. Under condition of independence between $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$, the random vector $\mathbf{X} = (\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)})$ is again extreme and one can detail the expression of its extreme value copula. The formal statement of this result is stated in the next proposition.

Proposition 1. *Let $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$ be independent extreme value random vectors with extreme value copulae $C_\infty^{(O_1)}, \dots, C_\infty^{(O_G)}$. Then the function C_∞ defined as*

$$\begin{aligned} C_\infty : \ [0,1]^d &\longrightarrow [0,1] \\ \mathbf{u} &\longmapsto \Pi_{g=1}^G C_\infty^{(O_g)}(u^{(i_{g,1})}, \dots, u^{(i_{g,d_g})}), \end{aligned}$$

is an extreme value copula associated to the random vector $\mathbf{X} = (\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)})$.

As a result, a random vector $\mathbf{X}$ that exhibits (asymptotic) independence between extreme-valued subvectors therefore inherits this extreme-valued property. Using the definitions and notations so far introduced in this work, we now present the definition of our model.

Definition 1 (Asymptotic Independent-block model)**.** Let $(\mathbf{Z}_t, t \in \mathbb{Z})$ be a d-variate stationary random process with law F and $\mathbf{X}$ a random vector with extreme value distribution H. The random process $\mathbf{Z}_t$ is said to follow an AI-block model if $F \in D(H)$ and for every $g \in \{1, \dots, G\}$, $\mathbf{X}^{(O_g)} = (X^{(i_{g,1})}, \dots, X^{(i_{g,d_g})})$ are extreme value random vectors and $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$ are independent, that is $H = \Pi_{g=1}^G H^{(O_g)}$.

Notice that, when $G = 1$, the definition of AI-block models thus reduces to the process $(\mathbf{Z}_t, t \in \mathbb{Z})$ is in the domain of attraction of an extreme value distribution H.

Following Bunea et al. 2020, we introduce the following notation in our framework. We say that $\mathbf{Z}$ follows an AI-block model with a partition O, denoted $\mathbf{Z} \sim O$. We define the set $\mathcal{L}(\mathbf{Z}) = \{O : O \text{ is a partition of } \{1, \dots, d\} \text{ and } \mathbf{Z} \sim O\}$, which is nonempty and finite, and therefore has maximal elements. We introduce a partial order on partitions as follows: let $O = \{O_g\}_g$ and $\{S_{g'}\}_{g'}$ be two partitions of $\{1, \dots, d\}$. We say that S is a sub-partition of O if, for each g', there exists g such that $S_{g'} \subseteq O_g$. We define the partial order $\leq$ between two partitions O and S of $\{1, \dots, d\}$ as follows:

$$O \leq S, \text{ if } S \text{ is a sub-partition of } O. \tag{2}$$

For any partition $O = \{O_g\}_{1 \leq g \leq G}$, we write $a \overset{O}{\sim} b$ where $a, b \in \{1, \dots, d\}$ if there exists $g \in \{1, \dots, G\}$ such that $a, b \in O_g$.

Definition 2. For any two partitions O, S of $\{1, \dots, d\}$, we define $O \cap S$ as the partition induced by the equivalence relation $a \overset{O \cap S}{\sim} b$ if and only if $a \overset{O}{\sim} b$ and $a \overset{S}{\sim} b$.

Checking that $a \overset{O \cap S}{\sim} b$ is an equivalence relation is straightforward. With this definition, we have the following interesting properties that lead to the desired result, the identifiability of AI-block models.

Theorem 1. *Let $(\mathbf{Z}_t, t \in \mathbb{Z})$ be a stationary random process. The following properties hold:*

(i) Consider $O \leq S$. Then $\mathbf{Z} \sim S$ implies $\mathbf{Z} \sim O$,
(ii) $O \leq O \cap S$ and $S \leq O \cap S$,
(iii) $\mathbf{Z} \sim O$ and $\mathbf{Z} \sim S$ is equivalent to $\mathbf{Z} \sim O \cap S$,
(iv) The set $\mathcal{L}(\mathbf{Z})$ has a unique maximum $\bar{O}(\mathbf{Z})$, with respect to the partition partial order $\leq$ in (2).

The proof demonstrates that for any partition such that $\mathbf{Z}$ follows an AI-block model, there exists a maximal partition, denoted by $\bar{O}(\mathbf{Z})$, and its structure is intrinsic of the definition of the extreme value random vector $\mathbf{X}$. This partition, which represents the thinnest partition where $\mathbf{Z}$ asymptotically independent per block, matches our expectations for a reasonable clustering target in these models. With a slight abuse of notation, we will refer to $\bar{O}(\mathbf{Z})$ as $\bar{O}$ throughout the rest of this paper.

2.3 Extremal dependence structure for AI-block models

In extreme value theory, independence between the components $X^{(1)}, \dots, X^{(d)}$ of an extreme-value random vector $\mathbf{X} \in \mathbb{R}^d$ can be characterized in a useful way: according to Takahashi 1994, Theorem 2.2, total independence of $\mathbf{X}$ is equivalent to the existence of a vector $\mathbf{p} = (p^{(1)}, \dots, p^{(d)}) \in \mathbb{R}^d$ such that $H(\mathbf{p}) = H^{(1)}(p^{(1)}) \dots H^{(d)}(p^{(d)})$. This characterization were extended for the independence of a multivariate extreme value distribution to its multivariate marginals from Ferreira 2011, Proposition 2.1, i.e., it holds that $H(\mathbf{x}) = \Pi_{g=1}^{G} H^{(O_g)}(\mathbf{x}^{(O_g)})$ for every $\mathbf{x} \in \mathbb{R}^d$ if and only if there exists $\mathbf{p} \in \mathbb{R}^d$ such that $0 < H^{(O_g)}(\mathbf{p}^{(O_g)}) < 1$ for every $g \in \{1, \dots, G\}$ and $H(\mathbf{p}) = \Pi_{g=1}^{G} H^{(O_g)}(\mathbf{p}^{(O_g)})$. Another proof of this result involving the exponent measure is proposed in Appendix D.1. One direct application of this result in AI-block models is that $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$ are independent if and only if :

$$A\left(\frac{1}{d}, \dots, \frac{1}{d}\right) = \sum_{g=1}^{G} \frac{d_g}{d} A^{(O_g)}\left(\frac{1}{d_g}, \dots, \frac{1}{d_g}\right).$$

Definition 3 (Sum of Extremal COefficients (SECO))**.** The extremal coefficient of an extreme value random vector $\mathbf{X}$ is defined as (see Smith 1990):

$$\theta := \theta^{(\{1,\dots,d\})} = d\, A(d^{-1}, \dots, d^{-1}), \tag{3}$$

where A is the Pickands dependence function. For a partition $O = \{O_1, \dots, O_G\}$ of $\{1, \dots, d\}$, we define $\theta^{(O_g)} = d_g A^{(O_g)}(d_g^{-1}, \dots, d_g^{-1})$, as the extremal coefficient of the subvectors $\mathbf{X}^{(O_g)}$ where $d_g = |O_g|$ is the size of the set O_g and $A^{(O_g)}$ is the Pickands dependence function of $\mathbf{X}^{(O_g)}$. Using these coefficients, we define the following quantity SECO as

$$\mathrm{SECO}(O) = \sum_{g=1}^{G} \theta^{(O_g)} - \theta. \tag{4}$$

The extremal coefficient satisfies $1 \leq \theta \leq d$ where the lower and upper bounds correspond to the complete dependence and independence among maxima, respectively. The Sum of Extremal Coefficient (SECO) serves as a quantitative measure that assesses how much the sum of extremal coefficients for subvectors $\mathbf{X}^{(O_g)}$ deviates from the extremal coefficient of the full vector $\mathbf{X}$. When the SECO equals 0, it signifies that the subvectors $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$ form an independent partition. In other words, these subvectors exhibit asymptotic independence, irrespective of any underlying distributional assumptions. Therefore, the SECO, as defined in Equation (4), is a valuable tool for capturing the asymptotic independent block structure of the random vector $\mathbf{X}$, and it offers the dual advantages of computational feasibility and being free from parametric assumptions, as discussed in Section 3.4.

Additionally, we establish a condition based on the extremal dependence of each cluster, which allows us to introduce a straightforward yet robust algorithm. This algorithm facilitates the comparison of pairwise extreme dependence between vector components, enabling us to draw informed conclusions about the dependence structures using only pairwise comparisons. It provides a practical means of assessing and quantifying the relationships among the various components of the vector, aiding in the analysis of complex high-dimensional data.

Condition $\mathcal{B}$. For every $g \in \{1, \dots, G\}$, the extreme value random vector $\mathbf{X}^{(\bar{O}_g)}$, where $\bar{O}_g$ is the maximal element of $\mathcal{L}(\mathbf{Z})$, exhibits dependence between all components i.e.,

$$a \overset{\bar{O}}{\sim} b \implies \chi(a, s) > 0,\ \chi(b, s) > 0, \text{ where } s \in \{1, \dots, d\} \text{ such that } a \overset{\bar{O}}{\sim} s \text{ and } b \overset{\bar{O}}{\sim} s.$$

One sufficient condition to satisfy Condition $\mathcal{B}$ is to suppose that exponent measures of the extreme value random vectors $\mathbf{X}^{(\bar{O}_g)}$ have nonnegative Lebesgue densities on the nonnegative orthant $[0,\infty)^{d_g} \setminus \{\mathbf{0}^{(\bar{O}_g)}\}$, for every $g \in \{1,\dots,G\}$. This condition implies that the components within a cluster are together extremes. Various classes of tractable extreme value distributions satisfy Condition $\mathcal{B}$. These popular models, commonly used for statistical inference, include the asymmetric logistic model (Tawn 1990), the asymmetric Dirichlet model (Coles and Tawn 1991), the pairwise Beta model (Cooley, Davis, and Naveau 2010) or the Hüsler Reiss model (Hüsler and Reiss 1989).

3. Consistent estimation of minimaly separated clusters

3.1 Multivariate tail coefficient

Throughout this section, assume that we observe copies $\mathbf{Z}_1 \dots, \mathbf{Z}_n$ of the d-dimensional stationary random process $(\mathbf{Z}_t, t \in \mathbb{Z})$ an AI-block model as in Definition 1. The sample of size n of $(\mathbf{Z}_t, t \in \mathbb{Z})$ is divided into k blocks of length m, so that $k = \lfloor n/m \rfloor$, the integer part of n/m and there may be a remaining block of length $n - km$. For the i-th block, the maximum value in the j-th component is denoted by

$$M_{m,i}^{(j)} = \max\left\{ Z_t^{(j)} \,:\, t \in (im-m, im] \cap \mathbb{Z} \right\}.$$

Let us denote by $\mathbf{M}_{m,i} = (M_{m,i}^{(1)}, \dots, M_{m,i}^{(d)})$ the vector of the componentwise maxima in the i-th block. For a fixed block length m, the sequence of block maxima $(\mathbf{M}_{m,i})_i$ forms a stationary process that exhibits the same regularity of the process $(\mathbf{Z}_t, t \in \mathbb{Z})$. The distribution functions of block maxima are denoted by

$$F_m(\mathbf{x}) = \mathbb{P}\left\{\mathbf{M}_{m,1} \le \mathbf{x}\right\}, \quad F_m^{(j)}(X^{(j)}) = \mathbb{P}\left\{M_{m,1}^{(j)} \le X^{(j)}\right\},$$

with $\mathbf{x} \in \mathbb{R}^d$ and $j \in \{1,\dots,d\}$. Denote by $U_{m,1}^{(j)} = F_m^{(j)}(M_{m,1}^{(j)})$ the unknown uniform margin of $M_{m,1}^{(j)}$ with $j \in \{1,\dots,d\}$. Let C_m be the unique (as the margins of $\mathbf{M}_{m,1}$ are continuous) copula of F_m. Then, from Condition $\mathcal{A}$, C_m is in the domain-of-attraction of a copula C_∞. By Hsing 1989, Theorem 4.2, C_∞ is an extreme value copula if the time series $(\mathbf{Z}_t, t \in \mathbb{Z})$ is β-mixing.

One way to measure tail dependence for a d-dimensional extreme value random vector is through the use of the extremal coefficient, as defined in Equation (3). According to Schlather and Tawn 2002, the coefficient θ can be interpreted as the number of independent variables that are involved in the given random vector. Let $x \in \mathbb{R}$ and $\theta_m(x)$ be the extremal coefficient for the vector of maxima $\mathbf{M}_{m,1}$, which is defined by the following relation:

$$\mathbb{P}\left\{ \bigvee_{j=1}^{d} U_{m,1}^{(j)} \le x \right\} = \mathbb{P}\{U_{m,1}^{(1)} \le x\}^{\theta_m(x)}.$$

Under Condition $\mathcal{A}$, the coefficient $\theta_m(x)$ of the componentwise maxima $\mathbf{M}_{m,1}$ converges to the extremal coefficient θ of the random vector $\mathbf{X}$, that is:

$$\theta_m(x) \underset{m\to\infty}{\longrightarrow} \theta, \quad \forall x \in \mathbb{R}.$$

It is worth noting that θ is a constant since $\mathbf{X}$ is a multivariate extreme value distribution. To generalize the bivariate madogram for the random vectors $\mathbf{M}_{m,1}$ we follow the same approach as in Marcon et al. 2017; Boulin et al. 2022 and define:

$$\nu_m = \mathbb{E}\left[\bigvee_{j=1}^{d} U_{m,1}^{(j)} - \frac{1}{d}\sum_{j=1}^{d} U_{m,1}^{(j)} \right], \quad \nu = \mathbb{E}\left[\bigvee_{j=1}^{d} H^{(j)}(X^{(j)}) - \frac{1}{d}\sum_{j=1}^{d} H^{(j)}(X^{(j)}) \right]. \tag{5}$$

Condition $\mathcal{A}$ implies that the distribution of $\mathbf{M}_{m,1}$ is sub-asymptotically extreme valued. A common approach for estimating the extremal coefficient in this scenario consists of supposing that the sample follows exactly the extreme value distribution and to consider $\theta_m(x) := \theta_m$ a sub-asymptotic extremal coefficient which is constant for every x. Thus, we have

$$\theta_m = \frac{1/2 + \nu_m}{1/2 - \nu_m}, \quad 1 \leq \theta_m \leq d.$$

One issue with the sub-asymptotic extremal coefficient is that it is misspecified, as extreme value distributions only arise in the limit as the block size m tends to infinity, while in practice we must use a finite sample size. We study this misspecification error in Section 3.3. A plug-in estimation process can be obtained using:

$$\hat{\theta}_{n,m} = \frac{1/2 + \hat{\nu}_{n,m}}{1/2 - \hat{\nu}_{n,m}}, \tag{6}$$

where $\hat{\nu}_{n,m}$ is an estimate of ν_m obtained using:

$$\hat{\nu}_{n,m} = \frac{1}{k}\sum_{i=1}^{k}\left[\bigvee_{j=1}^{d}\hat{U}^{(j)}_{n,m,i} - \frac{1}{d}\sum_{j=1}^{d}\hat{U}^{(j)}_{n,m,i}\right], \tag{7}$$

and $(\hat{U}^{(j)}_{n,m,1}, \ldots, \hat{U}^{(j)}_{n,m,k})$ are the empirical counterparts of $(U^{(j)}_{m,1}, \ldots, U^{(j)}_{m,k})$ or, equivalently, scaled ranks of the sample. In the following, we provide non-asymptotic bounds for the error $|\hat{\nu}_{n,m} - \nu_m|$.

Proposition 2. *Let $(\mathbf{Z}_t, t \in \mathbb{Z})$ be a stationary process with algebraic φ-mixing distribution, $\varphi(n) \leq \lambda n^{-\zeta}$ where $\lambda > 0$, and $\zeta > 1$. Then the following concentration bound holds*

$$\mathbb{P}\left\{|\hat{\nu}_{n,m} - \nu_m| \geq C_1 k^{-1/2} + C_2 k^{-1} + t\right\} \leq (d + 2\sqrt{e})\exp\left\{-\frac{t^2 k}{C_3}\right\},$$

where k is the number of block maxima and C_1, C_2 and C_3 are constants depending only on ζ and λ.

The non-asymptotic analysis in Proposition 2 is stringent and requires the use of φ-mixing in order to apply Hoeffding and McDiarmid inequalities in a dependent setting.

3.2 Inference in AI-block models

In this section, we present an adapted version of the algorithm developed in Bunea et al. 2020 for clustering variables based on a metric on their covariances, named as CORD. Our adaptation involves the use of the extremal correlation as a measure of dependence between the extremes of two variables.

The SECO in (4) can be written in the bivariate setting as

$$\mathrm{SECO}(\{a, b\}) = 2 - \theta(a, b),$$

where for notational convenience, $\theta(a, b) := \theta^{(\{a,b\})}$ is the bivariate extremal coefficient between $X^{(a)}$ and $X^{(b)}$ as defined in (3). This metric has a range between 0 and 1, with the boundary cases representing asymptotic independence and comonotonic extremal dependence, respectively. In fact, the bivariate SECO is exactly equal to the extremal correlation χ defined in Coles, Heffernan, and Tawn 1999 as

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

whenever the limit exists. In particular, if $\mathbf{X}$ is a multivariate extreme-value distribution, then $\chi(a,b) = \chi_q(a,b)$ for $q \in (0,1)$. In an AI-block model, the statement

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

is equivalent to

$$\chi(a,b) = \chi(b,a) = 0, \quad \forall a \in O_g, \forall\, b \in O_h, \quad g \neq h. \tag{8}$$

Thus using Condition $\mathcal{B}$ and Equation (8), the extremal correlation is a sufficient statistic to recover clusters in an AI-block model. Condition $\mathcal{B}$ implies a particular relationship: $a \overset{\bar{O}}{\sim} b \implies \chi(a,b) > 0$. Furthermore, Equation (8) reveals:

$$a \overset{\bar{O}}{\not\sim} b \implies \chi(a,b) = 0.$$

Consequently, in an AI-block model, two variables $X^{(a)}$ and $X^{(b)}$ are considered part of the same cluster under Condition $\mathcal{B}$ if and only if $\chi(a,b) > 0$. For the estimation procedure, using tools introduced in the previous section, we give a sample version of the extremal correlation associated to $M_{m,1}^{(a)}$ and $M_{m,1}^{(b)}$ by

$$\hat{\chi}_{n,m}(a,b) = 2 - \hat{\theta}_{n,m}(a,b), \quad a,b \in \{1,\dots,d\},$$

where $\hat{\theta}_{n,m}(a,b)$ is the sampling version defined in (6) of $\theta(a,b)$. With some technical arguments, a concentration result estimate follows directly from Proposition 2.

We can represent the matrix of all extremal correlations as $\mathcal{X} = [\chi(a,b)]_{a=1,\dots,d,b=1,\dots,d}$. Additionally, we introduce its empirical counterpart, denoted as $\hat{\mathcal{X}}$. This version, $\hat{\mathcal{X}}$ incorporates elements $\hat{\chi}_{n,m}(a,b)$ for pairs $(a,b) \in \{1,\dots,d\}^2$. We present an algorithm, named ECO (Extremal COrrelation), which estimates the partition $\bar{O}$ using a dissimilarity metric based on the extremal correlation. This algorithm, outlined in Algorithm (ECO), does not require the specification of the number of groups G, as it is automatically estimated by the procedure. The algorithm complexity for computing the k vectors $\hat{\mathbf{U}}_{n,m,i} = (\hat{U}_{n,m,i}^{(1)}, \dots, \hat{U}_{n,m,i}^{(d)})$ for $i \in \{1,\dots,k\}$ is of order $O(dk\ln(k))$. Given the empirical ranks, computing $\hat{\mathcal{X}}$ and performing the algorithm require $O(d^2 \vee dk\ln(k))$ and $O(d^3)$ computations, respectively. So the overall complexity of the estimation procedure is $O(d^2(d \vee k\ln(k)))$.

In Appendix D.2, we provide conditions under the regularity of the process ensuring that our algorithm is asymptotically consistent. These conditions involve β-mixing coefficients which are less stringent than φ-mixing used in the next section. Unlike in asymptotic analysis where the choice of the threshold becomes trivial, in a non-asymptotic framework, the algorithm's performance is influenced by the parameter τ. In a non-asymptotic framework, when $\tau \approx 0$, the algorithm is prone to identifying the sole cluster as $\{1,\dots,d\}$, while a value of $\tau \approx 1$ suggests that the algorithm is likely to return the largest partition $\{\{1\},\dots,\{d\}\}$. Thus, the parameter τ serves as a threshold that determines the algorithm's tolerance to differentiate between the noise in the inference and the signal indicating asymptotic dependence. This discriminatory capability depends on factors such as the sample size n, the dimension d, and the proximity between the sub-asymptotic framework and the maximum domain of attraction. Consequently, selecting an appropriate threshold τ becomes a critical consideration. However, this challenge can be addressed through a non-asymptotic analysis of the algorithm, which we will discuss in the following section.

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 $\hat{O} = (\hat{O}_l)_l$

3.3 Estimation in growing dimensions

We provide consistency results for our algorithm, allowing estimation in the case of growing dimensions, by adding non asymptotic bounds on the probability of consistently estimating the maximal element $\bar{O}$ of an AI-block model. Furthermore, this result provides an answer for how to leverage τ in Algorithm (ECO). The difficulty of clustering in AI-block models can be assessed via the size of the Minimal Extremal COrrelation (MECO) separation between two variables in a same cluster:

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

In AI-block models, with Condition $\mathcal{B}$, we always have $\mathrm{MECO}(\mathcal{X}) > \eta$ with $\eta = 0$. However, a large value of η will be needed for retrieving consistently the partition $\bar{O}$ stationary observations. We are now ready to state the main result of this section.

Theorem 2. *We consider the AI-block model as defined in Definition 1 under Condition $\mathcal{B}$, and $(\mathbf{Z}_t, t \in \mathbb{Z})$ be a d-multivariate stationary process with algebraic φ-mixing distribution, $\varphi(n) \leq \lambda n^{-\zeta}$ where $\lambda > 0$ and $\zeta > 1$. Define*

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

Let (τ, η) be parameters fulfilling

$$\tau \geq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{(1+\gamma)\ln(d)}{k}},$$
$$\eta \geq d_m + C_1 k^{-1/2} + C_2 k^{-1} + C_3 \sqrt{\frac{(1+\gamma)\ln(d)}{k}} + \tau,$$

where C_1, C_2, C_3 are universal constants depending only on λ and ζ, k is the number of block maxima, and $\gamma > 0$. For a given $\mathcal{X}$ and its corresponding estimator $\hat{\mathcal{X}}$, if $\mathrm{MECO}(\mathcal{X}) > \eta$, then the output of Algorithm (ECO) is consistent, i.e.,

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

Unsurprising, as Theorem 2 is not concerned with asymptotics, we did not actually assume Condition $\mathcal{A}$. A link between $\mathbf{Z}$ and $\mathbf{X}$ is implicitly provided through the bias term d_m which measures the distance between $\chi_m(a,b)$ and $\chi(a,b)$. This quantity vanishes when Condition $\mathcal{A}$ holds as $m \to \infty$.

Some comments on the implications of Theorem 2 are in order. On a high level, larger dimension d and bias d_m lead to a higher threshold τ. The effects of the dimension d and the bias d_m are intuitive: larger dimension or more bias make the partition recovery problem more difficult. It is clear that the partition recovery problem becomes more difficult as the dimension or bias increases. This is reflected in the bound of the MECO value below which distinguish between noise and asymptotic independence is impossible by our algorithm. Thus, whereas the dimension d increases, the dependence between each component should be stronger in order to distinguish between the two. In other words, for alternatives that are sufficiently separated from the asymptotic independence case, the algorithm will be able to distinguish between asymptotic independence and noise at the $\sqrt{\ln(d)k^{-1}}$ scale. For a more quantitative discussion, our algorithm is able to recover clusters when the data dimension scales at a polynomial rate, i.e., $d = o(n^p)$, with $p > 0$ as η in Theorem 2 decreases with increasing n.

The order of the threshold τ involves known quantity such as d and k and a unknown parameter d_m. For the latter, there is no simple manner to choose optimally this parameter, as there is no simple way to determine how fast is the convergence to the asymptotic extreme behavior, or how far into the tail the asymptotic block dependence structure appears. In particular, Condition $\mathcal{A}$ does not contain any information about the rate of convergence of C_m to C_∞. More precise statements about this rate can be made with second order conditions. Let a regularly varying function $\Psi : \mathbb{N} \to (0,\infty)$ with coefficient of regular variation $\rho_\Psi < 0$ and a continuous non-zero function S on $[0,1]^d$ such that

$$C_m(\mathbf{u}) - C_\infty(\mathbf{u}) = \Psi(m)S(\mathbf{u}) + o(\Psi(m)), \quad \text{for } m \to \infty, \tag{9}$$

uniformly in $\mathbf{u} \in [0,1]^d$ (see, e.g., Bücher, Volgushev, and Zou 2019; Zou, Volgushev, and Bücher 2021 for a proper introduction to this condition). In this case, we can show that $d_m = O(\Psi(m))$. In the typical case $\Psi(m) = c\, t^{\rho_\Psi}$ with $c > 0$, choosing m proportional to $n^{1/(1-\rho_\Psi)}$ leads to the optimal convergence rate $n^{\rho_\Psi/(1-2\rho_\Psi)}$ (see Drees and Huang 1998). However, there is no simple way to know in advance or infer the value of ρ_Ψ and, in practice, it is advisable to use a data-driven procedure to select the threshold.

3.4 Data-driven selection of the threshold parameter

The performance of Algorithm (ECO) depends crucially on the value of the threshold parameter τ. In practice, it is advisable to use a data-driven procedure to select the threshold in Algorithm (ECO). The idea is to use the SECO criteria presented in Equation (4). Let $\mathbf{Z} \sim O$, given a partition $\hat{O} = \{\hat{O}_g\}_g$, we know from Ferreira 2011 that the SECO similarity given by

$$\mathrm{SECO}(\hat{O}) = \sum_g \theta^{(\hat{O}_g)} - \theta \tag{10}$$

is equal to 0 if and only if $\hat{O} \leq \bar{O}$. We thus construct a loss function given by the SECO where we evaluate its value over a grid of the τ values. The value of τ for which the SECO similarity has minimum values is also the value of τ for which we have consistent recovery of our clusters. The based estimator of the SECO in (10) is thus defined as

$$\widehat{\mathrm{SECO}}_{n,m}(\hat{O}) = \sum_g \hat{\theta}_{n,m}^{(\hat{O}_g)} - \hat{\theta}_{n,m}. \tag{11}$$

Let $\widehat{\mathcal{O}}$ be a collection of partitions computed with Algorithm (ECO), by varying τ around its theoretical optimal value, of order $(d_m + \sqrt{\ln(d)k^{-1}})$, on a fine grid. For any $\hat{O} \in \widehat{\mathcal{O}}$, we evaluate our SECO in (11) and keep the greatest threshold that minimizes this criteria. Proposition 3 offers theoretical support for this procedure.

Proposition 3. *We consider an AI-block model as in Definition 1, the partial order $\leq$ between two partitions in (2). Let $(\boldsymbol{Z}_t, t \in \mathbb{Z})$ be a d-multivariate stationary process with algebraic φ-mixing distribution, i.e, $\varphi(n) \leq \lambda n^{-\zeta}$ where $\lambda \geq 0$ and $\zeta \geq 1$. Let $\bar{O} = \{\bar{O}_1, \dots, \bar{O}_G\}$ be the thinnest partition given by Theorem 1 with corresponding sizes $d_1, \dots, d_G$. Let $\hat{O} = \{\hat{O}_1, \dots, \hat{O}_I\}$ be any partition of $\{1, \dots, d\}$ with corresponding sizes $d_1, \dots, d_I$. Define*

$$D_m = \max\left\{ \left| \sum_{g=1}^{G} \theta_m^{(\bar{O}_g)} - \sum_{g=1}^{G} \theta^{(\bar{O}_g)} \right|, \left| \sum_{i=1}^{I} \theta_m^{(\hat{O}_i)} - \sum_{i=1}^{I} \theta^{(\hat{O}_i)} \right| \right\},$$

Then, there exists a constant $c > 0$, such that, if $\hat{O} \not\leq \bar{O}$ and

$$\mathrm{SECO}(\hat{O}) > 2\left(D_m + c\sqrt{\frac{\ln(d)}{k}} \max(G, I) \max(\vee_{g=1}^{G} d_g^2, \vee_{i=1}^{I} d_i^2) \right), \tag{12}$$

it holds that

$$\mathbb{E}[\widehat{\mathrm{SECO}}_{n,m}(\bar{O})] < \mathbb{E}[\widehat{\mathrm{SECO}}_{n,m}(\hat{O})].$$

However, the bound presented in Equation (12) is overly pessimistic since it exhibits a polynomial growth with respect to cluster sizes. Nevertheless, when we consider the scenario where $n \to \infty$ with d being fixed and Condition $\mathcal{B}$, this condition simplifies to:

$$\mathrm{SECO}(\hat{O}) > 0,$$

which holds true for every $\hat{O} \not\leq \bar{O}$ (see Section 7 in the supplementary materials). Therefore, despite the pessimistic nature of this bound, the asymptotic relevance of choosing the threshold parameter based on data-driven approaches remains intact. Additionally, numerical studies provide support for the effectiveness of SECO as an appropriate criterion for determining the threshold parameter for a suitable number of data and for important cluster sizes (see Appendix A). Furthermore, we establish the weak convergence of an estimator for $\mathrm{SECO}(O)$ when $\mathbf{Z} \sim O$ (we refer to Appendix E.3 for detailed information).

4. Hypotheses discussion for a multivariate random persistent process

A trivial example of an AI-block model is given by a partition O such that $\mathbf{Z}^{(O_g)}$ is in domain of attraction of an extreme value random vector $H^{(O_g)}$, $g \in \{1, \dots, G\}$ such that $\mathbf{Z}^{(O_1)}, \dots, \mathbf{Z}^{(O_G)}$ are independent. In this simple model, block independent clusters are sub-asymptotic hence asymptotic and the peculiar dependence structure under study is not inherent of the tail behaviour of the random vector.

More interestingly, in this section we will focus on a process where the dependence between clusters disappears in the distribution tails. To this aim, we recall here a φ-algebraically mixing process. The interested reader is referred for instance to Bücher and Segers 2014. We show that Conditions $\mathcal{A}$ and $\mathcal{B}$ hold with a bit more work.

Let D denote a copula and consider i.i.d d-dimensional random vectors $\boldsymbol{Z}_0, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots$ from D and independent Bernoulli random variables $I_1, I_2, \dots$ i.i.d. with $\mathbb{P}\{I_t = 1\} = p \in (0, 1]$. For $t = 1, 2, \dots$, define the stationary random process $(\mathbf{Z}_t, t \in \mathbb{Z})$ by

$$\mathbf{Z}_t = \boldsymbol{\xi}_t \delta_1(I_t) + \boldsymbol{Z}_{t-1} \delta_0(I_t), \tag{13}$$

where we suppose without loss of generality that the process is defined for all $t \in \mathbb{Z}$ using stationarity. The persistence of the process $(\mathbf{Z}_t, t \in \mathbb{Z})$ arises from repeatable values in (13). From this persistence, $(\mathbf{Z}_t, t \in \mathbb{Z})$ is φ-mixing with coefficient of order $O((1-p)^n)$ Bücher and Segers 2014, Lemma B.1, hence algebraically mixing.

Assuming that the copula D belongs to the (i.i.d.) copula domain of attraction of an extreme value copula $D_\infty^{(iid)}$, denoted as

$$D_m(\mathbf{u}) = \{D(\mathbf{u}^{1/m})\}^m \longrightarrow D_\infty^{(iid)}(\mathbf{u}), \quad (m \to \infty).$$

Here, D_m represents the copula of the componentwise block maximum of size m based on the serially independent sequence $(\boldsymbol{\xi}_t, t \in \mathbb{N})$.

According to Bücher and Segers 2014, Proposition 4.1, if C_m denotes the copula of the componentwise block maximum of size m based on the sequence $(\mathbf{Z}_t, t \in \mathbb{N})$, then

$$C_m(\mathbf{u}) \underset{m\to\infty}{\longrightarrow} D_\infty^{(iid)}(\mathbf{u}), \quad \mathbf{u} \in [0,1]^d.$$

This implies that Condition $\mathcal{A}$ is satisfied.

Consider the multivariate outer power transform of a Clayton copula with parameters $\theta > 0$ and $\beta \geq 1$, defined as:

$$D(\mathbf{u}; \theta, \beta) = \left[1 + \left\{\sum_{j=1}^{d} (\{u^{(j)}\}^{-\theta} - 1)^\beta\right\}^{1/\beta}\right]^{-1/\theta}, \quad \mathbf{u} \in [0,1]^d.$$

The copula of multivariate componentwise maxima of an i.i.d. sample of size m from a continuous distribution with copula $D(\cdot; \theta, \beta)$ is given by:

$$\left\{D\left(\{u^{(1)}\}^{1/m}, \ldots, \{u^{(d)}\}^{1/m}; \theta, \beta\right)\right\}^m = D\left(u^{(1)}, \ldots, u^{(d)}; \theta/m, \beta\right), \tag{14}$$

As $m \to \infty$, this copula converges to the Logistic copula with shape parameter $\beta \geq 1$:

$$D_\infty^{(iid)}(\mathbf{u}) = D(\mathbf{u}; \beta) = \lim_{m\to\infty} D\left(u^{(1)}, \ldots, u^{(d)}; \theta/m, \beta\right) = \exp\left[-\left\{\sum_{j=1}^{d} (-\ln u^{(j)})^\beta\right\}^{1/\beta}\right],$$

uniformly in $\mathbf{u} \in [0,1]^d$. This result, originally stated in Bücher and Segers 2014, Proposition 4.3 for the bivariate case, can be extended to an arbitrary dimension without further arguments. Now, consider the following nested Archimedean copula given by:

$$D\left(D^{(O_1)}(\mathbf{u}^{(O_1)}; \theta, \beta_1), \ldots, D^{(O_G)}(\mathbf{u}^{(O_g)}; \theta, \beta_G); \theta, \beta_0\right). \tag{15}$$

We aim to show that this copula is in the domain of attraction of an AI-block model. That is the purpose of the proposition stated below.

Proposition 4. *Consider $1 \leq \beta_0 \leq \min\{\beta_1, \ldots, \beta_G\}$, then the nested Archimedean copula given in (15) is in the copula domain of attraction of an extreme value copula given by*

$$D\left(D^{(O_1)}(\mathbf{u}^{(O_1)}; \beta_1), \ldots, D^{(O_G)}(\mathbf{u}^{(O_G)}; \beta_G); \beta_0\right).$$

In particular, taking $\beta_0 = 1$ gives an AI-block model where extreme value random vectors $\mathbf{X}^{(O_g)}$ correspond to a Logistic copula with parameter shape β_g.

From the last conclusion of Proposition 4, we obtain Condition $\mathcal{A}$, that is $(\mathbf{Z}_t, t \in \mathbb{Z})$ in (13) is in max-domain of attraction of an AI-block model. Noticing that the exponent measure of each cluster is absolutely continuous with respect to the Lebesgue measure, Condition $\mathcal{B}$ is thus valid.

Remark 1. Notice that, using results from Bücher and Segers 2014; Zou, Volgushev, and Bücher 2021, in the i.i.d. case, i.e. $p = 1$, there exists an auxiliary function Ψ_D for D_m with $\Psi_D(m) = O(m^{-1})$. By using considerations after Equation (9), we thus obtain $d_m = O(m^{-1})$.

5. Real-data applications

5.1 Clustering brain extreme from EEG channel data

Epilepsy, a significant neurological disorder, manifests as recurring unprovoked seizures. These seizures represent uncontrolled and abnormal electricity activity in the brain, posing a negative impact on one's quality of life and potentially triggering comorbid conditions like depression and anxiety. During an episode of seizure, the patient may experience a loss of muscle control, which can result in accidents an injuries (see Strzelczyk et al. 2023).

One essential tool used in the diagnosis of epilepsy is electroencephalograms (EEGs). These EEGs are utilized to measure the electrical activity of the brain by employing a uniform array of electrodes. Each EEG channel is formed by calculating the potential difference between two electrodes and captures the combined potential of millions of neurons. The EEG plays a crucial role in capturing the intricate brain activity, especially during epileptic seizures, and requires analysis using statistical models. Currently, most analysis methods rely on Gaussian models that focus on the central tendencies of the data distribution (see, for example, Embleton, Knight, and Ombao 2020; Ombao, Von Sachs, and Guo 2005). However, a significant limitation of these approaches is their disregard for the fact that neuronal oscillations exhibit non-Gaussian probability distributions with heavy tails. To address this limitation, we employ AI-block models as a comprehensive framework to overcome the limitations of light-tailed Gaussian models and investigate the extreme neural behavior during an epileptic seizures.

The dataset used to evaluate our method comprises 916 hours of continuous scalp EEG data sampled at a rate of 256 Hz. This dataset was recorded from a total of 23 pediatric patients at Children's Hospital Boston, see, e.g., Shoeb 2009. We focus the analysis on the Patient number 5 which is the first patient where 40 hours of continuous scalp EEG were sampled without interruption. Throughout the recordings, the patient experienced a total of 5 events that were professionally identified as clinical seizures by experts. The pediatric EEG data used in this paper is contained within the CHB-MIT database, which can be downloaded from: https://physionet.org/content/chbmit/1.0.0/.

For each non-seizures and seizures events, we follow the same specific processing pipeline. First, we calculate the block maxima, then calibrate the threshold using the SECO metric, as suggested in Section 3.4 and supported by the numerical results in Appendix A. Finally, we perform the clustering task (see Algorithm (ECO)) using this adjusted threshold.

In the case of non-seizures records, we compute the block maxima using a duration of 4 minutes. Figure 1a illustrates the relationship between the SECO and the threshold τ. Two notable local minima are observed at $\tau = 0.24$ and $\tau = 0.4$. We execute the algorithm for both values and present the results for $\tau = 0.4$ as the results are more suited to AI-block models. Indeed, we obtain three clusters that demonstrate extreme dependence within the clusters while displaying weak extreme dependence in the block's off-diagonal (refer to Figure 1b). The spatial organisation of channel clusters is depicted in Figure 1c.

Regarding seizure events, as the time series spans only 558 seconds, we compute block maxima with a length of 5 seconds. Considering the heavy-tailed nature of oscillations during a seizure, we believe that the limited length of the block used would not introduce a significant bias with respect

to the domain of attraction. Figure 1d shows that the SECO is monotonically increasing without exhibiting a significant decline. Thus, the optimal selected threshold is the lowest value (in this case, $\tau = 0.1$), which results in the minimal cluster $\{1, \ldots, d\}$. This phenomenon is also reflected, in the extremal correlation matrix, where each channel exhibits strong extremal dependence with other channels. Consequently, the neurological disorder of the studied Patient 5 manifests simultaneous extremes across all channels, indicating generalized seizures with inter-channel communication.

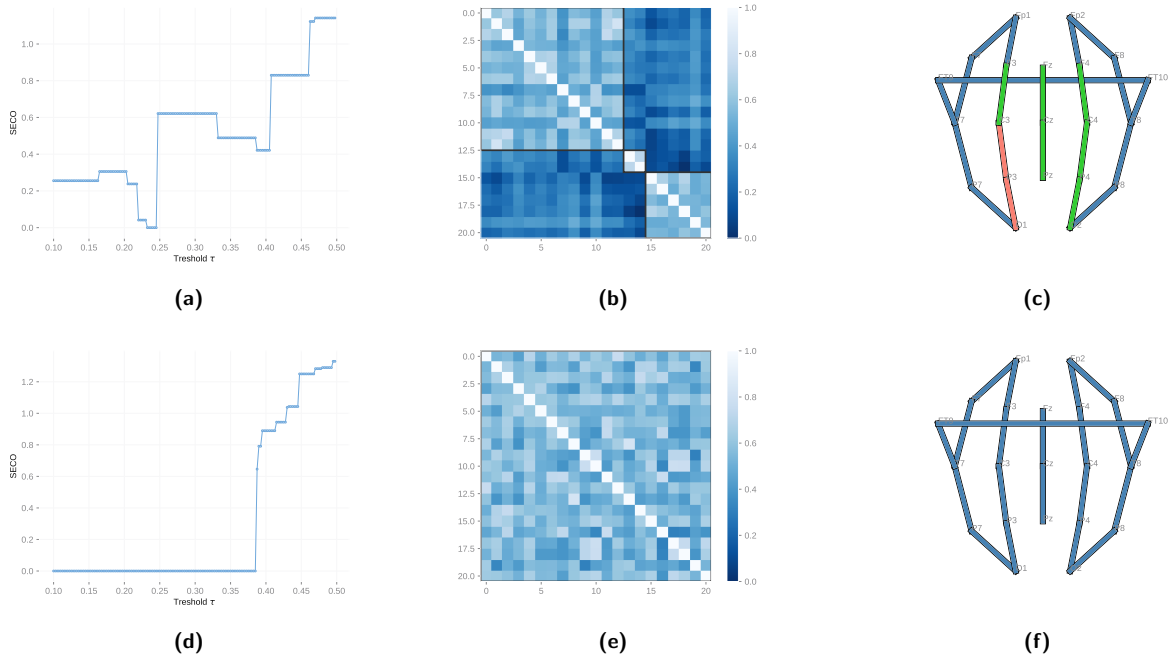

Figure 1. Clustering analysis on extreme brain activity derived from EEG channel data. The results are presented in the first and second rows, representing non-seizure and seizure events, respectively. The first column illustrates the behavior of the SECO metric as it relates to the threshold level, τ. The second column showcases the resulting clustering performed on the extremal correlation matrix using the optimal value of τ. Finally, the third column provides a spatial organisation of the clustered channels.

5.2 *Extremes on river network*

To demonstrate the novel regionalization method described in this paper, we employed biweekly maximum river discharge data, specifically, records collected over 14-day intervals, measured in (m^3/s). This dataset were sourced from a network of 1123 gauging stations strategically positioned across European rivers. The European Flood Awareness System (EFAS) provided these data, and they are accessible free of charge via the following website https://cds.climate.copernicus.eu/. EFAS primarily relies on a distributed hydrological model that operates on a grid-based system, focusing on extreme river basins. The model integrates various medium-range weather forecasts, including comprehensive sets from the Ensemble Prediction System (EPS). The dataset was generated by inputting gridded observational precipitation data, with a resolution of 5×5 km, into the LISFLOOD hydrological model across the EFAS domain. The temporal resolution utilized was a 24-hour time step, covering a span over 50 years.

For the calibration of the LISFLOOD within the EFAS framework, a total of 1137 stations from 215 different catchments across the Pan-European EFAS domain were used. From this list of stations with available coordinates, we extracted time-series data from the nearest cell where EFAS

data were accessible. However, in this pre-processing step, stations from Albania had to be excluded as the extracted time series were identical for those stations. Additionally, calibration stations from Iceland and Israel were removed since they were located far outside the domain. As a result, we were left with 1123 gauging stations, covering 10898 observed days of river discharge between 1991 and 2020. The biweekly block maxima approach yielded 783 observations.

Following the pipeline described in Section 5.1, in Figure 2a, the SECO is depicted as it evolves in relation to the threshold τ. The minimum value is attained at $\tau = 0.25$. Using this data-driven threshold, the Algorithm (ECO) is applied, resulting in 17 clusters, with 11 clusters comprising fewer than 20 stations. Figure 2b presents the resulting extremal correlation matrix, with clusters visually highlighted by squares. Within the clusters, there is evidence of asymptotic dependence, while moderate asymptotic dependence is observed in the off block-diagonal. Figure 2c provides a spatial representation of three main clusters. Notably, the clusters exhibit spatial concentration, despite the algorithm being unaware of their spatial dispersion. Overall, distinct clusters representing western, central, and northern Europe can be identified. It is crucial to emphasize that the northern Europe cluster includes stations situated in the Alps and the Pyrenees, which are geographically distant from the Scandinavian peninsula. Despite the geographical separation, these regions share mountainous terrain, and the simultaneous occurrence of extreme river discharges may be attributed to snow melting.

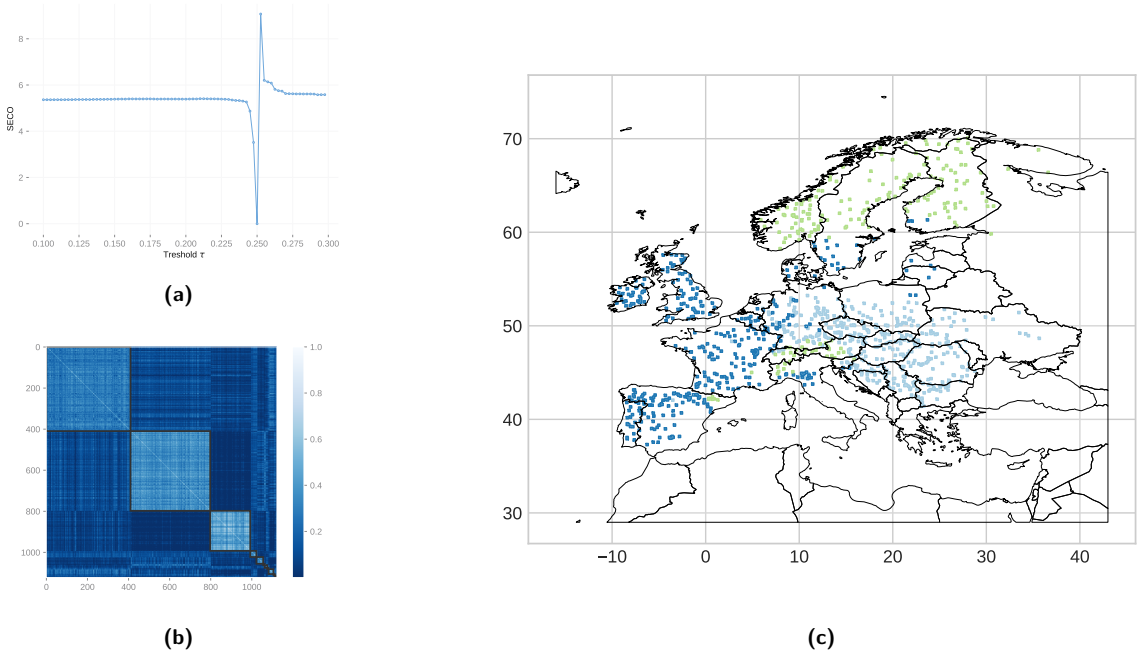

Figure 2. Clustering analysis on extreme river discharges on EFAS data. The first panel illustrates the behavior of the SECO metric as it relates to the threshold level, τ. The second panel showcases the resulting clustering performed on the extremal correlation matrix using the optimal value of τ. Finally, the third one provides a spatial representation of the clustered stations.

6. Conclusions

Our main focus in this work was to develop and analyze an algorithm for recovering clusters in AI-block models, and to understand how the dependence structure of maxima impacts the difficulty of clustering in these models. This is particularly challenging when we are dealing with high-dimensional

data and weakly dependent observations that are sub-asymptotically distributed. In order to better understand these phenomena, we ask stronger assumptions about the extremal dependence structure in our theoretical analysis. Specifically, we assume the asymptotic independence between blocks, which is the central assumption of AI-block models. This assumption enables us to examine the impact of the dependence structure and develop an efficient algorithm for recovering clusters in AI-block models. By employing this procedure, we can recover the clusters with high probability by employing a threshold that scales logarithmically with the dimension d. However, it remains important to explore the optimal achievable rate for recovering AI-block models.

In this paper, we find a bound for the minimal extremal correlation separation $\eta > 0$. A further goal is to find the minimum value η^* below which it is impossible, with high probability, to exactly recover $\bar{O}$ by any method. This question can be formally expressed using Le Cam's theory as follows:

$$\inf_{\hat{O}} \sup_{\mathcal{X} \in \mathbb{X}(\eta)} \mathbb{P}_{\mathcal{X}}(\hat{O} \neq \bar{O}) \geq \text{constant} > 0, \quad \forall\, \eta < \eta^*,$$

with $\mathbb{X}(\eta) = \{\mathcal{X}, \text{MECO}(\mathcal{X}) > \eta\}$ and the infimum is taken over all possible estimators. One possible direction to obtain such a result is to follow methods introduced by Drees 2001 for risk bounds of extreme value index. An interesting consequence of this result is to determine whether our procedure is optimal (in a minimax sense), i.e., whether the order of η^* and the one found in Theorem 2 are the same.

Acknowledgments

This work has been supported by the project ANR McLaren (ANR-20-CE23-0011). This work has been partially supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. This work was also supported by the french national programme LEFE/INSU

References

Agrawal, Rakesh, Ramakrishnan Srikant, et al. 1994. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, vldb,* 1215:487–499. Santiago, Chile.

Bador, Margot, Philippe Naveau, Eric Gilleland, Mercè Castellà, and Tatiana 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.

Balkema, August A, and Sidney I Resnick. 1977. Max-infinite divisibility. *Journal of Applied Probability* 14 (2): 309–319.

Beirlant, Jan, Yuri Goegebeur, Johan Segers, and Jozef L Teugels. 2004. *Statistics of extremes: theory and applications.* Vol. 558. John Wiley & Sons.

Bernard, Elsa, Philippe Naveau, Mathieu Vrac, and Olivier Mestre. 2013. Clustering of maxima: spatial dependencies among heavy rainfall in France. *Journal of climate* 26 (20): 7929–7937.

Boulin, Alexis, Elena Di Bernardino, Thomas Laloë, and Gwladys Toulemonde. 2022. Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework. *Journal of Multivariate Analysis* 192:105059. ISSN: 0047-259X. https://doi.org/https://doi.org/10.1016/j.jmva.2022.105059. https://www.sciencedirect.com/science/article/pii/S0047259X22000690.

Bücher, Axel, and Johan Segers. 2014. Extreme value copula estimation based on block maxima of a multivariate stationary time series. *Extremes* 17 (3): 495–528.

Bücher, Axel, Stanislav Volgushev, and Nan Zou. 2019. On second order conditions in the multivariate block maxima and peak over threshold method. *Journal of Multivariate Analysis* 173:604–619.

Bunea, Florentina, Christophe Giraud, Xi Luo, Martin Royer, and Nicolas Verzelen. 2020. Model assisted variable clustering: Minimax-optimal recovery and algorithms. *The Annals of Statistics* 48 (1): 111–137.

Chautru, Emilie. 2015. Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics* 9 (1): 383–418.

Chiapino, Maël, and Anne Sabourin. 2017. Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International workshop on new frontiers in mining complex patterns,* 132–147. Springer.

Chiapino, Maël, Anne Sabourin, and Johan Segers. 2019. Identifying groups of variables with the potential of being large simultaneously. *Extremes* 22:193–222.

Coles, Stuart, Janet Heffernan, and Jonathan Tawn. 1999. Dependence measures for extreme value analyses. *Extremes* 2 (4): 339–365.

Coles, Stuart G., and Jonathan A. Tawn. 1991. Modelling extreme multivariate events. *Journal of the Royal Statistical Society. Series B (Methodological)* 53 (2): 377–392.

Cooley, Dan, Philippe Naveau, and Paul Poncet. 2006. Variograms for spatial max-stable random fields. In *Dependence in probability and statistics,* 373–390. Springer.

Cooley, Daniel, Richard A. Davis, and Philippe Naveau. 2010. The pairwise beta distribution: a flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis* 101 (9): 2103–2117.

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

Drees, Holger. 2001. Minimax risk bounds in extreme value theory. *The Annals of Statistics* 29 (1): 266–294.

Drees, Holger, and Xin Huang. 1998. Best attainable rates of convergence for estimators of the stable tail dependence function. *Journal of Multivariate Analysis* 64 (1): 25–46. ISSN: 0047-259X. https://doi.org/https://doi.org/10.1006/jmva.1997.1708. https://www.sciencedirect.com/science/article/pii/S0047259X97917085.

Drees, Holger, and Anne Sabourin. 2021. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics* 15 (1): 908–943.

Embleton, Jonathan, Marina I Knight, and Hernando Ombao. 2020. Multiscale modelling of replicated nonstationary time series. *arXiv preprint arXiv:2005.09440.*

Engelke, Sebastian, and Adrien S Hitz. 2020. Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (4): 871–932.

Engelke, Sebastian, and Jevgenijs Ivanovs. 2021. Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application* 8:241–270.

Engelke, Sebastian, Jevgenijs Ivanovs, and Kirstin Strokorb. 2022. Graphical models for infinite measures with applications to extremes and Lévy processes. *arXiv preprint arXiv:2211.15769.*

Falk, M., J. Hüsler, and R.D. Reiss. 2010. *Laws of small numbers: extremes and rare events.* Springer Basel.

Fermanian, Jean-David, Dragan Radulovic, and Marten Wegkamp. 2004. Weak convergence of empirical copula processes. *Bernoulli* 10 (5): 847–860.

Ferreira, H. 2011. Dependence between two multivariate extremes. *Statistics & Probability Letters* 81 (5): 586–591. ISSN: 0167-7152. https://doi.org/https://doi.org/10.1016/j.spl.2011.01.014. https://www.sciencedirect.com/science/article/pii/S0167715211000216.

Gissibl, Nadine, and Claudia Klüppelberg. 2018. Max-linear models on directed acyclic graphs. *Bernoulli* 24 (4A): 2693–2720.

Goix, Nicolas, Anne Sabourin, and Stéphan Clémençon. 2016. Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial intelligence and statistics,* 75–83. PMLR.

Goix, Nicolas, Anne Sabourin, and Stephan Clémençon. 2017. Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis* 161:12–31.

Gudendorf, Gordon, and Johan Segers. 2010. Extreme-value copulas. In *Copula theory and its applications,* 127–145. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hofert, Marius, Raphaël Huser, and Avinash Prasad. 2018. Hierarchical Archimax copulas. *Journal of Multivariate Analysis* 167:195–211.

Hsing, Tailen. 1989. Extreme value theory for multivariate stationary sequences. *Journal of Multivariate Analysis* 29 (2): 274–291.

Huang, Xin. 1992. *Statistics of bivariate extreme values.* Thesis Publishers Amsterdam.

Hüsler, Jürg, and Rolf-Dieter Reiss. 1989. Maxima of normal random vectors: between independence and complete dependence. *Statistics & Probability Letters* 7 (4): 283–286.

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

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

Lee, Inbeom, Siyi Deng, and Yang Ning. 2021. Optimal variable clustering for high-dimensional matrix valued data. *arXiv preprint arXiv:2112.12909.*

Marcon, G., S.A. Padoan, P. Naveau, P. Muliere, and J. Segers. 2017. Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials. *Journal of Statistical Planning and Inference* 183:1–17.

Marius Hofert and Martin Mächler. 2011. Nested Archimedean copulas meet R: the nacopula package. *Journal of Statistical Software* 39 (9): 1–20.

Marshall, Albert W., and Ingram Olkin. 1983. Domains of Attraction of Multivariate Extreme Value Distributions. *The Annals of Probability* 11 (1): 168–177.

Meyer, Nicolas, and Olivier Wintenberger. 2021. Sparse regular variation. *Advances in Applied Probability* 53 (4): 1115–1148.

———. 2023. Multivariate sparse clustering for extremes. *Journal of the American Statistical Association,* no. just-accepted, 1–23.

Mohri, Mehryar, and Afshin Rostamizadeh. 2010. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research* 11 (26): 789–814. http://jmlr.org/papers/v11/mohri10a.html.

Naveau, Philippe, Armelle Guillou, Daniel Cooley, and Jean Diebolt. 2009. Modelling pairwise dependence of maxima in space. *Biometrika* 96 (1): 1–17.

Ombao, Hernando, Rainer Von Sachs, and Wensheng Guo. 2005. Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association* 100 (470): 519–531.

Pickands, James. 1981. Multivariate extreme value distribution. *Proceedings 43th, Session of International Statistical Institution, 1981* 49:859–878.

Pollard, David. 1981. Strong Consistency of K-Means Clustering. *The Annals of Statistics* 9 (1): 135–140. https://doi.org/10.1214/aos/1176345339. https://doi.org/10.1214/aos/1176345339.

Radulović, Dragan, Marten Wegkamp, and Yue Zhao. 2017. Weak convergence of empirical copula processes indexed by functions. *Bernoulli* 23 (4B): 3346–3384.

Resnick, S.I. 2008. *Extreme values, regular variation, and point processes.* Applied probability. Springer.

Rio, Emmanuel. 2017. *Asymptotic theory of weakly dependent random processes.* Vol. 80. Springer.

Rootzén, Holger, and Nader Tajvidi. 2006. Multivariate generalized pareto distributions. *Bernoulli* 12 (5): 917–930.

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

Schlather, Martin, and Jonathan Tawn. 2002. Inequalities for the extremal coefficients of multivariate extreme value distributions. *Extremes* 5 (1): 87–102.

Segers, Johan. 2020. One-versus multi-component regular variation and extremes of Markov trees. *Advances in Applied Probability* 52 (3): 855–878.

Shoeb, Ali Hossam. 2009. Application of machine learning to epileptic seizure onset detection and treatment. PhD diss., Massachusetts Institute of Technology.

Smith, Richard L. 1990. Max-stable processes and spatial extremes. *unpublished work.*

Strokorb, Kirstin. 2020. *Extremal independence old and new.*

Strzelczyk, Adam, Angel Aledo-Serrano, Antonietta Coppola, Adrien Didelot, Elizabeth Bates, Ricardo Sainz-Fuertes, and Charlotte Lawthom. 2023. The impact of epilepsy on quality of life: findings from a european survey. *Epilepsy & Behavior* 142:109179.

Takahashi, Rinya. 1987. Some properties of multivariate extreme value distributions and multivariate tail equivalence. *Annals of the Institute of Statistical Mathematics* 39:637–647.

———. 1994. Asymptotic independence and perfect dependence of vector components of multivariate extreme statistics. *Statistics & Probability Letters* 19 (1): 19–26.

Tawn, Jonathan A. 1990. Modelling multivariate extreme value distributions. *Biometrika* 77, no. 2 (June): 245–253.

Vaart, A. W., and Jon A Wellner. 1996. *Weak convergence.* New York, NY: Springer New York.

Zou, Nan, Stanislav Volgushev, and Axel Bücher. 2021. Multiple block sizes and overlapping blocks for multivariate time series extremes. *The Annals of Statistics* 49 (1): 295–320. https://doi.org/10.1214/20-AOS1957. https://doi.org/10.1214/20-AOS1957.

Appendix A. Numerical examples

Appendix A.1 Competitor clustering algorithms for extremes

In this section, we present some competitor algorithms: the spherical k-means (Chautru 2015; Janßen and Wan 2020) and hierarchical clustering using madogram as a dissimilarity (Bador et al. 2015; Bernard et al. 2013; Saunders, Stephenson, and Karoly 2021). The performance of the spherical k-means and hierarchical clustering will be compared with our Algorithm (ECO) in Appendix A.2.

The k-means procedure is a way to identify distinct groups within a population. This procedure involves partitioning a set of data into G groups (to be consistent with our notation). To do this, we first choose cluster centers $\psi_1, \dots, \psi_G$ for the points $\mathbf{Z}_1, \dots, \mathbf{Z}_n \in \mathbb{R}^d$ in order to minimize

$$W_n := \frac{1}{n} \sum_{i=1}^{n} \min_{g \in \{1,\dots,G\}} d(\mathbf{Z}_i, \psi_g),$$

where $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ is a distance function or, more generally, a dissimilarity function in $\mathbb{R}^d$. The motivation is to identify cluster centers such that distances of the observations to their nearest cluster center are minimized. Accordingly, all observations which are closest to the same cluster center are viewed as belonging to the same group.

While the original version of k-means uses the Euclidean distance, several alternatives choices of d have been suggested. As the extremal dependence structure can be described with the angular measure S (see Resnick 2008, Section 5 for details), a natural way to measure the distance between two points is by their angle. This corresponds to the spherical k-means clustering which is described as follow: for a given integer G, solve the following optimization problem

$$\frac{1}{n} \sum_{i=1}^{n} \min_{g \in \{1,\dots,G\}} d(\mathbf{Y}_i, \psi_g),$$

with $\mathbf{Y}_i$, i.i.d. observations from $\mathbf{Y}$, a random variable living on the unit sphere with law S. Consistency results with i.i.d. observations and for sufficiently many large observations had been

proved for this algorithm in Janßen and Wan 2020. The consistency result gives that the centroids obtained by minimizing the program above are close to the true centroids of the angular distribution.

In the framework of Bador et al. 2015; Bernard et al. 2013; Saunders, Stephenson, and Karoly 2021, the madogram is considered as a dissimilarity measure. This criterion can be read in the present context of block maxima method as

$$W_n = \frac{1}{k}\sum_{i=1}^{k} \min_{g\in\{1,\dots,G\}} \frac{1}{2}\left|\hat{\mathbf{U}}_{n,m,i} - \psi_g\right| = \int_{[0,1]^d} \min_{g\in\{1,\dots,G\}} \frac{1}{2}\left|\mathbf{u} - \psi_g\right| d\hat{C}_{n,m}(\mathbf{u}),$$

where $\hat{C}_{n,m}$ is the empirical copula defined as

$$\hat{C}_{n,m}(\mathbf{u}) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{\hat{\mathbf{U}}_{n,m,i}\leq \mathbf{u}\}}, \quad \mathbf{u}\in[0,1]^d. \tag{16}$$

For a copula C_m in the domain of attraction of an extreme value copula C_∞, let $\Psi = \{\psi_1,\dots,\psi_G\}$, be a set of cluster centers with $\psi_g \in \mathbb{R}^d$, $g\in\{1,\dots,G\}$ and consider the averaged distance from any observation to the closest element of Ψ as

$$W(\Psi, C) = \int_{[0,1]^d} \min_{\psi\in\Psi} \frac{1}{2}|\mathbf{u}-\psi| dC(\mathbf{u}).$$

To the best of our knowledge, consistency results for k-means procedure using the madogram have not yet been established. The following proposition tries to bridge this gap where the proof is given in Appendix E.2.

Proposition 5. *Let $(\mathbf{Z}_t, t\in\mathbb{Z})$ be a stationary multivariate random process with continuous univariate margins such that Conditions $\mathcal{A}$ in the main text and $\mathcal{C}$ hold. For each $\hat{C}_{n,m}$ in (16) and a given value $G\in\mathbb{N}$, denote by Ψ_G^n a random set which minimizes*

$$W(\Psi, \hat{C}_{n,m}) = \int_{[0,1]^d} \min_{\psi\in\Psi} \frac{1}{2}|\mathbf{u}-\psi| d\hat{C}_{n,m}(\mathbf{u}),$$

among all sets $\Psi\subset[0,1]^d$ with at most G elements. Accordingly, let us define Ψ_G the optimal set when we replace $\hat{C}_{n,m}$ by C and assume that for a given value of G, the set Ψ_G is uniquely determined. Thus Ψ_G^n converges almost surely to Ψ_G as $n\to\infty$.

From Proposition 5, the madogram seems to be a relevant dissimilarity to estimate the set of theoretical cluster centers with respect to the extreme value copula of $\mathbf{X}$. As far as we know, the madogram was used for clustering using the partitioning around medoids algorithm (Bador et al. 2015; Bernard et al. 2013) and the hierarchical clustering (Saunders, Stephenson, and Karoly 2021). For computational convenience, only the hierarchical clustering and spherical k-means are considered in the next Appendix A.2.

Appendix A.2 Numerical results

In this section, we investigate the finite-sample performance of our algorithm to retrieve clusters in AI-block models. We consider a number of AI-block models of increasing complexity where we compare the performance of our algorithm with state-of-the-art methods in literature, the Hierarchical Clustering (HC) using the madogram as dissimilarity and the spherical k-means (SKmeans) algorithms. We design three resulting partitions in the limit model C_∞:

E1 C_∞ is composed of two blocks O_1 and O_2, of equal lengths where $C_\infty^{(O_1)}$ and $C_\infty^{(O_2)}$ are Logistic extreme-value copulae with parameters set to $\beta_1 = \beta_2 = 10/7$.

E2 C_∞ is composed of $G = 5$ blocks of random sample sizes $d_1, \dots, d_5$ from a multinomial distribution with parameter $q_g = 0.5^g$ for $g \in \{1, \dots, 4\}$ and $q_5 = 1 - \sum_{g=1}^{4} q_g$. Each random vector is distributed according to a Logistic distribution where parameters $\beta_g = 10/7$ for $g \in \{1, \dots, 5\}$.

E3 We consider the same model as E2 where we add 5 singletons. Then we have 10 resulting clusters. Model with singletons are known to be the hardest model to recover in the clustering literature.

In Section 4, we consider observations from the model described in Equation (13). Here, the variable D is derived from a nested Archimedean copula, as indicated in Equation (15). Specifically, the outer Power Clayton copula with a parameter $\beta_0 = 1$ serves as the "mother" copula, while the outer Power Clayton copula with the same parameters $\beta_1 = \cdots = \beta_G = 10/7$ acts as the "childrens" copulae. It's worth noting that the subasymptotic copula D_m does not fall under the category of an extreme value copula. This can be observed by considering two observations, $u^{(i)}$ and $u^{(j)}$, belonging to the same cluster O_1. In this case, the nested Archimedean copula presented in Equation (15) takes the following form:

$$D^{(O_1)}(\mathbf{1}, u^{(i)}, u^{(j)}, \mathbf{1}; \theta, \beta_1),$$

where the margins for the indices outside of i and j are considered as 1. Consequently, the dependence is determined by an outer Power Clayton copula that does not exhibit max-stability. Similarly, when i and j belong to different clusters, the nested Archimedean copula in Equation (15) follows the expression:

$$D(\mathbf{1}, u^{(i)}, u^{(j)}, \mathbf{1}; \theta, 1),$$

representing a Clayton copula. It is worth noting that indices in different clusters exhibit dependence when the max-domain of attraction is not yet reached. This framework is particularly relevant as it allows us to evaluate the effectiveness of the proposed method in estimating the extremal dependence structure. We set $\theta = 1$ for every copula, as it does not alter the domain of attraction. Based on Proposition 4 and Bücher and Segers 2014, Proposition 4.1, we know that C_m falls within the max domain of attraction of the corresponding copulae C_∞ defined in Experiments E1-E3. In other words, it represents an AI-block model with a Logistic dependence structure for the marginals. We simulate them using the method proposed by the copula `R` package (Marius Hofert and Martin Mächler 2011). The goal of our algorithm is to cluster d variables in $\mathbb{R}^n$. Thus, to make comparisons, we transpose the dataset for the k-means algorithm in order to obtain centroids in $\mathbb{R}^d$. In contrast to our "blindfolded" algorithm that automatically infers the number of clusters, we need to specify it for SKmeans and HC. These procedures with this wisely chosen parameter are called "oracles". Several simulation frameworks are considered and detailed in the following.

F1 We first investigate the choice of the intermediate sequence m of the block length used for estimation. We let $m \in \{3, 6, \dots, 30\}$ with a fixed sample size $n = 10000$ and $k = \lfloor n/m \rfloor$.

F2 We compute the performance of the structure learning method for varying sample size n. Since the value of m which is required for consistent estimation is unknown in practice we choose $m = 20$.

F3 We show the relationship between the average SECO and exact recovery rate of the method presented in Section 3.4. We use the case $n = 16000$, $k = 800$ and $d = 1600$ to study the "large k, large d" of our approach.

In the simulation study, we use the fixed threshold $\alpha = 2 \times (1/m + \sqrt{\ln(d)/k})$ for F1 and F2 since our theoretical results given in Theorem 2 suggest the usage of a threshold proportional $d_m + \sqrt{\ln(d)/k}$ and we can show, in the i.i.d. settings (where $p = 1$) that $d_m = O(1/m)$ (see details in Appendix C.2). For F3, we vary α around its theoretical optimal value, on a fine grid. The specific parameter setting we employ involves setting $p = 0.9$, which is further detailed below and illustrated in Figure 3. In addition, we present the results of these evaluations in Figures 4, 5, and 6, showcasing the exact recovery rate for each algorithm while considering varying values of p within the range $\{0.5, 0.7, 1.0\}$, respectively. It is important to highlight that the observations are serially independent when $p = 1.0$.

Figure 3 states all the results we obtain from each experiment and framework considered in this numerical section. We plot exact recovery rate for Algorithm (ECO) with dimensions $d = 200$ and $d = 1600$. In the "large d" setting with $d = 1600$, we consider the performance of the HC algorithm using the madogram as a dissimilarity measure and the spherical k-means in the first two frameworks. Each experiment is performed using $p = 0.9$. As expected, the performance of our algorithm in F1 (see Figure 3, first row) is initially increasing in m, reaches a peak, and then decreases. This phenomenon depicts a trade-off between the sub-asymptotically regime and the accuracy of inference. Indeed, a large block's length m induces a lesser bias as we reach the domain of attraction. However, the number of block k is consequently decreasing and implies a high variance for the inference process. These joint phenomenona explain the parabolic form of the exact recovery rate for our algorithms for $d \in \{200, 1600\}$. Considering the framework F2 the performance of our algorithm is better as the number of block-maxima increases (see Figure 3, second row).

A classical pitfall for learning algorithms is high dimensional settings. Here, when the dimension increases from 200 to 1600, our algorithm consistently reports the maximal element $\bar{O}$ with a reasonable number of blocks. This is in accordance with our theoretical findings, as the difficulty of clustering in AI-block models, as quantified by η in Theorem 2, scales at a rate of $\sqrt{\ln(d)k^{-1}}$. This rate has a moderate impact on the dimension d. In the framework F3, the numerical studies in Figure 3 (third row) shows that the optimal ranges of τ values, for high exact recovery percentages, are also associated with low average SECO losses. This supports our data-driven choices of τ provided in Section 3.4.

We notice that the HC algorithm using the madogram as dissimilarity performs very well in each configuration even when the inference is strongly biased, i.e., the block length m is small hence we are far from the domain of attraction. This can be explained by the fact that madograms are lower when $a \overset{\bar{O}}{\sim} b$ and higher when $a \overset{\bar{O}}{\not\sim} b$. This is effectively true by construction of the madogram in the domain of attraction of $\mathbf{X}$ but it is even true in our considered sub-asymptotic framework. Hence, by construction of the HC, i.e., by merging the two closest centroids in terms of madogram, we obtain the correct partitioning of $\mathbf{X}$ even when the domain of attraction is not reached. To compare, our algorithm gives one and unique cluster, i.e., the vector is completely dependent, when the block's length m is too small and we are not yet in the domain of attraction of $\mathbf{X}$. This behavior is desirable as it corresponds to what it is effectively observed, the whole vector is dependent. This is a leading argument for model-based clustering which are designed for a specific model and where the inference remains coherent with the constructed target. One drawback of using HC with the madogram, as previously described, is the need to specify the number of groups G beforehand, which is not always straightforward. Despite this limitation, the HC procedure with the madogram performs well in retrieving clusters in AI-block models when the true number of clusters is known. Further researches can be lead in order to adapt our algorithm with a hierarchical design as proposed by Lee, Deng, and Ning 2021 for the algorithm of Bunea et al. 2020.

For the same reasons as for the HC case, the SKmeans performs well for Experiment (E1) and

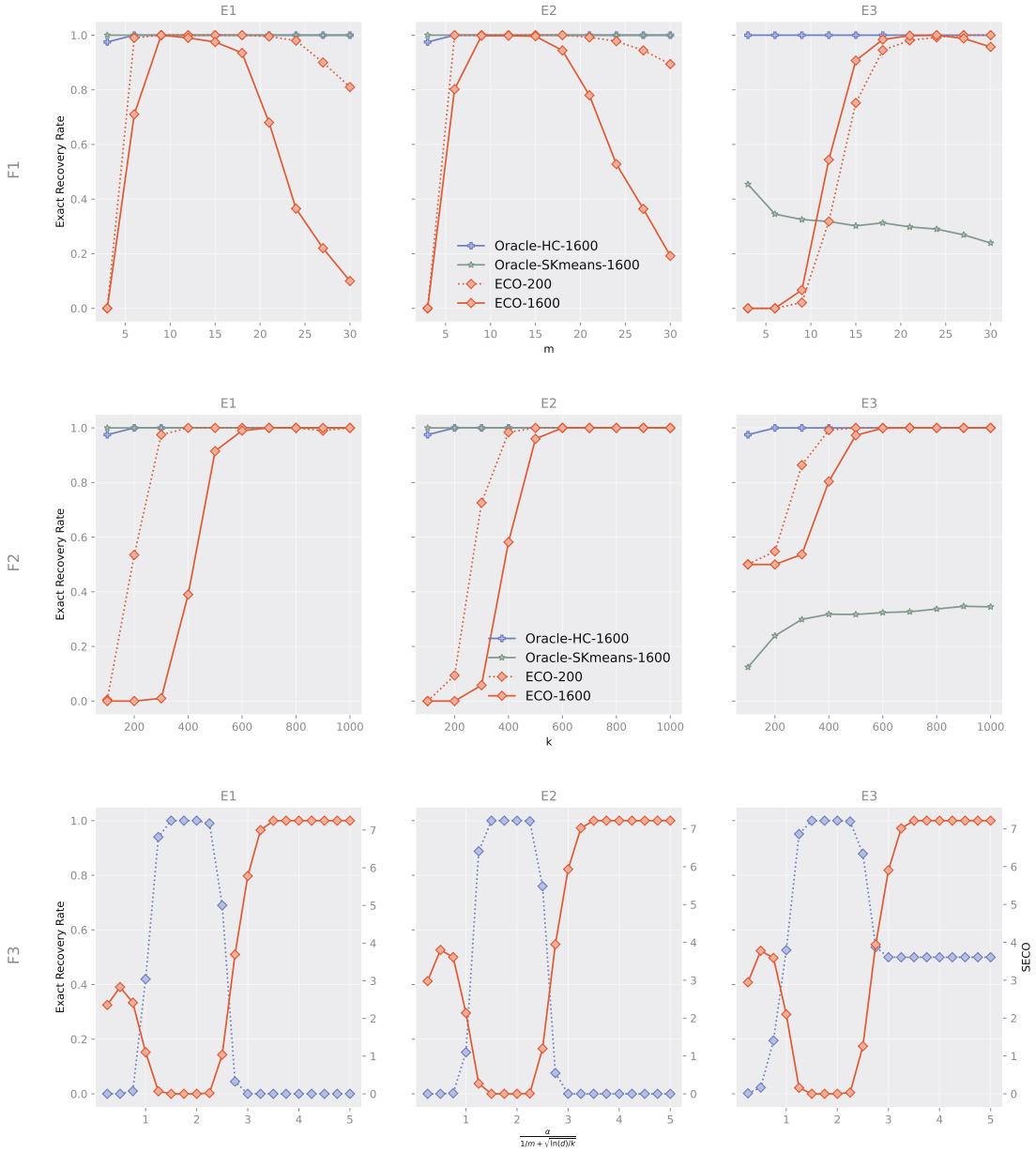

Figure 3. Simulation results with $p = 0.9$. From top to bottom: Framework F1, Framework F2, Framework F3. From left to right: Experiment E1, Experiment E2, Experiment E3. Exact recovery rate for our algorithm (red, diamond points), for the HC (blue, plus points) and the SKmeans (green, star points) for F1 and F2 across 100 runs. Dotted lines correspond to $d = 200$, solid lines to $d = 1600$. The threshold τ is taken as $2 \times (1/m + \sqrt{\ln(d)}/k)$. For F3, average SECO losses (red solid lines, circle points) and exact recovery percentages (blue dotted lines, diamond points) across 100 simulations. For better illustration, the SECO losses are standardized first by subtracting the minimal SECO loss in each figure, and the standardized SECO losses plus 1 are then plotted on the logarithmic scale.

Experiment (E2) for all considered values of m. However, when we consider Experiment (E3), its performance drastically decreases. Furthermore, the exact recovery rate decreases as m increases, which is not desirable in extreme settings. However, a rigorous method for choosing G is currently lacking and it remains an hyperparameter that must be chosen by the statistician. When the hyperparameter is known and equal to the true value, clusters are correctly inferred for Experiments

E1 and E2 for the HC algorithm and the SKmeans, but not for Experiment E3 for the SKmeans. Our algorithm, with the threshold specified in Theorem 2, can reach this level of performance depicted by the HC with madogram without specifying the number of clusters.

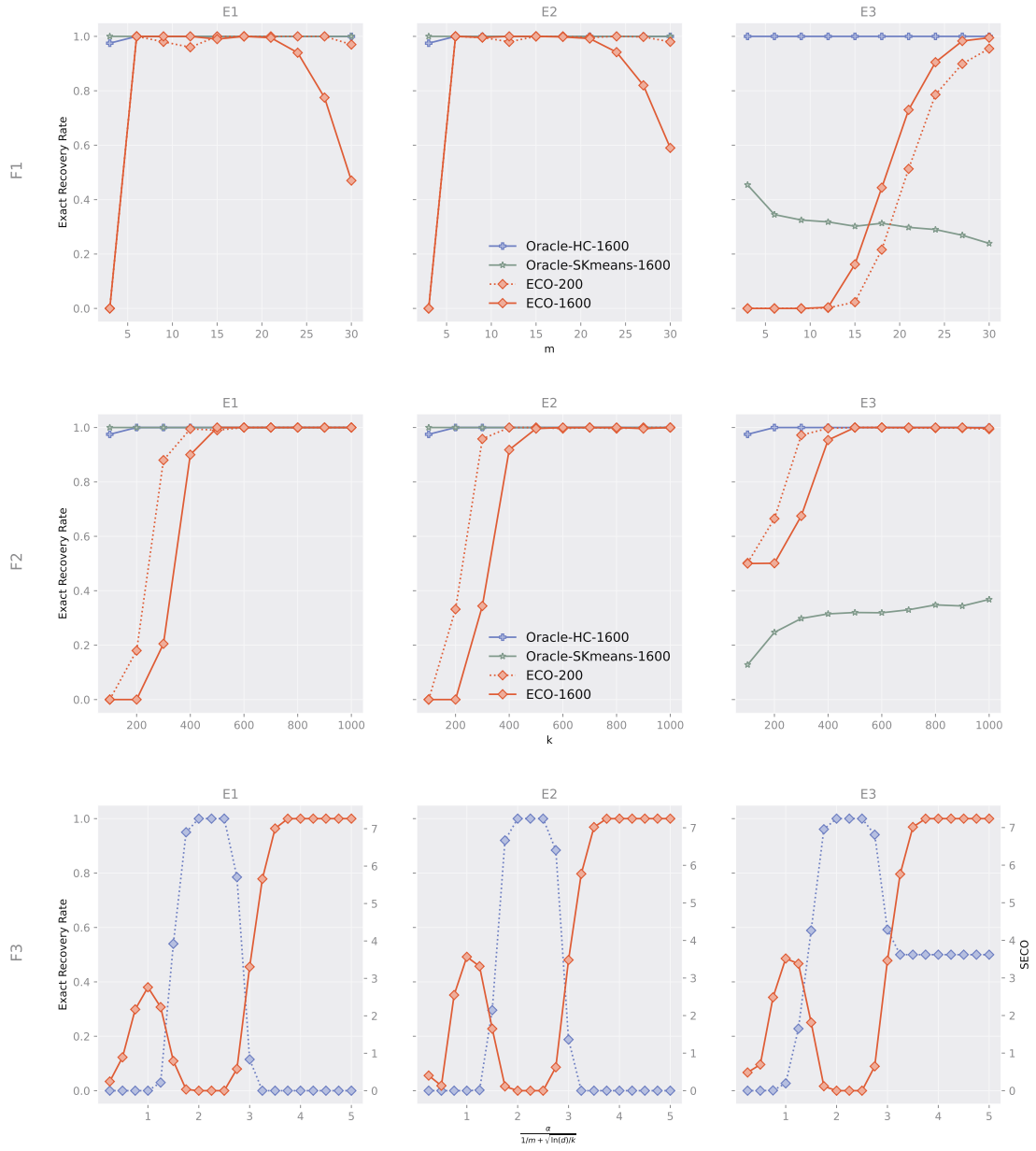

Figure 4. Simulation results for $p = 0.5$. From top to bottom: Framework F1, Framework F2, Framework F3. From left to right: Experiment E1, Experiment E2, Experiment E3. Exact recovery rate for our algorithm (red, diamond points), for the HC k-means (blue, plus points) and the SKmeans (green, star points) for F1 and F2 across 100 runs. Dotted lines correspond to $d = 200$, solid lines to $d = 1600$. The threshold τ is taken as $2 \times (1/m + \sqrt{\ln(d)}/k)$. For F3, average SECO losses (red solid lines, circle points) and exact recovery percentages (blue dotted lines, diamond points) across 100 simulations. For better illustration, the SECO losses are standardized first by subtracting the minimal SECO loss in each figure, and the standardized SECO losses plus 1 are then plotted on the logarithmic scale.

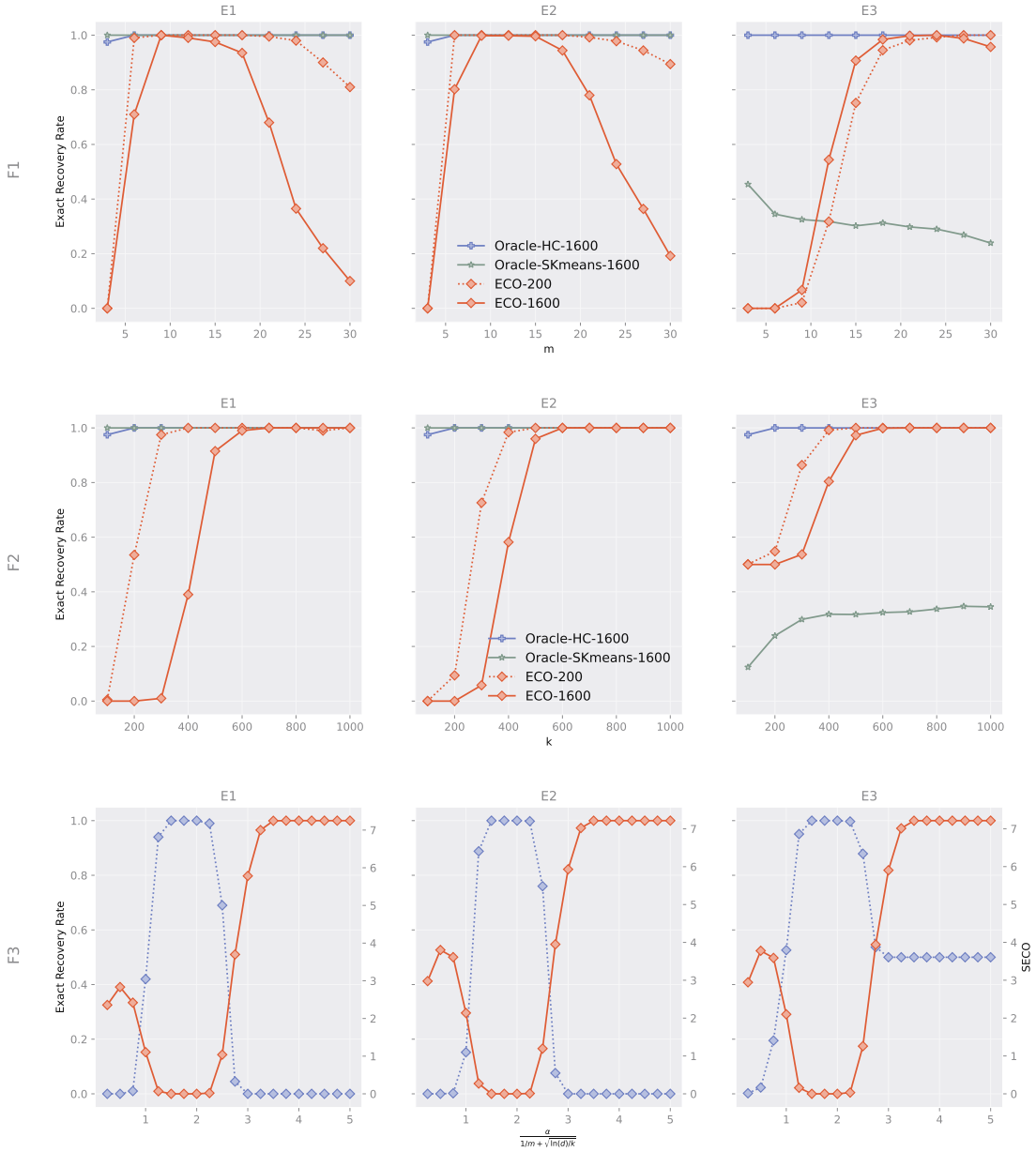

Figure 5. Simulation results for $p = 0.7$. From top to bottom: Framework F1, Framework F2, Framework F3. From left to right: Experiment E1, Experiment E2, Experiment E3. Exact recovery rate for our algorithm (red, diamond points), for the HC k-means (blue, plus points) and the SKmeans (green, star points) for F1 and F2 across 100 runs. Dotted lines correspond to $d = 200$, solid lines to $d = 1600$. The threshold τ is taken as $2 \times (1/m + \sqrt{\ln(d)}/k)$. For F3, average SECO losses (red solid lines, circle points) and exact recovery percentages (blue dotted lines, diamond points) across 100 simulations. For better illustration, the SECO losses are standardized first by subtracting the minimal SECO loss in each figure, and the standardized SECO losses plus 1 are then plotted on the logarithmic scale.

Appendix A.3 Additional competitors to AI block model detection

In the context of regular variation framework, the task of identifying the groups within d variables which can exhibit concomitant extreme translates to discerning the support of the exponent measure. The outcome of the DAMEX algorithm Goix, Sabourin, and Stéphan Clémençon 2016; Goix, Sabourin, and Stephan Clémençon 2017 will furnish a list of features denoted as α, which is a

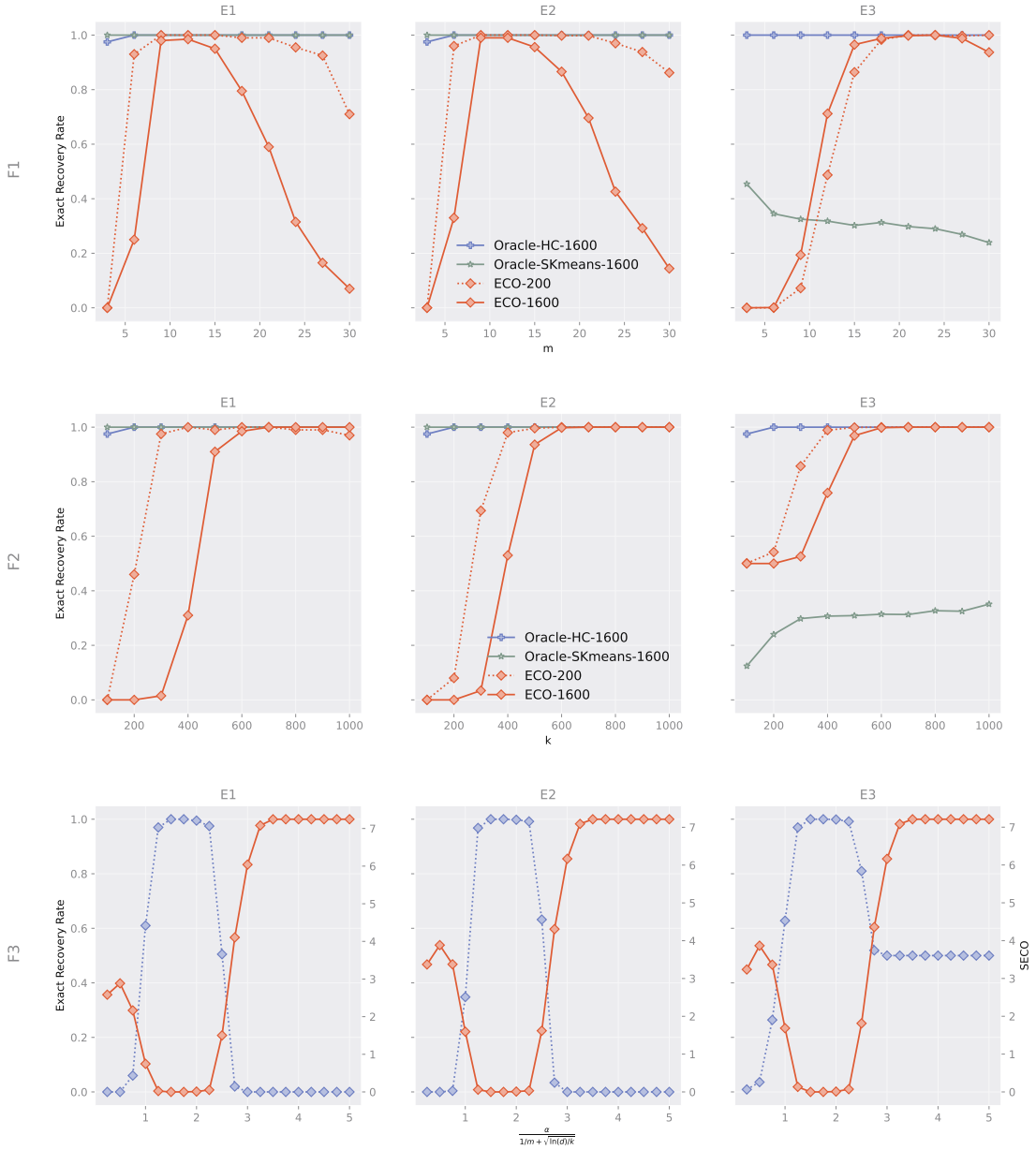

Figure 6. Simulation results for $p = 1.0$ (serially independent case). From top to bottom: Framework F1, Framework F2, Framework F3. From left to right: Experiment E1, Experiment E2, Experiment E3. Exact recovery rate for our algorithm (red, diamond points), for the HC k-means (blue, plus points) and the SKmeans (green, star points) for F1 and F2 across 100 runs. Dotted lines correspond to $d = 200$, solid lines to $d = 1600$. The threshold τ is taken as $2 \times (1/m + \sqrt{\ln(d)}/k)$. For F3, average SECO losses (red solid lines, circle points) and exact recovery percentages (blue dotted lines, diamond points) across 100 simulations. For better illustration, the SECO losses are standardized first by subtracting the minimal SECO loss in each figure, and the standardized SECO losses plus 1 are then plotted on the logarithmic scale.

subset of the set $\{1, \ldots, d\}$. These features possess an empirical exponent measure mass exceeding a user-defined threshold in specific cares. However, when the empirical version of the exponent measure is scattered over a large number of such cones, the DAMEX algorithm does not discover a clear-cut threshold structure.

To address this challenge, Chiapino and Sabourin 2017 suggest employing the Apriori algorithm

Agrawal, Srikant, et al. 1994 for mining frequent item set accompanied by a novel stopping criterion. This algorithm adopts a bottom-up strategy, commencing with individual elements (singletons), and gradually expanding the groups by one element at each step. This expansion occurs only if there is substantial evidence that all the components can exhibit extreme behavior simultaneously. The stopping criterion utilized a threshold-based and involves the empirical estimator of the conditional tail dependence coefficient.

This research was further extended by Chiapino, Sabourin, and Segers 2019, who introduced three additional stopping critera based on formal hypothesis testing. These tests rely on the empirical estimation of the conditional tail dependence coefficient, Hill's estimator and Peng's estimator. The tests are applied to each subface of a maximal face with mass, resulting in multiple testing issues and potentially lengthy execution times as the cluster sizes increases.

Our objective is to compare the performance of the DAMEX and CLEF algorithms in recovering the thinnest partition that represents an AI-block model for a given vector. Two variations of the CLEF algorithm are obtained by employing different criteria to identify subsets α as tail dependent, specifically using Hill's estimator and Peng's estimator. To conduct an experiment, we generate datasets according to the following procedure: the dimension is fixed at $d = 12$, and we can construct an unbiased AI-block model consisting of clusters with respective sizes $d_1 = 4$, $d_2 = 3$ and $d_3 = 5$ and logistic dependence with $\beta = 10/8$. Due to the computational complexity of the CLEF algorithm, which increases with cluster size, we limit our experiment to this range of dimensions. Additionally, for the sake of completeness, we include the output of the (ECO) algorithm in our comparison where we estimate the extremal coefficient and the extremal correlation using the peak-over-threshold approach:

$$\hat{\theta}_n(a,b) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{R_i^{(a)}>n-k+0.5,\, R_i^{(b)}>n-k+0.5\}}, \quad \hat{\chi}_n(a,b) = 2 - \hat{\theta}_n(a,b),$$

where $R_i^{(j)}$ denote the rank for the jth component in the observed sample.

We generate datasets of size $n \in \{100, 200, \ldots, 1000\}$. For each sample size, we simulate 100 independent datasets following the procedure outlined earlier. Our goal is to compare the performance of the five algorithms in recovering the thinnest partition. In the case of the CLEF algorithms utilizing Hill's and Peng's estimators, we set the confidence level δ to 0.005. For the CLEF algorithm, the threshold is chosen as 0.05. In the DAMEX algorithm, we retain the top 7 subsets with the highest empirical mass, and we set the subspace thickening parameter ϵ to 0.3. It is worth noting that the default value for ϵ is 0.1, but it yields poor performance. The threshold parameter in (ECO) algorithm is set using the data-driven method described in Section 3.4 in the main text. For each algorithm, we choose $k = 50$ as the number of retained greatest observed values in the sample.

Algorithm (ECO) consistently achieves the best overall scores for each sample size. However, it is important to note that methods being compared are not clustering algorithms and not specifically designed to recover groups with this particular structure. As mentioned earlier, the numerical framework has been simplified to its bare minimum for computational efficiency. Our clustering algorithm is designed to recover groups with a much larger number of entities, and the compared methods are not tailored for this purpose. In fact, for the competitors, only a small proportion of all $2^d - 1$ subsets has to be examined, while the computational complexity for each subset is low. The underlying sparsity assumption in our method is that there are only a small number of groups of variables that can exhibit extreme behavior simultaneously (referred as Sparsity 2a in Engelke and Ivanovs 2021). However the sparsity assumption in the compared methods is that each of these groups of concomitant extremes contains only a small number of variables (Sparsity 2b in

Engelke and Ivanovs 2021). These considerations help explain the performance of these competitor algorithms within a framework for which they were not originally designed.

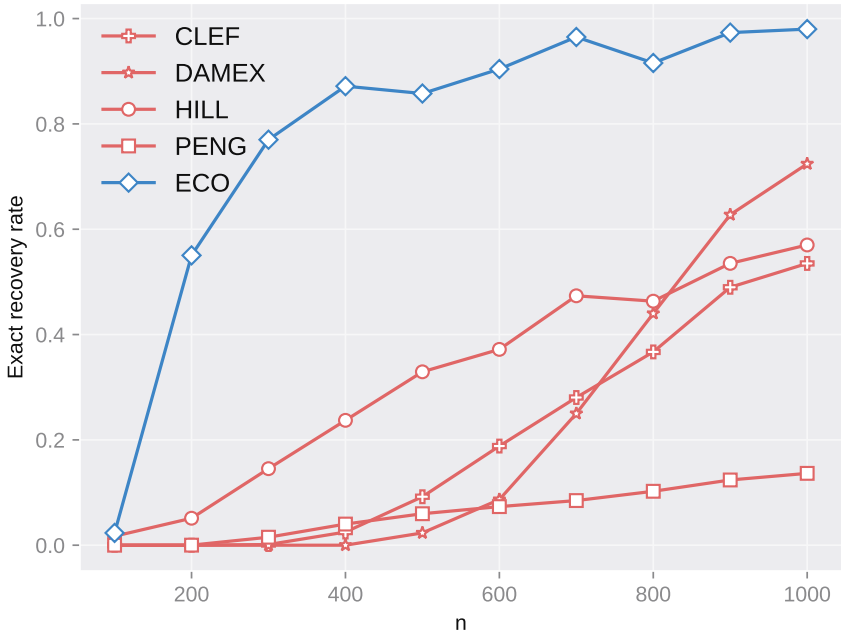

Figure 7. Simulation results for additional competitors of the ECO algorithm. Exact recovery rate for our algorithm (blue, diamond points), for the CLEF algorithm and its variants (red, plus, circle, square points) and the DAMEX (red, star points) across 100 runs. The threshold τ is taken using the data-driven approach described in Section 3.4 in the main text. In the case of CLEF variants (HILL, PENG), we set the confidence level set $\delta = 0.005$. For the CLEF algorithm, the threshold is set to 0.05. In the DAMEX algorithm, we retain the top 7 subsets with the highest empirical mass, and we set the subspace thickening parameter $\epsilon = 0.3$. For each algorithm, we choose the threshold $k = 50$.

Appendix B. Details on mixing coefficients

Consider $\mathbf{Z} = (Z^{(1)}, \ldots, Z^{(d)})$ and $\mathbf{Z}_t = (Z_t^{(1)}, \ldots, Z_t^{(d)})$, where $t \in \mathbb{Z}$ be respectively a d-dimensional random vector with law F and a strictly stationary multivariate random process distributed according to $\mathbf{Z}$. For the process $(\mathbf{Z}_t, t \in \mathbb{Z})$, let

$$\mathcal{F}_k = \sigma(\mathbf{Z}_t, t \leq k), \quad \text{and} \quad \mathcal{G}_k = \sigma(\mathbf{Z}_t, t \geq k),$$

be respectively the natural filtration and "reverse" filtration of $(\mathbf{Z}_t, t \in \mathbb{Z})$. Many types of mixing conditions exist in the literature. The weakest among those most commonly used is called strong or α-mixing. Specifically, for two σ-fields $\mathcal{A}_1$ and $\mathcal{A}_2$ of a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ the α-mixing coefficient of a multivariate random process is defined for $\ell \geq 1$

$$\alpha(\ell) = \sup_{t \in \mathbb{Z}} \alpha\left(\mathcal{F}_t, \mathcal{G}_{t+\ell}\right), \tag{17}$$

where

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

For any process $(\mathbf{Z}_t, t \in \mathbb{Z})$, let

$$\beta(\mathcal{A}_1, \mathcal{A}_2) = \sup \frac{1}{2} \sum_{i,j \in I \times J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|,$$

where the sup is taken over all finite partitions $(A_i)_{i\in I}$ and $(B_j)_{j\in J}$ of Ω with the sets A_i in $\mathcal{A}_1$ and the sets B_j in $\mathcal{A}_2$. The β-mixing (or completely regular) coefficient is defined as follows

$$\beta(\ell) = \sup_{t\in\mathbb{Z}} \beta(\mathcal{F}_t, \mathcal{G}_{t+\ell}). \tag{18}$$

By considering

$$\varphi(\mathcal{A}_1, \mathcal{A}_2) = \sup_{A_1, A_2 \in \mathcal{A}_1 \times \mathcal{A}_2, \mathbb{P}(A_1)\neq 0} |\mathbb{P}(A_2|A_1) - \mathbb{P}(A_1)|,$$

the φ-mixing coefficient is defined by

$$\varphi(\ell) = \sup_{t\in\mathbb{Z}} \varphi(\mathcal{F}_t, \mathcal{G}_{t+\ell}) \tag{19}$$

It should be noted that if the original process $(\mathbf{Z}_t, t \in \mathbb{Z})$ satisfies an α- or β- or φ-mixing condition, then the stationary process $(f(\mathbf{Z}_t), t \in \mathbb{Z})$ for a measurable function f also satisfies the same mixing condition. The α-mixing rate, β-mixing rate, and φ-mixing rate of the stationary process are all bounded by the corresponding rate of the original process. In terms of their order, the three mixing coefficients are related as follows:

$$\alpha(\ell) \leq \beta(\ell) \leq \varphi(\ell). \tag{20}$$

This means that the α-mixing coefficient is the weakest, followed by the β-mixing coefficient, and finally the φ-mixing coefficient is the strongest.

Appendix C. Proofs of main results

In the subsequent section of our materials, we employ the notation $(\mathbf{1}, \mathbf{x}^{(B)}, \mathbf{1})$ having its jth component equal to $x^{(j)}\mathbb{1}_{\{j\in B\}} + \mathbb{1}_{\{j\notin B\}}$. In a similar way, we note $(\mathbf{0}, \mathbf{x}^{(B)}, \mathbf{0})$ the vector in $\mathbb{R}^d$ which equals $x^{(j)}$ if $j \in B$ and 0 otherwise.

Appendix C.1 Proofs of Section 2

In Proposition 1, we prove that the function introduced in Section 2.2 is an extreme-value copula. We do this by showing that its margins are distributed uniformly on the unit interval [0,1] and that it is max-stable, which is a defining characteristic of extreme-value copulae.

Proof of Proposition 1 We first show that C is a copula function. It is clear that $C(\mathbf{u}) \in [0,1]$ for every $\mathbf{u} \in [0,1]^d$. We check that its univariate margins are uniformly distributed on $[0,1]$. Without loss of generality, take $u^{(i_{1,1})} \in [0,1]$ and let us compute

$$C(1, \ldots, u^{(i_{1,1})}, \ldots, 1) = C^{(O_1)}(u^{(i_{1,1})}, 1, \ldots, 1) = u^{(i_{1,1})}.$$

So C is a copula function. We now have to prove that C is an extreme-value copula. We recall that C is an extreme-value copula if and only if C is max-stable, that is for every $m \geq 1$

$$C(u^{(1)}, \ldots, u^{(d)}) = C(\{u^{(1)}\}^{1/m}, \ldots, \{u^{(d)}\}^{1/m})^m.$$

By definition, we have

$$C(\{u^{(1)}\}^{1/m}, \ldots, \{u^{(d)}\}^{1/m})^m = \Pi_{g=1}^{G} \left\{ C^{(O_g)} \left(\{u^{(i_{g,1})}\}^{1/m}, \ldots, \{u^{(i_{g,d_g})}\}^{1/m} \right) \right\}^m.$$

Using that $C^{(O_1)}, \dots, C^{(O_G)}$ are extreme-value copulae, thus max stable, we obtain

$$C(\{u^{(1)}\}^{1/m}, \dots, \{u^{(d)}\}^{1/m})^m = \Pi_{g=1}^{G} C^{(O_g)}\left(u^{(i_{g,1})}, \dots, u^{(i_{g,d_g})}\right) = C(u^{(1)}, \dots, u^{(d)}).$$

Thus C is an extreme-value copula. Finally, we prove that C is associated to the random vector $\mathbf{X} = (\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)})$, that is

$$\mathbb{P}\{\mathbf{X} \leq \mathbf{x}\} = C(H^{(1)}(x^{(1)}), \dots, H^{(d)}(x^{(d)})), \quad \mathbf{x} \in \mathbb{R}^d.$$

Using mutual independence between random vectors, we have

$$\begin{aligned}
\mathbb{P}\{\mathbf{X} \leq \mathbf{x}\} &= \Pi_{g=1}^{G} \mathbb{P}\left\{X^{(i_{g,1})} \leq x^{(i_{g,1})}, \dots, X^{(i_{g,d_g})} \leq x^{(i_{g,d_g})}\right\} \\
&= \Pi_{g=1}^{G} C^{(O_g)}\left(H^{(i_{g,1})}(x^{(i_{g,1})}), \dots, H^{(i_{g,d_g})}(x^{(i_{g,d_g})})\right) \\
&= C(H^{(1)}(x^{(1)}), \dots, H^{(d)}(x^{(d)})).
\end{aligned}$$

Hence the result. □

Theorem 1, proved below, establishes several fundamental properties of the set $\mathcal{L}(\mathbf{Z})$, including the fact that subpartitions of an element $O \in \mathcal{L}(\mathbf{Z})$ also belong to $\mathcal{L}(\mathbf{Z})$ (item (i)), the ordering of partitions and their intersections (item (ii)) and the stability of the intersection of two elements $O, S \in \mathcal{L}(\mathbf{Z})$ (item (iii)). Using these results, the theorem also provides an explicit construction of the unique maximal element $\bar{O}(\mathbf{Z})$ of $\mathcal{L}(\mathbf{Z})$ (see item (iv)).

Proof of Theorem 1 For (i), if $\mathbf{Z} \sim S$, then there exist an extreme value random vector with distribution H such that $F \in D(H)$ and a partition $S = \{S_1, \dots, S_G\}$ of $\{1, \dots, d\}$ which induces mutually independent random vectors $\mathbf{X}^{(S_1)}, \dots, \mathbf{X}^{(S_G)}$. As S is a sub-partition of O, it also generates a partition where vectors are mutually independent.

Now let us prove (ii), take $g \in \{1, \dots, G\}$ and $a, b \in (O \cap S)_g$, in particular $a \overset{O}{\sim} b$, thus there exists $g' \in \{1, \dots, G'\}$ such that $a, b \in O_{g'}$. The following inclusion $(O \cap S)_g \subseteq O_{g'}$ is hence obtained and the second statement follows.

The third result (iii) comes down from the definition for the direct sense and by (i) and (ii) for the reverse one. We now go to the last item of the theorem, i.e. item (iv). The set $\mathcal{L}(\mathbf{Z})$ is non-empty since the trivial partition $O = \{1, \dots, d\}$ belongs to $\mathcal{L}(\mathbf{Z})$. It is also a finite set, and we can enumerate it $\mathcal{L}(\mathbf{Z}) = \{O_1, \dots, O_M\}$. Define the sequence $O'_1, \dots, O'_M$ recursively according to

- $O'_1 = O_1$,
- $O'_g = O_g \cap O'_{g-1}$ for $g = 2, \dots, M$.

According to (iii), we have that by induction $O'_1, \dots O'_M \in \mathcal{L}(\mathbf{Z})$. In addition, we have both $O'_{g-1} \leq O'_g$ and $O_g \leq O'_g$, so by induction $O_1, \dots, O_g \leq O'_g$. Hence the partition $\bar{O}(\mathbf{Z}) := O'_M = O_1 \cap \dots \cap O_{M-1}$ is the maximum of $\mathcal{L}(\mathbf{Z})$. □

Appendix C.2 Proofs of Section 3

Denote by $C^o_{n,m}$ the empirical estimator of the copula C_m based on the (unobservable) sample $(U^{(j)}_{m,1}, \dots, U^{(j)}_{m,k})$ for $j \in \{1, \dots, d\}$. In Proposition 2 we state a concentration inequality for the madogram estimator. This inequality is obtained through two main steps, that are using classical concentration inequalities, such as Hoeffding and McDiarmid inequalites and chaining arguments in our specific framework of multivariate mixing random process.

Proof of Proposition 2 Let us define the following quantity

$$\hat{\nu}^o_{n,m} = \frac{1}{k}\sum_{i=1}^{k}\left[\bigvee_{j=1}^{d} U^{(j)}_{m,i} - \frac{1}{d}\sum_{j=1}^{d} U^{(j)}_{m,i}\right], \tag{21}$$

that is the madogram estimated through the sample $\mathbf{U}_{m,1},\dots,\mathbf{U}_{m,k}$. Then, the following bound is given:

$$\left|\hat{\nu}_{n,m} - \nu_m\right| \leq \left|\hat{\nu}_{n,m} - \hat{\nu}^o_{n,m}\right| + \left|\hat{\nu}^o_{n,m} - \nu_m\right|.$$

For the second term, using the triangle inequality, we obtain

$$\begin{aligned}\left|\hat{\nu}^o_{n,m} - \nu_m\right| &\leq \left|\frac{1}{k}\sum_{i=1}^{k}\left\{\bigvee_{j=1}^{d} U^{(j)}_{m,i} - \mathbb{E}\left[\bigvee_{j=1}^{d} U^{(j)}_{m,i}\right]\right\}\right| + \left|\frac{1}{k}\sum_{i=1}^{k}\left\{\frac{1}{d}\sum_{j=1}^{d} U^{(j)}_{m,i} - \mathbb{E}\left[\frac{1}{d}\sum_{j=1}^{d} U^{(j)}_{m,i}\right]\right\}\right| \\ &\triangleq E_1 + E_2,\end{aligned}$$

and for the first term,

$$\left|\hat{\nu}_{n,m} - \hat{\nu}^o_{n,m}\right| \leq 2 \sup_{j\in\{1,\dots,d\}}\sup_{x\in\mathbb{R}}\left|\hat{F}^{(j)}_{n,m}(x) - F^{(j)}_m(x)\right| \triangleq E_3.$$

The rest of this proof is devoted to control each term: E_1, E_2 and E_3. Notice that the sequences $(\bigvee_{j=1}^{d} U^{(j)}_{n,m,i})_{i=1}^{k}$, $(d^{-1}\sum_{j=1}^{d} U^{(j)}_{n,m,i})_{i=1}^{k}$ and $(\mathbb{1}_{\{M^{(j)}_{n,m,i}\leq x\}})_{i=1}^{k}$ share the same mixing regularity as $(\mathbf{Z}_t)_{t\in\mathbb{Z}}$ as measurable transformation of this process. Thus, they are in particular algebraically φ-mixing.

Control of the term E_1. For every $i\in\{1,\dots,k\}$, we have that $||\bigvee_{j=1}^{d} U^{(j)}_{n,m,i}||_\infty \leq 1$, by applying the Hoeffding's inequality for algebraically φ-mixing sequences (see Rio 2017, Corollary 2.1) we can control the following event, for $t>0$,

$$\mathbb{P}\left\{E_1 \geq t\right\} \leq \sqrt{e}\exp\left\{-\frac{t^2 k}{2(1+4\sum_{i=1}^{k-1}\varphi(i))}\right\}.$$

The term in the numerator can be bounded as

$$1 + 4\sum_{i=1}^{k}\varphi(k) \leq 1 + 4\sum_{i=1}^{k}\lambda i^{-\zeta} \leq 1 + 4\lambda\left(1 + \int_1^k x^{-\zeta}dx\right) = 1 + 4\lambda\left(1 + \frac{k^{1-\zeta}-1}{1-\zeta}\right).$$

Using the assumption $\zeta>1$, we can upper bound $k^{1-\zeta}$ by 1 and obtain

$$1 + 4\lambda\left(1 + \frac{k^{1-\zeta}-1}{1-\zeta}\right) \leq 1 + 4\lambda\left(1 + \frac{1}{\zeta-1}\right) = 1 + \frac{4\lambda\zeta}{\zeta-1}.$$

We thus obtain

$$\mathbb{P}\left\{E_1 \geq \frac{t}{3}\right\} \leq \sqrt{e}\exp\left\{-\frac{t^2 k}{C_3}\right\},$$

where $C_3>0$ is a constant depending on ζ and λ.

Control of the term E_2. This control is obtained with the same arguments used for E_1. Thus, we obtain, for $t > 0$,

$$\mathbb{P}\left\{E_2 \geq \frac{t}{3}\right\} \leq \sqrt{e}\exp\left\{-\frac{t^2 k}{C_3}\right\}.$$

Control of the term E_3. This bound is more technical. Before proceeding, we introduce some notations. For every $j \in \{1,\dots,d\}$, we define

$$\alpha_{n,m}^{(j)} = \left(\mathbb{P}_{n,m}^{(j)} - \mathbb{P}_m^{(j)}\right), \quad \beta_{n,m}^{(j)}(x) = \alpha_{n,m}^{(j)}(]-\infty, x]), \quad x \in \mathbb{R},$$

where $\mathbb{P}_{n,m}^{(j)}$ corresponds to the empirical measure for the sample $(M_{m,1}^{(j)},\dots,M_{m,k}^{(j)})$ and $\mathbb{P}_m^{(j)}$ is the law of the random variable $M_m^{(j)}$. To control the term E_3, we introduce chaining arguments as used in the proof of Proposition 7.1 of Rio 2017. Let be $j \in \{1,\dots,d\}$ fixed and N be some positive integer to be chosen later. For any real x such that $F_m^{(j)}(x) \neq 0$ and $F_m^{(j)}(x) \neq 1$, let us write $F_m^{(j)}(x)$ in base 2 :

$$F_m^{(j)}(x) = \sum_{l=1}^{N} b_l(x)2^{-l} + r_N(x), \text{ with } r_N(x) \in [0, 2^{-N}[$$

where $b_l = 0$ or $b_l = 1$. For any L in $[1,\dots,N]$, set

$$\Pi_L(x) = \sum_{l=1}^{L} b_l(x)2^{-l} \text{ and } i_L = \Pi_L(x)2^L.$$

Let the reals $(x_L)_L$ be chosen in such a way that $F_m^{(j)}(x_L) = \Pi_L(x)$. With these notations

$$\begin{aligned}\beta_{n,m}^{(j)}(x) =& \beta_{n,m}^{(j)}(\Pi_1(x)) + \beta_{n,m}^{(j)}(x) - \beta_{n,m}^{(j)}(\Pi_N(x)) \\ &+ \sum_{L=2}^{N}\left[\beta_{n,m}^{(j)}(\Pi_L(x)) - \beta_{n,m}^{(j)}(\Pi_{L-1}(x))\right].\end{aligned}$$

Let the reals $x_{L,i}$ be defined by $F_m^{(j)}(x_{L,i}) = i2^{-L}$. Using the above equality, we get that

$$\sup_{x\in\mathbb{R}}\left|\beta_{n,m}^{(j)}(x)\right| \leq \sum_{L=1}^{N}\Delta_L + \Delta_N^*,$$

with

$$\Delta_L = \sup_{i\in[1,2^L]}\left|\alpha_{n,m}^{(j)}(]x_{L,i-1}, x_{L,i}])\right| \text{ and } \Delta_N^* = \sup_{x\in\mathbb{R}}\left|\alpha_{n,m}^{(j)}(]\Pi_N(x), x])\right|.$$

From the inequalities

$$-2^{-N} \leq \alpha_{n,m}^{(j)}(]\Pi_N(x), x]) \leq \alpha_{n,m}^{(j)}(]\Pi_N(x), \Pi_N(x) + 2^{-N}]) + 2^{-N},$$

we get that

$$\Delta_N^* \leq \Delta_N + 2^{-N} \text{ and } \mathbb{E}\left[\sup_{x\in\mathbb{R}}|\beta_{n,m}^{(j)}(x)|\right] \leq 2\sum_{L=1}^{N}||\Delta_L||_1 + 2^{-N},$$

where $||\Delta_L||_1$ is the L^1-norm of Δ_L. Let N be the natural number such that $2^{N-1} < k \leq 2^N$. For this choice of N, we obtain

$$\mathbb{E}\left[\sup_{x\in\mathbb{R}}|\beta_{n,m}^{(j)}(x)|\right] \leq 2\sum_{L=1}^{N}||\Delta_L||_1 + k^{-1}.$$

Hence, using Rio 2017, Lemma 7.1 (where we divide by $\sqrt{k}$ the considering inequality in the lemma), we obtain that

$$\begin{aligned}\mathbb{E}\left[\sup_{x\in\mathbb{R}}|\beta_{n,m}^{(j)}(x)|\right] &\leq 2\frac{C_0}{\sqrt{k}}\sum_{L=1}^{N}\left(2^{-\frac{(\zeta-1)^2}{(4\zeta)^2}}\right)^L + k^{-1} \\ &\leq \frac{2}{\sqrt{k}}\frac{C_0}{1-2^{-\frac{(\zeta-1)^2}{(4\zeta)^2}}} + k^{-1} \triangleq C_1 k^{-1/2} + k^{-1},\end{aligned}$$

where C_0 and C_1 are constants depending on ζ and λ.

Now, fix $x \in \mathbb{R}$ and denote by $\Phi : \mathbb{R}^k \mapsto [0,1]$, the function defined by

$$\Phi(x_1,\ldots,x_k) = \sup_{x\in\mathbb{R}}\left|\frac{1}{k}\sum_{i=1}^{k}\mathbb{1}_{\{x_i\leq x\}} - F_m^{(j)}(x)\right|.$$

For $\mathbf{x},\mathbf{y} \in \mathbb{R}^k$, we obtain with some calculations:

$$|\Phi(\mathbf{x}) - \Phi(\mathbf{y})| \leq \sup_{x\in\mathbb{R}}\frac{1}{k}\sum_{i=1}^{k}\left|\mathbb{1}_{\{x_i\leq x\}} - \mathbb{1}_{\{y_i\leq x\}}\right| \leq \frac{1}{k}\sum_{i=1}^{k}\mathbb{1}_{\{x_i\neq y_i\}}.$$

Thus, Φ is k^{-1}-Lipschitz with respect to the Hamming distance. Under algebraically φ-mixing process, we may apply Mohri and Rostamizadeh 2010, Theorem 8 with $(M_{m,1}^{(j)},\ldots,M_{m,k}^{(j)})$, we obtain with probability at least $1-\exp\{-t^2k/||\Delta_k||_\infty^2\}$ where $||\Delta_k||_\infty \leq 1 + 4\sum_{i=1}^{k}\varphi(i)$

$$\sup_{x\in\mathbb{R}}\left|\hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x)\right| \leq \mathbb{E}\left[\sup_{x\in\mathbb{R}}\left|\beta_{n,m}^{(j)}(x)\right|\right] + \frac{t}{3} \leq C_1k^{-1/2} + C_2k^{-1} + \frac{t}{3}.$$

Thus, for a sufficiently large C_3, with probability at most $\exp\{-t^2k/C_3\}$

$$\sup_{x\in\mathbb{R}}\left|\hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x)\right| \geq C_1k^{-1/2} + k^{-1} + \frac{t}{3}.$$

Using Bonferroni inequality

$$\mathbb{P}\left\{E_3 \geq \frac{t}{3}\right\} \leq d\mathbb{P}\left\{\sup_{x\in\mathbb{R}}\left|\hat{F}_{n,m}^{(j)}(x) - F_m^{(j)}(x)\right| \geq t\right\},$$

we thus obtain a control bound for E_3. Assembling all the controls obtained for E_1, E_2 and E_3, we obtain the desired result.

□

The proof of Theorem 2 needs the following results : (1) an upper bound over the quantity $|\hat{\theta}_{n,m}(a,b) - \theta_m(a,b)|$ with respect to $|\hat{\nu}_{n,m}(a,b) - \nu_m(a,b)|$ to use the concentration inequality introduced in Proposition 2, (2) exhibit an event such that $\{\hat{O} = \bar{O}\}$. Lemmas 1 and 2 below address these two questions. Then, taking benefits of these results, we show that the probability of the exhibited event such that $\{\hat{O} = \bar{O}\}$ holds with high probability, as stated in Theorem 2.

Lemma 1. *Consider a pair $(a,b) \in \{1,\dots,d\}^2$, the following inequality holds:*

$$|\hat{\theta}_{n,m}(a,b) - \theta_m(a,b)| \le 9|\hat{\nu}_{n,m}(a,b) - \nu_m(a,b)|.$$

Proof of Lemma 1 We may write the respective quantities as $\theta = f(\nu(a,b))$ and $\hat{\theta}_{n,m} = f(\hat{\nu}_{n,m}(a,b))$ where f is a function defined as follows,

$$\begin{array}{rccc} f & : & [0,1/6] & \to & [1,2] \\ & & x & \mapsto & \frac{1/2+x}{1/2-x}, \end{array}$$

with $f(x) \in [1,2]$ by definition of the sub-asymptotic extremal coefficient θ_m. The domain of this function is restricted to the interval $[0,1/6]$ because we have $f(x) \le 2$, or

$$x + \frac{1}{2} \le 1 - 2\,x,$$

which holds if $x \le 1/6$. The inequality $f(x) \ge 1$ gives the positivity of the domain. In particular, $x < 1/2$ and thus $2^{-1} - x \ge 3^{-1} > 0$. Taking derivative of f, we find that

$$|f'(x)| = \frac{1}{(1/2-x)^2} \le 3^2, \quad x \in [0,1/6]\,.$$

Therefore, f is 9-Lipschitz continuous and we have

$$|\hat{\theta}_{n,m}(a,b) - \theta_m(a,b)| = |f(\hat{\nu}_{n,m}(a,b)) - f(\nu_m(a,b))| \le 9|\hat{\nu}_{n,m}(a,b) - \nu_m(a,b)|.$$

This completes the proof. □

Lemma 2. *Consider the AI-block model in Definition 1. Define*

$$\kappa = \sup_{a,b\in\{1,\dots,d\}} |\hat{\chi}_{n,m}(a,b) - \chi(a,b)|.$$

Consider parameters (τ,η) fulfilling

$$\tau \ge \kappa, \quad \eta \ge \kappa + \tau. \tag{22}$$

If $\mathrm{MECO}(\mathcal{X}) > \eta$*, then Algorithm (ECO) yields $\hat{O} = \bar{O}$.*

Proof of Lemma 2 If $a \overset{\bar{O}}{\not\sim} b$, then $\chi(a,b) = 0$ and

$$\hat{\chi}_{n,m}(a,b) = \hat{\chi}_{n,m}(a,b) - \chi(a,b) \le \kappa \le \tau.$$

Now, if $a \overset{\bar{O}}{\sim} b$, if $\mathcal{X} \in \mathbb{X}(\eta)$ then $\chi(a,b) > \kappa + \tau$ and

$$\kappa + \tau < \chi(a,b) - \hat{\chi}_{n,m}(a,b) + \hat{\chi}_{n,m}(a,b),$$

and thus $\hat{\chi}_{n,m}(a,b) > \tau$. In particular, under (22) and the separation condition $\mathrm{MECO}(\mathcal{X}) > \eta$, we have

$$a \overset{\bar{O}}{\sim} b \quad \Longleftrightarrow \quad \hat{\chi}_{n,m}(a,b) > \tau. \tag{23}$$

Let us prove the lemma by induction on the algorithm step l. We consider the algorithm at some step $l-1$ and assume that the algorithm was consistent up to this step, i.e. $\hat{O}_j = \bar{O}_j$ for $j = 1,\dots,l-1$.

If $\hat{\chi}_{n,m}(a_l,b_l) \le \tau$, then according to (23), no $b \in S$ is in the same group of a_l. Since the algorithm has been consistent up to this step l, it means that a_l is a singleton and $\hat{O}_l = \{a_l\}$.

If $\hat{\chi}_{n,m}(a_l,b_l) > \tau$, then $a_l \overset{\bar{O}}{\sim} b$ according to (23). Furthermore, the equivalence implies that $\hat{O}_l = S \cap \bar{O}_l$. Since the algorithm has been consistent up to this step, we have $\hat{O}_l = \bar{O}_l$. To conclude, the algorithm remains consistent at the step l and the result follows by induction.

□

Proof of Theorem 2 We have that for $t > 0$:

$$\mathbb{P}\left\{\sup_{a,b\in\{1,\dots,d\}} |\hat{\theta}_{n,m}(a,b) - \theta_m(a,b)| \geq t\right\} \leq d^2\mathbb{P}\left\{|\hat{\theta}_{n,m}(a,b) - \theta_m(a,b)| \geq t\right\}.$$

With probability at least $1 - 2(1+\sqrt{e})d^2\exp\{-t^2k/C_3\}$, using Proposition 2 and Lemma 1, one has

$$\sup_{a,b\in\{1,\dots,d\}} \left|\hat{\theta}_{n,m}(a,b) - \theta(a,b)\right| \leq d_m + C_1k^{-1/2} + C_2k^{-1} + t,$$

By considering $\delta \in]0,1[$ and solve the following equation

$$\frac{\delta}{d^2} = 2(1+\sqrt{e})\exp\left\{-\frac{kt^2}{C_3}\right\},$$

with respect to t gives that the event

$$\sup_{a,b\in\{1,\dots,d\}} \left|\hat{\theta}_{n,m}(a,b) - \theta(a,b)\right| \geq d_m + C_1k^{-1/2} + C_2k^{-1} + C_3\sqrt{\frac{1}{k}\ln\left(\frac{2(1+\sqrt{e})d^2}{\delta}\right)},$$

is of probability at most δ. Now, taking $\delta = 2(1+\sqrt{e})d^{-2\gamma}$, with $\gamma > 0$, we have

$$\sup_{a,b\in\{1,\dots,d\}} \left|\hat{\theta}_{n,m}(a,b) - \theta(a,b)\right| \leq d_m + C_1k^{-1/2} + C_2k^{-1} + C_3\sqrt{\frac{(1+\gamma)\ln(d)}{k}},$$

with probability at least $1 - 2(1+\sqrt{e})d^{-2\gamma}$ for C_3 sufficiently large. The result then follows from Lemma 2 along with Condition $\mathcal{B}$ and algebraically φ-mixing random process, since

$$\mathbb{P}\left\{\kappa \leq d_m + C_1k^{-1/2} + C_2k^{-1} + C_3\sqrt{\frac{(1+\gamma)\ln(d)}{k}}\right\} \geq 1 - 2(1+\sqrt{e})d^{-2\gamma},$$

and $\mathrm{MECO}(\mathcal{X}) > \eta$ by assumption. □

Therein, we prove the argument that were stated without proof in the paragraph next to Theorem 2. A condition of order two were introduced and we have state that $d_m = O(\Psi_m)$ can be shown. We propose a proof of this statement below.

Proof of $d_m = O(\Psi(m))$ Take $a \neq b$ fixed, we have, using Lemma 1

$$|\chi_m(a,b) - \chi(a,b)| = |\theta_m(a,b) - \theta(a,b)| \leq 9\,|\nu_m(a,b) - \nu(a,b)|\,,$$

where $\nu_m(a,b)$ (resp. $\nu(a,b)$) is the madogram computed between $M_m^{(a)}$ and $M_m^{(b)}$ (resp. between $X^{(a)}$ and $X^{(b)}$) and we use Lemma 1 to obtain the inequality. Using the results of Lemma 1 of Marcon et al. 2017, we have

$$\begin{aligned}\nu_m(a,b) - \nu(a,b) = \frac{1}{2}\left(\int_{[0,1]}(C_m - C_\infty)(\mathbf{1}, x^{(a)}, \mathbf{1})dx^{(a)} + \int_{[0,1]}(C_m - C_\infty)(\mathbf{1}, x^{(b)}, \mathbf{1})dx^{(b)}\right)\\ - \int_{[0,1]}(C_m - C_\infty)(1,\dots,\underbrace{x}_{a\text{th index}},1,\dots,1,\underbrace{x}_{b\text{th index}},\dots,1)dx,\end{aligned}$$

where the integration is taken respectively for the a-th, b-th and a,b-th components. Hence

$$|\nu_m(a,b)-\nu(a,b)| \le \frac{1}{2}\left(\int_{[0,1]}|(C_m-C_\infty)(\mathbf{1},x^{(a)},\mathbf{1})|dx^{(a)}+\int_{[0,1]}|(C_m-C_\infty)(\mathbf{1},x^{(b)},\mathbf{1})|dx^{(b)}\right)$$
$$+\int_{[0,1]}|(C_m-C_\infty)(1,\dots,\underbrace{x}_{a\text{th index}},1,\dots,1,\underbrace{x}_{b\text{th index}},\dots,1)|dx.$$

Using the second order condition in Equation (9) we obtain that $|C_m - C_\infty|(\mathbf{u}) = O(\Psi_m)$, uniformly in $\mathbf{u}\in[0,1]^d$. Hence the statement. □

Now, we prove the theoretical result giving support to our cross validation process.

Proof of Proposition 3 Using triangle inequality several times, we may obtain the following bound

$$\begin{aligned}\widehat{\mathrm{SECO}}_{n,m}(\bar{O})-\widehat{\mathrm{SECO}}_{n,m}(\hat{O}) &\le 2D_m+|\widehat{\mathrm{SECO}}_{n,m}(\bar{O})-\mathrm{SECO}_m(\bar{O})|\\ &\quad+|\widehat{\mathrm{SECO}}_{n,m}(\hat{O})-\mathrm{SECO}_m(\hat{O})|+\mathrm{SECO}(\bar{O})-\mathrm{SECO}(\hat{O})\\ &=: 2D_m+E_1+E_2+\mathrm{SECO}(\bar{O})-\mathrm{SECO}(\hat{O}).\end{aligned}$$

Taking expectancy, we now have

$$\mathbb{E}[\widehat{\mathrm{SECO}}_{n,m}(\bar{O})-\widehat{\mathrm{SECO}}_{n,m}(\hat{O})] \le 2D_m+\mathbb{E}[E_1]+\mathbb{E}[E_2]+\mathrm{SECO}(\bar{O})-\mathrm{SECO}(\hat{O}).$$

Using the same tool involved in the proof of Lemma 1, we can show

$$|\hat{\theta}_{n,m}^{(\bar{O}_g)}-\hat{\theta}_m^{(\bar{O}_g)}| \le (d_g+1)^2|\hat{\nu}_{n,m}^{(\bar{O}_g)}-\hat{\nu}_m^{(\bar{O}_g)}|,$$

Thus, using concentration bounds in Proposition 2, there exists a universal constant $K_1>0$ independent of n,k,m,t such that

$$\mathbb{P}\left\{|\hat{\theta}_{n,m}^{(\bar{O}_g)}-\hat{\theta}_m^{(\bar{O}_g)}|\ge t\right\} \le d_g\exp\left\{-\frac{t^2k}{K_1d_g^4}\right\}.$$

Now,

$$\begin{aligned}\mathbb{P}\left\{|\widehat{\mathrm{SECO}}_{n,m}(\bar{O})-\mathrm{SECO}_m(\bar{O})|\ge t\right\} &\le \sum_{g=1}^{G}\mathbb{P}\left\{|\hat{\theta}_{n,m}^{(\bar{O}_g)}-\hat{\theta}_m^{(\bar{O}_g)}|\ge\frac{t}{G}\right\}\\ &\le d\exp\left\{-\frac{t^2k}{K_1G^2\vee_{g=1}^{G}d_g^4}\right\}\end{aligned}$$

Thus, for every $\delta>0$, one obtains

$$\mathbb{E}[E_1]^2 \le \mathbb{E}[E_1^2] \le \delta+\int_\delta^\infty \mathbb{P}\left\{E_1>t^{1/2}\right\}dt \le \delta+d\int_\delta^\infty\exp\left\{-\frac{t}{2\sigma^2}\right\}dt,$$

where $\sigma^2=\frac{K_1G^2\vee_{g=1}^{G}d_g^4}{2k}$. Set $\delta=2\sigma^2\ln(d)$, we can obtain

$$\mathbb{E}[E_1]^2 \leq \delta + 2\sigma^2 = c^2 \frac{\ln(d) G^2 \vee_{g=1}^{G} d_g^4}{k}$$

with $c > 0$. Same results hold for $\mathbb{E}[E_2]$ with corresponding sizes, thus

$$\mathbb{E}[\widehat{\mathrm{SECO}}_{n,m}(\bar{O}) - \widehat{\mathrm{SECO}}_{n,m}(\hat{O})] \leq 2 \left(D_m + c\sqrt{\frac{\ln(d)}{k}} \max(G, I) \max(\vee_{g=1}^{G} d_g^2, \vee_{i=1}^{I} d_i^2) \right) + \mathrm{SECO}(\bar{O}) - \mathrm{SECO}(\hat{O}),$$

which is strictly negative by assumption. □

Appendix C.3 Proofs of Section 4

In the following we prove that the model introduced in Section 4 is in the domain of attraction of an AI-block model. This comes down from some elementary algebra where the fundamental argument is given by Bücher and Segers 2014, Proposition 4.2, from which the inspiration for the model was drawn thereof.

Proof of Proposition 4 We aim to show that the following quantity

$$\left| C_{\theta,\beta_0} \left(C_{\theta,\beta_1}^{(O_1)}(\{\mathbf{u}^{(O_1)}\}^{1/m}), \ldots, C_{\theta,\beta_G}^{(O_G)}(\{\mathbf{u}^{(O_G)}\}^{1/m}) \right)^m - C_{0,\beta_0} \left(C_{0,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)}), \ldots, C_{0,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)}) \right) \right|,$$

converges to 0 uniformly in $\mathbf{u} \in [0,1]^d$. Using Equation (14) in the main article, the latter term is equal to

$$E_{0,m} := \left| C_{\theta,\beta_0} \left(C_{\theta/m,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)})^{1/m}, \ldots, C_{\theta/m,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)})^{1/m} \right)^m - C_{0,\beta_0} \left(C_{0,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)}), \ldots, C_{0,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)}) \right) \right|.$$

Thus

$$\begin{aligned} E_{0,m} \leq & \left| C_{\theta,\beta_0} \left(C_{\theta/m,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)})^{1/m}, \ldots, C_{\theta/m,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)})^{1/m} \right)^m - C_{0,\beta_0} \left(C_{\theta/m,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)}), \ldots, C_{\theta/m,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)}) \right) \right| \\ & + \left| C_{0,\beta_0} \left(C_{\theta/m,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)}), \ldots, C_{\theta/m,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)}) \right) - C_{0,\beta_0} \left(C_{0,\beta_1}^{(O_1)}(\mathbf{u}^{(O_1)}), \ldots, C_{0,\beta_G}^{(O_G)}(\mathbf{u}^{(O_G)}) \right) \right| \\ & =: E_{1,m} + E_{2,m}. \end{aligned}$$

As $C_{\theta/m,\beta_0}$ converges uniformly to C_{0,β_0}, then, uniformly in $\mathbf{u} \in [0,1]^d$, $E_{1,m} \underset{m\to\infty}{\longrightarrow} 0$. Now, using Lipschitz property of the copula function, one has

$$E_{2,m} \leq \sum_{g=1}^{G} \left| C_{\theta/m,\beta_g}^{(O_g)}(\mathbf{u}^{(O_g)}) - C_{0,\beta_g}^{(O_g)}(\mathbf{u}^{(O_g)}) \right|,$$

which converges almost surely to 0 as $m \to \infty$. The limiting copula is an extreme value copula by $\beta_0 \leq \min\{\beta_1, \ldots, \beta_G\}$, see Example 3.8 of Hofert, Huser, and Prasad 2018. Hence the result. □

Appendix D. Additional results

Appendix D.1 Additional results of Section 2

Let $\mathbf{Z} \geq 0$ be a random vector, and for simplicity, let's assume that it has heavy-tailed marginal distributions with a common tail-index $\alpha > 0$. There are two distinct yet closely related classical approaches for describing the extreme values of the multivariate distribution of $\mathbf{Z}$.

The first approach focuses on scale-normalized componentwise maxima:

$$c_n^{-1} \bigvee_{i=1}^{n} \mathbf{Z}_i,$$

where $\mathbf{Z}_i$ are independent copies of $\mathbf{Z}$, and c_n is a scaling sequence. The limiting results are typically derived under the assumption of independence for the sake of consistency. However, they hold under more general conditions, such as mixing conditions (see, e.g., Hsing 1989). The only possible limit laws for such maxima are max-stable distributions with the following distribution function:

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigvee_{i=1}^{n} \mathbf{Z}_i \leq c_n \mathbf{u} \right\} = e^{-\Lambda([0,\mathbf{u}]^c)}, \quad \mathbf{u} \in \mathbb{R}^d + \setminus \mathbf{0},$$

where the exponent measure Λ is $(-\alpha)$-homogeneous.

The second approach examines the distribution of scale-normalized exceedances:

$$u^{-1}\, \mathbf{Z} \,|\, \bigvee_{j=1}^{d} Z^{(j)} > u,$$

which considers conditioning on the event that at least one component $Z^{(j)}$ exceeds a high threshold u. The only possible limits of these peak-over-thresholds as $u \to \infty$ are multivariate Pareto distributions (Rootzén and Tajvidi 2006). The probability laws of these distributions are induced by a homogeneous measure Λ on the set $\mathcal{L} = E \setminus [0,1]^d$, where $E = [0,\infty)^d \setminus \mathbf{0}$. The probability measure takes the form:

$$\mathbb{P}_{\mathcal{L}}(dy) = \frac{\Lambda(dy)}{\Lambda(\mathcal{L})}.$$

The exponent measure serves as a clear connection between these two approaches, as it characterizes the distribution function for both cases. In fact, the connection arises from a fundamental limiting result that establishes a link between the two approaches through regular variation. This result has been elegantly presented in Theorem 2.1.6 and Equation (2.3.1) in Kulik and Soulier 2020. The following proposition provides the form of the exponent measure when the random vectors $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ are independent, and it establishes the connection between AI-block models for the two approaches.

Proposition 6. *Suppose* $\mathbf{X}$ *is an extreme-value random vector with exponent measure* Λ *concentrating on* $E \setminus [0, \mathbf{x}]$ *where* $E = [0,\infty]^d \setminus \{\boldsymbol{0}\}$. *The following properties are equivalent:*

(i) The vectors $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ *are independent.*
(ii) The vectors are blockwise independent: for every $1 \leq g < h \leq G$

$$\mathbf{X}^{(O_g)} \text{ and } \mathbf{X}^{(O_h)}, \text{ are independent random vectors.}$$

(iii) The exponent measure Λ *concentrates on*

$$\bigcup_{g=1}^{G} \{\mathbf{0}\}^{d_1} \times \cdots \times]0, \infty[^{d_g} \times \cdots \times \{\mathbf{0}\}^{d_G}, \tag{24}$$

so that for $\mathbf{y} > \mathbf{0}$*,*

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

.

These conditions generalize straightforwardly those stated in Proposition 5.24 of Resnick 2008 (see Exercise 5.5.1 of the book aforementioned or the Lemma in Strokorb 2020).

Proof of Proposition 6

We will establish the result proceeding as *(iii)* $\implies$ *(i)* $\implies$ *(ii)* $\implies$ *(iii)* where we directly have *(i)* $\implies$ *(ii)*. Now for *(iii)* $\implies$ *(i)*, suppose Λ concentrates on the set (24). Then for $\mathbf{x} > 0$, noting $A_g(\mathbf{x}) = \{\mathbf{u} \in E, \exists a \in O_g, u_a > x_a\}$ for $g \in \{1, \ldots, G\}$, we obtain

$$\begin{aligned} -\ln H(\mathbf{x}) = \Lambda(E \setminus [0, \mathbf{x}]) &= \Lambda\left(\bigcup_{g=1}^{G} A_g(\mathbf{x}) \right) \\ &= \sum_{g=1}^{G} \Lambda(A_g(\mathbf{x})) + \sum_{g=2}^{G} (-1)^{g+1} \sum_{1 \leq i_1 < i_2 < \cdots < i_l \leq G} \Lambda(A_{i_1}(\mathbf{x}) \cap \cdots \cap A_{i_l}(\mathbf{x})), \end{aligned}$$

so that because of Equation (24) in the main paper,

$$-\ln H(\mathbf{x}) = \sum_{g=1}^{G} \Lambda(A_g(\mathbf{x})),$$

and we have $H(\mathbf{x}) = \Pi_{g=1}^{G} \exp\{-\Lambda(\{u \in E, \exists a \in O_g, u_a > x_a\})\} = \Pi_{g=1}^{G} H^{(O_g)}(\mathbf{x}^{(O_g)})$.

Thus H is a written as a product of the G distributions corresponding to random vectors $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$, as desired.

It remains to show *(ii)* $\implies$ *(iii)*. Set $Q^{(O_k)}(\mathbf{y}^{(O_k)}) = -\ln \mathbb{P}\{\mathbf{X}^{(O_k)} \leq \mathbf{y}^{(O_g)}\}$ for $k \in \{1, \ldots, G\}$. We have for $\mathbf{y} > 0$ that blockwise independence implies, with $k \neq l$,

$$Q^{(O_k)}(\mathbf{y}^{(O_k)}) + Q^{(O_l)}(\mathbf{y}^{(O_l)}) = -\ln \mathbb{P}\{\mathbf{X}^{(O_k)} \leq \mathbf{y}^{(O_k)}, \mathbf{X}^{(O_l)} \leq \mathbf{y}^{(O_l)}\}.$$

Since $H(\mathbf{x}) = \exp\{-\Lambda(E \setminus [\mathbf{0}, \mathbf{x}])\}$ for $\mathbf{x} > 0$, we have

$$\begin{aligned} Q^{(O_k)}(\mathbf{y}^{(O_k)}) + Q^{(O_l)}(\mathbf{y}^{(O_l)}) &= \Lambda(\{\mathbf{x}, \exists a \in O_k, x_a > y_a\} \cup \{\mathbf{x}, \exists b \in O_l, x_b > y_b\}) \\ &= \Lambda(\{\mathbf{x}, \exists a \in O_k, x_a > y_a\}) + \Lambda(\{\mathbf{x}, \exists b \in O_l, x_a > y_a\}) \\ &\quad - \Lambda(\{\mathbf{x}, \exists a \in O_k, x_a > y_a, \exists b \in O_l, x_b > y_b\}) \\ &= Q^{(O_k)}(\mathbf{y}^{(O_k)}) + Q^{(O_l)}(\mathbf{y}^{(O_l)}) \\ &\quad - \Lambda(\{\mathbf{x}, \exists a \in O_k, x_a > y_a, \exists b \in O_l, x_b > y_b\}), \end{aligned}$$

and thus

$$\Lambda(\{\mathbf{x}, \exists a \in O_k, x_a > y_a, \exists b \in O_l, x_b > y_b\}) = 0,$$

so that (iii) holds. This is equivalent to Λ concentrates on the set in Equation (24) in the main paper.

□

If $\mathbf{X}$ is an extreme value random vector, it is associated with a stable tail dependence function denoted by L. This function captures the tail dependence structure of the random vector and can be expressed as a specific integral with respect to the exponent measure (we refer to Section 8 of Beirlant et al. 2004). In the context of AI-block models, the tail dependence function takes the following form:

$$L\left(z^{(1)},\ldots,z^{(d)}\right)=\sum_{g=1}^{G}L^{(O_g)}\left(\mathbf{z}^{(O_g)}\right),\quad \mathbf{z}\in[0,\infty)^d, \tag{25}$$

where $L^{(O_1)},\ldots,L^{(O_G)}$ are the corresponding stable tail dependence functions with copulae $C_\infty^{(O_1)},\ldots,C_\infty^{(O_G)}$, respectively. This model is a specific form of the nested extreme value copula, as mentioned in the remark below and discussed in further detail in Hofert, Huser, and Prasad 2018.

Remark 2. Equation (25) can be rewritten as

$$L(\mathbf{z})=L_\Pi\left(L^{(O_1)}\left(z^{(O_1)}\right),\ldots,L^{(O_G)}\left(z^{(O_G)}\right)\right),$$

where $L_\Pi(z^{(1)},\ldots,z^{(G)})=\sum_{g=1}^{G}z^{(g)}$ is a stable tail dependence function corresponding to asymptotic independence. According to Proposition 1, C_∞ is an extreme value copula. Therefore, it follows that C_∞, which has the representation

$$C_\infty(\mathbf{u})=C_\Pi\left(C_\infty^{(O_1)}(\mathbf{u}^{(O_1)}),\ldots,C_\infty^{(O_G)}(\mathbf{u}^{(O_G)})\right),\quad C_\Pi=\Pi_{g=1}^{G}u^{(g)},$$

is also a nested extreme value copula, as defined in Hofert, Huser, and Prasad 2018.

Equation (25) can be restricted to the simplex, allowing us to express the stable tail dependence function in terms of the Pickands dependence function. Specifically, the Pickands dependence function A can be written as a convex combination of the Pickands dependence functions $A^{(O_1)},\ldots,A^{(O_G)}$ as follows:

$$\begin{aligned}A(t^{(1)},\ldots,t^{(d)})&=\frac{1}{z^{(1)}+\cdots+z^{(d)}}\left[\sum_{g=1}^{G}(z^{(i_{g,1})}+\cdots+z^{(i_{g,d_g})})A^{(O_g)}(\mathbf{t}^{(O_g)})\right]\\&=\sum_{g=1}^{G}w^{(O_g)}(\mathbf{t})A^{(O_g)}(\mathbf{t}^{(O_g)})=:A^{(O)}(t^{(1)},\ldots,t^{(d)}),\end{aligned} \tag{26}$$

with $t^{(j)}=z^{(j)}/(z^{(1)}+\cdots+z^{(d)})$ for $j\in\{2,\ldots,d\}$ and $t^{(1)}=1-(t^{(2)}+\cdots+t^{(d)})$, $w^{(O_g)}(\mathbf{t})=(z^{(i_{g,1})}+\cdots+z^{(i_{g,d_g})})/(z^{(1)}+\cdots+z^{(d)})$ for $g\in\{2,\ldots,G\}$ and $w^{(O_1)}(\mathbf{t})=1-(w^{(O_2)}(\mathbf{t})+\cdots+w^{(O_G)}(\mathbf{t}))$, $\mathbf{t}^{(O_g)}=(t^{(i_{g,1})},\ldots,t^{(i_{g,d_g})})$ where $t^{(i_{g,\ell})}=z^{(i_{g,\ell})}/(z^{(i_{g,1})}+\cdots+z^{(i_{g,d_g})})$ and $(i_{g,\ell})$ designates the ℓth variable in the gth cluster for $\ell\in\{1,\ldots,d_g\}$ and $g\in\{1,\ldots,G\}$. As a convex combination of Pickands dependence functions, A is itself a Pickands dependence function (see Falk, Hüsler, and Reiss 2010, Page 123).

In the context of independence between extreme random variables, it is well-known that the inequality $A(\mathbf{t})\le 1$ holds for $\mathbf{t}\in\Delta_{d-1}$, where A is the Pickands dependence function and equality stands if and only if the random variables are independent. This result extends to the case of random vectors, with the former case being a special case where $d_1=\cdots=d_G=1$.

Proposition 7. *Consider an extreme value random vector* $\mathbf{X} \in \mathbb{R}^d$ *with Pickands dependence function* A*. Let* $A^{(O)}$ *be as defined in* (26)*. For all* $\mathbf{t} \in \Delta_{d-1}$*, we have:*

$$\left(A^{(O)} - A\right)(\mathbf{t}) \geq 0,$$

with equality if and only if $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_G)}$ *are independent.*

We provide two methods for establishing this result: the first leverages the convexity and homogeneity of order one of the stable tail dependence function, while the second takes advantage of the associativity of extreme-value random vectors.

Proof of Proposition 7 For the first method, the stable tail dependence function L is subadditive as an homogeneous convex function under a cone, i.e.,

$$L(\mathbf{x}+\mathbf{y}) \leq L(\mathbf{x}) + L(\mathbf{y}),$$

for every $\mathbf{x}, \mathbf{y} \in [0,\infty)^d$. In particular, we obtain by induction on G

$$L\left(\sum_{g=1}^{G} \mathbf{x}^{(g)}\right) \leq \sum_{g=1}^{G} L(\mathbf{x}^{(g)}),$$

where $\mathbf{x}^{(g)} \in [0,\infty)^d$ and $g \in \{1,\ldots,G\}$. Consider now $\mathbf{z}^{(O_g)} = (\mathbf{0}, z^{(i_{g,1})}, \ldots, z^{(i_{g,d_g})}, \mathbf{0})$, we directly obtain using the equation above

$$L(\mathbf{z}) = L\left(\sum_{g=1}^{G} \mathbf{z}^{(O_g)}\right) \leq \sum_{g=1}^{G} L(\mathbf{z}^{(O_g)}) = \sum_{g=1}^{G} L^{(O_g)}(z^{(i_{g,1})}, \ldots, z^{(i_{g,d_g})}).$$

Translating the above inequality in terms of Pickands dependence function results on

$$\begin{aligned} A(\mathbf{t}) &\leq \sum_{g=1}^{G} \frac{1}{z^{(1)}+\cdots+z^{(d)}} L^{(O_g)}(z^{(i_{g,1})}, \ldots, z^{(i_{g,d_g})}) \\ &= \sum_{g=1}^{G} \frac{z^{(i_{g,1})}+\cdots+z^{(i_{g,d_g})}}{z^{(1)}+\cdots+z^{(d)}} A^{(O_g)}(t^{(i_{g,1})}, \ldots, t^{(i_{g,d_g})}), \end{aligned}$$

where $t^{(i)} = z^{(i)}/(z^{(1)}+\cdots+z^{(d)})$. Hence the result.

We can also prove this result by using the associativity of extreme-value distributions (see Marshall and Olkin 1983, Proposition 5.1 or Resnick 2008, Section 5.4.1), i.e.,

$$\mathbb{E}\left[f(\mathbf{X})g(\mathbf{X})\right] \geq \mathbb{E}\left[f(\mathbf{X})\right]\mathbb{E}\left[g(\mathbf{X})\right],$$

for every increasing (or decreasing) functions f, g. By induction on $G \in \mathbb{N}_*$,

$$\mathbb{E}\left[\Pi_{g=1}^{G} f^{(g)}(\mathbf{X})\right] \geq \Pi_{g=1}^{G} \mathbb{E}\left[f^{(g)}(\mathbf{X})\right]. \tag{27}$$

Take $f^{(g)}(\mathbf{x}) = \mathbb{1}_{\{]-\infty,\mathbf{x}^{(O_g)}]\}}$ for each $g \in \{1,\ldots G\}$, thus Equation (27) gives

$$C(H^{(1)}(x^{(1)}), \ldots, H^{(d)}(x^{(d)})) \geq \Pi_{g=1}^{G} C^{(O_g)}\left(H^{(O_g)}\left(\mathbf{x}^{(O_g)}\right)\right),$$

which can be restated in terms of stable tail dependence function as

$$L(\mathbf{z}) \leq \sum_{g=1}^{G} L^{(O_g)}(\mathbf{z}^{(O_g)}).$$

We obtain the statement expressing this inequality with Pickands dependence function. Finally, notice that (27) with $f^{(g)}(\mathbf{x}) = \mathbb{1}_{\{]-\infty,\mathbf{x}^{(O_g)}]\}}$ for each $g \in \{1, \dots G\}$ holds as an equality if and only if $\mathbf{X}^{(O_1)}, \dots, \mathbf{X}^{(O_G)}$ are independent random vectors. □

In the following paragraph, we give another proof of the extension of the results found in Takahashi 1987, 1994 made by Ferreira 2011, Proposition 2.1. Before going into details, we recall some useful expression of the dependence structure of extreme closely related to the notion of regular variation.

Let $\mathbf{X}$ be a regularly varying random vector in $\mathbb{R}_+^d$ with exponent measure Λ which is $(-\alpha)$-homogeneous, i.e. for $y > 0$ and A separated from $\mathbf{0}$, that is there exists an open set U such that $\mathbf{0} \in U$ and $U^c \subset A$, we have

$$\Lambda(yA) = y^{-\alpha}\Lambda(A).$$

Using the homogeneity of the exponent measure, we may define a probability measure Φ on $\Theta = S_d \cap [\mathbf{0}, \boldsymbol{\infty})$ where $S_d = \{\mathbf{x} \in \mathbb{R}^d, ||\mathbf{x}||^{-1}\mathbf{x}\}$ called the spectral measure associated to the norm $||\cdot||$ and defined by

$$\Phi(B) = \Lambda\left(\mathbf{z} \in E : ||\mathbf{z}|| > 1, \mathbf{z}||\mathbf{z}||^{-1} \in B\right)$$

for any Borel subset B of Θ (for a proper introduction to these notions, see Resnick 2008, Section 5.1 or Kulik and Soulier 2020, Section 2.2). The measure Φ is called the spectral measure. It is uniquely determined by the exponent measure Λ and the chosen norm. The homogeneity of Λ implies :

$$\Lambda\left(\mathbf{z} \in E : ||\mathbf{z}|| > r, \mathbf{z}||\mathbf{z}||^{-1} \in B\right) = r^{-1}\Phi(B),$$

for $0 < r < \infty$.

Proposition 8. *Let $\mathbf{X}$ be a regularly varying random vector in $\mathbb{R}_+^d$ with exponent measure Λ. Consider $O = \{O_1, \dots, O_g\}$ be a partition of $\{1, \dots, d\}$, then the following are equivalent:*

(i) Let $\Lambda^{(O_g)}$ be the restriction of the exponent measure to $\mathbb{R}_+^{(O_g)}$, we have

$$\Lambda = \sum_{g=1}^{G} \delta_0 \otimes \cdots \otimes \Lambda^{(O_g)} \otimes \cdots \otimes \delta_0.$$

(ii) The spectral measure Φ associated to the exponent measure Λ verifies

$$\Phi = \sum_{g=1}^{G} \delta_0 \otimes \cdots \otimes \Phi^{(O_g)} \otimes \cdots \otimes \delta_0 =: \Phi_\Pi, \tag{28}$$

where $\Phi^{(O_g)}$ is the restriction of Φ to $\Theta^{(O_g)} = S_{d_g}^{(O_g)} \cap [\mathbf{0}, \boldsymbol{\infty})$ with

$$S_{d_g}^{(O_g)} = \left\{\mathbf{x}^{(O_g)} \in \mathbb{R}^{(O_g)}, \mathbf{x}^{(O_g)}||\mathbf{x}^{(O_g)}||^{-1}\right\}$$

for $g \in \{1, \dots, G\}$.

(iii) There exists a $\mathbf{v} \in [0,\infty)^d$ *such that*

$$\int_\Theta \bigvee_{j=1}^d w^{(j)} v^{(j)} \Phi(d\mathbf{w}) = \sum_{g=1}^G \int_{\Theta^{(O_g)}} \bigvee_{j\in O_g} w^{(j)} v^{(j)} \Phi^{(O_g)}(d\mathbf{w}^{(O_g)}). \tag{29}$$

Proof of Proposition 8 The equivalence between (i) and (ii) falls down from definitions. The implication (ii) $\implies$ (iii) is trivial. We show now (iii) $\implies$ (ii) Notice that for every Borel set B of S_d, we have

$$\Phi(B) = \sum_{g=1}^G \Phi(B\cap\Theta^{(O_g)}) + \Phi\left(B\cap(\Theta\setminus\cup_{g=1}^G\Theta^{(O_g)})\right) \geq \sum_{g=1}^G \Phi(B\cap\Theta^{(O_g)}) = \Phi_\Pi(B).$$

The identity in Equation (29) can be rewritten as

$$\int_\Theta \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) = 0.$$

From above, we know that $(\Phi - \Phi_\Pi)$ defined a positive measure. For every Borel set B of S_d, we have

$$\int_B \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) \leq \int_\Theta \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) = 0.$$

Since the function $\mathbf{w} \mapsto \bigvee_{j=1}^d w^{(j)} v^{(j)}$ is strictly positive, continuous and defined on a compact set, we have that $\bigvee_{j=1}^d w^{(j)} v^{(j)} \geq c$ for a certain constant c strictly positive and we obtain

$$c(\Phi - \Phi)(B) \leq \int_B \bigvee_{j=1}^d w^{(j)} v^{(j)} (\Phi - \Phi_\Pi)(d\mathbf{w}) = 0.$$

The following identity is obtained

$$\Phi(B) = \Phi_\Pi(B),$$

since B is taken arbitrary from the Borelian of Θ, we conclude.

□

One can notice that the integrals defined in (29) can be rewritten with the help of stable tail dependence function, that is

$$L\left(v^{(1)},\dots,v^{(d)}\right) = \sum_{g=1}^G L^{(O_g)}\left(\mathbf{v}^{(O_g)}\right), \quad \mathbf{v}\in[0,\infty)^d,$$

since for every $\mathbf{x} \in [0,\infty)^d$

$$L(\mathbf{v}) = \int_\Theta \bigvee_{j=1}^d w^{(j)} v^{(j)} \Phi(d\mathbf{w}).$$

Appendix D.2 Additional results of Section 3

To establish the strong consistency of the estimator $\hat{\nu}_{n,m}$ in (7), certain conditions on the mixing coefficients must be satisfied.

Condition $\mathcal{C}$. Let $m_n = o(n)$. The series $\sum_{n\geq 1} \beta(m_n)$ is convergent, where β is defined in (18).

For the sake of notational simplicity, we will write $m = m_n$, $k = k_n$. The convergence of the series of β-mixing coefficients in Condition $\mathcal{C}$ is necessary to obtain the strong consistency of $\hat{\nu}_{n,m}$, and it can be achieved through the sufficiency condition of the Glivencko-Cantelli lemma for almost sure convergence.

Proposition 9. *Let $(\mathbf{Z}_t, t \in \mathbb{Z})$ be a stationary multivariate random process. Under Conditions $\mathcal{A}$ and $\mathcal{C}$, the madogram estimator in (7) is strongly consistent, i.e.,*

$$|\hat{\nu}_{n,m} - \nu| \xrightarrow[n\to\infty]{a.s.} 0,$$

with ν the theoretical madogram of the extreme value random vector $\mathbf{X}$ given in (5).

Let $C^o_{n,m}$ be the empirical estimator of the copula C_m based on the (unobservable) sample $(U^{(j)}_{m,1}, \ldots, U^{(j)}_{m,k})$ for $j \in \{1, \ldots, d\}$. The proof of Proposition 9 will use twice Lemma 3, which shows that $||C^o_{n,m} - C||_\infty$ converges almost surely to 0. The proof of this lemma is postponed to Appendix E.1 of supplementary results.

Proof of Proposition 9 We aim to show the following convergence

$$|\hat{\nu}_{n,m} - \nu| \xrightarrow[n\to\infty]{\text{a.s.}} 0,$$

where ν is the theoretical madogram of the extreme value random vector $\mathbf{X}$ given in (5) and $\hat{\nu}_{n,m}$ the madogram estimator in (7). Let us define the following quantity

$$\hat{\nu}^o_{n,m} = \frac{1}{k}\sum_{i=1}^{k}\left[\bigvee_{j=1}^{d} U^{(j)}_{m,i} - \frac{1}{d}\sum_{j=1}^{d} U^{(j)}_{m,i}\right], \tag{30}$$

that is the madogram estimated through the sample $\mathbf{U}_{m,1}, \ldots, \mathbf{U}_{m,k}$. Following Lemma A.1 of Marcon et al. 2017, we can show that

$$\hat{\nu}^o_{n,m} - \nu = \phi(C^o_{n,m} - C),$$

with $\phi : \ell^\infty([0,1]^d) \to \ell^\infty(\Delta_{d-1})$, $f \mapsto \phi(f)$ defined by

$$\phi(f) = \frac{1}{d}\sum_{j=1}^{d}\int_{[0,1]} f(1, \ldots, 1, \underbrace{u}_{j\text{-th component}}, 1, \ldots, 1)du - \int_{[0,1]} f(u, \ldots, u)du.$$

Using Conditions $\mathcal{A}$ and $\mathcal{C}$, by Lemma 3 in Appendix E.1, as $||C^o_{n,m} - C||_\infty$ converges almost surely to 0, we obtain that

$$\left|\hat{\nu}^o_{n,m} - \nu\right| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{31}$$

Furthermore, using the chain of inequalities and again Lemma 3 in Appendix E.1,

$$\begin{aligned}\left|\hat{\nu}_{n,m} - \hat{\nu}^o_{n,m}\right| &\leq 2 \sup_{j\in\{1,\ldots,d\}}\sup_{x\in\mathbb{R}}\left|\hat{F}^{(j)}_{n,m}(x) - F^{(j)}_m(x)\right| \\ &\leq 2 \sup_{j\in\{1,\ldots,d\}}\sup_{u\in[0,1]}\left|\frac{1}{k}\sum_{i=1}^{k}\mathbb{1}_{\{U^{(j)}_{m,i}\leq u\}} - u\right|.\end{aligned}$$

Then we obtain that

$$\left|\hat{\nu}_{n,m} - \hat{\nu}^o_{n,m}\right| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{32}$$

Now, write

$$|\hat{\nu}_{n,m} - \nu| \leq \left|\hat{\nu}_{n,m} - \nu^o_{n,m}\right| + \left|\hat{\nu}^o_{n,m} - \nu\right|,$$

and use Equations (31) and (32) to obtain the statement. □

We present here the strong consistency of our procedure when the dimension d is fixed the sample size n grows at infinity. The main technicality of the proof has already been tackled in Proposition 9 and we state the precise formulation of this theorem below.

Theorem 3. *Consider the AI-block model as defined in Definition 1 under Condition $\mathcal{B}$ and $(\mathbf{Z}_t, t \in \mathbb{Z})$ be a stationary multivariate random process. For a given $\mathcal{X}$ and its corresponding estimator $\hat{\mathcal{X}}$, if Conditions $\mathcal{A}$, $\mathcal{C}$ holds, then taking $\tau = 0$*

$$\mathbb{P}\left\{\hat{O} = \bar{O}\right\} = 1, \quad \text{as } n \to \infty.$$

Proof of Theorem 3 If a and b are not in the same cluster according to $\bar{O}$, i.e. $a \overset{\bar{O}}{\not\sim} b$, then $\chi(a,b) = 0$. Therefore, using Proposition 9 along with Conditions $\mathcal{A}$ and $\mathcal{C}$, we can conclude that almost surely

$$\lim_{n\to\infty} \hat{\chi}_{n,m}(a,b) = 0 \leq \tau.$$

Now, if $a \overset{\bar{O}}{\sim} b$, then $\chi(a,b) > 0$ and again by Propositions 9 and Conditions $\mathcal{A}$, $\mathcal{C}$, we obtain

$$\lim_{n\to\infty} \hat{\chi}_{n,m}(a,b) = \chi(a,b) > 0,$$

where the the strict positiveness is obtain through Condition $\mathcal{B}$, hence

$$a \overset{\bar{O}}{\sim} b \iff \lim_{n\to\infty} \hat{\chi}_{n,m}(a,b) > \tau.$$

Let us prove Theorem 3 by induction on the algorithm step l. We consider the algorithm at some step $l-1$ and assume that the algorithm was consistent up to this step, i.e. $\hat{O}_j = \bar{O}_j$ for $j = 1, \ldots, l-1$.

If $\lim_{n\to\infty} \hat{\chi}_{n,m}(a_l, b_l) = 0$, then no $b \in S$ is in the same group of a_l. Since the algorithm has been consistent up to this step l, it means that a_l is a singleton and $\hat{O}_l = \{a_l\}$.

If $\lim_{n\to\infty} \hat{\chi}_{n,m}(a_l, b_l) > \tau$, then $a_l \overset{\bar{O}}{\sim} b$. The equivalence above implies that $\hat{O}_l = S \cap \bar{O}_l$. Since the algorithm has been consistent up until this step, we know that $\hat{O}_l = \bar{O}_l$. Therefore, the algorithm remains consistent at step l with probability tending to one as $n \to \infty$, and Theorem 3 follows by induction. □

Appendix E. Further results

Appendix E.1 A usefull Glivenko-Cantelli result for the copula with known margins in a weakly dependent setting

In this section, we will prove an important auxiliary result: the empirical copula estimator $\hat{C}^o_{n,m}$ based on the weakly dependent sample $\mathbf{U}_{m,1}, \ldots, \mathbf{U}_{m,k}$ is uniformly strongly consistent towards the extreme value copula C. This result is a main tool to obtain important results in the paper such as Proposition 9, Theorem 3 and Proposition 5. For that purpose, the Berbee's coupling lemma is of prime interest (see, e.g., Rio 2017, Chapter 5) which gives an approximation of the original process by conveniently defined independent random variables.

Lemma 3. *Under conditions of Proposition 9, we have*

$$||C^o_{n,m} - C||_\infty \xrightarrow[n\to\infty]{a.s.} 0.$$

Lemma 3 Using triangle inequality, one obtain the following bound

$$||C^o_{n,m} - C||_\infty \leq ||C^o_{n,m} - C_m||_\infty + ||C_m - C||_\infty. \tag{33}$$

As $\{C_m, n \in \mathbb{N}\}$ is an equicontinuous class of functions (for every m, C_m is a copula hence a 1-Lipschitz function), defined on the compact set $[0,1]^d$ (by Tychonov's theorem) which converges pointwise to C by Condition $\mathcal{A}$. Then the convergence is uniform over $[0,1]^d$. Thus the second term of the RHS of Equation (33) converges to 0 almost surely.

Now, let us prove that $||C^o_{n,m} - C_m||_\infty$ converges almost surely to 0. By Berbee's coupling lemma (see Rio 2017, Theorem 6.1 or Bücher and Segers 2014, Theorem 3.1 for similar applications), we can construct inductively a sequence $(\bar{\mathbf{Z}}_{im+1}, \ldots, \bar{\mathbf{Z}}_{im+m})_{i\geq 0}$ such that the following three properties hold:

(i) $(\bar{\mathbf{Z}}_{im+1}, \ldots, \bar{\mathbf{Z}}_{im+m}) \overset{d}{=} (\mathbf{Z}_{im+1}, \ldots, \mathbf{Z}_{im+m})$ for any $i \geq 0$;
(ii) both $(\bar{\mathbf{Z}}_{2im+1}, \ldots, \bar{\mathbf{Z}}_{2im+m})_{i\geq 0}$ and $(\bar{\mathbf{Z}}_{(2i+1)m+1}, \ldots, \bar{\mathbf{Z}}_{(2i+1)m+m})_{i\geq 0}$ sequences are independent and identically distributed;
(iii) $\mathbb{P}\{(\bar{\mathbf{Z}}_{im+1}, \ldots, \bar{\mathbf{Z}}_{im+m}) \neq (\mathbf{Z}_{im+1}, \ldots, \mathbf{Z}_{im+m})\} \leq \beta(m)$.

Let $\bar{C}^o_{n,m}$ and $\bar{\mathbf{U}}_{m,i}$ be defined analogously to $C^o_{n,m}$ and $\mathbf{U}_{m,i}$ respectively but with $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ replaced with $\bar{\mathbf{Z}}_1, \ldots, \bar{\mathbf{Z}}_n$. Now write

$$C^o_{n,m}(\mathbf{u}) = \bar{C}^o_{n,m}(\mathbf{u}) + \left\{C^o_{n,m}(\mathbf{u}) - \bar{C}^o_{n,m}(\mathbf{u})\right\}. \tag{34}$$

We will show below that the term under brackets converges uniformly to 0 almost surely. Write $\bar{C}^o_{n,m}(\mathbf{u}) = \bar{C}^{o,\text{odd}}_{n,m}(\mathbf{u}) + \bar{C}^{o,\text{even}}_{n,m}(\mathbf{u})$ where $\bar{C}^{o,\text{odd}}_{n,m}(\mathbf{u})$ and $\bar{C}^{o,\text{even}}_{n,m}(\mathbf{u})$ are defined as sums over the odd and even summands of $\bar{C}^o_{n,m}(\mathbf{u})$, respectively. Since both of these sums are based on i.i.d. summands by properties (i) and (ii), we have $||\bar{C}^o_{n,m} - C_m||_\infty \xrightarrow[n\to\infty]{\text{a.s.}} 0$ using Glivenko-Cantelli (see Vaart and Wellner 1996, Chapter 2.5).

It remains to control the term under brackets on the right hand side of Equation (34), we have that

$$\begin{aligned}\left|C^o_{n,m}(\mathbf{u}) - \bar{C}^o_{n,m}(\mathbf{u})\right| &\leq \frac{1}{k}\sum_{i=1}^{k}\left|\mathbb{1}_{\{\bar{\mathbf{U}}_{m,i}\leq\mathbf{u}\}} - \mathbb{1}_{\{\mathbf{U}_{m,i}\leq\mathbf{u}\}}\right| \\ &\leq \frac{1}{k}\sum_{i=1}^{k}\mathbb{1}_{\{(\bar{Z}_{im+1},\ldots,\bar{Z}_{im+m})\neq(Z_{im+1},\ldots,Z_{im+m})\}}.\end{aligned}$$

Hence, using Markov's inequality and property (iii), we have

$$\mathbb{P}\left\{\sup_{u\in[0,1]^d}\left|\bar{C}^o_{n,m}(\mathbf{u}) - C^o_{n,m}(\mathbf{u})\right| > \epsilon\right\} \leq \frac{\beta(m)}{\epsilon}.$$

Thus by Condition $\mathcal{C}$,

$$\sum_{n\geq 1}\mathbb{P}\left\{\sup_{u\in[0,1]^d}\left|\bar{C}^o_{n,m}(\mathbf{u}) - C^o_{n,m}(\mathbf{u})\right| > \epsilon\right\} < \infty.$$

Applying Borel-Cantelli gives the desired convergence to 0 almost surely of the term under bracket in Equation (34). Gathering all results gives that the term $||C^o_{n,m} - C_m||_\infty$ converges almost surely to 0. Hence the statement using Equation (33). □

Appendix E.2 Proof of Proposition 5 in Appendix A.1

Proofs of consistency theorems for k-means clustering oftenly needs a uniform strong law of large numbers (SLLN) stated as

$$\sup_{g\in\mathcal{G}}\left|\int g d(\mathbb{P}_n-\mathbb{P})\right| \xrightarrow[n\to\infty]{a.s.} 0, \tag{35}$$

(see Section 4 of Pollard 1981), where $\mathbb{P}$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$, $\mathbb{P}_n$ the empirical measure and $\mathcal{G}$ a class of functions. Equation (35) is also stated as the class of functions $\mathcal{G}$ is Glivenko-Cantelli (see Chapter 2 of Vaart and Wellner 1996). In Proposition 3.3 of Janßen and Wan 2020, where $\mathbb{P}$ and $\mathbb{P}_n$ are replaced respectively by the angular measure S and its empirical counterpart S_n, it is shown that condition (35) holds. In our framework, we consider the extreme value copula C, the copula of block maxima C_m, its empirical counterpart $\hat{C}_{n,m}$ where the second is in the copula domain of attraction of C. The consistency of k-means clustering using madogram directly comes down from argments given in Janßen and Wan 2020 if we are able to state Equation (35) for a specific class of function $\mathcal{G}$ where the madogram belongs to.

For this purpose, the notion of bounded variation of functions and in particular the integration by part formula for Lebesgue-Stieltjes integral are of prime interest (see, for example, Fermanian, Radulovic, and Wegkamp 2004, Theorem 6). We will say that a function f is of bounded variation in the sense of Hardy-Krause if $V_{HK}(f)<\infty$ (see Radulović, Wegkamp, and Zhao 2017, Section 2 for a definition). Let us consider $\mathcal{G}$ as the class of functions which are continuous and $V_{HK}(g)<\infty$.

Proof of Proposition 5 We want to prove that for every $g\in\mathcal{G}$

$$\int g d\hat{C}_{n,m} \xrightarrow[n\to\infty]{a.s.} \int g dC. \tag{36}$$

Using integration by parts, we have that

$$\int g d\hat{C}_{n,m} = \Gamma(\hat{C}_{n,m}, g),$$

where $\Gamma(\cdot, g)$ is a linear and Lipschitz function, hence continuous. Now, if we can state that

$$||\hat{C}_{n,m}-C||_\infty \xrightarrow[n\to\infty]{a.s.} 0, \tag{37}$$

we obtain that $\Gamma(\hat{C}_{n,m},g)\xrightarrow[n\to\infty]{a.s.}\Gamma(C,g)$, using continuity, hence (36). Let us prove Equation (37). Using triangle inequality, we have

$$||\hat{C}_{n,m}-C||_\infty \le ||\hat{C}_{n,m}-C^o_{n,m}||_\infty + ||C^o_{n,m}-C||_\infty.$$

The second term in the right hand side converges almost surely to 0 by Lemma 3 (see Appendix E.1) with Conditions $\mathcal{A}$ and $\mathcal{C}$. It remains to work on the first term. Now, we have

$$\begin{aligned}||\hat{C}_{n,m}-C^o_{n,m}||_\infty &\le \frac{1}{k}\sum_{i=1}^k |\mathbb{1}_{\{\hat{U}^{(j)}_{n,m,i}\le u^{(j)},1\le j\le d\}} - \mathbb{1}_{\{U^{(j)}_{m,i}\le u^{(j)},1\le j\le d\}}| \\ &\le \frac{1}{k}\sum_{i=1}^k \prod_{j=1}^d \mathbb{1}_{\{\hat{U}^{(j)}_{n,m,i}\ne U^{(j)}_{m,i}\}}.\end{aligned}$$

Notice that, for every $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, d\}$

$$\begin{aligned}\mathbb{1}_{\{\hat{U}^{(j)}_{n,m,i} \neq U^{(j)}_{m,i}\}} &= \mathbb{1}_{\{|\hat{U}^{(j)}_{n,m,i} - U^{(j)}_{m,i}|>0\}} \\ &\leq \mathbb{1}_{\{\sup_{x\in\mathbb{R}} |\hat{F}^{(j)}_{n,m}(x) - F^{(j)}_{m}(x)|>0\}} = \mathbb{1}_{\{\sup_{u\in[0,1]} |\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{U^{(j)}_{m,i}\leq u\}} - u|>0\}}.\end{aligned}$$

Thus

$$||\hat{C}_{n,m} - C^{o}_{n,m}||_\infty \leq \Pi_{j=1}^{d} \mathbb{1}_{\{\sup_{u\in[0,1]} |\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{U^{(j)}_{m,i}\leq u\}} - u|>0\}}.$$

Denote by

$$A_n = \left\{ \omega \in \Omega : \sup_{u\in[0,1]} \left| \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{U^{(j)}_{m,i}\leq u\}} - u \right| (\omega) > 0 \right\},$$

$$A_{n,\epsilon} = \left\{ \omega \in \Omega : \sup_{u\in[0,1]} \left| \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{U^{(j)}_{m,i}\leq u\}} - u \right| (\omega) > \epsilon \right\},$$

where $\epsilon > 0$. Using Lemma 3 in Appendix E.1, we have that

$$\mathbb{P}\left\{ \limsup_{n\to\infty} A_{n,\epsilon} \right\} = 0.$$

Now, remark that $A_n = \bigcup_{N\geq 1} A_{n,1/N}$. We thus have

$$\mathbb{P}\left\{ \limsup_{n\to\infty} A_n \right\} = \mathbb{P}\left\{ \limsup_{n\to\infty} \bigcup_{N\geq 1} A_{n,1/N} \right\} = \mathbb{P}\left\{ \bigcup_{N\geq 1} \limsup_{n\to\infty} A_{n,1/N} \right\}.$$

Using σ-subadditivity of measures, we have

$$\mathbb{P}\left\{ \limsup_{n\to\infty} A_n \right\} \leq \sum_{N\geq 1} \mathbb{P}\left\{ \limsup_{n\to\infty} A_{n,1/N} \right\} = 0.$$

Hence $\mathbb{1}_{A_n} = 0$ almost surely as $n \to \infty$ which results on $||\hat{C}_{n,m} - C^{o}_{n,m}||_\infty$ converges almost surely to 0. We thus obtain (37) hence (36).

In order to prove Proposition 5, we prove in the following that the function

$$g(\mathbf{u}) = \min_{g\in\{1,\dots,G\}} \frac{1}{2} \sum_{j=1}^{d} |u^{(j)} - \psi_g^{(j)}|, \quad \psi_g \in [0,1]^d, \quad g \in \{1, \dots, G\}.$$

is of bounded variations in the sens of Hardy Krause (BHKV). Indeed, for every $j \in \{1, \dots, d\}$, the functions $(\mathbf{u}, \mathbf{v}) \mapsto u^{(j)}$ and $(\mathbf{u}, \mathbf{v}) \mapsto v^{(j)}$ are BHKV on $[0,1]^d \times [0,1]^d$ since it depends only on one variable and is monotone in this variable. As the difference between two BHKV functions are BHKV, it follows $\forall j \in \{1, \dots, d\}$ the function $(\mathbf{u}, \mathbf{v}) \mapsto u^{(j)} - v^{(j)}$ is BHKV on $[0,1]^d \times [0,1]^d$. Taking the absolute value preserves the BHKV property on $[0,1]^d$. It follows that $(\mathbf{u}, \mathbf{v}) \mapsto \sum_{j=1}^{d} |u^{(j)} - v^{(j)}|$ is BHKV as a sum of functions that are BHKV. Multiplying by a constant preserves the bounded variation property of the function. Finally, taking the min over a finite set of values in $[0,1]^d$ maintains the BHKV property as it holds for every $(\mathbf{u}, \mathbf{v}) \in [0,1]^d \times [0,1]^d$. Hence g is BHKV and clearly a continuous function. Thus using Equation (36) along with arguments from Janßen and Wan 2020, Theorem 3.1 we obtain the consistency the result.

□

Appendix E.3 Weak convergence of an estimator of $A^{(O)} - A$

We now state conditions on the block size m and the number of blocks k, as in Bücher and Segers 2014, to demonstrate the weak convergence of the empirical copula process based on the (unobservable) sample $(U^{(j)}_{n,m,1}, \dots, U^{(j)}_{n,m,k})$ for every $j \in \{1, \dots, d\}$ under mixing conditions. An additional condition will be required within the theorem to establish the weak convergence of the rank-based copula estimator under the same mixing conditions.

Condition $\mathcal{F}$. There exists a positive integer sequence ℓ_n such that the following statement holds:

(i) $m_n \to \infty$ and $m_n = o(n)$
(ii) $\ell_n \to \infty$ and $\ell_n = o(m_n)$
(iii) $k_n \alpha(\ell_n) = o(1)$ and $(m_n/\ell_n)\alpha(\ell_n) = o(1)$
(iv) $\sqrt{k_n}\beta(m_n) = o(1)$

For notational conveniency, we will write in the following $m_n = m$, $k_n = k$, $\ell_n = \ell$. Note that Condition $\mathcal{F}$ (iii) guarantees that the limit C is an extreme value copula by Hsing 1989, Theorem 4.2. As usual, the weak convergence of the empirical copula process stems down from the finite dimensional convergence and the asymptotic tightness of the process which then hold from Condition $\mathcal{F}$ (iii) and (iv) respectively. In order to apply Hadamard's differentiability to obtain the weak convergence of the empirical copula based on the sample's scaled ranks, we need a classical condition over the derivatives of the limit copula stated as follows.

Condition $\mathcal{G}$. For any $j \in \{1, \dots, d\}$, the jth first order partial derivative $\dot{C}^{(j)} = \partial C / \partial u^{(j)}$ exists and is continuous on $\{\mathbf{u} \in [0,1]^d, u^{(j)} \in (0,1)\}$.

The estimator of the Pickands dependence function that we present is based on the madogram concept (Cooley, Naveau, and Poncet 2006; Marcon et al. 2017), a notion borrowed from geostatistics in order to capture the spatial dependence structure. Our estimator is defined as

$$\hat{A}_{n,m}(\mathbf{t}) = \frac{\hat{\nu}_{n,m}(\mathbf{t}) + c(\mathbf{t})}{1 - \hat{\nu}_{n,m}(\mathbf{t}) - c(\mathbf{t})},$$

where

$$\hat{\nu}_{n,m}(\mathbf{t}) = \frac{1}{k}\sum_{i=1}^{k}\left[\bigvee_{j=1}^{d}\left\{\hat{U}^{(j)}_{n,m,j}\right\}^{1/t^{(j)}} - \frac{1}{d}\sum_{j=1}^{d}\left\{\hat{U}^{(j)}_{n,m,i}\right\}^{1/t^{(j)}}\right], \quad c(\mathbf{t}) = \frac{1}{d}\sum_{j=1}^{d}\frac{t^{(j)}}{1+t^{(j)}},$$

and $\hat{U}^{(j)}_{n,m,i} = \hat{F}^{(j)}_{n,m}(M^{(j)}_{m,i})$ corresponds to ranks scaled by k^{-1}. By convention, here $u^{1/0} = 0$ for $u \in (0,1)$. Let $g \in \{1, \dots, G\}$ and define

$$\hat{A}^{(O_g)}_{n,m}\left(\mathbf{t}^{(O_g)}\right) = \hat{A}_{n,m}\left(\mathbf{0}, \mathbf{t}^{(O_g)}, \mathbf{0}\right)$$

the empirical Pickands dependence function associated to the k-th subvector of $\mathbf{X}$p. We consider the empirical process of the difference between estimates of the Pickands dependence functions of subvectors $\mathbf{X}^{(O_g)}, g \in \{1, \dots, G\}$, and the estimator of the Pickands dependence function of $\mathbf{X}$:

$$\mathcal{E}_{nG}(\mathbf{t}) = \sqrt{k}\left(\hat{A}^{(O)}_{n,m}(\mathbf{t}) - \hat{A}_{n,m}(\mathbf{t})\right),$$

where $\hat{A}^{(O)}_{n,m}(\mathbf{t}) = \sum_{g=1}^{G} w^{(O_g)}(\mathbf{t})\hat{A}^{(O_g)}_{n,m}(\mathbf{t}^{(O_g)})$. Noticing that multiplying the above process by d and taking $\mathbf{t} = (d^{-1}, \dots, d^{-1})$ gives

$$\sqrt{k}\widehat{SECO}(O) = \sqrt{k}\left(\sum_{g=1}^{G}\hat{\theta}_{n,m}^{(O_g)} - \hat{\theta}_{n,m}\right).$$

Hence, the weak convergence of the above empirical process will immediately comes down from the one of the empirical process in $\mathcal{E}_{nG}$, as stated in the theorem below.

Theorem 4. *Consider the AI-block model in Definition 1 with a given partition O, i.e., $A = A^{(O)}$. Under Conditions $\mathcal{A}$, $\mathcal{F}$, $\mathcal{G}$ and $\sqrt{k}(C_m - C) \rightsquigarrow \Gamma$, the empirical process $\mathcal{E}_{nG}$ converges weakly in $\ell^\infty(\Delta_{d-1})$ to a tight Gaussian process having representation*

$$\begin{aligned}\mathcal{E}_G(\mathbf{t}) = (1 + A(\mathbf{t}))^2 \int_{[0,1]} (N_C + \Gamma)(u^{t^{(1)}}, \ldots, u^{t^{(d)}})du \\ - \sum_{g=1}^{G} w^{(O_g)}(\mathbf{t})\left(1 + A^{(O_g)}(\mathbf{t}^{(O_g)})\right)^2 \int_{[0,1]} (N_C + \Gamma)(\mathbf{1}, u^{t^{(i_{g,1})}}, \ldots, u^{t^{(i_{g,d_g})}}, \mathbf{1})du,\end{aligned}$$

where N_C is a continuous tight Gaussian process with representation

$$N_C(u^{(1)}, \ldots, u^{(d)}) = B_C(u^{(1)}, \ldots, u^{(d)}) - \sum_{j=1}^{d} \dot{C}_j(u^{(1)}, \ldots, u^{(d)})B_C(\mathbf{1}, u^{(j)}, \mathbf{1}),$$

and B_C is a continuous tight Gaussian process with covariance function

$$\mathrm{Cov}(B_C(\mathbf{u}), B_C(\mathbf{v})) = C(\mathbf{u} \wedge \mathbf{v}) - C(\mathbf{u})C(\mathbf{v}) \overset{\mathcal{H}_0}{=} C_\Pi(\mathbf{u} \wedge \mathbf{v}) - C_\Pi(\mathbf{u})C_\Pi(\mathbf{v}).$$

Theorem 4 The proof is straightforward, notice that by the triangle diagram in Figure 8

$$\mathcal{E}_{nG} = \psi \circ \phi\left(\sqrt{k}(\hat{A}_{n,m} - A)\right),$$

where ϕ is detailed as

$$\begin{array}{rccc}\phi & : & \ell^\infty(\Delta_{d-1}) & \to & \ell^\infty(\Delta_{d-1}) \otimes (\ell^\infty(\Delta_{d-1}), \ldots, \ell^\infty(\Delta_{d-1})) \\ & & x & \mapsto & (x, \phi_1(x), \ldots, \phi_G(x)),\end{array}$$

with for every $g \in \{1, \ldots, G\}$

$$\begin{array}{rccc}\phi_g & : & \ell^\infty(\Delta_{d-1}) & \to & \ell^\infty(S_d) \\ & & x & \mapsto & w^{(O_g)}(t^{(1)}, \ldots, t^{(G)})x(\mathbf{0}, t^{(i_{g,1})}, \ldots, t^{(i_{g,d_g})}, \mathbf{0}),\end{array}$$

and also

$$\begin{array}{rccc}\psi & : & \ell^\infty(\Delta_{d-1}) \otimes (\ell^\infty(\Delta_{d-1}), \ldots, \ell^\infty(\Delta_{d-1})) & \to & \ell^\infty(\Delta_{d-1}) \\ & & (x, \phi_1(x), \ldots, \phi_G(x)) & \mapsto & \sum_{g=1}^{G} \phi_g(x) - x.\end{array}$$

The function ϕ_g is a linear and bounded function hence continuous for every g, it follows that ϕ is continuous since each coordinate functions is continuous. As a linear and bounded function, ψ is also a continuous function. Noticing that,

$$(C_m - C)(\mathbf{1}, u, \mathbf{1}) = 0, \quad \forall n \in \mathbb{N},$$

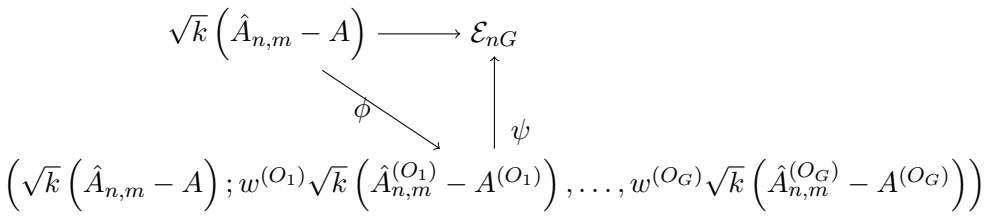

Figure 8. Commutative diagram of composition of function.

where $m = m_n$ is the block length for a sample size n. We thus have

$$\sqrt{k}(C_m - C)(\mathbf{1}, u, \mathbf{1}) \underset{n\to\infty}{\longrightarrow} 0,$$

with $k = k_n$ the number of blocks. Therefore $\Gamma(\mathbf{1}, u, \mathbf{1}) = 0$. Combining this equality with Bücher and Segers 2014, Corollary 3.6 and the same techniques as in the proof of Marcon et al. 2017, Theorem 2.4, we obtain along with Conditions $\mathcal{A}$, $\mathcal{F}$, $\mathcal{G}$

$$\sqrt{k}(\hat{A}_{n,m}(\mathbf{t}) - A(\mathbf{t})) \rightsquigarrow -(1 + A(\mathbf{t}))^2 \int_{[0,1]} (N_C + \Gamma)(u^{t^{(1)}}, \ldots, u^{t^{(d)}})du.$$

Applying the continuous mapping theorem for the weak convergence in $\ell^\infty(\Delta_{d-1})$ (Vaart and Wellner 1996, Theorem 1.3.6) leads the result. □