1 Introduction

Multivariate extremes arise when one or more of rare extremes events occur simultaneously. These events are of prime interest to assess natural hazard stemming from heavy rainfall, wind storms and earthqueakes since they are driven by joint extremes if a number meteorologial variables. The extension toward multivariate extrmes gives rise to the concept of tail dependence. It is well known from the classical theory that multivariate distributions can be decomposed into two distinct parts : the analysis of marginal distributions and the analysis of the dependence structure described by the copula function. Results from the extreme-value theory show that the possible dependence structures of extremes have to satisfy certain constraints. Indeed, the possible dependence structure may be described in various equivalent ways ([Resnick, 2008, Beirlant et al., 2004, De Haan et al., 2006]) : by the exponent measure Λ ([Balkema and Resnick, 1977]), by the Pickands dependence function A ([Pickands, 1981]), by the stable tail dependence function L ([Huang, 1992]), by the madogram ([Naveau et al., 2009]), by the extreme value copula C ([Gudendorf and Segers, 2010]). Estimating this extreme value copula is an important subject of the extreme litterature see, *e.g.* [Bücher et al., 2011, Gudendorf and Segers, 2012, Marcon et al., 2017, Escobar-Bach et al., 2018] to name a few.

The dependence structure between extreme observations can be complex and characterized by different notions from the ones arises in the classical theory. For this reason, recent works bring various notions to the framework of extreme such as sparsity ([Goix et al., 2015, Simpson et al., 2020, Meyer and Wintenberger, 2021]), conditional independence and graphical models ([Gissibl and Klüppelberg, 2018, Engelke and Hitz, 2020, Segers, 2020]), dimensionality reduction ([Chautru, 2015, Drees and Sabourin, 2021]) and unsupervised learning ([Cooley and Thibaud, 2019, Janßen and Wan, 2020]). In this work, we are concerned about variable clustering as a tool for learning the dependence structure of multivariate extreme and bridge important ideas from modern statistics and machine learning to the framework of extreme-value theory.

The problem of variable clustering is that of grouping similar coponents of a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$. Those groups are referred as clusters and unknown to the statistician who wants to recover them from $\mathbf{X}_1, \ldots, \mathbf{X}_n$, n independent copies of $\mathbf{X}$. Loosely speaking, we are able to perform clustering in two distinguish cases : by partitionning the set $\{1, \ldots, n\}$ of row indices or by partitioning with respect to column indices the set $\{1, \ldots, d\}$. The first problem will be designated as the data clustering problem whereas the second corresponds to the variable clustering problem discussed here so far. In data clustering, clusters are clouds of observations and corresponds to respective realizations of one of the mixture distribution, which is a distribution on the whole $\mathbb{R}^d$. In this framework with *i.i.d.* replications, [Pollard, 1981] shows the strong consistency of k-means clustering where the result was replicated in the context of extreme by [Janßen and Wan, 2020] for spherical k-means. In contrast, in variable clustering, the effort is to define cluster models relate to subsets of components X_j, of $\mathbf{X} \in \mathbb{R}^d$. In this framework, we do not longer want to cluster independent entities but, *a contrario*, to cluster those who are strongly dependent. Variable

clustering is of prime interest in weather extremes, with examples stemming from regionalisation, see [Bernard et al., 2013, Bador et al., 2015, Saunders et al., 2021], where one observes spatial phenoma at finitely many sites. An interesting special case is to cluster those sites according to their extremal dependencies. They include applications of K-means or hierarchical clustering with a dissimilariy designed for extremes. The statistical properties of those procedures have received a very limited amount of investigation. It is not currently known what probabilistic models on $\mathbf{X}$ can be estimated by these techniques. We will consider here model-based clustering where population-level clusters are clearly defined, offering interpretability and a benchmark to assess the performance of a peculiar clustering algorithm.

In this work, we propose the AI-block model as a model for variable clustering in extreme-value theory and show that the clusters given by this model are uniquely defined. We then motivate and develop an algorithm tailored to the model with the help of the SECO metric. We thus analyze its performance in terms of exact cluster recovery, for minimally separated clusters, under appropriately defined cluster separation metric.

Notations We use the following notations throughout the paper. All bold letters $\mathbf{x}$ corresponds to vector in $\mathbb{R}^d$. By considering $B \subseteq \{1, \dots, d\}$, we denote the $|B|$-subvector of $\mathbf{x}$ by $\mathbf{x}^{(B)} = (x_j)_{j \in B}$. Similarly, let G a cumulative distributive function on $[0,1]^d$, $G^{(B)}$ is defined as

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

where $(\mathbf{1}, \mathbf{x}^{(B)}, \mathbf{1})$ has jth component equals to $x_j \mathbb{1}_{\{j \in B\}} + \mathbb{1}_{\{j \notin B\}}$. In a similar way, we note $(\mathbf{0}, x^{(B)}, \mathbf{0})$ the vector in $\mathbb{R}^d$ which equals x_j if $j \in B$ and 0 otherwise. We will the d consecutive integer set starting from 1 as $[\![d]\!]$. Weak convergence of processes are denoted by '$\rightsquigarrow$'. The notation δ_x corresponds to the dirac measure at x. We define by $\mathbf{X} \in \mathbb{R}^d$ a random vector with law G. Let $O = \{O_k\}_{k=1,\dots,K}$ be a partition of $\{1, \dots, d\}$ into K groups and let $s : \{1, \dots, d\} \to \{1, \dots, K\}$ be an index assignement function defined by $O_k = \{a \in \{1, \dots, d\} : s(a) = k\} = \{i_{k,1}, \dots, i_{k,d_k}\}$ with $d_1 + \cdots + d_K = d$.

In the present paper, we present in Section 2 some background notions in extreme-value theory necessary for our analysis. All these notions introduced, we present AI-block models designed for variable clustering. All proofs are deferred to the Appendix.

2 A model for variable clustering

2.1- extreme value theory Consider $\mathbf{Z} = (Z_1, \dots, Z_d)$ a d-dimensional random vector with law F and $\mathbf{Z}_i = (Z_{i,1}, \dots, Z_{i,d})$, $i = 1, \dots, n$ be independent copies of $\mathbf{Z}$. By denoting the component-wise maxima by $\mathbf{M}_n = (\max_{i=1,\dots,n} Z_{i,1}, \dots, \max_{i=1,\dots,n} Z_{i,d})$. The vector $\mathbf{Z}$ is said to be in max-domain of attraction of the random vector $\mathbf{X} = (X_1, \dots, X_d)$, denoted as $F \in D(G)$, if for any

$\mathbf{x} = (x_1, \ldots, x_d)$,

$$\lim_{n\to\infty} \{F(\mathbf{a}_n \mathbf{x} + \mathbf{b}_n)\}^n = G(\mathbf{x}), \tag{1}$$

where $\mathbf{a}_n > \mathbf{0}$ (which means that $a_{i,n} > 0$ for every i and n) and $\mathbf{b}_n \in \mathbb{R}^d$. In this case, $\mathbf{X}$ is max-stable with GEV margins and we may write

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

where Λ is a Radon measure on the cone $E = [0, \infty)^d \setminus \{\mathbf{0}\}$. This condition is equivalent to the notion of regular variation, that is there exists a sequence $0 < a_n \to \infty$ and a limit measure such that

$$n\mathbb{P}\left\{a_n^{-1}\mathbf{X} \in \cdot\right\} \xrightarrow[n\to\infty]{v} \Lambda(\cdot),$$

with $\xrightarrow{v}$ denotes the vague convergence.

Those notions can be translated, as in the classical theory, in terms of copula. A d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$ follows the law of a multivariate extreme-value distribution if its one dimensional marginal distributions $G_j(x) = \mathbb{P}\{X_j \leq x\}$ for all $x \in \mathbb{R}$ and $j \in \{1, \ldots, d\}$ are GEV distributions, and the joint distribution can be written, for all $\mathbf{x} \in \mathbb{R}^d$, in the form

$$G(\mathbf{x}) = C\left(G_1(x_1), \ldots, G_d(x_d)\right), \tag{2}$$

where C is an extreme value copula, *i.e.*, for all $\mathbf{u} \in (0, 1]^d$

$$C(\mathbf{u}) = \exp\{-L\left(-\ln(u_1), \ldots, -\ln(u_d)\right)\},$$

with L is known as the stable tail dependence function (see [Gudendorf and Segers, 2010] for an overview of extreme value copulae). As it is a homogeneous function of order 1, *i.e.* $L(a\mathbf{z}) = aL(\mathbf{z})$ for all $a > 0$, we have, for all $z \in [0, \infty)^d$,

$$L(\mathbf{z}) = (z_1 + \cdots + z_d)A(\mathbf{t}),$$

with $t_j = z_j/(z_1 + \ldots z_d)$ for $j \in \{2, \ldots, d\}$ and $t_1 = 1 - t_2 - \cdots - t_d$, 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 quantitfy the extremal dependence among the element of 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 cases.

2.2- AI-block models Motivated by a rich set of applications, we consider variable clustering as the initial dimension reduction step applied to the observed vector $\mathbf{X} = (X_1, \ldots, X_d)$. These models are build on the assumption that the observed variables $\mathbf{X} = (X_1, \ldots, X_d)$ can be partionned into K unknown clusters $O = \{O_1, \ldots, O_K\}$ such that variables in the same cluster are as dependent as possible and the clusters are mutually independent. This structure of dependendence has been observed to hold, empirically, recent studies have shown that in many applications such as spatial precipitation (see [Le et al., 2018, Lalancette et al., 2021]) dependence tends to become weaker for the largest observations and asymptotic independence is an appropriate regime for sites reasonably far enough. We define here a population-level cluster as a group of variables that shares the same extremal dependence structure within the cluster and is independent from the other clusters.

To keep the presentation focus, let us consider $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_K)}$ be extreme value random vectors with extreme value copulae $C^{(O_1)}, \ldots, C^{(O_K)}$ respectively. We suppose that $\mathbf{X}^{(O_k)}$ and $\mathbf{X}^{(O_j)}$ are mutually asymptotically independent if $k \neq j$, *i.e.* the corresponding components of the limit vector in (1) are mutually independent. Let us define the following function :

$$\begin{array}{rccl} C_\Pi : & [0,1]^d & \longrightarrow & [0,1] \\ & \mathbf{u} & \longmapsto & \Pi_{k=1}^K C^{(O_k)}(u_{i_{k,1}}, \ldots, u_{i_{k,d_k}}). \end{array}$$

We want to show that C_Π is an extreme value copula associated to $\mathbf{X} = (\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_K)})$ in order that all objects we use below are well-defined, in particular the existence of a *stable tail dependence function* associated to C_Π.

Lemma 1. *C_Π is an extreme value copula associated to the random vector $\boldsymbol{X}$.*

A direct consequence of this Lemma is that if $\mathbf{X}$ admits a copula C_Π, it is an extreme value random vector. In particular, there exists a stable tail dependence function L say and one can show that L can be expressed in a convenient way such as :

$$L(z_1, \ldots, z_d) = \sum_{k=1}^{K} L^{(O_k)}\left(\mathbf{z}^{(O_k)}\right), \quad \mathbf{z} \in [0, \infty)^d, \tag{3}$$

where $L^{(O_1)}, \ldots, L^{(O_K)}$ are the stable tail dependence function associated to the extreme value copulae $C^{(O_1)}, \ldots, C^{(O_K)}$ respectively. Furthermore, this model is a particular form of the nested extreme value copula and is the object of the remark below.

Remark 1. *Equation (3) can restated as*

$$L(\boldsymbol{z}) = L^{(0)}\left(L^{(O_1)}\left(z^{(O_1)}\right), \ldots, L^{(O_K)}\left(z^{(O_K)}\right)\right),$$

where $L^{(0)}(z_1, \ldots, z_K) = \sum_{k=1}^K z_k$ is a stable tail dependence function *corresponding to the asymptotic independence. Since C is an extreme value copula by Lemma 1, we obtain that C is also a nested extreme value copula as formulated in [Hofert et al., 2018].*

The following remark state the form of the stable tail dependence function in terms of regular variation. This corresponding statement might be found in the litterature, see Lemma 7.2 of [Resnick, 2007] or Theorem 1.28 of [Lindskog, 2004] for example.

Remark 2. *Let $K = 2$, consider $\boldsymbol{X}^{(O_1)} \in \mathbb{R}^{d_1}$ and $\boldsymbol{X}^{(O_2)} \in \mathbb{R}^{d_2}$ defined in the same probability space, independent, and satisfy the regular variation assumption*

$$n\mathbb{P}\left\{a_n^{-1}\boldsymbol{X}^{(O_1)} \in \cdot\right\} \xrightarrow[n\to\infty]{v} \nu_1(\cdot), \quad n\mathbb{P}\left\{a_n^{-1}\boldsymbol{X}^{(O_2)} \in \cdot\right\} \xrightarrow[n\to\infty]{v} \nu_2(\cdot),$$

with the same sequence $0 < a_n \to \infty$. then the distribution tail of $\boldsymbol{X} = (\boldsymbol{X}^{(O_1)}, \boldsymbol{X}^{(O_2)})$ is also regularly varying with

$$n\mathbb{P}\left\{a_n^{-1}\boldsymbol{X} \in \cdot\right\} \xrightarrow[n\to\infty]{v} \nu(\cdot),$$

where

$$\nu(d\boldsymbol{x}^{(O_1)}, d\boldsymbol{x}^{(O_2)}) = \nu_1(d\boldsymbol{x}^{(O_1)})\delta_0(d\boldsymbol{x}^{(O_2)}) + \delta_0(d\boldsymbol{x}^{(O_1)})\nu_2(d\boldsymbol{x}^{(O_2)}).$$

With all notations and definitions previously introduced, we now are able to state the definition of the considered model here.

Definition 1 (Asymptotic Independence-block model)**.** *The random vector $\boldsymbol{X} = (\boldsymbol{X}^{(O_1)}, \ldots, \boldsymbol{X}^{(O_K)})$ follows an AI-block model if $\boldsymbol{X}^{(O_k)} = (X_{i_{k,1}}, \ldots, X_{i_{k,d_k}})$ are extreme value random vectors for $k \in \{1, \ldots, K\}$ and are mutually independent if $k \neq j$.*

Notice that, when $K = 1$, the definition of the AI-block model thus reduces to $\mathbf{X} = (X_1, \ldots, X_d)$ is an extreme value random vector, which is trivially obtained by the definition.

Let $\mathcal{L}(\mathbf{X}) = \{O : \mathbf{X} \text{ is AI-block model}\}$ which is nonempty and finite so it does have maximal elements, we note $\mathbf{X} \sim O$ if $\mathbf{X}$ follows an AI-block model. We introduce the following partial order on sets, let $O = \{O_k\}_k$, $S = \{S_{k'}\}_{k'}$ be two partitions of $\{1, \ldots, d\}$. We say that S is a sub-partition of O if for each k' there exists k such that $S_{k'} \subseteq O_k$. We define the partial order $\leq$ between two partitions O, S of $\{1, \ldots, d\}$ by $O \leq S$ if S is a sub-partition of O. For any partition $O = \{O_k\}_{1 \leq k \leq K}$, we write $j \overset{O}{\sim} k$ if there exist $k \in \{1, \ldots, K\}$ such that $j, k \in O_k$.

Definition 2. *For any two partitions O, S of $\{1, \ldots, d\}$, we define $O \cap S$ as the partition induced by the equivalence relation $j \overset{O\cap S}{\sim} k$ iff $j \overset{O}{\sim} k$ and $j \overset{S}{\sim} k$.*

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

Theorem 1. *Let $\boldsymbol{X}$ be an extreme value random vector, then the following properties hold :*

1. *Consider $O \leq S$. Then $\boldsymbol{X} \sim S$ implies $\boldsymbol{X} \sim O$.*
2. *$O \leq O \cap S$ and $S \leq O \cap S$*

3. *If* $\boldsymbol{X} \sim O$ *and* $\boldsymbol{X} \sim S$*, then* $\boldsymbol{X} \sim O \cap S$*.*

4. *The set* $\mathcal{L}(\boldsymbol{X})$ *has a unique maximum* $\bar{O}(\boldsymbol{X})$*, with respect to the partition partial order.*

We refer to Appendix for a proof of this Theorem. It shows that the maximum partition for which $\mathbf{X}$ is an AI-block model always exists and its structure is intrinsic of the definition of $\mathbf{X}$. The maximal partiton of $\bar{O}(\mathbf{X})$ matches our intuition regarding what would constitute a reasonable clustering target for these models, *i.e.* the finest partition $\bar{O}(\mathbf{X})$ where $\mathbf{X}$ is an AI-block model as stated in the introduction of [Ryabko, 2017]. With a slight abuse of notation, we will write $\bar{O}(\mathbf{X})$ as $\bar{O}$.

In an AI-block model, we may restrict Equation (3) to the simplex, this equation becomes equivalent to

$$\begin{aligned} A(t_1, \ldots, t_d) &= \frac{1}{z_1 + \cdots + z_d} \left[\sum_{k=1}^{K} (z_{i_{k,1}} + \cdots + z_{i_{k,d_k}}) A^{(O_k)}(\mathbf{t}^{(O_k)}) \right] \\ &= \sum_{k=1}^{K} w^{(k)}(\mathbf{t}) A^{(O_k)}(\mathbf{t}^{(O_k)}) =: A_O \end{aligned}$$

with $t_j = z_j/(z_1 + \cdots + z_d)$ for $j \in 2, \ldots, d$ and $t_1 = 1 - t_2 - \cdots - t_d$, $w^{(O_k)}(\mathbf{t}) = z_{i_{k,1}} + \cdots + z_{i_{k,d_k}}/(z_1 + \cdots + z_d)$ for $k \in \{2, \ldots, K\}$ and $w_1 = 1 - w^{(O_2)}(\mathbf{t}) - \cdots - w^{(O_K)}(\mathbf{t})$ and $t_{i_{k,l}} = z_{i_{k,l}}/(z_{i_{k,1}} + \cdots + z_{i_{k,d_k}})$ for $k \in \{1, \ldots, K\}$ and $l \in \{1, \ldots, d_k\}$. The function A is still a Pickands dependence function as a convex combination of Pickands dependence function (see p. 123 of [Falk et al., 2010]).

When considering independence between random variables, we know that $A(\mathbf{t}) \leq 1$ for $\mathbf{t} \in \Delta_{d-1}$ where inequality stands for asymptotic independence between all random variables. A more general statement can also be formulated considering random vectors, where the former case directly comes down by taking $d_1 = \cdots = d_K = 1$.

Proposition 1. *In general case, we have for every* $\boldsymbol{t} \in \Delta_{d-1}$*,*

$$(A_O - A)(\boldsymbol{t}) \geq 0,$$

with equality if and only if $\boldsymbol{X} \sim O$*.*

Remark 3. *A beautiful way to state asymptotic independence between random vectors is by the mean of the exponent measure. Taking notations of Remark 2,* $\boldsymbol{X}^{(O_1)}$ *is independent of* $\boldsymbol{X}^{(O_2)}$ *if and only if, for* $\boldsymbol{y} > \boldsymbol{0}$

$$\nu \left\{ \boldsymbol{x} \in E, \boldsymbol{x}^{(O_1)} > \boldsymbol{y}^{(O_1)}, \boldsymbol{x}^{(O_2)} > \boldsymbol{y}^{(O_2)} \right\} = 0.$$

Thus, the exponent measure ν *concentrates on*

$$]0, \infty[^{d_1} \times \{\boldsymbol{0}\}^{d_2} \cup \{\boldsymbol{0}\}^{d_1} \times]0, \infty[^{d_2}.$$

These conditions generalize straightforwadly those stated in Proposition 5.24 of [Resnick, 2008].

3 The SECO similarity metric

Let $\mathbf{X} \sim O$ has a block structure with associated Pickands A which can be expressed as a convex combination of the K Pickands $A^{(O_k)}$. One expect a consistent clustering could be possible when

$$A_O - A \tag{4}$$

is small enough. To motivate this metric, notice that in AI-block model, if a given partition $O \in \mathcal{L}(\mathbf{X})$, then (4) is equal to 0 while it is strictly greater than 0 if $O \notin \mathcal{L}(\mathbf{X})$. This statement is precised by the following proposition.

Proposition 2. *Let $\boldsymbol{X} \sim O$ and $S_{k'}$ an arbitrary partition of $[\![d]\!]$. We thus have for $\boldsymbol{t} \in \Delta_{d-1}$*

$$0 \leq \sum_{l=1}^{L} \sum_{k=1}^{K} w^{(O_k \cap S_{k'})}(\boldsymbol{t}) A^{(O_k)}(\boldsymbol{0}, \boldsymbol{t}^{(O_k \cap S_{k'})}, \boldsymbol{0}) - A(\boldsymbol{t}), \quad \forall \boldsymbol{t} \in \Delta_{d-1}, \tag{5}$$

where

$$w^{(O_k \cap S_{k'})}(\boldsymbol{t}) = \sum_{j \in S_{k'} \cap O_k} t_j, \quad \forall k \in \{1, \ldots, K\},\, k' \in \{1, \ldots, K'\}.$$

Furthermore, the right hand side of the inequality is equal to zero if and only if $S \leq O$.

Proposition 2 gives the theoretical value if $\mathbf{X} \sim O$ for an arbitrary partition which is stricly greater than 0 if the given partition does not belong to $\mathcal{L}(\mathbf{X})$. Now, using this result, one can think of an algorithm to recover the maximal element $\bar{O}$. Indeed, if the value of the Pickands dependence function is known for every $\mathbf{t} \in \Delta_{d-1}$, one has to evaluate this metric in (4) for a partition of $[\![d]\!]$ by $\hat{O}_1$ and $\hat{O}_2$ and to stop when it is equal to 0 and proceed inductively. However, the number of oracle calls will be equal to the Stirling number of the Second kind and will grows drastically as d increases. We thus need a metric that avoids to call for value for each partition of $\{1, \ldots, d\}$. Indeed, by Proposition 1, one has :

$$A(\mathbf{t}) \leq w^{([\![d-1]\!])}(\mathbf{t}) A\left(\frac{t_1}{w^{([\![d-1]\!])}(\mathbf{t})}, \frac{t_2}{w^{([\![d-1]\!])}(\mathbf{t})}, \ldots, \frac{t_{d-1}}{w^{([\![d-1]\!])}(\mathbf{t})}, 0\right) + t_d, \quad \mathbf{t} \in \Delta_{d-1}, \tag{6}$$

where inequalities holds if and only if $X_d \perp\!\!\!\perp (X_1, \ldots, X_{d-1})$. This statement holds whatever the chosen order but we choose the trivial one for notational convenience. This result gives us a metric to introduce an index in a candidate cluster. Indeed, take as candidate cluster $\hat{C} = \{1, \ldots, d-1\}$ and we ask if the element $\{d\}$ belongs to $\hat{C}$. Thus, with knowledge of the above equation, $\{d\}$ do not belongs to $\hat{C}$ if (6) holds as an equality for every $\mathbf{t} \in \Delta_{d-1}$.

For $\hat{O}_1$ and $\hat{O}_2$, we obtain an independent partition if and only if the criteria given in (4) is equal to 0. Another difficulty is this has to hold for every $\mathbf{t} \in \Delta_{d-1}$ which is not countable and implies computationaly infeasibility for every considered dimension $d \in \mathbb{N}_*$. We may further discretize the

simplex of $\mathbb{R}^d$ to have an approximation of (4). Take $\mathbf{t} \in \Delta_{d-1} \subset [0,1]^d$, we want to have at least one observation at distance less than 1 say, from $\mathbf{t}$, then we must increase the number N of observations as d increases. We thus need at least

$$N \geq \frac{\Gamma(d/2+1)}{\pi^{d/2}} \overset{d\to\infty}{\sim} \left(\frac{d}{2\pi e}\right)^{d/2} \sqrt{d\pi}$$

points in order to fill the hypercube $[0,1]^d$, thus Δ_{d-1}. This number of points grows exponentially fast with d. So, even with discretization, where we might obtain theoretical concentration bounds for the sup-norm over the simplex (using chaining method), the criteria would not be computable in practice.

However, in the framework of extremes, independence between the components $X_1, \ldots, X_d$ of an extreme value random bector $\mathbf{X} \in \mathbb{R}^d$ can be stated in a useful manner by the use of exponent measure (see Remark 3). One simple necessary and sufficient conditions is those stated in [Takahashi, 1987, Takahashi, 1994]. It is shown that

$$G(\mathbf{x}) = G_1(x_1) \ldots G_d(x_d), \quad \mathbf{x} \in \mathbb{R}^d, \tag{7}$$

that is $\mathbf{X}$ is totally independent, if and only if there exists $\mathbf{p} = (p_1, \ldots, p_d) \in \mathbb{R}^d$ such that (7) holds. Roughly speaking, in extreme-value theory if for at least one $\mathbf{p} \in \mathbb{R}^d$ Equation (7) holds, then it extends for every $\mathbf{x} \in \mathbb{R}^d$ using max-stability. An analogy of this statement in the non-extreme world could be similar to X_1 and X_2 are independent if

$$cov(f_1(X_1), f_2(X_2)) = 0, \tag{8}$$

holds for every continuous bounded functions f_1 and f_2 is equivalent to $\exists f, g$ a continuous and bounded function such that (8) is equal to 0. This statement is known to be false in general but stand as true for the Gaussian case (take $f = g = id$). The prrof of this statement is given in Appendix.

The result of [Takahashi, 1994] is extended to our framework by $G(\mathbf{x}) = G^{(O_1)}(\mathbf{x}^{(O_1)})G^{(O_2)}(\mathbf{x}^{(O_2)})$ for any $\mathbf{x} = (\mathbf{x}^{(O_1)}, \mathbf{x}^{(O_2)}) \in \mathbb{R}^d$ if and only if the equation holds for at least one $\mathbf{p} = (\mathbf{p}^{(O_1)}, \mathbf{p}^{(O_2)}) \in \mathbb{R}^d$. Moreover, this result belongs to those who were stated for single components of max-stable random vector $\mathbf{X}$ that can be extended for subvectors of max-stable random vector (see Lemma 6 of [Papastathopoulos and Strokorb, 2016] or Exercise 5.5.1 of [Resnick, 2008] for instance).

One direct application of this result is that $X^{(O_1)}$ and $X^{(O_2)}$ are independent if and only one has :

$$A\left(\frac{1}{d}, \ldots, \frac{1}{d}\right) = \frac{d_1}{d} A^{(O_1)}\left(\frac{1}{d_1}, \ldots, \frac{1}{d_1}\right) + \frac{d_2}{d} A^{(O_2)}\left(\frac{1}{d_2}, \ldots, \frac{1}{d_2}\right).$$

By denoting $\theta = d\, A(d^{-1}, \ldots, d^{-1})$ the so-called extremal coefficient, we restate the equation above

as the SECO (Sum of Extremal COefficients) metric

$$SECO(O_1, O_2) = \theta^{(O_1)} + \theta^{(O_2)} - \theta. \tag{9}$$

Loosely speaking, if $\mathbf{X}^{(O_1)}$ and $X^{(O_2)}$ are mutually independent, thus a characterizing property is that the extremal coefficient of $\mathbf{X}$ is written as the sum of the extremal coefficients $\theta^{(O_1)}$ and $\theta^{(O_2)}$ of $\mathbf{X}^{(O_1)}$ and $\mathbf{X}^{(O_2)}$ respectively.

Let $\bar{O}$ the unique maximal element of $\mathcal{L}(\mathbf{X})$ and consider $\hat{O}_1 \subset \bar{O}_1$. Consider now $j \in \{1, \ldots, d\} \setminus \hat{O}_1$ and we to know if $j \in \bar{O}_1$, again, using knowledge of the Pickands dependence function, the statistician can evaluate the following value

$$SECO(\hat{O}_1, j) = \theta^{(\hat{O}_1)} + 1 - \theta^{(\hat{O}_1 \cup \{j\})}. \tag{10}$$

The former term is equal to 0 if and only if $j \notin \bar{O}_1$ and strictly positive otherwise. This property is now being clear, the key quantity that quantifies the difficulty of clustering in AI-block models is MSECO that is defined as

$$MSECO(s_k) = \min_{\hat{O}_k \subseteq \bar{O}_k, |\hat{O}_k| = s_k} \max_{j \in \bar{O}_k \setminus \hat{O}_k} SECO(\hat{O}_k, j), \tag{11}$$

where $1 \leq s_k \leq d_k - 1$. Then, for each $k \in \{1, \ldots, K\}$ and $s_k \in \{1, \ldots, d_k - 1\}$ the larger is the value of $MSECO(s_k)$, then easier the cluster detection problem. We now state an assumption in order to recover our hidden clusters and guarantee the positivity of the MSECO metric.

Assumption A. *For every $k \in \{1, \ldots, K\}$, $\boldsymbol{X}^{(\bar{O}_k)}$ exhibits asymptotic dependence between all components.*

In AI-block models with Assumption A, we always have $MSECO(s_k) > \eta_{s_k+1}$ for each $k \in \{1, \ldots, K\}$ and $|\hat{O}_k| = s_k$ for every η_{s_k+1} with $\eta_{s_k+1} = 0$. However, a larger value of η_{s_k+1} for each size s_k will be needed for retrieving consistently the partition $\bar{O}$ from independent observations. Let A be the Pickands dependence function of $\mathbf{X} \sim O$. We define for every $k \in \{1, \ldots, K\}$ and $s_k \in \{1, \ldots, d_k - 1\}$

$$\mathcal{A}(\eta_{s_k}) = \left\{ A^{\bar{O}_k} : MSECO(s_k) > \eta_{s_k+1} \right\}.$$

A sufficiency condition in order that Assumption A is satisfied is to suppose that each exponent measure of the extreme value random vectors $\mathbf{X}^{(\bar{O}_k)}$ has a nonnegative Lebesgue density on the non negative orthant $[0, \infty)^{d_k} \setminus \{\mathbf{0}^{\bar{O}_k}\}$ for every $k \in \{1, \ldots, K\}$ (see [Engelke and Hitz, 2020] and Kirstin Strokorb's discussion contribution). Various classes of tractable extreme value distributions satisfy Assumption A. Popular models that are 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 et al., 2010]) or the Hüsler Reiss model ([Hüsler and

Reiss, 1989]).

Remark 4. *In its seminal work, [Ryabko, 2017] proposes a conditional independence test to decide whether an element j belongs to a cluster $\hat{O}_1$, say. Formally, one may ask if $\boldsymbol{X}^{(\hat{O}_1)} \perp\!\!\!\perp X_j | S \setminus (\hat{O}_1 \cup \{j\})$. However, conditional independence among extremes is relatively new and mainly designed for tree inference (see* e.g. *[Engelke and Hitz, 2020, Asenova et al., 2021, Segers, 2020, Engelke and Volgushev, 2020]). In those models, as in Gaussian graphical model (we refer to [Lauritzen, 1996] for an overview), one suppose that the extreme value random vector admits a Husler Reiss density which can be seen as a Gaussian extremal model with variogram. Interesting properties are also obtained such that conditional dependencies are encoded in the precision matrix of the Husler Reiss distribution.*

Highlighted by Remark 4, if we suppose, as in the litterature of extremal graphs, that $\mathbf{X}$ is absolutely continous with respect to the Lebesgue measure and admits an Husler Reiss density, then $\mathbf{X}$ exhibits extremal dependence between all its components. Thus, the resulting maximal element of the AI-block model is trivial and given by $\bar{O} = \{\{1\}, \dots, \{d\}\}$. In order to drop Assumption A, further works are needed and left for future investigations.

Remark 5. *In this remark we go outside the extreme framework, we suppose that $\boldsymbol{X}$ has a Gaussian copula distribution with zero mean, and copula function with parameters $\mu = 0$ and Σ, a correlation matrix. Recall that this implies that*

$$\boldsymbol{Y} := (Y_1, \dots, Y_d) := (h_1(X_1), \dots, h_d(X_d)) =: h(\boldsymbol{X}) \overset{d}{\sim} \mathcal{N}_d(0, \Sigma),$$

with $h_j = \Phi^{-1} \circ G_j$ for each $j \in \{1, \dots, d\}$, where Φ is the cumulative distribution function of a standard Gaussian random variable. One can show that

$$\boldsymbol{X}^{(O_1)} \perp\!\!\!\perp \boldsymbol{X}^{(O_2)} \iff |\Sigma^{(O_1)}||\Sigma^{(O_2)}| = |\Sigma|,$$

where $\Sigma^{(O_k)}$ is a sub-matrix of Σ where we only kept the ith rows and columns with $i \in O_k$, $k \in \{1, 2\}$. Also, conditional independencies are encoded in the correlation matrix such as $\boldsymbol{X}^{(O_1)} \perp\!\!\!\perp X_j | (S \setminus R)$, where $R = S \setminus (O_1 \cup \{j\})$ is equivalent to

$$|\Sigma^{(R)}||\Sigma| = |\Sigma^{(R \setminus \{j\})}||\Sigma^{(S \setminus \{j\})}|.$$

One can hope that these properties might find an equivalent statement in the case of Husler-Reiss for extreme frameworks. However, such a distributional assumption will lead to a trivial model as stated right after Remark 4.

We suppose here that the statistician has access to distribution $\mathbf{X}$ through the knowledge of A via an oracle. The proposed algorithm (see Algorithm 1) work as follows. It attempts to split the input set recursively into two independent clusters, until it is no longer possible. To split a set

in two, it starts with a candidate cluser $\hat{O}_1 = \{1\}$ and measures its discrepancy between the set $\hat{O}_2 = \{2, \ldots, d\}$ using the SECO metric. If $SECO(\hat{O}_1, \hat{O}_2)$ is already 0, then we have split the set into two independent clusters and can stop. Otherwise, the algorithm then takes element from $\hat{O}_2$ and shift its position if $SECO(\hat{O}_1, j) > 0$.

Algorithm 1 Recursive algorithm to cluster with K unknown

Data: The set $S = \{1, \ldots, d\}$, $l = 0$
Result: Cluster with K unknown, given an oracle
procedure ALG(S)
    while $S \neq \emptyset$ **do**
        $l = l + 1$
        $(\hat{C}, \hat{R}) = Split(S)$
        $\hat{O}_l = \hat{C}$
        $S = S \setminus \hat{R}$
    return $\{\hat{O}\} = \{\hat{O}_l\}_l$
procedure SPLIT(S)
    Initialize : $\hat{C} := \{1\}$, $\hat{R} := \{2, \ldots, d\}$.
    while $SECO(\hat{C}, \hat{R}) > 0$ **do**
        for $j \in \hat{R}$ **do**
            if $SECO(\hat{C}, j) > 0$ **then**
                move j from $\hat{R}$ to $\hat{C}$
                break for loop
            else
                j stays in $\hat{R}$
    return $\hat{C}$, $\hat{R}$

Theorem 2. *The algorithm outputs the correct clustering at $O(d^3)$ oracle calls.*

Proof We shall first show that the procedure for splitting a set into two indeed splits set into two independent sets, if and only if such two sets exists. First, if $SECO(\hat{O}_1, \hat{O}_2) = 0$, then $\mathbf{X}^{(O_1)} \perp\!\!\!\perp \mathbf{X}^{(O_2)}$ and the function terminates. In the opposite case, when $SECO(\hat{O}_1, \hat{O}_2) > 0$, by Assumption A, there exist an element $j \in S$ such $SECO(\hat{O}_1, j) > 0$. This process will continue as long as one element is not attached to its original cluster by Proposition 2. Since there are only finitely many elements in S, the while loop eventually terminates at d iteration. The algorithm will return $\hat{O}_1 = \bar{O}_1$ and $\hat{O}_2$ is independent of the first group and thus $SECO(\hat{O}_1, \hat{O}_2) = 0$. Finally, notice that if $(C_1, C_2) \perp\!\!\!\perp C_3$ and $C_1 \perp\!\!\!\perp C_2$ then also $C_1 \perp\!\!\!\perp C_2 \perp\!\!\!\perp C_3$, which means that by repeating the Split function recursively, we find the correct clustering. □

4 Consistent estimation of minimally separated clusters via SECO

In this section, we propose an estimation approach that utilizes nonparametric estimation of the Pickands dependence function and we use it to recover clusters. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be *i.i.d.* copies

of $\mathbf{X}$. The estimator that we present is based on the madogam concept, a notion borrowed from geostatistics in order to capture the spatial dependence structure. Our estimator is defined as

$$\hat{A}_n(\mathbf{t}) = \frac{\hat{\nu}_n(\mathbf{t}) + c(\mathbf{t})}{1 - \hat{\nu}_n(\mathbf{t}) - c(\mathbf{t})}, \tag{12}$$

where

$$\hat{\nu}_n(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n}\left[\bigvee_{j=1}^{d}\{G_{n,j}(X_{i,j})\}^{1/t_j} - \frac{1}{d}\sum_{j=1}^{d}\{G_{n,j}(X_{i,j})\}^{1/t_j}\right],$$
$$c(\mathbf{t}) = \frac{1}{d}\sum_{j=1}^{d}\frac{t_j}{1+t_j}.$$

by convention, here $u^{1/0} = 0$ for $u \in (0,1)$. One can use our results presented to design a new test of stochastic vectorial independence. Let $k \in \{1,\ldots,K\}$ and let $\hat{A}_n^{(O_k)}(\mathbf{t}^{(O_k)}) = \hat{A}_n(\mathbf{0}, \mathbf{t}^{(O_k)}, \mathbf{0})$ denotes the empirical Pickands dependence function associated to the k-th subvector of X. Then consider the empirical process

$$\mathcal{E}_{nK}(\mathbf{t}) = \sqrt{n}\left(\hat{A}_n(\mathbf{t}) - \hat{A}_{n\Sigma}(\mathbf{t})\right), \tag{13}$$

where $\hat{A}_{n\Sigma} = \sum_{k=1}^{K} w^{(O_k)}(\mathbf{t})\hat{A}_n^{(O_k)}(\mathbf{t}^{(O_k)})$. We define the set of hypotheses :

$$\mathcal{H}_0 : O_1,\ldots,O_K \text{ are mutually independent,} \qquad \mathcal{H}_1 : \exists j \neq k\, O_j, O_k \text{ are mutually dependent.}$$

Theorem below states the asymptotic behaviour of our statistic test $\mathcal{E}_{nK}$.

Theorem 3. *Under $\mathcal{H}_0$, the empirical process $\mathcal{E}_{nK}$ converges weakly in $\ell^\infty(\Delta_{d-1})$ to a tight Gaussian process having representation*

$$\mathcal{E}_K(\boldsymbol{t}) = -(1+A(\boldsymbol{t}))^2\int_{[0,1]} N_C(u^{t_1},\ldots,u^{t_d})du + \sum_{k=1}^{K} w^{(O_k)}(\boldsymbol{t})\left(1+A^{(O_k)}(\boldsymbol{t}^{(O_k)})\right)^2\int_{[0,1]} N_C(\boldsymbol{1},u^{t_{i,1}},\ldots,u^{t_{i,d_k}},\boldsymbol{1})du,$$

where N_C is a continuous tight Gaussian process with representation

$$N_C(u_1,\ldots,u_d) = B_C(u_1,\ldots,u_d) - \sum_{j=1}^{d}\dot{C}_j(u_1,\ldots,u_d)B_C(\boldsymbol{1},u_i,\boldsymbol{1}),$$

and B_C is a continuous tight Gaussian process with covariance function

$$cov(B_C(\boldsymbol{u}),B_C(\boldsymbol{v})) = C(\boldsymbol{u}\wedge\boldsymbol{v}) - C(\boldsymbol{u})C(\boldsymbol{v}) \overset{\boldsymbol{X}\sim\mathcal{L}(\boldsymbol{X})}{=} C_\Pi(\boldsymbol{u}\wedge\boldsymbol{v}) - C_\Pi(\boldsymbol{u})C_\Pi(\boldsymbol{v})$$

Using previous results stated in the present document, the test is equivalent to test the hypothesis under $\mathbf{t} = (d^{-1}, \ldots, d^{-1})$, that is, whether the extremal coefficient of $\mathbf{X}$ is equal to the sum of the other extremal coefficients of $\mathbf{X}^{(O_1)}, \ldots, \mathbf{X}^{(O_K)}$. Furthermore, the asymptotic variance is easily (but technical) computable under this specific point.

We can estimate $\bar{O}$ by applying a multiple test procedure. This procedure makes perfect sense when n is large, much larger than d. But when d is large, the procedure will leads to poor results and induce multiple testing procedure. One way to indentify clusters is to introduce a treshold α which, if greater, then j is assigned to the corresponding cluster. In order to obtain theoretical guarantees of exact recovery with high probability, concentration bounds are of prime interest. To do so, we estimate SECO metric in (10) using $\hat{\theta}_n = \hat{A}_n(d^{-1}, \ldots, d^{-1})$. A concentration inequality is stated here that will be of interest in further analysis.

Proposition 3. *For $t > 0$, we have*

$$\mathbb{P}\left\{|\hat{\theta}_n - \theta| \geq t\right\} \leq 4d \exp\left\{-\frac{nt^2}{128d^2}\right\}.$$

In this paragraph, we present an algorithm that recover clusters in AI-block models. We recall that we set an index $j \in \{1, \ldots, d\}$. As we do not have knowledge of A, we need to estimate it from n observed independent copy $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X}$. to a given estimator $\hat{A}_n$ of A we associate the estimation

$$\widehat{SECO}(\hat{O}_1, j) = \hat{\theta}_n^{(\hat{O}_1)} + 1 - \hat{\theta}_n^{(\hat{O}_1 \cup \{j\})},$$

of the SECO metric. We then estimate the partition $\hat{O}$ according to the split procedure. We

Algorithm 2 Split procedure with A unknown

```
procedure SPLIT(S, α, Â_n)
    Initialize : Ĉ := {1}, R̂ := {2, ..., d}, S = {0}
    while S ≠ ∅ do
        S = R̂
        for j ∈ R̂ do
            if SECO^(Ô_1, j) > α_{|Ô_1|+1} then
                Ĉ = Ĉ ∪ {j}, R̂ = R̂ \ {j}
                break for loop
            else
                R̂ = R̂
            S = S \ R̂
    return Ĉ, R̂
```

emphasize that this algorithm does not require as input the specification of the number K of groups. In the following, we provide conditions ensuring that $\hat{O} = \bar{O}$.

Proposition 4. *Consider the AI-block model with $d \geq 2$ and $K \leq d$. Let $U \subseteq \{1, \ldots, d\}$ with*

$|U| = s$ *for every* $s \in \{2, \dots, d\}$ *and define* $\tau_s = |\hat{\theta}_n^{(U)} - \theta^{(U)}|$ *and we consider parameters*

$$\alpha_s \geq 2\tau_s, \quad \eta_s \geq 4\tau_s. \tag{14}$$

Then if for every $k \in \{1, \dots, K\}$ *and* $s_k \in \{1, \dots, d-1\}$, $A^{(\bar{O}_k)} \in \mathcal{A}(\eta_{s_k})$ *and under Assumption A, our algorithm yields* $\hat{O} = \bar{O}$.

Corollary 1. *Let us consider parameters fulfilling*

$$\alpha_s \geq 8s\sqrt{\frac{2(1+\gamma)}{n}\ln\left(ds^{\frac{1}{1+\gamma}}\right)}, \quad \eta_s \geq 8s\sqrt{\frac{2(1+\gamma)}{n}\ln\left(ds^{\frac{1}{1+\gamma}}\right)} + \alpha_s,$$

for some $\gamma > 0$. *If* $\boldsymbol{X} \sim O$ *and for* $k \in \{1, \dots, K\}$ *and* $s_k \in \{1, \dots, d-1\}$, $A^{(\bar{O}_k)} \in \mathcal{A}(\eta_{s_k})$, *and under Assumption A, then the output of our algorithm applied to the estimator* $\hat{A}_n$, *is consistent :* $\hat{O} = \bar{O}$, *with probability higher than* $1 - d^{-3\gamma}$.

Remark 6. *Therefore, with* $\tau_s \leq C \times \sqrt{\frac{\ln(sd)}{n}}$, *with* $C > 0$ *a positive constant, thresholding* $\widehat{SECO}$ *at level* $2\tau_s$ *guarantees exact recovery, whenever the SECO separation* η_s *is at least* $4\tau_s$.

Note that above, independently of any distributional assumption, the $SECO$ metric and its related estimator $\widehat{SECO}$ is sufficient to recover the original clusters. This results is quite interesting, since it is stronger than what is known in the classical, non-extremal independence clustering [Ryabko, 2017]. Indeed, in classical theory, one ask the existence of density in order to compute the mutual information between two vector $\mathbf{X}^{(\hat{C})}$ and $\mathbf{X}^{(\hat{R})}$. Here, we just use a simple statistic, which can be estimated non parametrically and without the existence of densities to recover our clusters. A similar statement could be obtained for Gaussian models where the correlation coefficient between random variables gives knowledge about their dependences as the extremal coefficient gives for extreme value distributions. Such stronger results obtained in extreme value theory but not in the classical theory are also valid for trees, we refer to [Engelke and Volgushev, 2020] for further details.

Some comments on the implications of the above result are in order. On a high level, larger dimensions d lead to an higher bound of the threshold α_s for every considered cluster size s. This is also implicated for the bound of the $MSECO$ metric. Thus, have the dimension d increases, the dependence between each component should be stronger in order to distinguish between noise and a signal. Also, the threshold α_s increases with respect to its size. Thus, as our cluster grows and the algorithm is consistent up to this step, we set a greater threshold as we are capable to better distinguish between elements who are dependent of the candidate cluster and those who are not.

References

[Asenova et al., 2021] Asenova, S., Mazo, G., and Segers, J. (2021). Inference on extremal dependence in the domain of attraction of a structured hüsler–reiss distribution motivated by a markov tree with latent variables. *Extremes*, 24(3):461–500.

[Bador et al., 2015] Bador, M., Naveau, P., Gilleland, E., Castellà, M., and Arivelo, T. (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 and Resnick, 1977] Balkema, A. A. and Resnick, S. I. (1977). Max-infinite divisibility. *Journal of Applied Probability*, 14(2):309–319.

[Beirlant et al., 2004] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.

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

[Bücher et al., 2011] Bücher, A., Dette, H., and Volgushev, S. (2011). New estimators of the pickands dependence function and a test for extreme-value dependence. *The Annals of Statistics*, pages 1963–2006.

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

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

[Cooley et al., 2010] Cooley, D., Davis, R. A., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis*, 101(9):2103–2117.

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

[De Haan et al., 2006] De Haan, L., Ferreira, A., and Ferreira, A. (2006). *Extreme value theory: an introduction*, volume 21. Springer.

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

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

[Engelke and Volgushev, 2020] Engelke, S. and Volgushev, S. (2020). Structure learning for extremal tree models. *arXiv preprint arXiv:2012.06179*.

[Escobar-Bach et al., 2018] Escobar-Bach, M., Goegebeur, Y., and Guillou, A. (2018). Local robust estimation of the pickands dependence function. *The Annals of Statistics*, 46(6A):2806–2843.

[Falk et al., 2010] Falk, M., Hüsler, J., and Reiss, R. (2010). *Laws of Small Numbers: Extremes and Rare Events*. Springer Basel.

[Galambos, 1978] Galambos, J. (1978). The asymptotic theory of extreme order statistics. Technical report.

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

[Goix et al., 2015] Goix, N., Sabourin, A., Clémen, S., et al. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory*, pages 843–860. PMLR.

[Gudendorf and Segers, 2010] Gudendorf, G. and Segers, J. (2010). Extreme-value copulas. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, pages 127–145, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Gudendorf and Segers, 2012] Gudendorf, G. and Segers, J. (2012). Nonparametric estimation of multivariate extreme-value copulas. *Journal of Statistical Planning and Inference*, 142(12):3073–3085.

[Hofert et al., 2018] Hofert, M., Huser, R., and Prasad, A. (2018). Hierarchical archimax copulas. *Journal of Multivariate Analysis*, 167:195–211.

[Huang, 1992] Huang, X. (1992). *Statistics of bivariate extreme values*. Thesis Publishers Amsterdam.

[Hüsler and Reiss, 1989] Hüsler, J. and Reiss, R.-D. (1989). Maxima of normal random vectors: Between independence and complete dependence. *Statistics & Probability Letters*, 7(4):283–286.

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

[Lalancette et al., 2021] Lalancette, M., Engelke, S., and Volgushev, S. (2021). Rank-based estimation under asymptotic dependence and independence, with applications to spatial extremes. *The Annals of Statistics*, 49(5):2552–2576.

[Lauritzen, 1996] Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

[Le et al., 2018] Le, P. D., Davison, A. C., Engelke, S., Leonard, M., and Westra, S. (2018). Dependence properties of spatial rainfall extremes and areal reduction factors. *Journal of Hydrology*, 565:711–719.

[Lindskog, 2004] Lindskog, F. (2004). *Multivariate extremes and regular variation for stochastic processes*. PhD thesis, ETH Zurich.

[Marcon et al., 2017] Marcon, G., Padoan, S., Naveau, P., Muliere, P., and Segers, J. (2017). Multivariate nonparametric estimation of the pickands dependence function using bernstein polynomials. *Journal of Statistical Planning and Inference*, 183:1–17.

[Marshall and Olkin, 1983a] Marshall, A. W. and Olkin, I. (1983a). Domains of Attraction of Multivariate Extreme Value Distributions. *The Annals of Probability*, 11(1):168 – 177.

[Marshall and Olkin, 1983b] Marshall, A. W. and Olkin, I. (1983b). Domains of Attraction of Multivariate Extreme Value Distributions. *The Annals of Probability*, 11(1):168 – 177.

[Meyer and Wintenberger, 2021] Meyer, N. and Wintenberger, O. (2021). Sparse regular variation. *Advances in Applied Probability*, 53(4):1115–1148.

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

[Papastathopoulos and Strokorb, 2016] Papastathopoulos, I. and Strokorb, K. (2016). Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15.

[Pickands, 1981] Pickands, J. (1981). Multivariate extreme value distribution. *Proceedings 43th, Session of International Statistical Institution, 1981.*

[Pollard, 1981] Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140.

[Resnick, 2008] Resnick, S. (2008). *Extreme Values, Regular Variation, and Point Processes.* Applied probability. Springer.

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

[Ryabko, 2017] Ryabko, D. (2017). Independence clustering (without a matrix). In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

[Simpson et al., 2020] Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

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

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

[Tawn, 1990] Tawn, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253.

[van der Vaart et al., 1996] van der Vaart, A., van der Vaart, A., van der Vaart, A., and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

A Proofs

A.1 Proofs of main results

We show here that there exists a unique maximal element $\bar{O}$ in AI-block models. To do that, we show several items concerning the introduced the partition partial order and the partition induced by the equivalence relation $\overset{O\cap S}{\sim}$. Using these properties, we construct an explicit unique maximal element of $\mathcal{L}(\mathbf{X})$.

Proof of Theorem 1 For the first item, if $\mathbf{X} \sim O$, then the partition O induces mutually independent random vectors. As S is a sub-partition of O, it also generates a partition where vectors are mutually independent.

Now, take $k \in \{1, \dots, K\}$ and $j, k \in (O \cap S)_k$, we thus have in particular $j \overset{O}{\sim} k$, thus there exists $k' \in \{1, \dots, K'\}$ such that $a, b \in O'_k$. Thus $(O \cap S)_k \subseteq O_{k'}$. Hence the second statement.

The third result comes down from 1. and 2. We conclude the proof of this propostion by proving the fourth claim. The set $\mathcal{L}(\mathbf{X})$ is non-empty since the trivial partition $O = \{1, \dots, d\}$ belongs to $\mathcal{L}(\mathbf{X})$. It is also a finite set, and we can enumerate it $\mathcal{L}(\mathbf{X}) = \{O_1, \dots, O_m\}$. Define the sequence $O'_1, \dots, O'_M$ recursively according to

- $O'_1 = O_1$,
- $O'_k = O_k \cap O'_{k-1}$ for $k = 2, \dots, M$.

According to Theorem 1, we have that by induction $O'_1, \dots O'_M \in \mathcal{L}(\mathbf{X})$. In addition, we have both $O'_{k-1} \leq O'_k$ and $O_k \leq O'_k$, so by induction $O_1, \dots, O_k \leq O'_k$. Hence the partition $O^*(\mathbf{X}) := O'_M = O_1 \cap \dots \cap O_{M-1}$ is the maximum of $\mathcal{L}(\mathbf{X})$. □

We extend the well-known inequality $A \leq 1$ where inequality holds if and only if $\mathbf{X}$ is totally independent to the case of mutual independence, that is $A \leq A_O$ using notations of Section 2. Two ways are given to obtain this statement. The first use the convexity and homogeneity of order one

of the stable tail dependence function. The associativity of extreme-value random vectors are of a prime interest to obtain the statement in a second sketch.

Proof of Prosition 1 As L is an homogeneous and convex function under a cone, it is also subadditive, *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

$$L(\sum_{k=1}^{K} \mathbf{z}_k) \leq \sum_{k=1}^{K} L(\mathbf{z}_k),$$

where $\mathbf{x}_k \in [0,\infty)^d$ with $k \in \{1,\dots,K\}$. Consider now $\mathbf{x}_k = (\mathbf{0}, x_{i_{k,1}}, \dots, x_{i_{k,d_k}}, \mathbf{0})$, we directly obtain

$$L(\mathbf{z}) = L(\sum_{k=1}^{K} \mathbf{z}_k) \leq \sum_{k=1}^{K} L(\mathbf{z}_k) = \sum_{k=1}^{K} L^{(k)}(z_{i_{k,1}}, \dots, z_{i_{k,d_k}})$$

Expression this equation in terms of Pickands gives

$$A(\mathbf{t}) \leq \sum_{k=1}^{K} \frac{1}{z_1+\cdots+z_d} L^{(k)}(z_{i_{k,1}}, \dots, z_{i_{k,d_k}}) = \sum_{k=1}^{K} \frac{z_{i_{k,1}}+\cdots+z_{i_{k,d_k}}}{z_1+\cdots+z_d} A^{(O_k)}(t_{i_{k,1}}, \dots, t_{i_{k,d_k}}),$$

where $t_i = z_i/(z_1+\cdots+z_d)$. Hence the result.

Another proof is obtained using that extreme-value distributions are associated (see Proposition 5.1 of [Marshall and Olkin, 1983a] or Section 5.4.1 of [Resnick, 2008]), *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, we can prove that, $K \in \mathbb{N}_*$,

$$\mathbb{E}\left[\Pi_{k=1}^{K} f_k(\mathbf{X})\right] \geq \Pi_{k=1}^{K} \mathbb{E}\left[f_k(\mathbf{X})\right] \tag{15}$$

Take $f_k(\mathbf{x}) = \mathbb{1}_{\{]-\infty,\mathbf{x}_k]\}}$ for each $k \in \{1,\dots K\}$, thus Equation (15) gives

$$C(G_1(x_1), \dots, G_d(x_d)) \geq \Pi_{k=1}^{K} C^{(O_k)}\left(G^{(O_k)}\left(\mathbf{x}^{(O_k)}\right)\right),$$

which can be restated in terms of stable tail dependence function

$$L(\mathbf{z}) \leq \sum_{k=1}^{K} L^{(O_k)}(\mathbf{z}^{(O_k)}).$$

We obtain the statement expressing this inequality with Pickands dependence function.

For the rest of the proof, notice that (15) with $f_k(\mathbf{x}) = \mathbb{1}_{\{]-\infty,\mathbf{x}^{(O_k)}]\}}$ for each $k \in \{1,\dots K\}$ holds

as an inequality if and only if $\mathbf{X}^{(O_1)} \perp\!\!\!\perp \ldots \perp\!\!\!\perp \mathbf{X}^{(O_K)}$. □

We thus state the theoritical value taken by $\mathbf{X} \sim O$ for a given arbitrary partition called $\bigsqcup_{l=1}^{L} S_l$. We show that this value is strictly greater than 0 if and only if the considered partition does not induce mutually independent random vectors. Again, the proof makes of use of the associativity of extreme random vectors. To obtain the theoritical value, we decompose each element of the partition S_l, $l \in \{1, \ldots, L\}$ as a subset of $\bar{O}_k$. Using mutual independence, we thus obtain the closed form of the theoretical value through the Pickands dependence function.

Proof of Proposition 2 Note that $\mathbf{X}^{(S_l)} = (\mathbf{X}_{j_{l,1}}, \ldots, \mathbf{X}_{j_{l,D_l}})$ and $\mathbf{X}^{(S_m)} = (\mathbf{X}_{j_{m,1}}, \ldots, \mathbf{X}_{j_{m,D_m}})$ are not, in general independent where $l, m \in \{1, \ldots L\}$ because the partition $\bigsqcup_{l=1}^{L} S_l$ is arbitrarily given. Consider for some $l \in \{1, \ldots, L\}$,

$$f_l = \mathbb{1}_{\{]-\infty, \mathbf{x}^{(S_l)}]\}}, \quad \mathbf{x}^{(S_l)} = (x_{j_{l,1}}, \ldots, x_{j_{l,D_l}}).$$

We thus have in one hand

$$\mathbb{E}\left[\Pi_{l=1}^{L} f_l(\mathbf{X})\right] = \mathbb{P}\left\{X_1 \leq x_1, \ldots, X_d \leq x_d\right\} = \Pi_{k=1}^{K} C^{(O_k)}\left(G^{(O_k)}(\mathbf{x}^{(O_k)})\right),$$

on the other hand

$$\Pi_{l=1}^{L} \mathbb{E}\left[f_l(\mathbf{X})\right] = \Pi_{l=1}^{L} \mathbb{P}\left\{X_{j_{l,1}} \leq x_{j_{l,1}}, \ldots, X_{j_{l,D_l}} \leq x_{j_{l,D_l}}\right\}.$$

It is worth noticing that the above set can be rewriten as

$$\left\{X_{j_{l,1}} \leq x_{j_{l,1}}, \ldots, X_{j_{l,D_l}} \leq x_{j_{l,D_l}}\right\} = \bigcap_{k=1}^{K} \bigcap_{i \in S_l \cap O_k} \left\{X_i \leq x_i\right\}.$$

Notice that $\bigcap_{i \in \emptyset} \left\{X_i \leq x_i\right\} = \Omega$. As a subvector of $\mathbf{X}^{(k)}$, $\mathbf{X}_{kl} = (X_i)_{i \in S_l \cap O_k}$ is independent of $\mathbf{X}^{(j)}$ for $j \neq k$. Using this block-independence, we thus obtain :

$$\mathbb{P}\left\{X_{j_{l,1}} \leq x_{j_{l,1}}, \ldots, X_{j_{l,D_l}} \leq x_{j_{l,D_l}}\right\} = \Pi_{k=1}^{K} \mathbb{P}\left\{\bigcap_{i \in S_l \cap O_k} \left\{X_i \leq x_i\right\}\right\}.$$

Now using positive association, we obtain that

$$\Pi_{k=1}^{K} C^{(O_k)}\left(G^{(O_k)}(x^{(O_k)})\right) \geq \Pi_{l=1}^{L} \Pi_{k=1}^{K} C^{(O_k)}\left(\mathbf{x}_{S_l \cap O_k}\right),$$

where $\mathbf{x}_{kl} = (x_i)_{i \in S_l \cap O_k}$. Reexpressing the whole in terms of stable tail dependence function leads to :

$$\sum_{k=1}^{K} L^{(O_k)}(z_{i_{k,1}}, \ldots, z_{k,d_k}) \leq \sum_{l=1}^{L} \sum_{k=1}^{K} L^{(k)}(\mathbf{0}, \mathbf{z}^{(S_l \cap O_k)}, \mathbf{0}),$$

with $z_i = -\ln(G_i(x_i))$ for every $i \in \{1, \dots, d\}$. Dividing by $z_1 + \cdots + z_d$ gives

$$A_O(\mathbf{t}) = \sum_{k=1}^{K} w^{(O_k)}(\mathbf{t}) A^{(O_k)}(\mathbf{t}^{(O_k)}) \leq \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{1}{z_1 + \cdots + z_d} L^{(k)}(\mathbf{0}, \mathbf{z}^{(S_l \cap O_k)}, \mathbf{0}).$$

The right hand side of the equation can be written as $\forall k \in \{1, \dots, K\}$, $\forall l \in \{1, \dots, L\}$

$$\begin{aligned} \frac{1}{z_1 + \cdots + z_d} L^{(k)}(\mathbf{0}, \mathbf{z}^{(S_l \cap O_k)}, \mathbf{0}) &= \frac{\sum_{j \in S_l \cap O_k} z_j}{z_1 + \cdots + z_d} L^{(k)} \left(\mathbf{0}, \frac{\mathbf{z}^{(S_l \cap O_k)}}{\sum_{j \in S_l \cap O_k} z_j}, \mathbf{0} \right) \\ &\triangleq w^{(S_l \cap O_k)}(\mathbf{t}) A^{(O_k)}(\mathbf{0}, \mathbf{t}^{(S_l \cap O_k)}, \mathbf{0}). \end{aligned}$$

We thus obtain the statement.

Now, if $\forall k \in \{1, \dots, K\}$, $\exists l \in \{1, \dots, L\}$ such that $O_k \subseteq S_l$, then the random vectors $\mathbf{X}^{(S_l)} = (\mathbf{X}_{j_{l,1}}, \dots, \mathbf{X}_{j_{l,D_l}})$ and $\mathbf{X}^{(S_m)} = (\mathbf{X}_{j_{m,1}}, \dots, \mathbf{X}_{j_{m,D_m}})$ are now independent. If we suppose $L \geq K$, we thus have $S_1 = O_1, \dots, S_K = O_K$ and $S_l = \emptyset$ for $l > K$. The equality comes down from Lemma 1. Now, if $L < K$, we have with the same notations in the proof

$$\mathbb{E}\left[\Pi_{l=1}^{L} f_l(\mathbf{X})\right] = \Pi_{l=1}^{L} \mathbb{P}\left\{X_{j_{l,1}} \leq x_{j_{l,1}}, \dots, X_{j_{l,D_l}} \leq x_{j_{l,D_l}}\right\} = \Pi_{k=1}^{K} \mathbb{P}\left\{\mathbf{X}^{(k)} \leq \mathbf{x}_k\right\}.$$

Expressing this two terms as before, we obtain that $A = A_O$. For the converse, suppose that the right hand side of the inequality in (5) is equal to zero. Applying Lemma 1 gives that the arbitrary partition has the same value as A_O. That is saying that the random vectors inside $S_1, \dots S_L$ are mutually independent. If $L \geq K$, thus, apart for the K first clusters say for which $S_k = O_k$, the others are empty set. Now, if $L < K$, we group one or more O_k in a given cluster S_l without one overlaps to an other cluster S_j say (if it does, the value could not be equal as zero by Lemma 1). Hence the statement. □

To prove the weak convergence of our process given in Theorem 3, we make of use of empirical processes as stated in [van der Vaart et al., 1996].

Proof The proof is straightforward, notice that (see Figure 1)

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

where ϕ is detailed as

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

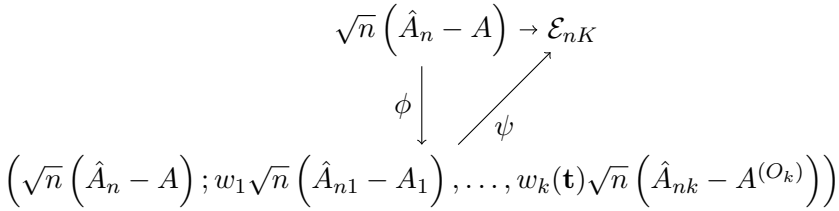

Figure 1: Diagram of composition of function.

with for every $k \in \{1, \ldots, K\}$

$$\begin{array}{rccc} \phi_k & : & \ell^\infty(\Delta_{d-1}) & \to & \ell^\infty(S_d) \\ & & x & \mapsto & x(\mathbf{0}, t_{i_{k,1}}, \ldots, t_{i_{k,d_k}}, \mathbf{0}). \end{array}$$

Thus ϕ_k is a linear and bounded function hence continuous, it follows that ϕ is continuous since each coordinate functions are continuous. Using that (see [Marcon et al., 2017])

$$\sqrt{n}(\hat{A}_n(\mathbf{t}) - A(\mathbf{t})) \rightsquigarrow -(1 + A(\mathbf{t}))^2 \int_{[0,1]} N_C(u^{t_1}, \ldots, u^{t_d}) du,$$

and applying the continuous mapping theorem for the weak convergence in $\ell^\infty(\Delta_{d-1})$ (Theorem 1.3.6 of [van der Vaart et al., 1996]) leads the result. □

We now present all tools used obtain the exact recovery of our algorithm, that is $\hat{O} = \bar{O}$. The result is obtained by induction on step l while we assume that the algorithm remains consistent at this step. If the algorithm wants to recover the first cluster $\bar{O}_1$ say, at step $l-1$ with an estimator $\hat{O}_1 \subset \bar{O}_1$ of size $|\hat{O}_1| = s_1 \leq d_1$. Take $j \in S \setminus \hat{O}_1$, we show that under conditions (14) and the cluster separation condition, that is $A^{(\bar{O}_1)} \in \mathcal{A}(\eta_{s_1+1})$, we have that :

$$j \in \bar{O}_1 \iff \widehat{SECO}(\hat{O}_1, j) > \alpha_{s_1+1}.$$

And thus the algorithm remains consistent at step l.

Proof of Proposition 4 Without loss of generality, take $\bar{O}_1$ and we will show under conditions (14) and the separation condition, our algorithm remain consistent with any size of $\hat{O}_1$. Suppose without loss of generality that $\hat{O}_1 = \{i_{1,1}\}$ and thus $|\hat{O}_1| = 1$. Consider $j \in S \setminus \hat{O}_1$. We then observe that $j \notin \bar{O}_1 \implies SECO(\hat{O}_1, j) = 0$, and thus

$$\widehat{SECO}(\hat{O}_1, j) \leq 2\tau_2.$$

When the group separation at size $s_1 = 1$ holds, that is $A^{(O_1)} \in \mathcal{A}(\eta_2)$, we have that

$$j \in \bar{O}_1 \implies SECO(\hat{O}_1, j) > 4\tau_2 \implies \widehat{SECO}(\hat{O}_1, j) > 2\tau_2.$$

In particular, we have

$$j \in \bar{O}_1 \quad \Longleftrightarrow \quad \widehat{SECO}(\hat{O}_1, j) > \alpha_2.$$

And the algorithm is consistent when $|\hat{O}_1| = 1$. Now we consider the algorithm at some step $l-1$ and assume that the algorithm was consistent up to this step, *i.e.* $\hat{O}_1 \subset \bar{O}_1$ with $|\hat{O}_1| = s_1 \leq d_1$. Consider $j \in S \setminus \hat{O}_1$, at this step, we have :

$$j \notin \bar{O}_1 \implies SECO(\hat{O}_1, j) = 0 \implies \widehat{SECO}(\hat{O}_1, j) \leq 2\tau_{s_1+1}.$$

Under the group separation condition $A^{(O_1)} \in \mathcal{A}(\eta_{s_1+1})$, we obtain that

$$j \in \bar{O}_1 \implies SECO(\hat{O}_1, j) > 4\tau_{s_1+1} \implies \widehat{SECO}(\hat{O}_1, j) > 2\tau_{s_1+1}.$$

Thus, under conditions (14) and the group separation condition, we have

$$j \in \bar{O}_1 \quad \Longleftrightarrow \quad \widehat{SECO}(\hat{O}_1, j) > \alpha_{s_1+1}.$$

Thus, the algorithm remains consistent at step l and exact recovery of the first cluster follows by induction.

Now, exact recovery of all cluster follows also by induction. For $K = 1$, the result is given below. Suppose that at some step $K-1$, the algorithm was consistent up to this step, that is $\hat{O}_j = \bar{O}_j$ for every $j \in \{1, \ldots, K\}$. Proceeding as below for the fist cluster, under conditions (14) and the separation condition, we obtain

$$\hat{O}_K = \bar{O}_K,$$

and the proposition follows by induction. □

By denoting by ALG, the set of $U \subseteq \{1, \ldots, d\}$ such that $|U| = s$ with $s \in \{1, \ldots, d-1\}$ and U is used by the split function, we will prove that Proposition 4 holds with high probabilty for given parameters. Indeed, by Theorem 1, we know that $|ALG| \leq d^3$. We thus need to specify some threshold τ_s such that $|\hat{\theta}_n^{(U)} - \theta^{(U)}| \leq \tau_s$ with high probability. To do so, we make use of concentration inequality stated in Proposition 3 that gives concentration of an estimator of the extremal coefficient for a given size.

Proof of Corollary 1 Following the notation introduced below, we have that for $t > 0$:

$$\mathbb{P}\left\{ \bigcup_{U \in ALG} |\hat{\theta}_n^{(U)} - \theta^{(U)}| \geq t \right\} \leq \sum_{U \in ALG} \mathbb{P}\left\{ |\hat{\theta}_n^{(U)} - \theta^{(U)}| \geq t \right\}.$$

Using Proposition 3, one has

$$\mathbb{P}\left\{ |\hat{\theta}_n^{(U)} - \theta^{(U)}| \geq t \right\} \leq 4s \exp\left\{ -\frac{nt^2}{128 s^2} \right\},$$

where $|U| = s$. By considering $\delta \in]0,1[$ and solve the following equation

$$\frac{\delta}{d^3} = 4s \exp\left\{-\frac{nt^2}{128s^2}\right\},$$

with respect to t gives that :

$$\mathbb{P}\left\{\bigcup_{U\in ALG} |\hat{\theta}_n^{(U)} - \theta^{(U)}| \geq 8s\sqrt{\frac{2}{n}\ln\left(\frac{2sd^3}{\delta}\right)}\right\} \leq \delta.$$

Now, taking $\delta = 4d^{-3A}$, we have that for every $U \in ALG$

$$|\hat{\theta}_n^{(U)} - \theta^{(U)}| \leq 8s\sqrt{\frac{2(1+A)}{n}\ln\left(s^{\frac{1}{1+A}}d\right)},$$

with probability higher than $1 - d^{-3A}$. The result then follows from Proposition 4, since for every $s \in \{1,\dots,d-1\}$, $\tau_s \leq 8s\sqrt{\frac{2(1+A)}{n}\ln\left(s^{\frac{1}{1+A}}d\right)}$ with probability higher than $1 - d^{-3A}$. □

B Proofs of auxiliary results

B.1 extreme-value copula

In this first lemma, we prove that the function introduced in Paragraph 2.2 is indeed an extreme-value copula. For the ease of reading, we recall here its definition

$$\begin{array}{rccl} C_\Pi : & [0,1]^d & \longrightarrow & [0,1] \\ & \mathbf{u} & \longmapsto & \Pi_{k=1}^K C^{(k)}(u_{i_{k,1}},\dots,u_{i_{k,d_k}}). \end{array}$$

To prove this statement, we show that each margins is indeed distributed uniformly on the unit segment $[0,1]$. Hence C is a copula function. In order to prove that C is an extreme-value copula, we show that C is max-stable as it is a characterizing property of extreme-value copula or, more generally, of extreme-value distribution.

Proof of Lemma 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 marginals are uniformly distributed on $[0,1]$. Without loss of generaly, take $u_{i_{1,1}} \in [0,1]$ and let us compute

$$C(1,\dots,u_{i_{1,1}},\dots,1) = C^{(1)}(u_{i_{1,1}},1,\dots,1)\Pi_{k=1}^K C^{(k)}(1,\dots,1) = C^{(1)}(u_{i_{1,1}},1,\dots,1) = u_{i_{1,1}}.$$

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 $m \geq 1$

$$C(u_1,\dots,u_d) = C(u_1^{1/m},\dots,u_d^{1/m})^m.$$

By definition, we have

$$C(u_1^{1/m},\dots,u_d^{1/m})^m = \left(\Pi_{k=1}^K C^{(k)}\left(u_{i_{k,1}}^{1/m},\dots,u_{i_{k,d_k}}^{1/m}\right)\right)^m = \Pi_{k=1}^K\left\{C^{(k)}\left(u_{i_{k,1}}^{1/m},\dots,u_{i_{k,d_k}}^{1/m}\right)\right\}^m.$$

Using that $C^{(1)},\dots,C^{(K)}$ are extreme-value copulae, thus max stable, we obtain

$$C(u_1^{1/m},\dots,u_d^{1/m})^m = \Pi_{k=1}^K C^{(k)}\left(u_{i_{k,1}},\dots,u_{i_{k,d_k}}\right) = C(u_1,\dots,u_d).$$

Thus C is an extreme-value copula. We end the proof by proving that C is associated to the random vector $\mathbf{X} = (\mathbf{X}^{(1)},\dots,\mathbf{X}^{(K)})$, that is

$$\mathbb{P}\left\{\mathbf{X}\le\mathbf{x}\right\} = C(G_1(x_1),\dots,G_d(x_d)),\quad \mathbf{x}\in\mathbb{R}^d.$$

Using mutual independence between random vectors, we have

$$\begin{aligned}\mathbb{P}\left\{\mathbf{X}\le\mathbf{x}\right\} &= \Pi_{k=1}^K\mathbb{P}\left\{X_{i_{k,1}}\le x_{i_{k,1}},\dots,X_{i_{k,d_k}}\le x_{i_{k,d_k}}\right\}\\ &= \Pi_{k=1}^K C^{(k)}\left(G_{i_{k,1}}(x_{i_{k,1}}),\dots,X_{i_{k,d_k}}\le x_{i_{k,d_k}}\right)\\ &= C(G_1(x_1),\dots,G_d(x_d)).\end{aligned}$$

Hence the result. □

B.2 Proof of proposition 3

Technical details of this proof will be subdivised in some lemmas of which the combined use will gives the statement of Proposition 3. The first lemma gives an upper bound of $|\hat{\theta}_n-\theta|$ with respect to $|\hat{\nu}_n(d^{-1},\dots,d^{-1})-\nu(d^{-1},\dots,d^{-1})|$. This follows from the link between the Pickands dependence function and the madogram.

Lemma 2. *We have,*

$$|\hat{\theta}_n-\theta|\le 4d|\hat{\nu}_n(d^{-1},\dots,d^{-1})-\nu(d^{-1},\dots,d^{-1})|.$$

Proof Fix $\mathbf{t}\in\Delta_{d-1}$, remember that $A(\mathbf{t}) = f(\nu(\mathbf{t}))$ and $\hat{A}_n(\mathbf{t}) = f(\hat{\nu}_n(\mathbf{t}))$, where $f:\mathbb{R}_+\to\mathbb{R}_+, x\mapsto(x+c(\mathbf{t}))/(1-x-c(\mathbf{t}))$ for every $\mathbf{t}\in\Delta_{d-1}$ and $c(\mathbf{t})$ is a constant equals to $d^{-1}\sum_{j=1}^d t_j/(1+t_j)$. Using that $A(\mathbf{t})\le 1$, we have that

$$\nu(\mathbf{t})+c(\mathbf{t})\le 1-\nu(\mathbf{t})-c(\mathbf{t}).$$

We obtain that

$$\nu(\mathbf{t})\le\frac{1}{2}-c(\mathbf{t})<1-c(\mathbf{t}).$$

In particular $1-\nu(\mathbf{t})-c(\mathbf{t}) \geq 2^{-1} > 0$. Now, taking derivation, we directly have for every $x \in [0, 2^{-1} - c(\mathbf{t})]$

$$|f'(x)| = \frac{1}{(1-x-c(\mathbf{t}))^2} \leq 4.$$

Thus, f is 4-Lipschitz on $[0, 1-c(\mathbf{t})]$ and in particular for $\mathbf{t} = (d^{-1}, \ldots, d^{-1})$

$$|\hat{A}_n(d^{-1}, \ldots, d^{-1}) - A(d^{-1}, \ldots, d^{-1})| \leq 4|\hat{\nu}_n(d^{-1}, \ldots, d^{-1}) - \nu(d^{-1}, \ldots, d^{-1})|$$

Multiply by d gives the statement. □

Now, we state a concentration inequality for the madogram estimator. This inequality is obtained through two main arguments, that are Hoeffding's inequality and the DKW inequality bound.

Lemma 3. *For $t > 0$ and a fixed $\boldsymbol{t} \in \Delta_{d-1}$, one has*

$$\mathbb{P}\left\{|\nu_n(\boldsymbol{t}) - \nu(\boldsymbol{t})| > t\right\} \leq 4d \exp\left\{-\frac{nt^2}{8}\right\}.$$

Proof observe that

$$|\hat{\nu}_n(\mathbf{t}) - \nu(\mathbf{t})| \leq |\hat{\nu}_n(\mathbf{t}) - \nu_n(\mathbf{t})| + |\nu_n(\mathbf{t}) - \nu(\mathbf{t})|,$$

where

$$\nu_n(\mathbf{t}) := \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \frac{1}{n}\left[\bigvee_{j=1}^{d} \{F_j(X_{i,j})\}^{1/t_j} - \frac{1}{d}\sum_{j=1}^{d} \{F_j(X_{i,j})\}^{1/t_j}\right].$$

As the following inequalities holds for every $i \in \{1, \ldots, n\}$

$$Y_i \leq \frac{(d-1)}{dn}.$$

Hoeffding's inequality applies and we obtain that

$$\mathbb{P}\left\{|\nu_n(\mathbf{t}) - \nu(\mathbf{t})| > t/2\right\} \leq 2\exp\left(-\frac{nd^2t^2}{4(d-1)^2}\right).$$

Furthermore, we have

$$|\hat{\nu}_n(\mathbf{t}) - \nu_n(\mathbf{t})| \leq 2 \sup_{j\in\{1,\ldots,d\}} \sup_{i\in\{1,\ldots,n\}} \left|\left\{\hat{F}_{n,j}(X_{i,j})\right\}^{1/t_j} - \{F_j(X_{i,j})\}^{1/t_j}\right|$$

Applying DKW inequality, we obtain

$$\mathbb{P}\left\{\sup_{j\in\{1,\ldots,d\}} \sup_{i\in\{1,\ldots,n\}} \left|\left\{\hat{F}_{n,j}(X_{i,j})\right\}^{1/t_j} - \{F_j(X_{i,j})\}^{1/t_j}\right| > \frac{t}{4}\right\} \leq 2d\exp\left(\frac{-nt^2}{8}\right).$$

We thus have for $d \geq 2$

$$\mathbb{P}\left\{|\hat{\nu}_n(\mathbf{t}) - \nu(\mathbf{t})| > t\right\} \leq 4d \exp\left(-\frac{nt^2}{8}\right).$$

Hence the statement. □

Combine Lemma 2 and Lemma 3 gives Proposition 3.

B.3 Asymptotic independence between multivariate extreme distribution

In this subsection, we extend the result given in Theorem 2.1 of [Takahashi, 1994] for asymptotic independence between extreme random vector. The used arguments are similar of those used in the proof in [Takahashi, 1994]. We make extensive use of the following result (see, for example [Marshall and Olkin, 1983b] and the proof of Theorem 5.3.1 of [Galambos, 1978]) *i.e.* $F \in D(G)$ is equivalent to

$$\lim_{n\to\infty} n\left\{1 - F(\mathbf{a}_n\mathbf{x} + \mathbf{b}_n)\right\} = -\ln G(\mathbf{x}) \tag{16}$$

for all $\mathbf{x}$ such that $0 < G(\mathbf{x}) < 1$. In this section, we denote by $\bar{F}$ the survival function of F.

Theorem 4. *Let F be a d-distribution function and let $G^{(O_i)}$ be a d_i-extreme-value distribution for $i = 1, 2$. Then for $\boldsymbol{a}_n > 0$ and $\boldsymbol{b}_n \in \mathbb{R}^d$*

$$\{F(\boldsymbol{a}_n\boldsymbol{x} + \boldsymbol{b}_n)\}^n \underset{n\to\infty}{\longrightarrow} G^{(O_1)}(\boldsymbol{x}^{(O_1)})G^{(O_2)}(\boldsymbol{x}^{(O_2)}) \tag{17}$$

if and only if

$$\left\{F^{(O_i)}(\boldsymbol{a}_n^{(O_i)}\boldsymbol{x}^{(O_1)} + b_n^{(O_i)})\right\}^n \underset{n\to\infty}{\longrightarrow} G^{(O_i)}(\boldsymbol{x}^{(O_i)}), \tag{18}$$

and there exists a $\boldsymbol{p} = (\boldsymbol{p}^{(O_1)}, \boldsymbol{p}^{(O_2)}) \in \mathbb{R}^d$ such that $0 < H^{(O_1)}(\boldsymbol{x}^{(O_1)}), H^{(O_2)}(\boldsymbol{x}^{(O_2)}) < 1$ and

$$\{F(\boldsymbol{a}_n\boldsymbol{x} + \boldsymbol{b}_n)\}^n \underset{n\to\infty}{\longrightarrow} G^{(O_1)}(\boldsymbol{p}^{(O_1)})G^{(O_2)}(\boldsymbol{p}^{(O_2)}) \tag{19}$$

Proof The proof follows exactly the same lines as in Theorem 2.1 of [Takahashi, 1994]. One substantial difference is emphasizes in Remark. For any $\mathbf{x} \in \mathbb{R}^d$, $0 < H^{(O_1)}(\mathbf{x}^{(O_1)}), H^{(O_2)}(\mathbf{x}^{(O_2)}) < 1$, there exists $s > 0$ such that $\{H^{(O_1)}(\mathbf{x}^{(O_1)})\}^{1/s} > H^{(O_1)}(\mathbf{p}^{(O_1)})$, $\{H^{(O_2)}(\mathbf{x}^{(O_2)})\}^{1/s} > H^{(O_2)}(\mathbf{p}^{(O_2)})$. By Equation (18)

$$\left\{F^{(O_i)}\left(\mathbf{a}_{\lfloor sn\rfloor}^{(O_i)}\mathbf{x}^{(O_i)} + b_{\lfloor sn\rfloor}^{(O_i)}\right)\right\}^{sn} \underset{n\to\infty}{\longrightarrow} H^{(O_i)}(\mathbf{x}^{(O_i)})$$

thus

$$\left\{F^{(O_i)}\left(\mathbf{a}_{\lfloor sn\rfloor}^{(O_i)}\mathbf{x}^{(O_i)} + b_{\lfloor sn\rfloor}^{(O_i)}\right)\right\}^{n} \underset{n\to\infty}{\longrightarrow} \left\{H^{(O_i)}(\mathbf{x}^{(O_i)})\right\}^{1/s}.$$

Notice that

$$\begin{aligned}\mathbb{P}\left\{\left[\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}, \mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right]^c\right\} &= \mathbb{P}\left\{\left[\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}\right]^c \cup \left[\mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right]^c\right\} \\ &= \mathbb{P}\left\{\left[\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}\right]^c\right\} + \mathbb{P}\left\{\left[\mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right]^c\right\} \\ &\quad - \mathbb{P}\left\{\left[\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}\right]^c \cap \left[\mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right]^c\right\}.\end{aligned}$$

We thus obtain

$$\begin{aligned}\mathbb{P}\left\{\left[\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}\right]^c \cap \left[\mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right]^c\right\} &= \left[1 - \mathbb{P}\left\{\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}\right\}\right] + \left[1 - \mathbb{P}\left\{\mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right\}\right] \\ &\quad - \left[1 - \mathbb{P}\left\{\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}, \mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}\right\}\right].\end{aligned}$$

Using Equations (18) and (19) we have, in joint hands, that

$$\begin{aligned}&n\left[1 - \mathbb{P}\left\{\mathbf{X}^{(O_i)} \leq \mathbf{a}_n^{(O_i)}\mathbf{p}^{(O_i)} + \mathbf{b}_n^{(O_i)}\right\}\right] \underset{n\to\infty}{\longrightarrow} -\ln G^{(O_i)}(\mathbf{p}^{(O_i)}), \quad i = 1, 2 \\ &n\left[1 - \mathbb{P}\left\{\mathbf{X} \leq \mathbf{a}_n\mathbf{p} + \mathbf{b}_n\right\}\right] \underset{n\to\infty}{\longrightarrow} -\ln G^{(O_1)}(\mathbf{p}^{(O_1)})G^{(O_2)}(\mathbf{p}^{(O_2)}).\end{aligned}$$

Thus,

$$\mathbb{P}\left\{\left[\mathbf{X}^{(O_1)} \leq \mathbf{a}^{(O_1)}\mathbf{p}^{(O_1)} + \mathbf{b}_n^{(O_1)}\right]^c \cap \left[\mathbf{X}^{(O_2)} \leq \mathbf{a}^{(O_2)}\mathbf{p}^{(O_2)} + \mathbf{b}_n^{(O_2)}\right]^c\right\} \underset{n\to\infty}{\longrightarrow} 0.$$

Using now that $\left\{\mathbf{X}^{(O_1)} > \mathbf{x}^{(O_1)}, \mathbf{X}^{(O_2)} > x^{(O_2)}\right\} \subset \left\{[\mathbf{X}^{(O_1)} \leq \mathbf{x}^{(O_1)}]^c \cap [\mathbf{X}^{(O_2)} \leq \mathbf{x}^{(O_2)}]^c\right\}$, we obtain

$$n\bar{F}(\mathbf{a}_n\mathbf{p} + \mathbf{b}_n) \underset{n\to\infty}{\longrightarrow} 0.$$

By $\mathbf{q} \geq \mathbf{p}$, we now have

$$0 \leq n\bar{F}(\mathbf{a}_n\mathbf{q} + \mathbf{b}_n) \leq n\bar{F}(\mathbf{a}_n\mathbf{p} + \mathbf{b}_n) \underset{n\to\infty}{\longrightarrow} 0.$$

The rest of the proof is similar to [Takahashi, 1994]. □

Remark 7. *When $d_1 = d_2 = 1$, we immediately have that :*

$$\mathbb{P}\{X_1 > x_1, X_2 > x_2\} = [1 - \mathbb{P}\{X_1 \leq x_1\}] + [1 - \mathbb{P}\{X_2 \leq x_2\}] - [1 - \mathbb{P}\{X_1 \leq x_1, X_2 \leq x_2\}].$$

We immediately obtain that, under the same hypotheses of Theorem 4 that

$$n\bar{F}(\boldsymbol{a}_n\boldsymbol{p} + \boldsymbol{b}_n) \underset{n\to\infty}{\longrightarrow} 0. \tag{20}$$

The arguments exposed in this remark are those used in the proof of Theorem 2.1 [Takahashi, 1994]. In our work framework, we do not directly obtain (20) but we can upper bound this quantity with respect to an other which indeedly converges to 0 as $n \to \infty$ in the framework of Theorem 4.