Multivariate sparse clustering for extremes

Nicolas Meyer, Olivier Wintenberger

Multivariate sparse clustering for extremes

Nicolas Meyer*
IMAG, Univ. Montpellier, CNRS, Montpellier, France
LEMON, Inria, Montpellier, France
and
Olivier Wintenberger†
Sorbonne Université, LPSM, F-75005, Paris, France
Wolfgang Pauli Institut, c/o Fakultät für Mathematik,
Universität Wien, 1090 Vienna, Austria

January 6, 2023

Abstract

Identifying directions where extreme events occur is a major challenge in multivariate extreme value analysis. In this paper, we use the concept of sparse regular variation introduced by Meyer and Wintenberger (2021) to infer the tail dependence of a random vector $\mathbf{X}$. This approach relies on the Euclidean projection onto the simplex which better exhibits the sparsity structure of the tail of $\mathbf{X}$ than the standard methods. Our procedure based on a rigorous methodology aims at capturing clusters of extremal coordinates of $\mathbf{X}$. It also includes the identification of the threshold above which the values taken by $\mathbf{X}$ are considered as extreme. We provide an efficient and scalable algorithm called MUSCLE and apply it on numerical examples to highlight the relevance of our findings. Finally we illustrate our approach with financial return data.

Keywords: Euclidean projection onto the simplex, model selection, multivariate extremes, regular variation

*nicolas.meyer@umontpellier.fr (corresponding author)
†olivier.wintenberger@sorbonne-universite.fr

1 Introduction

The aim of this article is to study the tail dependence of a random vector $\mathbf{X} \in \mathbb{R}_+^d$ with continuous marginals. In this context it is customary to assume that $\mathbf{X}$ is regularly varying (see e.g. Resnick (1987), Resnick (2007), Hult and Lindskog (2006)), i.e. that there exist $a_n \to \infty$ and a non-zero Radon measure μ on the Borel σ-field of $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ such that

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{v} \mu(\cdot)\,, \quad n \to \infty\,, \tag{1.1}$$

where $\xrightarrow{v}$ denotes the vague convergence in the space of nonnegative Radon measures on $[0,\infty]^d \setminus \{\mathbf{0}\}$. The limit measure μ is called the *tail measure* of the regularly varying vector $\mathbf{X}$. It satisfies the homogeneity property $\mu(tB) = t^{-\alpha}\mu(B)$, for any set B in $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ and any $t > 0$. The parameter α is called the *tail index* of $\mathbf{X}$. It highlights the intensity of the extremes. The smaller this index is, the heaviest the tail of $\mathbf{X}$ is likely to be.

It is often more convenient to decompose the former convergence into a radial and an angular part (see for instance Beirlant et al. (2006), Section 8.2.3): the regular variation property is equivalent to the convergence

$$\mathbb{P}((|\mathbf{X}|/t, \mathbf{X}/|\mathbf{X}|) \in \cdot \mid |\mathbf{X}| > t) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,, \tag{1.2}$$

where $\xrightarrow{w}$ denotes weak convergence, and where $\mathbf{\Theta}$ is a random vector on the positive unit sphere $\{\mathbf{x} \in [0,\infty)^d : |\mathbf{x}| = 1\}$ independent of the random variable Y which satisfies $\mathbb{P}(Y > y) = y^{-\alpha}$, $y > 1$. The random vector $\mathbf{\Theta}$ is called the spectral vector and its distribution $\mathbb{P}(\mathbf{\Theta} \in \cdot)$ the spectral measure. Its support indicates the directions supported by large events. The subspaces of the positive unit sphere on which the spectral vector puts mass correspond to the directions where large events are likely to appear. Note that the choice of the norm $|\cdot|$ in Equation (1.2) is arbitrary. In this article we choose the ℓ^1-norm and thus focus on the simplex $\mathbb{S}_+^{d-1} := \{\mathbf{x} \in [0,\infty)^d : x_1 + \cdots + x_d = 1\}$.

In order to study the support of the spectral measure we partition the simplex in terms of the nullity of some coordinates (Chautru (2015), Goix et al. (2017), Simpson et al. (2020)). For $\beta \subset \{1, \dots, d\}$ the subspace C_β is defined as

$$C_\beta = \{\mathbf{x} \in \mathbb{S}_+^{d-1} : x_i > 0 \text{ for } i \in \beta,\, x_i = 0 \text{ for } i \notin \beta\}\,. \tag{1.3}$$

This partition highlights the extremal structure of $\mathbf{X}$. For a given $\beta \subset \{1, \dots, d\}$ the inequality $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ implies that the marginals X_j, $j \in \beta$, are likely to take simultaneously large values while the ones for $j \in \beta^c$ are of smaller order. Hence the identification of clusters of directions β which concentrate the mass of the spectral measure brings out groups of coordinates which can be large together.

Highlighting such groups is at the core of several recent papers on multivariate extremes, all of them relying on some hyperparameters (Chiapino and Sabourin (2016), Goix et al. (2017), Chiapino et al. (2019), Simpson et al. (2020)). This approach faces a crucial issue, namely the difference of support between $\mathbf{\Theta}$ and $\mathbf{X}/|\mathbf{X}|$. Indeed, the spectral measure is likely to place mass on low-dimensional subspaces C_β, $\beta \neq \{1, \dots, d\}$. We say that this measure is *sparse* when the number of coordinates in the associated clusters β is small. Conversely, the distribution of the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ only concentrates on the subset $C_{\{1,\dots,d\}}$ since $\mathbf{X}$ has continuous marginals.

All the existing approaches proposed in the literature rely on nonstandard regular variation for which $\alpha = 1$ and all marginals are tail equivalent, possibly after a standardization. However, sparsity arises all the more for standard regular variation (1.2). In this case, it is possible that the marginals of $\mathbf{X}$ are not tail equivalent so that the support of the spectral measure is included in $\mathbb{S}_+^{r-1}$ for $r \ll d$. This is the approach we use in this article. For a comparison of standard and nonstandard regular variation we refer to Resnick (2007), Section 6.5.6.

In this article we provide a method which highlights the sparsity of the tail structure by exhibiting sparse clusters of extremal directions. By sparse clusters we mean groups of coordinates β which contain a reduced number of directions compared to d. We refer to this method as *sparse clustering*. The statistical procedure we propose to achieve this clustering relies on the framework of Meyer and Wintenberger (2021) which allows to circumvent the estimation's issue that arises with the spectral measure. The angular component $\mathbf{X}/|\mathbf{X}|$ in (1.2) is replaced by $\pi(\mathbf{X}/t)$, where π denotes the Euclidean projection onto $\mathbb{S}_+^{d-1}$ (Duchi et al. (2008), Kyrillidis et al. (2013), Condat (2016)). This substitution leads to the concept of *sparse regular variation*. A random vector $\mathbf{X}$ is said to be sparsely regularly varying if

$$\mathbb{P}((|\mathbf{X}|/t, \pi(\mathbf{X}/t)) \in \cdot \mid |\mathbf{X}| > t) \xrightarrow{w} \mathbb{P}((Y, \mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,, \tag{1.4}$$

where $\mathbf{Z}$ is a random vector on the simplex $\mathbb{S}_+^{d-1}$ and $\mathbb{P}(Y > y) = y^{-\alpha}$, $y > 1$. Meyer and Wintenberger (2021) proved that under mild assumptions both concepts of regular variation (1.2) and (1.4) are equivalent (see Theorem 1 in their article). In particular, the relation $\mathbf{Z} = \pi(Y\boldsymbol{\Theta})$ holds.

Similarly to the existing approaches with $\boldsymbol{\Theta}$, we are willing to capture the tail dependence of $\mathbf{X}$ via the identification of the clusters β which satisfy $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$. We call such β's the *extremal clusters*. They can be identified via the study of $\pi(\mathbf{X}/t)$ since the convergence $\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta)$ holds for any $\beta \subset \{1, \dots, d\}$ (see Meyer and Wintenberger (2021), Proposition 2). This encourages to consider for any β the quantity

$$T_{n,k}(\beta) = \sum_{j=1}^{k} \mathbb{1}\{\pi(\mathbf{X}_{(j)}/|\mathbf{X}_{(k+1)}|) \in C_\beta\}\,, \tag{1.5}$$

where $\mathbf{X}_1, \dots, \mathbf{X}_n$ is a sample of iid sparsely regularly varying random vectors, $k = k_n$ is an intermediate sequence called *level* which satisfies $k \to \infty$ and $k/n \to 0$, and $\mathbf{X}_{(j)}$ denotes

the observation with j-th largest norm: $|\mathbf{X}_{(1)}| \geq \cdots \geq |\mathbf{X}_{(n)}|$.

It turns out that the number of positive $T_{n,k}(\beta)$ often overestimates the total number of extremal clusters. We call the clusters which satisfy $T_{n,k}(\beta) > 0$ and $\mathbb{P}(\mathbf{Z} \in C_\beta) = 0$ the *biased clusters*. The approach we propose to reduce this bias relies on model selection. It consists in fitting a multinomial model to the data and to compare the Kullback-Leibler divergence between the data and this theoretical model. We obtain a minimization criterion based on a penalized likelihood similarly to Akaike's criterion (Akaike (1973)). This approach provides a way to select the appropriate number of extremal clusters for a given level k. This is the first step of our procedure, which we call the *bias selection*.

The second step then consists in extending the procedure in order to automatically select an appropriate level k. We call this step the *level selection*. Several authors have pointed out that choosing a reasonable level, or equivalently a reasonable threshold above which the data are considered as extreme, is a challenging task in practice. This issue is tackled in a few articles (Stărică (1999), Abdous and Ghoudi (2005), Kiriliouk et al. (2019), Wan and Davis (2019), see also the review on marginals threshold selection by Caiero and Gomes (2015)). It turns out that in the sparse regular variation framework the choice of such a level and the identification of the extremal clusters are closely related. Therefore our approach consists in extending the bias selection by including k as a parameter to tune. Since Akaike's procedure only holds for a constant sample size we have to adapt the standard approach to an extreme setting where the number of extremes varies. Therefore we include the non-extreme values in the model and separate the data into an *extreme* group and a *non-extreme* one. The procedure then provides a level k for which this separation is reasonable. To the best of our knowledge, our work is the first one which simultaneously tackles this issue with the study of tail dependence.

Outline of the paper The paper is organized as follows. Section 2 introduces the theoretical background on sparse regular variation and level selection that is needed throughout the paper. In Section 3 we introduce the statistical framework of our method and establish asymptotic results for the estimators of the probabilities $\mathbb{P}(\mathbf{Z} \in C_\beta)$. Section 4 details the methodology of our approach. We develop the two steps of the model selection, the bias selection and the level selection. In Section 5 we illustrate our findings on numerical results and compare our approach with the existing procedures proposed by Goix et al. (2017) and Simpson et al. (2020). Finally we illustrate our approach on financial data in Section 6. The proofs are given in the Supplementary Material.

2 Preliminaries

2.1 Notation

Symbols in bold such as $\mathbf{x} \in \mathbb{R}^d$ are column vectors with components denoted by x_j, $j \in \{1, \ldots, d\}$. Operations and relationships involving such vectors are meant componentwise. If $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, then $\mathrm{Diag}(\mathbf{x})$ or $\mathrm{Diag}(x_1, \ldots, x_d)$ denotes the diagonal matrix whose diagonal is $\mathbf{x}$. We denote by Id_s the identity matrix of $\mathbb{R}^s$. We define $\mathbb{R}^d_+ := \{\mathbf{x} \in \mathbb{R}^d : x_1 \geq 0, \ldots, x_d \geq 0\}$, $\mathbf{0} := (0, \ldots, 0)^\top \in \mathbb{R}^d$, and $\mathbf{1} := (1, \ldots, 1)^\top \in \mathbb{R}^d$. For $j = 1, \ldots, d$, $\mathbf{e}_j$ denotes the j-th vector of the canonical basis of $\mathbb{R}^d$. In all the paper we denote the ℓ^1-norm by $|\cdot|$. For $d \geq 1$ we denote by $\mathcal{P}_d$ the power set of $\{1, \ldots, d\}$ and by $\mathcal{P}_d^*$ the set $\mathcal{P}_d \setminus \{\emptyset\}$. If $\beta \in \mathcal{P}_d$ we denote by $|\beta|$ the number of coordinates in β.

2.2 Sparse regular variation

We consider a sparsely regularly varying random vector $\mathbf{X} \in \mathbb{R}_+^d$ as defined in (1.4) and focus on its angular component $\pi(\mathbf{X}/t)$:

$$\mathbb{P}\left(\pi(\mathbf{X}/t) \in \cdot \mid |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}(\mathbf{Z} \in \cdot)\,, \quad t \to \infty\,. \tag{2.1}$$

The orthogonal projection on the simplex enjoys many sparsity properties which justifies its use to study high-dimensional data. The vector $\pi(\mathbf{X}/t)$ may put mass in every subspace C_β even if $\mathbf{X}$ is almost surely positive. This is a key difference with the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ which shares the same sparsity properties as $\mathbf{X}$, and therefore always concentrates on the interior $C_{\{1,\ldots,d\}}$ of the simplex.

Remark 1. Our statistical methodology exhibits the choice of a level k which corresponds to the number of vectors among a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ which are considered as extreme. This is achieved by studying also the $n-k$ non-extreme vectors. In terms of the convergence (2.1) the latter vectors correspond to vectors whose norm is below the threshold $t = |\mathbf{X}_{(k+1)}|$. In order to propose a consistent methodology based on these non-extreme vectors we need to slightly modify the projection and to consider π as the Euclidean projection onto the unit positive ℓ^1-ball $\mathcal{B}_+^d = \{\mathbf{x} \in \mathbb{R}_+^d : x_1 + \ldots + x_d \leq 1\}$. It does not change the theory of sparse regular variation since projecting onto the sphere or the ball is equivalent for vectors with norm larger than 1. The only difference is that a vector $\mathbf{v}$ such that $|\mathbf{v}| < 1$ now satisfies $\pi(\mathbf{v}) = \mathbf{v}$.

Our aim is to infer the distribution of the angular vector $\mathbf{Z}$ in order to identify the extremal directions of $\mathbf{X}$. This is achieved by focusing on the probabilities $p^*(\beta) := \mathbb{P}(\mathbf{Z} \in C_\beta)$ for $\beta \in \mathcal{P}_d^*$. We define the set of extremal clusters

$$\mathcal{S}^*(\mathbf{Z}) := \{\beta : p^*(\beta) > 0\}\,, \tag{2.2}$$

and denote by s^* its cardinality. Meyer and Wintenberger (2021) proved that for any β we have the convergence

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to p^*(\beta)\,, \quad t \to \infty\,. \tag{2.3}$$

This convergence allows one to study the behavior of $\mathbf{Z}$ on the subsets C_β via the one of $\pi(\mathbf{X}/t)$. The aim of this paper is to build a statistical procedure to identify the extremal clusters $\beta \in \mathcal{S}^*(\mathbf{Z})$.

Example 1 (Discrete spectral measure)**.** *For $\beta \in \mathcal{P}_d^*$, we denote by $\mathbf{e}(\beta)$ the sum $\sum_{j\in\beta} \mathbf{e}_j$ so that the vector $\mathbf{e}(\beta)/|\beta|$ belongs to the simplex $\mathbb{S}_+^{d-1}$ (recall that $|\beta|$ corresponds to the length of the cluster β). We consider the following family of discrete distributions on the simplex:*

$$\sum_{\beta\in\mathcal{P}_d^*} c(\beta)\, \delta_{\mathbf{e}(\beta)/|\beta|}\,, \tag{2.4}$$

where $(c(\beta))_\beta$ is a probability vector on $\mathbb{R}^{2^d-1}$ (see Segers (2012), Example 3.3). Meyer and Wintenberger (2021) proved that in this case we have $\mathbf{Z} = \mathbf{\Theta}$ a.s. and that the family of distribution in (2.4) *is the only possible discrete distributions for $\mathbf{Z}$. For this type of distributions we have $\mathcal{S}^*(\mathbf{Z}) = \{\beta : c(\beta) > 0\}$.*

If we choose $c(\beta) = 0$ for all β's except the ones of length 1 *then the spectral measure becomes $\sum_{j=1}^d c_j\, \delta_{\mathbf{e}_j}$, $(c_j)_{1\le j\le d} \in \mathbb{S}_+^{d-1}$. This corresponds to* asymptotic independence *(see e.g. Ledford and Tawn (1996), Heffernan and Tawn (2004), de Haan and Ferreira (2006), Section 6.2).*

If $\mathbf{\Theta}$ places mass on a subset C_β then so does $\mathbf{Z}$, but the converse is not true. Thus the set of clusters we identify with our method includes the usual ones on which $\mathbf{\Theta}$ puts mass. However, the notion of maximal cluster (an extremal cluster which is not included

in another extremal one) defined by Meyer and Wintenberger (2021) coincide for $\mathbf{\Theta}$ and $\mathbf{Z}$ and links both types of clusters.

Example 2. *Consider a spectral measure in dimension* 2 *with* $\Theta_1 \sim \mathcal{U}(0,1)$. *Then the distribution of* $\mathbf{Z}$ *is given by* $Z_1 = \frac{1}{4}\delta_0 + \frac{1}{2}\mathcal{U}(0,1) + \frac{1}{4}\delta_1$, *see Meyer and Wintenberger (2021), Example 1. In this case the clusters* $\{1\}$ *and* $\{2\}$ *are extremal clusters for* $\mathbf{Z}$ *but not for* $\mathbf{\Theta}$. *The only maximal cluster for* $\mathbf{Z}$ *and* $\mathbf{\Theta}$ *is* $C_{\{1,2\}}$.

2.3 Impact of the level on the sparsity structure

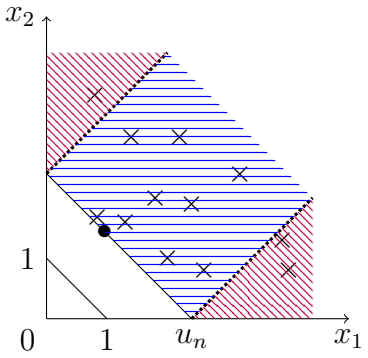
We briefly explain in this section how the choice of a threshold $t > 0$ influences the sparsity of the projected vector $\pi(\mathbf{x}/t)$ for $\mathbf{x} \in \mathbb{R}_+^d$. For $t > 0$, let us denote by π_t the Euclidean projection onto the positive sphere $\{\mathbf{x} \in \mathbb{R}_+^d : x_1 + \cdots + x_d = t\}$. The relation $\pi_t(\mathbf{x}) = t\pi(\mathbf{x}/t)$ implies that the sparsity structures of $\pi_t(\mathbf{x})$ and $\pi(\mathbf{x}/t)$ are the same. The number of null coordinates of the projected vector $\pi_t(\mathbf{x})$ strongly depends on the choice of t. Indeed, if t is close to $|\mathbf{x}|$, then $\pi_t(\mathbf{x})$ has only non-null coordinates (as soon as $\mathbf{x}$ itself has non-null coordinates). On the contrary, the vector $\pi_t(\mathbf{x})$ is sparse if $t \ll |\mathbf{x}|$.

Moving on to a statistical framework, we consider a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of iid random vectors in $\mathbb{R}_+^d$. It is common in extreme value theory to define a level $k = k_n$ satisfying $k \to \infty$ and $k/n \to 0$ (see e.g. de Haan and Ferreira (2006), Beirlant et al. (2006), Resnick (2007)). It leads to the choice of a threshold $t = u_n \to \infty$ such that

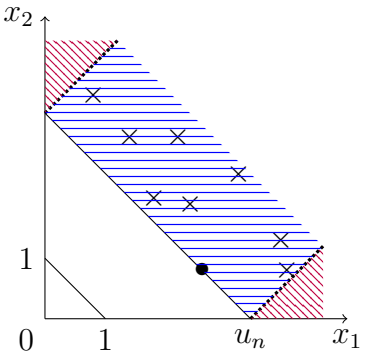
$$\frac{n}{k}\mathbb{P}(|\mathbf{X}| > u_n) \to 1\,, \quad n \to \infty\,. \tag{2.5}$$

The level k must be seen as the number of extreme vectors used for the statistical analysis. It is therefore natural to consider the k-largest vectors in terms of their norm, i.e. the vectors $\mathbf{X}_{(1)}, \cdots, \mathbf{X}_{(k)}$ where $|\mathbf{X}_{(j)}|$ denotes the j-th largest norm $|\mathbf{X}_{(1)}| \geq \cdots \geq |\mathbf{X}_{(n)}|$. Note

that since we assumed the marginals of $\mathbf{X}$ to be continuous, these inequalities are strict almost surely. This encourages to work with the random threshold $|\mathbf{X}_{(k+1)}|$. By Vervaat's Lemma (see Lemma 1.0.2 in de Haan and Ferreira (2006)), the assumption (2.5) implies that $|\mathbf{X}_{(k+1)}|/u_n$ converges to 1 in probability as $n \to \infty$.



(a) For $k = 12$ the points in the blue area are projected on the interior of the positive sphere $\{\mathbf{y} \in \mathbb{R}^d_+ : y_1 + \cdots + y_d = u_n\}$ while the ones in the red area are projected on the edges of this sphere.



(b) For $k = 8$ all points in the blue area are projected on the interior of the positive sphere $\{\mathbf{y} \in \mathbb{R}^d_+ : y_1 + \cdots + y_d = u_n\}$.

Figure 1: Influence of the level k on the sparsity structure of the data. The threshold u_n corresponds to the norm of the vector $\mathbf{X}_{(k+1)}$ which is represented by a bullet.

A small k corresponds to a large threshold u_n and vice versa. In this case only a few extreme vectors are kept for the statistical analysis and they are close to the threshold. Thus, these vectors are projected on subsets C_β with large $|\beta|$'s which means that the projected vectors are not very sparse. On the other hand, choosing a large k means choosing a low threshold u_n so that we move away from the extreme region. In this case the largest vectors are projected on subsets C_β with small $|\beta|$'s, i.e. the projected vectors are sparse.

We refer to Figure 1 for an illustration of these two cases. Following these remarks we have to make a balanced choice between providing a sparse structure for the data and staying in the extreme region.

3 Asymptotic analysis of the extremal clusters

We consider a sequence of iid sparsely regularly varying random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with generic distribution $\mathbf{X}$ and angular limit vector $\mathbf{Z}$. We also consider a level k satisfying $k \to \infty$ and $k/n \to 0$ and a threshold u_n such that (2.5) is satisfied. In order to identify the set $\mathcal{S}^*(\mathbf{Z})$ defined in (2.2) we provide suitable estimators for the probabilities $p^*(\beta)$, $\beta \in \mathcal{P}_d^*$. We define the estimators

$$T_n(x, \beta) = \sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{X}/u_n \in A(x,\beta)\}}\,, \quad \beta \in \mathcal{P}_d^*\,, \quad x > 0\,,$$

where $A(x, \beta) = \{\mathbf{y} \in \mathbb{R}_+^d : x|\mathbf{y}| > 1,\, \pi(x\mathbf{y}) \in C_\beta\}$ so that the estimator $T_{n,k}(\beta)$ defined in (1.5) satisfies $T_{n,k}(\beta) = T_n(u_n/|\mathbf{X}_{(k+1)}|, \beta)$. An empirical version of $\mathcal{S}^*(\mathbf{Z})$ is then given by

$$\widehat{\mathcal{S}}_n := \{\beta \in \mathcal{P}_d^* : T_{n,k}(\beta) > 0\}\,. \tag{3.1}$$

We denote by $\hat{s}_n$ the cardinality of this set. Finally, we define

$$p_n(\beta) = \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)\,, \quad \beta \in \mathcal{P}_d^*\,.$$

which converges to $p^*(\beta) \in \mathcal{P}_d^*$ for any β, see Equation (2.3).

3.1 The bias between $\widehat{\mathcal{S}}_n$ and $\mathcal{S}^*(\mathbf{Z})$

In this section we compare the set $\mathcal{S}^*(\mathbf{Z})$ with its empirical counterpart $\widehat{\mathcal{S}}_n$. We first establish the consistency of our estimator.

Proposition 1. *For any* $\beta \in \mathcal{P}_d^*$,

$$\frac{T_{n,k}(\beta)}{k} = \frac{1}{k}\sum_{j=1}^{k} \mathbb{1}_{\{\pi(\mathbf{x}_{(j)}/|\mathbf{x}_{(k+1)}|)\in C_\beta\}} \to p^*(\beta)\,, \quad n \to \infty\,, \tag{3.2}$$

in probability.

The proof relies on Proposition 2.2 of de Haan and Resnick (1993). It suffices to prove that $\mathbf{Z}$ does not put any mass on the boundary of $\{|\mathbf{x}|> 1\} \cap \pi^{-1}(C_\beta)$, which has already been established in Proposition 2 of Meyer and Wintenberger (2021).

Proposition 1 implies that if $p^*(\beta) = 0$, i.e. if $\mathbf{Z}$ does not place mass on the subset C_β, then $T_{n,k}(\beta)/k$ becomes smaller and smaller as n increases. Actually as soon as the dimension d is large a lot of $T_{n,k}(\beta)$'s are even equal to 0 since the number of extreme vectors, that is k, is far below the number of clusters, that is $2^d - 1$.

In order to study the bias between $\widehat{\mathcal{S}}_n$ and $\mathcal{S}^*(\mathbf{Z})$ we focus on the speed of convergence of $\mathbb{P}(T_{n,k}(\beta) = 0)$, and thus on the one of $\mathbb{P}(\mathbf{X}/u_n \in A(x,\beta)) = \mathbb{P}(\{\mathbf{y} \in \mathbb{R}_+^d : x|\mathbf{y}|> 1,\, \pi(x\mathbf{y}) \in C_\beta\})$. Meyer and Wintenberger (2021) established the equivalence

$$\pi(\mathbf{x}) \in C_\beta \quad \text{if and only if} \quad \forall i \in \beta^c,\, \forall j \in \beta,\, x_i \le \frac{|\mathbf{x}_\beta|}{|\beta|} < x_j\,.$$

In other words, all x_j, $j \in \beta$, should be of the same order, while the x_i, $i \in \beta^c$, should be of smaller order. We set $\mathbf{x}_{\beta,i} = \sum_{j\in\beta}(x_j - x_i)$ for any i and define

$$\mathbb{C}_\beta = \{\mathbf{x} \in \mathbb{R}_+^d : \min_{i\in\beta^c} \mathbf{x}_{\beta,i} \ge 0\} = \Big\{\mathbf{x} \in \mathbb{R}_+^d : \sum_{j\in\beta}(x_j - \max_{i\in\beta^c} x_i) \ge 0\Big\}\,,$$

which forms a cone of $\mathbb{R}_+^d$. Studying the convergence of $\mathbb{P}(\mathbf{X}/u_n \in A(x,\beta))$ then boils down to studying the asymptotic behavior of $\mathbf{X}$ on the cone $\mathbb{C}_\beta$. Based on the theory of hidden regular variation (HRV) by Lindskog et al. (2014), we make the following assumption on $\mathbf{X}$.

Assumption (HRV). For every $\beta \in \mathcal{P}_d^*$ the vector $\mathbf{X}$ is regularly varying on $\mathbb{R}_+^d \setminus \mathbb{C}_\beta$ with tail index $\alpha(\beta)$ and exponent measure μ_β satisfying

$$\mu_\beta(\{\mathbf{x} \in \mathbb{R}_+^d : \max_{i \in \beta} \mathbf{x}_{\beta,i} < 1, \min_{i \in \beta^c} \mathbf{x}_{\beta,i} \geq 1\}) > 0\,.$$

This assumption allows one to deal with the asymptotic behavior of $\mathbb{P}(T_{n,k}(\beta) = 0)$ even when $p^*(\beta) = 0$, as stated in the following lemma.

Lemma 1. *Under Assumption* **(HRV)** *we have for every* $\beta \in \mathcal{P}_d^*$,

$$\frac{\log \mathbb{P}(T_{n,k}(\beta) = 0)}{-kp_n(\beta)} \to 1\,, \quad n \to \infty\,.$$

Lemma 1 encourages to focus on the quantity $kp_n(\beta)$ and to consider the set

$$\mathcal{S}_\infty = \{\beta \in \mathcal{P}_d^* : kp_n(\beta) \to \infty \text{ as } n \to \infty\}\,. \tag{3.3}$$

We denote by s_∞ its cardinality. This set contains $\mathcal{S}^*(\mathbf{Z})$ so that we have the inequality $s^* \leq s_\infty$. Subsequently, Lemma 1 implies that

$$\mathbb{P}(\mathcal{S}_\infty \subset \widehat{\mathcal{S}}_n) = 1 - \mathbb{P}(\exists \beta \in \mathcal{S}_\infty,\, \beta \notin \widehat{\mathcal{S}}_n) \geq 1 - \sum_{\beta \in \mathcal{S}_\infty} \mathbb{P}(T_{n,k}(\beta) = 0) \to 1\,,$$

as $n \to \infty$. This leads to the following proposition.

Proposition 2. *Under Assumption* **(HRV)** *the inclusions*

$$\mathcal{S}^*(\mathbf{Z}) \subset \mathcal{S}_\infty \subset \widehat{\mathcal{S}}_n$$

hold true with probability converging to 1.

These inclusions highlight the fact that the observations $T_{n,k}(\beta)$ tend to overestimate the number of clusters β in $\mathcal{S}^*(\mathbf{Z})$. They imply that we only have a "one-side bias" composed

of clusters that appear empirically but which theoretically do not contain any mass. One of the main challenge of our study is the derivation of the asymptotic properties of $T_{n,k}(\beta)$ for biased clusters $\beta \in \widehat{\mathcal{S}}_n \setminus \mathcal{S}^*(\mathbf{Z})$.

By Lemma 1 the inclusion $\widehat{\mathcal{S}}_n \subset \mathcal{S}_\infty$ means that we only observe faces C_β for which $\mathbb{P}(T_{n,k}(\beta) = 0)$ does not decrease super-exponentially fast with n. We define the sets of admissible sequences (k_n) by

$$K = \{(k_n) : k_n \to \infty,\, k_n/n \to 0,\, \mathcal{S}_\infty = \widehat{\mathcal{S}}_n \text{ a.s. for all } n \text{ large enough}\}\,.$$

That K is non empty is a strong assumption equivalent to the fact that $\widehat{\mathcal{S}}_n$ converges a.s. to $\mathcal{S}_\infty$ for some sequence of levels (k_n). Combining the definition of K with Assumption **(HRV)** we can rely on the statistics $T_{n,k}(\beta)$, $(k_n) \in K$, which are non-null sufficiently often even when $p_n(\beta) \to p^*(\beta) = 0$, in order to quantify the bias.

3.2 Asymptotic normality

We now establish a convergence result for the joint distribution of $T_{n,k}(\beta)$ for $\beta \in \mathcal{S}_\infty$. This is achieved via the study of the joint distribution of $T_n(x, \beta)$ for $x \in [\frac{1}{1+\tau}, 1+\tau]$, $\tau > 0$. Having in mind the model selection proposed in Section 4 we consider for any $0 \le s < r \le s_\infty$ and any disjoint clusters $\beta_1, \dots, \beta_r \in \mathcal{S}_\infty$ the vectors

$$\mathbf{T}_n^{s,r}(x) = \Big(T_n(x,\beta_1), \dots, T_n(x,\beta_s), \sum_{j=s+1}^{r} T_n(x,\beta_j)\Big)^\top \in \mathbb{R}^{s+1}\,,$$

and

$$\mathbf{P}_n^{s,r}(x) = \Big(x^{\alpha(\beta_1)} p_n(\beta_1), \dots, x^{\alpha(\beta_s)} p_n(\beta_s), \sum_{j=s+1}^{r} x^{\alpha(\beta_j)} p_n(\beta_j)\Big)^\top \in \mathbb{R}^{s+1}\,.$$

For $\tau > 0$ we denote by $\ell^\infty([\frac{1}{1+\tau}, 1+\tau])$ the set of functions defined and bounded on $[\frac{1}{1+\tau}, 1+\tau]$.

Theorem 1. *Let Assumption* **(HRV)** *hold. Assume that there exists* $(k_n) \in K$ *and choose* u_n *such that* $k \sim n\mathbb{P}(|\mathbf{X}| > u_n)$ *as* $n \to \infty$.

1. *The following convergence holds in* $\ell^\infty([\frac{1}{1+\tau}, 1+\tau])$ *as* $n \to \infty$*:*

$$\left\{ \sqrt{k}\, \mathrm{Diag}(\mathbf{P}_n^{s,r}(x))^{-1/2} \left(\frac{\mathbf{T}_n^{s,r}(x)}{k} - \mathbb{E}\Big[\frac{\mathbf{T}_n^{s,r}(x)}{k} \Big] \right);\ (1+\tau)^{-1} \le x \le 1+\tau \right\}_{s<r} \xrightarrow{d} (\mathbf{N}^{s,r})_{s<r}\,, \qquad (3.4)$$

   where the constant limit process is identified to $\mathbf{N}^{s,r}$*, a standard centered multivariate Gaussian vector in* $\mathbb{R}^{s+1}$.

2. *If we assume moreover that for any* $\beta \in \mathcal{S}_\infty$*,*

$$\sup_{x \in [\frac{1}{1+\tau}, 1+\tau]} \sqrt{\frac{k}{p_n(\beta)}} \left| \frac{n}{k} \mathbb{P}(\mathbf{X}/u_n \in A(x,\beta)) - x^{\alpha(\beta)} p_n(\beta) \right| \to 0\,, \quad n \to \infty\,, \qquad (3.5)$$

   then we have

$$\left\{ \sqrt{k}\, \mathrm{Diag}(\mathbf{P}_n^{s,r}(x))^{-1/2} \left(\frac{\mathbf{T}_n^{s,r}(x)}{k} - \mathbf{P}_n^{s,r}(x) \right);\ (1+\tau)^{-1} \le x \le 1+\tau \right\}_{s<r} \xrightarrow{d} (\mathbf{N}^{s,r})_{s<r}\,, \qquad (3.6)$$

   in $\ell^\infty([\frac{1}{1+\tau}, 1+\tau])$ *as* $n \to \infty$.

Based on Theorem 1, we establish the asymptotic behavior of the estimators $T_{n,k}(\beta)$. We define

$$\mathbf{T}_{n,k}^{s,r} = \mathbf{T}_n^{s,r}(u_n/|\mathbf{X}|_{(k+1)}) = \Big(T_{n,k}(\beta_1), \ldots, T_{n,k}(\beta_s), \sum_{j=s+1}^{r} T_{n,k}(\beta_j) \Big)^\top \in \mathbb{R}^{s+1}\,.$$

Proposition 3. *Under the assumptions of Theorem 1, under* (3.5)*, and under the bias assumption*

$$\sqrt{k}(p_n(\beta) - p^*(\beta)) \to 0\,, \quad n \to \infty\,, \quad \beta \in \mathcal{S}_\infty\,, \qquad (3.7)$$

we have the convergence

$$\sqrt{k}\,\mathrm{Diag}(\mathbf{P}_n^{s,r}(1))^{-1/2}\left(\frac{\mathbf{T}_{n,k}^{s,r}}{k}-\mathbf{P}_n^{s,r}(1)\right)\xrightarrow{d}(Id_{s+1}-\sqrt{\mathbf{P}^{s,r}}\cdot\sqrt{\mathbf{P}^{s,r}}^{\top})\mathbf{N}\,,\quad n\to\infty\,,\quad(3.8)$$

where $\mathbf{N}\in\mathbb{R}^{s+1}$ *is a standard centered multivariate Gaussian vector, and where* $\mathbf{P}^{s,r}$ *is the limit vector of* $\mathbf{P}_n^{s,r}(1)$*:*

$$\mathbf{P}^{s,r}=\Big(p^*(\beta_1),\dots,p^*(\beta_s),\sum_{j=s+1}^{r}p^*(\beta_j)\Big)^{\top}=\lim_{n\to\infty}\Big(p_n(\beta_1),\dots,p_n(\beta_s),\sum_{j=s+1}^{r}p_n(\beta_j)\Big)^{\top}.$$

Remark 2. The bias assumption (3.7) holds for $\beta\in\mathcal{S}_\infty\setminus\mathcal{S}^*(\mathbf{Z})$ if $k=o(n^\kappa)$ as $n\to\infty$ where $\kappa>\frac{2(\alpha(\beta)-\alpha)}{2\alpha(\beta)-\alpha}$ for every $\beta\in\mathcal{S}_\infty\setminus\mathcal{S}^*(\mathbf{Z})$. We refer to the Supplementary Material for a proof.

Remark 3. For $r=s_\infty$ the matrix $Id_{s+1}-\sqrt{\mathbf{P}^{s,r}}\cdot\sqrt{\mathbf{P}^{s,r}}^{\top}$ is symmetric and satisfies

$$\begin{aligned}(Id_{s+1}-\sqrt{\mathbf{P}^{s,r}}\cdot\sqrt{\mathbf{P}^{s,r}}^{\top})^2&=Id_{s+1}-2\sqrt{\mathbf{P}^{s,r}}\cdot\sqrt{\mathbf{P}^{s,r}}^{\top}+(\sqrt{\mathbf{P}^{s,r}}\cdot\sqrt{\mathbf{P}^{s,r}}^{\top})^2\\&=Id_{s+1}-\sqrt{\mathbf{P}^{s,r}}\cdot\sqrt{\mathbf{P}^{s,r}}^{\top},\end{aligned}$$

since $\sqrt{\mathbf{P}^{s,r}}^{\top}\cdot\sqrt{\mathbf{P}^{s,r}}=\sum_{j=1}^{r}p^*(\beta_j)=1$. Therefore it corresponds to an orthogonal projection with rank s. Cochran's theorem then ensures that the ℓ^2-norm of the vector $(Id_{s+1}-\sqrt{\mathbf{P}^{s,r}}\sqrt{\mathbf{P}^{s,r}}^{\top})\mathbf{N}$ follows a chi-squared distribution with s degrees of freedom.

Going back to Proposition 3 we obtain the following convergence:

$$k\sum_{j=1}^{s}\frac{(T_{n,k}(\beta_j)/k-p_n(\beta_j))^2}{p_n(\beta_j)}+k\frac{[\sum_{j=s+1}^{r}(T_{n,k}(\beta_j)/k-p_n(\beta_j))]^2}{\sum_{j=s+1}^{r}p_n(\beta_j)}\xrightarrow{d}\psi(s)\,,\qquad(3.9)$$

where $\psi(s)$ follows a chi-squared distribution with s degrees of freedom. This convergence is useful to identify the parameter s in the bias selection, see Lemma 4 in the Supplementary Material.

4 Methodology

We develop in this section our methodology to estimate the set $\mathcal{S}^*(\mathbf{Z})$. We use the same notation as in Section 3.

4.1 Bias selection

We consider the vector $\mathbf{T}_{n,k} \in \mathbb{R}^{2^d-1}$ with components $T_{n,k}(\beta)$ whose distribution $\mathbf{P}_k$ is multinomial with probability weights $(p_n(\beta))_{\beta\in\mathcal{P}_d^*}$, and adding up to k. We propose a bias selection which consists in comparing the distribution $\mathbf{P}_k$ with the theoretical multinomial model $\mathbf{M}_k$ with 2^d-1 outcomes adding up to k and a probability vector $(p_1,\dots,p_s,p,\dots,p,0,\dots,0)^\top \in [0,1]^{2^d-1}$, with $p_1 \geq \cdots \geq p_s > p$ and $r-s$ components p satisfying $p_1+\cdots+p_s+(r-s)p = 1$. The parameters p_j model the probability that $\mathbf{Z}$ belongs to the associated subsets C_β while the parameter p models the probability that a biased cluster appears. We denote by $\mathbf{p}$ the vector $(p_1,\dots,p_s)^\top \in \mathcal{B}_+^s(0,1)$. The likelihood $L_{\mathbf{M}_k}$ of the model $\mathbf{M}_k$ is given by

$$L_{\mathbf{M}_k}(\mathbf{p};\mathbf{y}) = \frac{k!}{\prod_{i=1}^{2^d-1} y_i!} \prod_{i=1}^{s} p_i^{y_i} \prod_{i=s+1}^{r} \Big(\frac{1-\sum_{j=1}^s p_j}{r-s}\Big)^{y_i} \mathbb{1}_{\{y_{r+1}=\cdots=y_{2^d-1}=0\}}\,, \tag{4.1}$$

for any vector $\mathbf{p} \in \mathcal{B}_+^s(0,1) = \{\mathbf{u} \in \mathbb{R}_+^s : u_1+\cdots+u_s \leq 1\}$ and any $\mathbf{y} \in \mathbb{N}_0^{2^d-1}$ adding up to k, where $\mathbb{N}_0$ denotes the sets of non-negative integers.

The identification of the extremal clusters β in $\mathcal{S}^*(\mathbf{Z})$ is achieved by choosing the model $\mathbf{M}_k$ which best fits the sample $\mathbf{T}_{n,k}$. Following the AIC approach of Akaike (1973), we select the multinomial model which minimizes the expectation of the Kullback-Leibler (KL) divergence (see Kullback and Leibler (1951)) between the true distribution $\mathbf{P}_k$ and the model $\mathbf{M}_k$ evaluated at $\hat{\mathbf{p}}$, where $\hat{\mathbf{p}}$ denotes the maximum-likelihood estimator of $\mathbf{p}$.

Hence we consider the quantity

$$\mathbb{E}[KL(\mathbf{P}_k\|\mathbf{M}_k)|_{\mathbf{p}=\hat{\mathbf{p}}}] = \mathbb{E}[\log L_{\mathbf{P}_k}(\mathbf{T}_{n,k})] - \mathbb{E}[\mathbb{E}[\log L_{\mathbf{M}_k}(\mathbf{p};\mathbf{T}_{n,k})]|_{\mathbf{p}=\hat{\mathbf{p}}}]\,, \tag{4.2}$$

where $L_{\mathbf{P}_k}$ denotes the likelihood of the distribution $\mathbf{P}_k$. Theorem 2 below provides an asymptotic expansion of this quantity.

Before stating this result we compute the maximum likelihood estimator of the model $\mathbf{M}_k$. The first components of the model $\mathbf{M}_k$ being associated to the extremal clusters, we reorder the coordinates of the vector $\mathbf{T}_{n,k}$ so that its components are ordered in the decreasing order. Hence we define $T_{n,k,1} = \max_\beta T_{n,k}(\beta)$ and

$$T_{n,k,j} = \max\{T_{n,k}(\beta),\, \beta \in \mathcal{P}_d^*\} \setminus \{T_{n,k,1},\dots,T_{n,k,j-1}\}\,, \quad j = 2,\dots,2^d-1\,.$$

The expression in (4.1) is also maximal when r corresponds to the number $\hat{s}_n$ of clusters that appear empirically. This leads to the following expression of the log-likelihood $\log L_{\mathbf{M}_k}(\mathbf{p};\mathbf{T}_{n,k})$:

$$\log(k!\,) - \sum_{i=1}^{2^d-1}\log(T_{n,k,i}!\,) + \sum_{i=1}^{s} T_{n,k,i}\log(p_i) + \Big(\sum_{i=s+1}^{r} T_{n,k,i}\Big)\log\Big(\frac{1-\sum_{j=1}^s p_j}{r-s}\Big)\,. \tag{4.3}$$

The optimization of this quantity then provides the maximum likelihood estimator $\hat{\mathbf{p}} \in \mathbb{R}^s$ with components $\hat{p}_j := T_{n,k,j}/k$ for $1 \le j \le s$.

Theorem 2. *Under the assumptions of Proposition 3 the following convergence holds:*

$$\mathbb{E}[KL(\mathbf{P}_k\|\mathbf{M}_k)|_{\mathbf{p}=\hat{\mathbf{p}}}] - \mathbb{E}[\log L_{\mathbf{P}_k}(\mathbf{T}_{n,k})] + \mathbb{E}[\log L_{\mathbf{M}_k}(\hat{\mathbf{p}};\mathbf{T}_{n,k})] \to s\,, \quad n\to\infty\,.$$

Based on Theorem 2 we choose the model $\mathbf{M}_k$ which minimizes the quantity

$$-\log L_{\mathbf{M}_k}(\hat{\mathbf{p}};\mathbf{T}_{n,k}) + s\,. \tag{4.4}$$

Therefore for a given sequence $(k_n) \in K$ the bias selection procedure consists in choosing the parameter $\hat{s}(k)$ which minimizes this penalized log-likelihood.

4.2 Level selection

The second step of the model selection consists in considering k as a parameter which has to be estimated and tuned. It is therefore necessary to consider all observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and not only the extreme ones. We consider a vector $\mathbf{T}'_n \in \mathbb{R}^{2^d}$ such that

$$\mathcal{L}((T'_{n,1}, \ldots, T'_{n,2^d-1}) \mid T'_{n,2^d} = n - k) = \mathbf{T}_{n,k}\,.$$

The last component $T'_{n,2^d}$ corresponds to the number of non-extreme values of the sample. We assume that this vector follows a multinomial distribution $\mathbf{P}'_n$ with parameter n and probability vector $\mathbf{p}'_n = (q_n p_{n,1}, \ldots, q_n p_{n,2^d-1}, 1-q_n)^\top \in \mathbb{R}^{2^d}$.

Similarly to Section 4.1 we consider a multinomial model $\mathbf{M}'_n$ with probability vector given by $(q'p'_1, \ldots, q'p'_{s'}, q'p', \ldots, q'p', 0, \ldots, 0, 1-q')^\top \in \mathbb{R}^{2^d}$ with $p'_1 \geq \ldots \geq p'_{s'} > p'$ and $r'-s'$ components $q'p'$ satisfying the relation $p'_1 + \ldots + p'_s + (r'-s')p' = 1$. Here q' models the proportion of extreme vectors. We denote by $\mathbf{p}'$ the vector $(p'_1, \ldots, p'_{s'}, q')^\top \in \mathcal{B}_+^{s'} \times (0,1)$.

We consider the Kullback-Leibler divergence between $\mathbf{P}'_n$ and $\mathbf{M}'_n$ given by

$$KL(\mathbf{P}'_n \| \mathbf{M}'_n) = \mathbb{E}\left[\log\left(\frac{L_{\mathbf{P}'_n}(\mathbf{T}'_n)}{L_{\mathbf{M}'_n}(\mathbf{p}';\mathbf{T}'_n)}\right)\right] = \mathbb{E}[\log L_{\mathbf{P}'_n}(\mathbf{T}'_n)] - \mathbb{E}[\log L_{\mathbf{M}'_n}(\mathbf{p}';\mathbf{T}'_n)]\,, \quad (4.5)$$

where $L_{\mathbf{P}'_n}$ (resp. $L_{\mathbf{M}'_n}$) denotes the likelihood of the distribution $\mathbf{P}'_n$ (resp. $\mathbf{M}'_n$). Following the same ideas as in Section 4.1 and similarly to an AIC procedure we estimate the Kullback-Leibler divergence in Equation (4.5) by the estimator $KL(\mathbf{P}'_n \| \mathbf{M}'_n)|_{\widehat{\mathbf{p}'}}$ where $\widehat{\mathbf{p}'}$ denotes the maximum likelihood estimator of $\mathbf{p}'$.

We make the following assumptions.

(B1) For $k \in K$ and $\beta_j \in \mathcal{S}_\infty$ we have

$$\frac{\mathbb{E}[T_{n,n-T'_{n,2^d},j} \mid T'_{n,2^d}]}{n - T'_{n,2^d}} = \frac{\mathbb{E}[T_{n,k,j}]}{k} + O(1)\,, \quad n \to \infty\,.$$

(B2) For n sufficiently large, $k \in K$, there exist c, $C > 0$ such that $cnq_n \leq k \leq Cnq_n$.

Assumptions **(B1)** and **(B2)** allow one to control the bias between $T_{n,n-T'_{n,2^d},j} \mid T'_{n,2^d}$ and $T_{n,k,j}$, and between k and nq_n respectively.

The following theorem provides an asymptotic expansion of the expectation of this estimator.

Theorem 3. *Under* **(B1)**, **(B2)** *and the assumptions of Proposition 3 we have*

$$\mathbb{E}[KL(\mathbf{P}'_n \| \mathbf{M}'_n)|_{\widehat{\mathbf{p}'}}] = nq_n \Big(\frac{\mathbb{E}[-\log L_{\mathbf{M}_k}(\hat{\mathbf{p}}; \mathbf{T}_n)] + s}{k} + \log(k/n) \Big) + O(nq_n)\,, \quad n \to \infty\,.$$

Theorem 3 encourages to choose a level k which minimizes the penalized log-likelihood

$$\frac{-\log L_{\mathbf{M}_k}(\hat{\mathbf{p}}; \mathbf{T}_n) + s}{k} + \log\Big(\frac{k}{n}\Big)\,.$$

It turns out that the additive penalization term $\log(k/n)$ leads to numerical instability as k/n is small. To cope with this issue we upper bound it by $k/n - 1$. The level k which minimizes the criterion is then smaller than the one that appears with $\log(k/n)$. Thus it satisfies more likely the bias assumptions **(B1)**, **(B2)**. So in practice we choose a level k which minimizes the following penalized log-likelihood

$$\frac{-\log L_{\mathbf{M}_k}(\hat{\mathbf{p}}; \mathbf{T}_n) + s}{k} + \frac{k}{n}\,. \tag{4.6}$$

Note that the two steps of our procedure are clearly identified in the penalized log-likelihood. The term $-\log L_{\mathbf{M}_k}(\hat{\mathbf{p}}; \mathbf{T}_n) + s$ corresponds to the bias selection and the multiplicative factor and the additional term to the level one.

4.3 Algorithm: MUltivariate Sparse CLustering for Extremes (MUSCLE)

In practice we choose a large range $\mathcal{K}$ of k (often between 0.5% and 10% of n) and we compute the value of (4.6) for these k and for $s = 1, \ldots, \hat{s}_n$, where $\hat{s}_n$ depends on the chosen level k. We choose $\hat{k}$ which minimizes the penalized log-likelihood (4.6) and then choose $\hat{s}(\hat{k})$ which minimizes (4.6) for $k = \hat{k}$. Then we define $\widehat{\mathcal{S}^*}$ as the set gathering the $\hat{s}(\hat{k})$ clusters corresponding to the largest $T_{n,\hat{k}}(\beta)$'s. Finally we consider the probability vector $\hat{\boldsymbol{\zeta}}$ defined by

$$\hat{\zeta}(\beta) := \frac{T_{n,\hat{k}}(\beta)}{\sum_{\gamma \in \widehat{\mathcal{S}^*}} T_{n,\hat{k}}(\gamma)},$$

for $\beta \in \widehat{\mathcal{S}^*}$ and 0 elsewhere, as an estimator of $\mathbf{p}^*$. Our procedure entails the following parameter-free algorithm called *MUSCLE* for MUltivariate Sparse CLustering for Extremes.

Remark 4. While our procedure leads to the choice of a unique $\hat{k}$, we expect that this approach is not too sensitive to this choice. Therefore, it is relevant to plot the function $k \mapsto \hat{s}(k)$ which provides the chosen value of s for every $k \in \mathcal{K}$. We expect that this function is approximately constant around the chosen value $\hat{k}$.

5 Numerical results

5.1 Overview

The aim of the numerical results is to compare the extremal clusters given by MUSCLE with the theoretical ones in $\mathcal{S}^*(\mathbf{Z})$. To this end we compare the estimated probability

Algorithm 1: MUltivariate Sparse CLustering for Extremes (MUSCLE)

Data: A sample $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}_+^d$ and a range of values $\mathcal{K}$ for the level

Result: A list $\widehat{\mathcal{S}}^*$ of clusters β and the associated probability vector $\hat{\boldsymbol{\zeta}}$.

for $k \in \mathcal{K}$ **do**

Compute $u_n = |\mathbf{X}|_{(k+1)}$ the $(k+1)$-th largest norm;

Assign each $\pi(\mathbf{X}_j/u_n)$ the subsets C_β it belongs to;

Compute $T_{n,k}(\beta)$ for each $\beta \in \mathcal{P}_d^*$;

Compute the minimizer $\hat{s}(k)$ which minimizes the criterion given in Equation (4.4);

end

Choose $\hat{k}$ which minimizes (4.6) plugging in the minimal value in (4.4);

Output: $\widehat{\mathcal{S}}^* = \{$the clusters β associated to the $T_{n,\hat{k},1}, \ldots, T_{n,\hat{k},\hat{s}(\hat{k})}\}$ and $\hat{\boldsymbol{\zeta}}$ as above.

vector $\hat{\boldsymbol{\zeta}}$ with the theoretical one $\mathbf{p}^*$ via the Hellinger distance

$$h(\mathbf{p}^*, \hat{\boldsymbol{\zeta}}) = \frac{1}{\sqrt{2}} \Big[\sum_{\beta \in \mathcal{P}_d^*} (p^*(\beta)^{1/2} - \hat{\zeta}(\beta)^{1/2})^2 \Big]^{1/2}. \tag{5.1}$$

The closer $h(\mathbf{p}^*, \hat{\boldsymbol{\zeta}})$ is to 0, the better $\hat{\boldsymbol{\zeta}}$ estimates $\mathbf{p}^*$. In order to compare our method with some existing ones, we also compute the Hellinger distance between the true probabilities $\mathbb{P}(\boldsymbol{\Theta} \in C_\beta)$ and the estimated ones given by the algorithm called DAMEX of Goix et al. (2017) and the two methods of Simpson et al. (2020). We represent the mean Hellinger distance over $N = 100$ simulations. The parameters in the method of Goix et al. (2017) are chosen to be $\epsilon = 0.1$, $k = \sqrt{n}$, and $p = 0.1$, see the notation in their paper. Regarding the methods of Simpson et al. (2020) we use the parameters given by the authors in Section 4.2 of their paper, i.e. we set $\pi = 0.01$, and $p = 0.5$ and u_β to be the 0.75 quantile of observed

Q values in region C_β for the first method, and $\delta = 0.5$ and u_β to be the 0.85 quantile of observed Q values in region C_β for the first method. We refer to Simpson et al. (2020) for more insights on these parameters.

Remark 5. Contrary to the aforementioned methods, we recall that MUSCLE does not require any hyperparameter. This is a main advantage from a statistical and computational point of view. On the contrary, for the other methods these values could be tuned via cross-validation. For the numerical results we do not choose this approach and keep the values fixed by the authors of the cited papers.

In the following section we develop the example of a max-mixture distribution. The code related to this article can be found at `https://drive.google.com/drive/folders/11TvhbVMPXcSkxmdnnAySvZt64lpMKZqL?usp=sharing`. Another example related to asymptotic independence is given in the Supplementary Material.

5.2 Max-mixture distribution

For any $\beta \in \mathcal{P}_d^*$, let $\mathbf{A}_\beta \in \mathbb{R}_+^{|\beta|}$ be a random vector with standard Fréchet marginal distributions and with dependence structure given below, and let $\{\mathbf{A}_\beta : \beta \in \mathcal{P}_d^*\}$ be independent random vectors. Then the vector $\mathbf{X} = (X_1, \ldots, X_d)^\top$ whose components are defined via $X_j = \max_{\beta \in \mathcal{P}_d^* : i \in \beta} \lambda_{i,\beta} X_{j,\beta}$, with $\lambda_{i,\beta} \in [0,1]$ and $\sum_{\beta \in \mathcal{P}_d^* : i \in \beta} \lambda_{i,\beta} = 1$, has also standard Fréchet marginal distributions and is regularly varying.

For our simulations we consider the five-dimensional example introduced by Simpson et al. (2020) which we recall for completeness. We consider two bivariate Gaussian copulas with correlation parameter ρ and Fréchet marginals $\mathbf{A}_{\{1,2\}}$ and $\mathbf{A}_{\{4,5\}}$, and three extreme-value logistic copulas with dependence parameter α and Fréchet marginals $\mathbf{A}_{\{1,2,3\}}$, $\mathbf{A}_{\{3,4,5\}}$,

and $\mathbf{A}_{\{1,2,3,4,5\}}$. For $\rho < 1$, the Gaussian copula is asymptotically independent (see Example 1 for more insights on this notion) and thus the spectral measure defined in (1.2) concentrates on the subsets $C_{\{1\}}$, $C_{\{2\}}$, $C_{\{4\}}$, and $C_{\{5\}}$. For $\alpha \in (0,1)$ the logistic distribution is asymptotically dependent so that the spectral measure also places mass on the subsets $C_{\{1,2,3\}}$, $C_{\{3,4,5\}}$, and $C_{\{1,2,3,4,5\}}$. Following Simpson et al. (2020), we set

$$\lambda_{\{1,2\}} = (5,5)/7\,, \quad \lambda_{\{4,5\}} = (5,5)/7$$
$$\lambda_{\{1,2,3\}} = (1,1,3)/7\,, \quad \lambda_{\{3,4,5\}} = (3,1,1)/7\,, \quad \lambda_{\{1,2,3,4,5\}} = (1,1,1,1,1)/7\,,$$

so that equal mass is assigned to each of the seven aforementioned subsets. In order to compute the mass the distribution of $\mathbf{Z}$ assigns to every subset C_β we start from the distribution of $\boldsymbol{\Theta}$ and use Monte-Carlo simulation. We then compare these probabilities with their estimated ones $\hat{\boldsymbol{\zeta}}$ given by MUSCLE.

We run our algorithm for different values of $\rho \in \{0, 0.25, 0.5, 0.75\}$ and $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. Figures 2 and 3 shows the average mean Hellinger distance for our method, the one of Goix et al. (2017), and the two of Simpson et al. (2020) over 100 simulations. Our method provides a mean Hellinger distance which stabilizes between 0.2 and 0.3 for all values of ρ and α. For $\alpha \leq 0.7$ the distance slightly decreases with alpha, while it increases for $\alpha \geq 0.8$. The standard deviation is quite small for $\alpha \leq 0.7$ and then increases with α. Regarding the approach of Goix et al. (2017), the mean Hellinger distance tends to increase with α and with ρ. The smallest values is obtained for $\rho \in \{0, 0.25\}$ and for small α. The estimation particularly deteriorates for $\rho = 0.75$. Finally both methods proposed by Simpson et al. (2020) provide a mean Hellinger distance which increases with α and ρ. The second one seems to provide almost always better results than the first one.

While all methods provided by Goix et al. (2017) and Simpson et al. (2020) deteriorate when ρ or α increase, our procedure provides results which stabilize around a mean Hellinger

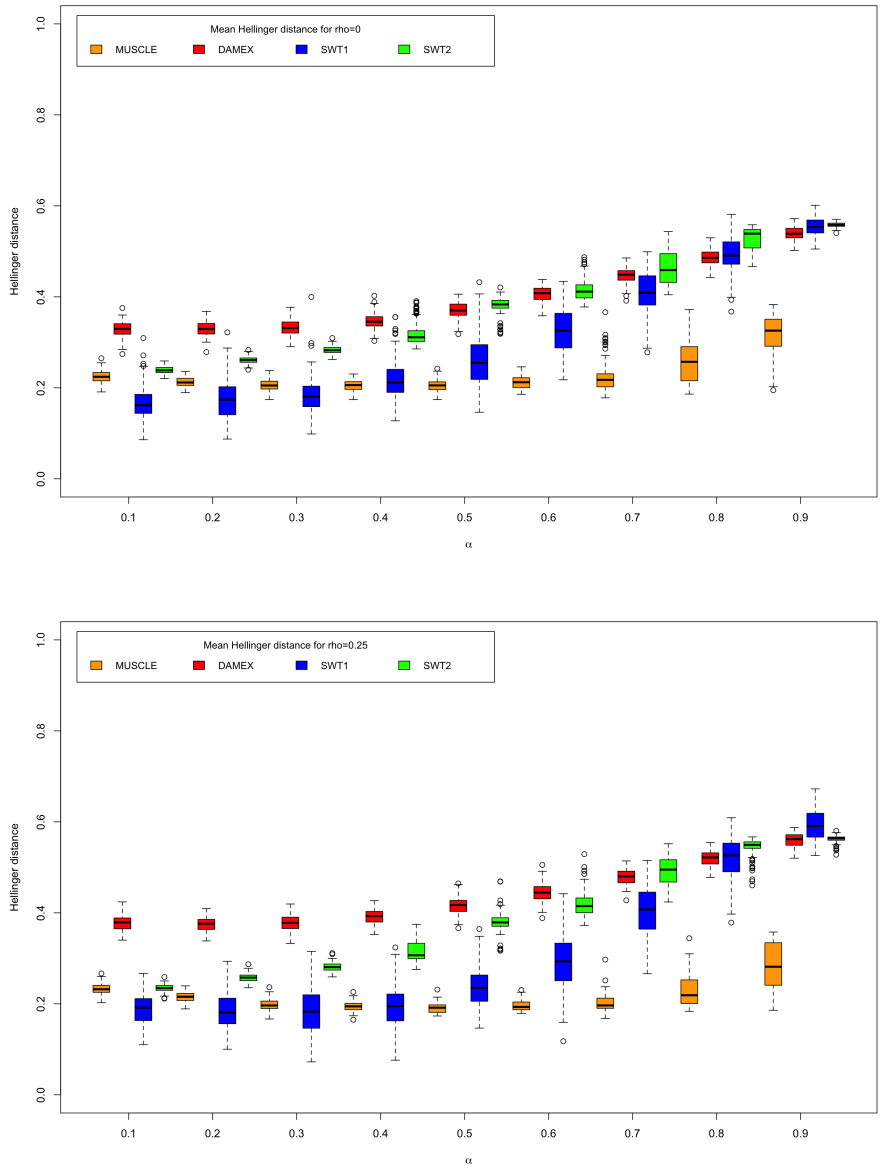


Figure 2: Mean Hellinger distance over 100 simulations for $\rho = 0$ (top) and $\rho = 0.25$ (bottom). The abbreviation SWT1 (resp. SWT2) refers to the first (resp. second) method of Simpson et al. (2020).

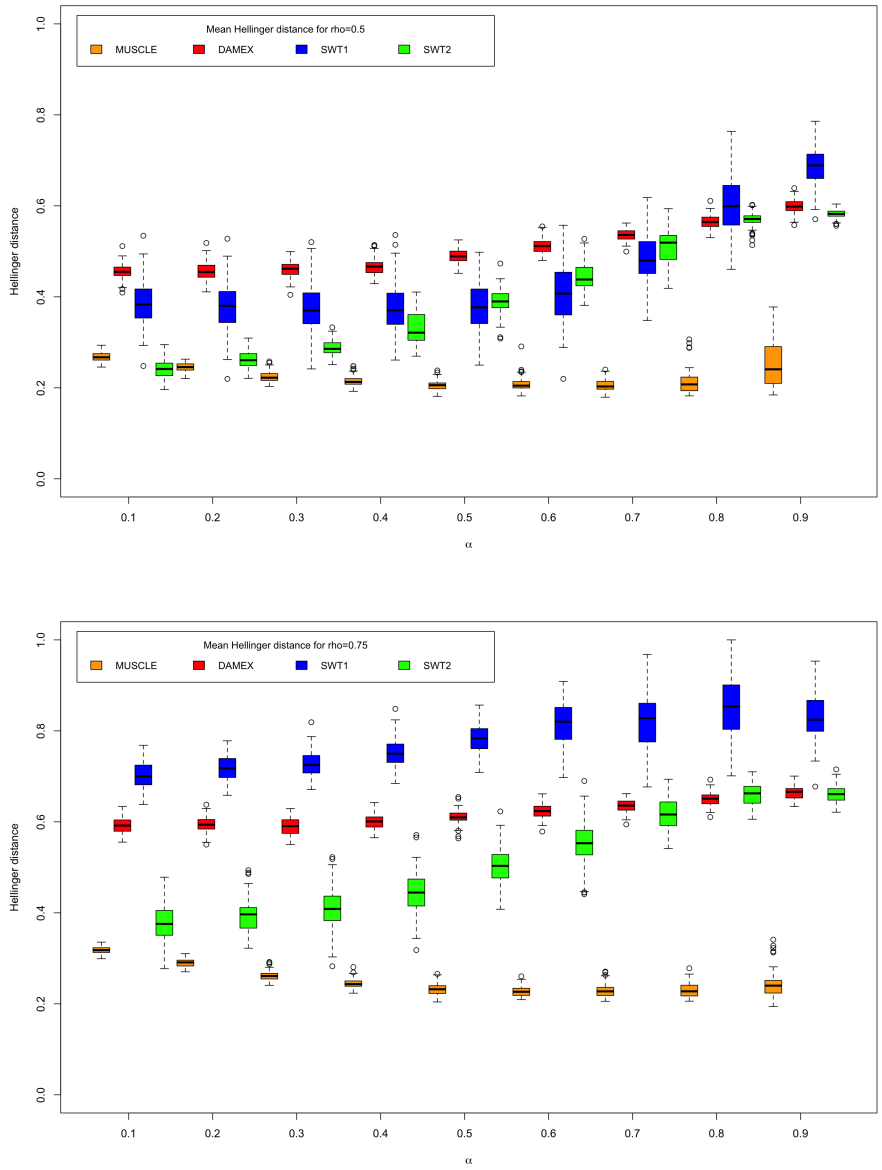


Figure 3: Mean Hellinger distance over 100 simulations for $\rho = 0.5$ (top) and $\rho = 0.75$ (bottom). The abbreviation SWT1 (resp. SWT2) refers to the first (resp. second) method of Simpson et al. (2020).

distance of 0.2. This distance is the smallest one for $\rho = 0.5$ and $\rho = 0.75$ for all α compared to the one of the three other methods. For small ρ, MUSCLE better performs for large α. For small α the second method of Simpson et al. (2020) provides better results than our approach, while its standard deviation is larger. It turns out that except for small α with $\rho = 0$ and $\rho = 0.25$ our algorithm better detects the extremal clusters.

6 Application to real-world data

6.1 Preprocessing for real-world data

In Section 5 we considered an example with standard Fréchet marginal distributions so that the tail index (see Equation (1.2)) of the considered vectors is equal to 1. The influence of this index on the extremal clusters has been studied on some numerical results by Meyer and Wintenberger (2021). It turns out that a large tail index does not provide accurate results while a small one highlights one-dimensional clusters, see Remark 11 in their article. A tail index of $\alpha = 1$ seems to provide the best results.

For real-world data the estimation of the tail index of a sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ is achieved with a Hill plot (Hill (1975)). It consists in plotting

$$\widehat{\alpha}(k) = \left(\frac{1}{k} \sum_{j=1}^{k} \log(|\mathbf{x}|_{(j)}) - \log(|\mathbf{x}|_{(k)}) \right)^{-1}, \quad k = 2, \dots, n\,,$$

where $|\mathbf{x}|_{(j)}$ denotes the order statistics of the norms $|\mathbf{x}_1|, \dots, |\mathbf{x}_n|$, i.e. $|\mathbf{x}|_{(1)} \geq \dots \geq |\mathbf{x}|_{(n)}$, and to choose $\hat{\alpha}$ as the value around which the plot stabilizes. Then, we consider the power transform $\mathbf{x}'_j = (\mathbf{x}_j)^{\hat{\alpha}}$. This transformation highlights the tail structure of the data without modifying the support of the spectral measure, see Meyer and Wintenberger (2021),

Remark 8. It differs from the standardization discussed in the introduction for which the vectors are normalized via a rank transform.

In the following section we apply MUSCLE to financial data. An application on wind speed data can be found in the Supplementary Material.

6.2 Extreme variability for financial data

The data set we use corresponds to the value-average daily returns of 49 industry portfolios compiled and posted as part of the Kenneth French Data Library. They are available at `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. A related study on a similar dataset has been conducted by Cooley and Thibaud (2019). We restrict our study to the period $1970-2019$ which provides $n = 12\,613$ observations denoted by $\mathbf{x}_1^{\text{obs}}, \dots \mathbf{x}_n^{\text{obs}} \in \mathbb{R}^{49}$. Our goal is to study the variability of these returns so that we take the componentwise absolute value $\mathbf{x}_j = |\mathbf{x}_j^{\text{obs}}|$ of the data. Thus, we study the non-negative vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ in $\mathbb{R}_+^d$ with $n = 12\,613$ and $d = 49$. Following Section 6.1, we consider the vectors $\mathbf{x}'_j = (\mathbf{x}_j)^{\hat{\alpha}}$, where $\hat{\alpha} = 2.99$ is the Hill estimator of the sample $|\mathbf{x}_1|, \dots, |\mathbf{x}_n|$.

Following Remark 4, we plot the evolution of the estimator of the Kullback-Leibler divergence in (4.6) as a function of k. We see on Figure 4 that this estimator decreases until it reaches a minimal value for $\hat{k} = 441$, before increasing for $k \geq \hat{k}$. The level $\hat{k}$ corresponds to a proportion $\hat{k}/n = 3\%$ and leads to a number of extremal clusters $\hat{s}(\hat{k}) = 14$. Contrary to the numerical results, we do not observe a range of k for which the minimal value $\hat{s}(k)$ remains approximately constant.

MUSCLE provides $\hat{s}(\hat{k}) = 14$ extremal clusters which gather 12 portfolios. These clusters and their inclusions are represented in Figure 5. The number of identified clusters is much smaller compared to the total number $2^{49} \approx 10^{15}$. Besides these clusters are

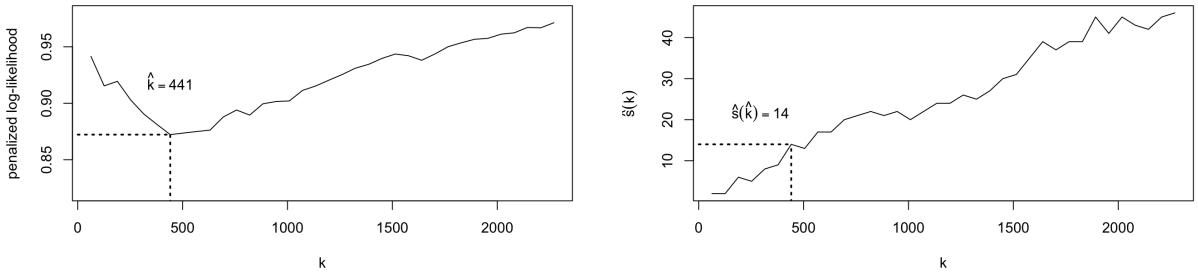


Figure 4: Evolution with respect to k of the penalized log-likelihood given in (4.6) (left) and of $\hat{s}_n(k)$ (right) for the financial data.

at most three-dimensional so that our procedure drastically reduces the dimension of the study. Most of the extremal portfolios which appear in the clusters correspond to "office/executive" sectors, such as Health, Software, Hardware, Banks, Finance, Electronic Equipment (Chips), Real Estate. Some other clusters group portfolios related to heavy industries, such as Steel, Coal, and Gold. The only clusters gathering a heavy industry and service sectors are {Coal, Banks} and {Coal, Banks, Fin}. The tail dependence of the variability of these different kinds of portfolios may result from the financing of the coal industry by several big banks, see Raval et al. (2020).

We conclude that the aforementioned 14 clusters given by MUSCLE correspond to subsets C_β which gather the mass of $\mathbf{Z}$. Among them, eight gather some mass of $\mathbf{Z}$ and are not included in larger subsets on which $\mathbf{Z}$ places mass. Following Meyer and Wintenberger (2021), Theorem 2, these *maximal* subsets also concentrate the mass of the spectral measure. We refer to Meyer and Wintenberger (2021), Section 3.2, for a discussion on maximal and non-maximal subsets. Standard approaches which hold for low-dimensional extremes can then be applied to these subsets, see Einmahl et al. (1993), Einmahl et al. (1997),

Einmahl and Segers (2009).

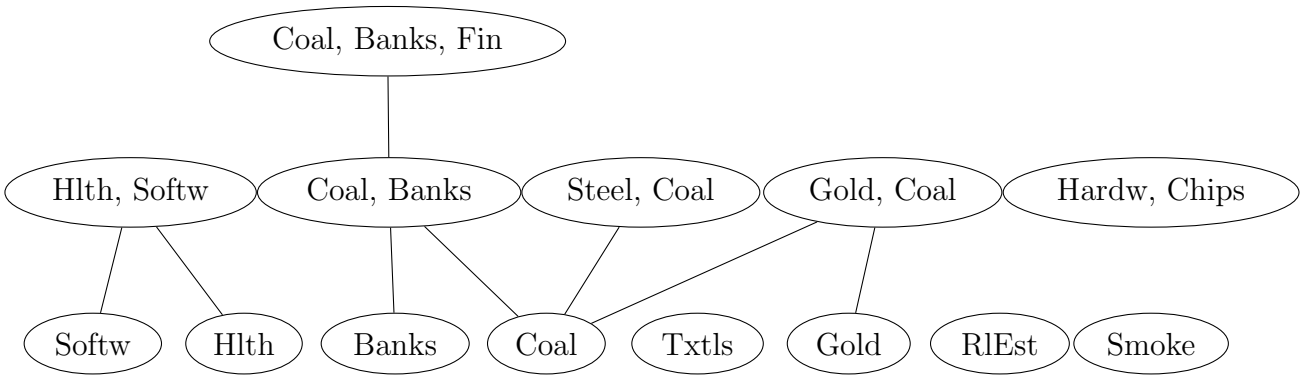


Figure 5: Representation of the 14 clusters and their inclusions. The abbreviations are the following ones: Softw = Computer Software, Txtls = Textiles, Hlth = Healthcare, RlEst = Real Estate, Hardw = Hardware, Chips = Electronic Equipment, Fin = Finance.

After removing the 12 extremal components we reapply MUSCLE to obtain the dependence structure of the non-extremal portfolios. The algorithm provides a unique cluster with all 37 remaining portfolios. Hence these portfolios tend to have a dependent tail structure: their extreme variability is strongly correlated.

7 Conclusion

The statistical analysis introduced in this article provides a new approach to detect the extremal directions of a multivariate random vector $\mathbf{X}$. This method relies on the notion of sparse regular variation which better highlights the tail dependence of $\mathbf{X}$. Several convergence results are established in Section 3 and are used to build a rigorous statistical method based on model selection. This approach provides not only the clusters of directions on which the extremes of $\mathbf{X}$ gather but also a reasonable threshold above which the

data are considered as extreme values. The latter issue has always been challenging and no theoretical-based procedure has been provided in a multivariate setting yet, even if it has been the subject of much attention in the literature. The choice of the directions is achieved with an AIC-type minimization whose penalization allows to reduce the number of selected subsets. Including the choice of the level k of the random threshold $|\mathbf{X}|_{(k)}$ then entails multiplicative and additive penalization terms. This approach leads to the parameter-free algorithm MUSCLE whose purpose is to recover the extremal clusters of a sample of iid sparsely regularly varying random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

The absence of any hyperparameter is a main difference with the existing methods (Goix et al. (2017), Simpson et al. (2020), Chiapino and Sabourin (2016), Chiapino et al. (2019)). Another main advantage of our procedure is that it is still efficient for large d. This follows from the expected linear-time algorithm introduced by Duchi et al. (2008) to compute the Euclidean projection.

The numerical experiments on max-mixture distributions provide promising results. Our algorithm provides better results than the ones of Goix et al. (2017) and Simpson et al. (2020) for ρ close to 1, or small ρ and α close to 1. Moreover the results do not vary a lot with ρ and α. Finally, the application of our algorithm on financial data highlights sparse clusters and thus reduces the dimension of the study. We obtain a sparse tail dependence structure for for the extreme variability of several industry portfolios. This reinforces the relevance of our approach for reducing the dimension in Extreme Value Theory.

Acknowledgments We are grateful to two referees for careful reading of the paper and for useful suggestions.

References

Abdous, B. and Ghoudi, K. (2005). Non-parametric estimators of multivariate extreme dependence functions. *Nonparametric Statistics* **17,** 915–935.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory,* 267–281.

Beirlant, J. and Goegebeur, Y. and Segers, J. and Teugels, J. L. (2006). *Statistics of Extremes: Theory and Applications.*, John Wiley & Sons Ltd., Chichester.

Bingham, N.H. and Goldie, C.M. and Teugels, J. L. (1987). *Regular Variation.*, Cambridge University Press, Cambridge.

Caiero F. and Gomes, M.I. (2015). Threshold selection in extreme value analysis. in *Extreme Value Modeling and Risk Analysis: Methods and Applications,* 71–89.

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

Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. *International Workshop on New Frontiers in Mining Complex Patterns,* 132–147.

Chiapino, M., Sabourin, A. and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes* **22,** 193–222.

Condat L. (2016). Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming* **158,** 575–585.

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

Duchi, J. and Shalev-Shwartz, S. and Singer, Y. and Chandra, T. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. *Proceedings of the 25th international conference on Machine learning*, 272–279.

Einmahl, J. and de Haan, L. and Huang, X. (1993). Estimating a multidimensional extreme-value distribution. *Journal of Multivariate Analysis* **47,** 35–47.

Einmahl, J. and de Haan, L. and Sinha, A.K. (1997). Estimating the spectral measure of an extreme value distribution. *Stochastic Processes and their Applications* **70,** 143–171.

Einmahl, J. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics* **37,** 2953–2989.

Goix, N. and Sabourin, A. and Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis* **161,** 12–31.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer, New-York.

de Haan, L. and Resnick, S.I. (1993). Estimating the limit distribution of multivariate extremes. *Stochastic Model* **9,** 275–309.

Heffernan, J. E and Tawn, J. A (1989). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66,** 497–546.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* **3,** 1163–1174.

Hult, H. and Lindskog, F. (1989). Regular variation for measures on metric spaces. *Publications de l'Institut Mathématique* **80,** 121–140.

Kiriliouk, A. and Rootzén, H. and Segers, J. and Wadsworth, J. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics* **61,** 123–135.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22,** 79–86.

Kyrillidis, A., Becker, S., Cevher, V. and Koch, C. (2013). Sparse projections onto the simplex. *International Conference on Machine Learning* **28,** 235–243.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika* **83,** 169–187.

Lindskog, F. and Resnick, S. I. and Roy, J. (2014). Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probability Survey* **11,** 270–314.

Massart, P. (2007). *Concentration inequalities and model selection*, Springer, Berlin.

Meyer, N. and Wintenberger, O. (2021). Sparse regular variation. *Advances in Applied Probability*, **53,** 1115 - 1148.

Raval, A. and Owen, W. and Hume, N. and Stephen, M. (2020). Biggest Banks Sustain Coal Financing despite Defunding Drive. *Financial Times*, 3 Sept. 2020.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer, New-York.

Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling.* Springer, New-York.

Segers, J. (2012). Max-stable models for multivariate extremes. *Revstat Statistical Journal* **10,** 61–82.

Simpson, E., Wadsworth, J. L and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika* **107,** 513–532.

Stărică, C. (1999). Multivariate extremes for models with constant conditional correlations. *Journal of Empirical Finance* **6,** 515–553.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics.* Springer Science & Business Media, New-York.

Wan, P. and Davis, R.A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes* **22,** 131–166.