Tail inference for high-dimensional data

Nicolas Meyer, Olivier Wintenberger

Tail inference for high-dimensional data

Nicolas Meyer*[1] and Olivier Wintenberger†[1]

[1]*Sorbonne Université, LPSM, F-75005, Paris, France*

July 22, 2020

Abstract

Identifying directions where severe events occur is a major challenge in multivariate Extreme Value Analysis. The support of the spectral measure of regularly varying vectors brings out which coordinates contribute to the extremes. This measure is defined via weak convergence which fails at providing an estimator of its support, especially in high dimension. The estimation of the support is all the more challenging since it relies on a threshold above which the data are considered to be extreme and the choice of such a threshold is still an open problem. In this paper we extend the framework of sparse regular variation introduced by Meyer and Wintenberger (2020) to infer tail dependence. This approach relies on the Euclidean projection onto the simplex which exhibits sparsity and reduces the dimension of the extremes' analysis. We provide an algorithmic approach based on model selection to tackle both the choice of an optimal threshold and the learning of relevant directions on which extreme events appear. We apply our method on numerical experiments to highlight the relevance of our findings. Finally we illustrate our approach with financial return data.

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

1 Introduction

Extreme events are often a consequence of the simultaneous occurrence of several large components. Identifying the joint extremal directions in the data is therefore crucial to evaluate the risk and predict future large events. In financial quantitative risk management, risk analyst seeks at detecting the probability that several firms incur together huge losses. In climate science it is important to identify areas which can be impacted simultaneously by a severe event (for instance a heavy rainfall, a heat wave, or a flood). In an oceanographic context the sea-level process can be explained by several factors like the tidal level, the mean-sea level, or the surge level (see Tawn (1992)). Therefore high sea-levels are often due to the simultaneous occurrence of extreme values among these components. These applications fit into the context of multivariate Extreme Value Theory (EVT) which provides models to study severe events and to assess the tail structure of a d-dimensional random vector $\mathbf{X}$ (see e.g. Resnick (1987), Beirlant et al. (2006), or Resnick (2007)).

nicolas.meyer@upmc.fr (corresponding author)
†*olivier.wintenberger@upmc.fr*

1.1 Regular variation

For a random vector $\mathbf{X}$ in $\mathbb{R}_+^d$ the purpose of EVT is to assess the tail structure of $\mathbf{X}$. To this end, it is customary to assume that $\mathbf{X}$ is regularly varying: There exist a positive sequence (a_n), $a_n \to \infty$ as $n \to \infty$, and a non-null Radon measure μ on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ such that

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

where $\xrightarrow{v}$ denotes vague convergence in the space of non-null Radon measures on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ (see for instance Resnick (2007), Theorem 6.1). The limit measure μ is called the *tail measure* and satisfies the homogeneity property $\mu(tA) = t^{-\alpha}\mu(A)$ for any set $A \subset \mathbb{R}_+^d \setminus \{\mathbf{0}\}$ and any $t > 0$. The parameter $\alpha > 0$ is called the *tail index*. Regular variation is used in various fields of applied probability or statistics (see Bingham et al. (1989) for an encyclopedia of results on this topic).

Regular variation provides an interpretable description of the extremal behavior of $\mathbf{X}$ via the tail measure. While the tail index highlights the intensity of the extremes (the smaller this index is, the heaviest the tail of $\mathbf{X}$ could be), the support of the tail measure indicates the directions on which large events occur. It is therefore more convenient to use a polar representation for μ in order to separately study the radial part and the angular part of $\mathbf{X}$, see Resnick (1986) and Beirlant et al. (2006), Section 8. Equation (1.1) is equivalent to the existence of a random vector $\mathbf{\Theta}$ on $\mathbb{S}_+^{d-1}$ and a Pareto(α)-distributed random variable Y independent of $\mathbf{\Theta}$ such that

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \frac{\mathbf{X}}{|\mathbf{X}|}\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,. \tag{1.2}$$

The random variable Y is the limit of the radial component $|\mathbf{X}|/t$. It models the intensity of extreme events and is characterized by the tail index α. The vector $\mathbf{\Theta}$, called the *spectral vector*, and its distribution $S = \mathbb{P}(\mathbf{\Theta} \in \cdot)$, the *spectral measure*, model the angular component of the tail of $\mathbf{X}$. The subspaces of the positive unit sphere on which the spectral vector concentrates correspond to the directions where large events are likely to appear. Note that the choice of the norm in Equation (1.2) is arbitrary. A convenient choice in our setting will be specify later.

From a statistical point of view, studying multivariate extremes consists in estimating the parameter α and the spectral measure. The former estimation is parametric and has been widely studied, for instance by Hill (1975), Smith (1987) or Beirlant et al. (1996). On the contrary providing accurate estimators of the spectral measure is a challenging problem, even more challenging in high dimensions. Until recently useful representations of the spectral measure have only been introduced in the bivariate case, see for instance Einmahl et al. (1993), Einmahl et al. (1997), Einmahl et al. (2001) and Einmahl and Segers (2009). Parametric approaches have also been proposed to tackle the study of extremes in moderate dimensions ($d \leq 10$) for instance by Coles and Tawn (1991) and Sabourin et al. (2013).

In higher dimensions only a small number of marginals is likely to be simultaneously large. This implies that there are many parts of the unit sphere on which the spectral measure could not put mass. In this case we say that this measure (or equivalently the spectral vector $\mathbf{\Theta}$) is *sparse*. In other words, the probability $\mathbb{P}(|\mathbf{\Theta}|_0 \ll d)$ is close to 1, where $|\cdot|_0$ denotes the ℓ^0-norm, that is, the number of non-null coordinates of a vector. Hence, the high-dimensional inference of the spectral measure boils down to the study of the low-dimensional subspaces on which the spectral measure concentrates. Regarding the estimation of the spectral measure's support, Lehtomaa and Resnick (2019) map the unit sphere to the $d-1$ dimensional space $[0,1]^{d-1}$ in order to partition it in equally sized rectangles. The study of the spectral measure's support is thus done with grid estimators. This estimation is combined with support testing to validate the proposed estimators. The authors apply their method to various examples as daily stock returns or rainfalls.

1.2 Dimension reduction in EVT

Since the complete support's estimation is often difficult to capture, a main objective is to identify the directions on which the spectral measure places mass. Different techniques have been recently proposed. A first type of approaches consists in adapting Principal Component Analysis (PCA) in an extremal setting. In this context Cooley and Thibaud (2019) define a particular inner product space on the positive orthant $\mathbb{R}_+^d$ in order to combine linear algebra and regular variation. This setting is then used to model the tail dependence through a matrix of pairwise tail dependence metrics. The authors apply their method to capture the extremal dependence in Swiss rainfall data and financial return data. Subsequently, Sabourin and Drees (2019) use PCA in a context of risk minimization. They assume that the vector space spanned by the support of the tail measure μ has dimension $p \ll d$. In order to identify this subspace the authors establish finite sample bounds on the reconstruction vector.

An approach related to clustering methods has been introduced in the seminal paper of Chautru (2015) who uses spherical data analysis to capture the tail dependence structure. In this context it is convenient to partition the space $\{\mathbf{x} \in \mathbb{R}_+^d,\ |\mathbf{x}|> 1\}$ or the positive unit sphere $\mathbb{S}_+^{d-1}$ in terms of the nullity of the coordinates. For $\beta \subset \{1, \dots, d\}$ the subspaces R_β and C_β are defined as

$$R_\beta = \{\mathbf{x} \in \mathbb{R}_+^d,\ |\mathbf{x}|> 1,\ x_i > 0 \text{ for } i \in \beta,\ x_i = 0 \text{ for } i \notin \beta\} \tag{1.3}$$

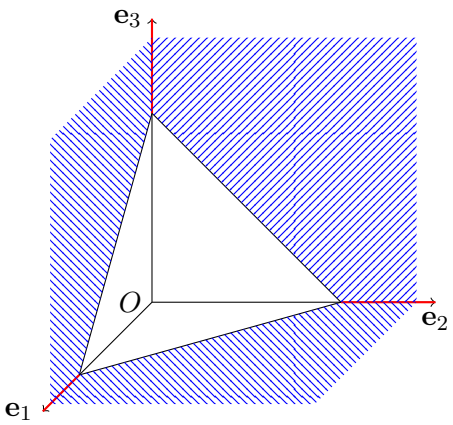
and

$$C_\beta = \{\mathbf{x} \in \mathbb{S}_+^{d-1},\ x_i > 0 \text{ for } i \in \beta,\ x_i = 0 \text{ for } i \notin \beta\}\,. \tag{1.4}$$

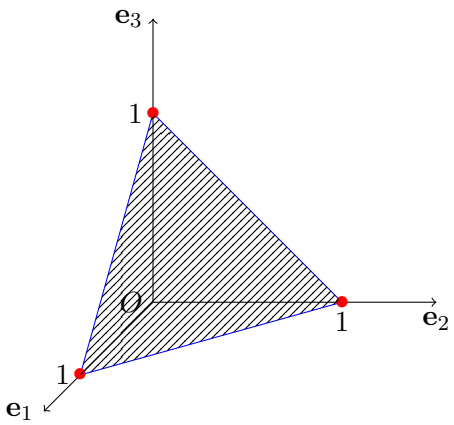
The subsets R_β (respectively C_β) are pairwise disjoint and form a partition of $\{\mathbf{x} \in \mathbb{R}_+^d,\ |\mathbf{x}|> 1\}$ (respectively $\mathbb{S}_+^{d-1}$):

$$\{\mathbf{x} \in \mathbb{R}_+^d,\ |\mathbf{x}|> 1\} = \bigsqcup_{\substack{\beta \subset \{1,\dots,d\} \\ \beta \neq \emptyset}} R_\beta \qquad \text{and} \qquad \mathbb{S}_+^{d-1} = \bigsqcup_{\substack{\beta \subset \{1,\dots,d\} \\ \beta \neq \emptyset}} C_\beta\,,$$

where $\bigsqcup$ denotes a disjoint union. An illustration of these subsets in dimension 3 is given in Figure 1.



(a) The subsets R_β in dimension 3 for the ℓ^1-norm. In red the one-dimensional subspaces spanned by $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. In shaded blue the two-dimensional ones. For readability purposes the full subspace $\{|\mathbf{x}|> 1,\ \forall i \in \{1,2,3\}, x_i > 0\}$ is not represented.



(b) The subspaces C_β in dimension 3 for to the ℓ^1-norm. In red the vectors $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. In blue the lines of the simplex's edges. The shaded part corresponds to the interior of the simplex.

Figure 1: The subspaces R_β and C_β in dimension 3 for the ℓ^1-norm.

Regarding the regularly varying random vector $\mathbf{X}$, these partitions highlight its extremal structure. Indeed, for a fixed $\beta \subset \{1, \dots, d\}$ the inequality $\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0$ means that it is likely to observe

simultaneously large values in the directions of β and small values in the directions of β^c. Identifying the subsets C_β allows then to bring out clusters of coordinates which can be large together. Hence, the main goal of the spectral measure's estimation consists in classifying the $2^d - 1$ probabilities $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ depending on their nullity or not. Several ideas have been developed on this topic. Goix et al. (2017) focus on the tail measure μ and estimate the mass this measure places on the subsets R_β. This estimation relies on a hyperparameter $\epsilon > 0$ and brings out a representation of the dependence structure. An algorithm called DAMEX (for *Detecting Anomalies among Multivariate EXtremes*) is proposed and reaches a complexity $O(dn \log n)$, where n corresponds to the number of data points. A similar method has been used by Chiapino and Sabourin (2016) and Chiapino et al. (2019) who provide other algorithms called CLEF (for *CLustering Extremal Features*) that gather the features that are likely to be extreme simultaneously. Simpson et al. (2019) use the concept of hidden regular variation introduced by Resnick (2002) and propose a set of parameters $(\tau_\beta)_\beta$ to describe the extremal behavior of the data on the subsets $(C_\beta)_\beta$ (see also Ledford and Tawn (1996) and Ledford and Tawn (1997)).

All these approaches are based on the standard definition of regular variation which does not provide a natural estimator for the spectral measure. Indeed, a subset C_β with $\beta \neq \{1, \ldots, d\}$ has empty interior so that $C_\beta \subset \partial C_\beta$. Hence, if $\mathbf{\Theta}$ satisfies $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ then C_β it is not a S-continuity set and the convergence in (1.2) fails. Empirically, this means that even if the spectral measure places mass on a low-dimensional subspace C_β, the vector $\mathbf{X}$ does not, except if it is degenerated in other directions. In order to circumvent this issue and to take the potential sparse structure of the spectral measure into account, we use the approach developed by Meyer and Wintenberger (2020). It consists in replacing the self-normalization which appears in the second component of Equation (1.2) by the Euclidean projection onto the simplex of $\mathbf{X}/t$. This projection has been widely studied in the learning community, see e.g. Duchi et al. (2008), Kyrillidis et al. (2013), or Liu and Ye (2009). This substitution provides a new angular limit which differs from the spectral vector $\mathbf{\Theta}$ and entails a new concept called *sparse regular variation* with which the sparsity structure of extreme events can be more easily captured.

1.3 Choice of the threshold via model selection

In a non-asymptotic context multivariate models in EVT consider Equation (1.2) as an approximation when the threshold t is "high enough". The choice of an optimal threshold in a statistical setting boils down to choosing a number k of vectors among the data which will be considered as extreme. The smaller k the closer we are from the theoretical framework. However it is still needed to keep a substantial number of extreme data to correctly learn the tail structure. This balanced choice is a major problem in EVT and no theoretical result has yet been obtained in a multivariate setting.

Several authors point out that the choice of an appropriate threshold for which the convergence in Equation (1.2) is a good approximation is a challenging task in practice (see for instance Rootzén et al. (2006)). This choice is however tackled in a few articles. Abdous and Ghoudi (2005) propose an automatic selection technique in the bivariate case which is based on a double kernel technique (see Devroye (1989)). A more abundant literature deals with marginals threshold selection, see Caeiro and Gomes (2015) for a review on these techniques. In a multivariate framework threshold selection for dependence models has been studied by Lee et al. (2015) who apply Bayesian selection and by Kiriliouk et al. (2019) who use stability properties of the multivariate Pareto distribution. Stărică (1999) provides an approach based on the homogeneity property $\mu(tA) = t^{-\alpha}\mu(A)$. The idea is to plot the function

$$u \mapsto \frac{\hat{\mu}(uA)}{u^{-\hat{\alpha}}\hat{\mu}(A)},$$

for different k, where $\hat{\mu}$ and $\hat{\alpha}$ are estimators of μ and α. This ratio should hover around 1 for a suitable choice of k and for u in a neighborhood of 1 (see also Resnick (2007), Section 9.4.2). Recently, Wan and Davis (2019) propose a threshold selection based on the independence of Y and $\mathbf{\Theta}$ in Equation (1.2).

Their procedure consists in testing if the components $|\mathbf{X}|/t$ and $\mathbf{X}/|\mathbf{X}|$ given $|\mathbf{X}|> t$ are independent. The threshold t is then chosen as the smallest for which independence holds.

In this non-asymptotic context the aim is to provide a method in order to identify an optimal value of the level k. To this end we define a family of models which depend on this level k and we provide a procedure which allows to select the most accurate model. This approach relies on model selection (Massart (2007)) and is similar to the minimization criterion developed by Akaike (1973). The proposed method actually tackles simultaneously the choice of this level with the estimation of the dependence structure.

1.4 Goal and outline of this paper

The purpose of this paper is to address the two following questions. Which directions $\beta \subset \{1, \ldots, d\}$ gather extreme events? Which threshold should be chosen to obtain a good approximation of the limit in (1.2)? To this end, we extend the theoretical framework developed by Meyer and Wintenberger (2020) to a statistical context. The idea is to use the notion of sparse regular variation which better highlights the tail structure of a random vector $\mathbf{X}$. We introduce convenient estimators for the proportion of extreme values in a direction $A \subset \mathbb{S}_+^{d-1}$. We apply these results on the subsets C_β for which we obtain a multivariate asymptotic normality. Identifying the most relevant subsets which gather extreme values is done via model selection. The idea is to deal simultaneously with the selection of these subsets and the choice of an optimal threshold. Hence we consider theoretical models which differ according to the number of C_β chosen and to the level k. Based on a sample of independent and identically distributed (iid) regularly varying random vectors we provide a procedure to select the sparse model which best fits the data. To the best of our knowledge our work is the first attempt to combine estimation of the tail dependence and choice of a threshold.

The paper is organized as follows. Section 2 introduces the theoretical background on multivariate EVT and model selection that is needed throughout the paper. We deal with the notions of regular variation and sparse regular variation and detail how the new way of projecting affects the convergence (1.2). We also discuss some convergence results and explain why this approach is useful to capture the sparsity structure of extreme events. In Section 3 we apply our theoretical results on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in order to introduce convenient estimators for the proportion of extreme values in a given subset A of $\mathbb{S}_+^{d-1}$. Univariate consistency and univariate asymptotic normality are proven. In Section 4 we restrict our study ot the subsets C_β and we extend the convergence results by considering all subsets simultaneously. Section 5 is devoted to the selection of the most relevant directions β and to the choice of an optimal level k. In Section 6 we illustrate our findings on different numerical results. The results are very promising since we obtain a good accuracy between the chosen model and the theoretical one. Finally we illustrate in Section 7 our approach on financial data.

2 Theoretical background

2.1 Notation

We introduce some standard notation that is used throughout the paper. 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 $\mathbf{x} \in \mathbb{R}^d$ is a vector, then $\mathrm{Diag}(\mathbf{x})$ denotes the diagonal matrix of $\mathcal{M}_d$ whose diagonal is $\mathbf{x}$. We define $\mathbb{R}_+^d = \{\mathbf{x} \in \mathbb{R}^d,\, x_1 \geq 0, \ldots, x_d \geq 0\}$, $\mathbf{e} = (1, \ldots, 1)^\top \in \mathbb{R}^d$, $\mathbf{0} = (0, \ldots, 0)^\top \in \mathbb{R}^d$. For $j = 1, \ldots, d$, $\mathbf{e}_j$ denotes the j-th vector of the canonical basis of $\mathbb{R}^d$. If $\mathbf{x} \in \mathbb{R}^d$ and $I = \{i_1, \ldots, i_r\} \subset \{1, \ldots, d\}$, then $\mathbf{x}_I$ denotes the vector $(x_{i_1}, ..., x_{i_r})$ of $\mathbb{R}^r$. For $p \in (1, \infty]$, we denote by $|\cdot|_p$ the ℓ^p-norm in $\mathbb{R}^d$. For the sake of simplicity, in all this article $|\cdot|$ denotes ℓ^1-norm and $\mathbb{S}_+^{d-1}$ the simplex $\{\mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = 1\}$. More generally $\mathbb{S}_+^{d-1}(z) = \{\mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = z\}$

for $z > 0$. For a set E, we denote by $\mathcal{P}(E)$ its power set: $\mathcal{P}(E) = \{A,\, A \subset E\}$, and we use the notation $\mathcal{P}^*(E) = \mathcal{P}(E)\setminus\{\emptyset\}$. If $E = \{1,\dots,r\}$, we simply write $\mathcal{P}_r = \mathcal{P}(\{1,\dots,r\})$ and $\mathcal{P}_r^* = \mathcal{P}(\{1,\dots,r\})\setminus\{\emptyset\}$. For a finite set E we denote by $|E|$ its cardinality. Finally we write $\xrightarrow{d}$ for the convergence in distribution, $\xrightarrow{w}$ for the weak convergence, and $\xrightarrow{v}$ for the vague convergence.

2.2 Regular variation and sparse regular variation

We consider a regularly varying random vector $\mathbf{X} \in \mathbb{R}_+^d$ satisfying convergence (1.1). The sparsity structure of the tail measure μ can be studied via the subsets R_β while the one of the spectral measure can be studied via the C_β's. The advantage of these subsets is that they are interpretable in terms of tail dependence while they reduce the dimension of the study. The main objective is then to identify the subspaces on which the spectral measure $S(\cdot) = \mathbb{P}(\mathbf{\Theta} \in \cdot)$ places mass, i.e. the ones which satisfy $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$. To this end, we define the set

$$\mathcal{S}(\mathbf{\Theta}) := \{\beta \in \mathcal{P}_d^*,\, \mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0\}\,. \tag{2.1}$$

In order to get a better understanding of the directions in $\mathcal{S}(\mathbf{\Theta})$ we introduce the notion of maximal direction (see Meyer and Wintenberger (2020), Definition 1).

Definition 1 (Maximal direction for $\mathbf{\Theta}$)**.** *Let $\beta \in \mathcal{P}_d^*$. We say that a direction β is* maximal *for $\mathbf{\Theta}$ if*

$$\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0 \quad \textit{and} \quad \mathbb{P}(\mathbf{\Theta} \in C_{\beta'}) = 0,\ \textit{for all } \beta' \supsetneq \beta\,. \tag{2.2}$$

It is straightforward to see that every direction β which satisfies $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ is included in a maximal one. In terms of extremes a direction β is maximal if it is likely that all components $j \in \beta$ are simultaneously extreme while there exists no set $\beta' \supsetneq \beta$ such that β' gathers extremes. Thus, identifying the maximal directions β allows to capture some trends in the tail dependence structure of $\mathbf{X}$.

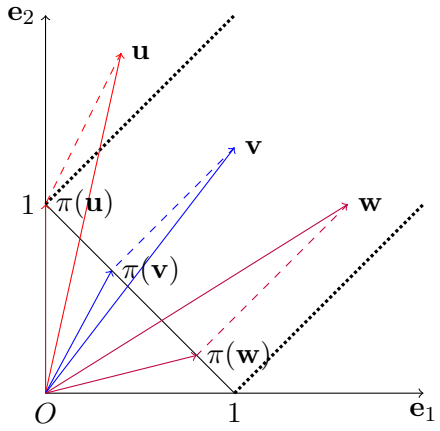
The estimation of the probabilities $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ can not be easily addressed based on the self-normalized vector $\mathbf{X}/|\mathbf{X}|$, see Section 1.2. We address this issue with the ideas of Meyer and Wintenberger (2020) replacing the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ by $\pi(\mathbf{X}/t)$ where π denotes the Euclidean projection onto the simplex (Gafni and Bertsekas (1984), Duchi et al. (2008), Kyrillidis et al. (2013), Condat (2016)). We use the device of Duchi et al. (2008). For $z > 0$ and $\mathbf{v} \in \mathbb{R}_+^d$ the projected vector $\pi_z(\mathbf{v})$ is the unique vector $\mathbf{w}$ of $\mathbb{S}_+^{d-1}(z)$ that minimizes the quantity $|\mathbf{w} - \mathbf{v}|_2^2$. This vector is defined by $w_i = (v_i - \lambda_{\mathbf{v},z})_+$, where $\lambda_{\mathbf{v},z} \in \mathbb{R}$ is the only constant satisfying the relation $\sum_{1\le i\le d}(v_i - \lambda_{\mathbf{v},z})_+ = z$. The Euclidean projection π_z onto the positive sphere $\mathbb{S}_+^{d-1}(z)$ is then defined as

$$\begin{array}{rccc} \pi_z & : & \mathbb{R}_+^d & \to & \mathbb{S}_+^{d-1}(z) \\ & & \mathbf{v} & \mapsto & \mathbf{w} = (\mathbf{v} - \lambda_{\mathbf{v},z})_+\,. \end{array}$$

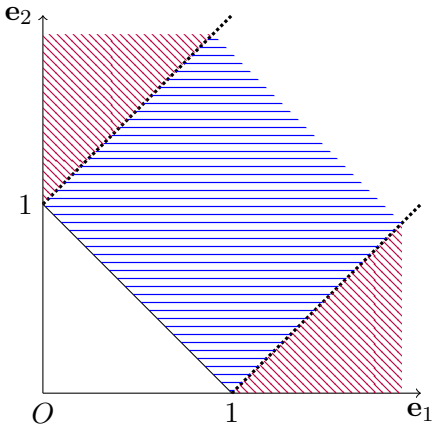
Duchi et al. (2008) provide an algorithm which computes $\pi_z(\mathbf{v})$ for $\mathbf{v} \in \mathbb{R}_+^d$ and $z > 0$. This algorithm is based on a median-search procedure whose expected time complexity is $O(d)$. We include this algorithm in Appendix A.

If $z = 1$ we shortly denote π for π_1. Note that the projection satisfies the relation $\pi_z(\mathbf{v}) = z\pi(\mathbf{v}/z)$ for all $\mathbf{v} \in \mathbb{R}_+^d$ and $z > 0$. This is why we mainly focus on the projection π onto the simplex $\mathbb{S}_+^{d-1}$. In this case we denote $\lambda_{\mathbf{v}} = \lambda_{\mathbf{v},1}$. An illustration of π in the bivariate case is given in Figure 2. It highlights the fundamental difference between π and the self-normalization on the subspaces C_β. For the latter, the subspaces $\{|\mathbf{x}|> 1,\, \mathbf{x}/|\mathbf{x}|\in C_\beta\} = R_\beta$ are of zero Lebesgue measure, as soon as $\beta \neq \{1,\dots,d\}$. On the contrary the subspaces $\{|\mathbf{x}|> 1,\, \pi(\mathbf{x}) \in C_\beta\}$ are of positive Lebesgue measure. In dimension 2 the purple shaded areas in Figure 2b correspond to the subspaces of $\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|> 1\}$ for which the vectors are projected on the axis $\mathbf{e}_1$ or $\mathbf{e}_2$. This topological difference between the sets R_β and $\pi^{-1}(C_\beta)$ entails that it is more likely that C_β is a $\pi(Y\mathbf{\Theta})$-continuity set than a $\mathbf{\Theta}$-one. Hence the projection π

allows to circumvent the issue that arise with the weak convergence in Equation (1.2) by substituting the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ by $\pi(\mathbf{X}/t)$. This encourages to introduce the following definition (see also Meyer and Wintenberger (2020)).



(a) Three vectors and their image by π. The dotted lines partition the space depending on the localization of the projected vectors: $\mathbf{e}_1$, $\mathbf{e}_2$, or the interior of the simplex.



(b) The preimages of the subsets C_β by π. In purple $\pi^{-1}(C_{\{1\}})$ and $\pi^{-1}(C_{\{12\}})$, and in blue $\pi^{-1}(C_{\{2\}})$.

Figure 2: Euclidean projection onto the simplex $\mathbb{S}_+^1$.

Definition 2 (Sparse regular variation)**.** *A random vector* $\mathbf{X}$ *on* $\mathbb{R}_+^d$ *is sparsely regularly varying if there exist a random vector* $\mathbf{Z}$ *defined on the simplex* $\mathbb{S}_+^{d-1}$ *and a non-degenerate random variable* Y *such that*

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \pi\left(\frac{\mathbf{X}}{t}\right)\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{d} \mathbb{P}((Y, \mathbf{Z}) \in \cdot), \quad t \to \infty\,. \tag{2.3}$$

In this case, there exists $\alpha > 0$ such that the random variable Y is Pareto(α)-distributed. The limit vector $\mathbf{Z}$ must be seen as the angular limit obtained after replacing the self-normalization by π. By continuity of π standard regular variation with tail index α and spectral vector $\mathbf{\Theta}$ implies sparse regular variation with tail index α and angular limit $\mathbf{Z} = \pi(Y\mathbf{\Theta})$.

In order to extend the concept of maximal direction for $\mathbf{Z}$ we adapt Definition 1.

Definition 3 (Maximal direction for $\mathbf{Z}$)**.** *Let* $\beta \in \mathcal{P}_d^*$*. We say that a direction* β *is* maximal *for* $\mathbf{Z}$ *if*

$$\mathbb{P}(\mathbf{Z} \in C_\beta) > 0 \quad \textit{and} \quad \mathbb{P}(\mathbf{Z} \in C_{\beta'}) = 0, \ \textit{for all } \beta' \supsetneq \beta\,. \tag{2.4}$$

In the rest of this section we discuss some theoretical results on the new angular limit vector $\mathbf{Z}$. The general idea is to prove that on the subsets C_β the behavior of $\mathbf{Z}$ is similar to the one of $\mathbf{\Theta}$ but that the estimation of $\mathbf{Z}$ on such sets can be done more easily. From now on we consider a regularly varying random vector $\mathbf{X}$ on $\mathbb{R}_+^d$ and set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. We gather here the results introduced by Meyer and Wintenberger (2020).

Proposition 1 (Meyer and Wintenberger (2020))**.** *Let* $\mathbf{X}$ *be a regularly varying random vector of* $\mathbb{R}_+^d$ *with spectral vector* $\mathbf{\Theta}$ *and tail index* $\alpha > 0$*. Set* $\mathbf{Z} = \pi(Y\mathbf{\Theta})$*, with* Y *a Pareto*(α) *random variable independent of* $\mathbf{\Theta}$*.*

1. *If A is a Borel subset of $\mathbb{S}_+^{d-1}$ satisfying $\mathbb{P}(Y\boldsymbol{\Theta} \in \partial\pi^{-1}(A)) = 0$, then*

$$\mathbb{P}(\pi(\mathbf{X}/t) \in A \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in A)\,, \quad t \to \infty\,. \tag{2.5}$$

In particular for any $\beta \in \mathcal{P}_d^*$, the following convergence holds:

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta)\,, \quad t \to \infty\,. \tag{2.6}$$

2. *For any $\beta \in \mathcal{P}_d^*$, if $\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0$, then $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$.*

3. *For any $\beta \in \mathcal{P}_d^*$, the direction β is maximal for $\boldsymbol{\Theta}$ if and only if it is maximal for $\mathbf{Z}$.*

Remark 1. *Note that the first convergence of Proposition 1 is a more general result than the one stated in Meyer and Wintenberger (2020) which only deals with the subsets C_β and another specific type of subsets but the generalization is straightforward.*

Let us briefly discuss the previous results. In order to capture the tail dependence structure of $\mathbf{X}$ we focus on the subsets C_β for $\beta \in \mathcal{P}_d^*$. We consider the projected vector $\pi(\mathbf{X}/t)$ instead of $\mathbf{X}/|\mathbf{X}|$ and then Equation (2.6) ensures that the angular component $\pi(\mathbf{X}/t)$ approximates well the limit vector $\mathbf{Z}$ on the subsets C_β. The behavior of $\mathbf{Z}$ on C_β has then to be related to the one of $\boldsymbol{\Theta}$ on the same subsets. This issue is addressed by the last two results which ensures that the distribution of $\mathbf{Z}$ places mass on C_β only if the distribution of $\boldsymbol{\Theta}$ does, and that the notion of maximal directions coincide for $\boldsymbol{\Theta}$ and $\mathbf{Z}$.

Therefore most of the work regarding the tail dependence of a regularly varying vector $\mathbf{X} \in \mathbb{R}_+^d$ can be done through the study of the behavior of the unit vector $\pi(\mathbf{X}/t) \mid |\mathbf{X}| > t$ on the subsets C_β. Our goal is thus to infer the support of the distribution of $\mathbf{Z}$ with the estimation of the probabilities $\mathbb{P}(\mathbf{Z} \in C_\beta)$ for $\beta \in \mathcal{P}_d^*$. Let us denote by $p(\beta)$ these $2^d - 1$ quantities. We are willing to decide which of these probabilities are positive. Similarly to the set $\mathcal{S}(\boldsymbol{\Theta})$ defined in (2.1) we define the set $\mathcal{S}(\mathbf{Z}) \subset \mathcal{P}_d^*$ as follows:

$$\mathcal{S}(\mathbf{Z}) := \{\beta \in \mathcal{P}_d^*,\, \mathbb{P}(\mathbf{Z} \in C_\beta) > 0\} = \{\beta \in \mathcal{P}_d^*,\, p(\beta) > 0\}\,, \tag{2.7}$$

and we denote by s^* its cardinality. In other words $\mathcal{S}(\mathbf{Z})$ gathers all directions β on which the angular vector $\mathbf{Z}$ puts mass. The aim of this paper is to build a statistical approach to decide which β's belong to $\mathcal{S}(\mathbf{Z})$. This method first relies on asymptotic results obtained for estimators of $p(\beta)$. The idea is then to use model selection to identify not only the most relevant features β but also an optimal threshold for which Equation (2.6) approximately holds.

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

$$S = \sum_{\beta\in\mathcal{P}_d^*} c(\beta)\, \delta_{|\beta|^{-1}\mathbf{e}(\beta)}\,, \tag{2.8}$$

where $(c(\beta))_\beta$ is a 2^d-1 vector with non-negative components adding up to 1 (see Segers (2012), Example 3.3). The distributions in (2.8) are stable after a multiplication by a positive random variable and a Euclidean projection onto the simplex. Hence, if $\boldsymbol{\Theta}$ follows a distribution of type (2.8), then $\mathbf{Z} = \boldsymbol{\Theta}$ a.s. This shows that (2.8) forms an accurate model for the angular vector $\mathbf{Z}$.

Example 2 (Asymptotic independence)**.** *If we choose $c(\beta) = 1/d$ for all β of length 1 in (2.8), then the spectral measure becomes*

$$S = d^{-1} \sum_{j=1}^{d} \delta_{\mathbf{e}_j}\,,$$

In this case we say that the spectral measure has the property of asymptotic independence *(see de Haan and Ferreira (2006), Section 6.2). Example 1 states that in this case* $\mathbf{Z} = \mathbf{\Theta}$ *a.s. This situation models data for which extreme events are likely to appear only because one coordinate is large. It has been widely studied since it provides models which are interpretable regarding extreme values and for which all calculations are feasible. There is an abundant literature on the subject, see e.g. Ledford and Tawn (1996), Heffernan and Tawn (2004), or Fougères and Soulier (2010).*

2.3 Model selection

The identification of the most relevant subsets C_β in $\mathcal{S}(\mathbf{Z})$ is done with model selection (Massart (2007)). In our context it boils down to choosing the theoretical distribution which best fits a given sample. A standard way to measure the distance between two distribution is given by the Kullback-Leibler divergence (Kullback and Leibler (1951)). For two distributions P and Q with respective density p and q with respect to a measure ν this divergence is given by

$$K(p,q) = \int p \log\left(\frac{p}{q}\right) \mathrm{d}\nu\,.$$

In this paper we will use this distance to compare the empirical repartition of the mass of $\mathbf{Z}$ on the $2^d - 1$ subsets C_β with the theoretical one. This means that we are willing to compare Categorical distributions. Thus, the Kullback-Leibler divergence is computed with respect to the counting measure and can be written as follows:

$$K(P,Q) = \mathbb{E}_{\xi\sim P}\Big[\log\Big(\frac{p(\xi)}{q(\xi)}\Big)\Big] = \sum_{x\in\Xi} x\log(p(x)/q(x))\,,$$

where Ξ denotes the support of ξ. The distribution P often describes the unknown underlying distribution of the data while Q is a distribution taken from a given family of models. Selection model then boils down to identifying which distribution Q is the closest to P, i.e. to choosing the distribution Q which minimizes the divergence $K(P,Q)$. The most common selection procedure is the one based on the Akaike Information Criterion (AIC, Akaike (1973)) which consists in comparing log-likelihoods and adding a penalization which constraints the number of parameters (in our context the number of relevant subsets C_β). This approach is developed in Section 5.

The model selection is conducted in a non-asymptotic setting which means that the convergence in Equation (2.6) is considered as an approximation for t large enough. The choice of this threshold is deeply related to the choice of the sphere on which the Euclidean projection is applied. For a vector $\mathbf{v} \in \mathbb{R}_+^d$ with ℓ^1-norm $|\mathbf{v}|$, the number of null coordinates of the projected vector $\pi_t(\mathbf{v})$ strongly depends on the choice of t. If t is close to $|\mathbf{v}|$ then $\pi_t(\mathbf{v})$ has almost only non-null coordinates (as soon as $\mathbf{v}$ itself has non-null coordinates). On the contrary, if $t \ll |\mathbf{v}|$ then the vector $\pi_t(\mathbf{v})$ becomes sparser. The impact of the threshold t on the sparsity of the projected vectors is illustrated in Figure 3. In a statistical context we want to study the tail behavior of n iid regularly varying random vectors. There, focusing on extreme values boils down to selecting only the vectors with the largest norms, that is vectors whose norm is above a given threshold. For a large threshold t, only extreme data are selected but many vectors are close to this threshold. This implies that these vectors are projected on subsets C_β with large $|\beta|$'s. Hence the projected vectors are not very sparse. On the other hand if we select a low threshold then we move away from the extreme regime. In this case the largest vectors are projected on subsets C_β with small $|\beta|$'s, i.e. the projected vectors are very sparse. Thus we have to make a balanced choice between providing a sparse structure for the data and staying in the extreme regime.

Remark 2. *Consider a sample* $\mathbf{X}_1, \dots, \mathbf{X}_n$ *of iid regularly varying random vectors with tail index* $\alpha > 0$*. This index models the behavior of* $|\mathbf{X}_1|, \dots, |\mathbf{X}_n|$*. If* α *is small, then distance between two extreme norms*

is large, which makes the choice of the threshold easier. Besides, the strict inequality in Equation (2.3) allows to obtain sparse projections for all large vectors. Indeed, it is customary in EVT to consider the k-th largest norms $|\mathbf{X}|_{(1)} \geq \ldots \geq |\mathbf{X}|_{(k)} \geq \ldots \geq |\mathbf{X}|_{(n)}$ *which means to choose a threshold* $t = |\mathbf{X}|_{(k+1)}$. *With a strict inequality the smallest vector* $|\mathbf{X}|_{(k)}$ *may be projected on a low-dimensional subspace which would not have been the case with a large inequality and a threshold* $t = |\mathbf{X}|_{(k)}$. *However, this strict inequality does not influence the selection of the level k.*

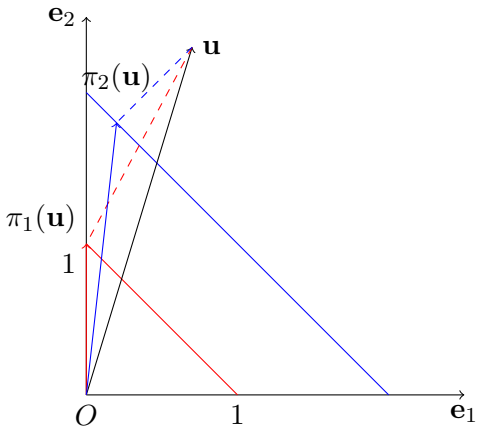


Figure 3: Consequence of the choice of the threshold on the sparsity. The image of the vector $\mathbf{u}$ is $\pi_1(\mathbf{u}) = (0, 1)$ with the threshold $z = 1$ while it is $\pi_2(\mathbf{u}) > 0$ with the threshold $z = 2$. The sparsity increases when the threshold decreases.

This entails that the selection of the most relevant C_β and the choice of an optimal threshold t in Equation (2.6) have to be done simultaneously. The idea is to extend the model selection to highlight which threshold provides the best estimation. For practical reasons it is often more convenient to focus on the number of exceedances k rather on the threshold z. Then the selection consists in choosing the appropriate number of extreme data. However, the AIC procedure relies on the minimization of a penalized maximum likelihood and holds for a constant sample size. Therefore the method has to be adapted in order to use it in an extreme setting. It is indeed necessary to include the non-extreme values in the models and to separate them into an *extreme* group and a *non-extreme* one. The idea is then to apply an AIC type procedure which highlights the k for which the separation is optimal. This method does not provided a generic choice of k for all data sets but suggests an ad hoc selection based on a penalized maximum likelihood minimization. This is the purpose of Section 5.

3 Asymptotic results

3.1 Statistical framework

From now on we consider a sequence of iid regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index α and spectral vector $\mathbf{\Theta}$. We also consider a Pareto(α)-distributed random variable Y independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. With these notations we have the following weak convergence:

$$\mathbb{P}\left(\pi(\mathbf{X}/t) \in \cdot \mid |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}(\mathbf{Z} \in \cdot)\,, \quad t \to \infty\,. \tag{3.1}$$

Our aim is to infer the behavior of the angular vector $\mathbf{Z}$ with the sample $(\mathbf{X}_n)_{n\in\mathbb{N}}$ by focusing on the probabilities $p(\beta) = \mathbb{P}(\mathbf{Z} \in C_\beta)$ for $\beta \in \mathcal{P}_d^*$ since they emphasize the extremal joint behavior of the components of $\mathbf{X}$ (see Subsection 2.2). Our aim is to classify which ones belong to $\mathcal{S}(\mathbf{Z})$ and

which ones do not. A standard assumption to provide a statistical approach in EVT is to consider a positive sequence $(u_n)_{n\in\mathbb{N}}$, $u_n \to \infty$, which plays the role of the threshold t in Equation (3.1) (see Beirlant et al. (2006), Section 9.4.1). This means that for $n \in \mathbb{N}$, the quantity u_n must be seen as the threshold above which the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are extreme. It is customary in EVT to define a level $k = k_n = n\mathbb{P}(|\mathbf{X}| > u_n)$ and to assume that $k_n \to \infty$ when $n \to \infty$. Note that the assumption $u_n \to \infty$ implies that $k_n/n = \mathbb{P}(|\mathbf{X}| > u_n) \to 0$ which means that k_n tends to infinity at a slower rate than n.

In order to identify the set $\mathcal{S}(\mathbf{Z})$ defined in (2.7), we need to provide suitable estimators for the quantities $p(\beta) = \mathbb{P}(\mathbf{Z} \in C_\beta)$. Following Equation (2.6), we know that the aforementioned probabilities appear as the limits of the pre-asymptotic quantities $\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$. Therefore, for a Borel subset A of $\mathbb{S}_+^{d-1}$, we set

$$\begin{aligned} p(A) &= \mathbb{P}(\mathbf{Z} \in A)\,, \\ p_n(A) &= \mathbb{P}(\pi(\mathbf{X}/u_n) \in A \mid |\mathbf{X}| > u_n)\,, \\ T_n(A) &= \sum_{j=1}^{n} \mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n\}\,. \end{aligned}$$

The random variable $T_n(A)$ corresponds to the data which are projected in the subset A among the extreme ones. For the sake of simplicity we keep our previous notations regarding C_β and write $p(\beta) = p(C_\beta)$, and similarly $p_n(\beta) = p_n(C_\beta)$ and $T_n(\beta) = T_n(C_\beta)$. The expectation of $T_n(A)$ is equal to $\mathbb{E}[T_n(A)] = k_n p_n(A)$. Besides, the first point of Proposition 1 implies that under a regularity assumption on A the probability $p_n(A)$ converges to $p(A)$. This encourages to estimate the probability $p(A)$ through a classical bias-variance decomposition:

$$\frac{T_n(A)}{k_n} - p(A) = \left[\frac{T_n(A)}{k_n} - p_n(A)\right] + [p_n(A) - p(A)]\,. \tag{3.2}$$

The first term is addressed in the following section. For the second one, Equation (2.5) ensures that it vanishes at infinity. However it is common to assume a stronger condition like $\sqrt{k_n}(p_n(A) - p(A)) \to 0$ as $n \to \infty$. This will be discussed more in detail in what follows.

Regarding the subsets C_β, the decomposition in Equation (3.2) highlights the fact that the study of $p(\beta)$ will be conducted through the analysis of the pre-asymptotic probabilities $p_n(\beta)$. In particular, we would like to define a similar set as $\mathcal{S}(\mathbf{Z})$ but for $p_n(\beta)$. In order to avoid that such a set depends on n, we replace the natural condition $p_n(\beta) > 0$ for all $n \geq 1$ by the stronger one $k_n p_n(\beta) \to \infty$. This leads to the following subset of features:

$$\mathcal{R}_k(\mathbf{Z}) = \{\beta \in \mathcal{P}_d^*,\, k_n p_n(\beta) \to \infty \text{ when } n \to \infty\}\,. \tag{3.3}$$

We denote by r^* the cardinality of $\mathcal{R}_k(\mathbf{Z})$. This definition implies two straightforward consequences. The first one is that for all feature $\beta \in \mathcal{R}_k(\mathbf{Z})$, the probability $p_n(\beta)$ is positive for n large enough. Second, we have the following inclusion: $\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z})$. In particular, the cardinalities of these sets satisfy the inequality $s^* \leq r^*$.

Outline of the statistical study The rest of this section is devoted to asymptotic results for the estimator $T_n(A)$ under some assumptions on the Borel set $A \subset \mathbb{S}_+^{d-1}$. A law of large numbers ensures the convergence in probability of $T_n(A)/k_n$ to $p(A)$ as n increases (Proposition 2). Then, a central limit theorem is established in order to exhibit a limit distribution for the previous estimators (Theorem 1). This latter convergence holds under an assumption of convergence of the bias $p_n(A) - p(A)$ and brings out a rate of convergence of order $\sqrt{k_n}$. Regarding the features β, the purpose of Section 4 is then to extend these results by considering all subsets C_β simultaneously.

3.2 A univariate approach

We start our statistical study with a proposition which establishes the consistency of the estimator $T_n(A)$ for a Borel subset $A \subset \mathbb{S}_+^{d-1}$. This result is proved for both pre-asymptotic probability $p_n(A)$ and asymptotic one $p(A)$. For this latter a regularity assumption on the Borel set A is needed.

Proposition 2. *We consider a sequence of iid regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index α and spectral vector $\mathbf{\Theta}$, a Pareto(α)-distributed random variable Y independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. We consider a threshold $u_n \to \infty$ and assume that $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$.*

1. *For all Borel set $A \subset \mathbb{S}_+^{d-1}$, the following convergence in probability holds:*

$$\frac{T_n(A)}{k_n} - p_n(A) \to 0\,, \quad n \to \infty\,. \tag{3.4}$$

2. *For all Borel set $A \subset \mathbb{S}_+^{d-1}$ such that $\mathbb{P}(Y\mathbf{\Theta} \in \partial\pi^{-1}(A)) = 0$ the following convergence in probability holds:*

$$\frac{T_n(A)}{k_n} \to p(A)\,, \quad n \to \infty\,. \tag{3.5}$$

3. *For all Borel set $A \subset \mathbb{S}_+^{d-1}$ such that $k_n p_n(A) \to \infty$, the following convergence in probability holds:*

$$\frac{T_n(A)}{k_n p_n(A)} \to 1\,, \quad n \to \infty\,. \tag{3.6}$$

The convergence in Equation (3.5) holds in particular for $A = C_\beta$ and ensures that the estimation of $p(\beta)$ can be done with $T_n(\beta)/k_n$. Besides the convergence (3.6) holds for all C_β such that $\beta \in \mathcal{R}_k(\mathbf{Z})$. Proposition 2 implies that if $p(\beta) = 0$, i.e. if $\mathbf{Z}$ does not place mass on the subset C_β, then $T_n(\beta)/k_n$ becomes smaller and smaller as n increases. Actually as soon as the dimension d is large a lot of $T_n(\beta)$'s are even equal to 0 since the level k_n is far below the number of subsets $2^d - 1$.

We now establish asymptotic normality for $T_n(A)$.

Theorem 1. *We consider a sequence of iid regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index α and spectral vector $\mathbf{\Theta}$, a Pareto(α)-distributed random variable Y independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. We consider a threshold $u_n \to \infty$ and assume that $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$. Finally, we fix a Borel set $A \in \mathbb{S}_+^{d-1}$.*

1. *If $k_n p_n(A) \to \infty$, then*

$$\sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p_n(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,. \tag{3.7}$$

2. *If $\mathbb{P}(Y\mathbf{\Theta} \in \partial\pi^{-1}(A)) = 0$ and $p(A) > 0$, then*

$$\sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,. \tag{3.8}$$

3. *If $\mathbb{P}(Y\mathbf{\Theta} \in \partial\pi^{-1}(A)) = 0$, $p(A) > 0$, and if we assume that*

$$\sqrt{k_n}(p_n(A) - p(A)) \to 0\,, \quad n \to \infty\,, \tag{3.9}$$

then

$$\sqrt{k_n}\frac{T_n(A)/k_n - p(A)}{\sqrt{p(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,. \tag{3.10}$$

Theorem 1 ensures that $T_n(A)/k_n$ is asymptotically normal as soon as $k_n p_n(A) \to \infty$. This assumption implies that $p_n(A)$ is positive for n large enough. It is for instance true as soon as $p(A) > 0$. From convergence (3.7) to convergence (3.8), the denominator $\sqrt{p_n(A)}$ has been replaced by $\sqrt{p(A)}$. This requires that $p_n(A) \to p(A) > 0$ which justifies the regularity assumption. Regarding the subsets C_β, the convergences (3.8) and (3.10) only hold for the features β such that $p(\beta) > 0$ i.e. for $\beta \in \mathcal{S}(\mathbf{Z})$. On the contrary the convergence (3.7) holds for the features β such that $k_n p_n(\beta) \to \infty$ i.e. for $\beta \in \mathcal{R}_k(\mathbf{Z})$.

The results obtained in Proposition 2 and in Theorem 1 highlight the asymptotic behavior of $T_n(A)$ when $A \subset \mathbb{S}_+^{d-1}$ is a fixed Borel set. Regarding the subsets C_β these results allow the features β to be studied individually. The next step is to establish results which address the question of the joint estimation of the probabilities $p(\beta)$.

Remark 3. *If we consider r features $\beta_1, \dots, \beta_r$ and if we assume that $k_n p_n(C_{\beta_j}) \to \infty$ for all $j = 1, \dots, r$, then we obtain that $k_n p_n(\cup_j C_{\beta_j}) \to \infty$. Thus, Theorem 1 yields to the following convergence*

$$\sqrt{k_n} \frac{T_n(\cup_j C_{\beta_j})/k_n - p_n(\cup_j C_{\beta_j})}{\sqrt{p_n(\cup_j C_{\beta_j})}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,,$$

which can be rephrased as follows

$$\sqrt{k_n} \frac{\sum_{j=1}^r T_n(C_{\beta_j})/k_n - \sum_{j=1}^r p_n(C_{\beta_j})}{\sqrt{\sum_{j=1}^r p_n(C_{\beta_j})}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,,$$

This convergence will be useful in Section 5.

4 Multivariate results

Moving on to a multivariate setting we now only focus on the subsets $A = C_\beta$, $\beta \in \mathcal{P}_d^*$, for which we study the behavior of $\mathbf{X}$ and $\mathbf{Z}$. Our main goal is identify the set of directions $\mathcal{S}(\mathbf{Z})$ defined in (2.7), i.e. to distinguish the positive probabilities $p(\beta)$ from the null ones. Such a distinction is done via the study of the estimators $T_n(A)$. Hence we consider the set

$$\widehat{\mathcal{S}}_n(\mathbf{Z}) = \{\beta \in \mathcal{P}_d^*,\, T_n(\beta) > 0\}\,, \tag{4.1}$$

and denote by $\hat{s}_n$ its cardinality.

4.1 Estimation of the set $\mathcal{S}(\mathbf{Z})$

Some properties of the empirical set $\widehat{\mathcal{S}}_n(\mathbf{Z})$ are developed here. First, for all $\beta \in \mathcal{S}(\mathbf{Z})$ the convergence $p_n(\beta) \to p(\beta) > 0$ implies that $p_n(\beta)$ is positive for n large enough. The corresponding result for $T_n(\beta)$ is a consequence of the following lemma.

Lemma 1. *For $\beta \in \mathcal{P}_d^*$, we have the following relation*

$$\log(\mathbb{P}(T_n(\beta) = 0)) \sim -k_n p_n(\beta)\,, \quad n \to \infty\,.$$

If $\beta \in \mathcal{R}_k(\mathbf{Z})$, then $-k_n p_n(\beta) \to -\infty$ and thus Lemma 1 implies that $\mathbb{P}(T_n(\beta) = 0) \to 0$ when $n \to \infty$. This proves that for all $\beta \in \mathcal{R}_k(\mathbf{Z})$ the estimator $T_n(\beta)$ is positive with probability converging to 1. In particular this remark holds for all $\beta \in \mathcal{S}(\mathbf{Z})$. In this case it means that if the vector $\mathbf{Z}$ places some mass in the direction β then at least one extreme observation appears in this direction.

A consequence of Lemma 1 is that

$$\mathbb{P}(\mathcal{R}_k(\mathbf{Z}) \subset \widehat{\mathcal{S}}_n(\mathbf{Z})) = 1 - \mathbb{P}(\exists \beta \in \mathcal{R}_k(\mathbf{Z}),\, \beta \notin \widehat{\mathcal{S}}_n(\mathbf{Z})) \geq 1 - \sum_{\beta \in \mathcal{R}_k(\mathbf{Z})} \mathbb{P}(T_n(\beta) = 0) \to 1\,,$$

when $n \to \infty$. Consequently since $\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z})$ we have the convergence

$$\mathbb{P}(\mathcal{S}(\mathbf{Z}) \subset \widehat{\mathcal{S}}_n(\mathbf{Z})) \to 1\,, \quad n \to \infty\,.$$

It is not so easy to obtain a converse inclusion between $\mathcal{R}_k(\mathbf{Z})$ and $\widehat{\mathcal{S}}_n(\mathbf{Z})$. However, if $\beta \notin \mathcal{S}(\mathbf{Z})$ then $p(\beta) = 0$. In this case Lemma 1 ensures that $\mathbb{P}(T_n(\beta) = 0) \to 1$ if and only if $k_n p_n(\beta) \to 0$. If this latter convergence holds for all $\beta \in \mathcal{S}(\mathbf{Z})^c$, then we obtain that

$$\mathbb{P}(\mathcal{S}(\mathbf{Z})^c \subset \widehat{\mathcal{S}}_n(\mathbf{Z})^c) = 1 - \mathbb{P}(\exists \beta \in \mathcal{S}(\mathbf{Z})^c,\, \beta \in \widehat{\mathcal{S}}_n(\mathbf{Z})) \geq 1 - \sum_{\beta \in \mathcal{S}(\mathbf{Z})^c} \mathbb{P}(T_n(\beta) > 0) \to 1\,,$$

when $n \to \infty$. We gather these results in the following proposition.

Proposition 3.

1. *With probability converging to 1, we have the inclusions*

$$\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z}) \subset \widehat{\mathcal{S}}_n(\mathbf{Z})\,.$$

2. *If for all $\beta \in \mathcal{S}(\mathbf{Z})^c$,*

$$k_n p_n(\beta) \to 0\,, \quad n \to \infty\,, \tag{4.2}$$

then with probability converging to 1, $\widehat{\mathcal{S}}_n(\mathbf{Z}) \subset \mathcal{S}(\mathbf{Z})$.

A consequence of the first point of Proposition 3 is that the cardinality of the sets $\mathcal{S}(\mathbf{Z})$, $\mathcal{R}_k(\mathbf{Z})$, and $\widehat{\mathcal{S}}_n(\mathbf{Z})$ are satisfying the inequality $s^* \leq r^* \leq \hat{s}_n$ with probability converging to 1. Regarding the second point, the assumption in (4.2) is quite strong compared to the one given in Equation (3.9) and implies in particular that $\mathcal{S}(\mathbf{Z}) = \mathcal{R}_k(\mathbf{Z})$. The numerical results introduced in Section 6 show that assumption (4.2) is not satisfied on simulated data. This is why unless stated otherwise we do not assume that 4.2 holds.

At this point the statistical setting is the following one. For a fixed n large enough we have a collection of directions β such that $T_n(\beta) > 0$ and the following inclusions are satisfied:

$$\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z}) \subset \{\beta \in \mathcal{P}_d^*,\, p_n(\beta) > 0\}\,, \tag{4.3}$$

and Equation (3.4) in Proposition 2 entails that

$$\mathbb{P}\Big(\{\beta \in \mathcal{P}_d^*,\, p_n(\beta) > 0\} \subset \widehat{\mathcal{S}}_n(\mathbf{Z})\Big) \to 1\,, \quad n \to \infty\,.$$

These inclusions highlight the fact that the observations tend to overestimate the number of directions β in $\mathcal{S}(\mathbf{Z})$. Therefore the goal is to the classiffy the ones that are indeed in $\mathcal{S}(\mathbf{Z})$ from the ones that are not. This means that we need to build a statistical method which brings out a cutoff dividing the empirical set $\widehat{\mathcal{S}}_n(\mathbf{Z})$ into two subsets: a first one corresponding to the directions β which belongs to $\mathcal{S}(\mathbf{Z})$ and a second one which contains the directions β which appear because of a possible bias between the probabilities arising from the non-asymptotic sample and the ones representing the theoretical asymptotic framework. In this context it is customary to use model selection (see Section 5). This needs to define an order on the different subsets C_β.

4.2 Ordering the β's

In order to study the common behavior of the components $T_n(\beta)$ we need to glue them together and to build a vector in $\mathbb{R}^{2^d-1}$. This can not be easily addressed since there is no specific order on $\mathcal{P}_d^*$. Therefore we need to fix an order between the β's, i.e. to define a bijection

$$\sigma : \{1, \dots, 2^d - 1\} \to \mathcal{P}_d^* \,.$$

The idea is to choose an order σ which takes into account the values of $p(\beta)$. However, such an order can only be introduced for $\beta \in \mathcal{S}(\mathbf{Z})$ since the other ones have all null probabilities. Therefore the bijection σ is defined in two steps. First, we consider the $s^* = |\mathcal{S}(\mathbf{Z})|$ first values. In order to define them without any ambiguity, we make the following assumption on $\mathbf{p}$.

Assumption 1. *For all $\beta, \beta' \in \mathcal{S}(\mathbf{Z})$, $p(\beta) \neq p(\beta')$.*

Under Assumption 1, we define $\sigma(j)$ for $j = 1, \dots, s^*$ by considering the probabilities in $\mathcal{S}(\mathbf{Z})$ in the decreasing order, that is

$$\begin{aligned}
\sigma(1) &= \operatorname*{arg\,max}_{\beta \in \mathcal{P}_d^*} p(\beta) = \operatorname*{arg\,max}_{\beta \in \mathcal{S}(\mathbf{Z})} p(\beta) \,, \\
\sigma(2) &= \operatorname*{arg\,max}_{\beta \in \mathcal{P}_d^* \setminus \sigma(1)} p(\beta) = \operatorname*{arg\,max}_{\beta \in \mathcal{S}(\mathbf{Z}) \setminus \sigma(1)} p(\beta) \,, \\
&\vdots \\
\sigma(s^*) &= \operatorname*{arg\,max}_{\beta \in \mathcal{P}_d^* \setminus \{\sigma(1), \dots, \sigma(s^*-1)\}} p(\beta) = \operatorname*{arg\,max}_{\beta \in \mathcal{S}(\mathbf{Z}) \setminus \{\sigma(1), \dots, \sigma(s^*-1)\}} p(\beta) \,.
\end{aligned} \tag{4.4}$$

Then we need to define $\sigma(j)$ for $j = s^* + 1, \dots, 2^d - 1$. Since the remaining values of $p(\beta)$ are null, no natural order appears here. Therefore, we fix an arbitrary order once and for all, i.e. we define distinct images $\sigma(j) \in \mathcal{P}_d^* \setminus \{\sigma(1), \dots, \sigma(s^*)\}$ for all $j = s^* + 1, \dots, 2^d - 1$. This order being now fixed for the rest of the article, all vectors of $\mathbb{R}^{2^d-1}$ whose components are indexed by $\mathcal{P}_d^*$ will be written based on this order. Moreover, we simplify the notations by setting $\beta_j = \sigma(j)$ for all $j = 1, \dots, 2^d - 1$.

With these considerations we define the vector $\mathbf{p} \in \mathbb{R}^{2^d-1}$ whose components are associated to the order defined in (4.4), i.e. $p_j = p(\beta_j) = p(\sigma(j))$. By construction the vector $\mathbf{p}$ satisfies

$$p_1 = p(\beta_1) > \ldots > p_{s^*} = p(\beta_{s^*}) > p_{s^*+1} = \ldots = p_{2^d-1} = 0 \,.$$

In particular we have the relation $\mathbf{p}_{\mathcal{S}(\mathbf{Z})} = \mathbf{p}_{\{1,\dots,s^*\}}$. We use the same order to define the vectors $\mathbf{T}_n$ and $\mathbf{p}_n$, whose components are given by

$$T_{n,\,j} = T_n(\beta_j) \quad \text{and} \quad p_{n,\,j} := p_n(\beta_j), \quad j = 1, \dots, 2^d - 1 \,. \tag{4.5}$$

Contrary to the components of the vector $\mathbf{p}$, the ones of the vectors $\mathbf{T}_n$ and $\mathbf{p}_n$ are not necessary ordered in a decreasing order. However this is asymptotically true, as stated in the following section.

4.3 Multivariate convergences

We discuss some convergence results for the random vector $\mathbf{T}_n$. With the order defined below, the components $T_{n,\,j}$, $p_{n,\,j}$, and p_j of the vectors $\mathbf{T}_n$, $\mathbf{p}_n$, and $\mathbf{p}$ are all the three associated to the same subset C_β. Therefore, the consistency of $\mathbf{T}_n$ is a straightforward extension of Proposition 2:

$$\frac{\mathbf{T}_n}{k_n} - \mathbf{p}_n \to 0 \,, \quad n \to \infty \,, \quad \text{in probability} \,, \tag{4.6}$$

and

$$\frac{\mathbf{T}_n}{k_n} \to \mathbf{p}\,, \quad n \to \infty\,, \quad \text{in probability}\,. \tag{4.7}$$

For $r \geq 1$ we consider the subset $\mathrm{Ord}_r = \{\mathbf{x} \in \mathbb{R}^r, x_1 \geq \ldots \geq x_r\}$ whose boundary is given by $\partial\mathrm{Ord}_r = \{\mathbf{x} \in \mathbb{R}^r, \exists j \neq k, x_j = x_k\}$. On the one hand the definition of the vector $\mathbf{p}$ ensures that $\mathbf{p} \in \mathrm{Ord}_{2^d-1}$. On the other hand Assumption 1 ensures that $\mathbf{p}_{\{1,\ldots,s^*\}} \notin \partial\mathrm{Ord}_{s^*}$. Since $\mathbf{T}_n/k_n \to \mathbf{p}$ in probability, it follows from the Portmanteau theorem that

$$\mathbb{P}(\mathbf{T}_{n,\{1,\ldots,s^*\}} \in \mathrm{Ord}_{s^*}) = \mathbb{P}(k_n^{-1}\mathbf{T}_{n,\{1,\ldots,s^*\}} \in \mathrm{Ord}_{s^*}) \to \mathbb{P}(\mathbf{p}_{\{1,\ldots,s^*\}} \in \mathrm{Ord}_{s^*}) = 1\,, \quad n \to \infty\,.$$

Hence, for $\delta > 0$ there exists n_0 such that for all $n \geq n_0$

$$\mathbb{P}(\mathbf{T}_{n,\{1,\ldots,s^*\}} \in \mathrm{Ord}_{s^*}) \geq 1 - \delta\,. \tag{4.8}$$

In other words, if n is large then the vector $\mathbf{T}_{n,\{1,\ldots,s^*\}}$ have its components ordered in the decreasing order with high probability.

In order to apply a model selection we need to obtain an asymptotic distribution for the vector $\mathbf{T}_n$. The idea is to extend the results obtained in Theorem 1. Recall that the convergence (3.7) in Theorem 1 holds only for subsets A such that $k_n p_n(A) \to \infty$. Therefore in order to obtain a multivariate convergence for the subsets C_β it is necessary to restrict ourselves to the directions $\beta \in \mathcal{R}_k(\mathbf{Z})$ where $\mathcal{R}_k(\mathbf{Z})$ is defined in (3.3). Consequently the restricted vectors $\mathbf{p}_{\mathcal{R}_k(\mathbf{Z})}$, $\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}$, and $\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}$ of $\mathbb{R}^{r^*}$ are considered.

Theorem 2. *We consider a sequence of iid regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index α and spectral vector $\mathbf{\Theta}$, a Pareto(α)-distributed random variable Y independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. We consider a threshold $u_n \to \infty$ and assume that $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$.*

1. *The following weak convergence on $\mathcal{R}_k(\mathbf{Z})$ holds:*

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{r^*})\,, \quad n \to \infty\,. \tag{4.9}$$

2. *The following weak convergence on $\mathcal{S}(\mathbf{Z})$ holds:*

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{S}(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{S}(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{s^*})\,, \quad n \to \infty\,. \tag{4.10}$$

3. *Moreover, if we assume that*

$$\forall \beta \in \mathcal{S}(\mathbf{Z})\,, \quad \sqrt{k_n}(p_n(\beta) - p(\beta)) \to 0\,, \quad n \to \infty\,, \tag{4.11}$$

then the following weak convergence on $\mathcal{S}(\mathbf{Z})$ holds:

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{S}(\mathbf{Z})}}{k_n} - \mathbf{p}_{\mathcal{S}(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{s^*})\,, \quad n \to \infty\,. \tag{4.12}$$

The multivariate convergence in (4.9) (respectively in (4.10) and in (4.12)) is the extension of the univariate convergence in (3.7) (respectively in (3.8) and in (3.10)). Similarly the bias assumption in (4.11) corresponds to the assumption in (3.9).

From Equation (4.9) we obtain that the vector

$$\mathbf{U}_n = \sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right)$$

satisfies the convergence

$$\mathbf{U}_n^\top \cdot \mathbf{U}_n = k_n \left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})} \right)^\top \mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1} \left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})} \right) \xrightarrow{d} \chi^2(r^*)\,, \quad (4.13)$$

when $n \to \infty$ and where $\chi^2(r^*)$ denotes a chi-squared distribution with r^* degrees of freedom. This convergence can be rephrased as follows:

$$k_n \sum_{j=1}^{r^*} \frac{(\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}/k_n - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^2}{\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}} \xrightarrow{d} \chi^2(r^*)\,, \quad n \to \infty\,. \quad (4.14)$$

Remark 4. *If we fix $s < r^*$, then the subsets $C_{\beta_{s+1}}, \dots, C_{\beta_{r^*}}$ satisfy the property*

$$\sum_{j=s+1}^{r^*} T_{n,j} = T_n(\cup_{j=s+1}^{r^*} C_{\beta_j})\,,$$

see Remark 3. Hence, the vector

$$\mathbf{U}_n(s) = \sqrt{k_n} \left(\frac{T_{n,1}/k_n - p_{n,1}}{\sqrt{p_{n,j}}}, \dots, \frac{T_{n,s}/k_n - p_{n,s}}{\sqrt{p_{n,s}}}, \frac{\sum_{j=s+1}^{r^*} (T_{n,s}/k_n - p_{n,s})}{\sum_{j=s+1}^{r^*} \sqrt{p_{n,s}}} \right)^\top$$

converges in distribution to a random vector of $\mathbb{R}^{s+1}$ with distribution $\mathcal{N}(0, Id_{s+1})$. Then, similarly to Equation (4.13), we have the convergence

$$\mathbf{U}_n(s)^\top \cdot \mathbf{U}_n(s) \xrightarrow{d} \chi^2(s+1)\,, \quad n \to \infty\,. \quad (4.15)$$

This convergence will be central in order to provide a suitable procedure for model selection, the key point of the method being to identify s^*.

5 Model selection

Based on the asymptotic results established in Section 4, we provide a model selection procedure which addresses two issues. The first one concerns the identification of the set $\mathcal{S}(\mathbf{Z})$ and the second one is the choice of an optimal level k_n. As already mentioned in Section 2.3 the choice of the threshold u_n has a direct impact on the result given by the Euclidean projection onto the simplex. Therefore the identification of $\mathcal{S}(\mathbf{Z})$ and the selection of an optimal level k_n are issues deeply related and which should be addressed simultaneously.

5.1 A multinomial model

In all what follows we assume that n is large enough. Thus it seems natural to assume that the inclusions in Proposition 3 holds, i.e. that there exists no direction β in $\mathcal{S}(\mathbf{Z})$ which does not belong to $\widehat{\mathcal{S}}_n(\mathbf{Z})$. Reciprocally, Assumption 4.2 may not hold which means that some observations could appear in a direction β on which the distribution of $\mathbf{Z}$ does not place mass. In this case it seems reasonable to assume that the quantity $T_n(\beta)$ associated to this direction is not very large. Since all the work is now done at a non-asymptotic level the strict inclusion mainly arises because of a possible bias which appears on the observations. All in all the sequence of inclusions in (4.3) implies that we make the following assumptions on the observations.

– If a feature β does not appear in $\widehat{\mathcal{S}}_n(\mathbf{Z})$, then we conclude that the distribution of $\mathbf{Z}$ does not place mass in this direction.

– If a feature β satisfies $T_n(\beta) \gg 0$, then we infer that $\mathbf{Z}$ concentrates on the associated subset C_β .

– If a feature β satisfies $T_n(\beta) \approx 0$, then it is likely that this direction appears in $\widehat{\mathcal{S}}_n(\mathbf{Z})$ only because of the bias which arises due to the non-asymptotic setting. There, we assume that the distribution of $\mathbf{Z}$ does not put mass in this direction.

The core of the study is now to provide a suitable procedure which classifies the directions β which appear in the last two cases. To this end we use Akaike Information Criterion in order to highlight the number of relevant directions β on which extreme values appear. Our goal is to identify which probabilistic model fits the best the data. We first work with a fixed level $k = k_n$ before dealing with the choice of this parameter. Recall from Equation (4.8) that for n large enough the probability $\mathbb{P}(T_{n,1} \geq T_{n,2} \geq \ldots \geq T_{n,s^*})$ is close to 1. From now on we fix a n large enough and work conditionally on this event. We also fix a level $k = k_n$. The random variables $T_{n,j}$ add up to k which encourages to introduce a multinomial model with $2^d - 1$ outcomes. We consider a particular multinomial model denoted by $\mathbf{M}(k; \tilde{\mathbf{p}})$, where the parameter $\tilde{\mathbf{p}}$ is defined as

$$\tilde{\mathbf{p}} = (\overbrace{\tilde{p}_1, \ldots, \tilde{p}_s, \underbrace{\tilde{p}, \ldots, \tilde{p}}_{r-s}, 0, \ldots, 0}^{2^d-1 \text{ components}}) ,$$

with $\tilde{p}_1 \geq \ldots \geq \tilde{p}_s, \tilde{p} \in (0, 1)$ satisfying the constraint

$$\tilde{p}_1 + \ldots + \tilde{p}_s + (r - s)\tilde{p} = 1 . \tag{5.1}$$

Such a model is entirely characterized by the parameters $\tilde{p}_1, \ldots, \tilde{p}_s, \tilde{p}$ and r. The model $\mathbf{M}(k; \tilde{\mathbf{p}})$ highlights the s relevant directions β which gather the mass of the distribution of $\mathbf{Z}$. The parameter $\tilde{p}$ models the bias and should therefore be considered as small and converging to zero when n increases. It underlines the idea that among the directions β which contain at least one observation some of them indeed belong to the support of $\mathbf{Z}$ while others only appear because of a bias. The first s directions correspond to the most relevant ones: It is likely to observe extreme events in the associated directions.

Estimation of the parameters The support of the distribution $\mathbf{M}(k; \tilde{\mathbf{p}})$ corresponds to the set of all $\mathbf{x} \in \mathbb{R}_+^{2^d-1}$ adding up to k. For such a $\mathbf{x}$ the likelihood $L_{\mathbf{M}(k;\tilde{\mathbf{p}})}$ of the model is given by

$$L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{x}) = \frac{k!}{\prod_{i=1}^{2^d-1} x_i!} \prod_{i=1}^{s} (\tilde{p}_i)^{x_i} \prod_{i=s+1}^{r} (\tilde{p})^{x_i} \mathbb{1}_{\{\forall j=r+1,\ldots,2^d-1,\, x_j=0\}} .$$

The condition $x_j = 0$ for all $j = r + 1, \ldots, 2^d - 1$ can be rewritten as

$$r \geq |\{j = 1, \ldots, 2^d - 1,\, x_j > 0\}| ,$$

so that $L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{x})$ is maximum for $r = |\{j = 1, \ldots, 2^d-1,\, x_j > 0\}|$. Then the log-likelihood $\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}$ evaluated in $\mathbf{T}_n$ is equal to

$$\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) = \log(k!) - \sum_{i=1}^{2^d-1} \log(T_{n,i}!) + \sum_{i=1}^{s} T_{n,i} \log(\tilde{p}_i) + \Big(\sum_{i=s+1}^{2^d-1} T_{n,i} \Big) \log(\tilde{p}) , \tag{5.2}$$

The optimization of this log-likelihood under the constraint (5.1) leads to the following maximum likelihood estimators:

$$\begin{aligned}
\widehat{r} &:= |\{j = 1, \ldots, 2^d - 1,\, T_{n,j} > 0\}| = \hat{s}_n , \\
\widehat{\tilde{p}} &:= \frac{\sum_{i=s+1}^{\hat{s}_n} T_{n,i}}{(\hat{s}_n - s)k} = \frac{\sum_{i=s+1}^{2^d-1} T_{n,i}}{(\hat{s}_n - s)k} , \\
\widehat{\tilde{p}}_j &:= \frac{T_{n,j}}{k} , \quad 1 \leq j \leq s .
\end{aligned}$$

5.2 An AIC approach for the model $\mathbf{M}(k;\tilde{\mathbf{p}})$

We still consider a fix level $k = k_n$ and the purpose of this section is to identify which model $\mathbf{M}(k;\tilde{\mathbf{p}})$ best fits the vector $\mathbf{T}_n$. The unknown distribution of this vector is denoted by $\mathbf{P}_k$ and the associated likelihood by $L_{\mathbf{P}_k}$. Recall that in the model $\mathbf{M}(k;\tilde{\mathbf{p}})$ the parameter r denotes the number of outcomes associated to a positive probability. Therefore, for a given sample $\mathbf{T}_n$ we only work with models satisfying $r = \hat{s}_n$. We also recall that $\widehat{\tilde{p}}_1, \dots, \widehat{\tilde{p}}_s, \widehat{\tilde{p}}$ denote the maximum likelihood estimators:

$$\widehat{\tilde{\mathbf{p}}} := \underset{\tilde{p}_1+\dots+\tilde{p}_s+(r-s)\tilde{p}=1}{\arg\max} L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\,. \tag{5.3}$$

We define the parameters $\tilde{\mathbf{p}}^* = (\tilde{p}_1^* + \tilde{p}^*, \dots, \tilde{p}_s^* + p^*, \tilde{p}^*)^\top \in \mathbb{R}^{s+1}$ as the optimum of the expectation of the log-likelihood $L_{\mathbf{M}(k;\tilde{\mathbf{p}})}$:

$$\tilde{\mathbf{p}}^* := \underset{\tilde{p}_1+\dots+\tilde{p}_s+(r-s)\tilde{p}=1}{\arg\max} \mathbb{E}[L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)]\,. \tag{5.4}$$

A similar computation as for the estimators $\widehat{\tilde{\mathbf{p}}}$ gives the relations

$$\forall j = 1,\dots,s\,, \quad \tilde{p}_j^* = p_{n,j}\,, \quad \text{and} \quad \tilde{p}^* = \frac{\sum_{i=s+1}^r p_{n,i}}{r-s}\,.$$

For all $j = 1,\dots 2^d-1$, we define

$$m_j = \min(\widehat{\tilde{p}}_j, \tilde{p}_j^*) = \min\Big(\frac{T_{n,j}}{k}, p_{n,j}\Big) \quad \text{and} \quad M_j = \max(\widehat{\tilde{p}}_j, \tilde{p}_j^*) = \max\Big(\frac{T_{n,j}}{k}, p_{n,j}\Big)\,.$$

We make the following assumption on these quantities.

Assumption 2. *For all* $j = 1,\dots,2^d-1$,

$$\frac{p_{n,j}}{m_j^2}\Big|\frac{M_j^2}{m_j^2} - 1\Big| \to 0\,, \qquad \textit{and} \qquad \frac{1}{m_j^2}\Big|\frac{T_{n,j}}{k} - p_{n,j}\Big| \to 0\,, \quad n\to\infty\,.$$

Note that this assumption is automatically satisfied for $j \le s^*$ since in this case m_j and M_j converge to $p_j > 0$. For $j > s^*$, the quantities $p_{n,j}$, m_j, and M_j converge to zero. Assumption 2 allows then to control their joint convergence.

In order to identify which model $\mathbf{M}(k;\,\tilde{\mathbf{p}})$ bets fits to the observation $\mathbf{T}_n$ we start by computing the Kullback-Leibler divergence between the true distribution $\mathbf{P}_k$ and the model $\mathbf{M}(k;\tilde{\mathbf{p}})$:

$$KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big) = \mathbb{E}\left[\log\left(\frac{L_{\mathbf{P}_k}(\mathbf{T}_n)}{L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)}\right)\right] = \mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\,. \tag{5.5}$$

This quantity must be seen as a function of $\tilde{\mathbf{p}}$. In particular the first term of the right-hand side is constant with respect to the parameter $\tilde{\mathbf{p}}$. Regarding the second term, Equation (5.2) entails that

$$\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big] = \log(k!\,) - \mathbb{E}\left[\sum_{i=1}^{2^d-1}\log(T_{n,i}!\,)\right] + k\sum_{i=1}^{s} p_{n,i}\log(\tilde{p}_i) + k\Big(\sum_{i=s+1}^{2^d-1} p_{n,i}\Big)\log(\tilde{p})\,. \tag{5.6}$$

Following the ideas of Akaike (1973), we select the model among the $\mathbf{M}(k;\,\tilde{\mathbf{p}})$ which minimizes the Kullback-Leibler divergence between the true distribution and its model evaluated at the maximum likelihood $\mathbf{M}(k;\,\widehat{\tilde{\mathbf{p}}})$:

$$KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}} = \mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\,. \tag{5.7}$$

The aim of this section is to provide an asymptotic approximation on the expectation of this criterion.

We first establish a Taylor expansion for the estimator in Equation (5.7).

Lemma 2. *There exists $c_1 \in (0,1)$ such that*

$$KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}} = KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*} + \frac{1}{2}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\,. \tag{5.8}$$

Since the quantity $\tilde{\mathbf{p}}^*$ is deterministic, the first term of the right-hand side in (5.8) can be written as follows

$$KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*} = \mathbb{E}\Big[\log L_{\mathbf{P}}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*;\mathbf{T}_n)\Big]\,.$$

The idea is then to provide a Taylor expansion of $\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*;\mathbf{T}_n)$ around the vector $\widehat{\tilde{\mathbf{p}}}$. This is the purpose of the following lemma.

Lemma 3. *There exists $c_2 \in (0,1)$ such that*

$$\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*;\mathbf{T}_n) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n) + \frac{1}{2}(\tilde{\mathbf{p}}^*-\widehat{\tilde{\mathbf{p}}})^\top \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(c_2\tilde{\mathbf{p}}^*+(1-c_2)\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)(\tilde{\mathbf{p}}^*-\widehat{\tilde{\mathbf{p}}})\,. \tag{5.9}$$

After taking the expectation with respect to $\widehat{\tilde{\mathbf{p}}}$ in Equations (5.8) and (5.9), and after combining these equations, we obtain the following expression for the expectation of the estimator in (5.7):

$$\begin{aligned}
&\mathbb{E}\Big[KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\Big] \qquad (5.10)\\
&= \mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big]\\
&+ \mathbb{E}\Big[\underbrace{(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)}_{(\star)}\Big]\\
&+ \frac{1}{2}\mathbb{E}\Big[(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\Big[-\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big|_{c_2\tilde{\mathbf{p}}^*+(1-c_2)\widehat{\tilde{\mathbf{p}}}} \qquad (5.11)\\
&\qquad\qquad + \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}\Big](\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\Big]\,.
\end{aligned}$$

The last two steps consist in dealing with (5.11) and with the term $(\star)$. For the first one, we prove that it converges to zero.

Lemma 4. *Under Assumption 2, the following convergence in probability holds:*

$$\sup_{(c_1,c_2)\in(0,1)^2}\frac{1}{k}\left|\frac{\partial^2\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}}{\partial\tilde{\mathbf{p}}^2}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*} - \mathbb{E}\left[\frac{\partial^2\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}}{\partial\tilde{\mathbf{p}}^2}(\tilde{\mathbf{p}};\mathbf{T}_n)\right]\Big|_{c_2\widehat{\tilde{\mathbf{p}}}+(1-c_2)\tilde{\mathbf{p}}^*}\right|_\infty \to 0\,,$$

when $n\to\infty$ and where $|\cdot|_\infty$ denotes the infinity norm.

Since $\sqrt{k}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)$ converges to a Gaussian distribution thanks to Theorem 2, the term in Equation (5.11) converges to zero when $n\to\infty$. Moving on to the term $(\star)$, we prove that it converges in distribution to a chi-square-distributed random variable with $s+1$ degrees of freedom.

Lemma 5. *For all $c\in(0,1)$, the following weak convergence holds:*

$$(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}\Big]\Big|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*) \xrightarrow{d} \chi(s+1)\,,\quad n\to\infty\,.$$

With Lemma 4 and Lemma 5, Equation (5.10) entails, for n large enough, the following approximation for the expectation of the estimator in Equation (5.7):

$$\begin{aligned}\mathbb{E}\Big[KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\Big] &\approx \mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big] + \mathbb{E}[\chi^2(s+1)]\\ &\approx \mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big] + (s+1)\,.\end{aligned}$$

The first term of the right-hand side is constant regarding the parameter $\tilde{\mathbf{p}}$. Hence, moving back to Equation (5.5), we estimate the quantity $\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]$ with

$$-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n) + (s+1)\,. \tag{5.12}$$

Therefore for a given level k the idea is to choose the parameter s which minimizes this quantity. This provides the s most relevant subsets C_β on which the distribution of the vector $\mathbf{Z}$ places mass. The last theoretical step of our study is to include the choice of $k = k_n$ in our procedure. This is the purpose of the following section.

5.3 A global model

The final step is to consider $k = k_n$ 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 still assume that n is large and we consider a vector $\mathbf{T}'_n$ which models the behavior of the data on the subsets C_β. We assume that this vector follows a multinomial distribution $\mathcal{M}(n;\mathbf{p}'_n)$ with

$$\mathbf{p}'_n = (q_n p_{n,1}, \ldots, q_n p_{n,1}, 1-q_n) \tag{5.13}$$

where the last components $T'_{n,2^d}$ is associated to the non-extreme values. Here, there is neither an underlying level k_n nor a threshold u_n. In order to refer to the work of the Section 5.2, that is, the model selection with a given level k, we stress the dependence in k by denoting the previous vector as $\mathbf{T}_n = \mathbf{T}_n(k) \in \mathbb{R}^{2^d-1}$. We make the following assumption on the quantity q_n.

Assumption 3. *$nq_n \log(n) \to 0$ when $n \to \infty$.*

To be consistent with the previous notations we denote by $\mathbf{P}'_n$ the distribution of $\mathbf{T}'_n$ and by $L_{\mathbf{P}'_n}$ its likelihood. The goal is to estimate the proportion $1-q_n$ of non-extreme values. This is achieved by extending the previous model $\mathbf{M}(k,\tilde{\mathbf{p}})$. We consider a multinomial model denoted by $\mathbf{M}'(n,\tilde{\mathbf{p}}')$ with 2^d outcomes and with a parameter $\tilde{\mathbf{p}}'$ defined as

$$\tilde{\mathbf{p}}' = (\overbrace{\tilde{q}'\tilde{p}'_1, \ldots, \tilde{q}'\tilde{p}'_{s'}, \underbrace{\tilde{q}'\tilde{p}', \ldots, \tilde{q}'\tilde{p}'}_{r'-s'}}^{2^d \text{ terms}}, 0, \ldots, 0, 1-\tilde{q}')\,,$$

where $\tilde{p}'_1 \geq \ldots \geq \tilde{p}'_{s'}, \tilde{p}', \tilde{q}' \in (0,1)$ are satisfying the constraint (5.1). The interpretation of this model is the same as for $\mathbf{M}(k,\tilde{\mathbf{p}})$, and the extra-parameter $\tilde{q}'$ models the proportion of extreme values taken among the data. The log-likelihood $\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}$ of this model is given by

$$\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n) = \log(n!) - \sum_{i=1}^{2^d} \log(T'_{n,i}!) + \sum_{i=1}^{s'} T'_{n,i} \log(\tilde{q}'\tilde{p}'_i) + \Big(\sum_{i=s'+1}^{2^d-1} T'_{n,i}\Big) \log(\tilde{p}'\tilde{q}') + T'_{n,2^d} \log(1-\tilde{q}')\,. \tag{5.14}$$

Similarly to Section 5.2, our aim is to select which model $\mathbf{M}'(n,\tilde{\mathbf{p}}')$ best fits the distribution of $\mathbf{T}'_n$. We fix $r'=\hat{s}_n$ and we write down the Kullback-Leibler divergence between $\mathbf{P}'_n$ and $\mathbf{M}'(n,\tilde{\mathbf{p}}')$ which can be decomposed as follows:

$$KL\Big(\mathbf{P}'_n\Big\|\mathbf{M}(n;\,\tilde{\mathbf{p}}')\Big)=\mathbb{E}\bigg[\log\bigg(\frac{L_{\mathbf{P}'_n}(\mathbf{T}'_n)}{L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n)}\bigg)\bigg]=\mathbb{E}\Big[\log L_{\mathbf{P}'_n}(\mathbf{T}'_n)\Big]-\mathbb{E}\Big[\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n)\Big]\,. \tag{5.15}$$

We focus on the second term of the right-hand side. The log-likelihood $\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}$ defined in Equation (5.14) can be decomposed in the following way:

$$\begin{aligned}\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n)&=\log((n-T'_{n,\,2^d})!\,)-\sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!\,)+\sum_{j=1}^{s'}T'_{n,\,j}\log(\tilde{p}'_j)+\log(\tilde{p}')\sum_{j=s'+1}^{2^d-1}T'_{n,\,j}\\&\quad+\log\bigg(\frac{n!}{(n-T'_{n,\,2^d})!}\bigg)-\log(T'_{n,\,2^d}!\,)+(n-T'_{n,\,2^d})\log(\tilde{q}')+T'_{n,\,2^d}\log(1-\tilde{q}')\\&=\log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}'_{n,\{1,\dots,2^d-1\}})+\phi(n,\tilde{q}',T'_{n,\,2^d})\,,\end{aligned}$$

where

$$\phi(n,\tilde{q}',T'_{n,\,2^d})=\log\bigg(\frac{n!}{(n-T'_{n,\,2^d})!}\bigg)-\log(T'_{n,\,2^d}!\,)+(n-T'_{n,\,2^d})\log(\tilde{q}')+T'_{n,\,2^d}\log(1-\tilde{q}')\,.$$

Following the same ideas as in Section 5.2 and similar to an AIC procedure we estimate the Kullback-Leibler divergence in Equation (5.15) by the estimator $KL(\mathbf{P}_n\|\mathbf{M}'(n;\,\tilde{\mathbf{p}}'))|_{\widehat{\tilde{\mathbf{p}}}'}$ where $\widehat{\tilde{\mathbf{p}}}'$ is the maximum likelihood estimator of $\tilde{\mathbf{p}}'$. The purpose of what follows is to study the expectation of this estimator:

$$\begin{aligned}\mathbb{E}\Big[KL\Big(\mathbf{P}'_n\Big\|\mathbf{M}'(n;\,\tilde{\mathbf{p}}')\Big)\Big|_{\widehat{\tilde{\mathbf{p}}}'}\Big]&=\mathbb{E}\Big[\log L_{\mathbf{P}_n}(\mathbf{T}'_n)\Big]+\mathbb{E}\bigg[\mathbb{E}\Big[-\log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}'_{n,\{1,\dots,2^d-1\}})\mid T'_{n,\,2^d}\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}}\bigg]\\&\quad-\mathbb{E}\bigg[\mathbb{E}\Big[\phi(n,\tilde{q}',T'_{n,\,2^d})\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}'}\bigg]\,.\end{aligned} \tag{5.16}$$

We refer to Appendix B for several calculations regarding the aforementioned quantity which lead to the following approximation:

$$\mathbb{E}\Big[KL\Big(\mathbf{P}'_n\Big\|\mathbf{M}'(n;\,\tilde{\mathbf{p}}')\Big)\Big|_{\widehat{\tilde{\mathbf{p}}}'}\Big]\approx C_n+C'_n\frac{1}{k}\Big(\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big]+(s+1)-k\log(1-k/n)\Big)\,,$$

where C_n and $C'_n>0$ are constants depending on n. Hence, the quantity

$$\frac{1}{k}\Big(-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)+(s+1)-k\log(1-k/n)\Big) \tag{5.17}$$

provides up to some constants an approximation of the Kullback-Leibler divergence in Equation (5.15) evaluated at the maximum likelihood estimator $\widehat{\tilde{p}}'=(\widehat{p},k/n)$. Practically speaking we choose a large range of k (often between 0.5% and 10% of n) and we compute the value of the previous estimator for these k and for $s=1,\dots,\hat{s}_n$. Then we choose the couple (k,s) which minimizes the quantity in (5.17).

This leads to the following algorithm.

Data: A sample $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}_+^d$ and a range of values K for the level
Result: A list $\widehat{\mathcal{S}}(\mathbf{Z})$ of directions β
for $k \in K$ **do**
  Compute $u_n = |\mathbf{X}|_{(k_n+1)}$ the (k_n+1)-th largest norm;
  Assign to each $\pi(\mathbf{X}_j/u_n)$ the subsets C_β it belongs to;
  Compute $\mathbf{T}_n(k)$;
  Compute the minimizer $\hat{s}(k)$ which minimizes the criterion given in Equation (5.12);
end
Minimize $\hat{k}$ of (5.17) plugging in the minimal value in (5.12);
Output: $\widehat{\mathcal{S}}(\mathbf{Z}) = \{\beta,\, T_{n,\,j}(\beta) > 0 \text{ for } j = 1, \dots, \hat{s}(\hat{k})\}$.

Algorithm 1: Tail inference for high-dimensional data.

Remark 5. *We should expect that a slight change in the choice of k does not modify the value of $\hat{s}(k)$. Indeed, as already mentioned, even if our procedure leads to the choice of a unique k it seems natural that all this approach is not too sensitive to this choice. Therefore, it is relevant to plot the function $k \mapsto \hat{s}(k)$. This plot represents the variation of the $\hat{s}$ which minimizes the estimator in (5.12) with k. If k slightly varies around its estimated optimal value $\hat{k}$, we should not observe huge variations of $\hat{s}(k)$. Idealistically, the value of $\hat{s}$ should be constant around the estimated optimal value of $\hat{k}$.*

6 Numerical results

In this section, we illustrate the performance of our method on different numerical examples. For each example, we generate data sets of size $n \in \{10^4, 3 \cdot 10^4, 7 \cdot 10^4\}$ and apply Algorithm 1. We repeat the procedure over $N = 100$ simulations and we compare the outcome $\widehat{\mathcal{S}}(\mathbf{Z})$ of our algorithm with the theoretical set $\mathcal{S}(\mathbf{Z})$. In this case, two types of errors can arise. We call error of Type 1 the selection of a feature β while the distribution of $\mathbf{Z}$ does not place mass in this direction, and we call error of Type 2 the absence of a feature β while this direction should appear theoretically. We summarize the results in a table which contains the following quantities: the average number of the two types of errors among the N simulations, the average number of relevant features s, the average level k and the associated average threshold u. The code can be found at https://github.com/meyernicolas/tail_inference_extremes.

6.1 Asymptotic independence

The first example is the same as the first one in Meyer and Wintenberger (2020), Section 5.2. There, sparse regular variation already provided good results. However, the method required two hyperparameters k and p which are automatically tuned here.

We consider an iid sequence of random vectors $\mathbf{N}_1, \dots, \mathbf{N}_n$ in $\mathbb{R}^{40}$ with generic random vector $\mathbf{N}$ whose distribution is a multivariate Gaussian distribution with all univariate marginals equal to $\mathcal{N}(0,1)$ and the correlations less than 1: $\mathbb{E}[N^i N^j] < 1$ for all $1 \le i \ne j \le d$. We transform the marginals with a rank transform which consists in considering the vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ such that the marginals X_i^j of $\mathbf{X}_i = (X_i^1, \dots, X_i^d)$ are defined as

$$X_i^j = \frac{1}{1 - \hat{F}_j(N_i^j)}\,, \quad 1 \le j \le d\,,$$

where $\hat{F}_j$ is the empirical version of the cumulative distribution function of $N_j \sim \mathcal{N}(0,1)$:

$$\hat{F}_j : x \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{N_i^j < x}\,. \tag{6.1}$$

This provides a sample of regularly varying random vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ and the assumption on the correlation leads to asymptotic independence (see Sibuya (1960)). This case has been discussed in Example 2. Equivalently, it means that $\mathbb{P}(\mathbf{\Theta} \in C_\beta) = \mathbb{P}(\mathbf{Z} \in C_\beta) = 1/d$ for all β such that $|\beta| = 1$ (and therefore $\mathbb{P}(\mathbf{\Theta} \in C_\beta) = \mathbb{P}(\mathbf{Z} \in C_\beta) = 0$ elsewhere). The aim of our procedure is then to recover these 40 maximal directions among the $2^{40} - 1 \approx 10^{12}$ subsets C_β.

Regarding the multivariate Gaussian random vectors $\mathbf{N}_1, \dots, \mathbf{N}_n$, the simulation of these vectors depends only on their covariance matrix. We proceed as follows. We generate a matrix $\mathbf{\Sigma}'$ with entries $\sigma'_{i,j}$ following independent uniform distributions on $(-1, 1)$. Then, we define the matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} := \mathrm{Diag}(\sigma'^{-1/2}_{1,1}, \dots, \sigma'^{-1/2}_{d,d}) \cdot \mathbf{\Sigma}'^{\mathrm{T}} \cdot \mathbf{\Sigma}' \cdot \mathrm{Diag}(\sigma'^{-1/2}_{1,1}, \dots, \sigma'^{-1/2}_{d,d}),$$

where $\mathrm{Diag}(\sigma_{1,1}^{-1/2}, \dots, \sigma_{d,d}^{-1/2})$ denotes the diagonal matrix of $\mathcal{M}_d(\mathbb{R})$ whose diagonal is given by the vector $(\sigma'^{-1/2}_{1,1}, \dots, \sigma'^{-1/2}_{d,d})$. This provides a covariance matrix with diagonal entries equal to 1 and off-diagonal entries less than 1. A given matrix $\mathbf{\Sigma}$ provides then a dependence structure for $\mathbf{N}_1, \dots, \mathbf{N}_n$ and thus for $\mathbf{X}_1, \dots, \mathbf{X}_n$. We generate $N_{\mathrm{model}} = 20$ different matrices $\mathbf{\Sigma}$, and for each of these dependence structures we generate $N = 100$ sample $\mathbf{N}_1, \dots, \mathbf{N}_n$.

Table 1 summarizes the different outcomes of our algorithm. Regarding the average number of errors, we observe that the ones of Type 1 slightly increase with n, while the ones of Type 2 are null. When the data size increases, it is more likely that our algorithm highlights a direction β which does not belong to $\mathcal{S}(\mathbf{Z})$. On the contrary, when n increases the average number of subsets that are not detected by our algorithm decreases. Despite these small variations, we obtain errors that are negligible regarding the total number of subsets $2^{40} - 1 \approx 10^{12}$. Regarding the level k, we see that it increases with n while the ratio k/n decreases. This fits with the standard context of EVT and reinforces our approach.

	Errors of Type 1	Errors of Type 2	Average value of s	Average value of the level k (and k/n)	Average value of the threshold u
$n_1 = 10^4$	1.90	0.00	41.90	630.05 (0.063)	904.91
$n_2 = 3 \cdot 10^4$	3.23	0.00	43.23	868.43 (0.029)	1728.46
$n_3 = 7 \cdot 10^4$	5.010	0.00	45.10	1180.38 (0.017)	2825.73

Table 1: Average outcomes of Algorithm 1 in an asymptotic independent case ($d = 40$).

Following Remark 5, we illustrate on an example the choice of k and s. We keep the same model with $n = n_1 = 10^4$. We plot in Figure 4 the variations of the quantity in Equation (5.17), that is, up to some constants, an estimator of the Kullback-Leibler divergence $KL(\mathbf{P}_n \| \mathbf{M}'(n, \tilde{\mathbf{p}}'))$. We identify on this simulation the optimal value of k which corresponds to $k = k_{\min} = 500$. To see if a slight change of k does not affect the choice of s, we plot in Figure 5 the variations of the optimal value of s with respect to k. For a large range of k the value of s chosen by the algorithm remains constant and close to the theoretical one $s^* = 40$ which means that there exists a range of values of k which provide the same s.

6.2 A dependent example

We consider a random vector $\mathbf{X} \in \mathbb{R}^{100}$ whose marginals are defined as follows:

$$\begin{aligned}
X_j &\sim \mathrm{Pareto}(1), \quad j = 1, \dots, 10, \\
(X_j, X_j + E_j) &\sim (\mathrm{Pareto}(1), X_j + \mathrm{Exp}(1)), \quad k = 11,\ 13,\ 15,\ 17,\ 19, \\
(X_j, X_j + E_j, X_j + E_{j+1}) &\sim (\mathrm{Pareto}(1), X_j + \mathrm{Exp}(1), X_j + \mathrm{Exp}(1)), \quad j = 20,\ 23,\ 26,\ 29,\ 32, \\
X_j &\sim \mathrm{Exp}(1), \quad j = 36, \dots, 100.
\end{aligned}$$

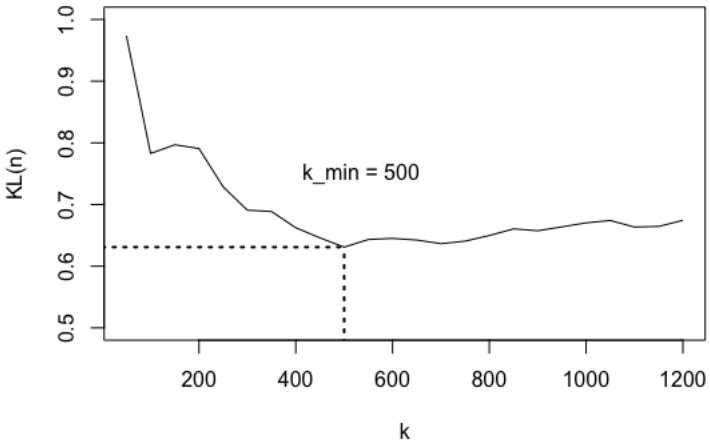


Figure 4: Evolution of the estimator of $KL(\mathbf{P}_n \| \mathbf{M}(n; \tilde{\mathbf{p}}'))$ in an asymptotic independence case.

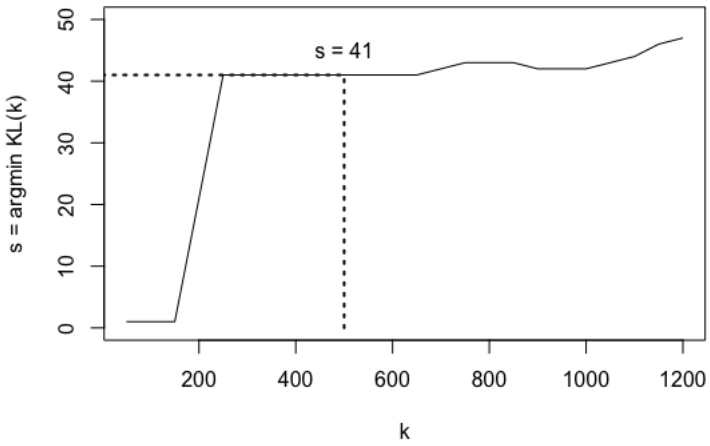


Figure 5: Evolution of the optimal value of s in an asymptotic independence case.

This implies that the spectral vector, and also the angular vector $\mathbf{Z}$, places mass on the following subsets:

$$\begin{aligned} C_{\{k\}} &= \mathbf{e}_k\,, \quad \text{for } k = 1, \ldots, 10\,, \\ C_{\{k,k+1\}}\,, &\quad \text{for } k = 11, 13, 15, 17, 19\,, \\ C_{\{k,k+1,k+2\}}\,, &\quad \text{for } k = 20, 23, 26, 29, 32\,. \end{aligned}$$

Our goal is then to identify the $d_{\text{indep}} = 10$ one-dimensional subsets, the $d_{\text{dep1}} = 5$ two-dimensional subsets, and the $d_{\text{dep2}} = 5$ three-dimensional subsets. Thus in this example $s^* = 20$.

Table 2 summarizes the two types of errors averaged over the N simulations, as well as the average number of relevant features s, the average level k and the associated average threshold u. The errors of

Type 2 decreases when n increases, which makes sense: With not enough data, our procedure fails to identify all relevant directions. But this issue vanishes when n becomes large. For the errors of Type 1, it seems that their number slightly increases with n. If n is large, it is possible to capture a direction that should not be taken into account. However, the average error is negligible regarding the total number of possible directions, that is, $2^{100} - 1 \sim 10^{30}$. In this example, we also observe that the chosen k increases with n, while the ratio k/n tends to decrease.

	Errors of Type 1	Errors of Type 2	Average value of s	Average value of the level k (and k/n)	Average value of the threshold u
$n_1 = 4 \cdot 10^3$	0.05	25.29	14.75	170 (0.043)	1368
$n_2 = 7 \cdot 10^3$	0.09	1.66	18.42	262 (0.037)	1313
$n_3 = 10^4$	0.25	0.82	19.42	304 (0.030)	1504

Table 2: Average number of errors in a dependent case ($d = 100$).

As for the independent case, and following Remark 5, we illustrate on an example the choice of k and s for this dependent case with $n = n_2 = 10^4$. There, we plot in Figure 6 the variations of the quantity in Equation (5.17), that is, up to some constant, an estimator of Akaike's criterion $KL(\mathbf{P}_n \| \mathbf{M}'(n, \widehat{\mathbf{p}}'))$. This simulation leads to a choice of $k = 250$ and provides a graph for which the minimum is well identified. Figure 7 shows that the optimal value of s remains constant around $k = 250$. As for the previous case, we conclude that a slight variation of k does not impact the choice of s.

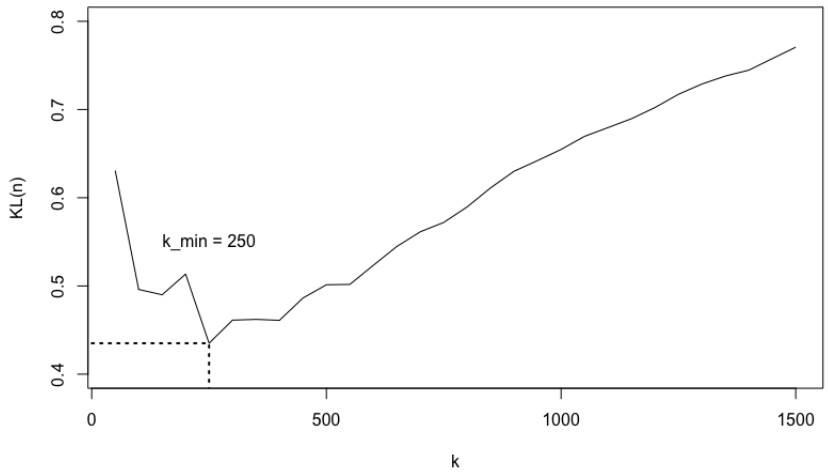


Figure 6: Evolution of the estimator of $KL(\mathbf{P}_n \| \mathbf{M}(n; \tilde{\mathbf{p}}'))$ in a dependent case.

7 Application to extreme variability for financial data

We apply our procedure to 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. 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 absolute

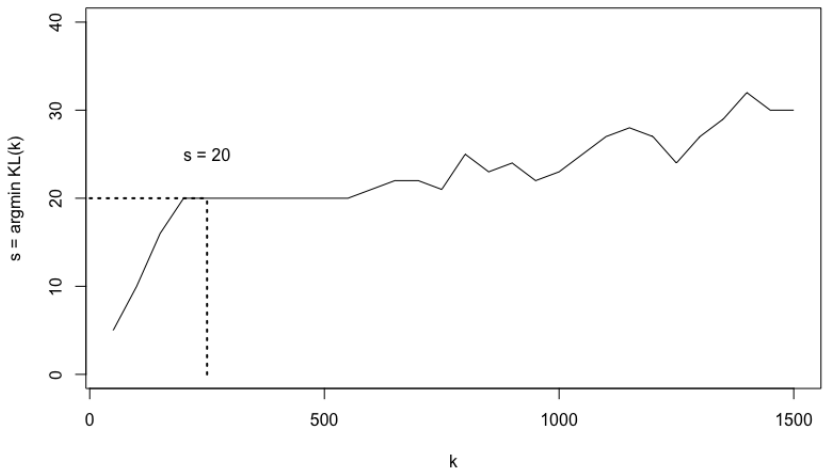


Figure 7: Evolution of the optimal value of s in a dependent case.

value of the data, i.e. we define $\mathbf{x}'_j = \max(\mathbf{x}_j^{\text{obs}}, -\mathbf{x}_j^{\text{obs}})$, where the maximum is meant componentwise. Thus, we study the non-negative vectors $\mathbf{x}'_1, \dots, \mathbf{x}'_n$ in $\mathbb{R}^d_+$ with $n = 12\,613$ and $d = 49$. The iid assumption may be not reasonable but we work as if it was satisfied.

A common preliminary step in the study of extreme values consists in the identification of the tail index α' of the sample $\mathbf{x}'_1, \dots, \mathbf{x}'_n$ which arises in the limit in Equation (1.2). An estimation of this index can be done with a Hill (1975) plot which represents the evolution of the estimation

$$\widehat{\alpha'} = \left(\frac{1}{k} \sum_{j=1}^{k} \log(|\mathbf{x}'|_{(j)}) - \log(|\mathbf{x}'|_{(k)}) \right)^{-1}$$

when $k = 2, \dots, n$, and where $|\mathbf{x}'|_{(j)}$ denotes the order statistics of the norms $|\mathbf{x}'_1|, \dots, |\mathbf{x}'_n|$, i.e.

$$|\mathbf{x}'|_{(1)} \geq \dots \geq |\mathbf{x}'|_{(n)}\,.$$

This plot is given in Figure 9. We observe that the estimator stabilizes around a value of 3.12 which corresponds to the horizontal line. In this context, the idea is to estimate the parameter α' with $\widehat{\alpha'} = 3.12$. Regarding the sample $\mathbf{x}'_1, \dots, \mathbf{x}'_n$, this empirical value of $\widehat{\alpha'}$ means that the tail of these vectors is not very heavy. In this case we observe that the choice of an optimal k is not obvious, see Figure 8. Then, in order to better highlight the extreme behavior of our data and following Remark 2, we consider the vectors $\mathbf{x}_j = (\mathbf{x}'_j)^{\widehat{\alpha'}}$. The tail index α of the sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ is then equal to 1. This transformation allows to work with distributions with heavy tails while it does not change their dependence structure.

We apply Algorithm 1 to the data $\mathbf{x}_1, \dots, \mathbf{x}_n$. As for the numerical examples in Section 6, we plot the evolution of the estimator of the Kullback-Leibler divergence in (5.17) as a function of k. We see on Figure 10 that the minimum of this estimator is well identified and corresponds to a choice of $k = 505$.

This choice of k leads to the estimation $\hat{s} = 16$. Contrary to the numerical examples, Figure 11 does not provide a range of k on which the value of s remains constant. For $\hat{s} = 16$, the directions β which appear are the following ones:

– eleven one-dimensional directions corresponding to the industries Gold, Coal, Smoke, Textiles, Computer Software, Healthcare, Real Estates, Banks, Computers, Entertainment, Soda,

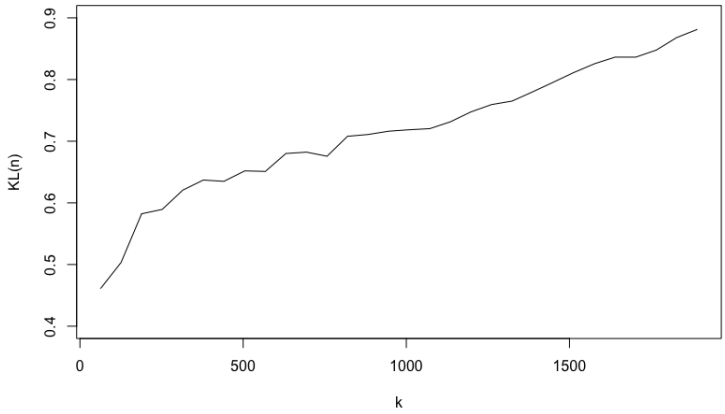


Figure 8: Evolution of the estimator of $KL(\mathbf{P}_n \| \mathbf{M}(n; \tilde{\mathbf{p}}'))$ for the financial data before preprocessing.

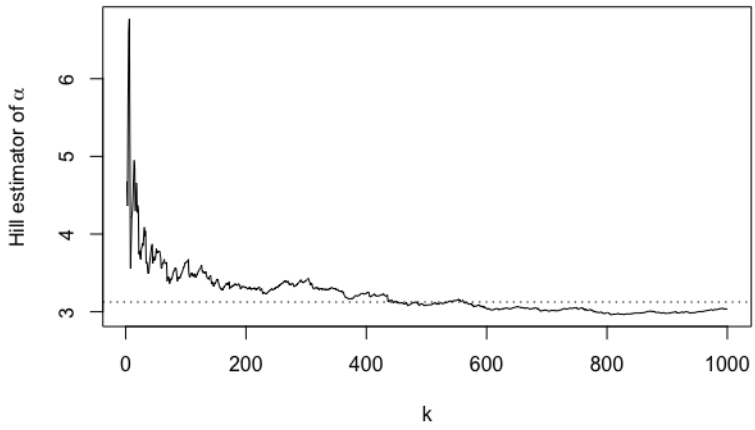


Figure 9: Evolution of the Hill estimator $\hat{\alpha}$ for financial data.

– four two-dimensional directions which gather respectively the couples of industries {Gold, Coal}, {Healthcare, Software}, {Steel, Coal}, and {Banks, Finance},

– a three-dimensional direction which gathers the triplet of industries Coal, Computers, and Electronic Equipment.

We identify 14 different industries which contribute to the extreme variability of the data. Some intuitive groups of variables appear, as the couple Finance and Banks or Steel and Coal. We conclude that the aforementioned subsets concentrate the mass of the angular vector $\mathbf{Z}$. Among these 16 subsets, 10 of them are maximal subsets for $\mathbf{Z}$ and thus for $\mathbf{\Theta}$ thanks to the third point of Proposition 1. These groups of coordinates correspond to the non-dashed ellipses in Figure 12. Separate studies should be conducted on these groups of portfolios, for which standard approaches for low-dimensional extremes can be applied, see Coles and Tawn (1991), Einmahl et al. (1993), Einmahl et al. (1997), Einmahl and Segers (2009).

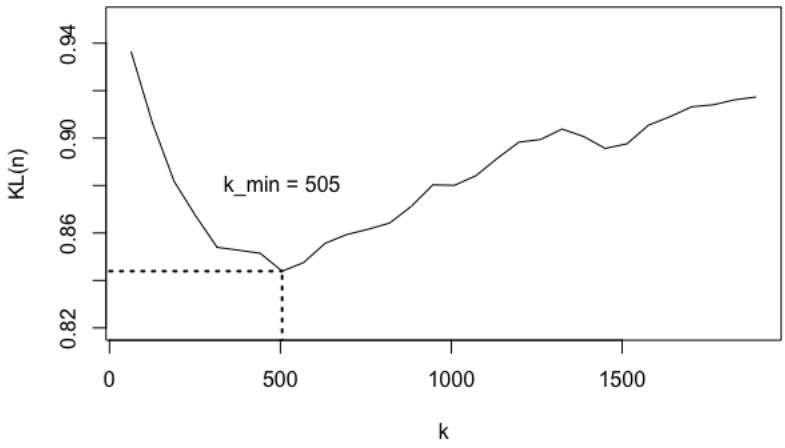


Figure 10: Evolution of the estimator of $KL(\mathbf{P}_n \| \mathbf{M}(n; \tilde{\mathbf{p}}'))$ for the financial data.

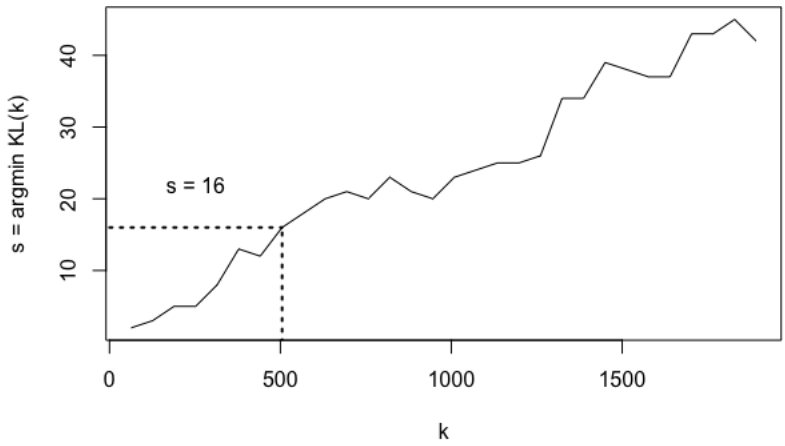


Figure 11: Evolution of the optimal value of s for financial data.

8 Discussion

The statistical analysis introduced in this article highlights a new method whose goal is double: to capture the tail structure of a regularly varying random vector and to provide a statistical method for the choice of the level k. 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. Regarding the tail estimation, most of the existing methods only tackle low-dimensional framework.

Our approach based on sparse regular variation relies on several asymptotic results (Proposition 2, Theorem 1 and Theorem 2) which ensure consistency and asymptotic normality of given estimators. The model selection proposed in Section 5 manages to tackle simultaneously the detection of the extremal

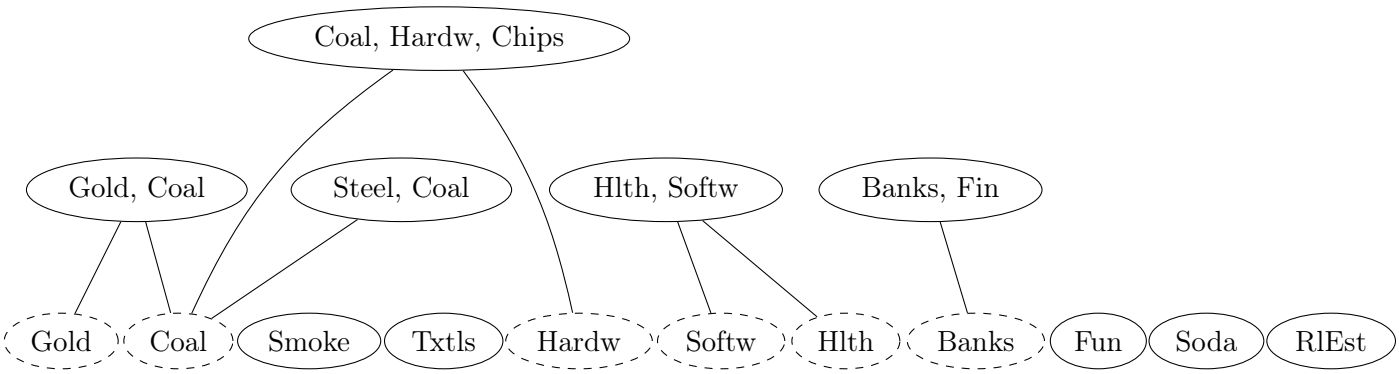


Figure 12: Representation of the 16 features and their inclusions. The abbreviations are the following ones: Hardw = Hardware, Chips = Electronic Equipment, Hlth = Healthcare, Softw = Computer Software, Fin = Finance, Txtls = Textiles, RlEst = Real Estate, Fun = Entertainment. The dashed ellipses correspond to non-maximal directions.

directions and the choice of a level. Regarding the tail dependence, our method highlights it with the identification of the most relevant directions β. The selection is done with an AIC-type minimization whose penalization allows to reduce the number of selected subsets. Including the choice of an appropriate level k entails then a multiplicative penalization. Following the philosophical mantra of EVT, "let the tail speaks for itself", we provide an ad hoc estimation procedure in which the data partitions itself into two categories, an extreme one and non-extreme one. This approach leads to an algorithm whose purpose is to recover the extremal directions of a sample of iid regularly varying random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

The numerical simulations of our algorithm provide promising results for asymptotic independent cases but also for dependent cases. In particular we manage to deal with high-dimensional data, at least compared to the examples given in the EVT literature so far. This procedure is indeed tested on samples of dimension $d = 100$ for which the number of possible directions is $2^{100} - 1$. Despite this very high-dimensional setting we succeed in capturing the extremal directions and do not recover many exta-subsets. In particular, the numerical results show that our approach seems quite accurate for asymptotically independent data. The real-world example we develop reinforces the relevance of our approach. We obtain a sparse structure for the extreme variability of the portfolios.

The proposed method provides good results when the dimension d is large. Paradoxically, it does not seem to be effective for small values of d. Indeed, some numerical results not exposed here show that in this case our algorithm does not manage to capture the theoretical subsets C_β very well. Actually, it overestimates the number of relevant directions which implies a large error of Type 1, while the error of Type 2 is moderate. This issue is a good starting point for future work appealing for a study of the bias in small, moderate and large dimension settings. Besides, the study of non-maximal directions should also be at the core of future reaserch since there is no guarantee that the spectral measure actually places mass in such directions. A more deeper study of the behavior of $\mathbf{Z}$ in these directions should therefore be conducted.

A Algorithm

We introduce here the linear-time algorithm given in Duchi et al. (2008). It is based on a random selection of the coordinates.

Data: A vector $\mathbf{v} \in \mathbb{R}_+^d$ and a scalar $z > 0$
Result: The projected vector $\mathbf{w} = \pi(\mathbf{v})$
Initialize $U = \{1, \dots, d\}$, $s = 0$, $\rho = 0$;
while $U \neq \emptyset$ **do**
  Pick $k \in U$ at random;
  Partition U: $G = \{j \in U,\, v_j \geq v_k\}$ and $L = \{j \in U,\, v_j < v_k\}$;
  Calculate $\Delta\rho = |G|, \quad \Delta s = \sum_{j \in G} v_j$;
  if $(s + \Delta s) - (\rho + \Delta\rho)v_k < z$ **then**
    $s = s + \Delta s$;
    $\rho = \rho + \Delta\rho$;
    $U \leftarrow L$;
  else
    $U \leftarrow G \setminus \{k\}$;
  end
end
Set $\eta = (s - z)/\rho$;
Output: $\mathbf{w}$ s.t. $w_i = v_i - \eta$.

Algorithm 2: Linear time projection onto the positive sphere $\mathbb{S}_+^{d-1}(z)$.

B Some calculations for the global model

We start from Equation (5.16). The first term on the right-hand side is a constant. For the second term we start with Equation (5.2) and remark that the log terms

$$-\mathbb{E}\Big[\log((n - T'_{n,2^d})!\,) - \sum_{j=1}^{2^d-1} \log(T'_{n,j}!\,)\Big]$$

are constant. The idea is then to condition the remainder with respect to $T'_{n,2^d}$ in order to apply the results of the previous section. Since $\mathbf{T}'_n$ follows a multinomial distribution with parameter $\mathbf{p}'_n$ given in (5.13), the conditional distribution of $(T'_{n,1}, \dots, T'_{n,2^d-1})^\top \mid T'_{n,2^d}$ is $\mathcal{M}(n - T'_{n,2^d}, \mathbf{p}_n)$. This entails in particular that $\mathbb{E}[T'_{n,j} \mid T'_{n,2^d}] = (n - T'_{n,2^d})p_{n,j}$.

Hence, after removing the log terms, we consider the quantity

$$\begin{aligned}
&\mathbb{E}\Big[-\log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}'_{n,\{1,\dots,2^d-1\}}) + \log((n-T'_{n,\,2^d})!\,) - \sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!\,) \mid T'_{n,\,2^d}\Big] \\
&\quad = \sum_{j=1}^{s}\mathbb{E}[T'_{n,\,j}\mid T'_{n,\,2^d}]\log(\tilde{p}_j) + \log(\tilde{p})\sum_{j=s+1}^{2^d-1}\mathbb{E}[T'_{n,\,j}\mid T'_{n,\,2^d}] \\
&\quad = (n-T'_{n,\,2^d})\Big(\sum_{j=1}^{s}p_{n,\,j}\log(\tilde{p}_j) + \log(\tilde{p})\sum_{j=s+1}^{2^d-1}p_{n,\,j}\Big) \\
&\quad = \frac{n-T'_{n,\,2^d}}{k}\Big(k\sum_{j=1}^{s}p_{n,\,j}\log(\tilde{p}_j) + k\log(\tilde{p})\sum_{j=s+1}^{2^d-1}p_{n,\,j}\Big) \\
&\quad = \frac{n-T'_{n,\,2^d}}{k}\left(\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n(k))\Big] + \log(k!\,) - \mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T_{n,\,j}(k)!\,)\Big]\right),
\end{aligned}$$

from (5.6) since $\mathbf{T}_n(k)$ is given $T'_{n,\,2^d} = n-k$ implicitly in the model $\mathbf{M}(k;\tilde{\mathbf{p}})$. Then, we evaluate the previous quantity in $\widehat{\tilde{\mathbf{p}}}'$ and take its expectation:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\Big[-\log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}'_{n,\{1,\dots,2^d-1\}}) + \log((n-T'_{n,\,2^d})!\,) - \sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!\,) \mid T'_{n,\,2^d}\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}}\right] \\
&\quad = \frac{n(1-q_n)}{k}\left(\mathbb{E}\left[\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n(k))\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}}\right] + \log(k!\,) - \mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T_{n,\,j}(k)!\,)\Big]\right).
\end{aligned}$$

For the third term in Equation (5.16), we have

$$\begin{aligned}
\mathbb{E}\Big[\phi(n,\tilde{q}',T'_{n,\,2^d})\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}'} &= \mathbb{E}\Big[\log\left(\frac{n!}{(n-T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!\,)\Big] + \mathbb{E}\Big[(n-T'_{n,\,2^d})\log(\tilde{q}') + T'_{n,\,2^d}\log(1-\tilde{q}')\Big]\Big|_{\widehat{\tilde{q}}'} \\
&= \mathbb{E}\Big[\log\left(\frac{n!}{(n-T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!\,)\Big] + nq_n\log(k/n) + n(1-q_n)\log(1-k/n)\,,
\end{aligned}$$

as we use the estimator $\widehat{\tilde{q}}' = k/n$.

We consider a large n and we use the results of the previous section. Then, we estimate the Kullback-Leibler divergence $KL(\mathbf{P}_n\|\mathbf{M}(n;\,\tilde{\mathbf{p}}'))$ in (5.15) by the unbiased estimator $\mathbb{E}[KL(\mathbf{P}_n\|\mathbf{M}(n;\,\tilde{\mathbf{p}}'))|_{\widehat{\tilde{\mathbf{p}}}'}]$ in (5.16) which can be approximated by the following quantity

$$\mathbb{E}\Big[\log L_{\mathbf{P}'_n}(\mathbf{T}'_n)\Big] + \frac{n(1-q_n)}{k}\Big(\mathbb{E}\Big[-\log L_{\mathbf{M}'(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n(k))\Big] + (s+1)\Big) + R_{n,\,k}\,, \tag{B.1}$$

where $R_{n,\,k}$ is defined as

$$\begin{aligned}
R_{n,\,k} &= -\mathbb{E}\Big[\log\left(\frac{n!}{(n-T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!\,)\Big] - nq_n\log(k/n) - n(1-q_n)\log(1-k/n) \\
&\quad + \frac{n(1-q_n)}{k}\Big(\log(k!\,) - \mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T_{n,\,j}(k)!\,)\Big]\Big) + \mathbb{E}\Big[-\log((n-T'_{n,\,2^d})!\,) + \sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!\,)\Big]\,.
\end{aligned}$$

After withdrawing the terms which are constant with respect to k, it remains

$$-nq_n \log(k/n) - n(1-q_n)\log(1-k/n) + \frac{n(1-q_n)}{k}\Big(\log(k!\,) - \mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T_{n,j}(k)!\,)\Big]\Big)\,. \tag{B.2}$$

With Proposition 2, we have the convergence in probability $T_{n,j}(k)/k - p_{n,j} \to 0$ when $n \to \infty$. Then, Stirling's approximation entails the following approximation:

$$\begin{aligned}
\log(k!\,) - \mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T_{n,j}(k)!\,)\Big] &\approx k\log(k) - k - \sum_{j=1}^{2^d-1}\Big(kp_{n,j}\log(kp_{n,j}) - kp_{n,j}\Big)\\
&\approx k\log(k) - \sum_{j=1}^{2^d-1} kp_{n,j}\log(k) - \sum_{j=1}^{2^d-1} kp_{n,j}\log(p_{n,j})\\
&\approx -k\sum_{j=1}^{2^d-1} p_{n,j}\log(p_{n,j})\,,
\end{aligned}$$

where we use that the $p_{n,j}$ add up to 1. This implies that the last term in (B.2) is approximately constant. Regarding the first one, Assumption 3 implies that

$$|-nq_n\log(k/n)| \leq nq_n\log(n) \to 0\,, \quad n\to\infty\,.$$

The only term remaining which has not been neglected in $R_{n,k}$ is then $-n(1-q_n)\log(1-k/n)$.

C Proofs

Proof of Proposition 2. We use the Weak Law of Large Numbers for triangular array (see for instance Feller (1971)). For a Borel set $A \subset \mathbb{S}_+^{d-1}$ we set $Y_{j,n} = k_n^{-1}\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n\}$. Then, in order to obtain the convergence in probability

$$\sum_{j=1}^{n}(Y_{j,n} - \mathbb{E}[Y_{j,n}]) \to 0\,, \quad n\to\infty\,,$$

it suffices to show that $\sup_{n,j} E[Y_{j,n}^2] < \infty$. Starting from the relation

$$E[Y_{j,n}^2] = \frac{\mathbb{P}(\pi(\mathbf{X}_j/u_n)\in A,\, |\mathbf{X}_j| > u_n)}{k_n^2} = \frac{p_n(A)}{nk_n}\,,$$

we obtain that $\sup_{n,j} E[Y_{j,n}^2] \leq k_n^{-1}$. Since $k_n \to \infty$, this implies the convergence in probability

$$\frac{T_n(A)}{k_n} - p_n(A) \to 0\,, \quad n\to\infty\,. \tag{C.1}$$

The convergence in Equation (3.6) can be established in a similar way. Finally, the convergence in Equation (3.5) is just a consequence of Equations (3.4) and (2.5). □

Proof of Theorem 1. For a Borel set A of $\mathbb{S}_+^{d-1}$, we define

$$V_{j,n} = \sigma_n^{-1}\Big(\mathbb{1}\{\pi(\mathbf{X}_j/u_n)\in A,\, |\mathbf{X}_j| > u_n\} - \frac{k_n}{n}p_n(A)\Big)\,,$$

where

$$\sigma_n^2 = n\mathbb{P}(\pi(\mathbf{X}/u_n) \in A,\, |\mathbf{X}| > u_n)[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in A,\, |\mathbf{X}| > u_n)]\,.$$

The variable $V_{j,n}$ satisfies the relations $\mathbb{E}[V_{j,n}] = 0$ and $\mathrm{Var}(V_{j,n}) = 1/n$. Then, in order to prove the convergence

$$\frac{\sum_{j=1}^n (V_{j,n} - \mathbb{E}[V_{j,n}])}{\sqrt{\sum_{j=1}^n \mathrm{Var}(V_{j,n})}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,,$$

it suffices to show that Lindeberg's condition holds:

$$\sum_{j=1}^n \mathbb{E}\Big[V_{j,n}^2 \mathbb{1}_{\{V_{j,n} > \epsilon\}}\Big] = n\mathbb{E}\Big[V_{1,n}^2 \mathbb{1}_{\{V_{1,n} > \epsilon\}}\Big] \to 0\,, \quad n \to \infty\,, \tag{C.2}$$

for all $\epsilon > 0$. Thus, fix $\epsilon > 0$. On the one hand, the variance σ_n^2 is equivalent to $(k_n p_n(A))^2$ which converges to ∞ by assumption. On the other hand, $|\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n\} - \frac{k_n}{n} p_n(A)|$ is always bounded by 1. Hence, for n large enough, the inequality $V_{1,n} \le \epsilon$ is always satisfied. This proves that the condition in (C.2) holds and then implies that

$$\frac{\sum_{j=1}^n (V_{j,n} - \mathbb{E}[V_{j,n}])}{\sqrt{\sum_{j=1}^n \mathrm{Var}(V_{j,n})}} = \frac{T_n(A) - k_n p_n(A)}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,.$$

Finally, Slutsky's theorem allows to replace σ_n by $\sqrt{k_n p_n(A)}$, which yields to the following convergence

$$\sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p_n(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,.$$

For the convergence (3.8), the regularity assumption implies that $p_n(A) \to p(A) > 0$ when $n \to \infty$. Therefore, an application of Slutsky's theorem allows to conclude.

In order to prove (3.10), we decompose the previous ratio in the following way:

$$\sqrt{k_n}\frac{T_n(A)/k_n - p(A)}{\sqrt{p_n(A)}} = \sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p_n(A)}} + \sqrt{k_n}\frac{p_n(A) - p(A)}{\sqrt{p_n(A)}}\,.$$

It is then sufficient to show that the second term goes to 0 as $n \to \infty$. This is true thanks to the bias assumption (3.9) and since the denominator $\sqrt{p_n(A)}$ converges to a positive limit. □

Proof of Lemma 1. Recall that $T_n(\beta) = \sum_{j=1}^n \mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\}$ where the $\mathbf{X}_j$'s are iid. This implies that

$$\begin{aligned}
\mathbb{P}(T_n(\beta) = 0) &= \mathbb{P}(\forall j = 1, \dots, n,\, \pi(\mathbf{X}_j/u_n) \notin C_\beta \text{ or } |\mathbf{X}_j| \le u_n) \\
&= [1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| \le u_n)]^n \\
&= \exp\left(n \log[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n)]\right).
\end{aligned}$$

Hence, since $\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n) \to 0$, we obtain the Taylor expansion

$$\log[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n)] \sim -\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n)\,, \quad n \to \infty\,.$$

Finally, we write $n\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n) = k_n \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$ which gives the desired result. □

Proof of Theorem 2. We consider the vector $\mathbf{V}_{n,\mathcal{R}_k(\mathbf{Z})} \in \mathbb{R}^{r^*}$ whose components are

$$V_{n,\beta} = \frac{1}{\sqrt{k_n p_n(\beta)}}\Big(\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\} - \frac{k_n}{n} p_n(\beta)\Big)\,.$$

This vector has null expectation. We denote by $\Sigma_n \in \mathcal{M}_{r^*}(\mathbb{R})$ its covariance matrix. First, the diagonal entries correspond to the variance of a Bernoulli distribution, i.e.

$$\Sigma_n(\beta,\beta) = \frac{1}{k_n p_n(\beta)}\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)] = \frac{1}{n} - \frac{k_n}{n^2} p_n(\beta)\,.$$

Second, the non-diagonal entries can be computed as follows:

$$\begin{aligned}
\Sigma_n(\beta,\beta') &= \mathbb{E}[V_{n,\beta} V_{n,\beta'}] \\
&= \frac{1}{\sqrt{k_n p_n(\beta)}}\frac{1}{\sqrt{k_n p_n(\beta')}}\Big(\mathbb{E}[\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\}\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_{\beta'},\, |\mathbf{X}| > u_n\}] \\
&\quad - \frac{k_n}{n} p_n(\beta)\mathbb{E}[\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_{\beta'},\, |\mathbf{X}| > u_n\}] - \frac{k_n}{n} p_n(\beta')\mathbb{E}[\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\}] \\
&\qquad\qquad + \frac{k_n^2}{n^2} p_n(\beta) p_n(\beta')\Big) \\
&= -\frac{1}{k_n\sqrt{p_n(\beta)p_n(\beta')}}\frac{k_n^2}{n^2} p_n(\beta) p_n(\beta') \\
&= -\frac{k_n}{n^2}\sqrt{p_n(\beta)p_n(\beta')}\,.
\end{aligned}$$

Hence, the covariance matrix Σ_n can be written as

$$\Sigma_n = \frac{1}{n} Id_{r^*} - \frac{k_n}{n^2}\sqrt{\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}} \cdot \sqrt{\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}}^{\top}\,,$$

where the square root is meant componentwise. In particular, $n\Sigma_n \to Id_{r^*}$ when $n \to \infty$.

Consider now a triangular array $\mathbf{V}_{n,1}, \ldots, \mathbf{V}_{n,n}$ with the same distribution as $\mathbf{V}_{n,\mathcal{R}_k(\mathbf{Z})}$. We prove that this triangular array satisfies Lindeberg's condition:

$$\sum_{j=1}^{n} \mathbb{E}\left[\frac{1}{k_n}\max_\beta \frac{1}{p_n(\beta)}\left|\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\} - \frac{k_n}{n} p_n(\beta)\right|^2 \mathbb{1}_{\{\max_\beta |V_{n,j,\beta}| > \epsilon\}}\right] \to 0\,, \quad n \to \infty\,,$$

for all $\epsilon > 0$, or equivalently that

$$\mathbb{E}\left[\frac{n}{k_n}\max_\beta \frac{1}{p_n(\beta)}\left|\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\} - \frac{k_n}{n} p_n(\beta)\right|^2 \mathbb{1}_{\{\max_\beta |V_{n,\beta}| > \epsilon\}}\right] \to 0\,, \quad n \to \infty\,. \tag{C.3}$$

Fix $\epsilon > 0$. Recall that $\mathcal{R}_k(\mathbf{Z})$ gathers all features β such that $k_n p_n(\beta) \to \infty$. Thus, there exists n_0 such that for all $n \geq n_0$,

$$\max_{\beta \in \mathcal{R}_k(\mathbf{Z})}\left|\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\} - \frac{k_n}{n} p_n(\beta)\right| \leq \epsilon k_n \min_{\beta \in \mathcal{R}_k(\mathbf{Z})} p_n(\beta)\,,$$

since the term on the left-hand side is always bounded by 1. This implies that for n large enough, the inequality $\max_\beta |V_{n,\beta}| > \epsilon$ is never satisfied. Hence, Lindeberg's condition in (C.3) holds and yields to the following convergence

$$\sum_{j=1}^{n} \mathbf{V}_{n,j} \xrightarrow{d} \mathcal{N}(0, Id_{r^*})\,, \quad n \to \infty\,.$$

This convergence can be rephrased as

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n}-\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right)\xrightarrow{d}\mathcal{N}(0,Id_{r^*})\,,\quad n\to\infty\,,$$

which proves (4.9).

To obtain the convergence in (4.10), it suffices to restrict the previous convergence to the coordinates $\beta\in\mathcal{S}(\mathbf{Z})$ and to notice that

$$\mathrm{Diag}(\mathbf{p}_{n,\mathcal{S}(\mathbf{Z})})^{1/2}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}\to Id_{s^*}\,,\quad n\to\infty\,.$$

Finally, to prove (4.12) it suffices to show that $\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}(\mathbf{p}_{n,\mathcal{S}(\mathbf{Z})}-\mathbf{p}_{\mathcal{S}(\mathbf{Z})})\to 0$ which is true under assumption (4.11). □

The two following lemmas are a consequence the following result known as "Cauchy's Mean-Value Theorem" (see Hille (1964) for a proof).

Lemma 6. *Let f and g be two continuous functions on the closed interval $[a,b]$, $a<b$, and differentiable on the open interval (a,b). Then there exists some $c\in(a,b)$ such that*

$$(f(b)-f(a))g'(c)=(g(b)-g(a))f'(c)\,.$$

Proof of Lemma 2. Let f be the function defined as $f(t)=h(t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*)$ for $t\in[0,1]$, where h is defined as

$$h(\tilde{\mathbf{p}})=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)+\frac{\partial}{\partial\tilde{\mathbf{p}}}KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}})\,.$$

Some short calculations give the following relations:

$$\begin{aligned}
f(1)&=h(\widehat{\tilde{\mathbf{p}}})=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\,,\\
f(0)&=h(\tilde{\mathbf{p}}^*)=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}+\frac{\partial}{\partial\tilde{\mathbf{p}}}KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}-\underbrace{\frac{\partial}{\partial\tilde{\mathbf{p}}}\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}}_{=0\text{ by definition of }\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}\,,\\
f'(t)&=\frac{\partial h}{\partial\tilde{\mathbf{p}}}(t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=(\widehat{\tilde{\mathbf{p}}}-[t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*])^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=(1-t)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\,.
\end{aligned}$$

We apply Lemma 6 to the functions f and $g:t\mapsto(t-1)^2$. There exists $c_1\in(0,1)$ such that $(f(1)-f(0))g'(c_1)=(g(1)-g(0))f'(c_1)$, i.e.

$$\begin{aligned}
&\left(KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}-KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}\right)2(c_1-1)\\
&\qquad=(1-c_1)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\,.
\end{aligned}$$

Simplifying by $2(c_1-1)\neq 0$ gives the desired result. □

Proof of Lemma 3. Consider $f(t) = h(t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}})$, for $t \in [0,1]$ where h is defined as

$$h(\tilde{\mathbf{p}}) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) + \frac{\partial}{\partial \tilde{\mathbf{p}}} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \tilde{\mathbf{p}})\,.$$

Some short calculations give the following relations:

$$\begin{aligned}
f(1) &= h(\tilde{\mathbf{p}}^*) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*; \mathbf{T}_n)\,,\\
f(0) &= h(\widehat{\tilde{\mathbf{p}}}) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n) + \underbrace{\frac{\partial}{\partial \tilde{\mathbf{p}}} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)}_{=0 \text{ by definition of } \widehat{\tilde{\mathbf{p}}}}(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\,,\\
f'(t) &= \frac{\partial h}{\partial \tilde{\mathbf{p}}}(t\tilde{\mathbf{p}}^* + (1-t)t\widehat{\tilde{\mathbf{p}}})(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\\
&= (\tilde{\mathbf{p}}^* - [t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}}])^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\\
&= (1-t)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\,.
\end{aligned}$$

We apply Lemma 6 to the functions f and $g : t \mapsto (t-1)^2$. There exists $c_2 \in (0,1)$ such that $f(1) - f(0))g'(c_2) = (g(1) - g(0))f'(c_2)$, i.e.

$$\begin{aligned}
\Big(\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*; \mathbf{T}_n) - \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)\Big) 2(c_2 - 1)&\\
= -(1-c_2)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(c_2\tilde{\mathbf{p}}^* + (1-c_2)\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\,.&
\end{aligned}$$

A simplification by $2(c_2 - 1) \neq 0$ leads to the desired result. □

For the following two lemmas, we define the functions ψ_j and ψ as follows:

$$\psi_j(c) = c\widehat{\tilde{p}}_j + (1-c)\tilde{p}_j^* = c\frac{T_{n,j}}{k} + (1-c)p_{n,j}\,, \quad j = 1, \dots, s\,,$$

and

$$\psi(c) = c\widehat{\tilde{p}} + (1-c)\tilde{p}^* = \frac{1}{r-s}\sum_{j=s+1}^{r} \psi_j(c)\,.$$

The third point of Proposition 2 implies that

$$\frac{\psi_j(c)}{p_{n,j}} = c\frac{T_{n,j}}{kp_{n,j}} + (1-c) \to c + (1-c) = 1\,, \quad n \to \infty\,. \tag{C.4}$$

Besides, the functions ψ_j and ψ satisfy the relations

$$\inf_{c\in(0,1)} \psi_j(c) = \min(\widehat{\tilde{p}}_j, \tilde{p}_j^*) = m_j \quad \text{and} \quad \sup_{c\in(0,1)} \psi_j(c) = \max(\widehat{\tilde{p}}_j, \tilde{p}_j^*) = M_j\,, \quad j = 1, \dots, s\,,$$

and we define

$$m := \inf_{c\in(0,1)} \psi(c) = \min(\widehat{\tilde{p}}, \tilde{p}^*) \quad \text{and} \quad M := \sup_{c\in(0,1)} \psi(c) = \max(\widehat{\tilde{p}}, \tilde{p}^*)\,.$$

Proof of Lemma 4. We differentiate twice the expression in Equation (5.2) with respect to the vector $\tilde{\mathbf{p}}$. This leads to the following Hessian matrix:

$$-\frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) = \begin{pmatrix} \frac{T_{n,1}}{\tilde{p}_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{T_{n,2}}{\tilde{p}_2^2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \frac{\sum_{j=s+1}^r T_{n,j}}{\tilde{p}^2} \end{pmatrix}.$$

Then, our goal is to prove that

$$\forall j = 1, \dots, s\,, \qquad \sup_{(c_1,c_2)\in(0,1)^2} \left| \frac{T_{n,j}}{k\psi_j(c_1)^2} - \frac{p_{n,j}}{\psi_j(c_2)^2} \right| \to 0\,, \tag{C.5}$$

$$\text{and} \qquad \sup_{(c_1,c_2)\in(0,1)^2} \left| \frac{\sum_{j=s+1}^r T_{n,j}}{(r-s)k\psi(c_1)^2} - \frac{\sum_{j=s+1}^r p_{n,j}}{(r-s)\psi(c_2)^2} \right| \to 0\,. \tag{C.6}$$

Regarding (C.5), we write

$$\begin{aligned} \left| \frac{T_{n,j}}{k\psi_j(c_1)^2} - \frac{p_{n,j}}{\psi_j(c_2)^2} \right| &\le \frac{1}{\psi_j(c_1)^2} \left| \frac{T_{n,j}}{k} - p_{n,j} \right| + p_{n,j} \left| \frac{1}{\psi_j(c_1)^2} - \frac{1}{\psi_j(c_2)^2} \right| \\ &\le \frac{1}{m_j^2} \left| \frac{T_{n,j}}{k} - p_{n,j} \right| + \frac{p_{n,j}}{\psi_j(c_1)^2\psi_j(c_2)^2} \left| \psi_j(c_2)^2 - \psi_j(c_1)^2 \right| \\ &\le \frac{1}{m_j^2} \left| \frac{T_{n,j}}{k} - p_{n,j} \right| + \frac{p_{n,j}}{m_j^4} \left| M_j^2 - m_j^2 \right|, \end{aligned}$$

and thus we obtain that

$$\sup_{(c_1,c_2)\in(0,1)^2} \left| \frac{T_{n,j}}{k\psi_j(c_1)^2} - \frac{p_{n,j}}{\psi_j(c_2)^2} \right| \le \frac{1}{m_j^2} \left| \frac{T_{n,j}}{k} - p_{n,j} \right| + \frac{p_{n,j}}{m_j^4} \left| M_j^2 - m_j^2 \right| \to 0\,, \quad n \to \infty\,,$$

where the convergence of both terms results from Assumption 2.

We move on to the term (C.6). For all $c \in (0,1)$ we have the following inequalities

$$\begin{aligned} \left| \frac{\sum_{j=s+1}^r T_{n,j}}{k(r-s)\psi(c_1)^2} - \frac{\sum_{j=s+1}^r p_{n,j}}{(r-s)\psi(c_2)^2} \right| &\le \frac{1}{\psi(c_1)^2} \left| \frac{\sum_{j=s+1}^r T_{n,j}}{k} - \sum_{j=s+1}^r p_{n,j} \right| + \frac{\sum_{j=s+1}^r p_{n,j}}{\psi(c_1)^2\psi(c_2)^2} \left| \psi(c_2)^2 - \psi(c_1)^2 \right| \\ &\le \frac{1}{(\sum_{j=s+1}^r m_j)^2} \sum_{j=s+1}^r \left| \frac{T_{n,j}}{k} - p_{n,j} \right| + \frac{\sum_{j=s+1}^r p_{n,j}}{(\sum_{j=s+1}^r m_j)^4} \left| \Big(\sum_{j=s+1}^r M_j \Big)^2 - \Big(\sum_{j=s+1}^r m_j \Big)^2 \right|, \end{aligned}$$

which converges to zero thanks to Assumption 2. □

Proof of Lemma 5. We start with Equation (5.2) and take the expectation of both sides:

$$\mathbb{E}[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n)] = -\log(k!) + \sum_{j=1}^{2^d-1} \mathbb{E}[\log(T_j!)] - \sum_{j=1}^s k p_{n,j} \log(\tilde{p}_j) - \Big(\sum_{j=s+1}^r k p_{n,j} \Big) \log(\tilde{p})\,.$$

Then, differentiating twice this expression with respect to the vector $\tilde{\mathbf{p}}$ leads to the following Hessian matrix:

$$\frac{\partial^2}{\partial \tilde{\mathbf{p}}^2}\mathbb{E}[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)] = \begin{pmatrix} \frac{kp_{n,1}}{\tilde{p}_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{kp_{n,2}}{\tilde{p}_2^2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \frac{\sum_{j=s+1}^r kp_{n,j}}{\tilde{p}^2} \end{pmatrix}.$$

Then, for $c \in (0,1)$, we write

$$\begin{aligned}(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2}\mathbb{E}[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)]|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*)& \\ = \sum_{j=1}^s \frac{k(\widehat{\tilde{p}}_j - \tilde{p}_j^*)^2 p_{n,j}}{\psi_j(c)^2} + \frac{\sum_{j=s+1}^r k(\widehat{\tilde{p}} - \tilde{p}^*)^2 p_{n,j}}{\psi(c)^2} & \quad (C.7)\\ = \sum_{j=1}^s \frac{k(T_{n,j}/k - p_{n,j})^2}{p_{n,j}}\frac{p_{n,j}^2}{\psi_j(c)^2} + k\frac{(\sum_{j=s+1}^r T_{n,j}/k - p_{n,j})^2}{\sum_{j=s+1}^r p_{n,j}}\frac{\sum_{j=s+1}^r p_{n,j}}{(r-s)^2\psi(c)^2}. & \quad (C.8)\end{aligned}$$

Following Equation (C.4), we know that $\psi_j(c)/p_{n,j}$ and $\sum_{j=s+1}^r p_{n,j}/[(r-s)\psi(c)]$ converge to 1 when $n \to \infty$, and thus Equation (4.15) and Slutsky's theorem yield to the following convergence:

$$(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2}\mathbb{E}[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)]|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*) \xrightarrow{d} \chi^2(s+1), \quad n \to \infty.$$

□

References

Abdous, B. and Ghoudi, K. (2005). Non-parametric estimators of multivariate extreme dependence functions. *Nonparametric Statistics*, 17(8):915–935.

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

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., and Ferro, C. (2006). *Statistics of Extremes: Theory and Applications*. Wiley.

Beirlant, J., Vynckier, P., and Teugels, J. (1996). Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular variation*. Cambridge University Press.

Caeiro, F. and Gomes, M. I. (2015). Threshold selection in extreme value analysis. *Extreme Value Modeling and Risk Analysis: Methods and Applications*, pages 69–86.

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

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

Chiapino, M., Sabourin, A., and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222.

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.

Condat, L. (2016). Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1-2):575–585.

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

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

Devroye, L. (1989). The double kernel method in density estimation. In *Annales de l'IHP Probabilités et statistiques*, volume 25, pages 533–580.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM.

Einmahl, J., de Haan, L., and Huang, X. (1993). Estimating a multidimensional extreme-value distribution. *Journal of Multivariate Analysis*, 47(1):35–47.

Einmahl, J., de Haan, L., and Piterbarg, V. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

Einmahl, J., de Haan, L., and Sinha, A. (1997). Estimating the spectral measure of an extreme value distribution. *Stochastic Processes and their Applications*, 70(2):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(5B):2953–2989.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume II. Wiley, New-York, second edition.

Fougères, A.-L. and Soulier, P. (2010). Limit conditional distributions for bivariate vectors with polar representation. *Stochastic models*, 26(1):54–77.

Gafni, E. M. and Bertsekas, D. P. (1984). Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964.

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

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hille, E. (1964). *Analysis*, volume 1. Blaisdell, New-York.

Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. and Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):475–499.

Lee, J., Fan, Y., and Sisson, S. A. (2015). Bayesian threshold selection for extremal models using measures of surprise. *Computational Statistics & Data Analysis*, 85:84–99.

Lehtomaa, J. and Resnick, S. (2019). Asymptotic independence and support detection techniques for heavy-tailed multivariate data. *arXiv: 1904.00917*.

Liu, J. and Ye, J. (2009). Efficient euclidean projections in linear time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 657–664, New-York.

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

Meyer, N. and Wintenberger, O. (2020). Detection of extremal directions vie Euclidean projections. arXiv: 1907.00686.

Resnick, S. (1986). Point processes, regular variation and weak convergence. *Advances in Applied Probability*, 18(1):66–138.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5(4):303–336.

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.

Rootzén, H., Tajvidi, N., et al. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930.

Sabourin, A. and Drees, H. (2019). Principal component analysis for multivariate extremes. *arXiv:1906.11043*.

Sabourin, A., Naveau, P., and Fougeres, A.-L. (2013). Bayesian model averaging for multivariate extremes. *Extremes*, 16(3):325–350.

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

Sibuya, M. (1960). Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210.

Simpson, E., Wadsworth, J., and Tawn, J. (2019). Determining the dependence structure of multivariate extremes. *arXiv:1809.01606*.

Smith, R. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174–1207.

Stărică, C. (1999). Multivariate extremes for models with constant conditional correlations. *Journal of Empirical Finance*, 6(5):515–553.

Tawn, J. A. (1992). Estimating probabilities of extreme sea-levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):77–93.

Wan, P. and Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166.