Electronic Journal of Statistics
Vol. 9 (2015) 383–418
ISSN: 1935-7524
DOI: 10.1214/15-EJS1002

Dimension reduction in multivariate extreme value analysis

Emilie Chautru

Mines ParisTech, PSL Research University, Centre de géosciences
35 rue St Honoré, 77300 Fontainebleau, France
e-mail: emilie.chautru@mines-paristech.fr

Abstract: Non-parametric assessment of extreme dependence structures between an arbitrary number of variables, though quite well-established in dimension 2 and recently extended to moderate dimensions such as 5, still represents a statistical challenge in larger dimensions. Here, we propose a novel approach that combines clustering techniques with angular/spectral measure analysis to find groups of variables (not necessarily disjoint) exhibiting asymptotic dependence, thereby reducing the dimension of the initial problem. A heuristic criterion is proposed to choose the threshold over which it is acceptable to consider observations as extreme and the appropriate number of clusters. When empirically evaluated through numerical experiments, the approach we promote here is found to be very efficient under some regularity constraints, even in dimension 20. For illustration purpose, we also carry out a case study in dietary risk assessment.

MSC 2010 subject classifications: 62H12, 62H30, 62G32.
Keywords and phrases: Angular/spectral measure, dimension reduction, latent variable, mixture model, extreme dependence, multivariate extremes.

Received December 2013.

1. Introduction

High dimension raises important issues in applied multivariate statistics; while sample sizes are finite, the set on which probability measures are defined can be so large that extrapolation is intricate. Referred to as the curse of dimensionality (Donoho, 2000; Massart, 1989), this phenomenon makes the variance of classical estimators explode, thereby impeding inference. In extreme value analysis, the quality of estimation is all the more degraded as it is not carried out on the entire sample, but on some relatively small number of largest observations that are considered representative of the tail of the distribution. Whereas a plethora of techniques has been developed in the field of statistical learning to overcome this issue (Friedman et al., 2009), multivariate extremes in dimensions larger that 2 are still handled with difficulty. It is the main purpose of the present paper to address this issue, by developing a non-parametric technique for identifying groups of variables (not necessarily disjoint) exhibiting asymptotic dependence. Beyond a possible overall description of the tail dependence structure, when these classes are of small dimension, this method would enable further and more efficient assessment of multivariate tails. It combines recent statistical learning

algorithms with multivariate extreme value theory (MEVT). From a practical perspective, it should be also pointed out that it includes a heuristic criterion to help select the sub-sample of extreme observations on which inference should be performed.

From a theoretical perspective, non-parametric assessment of multivariate extreme dependencies is already well documented. Under the mild assumption that there exists a tail dependence function, focus is usually on an *angular measure*, often referred to as the *spectral measure*, which characterizes extreme dependencies. In the bivariate setting, many estimators were proposed to assess this angular measure (Beirlant et al., 2004; Einmahl and Segers, 2009; Einmahl et al., 2001; Resnick, 2007). Bayesian models have also flourished (Boldi and Davison, 2007; Guillotte et al., 2011), in which vein Sabourin and Naveau (2014) recently proposed a novel algorithm that handles moderate dimensions. Unfortunately, their technique is only efficient when all variables considered are asymptotically dependent; higher-complexity angular measures may not be studied with their method. Hence, were we able to first identify groups of dependent variables in regard to their extreme behavior, the aforementioned estimators would enable more precise estimation up to dimension 5. Lately, Haug et al. (2009) have adapted a classical dimension reduction method, Principal Components Analysis (PCA), to multivariate extremes analysis. Under an elliptical copula assumption, they recover the set of straight lines summarizing best the extreme covariance function, thereby leading to a clustering of variables based on extreme dependence. Following in their footsteps, we propose to borrow algorithms from statistical learning to achieve dimension reduction, without making any parametric assumption in contrast. Our goal is to identify and interpret hopefully small groups of asymptotically dependent variables, possibly overlapping. For this, we exhibit a natural mixture model of the angular measure that corresponds to a partition of the space it lives in: the simplex. Useful properties arising from this setting are revealed and exploited. Inference aims at recovering the components of the mixture, which define all groups of variables exhibiting asymptotic dependence. Mimicking classical non-parametric angular measure estimation, it focuses on the cloud of observation angles related to the L_2-norm. In a first step, they are projected on a space with drastically lower dimension by using a recent algorithm that adapts PCA to Riemannian manifolds (Jung et al., 2012). Then, identification of the groups of interest is subsequently achieved by implementing an appropriate clustering technique on the obtained sub-space (Dhillon et al., 2002). To illustrate the assets and liabilities of our method, we perform numerical experiments and conduct a real case study for *dietary risk assessment*.

The paper is organized as follows: we start off in Section 2 by recalling a few basic notions in angular measure analysis and introducing the main hypotheses, subsequently used throughout the methodological part of our work in Section 3. There, we introduce a mixture model for the angular probability measure and emphasize the ensuing fruitful properties it enjoys, when viewed as a latent variable model. Then we turn to the practical aspects of the approach we promote and depict our strategy for statistical inference under the assumed model, based

on dimension reduction techniques, in Section 4. It is supported by numerical experiments carried through in Section 5, and subsequently applied for illustration purposes to dietary risk assessment in Section 6. In view of both simulation and case study results, assets, liabilities and natural extensions of our method are finally listed and discussed in Section 7.

2. Angular measures

Throughout this article, we consider $\mathbf{X} := (X_1, \ldots, X_d)$ a d-dimensional random vector, $d \geqslant 2$, with Lebesgue-dominated probability distribution $\mathbb{P}$ on the positive orthant $\mathcal{C} := [0, +\infty]^d$ and cumulative distribution function (cdf) F, whose tail structure we wish to assess. For all j in $\{1, \ldots, d\}$, we denote by $\mathbb{P}_j$ the j-th 1-dimensional marginal distribution of $\mathbb{P}$, *i.e.* the probability distribution of X_j, with corresponding *continuous* cdf $F_j(x) := \mathbb{P}_j([0,\ x])$, $x \geqslant 0$. Statistical inference on the extreme behavior of F will be based on the observation of a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $n > 1$ independent copies of $\mathbf{X}$ (we shall write $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$ for $1 \leqslant i \leqslant n$). Also define the random vector $\mathbf{Z} := (Z_1, \ldots, Z_d)$ of standardized components

$$Z_j := \frac{1}{1 - F_j(X_j)},\ j \in \{1, \ldots, d\}, \tag{1}$$

and the corresponding transformed sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. All Z_j, $j \in \{1, \ldots, d\}$, are standard Pareto distributed, *i.e.* for all $x \geqslant 1$ we have $\mathbb{P}(Z_j > x) = x^{-1}$.

It is customary in MEVT to assume that the cdf F of $\mathbf{X}$ is in the maximum domain of attraction (MDA) of a multivariate extreme value distribution (EVD) G, *i.e.* for all $j \in \{1, \ldots, d\}$ there exists sequences $a_{n,j} > 0$ and $b_{n,j} \in \mathbb{R}$ such that

$$\mathbb{P}\left(\frac{\max\limits_{1 \leqslant i \leqslant n} X_{i,1} - b_{n,1}}{a_{n,1}} \leqslant x_1, \ldots, \frac{\max\limits_{1 \leqslant i \leqslant n} X_{i,d} - b_{n,d}}{a_{n,d}} \leqslant x_d \right) \underset{n \to +\infty}{\longrightarrow} G(\mathbf{x}) \tag{2}$$

for all continuity points $\mathbf{x} := (x_1, \ldots, x_d) \in \mathbb{R}^d$ of G, where the d marginals of G are univariate extreme value distributions (*cf. e.g.* Beirlant et al., 2004, Section 2.1). Here, it is assumed that there exists a Radon measure μ called *exponent measure*, not identically zero and not degenerate at a point, concentrated on the blunt convex cone $\mathcal{C}_\star := [0, +\infty]^d \backslash \{\mathbf{0}\}$ such that

$$t\, \mathbb{P}\left(\frac{\mathbf{Z}}{t} \in . \right) \underset{t \to +\infty}{\overset{v}{\longrightarrow}} \mu(.). \tag{3}$$

In this equation, $\mathbf{0}$ denotes the null vector in $\mathbb{R}^d$ and the notation "$\overset{v}{\longrightarrow}$" stands for the vague convergence of measures in $\mathcal{C}_\star$: for all continuous functions with compact support $f : \mathcal{C}_\star \to \mathbb{R}_+$,

$$t\, \mathbb{E}\left(f\left(\frac{\mathbf{Z}}{t} \right) \right) \underset{t \to +\infty}{\longrightarrow} \int_{\mathcal{C}_\star} f\, d\mu.$$

The distribution of a random vector $\mathbf{Z}$ that fulfills Eq. (3) is said to be *regularly varying* in the multivariate sense. Actually, when combined with the assumption that for all $j \in \{1, \ldots, d\}$ the marginal F_j is in the maximum domain of attraction of a univariate EVD G_j, Eq. (3) implies Eq. (2). Here we do not require the existence of univariate maximum domains of attraction.

The exponent measure exhaustively describes the extreme dependence structure between the random variables $X_1, \ldots, X_d$, see for instance Section 8.2.3 in Beirlant et al. (2004) or Section 6.5.6 in Resnick (2007). It is homogeneous, *i.e.*

$$\text{for all } 0 < s < +\infty \text{ and Borel subsets } B \text{ of } \mathcal{C}_\star : \quad \mu(s\,B) = s^{-1}\,\mu(B), \tag{4}$$

and fulfills d marginal constraints expressing the nature of the marginal survival functions, namely

$$\text{for all } j = 1, \ldots, d \text{ and } 0 < z < +\infty : \quad \mu\left(\{\mathbf{x} \in \mathcal{C}_\star : x_j > z\}\right) = z^{-1}, \tag{5}$$

see for instance Section 8.2.2 in Beirlant et al. (2004) and Section 6.1.4 in Resnick (2007). When switching to pseudo-polar coordinates, μ enjoys a particularly useful representation. Choose two norms $\|.\|_{(1)}$ and $\|.\|_{(2)}$ on $\mathbb{R}^d$ with corresponding unit hyperspheres $\mathbb{S}_{(1)}$ and $\mathbb{S}_{(2)}$ and define the following mapping:

$$T : \begin{pmatrix} \mathcal{C}_\star & \longrightarrow & (0, +\infty] \times \mathbb{S}_{(2)} \\ \mathbf{x} & \longmapsto & (\rho, \boldsymbol{\omega}) = \left(\|\mathbf{x}\|_{(1)},\, \mathbf{x}/\|\mathbf{x}\|_{(2)}\right) \end{pmatrix},$$

with $T^{-1}(\rho, \boldsymbol{\omega}) = \rho\,\boldsymbol{\omega}/\|\boldsymbol{\omega}\|_{(1)} = \mathbf{x}$. Typical choices of norms include the L_p-norm or the sup-norm L_∞. Then, the homogeneity property stated in Eq. (4) implies

$$\mu \circ T^{-1} := \mu_{-1} \otimes S,$$

where the radius measure μ_{-1}, defined on $(0, +\infty]$, is such that for all $x > 0$, $\mu_{-1}((x, +\infty]) = x^{-1}$, and the angle measure S, referred to as the *angular* (or *spectral*) *measure*, has support on $\Omega := \mathbb{S}_{(2)} \cap \mathcal{C}_\star$ and satisfies

$$S(B) = \mu\left(\{\mathbf{x} \in \mathcal{C}_\star : \|\mathbf{x}\|_{(1)} \geqslant 1,\ \mathbf{x}/\|\mathbf{x}\|_{(2)} \in B\}\right) \tag{6}$$

for all Borel subsets B of Ω. A simple normalization of S yields the so-termed *angular probability measure* Q on Ω,

$$Q := \frac{S}{S(\Omega)}. \tag{7}$$

Set $\boldsymbol{\omega} = \mathbf{Z}/\|\mathbf{Z}\|_{(2)}$ and $\rho = \|\mathbf{Z}\|_{(1)}$, then Equations (3), (6) and (7) imply

$$t\,\mathbb{P}\left(\boldsymbol{\omega} \in .,\ \rho \geqslant t\right) \xrightarrow[t \to +\infty]{v} S(.), \tag{8}$$

$$\mathbb{P}\left(\boldsymbol{\omega} \in . \,\middle|\, \rho \geqslant t\right) \xrightarrow[t \to +\infty]{\mathcal{D}} Q(.), \tag{9}$$

where "$\xrightarrow{\mathcal{D}}$" stands for the convergence in distribution. In words, Q is the limit distribution of the angles when the radius gets infinitely large. It thereby

encapsulates the extreme (or asymptotic) dependence structure between the d variables in dimension $d-1$. Observe that Eq. (5) can be expressed in terms of moment constraints for S and Q respectively: for all j in $\{1,\ldots,d\}$,

$$\int_\Omega \frac{\omega_j}{\|\boldsymbol{\omega}\|_{(1)}}\, S(d\boldsymbol{\omega}) = 1 \qquad \text{and} \qquad \int_\Omega \frac{\omega_j}{\|\boldsymbol{\omega}\|_{(1)}}\, Q(d\boldsymbol{\omega}) = \frac{1}{S(\Omega)}. \tag{10}$$

3. Mixture model of the angular probability measure

The extreme dependence structure between the variables $X_1,\ldots,X_d$ can be expressed in terms of the geometry of the support of Q (or S), which we denote by $supp(Q)$. Indeed, recall that $supp(Q)$ is included in Ω, the positive orthant of the unit hypersphere $\mathbb{S}_{(2)}$, or the *simplex* associated with $\|.\|_{(2)}$. The latter can be partitioned into 2^d-1 non-empty and disjoint *open faces* with dimensions ranging from 0 up to $d-1$. They are identified by the collections of indexes corresponding to the non-empty subsets of $\{1,\ldots,d\}$. Let $\mathcal{P}_d$ denote the power set of $\{1,\ldots,d\}$ and $\mathcal{P}_d^\star := \mathcal{P}_d\backslash\{\varnothing\}$. For any $h \in \mathcal{P}_d^\star$, the open face identified by the set of indexes h is written

$$\Omega_h := \{\boldsymbol{\omega} \in \Omega : \boldsymbol{\omega}_j > 0 \ \forall\, j \in h,\, \boldsymbol{\omega}_j = 0 \ \forall\, j \notin h\}\,.$$

It is of dimension $\#h-1$, where $\#h$ denotes the cardinal of h. See Figure 1 for an illustration in dimension 3. By convention, the set referring to the empty face is denoted by $\Omega_\varnothing := \varnothing$.

Given this decomposition, for any $h \in \mathcal{P}_d^\star$, $Q(\Omega_h) \neq 0$ *implies* that all X_j such that $j \in h$ exhibit *asymptotic dependence* (see Beirlant et al., 2004, Section 8.2.3). Observe that the converse is not necessarily true: $Q(\Omega_h) = 0$ does *not* imply the asymptotic independence of $\{X_j : j \in h\}$. For instance, take X_1, X_2, X_3

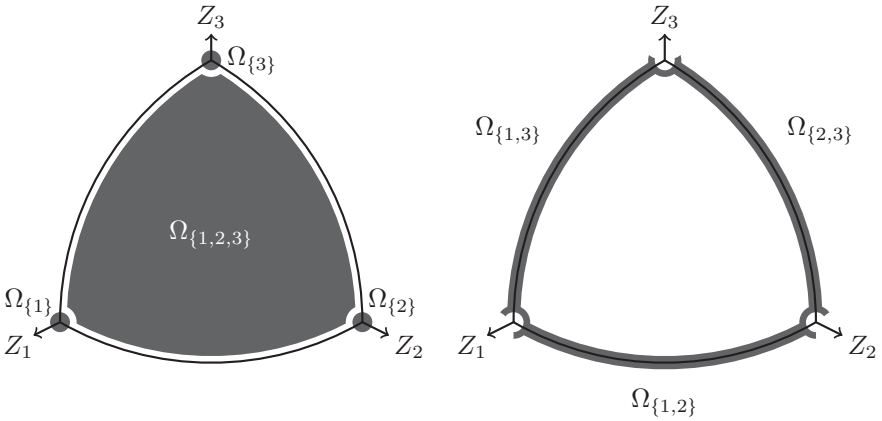

FIG 1. *The* 7 *nonempty open faces in the* L_2*-norm simplex* Ω *in* $\mathbb{R}^3$*: the* 3 *vertices (left), the* 3 *edges (right), and the interior (left).*

three asymptotically dependent variables. Then Q has no mass on $\Omega_{\{1,2\}}$ (only on $\Omega_{\{1,2,3\}}$) even though X_1 and X_2 are asymptotically dependent. Therefore, recovering the set of faces that intersect with the support of Q suffices to identify the sets of variables that are dependent in the extremes, but only then is it possible to deduce which ones are not. This motivates the following *mixture model*:

$$Q(.) = \sum_{h \in \mathcal{P}_d} \pi_h \, Q_h(.), \tag{11}$$

where for all $h \in \mathcal{P}_d$, $\pi_h := Q(\Omega_h)$ and

$$Q_h(.) := \begin{cases} Q(. \cap \Omega_h)/Q(\Omega_h) & \text{if } Q(\Omega_h) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

In addition, denote by $\mathcal{H}$ the set made of all open faces intersecting $supp(Q)$:

$$\mathcal{H} := \{h \in \mathcal{P}_d \, : \, Q(\Omega_h) \neq 0\} \, .$$

Then for all $h \in \mathcal{H}$, Q_h is by definition a probability distribution on Ω_h. Even so, it cannot be interpreted directly as the angular measure of $\{Z_j : j \in h\}$ in the sense that it does not fulfill the marginal condition stated in Eq. (10) (see Section A.1, p.412). Obviously, we have $\pi_h \in [0,1]$ for all $h \in \mathcal{P}_d^\star$, $\pi_\varnothing = 0$ and $\sum_{h \in \mathcal{P}_d} \pi_h = \sum_{h \in \mathcal{H}} \pi_h = 1$.

Following in the footsteps of standard mixture model analysis (see for instance McLachlan and Peel, 2000), we shall assume that there exists an intrinsic (unknown) clustering of the data into $\#\mathcal{H}$ classes leading to an identification of the set of interest $\mathcal{H}$.

Assumption 3.1. There exists a random vector of indicators $\boldsymbol{\lambda} \in \{0,1\}^{2^d}$ that has Categorical distribution with parameters $(p_h)_{h \in \mathcal{P}_d} \in [0,1]^{2^d}$, $\sum_{h \in \mathcal{P}_d} p_h = 1$, such that:

(*i*) for all $h \in \mathcal{P}_d$, $(\pi_h \neq 0) \Rightarrow (p_h \neq 0)$,
(*ii*) for all $h \in \mathcal{P}_d$, Q_h is the angular distribution of $\mathbf{Z} | \lambda_h = 1$, *i.e.*

$$\mathbb{P}\left(\boldsymbol{\omega} \in . \, \middle| \, \rho \geqslant t, \, \lambda_h = 1\right) \xrightarrow[t \to +\infty]{\mathcal{D}} \begin{cases} Q_h(.) & \text{if } h \in \mathcal{H}, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

(*iii*) $\forall \; h \in \mathcal{P}_d^\star, \; (\pi_h = 0) \Rightarrow (p_h = 0)$.

We conjecture that in the present setting, Eq. (3) is a sufficient condition for Assumption 3.1 to be fulfilled.

Remark 3.1. It is always possible to construct a categorical vector $\boldsymbol{\lambda} \in \{0,1\}^{2^d}$ that meets the requirements (*i*)–(*iii*) of Assumption 3.1 from a categorical vector $\widetilde{\boldsymbol{\lambda}} \in \{0,1\}^{2^d}$ that only fulfills conditions (*i*) and (*ii*). Indeed, let $(\widetilde{p}_h)_{h \in \mathcal{P}_d}$ be the parameters of the distribution of such a random vector $\widetilde{\boldsymbol{\lambda}}$. Consider the sets

$\mathcal{A}_{0,0} := \{h \in \mathcal{P}_d : \pi_h = 0, p_h = 0\}$, $\mathcal{A}_{\star,\star} := \{h \in \mathcal{P}_d : \pi_h \neq 0, p_h \neq 0\}$ and $\mathcal{A}_{0,\star} := \{h \in \mathcal{P}_d : \pi_h = 0, p_h \neq 0\}$ and define $\boldsymbol{\lambda} \in \{0,1\}^{2^d}$ with components

$$\lambda_h := \begin{cases} \widetilde{\lambda}_h & \text{if } h \in \mathcal{A}_{0,0} \cup \mathcal{A}_{\star,\star}, \\ 0 & \text{if } h \in \mathcal{A}_{0,\star} \backslash \{\varnothing\}, \\ \mathbb{I}\left\{ \sum_{\ell \in \mathcal{A}_{0,\star}} \widetilde{\lambda}_\ell \geqslant 1 \right\} & \text{if } h = \varnothing, \end{cases}$$

where $\mathbb{I}\{.\}$ is the indicator function. Then $\boldsymbol{\lambda}$ has Categorical distribution with parameters $(p_h)_{h \in \mathcal{P}_d}$ defined for all $h \in \mathcal{P}_d$ as

$$p_h = \begin{cases} \widetilde{p}_h & \text{if } h \in \mathcal{A}_{0,0} \cup \mathcal{A}_{\star,\star}, \\ 0 & \text{if } h \in \mathcal{A}_{0,\star} \backslash \{\varnothing\}, \\ \sum_{\ell \in \mathcal{A}_{0,\star}} \widetilde{p}_\ell & \text{if } h = \varnothing, \end{cases}$$

and it fulfills all three points of Assumption 3.1.

Some useful properties can be established in such a setting. We display here two results which are subsequently exploited for inference, as shall be seen in the next section. Proofs and technical details are deferred to the appendix, in Sections A.1 and A.2 respectively. The proposition below exhibits the asymptotic behavior of conditional marginals under the latent variable model.

Proposition 3.1. *We place ourselves in the framework of Section 3 and denote by $\mathcal{H}(j)$ the set $\{h \in \mathcal{H} : j \in h\}$. Then, for all $j \in \{1, \ldots, d\}$, $h \in \mathcal{P}_d$, $x \geqslant 1$,*

$$t\,\mathbb{P}\left(Z_j > x\,t \,\middle|\, \lambda_h = 1\right) \underset{t \to +\infty}{\longrightarrow} c_{j,h}\, x^{-1}, \tag{13}$$

where $\sum_{h \in \mathcal{P}_d} p_h\, c_{j,h} = 1$ and $c_{j,h} \in [0, 1/p_h]$ is non-null if and only if $h \in \mathcal{H}(j)$.

Therefore, nonempty open faces intersecting $supp(Q)$ are identifiable by remaining only in the univariate level. In particular, the following result reveals that there exists a function, the asymptotic behavior of which enables the characterization of $\{\mathcal{H}(j),\ 1 \leqslant j \leqslant d\}$, and by extension of $\mathcal{H}$.

Proposition 3.2. *We place ourselves in the framework of Proposition 3.1.*
For all $j \in \{1, \ldots, d\}$, $h \in \mathcal{P}_d$, $x \geqslant 1$, define the functional

$$\kappa_{j,h}(t) := \int_1^{+\infty} t\,\mathbb{P}\left(\rho \geqslant t\right)\,\mathbb{P}\left(Z_j > x\,t \,\middle|\, \rho \geqslant t,\ \lambda_h = 1\right)\,dx,$$

and assume that there exists some constants $\gamma^\star \in (0,1)$, $c^\star \geqslant 0$ and $t^\star > 1$, such that for all $j \in \{1, \ldots, d\}$, $h \notin \mathcal{H}(j)$,

$$\forall\, x > 1, \quad (t > t^\star) \Rightarrow \left(\frac{\mathbb{P}\left(Z_j > x\,t \,\middle|\, \lambda_h = 1\right)}{\mathbb{P}\left(Z_j > t \,\middle|\, \lambda_h = 1\right)} \leqslant c^\star\, x^{-1/\gamma^\star} \right).$$

Then

$$\kappa_{j,h}(t) \underset{t\to+\infty}{\longrightarrow} \begin{cases} +\infty & \text{if } h \in \mathcal{H}(j) \\ 0 & \text{if } h \notin \mathcal{H}(j) \cup \{\varnothing\} \end{cases} \quad \text{and} \quad \forall\, t > t^\star,\ \kappa_{j,\varnothing}(t) < +\infty. \quad (14)$$

As a consequence, for fixed dimension $j \in \{1, \ldots, d\}$, the set $\mathcal{H}(j)$ consists of all sets of indexes h such that $\kappa_{j,h}(t)$ diverges as t tends to infinity, instead of converging towards a finite constant, possibly zero. Since for all $h \in \mathcal{H}$ we can write $h = \{j : h \in \mathcal{H}(j)\}$, we have that $\mathcal{H}$ is the set of all $h \in \mathcal{P}_d^\star$ such that there exists at least one index $j \in \{1, \ldots, d\}$ for which $\kappa_{j,h}(t) \to +\infty$ as $t \to +\infty$.

Remark 3.2. The assumption in Proposition 3.2 simply requires that the extreme dependence structure is reached at a reasonably fast rate. It can be directly linked to the concept of hidden regular variation introduced in Das and Resnick (2011); Das et al. (2013); Heffernan and Resnick (2005); Resnick (2002, 2007, 2008). A function $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index α if for all $x > 0$, $f(xt)/f(t) \to x^\alpha$ as $t \to +\infty$. Roughly speaking, if the distribution of $\mathbf{Z}$ had hidden regular variation, there would be an angular measure on $\Omega \backslash \bigcup_{h\in\mathcal{H}} \Omega_h$ when making the radius increase with some regularly varying function $b(t) = o(t)$ with index $1/\alpha \leqslant 1$ instead of t. In that case, our assumption guarantees that $1/\alpha \leqslant \gamma^\star < 1$, which is a rational condition for hidden regular variation not to be mistaken for multivariate regular variation in practice.

The proposed approach to statistical inference is based on Proposition 3.2 as explained in the next section. Numerical experiments illustrating the relevance of the method we promote here are subsequently presented in Section 5.

4. Statistical inference

Relying on the probabilistic framework detailed in Sections 2 and 3, we now review the various steps of the proposed methodology to assess the dependence structure governing the extreme values of $X_1, \ldots, X_d$. The ensuing algorithm, which combines techniques borrowed from multivariate extreme value theory with clustering procedures, is depicted step by step in the next paragraphs.

Just as in classical angular measure assessment, we consider that for some high enough threshold t, asymptotic relations such as in Equations (8), (9), (13) and (14) are sufficiently well approached to enable estimation. In keeping with the literature, we use $t = n/k$, where k represents a number of upper radii. Asymptotic statistical results in EVT are typically obtained under the assumption that $k = k_n$ is an intermediate sequence such that $k_n \uparrow +\infty$ and $k_n/n \to 0$ as $n \to +\infty$ (see for instance Beirlant et al., 2004, Resnick, 2007, or De Haan and Ferreira, 2006). For fixed n, this suggests to pick a number k large enough to get a reasonable variance but also small enough to avoid biasing the estimation with non-tail observations. Then, the analysis of extremes is carried on the set of most extreme observations $\aleph_k := \{i \in \{1, \ldots, n\} : \hat{\rho}_i \geqslant n/k\}$, with cardinal $\#\aleph_k =: n_k$.

4.1. Estimation of the marginals

From the beginning, we have worked with the standardized vector $\mathbf{Z}$ defined in Eq. (1) instead of the vector of interest $\mathbf{X}$. In practice $\mathbf{Z}$ cannot be computed directly since it depends on the unknown marginals $F_1, \ldots, F_d$; it has to be estimated. To avoid restrictive hypotheses, we privilege here a non-parametric procedure, usually referred to as the *rank transform*: for all units $i \in \{1, \ldots, n\}$ and dimensions $j \in \{1, \ldots, d\}$, set

$$\widehat{F}_j(X_{i,j}) = \frac{1}{n} \sum_{\ell=1}^{n} \mathbb{I}\{X_{\ell,j} < X_{i,j}\},$$

and pursue the analysis with $\widehat{Z}_{i,j} = 1/(1 - \widehat{F}_j(X_{i,j}))$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant d$ (Beirlant et al., 2004; Einmahl and Segers, 2009; Einmahl et al., 2001; Resnick, 2007). Angles and radii are subsequently denoted by $\hat{\boldsymbol{\omega}}_i$ and $\hat{\rho}_i$ respectively. For geometrical reasons explained in the next subsection, we set $\|.\|_{(2)}$ as the L_2-norm. In addition, we use the L_∞-norm for $\|.\|_{(1)}$. Observe that whereas it is unimportant regarding the angle, selecting a specific norm for the radius can have major implications. This is due to the selection process of tail observations, defined as those with radius larger than n/k; clearly, different norms are bound to produce different sub-samples (Einmahl and Segers, 2009). However, such issues go beyond the scope of our analysis and are not discussed further here.

4.2. Principal Nested Spheres

Because of the curse of dimensionality, clustering in high dimensions can be problematic. Thus, it is customary in statistical learning to start by projecting the data on a manifold of smaller dimension in order to reduce the noise and synthesize the information. Here, we propose to mimic a classical approach in statistical learning, namely Principal Components Analysis (PCA, Friedman et al., 2009). We work on the angles instead of the raw data and set $\|.\|_{(2)}$ as the L_2-norm, with unit hypershpere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. This enables the use of algorithms that respect the intrinsic distance of $\mathbb{S}^{d-1}$, like the Principal Nested Spheres (PNS) technique developed by Jung et al. (2012). It consists of an iterative projection of the data on sub-spheres of smaller and smaller dimension, called PNS, which are then identified with the unit spheres $\mathbb{S}^{d-2}, \ldots, \mathbb{S}^1$. Sub-spheres of the unit circle being points, the last PNS (of dimension 0) corresponds to the Fréchet mean defined in Algorithm 1 (which is not necessarily unique, see Jung et al., 2012, p.555). More formally, let $\ell \in \{1, \ldots, d-1\}$. The geodesic distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ of $\mathbb{S}^\ell$ (the unit sphere in $\mathbb{R}^{\ell+1}$) is written $d_G^\ell(x, y) = \arccos \mathbf{x}'\mathbf{y}$, where $\mathbf{x}'$ stands for the transpose of the vector $\mathbf{x}$. Any $(\ell-1)$-dimensional sub-sphere $A_{\ell-1}$ in $\mathbb{S}^\ell$ is identified by a center $\mathbf{v} \in \mathbb{S}^\ell$ and a radius $r \in (0, \pi/2]$: $A_{\ell-1} := A_{\ell-1}(\mathbf{v}, r) := \left\{\mathbf{x} \in \mathbb{S}^\ell : d_G^\ell(\mathbf{v}, \mathbf{x}) = r\right\}$. For any $\mathbf{v} \in \mathbb{S}^\ell$, we denote by $R(\mathbf{v})$ a $(\ell+1) \times (\ell+1)$ rotation matrix that moves $\mathbf{v}$

Algorithm 1 Principal Nested Spheres

Input: • $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{S}^{d-1}$, $n \geqslant 2$, $d \geqslant 3$ ▷ The data
• $v^\star \in (0,1)$ ▷ A threshold

Procedure PNS$(\mathbf{x}_1, \ldots, \mathbf{x}_n)$

Initialization: $\left(\mathbf{x}_1^{(d-1)}, \ldots, \mathbf{x}_n^{(d-1)}\right) \leftarrow (\mathbf{x}_1, \ldots, \mathbf{x}_n)$

For $\ell \in \{d-1, \ldots, 2\}$ **do**

1: $(\mathbf{v}_\ell, r_\ell) \leftarrow \underset{\substack{\mathbf{v} \in \mathbb{S}^\ell \\ r \in (0, \frac{\pi}{2}]}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(d_G^\ell\left(\mathbf{x}_i^{(\ell)}, \mathbf{v}\right) - r\right)^2$ ▷ Find the best fitting sub-sphere in $\mathbb{S}^\ell$

2: $\forall\, i \in \{1, \ldots, n\} : \xi_i^{(\ell)} \leftarrow \left(\prod_{j=\ell}^{d-1} \sin(r_j)\right)\left(d_G^\ell\left(\mathbf{x}_i^{(\ell)}, \mathbf{v}_\ell\right) - r_\ell\right)$ ▷ Stock the scaled residuals

3: $\forall\, i \in \{1, \ldots, n\} : \tilde{\mathbf{x}}_i^{(\ell)} \leftarrow P\left(\mathbf{x}_i^{(\ell)} \mid \mathbf{v}_\ell, r_\ell\right)$ ▷ Project the data on the sub-sphere

4: $\forall\, i \in \{1, \ldots, n\} : \mathbf{x}_i^{(\ell-1)} \leftarrow f_\ell\left(\tilde{\mathbf{x}}_i^{(\ell)} \mid \mathbf{v}_\ell\right)$ ▷ Identify the sub-sphere with $\mathbb{S}^{\ell-1}$

For $\ell = 1$ **do**

1: $(\mathbf{v}_1, r_1) \leftarrow \left(\underset{\mathbf{v} \in \mathbb{S}^1}{\operatorname{argmin}} \sum_{i=1}^{n} d_G^1\left(\mathbf{x}_i^{(1)}, \mathbf{v}\right)^2, 0\right)$ ▷ Find the Fréchet mean $\mathbf{v}_1$

2: $\forall\, i \in \{1, \ldots, n\} : \xi_i^{(1)} \leftarrow \left(\prod_{j=1}^{d-1} \sin(r_j)\right)\left(d_G^1\left(\mathbf{x}_i^{(1)}, \mathbf{v}_1\right)\right)$ ▷ Stock the scaled residuals

Return $\left(\mathbf{x}_i^{(\ell)}\right)_{\substack{1 \leqslant i \leqslant n \\ 2 \leqslant \ell \leqslant d-1}}, \left(\xi_i^{(\ell)}\right)_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant \ell \leqslant d-1}}$

End procedure

Procedure SELECTPNS(PNS$(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $v^\star$)

Initialization: • $\ell \leftarrow 1$
• $v_0 \leftarrow V(1)$ ▷ Relative variance of the 1st PNS
• $v_1 \leftarrow V(2)$ ▷ Relative variance of the 2nd PNS

While $v_0 \geqslant v^\star$ **and** $v_1 \geqslant v^\star$ **do**

1: $\ell \leftarrow \ell + 1$
2: $v_0 \leftarrow v_1$
3: $v_1 \leftarrow V(\ell)$ ▷ Relative variance of the ℓ-th PNS

Return $\left(\mathbf{x}_i^{(\ell)}\right)_{1 \leqslant i \leqslant n}$ ▷ The data projected on the selected PNS $\mathbb{S}^\ell$

End procedure

to the north pole and by $R^-(\mathbf{v})$ the $\ell \times (\ell+1)$ matrix consisting of the first ℓ rows of $R(\mathbf{v})$ (see Jung et al., 2012, p. 567 for more details on $R(\mathbf{v})$). A sub-sphere $A_{\ell-1} \in \mathbb{S}^\ell$ can be identified with the unit sphere $\mathbb{S}^{\ell-1}$ using the function $f_\ell(. \mid \mathbf{v}, r) : A_{\ell-1}(\mathbf{v}, r) \to \mathbb{S}^{\ell-1}$ such that $f_\ell(\mathbf{x} \mid \mathbf{v}, r) = \sin(r)^{-1} R^-(\mathbf{v})\,\mathbf{x}$ for all $\mathbf{x} \in \mathbb{S}^\ell$. The projection of any point $\mathbf{x} \in \mathbb{S}^\ell \backslash \{-\mathbf{v}, \mathbf{v}\}$ on a sub-sphere $A_{\ell-1}(\mathbf{v}, r)$

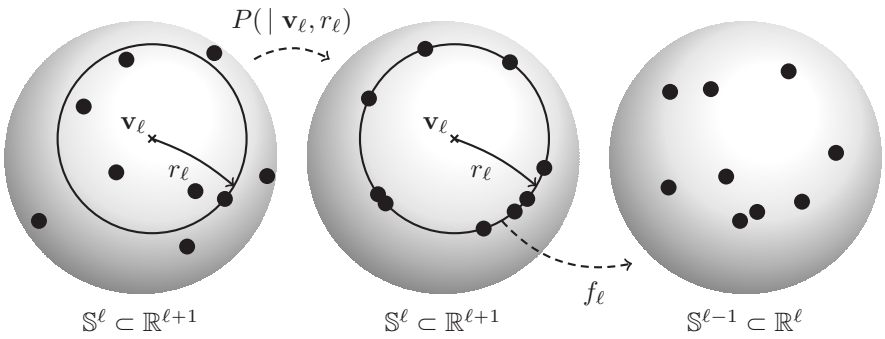

FIG 2. *Schematic illustration of steps 1, 3 and 4 of the first* ***For*** *loop in Algorithm 1 at iteration* $\ell \in \{d-1, \ldots, 2\}$.

is denoted by

$$\tilde{\mathbf{x}} := P(\mathbf{x} \mid \mathbf{v}, r) = \frac{\sin(r)\,\mathbf{x} + \sin(d_G^\ell(\mathbf{x}, \mathbf{v}) - r)\,\mathbf{v}}{\sin(d_G^\ell(\mathbf{x}, \mathbf{v}))}.$$

The procedure PNS() in Algorithm 1 summarizes the main steps of the PNS algorithm of Jung et al. (2012), the code of which was made available by its authors at `http://www.stat.pitt.edu/sungkyu/MiscPage.html`. An illustration of the key steps is given in Figure 2. Practical issues like the possible multiplicity of Fréchet means go beyond the scope of our work and are not discussed here. In the end it is practical to restrict the rest of the analysis to one of the $d-2$ PNS of positive dimension, chosen for instance by the simple rule-of-thumb procedure SELECTPNS() in Algorithm 1. It relies on the calculation of the relative variance encapsulated in each PNS (Jung et al., 2012, Section 2.4): for all $\ell \in \{1, \ldots, d-1\}$,

$$V(\ell) := \frac{\sum_{i=1}^{n} \left(\xi_i^{(\ell)}\right)^2}{\sum_{j=1}^{d-1} \sum_{i=1}^{n} \left(\xi_i^{(j)}\right)^2},$$

where $\xi_i^{(\ell)}$ is the scaled residual of the projection of observation i from $\mathbb{S}^\ell$ onto $\mathbb{S}^{\ell-1}$ defined in Algorithm 1 – scaling the residuals is required to compare geodesic distances in different spaces (Jung et al., 2012, Section 4). In other words, $V(\ell)$ is supposed to give an indication of the level of data variability explained by the ℓ-th PNS. Then, the SELECTPNS() procedure simply consists in recursively checking that $V(\ell)$ is greater than a user defined threshold for $\ell = 1, 2, \ldots$ and select the first PNS after which the condition is no longer satisfied; we stop as soon as the gain in explained variance is considered too low to justify raising the dimension. As was pointed out by Jung et al. (2012), projecting the data onto spheres of very low dimension like $\mathbb{S}^1$ and $\mathbb{S}^2$ usually suffices to explain most of the the variability.

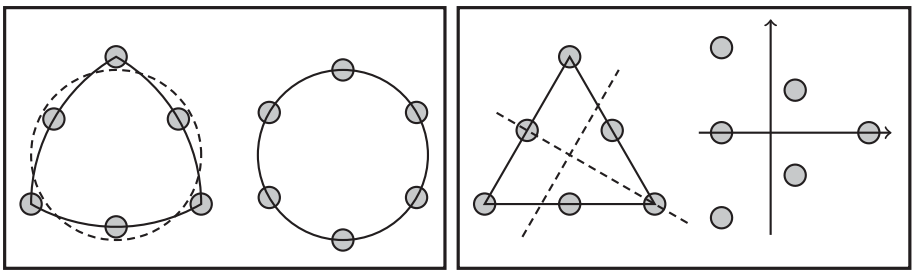

FIG 3. *Example in* $\mathbb{R}^3$ *where the angles are concentrated around the vertices and the middle of the edges: using the* L_2*-norm and applying PNS gives a good representation of the data in dimension* 1 *(left block); using the* L_1*-norm and applying PCA gives a good representation of the data in dimension* 2 *but not in dimension* 1 *(right block).*

Other algorithms could have been used to project the data on manifolds of smaller dimension, see for instance Fletcher et al. (2004); Huckemann and Ziezold (2006); Jung et al. (2011). Had the angles been calculated with the L_1-norm, PCA would have been an acceptable choice too. Our preference for PNS is justified by its ability to synthesize the structure of the data in very small dimensions by taking into account the intrinsic property that the angles are of norm 1. Jung et al. (2012, p. 562–564) observed on real data analyses that when the observations are located on a Riemannian manifold, *"Euclidean principal component analysis, which completely ignores the manifold nature of the data, gives the worst performance"*, whereas *"the principal nested spheres capture more interesting variability in fewer components"*. In the present context, although there is no absolute guarantee that PNS is more suitable than PCA, it is nonetheless possible to find configurations where this postulate is verified, like shown in Figure 3.

4.3. Spherical k-means

The projected data obtained by running Algorithm 1 can be organized into groups that will later serve to assess the set $\mathcal{H}$. For this, we propose to use an accurate clustering procedure such as spherical k-means (Dhillon et al., 2002; Maitra and Ramler, 2010), based again on the geodesic distance, which we describe hereinafter. Choose a number M of clusters and let $m \in \{1, \dots, M\}$ and $\ell \in \{1, \dots, d-1\}$. For any set of $n \geqslant 1$ points $\mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_n)$ on the unit sphere $\mathbb{S}^\ell \subset \mathbb{R}^{\ell+1}$ such that for all $i \in \{1, \dots, n\}$, $\mathbf{x}_i := (x_{i,1}, \dots, x_{i,\ell+1})'$, define the barycenter function

$$B(\mathbf{x}) := \left(\frac{1}{n} \sum_{i=1}^{n} x_{i,1}, \dots, \frac{1}{n} \sum_{i=1}^{n} x_{i,\ell} \right)'$$

and $SB(\mathbf{x}) := B(\mathbf{x})/\|B(\mathbf{x})\|_2$ its projection on $\mathbb{S}^\ell$. Let $I_{i,m}$ be the binary variable that indicates whether observation i belongs to cluster m ($I_{i,m} = 1$) or not

Algorithm 2 Spherical k-means

Input: • $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{S}^\ell$, $\ell \geqslant 1$, $n \geqslant 2$ ▷ The data
• $M \in \{1, \ldots, n\}$ ▷ The number of clusters
• $v_G^\star$ ▷ A user-defined tolerance

Procedure $\textsc{Skm}\big((\mathbf{x}_1, \ldots, \mathbf{x}_n),\, M,\, v_G^\star\big)$

Initialization:

1: $\mathrm{T} \leftarrow 0$
2: $v_0 \leftarrow -\infty$
3: $\mathbf{c}_1^{\{0\}} \leftarrow \underset{\mathbf{c} \in \mathbb{S}^\ell}{\operatorname{argmin}} \sum_{i=1}^{n} d_G^\ell\left(\mathbf{x}_i, \mathbf{c}\right)^2$ ▷ Initial concept vectors
4: $\forall\, m \in \{2, \ldots, M\} : \mathbf{c}_m^{(0)} \leftarrow \underset{\{\mathbf{x}_i, 1 \leqslant i \leqslant n\}}{\operatorname{argmax}} d_G^\ell\left(\mathbf{x}_i, SB\left(\mathbf{c}_1^{(0)}, \ldots, \mathbf{c}_m^{(0)}\right)\right)^2$
5: $\forall\, i \in \{1, \ldots, n\}\ \forall\, m \in \{1, \ldots, M\} : I_{i,m}^{(0)} \leftarrow \underset{1 \leqslant j \leqslant M}{\operatorname{argmin}} d_G^\ell\left(\mathbf{x}_i, \mathbf{c}_j^{(0)}\right)^2$ ▷ Cluster indicators
6: $v_1 \leftarrow V_G^\ell\left(\left(\mathbf{I}_m^{(0)}\right)_{1 \leqslant m \leqslant M}\right)$ ▷ Intra-class geodesic variance

While $v_1 - v_0 \geqslant v_G^\star$ **do**

1: $v_0 \leftarrow v_1$
2: $\forall m \in \{1, \ldots, M\} : \mathbf{c}_m^{(\mathrm{T}+1)} \leftarrow SB\left(\mathbf{x}\,\mathbf{I}_m^{(\mathrm{T})}\right)$ ▷ Concept vectors
3: $\forall\, i \in \{1, \ldots, n\}\ \forall m \in \{1, \ldots, M\} : I_{i,m}^{(\mathrm{T}+1)} \leftarrow \underset{1 \leqslant j \leqslant M}{\operatorname{argmin}} d_G^\ell\left(\mathbf{x}_i, \mathbf{c}_j^{(\mathrm{T}+1)}\right)^2$ ▷ Cluster indicators
4: $v_1 \leftarrow V_G^\ell\left(\left(\mathbf{I}_m^{(\mathrm{T})}\right)_{1 \leqslant m \leqslant M}\right)$ ▷ Intra-class geodesic variance
5: $\mathrm{T} \leftarrow \mathrm{T} + 1$

Return $\left(I_{i,m}^{(\mathrm{T})}\right)_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant m \leqslant M}}$ ▷ The final cluster indicators

End procedure

$(I_{i,m} = 0)$. For all $m \in \{1, \ldots, M\}$, set $\mathbf{x}\,\mathbf{I}_m := (\mathbf{x}_1\, I_{1,m}, \ldots, \mathbf{x}_n\, I_{n,m})$. The normalized barycenter of class m is called the *concept vector* and is denoted by $\mathbf{c}_m := SB(\mathbf{x}\,\mathbf{I}_m)$. The objective of the spherical k-means algorithm is to find the collection of cluster indicators that minimizes the intra-class geodesic variance

$$V_G^\ell\left((\mathbf{I}_m)_{1 \leqslant m \leqslant M}\right) := \sum_{m=1}^{M} \sum_{i=1}^{n} \left(d_G^\ell(\mathbf{x}_i, \mathbf{c}_m)\, I_{i,m}\right)^2.$$

This is achieved in the manner depicted in Algorithm 2, which corresponds to the function `skmeans` with option `start` = "`S`" in the `R` package `skmeans`. The option "`S`" specifies the method used to choose the initial concept vectors (steps 3 and 4 of the initialization of $\textsc{Skm}()$ in Algorithm 2). It produces an initial clustering with centers as scattered as possible. Obviously, many other initialization techniques may have been applied, *e.g.* picking the first concept vectors at random. Such considerations are disregarded here. The main advantage of the

spherical k-means algorithm is that it is very simple to implement. However, it can often happen that it remains stuck at a local minimum of the intra-class geodesic variance function. To counteract this undesirable effect, many refinements have been proposed in the literature (see for instance Dhillon et al., 2002 and the references therein), which are overlooked here.

Obviously, many other natural techniques in mixture models analysis could have been adopted (McLachlan and Peel, 2000); our preference for geometrical methods is based on a strong belief that Riemannian geometry is a key concept for understanding the structure of the angular (probability) measure, as suggested by the encouraging results of the numerical experiments conducted in Section 5.

4.4. Estimation of $\mathcal{H}$

The clustering obtained after the successive application of Algorithms 1 and 2 can now serve to assess $\mathcal{H}$ in the following manner. Let $k \in \{1, \dots, n\}$ and $M \in \{1, \dots, n_k \wedge 2^d - 1\}$, where $n_k \wedge 2^d - 1 := \min(n_k, 2^d - 1)$ (notice that M depends on k). Assume that we have at our disposal a clustering into M groups of the set of most extreme observations $\aleph_k$ identified by the binary variables $I_{i,m}$, $i \in \aleph_k$, $m \in \{1, \dots, M\}$ obtained with Algorithm 2. Now denote by $\mathcal{F}_M$ the set of injective functions

$$f_M : \begin{pmatrix} \{1, \dots, M\} & \longrightarrow & \mathcal{P}_d \\ m & \longmapsto & h \end{pmatrix}$$

with image $Im(f_M)$. Provided that $M = \#\mathcal{H}$, our hope is that there exists f_M in $\mathcal{F}_{\mathcal{M}}$ such that $Im(f_M) = \mathcal{H}$, *i.e.* $\mathcal{H} = \{f_M(1), \dots, f_M(M)\}$. In that case f_M would be a bijection between $\{1, \dots, M\}$ and $\mathcal{H}$ and, by virtue of Proposition 3.2, we would have the M equalities

$$f_M(m) = \left\{ j \in \{1, \dots, d\} : \kappa_{j, f_M(m)} \underset{t \to +\infty}{\longrightarrow} +\infty \right\}, \quad m \in \{1, \dots, M\}.$$

Assuming that such a mapping exists, we propose to assess it with the statistic $\widehat{f}_M$ constructed as follows. Consider that $I_{i,m}$ is an estimator of $\lambda_{i, f_M(m)}$ and for all $m \in \{1, \dots, M\}$, $j \in \{1, \dots, d\}$, set

$$\widehat{\kappa}_{j,m}(k) := \int_1^{+\infty} \frac{1}{k} \frac{n_k}{n_m} \sum_{i \in \aleph_k} \mathbb{I}\left\{ \widehat{Z}_{i,j} > x \frac{n}{k}, \, I_{i,m} = 1 \right\} dx,$$

where $n_m := \sum_{i \in \aleph_k} \mathbb{I}\{\widehat{I}_{i,m} = 1\}$ is the size of cluster m. This statistic, detailed in Section A.3, can be viewed as the empirical counterpart of $\kappa_{j, f_M(m)}(n/k)$. Then, for all $m \in \{1, \dots, M\}$ define

$$\widehat{f}_M(m) := \{ j \in \{1, \dots, d\} \, : \, \widehat{\kappa}_{j,m}(k) \gg 0 \}.$$

To decide which couples (j, m) fulfill the condition $\widehat{\kappa}_{j,m}(k) \gg 0$, we perform a scree test-like analysis (Cattell, 1966) described in Algorithm 3.

Algorithm 3 Estimation of f_M

Input: • $(k, M) \in \{1, \dots, n\} \times \{1, \dots, n_k \wedge 2^d - 1\}$

• $\left(\hat{\mathbf{Z}}_i\right)_{i \in \aleph_k} \in \left(\mathbb{R}^d\right)^{n_k}$ ▷ The standardized data with largest radii

• $(I_{i,m})_{\substack{i \in \aleph_k \\ 1 \leqslant m \leqslant M}} \in \{0,1\}^{n_k \times M}$ ▷ The cluster indicators

Procedure $\textsc{Fm}\Big((k, M), \left(\hat{\mathbf{Z}}_i\right)_{i \in \aleph_k}, (I_{i,m})_{\substack{i \in \aleph_k \\ 1 \leqslant m \leqslant M}}\Big)$

$\left(\kappa^1, \dots, \kappa^{M \times d}\right) \leftarrow \left(\hat{\kappa}_{1,1}(k), \dots, \hat{\kappa}_{d,M}(k)\right)$ ▷ Calculate all $(\hat{\kappa}_{j,m}(k))_{\substack{1 \leqslant j \leqslant d \\ 1 \leqslant m \leqslant M}}$

$\left(\kappa^{(1)}, \dots, \kappa^{(M \times d)}\right) \leftarrow \mathbf{sort}\left(\hat{\kappa}^1, \dots, \hat{\kappa}^{M \times d}\right)$ ▷ Sort the sequence

$\mathrm{T}^\star \leftarrow \underset{1 \leqslant \mathrm{T} \leqslant M \times d}{\operatorname{argmax}} \kappa^{(\mathrm{T}+1)} - \kappa^{(\mathrm{T})}$ ▷ Find the biggest jump

$\forall\, m \in \{1, \dots, M\} : \hat{f}_M(m) \leftarrow \left\{ j \in \{1, \dots, d\} : \hat{\kappa}_{j,m}(k) > \kappa^{(\mathrm{T}^\star)} \right\}$ ▷ Assume $\hat{\kappa}_{j,m}(k) \gg 0$ if $\hat{\kappa}_{j,m}(k) > \kappa^{(\mathrm{T}^\star)}$

Return $\left(\hat{f}_M(m)\right)_{1 \leqslant m \leqslant M}$

End procedure

This procedure only makes sense if $M = \#\mathcal{H}$ and k is large enough to assume that $\hat{\kappa}_{j,m}(k)$ approximates the limit of $\kappa_{j,h}(t)$ as $t \to +\infty$. Since in practice $\#\mathcal{H}$ is usually unknown and k has to be picked at hand, we develop a heuristic criterion to measure the quality of a clustering, given some $k \in \{1, \dots, n\}$ and $M \in \{1, \dots, n_k \wedge 2^d - 1\}$. Set $\widehat{\mathcal{H}}_M(j) := \left\{ \hat{f}_M(m) \,:\, j \in \hat{f}_M(m),\, m \in \{1, \dots, M\} \right\}$ and consider the statistic

$$\Upsilon(k, M) = \sum_{j=1}^{d} \sum_{m=1}^{M} (-1)^{\mathbb{I}\left\{ \hat{f}_M(m) \notin \widehat{\mathcal{H}}_M(j) \right\}} \, \hat{\kappa}_{j,m}(k).$$

It is built from Eq. (14): after having computed $\hat{\kappa}_{j,m}(k)$ on all $m \in \{1, \dots, M\}$, we add up all quantities corresponding to $\hat{f}_M(m) \in \widehat{\mathcal{H}}_M(j)$ (which should be large) and subtract the others (supposedly close to zero). When (k, M) provides an accurate clustering of the data, $\Upsilon(k, M)$ is expected to reach high values. To avoid possible practical errors, we further refine this criterion with some additional constraints. Specifically, classes should contain more than 1 individual, groups should each identify a different open face and no set $\widehat{\mathcal{H}}_M(j)$, $1 \leqslant j \leqslant d$, should be empty. Observe that while the first two conditions are just common sense, the last one is necessary to guarantee that the marginal distributions of the standardized data are standard Pareto. Finally, we retain the partition inherited from $(k^\star, M^\star) := \operatorname{argmax}_{(k,M)} \widetilde{\Upsilon}(k, M)$, where

$$\widetilde{\Upsilon}(k, M) := \Upsilon(k, M) \times \prod_{m=1}^{M} \mathbb{I}\{n_m > 1\} \times \prod_{j=1}^{d} \mathbb{I}\left\{ \widehat{\mathcal{H}}_M(j) \neq \varnothing \right\} \times \prod_{1 \leqslant m \neq m' \leqslant M} \mathbb{I}\left\{ \hat{f}_M(m) \neq \hat{f}_M(m') \right\}. \tag{15}$$

Observe that the condition $\mathbb{I}\{n_m > 1\}$ makes it impossible for our procedure to work if $\#\mathcal{H}$ exceeds $2 \times n_k$. This is not surprising: complex underlying models demand large datasets to be assessed. Finding the maximum in Eq. (15) requires the computation of $\Upsilon(k, M)$ for all (k, M) in $\{1, \ldots, n\} \times \{1, \ldots, n_k \wedge 2^d - 1\}$, which can be time-consuming from an algorithmic point of view. As a result, we recommend to restrict the analysis to k comprised between 10 and $\lfloor n \times 0.3 \rfloor$, where $\lfloor . \rfloor$ denotes the floor function. In addition, to reduce the number of calculations required to choose M given some fixed k, we propose to iteratively compute $\widetilde{\Upsilon}(k, M)$ for $M = 1, 2, \ldots$ and stop as soon as the next 5 iterations cease improving it (steps 3 to 19 of the **For** loop in Algorithm 4). The complete procedure involving the estimation of the marginals, the PNS, the spherical k-means and the estimation of $\mathcal{H}$ is synthesized in Algorithm 4.

On account of the nice properties of the rank transform (Das and Resnick, 2011; Heffernan and Resnick, 2005; Resnick, 2007), we can reasonably hope that these statistical objects converge to the true quantities they approximate as $n \to +\infty$, provided at least that $k^\star$ fulfills the usual conditions $k^\star = k_n^\star \uparrow +\infty$ and $k_n^\star / n \to 0$ as $n \to +\infty$. Unfortunately, due to the lack of probabilistic results on PNS and spherical k-means, which were originally introduced as geometrical techniques, we cannot provide here a thorough asymptotic analysis of the solution output by the statistical procedure described above. Nonetheless, as shall be seen in the next section, numerical experiments provide strong empirical evidence of the efficiency of the approach we propose.

5. Numerical experiments

We tested our method through a number of numerical experiments, for various values of n, d, and $\#\mathcal{H}$. In doing so, we tried to handle various types of extreme dependence structures, to illustrate the impact of the complexity of $supp(Q)$ on our algorithm. In the next two subsections, we first describe the different scenarios analyzed, then present and comment on the simulation results.

5.1. Settings

We generated n i.i.d. copies of a d-dimensional random vector $(X_1, \ldots, X_d)$ with varying degrees of extreme dependence. Observations were drawn using the general asymmetric multivariate logistic model: for any $\mathbf{x} = (x_1, \ldots, x_d)$ in $\mathbb{R}_+^d$, for all $j \in \{1, \ldots, d\}$ define $y_j(x_j) := (1 + \gamma_j (x_j - b_j)/a_j)^{-1/\gamma_j}$ if $1 + \gamma_j (x_j - b_j)/a_j$ is positive, and set $\#\mathcal{H}, d \geqslant 1$. The data was generated according to the following distribution:

$$\mathbb{P}(X_1 \leqslant x_1, \ldots, X_d \leqslant x_d) = \exp\left\{ - \sum_{h \in \mathcal{H}} \left(\sum_{j \in h} y_j(x_j)^{1/r_h} \right)^{r_h} \right\},$$

where $r_h \in (0, 1]$ is a parameter that controls the strength of the asymptotic dependence within the group of variables indicated by h. In particular, $r_h = 1$

Algorithm 4 Dimension reduction of multivariate extreme values

Input: • $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ ▷ The original dataset
• $v^\star \in (0,1)$ ▷ A threshold for Algorithm 1
• $v_G^\star \in (0,1)$ ▷ A tolerance for Algorithm 2

Procedure DimRedMev $\left((\mathbf{X}_1, \ldots, \mathbf{X}_n),\, v^\star,\, v_G^\star\right)$

Initialization

$\forall\, i \in \{1, \ldots, n\}\ \forall\, j \in \{1, \ldots, d\} : \widehat{Z}_{i,j} \leftarrow \left(1 - \dfrac{1}{n} \displaystyle\sum_{\ell=1}^{n} \mathbb{I}\left\{X_{\ell,j} < X_{i,j}\right\}\right)^{-1}$ ▷ Standardized data

$\forall\, i \in \{1, \ldots, n\} : (\widehat{\boldsymbol{\omega}}_i,\, \widehat{\rho}_i) \leftarrow \left(\widehat{\mathbf{Z}}_i / \|\widehat{\mathbf{Z}}_i\|_2,\, \|\widehat{\mathbf{Z}}_i\|_\infty\right)$ ▷ Pseudo-polar coordinates

$\mathcal{K} \leftarrow \{1, \ldots, \lfloor n \times 0.3 \rfloor\}$; $(M_1, \ldots, M_{\lfloor n \times 0.3 \rfloor}) \leftarrow (1, \ldots, 1)$

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

$\aleph_k \leftarrow \{i \in \{1, \ldots, n\} : \widehat{\rho}_i \geqslant n/k\}$ ▷ Set of observations with largest radii

$\left(\widehat{\boldsymbol{\omega}}_i^\dagger\right)_{i \in \aleph_k} \leftarrow$ SelectPns $\left(\text{Pns}\left((\widehat{\boldsymbol{\omega}}_i)_{i \in \aleph_k}\right),\, v^\star\right)$ ▷ PNS (Algorithm 1)

$M \leftarrow 0$; $(\widetilde{\Upsilon}_0, \ldots, \widetilde{\Upsilon}_5) \leftarrow (0, 1, 0, 0, 0, 0)$; $\text{T} \leftarrow 0$ ▷ Initialize the search for an optimal M

while $\widetilde{\Upsilon}_0 < \max\limits_{\ell \neq 0} \widetilde{\Upsilon}_\ell$ **and** $\widetilde{\Upsilon}_1 \geqslant \min\limits_{\ell \neq 1} \widetilde{\Upsilon}_\ell$ **do**
  $\ell \leftarrow 1$
  while $\ell \leqslant 5$ **do**
    $M \leftarrow M + 1$
    if $\text{T} > 0$ **then**
      $(\widetilde{\Upsilon}_0, \ldots, \widetilde{\Upsilon}_5) \leftarrow (\widetilde{\Upsilon}_1, \ldots, \widetilde{\Upsilon}_5, 0)$; $\ell \leftarrow 5$
    end if
    $(I_{i,m})_{\substack{i \in \aleph_k \\ 1 \leqslant m \leqslant M}} \leftarrow$ Skm $\left(\left(\widehat{\boldsymbol{\omega}}_i^\dagger\right)_{i \in \aleph_k},\, M,\, v_G^\star\right)$ ▷ Spherical k-means (Algorithm 2)
    $\left(\widehat{f}_M(m)\right)_{1 \leqslant m \leqslant M} \leftarrow$ Fm $\left((k, M),\, \left(\widehat{\mathbf{Z}}_i\right)_{i \in \aleph_k},\, (I_{i,m})_{\substack{i \in \aleph_k \\ 1 \leqslant m \leqslant M}}\right)$ ▷ Estimate $\mathcal{H}$ (Algorithm 3)
    $\forall\, j \in \{1, \ldots, d\} : \widehat{\mathcal{H}}_M(j) \leftarrow \left\{\widehat{f}_M(m) \,:\, j \in \widehat{f}_M(m),\, 1 \leqslant m \leqslant M\right\}$ ▷ Estimate all $\widehat{\mathcal{H}}_M(j)$
    $\widetilde{\Upsilon}_\ell \leftarrow \widetilde{\Upsilon}(k, M)$ ▷ Quality of the clustering
    $\ell \leftarrow \ell + 1$
  end while
  $\text{T} \leftarrow \text{T} + 1$
end while

$M_k \leftarrow M - 5$ ▷ Optimal M for fixed k

$(k^\star, M^\star) \leftarrow \underset{k \in \mathcal{K}}{\text{argmax}} \widetilde{\Upsilon}(k, M_k)$ ▷ Optimal couple (k, M)

Return $k^\star,\ \left(\widehat{f}_{M^\star}(m)\right)_{1 \leqslant m \leqslant M^\star}$

End procedure

TABLE 1
List of scenarios considered in our numerical experiments; open faces intersecting with supp(Q) are filled in, with the corresponding extreme dependence coefficient r_h

Scenario 1 $\#\mathcal{H} = 2$ $d = 20$		**Scenario 2** $\#\mathcal{H} = 4$ $d = 20$		**Scenario 3** $\#\mathcal{H} = 4$ $d = 6$	
$h \in \mathcal{H}$	r_h	$h \in \mathcal{H}$	r_h	$h \in \mathcal{H}$	r_h
$\{1, 2, 3\}$	0.1	$\{1, 2\}$	0.1	$\{1, 2\}$	0.1
$\{4, \ldots, 20\}$	0.1	$\{2, 3\}$	0.1	$\{2, 3\}$	0.1
		$\{4\}$	1	$\{4\}$	1
		$\{5, \ldots, 20\}$	0.2	$\{5, 6\}$	0.2

gives asymptotic independence, whereas asymptotic perfect dependence occurs when $r_h \approx 0$. Observe that with such a model, though it was not required in our analysis, the marginal distributions of $X_1, \ldots, X_d$ are extreme values distributions, which naturally belong to their own maximum domain of attraction. Our preference for this model was purely practical: simulations were performed using function `rmvevd` in `R` package `evd` (Stephenson, 2003). We repeated 100 trials of Algorithm 4 under 3 scenarios listed in Table 1. Notice that in accordance with what was pointed out in the introduction and suggested in Sections 3 and 4, in scenarios 2 and 3 some classes overlap ($\{1, 2\} \cap \{2, 3\} = \{2\}$).

To limit computation time, we tested its performance on 5 different sample sizes, namely $n = 500, 1000, 5000, 10\,000$, and 10 thresholds $t = n/k$, with $k = n \times 0.001, n \times 0.002, \ldots, n \times 0.01$. For the same reason, we disregarded situations where $d > 20$. However, in MEVT, $d = 6$ and $d = 20$ can already be considered as high dimensions.

5.2. Results

Results are displayed in Table 2, Table 3 and Table 4. The highlighted row reports the number of trials where we managed to exactly recover the set of open faces intersecting with the support of the angular probability measure. As expected, in all scenarios, results improve when n increases, and success rates become particularly satisfactory as soon as $n \geqslant 5000$, for they then exceed 85% in all 3 scenarios. The best performance is obtained in scenario 1, where $d = 20$ and $\#\mathcal{H} = 2$. Indeed, even with a very small sample ($n = 500$), only 10 trials out of 100 fail to recover the true decomposition of $supp(Q)$, while in scenario 3, where $d = 6$ and $\#\mathcal{H} = 4$, this rate never goes below 12% for any n. This

TABLE 2
Results of our numerical experiments in ***scenario 1****, repeated on* 100 *trials*

		n			
	Estimation of $\mathcal{H}$	500	1000	5000	10 000
Accurate sets	$\{1, 2, 3\}, \{4, \ldots, 20\}$	90	96	100	100
Other inaccurate sets		10	4	0	0

TABLE 3
Results of our numerical experiments in ***scenario 2****, repeated on* 100 *trials*

		n			
	Estimation of $\mathcal{H}$	500	1000	5000	10 000
Accurate sets	$\{1,2\},\{2,3\},\{4\},\{5,\ldots,20\}$	54	72	97	93
Extra sets with cardinal 1	$\{1,2\},\{2,3\},\{4\},\{5,\ldots,20\},\{1\}$	2	3	0	0
	$\{1,2\},\{2,3\},\{4\},\{5,\ldots,20\},\{3\}$	7	2	0	1
	$\{1,2\},\{2,3\},\{4\},\{5,\ldots,20\},\{1\},\{3\}$	1	0	0	0
Missing set $\{1,2\}$ or $\{2,3\}$	$\{1,2\},\{3\},\{4\},\{5,\ldots,20\}$	13	12	0	1
	$\{1\},\{1,2\},\{3\},\{4\},\{5,\ldots,20\}$	1	0	0	0
	$\{1\},\{2,3\},\{3\},\{4\},\{5,\ldots,20\}$	8	5	1	1
	$\{1\},\{2,3\},\{4\},\{5,\ldots,20\}$	0	0	0	0
Other inaccurate sets		14	6	2	4

TABLE 4
Results of our numerical experiments in ***scenario 3****, repeated on* 100 *trials*

		n			
	Estimation of $\mathcal{H}$	500	1000	5000	10 000
Accurate sets	$\{1,2\},\{2,3\},\{4\},\{5,6\}$	39	65	85	88
Extra sets with cardinal 1	$\{1,2\},\{2,3\},\{4\},\{5,6\},\{1\}$	0	1	0	0
	$\{1,2\},\{2,3\},\{4\},\{5,6\},\{3\}$	1	1	0	0
	$\{1,2\},\{2,3\},\{4\},\{5,6\},\{1\},\{3\}$	1	0	0	0
Missing set $\{1,2\}$ or $\{2,3\}$	$\{1,2\},\{3\},\{4\},\{5,6\}$	16	11	6	2
	$\{1\},\{1,2\},\{3\},\{4\},\{5,6\}$	0	0	0	0
	$\{1\},\{2,3\},\{3\},\{4\},\{5,6\}$	23	14	5	6
	$\{1\},\{2,3\},\{4\},\{5,6\}$	0	0	0	1
Other inaccurate sets		20	8	4	3

suggests that rather than the dimension, the complexity of $supp(Q)$ may be one of the principal determinants of the performance of our procedure. Actually, given two angular probability measures with equivalently complex supports, increasing dimensionality can produce better outcomes. This is the case with scenarios 2 and 3, where $supp(Q)$ is contained on small subsets of 4 open faces, but $d = 6$ in the former while $d = 20$ in the latter. These results are not surprising and illustrate a typical phenomenon called the blessing of dimensionality (Donoho, 2000); as d increases, observations occur in relatively small subsets of the original space and are therefore easier to detect and separate. This property is the basis for common techniques in statistical learning, such as the widely celebrated support vector machine (Friedman et al., 2009, Chapter 12), which projects the data onto some space with higher dimension in which they are well divided. In our numerical experiments, switching from scenario 3 to scenario 2 significantly reduces the risk of overriding either $\Omega(\{1,2\})$ or $\Omega(\{2,3\})$, which are very close to one another in the unit hypersphere and may be wrongfully confused during the PNS procedure. Observe nonetheless that these simulations were performed for very small values of parameter r_h, *i.e.* all dependencies were

strong. Since we used the multivariate logistic model, this means that for all $h \in \mathcal{H}$, subsets $supp(Q) \cap \Omega_h$ did not cover the entire open faces Ω_h but were concentrated around small neighborhoods of one of their points. Had we considered less obvious extreme dependencies, these results would have probably been significantly degraded. This remark can be linked to the influence of the hidden angular/spectral measure on inference (Resnick, 2002), which controls the rate at which extreme structure is reached and thus dangerously impacts statistical analysis if the chosen threshold n/k is too small.

In fine, these results are quite encouraging, and underline the usefulness of algorithms from the field of statistical learning for MEVT, provided of course that the underlying model is not too complex.

6. Application to dietary risk assessment

While eating is the privileged way of providing the necessary nutrients for the human organism, it also conveys toxic elements that, due to various environmental causes, contaminate the food. When consumed over certain tolerable doses, called *dietary intake limits* (DIL), these toxic elements can have a non-negligible impact on health. Similar phenomena also occur when diets are either too rich or too poor in nutrients. More importantly, further noxious effects may be caused by possible interactions between elements that are ingested simultaneously (Carpenter et al., 2002). For international institutes concerned about public health issues such as the WHO (World Health Organization), FAO (Food and Agriculture Organization), UNEP (United Nations Environment Program), EFSA (European Food Safety Authority) or for national agencies such as the Anses (the French Agency for Food, Environmental and Occupational Health & Safety), it is then of major interest to identify cocktails of food chemicals to which populations are indeed highly exposed. Extreme value theory has already proven useful to assess the probability of getting over a single dietary intake limit, in both univariate (Tressou et al., 2004) and bivariate settings (Paulo et al., 2006). Here, we propose to apply Algorithm 4 to examine the relationships between high simultaneous long-term exposure to 6 common nutrients and contaminants, namely iron (Fe), calcium (Ca), sodium (Na), methylmercury (MeHg), cadmium (Cd) and dioxins and dioxin-like polychlorinated biphenyls (PCB-DL). Their long-term toxicity is well-known, see for instance Anses (2011) and Carpenter et al. (2002). Methylmercury, cadmium, and PCB-DL are three contaminants found mainly in seafood products. While cadmium was recognized in 2004 as a type 2 carcinogen by the European Union, methylmercury and PCB-DL can attack the nervous system. Sodium, calcium and iron are three minerals principally found in animal products such as meat or dairy products. Long-term over-exposure to these nutrients is also harmful, *e.g.* consuming too much calcium can provoke urinary and renal calculi and excessive ingestion of sodium favors cardiac issues. As for iron, some studies have underlined a probable link between its excessive ingestion and Parkinson disease (Jenner et al., 1992). The current knowledge about possible synergistic effects between these chemicals, which may increase

sanitary risks, is still quite poor, due to the complexity of these phenomena. Only methylmercury and PCB-DL have been studied jointly, and their simultaneous consumption was observed to amplify health issues in a number of experimental surveys (Bemis and Seegal, 1999; Carpenter et al., 2002). Henceforth, recovering groups of nutrients or contaminants to which the population is observed to be simultaneously over-exposed can help orient future biological and chemical research, which would in turn provide a better understanding of dietary risks. This is the purpose, for instance, of the PERICLES research program (PEsticide Residue In vitro Combined Level of Exposure Study), recently launched by the Anses to identify and quantify the risk due to the exposure to mixtures of pesticides (Béchaux et al., 2013; Crépet and Tressou, 2011; Crépet et al., 2013; Crépet et al., 2013). In terms of statistical analysis, thus reducing the dimension would also enable a more accurate estimation of the complex relationships between these types of exposure. Indeed, even though they are clearly linked by the type of food (fish or meat) introduced in the diet, there are differences of composition between species – like tuna or salmon – that can imply independence between types of extreme long-term exposure. In particular, exceedance of the DIL of more than 3 of these elements are never observed in the data. Because of the variety of individual dietary habits and the complexity of the contamination process, simultaneous types of high exposure are not an obvious phenomenon, are rarely observed, and need to be analyzed in detail.

6.1. Description of the data

Our vectors of 6 types of exposure were calculated on the $n := 2488$ non-pregnant, non-lactating adults of the INCA2 database for which no important variable was missing. Excluding pregnant and lactating women is due to the specificity of their dietary needs, which significantly differ from those of the rest of the population. INCA2 is a nation-wide survey conducted by the Anses from 2005 to 2007 (Afssa, 2009). Carried out in collaboration with the French National Institute of Statistics and Economic Studies (INSEE), it took inventory of the amounts of 1342 foods eaten by 2624 adults during 7 consecutive days. Hence, we consider weekly exposure. Levels of nutrients within each of these 1342 food items were given in the CIQUAL database (Anses, 2008), and equivalents for contaminants were found in TDS2 (Anses, 2011). Both tables result from surveys conducted by the Anses and were designed to match the food nomenclature of INCA2. Then, by simply multiplying quantities of food with the corresponding average amounts of chemical elements contained, we obtained the vectors of exposure $\mathbf{X}_1, \ldots, \mathbf{X}_n$ that are to be examined.

6.2. Analysis of extreme dependencies

We applied the DimRedMev() procedure of Algorithm 4 on the sample of exposure $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with hyper-parameters $v^\star = 0.1$ and $v_G^\star = 1.5 \times 10^{-8}$ (the

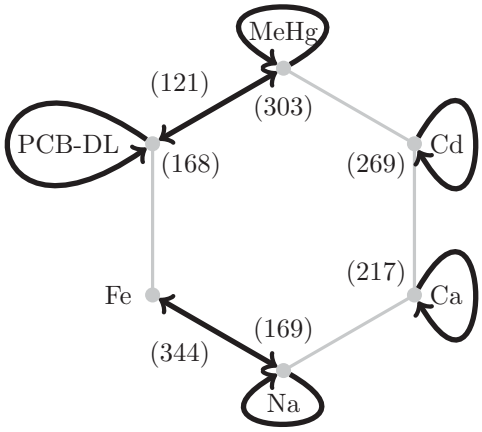

FIG 4. *Dependence structure between the* 6 *nutrients and contaminants of interest on* $k^\star = 564$*; arrows indicate extreme dependencies and the number of observations within each class is given in parentheses.*

TABLE 5
Number of times extreme dependencies occur among all thresholds $t = n/k$, $k \in \{10, \dots, \lfloor n \times 0.3 \rfloor\}$ *(in %)*

MeHg	MeHg & PCB-DL	PCB-DL	Cd	Ca	Fe & Na	Na
7.60	97.15	45.32	95.52	78.83	50.88	48.85

default in `R` package `skmeans`). The resulting dependence structure is represented in Figure 4. To get further confidence in this outcome, we summarize in Table 5 the strongest relationships that were found over all thresholds $t = n/k$. The evolution of $\widetilde{\Upsilon}(k, M_k)$ with k is displayed on Figure 5. Here M_k refers to the optimal number of clusters for fixed k defined on step 19 of Algorithm 4. Our criterion reaches its maximum when $k = 564$, *i.e.* when calculations are based on the 1591 observations with largest radii.

In fact, the dependence structure represented in Figure 4 is found on all 16 largest values of $\widetilde{\Upsilon}(k, M_k)$. The corresponding number of largest values k can be divided into two groups, one where k is in a neighborhood of 360, and another where k is around 560, as illustrated by the highlighted regions in Figure 5. Moreover, Table 5 shows that some dependencies are spotted whatever the number of largest values. In particular, methylmercury is almost always associated to PCB-DL, while cadmium and calcium get separated from all other chemicals. Concerning iron and sodium, uncertainty remains quite high, and a complementary bivariate analysis seems necessary to confirm the nature of their relationship. Figure 6 shows the estimated bivariate angular probability measures of joint exposure first to MeHg and PCB-DL, then to Fe and Na. They were obtained using the maximum empirical likelihood (abbreviated MEL) approach of Einmahl and Segers (2009). Clearly, the strong asymptotic dependence between

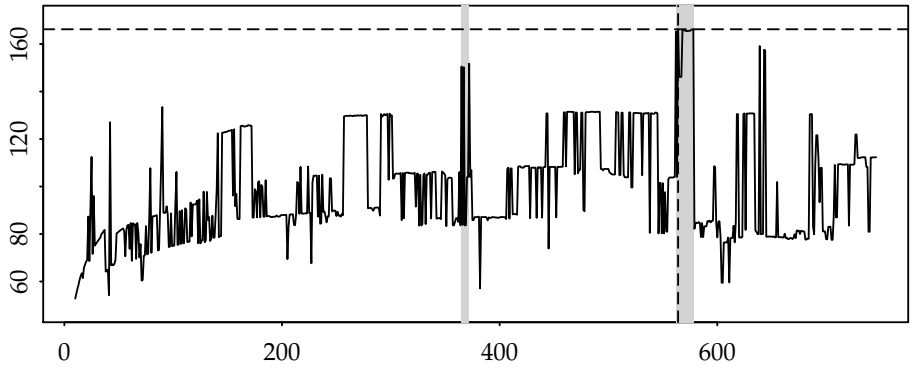

FIG 5. *Evolution in log-scale of* $\widetilde{\Upsilon}(k, M_k)$ *with the number of largest values* k*; the two dashed lines indicate the location of* $\widetilde{\Upsilon}(k^\star, M^\star)$*, while the grayed areas highlight regions where the* 16 *best criteria are obtained.*

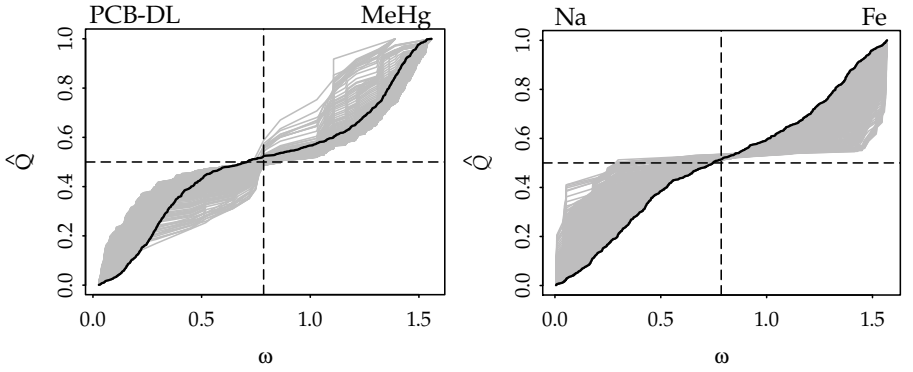

FIG 6. *MEL estimator of the bivariate angular probability measure* Q*, obtained for various values of* k *(grey lines) up to* $k = 564$ *(black line), the optimal number of largest values selected by our criterion; the horizontal dashed line represents asymptotic independence, and the vertical one perfect asymptotic dependence.*

methylmercury and PCB-DL is confirmed, on whatever value of k the estimation may be carried out. The presence of a sub-population reaching extreme exposure to PCB-DL alone is also suggested by the form of $\hat{Q}$, which gets close to vertical height on the extreme left part of the plot, for many values of k. However, methylmercury does not exhibit such a behavior, and given that a specific class of independent exposure to MeHg only occurs for 7.60% of the largest values, we decide to disregard it. In terms of dietary habits, getting two clusters of individuals, one highly exposed to both MeHg and PCB-DL and another solely to PCB-DL, makes perfect sense. Contrary to PCB-DL, methylmercury is a contaminant found exclusively in seafood products. Hence, it is possible to get over-exposed to PCB-DL without ingesting high amounts of MeHg. As for iron

and sodium, according to the evolution of $\hat{Q}$ with k shown on Figure 6, if these two types of exposure exhibit asymptotic dependence, the latter is clearly weak. In fact, we are more inclined to believe in the presence of a mixture of three sub-populations, one ingesting high amounts of both Fe and Na, and the other two getting over-exposed to only one of these nutrients. It is also possible that $k = 564$ being quite high, the relationship appearing in Figure 4 corresponds not to extreme but moderately high levels of exposures. This inconclusive example suggests that extending our approach to the analysis of the hidden angular measure (Resnick, 2002, 2008) would be of major interest.

7. Discussion

Non-parametric analysis of extreme dependencies via the angular measure in high dimension d is still an open issue in multivariate extreme value theory. Though the bivariate setting has already been thoroughly investigated (Beirlant et al., 2004; de Haan, 1985; Einmahl and Segers, 2009; Einmahl et al., 2001; Guillotte et al., 2011; Resnick, 2007), and moderate dimensions are now accessible when all variables are asymptotically dependent (Sabourin and Naveau, 2014), the matter is still unresolved for $d > 5$. Following in the footsteps of Haug et al. (2009), who adapted Principal Components Analysis to extreme dependence assessment, we proposed a method combining multivariate extreme value theory with statistical learning and data mining standards so as to identify sub-groups of variables exhibiting asymptotic dependence. Once these clusters are identified, if they each encompass less than 5 variables, it becomes possible to further estimate the corresponding sub-parts of the angular measure with any existing method, for instance those cited herein-before.

We started in Section 3 by developing the theoretical context under which our approach was constructed. In a non-parametric setting, we focused our attention on the angular probability measure Q. After recalling that it can be viewed as the limit distribution of observation angles given that their radius is getting infinitely large, we underlined the adequacy between the geometry of its support on the positive orthant of the unit hypersphere and the nature of extreme dependencies. Therefore, we proposed a natural model of the angular probability measure as a mixture of angular distributions with supports on each of the $2^d - 1$ non-empty open faces of the simplex. Tackled from a latent variable point of view, this model provided particularly useful properties, formulated in Proposition 3.1 and Proposition 3.2. In particular, we showed that open faces intersecting the support of Q, namely $\{\Omega_h,\ h \in \mathcal{H}\}$, could be identified by means of a simple functional $\kappa_{j,h}(t)$, $1 \leqslant j \leqslant d$, $h \in \mathcal{P}_d$, introduced in Proposition 3.2.

Then, we moved to the practical part of our method in Section 4. Because we had pinpointed the major role of geometry for analyzing angular measures in the preceding section, we adopted geometrical techniques suited for Riemannian objects such as the unit hypersphere for statistical inference. Borrowed from the statistical learning field, they consist in first projecting the initial cloud of points on a lower-dimensional space by means of the Principal Nested Spheres

algorithm of Jung et al. (2012) to reduce the noise, then clustering the obtained data with spherical k-means (Dhillon et al., 2002; Maitra and Ramler, 2010). Resulting clusters were then used to assess $\mathcal{H}$, the set of open faces that intersect $supp(Q)$. Heuristics were constructed based on the empirical counterpart of the functional $\kappa_{j,h}(t)$, in particular to select both the appropriate numbers of groups of dependent variables and of "extreme" observations. Unfortunately, due mainly to the absence of probabilistic analysis of PNS and spherical k-means in the literature, we were not able to provide asymptotic results about the aforementioned objects (this is the object of an ongoing work). Hence, assets and liabilities of our technique were discussed based solely on numerical experiments.

In Section 5, we tested our method on a set of simulated databases. Three scenarios were considered varying in dimension d, the number $\#\mathcal{H}$ of open faces charged by the angular measure and the complexity of $supp(Q)$. In spite of a clearly improvable practical algorithm, the encouraging results we obtained enabled us to define which characteristics of Q have most influence on estimation. In particular, we saw that unlike $\#\mathcal{H}$, d is of negligible importance to the complexity of $supp(Q)$ and the strength of extreme dependence. The closer $\{\Omega_h, h \in \mathcal{H}\}$ are to one another (*e.g.* both $\Omega(\{1,2\})$ and $\Omega(\{2,3\})$ intersect $supp(Q)$), the harder it seems to be to separate and correctly identify each of them. Though they were not considered in the simulations, we added some comments on rates of convergence to the asymptotic dependence structure that were sensed as a determining factor in assessment efficiency. Specifically, we insisted on the role that the hidden angular measure may play when selecting an optimal number of largest values and suggested the interest of generalizing our approach to its analysis. In fine, what emerged from these numerical experiments is that improvement in the estimation of the angular measure can be hoped for, provided that some regularity and sparsity hypotheses are fulfilled.

Further insight into our method was provided by a case-study illustration. Applied to real databases about exposures to 6 food contaminants, it produced stable outcomes, thereby giving confidence in the results. We were able to conclude that only two pairs of chemicals are actually linked in extremes, namely methylmercury and PCB-DL on the one hand, and iron and sodium on the other hand. These associations were confirmed by further computing the MEL estimator of Einmahl and Segers (2009) on these two pairs of variables. In addition, our method spotted a configuration usually hard to notice with traditional estimators, but quite natural given the underlying mixture model on which we based the analysis: it underlined the presence of a mixture of populations, some being jointly over-exposed to a couple of elements, while others ingest high quantities of only one of them (PCB-DL or Na). In terms of public health implications, this means that people who are over-exposed to methylmercury tend to ingest simultaneously high amounts of dioxins and PCB. Knowing that these two toxicants have similar noxious effects on the human organism (Fischer et al., 2008; Weihe et al., 1996), and that when combined, synergistic effects can occur (Bemis and Seegal, 1999; Carpenter et al., 2002), this suggests paying particular attention to the populations that do not respect the corresponding DIL. It also

justifies the need for specific research on potential combined effects of these two contaminants, which would help in assessing the sanitary risks brought upon the concerned population.

In view of these results, one advantage of our multivariate approach is that people in the data are dispatched into multiple classes that embody different types of extreme dependencies. In our case-study example, it facilitates the understanding of over-exposure categories by allowing classical discriminant analyses. An interesting alternative would be to model the various π_h appearing in the mixture model of the angular probability measure in function of auxiliary covariates, *e.g.* some sociologic or economic variables here. More than providing easily interpretable results, this would probably increase the performance of our procedure by helping discriminate between the various clusters. Such generalizations of the present work will be the subject of further investigation in the near future.

Acknowledgements

This work is part of a PhD thesis realized under the supervision of Pr. Stéphan Clémençon (Télécom ParisTech, France) and Jean-Luc Volatier (Anses, France) with the financial support of the INRA (the French National Institute for Agricultural Research), Anses (the French Agency for Food, Environmental and Occupational Health & Safety), Ensai (National School for Statistics and Information Analysis in France) and the University of Cergy-Pontoise.

I would like to thank everyone who contributed to this work, especially Pr. J. Segers for his judicious remarks. I am also grateful to the editor and both anonymous referees for their helpful and detailed comments.

Appendix

A.1. Proof of Proposition 3.1

We shall start this proof by exhibiting two preliminary results. The first one, given in the lemma below, states that π_h can be viewed as the limit probability that λ_h equals 1, $h \in \mathcal{P}_d$, when the radius becomes infinitely large.

Lemma A.1. *Consider the same framework as in Proposition 3.1, then for all* $h \in \mathcal{P}_d$,

$$\mathbb{P}\left(\lambda_h = 1 \,\middle|\, \rho \geqslant t\right) \underset{t \to +\infty}{\longrightarrow} \pi_h.$$

Proof. First of all, extend Q to the whole sphere by setting $Q\left(\mathbb{S}_{(2)} \backslash \Omega\right) = 0$, then consider the following neighborhoods of each of the $2^d - 1$ open faces of the simplex: for any $\epsilon > 0$, $h \in \mathcal{P}_d^\star$ and the geodesic distance $d_G^{d-1}(.,.)$ on Ω, set

$$\mathcal{V}_\epsilon(\Omega_h) := \left\{\boldsymbol{\omega} \in \mathbb{S}_{(2)} : \inf\left\{d_G^{d-1}(\boldsymbol{\omega}, \mathbf{x}),\, \mathbf{x} \in \Omega_h\right\} \leqslant \epsilon\right\}.$$

We shall prove that for all $h \in \mathcal{P}_d^\star$,

$$\lim_{t\to+\infty} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t\right) = Q\left(\mathcal{V}_\epsilon(\Omega_h)\right), \tag{A.16}$$

for an arbitrary small ϵ. This result can be obtained by applying the Portmanteau theorem to Eq. (9), provided that we find at least a decreasing sequence of positive constants $\epsilon_1, \epsilon_2, \ldots$ that tends to 0 such that for any $m \geqslant 1$ and open face Ω_h, the frontier of $\mathcal{V}_{\epsilon_m}(\Omega_h)$ has null measure relative to Q. Since Q is a finite measure, its associated cdf admits at most countably many discontinuity sets, hence the requirement is met.

Now we shall prove that for all $h \in \mathcal{P}_d^\star$,

$$\lim_{\epsilon\to 0} Q\left(\mathcal{V}_\epsilon(\Omega_h)\right) = Q\left(\overline{\Omega_h}\right), \tag{A.17}$$

where $\overline{\Omega_h}$ denotes the closure of Ω_h in Ω. Observe that $(\mathcal{V}_\epsilon(\Omega_h))_\epsilon$ is a decreasing sequence of sets that tends to $\overline{\Omega_h}$ as ϵ tends to 0. Therefore, Eq. (A.17) can be deduced from the monotone continuity property of probability distributions.

By combining Eq. (A.16) and Eq. (A.17), we obtain for all $h \in \mathcal{P}_d^\star$:

$$\lim_{\epsilon\to 0}\lim_{t\to+\infty} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t\right) = Q\left(\overline{\Omega_h}\right). \tag{A.18}$$

Now let $\mathcal{D}_j$, $j \in \{1, \ldots, d\}$ denote the set $\{h \in \mathcal{P}_d^\star : \#h = j\}$. We shall prove Lemma A.1 by strong induction, starting with $j = 1$. First, observe that for all $h \in \mathcal{D}_1$ we have $\overline{\Omega_h} = \Omega_h$ and that the events $\{\lambda_h = 1,\ h \in \mathcal{P}_d\}$ are disjoint by construction. Hence, for any $h \in \mathcal{D}_1$, Eq. (A.18) can be rewritten as follows:

$$\begin{aligned} Q(\Omega_h) &= \lim_{\epsilon\to 0}\lim_{t\to+\infty} \sum_{\ell\in\mathcal{P}_d} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_\ell = 1\right) \mathbb{P}\left(\lambda_\ell = 1 \,\middle|\, \rho \geqslant t\right) \\ &= \lim_{\epsilon\to 0}\lim_{t\to+\infty} \Bigg(\mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_h = 1\right) \mathbb{P}\left(\lambda_h = 1 \,\middle|\, \rho \geqslant t\right) \\ &\quad + \sum_{\substack{\ell\in\mathcal{P}_d \\ \ell\neq h}} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_\ell = 1\right) \mathbb{P}\left(\lambda_\ell = 1 \,\middle|\, \rho \geqslant t\right) \Bigg). \end{aligned}$$

Since Eq. (12) ensures that $\lim_{t\to+\infty} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_h = 1\right) = 1$ for all $\epsilon > 0$ and that for all $\ell \neq h$, $\lim_{\epsilon\to 0}\lim_{t\to+\infty} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_\ell = 1\right) = 0$, we can conclude that

$$\lim_{\epsilon\to 0}\lim_{t\to+\infty} \mathbb{P}\left(\lambda_h = 1 \,\middle|\, \rho \geqslant t\right) = \lim_{t\to+\infty} \mathbb{P}\left(\lambda_h = 1 \,\middle|\, \rho \geqslant t\right) = Q(\Omega_h).$$

Lemma A.1 is thus true for all $h \in \mathcal{D}_1$. Now fix some $J \in \{2, \ldots, d-1\}$ and assume that it holds for all $h \in \bigcup_{j=1}^{J} \mathcal{D}_j$. Set

$$\mathcal{F}_h := \left\{ \ell \in \mathcal{P}_d^\star : \Omega_\ell \in \overline{\Omega_h} \setminus \overset{\circ}{\overline{\Omega_h}} \right\},$$

where $\overset{\circ}{\overline{\Omega_h}}$ denotes the reunion of all open subsets of Ω_h. Using the same arguments as before, for all $h \in \mathcal{D}_{J+1}$ we have:

$$Q(\overline{\Omega_h}) = \lim_{\epsilon \to 0} \lim_{t \to +\infty} \Bigg(\mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_h = 1\right) \mathbb{P}\left(\lambda_h = 1 \,\middle|\, \rho \geqslant t\right) + \sum_{\ell \in \mathcal{F}_h} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_\ell = 1\right) \mathbb{P}\left(\lambda_\ell = 1 \,\middle|\, \rho \geqslant t\right) + \sum_{\substack{\ell \notin \mathcal{F}_h \\ \ell \neq h}} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_\ell = 1\right) \mathbb{P}\left(\lambda_\ell = 1 \,\middle|\, \rho \geqslant t\right) \Bigg).$$

Invoking again Eq. (12), $\lim_{t \to +\infty} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_h = 1\right) = 1$ for all $\epsilon > 0$ and for all $\ell \neq h$,

$$\lim_{\epsilon \to 0} \lim_{t \to +\infty} \mathbb{P}\left(\boldsymbol{\omega} \in \mathcal{V}_\epsilon(\Omega_h) \,\middle|\, \rho \geqslant t,\, \lambda_\ell = 1\right) = \begin{cases} 1 & \text{if } \ell \in \mathcal{F}_h, \\ 0 & \text{if } \ell \notin \mathcal{F}_h. \end{cases}$$

Combined with the induction hypothesis, these equations entail

$$\lim_{t \to +\infty} \mathbb{P}\left(\lambda_h = 1 \,\middle|\, \rho \geqslant t\right) = Q(\overline{\Omega_h}) - \sum_{\ell \in \mathcal{F}_h} \pi_\ell = \pi_h.$$

Now that we have proved Lemma A.1 for all $h \in \mathcal{P}_d^\star$, there remains to check that $\lim_{t \to +\infty} \mathbb{P}\left(\lambda_\varnothing = 1 \mid \rho \geqslant t\right) = \pi_\varnothing = 0$. By construction the distribution of $\boldsymbol{\lambda}$ is Categorical with parameters $(p_h)_{h \in \mathcal{P}_d}$, therefore the events $\{\lambda_h = 1\}_{h \in \mathcal{P}_d}$ are mutually exclusive and $\{\lambda_\varnothing = 0\} = \bigcup_{h \in \mathcal{P}_d^\star} \{\lambda_h = 1\}$. As a result, we have:

$$\begin{aligned} \mathbb{P}\left(\lambda_\varnothing = 1 \mid \rho \geqslant t\right) &= 1 - \mathbb{P}\left(\lambda_\varnothing = 0 \mid \rho \geqslant t\right) = 1 - \mathbb{P}\left(\bigcup_{h \in \mathcal{P}_d^\star} \{\lambda_h = 1\} \,\middle|\, \rho \geqslant t \right) \\ &= 1 - \sum_{h \in \mathcal{P}_d^\star} \mathbb{P}\left(\lambda_h = 1 \mid \rho \geqslant t\right) \underset{t \to +\infty}{\longrightarrow} 1 - \sum_{h \in \mathcal{P}_d^\star} \pi_h = \pi_\varnothing = 0. \end{aligned}$$

This concludes the proof. □

The second preliminary result in the lemma below states that the distribution of vector $\mathbf{Z}$ given that $\lambda_h = 1$ is multivariate regularly varying when $h \in \mathcal{H}$.

Lemma A.2. *Consider the same framework as in Proposition 3.1, then for all $h \in \mathcal{H}$, there is a Radon measure μ_h, non identically zero and not degenerate at a point, concentrated on the blunt convex cone $\mathcal{C}_h := \{\mathbf{x} \in \mathcal{C}_\star : \mathbf{x}/\|\mathbf{x}\|_{(2)} \in \Omega_h\}$, such that*

$$t\, \mathbb{P}\left(\frac{\mathbf{Z}}{t} \in . \;\middle|\; \lambda_h = 1 \right) \underset{t \to +\infty}{\overset{v}{\longrightarrow}} \mu_h(.).$$

Proof. By Lemma A.1, Eq. (8) and Eq. (12), we have that

$$t\,\mathbb{P}\left(\boldsymbol{\omega}\in .,\ \rho\geqslant t\ \middle|\ \lambda_h=1\right)\ \xrightarrow[t\to+\infty]{v}\ S_h(.):=\begin{cases}S(.\cap\Omega_h)/p_h & \text{if } h\in\mathcal{H},\\ 0 & \text{otherwise,}\end{cases}$$

where by definition,

$$\begin{aligned}S(.\cap\Omega_h)&=\mu\left(\{\mathbf{x}\in\mathcal{C}_\star:\|\mathbf{x}\|_{(1)}\geqslant 1,\ \mathbf{x}/\|\mathbf{x}\|_{(2)}\in .\cap\Omega_h\}\right)\\&=\mu\left(\{\mathbf{x}\in\mathcal{C}_\star:\|\mathbf{x}\|_{(1)}\geqslant 1,\ \mathbf{x}/\|\mathbf{x}\|_{(2)}\in .\}\right.\\&\qquad\left.\cap\{\mathbf{x}\in\mathcal{C}_\star:\mathbf{x}/\|\mathbf{x}\|_{(2)}\in\Omega_h\}\right).\end{aligned}$$

Recall that $\mathcal{C}_h:=\{\mathbf{x}\in\mathcal{C}_\star:\mathbf{x}/\|\mathbf{x}\|_{(2)}\in\Omega_h\}$, and set

$$\mu_h(.):=\begin{cases}\mu(.\cap\mathcal{C}_h)/p_h & \text{if } h\in\mathcal{H}\\ 0 & \text{otherwise,}\end{cases}$$

then we can rewrite S_h in function of μ_h as below:

$$S_h(.)=\mu_h(\{\mathbf{x}\in\mathcal{C}_\star:\|\mathbf{x}\|_{(1)}\geqslant 1,\ \mathbf{x}/\|\mathbf{x}\|_{(2)}\in .\}).$$

Since $\mathcal{C}_h$ is a cone, the homogeneity property of μ stated in Eq. (4) is passed on μ_h, $h\in\mathcal{H}$. Indeed, for all $0<s<+\infty$ and Borel subsets B of $\mathcal{C}_\star$,

$$\begin{aligned}\mu_h(sB)&=\mu\left((sB)\cap(\mathcal{C}_h)\right)/p_h=\mu\left(s\left(B\cap\mathcal{C}_h\right)\right)/p_h\\&=s^{-1}\mu\left(B\cap\mathcal{C}_h\right)/p_h=s^{-1}\mu_h(B).\end{aligned}$$

According to Theorem 6.1 in Resnick (2007), it naturally follows that

$$t\,\mathbb{P}\left(\frac{\mathbf{Z}}{t}\in .\ \middle|\ \lambda_h=1\right)\ \xrightarrow[t\to+\infty]{v}\ \mu_h(.),$$

where, just like μ, μ_h can be written as the product of a measure on the radius with a measure on the angles when switching to pseudo-polar coordinates:

$$\mu_h\circ T^{-1}=\mu_{-1}\times S_h.$$

□

We can now tackle the proof of Proposition 3.1. Going back to the marginal level, multivariate regular variation of conditional distributions gives for all $x\geqslant 1$, $1\leqslant j\leqslant d$,

$$t\,\mathbb{P}\left(\frac{Z_j}{t}>x\ \middle|\ \lambda_h=1\right)\ \xrightarrow[t\to+\infty]{v}\ \mu_h(\{\mathbf{z}\in\mathcal{C}_\star:z_j>x\}).$$

Notice that we now have a null limit for all $h\notin\mathcal{H}(j)$. Furthermore, for all $h\in\mathcal{H}(j)$, we have

$$\mu_h(\{\mathbf{z}\in\mathcal{C}_\star:z_j>x\})=\int_\Omega\int_{(0,+\infty]}\mathbb{I}\left\{\rho\,\frac{\omega_j}{\|\boldsymbol{\omega}\|_{(1)}}>x\right\}\mu_{-1}(d\rho)\,S_h(d\boldsymbol{\omega})$$

$$= x^{-1} \underbrace{\int_\Omega \frac{\omega_j}{\|\boldsymbol{\omega}\|_{(1)}} S_h(d\boldsymbol{\omega})}_{c_{j,h}}.$$

Hence, for all $h \in \mathcal{P}_d$, $j \in \{1, \dots, d\}$, $x \geqslant 1$, we can write

$$t\,\mathbb{P}\left(Z_j > x\,t \mid \lambda_h = 1\right) \underset{t\to+\infty}{\longrightarrow} c_{j,h}\,x^{-1},$$

where $c_{j,h} > 0$ when $h \in \mathcal{H}(j)$, and $c_{j,h} = 0$ otherwise. Based on the marginal constraints on S stated in Eq. (10) and because $\{\Omega_h\}_{h\in\mathcal{P}_d}$ forms a partition of Ω, we have that $c_{j,h} \in [0, 1/p_h]$ for all $h \in \mathcal{P}_d$ and $\sum_{h\in\mathcal{P}_d} p_h\,c_{j,h} = \sum_{h\in\mathcal{H}(j)} p_h\,c_{j,h} = 1$.

A.2. Proof of Proposition 3.2

We shall handle the situations where $h \in \mathcal{H}(j)$ and $h \notin \mathcal{H}(j)$ separately. To simplify notations, for all $h \in \mathcal{P}_d$ and $x \geqslant 0$ we will denote by $\bar{F}_{j,h}(x)$ the conditional probability that Z_j exceeds x given λ_h equals 1:

$$\bar{F}_{j,h}(x) := \mathbb{P}\left(Z_j > x \mid \lambda_h = 1\right).$$

- $h \in \mathcal{H}(j)$ **:** $p_h \neq 0$ **and** $\pi_h \neq 0$

From Eq. (13) in Proposition 3.1, it is straightforward that $\bar{F}_{j,h}$ is regularly varying with index -1, *i.e.* for any $x \geqslant 1$,

$$\frac{\bar{F}_{j,h}(x\,t)}{\bar{F}_{j,h}(t)} \underset{t\to+\infty}{\longrightarrow} x^{-1}.$$

Hence, $\bar{F}_{j,h}$ may be written as follows:

$$\bar{F}_{j,h}(x) = x^{-1}\,L_{j,h}(x),$$

where $L_{j,h}(x)$ is a slowly varying function ($\forall\, s > 0$, $L_{j,h}(s\,x)/L_{j,h}(x) \longrightarrow_{x\to+\infty} 1$) that converges to $c_{j,h}$ as $x \to +\infty$.

Remark A.1. Define $x^\star_{j,h} := \inf\{x \geqslant 1 : \bar{F}_{j,h}(x) = 0\}$, the right endpoint of the survival function $\bar{F}_{j,h}$ for any $j \in \{1, \dots, d\}$ and any $h \in \mathcal{P}_d$. Then for all $h \in \mathcal{H}(j)$, $x^\star_{j,h} = +\infty$, that is $\forall t \geqslant 1$, $\bar{F}_{j,h}(t) > 0$.

Since Bayes' formula gives

$$\begin{aligned}\kappa_{j,h}(t) &:= \int_1^{+\infty} t\,\mathbb{P}\left(\rho \geqslant t\right)\,\mathbb{P}\left(Z_j > x\,t \mid \rho \geqslant t,\,\lambda_h = 1\right)\,dx \\ &= \frac{t\,p_h}{\mathbb{P}\left(\lambda_h = 1 \mid \rho \geqslant t\right)} \int_1^{=} +\infty \mathbb{P}\left(Z_j > x\,t,\,\rho \geqslant t \mid \lambda_h = 1\right)\,dx,\end{aligned}$$

and for all $x \geqslant 1$, $\mathbb{P}\left(Z_j > x\,t,\,\rho \geqslant t \mid \lambda_h = 1\right) = \bar{F}_{j,h}(x\,t)$, we have

$$\kappa_{j,h}(t) = \frac{t\,\bar{F}_{j,h}(t)\,p_h}{\mathbb{P}\left(\lambda_h = 1 \mid \rho \geqslant t\right)} \int_1^{+\infty} \frac{\bar{F}_{j,h}(x\,t)}{\bar{F}_{j,h}(t)}\,dx$$

$$= \frac{L_{j,h}(t)\, p_h}{\mathbb{P}\left(\lambda_h = 1 \mid \rho \geqslant t\right)} \int_1^{+\infty} x^{-1}\, \frac{L_{j,h}(x\,t)}{L_{j,h}(t)}\, dx.$$

Fix some $\epsilon > 0$, small enough to verify $c_{j,h} - \epsilon > 0$, and some $t_\epsilon > 0$ such that $\forall\, t \geqslant t_\epsilon$, we have simultaneously $\left|\mathbb{P}\left(\lambda_h = 1 \mid \rho \geqslant t\right) - \pi_h\right| < \epsilon$ (Lemma A.1) and $\left|L(t) - c_{j,h}\right| < \epsilon$. Obviously, as soon as $t \geqslant t_\epsilon$, we also have $\left|L(x\,t) - c_{j,h}\right| < \epsilon$ for all $x \geqslant 1$, and

$$0 < \frac{c_{j,h} - \epsilon}{c_{j,h} + \epsilon} < \frac{L_{j,h}(x\,t)}{L_{j,h}(t)}.$$

Hence, $\forall\, t \geqslant t_\epsilon$,

$$\kappa_{j,h}(t) > \frac{(c_{j,h} - \epsilon)^2\, p_h}{(\pi_h + \epsilon)(c_{j,h} + \epsilon)} \int_1^{+\infty} x^{-1}\, dx = +\infty,$$

or equivalently $\kappa_{j,h}(t) \longrightarrow_{t \to +\infty} +\infty$.

- $h \in \mathcal{H} \backslash \mathcal{H}(j)$ **:** $p_h \neq 0$ **and** $\pi_h \neq 0$

Contrary to the case where $h \in \mathcal{H}(j)$, the conditional cdf $\bar{F}_{j,h}$ can have either finite or infinite right endpoint. When its support is bounded, relying on the Bayes decomposition exhibited in the previous paragraph, the desired result is straightforward: because there exists some $t_0 > 1$ such that for all $t > t_0$, $\bar{F}_{j,h}(t) = 0$, then as $t \to +\infty$, the integral also becomes null. If on the contrary, $\bar{F}_{j,h} > 0$ for all $t \geqslant 1$, then, as previously, we can rewrite the quantity of interest in the following form:

$$\kappa_{j,h}(t) = \frac{t\, p_h\, \bar{F}_{j,h}(t)}{\mathbb{P}\left(\lambda_h = 1 \mid \rho \geqslant t\right)} \int_1^{+\infty} \frac{\bar{F}_{j,h}(x\,t)}{\bar{F}_{j,h}(t)}\, dx.$$

Since as t tends to infinity $t\, \bar{F}_{j,h}(t)$ tends to 0 (Proposition 3.1), $\mathbb{P}\left(\lambda_{i,h} \mid \rho_i \geqslant t\right)$ tends to $\pi_h > 0$ (Lemma A.1) and since $p_h > 0$, for the integral of interest to converge to 0 it suffices to prove that there exists some $t_0 > 1$ such that for all $t > t_0$,

$$\int_1^{+\infty} \frac{\bar{F}_{j,h}(x\,t)}{\bar{F}_{j,h}(t)}\, dx < +\infty.$$

According to the assumption in Proposition 3.2, there exists some constants $\gamma^\star \in (0,1)$, $c^\star \geqslant 0$ and $t^\star > 1$ such that

$$(t > t^\star) \Rightarrow \left(\frac{\bar{F}_{j,h}(x\,t)}{\bar{F}_{j,h}(t)} \leqslant c^\star\, x^{-1/\gamma^\star}\right).$$

Hence, for all $t > t^\star$,

$$\int_1^{+\infty} \frac{\bar{F}_{j,h}(x\,t)}{\bar{F}_{j,h}(t)}\, dx \leqslant c^\star \int_1^{+\infty} x^{-1/\gamma^\star}\, dx = \frac{-c^\star}{1 - 1/\gamma^\star} < +\infty,$$

which produces the desired outcome.

- $h \notin \mathcal{H} \cup \{\varnothing\}$ **:** $p_h = \pi_h = 0$

  By definition, for all $h \in \mathcal{P}_d^\star$, the equivalence below holds true:

$$(h \in \mathcal{H}^c \backslash \{\varnothing\}) \Leftrightarrow (\pi_h = 0) \Leftrightarrow (p_h = 0).$$

Consequently, when $h \notin \mathcal{H} \cup \{\varnothing\}$, we have $\mathbb{P}\left(Z_j > x \,\middle|\, \rho \geqslant t,\, \lambda_h = 1\right) = 0$ for all $x \geqslant 0$, and by extension

$$\int_1^{+\infty} t\,\mathbb{P}\left(\rho \geqslant t\right) \mathbb{P}\left(Z_j > x\,t \,\middle|\, \rho \geqslant t,\, \lambda_h = 1\right) dx = 0,$$

for all $t > 0$. This remains true as $t \to +\infty$.

- $h = \varnothing$ **:** $p_h \neq 0$ **and** $\pi_h = 0$

  Let us start again with the following decomposition:

$$\kappa_{j,\varnothing}(t) = \frac{t\,p_\varnothing\,\bar{F}_{j,\varnothing}(t)}{\mathbb{P}\left(\lambda_\varnothing = 1 \,\middle|\, \rho \geqslant t\right)} \int_1^{+\infty} \frac{\bar{F}_{j,\varnothing}(x\,t)}{\bar{F}_{j,\varnothing}(t)}\, dx.$$

Contrary to the case where $h \in \mathcal{H}\backslash\mathcal{H}(j)$, we cannot guarantee the convergence of $\kappa_{j,\varnothing}(t)$ to 0 as t grows to infinity, since $\mathbb{P}\left(\lambda_\varnothing = 1 \,\middle|\, \rho \geqslant t\right)$ now tends to 0 instead of a positive constant. Nonetheless, it is still possible to prove that it does not diverge to $+\infty$. Indeed, notice that

$$\frac{t\,p_\varnothing\,\bar{F}_{j,\varnothing}(t)}{\mathbb{P}\left(\lambda_\varnothing = 1 \,\middle|\, \rho \geqslant t\right)} = \frac{t\,\mathbb{P}\left(\rho \geqslant t\right)\,\bar{F}_{j,\varnothing}(t)}{\mathbb{P}\left(\rho \geqslant t \,\middle|\, \lambda_\varnothing = 1\right)},$$

and that $\bar{F}_{j,\varnothing}(t) \leqslant \mathbb{P}\left(\rho \geqslant t \,\middle|\, \lambda_\varnothing = 1\right)$. Hence,

$$\kappa_{j,\varnothing}(t) \leqslant t\,\mathbb{P}\left(\rho \geqslant t\right) \int_1^{+\infty} \frac{\bar{F}_{j,\varnothing}(x\,t)}{\bar{F}_{j,\varnothing}(t)}\, dx.$$

We have already seen that according to the assumption in Proposition 3.2, there exists some constants $\gamma^\star \in (0,1)$, $c^\star \geqslant 0$ and $t^\star > 1$ such that for all $t > t^\star$,

$$\int_1^{+\infty} \frac{\bar{F}_{j,\varnothing}(x\,t)}{\bar{F}_{j,\varnothing}(t)}\, dx \leqslant \frac{-c^\star}{1 - 1/\gamma^\star}.$$

Moreover, by virtue of Eq. (8), for all $\epsilon > 0$ there exists some $t_\epsilon > 0$ such that for all $t > t_\epsilon$, $|t\,\mathbb{P}\left(\rho \geqslant t\right) - S(\Omega)| < \epsilon$. Fix some $\epsilon > 0$ and set $\epsilon^\star := -\epsilon\,(1 - 1/\gamma^\star)/c^\star$, then for all $t > \max(t^\star, t_\epsilon)$, we have

$$\kappa_{j,\varnothing}(t) \leqslant \frac{-S(\Omega)\,c^\star}{1 - 1/\gamma^\star} + \epsilon^\star < +\infty.$$

Observe that the smaller $\gamma^\star$, *i.e.* the faster the limit dependence structure is reached, the smaller the bound of $\kappa_{j,\varnothing}(t)$. Ideally, when all $\bar{F}_{j,\varnothing}$, $1 \leqslant j \leqslant d$, are rapidly varying, *i.e.* $c^\star = 0$, we obtain the same result as in the case where $h \in \mathcal{H}\backslash\mathcal{H}(j)$. This would correspond in fact to the absence of hidden regular variation, like mentioned in Section 5 and Section 7 (Heffernan and Resnick, 2005; Resnick, 2002, 2008).

A.3. About $\widehat{\kappa}_{j,\ell}(k)$

For the sake of clarity, we give here a more explicit version of the statistic $\widehat{\kappa}_{j,m}(k)$, which was defined for all $1 \leqslant m \leqslant M,\ 1 \leqslant j \leqslant d$ as

$$\widehat{\kappa}_{j,m}(k) := \int_1^{+\infty} \frac{1}{k}\,\frac{n_k}{n_m} \sum_{i \in \aleph_k} \mathbb{I}\left\{\widehat{Z}_{i,j} > x\,\frac{n}{k},\, I_{i,m} = 1\right\} dx,$$

where $n_k := \#\aleph_k$ is the number of observations the radius of which exceeds n/k, and $n_m := \sum_{i \in \aleph_k} \mathbb{I}\{I_{i,m} = 1\}$ the size of class m. Recall that m is supposed to refer to some $h \in \mathcal{P}_d^\star$ via the function $f_M(m)$. Let us begin by considering that k is fixed, and set

$$g_{j,m}(x) = \frac{1}{k}\,\frac{n_k}{n_m} \sum_{i \in \aleph_k} \mathbb{I}\left\{\widehat{Z}_{i,j} > x\,\frac{n}{k},\, I_{i,m} = 1\right\},\ 1 \leqslant m \leqslant M,\ 1 \leqslant j \leqslant d.$$

Our statistic of interest, $\widehat{\kappa}_{j,m}(k)$ is none other than the integral over $x \geqslant 1$ of $g_{j,m}(x)$. Actually, because it relies on a finite set of $n \leqslant 1$ observations, $g_{j,m}(x)$ is a step function with support on $[\min_{1 \leqslant i \leqslant n} \widehat{Z}_{i,j}\, k/n, \max_{1 \leqslant i \leqslant n} \widehat{Z}_{i,j}\, k/n]$. As $g_{j,m}$ only takes into account observations verifying $I_{i,m} = 1$, we denote by $(\widehat{Z}^m_{1,j}, \ldots, \widehat{Z}^m_{n_m,j})$ the sub-sample of n_m observations within $(\widehat{Z}_{1,j}, \ldots, \widehat{Z}_{n,j})$, for any $j \in \{1, \ldots, d\}$. Denote by $\widehat{Z}^m_{(1,j)} < \ldots < \widehat{Z}^m_{(n_m,j)}$ the corresponding ordered statistics and consider

$$\widehat{Z}^m_{(n_m - u^\star, j)} = \inf\left\{\widehat{Z}^m_{i,j}, 1 \leqslant i \leqslant n_m \,:\, \widehat{Z}^m_{i,j} \geqslant \frac{n}{k}\right\},$$

the smallest observation $\widehat{Z}^m_{i,j}$ that exceeds n/k. Arbitrarily set $\widehat{Z}^m_{(n_m - u^\star - 1, j)} = 1$, then $g_{j,m}$ can be expressed as follows:

$$g_{j,m}(x) = \frac{1}{k}\,\frac{n_k}{n_m} \sum_{u=1}^{u^\star+1} u\, \mathbb{I}\left\{x \in \left[\widehat{Z}^m_{(n_m-u,j)}\,\frac{k}{n},\, \widehat{Z}^m_{(n_m-u+1,j)}\,\frac{k}{n}\right)\right\}.$$

In particular, when $x \geqslant \widehat{Z}^m_{(n_m,j)}\, k/n$, there is no $\widehat{Z}^m_{i,j}$, $1 \leqslant i \leqslant n_m$, such that $\widehat{Z}^m_{i,j}\, k/n > x$, and conversely, when $x \in [1, \widehat{Z}^m_{(n_m - u^\star, j)}\, k/n)$, there are exactly $u^\star + 1$ observations $\widehat{Z}^m_{i,j}$ in the sub-sample defined by $I_{i,m} = 1$ that exceed $x\, n/k$. Therefore, the integral of $g_{j,m}(x)$ over all $x \geqslant 1$ verifies

$$\begin{aligned}\int_1^{\max_{1 \leqslant i \leqslant n} \widehat{Z}^m_{i,j}\, k/n} g_{j,m}(x)\, dx &= \widehat{\kappa}_{j,m}(k) \\ &= \frac{n_k}{n}\,\frac{1}{n_m} \sum_{u=1}^{u^\star+1} u\, \left(\widehat{Z}^m_{(n_m-u+1,j)} - \widehat{Z}^m_{(n_m-u,j)}\right).\end{aligned}$$

References

Afssa, INCA 2 (2006–2007), Individual and National Study on Food Consumption, 2009. Summary and full report at https://www.anses.fr/sites/default/files/documents/PASER-Sy-INCA2.pdf, https://www.anses.fr/sites/default/files/documents/PASER-Ra-INCA2.pdf.

Anses, French food composition table, 2008. Official webpage https://pro.anses.fr/TableCIQUAL/index.htm.

Anses, Second French Total Diet Study (TDS 2), 2011. Volumes 1 and 2 at https://www.anses.fr/sites/default/files/documents/PASER2006sa0361Ra1EN.pdf, https://www.anses.fr/sites/default/files/documents/PASER2006sa0361Ra2EN.pdf.

Béchaux, C., Zetlaoui, M., Tressou, J., Leblanc, J.C., Héraud, F., and Crépet, A., Identification of pesticide mixtures and connection between combined exposure and diet. *Food Chem. Toxicol.*, 59:191–198, 2013.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J., *Statistics of extremes: theory and applications*. John Wiley & Sons Inc, 2004. ISBN 0471976474. MR2108013

Bemis, J.C. and Seegal, R.F., Polychlorinated biphenyls and methylmercury act synergistically to reduce rat brain dopamine content in vitro. *Environ. Health Persp.*, 107(11):879, 1999.

Boldi, M.O. and Davison, A.C., A mixture model for multivariate extremes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2):217–229, 2007. MR2325273

Carpenter, D.O., Arcaro, K., and Spink, D.C., Understanding the human health effects of chemical mixtures. *Environ. Health Persp.*, 110(Suppl 1):25, 2002.

Cattell, R.B., The scree test for the number of factors. *Multivar. Behav. Res.*, 1(2):245–276, 1966.

Crépet, A. and Tressou, J., Bayesian nonparametric model with clustering individual co-exposure to pesticides found in the French diet. *Bayesian Anal.*, 6(1):127–144, 2011. MR2781810

Crépet, A., Héraud, F., Béchaux, C., Gouze, M.E., Pierlot, S., Fastier, A., Leblanc, J.C., Le Hégarat, L., Takakura, N., Fessard, V., et al., The PERICLES research program: an integrated approach to characterize the combined effects of mixtures of pesticide residues to which the French population is exposed. *Toxicology*, 313(2):83–93, 2013.

Crépet, A., Tressou, J., Graillot, V., Béchaux, , Pierlot, S., Héraud, F., and Leblanc, J.C., Identification of the main pesticide residue cocktails to which the French population is exposed. *Environ. Res.*, 126:125–133, 2013.

Das, B. and Resnick, S.I., Detecting a conditional extreme value model. *Extremes*, 14(1):29–61, 2011. MR2775870

Das, B., Mitra, A., and Resnick, S.I., Living on the multidimensional edge: seeking hidden risks using regular variation. *Adv. in Appl. Probab.*, 45(1):139–163, 2013. MR3077544

DE HAAN, L., Extremes in higher dimensions: the model and some statistics. In *Proceedings 45th Session of the I.S.I. paper 26.3, Amsterdam*, 1985. MR0886266

DE HAAN, L. and FERREIRA, A., *Extreme value theory: an introduction*. Springer Verlag, 2006. ISBN 978-0-387-23946-0. MR2234156

DHILLON, I.S., GUAN, Y., and KOGAN, J., Iterative clustering of high dimensional text data augmented by local search. In *Proc. 2002 IEEE Int. Conf. Data Mining*, pages 131–138. IEEE, 2002.

DONOHO, D.L., High-dimensional data analysis: the curses and blessings of dimensionality. *AMS Math. Challenges Lecture*, pages 1–32, 2000.

EINMAHL, J.H.J. and SEGERS, J., Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *Ann. Statist.*, 37(5B):2953–2989, 2009. MR2541452

EINMAHL, J.H.J., DE HAAN, L., and PITERBARG, V.I., Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Statist.*, 29(5):1401–1423, 2001. MR1873336

FISCHER, C., FREDRIKSSON, A., and ERIKSSON, P., Neonatal co-exposure to low doses of an ortho-PCB (PCB 153) and methylmercury exacerbate defective developmental neurobehavior in mice. *Toxicology*, 244(2–3):157–165, 2008.

FLETCHER, P.T., LU, C., PIZER, S.M., and JOSHI, S., Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.*, 23(8):995–1005, 2004.

FRIEDMAN, J., HASTIE, T., and TIBSHIRANI, R., *The elements of statistical learning*. Springer Series in Statistics, 2009. ISBN 978-0-387-84857-0. MR2722294

GUILLOTTE, S., PERRON, F., and SEGERS, J., Non-parametric Bayesian inference on bivariate extremes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(3):377–406, 2011. MR2815781

HAUG, S., KLÜPPELBERG, C., and KUHN, G., Dimension reduction based on extreme dependence. 2009.

HEFFERNAN, J. and RESNICK, S.I., Hidden regular variation and the rank transform. *Adv. in Appl. Probab.*, 37(2):393–414, 2005. MR2144559

HUCKEMANN, S. and ZIEZOLD, H., Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab.*, 38(2):99–319, 2006. MR2264946

JENNER, P., SCHAPIRA, A.H.V., and MARSDEN, C.D., New insights into the cause of Parkinson's disease. *Neurology*, 42(12):2241, 1992.

JUNG, S., FOSKEY, M., and MARRON, J.S., Principal arc analysis on direct product manifolds. *Ann. Appl. Statist.*, 5(1):578–603, 2011. MR2810410

JUNG, S., DRYDEN, I.L., and MARRON, J.S., Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, 2012. MR2966769

MAITRA, R. and RAMLER, I.P., A k-mean-directions algorithm for fast clustering of data on the sphere. *J. Comput. Graph. Statist.*, 19(2):377–396, 2010. MR2758308

Massart, P., Strong approximation for multivariate empirical and related processes, via KMT constructions. *Ann. Probab.*, 17(1):266–291, 1989. MR0972785

McLachlan, G. and Peel, D., *Finite mixture models*, volume 299. Wiley-Interscience, 2000. ISBN 9780471006268. MR1789474

Paulo, M.J., Van der Voet, H., Wood, J.C., Marion, G.R., and Van Klaveren, J.D., Analysis of multivariate extreme intakes of food chemicals. *Food Chem. Toxicol.*, 44(7):994–1005, 2006. ISSN 0278-6915.

Resnick, S.I., Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5(4):303–336, 2002. MR2002121

Resnick, S.I., *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Verlag, 2007. ISBN 978-0-387-24272-9. MR2271424

Resnick, S.I., Multivariate regular variation on cones: application to extreme values, hidden regular variation and conditioned limit laws. *Stochastics*, 80(2–3):269–298, 2008. MR2402168

Sabourin, A. and Naveau, P., Bayesian Dirichlet mixture model for multivariate extremes: a re-parametrization. *Comput. Statist. Data Anal.*, 71:542–567, 2014. MR3131989

Stephenson, A., Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59, 2003. MR2021592

Tressou, J., Crépet, A., Bertail, P., Feinberg, M.H., and Leblanc, J.C., Probabilistic exposure assessment to food chemicals based on extreme value theory. Application to heavy metals from fish and sea products. *Food Chem. Toxicol.*, 42(8):1349–1358, 2004.

Weihe, P., Grandjean, P., Debes, F., and White, R., Health implications for Faroe Islanders of heavy metals and PCBs from pilot whales. *Sci. Total Environ.*, 186(1–2):141–148, 1996.