Electronic Journal of Statistics
Vol. 14 (2020) 1211–1233
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1689

k-means clustering of extremes

Anja Janßen

KTH Royal Institute of Technology, Stockholm
Department of Mathematics, Lindstedtsvägen 25, 20 400 Stockholm, Sweden
e-mail: anjaj@kth.se

Phyllis Wan

Erasmus University Rotterdam
Erasmus School of Economics, Burg. Oudlaan 50, 3062 PA Rotterdam, the Netherlands
e-mail: wan@ese.eur.nl

Abstract: The k-means clustering algorithm and its variant, the spherical k-means clustering, are among the most important and popular methods in unsupervised learning and pattern detection. In this paper, we explore how the spherical k-means algorithm can be applied in the analysis of only the extremal observations from a data set. By making use of multivariate extreme value analysis we show how it can be adopted to find "prototypes" of extremal dependence and derive a consistency result for our suggested estimator. In the special case of max-linear models we show furthermore that our procedure provides an alternative way of statistical inference for this class of models. Finally, we provide data examples which show that our method is able to find relevant patterns in extremal observations and allows us to classify extremal events.

MSC 2010 subject classifications: Primary 62G32; secondary 62H30, 60G70.
Keywords and phrases: dimension reduction, extreme value statistics, k-means clustering, spectral measure.

Received May 2019.

Contents

1. Introduction

When looking at multivariate and in particular high-dimensional data, a key aspect is to detect structures and patterns in the observations so as to simplify their complexity. This task has led to the development of an abundance of procedures in the field of unsupervised learning, see for example Hastie et al. (2009) for an overview. Usually the analysis looks for results that apply to most observations at hand and hence focuses on describing the bulk of the data. On the other hand, when *extremal* observations are of interest, a different approach needs to be taken. In this paper we consider the complexity reduction of extremal observations via clustering.

For a specific dataset, a naive implementation is to choose the observations with the largest norm (thereby considering them as extremal observations) and apply a clustering algorithm to them. However, this is inefficient as extremal points are typically spread out in space. In the presence of heavy-tailed observations, most classical clustering algorithms would have further problems from the possibly infinite second moments. In order to allow for robust and meaningful estimation, one should incorporate structural results about the particular kind of data at hand into the estimation procedure.

Multivariate extreme value theory (MEVT) provides us with such a framework and has useful applications in a wide range of disciplines, such as finance and climate science, see for example Fougères (2003) and Davison et al. (2012) for overviews. Most of the current parametric models and estimation methods focus on the bivariate or lower dimensional scenario and are difficult to generalize to higher dimensions due to either lack of flexibility or heavy computation loads, see Davison and Huser (2015).

Recently there have been a few attempts to adapt complexity reduction to extremal dependence. One way of doing so is applying classical dimension reduction techniques to (transformed) extremal observations, in the form of principal component analysis and related covariance matrix decomposition techniques, see Haug et al. (2015) and Cooley and Thibaud (2019), or empirical basis functions, see Morris et al. (2019). Another direction of research is aimed at dividing the parameter space into lower dimensional subspaces using the phenomenon of asymptotic independence (meaning that in extremal events there are often only a few components which are large at the same time), see Goix et al. (2017) and Chiapino et al. (2019). Chautru (2015) identifies relevant subspaces by first reducing the dimension of projected observations by a spherical principal components procedure, then clustering the projected data using spherical k-means, and as a last step attributes a lower-dimensional subspace to each cluster. Finally, Bernard et al. (2013) presents a classification approach with special emphasis on a spatial decomposition of separated regional clusters for extremal events. The methodology is based on a k-means like algorithm, where distances between two stations are measured by their F-madogram.

The usage of k-means estimation in extremes is therefore not entirely new (see also Einmahl et al. (2012), where the procedure is mentioned to produce starting values for numerical estimation algorithms), but so far it has only

been applied as an intermediate step towards a specific goal and its theoretical properties have not been explored. The aim of this paper is twofold: First, we provide the theoretical background as to how a k-means algorithm applied to extremal observations can be constructed as a consistent estimator of theoretical extremal cluster centers, see Theorem 3.1. Second, we demonstrate that these cluster centers themselves can be seen as prototypes of directions of extremal events and the algorithm therefore provides a comprehensive, computationally fast and robust procedure to interpret observed extremes. As a side effect we demonstrate that our procedure can be seen as an alternative, consistent way of estimating relevant components of max-linear models, which recently gained popularity in applications due to their relationship to causal models for extremes as implied by directed acyclic graph (DAG) models, see Gissibl (2018); Gissibl and Klüppelberg (2018); Gissibl et al. (2018).

The paper is organized as follows: Section 2 provides a short background on the two main components of our method, MEVT and the spherical k-means algorithm. In Section 3, we present a general consistency result for the spherical k-means algorithm in the extremal setting and construct a non-parametric estimator for the theoretical cluster centers. The application of our procedure to the particular class of max-linear models is outlined in Section 4. In Section 5, three data examples illustrate the application and interpretation of the method. We finish with a short discussion of our contribution in Section 6.

2. Background

2.1. Multivariate extreme value theory

In studying the extremal behavior of a random vector, a general assumption is that the componentwise maxima, generated from i.i.d. copies of this vector, converge jointly to a non-degenerate limit distribution after proper linear normalization. More formally, let $(X_1^i, \dots, X_d^i), i \in \mathbb{N}$, be i.i.d. copies of the random vector $\mathbf{X} = (X_1, \dots, X_d)$. We assume that there exist sequences of constants $a_j^n > 0, b_j^n \in \mathbb{R}, 1 \le j \le d, n \in \mathbb{N}$ and a (in each margin non-degenerate) distribution function G, such that

$$\begin{aligned} &\lim_{n\to\infty} P\left(\frac{\max_{i=1,\dots,n} X_1^i - b_1^n}{a_1^n} \le x_1, \dots, \frac{\max_{i=1,\dots,n} X_d^i - b_d^n}{a_d^n} \le x_d\right) \\ &= G(x_1, \dots, x_d), \end{aligned} \tag{2.1}$$

for all continuity points $(x_1, \dots, x_d)$ of G. We say that the distribution of $\mathbf{X}$ is *in the max-domain of attraction of the extreme value distribution* G. A central result of extreme value theory is that this convergence can be broken down into two separate components. First, all marginal distribution functions $G_j, 1 \le j \le d$, of G are of the form

$$G_j(x) = \exp\left(-\left(1 + \gamma_j \frac{x - \mu_j}{\sigma_j}\right)^{-1/\gamma_j}\right), \quad 1 + \gamma_j \frac{x - \mu_j}{\sigma_j} > 0, \tag{2.2}$$

with $\gamma_j, \mu_j \in \mathbb{R}, \sigma_j > 0$, where for $\gamma_j = 0$ the right hand side is interpreted as $\exp(-\exp(-(x-\mu_j)/\sigma_j)), x \in \mathbb{R}$. The parameter γ_j, known as the *extreme value index*, is the most important parameter in describing the univariate extremal behavior of component j. Here a wealth of statistical procedures exists for univariate extremes, see de Haan and Ferreira (2007), Chapters 3 and 4, for an overview. Second, let F_j be the (continuous) marginal distribution function of X_j, $j = 1, \ldots, d$, then the convergence in (2.1) holds if and only if the standardized vector

$$\mathbf{Y} = \left(\frac{1}{1 - F_1(X_1)}, \ldots, \frac{1}{1 - F_d(X_d)} \right) \tag{2.3}$$

satisfies

$$\lim_{u \to \infty} P\left(\frac{\mathbf{Y}}{\|\mathbf{Y}\|} \in B \,\middle|\, \|\mathbf{Y}\| > u \right) = S(B), \tag{2.4}$$

for a probability measure S on $\mathbb{S}_+^{d-1} := \{\mathbf{x} \in [0,\infty)^d : \|\mathbf{x}\| = 1\}$, where $\|\cdot\|$ stands for an arbitrary but fixed norm, and B any S-continuity-Borel-set, see Beirlant et al. (2006), Chapter 8. The measure S in (2.4) thus describes the limiting behavior of the directions that we see in extremal observations from $\mathbf{Y}$. Furthermore, the measure S can be obtained from the limiting behavior of maxima as described in the following. The transformed $\mathbf{Y}$ satisfies

$$\lim_{n \to \infty} P\left(\frac{\max_{i=1,\ldots,n} Y_1^i}{n} \le x_1, \ldots, \frac{\max_{i=1,\ldots,n} Y_d^i}{n} \le x_d \right) = G_0(x_1, \ldots, x_d),$$

for i.i.d. copies $(Y_1^i, \ldots, Y_d^i), i \in \mathbb{N}$, of $\mathbf{Y}$, where G_0 is an extreme-value distribution with standard Fréchet margins (i.e., $\gamma_j = \sigma_j = 1, \mu_j = 0$ in (2.2)). For G_0 there exists a so-called *exponent measure* ν such that

$$G_0(x_1, \ldots, x_d) = \exp(-\nu\{(u_1, \ldots, u_d) \in [0,\infty)^d : \exists j : u_j > x_j\}) \tag{2.5}$$

for all $(x_1, \ldots, x_d) \in [0,\infty)^d$. This exponent measure is homogeneous of degree -1 and there exists a constant $c > 0$ such that

$$\nu\{\mathbf{u} \in [0,\infty)^d : \mathbf{u}/\|\mathbf{u}\| \in B, \|\mathbf{u}\| > y\} = cy^{-1} S(B) \tag{2.6}$$

for Borel sets $B \subset \mathbb{S}_+^{d-1}$ and $y > 0$. The same measure S appearing in (2.4) and (2.6) is called the *spectral measure* of $\mathbf{X}$ or $\mathbf{Y}$ and by the above it uniquely describes the dependence structure (or copula) of both exceedances and maxima. Due to the marginal standardization of the vector $\mathbf{Y}$ there always exists a constant $c > 0$ such that

$$\int_{\mathbb{S}_+^{d-1}} x_j S(d\mathbf{x}) = c \tag{2.7}$$

for all $j = 1, \ldots, d$. In this analysis, we are less interested in the marginal extremal behavior of individual components and more concerned with the extremal dependence structures. Hence we are interested in the structure of S.

Note that even for a vector $\mathbf{X}$ whose marginal distributions are not in the domain of attraction of a univariate extreme value distribution, it can still make sense to define (2.3) and look at (2.4) in order to describe its extremal behavior. On the other hand, there may also be situations where one does not want to standardize marginals first but treats observations as coming from a random vector $\mathbf{Y}$ which satisfies (2.4) with a spectral measure S that does not necessarily satisfy (2.7). Both in the un-standardized or the standardized case, the measure S tells us about the angle or direction of an observation that is considered extreme, either in the individual scale of each component or on a uniform scale. If there exist small sets which receive relatively high probabilities under S, these sets can be seen as "typical" directions for an extremal event. The idea of this paper is to identify these sets without assuming a specific underlying model, thereby identifying extremal patterns in a non-parametric way.

We have until now not specified which norm we meant when writing $\|\cdot\|$. The general equivalence between convergence of multivariate maxima in (2.1) and marginal convergences together with convergence of exceedances as described in (2.4) holds for any choice of norm $\|\cdot\|$, although the particular choice will of course affect the specific form of the spectral measure S. Depending on the particular application, there may be different possible choices for the particular norm. The main results in Section 3 are formulated in a general way that hold for any choice of norm. For our simulations in Section 4 and data examples in Section 5 we use the Euclidean norm $\|\cdot\|_2$, since this norm facilitates the interpretation of the spherical k-means algorithm, which is explained in the next section.

2.2. k-means and spherical k-means

The k-means clustering procedure is a way to identify distinct groups within a population. The name was first introduced in MacQueen (1967) although the ideas behind the algorithm date back further, see Bock (2008). The motivation is to identify cluster centers such that distances of the observations to their nearest cluster centers are minimized. Accordingly, all observations which are closest to the same cluster center are viewed as belonging to the same group.

In the following, let $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ be a distance function or, more generally, a dissimilarity function in $\mathbb{R}^d$ (see Gan et al. (2007), Chapter 6). For a probability measure P on $\mathbb{B}(\mathbb{R}^d)$, where $\mathbb{B}(\mathcal{C})$ stands for the Borel σ-algebra on the topological space $\mathcal{C}$, and a set $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$, $\mathbf{a}_i \in \mathbb{R}^d$ for $i = 1, \ldots, k$ and $k \in \mathbb{N}$, one can introduce the averaged distance from any observation to the closest element of A as

$$W(A, P) := \int_{\mathbb{R}^d} \min_{\mathbf{a} \in A} d(\mathbf{x}, \mathbf{a}) P(d\mathbf{x}) \in [0, \infty]. \tag{2.8}$$

For given P and k, a set A_k which minimizes $W(A, P)$ among all A with $|A| \leq k$, where $|A|$ stands for the cardinality of the (finite) set A, can be seen as

a set of theoretical cluster centers. Note that the set may not necessarily be unique.

If we replace P by its sample version P_n (i.e. the measure that places mass $1/n$ on each observation $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of a sample) in (2.8), and derive an accordingly optimal set A_k^n, its components minimize the sum of the distances from every observation to its nearest cluster center. While the original version of k-means uses the Euclidean distance, several alternative choices for d have been suggested. Recall that our central interest is in the spectral measure S in (2.4), which resides on the unit sphere $\mathbb{S}_+^{d-1}$, so that a natural way to measure the distance between two points is by their angle. This corresponds to the spherical k-means procedure of Dhillon and Modha (2001), which define $d(\cdot)$ in terms of angular dissimilarity,

$$d(\mathbf{x},\mathbf{y}) = d_\varphi(\mathbf{x},\mathbf{y}) := 1 - \cos(\mathbf{x},\mathbf{y}) = 1 - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} = 1 - \frac{\sum_{j=1}^d x_j y_j}{\sqrt{\sum_{j=1}^d x_j^2}\sqrt{\sum_{j=1}^d y_j^2}} \tag{2.9}$$

for $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$. If one restricts $\mathbf{x}$ and $\mathbf{y}$ to the Euclidean unit sphere, then this simplifies to $d_\varphi(\mathbf{x},\mathbf{y}) = 1 - \langle \mathbf{x},\mathbf{y}\rangle$ and has the advantage that the cosine dissimilarity can equally well be interpreted as Euclidean distance or spherical distance between two points, since bijections exist between all three measures.

In Section 3, we introduce the main results of this paper. Note that in order to allow for more flexibility in the choice of a suitable distance measure, the results only assume that $d : \mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1} \to [0,\infty)$ is a continuous function such that a unique minimizing set in (2.8) exists. If one prefers to study the spectral measure on a non-Euclidean unit sphere, the angular dissimilarity in (2.9) may be replaced by a more suitable function.

It should be noted that finding the optimal cluster centers for a given P can be an NP-hard problem (see Mahajan et al. (2012)) and the known iterative algorithms often depend crucially on the initial cluster centers, see Bradley and Fayyad (1998). For the examples and simulations in Sections 4 and 5 we rely on the `R`-package `skmeans` by Hornik et al. (2012), which provides short run-times and stable results.

3. Main result

In this section we formally introduce our estimation procedure which will, for a given sample, provide a set of empirical cluster centers. Each center can then be interpreted as a "dependence prototype" for a particular class of an extremal event. In brief, our procedure looks as follows:

1. With the help of the empirical distribution function, transform a sample from the distribution of $\mathbf{X}$ into (approximately) a sample from $\mathbf{Y}$ as in (2.3).
2. Choose a fraction of the latter that only keeps the transformed observations with largest norm.

3. For the chosen subsample, project the transformed observations onto the corresponding unit sphere.
4. Apply a spherical k-means procedure to the projected observations.

Note that steps 1.-3. in the above procedure generate a "pseudo-sample" from the spectral measure S of standardized observations. For the actual statistical inference of S, there exist methods in the literature from non-parametric (e.g. Einmahl et al. (2001), Einmahl and Segers (2009)), over semiparametric (e.g. Einmahl et al. (1997)) to fully parametric procedures (e.g. Coles and Tawn (1991)), for standardized and non-standardized data. The following theorem is therefore formulated in a way such that it holds for any estimator of the spectral measure as long as it is weakly or strongly consistent.

Theorem 3.1. *Assume that S is a probability measure on $\mathbb{B}(\mathbb{S}_+^{d-1})$ and that $S_n, n \in \mathbb{N}$, is a sequence of random probability measures on $\mathbb{B}(\mathbb{S}_+^{d-1})$ defined on a common probability space $(\Omega, \mathcal{A}, P)$. Furthermore, assume that $d : \mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1} \to [0, \infty)$ is a continuous function.*

For each S_n and a given value of $k \in \mathbb{N}$, denote by A_k^n a random set which minimizes

$$W(A, S_n) := \int_{\mathbb{S}_+^{d-1}} \min_{\mathbf{a} \in A} d(\mathbf{x}, \mathbf{a}) S_n(d\mathbf{x}) \tag{3.1}$$

among all sets $A \subset \mathbb{S}_+^{d-1}$ with at most k elements. Accordingly, if we replace S_n by S, denote the optimal set by A_k, and assume that for a given value of k, the set A_k is uniquely determined.

a) If $\int_{\mathbb{S}_+^{d-1}} f(\mathbf{x}) S_n(d\mathbf{x}) \to \int_{\mathbb{S}_+^{d-1}} f(\mathbf{x}) S(d\mathbf{x}), n \to \infty$, in probability for all continuous functions $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$, then A_k^n converges in probability to A_k as $n \to \infty$.

b) If $\int_{\mathbb{S}_+^{d-1}} f(\mathbf{x}) S_n(d\mathbf{x}) \to \int_{\mathbb{S}_+^{d-1}} f(\mathbf{x}) S(d\mathbf{x}), n \to \infty$, almost surely for all continuous functions $f : \mathbb{S}_+^{d-1} \to \mathbb{R}$, then A_k^n converges almost surely to A_k as $n \to \infty$.

Remark 3.2. *In the above theorem, the convergence of sets is formally meant in the Hausdorff distance d_H, but since all involved sets have only finitely many elements, it implies pointwise convergence of elements after a suitable reordering.*

Proof of Theorem 3.1. The argumentation below is similar to the one used in Pollard (1981, 1982), and the crucial ingredients for the proof are the continuity of d and the compactness of $\mathbb{S}_+^{d-1}$.

1. Set $\mathcal{E}_k := \{B \subset \mathbb{S}_+^{d-1}, |B| \leq k\}$. Continuity of d and compactness of $\mathbb{S}_+^{d-1}$ imply that $W_S : B \mapsto W(B, S)$ is continuous with respect to the Hausdorff-metric d_H on $\mathcal{E}_k$. As $\mathbb{S}_+^{d-1}$ is compact, so is $\mathcal{E}_k$ with respect to the Hausdorff-metric. Since A_k is uniquely determined, there exists for any neighbourhood $N \subset \mathcal{E}_k$ of A_k an $\eta > 0$ such that

$$W(B, S) > W(A_k, S) + \eta$$

for all $B \notin N$. Therefore, for all $\epsilon > 0$ there exists a $\delta > 0$ such that $|W(B,S) - W(A_k,S)| < \delta$ implies that $d_H(B,A_k) < \epsilon$. In the remainder of the proof we now show that

$$\begin{aligned}&|W(A_k^n,S) - W(A_k,S)|\\ \leq &|W(A_k^n,S) - W(A_k^n,S_n)| + |W(A_k^n,S_n) - W(A_k,S)| \to 0\end{aligned}$$

in the respective mode of convergence and the statements in a) and b) follow then from the above.

2. We first show that $|W(A_k^n,S) - W(A_k^n,S_n)| \to 0$.
As $\mathbb{S}_+^{d-1}$ is compact, we can, for a given $\epsilon > 0$, find m and $B_1,\dots,B_m \in \mathcal{E}_k$ such that

$$\min_{i=1,\dots,m} d_H(B,B_i) < \epsilon \tag{3.2}$$

for all $B \in \mathcal{E}_k$. By continuity of d and compactness of $\mathbb{S}_+^{d-1}$ the family of functions $B \mapsto \min_{\mathbf{b}\in B} d(\mathbf{x},\mathbf{b})$, $\mathbf{x} \in \mathbb{S}_+^{d-1}$, is uniformly equicontinuous with respect to the Hausdorff-metric and so with a suitable choice of ϵ in (3.2) this implies

$$\min_{i=1,\dots,m} \sup_{\mathbf{x}\in\mathbb{S}_+^{d-1}} |\min_{\mathbf{b}\in B} d(\mathbf{x},\mathbf{b}) - \min_{\mathbf{b}\in B_i} d(\mathbf{x},\mathbf{b})| < \delta$$

for a given $\delta > 0$ and all $B \in \mathcal{E}_k$.
From this we get that for each $B \in \mathcal{E}_k$ there exists one $i \in \{1,\dots,m\}$ such that

$$|W(B,P) - W(B_i,P)| \leq \int_{\mathbb{S}_+^{d-1}} |\min_{\mathbf{b}\in B} d(\mathbf{x},\mathbf{b}) - \min_{\mathbf{b}\in B_i} d(\mathbf{x},\mathbf{b})| P(d\mathbf{x}) < \delta$$

for all probability measures P on $\mathbb{B}(\mathbb{S}_+^{d-1})$.
Now, assumption a) or b) implies by continuity of d that

$$\max_{i=1,\dots,m} |W(B_i,S_n) - W(B_i,S)| \to 0, \quad n \to \infty, \tag{3.3}$$

in the respective mode of convergence.
Together, this gives

$$\begin{aligned}&|W(B,S_n) - W(B,S)|\\ \leq &\min_{i=1,\dots,m} |W(B,S_n) - W(B_i,S_n)\\ &\qquad + W(B_i,S_n) - W(B_i,S) + W(B_i,S) - W(B,S)|\\ \leq &\max_{i=1,\dots,m} |W(B_i,S_n) - W(B_i,S)| + 2\delta\end{aligned}$$

for all $B \in \mathcal{E}_k$ and thus

$$\sup_{B\in\mathcal{E}_k} |W(B,S_n) - W(B,S)| \to 0 \tag{3.4}$$

in the respective mode of convergence, which implies that $|W(A_k^n, S) - W(A_k^n, S_n)| \to 0$.

3. Finally, we show that $|W(A_k^n, S_n) - W(A_k, S)| \to 0$.
   Convergence (3.4) implies that

$$|\inf_{B\in\mathcal{E}_k} W(B, S_n) - \inf_{B\in\mathcal{E}_k} W(B, S)| \to 0.$$

   So,

$$|W(A_k^n, S_n) - W(A_k, S)| \to 0,$$

   in the respective mode of convergence. This concludes the proof. □

Since the idea of our approach is to detect general patterns without relying on a particular model, we choose for the rest of the analysis a straightforward non-parametric estimator of the spectral measure of standardized observations. This is a natural empirical counterpart to (2.4) and a slight modification of the estimator introduced in Einmahl et al. (2001).

First note that the spectral measure of $\mathbf{X}$ is defined in terms of the random vector $\mathbf{Y}$ from (2.3), but that we do not know the marginal distribution functions $F_j, 1 \le j \le d$. A solution is to replace F_j by the left-continuous version of the empirical distribution function

$$F_{j,n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_j^i < x\}}, \quad x \in \mathbb{R},$$

and transform the observations to

$$\hat{\mathbf{Y}}_i = (\hat{Y}_1^i, \ldots, \hat{Y}_d^i), \quad \text{with } \hat{Y}_j^i := (1 - F_{j,n}(X_j^i))^{-1}. \tag{3.5}$$

The empirical counterpart of (2.4) then motivates the estimator

$$\hat{S}_n(B) := \frac{\sum_{i=1}^n \mathbb{1}_{\left\{\|\hat{\mathbf{Y}}_i\| \ge \frac{n}{l_n}, \frac{\hat{\mathbf{Y}}_i}{\|\hat{\mathbf{Y}}_i\|} \in B\right\}}}{\sum_{i=1}^n \mathbb{1}_{\left\{\|\hat{\mathbf{Y}}_i\| \ge \frac{n}{l_n}\right\}}}, \tag{3.6}$$

for Borel subsets B of $\mathbb{S}_+^{d-1}$, where $l_n \in \mathbb{N}$ affects how many observations will be used for the estimator. Note that due to their definition the observed components of the $\hat{\mathbf{Y}}_i's$ will have values in $\{1, n/(n-1), \ldots, n/2, n\}$. Therefore, the value of l_n should grow with n in order to obtain a consistent estimator, but the growth rate should also not be too fast in order to catch only the extremal observations. Proposition 3.3 below gives necessary assumptions on l_n for the weak and strong consistency of this estimator.

Proposition 3.3. *Assume that $\mathbf{X}_i = (X_1^i, \ldots, X_d^i), i \in \mathbb{N}$, are i.i.d. copies of a vector $\mathbf{X}$, such that the transformed vector $\mathbf{Y}$ from (2.3) satisfies (2.4) with spectral measure S. For $S_n = \hat{S}_n$ as in (3.6), define the sets A_k^n and A_k as in Theorem 3.1 and assume that the set A_k is uniquely determined for a given value of k. Then, as $n \to \infty$,*

a) if $l_n/n \to 0$ and $l_n \to \infty$ the sets A_k^n converge to A_k in probability;
b) if $l_n/n \to 0$ and $l_n/\log(\log(n)) \to \infty$ the sets A_k^n converge to A_k almost surely.

Proof. The proposition follows from Theorem 3.1 if we can verify that the sequence of random measures S_n satisfies the corresponding assumptions. To see this, we argue similar to Einmahl et al. (2001), Theorem 1. We start by looking at the estimator

$$\hat{l}(x_1,\dots,x_d) := \frac{1}{l_n}\sum_{i=1}^{n} \mathbb{1}_{\{\exists j:\hat{Y}_j^i > \frac{n}{l_n x_j}\}}$$

for the so-called stable tail dependence function

$$l(x_1,\dots,x_d) := \nu\{(u_1,\dots,u_d) \in [0,\infty)^d : \exists j : u_j > 1/x_j\}, \tag{3.7}$$

with ν as introduced in (2.5). For $l_n/n \to 0, l_n \to \infty$ the estimator converges in probability (see Huang (1992), Theorem 1 in Chapter 2 or Theorem 7.2.1 in de Haan and Ferreira (2007)) and for $l_n/n \to 0, l_n/\log(\log(n)) \to \infty$ it converges almost surely (see Qi (1997), Theorem 1.2) for all $x_1,\dots,x_d \in (0,\infty)^d$. This pointwise convergence can be extended to show that

$$\frac{1}{l_n}\sum_{i=1}^{n} \mathbb{1}_{\left\{\|\hat{\mathbf{Y}}_i\|_\infty > \frac{n}{l_n c}, \hat{\mathbf{Y}}_i \frac{l_n}{n} \in \hat{B}\right\}} \to \nu\left\{\mathbf{u} \in [0,\infty)^d : \|\mathbf{u}\|_\infty > 1/c, \mathbf{u} \in \hat{B}\right\}$$

for each $c > 0$ and continuity set $\hat{B} \subset [0,\infty)^d$ in the respective mode of convergence, see the proof of Theorem 1 in Einmahl et al. (2001) for details. Due to equivalence of norms there exists a $c_0 > 0$ such that $\|\mathbf{x}\| > 1$ implies $\|\mathbf{x}\|_\infty > 1/c_0$ for the chosen norm $\|\cdot\|$. Set now $\hat{B} = \{\mathbf{u} \in [0,\infty)^d : \|\mathbf{u}\| > 1, \mathbf{u}/\|\mathbf{u}\| \in B\}$ for an S-continuity set $B \subset \mathbb{S}_+^{d-1}$ and $A = \{\mathbf{u} \in [0,\infty)^d : \|\mathbf{u}\| > 1\}$ so that

$$\begin{aligned}\hat{S}_n(B) &= \frac{\sum_{i=1}^{n} \mathbb{1}_{\left\{\|\hat{\mathbf{Y}}_i\|_\infty > \frac{n}{l_n c_0}, \hat{\mathbf{Y}}_i \frac{l_n}{n} \in \hat{B}\right\}}}{\sum_{i=1}^{n} \mathbb{1}_{\left\{\|\hat{\mathbf{Y}}_i\|_\infty > \frac{n}{l_n c_0}, \hat{\mathbf{Y}}_i \frac{l_n}{n} \in A\right\}}} \\ &\to \frac{\nu\left\{\mathbf{u} \in [0,\infty)^d : \|\mathbf{u}\|_\infty > 1/c_0, \mathbf{u} \in \hat{B}\right\}}{\nu\left\{\mathbf{u} \in [0,\infty)^d : \|\mathbf{u}\|_\infty > 1/c_0, \mathbf{u} \in A\right\}} = S(B)\end{aligned}$$

again in the respective mode of convergence. This implies the weak convergence either in probability or almost surely of $\hat{S}_n$ to S (see again Einmahl et al. (2001), Theorem 1) and thereby that the assumption a) or b), respectively, of Theorem 3.1 is satisfied. □

4. Application to max-linear models

The idea behind the decomposition of a spectral measure into k clusters is motivated by the special case where the spectral measure is clearly concentrated around k different centers. An idealized example is provided by the max-linear

model where a spectral measure has only mass on k different points. In the following we will demonstrate how our k-means procedure can be seen as an alternative way of estimating the corresponding model parameters.

A max-linear model consists of k different so-called *factors* $\mathbf{b}_i = (b_1^i, \dots, b_d^i) \in [0, \infty)^d, i = 1, \dots, k$ from which a random vector is generated by

$$\mathbf{X} = (X_1, \dots, X_d) = \left(\max_{i=1,\dots,k} b_1^i Z_i, \dots, \max_{i=1,\dots,k} b_d^i Z_i \right), \tag{4.1}$$

where $Z_1, \dots, Z_k$ are i.i.d. random variables with the same heavy-tailed distribution. The most common choice for this distribution is a standard Fréchet-distribution. Furthermore, one typically assumes that

$$\sum_{i=1}^{k} b_j^i = 1 \quad \text{for all } j = 1, \dots, d, \tag{4.2}$$

such that all margins of $\mathbf{X}$ are standard Fréchet as well.

Looking at (4.1) it is clear that the largest observations of $\mathbf{X}$ are due to a large observation of a Z_i and therefore the factors $\mathbf{b}_i$ determine the possible directions of extremal observations. In fact one can show that the spectral measure S concentrates on the points $\mathbf{a}_i = \mathbf{b}_i / \|\mathbf{b}_i\|$ with corresponding probabilities $p_i = \|\mathbf{b}_i\| / (\sum_{l=1}^{k} \|\mathbf{b}_l\|), 1 \le i \le k$. On the other hand, for each discrete spectral measure with mass concentrated on k points there exists a max-linear model with k factors which results in this given spectral measure, see Yuen and Stoev (2014). It is also shown that any given dependence structure of extremes can be approximated arbitrarily well by spectral measures generated from max-linear models if one allows the number of factors to grow, see Fougères et al. (2013). Furthermore, it was recently shown in Gissibl (2018); Gissibl and Klüppelberg (2018); Gissibl et al. (2018) that max-linear models also evolve from a natural modeling of extremal dependence generated from a directed acyclic graph of components, thereby allowing the modeling and detection of causality in extreme events.

Parameter estimation of max-linear models has proven to be a difficult task due to the fact that no spectral density exists which excludes standard maximum likelihood procedures. Instead, Einmahl et al. (2012), Einmahl et al. (2016) and Einmahl et al. (2018) use a least squares estimator based on the stable tail dependence function to estimate parameters from extremal observations and Yuen and Stoev (2014) construct a least squares estimator based on the joint distribution function and make use of all observations. In the following we illustrate how the k-means procedure serves as an alternative and effective way of inference for max-linear models.

From the above it follows that the discrete spectral measure, i.e., the points $\mathbf{a}_1, \dots, \mathbf{a}_k \in \mathbb{S}_+^{d-1}$ on which the spectral measure S is concentrated, and the corresponding probabilities $p_1, \dots, p_k$, are parametrizing a max-linear model. For such a spectral measure, it is clear that $W(A, S)$ as defined in (2.8) is minimized by choosing $A = \{\mathbf{a}_1, \dots, \mathbf{a}_k\}$ and A is uniquely determined. For an

estimator S_n of S, which satisfies the assumptions from the previous section, the k-means cluster centers can therefore be seen as consistent estimators of $\mathbf{a}_1, \ldots, \mathbf{a}_k$. This consistency also implies that the percentages of points which are classified as belonging to cluster i converge to the corresponding probability $p_i, 1 \leq i \leq k$. Especially if one is interested in using the max-linear model as an approximation for the largest observations only, the estimation of the $\mathbf{a}_i$'s and p_i's can be seen as an alternative to the estimation of the components $\mathbf{b}_i$'s.

In order to compare the spherical k-means procedure with the previously mentioned approaches we set up a small simulation study, where we generate a random parameter constellation for a max-linear model with d dimensions and k factors. On the one hand, we estimate the factor coefficients according to Einmahl et al. (2016) and Einmahl et al. (2018), as provided by the `R`-package Kiriliouk (2016), and Yuen and Stoev (2014), as provided by Yuen (2015), and derive the estimators for $\mathbf{a}_1, \ldots, \mathbf{a}_k$ and $p_1, \ldots, p_k$ from the resulting estimated spectral measures. On the other hand, we apply directly the k-means estimator which provides estimators for the cluster centers $\hat{\mathbf{a}}_i$, and we estimate $p_1, \ldots, p_k$ by the the percentage of extreme observations that are classified as belonging to cluster i. For all procedures, the estimator $\hat{S}$ of the spectral measure S is determined by k points of mass, $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_k$ and corresponding probabilities $\hat{p_1}, \ldots, \hat{p_k}$. The difference between the true spectral measure S and the corresponding estimator is evaluated by two criteria. The first criterion evaluates

$$d_s(S, \hat{S}) := \min_{\pi : \pi \text{ is permutation of } \{1, \ldots, k\}} \sqrt{\sum_{i=1}^{k} \|\hat{a}_{\pi(i)} - a_i\|_2^2},$$

which can be seen as a distance measure similar to the Hausdorff distance applied to finite sets, but taking into account all distances of the matched vectors instead of only the maximal one. This gives an idea about how well the estimator identifies possible extremal directions, but does not take into account their frequencies. We also look at a metric on the space of probability measures, where we use the Wasserstein metric with $p = 1$, which is defined as

$$W_1(S, \hat{S}) := \inf_{P \in \Gamma(S, \hat{S})} \int_{\mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1}} \|\mathbf{x} - \mathbf{y}\|_2 P(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}),$$

where $\Gamma(S, \hat{S})$ is the set of all probability measures on $\mathbb{S}_+^{d-1} \times \mathbb{S}_+^{d-1}$ with first marginal S and second marginal $\hat{S}$. We use the `R`-package `transport`, see Schuhmacher et al. (2019), for evaluation of Wasserstein distances.

In the first set of simulations, we assume the number k of factors (see (4.1)) to be known and therefore the number of cluster centers in our algorithm is fixed as k. For each combination of dimension d and number of factors k we randomly generate 100 model specifications and for each model specification we generate 1000 observations. The random factors are specified as below, where $U_i, i \in \mathbb{N}$, stand for i.i.d. random variables with uniform distribution on $[0, 1]$.

TABLE 1
Comparison of $d_s(S, \hat{S})$ for different values of d and k. Mean over 100 simulations, standard deviation in brackets.

Method	$d = 4, k = 2$	$d = 4, k = 6$	$d = k = 6$	$d = 10, k = 6$
spherical k-means	0.0562(0.0276)	0.4042(0.2693)	0.3376(0.2050)	0.3711(0.2306)
Einmahl et al. '16	0.0728(0.0415)	0.5255(0.3198)	0.4295(0.2506)	0.4264(0.2694)
Einmahl et al. '18	0.0605(0.0309)	0.4868(0.3001)	0.3971(0.2271)	0.4103(0.2312)
Yuen & Stoev '14	0.1370(0.1905)	1.1049(0.2052)	1.3808(0.1541)	1.8755(0.1582)

TABLE 2
Comparison of $W_1(S, \hat{S})$ for different values of d and k. Mean over 100 simulations, standard deviation in brackets.

Method	$d = 4, k = 2$	$d = 4, k = 6$	$d = k = 6$	$d = 10, k = 6$
spherical k-means	0.0450(0.0206)	0.1310(0.0285)	0.1393(0.0302)	0.1578(0.0332)
Einmahl et al. '16	0.0629(0.0351)	0.1445(0.0344)	0.1464(0.0328)	0.1592(0.0345)
Einmahl et al. '18	0.0458(0.0219)	0.1310(0.0314)	0.1386(0.0329)	0.1569(0.0348)
Yuen & Stoev '14	0.0725(0.0977)	0.3336(0.0876)	0.4860(0.0625)	0.6764(0.0648)

We only state the first $k-1$ factors, as the last factor is always determined from the first ones by the standardization assumption (4.2).

- $d = 4$, $k = 2$: First factor is $(U_1, U_2, U_3, U_4)/2$.
- $d = 4$, $k = 6$: First five factors are $(U_1, U_2, U_3, U_4)/3$, $(U_5, 0, U_6, 0)/3$, $(0, U_7, 0, U_8)/3$, $(U_9, U_{10}, 0, 0)/3$, $(0, 0, U_{11}, U_{12})/3$.
- $d = k = 6$: First five factors are $(U_1, \ldots, U_6)/3$, $(0, U_7, 0, U_8, 0, U_9)/3$, $(U_{10}, 0, U_{11}, 0, U_{12}, 0)/3$, $(0, 0, 0, U_{13}, U_{14}, U_{15})/3$, $(U_{16}, U_{17}, U_{18}, 0, 0, 0)$.
- $d = 10$, $k = 6$: First five factors are $(U_1, \ldots, U_{10})/2, (U_{11}, U_{12}, 0, \ldots, 0)/2$, $(0, 0, U_{13}, U_{14}, 0, 0, 0, 0, 0, 0)/2$, $(0, 0, 0, 0, U_{15}, U_{16}, 0, 0, 0, 0)/2$, $(0, 0, 0, 0, 0, 0, 0, U_{17}, U_{18}, U_{19}, U_{20})/2$.

For the method from Yuen and Stoev (2014) we use all observations, for the others only the 100 with largest norm. The grid for the estimator from Einmahl et al. (2018) includes all d-dimensional vectors with entries from the set $\{0, 1/3, 2/3, 1\}$ and exactly 2 non-zero entries. A finer grid would have implied very long run times. For those three estimators, the fact that there are 0's in the factors and knowledge of their positions is not used for the estimation, so there are $d \cdot (k-1)$ parameters to estimate. For the (spherical) k-means procedure, the 100 observations with largest Euclidean norm have been projected on the Euclidean unit sphere and we applied the procedure `skmeans` from Hornik et al. (2012).

The average values of $d_s(S, \hat{S})$ and $W_1(S, \hat{S})$ over all 100 realizations for different constellations are shown in Tables 1 and 2, with the observed standard deviations in brackets. It can be seen that the locations of the points of mass of the spectral measure are most precisely estimated by the spherical k-means procedure. The accuracy in estimating the spectral measure is very similar for all methods except for the method from Yuen and Stoev (2014), which was constructed for an overall good fit but with little focus on extremes.

Since the number of factors is usually not known a priori, we also look at two misspecified models, where in both cases the model was fitted as if there

TABLE 3
Comparison of $W_1(S, \hat{S})$ for true model parameters $d = 4, k = 2$ and $d = 4, k = 6$ but with models fitted to $k = 3$ instead. Mean over 100 simulations, standard deviation in brackets.

Method	$d = 4, k = 2$	$d = 4, k = 6$
spherical k-means	0.0504(0.0226)	0.2746(0.0537)
Einmahl et al. '16	0.0571(0.0270)	0.3225(0.0646)
Einmahl et al. '18	0.0532(0.0230)	0.2921(0.0663)
Yuen & Stoev '14	0.0755(0.0649)	0.4230(0.1039)

were $k = 3$ factors, but the true value of k is 2 or 6. The random models are generated as described previously for the respective constellations of d and k.

In Table 3 we see the results for the two misspecified models as measured in the W_1-metric of estimated and true spectral measure. In the two examples, the spherical k-means procedure copes better with the fact that our model is misspecified.

Regarding the numerical implementation of the estimators Einmahl et al. (2016) and Einmahl et al. (2018) we noted in our simulations that the first depends for larger values of d and k heavily on the starting parameters of the algorithm, where we choose the starting value for factor parameters from the spherical k-means estimator, as also suggested in Einmahl et al. (2012). The procedure for the estimator from Einmahl et al. (2018) has very long runtimes, even with the relatively coarse grid that we use.

We conclude from the simulations that the spherical k-means procedure is, for the chosen examples and metrics, superior or competitive to (and usually faster than) methods for estimating max-linear models if one is mainly interested in the resulting spectral measure of observations.

5. Data examples

In the following we apply our procedure to three different data sets with dimensions 5, 30 and 38. For all data sets we observe that in each estimated cluster center $\mathbf{a}_i, i = 1, \ldots, k$, there are many components with values close to 0 and only a few with significantly positive entries. This hints at the phenomenon of asymptotic independence. In extreme value theory, two random variables X, Y with distributions F_X, F_Y are called *asymptotically independent* if

$$\lim_{u \nearrow 1} P(X > F_X^{-1}(u) | Y > F_Y^{-1}(u)) = 0,$$

(assuming that the limit exists) and *asymptotically dependent* otherwise. Detecting the groups of random variables which share asymptotic dependencies and classify them accordingly was the main aim of Chautru (2015). Our analysis allows for a more gradual view on dependencies since looking at the estimated clusters and observed differences in cluster components gives an idea about strong and weak hints towards asymptotic dependence or independence. For groups of asymptotically dependent variables it furthermore allows to identify patterns within those groups.

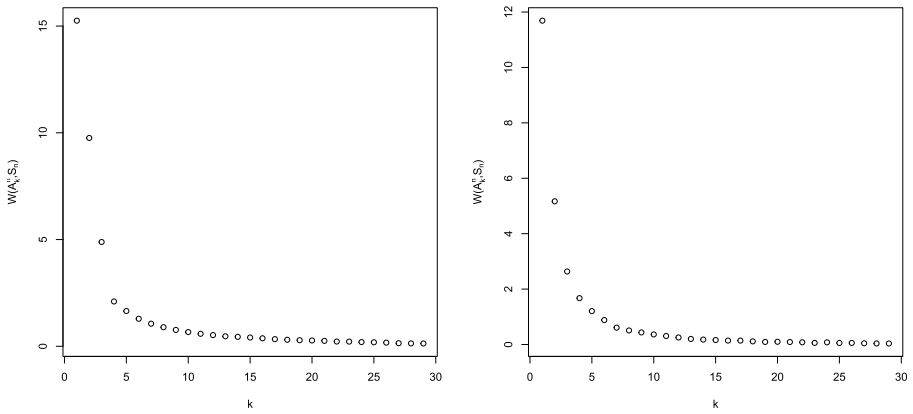

FIG 1. *The value of the minimized mean distance $W(A_k^n, S_n)$, see (3.1), for different values of k in the air pollution data sets. Left: summer data. Right: winter data. For values of k larger than 5 the curve becomes rather flat.*

5.1. Air pollution data

We start with the air pollution data which has been analyzed in Heffernan and Tawn (2004). This dataset is of relatively small dimension and consists of daily measurements of five air pollutants in the city center of Leeds (U.K.), collected between 1994 and 1998 and split up into summer and winter months which gives 578 and 532 observations, respectively. The data is available via the `R`-package `texmex`, see Southworth et al. (2018). We apply the four-step procedure as outlined in the beginning of Section 3, with the transformation of marginals as described in (3.5). For this data set, we use the 10% of transformed observations with largest Euclidean norm, project them on the unit sphere and apply the spherical k-means procedure. For this and the other two data examples we use the command `skmeans` with method `pclust` and additional parameters `nruns = 1000, maxchains=100` from Hornik et al. (2012).

The first step is now to determine suitable values of the number of clusters k. A common way of doing this is creating a so-called "elbow plot" by plotting the minimized distance $W(A_k^n, S_n)$ (recall from (3.1)) against k, see Figure 1 for the summer and winter data. Note that $W(A_k^n, S_n)$ necessarily decreases with the increase of k. In the plot one usually looks for a k such that for larger values the decrease becomes insignificant, but we note that there is no clear theoretical criterion for an optimal choice of k. Our goal is to use the algorithm as a tool to explore the structure in the data and we stress that it often makes sense to look at different values of k in the analysis.

From the elbow plots of the summer and winter data we decide to pursue our analysis for both $k = 4$ and $k = 5$. For these choices of k, we illustrate the cluster centers $\mathbf{a}_1, \ldots, \mathbf{a}_k$ in colored heat maps in Figures 2 and 3.

In the graphs, the cluster centers are re-normalized such that the maximum

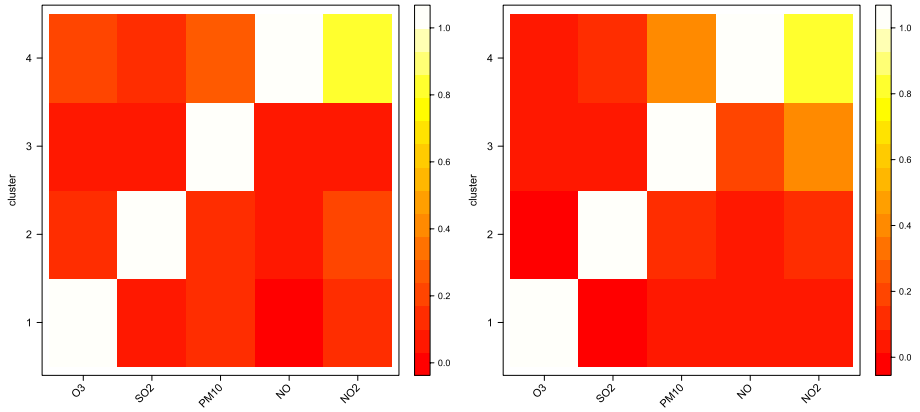

FIG 2. *The k-means clustering result on the air pollution data for $k = 4$. Left: summer data. Right: winter data. Here, each row in the corresponding picture corresponds to one of the four estimated cluster centers, where values have been normalized according to (5.1). That means the lighter the colour in a box, the larger is the value of the corresponding component relative to all other components in this cluster center.*

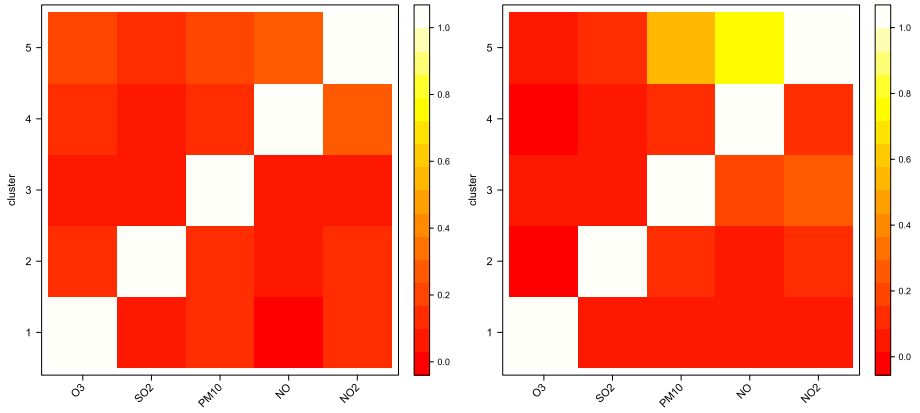

FIG 3. *The k-means clustering result on the air pollution data for $k = 5$. Left: summer data. Right: winter data. Here, each row in the corresponding picture corresponds to one of the five estimated cluster centers, where values have been normalized according to (5.1). That means the lighter the colour in a box, the larger is the value of the corresponding component relative to all other components in this cluster center.*

component is scaled to 1 to provide better visual comparison, i.e.,

$$\mathbf{a}_i = (a_1^i, \ldots, a_d^i) \to \left(\frac{a_1^i}{\max_{h=1,\ldots,d}\{a_h^i\}}, \ldots, \frac{a_d^i}{\max_{h=1,\ldots,d}\{a_h^i\}} \right), \quad i = 1, \ldots, k. \tag{5.1}$$

Hence in each heat map, each row corresponds to a cluster center and the white/bright yellow entries indicate the components that are largest.

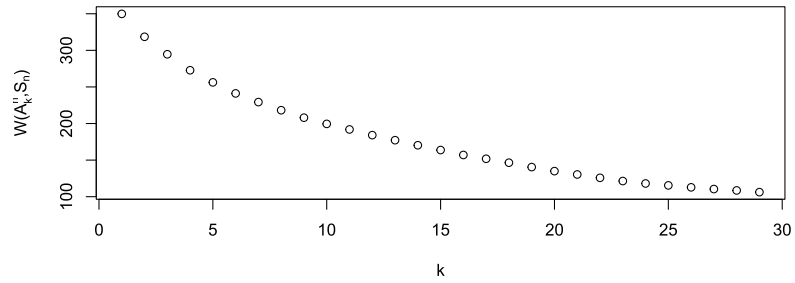

FIG 4. *The value of the minimized mean distance $W(A_k^n, S_n)$, see (2.8), for different values of k in the financial portfolio loss data. There exists no clear indication for an optimal value of k.*

For $k = 5$ we see that for the summer data each cluster has exactly one large component, hinting at asymptotic independence between all air pollutants. We note that due to the marginal transformation in (3.5) and the standardization property (2.7) of the spectral measure S, the components of our projections on the unit sphere should all have approximately the same expected value. Therefore, each component should correspond to a large entry in at least one cluster center. As a result, we note that for $k = 4$ there has to be at least one cluster center where at least two components are large. We can interpret cluster center 4 in the summer data for $k = 4$ in the way that NO and NO_2 are the air pollutants which are most likely to occur at extreme levels simultaneously. Both for $k = 4$ and for $k = 5$ we can identify a tendency for particulate matter (PM_{10}), NO and NO_2 to occur jointly at extreme levels, but only in winter. This is in line with the conclusions in Heffernan and Tawn (2004) who state asymptotic independence for all components except for PM_{10}, NO and NO_2 in winter.

5.2. Financial portfolio losses

In this example, we consider the 'value-averaged' daily returns of 30 industry portfolios compiled and posted as part of the Kenneth French Data Library. The data in consideration span between 1950–2015 with $n = 16694$ observations. This is the same dataset as analyzed in Cooley and Thibaud (2019), where dependencies in extreme losses were explored with a method related to principle component analysis. Here we attempt to recover more information using our method. Since we are interested in extremal losses we first multiply all returns by -1. After that we use the same procedure as for the previous data set, but this time we only look at the transformed observations with the largest 5% of Euclidean norms.

We create again an "elbow plot", see Figure 4. Here it is rather difficult to find a concrete value for k so we compare the analysis for $k = 5$ and $k = 10$.

The top graph of Figure 5 illustrates the cluster centers with $k = 5$. We

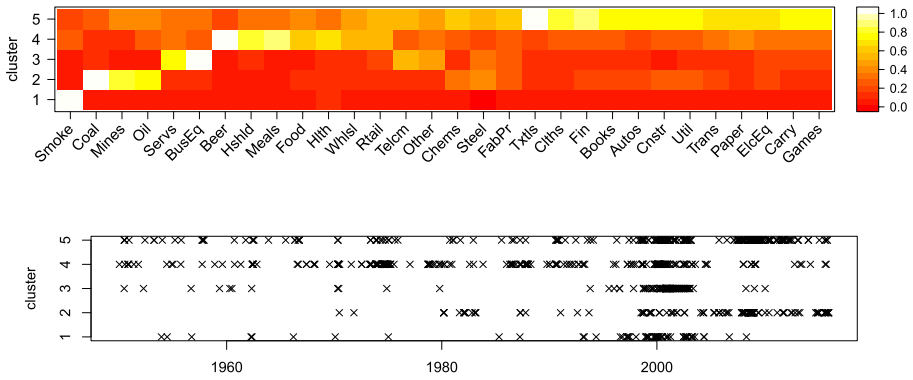

FIG 5. *The k-means clustering result on the financial portfolio loss data for $k = 5$. Top: Each row in the corresponding picture corresponds to one of the five estimated cluster centers, where values have been normalized according to* (5.1). *That means the lighter the colour in a box, the larger is the value of the corresponding component relative to all other components in this cluster center. Bottom: cluster classification result vs. time.*

can see that the clusters clearly separate the categories into different sectors. Cluster 1 signals the asymptotic independence of the tobacco industry to all other categories. Cluster 2 focuses on the energy and material sectors. Cluster 3 consists of business and IT related industries. Cluster 4 consists of the consumer oriented industries and Cluster 5 encompasses the rest.

The bottom graph of Figure 5, which shows the time points of the extreme losses in each cluster, also provides interesting insights. The extreme losses for Cluster 1 occurred around 2000, when massive lawsuits surged against tobacco companies. For Cluster 2, since the turn of the millennium, the U.S. coal and mining industries have been more heavily affected by government regulations, the rise of alternative sources of energy and foreign imports, which led to many struggles in the industry. The timeline for Cluster 3 clearly indicates the dot-com bubble in the late nineties. The consumer goods of Cluster 4 were heavily affected by U.S. recessions, most prominently by the oil crisis in 1973. Cluster 5 can be explained as widespread effects of the dot-com bubble and financial crisis.

Figure 6 shows the same set of results for $k = 10$. There is a clear pattern where most cluster centers have only one or few large components, hinting at an overall strong level of asymptotic independence. We can clearly identify sectors which are more prone to exhibiting joint losses, and that many of the connections from the analysis with $k = 5$ remain. Still tightly linked are the consumer sector (Cluster 5), the energy sector (Cluster 7), the business and IT sectors (Cluster 8) and a wide collection of industries linked to the financial industry (Cluster 10). The timeline plot also shows similar patterns to the previous results and clearly identifies the major events in the financial market history.

We would like to point out that a time series over such a long horizon cannot

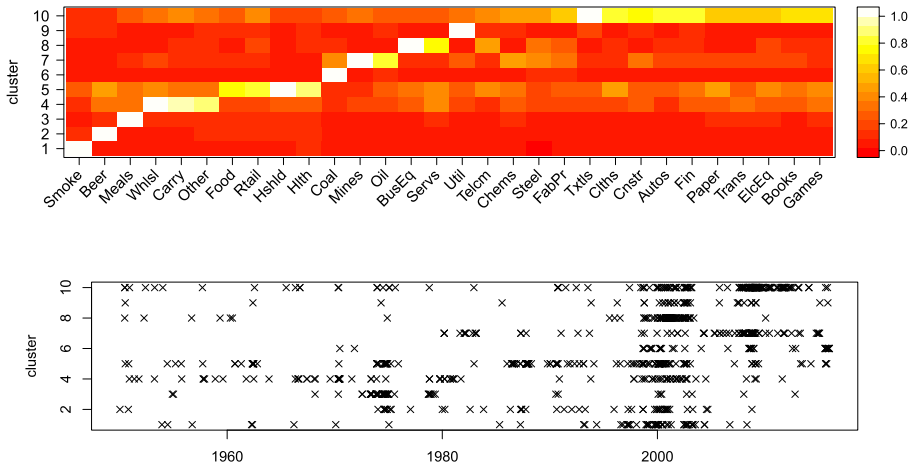

FIG 6. *The k-means clustering result on the financial portfolio loss data for* $k = 10$. *Top: Each row in the corresponding picture corresponds to one of the ten estimated cluster centers, where values have been normalized according to* (5.1). *That means the lighter the colour in a box, the larger is the value of the corresponding component relative to all other components in this cluster center. Bottom: cluster classification result vs. time.*

be safely assumed to be stationary and in fact the cluster classification results vs. time indicate just that, namely that the dependence structure of extremal observations changes over time. One can therefore not estimate cluster centers of the current or even future spectral measure but rather of the time averaged measure over the last 65 years. Futhermore, our estimation procedure for the spectral measure suffers from the fact that our observations are not i.i.d., in contrast to the assumptions of Section 3. This problem also applies to the example from Section 5.1.

5.3. Dietary intakes data

In this section we look at the dietary interview from the 2015–2016 NHANES report, available at `https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DR1TOT_I.XPT`. The interview component, called "What We Eat in America", recorded the food and beverage consumed by all participants during the 24 hours period prior to the interview. The resulting dataset describes the nutrients information calculated from these observations. We are interested in the dependency of 38 chosen nutrients in high-level intakes, as high doses of some of the components can have negative health effects. See also Chautru (2015) for the analysis of a similar, but smaller data set.

We derive again an estimator for the spectral measure by transforming observations with the help of the empirical distribution function and keeping the transformed observations whose Euclidean norm belongs to the largest 5%. The choice of k is again ambiguous in this data set and the elbow plot is similar

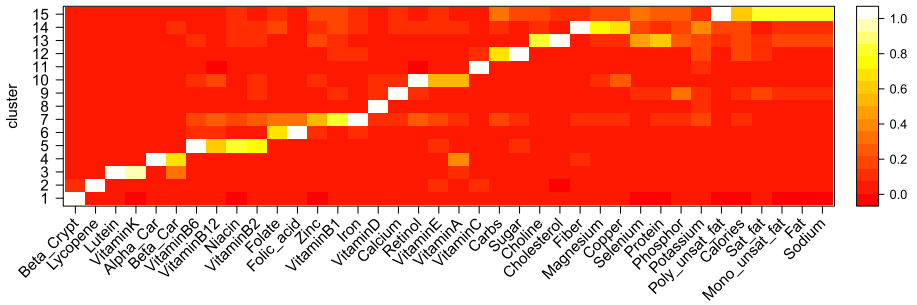

FIG 7. *The k-means cluster centers in the dietary intakes data for* $k = 15$. *Top: Each row in the corresponding picture corresponds to one of the 15 estimated cluster centers, where values have been normalized according to* (5.1). *That means the lighter the colour in a box, the larger is the value of the corresponding component relative to all other components in this cluster center.*

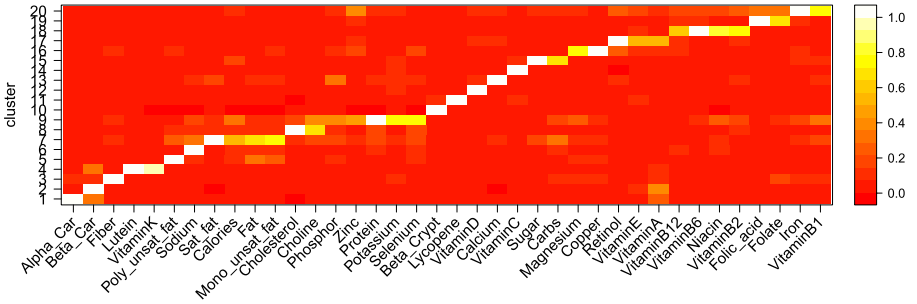

FIG 8. *The k-means cluster centers in the dietary intakes data for* $k = 20$. *Top: Each row in the corresponding picture corresponds to one of the 20 estimated cluster centers, where values have been normalized according to* (5.1). *That means the lighter the colour in a box, the larger is the value of the corresponding component relative to all other components in this cluster center.*

to that in Figure 4. What is clear is that as k increases the number of cluster centers with only one large component increases, again pointing at asymptotic independence of most of the nutrients. Significant clusters for several values of k can nevertheless be identified as clusters formed by carbs and sugar, by vitamin B_2, Vitamin B_6, Vitamin B_{12} and niacin, by lutein and vitamin K, by iron and vitamin B_1 and finally by fat together with fatty acids which goes also hand in hand with high values of intaken calories.

6. Summary and discussion

In this paper, we introduced a new procedure to describe and analyze extremal dependence of random vectors in a concise way. The key idea behind the method is the application of spherical k-means procedure to the estimated spectral mea-

sure. In Section 3, we showed that under suitable convergence of the estimated spectral measure, the corresponding empirical cluster centers converge to their theoretical counterparts. This provides a new inference procedure for the class of max-linear models which we showed to be competitive or even superior to existing methods, see Section 4. Finally, our data examples in Section 5 illustrated how cluster centers can be interpreted as "extremal prototypes" which reveal dependence structures in the largest observations.

Cluster centers with mainly one large component hint at asymptotic independence between components, while cluster centers with several components being large at the same time hint at asymptotic dependence. Though we do not provide a rigorous statistical test, our method can be used as an explorative first step to give a quick overview over dependence structures. Plots of the cluster centers like the ones presented in Section 5 provide a visual summary which is easily accessible. Based on the results of the suggested procedure, a more formal analysis, for example by fitting a suitable parametric model to the spectral measure, can be applied as a next step. We would like to point out that the selection of a threshold to identify the largest observations and the choice of a suitable value of k are crucial and that one should check robustness of results with respect to these parameters.

We demonstrated that our method could provide insights on data examples of moderate dimensions. One of the future directions would be to accommodate examples with higher dimensions, possibly by combining multiple dimension reduction techniques, such as done in Chautru (2015), or using building blocks from Cooley and Thibaud (2019) and Morris et al. (2019).

Acknowledgements

The authors thank Dan Cooley, Clément Dombry, Holger Drees, Henrik Hult and Chen Zhou for valuable discussions and suggestions regarding this manuscript.

References

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006. MR2108013

E. Bernard, P. Naveau, M. Vrac, and O. Mestre. Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of Climate*, 26(20):7929–7937, 2013.

H.-H. Bock. Origins and extensions of the k-means algorithm in cluster analysis. *Journal Electronique d'Histoire des Probabilités et de la Statistique Electronic Journal for History of Probability and Statistics*, 4:48–49, 2008. MR2471974

P. S. Bradley and U. M. Fayyad. Refining initial points for k-means clustering. In *ICML*, volume 98, pages 91–99. Citeseer, 1998.

E. Chautru. Dimension reduction in multivariate extreme value analysis. *Electron. J. Statist.*, 9(1):383–418, 2015. MR3323204

M. Chiapino, A. Sabourin, and J. Segers. Identifying groups of variables with the potential of being large simultaneously. *Extremes*, Jan 2019. ISSN 1572-915X. MR3949044

S. G. Coles and J. A. Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):377–392, 1991. MR1108334

D. Cooley and E. Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106:587–604, 2019. MR3992391

A. Davison and R. Huser. Statistics of extremes. *Annual Review of Statistics and Its Application*, 2(1):203–235, 2015.

A. C. Davison, S. A. Padoan, M. Ribatet, et al. Statistical modeling of spatial extremes. *Statistical science*, 27(2):161–186, 2012. MR2963980

L. de Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007. MR2234156

I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175, 2001.

J. H. Einmahl and J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989, 2009. MR2541452

J. H. Einmahl, L. de Haan, and A. K. Sinha. Estimating the spectral measure of an extreme value distribution. *Stochastic Processes and their Applications*, 70(2):143–171, 1997. MR1475660

J. H. Einmahl, L. de Haan, and V. I. Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Statist.*, 29(5):1401–1423, 10 2001. MR1873336

J. H. Einmahl, A. Krajina, and J. Segers. An M-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764–1793, 2012. MR3015043

J. H. Einmahl, A. Kiriliouk, A. Krajina, and J. Segers. An M-estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):275–298, 2016. MR3453656

J. H. Einmahl, A. Kiriliouk, and J. Segers. A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, pages 1–29, 2018. MR3800297

A.-L. Fougères. Multivariate extremes. In *Extreme values in finance, telecommunications, and the environment*, pages 373–388. Chapman and Hall/CRC, 2003.

A.-L. Fougères, C. Mercadier, and J. P. Nolan. Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, 116:109–129, 2013. MR3049895

G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Siam, 2007. MR2331172

N. Gissibl. *Graphical Modeling of Extremes: Max-linear Models on Directed Acyclic Graphs*. PhD thesis, Technical University of Munich, 2018.

N. Gissibl and C. Klüppelberg. Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693–2720, 2018. MR3779699

N. Gissibl, C. Klüppelberg, and M. Otto. Tail dependence of recursive max-linear models with regularly varying noise variables. *Econometrics and statistics*, 6:149–167, 2018. MR3797980

N. Goix, A. Sabourin, and S. Clémençon. Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12 – 31, 2017. ISSN 0047-259X. MR3698112

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. MR2722294

S. Haug, C. Klüppelberg, and G. Kuhn. Copula structure analysis based on extreme dependence. *Statistics and Its Interface*, 8:93–107, 2015. MR3320392

J. E. Heffernan and J. A. Tawn. A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546, 2004. MR2088289

K. Hornik, I. Feinerer, M. Kober, and C. Buchta. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012. .

X. Huang. *Statistics of bivariate extreme values*. PhD thesis, Erasmus University Rotterdam, 1992.

A. Kiriliouk. *tailDepFun: Minimum Distance Estimation of Tail Dependence Models*, 2016. URL https://CRAN.R-project.org/package=tailDepFun. R package version 1.0.0.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press. MR0214227

M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. *Theoretical Computer Science*, 442:13–21, 2012. MR2927097

S. A. Morris, B. J. Reich, and E. Thibaud. Exploration and inference in spatial extremes using empirical basis functions. *Journal of Agricultural, Biological and Environmental Statistics*, 24:555–572, 2019. MR4020704

D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135–140, 01 1981. URL https://doi.org/10.1214/aos/1176345339.

D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205, 1982. MR0651814

Y. Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. *Acta Mathematicae Applicatae Sinica*, 13(2):167–175, Apr 1997. ISSN 1618-3932. MR1443948

D. Schuhmacher, B. Bähre, C. Gottschlich, V. Hartmann, F. Heinemann, and B. Schmitzer. *transport: Computation of Optimal Transport Plans and Wasserstein Distances*, 2019. URL https://cran.r-project.org/package=transport. R package version 0.11-1.

H. Southworth, J. E. Heffernan, and P. D. Metcalfe. *texmex: Statistical modelling of extreme values*, 2018. R package version 2.4.2.

R. Yuen. R-code for fitting max-linear models via minimum CRPS. http://hdl.handle.net/2027.42/110774, 2015.

R. Yuen and S. Stoev. CRPS M-estimation for max-stable models. *Extremes*, 17(3):387–410, 2014. MR3252818