Submitted to the Annals of Statistics

KERNEL PCA FOR MULTIVARIATE EXTREMES

BY MARCO AVELLA MEDINA[1,a], RICHARD
A. DAVIS[1,b] AND GENNADY SAMORODNITSKY[2,c]

[1]*Department of Statistics, Columbia University,* [a]*marco.avella@columbia.edu;* [b]*rdavis@stat.columbia.edu*

[2]*School of Operations Research and Information Engineering, Cornell University,* [c]*gs18@cornell.edu*

We propose kernel PCA as a method for analyzing the dependence structure of multivariate extremes and demonstrate that it can be a powerful tool for clustering and dimension reduction. Our work provides some theoretical insight into the preimages obtained by kernel PCA, demonstrating that under certain conditions they can effectively identify clusters in the data. We build on these new insights to characterize rigorously the performance of kernel PCA based on an extremal sample, i.e., the angular part of random vectors for which the radius exceeds a large threshold. More specifically, we focus on the asymptotic dependence of multivariate extremes characterized by the angular or spectral measure in extreme value theory and provide a careful analysis in the case where the extremes are generated from a linear factor model. We give theoretical guarantees on the performance of kernel PCA preimages of such extremes by leveraging their asymptotic distribution together with Davis-Kahan perturbation bounds. Our theoretical findings are complemented with numerical experiments illustrating the finite sample performance of our methods.

1. Introduction.

The study and modeling of the behavior of extremes continues to generate increasing interest from scientists in a variety of fields including environmental, industrial, economic and social media related activities. While extremes are reasonably well understood for univariate and low dimensional data, it remains very challenging to model multivariate extremes when one or more of rare extreme events may occur simultaneously. An important recent line of work in multivariate extreme value theory seeks to connect this literature to ideas from modern statistics and machine learning. This task is not at all trivial since the dependence structure between extreme observations can be very complex and involve notions of dependence that differ from the typical ones arising in the non-extreme world. Work in this direction has included adapting various notions of sparsity for extremes [18, 24, 32], concentration inequalities [17, 8], conditional independence [12, 14], causality [16, 10] and unsupervised learning [7, 9, 21, 11, 1, 20, 15, 28], to name a few important examples. See also [13] for a review of recent developments in the literature of multivariate extremes. Our work is aligned with this direction of research as we propose kernel PCA as a preprocessing tool that facilitates clustering multivariate extremes.

The covariance matrix plays a central role in non-extremal statistics as it is widely used to quantify the linear dependence among random variables. The eigen-decomposition of the covariance matrix is the building block of principal components analysis (PCA), which in turn is one of the most popular dimension reduction techniques in statistics. It can be used to find low dimensional projections of p-dimensional data into the linear subspace spanned by $k < p$ eigenvectors associated with the k largest eigenvalues of the empirical covariance matrix. This projection corresponds to the best k-dimensional projection in the sense of being

MSC2020 subject classifications: Primary ???, ???; secondary ???.

Keywords and phrases: Angular measure, heavy tails, Kernel PCA, perturbation bounds, regular variation, reproducing kernel Hilbert space.

the one that retains the most variance present in the original data. Kernel PCA is a nonlinear generalization of PCA that first lifts the original data to a space of functions and then produces low dimensional projections in this function space. This representation can help extract nonlinear structures in the data [30], can be used for data denoising [25] and extracting high dimensional features for regression and classification tasks [29, 22, 6]. More extensive references to kernel methods can be found in the books by [33] and [34].

In this work we use kernel PCA as a denoising tool for multivariate extremes. More specifically, we perform kernel PCA on a subset of observations that is viewed as extremes and we reconstruct the preimages of the kernel PCA projections of this extremal subsample. While kernel PCA projections live in a function space, their preimages live in the original space of the data and constitute our main objects of interest.

In our analysis, we first provide some general insights showing that kernel PCA preimages naturally cluster in finite subsets of points when there are also some clusters in the kernel space. We believe these insights are interesting in their own right and complement existing work on kernel PCA preimages; see [19] for an overview of this literature. We remark that our analysis also complements existing theoretical results regarding the convergence of the spectrum of the empirical covariance operator used in the construction of kernel PCA to a population covariance operator [31, 39, 5].

We then consider the case of an extremal sample and utilize tools from multivariate extreme value theory for analyzing the clustering properties of kernel PCA preimages. In particular, we use multivariate regular variation as a modeling tool since it is closely connected to asymptotic characterizations of multivariate extreme value distributions [26, 27]. We provide a detailed analysis of the clustering properties of kernel PCA preimages for multivariate extremes generated by the linear factor model recently introduced by [1]. We leverage the asymptotic distributions and rates of convergence derived in the latter work in out perturbation analysis of kernel PCA. Since the spectral measure of this model is discrete, the kernel PCA preimages converge to a well separated set of points in this case. We establish rates of convergence of the kernel matrix defining these preimages and show that they depend on the tail index of the extremes as well as on the smoothness, around the origin, of the kernel function used for kernel PCA. In the process of establishing these convergence results, we obtain an asymptotic characterization of the point process giving rise to the different clusters of extremes in the linear factor model. We believe this is an interesting side product of our analysis that is likely to be useful in other contexts.

2. Background on Kernel PCA.

Kernel principal component analysis builds on the framework of Reproducing Kernel Hilbert Spaces (RKHS). We will review some basic notation and concepts that will be needed in order to describe how kernel PCA gets a low dimensional representation of the data after being lifted to a RKHS.

DEFINITION 1. *Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on a space $\mathcal{X}$ with inner product and norm denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ respectively. A function $\kappa : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is called reproducing kernel if (i) for all $x\in\mathcal{X}$, $\kappa(\cdot,x)\in\mathcal{H}$ and (ii) for all $f\in\mathcal{H}$, we have $f(x)=\langle f,\kappa(\cdot,x)\rangle$ for all $x\in\mathcal{X}$.*

If $\mathcal{H}$ admits a reproducing kernel, then it is called a reproducing kernel Hilbert space.

By the Moore-Aronszajn Theorem, each symmetric, non-negative definite kernel function $\kappa(\cdot,\cdot)$ on $\mathcal{X}\times\mathcal{X}$ can be identified uniquely with a RKHS of real-valued functions on $\mathcal{X}$ for which it is the reproducing kernel. For simplicity, in what follows let us denote this RKHS by $\mathcal{H}$.

The map $\phi : x \mapsto \kappa(\cdot, x)$ from $\mathcal{X}$ to $\mathcal{H}$ is commonly called the feature map. It follows from the reproducing kernel property that $f(x) = \langle f, \phi(x)\rangle$ for all $f \in \mathcal{H}$ and

$$\langle \phi(x), \phi(y)\rangle = \langle \kappa(\cdot, x), \kappa(\cdot, y)\rangle = \kappa(x, y), \quad \forall x, y, \in \mathcal{X}.$$

We are now well equipped to describe the methodology of kernel PCA. Assume that we have some observed data $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X} \subset \mathbb{R}^d$ and define the empirical covariance operator $\mathscr{C}_n : \mathcal{H} \mapsto \mathcal{H}$ as

$$\mathscr{C}_n f = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\langle \phi(\mathbf{x}_i), f\rangle = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i)\phi(\mathbf{x}_i).$$

The covariance operator[1] is positive definite and admits the spectral representation

$$\mathscr{C}_n f = \sum_{i=1}^{n} \lambda_i \langle \varphi_i, f\rangle \varphi_i,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues of $\mathscr{C}_n$ and $\varphi_1, \dots, \varphi_n \in \mathcal{H}$ are their corresponding normalized eigenfunctions. We are interested in finding the principal components i.e., a lower dimensional representation of arbitrary functions in $\mathcal{H}$ using the first $m < n$ eigenfunctions of $\mathscr{C}_n$. More specifically, given a function $f \in \mathcal{H}$, we want to find its projection into the m-dimensional subspace spanned by $\{\varphi_i\}_{i=1}^{m}$ i.e.

$$\mathcal{P}_m f = \sum_{j=1}^{m} \lambda_j \langle \varphi_j, f\rangle \varphi_j.$$

The above projection can be computed efficiently in practice as one can avoid working directly in $\mathcal{H}$ by noting that the eigenvalues of $\mathscr{C}_n$ coincide with the eigenvalues of the kernel matrix $C_n = \frac{1}{n}\{\kappa(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^{n}$. If $\mathbf{v}_j = (v_{j1}, \dots, v_{jn})^\top$ denotes the jth eigenvector of C_n, then the jth (unnormalized) eigenfunction of $\mathscr{C}_n$ can be computed by

$$\varphi_j = \sum_{i=1}^{n} v_{ji}\phi(\mathbf{x}_i). \tag{1}$$

From this last formula and the reproducing kernel property we obtain

$$\mathcal{P}_m f = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} f(\mathbf{x}_k) v_{ij} v_{ik} \phi(\mathbf{x}_j).$$

In what follows we focus on the properties of the preimages of the lower dimensional projection of the feature map i.e., $\mathcal{P}_m\phi$. Intuitively, we would like to understand the impact of the steps: a) lifting the data from the original space $\mathcal{X}$ to a function space b) choosing a lower dimensional representation of the lifted data in this richer space, and c) identifying the map of those projections back to the original data space $\mathcal{X}$.

We finish this section by mentioning a connection between RKHSs and Gaussian random fields that often provides a convenient and appealing framework to view RKHS and kernel PCA. Specifically, if κ is a stationary symmetric non-negative definite kernel on $\mathbb{R}^d$ (that is, $\kappa(\mathbf{x}, \mathbf{y})$ depends only on $\mathbf{x} - \mathbf{y}$), and $\boldsymbol{G} = \big(G(\mathbf{x}), \, \mathbf{x} \in \mathbb{R}^d\big)$ is a centered stationary Gaussian random field, with covariance function κ, then the RKHS corresponding to κ can be identified with the closure in L^2 of the space of finite linear combinations of the values of the field at different points (here each $\kappa(\cdot, \mathbf{x})$ is identified with $G(\mathbf{x})$). This RKHS can also be identified

[1] We note that the covariance operator is sometimes defined using the centered feature map $\bar{\phi}(\mathbf{x}_i) = \phi(\mathbf{x}_i) - \frac{1}{n}\sum_{j=1}^{n}\phi(\mathbf{x}_j)$ instead of the non-centered feature map $\phi(\mathbf{x}_i)$. This leads to minor changes to the expressions that we give for the eigenfunctions and eigenvalues.

with the subspace of $L^2(\mu)$ consisting of functions with even real parts and odd imaginary parts. Here μ is the spectral measure of the random field $\boldsymbol{G}$, i.e., a finite symmetric measure on $\mathbb{R}^d$ such that

$$\kappa(\mathbf{x}) = \int_{\mathbb{R}^d} e^{i\langle \mathbf{x},\mathbf{z}\rangle}\,\mu(d\mathbf{z}),\ \mathbf{x} \in \mathbb{R}^d$$

(since κ is stationary, we are using a one-argument notation). Here $\kappa(\cdot,\mathbf{x})$ is identified with $e^{i\langle \cdot,\mathbf{x}\rangle}$. See e.g., [36].

3. Some insights into kernel PCA preimages. Let $\mathcal{S}$ be a subset of $\mathbb{R}^d$, possibly the unit sphere in $\mathbb{R}^d$. Let $S_0 \subset \mathcal{S}$ be a small subset of $\mathcal{S}$, potentially of a smaller dimension. Let $T = \{\mathbf{t}_1,\dots,\mathbf{t}_{n_1},\mathbf{t}_{n_1+1},\dots,\mathbf{t}_{n_1+n_2}\}$ be a collection of points in $\mathcal{S}$, such that $\mathbf{t}_1,\dots,\mathbf{t}_{n_1}$ lie in or near S_0, while $\mathbf{t}_{n_1+1},\dots,\mathbf{t}_{n_1+n_2}$ lie at some distance from S_0. Assume, further, that the points $\mathbf{t}_{n_1+1},\dots,\mathbf{t}_{n_1+n_2}$ are dispersed, and that n_2 is not too large in comparison with n_1. In the following we will focus on RKHS defined by stationary kernels of the form $\kappa(\mathbf{x},\mathbf{y}) = R(\mathbf{x}-\mathbf{y})$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^d$ and some continuous non-negative definite $R : \mathbb{R}^d \mapsto \mathbb{R}$ (which may be thought of as the covariance function of a stationary Gaussian random field).

3.1. *Discrete signal case.* Let us first consider a special case. Suppose that $n_2 = 0$ and that $S_0 = \{\mathbf{s}_1,\dots,\mathbf{s}_K\}$ is a finite collection of points in $\mathcal{S}$. Furthermore, suppose that the points $\mathbf{s}_1,\dots,\mathbf{s}_K$ are well separated, in the sense that $R(0) = 1$ while $|R(\mathbf{s}_i - \mathbf{s}_j)|$ is small if $i \neq j$. Let us suppose, further, that m_1 of the points $\mathbf{t}_1,\dots,\mathbf{t}_{n_1}$ equal $\mathbf{s}_1$, m_2 of the points $\mathbf{t}_1,\dots,\mathbf{t}_{n_1}$ equal $\mathbf{s}_2$, etc. In particular, $n_1 = m_1 + m_2 + \dots + m_K$. We will assume for simplicity that the first $m_1$points $\mathbf{t}_1,\dots,\mathbf{t}_{m_1}$ equal to $\mathbf{s}_1$, the next m_2 points are equal to $\mathbf{s}_2$, etc. In this case the matrix $C_n = \frac{1}{n}\{R(\mathbf{s}_i - \mathbf{s}_j)\}_{i,j=1}^n$ becomes a block matrix with K^2 blocks. The block (k_1,k_2), $k_1,k_2 = 1,\dots,K$ has m_{k_1} rows and m_{k_2} columns and $m_{k_1}m_{k_2}$ identical entries equal to

$$\frac{1}{(m_1 + \dots + m_K)} R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}). \tag{2}$$

In particular, an eigenvector $\mathbf{v}_j$ of C_n corresponding to any eigenvalue λ_j will have the form

$$\mathbf{v}_j = \big(b_1,\dots,b_1,b_2,\dots,b_2,\dots,b_K,\dots,b_K\big)^{\top}, \tag{3}$$

with each b_i repeated m_i times.

Returning to the general case, let $\lambda_1 \geq \dots \geq \lambda_m \geq 0$ be the m largest eigenvalues of $\mathcal{C}_n$ and let $\varphi_1,\dots,\varphi_m$ be the corresponding eigenfunctions. Suppose that we now get a new point $\mathbf{w} \in \mathcal{S}$. We define its kernel PCA preimage as

$$T(\mathbf{w}) = \text{argmin}_{\mathbf{v}\in\mathcal{S}} \|\phi(\mathbf{v}) - \mathcal{P}_m\phi(\mathbf{w})\|. \tag{4}$$

Note that $\|\phi(\mathbf{v})\|^2 = R(0)$ is independent of $\mathbf{v}$. Therefore, (4) reduces to

$$\begin{aligned} T(\mathbf{w}) &= \text{argmax}_{\mathbf{v}\in\mathcal{S}} \langle \phi(\mathbf{v}), \mathcal{P}_m\phi(\mathbf{w})\rangle \\ &= \text{argmax}_{\mathbf{v}\in\mathcal{S}} \sum_{k=1}^{m} \sum_{\mathbf{t}_j\in T} v_{kj} R(\mathbf{w} - \mathbf{t}_j) \sum_{\mathbf{t}_j\in T} v_{kj} R(\mathbf{v} - \mathbf{t}_j). \end{aligned} \tag{5}$$

EXAMPLE 1. *To get a feeling of what is happening let us consider the case* $m = 1$. *In this case the problem* (5) *becomes*

$$T(\mathbf{w}) = \operatorname{argmax}_{\mathbf{v}\in\mathcal{S}} \sum_{\mathbf{t}_j\in T} v_{1j} R(\mathbf{w}-\mathbf{t}_j) \sum_{\mathbf{t}_j\in T} v_{1j} R(\mathbf{v}-\mathbf{t}_j).$$

This means that

$$T(\mathbf{w}) = \begin{cases} \operatorname{argmax}_{\mathbf{v}\in\mathcal{S}} \sum_{\mathbf{t}_j\in T} v_{1j}R(\mathbf{v}-\mathbf{t}_j), \textit{ if } \sum_{\mathbf{t}_j\in T} v_{1j}R(\mathbf{w}-\mathbf{t}_j) > 0, \\ \operatorname{argmin}_{\mathbf{v}\in\mathcal{S}} \sum_{\mathbf{t}_j\in T} v_{1j}R(\mathbf{v}-\mathbf{t}_j), \textit{ if } \sum_{\mathbf{t}_j\in T} v_{1j}R(\mathbf{w}-\mathbf{t}_j) < 0. \end{cases} \tag{6}$$

That is, most of the points $\mathbf{w}$ *get mapped to one of the two points in* S_0 *that achieve the minimum and maximum* (6). *If we return to the special case considered in* (2) *and* (3), *then*

$$\sum_{\mathbf{t}_j\in T} v_{1j}R(\mathbf{v}-\mathbf{t}_j) = \sum_{k=1}^{K} m_k b_k R(\mathbf{v}-\mathbf{s}_k).$$

Since we are assuming that the points $\mathbf{s}_1,\dots,\mathbf{s}_K$ *are well separated, it is likely that the maximum of that expression will be achieved near the point* $\mathbf{s}_k$ *with the largest value of* $m_k b_k$, *and the minumum will be achieved near the point* $\mathbf{s}_k$ *with the smallest value of* $m_k b_k$. *That is, most of the points* $\mathbf{w}$ *are likely to get mapped close to one of these two points in the set* S_0.

3.2. *Discrete signal with noise.* Now we consider the generic case where $S_0 = \{\mathbf{s}_1,\dots,\mathbf{s}_K\}$ is still a finite collection of well separated points in $\mathcal{S}$. However, now $n_2 > 0$ and the points $\mathbf{t}_{n_1+1},\dots,\mathbf{t}_{n_1+n_2}$ lie at some distance from S_0, and do not concentrate too much themselves. Now the points $\mathbf{t}_1,\dots,\mathbf{t}_{n_1}$ are not necessarily exactly equal to one of the points in S_0, but are only lie nearby. Specifically, we assume that m_1 of the points $\mathbf{t}_1,\dots,\mathbf{t}_{n_1}$ are near $\mathbf{s}_1$, m_2 of the points $\mathbf{t}_1,\dots,\mathbf{t}_{n_1}$ are near $\mathbf{s}_2$, etc. We still have $n_1 = m_1 + m_2 + \cdots + m_K$. This time the covariance matrix C_n will have 4 distinct parts whose structure we now describe.

Recall that each $\mathbf{t}_j$, $j = 1,\dots,n_1$ is near one of the points in S_0. We preserve the numbering we used in (2) and (3). That is, we write

$$\mathbf{t}_j = \mathbf{s}_k + \mathbf{r}_j \ \text{ if } \ m_1 + \cdots + m_{k-1} < j \le m_1 + \cdots + m_{k-1} + m_k, \tag{7}$$

$j = 1,\dots,n_1,\ k = 1,\dots,K$, and assume that $\|\mathbf{r}_j\|$ is small ($\mathbf{r}_j$ does not need to lie in $\mathcal{S}$). The matrix C_n will have an $n_1 \times n_1$ block matrix in the top left corner with K^2 blocks, whose entries are perturbations of the entries entries described in (2). Specifically, the block (k_1,k_2), $k_1,k_2 = 1,\dots,K$ in that matrix has m_{k_1} rows and m_{k_2} columns, and the entry in the position (i,j) within that block can be written as

$$c_{ij} = \frac{1}{m_1 + \cdots + m_K} R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) + \delta_{ij}, \tag{8}$$

with

$$\delta_{ij} = \frac{1}{n_1+n_2}\Big(R\big(\mathbf{s}_{k_1} - \mathbf{s}_{k_2} + \mathbf{r}_i - \mathbf{r}_j\big) - R\big(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}\big)\Big). \tag{9}$$

We will view this matrix C_n as resulting from a perturbation of the matrix $C_n^{(0)}$ of the same size as C_n. The matrix $C_n^{(0)}$ has an $n_1 \times n_1$ block matrix in the top left corner with K^2 blocks, whose entries are described in (2). The rest of the entries of the matrix $C_n^{(0)}$ are equal to zero. Let

$$\Delta = C_n - C_n^{(0)}$$

be the perturbation. Then Δ has an $n_1 \times n_1$ block matrix in the top left corner with K^2 blocks, whose entries are δ_{ij} in (9). The rest of the entries of the matrix Δ have the general form

$$\frac{1}{n_1 + n_2} R(\mathbf{t}_i - \mathbf{t}_j).$$

We start by noticing that the non-zero eigenvalues of the matrix $C_n^{(0)}$ coincide with the non-zero eigenvalues of the $n_1 \times n_1$-matrix in its top left corner. Furthermore, the corresponding eigenvectors of $C_n^{(0)}$ result from taking the eigenvectors of the latter matrix and appending to them n_2 zero entries. We note that the $n_1 \times n_1$-matrix in the top left corner of $C_n^{(0)}$ represents the special situation (2) and we have some understanding of why our procedure results in points being mapped close to the set S_0.

The true matrix is, of course, C_n and not $C_n^{(0)}$. Our plan is to use the Davis-Kahan theorem to check that the eigenvectors of C_n corresponding to its top eigenvalues are not far from the eigenvectors of $C_n^{(0)}$ corresponding to its top eigenvalues. If this is the case, then the eigenfunctions (1) corresponding to the top eigenvalues of the matrix C_n will be close to the eigenfunctions (1) corresponding to the top eigenvalues of the matrix $C_n^{(0)}$, and so the algorithm will still map most of the points to lie close to the set S_0.

We will use a version of the Davis-Kahan theorem given in [37], which says that, if $\lambda_1, \cdots, \lambda_m$ are the top eigenvalues of C_n and $\lambda_1^{(0)}, \cdots, \lambda_m^{(0)}$ are the top eigenvalues of $C_n^{(0)}$, then the corresponding orthonormal eigenvectors $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(m)}$ and $\mathbf{v}^{(0,1)}, \dots, \mathbf{v}^{(0,m)}$ are close in the following sense. Let V and $V^{(0)}$ be $(n_1 + n_2) \times m$ matrices with columns $\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(m)}$ and $\mathbf{v}^{(0,1)}, \dots, \mathbf{v}^{(0,m)}$, correspondingly. Then there is an orthogonal $m \times m$ matrix O such that

$$\|VO - V^{(0)}\|_{\mathrm{F}} \le \frac{2\min\big(m^{1/2}\|\Delta\|_{\mathrm{op}}, \|\Delta\|_{\mathrm{F}}\big)}{\lambda_m^{(0)} - \lambda_{m+1}^{(0)}}. \tag{10}$$

Here $\|A\|_{\mathrm{F}}$ and $\|A\|_{\mathrm{op}}$ are, respectively, the Frobenius norm and the operator norm of a matrix A. This is Theorem 2 in [37].

Since the orthogonal matrix O in (10) plays no role in the projection onto the subspace spanned by the eigenfunctions corresponding to the top eigenvectors, we need to check that the bound in the right hand side of (10) is small. Assuming that the covariance matrix of the Gaussian random field is nonsingular, in the generic case the matrix $C_n^{(0)}$ will have K nonzero eigenvalues. Furthermore, the size of these K eigenvalues should be comparable to the size of the entries in the matrix $C_n^{(0)}$, which by (2) are of order $1/n$. It is likely that also the distances between these eigenvalues are of the same order. This, of course, means that we should choose $m \le K$ and, ideally, $m = K$. In this case we expect the denominator in the right hand side of (10) to be of order $1/n$.

Now consider the numerator in the right hand side of (10). One can expect that the norms appearing there would be comparable to the size of the entries in the matrix Δ. Notice that the entries in this matrix that are not in the $n_1 \times n_1$ block matrix in the top left corner are still of the order $1/n$, but because the points $\{\mathbf{s}_k\}$ are well separated, these entries will be small in comparison with the denominator in the right hand side of (10). Finally, the entries in the $n_1 \times n_1$ block matrix in the top left corner of the matrix Δ will be small if the perturbations $\{r_i\}$ are small, and this should be established on the case-by-case basis, for different data-producing mechanisms. In the next section we apply this idea to the extremes of a heavy tailed linear factor model.

4. Applications to the linear factor model.

Consider the linear factor model

$$\mathbf{X} = A\mathbf{Z}, \tag{11}$$

where A is a $d \times p$ matrix of non-negative elements and $\mathbf{Z}$ is a p-dimensional random vector of factors consisting of independent and identically distributed non-negative random variables[2], that have asymptotically Pareto tails, i.e.,

$$\mathbb{P}(Z_1 > z) \sim c_\alpha z^{-\alpha}, \quad \text{as } z \to \infty, \tag{12}$$

for some $\alpha > 0$ and $c_\alpha > 0$. As pointed out in [1], it follows immediately from (11) and (12) (see, for example, [3], Proposition A.1) that $\mathbf{X}$ is a multivariate regularly varying random vector satisfying

$$\lim_{x\to\infty} \mathbb{P}\left(\frac{\mathbf{X}}{\|\mathbf{X}\|} \in \cdot \mid \|\mathbf{X}\| > x \right) \Rightarrow \Gamma(\cdot)\,,$$

where $\Rightarrow$ denotes weak convergence on the unit sphere $\mathbb{S}^{d-1}$, Γ is the discrete probability measure on $\mathbb{S}^{d-1}$ given by

$$\Gamma(\cdot) = w^{-1} \sum_{k=1}^{p} \|\mathbf{a}^{(k)}\|^\alpha \delta_{\frac{\mathbf{a}^{(k)}}{\|\mathbf{a}^{(k)}\|}}(\cdot)\,,$$

where $\delta_x(\cdot)$ is the Dirac measure that puts unit mass at x, $\mathbf{a}^{(k)}$ is the kth column of the matrix A, $k = 1, \dots, p$, and

$$w = \sum_{k=1}^{p} \|\mathbf{a}^{(k)}\|^\alpha\,. \tag{13}$$

Based on a random sample of iid copies of $\mathbf{X}_1, \dots, \mathbf{X}_n$ of $\mathbf{X}$ as above, we expect for large n, the *angular parts* $\mathbf{X}_i/\|\mathbf{X}_i\|$ of the sample for which $\|\mathbf{X}_i\|$ is large, to cluster around the points

$$\mathbf{s}_k = \frac{\mathbf{a}^{(k)}}{\|\mathbf{a}^{(k)}\|}, \; k = 1, \dots, p,$$

as in Section 3.2.

In order to understand how well the kernel PCA algorithm works for extreme values in this model, we will analyze the corresponding perturbation matrix Δ. We start with the $n_1 \times n_1$ block matrix in the top left corner of Δ, whose entries are given by (9). We call this matrix Δ_B.

The first result that we will need in the study of Δ_B is a characterization of the convergence of the covariance function evaluated at the difference of directions of extreme observations. We will assume that for some $\theta \in [1, 2]$ and $d_\theta > 0$,

$$R(\mathbf{0}) - R(\mathbf{x}) \sim d_\theta \|\mathbf{x}\|^\theta \quad \text{as } \mathbf{x} \to \mathbf{0}. \tag{14}$$

Common choices of R are the exponential covariance function $R(\mathbf{x}) = \exp\{-\gamma\|\mathbf{x}\|\}$ for which $\theta = 1$, and the Gaussian covariance function $R(\mathbf{x}) = \exp\{-\gamma\|\mathbf{x}\|^2\}$, for which $\theta = 2$.

[2] The results of this section can be extend with simple but tedious modifications to symmetric regularly varying iid random variables and a real-valued A.

THEOREM 1. *Let (u_n) be a sequence of levels such that $u_n \to \infty$ and suppose the covariance function R satisfies* (14) *and is continuously differentiable outside of the origin.*

(i) *Let $i \neq j$ belong to a diagonal block (k,k) of the matrix Δ_B, $k = 1,\dots,p$. Then for any $i \neq j$, computing the law in the left hand side as the conditional law given the event $\{\|\mathbf{X}_i\| > u_n, \|\mathbf{X}_j\| > u_n,\ Z_{ik} > u_n/w^{1/\alpha},\ Z_{jk} > u_n/w^{1/\alpha}\}$, ($w$ defined in* (13)*),*

$$u_n^{\theta}\Big[R(\mathbf{0}) - R\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{X}_j/\|\mathbf{X}_j\|\big)\Big] \Rightarrow d_\theta \|\mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)}\|^{\theta}, \tag{15}$$

where $\mathbf{S}_1^{(k)}, \mathbf{S}_2^{(k)}$ are independent random vectors with the common law defined as the law of

$$\frac{1}{w_k^2 W_\alpha}\big(S^*_{1,-k}, \dots, S^*_{d,-k}\big)^T, \tag{16}$$

with $w_k = \|\mathbf{a}^{(k)}\|$. Here, with $\mathbf{X}$ as in (11)*, we set*

$$S^*_{l,-k} = \sum_{r=1}^{d}\big(a_{rk}^2 X_{l,-k} - a_{lk}a_{rk}X_{r,-k}\big)$$

with

$$X_{l,-k} = X_l - a_{lk}Z_k,\quad l = 1,\dots,d.$$

Furthermore, W_α is standard Pareto(α) random variable ($\mathbb{P}(W_\alpha > x) = x^{-\alpha}$, $x \geq 1$), independent of $Z_1,\dots,Z_p$.

(ii) *Let (i,j) belong to a block (k_1,k_2) of the matrix Δ_B, $k_1 \neq k_2$, $k_1,k_2 = 1,\dots,p$. Then, for any $i \neq j$, computing the law in the left hand side as the conditional law given the event $\{\|\mathbf{X}_i\| > u_n, \|\mathbf{X}_j\| > u_n,\ Z_{ik_1} > u_n/w^{1/\alpha},\ Z_{jk_2} > u_n/w^{1/\alpha}\}$,*

$$u_n\Big[R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) - R\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{X}_j/\|\mathbf{X}_j\|\big)\Big] \Rightarrow \langle \nabla R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}), \mathbf{S}_1^{(k_2)} - \mathbf{S}_2^{(k_1)}\rangle, \tag{17}$$

with $\mathbf{S}_1^{(k_1)}, \mathbf{S}_2^{(k_2)}$ independent and distributed as above.

PROOF. (i) Write

$$\begin{aligned}&R(\mathbf{0}) - R\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{X}_j/\|\mathbf{X}_j\|\big)\\ =&R(\mathbf{0}) - R\Big[u_n^{-1}\Big(u_n\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{s}_k\big) - u_n\big(\mathbf{X}_j/\|\mathbf{X}_j\| - \mathbf{s}_k\big)\Big)\Big].\end{aligned}$$

By Theorem 4.1 in [1] (which holds under the sole assumption $u_n \to \infty$), we have

$$u_n\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{s}_k\big) \Rightarrow \mathbf{S}_1^{(k)} \tag{18}$$

weakly in $\mathbb{R}^d$. The same is true when i is replaced by j and by independence and (14), this implies (15) using the delta method.

(ii) Write

$$\begin{aligned}&R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) - R\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{X}_j/\|\mathbf{X}_j\|\big)\\ =&R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) - R\Big[\mathbf{s}_{k_1} - \mathbf{s}_{k_2} + u_n^{-1}\Big(u_n\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{s}_{k_1}\big) - u_n\big(\mathbf{X}_j/\|\mathbf{X}_j\| - \mathbf{s}_{k_2}\big)\Big)\Big].\end{aligned}$$

Now (17) follows from the differentiability of R outside of the origin and (18), once again using the delta method.

□

Recall that we would like to understand the magnitude of the matrix norms appearing in the numerator of (10). As explained above, we expect the two norms to be of the same order, so we will consider the Frobenius norm of Δ. We start with the matrix Δ_B. Theorems 2, 4 and 5 below constitute the main results regarding the asymptotic behaviour of the Frobenius norm of Δ_B under the linear factor model. These theorems highlight that there are three different regimes resulting from the tail index of the underlying factors and the smoothness of the chosen kernel function. In particular, the regime $\alpha < 2\theta$ requires us to establish a new point process convergence result stated in Theorem 3. It serves as an important technical tool in the proof of Theorems 4 and 5, and we believe that it can be of broader interest.

It will be convenient to introduce some notation used in [1]. For $n = 1, 2, \dots$, we define the set of indexes corresponding to extreme observations

$$\mathcal{I}_n = \{i = 1, \dots, n : \|\mathbf{X}_i\| > u_n\},$$

and denote its cardinality by $N_n = \mathrm{card}(\mathcal{I}_n)$. Now assuming the thresholds (u_n) satisfy the standard assumptions

$$u_n \to \infty,\ \ n^{-1/\alpha} u_n \to 0,\ \ n \to \infty\,, \tag{19}$$

which imply $n\mathbb{P}(\|\mathbf{X}\| > u_n) \sim c_\alpha (n^{-1/\alpha} u_n)^{-\alpha} \to \infty$, it follows from (12) and (19) (see [1]) that the mean and variance of $N_n/(nu_n^{-\alpha})$ converge to cw and 0, respectively and hence that

$$N_n/(nu_n^{-\alpha}) \overset{P}{\to} cw, \text{ as } n \to \infty. \tag{20}$$

Let $\mathcal{I}_n^{(k)}$ denote the collection of indexes of extremes caused by the kth factor Z_k, $k = 1, \dots, p$. Formally,

$$\mathcal{I}_n^{(k)} = \{i = 1, \dots, n : \|\mathbf{X}_i\| > u_n,\ Z_{ik} > u_n/w^{1/\alpha}\},\ k = 1, \dots, p$$

(this is (4.22) in [1]). We denote $N_n^{(k)} = \mathrm{card}(\mathcal{I}_n^{(k)})$, $k = 1, \dots, p$. It turns out that with probability converging to 1, the sets $\mathcal{I}_n^{(1)}, \dots, \mathcal{I}_n^{(p)}$ are disjoint and $N_n = N_n^{(1)} + \dots + N_n^{(p)}$; see Lemma 4.4 in [1]. Furthermore, by (4.24) in [1],

$$N_n^{(k)}/(nu_n^{-\alpha}) \overset{P}{\to} c_\alpha w_k^\alpha,\ k = 1, \dots, p, \text{ as } n \to \infty, \tag{21}$$

where we recall

$$w = \sum_{k=1}^{p} \|\mathbf{a}^{(k)}\|^\alpha = \sum_{k=1}^{p} w_k^\alpha.$$

Using this notation, we see that

$$\|\Delta_B\|_{\mathrm{F}} = \left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \delta_{ij}^2 \right)^{1/2} = \left(\sum_{k_1=1}^{p} \sum_{k_2=1}^{p} \sum_{i \in \mathcal{I}_n^{(k_1)}} \sum_{j \in \mathcal{I}_n^{(k_2)}} \delta_{ij}^2 \right)^{1/2}$$

$$=: \left(\sum_{k_1=1}^{p} \sum_{k_2=1}^{p} F_{k_1,k_2}(n) \right)^{1/2}. \tag{22}$$

Further note that $F_{k_1,k_2}(n)$ can be written as a U-statistic of the form

$$F_{k_1,k_2}(n) = \frac{1}{N_n^2} \sum_{i \in \mathcal{I}_n^{(k_1)}} \sum_{j \in \mathcal{I}_n^{(k_2)}} \left[R\left(\frac{\mathbf{X}_i}{\|\mathbf{X}_i\|} - \frac{\mathbf{X}_j}{\|\mathbf{X}_j\|} \right) - R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) \right]^2$$

$$
\begin{aligned}
&= \frac{1}{N_n^2} \sum_{i=1}^{N_n^{(k_1)}} \sum_{j=1}^{N_n^{(k_2)}} \left[R\left(\mathbf{Y}_i^{(k_1)} - \mathbf{Y}_j^{(k_2)} \right) - R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) \right]^2 \\
&= \frac{N_n^{(k_1)} N_n^{(k_2)}}{N_n^2} \frac{1}{N_n^{(k_1)} N_n^{(k_2)}} \sum_{i=1}^{N_n^{(k_1)}} \sum_{j=1}^{N_n^{(k_2)}} \left[R\left(\mathbf{Y}_i^{(k_1)} - \mathbf{Y}_j^{(k_2)} \right) - R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}) \right]^2 \\
&=: \frac{N_n^{(k_1)} N_n^{(k_2)}}{N_n^2} G_{k_1,k_2}(n). \qquad (23)
\end{aligned}
$$

In (23), for each $k = 1,\dots,p$, we enumerate $\mathbf{X}_i/\|\mathbf{X}_i\|$, $i \in \mathcal{I}_n^{(k)}$ as $\mathbf{Y}_i^{(k)}$, $i = 1,\dots,N_n^{(k)}$, a sample on $\mathbb{S}^{d-1}$ of random size $N_n^{(k)}$. We also have that

$$
\frac{N_n^{(k_1)} N_n^{(k_2)}}{N_n^2} \xrightarrow{P} \frac{w_{k_1}^\alpha w_{k_2}^\alpha}{w^2} \text{ as } n \to \infty. \tag{24}
$$

Theorem 1 makes it reasonable to expect that under some assumptions it should be true that

$$
u_n^{2\theta} G_{k,k}(n) \to d_\theta^2 \mathbb{E}\left\|\mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)}\right\|^{2\theta}, \tag{25}
$$

and that for $k_1 \neq k_2$,

$$
u_n^2 G_{k_1,k_2}(n) \to \mathbb{E}\left[\langle \nabla R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}), \mathbf{S}_1^{(k_2)} - \mathbf{S}_2^{(k_1)} \rangle\right]^2, \tag{26}
$$

at least in probability. At the very least (25) requires $\mathbb{E}\left\|\mathbf{S}_1^{(k)}\right\|^{2\theta} < \infty$, while (26) requires $\mathbb{E}\left\|\mathbf{S}_1^{(k)}\right\|^2 < \infty$. Since $\theta \geq 1$, we will assume that

$$
\alpha > 2\theta. \tag{27}
$$

The following statement formalizes this intuition and characterizes the behavior of $\|\Delta_B\|_{\mathrm{F}}$ when the tails are not too heavy i.e., when (27) holds. The proofs of this and the subsequent theorems in this section are given in the Appendix .

THEOREM 2. *Suppose that* (27) *holds. Then,*

$$
\begin{aligned}
u_n^2 \|\Delta_B\|_{\mathrm{F}} \xrightarrow{P} \frac{1}{w} \Bigg(& \sum_{k=1}^{p} \left(d_* w_k^\alpha\right)^2 \mathbb{E}\left\|\mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)}\right\|^2 \\
& + \sum_{\substack{k_1=1 \\ k_1 \neq k_2}}^{p} \sum_{k_2=1}^{p} w_{k_1}^\alpha w_{k_2}^\alpha \mathbb{E}\left[\langle \nabla R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}), \mathbf{S}_1^{(k_2)} - \mathbf{S}_2^{(k_1)} \rangle\right]^2 \Bigg)^{1/2}, \textit{as } n \to \infty,
\end{aligned} \tag{28}
$$

where $d_* = d_\theta$ if $\theta = 1$ and $d_* = 0$ if $\theta > 1$.

Note that in case $\theta > 1$ (impliying that the covariance function $R(\cdot)$ is differentiable at the origin) only the off-diagonal terms contribute to the asymptotic behavior of the Frobenius norm. In the scenario

$$
\alpha < 2\theta\,, \tag{29}
$$

the analysis for the diagonal and off-diagonal terms is different. We will show that under the following additional assumption on the sequence of levels,

$$
n^{-1/\alpha} u_n^2 \to \infty, \;\; n \to \infty. \tag{30}
$$

upon proper rescaling, these terms converge in distribution to $\alpha/(2\theta)$-stable positive random variables.

We start with a result on convergence of a certain sequence of point processes that may be of independent interest. For every $k = 1,\dots,p$ and $n \geq 1$ we define a point process

$$M_n^{(k)} = \sum_{i=1}^{N_n^{(k)}} \delta_{u_n^2 n^{-1/\alpha}(\mathbf{Y}_i^{(k)} - \mathbf{s}_k)}.$$

THEOREM 3. *Suppose that* (29) *holds, and that the sequence of levels satisfies* (19) *and* (30). *Then,*

$$M_n^{(k)} \Rightarrow M_\alpha^{(k)}, \quad n \to \infty, \tag{31}$$

weakly in the vague topology on $\mathbb{R}^d \setminus \{0\}$, where $M_\alpha^{(k)}$ is a Poisson point process on $\mathbb{R}^d$ with mean measure

$$m_\alpha^{(k)}(\cdot) = (c_\alpha^2 w_k^2/2) \sum_{j \neq k} m_{\alpha,j}(\cdot),$$

and for $j = 1,\dots,p$, $j \neq k$,

$$m_{\alpha,j} = \int_0^\infty \alpha y^{-(1+\alpha)} \delta_{y\mathbf{b}^{(j,k)}}(\cdot)\,dy.$$

Here $\mathbf{b}^{(j,k)} = \big(b_1^{(j,k)},\dots,b_d^{(j,k)}\big)^\top$, and for $r = 1,\dots,d$,

$$b_r^{(j,k)} = w_k \left(a_{rj} - \frac{a_{rk}}{w_k^2} \sum_{m=1}^{d} a_{mj} a_{mk} \right).$$

Moreover, the convergence in (31) *is joint in $k = 1,\dots,p$, i.e.,*

$$\big(M_n^{(1)},\dots,M_n^{(p)}\big) \Rightarrow \big(M_\alpha^{(1)},\dots,M_\alpha^{(p)}\big), \quad n \to \infty,$$

where $M_\alpha^{(1)},\dots,M_\alpha^{(p)}$ are independent Poisson point processes.

The asymptotic behaviour of the Frobenius norm of the perturbation matrix Δ_B can be now expressed via integrals with respect to the limit point processes $M_\alpha^{(k)}$. We consider two separate cases.

THEOREM 4. *Suppose that $\alpha < 2$ and that the sequence of levels satisfies* (19) *and* (30). *Then,*

$$\big(u_n^{2-\alpha/2} n^{-1/\alpha+1/2}\big)\|\Delta_B\|_{\mathrm{F}} \Rightarrow \Bigg(\frac{2d_*^2}{c_\alpha} \sum_{k=1}^{p} w_k^\alpha \int_{\mathbb{R}^d} \|x\|^{2\theta} M_\alpha^{(k)}(d\mathbf{x}) + \frac{1}{w^2 c_\alpha} \sum_{\substack{k_1=1 \\ k_1 \neq k_2}}^{p} \sum_{k_2=1}^{p} \int_{\mathbb{R}^d} \Big[\big(\nabla R(s_{k_1} - s_{k_2}), x\big)\Big]^2 \big(w_{k_2}^\alpha M_\alpha^{(k_1)} + w_{k_1}^\alpha M_\alpha^{(k_2)}\big)(d\mathbf{x}) \Bigg)^{1/2},$$

where $d_* = d_\theta$ if $\theta = 1$ and $d_* = 0$ if $\theta > 1$.

Once again the off-diagonal terms dominate for the case $\theta > 1$.

In the final situation we consider in this section we have, once again, convergence in probability.

THEOREM 5. *Suppose that* $2 < \alpha < 2\theta$ *and that the sequence of levels satisfies* (19) *and* (30). *Then*

$$u_n \|\Delta_B\|_{\mathrm{F}} \to \frac{1}{w}\left(\sum_{k_1=1}^{p} \sum_{\substack{k_2=1\\ k_1\neq k_2}}^{p} {k_1}^{\alpha} w_{k_2}^{\alpha} \mathbb{E}\big[\big(\nabla R(s_{k_1}-s_{k_2}), S_1^{(k_1)} - S_1^{(k_2)}\big)\big]^2\right)^{1/2}$$

in probability.

The theoretical results of this section rigorously establish the precise rate at which the Frobenius norm of the perturbation matrix vanishes when the sample size tends to infinity, when performing kernel PCA on data generated from a linear factor model. This explains why kernel PCA preimages correctly find the underlying clusters of extremes in this model. Our numerical experiments will show that empirically kernel PCA performs well in many additional settings, including models where the angular measure is continuous. Our experiments rely on the gradient-based optimization framework discussed next.

5. Computational considerations. Our task is to obtain low-dimensional kernel PCA representations of the observations in their natural domain (the unit sphere in the case of extremes), usually referred to as kernel PCA preimages (of the low-dimensional representation of the observations in the RKHS). There have been numerous proposals for recovering such kernel PCA preimages, including a fixed point iteration in [25], the multidimensional scaling-based procedure of [23], and penalized methods of [2, 38], among others; see [19] for an overview.

We formally define a preimage as a solution to the optimization problem (5), which we repeat for convenience here:

$$T(\mathbf{w}) = \mathrm{argmax}_{\mathbf{v}\in\mathcal{S}} \sum_{k=1}^{m} \sum_{\mathbf{t}_j\in T} v_{kj} R(\mathbf{w}-\mathbf{t}_j) \sum_{\mathbf{t}_j\in T} v_{kj} R(\mathbf{v}-\mathbf{t}_j). \tag{32}$$

One can in principle solve the problem above by Monte Carlo up to arbitrary numerical precision. However, since this involves maximization over a potentially high-dimensional sphere is involved, computational issues are important.

In our implementation we employed projected gradient descent, which is a standard algorithm for optimizing smooth objective functions under convex constraints. Denoting the function being maximized in (32) by $f(\mathbf{v})$, the algorithm is defined by the iterates

$$\mathbf{v}^{(k)} = \Pi_{\mathcal{S}}(\mathbf{v}^{(k-1)} - \eta\nabla f(\mathbf{v}^{(k-1)})),$$

where $\eta > 0$ is a fixed step-size parameter, ∇ is the gradient and the projection operator Π to $\mathcal{S}$ is defined as

$$\Pi_{\mathcal{S}}(\mathbf{x}) = \underset{\mathbf{y}\in\mathcal{S}}{\mathrm{argmin}}\, \|\mathbf{x}-\mathbf{y}\|_2.$$

In the sequel we use the Gaussian kernel $R(\mathbf{x}) = \exp(-\gamma\|\mathbf{x}\|_2^2)$, as it is perhaps the most popular kernel function used in machine learning.

In order to implement this algorithm, one also needs to calculate the Lipschitz constant β of the function $f(\mathbf{v})$ in order to set a stepsize $\eta \leq 1/\beta$.

A direct calculation shows that, in the case of the Gaussian kernel, it suffices to set the stepsize to be

$$\eta = \left(2\left\|\sum_{k=1}^{m}\sum_{\mathbf{t}_j\in T} \mathbf{v}_j v_{kj} R(\mathbf{w}-\mathbf{t}_j)\right\|\right)^{-1}$$

in order to ensure the converge of projected gradient descent. In the last equation $\mathbf{v}_j = (v_{j1},\ldots,v_{jd})^\top$ denotes the jth eigenvector of the kernel matrix $C_n = \{R(\mathbf{x}_i - \mathbf{x}_j)\}_{i,j=1}^n$.

6. Empirical study. We have chosen several numerical examples to illustrate the performance of kernel PCA in scenarios that go well beyond the linear factor model studied in detail in Section 4. In particular, we consider a contaminated linear factor model and three examples where the spectral measure is continuous. In all the examples considered below we compute weighted adjacency matrices using the Gaussian kernel $\kappa(\mathbf{x},\mathbf{y}) = \exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$ with $\gamma = 1$ unless explicitly stated otherwise. We select the number m of the largest eigenvalues of the covariance operator to use kernel PCA as suggested by the screeplots of the kernel matrix. In most of the examples with a well-defined true number of clusters of extremes, the choice of m suggested by the screeplot matched that number. In all our examples we generate $10,000$ observations and take a sample of extremes defined as the 200 observations with the largest Euclidean norms.

In all the examples below the random vector of interest $\mathbf{X}$ is regularly varying, a commmon assumption in studying heavy-tailed data. This means $\|\mathbf{X}\|$ is regularly varying with index $\alpha > 0$ and the angular part $\mathbf{X}/\|\mathbf{X}\|$ is independent of the radius as the radius becomes large. Formally,

$$\lim_{r\to\infty} \mathbb{P}\big(\mathbf{X}/\|\mathbf{X}\| \in \cdot \mid \|\mathbf{X}\| > r\big) \Rightarrow \Gamma(\cdot)$$

and

$$\lim_{r\to\infty} \frac{\mathbb{P}\big(\|\mathbf{X}\| > rx\big)}{\mathbb{P}\big(\|\mathbf{X}\| > r\big)} = x^{-\alpha}$$

for all $x > 0$. The limit probability measure Γ is called the *angular measure* or *spectral measure* and describes how likely the extremal observations are to point in different directions. In other words, the angular measure describes the limiting extremal angle for high threshold exceedances that correspond to large $\|\mathbf{X}\|$. The support of this measure is particularly important since it shows which directions of the extremes are feasible and which are not feasible.

6.1. *Contaminated linear factor model.* The extremes from the linear factor model (11) have a discrete spectral measure and we expect kernel PCA applied to these extremes to concentrate the extremes near the atoms of the spectral measure with the largest masses. What happens if we"contaminate" this spectral measure by a small continuous component? We investigate this question empirically by considering the extremes arising from the model

$$\mathbf{X}_i = A\mathbf{Z}_i + \sigma\varepsilon_i, \quad i = 1,\dots,n,$$

with $\{\mathbf{Z}_i\}_{i=1}^n$ i.i.d. copies of the vector $\mathbf{Z}$ in (11), and ε is the "contamination" vector. In this case $\sigma \geq 0$ regulates the level of contamination. We choose the vector $\mathbf{Z}$ to consist of i.i.d. standard Fréchet[3] components and obtain ε by multiplying the element-wise absolute values of a standard p-dimensional normal random vector by univariate standard Fréchet random variable (with all random objects independent). The contamination adds a uniform component to the spectral measure, the weight of which is proportional to σ. In our simulation we take $d = 4$, $p = 2$, and use the matrix

$$A = \begin{pmatrix} 0.1 & 0.9 \\ 0.2 & 0.8 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix} \tag{33}$$

and choose $\sigma = 1$, which leads to a sample of extremes where approximately half can be assigned to the signal of the latent factors.

[3] The standard Fréchet distribution is $F(x) = e^{-x^{-1}}$, $x \geq 0$.

As the screeplot in Figure 1 indicates, we use kernel PCA with $m = 2$. Notice that the preimages of the kernel PCA shown on the right panel of Figure 2 show the two-dimensional nature of the support of the uncontaminated spectral measure much more clearly than the original extremes do. In particular, we see clearly that the extremes generated from the linear factor model are mapped close to two points while most of the extremes due to the“noise” are mapped in between these two points.

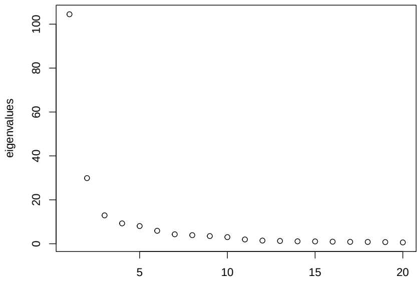
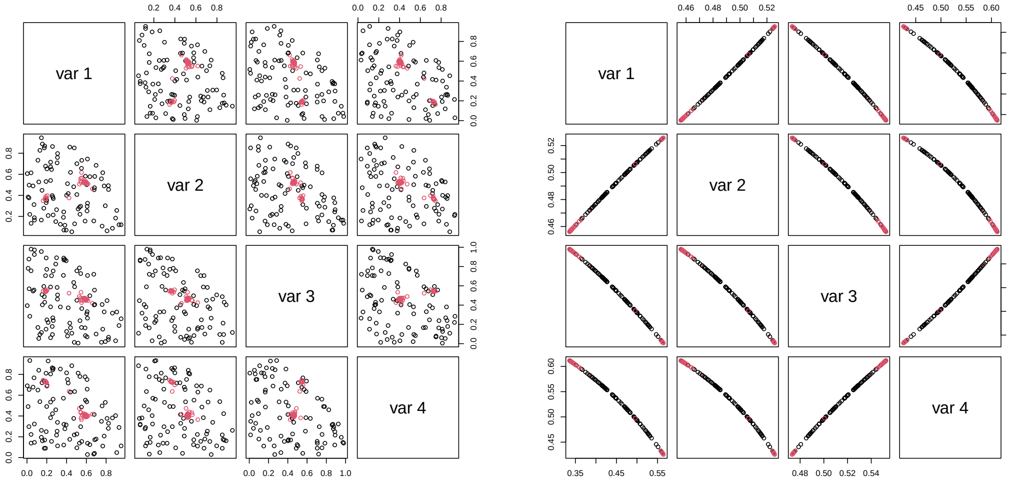


FIG 1. *Largest 20 eigenvalues of the kernel matrix used to run kernel PCA on extremes from a contaminated linear factor model*



(a) Contaminated linear factor model data

(b) Preimages

FIG 2. *Pairwise scatterplots of the angular part of the extremes generated from a contaminated linear factor model and their corresponding kernel PCA preimages. The red points denote extremes attributed to the signal $A\mathbf{Z}_i$. The black points denote extremes attributed to the noise $\sigma\varepsilon_i$.*

6.2. *Spiked angular Gaussian model.* We consider extremes arising from the model

$$\mathbf{X}_i = u_i \mathbf{N}_i + \sigma \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n,$$

where $\{u_i\}_{i=1}^n$ are i.i.d. univariate standard Fréchet random variables, $\{\mathbf{N}_i\}_{i=1}^n$ are i.i.d. d-dimensional centered normal vectors with covariance matrix of the form

$$\Sigma = \sum_{k=1}^{p} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top + \sigma_0^2 I_d \tag{34}$$

for $1 \leq p \leq d$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and the vectors $\mathbf{v}_1, \dots, \mathbf{v}_p$ are orthonormal, and the terms $\{\sigma \boldsymbol{\varepsilon}_i\}_{i=1}^n$ are "contamination" terms of the same type as in the contaminated linear factor model example. The covariance matrix Σ in (34) is a popular model that received a lot of attention in the machine learning and high dimensional (but non-extreme) statistics in recent years.

We note that when $\sigma = 0$, the spectral distribution on the d-dimensional sphere is given by

$$\Gamma(\cdot) = \mathbb{E}\left(\|\mathbf{N}\| \delta_{\frac{\mathbf{N}}{\|\mathbf{N}\|}}(\cdot) \right) / C, \tag{35}$$

where $C = \mathbb{E}\|\mathbf{N}\|$ (see equation (6.4) in [1]). Assuming the model in (34), the spectral distribution Γ has a spiked angular central Gaussian distribution[4] on the d-dimensional sphere $\mathbb{S}^{d-1}$ with density function given by

$$g(\boldsymbol{\omega}; \Sigma) = C^{-1} \frac{2\pi^{d/2}}{\Gamma(d/2)} |\Sigma|^{-1/2} (\boldsymbol{\omega}^\top \Sigma^{-1} \boldsymbol{\omega})^{-(d+1)/2}, \quad \boldsymbol{\omega} \in \mathbb{S}^{d-1}.$$

Intuitively, this model generates r clusters of extremes corresponding to higher density regions of the angular Gaussian distribution given by the principal directions of Σ.

In our experiment we take $d = 4$ and $p = 2$, and use $\Sigma = BB^\top$ where

$$B^\top = \begin{pmatrix} 0.1 \, 0.2 \, 0.3 \, 0.4 \\ 0.9 \, 0.8 \, 0.7 \, 0.6 \end{pmatrix},$$

so that $\mathbf{v}_1$ and $\mathbf{v}_2$ are the left singular vectors of B. We have chosen $\sigma = 0.1$ and $\sigma_0 = 1$.

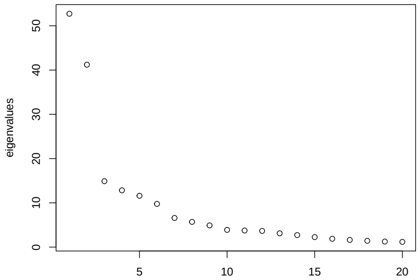
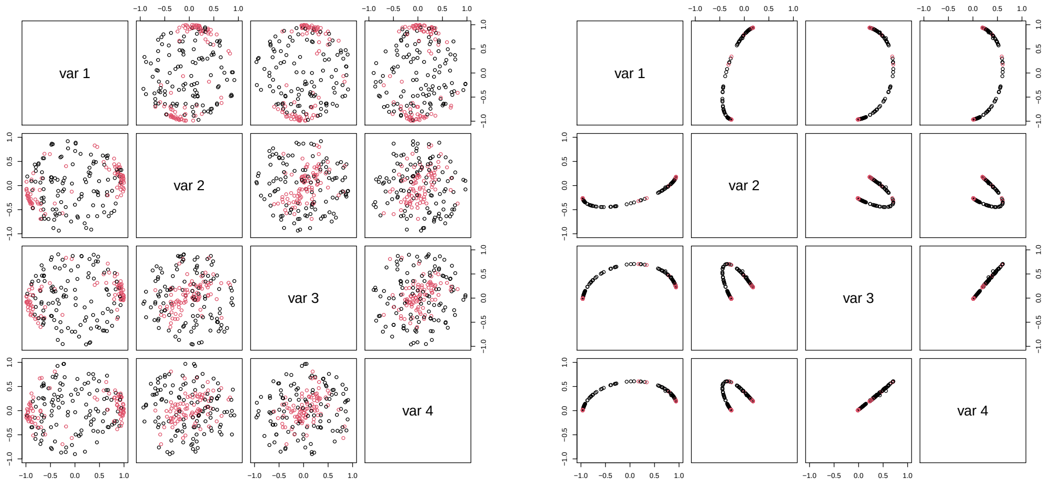


FIG 3. *Largest 20 eigenvalues of the kernel matrix used to run kernel PCA for the 4 dimensional contaminated spiked angular Gaussian model.*

According to the screeplot of Figure 3 we use kernel PCA with $m = 2$. Once again, the dramatic dimension reduction of the support of the extremes after going through kernel PCA is clear in Figure 4.

[4] Note that this is not exactly the same angular Gaussian distribution of [35] because of the exponent $-(d+1)/2$ due to the presence of $\|\mathbf{N}\|$ in (35).



(a) Contaminated spiked angular Gaussian data

(b) Preimages, $\gamma = 0.5$

FIG 4. *Pairwise scatterplots of the extremes generated from the spiked angular Gaussian model with $d = 4$ and $r = 2$. The red points denote extremes attributed to the signal $u_i\mathbf{Z}_i$. The black points denote extremes attributed to the noise $\sigma\varepsilon_i$.*

6.3. *Approximate subspace model: regularly varying circle.* We also consider a model where the ambient dimension is 5 but the signal is driven by a 3-dimensional regularly varying vector with spectral measure supported on a circle. More specifically, we consider the model

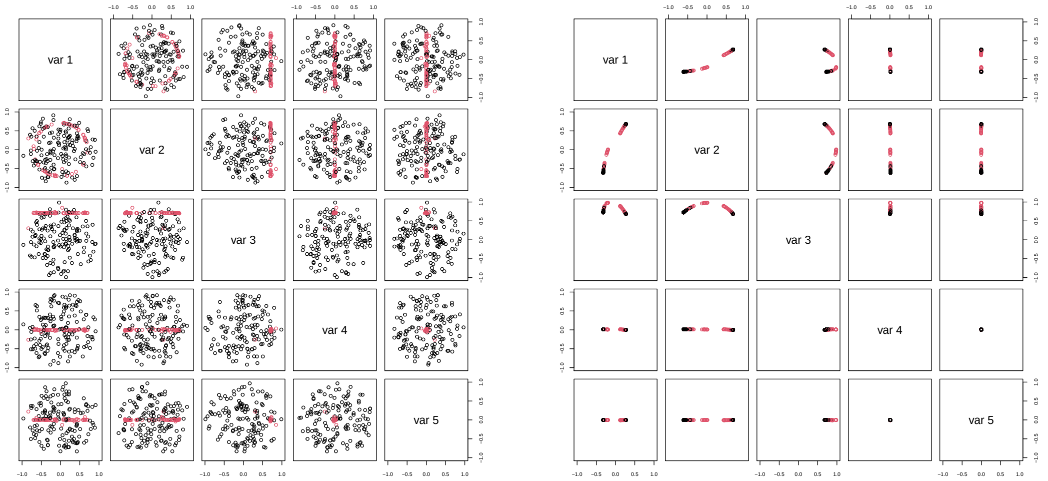
$$\mathbf{X}_i = \mathbf{Z}_i + \sigma\boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n,$$

where $\mathbf{Z}_i \in \mathbb{R}^5$ is such that its last 2 components are 0 and its first 3 entries are of the form

$$Z_{i1} = Y_i G_{i1}, \quad Z_{i1} = Y_i G_{i2}, \quad Z_{i3} = Y_i\{(G_{i1}^2 + G_{i2}^2)^{1/2}\},$$

where Y_i is an i.i.d. sequence of standard Fréchet random variables and $\boldsymbol{G}_i$ is a sequence of bivariate standard i.i.d. normal random variables. It follows that $\mathbf{Z}_i$ is regularly varying, and its spectral measure is uniform on the circle $\{(z_1, z_2, z_3) : z_1^2 + z_2^2 = 1/2,\ z_3 = 1/\sqrt{2}\}$. The constant $\sigma > 0$ regulates the signal to noise ratio and $\boldsymbol{\varepsilon}_i$ is noise vector obtained by multiplying a univariate independent standard Fréchet with an independent p-dimensional composed by the absolute value of i.i.d. standard normals. In our example we chose $\sigma = 2$ which leads to about about 60 out of 200 extremes generated from the signal of the circle. We see from Figure 5 that very much like the last two examples, the kernel PCA preimage map the data to a lower dimensional subspace where distinguished the locations of the signal and the noise terms. In particular, we observe that the preimages correctly identify the structure of the subspace, mapping to variables 4 and 5 collapse to one point and identifying the straight lines of the signal as seen in the original colored data points. The subspace corresponding to the circle is distorted but the method recognizes seem to recognize that there is lower dimensional structure in the first 3 coordinates.

6.4. *Extremes from time series: ARCH(1) process.* When the extremes arise from a time series model, the independence assumption is, generally, violated. Here we consider the square of a standard *integrated* ARCH(1) process that follows the recursions, $Y_t = (1 + Y_{t-1})Z_t^2$, where $\{Z_t\}$ is an i.i.d. sequence of standard Gaussian random variables.



(a) Noisy subspace data

(b) Preimages

Fig 5. *Pairwise scatterplots of the extremes generated from the noisy circle subspace model. The red points denote extremes attributed to the subspace signal. The black points denote extremes attributed to the noise $\sigma\varepsilon_i$.*

There is a unique stationary solution $\{Y_t\}$ to these recursions such that Y_t has asymptotically Pareto tails with $\alpha = 1$ (see [4] and [3] for more details). Here we consider the two-dimensional vector $\mathbf{X}_t = (Y_{t-1}, Y_t)^\top$, whose tails are asymptotically equivalent to those of the vector $(1, Z_t^2)^\top Y_{t-1}$. By an application of Breiman's lemma [4, Proposition A.1], the spectral distribution of this vector on $[0, \pi/2]$ is given by

$$\Gamma(\cdot) = \mathbb{E}\left(|1 + Z_t^4|^{1/2}\delta_{\arctan(Z_t^2)}(\cdot)\right)/\mathbb{E}|1 + Z_t^4|^{1/2} .$$

The spectral measure can be shown to be bimodal as is also suggested by a kernel density estimator obtained from the angles of the empirical sample of extremes displayed in Figure 6. We look for clusters in the extremes of $\|\mathbf{X}_t\|_2$ and note that the spectral measure of this example is continuous and hence it is less clear what the correct number of clusters should be. The preimages shown in Figure 7 where obtained with kernel PCA with $m = 3$, as suggested by the screeplot. The bottom right plot displays reveals that the distribution of the angles of the preimages remains bimodal with slightly more pronounced modes with more mass in between them. This is also reflected in the top right plot where one can perceive a third cluster of preimages.

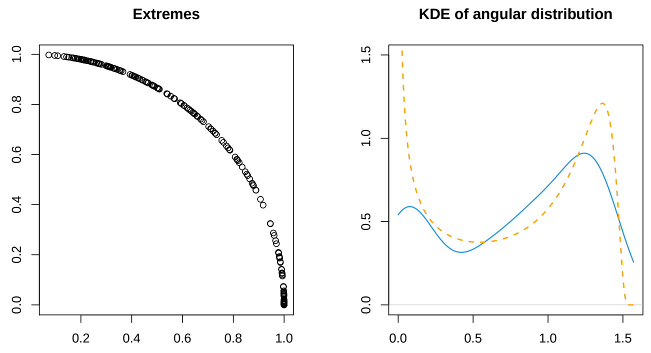


FIG 6. *The plot on the left shows the empirical sample of extremes $\mathbf{X}_t/\|\mathbf{X}_t\|$. The right hand side shows two density functions. The solid blue line is the kernel density estimator of the angles of the extremes fitted with default values of the* R *function* `density`$(\cdot)$ *i.e., using the Gaussian kernel with Silverman's rule of thumb. The dashed orange line corresponds to the theoretical spectral density.*

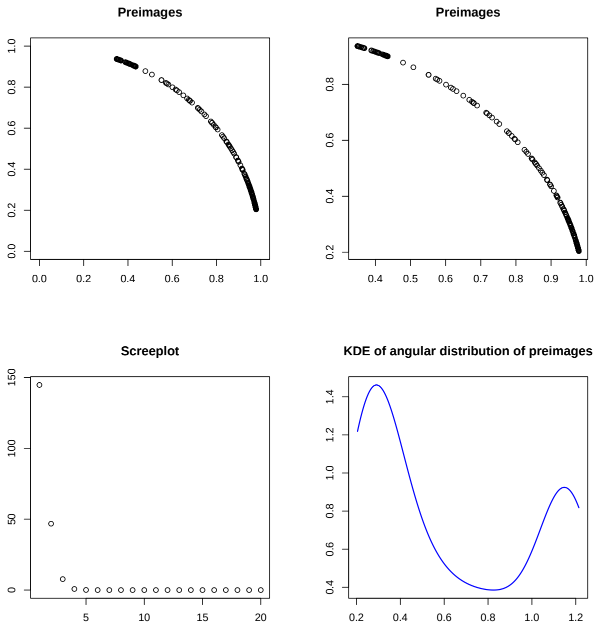


FIG 7. *The top plots show the kernel PCA preimages in the original scale of the data and on the scale of the preimages. The bottom plot shows the screeplot of the kernel matrix and the kernel density estimator of the angle of the preimages fitted with default values of the* R *function* `density`$(\cdot)$.

SUPPLEMENT.

Proof of Theorem 2. We start by considering the "diagonal terms" in (22) and note that for $k = 1, \ldots, p$,

$$F_{k,k}(n) = \frac{1}{N_n^2} \sum_{i \in \mathcal{I}_n^{(k)}} \sum_{j \in \mathcal{I}_n^{(k)}} \Big[R\big(\mathbf{X}_i/\|\mathbf{X}_i\| - \mathbf{X}_j/\|\mathbf{X}_j\|\big) - R(0) \Big]^2$$

$$= \frac{1}{N_n^2} \sum_{i=1}^{N_n^{(k)}} \sum_{j=1}^{N_n^{(k)}} \left[R\big(\mathbf{Y}_i^{(k)} - \mathbf{Y}_j^{(k)}\big) - R(0) \right]^2 = \frac{(N_n^{(k)})^2}{N_n^2} G_{k,k}(n) \,.$$

Recall that the law of each $\mathbf{Y}_i^{(k)}$, $i = 1, \dots, N_n^{(k)}$ is the conditional law of $\mathbf{X}/\|\mathbf{X}\|$ given $\|\mathbf{X}\| > u_n$, $Z_k > u_n/w^{1/\alpha}$, different $\mathbf{Y}_i^{(k)}$ are independent, and also independent of N_n and $N_n^{(k)}$. For a large $M > 0$ write

$$\begin{aligned} G_{k,k}(n) =& \frac{1}{(N_n^{(k)})^2} \sum_{i=1}^{N_n^{(k)}} \sum_{j=1}^{N_n^{(k)}} \left[R\left(\mathbf{Y}_i^{(k)} - \mathbf{Y}_j^{(k)}\right) - R(0) \right]^2 \mathbf{1}\left(\left\| \mathbf{Y}_i^{(k)} - \mathbf{Y}_j^{(k)} \right\| > M u_n^{-1} \right) \\ &+ \frac{1}{(N_n^{(k)})^2} \sum_{i=1}^{N_n^{(k)}} \sum_{j=1}^{N_n^{(k)}} \left[R\left(\mathbf{Y}_i^{(k)} - \mathbf{Y}_j^{(k)}\right) - R(0) \right]^2 \mathbf{1}\left(\left\| \mathbf{Y}_i^{(k)} - \mathbf{Y}_j^{(k)} \right\| \le M u_n^{-1} \right) \\ &=: G_{k,k}^{>M}(n) + G_{k,k}^{\le M}(n). \end{aligned}$$

Clearly,

$$\begin{aligned} &\mathbb{E}[u_n^{2\theta} G_{k,k}^{>M}(n)] \\ &\quad \le u_n^{2\theta} \mathbb{E}\left\{ \left[R\left(\mathbf{Y}_1^{(k)} - \mathbf{Y}_2^{(k)}\right) - R(0) \right]^2 \mathbf{1}\left(\left\| \mathbf{Y}_1^{(k)} - \mathbf{Y}_2^{(k)} \right\| > M u_n^{-1} \right) \right\}. \end{aligned}$$

If we can show that

$$\sup_n u_n^{(2+\varepsilon)\theta} \mathbb{E}\Big[R\big(\mathbf{Y}_1^{(k)} - \mathbf{Y}_2^{(k)}\big) - R(0) \Big]^{2+\varepsilon} < \infty \tag{36}$$

for some $\varepsilon > 0$ (in which case, $u_n^{2\theta} \left(R(\mathbf{Y}_1^{(k)} - \mathbf{Y}_2^{(k)}) - R(0) \right)^2$ is uniformly integrable), it will follow from Theorem 1 that

$$\limsup_{n\to\infty} \mathbb{E}\left[u_n^{2\theta} G_{k,k}^{>M}(n) \right] \le d_\theta^2 \mathbb{E}\Big\{ \big\| \mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)} \big\|^{2\theta} \mathbf{1}\big(\big\| \mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)} \big\| > M/d_\theta \big) \Big\}. \tag{37}$$

By (14) the bound (36) will follow once we check that

$$\sup_n u_n^{(2+\varepsilon)\theta} \mathbb{E}\left[\left\| \mathbf{Y}_1^{(k)} - \mathbf{s}_k \right\|^{(2+\varepsilon)\theta} \right] < \infty. \tag{38}$$

Using the notation in (4.13) and (4.14) in [1] it suffices to check that for any $l = 1, \dots, d$,

$$\sup_n \mathbb{E}\left[\left| w_j^2 \left(\sum_{m=1}^{p} a_{lm} Z_m \right)^2 - a_{lj}^2 \|\mathbf{X}\|^2 \right|^{(2+\varepsilon)\theta} \right] < \infty.$$

The expectation, however, is independent of n and is finite for $\epsilon > 0$ sufficiently small since $\alpha > 2\theta$ by assumption. This establishes (37), whence

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[u_n^{2\theta} G_{k,k}^{>M}(n) \right] = 0. \tag{39}$$

The same argument shows that for every fixed $M > 0$,

$$\lim_{n\to\infty} \mathbb{E} u_n^{2\theta} G_{k,k}^{\le M}(n) = d_\theta^2 E\Big\{ \big\| \mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)} \big\|^{2\theta} \mathbf{1}\big(\big\| \mathbf{S}_1^{(k)} - \mathbf{S}_2^{(k)} \big\| \le M/d_\theta \big) \Big\}. \tag{40}$$

Furthermore,

$$\mathbb{E}\big(u_n^{2\theta}G_{k,k}^{\le M}(n)\big)^2 = \mathbb{E}\left(\frac{N_n^{(k)}-1}{(N_n^{(k)})^3}\mathbf{1}\left(N_n^{(k)}\ge 1\right)\right)E_n^{(1)} + 4\mathbb{E}\left(\frac{(N_n^{(k)}-1)^2}{(N_n^{(k)})^3}\mathbf{1}\left(N_n^{(k)}\ge 1\right)\right)E_n^{(2)}$$

$$+\,\mathbb{E}\left(\frac{(N_n^{(k)}-1)(N_n^{(k)}-2)(N_n^{(k)}-3)}{(N_n^{(k)})^3}\mathbf{1}\left(N_n^{(k)}\ge 1\right)\right)E_n^{(3)}, \tag{41}$$

where

$$E_n^{(1)} = \mathbb{E}\left\{u_n^{4\theta}\left[R\left(\mathbf{Y}_1^{(k)}-\mathbf{Y}_2^{(k)}\right)-R(0)\right]^4\mathbf{1}\left(\left\|\mathbf{Y}_1^{(k)}-\mathbf{Y}_2^{(k)}\right\|\le Mu_n^{-1}\right)\right\}$$

$$E_n^{(2)} = \mathbb{E}\left\{u_n^{4\theta}\left[R\left(\mathbf{Y}_1^{(k)}-\mathbf{Y}_2^{(k)}\right)-R(0)\right]^2\left[R\left(\mathbf{Y}_1^{(k)}-\mathbf{Y}_3^{(k)}\right)-R(0)\right]^2\right.$$
$$\left.\cdot\mathbf{1}\left(\left\|\mathbf{Y}_1^{(k)}-\mathbf{Y}_2^{(k)}\right\|\le Mu_n^{-1},\ \left\|\mathbf{Y}_1^{(k)}-\mathbf{Y}_3^{(k)}\right\|\le Mu_n^{-1}\right)\right\}$$

$$E_n^{(3)} = \left(\mathbb{E}\left\{u_n^{4\theta}\left[R\left(\mathbf{Y}_1^{(k)}-\mathbf{Y}_2^{(k)}\right)-R(0)\right]^2\mathbf{1}\left(\left\|\mathbf{Y}_1^{(k)}-\mathbf{Y}_2^{(k)}\right\|\le Mu_n^{-1}\right)\right\}\right)^2.$$

It follows from (14) that, for a fixed M, $E_n^{(1)}$ and $E_n^{(2)}$ are bounded by an M-dependent constant, so the first two terms in the right hand side of (41) vanish in the limit. Furthermore, it follows by (36) that

$$E_n^{(3)} \to \left(d_\theta^2\mathbb{E}\big\{\|\mathbf{S}_1^{(k)}-\mathbf{S}_2^{(k)}\|^{2\theta}\mathbf{1}\big(\|\mathbf{S}_1^{(k)}-\mathbf{S}_2^{(k)}\|\le M/d_\theta\big)\big\}\right)^2.$$

Therefore,

$$\lim_{n\to\infty}\mathbb{E}\big(u_n^{2\theta}G_{k,k}^{\le M}(n)\big)^2 = \big(\lim_{n\to\infty}\mathbb{E}u_n^{2\theta}G_{k,k}^{\le M}(n)\big)^2$$

and consequently

$$\lim_{n\to\infty}\mathrm{Var}\big(u_n^{2\theta}G_{k,k}^{\le M}(n)\big) = 0 \tag{42}$$

It follows from (40) and (42) that

$$u_n^{2\theta}G_{k,k}^{\le M}(n)\overset{P}{\to} d_\theta^2\mathbb{E}\big\{\|\mathbf{S}_1^{(k)}-\mathbf{S}_2^{(k)}\|^{2\theta}\mathbf{1}\big(\|\mathbf{S}_1^{(k)}-\mathbf{S}_2^{(k)}\|\le M/d_\theta\big)\big\},\ \text{as } n\to\infty. \tag{43}$$

Now (25) follows easily from (39) and (43). It follows from (25) and (21) that

$$u_n^{2\theta}F_{k,k}(n)\overset{P}{\to}\left(\frac{d_\theta w_k^\alpha}{w}\right)^2\mathbb{E}\|\mathbf{S}_1^{(k)}-\mathbf{S}_2^{(k)}\|^{2\theta},\ \text{as } n\to\infty. \tag{44}$$

We now consider the “off-diagonal terms” in (22) as described in (23). We claim that, under (27),

$$u_n^2G_{k_1,k_2}(n)\overset{P}{\to}\mathbb{E}\big[\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\big(\mathbf{S}_1^{(k_1)}-\mathbf{S}_1^{(k_2)}\big)\big]^2,\ \text{as } n\to\infty. \tag{45}$$

The argument is similar to the argument we used to prove (25). Once again we write for a large $M>0$

$$G_{k_1,k_2}(n) = \frac{1}{N_n^{(k_1)}N_n^{(k_2)}}\sum_{i=1}^{N_n^{(k_1)}}\sum_{j=1}^{N_n^{(k_2)}}\left[R\left(\mathbf{Y}_i^{(k_1)}-\mathbf{Y}_j^{(k_2)}\right)-R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\right]^2$$
$$\cdot\,\mathbf{1}\left(\left\|\mathbf{Y}_i^{(k_1)}-\mathbf{Y}_j^{(k_2)}-(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\right\|>Mu_n^{-1}\right)$$

$$
\begin{aligned}
&+\frac{1}{N_n^{(k_1)}N_n^{(k_2)}}\sum_{i=1}^{N_n^{(k_1)}}\sum_{j=1}^{N_n^{(k_2)}}\left[R\left(\mathbf{Y}_i^{(k_1)}-\mathbf{Y}_j^{(k_2)}\right)-R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\right]^2\\
&\qquad\qquad\qquad\qquad\cdot\mathbf{1}\left(\left\|\mathbf{Y}_i^{(k_1)}-\mathbf{Y}_j^{(k_2)}\right\|\le Mu_n^{-1}\right)\\
&=:G_{k_1,k_2}^{>M}(n)+G_{k_1,k_2}^{\le M}(n), \qquad (46)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[u_n^2G_{k_1,k_2}^{>M}(n)\right]\\
&=u_n^2\mathbb{E}\left\{\left[R\left(\mathbf{Y}_1^{(k_1)}-\mathbf{Y}_2^{(k_2)}\right)-R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\right]^2\cdot\mathbf{1}\left(\left\|\mathbf{Y}_1^{(k_1)}-\mathbf{Y}_2^{(k_2)}\right\|>Mu_n^{-1}\right)\right\}.
\end{aligned}
$$

By Theorem 1, if we can show that for some $\varepsilon>0$,

$$
\sup_n u_n^{2+\varepsilon}\mathbb{E}\left|R\left(\mathbf{Y}_1^{(k_1)}-\mathbf{Y}_2^{(k_2)}\right)-R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\right|^{2+\varepsilon}<\infty, \qquad (47)
$$

then it would follow that

$$
\lim_{n\to\infty}\mathbb{E}\left[u_n^2G_{k_1,k_2}^{>M}(n)\right]=\mathbb{E}\left[\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\big(\mathbf{S}_1^{(k_1)}-\mathbf{S}_2^{(k_2)}\big)\mathbf{1}\big(\|\mathbf{S}_1^{(k_1)}-\mathbf{S}_2^{(k_2)}\|>M/d_\theta\big)\right]^2. \qquad (48)
$$

However, because of the assumed continuous differentiability of R outside of the origin, (47) follows immediately from (38). Therefore, (48) holds. The rest of the argument for (45) is the same as for the "diagonal terms", and it follows from (45) that

$$
u_n^2F_{k_1,k_2}(n)\overset{P}{\to}\frac{w_{k_1}^\alpha w_{k_2}^\alpha}{w^2}\mathbb{E}\left[\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2})\big(\mathbf{S}_1^{(k_1)}-\mathbf{S}_2^{(k_2)}\big)\right]^2,\ \text{as } n\to\infty. \qquad (49)
$$

This completes the proof. □

Proof of Theorem 3. In order to establish the convergence in (31), it suffices to show that the corresponding Laplace functionals converge. That is, it is enough to prove that

$$
\mathbb{E}\left[\exp\{-M_n(f)\}\right]\to\mathbb{E}\left[\exp\{-M_\alpha(f)\}\right], \qquad (50)
$$

for any nonnegative, bounded continuous function f with support outside a neighborhood of the origin. Here $M(f)$ is shorthand for $\int_{\mathbb{R}^d\setminus\{0\}}f(\mathbf{x})M(d\mathbf{x})$; see Theorem 5.1 in [26] (this book can be consulted also for more details on Poisson processes and weak convergence to point processes).

Adapting Theorem 5.3 in [26] to random sums of point measures and using the fact that

$$
N_n^{(k)}\sim c_\alpha w_k^\alpha u_n^{-\alpha}n\ \text{ in probability as } n\to\infty.,
$$

it suffices to show that the intensity measures converges vaguely, i.e.,

$$
c_\alpha w_k^\alpha u_n^{-\alpha}n\mathbb{P}\big(u_n^2n^{-1/\alpha}\mathbf{Y}_0)\in\cdot\big)\overset{v}{\to}m_\alpha^{(k)}(\cdot), \qquad (51)
$$

where $\overset{v}{\to}$ denotes vague convergence of measures on $\mathbb{R}^p\setminus\{0\}$ and $\mathbf{Y}_0=\mathbf{Y}_1^{(k)}-\mathbf{s}_k$. To ease notation we write $\mathbb{P}_{k,n}(\cdot)=\mathbb{P}\left(\cdot\mid\|\mathbf{X}\|>u_n,Z_k>u_n/w^{1/\alpha}\right)$ and $\mathbb{E}_{k,n}$ to be the corresponding conditional expectation. We start by showing that

$$
\begin{aligned}
\nu_n(A\times B)&:=c_\alpha w_k^\alpha u_n^{-\alpha}n\mathbb{P}_{k,n}\left((u_nn^{-1/\alpha}\mathbf{Z}_{-k},u_n^{-1}Z_k)\in A\times B\right)\\
&\overset{v}{\to}c_\alpha w_k^\alpha\nu(A)\mu(B), \qquad (52)
\end{aligned}
$$

for all *bounded Borel sets* $A \times B \subset ([0,\infty)^{p-1} \setminus \{0\}) \times [w_k^{-1},\infty)$ that are also continuity sets of the limit measure, where ν is the measure on $[0,\infty]^{p-1} \setminus \{0\}$ given by

$$\nu(A) = \sum_{j \neq k} \int_0^\infty \mathbf{1}_{A^{(j)}}(x) \alpha x^{-\alpha-1}\, dx\,,$$

$\mathbf{Z}_{-k}$ is the vector $(Z_1 \dots, Z_p)^\top$ with the kth component omitted, $A^{(j)}$ is the intersection of the set A with the jth coordinate axis, and $\mu(x,\infty) = w_k^{-\alpha} x^{-\alpha}$, $x \geq w_k^{-1}$. To prove (52), consider $j=1$, $A = (x,\infty) \times \mathbb{R}^{p-2}$, $B = (y,\infty)$ with $y \geq w_k^{-1}$ and $k \neq 1$. Then

$$\begin{aligned} \nu_n(A \times B) &= c_\alpha w_k^\alpha u_n^{-\alpha} n \mathbb{P}_{k,n}\left(u_n n^{-1/\alpha} Z_1 > x, u_n^{-1} Z_k > y\right) \\ &\sim c_\alpha w_k^\alpha u_n^{-\alpha} n c_\alpha x^{-\alpha} u_n^\alpha n^{-1} w_k^{-\alpha} y^{-\alpha} \\ &\to c_\alpha^2 w_k^\alpha x^{-\alpha} w_k^{-\alpha} y^{-\alpha} \\ &= c_\alpha^2 w_k^\alpha \nu(A) \mu(B)\,. \end{aligned}$$

The argument for general choices of sets A and B is similar.

We now show how to derive (51) from (52) using the continuous mapping theorem. On the event $\|\mathbf{X}\| > u_n, Z_k > u_n / w^{1/\alpha}$, it follows that $\|\mathbf{X}\| \sim w_k |Z_k|$, so the expression in (4.13) of [1] corresponding to $(V_1, \dots, V_p)^\top = u_n \mathbf{Y}_0$ can be written as

$$V_l = u_n \frac{2a_{lk} Z_k T_{lk} + w_k X_{l,-k}^2 - a_{lk}^2 \sum_{i=1}^d X_{i,-k}^2}{2 w_k^2 a_{lk} Z_k^2} (1 + o(1))\,, \tag{53}$$

where

$$\begin{aligned} T_{lk} &= w_k^2 X_{l,-k} - a_{lk} \sum_{i=1}^d a_{ik} X_{i,-k} \\ &= w_k^2 \sum_{j \neq k} b_l^{(j,k)} Z_j\,, \end{aligned}$$

and

$$\hat{b}_l^{(j,k)} = a_{lj} - \frac{a_{lk}}{w_k^2} \sum_{i=1}^d a_{ik} a_{ij}\,.$$

Note that for any $i = 1, \dots, d$, and $c > 0$ a constant that may change from line to line, we have by (56) below, with $z = c n^{1/\alpha} u_n^{-1}$,

$$\begin{aligned} u_n^{-\alpha} n \mathbb{P}_{k,n}\left(u_n \frac{X_{i,-k}^2}{Z_k^2} > c n^{1/\alpha} u_n^{-1} \right) &\leq C u_n^{-\alpha} n u_n^{-\alpha/2} n^{-1/2} u_n^{\alpha/2} \\ &\sim C u_n^{-\alpha} n^{1/2} \to 0 \end{aligned}$$

as $n \to \infty$ by (30). Therefore, writing $\tilde{\mathbf{V}} = (\tilde{V}_1, \dots, \tilde{V}_l)^T$ with

$$\tilde{V}_l = \frac{u_n T_{lk}}{w_k^2 Z_k} = u_n \frac{\sum_{j \neq k} \hat{b}_l^{(j,k)} Z_j}{Z_k}, \quad l = 1, \dots, d,$$

we have

$$c_\alpha w_k^\alpha u_n^{-\alpha} n \mathbb{P}\big(u_n^2 n^{-1/\alpha} \mathbf{Y}_0 \in \cdot\big) = c_\alpha w_k^\alpha u_n^{-\alpha} n \mathbb{P}_{k,n}(u_n n^{-1/\alpha} \tilde{\mathbf{V}} \in \cdot) + o(1)\,. \tag{54}$$

So to finish the proof, it suffices to show

$$c_\alpha w_k^\alpha u_n^{-\alpha} n \mathbb{P}_{k,n}(u_n n^{-1/\alpha} \tilde{\mathbf{V}} \in \cdot) \xrightarrow{v} m_\alpha^{(k)}(\cdot)\,. \tag{55}$$

Consider the continuous mapping from $E := ([0,\infty)^{p-1} \setminus \{0\}) \times [w_k^{-1}, \infty)$ to $\mathbb{R}^d \setminus \{0\}$ given by

$$T(z_{-k}, y) = \left(\sum_{j \neq k} \hat{b}_l^{(j,k)} z_j / y \, , l = 1, \dots, d \right)$$
$$= \sum_{j \neq k} \hat{\mathbf{b}}^{(j,k)} z_j / y \, ,$$

where $\hat{\mathbf{b}}^{(j,k)} = \left(\hat{b}_1^{(j,k)}, \dots, \hat{b}_d^{(j,k)} \right)^\top$. Since this mapping has the property that for a compact set K in E, $T^{-1}(K)$ is also compact in E, it follows from Proposition 5.2 in [26] that for any Borel set A in $\mathbb{R}^d$ bounded away from 0 with $\nu \times \mu(\partial(T^{-1}(A)) = 0$, the left hand side of (55) is

$$\begin{aligned} \nu_n \circ T^{-1}(A) &= c_\alpha w_k^\alpha u_n^{-\alpha} n \mathbb{P}_{k,n} \left(T(u_n n^{-1/\alpha} \mathbf{Z}_{-k}, u_n^{-1} Z_k) \in A \right) \\ &= c_\alpha w_k^\alpha u_n^{-\alpha} n \mathbb{P}_{k,n} \left(\frac{u_n n^{-1/\alpha} \sum_{j \neq k} \hat{\mathbf{b}}^{(j,k)} Z_j}{u_n^{-1} Z_k} \in A \right) \\ &\to c_\alpha^2 w_k^\alpha (\nu \times \mu) \circ T^{-1}(A) \, . \end{aligned}$$

Finally, by Fubini's theorem,

$$\begin{aligned} c_\alpha^2 w_k^\alpha (\nu \times \mu) \circ T^{-1}(\cdot) &= c_\alpha^2 w_k^\alpha \sum_{j \neq k} \int_{w_k^{-1}}^\infty \left(\int_0^\infty \alpha z^{-\alpha-1} \delta_{\hat{\mathbf{b}}^{(j,k)} z/y}(\cdot) \, dz \right) \alpha w_k^{-\alpha} y^{-\alpha-1} dy \, , \\ &= c_\alpha^2 w_k^\alpha \sum_{j \neq k} \int_0^\infty \alpha z^{-\alpha-1} \delta_{\hat{\mathbf{b}}^{(j,k)} z}(\cdot) \, dz \int_{w_k^{-1}}^\infty \alpha w_k^{-\alpha} y^{-2\alpha-1} dy \, , \\ &= (c_\alpha^2 w_k^{2\alpha}/2) \sum_{j \neq k} \int_0^\infty \alpha z^{-\alpha-1} \delta_{\hat{\mathbf{b}}^{(j,k)} z}(\cdot) \, dz = m_\alpha^{(k)}(\cdot) \, , \end{aligned}$$

proving (54).

Finally, the conditional independence of the extremes along with the laws of large numbers in (20) and (24) show the required joint convergence along with the independence in the limit. □

Below is a statement, needed in the sequel. A part of it was already used in the proof of Theorem 3.

PROPOSITION 1. *Under the assumptions of Theorem 3, there is $C > 0$ independent of n and k such that for all $z > 0$,*

$$\mathbb{P}_{k,n} \left(u_n X_{i,-k}^2 / Z_k^2 > z \right) \le C u_n^{-\alpha/2} z^{-\alpha/2} \, , \tag{56}$$

$$\mathbb{P}_{k,n} \left(u_n |T_{lk}| / Z_k > z \right) \le C z^{-\alpha}, \tag{57}$$

$$\mathbb{E}_{k,n} \left(|V_l|^{2\theta} \mathbf{1}\{|V_l| \le z\} \right) \le C z^{2\theta-\alpha} \, . \tag{58}$$

PROOF. Since $|X_{i,-k}| \le a^* \sum_{j \neq k} Z_j$, where $a^* = \max\{a_{ij}, i = 1, \ldots, d; j = 1, \ldots, p\}$, it follows that

$$\mathbb{P}(X_{i,-k} > z) \le \sum_{j \neq k} \mathbb{P}(a^* Z_j > z/p) \le C z^{-\alpha}$$

for all z and some $C > 0$ by assumption (12). Thus,

$$\begin{aligned}\mathbb{P}_{k,n}\left(u_n X_{i,-k}^2 / Z_k^2 > z\right) &\le \frac{\mathbb{P}\left(X_{i,-k}^2 > z u_n^{-1} Z_k^2, Z_k > u_n / w^{1/\alpha}\right)}{\mathbb{P}\left(\|\mathbf{X}\| > u_n, Z_k > u_n / w^{1/\alpha}\right)} \\ &\le \frac{\mathbb{P}\left(X_{i,-k} > z^{1/2} u_n^{1/2} / w^{1/\alpha}\right) \mathbb{P}\left(Z_k > u_n / w^{1/\alpha}\right)}{\mathbb{P}\left(\|\mathbf{X}\| > u_n, Z_k > u_n / w^{1/\alpha}\right)} \\ &\le C u_n^{-\alpha/2} z^{-\alpha/2},\end{aligned}$$

where C may change value from line to line, and we have used the relation, $\mathbb{P}(Z_k > u_n / w^{1/\alpha}) / \mathbb{P}(\|\mathbf{X}\| > u_n, Z_k > u_n / w^{1/\alpha}) \to w_k^{-\alpha}$. This proves (56). The same argument shows that

$$\begin{aligned}\mathbb{P}_{k,n}\left(u_n |T_{lk}| / Z_k > z\right) &\le \frac{\mathbb{P}(|T_{lk}| > z / w^{1/\alpha}) \mathbb{P}(Z_k > u_n / w^{1/\alpha})}{\mathbb{P}\left(\|\mathbf{X}\| > u_n, Z_k > u_n / w^{1/\alpha}\right)} \\ &\le C z^{-\alpha},\end{aligned}$$

for all $z \ge 0$, proving (57). Finally, it is straightforward to see from (56), (57) and (53) that

$$\mathbb{P}_{k,n}(|V_l| > z) \le C z^{-\alpha}, \tag{59}$$

for some constant $C > 0$. Therefore,

$$\begin{aligned}\mathbb{E}_{k,n}\left(|V_l|^{2\theta} \mathbf{1}\{|V_l| \le z\}\right) &\le 2\theta \int_0^z u^{2\theta-1} \mathbb{P}_{k,n}(|V_l| > u)\, du \\ &\le C \int_0^z u^{2\theta-1} u^{-\alpha}\, du = C z^{2\theta-\alpha}.\end{aligned}$$

□

Proof of Theorem 4. We start with the "diagonal terms". It follows from (14) and Theorem 1 that

$$F_{k,k}(n) = d_\theta^2 \frac{1}{N_n^2} \sum_{i \in \mathcal{I}_n^{(k)}} \sum_{j \in \mathcal{I}_n^{(k)}} \|\mathbf{Y}_i^{(k)} - \mathbf{Y}_j^{(k)}\|^{2\theta} + o_P(1). \tag{60}$$

By (20),

$$u_n^{4\theta-\alpha} n^{1-2\theta/\alpha} F_{k,k}(n) \sim \frac{d_\theta^2}{w^2 c_\alpha^2} u_n^\alpha n^{-1} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{x})\, M_n^{(k)}(d\mathbf{y}). \tag{61}$$

Notice that the scaling for $F_{kk}(n)$ is $u_n^{4\theta-\alpha} n^{1-2\theta/\alpha} = (u_n^2 n^{-1/\alpha})^{2\theta} u_n^{-\alpha} n \to \infty$.

For $\varepsilon > 0$, set $B_\epsilon = \{\mathbf{y} : \|\mathbf{y}\| > \varepsilon\}$. Take $0 < \varepsilon' < \varepsilon$, with ε' much smaller than ε. By symmetry we can write

$$\begin{aligned}\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{x})\, M_n^{(k)}(d\mathbf{y}) &= 2 \int_{B_\varepsilon} \int_{B_{\varepsilon'}^c} + 2 \int_{B_\varepsilon} \int_{B_{\varepsilon'} \cap B_\varepsilon^c} + \int_{B_\varepsilon} \int_{B_\varepsilon} + \int_{B_\varepsilon^c} \int_{B_\varepsilon^c} \\ &=: T_{\varepsilon,\varepsilon',n}^{(1)} + T_{\varepsilon,\varepsilon',n}^{(2)} + T_{\varepsilon,n}^{(3)} + T_{\varepsilon,n}^{(4)}.\end{aligned}$$

It will turn out that the asymptotic behaviour of $F_{kk}(n)$ will be determined by $T^{(1)}_{\varepsilon,\varepsilon',n}$. To treat $T^{(1)}_{\varepsilon,\varepsilon',n}$, we use a simple fact: if $\|\mathbf{b}\|/\|\mathbf{a}\| \le \delta \in (0,1)$, then

$$(1-\delta)^{2\theta}\|\mathbf{a}\|^{2\theta} \le \|\mathbf{a}+\mathbf{b}\|^{2\theta} \le (1+\delta)^{2\theta}\|\mathbf{a}\|^{2\theta}.$$

By taking $\delta = \varepsilon'/\varepsilon$, we obtain

(62)
$$u_n^\alpha n^{-1} T^{(1)}_{\varepsilon,\varepsilon',n} \in \left[\left(1-\varepsilon'/\varepsilon\right)^{2\theta}, \left(1+\varepsilon'/\varepsilon\right)^{2\theta}\right] u_n^\alpha n^{-1} 2\int_{B_\varepsilon} \|\mathbf{y}\|^{2\theta}\left(\int_{B^c_{\varepsilon'}} M_n^{(k)}(d\mathbf{x})\right) M_n^{(k)}(d\mathbf{y}).$$

Since $u_n^\alpha n^{-1} M_n^{(k)}(B^c_{\varepsilon'}) = u_n^\alpha n^{-1}(N_n^{(k)} - M_n^{(k)}(B_{\varepsilon'}))$ and $M_n^{(k)}(B_{\varepsilon'}) \Rightarrow M_\alpha^{(k)}(B_{\varepsilon'}) < \infty$ a.s., we conclude that

$$\begin{aligned} u_n^\alpha n^{-1} 2\int_{B_\varepsilon} \|\mathbf{y}\|^{2\theta}\left(\int_{B^c_{\varepsilon'}} M_n^{(k)}(d\mathbf{x})\right) M_n^{(k)}(d\mathbf{y}) &= 2\int_{B_\varepsilon} \|\mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{y})\, u_n^\alpha n^{-1} M_n^{(k)}(B^c_{\varepsilon'}) \\ &\Rightarrow 2c_\alpha w_k^\alpha \int_{\|\mathbf{y}\|>\varepsilon} \|\mathbf{y}\|^{2\theta} M_\alpha^{(k)}(d\mathbf{y}). \qquad (63) \end{aligned}$$

From the weak convergence in (31), it follows directly that

$$2\int_{B_\varepsilon}\int_{B_{\varepsilon'}\cap B^c_\varepsilon} \|\mathbf{x}-\mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{x})\, M_n^{(k)}(d\mathbf{y}) \Rightarrow 2\int_{\|\mathbf{y}\|>\varepsilon}\int_{\varepsilon'<\|\mathbf{x}\|\le\varepsilon} \|\mathbf{x}-\mathbf{y}\|^{2\theta} M_\alpha^{(k)}(d\mathbf{x})\, M_\alpha^{(k)}(d\mathbf{y})$$

and hence

$$u_n^\alpha n^{-1} T^{(2)}_{\varepsilon,\varepsilon',n} \xrightarrow{P} 0.$$

Combining this result with (63), we obtain,

$$u_n^\alpha n^{-1}(T^{(1)}_{\varepsilon,\varepsilon',n} + T^{(2)}_{\varepsilon,\varepsilon',n}) \Rightarrow 2c_\alpha w_k^\alpha \int_{\|\mathbf{y}\|>\varepsilon} \|\mathbf{y}\|^{2\theta} M_\alpha^{(k)}(d\mathbf{y}). \qquad (64)$$

Turning to $T^{(3)}_{\varepsilon,n}$, we have once again from the point process convergence in (31),

$$T^{(3)}_{\varepsilon,n} = \int_{B_\varepsilon}\left(\int_{B_\varepsilon} \|\mathbf{y}-\mathbf{x}\|^{2\theta} M_n^{(k)}(d\mathbf{x})\right) M_n^{(k)}(d\mathbf{y}) \Rightarrow \int_{B_\varepsilon}\left(\int_{B_\varepsilon} \|\mathbf{y}-\mathbf{x}\|^{2\theta} M_\alpha^{(k)}(d\mathbf{x})\right) M_\alpha^{(k)}(d\mathbf{y}),$$

from which we conclude,

$$u_n^\alpha n^{-1} T^{(3)}_{\varepsilon,n} \xrightarrow{P} 0. \qquad (65)$$

Finally we handle the last term $T^{(4)}_{\varepsilon,n}$. We have

$$u_n^\alpha n^{-1} T^{(4)}_{\varepsilon,n} \le 2\int_{B_\varepsilon} \|\mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{y})\left(u_n^\alpha n^{-1}\right) N_n^{(k)} \qquad (66)$$

and, since $u_n^{-\alpha} n^{-1} N_n^{(k)} \xrightarrow{P} c_\alpha w_k^\alpha$, we restrict attention to the integral. By (58) with $z = \epsilon u_n^{-1} n^{1/\alpha}$,

$$\begin{aligned} \mathbb{E}\left(\int_{B_\varepsilon} \|\mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{y})\right) &\sim w_k^\alpha c_\alpha u_n^{2\theta-\alpha} n^{1-2\theta/\alpha} \mathbb{E}_{k,n}\left(\|\mathbf{V}\|^{2\theta}\mathbf{1}(\|\mathbf{V}\| \le \epsilon u_n^{-1} n^{1/\alpha})\right) \\ &\le C\epsilon^{2\theta-\alpha}, \end{aligned}$$

and hence

$$\lim_{\epsilon\to 0}\limsup_{n\to\infty} \mathbb{E}\left(\int_{B_\varepsilon} \|\mathbf{y}\|^{2\theta} M_n^{(k)}(d\mathbf{y})\right) = 0.$$

This, in turn, implies from (66) that for any $\eta > 0$,

$$\lim_{\epsilon\to 0}\limsup_{n\to\infty}\mathbb{P}(u_n^{\alpha}n^{-1}T_{\varepsilon,n}^{(4)} > \eta) = 0. \tag{67}$$

Combining (63), (64), (65), and (67), we obtain

$$u_n^{4\theta-\alpha}n^{1-2\theta/\alpha}F_{k,k}(n) \Rightarrow \frac{2d_\theta^2 w_k^\alpha}{w^2 c_\alpha}\int_{\|\mathbf{y}\|>0}\|\mathbf{y}\|^{2\theta}M_\alpha^{(k)}(d\mathbf{y})\,. \tag{68}$$

The “off-diagonal terms” can be treated in a similar fashion. It follows from the continuous differentiability of R outside the origin and Theorem 1 that, in the obvious notation,

(69)

$$\begin{aligned}F_{k_1,k_2}(n) &= \frac{1}{N_n^2}\sum_{i=1}^{N_n^{(k_1)}}\sum_{j=1}^{N_n^{(k_2)}}\Big[R\big(\mathbf{Y}_i^{(k_1)}-\mathbf{Y}_j^{(k_2)}\big)-R\big(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}\big)\Big]^2\\ &= \frac{1}{N_n^2}\sum_{i=1}^{N_n^{(k_1)}}\sum_{j=1}^{N_n^{(k_2)}}\Big[\big\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),(\mathbf{Y}_i^{(k_1)}-\mathbf{s}_{k_1})-(\mathbf{Y}_j^{(k_2)}-\mathbf{s}_{k_2})\big\rangle\Big]^2+o_P(1)\,.\end{aligned}$$

Arguing as before (see (61)),

$$u_n^{4-\alpha}n^{1-2/\alpha}F_{k_1,k_2}(n) \sim \frac{1}{w^2c_\alpha^2}u_n^\alpha n^{-1}\int_{\mathbb{R}^d\times\mathbb{R}^d}[\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),\mathbf{x}-\mathbf{y}\rangle]^2\,M_n^{(k_1)}(d\mathbf{x})\,M_n^{(k_2)}(d\mathbf{y})\,.$$

We the same notation, as above we write

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d} &= \int_{B_\varepsilon}\int_{B_{\varepsilon'}^c}+\int_{B_\varepsilon}\int_{B_{\varepsilon'}\cap B_\varepsilon^c}+\int_{B_{\varepsilon'}^c}\int_{B_\varepsilon}+\int_{B_{\varepsilon'}\cap B_\varepsilon^c}\int_{B_\varepsilon}+\int_{B_\varepsilon}\int_{B_\varepsilon}+\int_{B_\varepsilon^c}\int_{B_\varepsilon^c}\\ &=: T_{\varepsilon,\varepsilon',n}^{(1,1)}+T_{\varepsilon,\varepsilon',n}^{(1,2)}+T_{\varepsilon,\varepsilon',n}^{(2,1)}+T_{\varepsilon,\varepsilon',n}^{(2,2)}+T_{\varepsilon,n}^{(3)}+T_{\varepsilon,n}^{(4)}\,.\end{aligned}$$

These six terms are treated in nearly the same way as earlier for the analogous decomposition for $F_{kk}(n)$. In this vein, we have for the first term

(70)

$$u_n^\alpha n^{-1}T_{\varepsilon,\varepsilon',n}^{(1,1)} \in \Big[\big(1-\varepsilon'/\varepsilon\big)^2,\big(1+\varepsilon'/\varepsilon\big)^2\Big]u_n^\alpha n^{-1}\int_{B_\varepsilon}\Big[\big\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),\mathbf{x}\big\rangle\Big]^2 M_n^{(k_1)}(d\mathbf{x})M_n^{(k_2)}(B_{\varepsilon'}^c)\,.$$

Since $u_n^\alpha n^{-1}M_n^{(k_2)}(B_{\varepsilon'}^c) \sim u_n^\alpha n^{-1}N_n^{(k_2)} \xrightarrow{P} c_\alpha w_{k_2}^\alpha$, we have by the weak convergence in (31)

$$\begin{aligned}&u_n^\alpha n^{-1}\int_{B_\varepsilon}[\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),\mathbf{x}\rangle]^2\,M_n^{(k_1)}(d\mathbf{x})M_n^{(k_2)}(B_{\varepsilon'}^c)\\ &\Rightarrow c_\alpha w_{k_2}^\alpha\int_{B_\varepsilon}[\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),\mathbf{x}\rangle]^2\,M_\alpha^{(k_1)}(d\mathbf{x})\,.\end{aligned} \tag{71}$$

Next, a straightforward application of the weak convergence in (31), gives

$$\begin{aligned}&\int_{B_\varepsilon}\left(\int_{B_{\varepsilon'}\cap B_\varepsilon^c}[\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),\mathbf{x}-\mathbf{y}\rangle]^2\,M_n^{(k_2)}(d\mathbf{y})\right)M_n^{(k_1)}(d\mathbf{x})\\ &\Rightarrow\int_{\|\mathbf{x}\|>\varepsilon}\left(\int_{\varepsilon'<\|\mathbf{y}\|\le\varepsilon}[\langle\nabla R(\mathbf{s}_{k_1}-\mathbf{s}_{k_2}),\mathbf{x}-\mathbf{y}\rangle]^2\,M_\alpha^{(k_2)}(d\mathbf{y})\right)M_\alpha^{(k_1)}(d\mathbf{x})\end{aligned}$$

and hence

$$u_n^\alpha n^{-1}T_{\varepsilon,\varepsilon',n}^{(1,2)} \xrightarrow{P} 0\,. \tag{72}$$

Interchanging the roles of $\mathbf{x}$ and $\mathbf{y}$, the terms $T_{\varepsilon,\varepsilon',n}^{(1,2)}$ and $T_{\varepsilon,\varepsilon',n}^{(2,2)}$ are handled in exactly the same way. The same reasoning (see (65)) also shows that

$$u_n^{-\alpha}n^{-1}T_{\varepsilon,n}^{(3)} \xrightarrow{P} 0. \tag{73}$$

Finally, turning to $T^{(4)}_{\varepsilon,n}$, we have by the Cauchy-Schwarz inequality,

$$u_n^\alpha n^{-1} T^{(4)}_{\varepsilon,n} \le C\left(\int_{B_\varepsilon} \|\mathbf{y}\|^2 M_n^{(k_1)}(d\mathbf{y}) \left(u_n^\alpha n^{-1}\right) N_n^{(k_2)} + \int_{B_\varepsilon} \|\mathbf{y}\|^2 M_n^{(k_2)}(d\mathbf{y}) \left(u_n^\alpha n^{-1}\right) N_n^{(k_1)} \right)$$

for some constant $C > 0$. But now the exact same argument leading to (67) (with $\theta = 1$) can be applied to each of the summands from which it follows that for any $\eta > 0$

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \mathbb{P}(u_n^\alpha n^{-1} T^{(4)}_{\varepsilon,n} > \eta) = 0\,. \tag{74}$$

Combining the results in (70), (71), (72), (73), and (74), we conclude that

$$\left(u_n^{4-\alpha} n^{1-2/\alpha}\right) F_{k_1,k_2}(n) \Rightarrow \frac{1}{w^2 c_\alpha} \int_{\mathbb{R}^d} \Big[\langle \nabla R(\mathbf{s}_{k_1} - \mathbf{s}_{k_2}), \mathbf{x}\rangle\Big]^2 \left(w_{k_2}^2 M_\alpha^{(k_1)} + w_{k_1}^2 M_\alpha^{(k_2)}\right)(d\mathbf{x}). \tag{75}$$

By Theorem 3 the convergence in (75) and (68) is joint in k_1, k_2, and the claim of the theorem follows. Note that if $\theta = 1$, then the scaling for $F_{k_1,k_2}(n)$ is the same as that for F_{kk}. On the other hand if $\theta > 1$, then $\left(u_n^{4-\alpha} n^{1-2/\alpha}\right) F_{k_1,k_2}(n) \overset{P}{\to} 0$ so that the diagonal terms are of smaller order. □

Proof of Theorem 5. In this case the order of magnitude of the "diagonal terms" is still given by (44) while the order of magnitude of the "off-diagonal terms" is given by (49), because the latter statement requires only that $\alpha > 2$. We claim that

$$u_n^{-4\theta+\alpha} n^{2\theta/\alpha - 1} \ll u_n^{-2}. \tag{76}$$

Indeed, (76) is equivalent to

$$n \ll u_n^{(4\theta-\alpha-2)/(2\theta/\alpha-1)},$$

which is a true statement due to (30) and the fact that

$$\frac{4\theta - \alpha - 2}{2\theta/\alpha - 1} > 2\alpha$$

as implied by the condition $2 < \alpha < 2\theta$. It follows from (76) that the "off-diagonal terms" are of a larger order of magnitude than the "diagonal terms". □

REFERENCES

[1] Avella Medina, M., Davis, R. A. and Samorodnitsky, G. (2021). Spectral learning of multivariate extremes. *arXiv preprint arXiv:2111.07799*.

[2] Bakir, G. H., Weston, J. and Schölkopf, B. (2004). Learning to find pre-images. *Advances in Neural Information Processing Systems* **16** 449–456.

[3] Basrak, B., Davis, R. A. and Mikosch, T. (2002). A characterization of multivariate regular variation. *Annals of Applied Probability* **12** 908–920.

[4] Basrak, B., Davis, R. A. and Mikosch, T. (2002). Regular variation of GARCH processes. *Stochastic Processes and their Applications* **99** 95–115.

[5] Blanchard, G., Bousquet, O. and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning* **66** 259–294.

[6] Braun, M. L., Buhmann, J. M. and Müller, K.-R. (2008). On relevant dimensions in kernel feature spaces. *Journal of Machine Learning Research* **9** 1875–1908.

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

[8] Clémençon, S., Jalalzai, H., Sabourin, A. and Segers, J. (2021). Concentration bounds for the empirical angular measure with statistical learning applications. *arXiv preprint arXiv:2104.03966*.

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

[10] DEUBER, D., LI, J., ENGELKE, S. and MAATHUIS, M. H. (2021). Estimation and Inference of Extremal Quantile Treatment Effects for Heavy-Tailed Distributions. *arXiv preprint arXiv:2110.06627*.

[11] DREES, H. and SABOURIN, A. (2021). Principal component analysis for multivariate extremes. *Electronic Journal of Statistics* **15** 908–943.

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

[13] ENGELKE, S. and IVANOVS, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application* **8** 241–270.

[14] ENGELKE, S., LALANCETTE, M. and VOLGUSHEV, S. (2022). Learning extremal graphical models in high dimensions. *arXiv preprint arXiv:2111.00840*.

[15] FOMICHOV, V. and IVANOVS, J. (2022). Spherical clustering in detection of groups of concomitant extremes. *Biometrika*.

[16] GNECCO, N., MEINSHAUSEN, N., PETERS, J. and ENGELKE, S. (2021). Causal discovery in heavy-tailed models. *Annals of Statistics* **49** 1755–1778.

[17] GOIX, N., SABOURIN, A. and CLÉMEN, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory* 843–860. PMLR.

[18] GOIX, N., SABOURIN, A. and CLÉMENÇON, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis* **161** 12–31.

[19] HONEINE, P. and RICHARD, C. (2011). Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine* **28** 77–88.

[20] JALALZAI, H. and LELUC, R. (2021). Feature Clustering for Support Identification in Extreme Regions. In *International Conference on Machine Learning* 4733–4743. PMLR.

[21] JANSSEN, A. and WAN, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics* **14** 1211–1233.

[22] KIM, K. I., JUNG, K. and KIM, H. J. (2002). Face recognition using kernel principal component analysis. *IEEE signal processing letters* **9** 40–42.

[23] KWOK, J. T. Y. and TSANG, I. W. H. (2004). The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks* **15** 1517–1525.

[24] MEYER, N. and WINTENBERGER, O. (2019). Sparse regular variation. *arXiv preprint arXiv:1907.00686*.

[25] MIKA, S., SCHÖLKOPF, B., SMOLA, A., MÜLLER, K.-R., SCHOLZ, M. and RÄTSCH, G. (1998). Kernel PCA and de-noising in feature spaces. *Advances in Neural Information Processing Systems* **11**.

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

[27] RESNICK, S. I. (2008). *Extreme values, regular variation, and point processes* **4**. Springer Science & Business Media.

[28] ROHRBECK, C. and COOLEY, D. (2022). Simulating flood event sets using extremal principal components. *Annals of Applied Statistics (to appear)*.

[29] ROSIPAL, R., GIROLAMI, M., TREJO, L. J. and CICHOCKI, A. (2001). Kernel PCA for feature extraction and de-noising in nonlinear regression. *Neural Computing & Applications* **10** 231–243.

[30] SCHÖLKOPF, B., SMOLA, A. and MÜLLER, K.-R. (1997). Kernel principal component analysis. In *International Conference on Artificial Neural Networks* 583–588. Springer.

[31] SHAWE-TAYLOR, J., WILLIAMS, C. K., CRISTIANINI, N. and KANDOLA, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Transactions on Information Theory* **51** 2510–2522.

[32] SIMPSON, E. S., WADSWORTH, J. L. and TAWN, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika* **107** 513–532.

[33] SMOLA, A. J. and SCHÖLKOPF, B. (1998). *Learning with Kernels* **4**. Citeseer.

[34] STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

[35] TYLER, D. E. (1987). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika* **74** 579–589.

[36] VAN ZANTEN, J. H. and VAN DER VAART, A. W. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh* 200–222. Institute of Mathematical Statistics.

[37] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.

[38] ZHENG, W.-S., LAI, J. and YUEN, P. C. (2010). Penalized preimage learning in kernel principal component analysis. *IEEE Transactions on Neural Networks* **21** 551–570.

[39] ZWALD, L. and BLANCHARD, G. (2005). On the convergence of eigenspaces in kernel principal component analysis. *Advances in Neural Information Processing Systems* **18**.