Electronic Journal of Statistics
Vol. 15 (2021) 908–943
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1803

Principal component analysis for multivariate extremes

Holger Drees[1] **and Anne Sabourin**[2]

[1] *University of Hamburg, Department of Mathematics, Germany*
e-mail: holger.drees@uni-hamburg.de

[2] *LTCI, Télécom Paris, Institut polytechnique de Paris, France*
e-mail: anne.sabourin@telecom-paris.fr

Abstract: In the probabilistic framework of multivariate regular variation, the first order behavior of heavy-tailed random vectors above large radial thresholds is ruled by a homogeneous limit measure. For a high dimensional vector, a reasonable assumption is that the support of this measure is concentrated on a lower dimensional subspace, meaning that certain linear combinations of the components are much likelier to be large than others. Identifying this subspace and thus reducing the dimension will facilitate a refined statistical analysis. In this work we apply Principal Component Analysis (PCA) to a re-scaled version of radially thresholded observations.

Within the statistical learning framework of empirical risk minimization, our main focus is to analyze the squared reconstruction error for the exceedances over large radial thresholds. We prove that the empirical risk converges to the true risk, uniformly over all projection subspaces. As a consequence, the best projection subspace is shown to converge in probability to the optimal one, in terms of the Hausdorff distance between their intersections with the unit sphere. In addition, if the exceedances are rescaled to the unit ball, we obtain finite sample uniform guarantees to the reconstruction error pertaining to the estimated projection subspace. Numerical experiments illustrate the capability of the proposed framework to improve estimators of extreme value parameters.

MSC2020 subject classifications: Primary 62G32; secondary 62H25.
Keywords and phrases: Dimension reduction, empirical risk minimization, multivariate extreme value analysis, multivariate regular variation, Principal Component Analysis.

Received October 2019.

1. Introduction

If one wants to analyze the tail behavior of an $\mathbb{R}^d$-valued random vector $X = (X^1, \ldots, X^d)$ one usually assumes that X is regularly varying (if necessary after a standardization of the marginal distributions), *i.e.* there exists a non-zero measure μ on $\mathbb{R}^d \setminus \{0\}$ such that

$$\mu_t(B) := \frac{\mathbb{P}(X \in tB)}{\mathbb{P}(\|X\| > t)} \xrightarrow[t\to\infty]{} \mu(B) < \infty \tag{1.1}$$

for all μ-continuous Borel sets B that are bounded away from the origin. This definition does not depend on the choice of the norm $\|\cdot\|$ on $\mathbb{R}^d$, but in what fol-

lows we only consider the Euclidean norm. Convergence (1.1) may be understood as a generalization to arbitrary dimension of a heavy-tail assumption regarding a real-valued random variable. This mathematical framework is particularly useful in situations where the focus is on 'tail events' of the kind $\{X \in C\}$ where the distance to the origin $u = \inf\{\|x\| : x \in C\}$ is large, for some norm $\|\cdot\|$. In a risk management context, the probability of such tail events is of crucial importance. If the distance u is so large that few or no data are available in the considered region C, all attempts to resort to empirical estimation are in vain. One common idea behind statistical methods based on Extreme Value Theory (EVT) is to use a small proportion of the available data (those with a comparatively large norm) to learn an estimate for μ, which may be used for quantifying the probability of tail events.

1.1. Regular variation

A standard reference concerning the probabilistic aspects of regular variation in the setting of EVT is [28], see also [27] for application-oriented examples. Regular variation for Borel measures on Polish spaces has since been revisited in [19] and [23]. It is well known that if convergence (1.1) holds true, then the limit measure μ is homogeneous of order $-\alpha$ for some $\alpha > 0$. Moreover, the norm $\|X\|$ is regularly varying, too: $\mathbb{P}\{\|X\| > tx\}/\mathbb{P}\{\|X\| > t\} \to x^{-\alpha}$ as $t \to \infty$ for all $x > 0$.

Because the limit measure is homogeneous, after a polar transformation, it can be decomposed into a so-called spectral (or angular) probability measure H and an independent radial component, that is

$$\mu\Big\{x \in \mathbb{R}^d \ : \ \|x\| > r, \frac{x}{\|x\|} \in A\Big\} = r^{-\alpha} H(A), \tag{1.2}$$

for all $r > 0$ and all Borel subsets A of the unit sphere. Whereas the literature on the design and the asymptotics of flexible multivariate parametric or nonparametric models for μ or integrated versions of it is plentiful (see *e.g.* [32, 13, 9, 15, 29], or [2] and the references therein), the issue of how to escape the curse of dimensionality has only recently been raised (see below). One reason for this may be that a major field of application of EVT concerns environmental, spatial extremes such as heavy rainfalls, heat waves, droughts or floods. In this context, max-stable or generalized Pareto spatial models are widely used ([25, 12, 30]) which have built in a priori information about the spatial dependence structure, thus reducing the effective dimension.

1.2. Dimensionality reduction for extreme values, a brief overview

For applications such as *e.g.* anomaly detection or network monitoring where no particular structure is known a priori, dimension reduction suggests itself as a preliminary step before implementing any kind of statistical procedure. This subject has recently received increasing attention. If d is moderate or large, the measure μ (and hence H) will often exhibit some 'sparse' structure. For example, if some of the components of X are asymptotically independent, *i.e.*

for some index set $I \subset \{1,\dots,d\}$ of size $|I| \in \{1,\dots,d-1\}$

$$\mathbb{P}\Big\{ \max_{i\in I} |X^i| > t, \max_{i\notin I} |X^i| > t \Big\} = o\Big(\mathbb{P}\Big\{ \max_{1\le i\le d} |X^i| > t \Big\}\Big),$$

then μ is concentrated on $\{x = (x^1,\dots,x^d) : \max_{i\in I} |x^i| = 0$ or $\max_{i\notin I} |x^i| = 0\}$. More generally, one may consider the case where only a small number of subsets of components $\{I_k \subset \{1,\dots,d\}, k = 1,\dots,K\}$ are likely to be large simultaneously, while the other components remain small. Here, 'small number' is understood relatively to the $2^d - 1$ non empty possible subsets of components. This setting applies *e.g.* to heavy rainfalls in a spatial setting (storms are usually localized, so that neighboring sites are more likely to be jointly impacted) or of shocks to different assets of a financial portfolio. [5] proposes a clustering approach combined with spherical data analysis to detect structures of this type. [16, 17] propose an algorithm with moderate computational cost (linear in the dimension and the sample size) and finite sample uniform guarantees. Their error bounds are linear in d and scale as $1/\sqrt{k}$, where k is the number of order statistics of each component that are considered extreme during the training step. A refinement of the latter framework is proposed in the yet unpublished work of [34]. [6] and [7] aim at identifying subgroups of components for which the probability of a joint excess over a large quantile is not negligible compared to that of an excess by a single component. [10] use graphical models to reduce the complexity of the extremal dependence structure. In a regression context, [14] sets up a mathematical framework for tail dimension reduction suited to the case where the distribution of the target variable above high thresholds only depends on the projection of the covariates on a lower dimensional subspace. Consistency of k-means clustering applied to the most extreme observations of a data set has recently been proven in [20]. An overview of different concepts of sparsity in EVT can be found in [11].

1.3. Principal component analysis (PCA) and support identification

Here we focus on finding a linear subspace on which μ is (nearly) concentrated. In a classical setting, when $\|X\|$ has finite second moments, PCA ([1]) is the method of choice to determine such supporting linear subspaces if *i.i.d.* random vectors X_i, $1 \le i \le n$, with the same distribution as X are observed. Theoretical guarantees obtained so far concern the reconstruction error ([21, 33, 4, 22, 26]) or the approximation error for the eigenspaces of the covariance matrix ([37]), under the assumption that the sample space (or the feature space for Kernel-PCA) has finite diameter or that sufficiently high order moments exist.

For motivation of our version of PCA, it is useful to keep the following working hypothesis in mind, although it is not required for most results to hold.

Hypothesis 1. *The vector space* $V_0 = \operatorname{span}(\operatorname{supp}\mu)$ *generated by the support of* μ *has dimension* $p < d$.

Note that then the points $(X_i/t)\mathbf{1}\{\|X_i\| > t\}$ are more and more concentrated on a neighborhood of V_0 as t increases, but usually they will not lie on V_0. If

the dimension p of V_0 is known, then it suggests itself to approximate V_0 by the subspace of dimension p that is 'closest' in expectation to these points.

In PCA one measures the closeness by the squared Euclidean distance which hugely alleviates the optimization problem as one may work with orthogonal projections in the Hilbert space L_2. However, this approach requires finite second moments which cannot be taken for granted in the above setting. Indeed, if $\alpha < 2$ then $\mathbb{E}(\|X_i\|^2) = \infty$. Hence, we will instead consider re-scaled vectors

$$\Theta_i := \omega(X_i)X_i, \quad 1 \le i \le n, \tag{1.3}$$

where $\omega : \mathbb{R}^d \to (0, \infty)$ is a suitable scaling function. The most common choice is $\omega(x) = 1/\|x\|$, leading to Θ_i on the unit sphere which describes the direction of X_i, and we will focus on this re-scaling when we derive finite sample bounds on the reconstruction error (see Section 3). However, consistency results will be proved for considerably more general scaling functions; cf. Section 2.

To the best of our knowledge, the only existing work considering PCA properly speaking for high dimensional extremes is the paper [8]. The authors discuss a transformation mapping negative observations to small positive ones and apply PCA in this transformed space. They also use a preliminary re-scaling involving the norm of the transformed vector. They illustrate their approach with simulations and real data examples, without deriving theoretical statistical guarantees.

1.4. Notation and risk minimization setting

To give a formal description of our method, we first introduce some notation. All random variables are defined on some probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$; the expectation with respect to $\mathbb{P}$ is denoted by $\mathbb{E}$. For $x \in \mathbb{R}^d$ and $t > 0$, let

$$\begin{aligned} \theta(x) &= \omega(x)x, \\ \theta_t(x) &= \omega(x)x\mathbf{1}\{\|x\| > t\}, \\ \Theta &= \theta(X) = \omega(X)X, \\ \Theta_t &= \theta_t(X) = \Theta\mathbf{1}\{\|X\| > t\}. \end{aligned} \tag{1.4}$$

By P we denote the distribution of X and by P_t its conditional distribution given that $\|X\| > t$, *i.e.* $P_t(\cdot) = \mathbb{P}(X \in \cdot \mid \|X\| > t\}$. Then $P_\infty := \mu|_{(B_1(0))^c}$ is the weak limit of $P_t(t\cdot)$ (with $B_1(0)$ denoting the closed unit ball); cf. (1.1).

For any probability measure Q and any Q-integrable function f, we denote the expectation of f with respect to Q by Qf or $Q(f)$. By $\mathbb{E}_t$ we denote the conditional expectation (with respect to $\mathbb{P}$) given $\|X\| > t$ so that $\mathbb{E}_t(f(X)) = P_t(f)$, provided the expectations exist.

For any linear subspace $V \subset \mathbb{R}^d$, let $\mathbf{\Pi}_V$ be the orthogonal projection onto V (or the associated projection matrix), and let $\mathbf{\Pi}_V^\perp$ be the orthogonal projection onto the orthogonal complement $V^\perp$ of V.

To apply PCA to the re-scaled vectors, we have to assume that the scaling function ω is chosen such that $\mathbb{E}(\|\Theta\|^2) = P(\|\theta\|^2) < \infty$ and $P_\infty(\|\theta\|^2) < \infty$. Note that this condition is always fulfilled if there exist $\beta > 1 - \alpha/2$ and $c > 0$

such that $\omega(x) \le c\|x\|^{-\beta}$ for all $x \in \mathbb{R}^d$. For simplicity's sake, in what follows we will impose the following stronger homogeneity condition:

$$\exists\, \beta \in \Big(1 - \frac{\alpha}{2}, 1\Big] \;\forall\, \lambda > 0, x \in \mathbb{R}^d : \quad \omega(\lambda x) = \lambda^{-\beta}\omega(x) \\ \text{and} \quad c_\omega := \sup_{x \in \mathbb{S}^{d-1}} \omega(x) < \infty, \tag{1.5}$$

where $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ denotes the unit sphere. Note that then

$$\|\theta(x)\| \le c_\omega \|x\|^{1-\beta}. \tag{1.6}$$

The choice $\omega(x) = \|x\|^{-\beta}$ seems natural, but different choices allow for focusing on particular aspects of the extreme value behavior. For instance, if one is only interested in the positive components of X, one may choose $\omega(x) = \|x\|^{-\beta}\mathbf{1}_{[0,\infty)^d}(x)$.

Hypothesis 1 is equivalent to the statement that $\inf_{V:\dim(V)=p} R_\infty(V) = 0$ and $\inf_{V:\dim(V)=p'} R_\infty(V) > 0$ for all $p' < p$ where

$$R_\infty(V) := P_\infty \|\mathbf{\Pi}_V \theta - \theta\|^2 = P_\infty \|\mathbf{\Pi}_V^\perp \theta\|^2$$

and the infima are taken over all linear subspaces of the specified dimension (cf. Lemma 2.5). The risk R_∞ may be interpreted as the expected reconstruction error in the limit model if the re-scaled observation Θ is replaced with its lower dimensional approximation $\mathbf{\Pi}_V \Theta$. Since $P_t(t\cdot) \to P_\infty(\cdot)$ weakly, one may approximate V_0 by a subspace $V_t^* = V_t^{p*}$ of dimension p that minimizes the conditional risk

$$R_t(V) := P_t\big(\|\mathbf{\Pi}_V^\perp \theta)\|^2\big) = \mathbb{E}_t\left(\|\mathbf{\Pi}_V^\perp \Theta\|^2\right) \tag{1.7}$$

given that $\|X\|$ exceeds a high threshold $t > 0$. Note that V_t^* may be of interest even if Hypothesis 1 only holds approximately, in the sense that P_∞ concentrates most of its mass on a small neighborhood of a p-dimensional subspace.

It is natural to 'estimate' V_t^* (and thus V_0) by a minimizer of the corresponding empirical risk

$$\hat{R}_t(V) := \frac{1}{N_t} \sum_{i=1}^n \|\mathbf{\Pi}_V^\perp \Theta_i\|^2 \mathbf{1}\{\|X_i\| > t\} \quad \text{with} \quad N_t := \sum_{i=1}^n \mathbf{1}\{\|X_i\| > t\}.$$

Here the threshold t must be chosen suitably, depending on the sample size. To this end, often order statistics of the norms of the observed vectors are used, and we follow this approach. Let $X_{(j)} = X_{\sigma(j)}$ where σ is a permutation of indices such that $\|X_{(1)}\| \ge \|X_{(2)}\| \ge \cdots \ge \|X_{(n)}\|$. (For brevity, we suppress the dependence on n in our notation of order statistics.) For $1 \le k \le n$, denote by $\hat{t}_{n,k} = \|X_{(k+1)}\|$ the empirical quantile of level $1 - k/n$ for $\|X\|$. We define the empirical risk for the subspace V related to the k largest observations as

$$\hat{R}_{n,k}(V) := \hat{R}_{\hat{t}_{n,k}}(V) = \frac{1}{k} \sum_{i=1}^n \|\mathbf{\Pi}_V^\perp \Theta_{i,\hat{t}_{n,k}}\|^2 \tag{1.8}$$

where $\Theta_{i,t} = \theta_t(X_i)$ in accordance with the notation introduced in (1.4). Here and throughout the paper, we suppose that the *c.d.f.* of $\|X\|$ is continuous in the tail to avoid technicalities. Then we may assume w.l.o.g. that there are no ties and thus exactly k observations have norm larger than $\hat{t}_{n,k}$. A minimizer of $\hat{R}_{n,k}(V)$ among all linear subspaces of dimension p will be denoted by $\hat{V}_n = \hat{V}_n^p$. It is the main goal of the present paper to analyze the asymptotic and the finite sample behavior of the empirical risk $\hat{R}_{n,k}(V)$ and its minimizer $\hat{V}_n$.

1.5. Outline

In Section 2 we will first show that the minimizer of the risk R_t based on a finite threshold t converges to the minimizer of the limit risk R_∞, and thus under Hypothesis 1 to V_0, as $t \to \infty$. Moreover, we show consistency of the empirical risk minimizer $\hat{V}_n$ under condition (1.5). In Section 3, we derive non-asymptotic uniform bounds on the difference between the empirical risk and its theoretical counterpart for the most important scaling $\omega(x) = 1/\|x\|$. Furthermore, we construct uniform confidence bands for $R_t(V)$. The results obtained in a simulation study are reported in Section 4. In particular, we explore the choice of the dimension p based on empirical risk plots and the effect of a PCA projection on estimators of probabilities expressed in terms of the spectral measure H. All proofs and technical lemmas are postponed to Section 5, while an appendix contains some details about the proof of a modification of a result by [4].

2. Consistency of risk minimizers

In this section, we first discuss how to calculate minimizers of the conditional risk R_t given $\|X\| > t$ and the empirical risk $\hat{R}_{n,k}$. Moreover, we prove that these converge in some sense towards a minimizer of R_∞.

It is well known that a point of minimum of $V \mapsto \mathbb{E}\,\|\mathbf{\Pi}_V^\perp Y\|^2$ can be derived from the spectral analysis of the matrix of second (mixed) moments of Y:

Lemma 2.1. *(i) Let Y be an $\mathbb{R}^d$-valued random vector with $\mathbb{E}(\|Y\|^2) < \infty$ and $\Sigma := \mathbb{E}(YY^\top)$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ denote the eigenvalues of Σ with corresponding orthogonal eigenvectors $x_1, \ldots, x_d$. Then $V^* = \operatorname{span}(x_1, \ldots, x_p)$ minimizes $\mathbb{E}(\|\mathbf{\Pi}_V^\perp Y\|^2)$ among all linear subspaces V of dimension p. In the case $\lambda_p > \lambda_{p+1}$ it is the unique minimizer.*

(ii) If the scaling condition (1.5) holds and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ denote the eigenvalues of $\Sigma_t := \mathbb{E}_t(\Theta\Theta^\top)$ with corresponding orthogonal eigenvectors $x_1, \ldots, x_d$, then $V^* = \operatorname{span}(x_1, \ldots, x_p)$ minimizes $R_t(V)$ among all linear subspaces V of dimension p. It is the unique minimizer if $\lambda_p > \lambda_{p+1}$.

(iii) If the scaling condition (1.5) holds and $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,d} \geq 0$ denote the eigenvalues of $\Sigma_{n,k} := k^{-1} \sum_{i=1}^n (\Theta_{i,\hat{t}_{n,k}} \Theta_{i,\hat{t}_{n,k}}^\top)$ with corresponding orthogonal eigenvectors $x_{n,1}, \ldots, x_{n,d}$, then

$$\hat{V}_n = \operatorname{span}(x_{n,1}, \ldots, x_{n,p})$$

minimizes $\hat{R}_{n,k}(V)$ among all linear subspaces V of dimension p.

A proof of assertion (i) can *e.g.* be found in [31], Theorem 5.3, where also other optimality properties of the minimizers are given. Both the other results follow directly by an application of (i) with Y equal to Θ conditional on $\|X\| > t$, respectively a random variable according to the empirical distribution of those Θ_i for which $\|X_i\| > \hat{t}_{n,k}$. If $\lambda_p = \lambda_{p+1}$, then the minimizer is not unique. With $m = \min\{i \in \{1,\dots,p\} : \lambda_i = \lambda_p\}$ any minimizer V_t^* of R_t can be represented as $V_t^* = \mathrm{span}(x_1,\dots,x_{m-1},\tilde{x}_m,\dots,\tilde{x}_p)$ where $\tilde{x}_m,\dots,\tilde{x}_p$ are orthogonal eigenvectors to the eigenvalue λ_p and all these subspaces are minimizers. An analogous statement holds for the empirical risk.

2.1. Asymptotic behavior of the conditional risk and its minimizer

Here we discuss the relationship between R_t and R_∞ and their respective minimizers.

Proposition 2.2. *Suppose that ω fulfills condition (1.5). Then, for any subspace V of $\mathbb{R}^d$, the suitably standardized associated finite threshold risk converges:*

$$\lim_{t\to\infty} t^{2(\beta-1)} R_t(V) = R_\infty(V).$$

In view of Proposition 2.2, one may ask whether a minimizer of $\tilde{R}_t := t^{2(\beta-1)}R_t$ (which of course is also a minimizer of R_t) converges in some sense to a minimizer of R_∞. Denote by $\mathcal{V}_p$ the set of all subspaces of $\mathbb{R}^d$ of dimension p, endowed with the metric

$$\rho(V,W) := |||\mathbf{\Pi}_V - \mathbf{\Pi}_W||| = \left|\left|\left|\mathbf{\Pi}_V^\perp - \mathbf{\Pi}_W^\perp\right|\right|\right| := \sup_{x\in\mathbb{S}^{d-1}} \|\mathbf{\Pi}_V^\perp x - \mathbf{\Pi}_W^\perp x\|,$$

pertaining to the operator norm $|||\cdot|||$ of the projections.

Remark 2.3. Note that $\rho(V,W)$ also gives an upper bound on the Hausdorff distance between $V\cap\mathbb{S}^{d-1}$ and $W\cap\mathbb{S}^{d-1}$. To see this, let $x^* \in V\cap\mathbb{S}^{d-1}$ and $y^* \in W\cap\mathbb{S}^{d-1}$ be such that the Hausdorff distance equals $\inf_{y\in W\cap\mathbb{S}^{d-1}} \|x^* - y\| = \|x^* - y^*\|$. Then $y^* = \mathbf{\Pi}_W x^*/\|\mathbf{\Pi}_W x^*\|$, $\|x^* - \mathbf{\Pi}_W x^*\| \le \rho(V,W)$ and $\|\mathbf{\Pi}_W x^*\|^2 \ge 1 - (\rho(V,W))^2$. Hence

$$\begin{aligned}
\|x^* - y^*\|^2 &= \|x^* - \mathbf{\Pi}_W x^*\|^2 + \|\mathbf{\Pi}_W x^* - y^*\|^2 \\
&\le (\rho(V,W))^2 + (1 - \|\mathbf{\Pi}_W x^*\|)^2 \\
&\le (\rho(V,W))^2 + \left(1 - \sqrt{1-(\rho(V,W))^2}\right)^2 \\
&= 2\left(1 - \sqrt{1-(\rho(V,W))^2}\right).
\end{aligned}$$

□

It can be shown that $\mathcal{V}_p$ is compact w.r.t. ρ (see Lemma 5.2) and that the normalized conditional risk functions $\tilde{R}_t$ are uniformly Lipschitz continuous (Lemma 5.3), from which the convergence of the risk minimizers follows by standard arguments.

Theorem 2.4. *Suppose that ω satisfies condition (1.5) and that R_∞ has a unique minimizer V_∞^* in $\mathcal{V}_p$. Then, for any minimizer V_t^* of R_t in $\mathcal{V}_p$, one has*

$$\lim_{t\to\infty} \rho(V_t^*, V_\infty^*) = 0.$$

Under Hypothesis 1, V_0 is the unique minimizer of R_∞ over $\mathcal{V}_p$, that is if we minimize the risk over linear subspaces with the correct dimension, as the following result shows. Hence in this case, V_t^* converges to V_0.

Lemma 2.5. *Under Hypothesis 1, for any subspace $V \subset \mathbb{R}^d$ of arbitrary dimension one has*

$$R_\infty(V) = 0 \iff V_0 \subset V.$$

Thus, V_0 is the unique minimizer of R_∞ in $\mathcal{V}_p$, whereas on $\mathcal{V}_{\tilde{p}}$ with $\tilde{p} > p$ the points of minimum of the limit risk R_∞ are not unique.

Proof. If $V_0 \subset V$ then $V^\perp \subset V_0^\perp$. By Hypothesis 1, P_∞ is concentrated on V_0, which implies $R_\infty(V) = P_\infty \|\mathbf{\Pi}_V^\perp \theta\|^2 \le P_\infty \|\mathbf{\Pi}_{V_0}^\perp \theta\|^2 = 0$.

Conversely, if $R_\infty(V) = 0$, then $1 = P_\infty\{\mathbf{\Pi}_V^\perp \theta = 0\} = P_\infty(V)$. By definition of P_∞ and the homogeneity of μ, this means that the support of μ must be a subset of V and thus $V_0 \subset V$. □

2.2. Convergence of the empirical risk and its minimizer

We now establish analogous consistency results for the empirical risk $\hat{R}_{n,k}$ and its minimizer. In what follows, let $F_{\|X\|}$ be the *c.d.f.* of $\|X\|$, $F_{\|X\|}^{\leftarrow}$ its generalized inverse (quantile function) and define

$$t_{n,k} := F_{\|X\|}^{\leftarrow}(1 - k/n). \tag{2.1}$$

We start with consistency of the standardized empirical risk.

Proposition 2.6. *If ω satisfies condition (1.5), then $t_{n,k}^{2(\beta-1)} \hat{R}_{n,k}(V) \to R_\infty(V)$ in probability for all linear subspaces V of $\mathbb{R}^d$.*

The following main result of this section states the consistency of the empirical risk minimizer.

Theorem 2.7. *If ω satisfies condition (1.5) and R_∞ has a unique minimizer V_∞^* in $\mathcal{V}_p$, then $\rho(\hat{V}_n, V_\infty^*) \to 0$ in probability for all minimizers $\hat{V}_n$ of $\hat{R}_{n,k}$ in $\mathcal{V}_p$.*

So far, we have proved weak consistency of both the standardized empirical risk and the empirical risk minimizer under mild assumptions on the scaling function ω. However, the rates of convergence may be arbitrarily slow. In the next section, we establish bounds on the recovery risk under stronger conditions.

3. Uniform risk bounds

Since a minimizer $\hat{V}_t$ of the empirical risk $\hat{R}_t$ (or $\hat{V}_n$ of $\hat{R}_{n,k}$) differs from the minimizer of the true risk R_t, usually the so-called excess risk $R_t(\hat{V}_t) - \inf_{V \in \mathcal{V}_p} R_t(V)$ will be strictly positive. We follow the common approach in the theory of risk minimization to bound the excess risk by deriving uniform bounds on $|\hat{R}_t(V) - R_t(V)|$ which hold with high probability for a fixed sample size n. If these uniform bounds can be calculated from the observed data, they may also be used to construct confidence intervals for the reconstruction error $R_t(\hat{V}_t)$ resp. $R_{t_{n,k}}(\hat{V}_n)$.

As condition (1.5) does not guarantee any finite moments of Θ of order greater than 2, and tight concentration inequalities are available only for subgaussian distributions, we now assume that the scaling function ω satisfies the following stronger condition:

$$\omega(x) \le \frac{1}{\|x\|}, \quad \forall\, x \in \mathbb{R}^d, \tag{3.1}$$

so that $\|\theta(x)\| \le 1$ for all $x \in \mathbb{R}^d$.

For classical PCA (and a kernel version thereof), [33] established uniform risk bounds for bounded random vectors Z_i, which were improved by the following result by [4]. Assume $\|Z_i\| \le 1$, and denote the empirical matrix of second (mixed) moments by $\hat{\Sigma}_n$ and the Hilbert-Schmidt norm on the space of matrices by $\|\cdot\|_{HS}$. Then, with probability greater than $1-\delta$,

$$\Big|\frac{1}{n}\sum_{i=1}^n \|\mathbf{\Pi}_V^\perp Z_i\|^2 - \mathbb{E}\,\|\mathbf{\Pi}_V^\perp Z\|^2\Big| \le \Big[\frac{p}{n-1}\Big(\frac{1}{n}\sum_{i=1}^n \|Z_i\|^4 - \|\hat{\Sigma}_n\|_{HS}\Big)\Big]^{1/2} + \Big(\frac{\log(3/\delta)}{2n}\Big)^{1/2} + \Big(\frac{p^2\log(3/\delta)}{n^3}\Big)^{1/4}$$

for all $V \in \mathcal{V}_p$. One may try to derive uniform risk bounds in our extreme value setting by applying this result to the random variables $Z_i = \Theta_{i,t} = \Theta_i \mathbf{1}\{\|X_i\| > t\}$, so that the left hand side is approximately equal to $\pi_t|\hat{R}_t(V) - R_t(V)|$ with

$$\pi_t := \mathbb{P}\{\|X\| > t\}$$

if one ignores the difference between N_t and its expectation $n\pi_t$. In the case $\pi_t = o(n^{-1/2})$, however, the above upper bound will not even converge to 0 when it is divided by π_t because of the second term. Hence this direct approach does not give meaningful bounds for $|\hat{R}_t(V) - R_t(V)|$.

The reason for this inconsistency is that, unlike in the classical setting, most of the random variables Z_i will vanish as t increases, and the concentration inequalities used in the proofs of the aforementioned bounds are too crude in such a situation. However, we will take up ideas used by [4], with appropriate modifications, to derive much tighter uniform bounds on $|\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)|$. Furthermore, we will derive uniform bounds on $|\hat{R}_t(V) - R_t(V)|$ which hold conditionally on $N_t = \ell$ and depend only on the data. These can then be used to construct confidence bands for $R_t(V)$.

If, for the time being, one neglects the difference between the empirical $(1 - k/n)$-quantile of $\|X\|$ (*i.e.* $\hat{t}_{n,k}$) and the true quantile $t_{n,k}$, then $\hat{R}_{n,k}(V)$ can be approximated by $\bar{R}_{t_{n,k}}(V)$ where

$$\bar{R}_t(V) := \frac{1}{n\pi_t} \sum_{i=1}^{n} \|\mathbf{\Pi}_V^{\perp} \Theta_{i,t}\|^2. \tag{3.2}$$

Denote the empirical distribution of the observed random vectors X_i, $1 \le i \le n$, by P_n, and recall the notation P for the distribution of X. For any threshold $t > 0$, the maximal difference between the approximate empirical risk $\bar{R}_t(V)$ and the true risk $R_t(V)$ can be rewritten as

$$\begin{aligned}
\sup_{V \in \mathcal{V}_p} |\bar{R}_t(V) - R_t(V)| &= \sup_{V \in \mathcal{V}_p} \frac{1}{\pi_t} \Big| \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{\Pi}_V^{\perp} \Theta_{i,t}\|^2 - \mathbb{E}\, \|\mathbf{\Pi}_V^{\perp} \Theta_t\|^2 \Big| \\
&= \frac{1}{\pi_t} \sup_{V \in \mathcal{V}_p} \big| (P_n - P) \|\mathbf{\Pi}_V^{\perp} \theta_t\|^2 \big| \\
&= \frac{1}{\pi_t} \max(\varphi_t^+(X_1, \ldots, X_n), \varphi_t^-(X_1, \ldots, X_n))
\end{aligned} \tag{3.3}$$

with

$$\varphi_t^{\pm}(x_1, \ldots, x_n) := \sup_{V \in \mathcal{V}_p} \pm \Big(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{\Pi}_V^{\perp} \theta_t(x_i)\|^2 - P \|\mathbf{\Pi}_V^{\perp} \theta_t\|^2 \Big). \tag{3.4}$$

To derive uniform bounds on on $|\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)|$, we thus first analyze the error of the approximation of $\hat{R}_{n,k}$ by $\bar{R}_{t_{n,k}}$ (Lemma 5.5) and then bound $\varphi^{\pm}(X_1, \ldots, X_n)$. For the latter step, we employ a version of the bounded difference inequality to conclude a concentration inequality for $\varphi^{\pm}(X_1, \ldots, X_n)$ (Lemma 5.7) and combine this with an upper bound on its expectation (Lemma 5.8). This approach leads to our first uniform bound on the difference between the empirical risk and its theoretical counterpart.

Theorem 3.1. *If* (3.1) *holds, then for all* $u, v > 0$

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{V \in \mathcal{V}_p} |\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)| \ge \Big[\frac{p \wedge (d-p)}{k} S_{t_{n,k}} \Big]^{1/2} + u + v \Big\} \\
&\qquad \le 2 \exp\Big(- \frac{ku^2}{2(1 + k/n + u/3)} \Big) + 2 \exp\Big(- \frac{kv^2}{2(1 + v/3)} \Big)
\end{aligned} \tag{3.5}$$

with $S_t := \mathbb{E}_t \|\Theta\|^4 - \pi_t \operatorname{tr}(\Sigma_t^2)$ *and* $\Sigma_t := \mathbb{E}_t(\Theta\Theta^{\top})$.

In particular, with probability greater than or equal to $1 - \delta$,

$$\begin{aligned}
&\sup_{V \in \mathcal{V}_p} |\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)| \\
&\qquad \le \Big[\frac{p \wedge (d-p)}{k} S_{t_{n,k}} \Big]^{1/2} + \frac{2 \log(4/\delta)}{3k} + \Big[\Big(\frac{\log(4/\delta)}{3k} \Big)^2
\end{aligned}$$

$$+\frac{2}{k}(1+k/n)\log(4/\delta)\Big]^{1/2}+\Big[\Big(\frac{\log(4/\delta)}{3k}\Big)^2+\frac{2}{k}\log(4/\delta)\Big]^{1/2}$$
$$\le\Big[\frac{p\wedge(d-p)}{k}S_{t_{n,k}}\Big]^{1/2}+\Big[\frac{8}{k}(1+k/n)\log(4/\delta)\Big]^{1/2}+\frac{4\log(4/\delta)}{3k}. \quad (3.6)$$

Note that (3.6) also implies an upper bound on the excess risk:

$$\begin{aligned}
&R_{t_{n,k}}(\hat V_n)-\inf_{V\in\mathcal{V}_p}R_{t_{n,k}}(V)\\
&\le \hat R_{n,k}(\hat V_n)-R_{t_{n,k}}(V^*_{t_{n,k}})+\sup_{V\in\mathcal{V}_p}|\hat R_{n,k}(V)-R_{t_{n,k}}(V)|\\
&\le \hat R_{n,k}(V^*_{t_{n,k}})-R_{t_{n,k}}(V^*_{t_{n,k}})+\sup_{V\in\mathcal{V}_p}|\hat R_{n,k}(V)-R_{t_{n,k}}(V)|\\
&\le 2\sup_{V\in\mathcal{V}_p}|\hat R_{n,k}(V)-R_{t_{n,k}}(V)|.
\end{aligned}$$

Remark 3.2. In the case $\omega(x)=1/\|x\|$, the upper bound in (3.6) simplifies to

$$\Big[\frac{p\wedge(d-p)}{k}\big(1-(k/n)\operatorname{tr}(\Sigma^2_{t_{n,k}})\big)\Big]^{1/2}+\Big[\frac{8}{k}(1+k/n)\log(4/\delta)\Big]^{1/2}+\frac{4\log(4/\delta)}{3k}. \quad \square$$

Note that the upper bound in Theorem 3.1 cannot be calculated from the data and can thus not directly be used to construct confidence intervals for the true reconstruction error $R_{t_{n,k}}(\hat V_n)$ or the minimal reconstruction error $\inf_{V\in\mathcal{V}_p}R_{t_{n,k}}(V)$. Therefore, we derive data-dependent bounds directly from (a minor improvement of) the bound established by [4]. However, this result will be applied to the conditional distribution of Θ given $\|X\|>t$ and the resulting bound is to be interpreted conditionally on the number N_t of exceedances over the chosen threshold t.

Theorem 3.3. *If condition* (3.1) *is met, for all* $\ell>1, u, v>0$,

$$\mathbb{P}\Big(\sup_{V\in\mathcal{V}_p}|\hat R_t(V)-R_t(V)|\ge\Big[(p\wedge(d-p))\Big(\frac{\tilde S_t}{\ell-1}+\frac{v}{\ell}\Big)\Big]^{1/2}+u\,\Big|\,N_t=\ell\Big)$$
$$\le 2\exp\big(-2\ell u^2\big)+\exp\big(-\lfloor\ell/2\rfloor v^2/2\big)$$

with $\tilde S_t:=N_t^{-1}\sum_{i=1}^n\|\Theta_{i,t}\|^4-\operatorname{tr}\Big((N_t^{-1}\sum_{i=1}^n\Theta_{i,t}\Theta_{i,t}^\top)^2\Big)$ *and* $\lfloor x\rfloor:=\max\{k\in\mathbb{Z}:k\le x\}$.

If, for all $\ell>1$, *constants* $u_\ell, v_\ell>0$ *are chosen such that* $2\exp\big(-2\ell u_\ell^2\big)+\exp\big(-\lfloor\ell/2\rfloor v_\ell^2/2\big)=1-\alpha$, *then*

$$I_\ell(V):=\big[\hat R_t(V)-B_{t,\ell},\hat R_t(V)+B_{t,\ell}\big]\cap[0,\infty)$$

with

$$B_{t,\ell}:=\Big[(p\wedge(d-p))\Big(\frac{\tilde S_t}{\ell-1}+\frac{v_\ell}{\ell}\Big)\Big]^{1/2}+u_\ell$$

defines a uniform level α *confidence band for* $R_t(V)$, $V\in\mathcal{V}_p$, *conditionally on* $N_t=\ell$. *If one defines* $I_0(V)=I_1(V)=[0,\infty)$, *then* $I_{N_t}(V)$ *defines a uniform level* α *confidence band for* $R_t(V)$, $V\in\mathcal{V}_p$ *(unconditionally).*

Remark 3.4. In the statement about the confidence bands one may replace $B_{t,\ell}$ with

$$\tilde{B}_{t,\ell} := \Big[(p \wedge (d-p))\frac{\tilde{S}_t}{\ell - 1}\Big]^{1/2} + \Big[(p \wedge (d-p))\frac{v_\ell}{\ell}\Big]^{1/2} + u_\ell.$$

This half width of a confidence band is more suitable for (numerical) minimization (as a function of u_ℓ and v_ℓ) under the constraint $2\exp\big(-2\ell u_\ell^2\big) + \exp(-\lfloor \ell/2 \rfloor v_\ell^2/2) = 1 - \alpha$. □

Remark 3.5. The conditional approach employed in Theorem 3.3 can also be used to obtain a uniform risk bound similar to the one in Theorem 3.1:

$$\begin{aligned}\mathbb{P}\Big\{ \sup_{V \in \mathcal{V}_p} |\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)| \geq \Big[(1+v)\frac{p \wedge (d-p)}{k} S^*_{t_{n,k}}\Big]^{1/2} + u + 2v \Big\} \\ \leq 2\exp\Big(-\frac{2ku^2}{1+v}\Big) + 2\exp\Big(-\frac{kv^2}{2(1+v/3)}\Big)\end{aligned}$$

with $S_t^* := \mathbb{E}_t \|\Theta\|^4 - \operatorname{tr}(\Sigma_t^2)$. A comparison with Theorem 3.1 reveals that the new bound may be tighter if $S^*_{t_{n,k}}$ is substantially smaller than $S_{t_{n,k}}$. This will be the case if k/n is small and $\operatorname{tr}\big((\mathbb{E}_t\, \Theta\Theta^\top)^2\big)$ is not much smaller than $\mathbb{E}_t \|\Theta\|^4$. □

So far, we have compared empirical risks with the true risk R_t for finite thresholds t. A comparison with the limit risk R_∞ would require second order refinements of our basic assumption (1.1). Let $\Sigma_t := \mathbb{E}_t(\Theta\Theta^\top) = P_t(\theta\theta^\top)$ and $\Sigma_\infty = P_\infty(\theta\theta^\top)$. Denote the eigenvalues of $\Sigma_t - \Sigma_\infty$ by $\lambda^\Delta_{t,1} \geq \lambda^\Delta_{t,2} \geq \ldots \geq \lambda^\Delta_{t,n}$. Then standard calculations from classical PCA show that

$$\sup_{V \in \mathcal{V}_p} R_t(V) - R_\infty(V) = \sup_U \operatorname{tr}(U^\top(\Sigma_t - \Sigma_\infty)U) = \sum_{i=1}^{d-p} \lambda^\Delta_{t,i}$$

where the second supremum is taken over all $(d \times (d-p))$-matrices with orthogonal columns. Likewise, $\sup_{V \in \mathcal{V}_p} R_\infty(V) - R_t(V) = -\sum_{i=1}^{d-p} \lambda^\Delta_{t,d+1-i}$ and hence

$$\begin{aligned}\sup_{V \in \mathcal{V}_p} |R_t(V) - R_\infty(V)| &\leq \max\Big(\Big|\sum_{i=1}^{d-p} \lambda^\Delta_{t,i}\Big|, \Big|\sum_{i=p+1}^{d} \lambda^\Delta_{t,i}\Big|\Big) \\ &= \max\Big(\Big|\sum_{i=1}^{p} \lambda^\Delta_{t,i}\Big|, \Big|\sum_{i=d-p+1}^{d} \lambda^\Delta_{t,i}\Big|\Big).\end{aligned}$$

Therefore, bounds on the difference between empirical risks and the limit risk require additional assumptions on the spectrum of the difference $\Sigma_t - \Sigma_\infty$ between the matrix of second moments for the re-scaled exceedances over the threshold t and the corresponding matrix in the limit model.

If one merely wants to compare the minimum risk for finite thresholds with the minimum limit risk, which equal the sums of $d-p$ smallest eigenvalues of Σ_t resp. Σ_∞, then somewhat weaker assumptions on the convergence of

the spectrum of Σ_t and Σ_∞ are needed. In particular, under Hypothesis 1, $\inf_{V\in\mathcal{V}_p} R_t(V) - \inf_{V\in\mathcal{V}_p} R_\infty(V)$ equals the sum of the smallest $d-p$ eigenvalues of Σ_t.

4. Simulation study

4.1. The setting

We investigate the performance of our PCA procedure. In particular, we examine how the standard non-parametric estimator of the spectral measure (defined via (1.2)) based on the k largest observations

$$\hat{H}_{n,k} := \frac{1}{k}\sum_{i=1}^{n} \delta_{\theta_{t_{n,k}}(X_i)}$$

(with $\theta(x) = x/\|x\|$) is influenced if the data is first projected onto a lower dimensional subspace using PCA:

$$\hat{H}^{PCA}_{n,k} := \frac{1}{k}\sum_{i=1}^{n} \delta_{\theta_{t_{n,k}}(\mathbf{\Pi}_V^{\perp} X_i)}.$$

Here, δ_y is the Dirac measure with point mass at y and V denotes the subspace picked by PCA based on the same number k of largest observations. It will turn out that sometimes it is advisable to use a smaller number $\tilde{k}$ for the PCA procedure; the resulting estimator of the spectral measure will be denoted by $\hat{H}^{PCA}_{n,k,\tilde{k}}$.

We simulate from different models of d-dimensional regularly varying vectors for which the spectral measure is (approximately) concentrated on a p-dimensional subspace. Since PCA is equivariant under rotations, w.l.o.g. we assume that this subspace is spanned by the first p unit vectors. The dependence between these p components is either described by a so-called Dirichlet model or by a Gumbel copula $C_\vartheta(x) = \exp\big(-\big(\sum_{i=1}^{p}(-\log x^i)^\vartheta\big)^{1/\vartheta}\big)$. In Example 3.6 of [32] it is described how to simulate from the former model, while the data according to the second model is generated using the transformation method proposed by [35]. The marginal distributions are chosen as Fréchet with *c.d.f.* $\exp(-x^{-\alpha})$, $\alpha \in \{1, 2\}$.

In addition, we have simulated observations from a Dirichlet model which are then rotated in the plane spanned by two randomly chosen coordinates, one of them among the first p coordinates, the other among the last $d-p$. The rotation angle is uniformly distributed on the interval $[-\pi/10, \pi/10]$. Note that, unlike in the first two models, Hypothesis 1 is not fulfilled here. By this means we evaluate how sensitive PCA is to moderate deviations from the ideal situation.

In all cases, we add the modulus of a d-dimensional multivariate normal vector with suitable variances and constant correlations 0.2. This way, it is ensured that the support of the exceedances over high thresholds is not fully concentrated on

the p-dimensional subspace. The variances are chosen equal to $10^5/d$ for $\alpha = 1$ (*i.e.*, if we start with unit Fréchet margins) and equal to $10/d$ for $\alpha = 2$, so that the sparsity assumption becomes apparent for the most extreme observations, whereas large yet less extreme data points are more spread out.

In all settings, we simulate samples of size $n = 1000$ and examine the performance of the PCA procedure based on $\Theta = X/\|X\|$ for the k vectors with largest norms for $k \in \{5, 10, 15, \ldots, 200\}$. The results reported here are based on 1000 simulations in each setting.

Write x^j for the jth coordinate of a p-dimensional vector x. To measure the performance of the spectral estimators, we calculate the errors of the resulting estimators of the following probabilities in the limit model, which can be expressed in terms of the spectral measure:

(i) $\lim_{u\to\infty} \mathbb{P}(p^{-1}\sum_{1\le j\le p} X^j/\|X\| > t_{(i)} \mid \|X\| > u)$
$= H\{x : p^{-1}\sum_{j=1}^{p} x^j > t_{(i)}\}$ for some $t_{(i)} \in (0, p^{-1/2})$

(ii) $\lim_{u\to\infty} \mathbb{P}(\min_{1\le j\le p} X^j > u, \max_{p+1\le j\le d} X^j \le u \mid \|X\| > u)$
$= \int \left((\min_{1\le j\le p} x^j)^\alpha - (\max_{p+1\le j\le d} x^j)^\alpha\right)^+ H(dx)$

(iii) $\lim_{u\to\infty} \mathbb{P}(X^1 > u \mid \max_{1\le j\le d} X^j > u)$
$= \int (x^1)^\alpha H(dx) / \int (\max_{1\le j\le p} x^j)^\alpha H(dx)$

(iv) $\lim_{u\to\infty} \mathbb{P}(\min_{1\le j\le d} X^j > u \mid \|X\| > u) = \int (\min_{1\le j\le d} x^j)^\alpha H(dx)$

The first probability is related to the *c.d.f.* of the mean contribution of the first p coordinates to the norm of the random vector, thus quantifying, in some sense, how strongly the norm is spread over the coordinates. Probability (ii) indicates how likely it is that the first p components are all large, while this is not true for any of the other components, given that the norm of the vector is large. Probability (iii) specifies how likely it is that the first component is extreme, given that any component is extreme. In a financial context, such probabilities are used to quantify how strongly a specific market participant is exposed to a failure of any market participant. Finally, probability (iv) specifies the minimal contribution of any coordinate to the norm. Note that under Hypothesis 1 this probability equals 0. The other true values are determined by Monte Carlo simulations with sample size of at least 10^7, unless they can be easily calculated analytically; the approximation error is always smaller than 10^{-3}. Throughout, we assume α to be known since we are interested in the effect of the PCA procedure on the estimator of the spectral measure, which should not be compounded with the estimation error of the tail index.

We first investigate the performance of the estimators in models of moderate dimension ($d = 5, p = 2$), before we examine high dimensional models ($d = 100, p = 5$).

4.2. Moderate dimensions

Throughout this subsection, vectors of dimension $d = 10$ are considered whose extremes are approximately concentrated on a two-dimensional subspace.

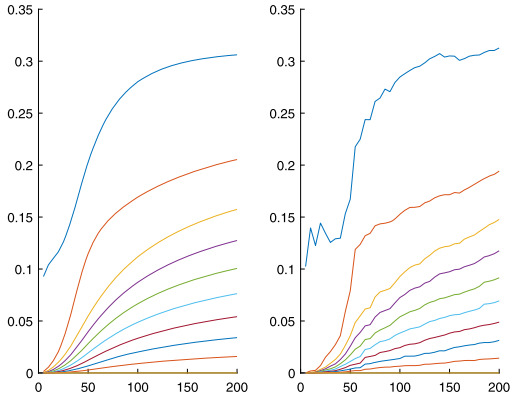

FIG 1. *Mean empirical risk (left) and empirical risk for one sample (right) versus k for PCA projecting onto a subspace of dimension* $1 \le \tilde{p} \le 10$ *in the Dirichlet model with parameter* 3, $p = 2$ *and* $d = 10$

We first discuss the simulation results for the Dirichlet model with all Dirichlet parameters α_i, $1 \le i \le p$, equal to 3 and unit Fréchet margins. Figure 1 shows the mean empirical risk in the left plot as a function of k for the PCA that projects onto a subspace of dimension $\tilde{p} \in \{1, \ldots, 10\}$. Since the mean empirical risk cannot be observed if one analyzes a given data set, the right plot shows the corresponding empirical risk for a single data set. The structure of both plots is very similar: essentially, the mean empirical risk curves are just a bit smoother. For this reason, in the remaining settings, we will only report the mean empirical risk.

It is obvious from the risk plot that $\tilde{p} = 2$ is a good choice, since there is a big gap to the empirical risk for $\tilde{p} = 1$, whereas the empirical risk almost vanishes for small k and $\tilde{p} = 2$, and the risk decreases more regularly for values $\tilde{p} > 2$, with no obvious structural breaks. The growing influence of the multivariate normal component as k increases is manifest in these plots, since the empirical risk quickly increases with k for all choices of $\tilde{p}$. This suggests to choose k rather small to detect the sparsity in the model, a finding which will be corroborated in the analysis of the estimator of the spectral measure below.

In Figure 2, the mean operator norm of the difference between the projection onto the true support of the limit measure μ and the projection onto the subspace of dimension 2 chosen by PCA is plotted versus k. Again it becomes obvious that for less extreme observations the approximation by a lower dimensional vector is rather poor, which leads to a larger error for the projection matrix estimated from these data. For $k = 80$, the norm has almost reached its maximal value. However, one should keep in mind that the operator norm measures the maximal distance between the projection of some vector $y \in \mathbb{S}^{d-1}$ onto the estimated respectively the true subspace. If the underlying distribution of $X/\|X\|$ puts little mass on vectors y for which the distance is large, the true risk corresponding to the estimated subspace may still be small.

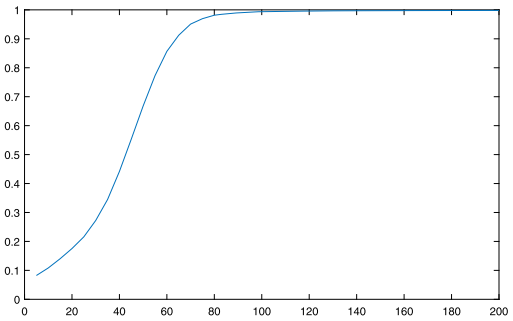

FIG 2. *Mean operator norm of the difference between the projection onto the true subspace and the projection onto the two-dimensional subspace picked by PCA as a function of k in the Dirichlet model with parameter* 3, $p = 2$ *and* $d = 10$

Next we consider the estimators of the probabilities (i)–(iv), obtained by replacing the spectral measure H either with $\hat{H}_{n,k}$ or $\hat{H}^{PCA}_{n,k}$. Since the PCA estimator of the subspace supporting μ quickly deteriorates as k increases, in addition we consider the estimators resulting from $\hat{H}^{PCA}_{n,k,10}$, that uses just the largest 10 observations to estimate the supporting subspace.

Figure 3 displays the root mean squared errors (RMSE) of the resulting estimators as a function of k. For very small values of k, all estimators perform similarly. For probability (i) with $t_{(i)} = 0.65$ (leading to a true value of about 0.684), both PCA based estimators have a considerably smaller RMSE than the standard estimator for most k. In particular, the PCA based method using just 10 largest observations to estimate the support of the spectral measure clearly outperforms both other estimators (almost) irrespective of the number of observations used for estimation of the spectral measure.

For the estimation of probability (ii) (≈ 0.309), the standard non-parametric estimator performs best for $k \leq 40$. The classical PCA using the same number of order statistics in both steps performs better for larger values of k and its minimum RMSE is a bit lower than that of the standard estimator. The PCA based estimator which determines the support of μ from the largest 10 observations has a very stable RMSE, but its minimum is much larger than that both of the other estimators.

In case (iii) (with true value of about 0.770), the RMSE of the standard estimator and the estimator based on $\hat{H}^{PCA}_{n,k,10}$ are very similar for k up to about 80, but the latter is remarkably insensitive to the choice of k up to 200. This feature might be useful in practical applications where the selection of k is often tricky. In contrast, the PCA based procedure that uses the same number of largest observations in both steps is even more sensitive to this choice than the standard estimator.

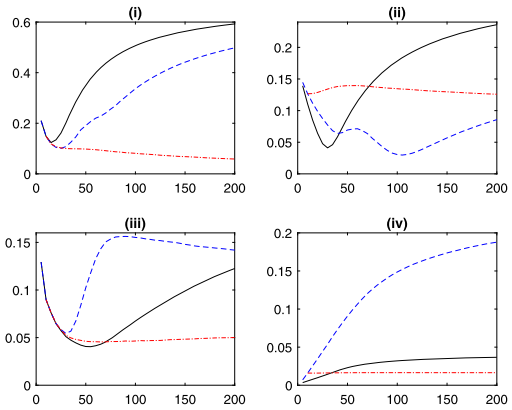

FIG 3. *RMSE of the estimators of the probabilities (i)–(iv) based on* $\hat{H}_{n,k}$ *(black, solid),* $\hat{H}^{PCA}_{n,k}$ *(blue, dashed) and* $\hat{H}^{PCA}_{n,k,10}$ *(red, dash-dotted) versus* k *in the Dirichlet model with parameter* 3, $p = 2$ *and* $d = 10$

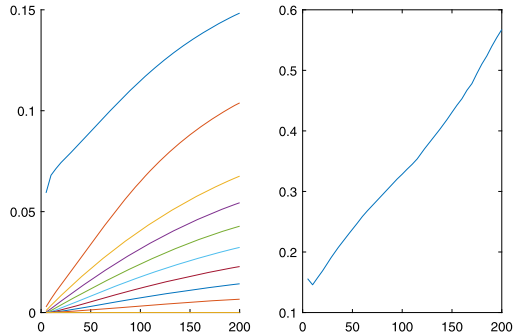

FIG 4. *Mean empirical risk for PCA projecting onto a subspace of dimension* $1 \le \tilde{p} \le 10$ *(left) and mean operator norm of the difference between the projection onto the true subspace and the projection onto the subspace picked by PCA with* $\tilde{p} = p$ *(right) versus* k *in the Gumbel model with parameter* $\vartheta = 2$, $p = 2$ *and* $d = 10$

Similarly, the classical PCA estimator of probability (iv) strongly depends on the choice of k while both other estimators stably have a very low error.

Next we consider the model whose extremal behavior is described by the Gumbel copula with $\vartheta = 2$ and Fréchet marginal distributions with *c.d.f.* $F(x) = \exp(-x^{-2})$, $x > 0$. The mean empirical risk and the mean operator norm of the difference matrix are shown in Figure 4. Overall the picture is similar as for the Dirichlet model, but the operator norm increases more slowly with k. Based on the left plot, one will choose $\tilde{p} = 2$.

Figure 5 displays the RMSE of the estimators of (i)–(iv) with PCA projecting on two-dimensional subspaces. Here $t_{(i)} = 0.7$ and the true values for (i)–(iv) are 0.3813, 0.083, $1/\sqrt{2}$ and 0. The relative performance of the estimators for

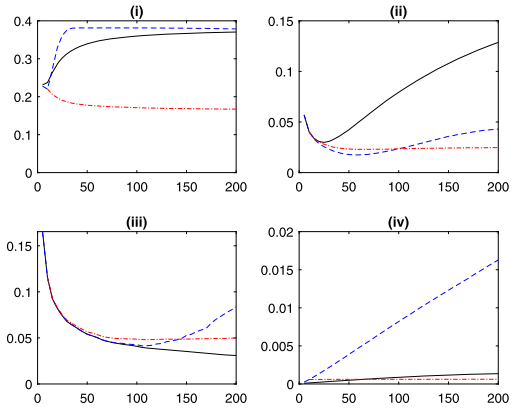

FIG 5. *RMSE of the estimators of the probabilities (i)–(iv) based on* $\hat{H}_{n,k}$ *(black, solid),* $\hat{H}^{PCA}_{n,k}$ *(blue, dashed) and* $\hat{H}^{PCA}_{n,k,10}$ *(red, dash-dotted) vs.* k *in the Gumbel model with parameter* $\vartheta = 2$*,* $p = 2$ *and* $d = 10$

probability (iv) is very similar to the one in the Dirichlet model. The same is true for the standard estimator of (i) and the PCA estimator that uses just 10 largest observations for estimating the support, but now the PCA estimator that uses the same number k in both steps performs slightly worse than the standard estimator. In contrast, both PCA based estimators of probability (ii) outperform the standard estimator, while for probability (iii) all three estimators perform similarly for k up to 100 where the usual PCA based estimator starts to deteriorate. Again, for all probabilities, the RMSE of the estimator resulting from $\hat{H}^{PCA}_{n,k,10}$ is remarkably insensitive against the choice of k.

Finally, we turn to the disturbed Dirichlet model where the observations are randomly rotated by an angular up to $\pi/10$, leading to true values for (i)–(iv) of 0.653 (with $t_{(i)} = 0.65$), 0.185, 0.770 and 0. The corresponding plots are shown in the Figures 6 and 7. In view of the empirical risk, the choices $\tilde{p} \in \{2, 3\}$ seem reasonable.

Again, the PCA procedure that uses the same largest observations in both steps performs better for the larger choice of $\tilde{p}$, whereas the performance of the other PCA procedure improves only for (ii), while it does not change much for (iii) and it deteriorates a bit for (i) and (iv). The PCA estimators perform better than the standard procedure for probability (i) and for (iii) if k is large (for classical PCA only if $\tilde{p} = 3$), whereas for (ii) overall the estimators perform similarly well with the standard procedure performing better for small values of k and the PCA estimators for larger values.

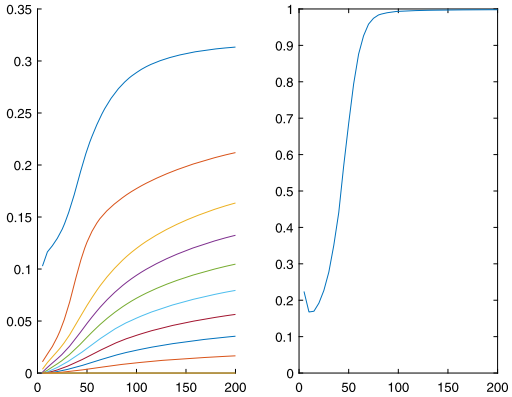

FIG 6. *Mean empirical risk for PCA projecting onto a subspace of dimension* $1 \le \tilde{p} \le 10$ *(left) and mean operator norm of the difference between the projection onto the true subspace and the projection onto the subspace picked by PCA with* $\tilde{p} = p$ *(right) versus* k *for randomly rotated Dirichlet observations with parameter* 3, $p = 2$ *and* $d = 10$

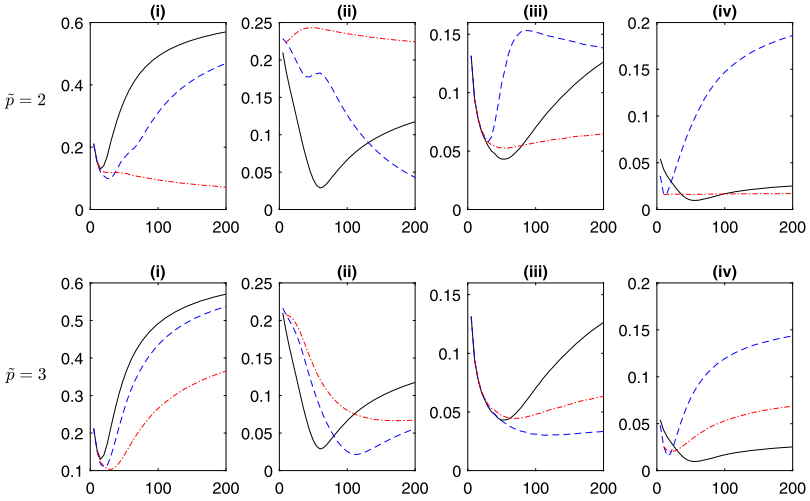

FIG 7. *RMSE of the estimators of the probabilities (i)–(iv) based on* $\hat{H}_{n,k}$ *(black, solid),* $\hat{H}^{PCA}_{n,k}$ *(blue, dashed) and* $\hat{H}^{PCA}_{n,k,10}$ *(red, dash-dotted) vs.* k *for randomly rotated Dirichlet observations with parameter* 3, $p = 2$ *and* $d = 10$*; the upper plots correspond to PCA projections on subspaces of dimension* $\tilde{p} = 2$*, the lower to* $\tilde{p} = 3$

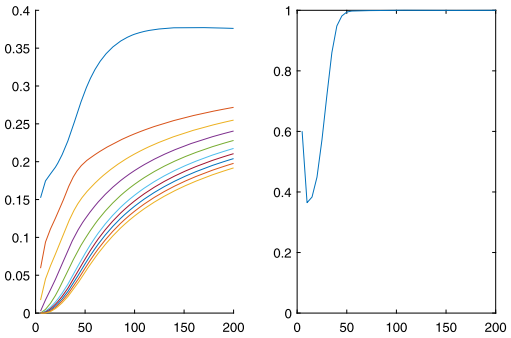

FIG 8. *Mean empirical risk for PCA projecting onto a subspace of dimension* $1 \leq \tilde{p} \leq 10$ *(left) and mean operator norm of the difference between the projection onto the true subspace and the projection onto the subspace picked by PCA with* $\tilde{p} = p$ *(right) versus* k *in the Dirichlet model with parameter* 3, $p = 5$ *and* $d = 100$

4.3. High dimensional models

We now compare the estimators when random vectors of dimension $d = 100$ are observed whose extremes are concentrated near a $p = 5$ dimensional subspace.

Again, we start with the Dirichlet model, for which the mean empirical risk for PCA projecting on a subspace of dimension $\tilde{p} \in \{1, \ldots, 10\}$ are shown in the left plot of Figure 8 and the mean operator norm of the difference between the estimated and the true projection matrix in the right plot. Here the choice of an appropriate dimension based on the empirical risk plot is less obvious than in the lower dimensional setting, but one should clearly arrive at some value $\tilde{p}$ between 4 and 6 and choose k not much larger than 50 for estimating the support of the limit measure.

Figure 9 shows the RMSE of the different estimators of the probabilities (i)–(iv) with $t_{(i)} = 0.4$ and true values $0.573, 0.072, 0.584$ and 0, respectively. Here, we have used PCA with $\tilde{p} = 4$ in the upper row, $\tilde{p} = 5$ in the mid row and $\tilde{p} = 6$ in the lower row. As expected, in most cases the PCA procedures perform worse when they project on too low dimensional subspaces, yet in the cases (i) and (iv) the differences are moderate. At first glance somewhat surprisingly, overall the PCA procedures exhibit a better behavior for $\tilde{p} = 6$ than for the "correct" value $\tilde{p} = 5$. This may be explained by the fact that the extra dimension offers the opportunity to compensate for the difference between the subspaces minimizing the true resp. the empirical risk. This difference is expected to be larger if the dimension of the observed vectors is large, as can also be seen from the right plot in Figure 8.

Again, the PCA based estimators for probability (i) outperform the standard procedure, but the other probabilities are more accurately estimated by the standard procedures if $\tilde{p} \leq 5$ (though all estimators of (iv) perform reasonably well). For $\tilde{p} = 6$, the RMSE of both variants of PCA based estimators of (ii) are very similar with a minimum value which is somewhat smaller than

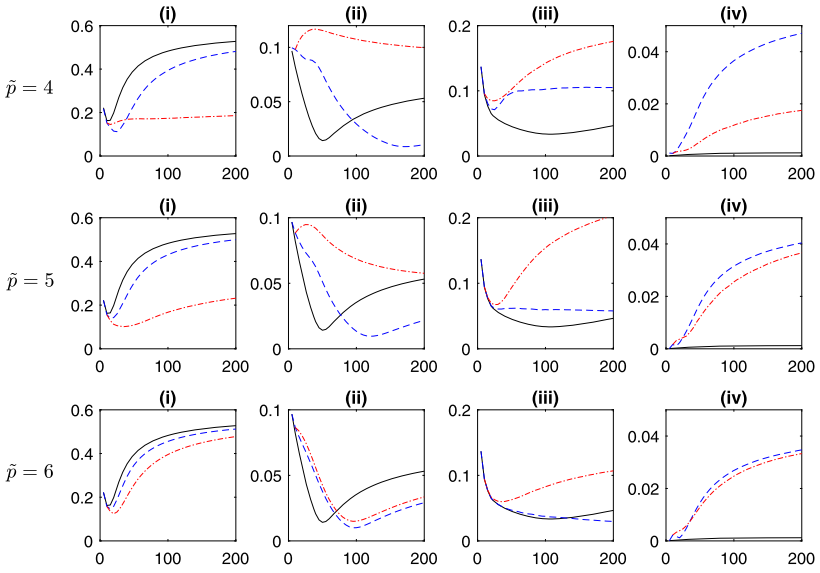

FIG 9. *RMSE of the estimators of the probabilities (i)–(iv) based on* $\hat{H}_{n,k}$ *(black, solid),* $\hat{H}^{PCA}_{n,k}$ *(blue, dashed) and* $\hat{H}^{PCA}_{n,k,10}$ *(red, dash-dotted) vs.* k *in the Dirichlet model with parameter* 3, $p = 5$ *and* $d = 100$*; the upper plots correspond to PCA projections on subspaces of dimension* $\tilde{p} = 4$*, the middle to* $\tilde{p} = 5$*, and the lower to* $\tilde{p} = 6$

the minimum RMSE of the standard estimator. The performance of the standard estimator and the one based on classical PCA are almost identical for the probability (iii), while the estimator with PCA based on just $k = 10$ largest observations is less accurate, probably because it is difficult to estimate a subspace of dimension 6 based on just 10 observations. It might help to increase the number of largest observations used to estimate the supporting subspace with the dimension d, but we do not explore this idea here in order not to overload the presentation.

For the high dimensional Gumbel model with $d = 100$ and $p = 5$, by and large the findings are similar to the ones observed for the Dirichlet model so that we do not show the corresponding plots. However, in this model $\tilde{p} = 4$ can be ruled out by the empirical risk plot and both PCA based estimators outperform the standard estimator of (ii).

4.4. Conclusion

To sum up, while the PCA step does not always improve the estimator of the spectral measure, for probability (i) the resulting estimators are superior to the standard estimator and in most other cases they seem competitive if $\tilde{p}$ is chosen appropriately. To this end, the plot of the empirical risk is a very useful tool; this is particularly true for observations with moderate dimensions. For

higher dimensional data, there may be some ambiguity about the dimension of the subspace onto which the data should be projected. In case of doubt, it is advisable to choose a higher dimensional subspace, in particular for the PCA method that uses the same number of largest observations to estimate the support and to calculate the estimator of the spectral measure. The PCA estimators that determine the support based only on the largest 10 observations often exhibit a desirable insensitivity to the choice of largest observations used to estimate the spectral measure, which makes them easier to apply in practice.

5. Proofs

5.1. Proofs to Section 2

The following technical lemma comes in handy for the asymptotic analysis of the conditional risk.

Lemma 5.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function that is locally bounded, P_∞-a.e. continuous and satisfies $\limsup_{\|x\|\to\infty} |f(x)|\|x\|^{-\tilde\alpha} < \infty$ for some $\tilde\alpha < \alpha$. Then $\lim_{t\to\infty} \int f(x/t)\, P_t(dx) = \int f(x)\, P_\infty(dx)$.*

Proof. According to (1.1),

$$P_t(t\,\cdot) = \mathbb{P}(X \in t\,\cdot \mid \|X\| > t) \to \mu|_{(B_1(0))^c}(\,\cdot\,) = P_\infty(\,\cdot\,)$$

weakly. Let Y_t and Y_∞ be random vectors with distribution $P_t(t\,\cdot)$ and P_∞, respectively. Since $\int f(x/t)\, P_t(dx) = \mathbb{E}\, f(Y_t)$ and $\int f(x)\, P_\infty(dx) = \mathbb{E}(Y_\infty)$, the assertion follows if the $f(Y_t)$ are asymptotically uniformly integrable (see [36], Theorem 2.20).

By assumption $f(Y_t)$ can be bounded by a multiple of $1 + \|Y_t\|^{\tilde\alpha}$. Now, for all $\tau \in [0, \alpha)$ and $t \geq t_0$ for some sufficiently large t_0, integration by parts, regular variation of $u \mapsto u^{\tau-1}\mathbb{P}\{\|X\| > u\}$ with index $\tau - \alpha - 1$ and Karamata's theorem (see [3], Theorem 1.6.1) yield

$$\begin{aligned}
\mathbb{E}\,\|Y_t\|^\tau &= \int \|x/t\|^\tau\, P_t(dx) \\
&= \frac{t^{-\tau}}{\mathbb{P}\{\|X\| > t\}} \int_t^\infty u^\tau\, \mathbb{P}^{\|X\|}(du) \\
&= \frac{\tau}{t^\tau \mathbb{P}\{\|X\| > t\}} \int_t^\infty u^{\tau-1} \mathbb{P}\{\|X\| > u\}\, du \\
&\leq 2\frac{\tau}{\alpha - \tau}. \qquad (5.1)
\end{aligned}$$

In particular, $\sup_{t\geq t_0} \mathbb{E}\,\|Y_t\|^{\tilde\alpha(1+\varepsilon)} < \infty$ for $\varepsilon \in (0, \alpha/\tilde\alpha - 1)$, so that $\|Y_t\|^{\tilde\alpha}$ and thus $f(Y_t)$ are asymptotically uniformly integrable. □

Proof of Proposition 2.2. Note that by the homogeneity of θ and (1.6),

$$t^{2(\beta-1)} R_t(V) = P_t(\|\mathbf{\Pi}_V^\perp t^{\beta-1}\theta\|^2) = \int f(x/t)\, P_t(dx)$$

with $f(x) := \|\mathbf{\Pi}_V^\perp \theta(x)\|^2 \le c_\omega^2 \|x\|^{2(1-\beta)}$. Since $2(1-\beta) < \alpha$, Lemma 5.1 yields the assertion. □

The following two lemmas are crucial to prove convergence of the minimizers of R_t.

Lemma 5.2. *The set $\mathcal{V}_p$ of p-dimensional linear subspaces of $\mathbb{R}^d$ is compact w.r.t. ρ.*

Proof. We have to show that any sequence $(V_n)_{n\in\mathbb{N}}$ in $\mathcal{V}_p$ has a convergent subsequence. For each n, let $(u_{1,n}, \ldots, u_{p,n})$ be an orthonormal basis for V_n so that $\mathbf{\Pi}_{V_n} x = U_n U_n^\top x$ where U_n denotes the matrix with columns $u_{j,n}$. The vectors $(u_{j,n})_{1\le j\le p}$ belong to the compact set $(\mathbb{S}^{d-1})^p$. Thus there exists a subsequence n_ℓ such that $u_{j,n_\ell} \to u_j^0$ for all $1 \le j \le p$. Since for all n, $\langle u_{j,n}, u_{i,n}\rangle = \delta_{i,j}$, we also have $\langle u_j^0, u_i^0\rangle = \delta_{i,j}$ and the $u_j^0, j \le p$, form an orthonormal family in $\mathbb{R}^d$. Let V^0 be the space generated by the u_j^0's and denote by U_0 the matrix with these columns. Then V^0 has dimension p, *i.e.* $V^0 \in \mathcal{V}_p$, and by construction

$$\rho(V_{n_\ell}, V^0) = \sup_{x\in\mathbb{S}^{d-1}} \|(U_{n_\ell} U_{n_\ell}^\top - U_0 U_0^\top)x\| \to 0.$$

which proves the claimed compactness. □

Lemma 5.3. *If ω satisfies condition (1.5), then for sufficiently large t_0, the standardized risks $\tilde{R}_t = t^{2(\beta-1)} R_t$, $t \ge t_0$, are Lipschitz continuous w.r.t. ρ with a Lipschitz constant independent of t.*

Proof. First note that $\big|\|\mathbf{\Pi}_V^\perp\theta(x)\| - \|\mathbf{\Pi}_W^\perp\theta(x)\|\big| \le \|\mathbf{\Pi}_V^\perp\theta(x) - \mathbf{\Pi}_W^\perp\theta(x)\| \le \|\theta(x)\|\rho(V,W)$. Recall the definition of Y_t given in the proof of Lemma 5.1. From (1.6) and (5.1) one may conclude that, for all $\gamma \in (0, \alpha/(1-\beta))$, eventually

$$t^{\gamma(\beta-1)} P_t\|\theta\|^\gamma \le c_\omega^\gamma\, \mathbb{E}_t\, \|X/t\|^{\gamma(1-\beta)} = c_\omega^\gamma\, \mathbb{E}\, \|Y_t\|^{\gamma(1-\beta)} \le 2c_\omega^\gamma \frac{\gamma(1-\beta)}{\alpha - \gamma(1-\beta)}. \tag{5.2}$$

Hence, for all subspaces V, W of $\mathbb{R}^d$

$$\begin{aligned}
|\tilde{R}_t(V) - \tilde{R}_t(W)| &= t^{2(\beta-1)} \Big| P_t\|\mathbf{\Pi}_V^\perp\theta\|^2 - P_t\|\mathbf{\Pi}_W^\perp\theta\|^2 \Big| \\
&\le t^{2(\beta-1)} P_t\Big(\big|\|\mathbf{\Pi}_V^\perp\theta\| - \|\mathbf{\Pi}_W^\perp\theta\|\big| \cdot \big(\|\mathbf{\Pi}_V^\perp\theta\| + \|\mathbf{\Pi}_W^\perp\theta\|\big)\Big) \\
&\le 2t^{2(\beta-1)} P_t\|\theta\|^2 \rho(V,W) \\
&\le 4c_\omega^2 \frac{2(1-\beta)}{\alpha - 2(1-\beta)} \rho(V,W)
\end{aligned}$$

which proves the assertion. □

Proof of Theorem 2.4. Suppose the assertion of the theorem were wrong. By the compactness of $\mathcal{V}_p$, then there exist a sequence $t_n \to \infty$ such that $V_{t_n}^*$ converges to some $V_\infty \neq V_\infty^*$. By Lemma 5.3, $|\tilde{R}_{t_n}(V_{t_n}^*) - \tilde{R}_{t_n}(V_\infty)| \to 0$, and by Proposition 2.2 $|\tilde{R}_{t_n}(V_\infty) - R_\infty(V_\infty)| \to 0$ and $|\tilde{R}_{t_n}(V_\infty^*) - R_\infty(V_\infty^*)| \to 0$. Hence, for $\varepsilon := R_\infty(V_\infty) - R_\infty(V_\infty^*)$, which is strictly positive by assumption, and sufficiently large n, one may conclude a contradiction:

$$\begin{aligned} R_\infty(V_\infty) &\le \tilde{R}_{t_n}(V_\infty) + \frac{\varepsilon}{4} \le \tilde{R}_{t_n}(V_{t_n}^*) + \frac{\varepsilon}{2} \\ &\le \tilde{R}_{t_n}(V_\infty^*) + \frac{\varepsilon}{2} \le R_\infty(V_\infty^*) + \frac{3\varepsilon}{4} < R_\infty(V_\infty). \end{aligned}$$

Therefore, the assertion must be correct. □

Proof of Proposition 2.6. For simplicity, we assume that $F_{\|X\|}$ is continuous in the tail (so that there are no ties among the observed norms), but the proof can be easily generalized using standard techniques from the theory of regular varying functions. First we want to replace the random threshold $\hat{t}_{n,k}$ with $t_{n,k}$ in the definition of $\hat{R}_{n,k}$. Since $\mathbf{\Pi}_V^\perp$ is a contraction, the Hölder inequality yields

$$\begin{aligned} & t_{n,k}^{2(\beta-1)} \Big| \hat{R}_{n,k}(V) - \frac{1}{k} \sum_{i=1}^n \|\mathbf{\Pi}_V^\perp \Theta_i\|^2 \mathbf{1}\{\|X_i\| > t_{n,k}\} \Big| \\ &\le \frac{1}{k} t_{n,k}^{2(\beta-1)} \sum_{i=1}^n \|\mathbf{\Pi}_V^\perp \Theta_i\|^2 \big| \mathbf{1}\{\|X_i\| > \hat{t}_{n,k}\} - \mathbf{1}\{\|X_i\| > t_{n,k}\} \big| \qquad (5.3) \\ &\le \Big[\frac{1}{k} \sum_{i=1}^n t_{n,k}^{(2+\eta)(\beta-1)} \|\Theta_i\|^{2+\eta} \mathbf{1}\{\|X_i\| > t_{n,k} \wedge \hat{t}_{n,k}\} \Big]^{2/(2+\eta)} \\ &\quad \cdot \Big[\frac{1}{k} \sum_{i=1}^n \big| \mathbf{1}\{\|X_i\| > \hat{t}_{n,k}\} - \mathbf{1}\{\|X_i\| > t_{n,k}\} \big|^{(2+\eta)/\eta} \Big]^{\eta/(2+\eta)}. \end{aligned}$$

where $\eta > 0$ is chosen such that $(2+\eta)(1-\beta) < \alpha$.

It is well known that $\hat{t}_{n,k}/t_{n,k} \to 1$ in probability. Thus there exists a sequence $\delta_n \downarrow 0$ such that $P\{\hat{t}_{n,k} > (1-\delta_n) t_{n,k}\} \to 0$. By (5.2) and the regular variation of $1 - F_{\|X\|}$

$$\begin{aligned} & \mathbb{E}\Big(t_{n,k}^{(2+\eta)(\beta-1)} \|\Theta\|^{2+\eta} \mathbf{1}\{\|X_i\| > (1-\delta_n) t_{n,k}\} \Big) \\ &= t_{n,k}^{(2+\eta)(\beta-1)} P_{(1-\delta_n)t_{n,k}} \|\theta\|^{2+\eta} \big(1 - F_{\|X\|}((1-\delta_n) t_{n,k})\big) \\ &= O\big(1 - F_{\|X\|}(t_{n,k})\big) = O(k/n). \end{aligned}$$

In particular, $k^{-1} \sum_{i=1}^n t_{n,k}^{(2+\eta)(\beta-1)} \|\Theta_i\|^{2+\eta} \mathbf{1}\{\|X_i\| > t_{n,k} \wedge \hat{t}_{n,k}\}$ is stochastically bounded.

Furthermore,

$$\sum_{i=1}^n \big| \mathbf{1}\{\|X_i\| > \hat{t}_{n,k}\} - \mathbf{1}\{\|X_i\| > t_{n,k}\} \big|^{(2+\eta)/\eta} = \Big| \sum_{i=1}^n \mathbf{1}\{\|X_i\| > t_{n,k}\} - k \Big|,$$

because there exist exactly k exceedances of $\hat{t}_{n,k}$, and either all non-vanishing differences of the indicator functions equal 1 or all equal -1, depending on whether $\hat{t}_{n,k} < t_{n,k}$ or $\hat{t}_{n,k} > t_{n,k}$. Now, the last sum is binomially distributed with parameters n and k/n. By the central limit theorem for triangular arrays, the right hand side is of stochastic order $k^{1/2}$.

A combination of these results show that

$$t_{n,k}^{2(\beta-1)}\Big|\hat{R}_{n,k}(V) - \frac{1}{k}\sum_{i=1}^{n}\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| > t_{n,k}\}\Big| = O_P\big(k^{-\eta/(2(2+\eta))}\big) = o_P(1) \tag{5.4}$$

uniformly for all subspaces V.

In view of Proposition 2.2, it thus suffices to show that

$$\begin{aligned}
&t_{n,k}^{2(\beta-1)}\Big|\frac{1}{k}\sum_{i=1}^{n}\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| > t_{n,k}\} - R_{t_{n,k}}(V)\Big| \\
&\quad\le t_{n,k}^{2(\beta-1)}\Big|\frac{1}{k}\sum_{i=1}^{n}\Big(\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| \in (t_{n,k}, d_{n,k}]\} \\
&\qquad\qquad - \mathbb{E}\left(\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| \in (t_{n,k}, d_{n,k}]\}\right)\Big)\Big| \\
&\qquad + t_{n,k}^{2(\beta-1)}\Big|\frac{1}{k}\sum_{i=1}^{n}\Big(\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| > d_{n,k}\} \\
&\qquad\qquad - \mathbb{E}\left(\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| > d_{n,k}\}\right)\Big)\Big| \\
&\quad=: T_{n,1} + T_{n,2} \to 0
\end{aligned}$$

in probability, with $d_{n,k} := (\log k)t_{n,k}$.

Let $\alpha^* := 4(1-\beta) \vee (\alpha+1)$. Since $\mathbf{\Pi}_V^{\perp}$ is a contraction, (1.6) implies

$$\begin{aligned}
\mathbb{E}(T_{n,1}^2) &= \frac{n}{k^2}t_{n,k}^{4(\beta-1)}\,\mathbf{Var}\left(\|\mathbf{\Pi}_V^{\perp}\Theta\|^2\mathbf{1}\{\|X\| \in (t_{n,k}, d_{n,k}]\}\right) \\
&\le \frac{n}{k^2}t_{n,k}^{-\alpha^*}c_\omega^4\,\mathbb{E}\left(\|X\|^{\alpha^*}\mathbf{1}\{\|X\| \in (t_{n,k}, d_{n,k}]\}\right).
\end{aligned}$$

Similarly as in the proof of Lemma 5.1, we can bound the expectation using integration by parts and Karamata's theorem:

$$\begin{aligned}
&\mathbb{E}\left(\|X\|^{\alpha^*}\mathbf{1}\{\|X\| \in (t_{n,k}, d_{n,k}]\}\right) \\
&\quad\le \int_0^{d_{n,k}} z^{\alpha^*}\,\mathbb{P}^{\|X\|}(dz) \\
&\quad= -d_{n,k}^{\alpha^*}(1 - F_{\|X\|}(d_{n,k})) + \alpha^*\int_0^{d_{n,k}} z^{\alpha^*-1}(1 - F_{\|X\|}(z))\,dz \\
&\quad= d_{n,k}^{\alpha^*}(1 - F_{\|X\|}(d_{n,k}))\Big(\frac{\alpha^*}{\alpha^* - \alpha} - 1 + o(1)\Big) \\
&\quad\le \frac{2\alpha}{\alpha^* - \alpha}d_{n,k}^{\alpha^*}(1 - F_{\|X\|}(d_{n,k}))
\end{aligned}$$

for sufficiently large n. Therefore, by the choice of $d_{n,k}$,

$$\mathbb{E}(T_{n,1}^2) = O\Big(\frac{n}{k^2}(\log k)^{\alpha^*}(1 - F_{\|X\|}(d_{n,k}))\Big) = o\Big(\frac{(\log k)^{\alpha^*}}{k}\Big) = o(1),$$

which implies the convergence in probability of $T_{n,1}$.

For the second term, we may similarly conclude from (5.2) and the definition of $t_{n,k}$ that

$$\begin{aligned}\mathbb{E}(T_{n,2}) &\le \frac{n}{k} t_{n,k}^{2(\beta-1)} 2\,\mathbb{E}\left(\|\mathbf{\Pi}_V^{\perp}\Theta\|^2 \mathbf{1}\{\|X\| > d_{n,k}\}\right) \\ &\le 2\frac{n}{k} t_{n,k}^{2(\beta-1)}\,\mathbb{E}_{d_{n,k}}\left(\|\Theta\|^2\right)(1 - F_{\|X\|}(d_{n,k})) \\ &\le \frac{8(1-\beta)c_\omega^2}{\alpha - 2(1-\beta)} \cdot \frac{d_{n,k}^{2(1-\beta)}(1 - F_{\|X\|}(d_{n,k}))}{t_{n,k}^{2(1-\beta)}(1 - F_{\|X\|}(t_{n,k}))}.\end{aligned}$$

Because $t \mapsto t^{2(1-\beta)}(1-F_{\|X\|}(t))$ is regularly varying with index $2(1-\beta)-\alpha < 0$ and $t_{n,k} = o(d_{n,k})$, the right hand side tends to 0. Thus, also $T_{n,2}$ converges to 0 in probability, which concludes the proof. □

Similarly as in the analysis of the conditional risk, the following result on the equicontinuity in probability of the standardized empirical risk is central to the consistency of the empirical risk minimizer.

Lemma 5.4. *If ω satisfies condition (1.5), then for all $\varepsilon > 0$ there exists $\delta > 0$ such that for sufficiently large n*

$$P\Big\{\sup_{V,W\in\mathcal{V}_p:\rho(V,W)\le\delta} t_{n,k}^{2(\beta-1)}|\hat{R}_{n,k}(V) - \hat{R}_{n,k}(W)| > \varepsilon\Big\} \le \varepsilon.$$

Proof. First note that in view of (5.4), it suffices to prove the assertion with $\hat{R}_{n,k}(V)$ replaced by $k^{-1}\sum_{i=1}^n \|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2\mathbf{1}\{\|X_i\| > t_{n,k}\}$ and $\hat{R}_{n,k}(W)$ replaced by the analogous expression.

Similarly as in the proof of Lemma 5.3, we have

$$\begin{aligned}&\frac{1}{k}\sum_{i=1}^n \left(\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2 - \|\mathbf{\Pi}_W^{\perp}\Theta_i\|^2\right)\mathbf{1}\{\|X_i\| > t_{n,k}\} \\ &\quad\le \frac{2}{k}\sum_{i=1}^n \|\mathbf{\Pi}_V^{\perp}\Theta_i - \mathbf{\Pi}_W^{\perp}\Theta_i\| \cdot \|\Theta_i\|\mathbf{1}\{\|X_i\| > t_{n,k}\} \\ &\quad\le \frac{2}{k}\rho(V,W)\sum_{i=1}^n \|\Theta_i\|^2\mathbf{1}\{\|X_i\| > t_{n,k}\} \\ &\quad\le \frac{2}{k}\rho(V,W)c_\omega^2\sum_{i=1}^n \|X_i\|^{2(1-\beta)}\mathbf{1}\{\|X_i\| > t_{n,k}\}\end{aligned}$$

for n sufficiently large. Hence, by Markov's inequality, (5.2) and the definition of $t_{n,k}$,

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{V,W\in\mathcal{V}_p:\rho(V,W)\le\delta} \frac{1}{k} t_{n,k}^{2(\beta-1)} \sum_{i=1}^{n} \big(\|\mathbf{\Pi}_V^{\perp}\Theta_i\|^2 - \|\mathbf{\Pi}_W^{\perp}\Theta_i\|^2\big)\mathbf{1}\{\|X_i\| > t_{n,k}\} > \varepsilon \Big\} \\
&\quad \le \frac{2c_\omega^2\delta n}{\varepsilon k}\,\mathbb{E}\left(\Big(\frac{\|X\|}{t_{n,k}}\Big)^{2(1-\beta)}\mathbf{1}\{\|X\| > t_{n,k}\}\right) \\
&\quad = \frac{2c_\omega^2\delta}{\varepsilon}\,\mathbb{E}_{t_{n,k}}\Big(\frac{\|X\|}{t_{n,k}}\Big)^{2(1-\beta)} \\
&\quad \le \frac{8(1-\beta)c_\omega^2\delta}{\varepsilon(\alpha-2(1-\beta))} = \varepsilon
\end{aligned}$$

for $\delta := \varepsilon^2(\alpha-2(1-\beta))/\big(8(1-\beta)c_\omega^2\big)$. □

Proof of Theorem 2.7. Let $\tilde{R}_{n,k} := t_{n,k}^{2(\beta-1)}\hat{R}_{n,k}$. Fix an arbitrary $\varepsilon > 0$ and let $\mathcal{M} := \{W \in \mathcal{V}_p : \rho(V_\infty^*, W) \ge \varepsilon/2\}$. In view of Proposition 2.2 and Lemma 5.3, it is easily seen that R_∞ is Lipschitz continuous w.r.t. ρ. Moreover, by Lemma 5.2, $\mathcal{M}$ is a closed subset of a compact set and thus compact, too. Hence, $\eta := \inf_{W\in\mathcal{M}} R_\infty(W) - R_\infty(V_\infty^*) > 0$, since the infimum is attained and V_∞^* is the unique minimizer of R_∞.

According to Lemma 5.4, there exists $\delta \le \varepsilon/2$ and n_0 such that for all $n \ge n_0$ with probability greater than $1-\varepsilon/4$ one has

$$|\tilde{R}_{n,k}(V) - \tilde{R}_{n,k}(W)| \le \eta/4$$

for all $V, W \in \mathcal{V}_p$ such that $\rho(V,W) \le \delta$. Since $\mathcal{V}_p$ is compact, there exists a finite cover of $\mathcal{V}_p$ by open balls with radius δ and centers $W_1, \ldots, W_m$, say. By Proposition 2.6, there exists $n_1 \ge n_0$ such that with probability greater than $1-\varepsilon/2$

$$\begin{aligned}
&|\tilde{R}_{n,k}(W_j) - R_\infty(W_j)| \le \eta/4, \quad \forall\, 1 \le j \le m, \\
&|\tilde{R}_{n,k}(V_\infty^*) - R_\infty(V_\infty^*)| \le \eta/4.
\end{aligned}$$

Hence, there exists a (random) index $j \in \{1,\ldots,m\}$ such that $\rho(\hat{V}_n, W_j) < \delta \le \varepsilon/2$, and, on a set with probability greater than $1-\varepsilon$,

$$R_\infty(W_j) \le \tilde{R}_{n,k}(W_j) + \frac{\eta}{4} \le \tilde{R}_{n,k}(\hat{V}_n) + \frac{\eta}{2} \le \tilde{R}_{n,k}(V_\infty^*) + \frac{\eta}{2} \le R_\infty(V_\infty^*) + \frac{3\eta}{4}.$$

By the definition of η, this implies $W_j \notin \mathcal{M}$ and thus

$$\rho(\hat{V}_n, V_\infty) \le \rho(\hat{V}_n, W_j) + \rho(W_j, V_\infty^*) < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this concludes the proof. □

5.2. Proofs to Section 3

The proofs rely on some well-known facts about Hilbert spaces, which we recall in a version specialized to the present setting. Let $(e_i)_{1\le i\le d}$ be an arbitrary orthonormal basis of $\mathbb{R}^d$ and denote by $\langle\cdot,\cdot\rangle$ the usual inner product on $\mathbb{R}^d$. The space of linear operators from $\mathbb{R}^d$ to $\mathbb{R}^d$ (*i.e.*, $d\times d$-matrices) equipped with the inner product $\langle A,B\rangle_{HS}:=\sum_{i=1}^d\langle Ae_i,Be_i\rangle$ (which does not depend on the chosen orthonormal basis) is a Hilbert space. The corresponding Hilbert Schmidt norm can be expressed as $\|A\|_{HS}=\big(\sum_{i=1}^d\|Ae_i\|^2\big)^{1/2}=\big(\operatorname{tr}(AA^\top)\big)^{1/2}$ with tr denoting the trace operator. If, for any subspace W of $\mathbb{R}^d$, the first $\dim W$ vectors e_i form an orthonormal basis of W, then one sees that

$$\|\mathbf{\Pi}_W\|_{HS}=\sqrt{\dim W}. \tag{5.5}$$

Moreover, direct calculations show that

$$\langle Ay,x\rangle=\langle A,xy^\top\rangle_{HS}. \tag{5.6}$$

Finally, for independent centered random matrices A_i, $1\le i\le n$, one has

$$\mathbb{E}\Big\|\sum_{i=1}^n A_i\Big\|_{HS}^2=\sum_{i=1}^n\mathbb{E}\,\|A_i\|_{HS}^2. \tag{5.7}$$

We start with a uniform bound on the difference between $\hat R_{n,k}$ and $\bar R_{t_{n,k}}$, defined in (3.2).

Lemma 5.5. *If (3.1) holds, then*

$$\mathbb{P}\Big\{\sup_{V\in\mathcal{V}_p}|\hat R_{n,k}(V)-\bar R_{t_{n,k}}(V)|\ge v\Big\}\le 2\exp\Big(-\frac{kv^2}{2(1+v/3)}\Big).$$

Proof. In view of (5.3) and $\pi_{t_{n,k}}=k/n$, we have for all $V\in\mathcal{V}_p$,

$$\begin{aligned}|\hat R_{n,k}(V)-\bar R_{t_{n,k}}(V)|&=\frac1k\Big|\sum_{i=1}^n\|\mathbf{\Pi}_V^\perp\Theta_i\|^2(\mathbf{1}\{\|X_i\|>\hat t_{n,k}\}-\mathbf{1}\{\|X_i\|>t_{n,k}\})\Big|\\&\le\frac1k\Big|\sum_{i=1}^n\mathbf{1}\{\|X_i\|>t_{n,k}\}-k\Big|.\end{aligned}$$

By Bernstein's inequality ([24, Theorem 2.7]), it follows that

$$\mathbb{P}\Big\{\sup_{V\in\mathcal{V}_p}|\hat R_{n,k}(V)-\bar R_{t_{n,k}}(V)|\ge v\Big\}\le 2\exp\Big(-\frac{(kv)^2}{2(k(1-k/n)+kv/3)}\Big)$$

and hence the assertion. □

We next prove concentration inequalities for $\varphi_t^\pm(X_1,\dots,X_n)$ defined in (3.4), using a version of the bounded difference inequality by [24, Theorem 3.8], which we recall for convenience. For brevity's sake, in what follows we use the notation $x_{i:j}:=(x_i,\dots,x_j)$ for a subvector of $(x_1,\dots,x_n)$.

Theorem 5.6. *Let $X_{1:n} = (X_1, \ldots, X_n)$ be an i.i.d. sample taking its values in some space E and $\varphi : E^n \to \mathbb{R}$ be any measurable function. Consider the positive deviation functions defined on E^m, for any $1 \le m \le n$, by*

$$h_m(x_{1:m}) = \mathbb{E}\left(\varphi(x_{1:m}, X_{m+1:n}) - \varphi(x_{1:m-1}, X_{m:n})\right).$$

Denote their maximum by

$$\mathrm{maxdev}^+ = \max_{1\le m\le n} \sup_{x_{1:m}\in E^m} h_m(x_{1:m}), \tag{5.8}$$

and the maximal summed variance by

$$\hat{v} = \sup_{x_{1:n}\in E^n} \sum_{m=1}^{n} \mathbf{Var}\, h_m(x_{1:m-1}, X_m). \tag{5.9}$$

If both maxdev^+ *and* $\hat{v}$ *are finite then, for all* $u \ge 0$,

$$\mathbb{P}\big\{\varphi(X_{1:n}) - \mathbb{E}\,\varphi(X_{1:n}) \ge u\big\} \le \exp\left(-\frac{u^2}{2(\hat{v} + \mathrm{maxdev}^+ u/3)}\right).$$

Lemma 5.7. *Under (3.1), one has, for all* $u > 0$,

$$\mathbb{P}\Big\{\varphi_t^{\pm}(X_1, \ldots, X_n) \ge \mathbb{E}\,\varphi_t^{\pm}(X_1, \ldots, X_n) + u\Big\} \le \exp\Big(-\frac{nu^2}{2(\pi_t(1+\pi_t) + u/3)}\Big).$$

Proof. The assertion follows immediately from Theorem 5.6 applied to $\varphi_t^{\pm}$ and the following bounds:

$$\begin{aligned} h_m(x_{1:m}) &= \mathbb{E}\left(\varphi_t^{\pm}(x_{1:m}, X_{m+1:n}) - \varphi_t^{\pm}(x_{1:m-1}, X_{m:n})\right) \\ &\le \frac{1}{n}\,\mathbb{E}\left(\sup_{V\in\mathcal{V}_p} \left|\|\mathbf{\Pi}_V^{\perp}\theta_t(X_m)\|^2 - \|\mathbf{\Pi}_V^{\perp}\theta_t(x_m)\|^2\right|\right) \\ &\le \frac{1}{n}\,\mathbb{E}\left(\mathbf{1}\{\|X_m\| > t\| \text{ or } \|x_m\| > t\}\right) \\ &= \frac{1}{n}(\pi_t \vee \mathbf{1}\{\|x_m\| > t\}) \\ &\le \frac{1}{n} \end{aligned}$$

and

$$\begin{aligned} \sum_{m=1}^{n} \mathbf{Var}\, h_m(x_{1:m-1}, X_m) &\le \sum_{m=1}^{n} \mathbb{E}\, h_m^2(x_{1:m-1}, X_m) \\ &\le \frac{1}{n}\,\mathbb{E}(\pi_t \vee \mathbf{1}\{\|X\| > t\})^2 \\ &= \frac{\pi_t^2(1-\pi_t) + \pi_t}{n} \\ &\le \frac{\pi_t(1+\pi_t)}{n}. \end{aligned}$$

□

Next, the expectations $\mathbb{E}\,\varphi_t^{\pm}(X_1,\dots,X_n)$ are bounded using arguments from [4].

Lemma 5.8.

$$\mathbb{E}\,\varphi_t^{\pm}(X_1,\dots,X_n) \le \Big[\frac{p\wedge(d-p)}{n}\pi_t S_t\Big]^{1/2}$$

with S_t defined in Theorem 3.1, provided condition (3.1) is satisfied.

Proof. Since, by (5.6), $\|\mathbf{\Pi}_W x\|^2 = \langle \mathbf{\Pi}_W x, x\rangle = \langle \mathbf{\Pi}_W, xx^\top\rangle_{HS}$ for any linear subspace W and any $x\in\mathbb{R}^d$, using the bilinearity of the inner product and the Cauchy-Schwarz inequality in the Hilbert-Schmidt space, we obtain

$$\begin{aligned}\pm(P_n-P)(\|\mathbf{\Pi}_V^{\perp}\theta_t\|^2) &= \Big\langle \mathbf{\Pi}_V^{\perp}, \pm(P_n-P)(\theta_t\theta_t^\top)\Big\rangle_{HS}\\ &\le \|\mathbf{\Pi}_V^{\perp}\|_{HS}\|(P_n-P)(\theta_t\theta_t^\top)\|_{HS}.\end{aligned}$$

Using (5.5) and taking the supremum over all $V\in\mathcal{V}_p$ and the expectation, one arrives at

$$\mathbb{E}\,\varphi_t^{\pm}(X_1,\dots,X_n) \le \sqrt{d-p}\,\mathbb{E}\,\|(P_n-P)(\theta_t\theta_t^\top)\|_{HS}. \tag{5.10}$$

One the other hand, by first rewriting $\|\mathbf{\Pi}_V^{\perp}\theta_t\|^2 = \|\theta_t\|^2 - \|\mathbf{\Pi}_V\theta_t\|^2$, analogously one obtains

$$\begin{aligned}\mathbb{E}\,\varphi_t^{\pm}(X_1,\dots,X_n) &= \mathbb{E}\Big(\sup_{V\in\mathcal{V}_p}\pm(P_n-P)(\|\mathbf{\Pi}_V^{\perp}\theta_t\|^2)\Big)\\ &= \mathbb{E}\,\big((P_n-P)\|\theta_t\|^2\big) + \mathbb{E}\Big(\sup_{V\in\mathcal{V}_p}\mp(P_n-P)(\|\mathbf{\Pi}_V\theta_t\|^2)\Big)\\ &\le 0 + \sup_{V\in\mathcal{V}_p}\|\mathbf{\Pi}_V\|_{HS}\,\mathbb{E}\,\|(P_n-P)(\theta_t\theta_t^\top)\|_{HS}\\ &\le \sqrt{p}\,\mathbb{E}\,\|(P_n-P)(\theta_t\theta_t^\top)\|_{HS}.\end{aligned} \tag{5.11}$$

Now, by the Jensen's inequality and (5.7),

$$\begin{aligned}\mathbb{E}\,\|(P_n-P)(\theta_t\theta_t^\top)\|_{HS} &\le \big(\mathbb{E}\,\|(P_n-P)(\theta_t\theta_t^\top)\|_{HS}^2\big)^{1/2}\\ &= \Big(\mathbb{E}\,\Big\|\frac{1}{n}\sum_{i=1}^n\big(\Theta_{i,t}\Theta_{i,t}^\top - \mathbb{E}(\Theta_t\Theta_t^\top)\big)\Big\|_{HS}^2\Big)^{1/2}\\ &= \Big(\frac{1}{n}\,\mathbb{E}\,\|\Theta_t\Theta_t^\top - \mathbb{E}(\Theta_t\Theta_t^\top)\|_{HS}^2\Big)^{1/2}.\end{aligned}$$

Combining this with (5.10) and (5.11), we arrive at

$$\mathbb{E}\,\varphi_t^{\pm}(X_1,\dots,X_n) \le \Big[\frac{p\wedge(d-p)}{n}\,\mathbb{E}\,\|\Theta_t\Theta_t^\top - \mathbb{E}(\Theta_t\Theta_t^\top)\|_{HS}^2\Big]^{1/2}. \tag{5.12}$$

It remains to show that $\mathbb{E}\,\|\Theta_t\Theta_t^\top - \mathbb{E}(\Theta_t\Theta_t^\top)\|_{HS}^2 = \pi_t\big(\mathbb{E}_t\,\|\Theta\|^4 - \pi_t\,\mathrm{tr}(\Sigma_t^2)\big)$. From the representation of the Hilbert Schmidt norm by the trace operator and the linearity of the latter, one may conclude by direct calculations that

$$\mathbb{E}\,\|\Theta_t\Theta_t^\top - \mathbb{E}(\Theta_t\Theta_t^\top)\|_{HS}^2 = \mathrm{tr}\,\big(\mathbb{E}(\Theta_t\Theta_t^\top - \mathbb{E}\,\Theta_t\Theta_t^\top)^2\big)$$

$$
\begin{aligned}
&= \operatorname{tr}\left(\mathbb{E}(\Theta_t\Theta_t^\top)^2\right) - \operatorname{tr}\left((\mathbb{E}(\Theta_t\Theta_t^\top))^2\right)\\
&= \operatorname{tr}\left(\pi_t\,\mathbb{E}_t(\Theta\Theta^\top)^2\right) - \operatorname{tr}\left((\pi_t\,\mathbb{E}_t\,\Theta\Theta^\top)^2\right)\\
&= \pi_t\,\mathbb{E}_t\operatorname{tr}\left((\Theta\Theta^\top)^2\right) - \pi_t^2\operatorname{tr}(\Sigma_t^2).
\end{aligned}
$$

Hence the assertion follows from

$$
\operatorname{tr}\left((\Theta\Theta^\top)^2\right) = \sum_{j=1}^{d}\|\Theta\Theta^\top e_j\|^2 = \sum_{j=1}^{d}\sum_{l=1}^{d}\left(\Theta^{(l)}\Theta^{(j)}\right)^2 = \|\Theta\|^4
$$

with e_j denoting the jth unit vector. □

Proof of Theorem 3.1. With $\bar{R}_t(V)$ defined in (3.2), we have

$$
\begin{aligned}
&\sup_{V\in\mathcal{V}_p}|\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)|\\
&\le \sup_{V\in\mathcal{V}_p}|\hat{R}_{n,k}(V) - \bar{R}_{t_{n,k}}(V)| + \sup_{V\in\mathcal{V}_p}|\bar{R}_{t_{n,k}}(V) - R_{t_{n,k}}(V)|.
\end{aligned}
$$

Recall from (3.3) that the second term can be written as

$$
\sup_{V\in\mathcal{V}_p}|\bar{R}_{t_{n,k}}(V) - R_{t_{n,k}}(V)| = \frac{n}{k}\max(\varphi^+_{t_{n,k}}(X_{1:n}), \varphi^-_{t_{n,k}}(X_{1:n})).
$$

Hence, Lemma 5.7, Lemma 5.8 and $\pi_{t_{n,k}} = k/n$ immediately yield

$$
\begin{aligned}
\mathbb{P}\Big\{\sup_{V\in\mathcal{V}_p}|\bar{R}_{t_{n,k}}(V) - R_{t_{n,k}}(V)| \ge \Big[\frac{p\wedge(d-p)}{k}S_{t_{n,k}}\Big]^{1/2} + u\Big\}&\\
\le 2\exp\Big(-\frac{n(uk/n)^2}{2(k/n(1+k/n)+uk/(3n))}\Big)&\\
= 2\exp\Big(-\frac{ku^2}{2(1+k/n+u/3)}\Big).&
\end{aligned}
$$

Combine this with the bound on the first term given by Lemma 5.5 to conclude the proof of the first assertion.

Check that for

$$
\begin{aligned}
u &:= \frac{\log(4/\delta)}{3k} + \Big[\Big(\frac{\log(4/\delta)}{3k}\Big)^2 + \frac{2}{k}(1+k/n)\log(4/\delta)\Big]^{1/2}\\
v &:= \frac{\log(4/\delta)}{3k} + \Big[\Big(\frac{\log(4/\delta)}{3k}\Big)^2 + \frac{2}{k}\log(4/\delta)\Big]^{1/2}
\end{aligned}
$$

both exponential expressions on the right hand side of (3.5) equal $\delta/4$, and so the upper bound equals δ. Hence the remaining assertions follow from $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$. □

Proof of Theorem 3.3. Define *i.i.d.* random vectors Z_i whose distribution equals the conditional distribution of Θ given $\|X\| > t$. Recall that $\Theta_{(i)} := \theta(X_{(i)})$ where $X_{(i)}$ is the vector X_j with the ith largest norm among $X_1,\ldots,X_n$. Then,

conditionally on $N_t = \ell$, the joint distribution of the empirical risk $\hat{R}_t(V)$ and $\Theta_{(1)}, \dots \Theta_{(\ell)}$ equals the joint distribution of $\ell^{-1} \sum_{i=1}^{\ell} \|\mathbf{\Pi}_V^{\perp} Z_i\|^2$ and the order statistics of $Z_1, \dots, Z_\ell$. Therefore, the proof of Theorem 3.1 of [4] (with $M = 1$ and $L = 2$) combined with arguments given in the proof of Lemma 5.8 show that with probability at least $1 - 2\exp\big(-2\ell u^2\big) - \exp\big(-\lfloor \ell/2 \rfloor v^2/2\big)$ conditionally to $N_t = \ell$,

$$\sup_{V \in \mathcal{V}_p} |\hat{R}_t(V) - R_t(V)| < \Big[\frac{p \wedge (d-p)}{2\ell}\Big(\frac{1}{\ell(\ell-1)} \sum_{i,j=1}^{\ell} \|\Theta_{(i)}\Theta_{(i)}^\top - \Theta_{(j)}\Theta_{(j)}^\top\|_{HS}^2 + 2v\Big)\Big]^{1/2} + u \qquad (5.13)$$

Since the proof of Theorem 3.1 of [4] is quite tersely formulated in a more abstract setting and it contains a minor inaccuracy, for convenience we give more details of the proof of (5.13) in the Appendix.

Similarly as in the proof of Lemma 5.8, the first assertion thus follows from

$$\begin{aligned}
&\sum_{i,j=1}^{\ell} \|\Theta_{(i)}\Theta_{(i)}^\top - \Theta_{(j)}\Theta_{(j)}^\top\|_{HS}^2 \\
&= 2\ell \sum_{i=1}^{\ell} \|\Theta_{(i)}\Theta_{(i)}^\top\|_{HS}^2 - 2 \sum_{i,j=1}^{\ell} \langle \Theta_{(i)}\Theta_{(i)}^\top, \Theta_{(j)}\Theta_{(j)}^\top \rangle_{HS} \\
&= 2\ell \sum_{i=1}^{\ell} \|\Theta_{(i)}\|^4 - 2\Big\| \sum_{i=1}^{\ell} \Theta_{(i)}\Theta_{(i)}^\top \Big\|_{HS}^2 \\
&= 2\ell^2 \Big(\frac{1}{\ell} \sum_{i=1}^{\ell} \|\Theta_{(i)}\|^4 - \mathrm{tr}\Big(\Big(\frac{1}{\ell} \sum_{i=1}^{\ell} \Theta_{(i)}\Theta_{(i)}^\top \Big)^2 \Big) \Big) \\
&= 2\ell^2 \Big(\frac{1}{\ell} \sum_{i=1}^{n} \|\Theta_{i,t}\|^4 - \mathrm{tr}\Big(\Big(\frac{1}{\ell} \sum_{i=1}^{n} \Theta_{i,t}\Theta_{i,t}^\top \Big)^2 \Big) \Big)
\end{aligned}$$

where in the last step we have used that, on $\{N_t = \ell\}$, the set of non-vanishing vectors $\Theta_{i,t}$ equals the set of non-vanishing random vectors $\Theta_{(i)}$.

The remaining assertions are now obvious. □

Proof of Remark 3.5. The (modified) proof of Theorem 3.1 of [4] shows that

$$\mathbb{P}\Big(\sup_{V \in \mathcal{V}_p} |\hat{R}_t(V) - R_t(V)| \geq \Big[\frac{p \wedge (d-p)}{\ell} S_t^*\Big]^{1/2} + u \,\Big|\, N_t = \ell \Big) \leq 2\exp\big(-2\ell u^2\big)$$

with $S_t^* := \mathbb{E}_t \|\Theta\|^4 - \mathrm{tr}(\Sigma_t^2)$ (cf. (A.1) and (A.2)). Observe that $\bar{R}_t(V) = N_t \hat{R}_t(V)/(n\pi_t)$. On the set $M_t(v) := \{|N_t - n\pi_t| \leq n\pi_t v\}$, one thus has

$$\sup_{V \in \mathcal{V}_p} |\bar{R}_t(V) - R_t(V)| \leq \frac{N_t}{n\pi_t} \sup_{V \in \mathcal{V}_p} |\hat{R}_t(V) - R_t(V)| + v,$$

since $R_t(V) \le 1$. Moreover, for $t = t_{n,k}$, it is shown in the proof of Theorem 3.1 that $\sup_{V\in\mathcal{V}_p} |\hat{R}_{n,k}(V) - \bar{R}_{t_{n,k}}(V)| \le v$ on the set $M_{t_{n,k}} = \{N_{t_{n,k}} \in [k(1-v), k(1+v)]\}$ and that $\mathbb{P}(M^c_{t_{n,k}}) \le 2\exp\big(-kv^2/(2(1+v/3))\big)$. Hence,

$$
\begin{aligned}
&\mathbb{P}\Big\{ \sup_{V\in\mathcal{V}_p} |\hat{R}_{n,k}(V) - R_{t_{n,k}}(V)| \ge \Big[(1+v)\frac{p\wedge(d-p)}{k} S^*_{t_{n,k}}\Big]^{1/2} + u + 2v\Big\} \\
&\le \mathbb{P}\Big(M_{t_{n,k}} \cap \\
&\qquad \Big\{ \sup_{V\in\mathcal{V}_p} |\bar{R}_{t_{n,k}}(V) - R_{t_{n,k}}(V)| \ge \Big[(1+v)\frac{p\wedge(d-p)}{k} S^*_{t_{n,k}}\Big]^{1/2} + u + v\Big\}\Big) \\
&\quad + \mathbb{P}(M^c_{t_{n,k}}) \\
&\le \mathbb{P}\Big(M_{t_{n,k}} \cap \Big\{ \sup_{V\in\mathcal{V}_p} |\hat{R}_{t_{n,k}}(V) - R_{t_{n,k}}(V)| \ge \Big[\frac{p\wedge(d-p)}{N_{t_{n,k}}} S^*_{t_{n,k}}\Big]^{1/2} + \frac{ku}{N_{t_{n,k}}}\Big\}\Big) \\
&\quad + \mathbb{P}(M^c_{t_{n,k}}) \\
&\le \sum_{\ell=\lceil k(1-v)\rceil}^{\lfloor k(1+v)\rfloor} 2\exp\big(-2\ell(ku/\ell)^2\big)\mathbb{P}\{N_{t_{n,k}} = \ell\} + \mathbb{P}(M^c_{t_{n,k}}) \\
&\le 2\exp\Big(-\frac{2ku^2}{1+v}\Big) + 2\exp\Big(-\frac{kv^2}{2(1+v/3)}\Big).
\end{aligned}
$$

□

Appendix: Details of the proof of (5.13)

Recall that Z_i are iid random variables whose distribution equals the conditional distribution of Θ given $\|X\| > t$. Let

$$
\phi^{\pm}(z_1,\dots,z_\ell) := \sup_{V\in\mathcal{V}_p} \pm\Big(\frac{1}{\ell}\sum_{i=1}^{\ell} \|\mathbf{\Pi}^{\perp}_V z_i\|^2 - P\|\mathbf{\Pi}^{\perp}_V Z_1\|^2\Big).
$$

First note that

$$
\big|\phi^{\pm}(z_{1:\ell}) - \phi^{\pm}(z_{1:i-1}, \tilde{z}_i, z_{i+1:\ell})\big| \le \sup_{V\in\mathcal{V}_p} \frac{1}{\ell}\big|\|\mathbf{\Pi}^{\perp}_V z_i\|^2 - \|\mathbf{\Pi}^{\perp}_V \tilde{z}_i\|^2\big| \le \frac{1}{\ell}
$$

for all $z, \tilde{z} \in B_1(0)$. Thus a simple version of the bounded difference inequality (see, *e.g.*, Theorem 3.1 of [24]) gives

$$
\mathbb{P}\big\{\phi^{\pm}(Z_{1:\ell}) - \mathbb{E}\,\phi^{\pm}(Z_{1:\ell}) \ge u\big\} \le \exp(-2\ell u^2), \quad \forall u > 0. \tag{A.1}
$$

Exactly in the same way as in the proof of Lemma 5.8 (cf. (5.12)), one obtains

$$
\begin{aligned}
\mathbb{E}\,\phi^{\pm}(Z_{1:\ell}) &\le \Big[\frac{p\wedge(d-p)}{\ell}\,\mathbb{E}\,\|ZZ^\top - \mathbb{E}\,ZZ^\top\|^2_{HS}\Big]^{1/2} \\
&= \Big[\frac{p\wedge(d-p)}{\ell}\big(\mathbb{E}\,\|ZZ^\top\|^2_{HS} - \|\,\mathbb{E}\,ZZ^\top\|^2_{HS}\big)\Big]^{1/2}.
\end{aligned} \tag{A.2}
$$

Let $\tilde{Z}$ be an independent copy of Z. Then

$$\mathbb{E}\,\|ZZ^\top - \tilde{Z}\tilde{Z}^\top\|_{HS}^2 = 2\,\mathbb{E}\,\|ZZ^\top\|_{HS}^2 - 2\,\mathbb{E}\langle ZZ^\top, \tilde{Z}\tilde{Z}^\top\rangle_{HS}$$

with

$$\begin{aligned}\mathbb{E}\langle ZZ^\top, \tilde{Z}\tilde{Z}^\top\rangle_{HS} &= \mathbb{E}\left(\mathbb{E}(\langle ZZ^\top, \tilde{Z}\tilde{Z}^\top\rangle_{HS} \mid Z)\right)\\ &= \mathbb{E}\langle ZZ^\top, \mathbb{E}\,\tilde{Z}\tilde{Z}^\top\rangle_{HS} = \|\,\mathbb{E}\,ZZ^\top\|_{HS}^2.\end{aligned}$$

To sum up, so far we have shown that, for all $u \geq 0$,

$$\mathbb{P}\Big\{\phi^{\pm}(Z_{1:\ell}) \geq \Big[\frac{p \wedge (d-p)}{2\ell}\,\mathbb{E}\,\|ZZ^\top - \tilde{Z}\tilde{Z}^\top\|_{HS}^2\Big]^{1/2} + u\Big\} \leq \exp(-2\ell u^2).$$

Next consider the U-statistic $U := (\ell(\ell-1))^{-1}\sum_{i,j=1}^{\ell} g(Z_i, Z_j)$ with

$$g(z,\tilde{z}) := \|zz^\top - \tilde{z}\tilde{z}^\top\|_{HS}^2 \leq (\|zz^\top\|_{HS} + \|\tilde{z}\tilde{z}^\top\|_{HS})^2 \leq 4.$$

By equation (5.7) of [18], one has

$$\mathbb{P}\{U - \mathbb{E}\,U \geq 2v\} \leq \exp\big(-2\lfloor \ell/2\rfloor (2v)^2/16\big) = \exp\big(-\lfloor \ell/2\rfloor v^2/2\big), \quad \forall v \geq 0,$$

with $\mathbb{E}\,U = \mathbb{E}\,\|ZZ^\top - \tilde{Z}\tilde{Z}^\top\|_{HS}^2$. Hence,

$$\begin{aligned}&\mathbb{P}\Big\{\max\big(\phi^+(Z_{1:\ell}), \phi^-(Z_{1:\ell})\big) \geq\\ &\qquad \Big[\frac{p\wedge(d-p)}{2\ell}\Big(\frac{1}{\ell(\ell-1)}\sum_{i,j=1}^{\ell}\|Z_iZ_i^\top - Z_jZ_j^\top\|_{HS}^2 + 2v\Big)\Big]^{1/2} + u\Big\}\\ &\leq 2\exp(-2\ell u^2) + \exp\big(-\lfloor \ell/2\rfloor v^2/2\big), \quad \forall u, v \geq 0.\end{aligned}$$

This, in turn, is equivalent to (5.13), because the joint distribution of $\max\big(\phi^+(Z_{1:\ell}),\ \phi^-(Z_{1:\ell})\big)$ and $\sum_{i,j=1}^{\ell}\|Z_iZ_i^\top - Z_jZ_j^\top\|_{HS}^2$ is the same as the joint conditional distribution of $\sup_{V\in\mathcal{V}_p}|\hat{R}_t(V) - R_t(V)|$ and $\sum_{i,j=1}^{\ell}\|\Theta_{(i)}\Theta_{(i)}^\top - \Theta_{(j)}\Theta_{(j)}^\top\|_{HS}^2$, given $N_t = \ell$.

Acknowledgements

Holger Drees was partly supported by DFG grant DR271/6-2. Anne Sabourin was partly supported by the industrial chairs 'Stress testing' from École Polytechnique and BNP Paribas and DSAIDS from Télécom Paris. Questions raised by an anonymous referee have lead to an improvement of the presentation.

References

[1] Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148. MR0145620

[2] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons. MR2108013
[3] Bingham, N., Goldie, C., and Teugels, J. (1987). *Regular Variation*. Cambridge University Press. MR0898871
[4] Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294.
[5] Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9(1):383–418. MR3323204
[6] Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer.
[7] Chiapino, M., Sabourin, A., and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222. MR3949044
[8] Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604. MR3992391
[9] Einmahl, J. H., de Haan, L., and Piterbarg, V. I. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423. MR1873336
[10] Engelke, S. and Hitz, A. S. (2018). Graphical models for extremes. *arXiv preprint* arXiv:1812.01734. MR4136498
[11] Engelke, S. and Ivanovs, J. (2020). Sparse structures for multivariate extremes. *arXiv preprint* arXiv:2004.12182.
[12] Ferreira, A. and de Haan, L. (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717–1737. MR3263087
[13] Fougères, A.-L., de Haan, L., and Mercadier, C. (2015). Bias correction in multivariate extremes. *The Annals of Statistics*, 43(2):903–934. MR3325714
[14] Gardes, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1):57–95. MR3764611
[15] Genest, C. and Segers, J. (2009). Rank-based inference for bivariate extreme-value copulas. *The Annals of Statistics*, 37(5B):2990–3022. MR2541453
[16] Goix, N., Sabourin, A., and Clémençon, S. (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *AISTATS*, pages 75–83. MR3698112
[17] 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. MR3698112
[18] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30. MR0144363
[19] Hult, H. and Lindskog, F. (2006). Regular variation for measures on metric spaces. *Publ. Inst. Math.(Beograd) (NS)*, 80(94):121–140. MR2281910

[20] Janßen, A. and Wan, P. (2020). k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233. MR4071364

[21] Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167. MR1781185

[22] Koltchinskii, V. and Lounici, K. (2017). New asymptotic results in principal component analysis. *Sankhya A*, 79(2):254–297. MR3707422

[23] Lindskog, F., Resnick, S. I., and Roy, J. (2014). Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probability Surveys*, 11:270–314. MR3271332

[24] McDiarmid, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer. MR1678578

[25] Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489):263–277. MR2757202

[26] Reiß, M. and Wahl, M. (2020). Nonasymptotic upper bounds for the reconstruction error of pca. *Annals of Statistics*, 48(2):1098–1123. MR4102689

[27] Resnick, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media. MR2271424

[28] Resnick, S. I. (2013). *Extreme Values, Regular Variation and Point Processes*. Springer. MR2364939

[29] Rootzén, H., Segers, J., and Wadsworth, J. L. (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131. MR3768756

[30] Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes*, 5(1):33–44. MR1947786

[31] Seber, G. A. F. (1984). *Multivariate Observations*. Wiley. MR0746474

[32] Segers, J. (2012). Max-stable models for multivariate extremes. *REVSTAT*, 10:61–82. MR2912371

[33] 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. *Information Theory, IEEE Transactions on*, 51(7):2510–2522. MR2246374

[34] Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2019). Determining the dependence structure of multivariate extremes. *arXiv preprint* arXiv:1809.01606. MR4138974

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

[36] Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press. MR1652247

[37] Zwald, L. and Blanchard, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems*, pages 1649–1656.