The Annals of Statistics
2020, Vol. 48, No. 2, 1098–1123
https://doi.org/10.1214/19-AOS1839

NONASYMPTOTIC UPPER BOUNDS FOR THE RECONSTRUCTION ERROR OF PCA

BY MARKUS REISS* AND MARTIN WAHL**

*Institut für Mathematik, Humboldt-Universität zu Berlin, *mreiss@math.hu-berlin.de; **martin.wahl@math.hu-berlin.de*

We analyse the reconstruction error of principal component analysis (PCA) and prove nonasymptotic upper bounds for the corresponding excess risk. These bounds unify and improve existing upper bounds from the literature. In particular, they give oracle inequalities under mild eigenvalue conditions. The bounds reveal that the excess risk differs significantly from usually considered subspace distances based on canonical angles. Our approach relies on the analysis of empirical spectral projectors combined with concentration inequalities for weighted empirical covariance operators and empirical eigenvalues.

1. Introduction. Principal component analysis (PCA) and variants like functional PCA or kernel PCA are standard tools in high-dimensional statistics and unsupervised learning; see, for example, Jolliffe [16], Horváth and Kokoszka [12] and Schölkopf and Smola [29] for an overview. Usually, they are employed as a first step to reduce the high dimensionality of the data before methods for the specific task come into play. The basic motivation for this work is that the understanding of the error incurred by PCA in high dimensions is so far limited. In fact, Blanchard, Bousquet and Zwald [6] exhibit upper bounds for the excess risk of the reconstruction error which give different rates in sample size and dimensionality depending on spectral properties of the covariance operator, and thus exhibit complex facets of this classical statistical method. By combining spectral projector calculus with concentration inequalities, we are able to give tight bounds for the excess risk which clarify the underlying error structure. This gives rise to oracle risk bounds which in wide generality prove that the error due to projecting on empirical principal components is negligible compared to the error due to optimal dimension reduction via the population version of PCA.

We include functional PCA and kernel PCA in the standard multivariate PCA setting by allowing for general Hilbert spaces $\mathcal{H}$. PCA is commonly derived by minimising the reconstruction error $\mathbb{E}[\|X - PX\|^2]$ over all orthogonal projections P of rank d, where X is an $\mathcal{H}$-valued random variable and d is a given dimension. Replacing the population covariance Σ by an empirical covariance $\hat{\Sigma}$, PCA computes the orthogonal projection $\hat{P}_{\leq d}$ onto the eigenspace of the d leading eigenvalues of $\hat{\Sigma}$. Put differently, $\hat{P}_{\leq d}$ minimises the empirical reconstruction error and it is natural to measure its performance by the excess risk $\mathcal{E}_d^{\mathrm{PCA}}$, that is, by the difference between the reconstruction errors of $\hat{P}_{\leq d}$ and the overall minimiser $P_{\leq d}$. It is easy to see that $\mathcal{E}_d^{\mathrm{PCA}} = \langle \Sigma, P_{\leq d} - \hat{P}_{\leq d} \rangle$ holds with respect to the Hilbert–Schmidt scalar product.

Comparing the excess risk $\mathcal{E}_d^{\mathrm{PCA}}$ to the Hilbert–Schmidt distance $\|\hat{P}_{\leq d} - P_{\leq d}\|_2$, which is up to a constant equal to the l^2-norm of the sines of the canonical angles between the corresponding subspaces, the main difference is that $\mathcal{E}_d^{\mathrm{PCA}}$ remains small if $\hat{P}_{\leq d}$ projects into

Received March 2018; revised November 2018.
MSC2010 subject classifications. Primary 62H25; secondary 15A42, 60F10.
Key words and phrases. Principal component analysis, reconstruction error, excess risk, spectral projectors, concentration inequalities.

eigenspaces with eigenvalues that are not much smaller than the d largest ones. In the extreme case $\lambda_d = \lambda_{d+1}$, where the dth and $(d+1)$st largest eigenvalues coincide, the Hilbert–Schmidt distance is not even uniquely defined. Statistically, the reconstruction error is not only the basis for the very definition of PCA, but it is also more adequate for many tasks like reconstruction and prediction than the Hilbert–Schmidt distance. A typical example is given by the prediction error of principal component regression, for which Wahl [35] establishes a clear connection with the excess risk of PCA. Mathematically, an arbitrarily small spectral gap $\lambda_d - \lambda_{d+1}$ requires new techniques because spectral perturbation results deteriorate as the spectral gap shrinks. Our aim is to treat even the isotropic case $\Sigma = \sigma^2 I$, where the covariance is a multiple of the identity matrix and $\mathcal{E}_d^{\mathrm{PCA}} = 0$ holds.

Classical results for PCA provide limit theorems for the empirical eigenvalues and eigenvectors when the sample size n tends to infinity; see, for example, Anderson [2] and Dauxois, Pousse and Romain [9]. For the Hilbert–Schmidt distance, the most well-known result is the Davis–Kahan $\sin\Theta$ theorem [10], which gives an upper bound in terms of the eigenvalue separation and the Hilbert–Schmidt norm of $\hat{\Sigma} - \Sigma$. In many cases, more precise bounds can be derived using higher-order spectral perturbation results. Nadler [27] obtains nonasymptotic bounds for the spiked covariance model and studies phase transitions when dimension and sample size tend to infinity simultaneously. Mas and Ruymgaart [25] and Jirak [14] ask for near-optimal bounds for functional PCA with exponential or polynomial spectral decay. Koltchinskii and Lounici [18–21] derive tight concentration bounds for the operator norm of $\hat{\Sigma} - \Sigma$ and study empirical spectral projectors in the so-called effective rank setting.

Bounds for the reconstruction error using the theory of empirical risk minimisation (ERM) are derived by Shawe-Taylor et al. [30, 31] and Blanchard, Bousquet and Zwald [6]. While [31] only establishes a slow $n^{-1/2}$-rate, in [6] the existence of faster rates, difficult to quantify explicitly, is discovered. We take up the ERM approach in Section 2.2 below and establish by a simple recursion argument upper bounds, based on an interplay between a slow $n^{-1/2}$-rate and a fast n^{-1}-rate. These bounds clarify and partly improve the existing theory, while the proofs are short and transparent such that they have a value on their own.

Yet, we observe that the basic inequality of ERM prevents us from deriving good bounds in basic settings like isotropic covariance. In order to obtain tight bounds in more generality for $\mathcal{E}_d^{\mathrm{PCA}}$ in Section 2.3, we employ a more sophisticated recursion argument in combination with concentration inequalities for weighted empirical covariance operators and empirical eigenvalues. This is achieved by an algebraic projector-based calculus that allows us to take advantage of the presence of the true covariance Σ in the expression for the excess risk and to avoid difficulties arising from a straightforward application of standard perturbation theory; compare Remarks 3.3, 3.4 and 3.15 for more details. Considering standard examples in high-dimensional statistics and functional data analysis like spiked covariance models and exponential or polynomial eigenvalue decay, Section 2.4 shows how the general bounds apply and that existing bounds in the literature can be rediscovered and in some important aspects improved. The overall finding is that in all of these cases a tight oracle inequality holds.

Finally, we discuss in Section 2.5 how our results can be transferred to the subspace distance and, for instance, how the projector calculus yields the Davis–Kahan $\sin\Theta$ theorem and other spectral perturbation results in a straightforward manner. Moreover, a CLT for the excess risk is presented for fixed dimensions, acting as a benchmark for the high-dimensional results and revealing a surprising inhomogeneity of the excess risk with respect to the eigenvalue spacings. We also state the concentration inequalities for individual empirical eigenvalues that might be of independent interest. Section 3 supplies the main tools from projector-based calculus, ω-wise error decompositions and concentration inequalities. Section 4 is devoted to the proofs. The Supplementary Material [28] contains additional proofs and extensions via linear expansions.

2. Main results.

2.1. *The reconstruction error of PCA.* Let X be a centered random variable taking values in a separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ of dimension $p \in \mathbb{N}\cup\{+\infty\}$ and let $\|\cdot\|$ denote the norm on $\mathcal{H}$ defined by $\|u\| = \sqrt{\langle u, u\rangle}$.

ASSUMPTION 2.1. Suppose that X is sub-Gaussian, meaning that $\mathbb{E}[\|X\|^2]$ is finite and that there is a constant C_1 with

$$\big\|\langle X, u\rangle\big\|_{\psi_2} := \sup_{k\geq 1} k^{-1/2}\mathbb{E}\big[\big|\langle X, u\rangle\big|^k\big]^{1/k} \leq C_1\mathbb{E}\big[\langle X, u\rangle^2\big]^{1/2}$$

for all $u \in \mathcal{H}$.

If X is Gaussian, then it is easy to see that Assumption 2.1 holds with $C_1 = 1$ (cf. [33], the first formula in equation (5.6)).

The covariance operator of X is denoted by

$$\Sigma = \mathbb{E}[X \otimes X].$$

By the spectral theorem, there exists a sequence $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ of positive eigenvalues (which is either finite or converges to zero) together with an orthonormal system of eigenvectors $u_1, u_2, \ldots$ such that Σ has the spectral representation

$$\Sigma = \sum_{j\geq 1} \lambda_j P_j,$$

with rank-one projectors $P_j = u_j \otimes u_j$, where $(u \otimes v)x = \langle v, x\rangle u$, $x \in \mathcal{H}$. Note that the choice of u_j and P_j is nonunique in case of multiple eigenvalues λ_j.

Without loss of generality, we shall assume that the eigenvectors $u_1, u_2, \ldots$ form an orthonormal basis of $\mathcal{H}$ such that $\sum_{j\geq 1} P_j = I$. We write

$$P_{\leq d} = \sum_{j\leq d} P_j, \qquad P_{>d} = I - P_{\leq d} = \sum_{k>d} P_k$$

for the orthogonal projections onto the linear subspace spanned by the first d eigenvectors of Σ, and onto its orthogonal complement.

Let $X_1, \ldots, X_n$ be n independent copies of X and let

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i$$

be the sample covariance. Again, there exists a sequence $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq 0$ of eigenvalues together with an orthonormal basis of eigenvectors $\hat{u}_1, \hat{u}_2, \ldots$ such that we can write

$$\hat{\Sigma} = \sum_{j\geq 1} \hat{\lambda}_j \hat{P}_j \quad \text{with } \hat{P}_j = \hat{u}_j \otimes \hat{u}_j$$

and

$$\hat{P}_{\leq d} = \sum_{j\leq d} \hat{P}_j, \qquad \hat{P}_{>d} = I - \hat{P}_{\leq d} = \sum_{k>d} \hat{P}_k.$$

For linear operators $S, T : \mathcal{H} \to \mathcal{H}$, we make use of trace and adjoint $\operatorname{tr}(S)$, S^* to define the Hilbert–Schmidt or Frobenius norm and scalar product

$$\|S\|_2^2 = \operatorname{tr}\big(S^*S\big), \qquad \langle S, T\rangle = \operatorname{tr}\big(S^*T\big)$$

as well as the operator norm $\|S\|_\infty = \max_{u\in\mathcal{H},\|u\|=1}\|Su\|$. For covariance operators Σ, this gives $\|\Sigma\|_\infty = \lambda_1$ and $\|\Sigma\|_2^2 = \sum_{j\geq 1}\lambda_j^2$. Under Assumption 2.1, Σ is a trace class operator (see, e.g., [32], Theorem III.2.3) and all quantities are indeed well-defined. In addition, for $r \geq 1$, we use the abbreviations $\mathrm{tr}_{>r}(\Sigma)$ and $\mathrm{tr}_{\geq r}(\Sigma)$ for $\sum_{j>r}\lambda_j$ and $\sum_{j\geq r}\lambda_j$, respectively.

Introducing the class

$$\mathcal{P}_d = \{P : \mathcal{H} \to \mathcal{H} | P \text{ is orthogonal projection of rank } d\},$$

the (population) reconstruction error of $P \in \mathcal{P}_d$ is defined by

$$R(P) = \mathbb{E}\big[\|X - PX\|^2\big] = \langle \Sigma, I - P\rangle.$$

The fundamental idea behind PCA is that $P_{\leq d}$ satisfies

$$P_{\leq d} \in \operatorname*{argmin}_{P\in\mathcal{P}_d} R(P), \qquad R(P_{\leq d}) = \mathrm{tr}_{>d}(\Sigma). \tag{2.1}$$

Similarly, the empirical reconstruction error of $P \in \mathcal{P}_d$ is defined by

$$R_n(P) = \frac{1}{n}\sum_{i=1}^n \|X_i - PX_i\|^2 = \langle \hat{\Sigma}, I - P\rangle,$$

and we have

$$\hat{P}_{\leq d} \in \operatorname*{argmin}_{P\in\mathcal{P}_d} R_n(P). \tag{2.2}$$

The excess risk of the PCA projector $\hat{P}_{\leq d}$ is thus given by

$$\mathcal{E}_d^{\mathrm{PCA}} := R(\hat{P}_{\leq d}) - R(P_{\leq d}) = \langle \Sigma, P_{\leq d} - \hat{P}_{\leq d}\rangle. \tag{2.3}$$

By (2.1), the excess risk $\mathcal{E}_d^{\mathrm{PCA}}$ defines a nonnegative loss function in the decision-theoretic sense for the estimator $\hat{P}_{\leq d}$ under the parameter Σ. Our main objective is to find nonasymptotic bounds for

$$\mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] = \mathbb{E}R(\hat{P}_{\leq d}) - \min_{P\in\mathcal{P}_d} R(P),$$

the decision-theoretic risk. In some situations, we also consider the problem of deriving standard oracle inequalities, by allowing to replace the constant 1 in front of the minimum by a larger constant.

Throughout the paper, c and C denote constants. We make the convention that these constants are not necessarily the same at each occurrence. They usually depend on C_1 from Assumption 2.1. For our expectation bounds, we make C more explicit by using the constant C_2, where $C_2 > 0$ is the smallest constant such that

$$\mathbb{E}\big[\|P_j(\Sigma - \hat{\Sigma})P_k\|_2^2\big] \leq C_2\delta\lambda_j\lambda_k/n \tag{2.4}$$

with $\delta = 1$ if $j \neq k$ and $\delta = 2$ otherwise. It is easy to check that (2.4) holds with $C_2 \leq 16C_1^4$ and that for X Gaussian (2.4) holds with $C_2 = 1$.

2.2. *ERM-bounds for the excess risk.* A natural approach to derive upper bounds for the excess risk is to follow the standard theory of empirical risk minimisation (ERM). The important basic inequality in ERM is

$$0 \leq \langle \Sigma, P_{\leq d} - \hat{P}_{\leq d}\rangle \leq \langle \Sigma - \hat{\Sigma}, P_{\leq d} - \hat{P}_{\leq d}\rangle = \langle \Delta, P_{\leq d} - \hat{P}_{\leq d}\rangle \tag{2.5}$$

with

$$\Delta = \Sigma - \hat{\Sigma},$$

which follows from (2.1) and (2.2). This route has been taken by Blanchard, Bousquet and Zwald [6], who applied sophisticated arguments from empirical process theory, based on Bartlett, Bousquet and Mendelson [3]. Let us derive some simple nonasymptotic expectation bounds from (2.5), which will set the stage for more refined results later.

PROPOSITION 2.2. *We have*

$$\mathcal{E}_d^{\mathrm{PCA}} \le \min\Big(\sqrt{2d}\|\Delta\|_2, \frac{2\|\Delta\|_2^2}{\lambda_d - \lambda_{d+1}}\Big)$$

with the convention $x/0 := \infty$. *With Assumption* 2.1,

$$\mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \le \min\Big(\frac{\sqrt{4C_2 d}\,\mathrm{tr}(\Sigma)}{\sqrt{n}}, \frac{4C_2\,\mathrm{tr}^2(\Sigma)}{n(\lambda_d - \lambda_{d+1})}\Big) \tag{2.6}$$

follows, *where* C_2 *is the constant in* (2.4).

REMARK 2.3. By (2.1), the left-hand side in (2.5) does not depend on the choice of $P_{\le d}$ if $\lambda_d = \lambda_{d+1}$, while the right-hand side in general does. Nevertheless, since the actual choice of $P_{\le d}$ does not alter the final result in Proposition 2.2, we let this choice unspecified and make the convention that the P_j have been fixed in advance.

REMARK 2.4. Extending the terminology of [6], we call the first and the second part in (2.6) global and local bound, respectively, referring to the dependence on specific spectral gaps or not. The expected excess risk is thus bounded by a slow global $n^{-1/2}$-rate as well as by a fast local n^{-1}-rate which depends on the spectral gap $\lambda_d - \lambda_{d+1}$. For (2.6) to hold, only the fourth moment bound (2.4) is required instead of the full Assumption 2.1.

PROOF OF PROPOSITION 2.2. From (2.5), we obtain

$$(\mathcal{E}_d^{\mathrm{PCA}})^2 \le \|\Delta\|_2^2 \|P_{\le d} - \hat{P}_{\le d}\|_2^2 \tag{2.7}$$

by the Cauchy–Schwarz inequality. Since orthogonal projectors are idempotent and self-adjoint, we have $\langle P_{\le d}, \hat{P}_{\le d}\rangle = \|P_{\le d}\hat{P}_{\le d}\|_2^2 \ge 0$, and thus

$$\|P_{\le d} - \hat{P}_{\le d}\|_2^2 = 2\big(d - \langle P_{\le d}, \hat{P}_{\le d}\rangle\big) \le 2d.$$

Insertion into (2.7) yields the first part of the bound. The second part of the bound follows from a short recursion argument. Indeed, we have

$$\|P_{\le d} - \hat{P}_{\le d}\|_2^2 \le \frac{2\mathcal{E}_d^{\mathrm{PCA}}}{\lambda_d - \lambda_{d+1}},$$

which is a variant of the Davis–Kahan inequality and follows by simple projector calculus; see Lemma 2.6 and (2.21) below. We obtain

$$(\mathcal{E}_d^{\mathrm{PCA}})^2 \le \|\Delta\|_2^2 \frac{2\mathcal{E}_d^{\mathrm{PCA}}}{\lambda_d - \lambda_{d+1}}.$$

This yields the second part of the bound. Finally, the expectation bound (2.6) follows from inserting (2.4). □

The global rate can be improved by using the variational characterisation of partial traces again. In the case $\Sigma = I + x P_{\le d}$, for instance, the global rate $p\sqrt{d/n}$ of Proposition 2.2 is improved to $d\sqrt{p/n}$. The latter is optimal for $d \le p/2$ and spectral gap $x = \sqrt{p/n}$; see the lower bound (2.19) below.

PROPOSITION 2.5. *Grant Assumption 2.1. Then we have*

$$\mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \le C \sum_{j \le d} \max\left(\sqrt{\frac{\lambda_j \operatorname{tr}_{\ge j}(\Sigma)}{n}}, \frac{\operatorname{tr}_{\ge j}(\Sigma)}{n}\right),$$

where $C > 0$ is a constant depending only on C_1.

PROOF. Using (2.5), we have

$$\mathcal{E}_d^{\mathrm{PCA}} \le \langle \Delta, P_{\le d} \rangle + \sup_{P \in \mathcal{P}_d} \langle -\Delta, P \rangle. \tag{2.8}$$

By the variational characterisation of partial traces (cf. (2.1), (2.2)) and the min-max characterisation of eigenvalues (see, e.g., [22], Chapter 28), we get

$$\sup_{P \in \mathcal{P}_d} \langle -\Delta, P \rangle \le \sum_{j \le d} \| P_{\ge j} \Delta P_{\ge j} \|_\infty.$$

Noting that $\mathbb{E}[\langle \Delta, P_{\le d} \rangle] = 0$, we conclude that

$$\mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \le \sum_{j \le d} \mathbb{E}[\| P_{\ge j} \Delta P_{\ge j} \|_\infty].$$

Finally, we apply the moment bound for sample covariance operators obtained by Koltchinskii and Lounici [19]. Consider $X' = P_{\ge j} X$, $X_i' = P_{\ge j} X_i$ which again satisfy Assumption 2.1 (with the same constant C_1) and lead to the covariance and the sample covariance

$$\Sigma' = P_{\ge j} \Sigma P_{\ge j}, \qquad \hat{\Sigma}' = P_{\ge j} \hat{\Sigma} P_{\ge j}. \tag{2.9}$$

Since Σ' has trace $\operatorname{tr}_{\ge j}(\Sigma)$ and operator norm $\lambda_1' = \lambda_j$, [19], Theorem 4, applied to $\Delta' = \Sigma' - \hat{\Sigma}'$ gives

$$\mathbb{E}[\| P_{\ge j} \Delta P_{\ge j} \|_\infty] \le C \max\left(\sqrt{\frac{\lambda_j \operatorname{tr}_{\ge j}(\Sigma)}{n}}, \frac{\operatorname{tr}_{\ge j}(\Sigma)}{n}\right),$$

where C is a constant depending only on C_1, and the claim follows. □

The bounds in Propositions 2.2 and 2.5 exhibit nicely the interplay between the global $n^{-1/2}$-rate and the local n^{-1}-rate. At first glance, it is surprising that the bounds derived via the basic ERM-inequality may nevertheless be suboptimal. For the simple isotropic case $\Sigma = \sigma^2 I$ (enforcing a finite dimension p) with $\mathcal{E}_d^{\mathrm{PCA}} = 0$, they only provide an upper bound of order $d\sqrt{p/n}$. The reason is an asymmetry with the risk $\langle \hat{\Sigma}, \hat{P}_{\le d} - P_{\le d} \rangle$ with the population and empirical versions exchanged, which may be much larger than the excess risk.

For the lower bound model $\Sigma = I + x P_{\le d}$ with $n = 1000$, $p = 50$ and $d = 3$, Figure 1 displays the expectation (obtained from accurate Monte Carlo simulations) of the upper bound from the basic inequality (2.5) (dashed-dotted line) and the upper bound (2.8), used for proving Proposition 2.5 (dotted line), compared to the expected excess risk (solid line). In addition, Figure 1 displays the upper bound obtained in (2.18) with $C = 1.1$, taking into account Remark 3.7 (dashed line). This new upper bound captures correctly the small excess risk for small spectral gaps x.

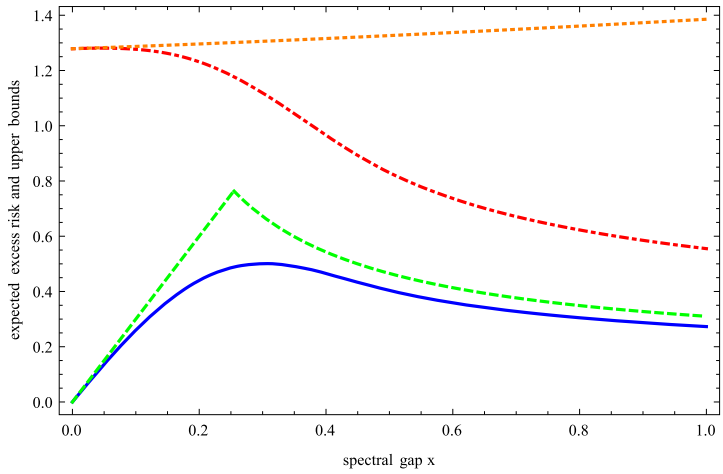

FIG. 1. *Expected excess risk* (*solid*) *and its upper bounds from* (2.18) (*dashed*), (2.5) (*dashed-dotted*), *and* (2.8) (*dotted*) *as functions of the spectral gap*.

2.3. *New bounds for the excess risk.* All results presented are proved in Section 4 below. The following representation of the excess risk is fundamental for the new bounds.

LEMMA 2.6. *For any* $\mu \in \mathbb{R}$, *we have*

$$\mathcal{E}_d^{\mathrm{PCA}} = \sum_{j \le d} (\lambda_j - \mu) \| P_j \hat{P}_{>d} \|_2^2 + \sum_{k > d} (\mu - \lambda_k) \| P_k \hat{P}_{\le d} \|_2^2.$$

It turns out that the two risk parts exhibit a different behaviour and we shall bound them separately. Therefore, we introduce

$$\mathcal{E}_{\le d}^{\mathrm{PCA}}(\mu) = \sum_{j \le d} (\lambda_j - \mu) \| P_j \hat{P}_{>d} \|_2^2, \qquad \mathcal{E}_{>d}^{\mathrm{PCA}}(\mu) = \sum_{k > d} (\mu - \lambda_k) \| P_k \hat{P}_{\le d} \|_2^2.$$

Usually, we shall choose $\mu \in [\lambda_{d+1}, \lambda_d]$ such that all terms are positive, but sometimes it pays off to choose a different value. Our first main result is as follows.

PROPOSITION 2.7. *Grant Assumption* 2.1 *and let* $\mu \in [\lambda_{d+1}, \lambda_d]$. *Then for all* $r = 0, \dots, d$ *we have*

$$\mathbb{E}\big[\mathcal{E}_{\le d}^{\mathrm{PCA}}(\mu)\big] \le C \sum_{j \le r} (\lambda_j - \mu) \frac{\lambda_j \operatorname{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})^2} + \sum_{j=r+1}^{d \wedge (r+p-d)} (\lambda_j - \mu)$$

with $C = 8C_2 + 8C_3^2$, *where* C_2 *and* C_3 *are given in* (2.4) *and* (3.16), *respectively. Moreover, if* $d \le n/(16C_3^2)$, *then for all* $l = d+1, \dots, p+1$ *we have*

$$\mathbb{E}\big[\mathcal{E}_{>d}^{\mathrm{PCA}}(\mu)\big] \le C \sum_{k \ge l} (\mu - \lambda_k) \frac{\lambda_k \operatorname{tr}(\Sigma)}{n(\lambda_d - \lambda_k)^2} + \sum_{k=(d+1)\vee(l-d)}^{l-1} (\mu - \lambda_k) + R$$

with remainder term $R = (\mu - \lambda_p) e^{-n/(32C_3^2)}$. *For* $p = \infty$, *we understand* $\lambda_p = 0$ *and* $l \in \{k \in \mathbb{N} \,|\, k \ge d+1\} \cup \{+\infty\}$ *and for* $l = p = \infty$ *we understand* $\sum_{k=(l-d)\vee(d+1)}^{l-1} (\mu - \lambda_k) = d\mu$.

Bounds of the same order can be derived for the L^p-norms of $\mathcal{E}_{\le d}^{\mathrm{PCA}}(\mu)$ and $\mathcal{E}_{>d}^{\mathrm{PCA}}(\mu)$ with a constant C depending additionally on p; see, for example, Lemma 4.3 for the additional arguments needed in the case $p = 2$. By simple arguments, we obtain the following corollary.

COROLLARY 2.8. *Grant Assumption 2.1 and let $\mu \in [\lambda_{d+1}, \lambda_d]$. Then we have*

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_{\le d}(\mu)\big] \le \sum_{j \le d} \min\bigg(C \frac{\lambda_j \operatorname{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})}, \lambda_j - \lambda_{d+1} \bigg).$$

Moreover, if $d \le n/(16C_3^2)$, then

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_{>d}(\mu)\big] \le \sum_{k > d} \min\bigg(C \frac{\lambda_k \operatorname{tr}(\Sigma)}{n(\lambda_d - \lambda_k)}, \lambda_d - \lambda_k \bigg) + (\lambda_d - \lambda_p) e^{-\frac{n}{32C_3^2}}.$$

In both inequalities, we have $C = 8C_2 + 8C_3^2$.

Summing up the inequalities in Proposition 2.7 leads to an upper bound for $\mathbb{E}[\mathcal{E}^{\mathrm{PCA}}_d]$ which improves the local bound of Proposition 2.2 and gives the value 0 in the isotropic case $\Sigma = \sigma^2 I$. Furthermore, global bounds emerge as trade-off between the two terms involved in the upper bounds. More precisely, we have the following.

THEOREM 2.9. *Grant Assumption 2.1 and suppose $d \le n/(16C_3^2)$. Then we have the local bound*

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_d\big] \le C \sum_{\substack{j \le d: \\ \lambda_j > \lambda_{d+1}}} \frac{\lambda_j \operatorname{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})} + C \sum_{\substack{k > d: \\ \lambda_k < \lambda_d}} \frac{\lambda_k \operatorname{tr}(\Sigma)}{n(\lambda_d - \lambda_k)} + (\lambda_d - \lambda_p) e^{-\frac{n}{32C_3^2}}$$

and the global bound

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_d\big] \le \sum_{j \le d} \sqrt{\frac{C\lambda_j \operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{C d \operatorname{tr}_{>d}(\Sigma) \operatorname{tr}(\Sigma)}{n}}. \tag{2.10}$$

In both inequalities, we have $C = 8C_2 + 8C_3^2$.

For our second main result, we impose additional eigenvalue conditions, and thus improve the first bound of Proposition 2.7. A main feature is that the full trace of Σ can be replaced by the partial trace $\operatorname{tr}_{>s}(\Sigma)$, which in the case $s = d$ coincides with the oracle reconstruction error.

PROPOSITION 2.10. *Grant Assumption 2.1. Then for all indices $s = 1, \dots, d$ such that*

$$\frac{\lambda_s}{\lambda_s - \lambda_{d+1}} \sum_{j \le s} \frac{\lambda_j}{\lambda_j - \lambda_{d+1}} \le n/\big(16C_3^2\big) \tag{2.11}$$

and all $r = 0, \dots, s$, we have

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_{\le d}(\lambda_{d+1})\big] \le C \sum_{j \le r} \frac{\lambda_j \operatorname{tr}_{>s}(\Sigma)}{n(\lambda_j - \lambda_{d+1})} + 2 \sum_{r < j \le d} (\lambda_j - \lambda_{d+1}) + R$$

with $C = 16C_2 + 8C_3^2$ and remainder term given by

$$R = 1024 C_1^4 \sum_{j \le r} \frac{\lambda_j \operatorname{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})} e^{-\frac{n(\lambda_s - \lambda_{d+1})^2}{(4C_3\lambda_s)^2}}.$$

In the special case $\lambda_{d+1} = \cdots = \lambda_p$ (compare the spiked covariance model below), we have $\mathcal{E}^{\mathrm{PCA}}_{>d}(\lambda_{d+1}) = 0$, and thus Proposition 2.10 yields an upper bound for the whole excess risk. In the general case, we still have the following consequence.

THEOREM 2.11. *Grant Assumption 2.1 and suppose* $\lambda_d - \lambda_{d+1} \geq c_1(\lambda_d - \lambda_p)$ *with* $c_1 > 0$. *If* (2.11) *holds with* $s = d$, *then we have the local bound*

$$\mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] \leq \frac{C}{c_1 n}\big(\mathrm{tr}_{>d}(\Sigma) + \mathrm{tr}(\Sigma)e^{-c_1^2 n(\lambda_d-\lambda_p)^2/(C\lambda_d^2)}\big)\sum_{j\leq d}\frac{\lambda_j}{\lambda_j - \lambda_{d+1}}.$$

Moreover, if $s \leq d$ *is the largest number such that* (2.11) *is satisfied* (*and* $s = 0$ *if such a number does not exist*), *then we have the global bound*

$$\mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] \leq \frac{C}{c_1\sqrt{n}}\big(\sqrt{\mathrm{tr}_{>s}(\Sigma)} + \sqrt{\mathrm{tr}(\Sigma)}e^{-c_1^2 n(\lambda_s-\lambda_p)^2/(C\lambda_s^2)}\big)\sum_{j\leq d}\sqrt{\lambda_j}.$$

In both inequalities, C *is a constant depending only on* C_1.

Finally, observe that upper bounds for the expectation of the excess risk $\mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \leq r.h.s.$ can be equivalently formulated as exact oracle inequalities $\mathbb{E}[R(\hat{P}_{\leq d})] \leq \min_{P\in\mathcal{P}_d} R(P) + r.h.s.$ If we give up the constant 1 in front of the minimum, Proposition 2.10 also leads to a third type of bound.

THEOREM 2.12. *Grant Assumption 2.1. Then for all indices* $s = 1, \ldots, d$ *such that* (2.11) *holds*, *we have*

$$\mathbb{E}\big[R(\hat{P}_{\leq d})\big] \leq C\,\mathrm{tr}_{>s}(\Sigma) + C\,\mathrm{tr}(\Sigma)e^{-n(\lambda_s-\lambda_{d+1})^2/(4C_3\lambda_s)^2}$$

with a constant $C > 0$ *depending only on* C_1.

If (2.11) holds with $s = d$, then $\mathrm{tr}_{>d}(\Sigma) = \inf_{P\in\mathcal{P}_d} R(P)$ and we obtain a standard oracle inequality with an exponentially small remainder term.

2.4. *Applications.* Let us illustrate our different upper bounds for three main classes of eigenvalue behaviour: exponential decay, polynomial decay and a simple spiked covariance model. Eigenvalue structures such as exponential or polynomial decay are typically considered in the context of functional data (see, e.g., [11, 14, 25]). Spiked covariance models are often studied in the context of high-dimensional data [8, 15, 34].

Exponential decay. Assume for some $\alpha > 0$

$$\lambda_j = e^{-\alpha j}, \quad j \geq 1. \tag{2.12}$$

Then we have $\lambda_j - \lambda_{d+1} \geq (1 - e^{-\alpha})\lambda_j$ for every $j \leq d$ and (4.4) below gives $\mathcal{E}_d^{\mathrm{PCA}} \leq (1 - e^{-\alpha})^{-1}\mathcal{E}_{\leq d}^{\mathrm{PCA}}(\lambda_{d+1})$. Hence, Corollary 2.8 implies

$$\mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] \leq C\sum_{j\leq d}\min\big(1/n, e^{-\alpha j}\big) \leq C\frac{d \wedge \log(en)}{n},$$

where C (not the same at each occurrence) is a constant depending only on C_1 and α. This bound improves the local bound in Proposition 2.2 (which gives $Ce^{\alpha d}/n$) and the bounds in Theorems 3.2 and 3.4 of [6], respectively.

Next, we show that this result can be much improved by applying the local bound in Theorem 2.11. Indeed, the left-hand side of (2.11) with $s = d$ can be bounded by $d(1 - e^{-\alpha})^{-2}$. Thus, assuming that this value is smaller than $n/(16C_3^2)$, we can apply the local bound in Theorem 2.11. The main term is bounded by $C(1 - e^{-\alpha})^{-3}dn^{-1}e^{-\alpha(d+1)}$ and the

remainder term by $1024C_1^4(1-e^{-\alpha})^{-2}n^{-1}\exp(-n(1-e^{-\alpha})^2/(16C_3^2))$. We conclude that there are constants $c, C>0$ depending only on C_1 and α such that

$$\mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] \le C\frac{de^{-\alpha d}}{n}, \tag{2.13}$$

provided that $d \le cn$. Noting for the population reconstruction error

$$R(P_{\le d}) = \sum_{k>d} e^{-\alpha k} = e^{-\alpha}\big(1-e^{-\alpha}\big)^{-1}e^{-\alpha d},$$

we see that the excess risk is smaller than the oracle risk, provided that $d \le cn$. In the Supplementary Material [28], we derive linear expansions for the excess risk, implying that (2.13) is indeed sharp. In fact, (B.17) in [28] says that for X Gaussian, there are constants $c, C>0$ depending only on α such that $\mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \ge C^{-1}de^{-\alpha d}n^{-1}$, provided that $d \le cn$.

Polynomial decay. Assume for some $\alpha > 1$

$$\lambda_j = j^{-\alpha}, \quad j \ge 1. \tag{2.14}$$

Then the local bound in Theorem 2.9 and the inequalities

$$\sum_{j\le d}\frac{\lambda_j}{\lambda_j-\lambda_{d+1}} \le Cd\log(ed), \qquad \sum_{k>d}\frac{\lambda_k}{\lambda_d-\lambda_k} \le Cd\log(ed) \tag{2.15}$$

from (A.19) in the Supplementary Material [28] yield that there are constants $c, C>0$ depending only on C_1 and α such that

$$\mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] \le C\frac{d\log(ed)}{n}$$

for all $d \le cn$. This already improves the results obtained in [6], Section 5, where a rate strictly between $n^{-1/2}$ and n^{-1} is derived.

Again, for large d, this result can be much improved by using Theorems 2.11 and 2.12. Choosing $s = \lfloor d/2 \rfloor$, there is a constant c depending only on C_1 and α such that Condition (2.11) is satisfied if $d \le cn$. Thus, Theorem 2.12 yields

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}\big] \le C\operatorname{tr}_{>\lfloor d/2\rfloor}(\Sigma) + Ce^{-n/C} \le Cd^{1-\alpha} + Ce^{-n/C}, \tag{2.16}$$

provided that $d \le cn$. Noting for the population reconstruction error

$$R(P_{\le d}) = \sum_{k>d} k^{-\alpha} \ge cd^{1-\alpha},$$

we see from (2.16) that for $d \le cn$ the excess risk is always smaller than a constant times the oracle risk.

Similarly, Proposition 2.10 (applied with $r=s=d$), Theorem 2.11, and (2.15) yield

$$\mathbb{E}\big[\mathcal{E}_{\le d}^{\mathrm{PCA}}(\lambda_{d+1})\big] \le C\frac{d^{2-\alpha}\log(ed)}{n}, \qquad \mathbb{E}\big[\mathcal{E}_d^{\mathrm{PCA}}\big] \le C\frac{d^{3-\alpha}\log(ed)}{n},$$

provided that $d^2\log(ed) \le cn$, where $c, C>0$ are constants depending only on C_1 and α. In Appendix B.3 in the Supplementary Material [28], we show that the first inequality also holds without the $\log(ed)$ term, and that the second inequality can be improved to the sharp bound $Cd^{2-\alpha}n^{-1}$, yet under a more restrictive condition on d. This leads to the conjecture that for the excess risk the bound $Cd^{2-\alpha}n^{-1}$ holds in the larger regime $d^2\log(ed) \le cn$.

Spiked covariance model. Let Θ be the class of all symmetric matrices whose eigenvalues satisfy

$$1+\kappa x \geq \lambda_1 \geq \cdots \geq \lambda_d \geq 1+x \quad \text{and} \quad \lambda_{d+1} = \cdots = \lambda_p = 1, \tag{2.17}$$

where $x \geq 0$ and $\kappa > 1$. Then it holds

$$\sup_{\Sigma\in\Theta} \mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \leq \min\biggl(C\kappa\frac{(1+\kappa x)d(p-d)}{nx}, d\kappa x, (p-d)\kappa x\biggr) + \kappa x e^{-\frac{n}{32C_3^2}}, \tag{2.18}$$

provided that $d \leq cn$, where $c, C > 0$ are constants depending only on C_1. Considering separately the cases $x \leq c$ and $x > c$, we see that the excess risk is always smaller than the oracle risk $R(P_{\leq d}) = p - d$. To prove (2.18), it suffices to apply Proposition 2.7. Indeed, the claim follows from applying either Lemma 2.6 with $\mu = 1$ and the first inequality in Proposition 2.7 or Lemma 2.6 with $\mu = 1+\kappa x$ and the second inequality in Proposition 2.7 (depending on whether $(1+\kappa x)d \leq p-d$ or $(1+\kappa x)d > p-d$).

In fact, since

$$x\|P_{\leq d} - \hat{P}_{\leq d}\|_2^2 \leq 2\mathcal{E}_d^{\mathrm{PCA}} \leq \kappa x\|P_{\leq d} - \hat{P}_{\leq d}\|_2^2,$$

(2.18) is equivalent to a result by Cai, Ma and Wu [8], Theorem 9. Moreover, their minimax lower bound [8], Theorem 8 (see also Vu and Lei [34], Theorem A.2) gives

$$\inf_{\hat{P}_{\leq d}} \sup_{\Sigma\in\Theta} \mathbb{E}\bigl[\langle \Sigma, P_{\leq d} - \hat{P}_{\leq d}\rangle\bigr] \geq c\min\biggl(\frac{(1+x)d(p-d)}{nx}, dx, (p-d)x\biggr), \tag{2.19}$$

where the infimum is taken over all estimators $\hat{P}_{\leq d}$ based on $X_1, \ldots, X_n$ with values in $\mathcal{P}_d$ and $c > 0$ is a constant.

Oracle inequality. One interesting conclusion in the above typical situations is a nonasymptotic bound by the oracle risk, more precisely the following.

COROLLARY 2.13. *In the cases* (2.12), (2.14) *and* (2.17), *there are constants* $c, C > 0$ *depending only on* C_1, α, *and* κ *such that the oracle inequality*

$$\mathbb{E}\bigl[R(\hat{P}_{\leq d})\bigr] \leq C \cdot R(P_{\leq d}),$$

holds for all $d \leq cn$.

2.5. *Discussion.* Let us review some connections and implications.

Subspace distance versus excess risk. Many results cover the Hilbert–Schmidt distance $\|\hat{P}_{\leq d} - P_{\leq d}\|_2$, which has a geometric interpretation in terms of canonical angles. In this direction, the most well-known bound is the Davis–Kahan $\sin\Theta$ theorem; see, for example, Yu, Wang and Samworth [36] for a recent statistical account. More accurate bounds are derived, for example, in Mas and Ruymgaart [25] in a functional setting and in Vu and Lei [34] and Cai, Ma and Wu [8] in a high-dimensional sparse setting.

The squared Hilbert–Schmidt distance can be written as

$$\|\hat{P}_{\leq d} - P_{\leq d}\|_2^2 = 2\sum_{j\leq d}\|P_j\hat{P}_{>d}\|_2^2 = 2\sum_{k>d}\|P_k\hat{P}_{\leq d}\|_2^2; \tag{2.20}$$

see, for example, the proof of Lemma 2.6. Compared to

$$\mathcal{E}_d^{\mathrm{PCA}} = \sum_{j \le d} (\lambda_j - \lambda_{d+1}) \| P_j \hat{P}_{>d} \|_2^2 + \sum_{k > d} (\lambda_{d+1} - \lambda_k) \| P_k \hat{P}_{\le d} \|_2^2$$

from Lemma 2.6, we see that the squared Hilbert–Schmidt distance and the excess risk differ in the weighting of the projector norms. In fact, we obtain

$$\frac{2\mathcal{E}_{\le d}^{\mathrm{PCA}}(\lambda_{d+1})}{\lambda_1 - \lambda_{d+1}} \le \| \hat{P}_{\le d} - P_{\le d} \|_2^2 \le \frac{2\mathcal{E}_{\le d}^{\mathrm{PCA}}(\lambda_{d+1})}{\lambda_d - \lambda_{d+1}} \le \frac{2\mathcal{E}_d^{\mathrm{PCA}}}{\lambda_d - \lambda_{d+1}}. \tag{2.21}$$

This means that all excess risk bounds a fortiori imply bounds on the Hilbert–Schmidt distance up to a spectral gap factor. For instance, in our setting, (2.21) implies most versions of the Davis–Kahan $\sin\Theta$ theorem, for example, those in [36], by using the basic inequality $\langle \Delta, P_{\le d} - \hat{P}_{\le d} \rangle \ge \mathcal{E}_d^{\mathrm{PCA}}$ in (2.5) and bounding the scalar product by a Cauchy–Schwarz or operator norm inequality. In contrast, the first inequality in (2.21) does not lead to good bounds for the excess risk when $\lambda_d - \lambda_{d+1}$ is small relative to $\lambda_1 - \lambda_{d+1}$. In the extreme case $\lambda_d = \lambda_{d+1}$, the Hilbert–Schmidt distance depends on the choice of (u_d, u_{d+1}) and is thus not even well-defined. A more sophisticated version of (2.21) is derived in Appendix B in the Supplementary Material [28].

Finally, note that the Hilbert–Schmidt distance and the excess risk have different applications. For instance, bounds for the Hilbert–Schmidt distance $\| \hat{P}_j - P_j \|_2$ are fundamental in the analysis of several testing algorithms; see, for example, Horváth and Kokoszka [12]. On the other hand, the excess risk is more adequate for tasks like reconstruction and prediction; see, for example, Wahl [35] for the case of the prediction error of principal component regression.

Asymptotic versus nonasymptotic. For the Hilbert–Schmidt distance, it is known that for $\mathcal{H} = \mathbb{R}^p$ and $X \sim N(0, \Sigma)$ with fixed Σ in the case $\lambda_d > \lambda_{d+1}$,

$$n \| \hat{P}_{\le d} - P_{\le d} \|_2^2 \xrightarrow{d} 2 \sum_{j \le d, k > d} \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2} g_{jk}^2 \tag{2.22}$$

holds as $n \to \infty$, where $(g_{jk})_{j \le d < k}$ is an array of independent standard Gaussian random variables; see, for example, Dauxois, Pousse and Romain [9] and also Koltchinskii and Lounici [18, 19]. The projector calculus developed in Section 3.1 allows to obtain readily the analogue of the asymptotic result (2.22) for the excess risk $\mathcal{E}_d^{\mathrm{PCA}}$ without any spectral gap condition. More precisely, we prove in Appendix A.7 in the Supplementary Material [28] the following.

PROPOSITION 2.14. *Let $\mathcal{H} = \mathbb{R}^p$ and $X \sim N(0, \Sigma)$ with Σ fixed. As $n \to \infty$, we have for the excess risk $\mathcal{E}_{d,n}^{\mathrm{PCA}} = \mathcal{E}_d^{\mathrm{PCA}}$,*

$$n \mathcal{E}_{d,n}^{\mathrm{PCA}} \xrightarrow{d} \sum_{\substack{j \le d, k > d: \\ \lambda_j > \lambda_k}} \frac{\lambda_j \lambda_k}{\lambda_j - \lambda_k} g_{jk}^2,$$

where $(g_{jk})_{j \le d < k}$ are independent standard Gaussian random variables.

We see that the excess risk converges with n^{-1}-rate also in the case $\lambda_d = \lambda_{d+1}$. Note, however, that the convergence cannot be uniform in the parameter Σ in view of the discontinuity of the right-hand side in (λ_j). This clearly underpins the need for nonasymptotic upper bounds for the excess risk.

In certain examples, including the spiked covariance model and exponential decay of eigenvalues, the eigenvalue expression in Proposition 2.10 (with $r = s = d$) coincides with the one in Proposition 2.14. In the general case, including polynomial decay, the eigenvalue expressions differ. In Appendix B in the Supplementary Material, we derive nonasymptotic bounds which give the asymptotic leading terms in (2.22) and Proposition 2.14, by using linear expansions for $\hat{P}_{\le d}$ and $\hat{P}_{>d}$. These bounds, however, require stronger eigenvalue conditions (including $\lambda_d > \lambda_{d+1}$). In contrast, our main results in Section 2.3 also apply to the case of small or vanishing spectral gaps.

Eigenvalue concentration. We obtain deviation inequalities for empirical eigenvalues which are of independent interest. Concentration inequalities for eigenvalues using tools from measure concentration are widespread; see, for example, [1, 4, 7, 23, 24, 26]. The main difference to our deviation inequalities is that we take into account the local eigenvalue structure. For instance, from Propositions 3.10 and 3.13, we get the following theorem.

THEOREM 2.15. *Grant Assumption 2.1. Then there is a constant $c > 0$ depending only on C_1 such that for all $y > 0$ satisfying*

$$\frac{1}{n(y \wedge 1)} \sum_{k>d} \frac{\lambda_k}{\lambda_d - \lambda_k + y\lambda_d} \le 1/(2C_3^2)$$

we have

$$\mathbb{P}(\hat{\lambda}_d - \lambda_d > y\lambda_d) \le e^{1-cn(y\wedge y^2)}.$$

Moreover, for all $y > 0$ satisfying

$$\frac{1}{n(y \wedge 1)} \sum_{j<d} \frac{\lambda_j}{\lambda_j - \lambda_d + y\lambda_d} \le 1/(2C_3^2)$$

we have

$$\mathbb{P}(\hat{\lambda}_d - \lambda_d < -y\lambda_d) \le e^{1-cn(y\wedge y^2)}.$$

If λ_d is a simple eigenvalue, then Theorem 2.15 can be seen as a nonasymptotic version of the classical central limit theorem $\sqrt{n}(\hat{\lambda}_d/\lambda_d - 1) \to \mathcal{N}(0, 2)$ which holds for X Gaussian; compare Anderson [2], Theorem 13.5.1, and Dauxois, Pousse and Romain [9], Proposition 8. Moreover, the conditions imposed are related to $\mathbb{E}[\hat{\lambda}_d]$ by the following asymptotic expansion (see, e.g., [27], equation (2.22)):

$$\mathbb{E}[\hat{\lambda}_d/\lambda_d] - 1 = \frac{1}{n} \sum_{k\neq d} \frac{\lambda_k}{\lambda_d - \lambda_k} + \cdots.$$

A discussion how the eigenvalue conditions in Theorem 2.15 improve upon standard conditions from the literature is given in Remark 3.15.

3. Main tools.

3.1. *Projector-based calculus.* In this section, we present two perturbation formulas, which together with the representation of the excess risk given in Lemma 2.6 form the basis of our analysis of the excess risk.

LEMMA 3.1. *For $j \le d$, we have*

$$\|P_j \hat{P}_{>d}\|_2^2 = \sum_{k>d} \frac{\|P_j \Delta \hat{P}_k\|_2^2}{(\lambda_j - \hat{\lambda}_k)^2},$$

and for $k > d$, we have

$$\|P_k \hat{P}_{\le d}\|_2^2 = \sum_{j \le d} \frac{\|P_k \Delta \hat{P}_j\|_2^2}{(\hat{\lambda}_j - \lambda_k)^2}.$$

Both identities hold provided that all denominators are nonzero.

PROOF. The main ingredient is the formula

$$P_j \hat{P}_k = \frac{1}{\lambda_j - \hat{\lambda}_k} P_j \Delta \hat{P}_k, \tag{3.1}$$

which follows from inserting the spectral representations of Σ and $\hat{\Sigma}$ into the right-hand side. Indeed,

$$P_j \Delta \hat{P}_k = \sum_{l \ge 1} \lambda_l P_j P_l \hat{P}_k - \sum_{l \ge 1} \hat{\lambda}_l P_j \hat{P}_l \hat{P}_k = (\lambda_j - \hat{\lambda}_k) P_j \hat{P}_k.$$

The first claim now follows from inserting (3.1) into the identity

$$\|P_j \hat{P}_{>d}\|_2^2 = \sum_{k>d} \|P_j \hat{P}_k\|_2^2.$$

The second claim follows similarly by switching j and k and summation over j. □

Identity (3.1) can be seen as a basic building block to derive expansions for empirical spectral projectors. Indeed, using (3.1), we get

$$P_j \hat{P}_{>d} = \sum_{k>d} \frac{P_j \Delta \hat{P}_k}{\lambda_j - \hat{\lambda}_k} \tag{3.2}$$

and a similar formula for $P_k \hat{P}_{\le d}$, leading to

$$\hat{P}_{>d} - P_{>d} = P_{\le d} \hat{P}_{>d} - P_{>d} \hat{P}_{\le d} = \sum_{j \le d} \sum_{k>d} \left(\frac{P_j \Delta \hat{P}_k}{\lambda_j - \hat{\lambda}_k} + \frac{P_k \Delta \hat{P}_j}{\hat{\lambda}_j - \lambda_k} \right). \tag{3.3}$$

The following lemma immediately leads to a linear expansion of $\hat{P}_{>d}$.

LEMMA 3.2. *For $j \le d$, we have*

$$\begin{aligned} P_j \hat{P}_{>d} = {} & \sum_{k>d} \frac{P_j \Delta P_k}{\lambda_j - \lambda_k} + \sum_{k \le d} \sum_{l>d} \frac{P_j \Delta P_k \Delta \hat{P}_l}{(\lambda_j - \hat{\lambda}_l)(\lambda_k - \hat{\lambda}_l)} \\ & + \sum_{k>d} \sum_{l \le d} \frac{P_j \Delta P_k \Delta \hat{P}_l}{(\lambda_j - \lambda_k)(\hat{\lambda}_l - \lambda_k)} - \sum_{k>d} \sum_{l>d} \frac{P_j \Delta P_k \Delta \hat{P}_l}{(\lambda_j - \hat{\lambda}_l)(\lambda_j - \lambda_k)} \end{aligned}$$

and for $k > d$, we have

$$\begin{aligned} P_k \hat{P}_{\le d} = {} & \sum_{j \le d} \frac{P_k \Delta P_j}{\lambda_k - \lambda_j} + \sum_{j>d} \sum_{l \le d} \frac{P_k \Delta P_j \Delta \hat{P}_l}{(\lambda_k - \hat{\lambda}_l)(\lambda_j - \hat{\lambda}_l)} \\ & + \sum_{j \le d} \sum_{l>d} \frac{P_k \Delta P_j \Delta \hat{P}_l}{(\lambda_k - \lambda_j)(\hat{\lambda}_l - \lambda_j)} - \sum_{j \le d} \sum_{l \le d} \frac{P_k \Delta P_j \Delta \hat{P}_l}{(\lambda_k - \lambda_j)(\lambda_k - \hat{\lambda}_l)}. \end{aligned}$$

Both identities hold provided that all denominators are nonzero.

PROOF. We only prove the first identity, since the second one follows by the same line of arguments. First, using (3.2) and the identity $I = P_{\le d} + P_{>d} = \hat{P}_{\le d} + \hat{P}_{>d}$, we have

$$P_j \hat{P}_{>d} = \sum_{l>d} \frac{P_j \Delta \hat{P}_l}{\lambda_j - \hat{\lambda}_l} = \sum_{l>d} \frac{P_j \Delta P_{\le d} \hat{P}_l}{\lambda_j - \hat{\lambda}_l} + \sum_{l>d} \frac{P_j \Delta P_{>d} \hat{P}_l}{\lambda_j - \hat{\lambda}_l}$$

and

$$\sum_{k>d} \frac{P_j \Delta P_k}{\lambda_j - \lambda_k} = \sum_{k>d} \frac{P_j \Delta P_k \hat{P}_{\le d}}{\lambda_j - \lambda_k} + \sum_{k>d} \frac{P_j \Delta P_k \hat{P}_{>d}}{\lambda_j - \lambda_k}.$$

Thus

$$\begin{aligned} P_j \hat{P}_{>d} = {} & \sum_{k>d} \frac{P_j \Delta P_k}{\lambda_j - \lambda_k} + \sum_{l>d} \frac{P_j \Delta P_{\le d} \hat{P}_l}{\lambda_j - \hat{\lambda}_l} - \sum_{k>d} \frac{P_j \Delta P_k \hat{P}_{\le d}}{\lambda_j - \lambda_k} \\ & + \Bigg(\sum_{l>d} \frac{P_j \Delta P_{>d} \hat{P}_l}{\lambda_j - \hat{\lambda}_l} - \sum_{k>d} \frac{P_j \Delta P_k \hat{P}_{>d}}{\lambda_j - \lambda_k} \Bigg). \end{aligned} \tag{3.4}$$

Using (3.1), we get

$$\sum_{l>d} \frac{P_j \Delta P_{\le d} \hat{P}_l}{\lambda_j - \hat{\lambda}_l} = \sum_{k \le d} \sum_{l>d} \frac{P_j \Delta P_k \Delta \hat{P}_l}{(\lambda_j - \hat{\lambda}_l)(\lambda_k - \hat{\lambda}_l)}$$

and

$$-\sum_{k>d} \frac{P_j \Delta P_k \hat{P}_{\le d}}{\lambda_j - \lambda_k} = \sum_{k>d} \sum_{l \le d} \frac{P_j \Delta P_k \Delta \hat{P}_l}{(\lambda_j - \lambda_k)(\hat{\lambda}_l - \lambda_k)}.$$

Moreover, again using (3.1), the term in parentheses in (3.4) is equal to

$$\begin{aligned} & \sum_{l>d} \frac{P_j \Delta P_{>d} \hat{P}_l}{\lambda_j - \hat{\lambda}_l} - \sum_{k>d} \frac{P_j \Delta P_k \hat{P}_{>d}}{\lambda_j - \lambda_k} \\ & \quad = -\sum_{k>d} \sum_{l>d} \frac{\lambda_k - \hat{\lambda}_l}{(\lambda_j - \hat{\lambda}_l)(\lambda_j - \lambda_k)} P_j \Delta P_k \hat{P}_l \\ & \quad = -\sum_{k>d} \sum_{l>d} \frac{1}{(\lambda_j - \hat{\lambda}_l)(\lambda_j - \lambda_k)} P_j \Delta P_k \Delta \hat{P}_l, \end{aligned}$$

and the claim follows. □

REMARK 3.3. Note that compared to (3.2), where only spectral gaps between j and $k > d$ appear, the first formula in Lemma 3.2 includes all spectral gaps between $k > d$ and $l \le d$, even in the case $j = 1$. Since we are also interested in the case of small spectral gaps (including $\lambda_d = \lambda_{d+1}$), our main analysis of the excess risk will be based on Lemma 3.1. Lemma 3.2 will be important to derive linear expansions for the excess risk under stronger eigenvalue conditions.

REMARK 3.4. Usually, expansions for spectral projectors are obtained by the Cauchy integral representation for spectral projectors in combination with the second resolvent equation (resp., the second Neumann series); see, for example, Kato [17]. The difference of Lemmas 3.1 and 3.2 to the formulation in, for example, [18], Lemma 2, or [13], Theorem 5.1.4, is the form of the remainder term. In [13, 18], the remainder term is given by an integral over the resolvent, while the above results lead to an algebraic form of the remainder term. In Section 3.2 and Appendix B in the Supplementary Material [28], we will use these algebraic expressions to establish recursion arguments.

3.2. *Error decompositions.* In this section, we prove deterministic upper bounds for the excess risk which form the basis of our new upper bounds in Section 2.3. For $\mathcal{E}^{\mathrm{PCA}}_{\le d}(\mu)$, we split the sum into indices $j \le r$, where we expect the spectral gaps $\lambda_j - \lambda_{d+1}$ to be large, meaning that we can insert the perturbation formulas from Lemma 3.1, and into indices $r < j \le d$, where we expect the spectral gaps $\lambda_j - \mu$ to be small, meaning that wrong projections do not incur a large error. The terms of the first sum can then be controlled by a recursion argument.

PROPOSITION 3.5. *For $\mu \in [\lambda_{d+1}, \lambda_d]$ and $r = 0, \dots, d$, we have*

$$\begin{aligned}\mathcal{E}^{\mathrm{PCA}}_{\le d}(\mu) \le{}& 4\sum_{j\le r}(\lambda_j - \mu)\frac{\|P_j \Delta \hat{P}_{>d}\|_2^2}{(\lambda_j - \lambda_{d+1})^2} + \sum_{j=r+1}^{d\wedge(r+p-d)}(\lambda_j - \mu) \\ &+ \sum_{j\le r}(\lambda_j - \mu)\mathbb{1}\big(\hat{\lambda}_{d+1} - \lambda_{d+1} > (\lambda_j - \lambda_{d+1})/2\big).\end{aligned} \tag{3.5}$$

Furthermore, for $s = r, \dots, d$ and the weighted projector

$$S_{\le s} = S_{\le s}(\mu) = \sum_{j\le s}\frac{1}{\sqrt{\lambda_j - \mu}}P_j \tag{3.6}$$

(assuming $\lambda_s > \mu$) we obtain

$$\begin{aligned}\mathcal{E}^{\mathrm{PCA}}_{\le d}(\mu) \le{}& 16\sum_{j\le r}(\lambda_j - \mu)\frac{\|P_j \Delta P_{>s}\|_2^2}{(\lambda_j - \lambda_{d+1})^2} + 2\sum_{j=r+1}^{d\wedge(r+p-d)}(\lambda_j - \mu) \\ &+ 2\sum_{j\le r}(\lambda_j - \mu)\mathbb{1}\big(\hat{\lambda}_{d+1} - \lambda_{d+1} > (\lambda_j - \lambda_{d+1})/2\big) \\ &+ 8\sum_{j\le r}(\lambda_j - \mu)\frac{\|P_j \Delta\|_2^2}{(\lambda_j - \lambda_{d+1})^2}\mathbb{1}\big(\|S_{\le s}\Delta S_{\le s}\|_\infty > 1/4\big).\end{aligned} \tag{3.7}$$

REMARK 3.6. Note that the convention of Remark 2.3 is still in force. For certain values of r and s, the upper bounds in Proposition 3.5 may depend on the choice of the P_j. The actual choices, however, do not alter the final results in Section 2.3.

REMARK 3.7. The constants are chosen for simplicity. For each $\varepsilon > 0$, the constant 16 in (3.7) can be replaced by $1 + \varepsilon$ provided that the constants $1/2$ and $1/4$ in the definition of the events are replaced by bigger constants depending on ε.

PROOF. Using $\|P_j \hat{P}_{>d}\|_2^2 \le 1$ and $\sum_{j=r+1}^{d}\|P_j \hat{P}_{>d}\|_2^2 \le p - d$, we obtain

$$\mathcal{E}^{\mathrm{PCA}}_{\le d}(\mu) \le \sum_{j\le r}(\lambda_j - \mu)\|P_j \hat{P}_{>d}\|_2^2 + \sum_{j=r+1}^{d\wedge(r+p-d)}(\lambda_j - \mu). \tag{3.8}$$

By Lemma 3.1, we have

$$\|P_j \hat{P}_{>d}\|_2^2 = \sum_{k>d}\frac{\|P_j \Delta \hat{P}_k\|_2^2}{(\lambda_j - \hat{\lambda}_k)^2}.$$

Moreover, on the event

$$\{\hat{\lambda}_{d+1} - \lambda_{d+1} \le (\lambda_j - \lambda_{d+1})/2\} = \{\lambda_j - \hat{\lambda}_{d+1} \ge (\lambda_j - \lambda_{d+1})/2\}$$

we can bound

$$\|P_j\hat{P}_{>d}\|_2^2 \le \sum_{k>d} 4\frac{\|P_j\Delta\hat{P}_k\|_2^2}{(\lambda_j-\lambda_{d+1})^2} = 4\frac{\|P_j\Delta\hat{P}_{>d}\|_2^2}{(\lambda_j-\lambda_{d+1})^2}. \tag{3.9}$$

By (3.9) and $\|P_j\hat{P}_{>d}\|_2^2 \le 1$, we conclude that

$$\|P_j\hat{P}_{>d}\|_2^2 \le 4\frac{\|P_j\Delta\hat{P}_{>d}\|_2^2}{(\lambda_j-\lambda_{d+1})^2} + \mathbb{1}\big(\hat{\lambda}_{d+1}-\lambda_{d+1} > (\lambda_j-\lambda_{d+1})/2\big). \tag{3.10}$$

Inserting (3.10) into (3.8), we obtain the first claim (3.5). The second claim follows from an additional recursion argument. For this, we introduce

$$R_{\le s} = R_{\le s}(\mu) = \sum_{j\le s}\sqrt{\lambda_j-\mu}\,P_j, \tag{3.11}$$

which satisfies the identities $S_{\le s}R_{\le s} = P_{\le s}$ and

$$\sum_{j\le s}(\lambda_j-\mu)\|P_j\hat{P}_{>d}\|_2^2 = \|R_{\le s}\hat{P}_{>d}\|_2^2. \tag{3.12}$$

Then we have

$$\begin{aligned}
&\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta\hat{P}_{>d}\|_2^2}{(\lambda_j-\lambda_{d+1})^2}\\
&\quad\le 2\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta P_{>s}\hat{P}_{>d}\|_2^2}{(\lambda_j-\lambda_{d+1})^2} + 2\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta P_{\le s}\hat{P}_{>d}\|_2^2}{(\lambda_j-\lambda_{d+1})^2}\\
&\quad\le 2\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta P_{>s}\|_2^2}{(\lambda_j-\lambda_{d+1})^2} + 2\sum_{j\le r}\frac{\|P_j\Delta P_{\le s}\hat{P}_{>d}\|_2^2}{\lambda_j-\mu}\\
&\quad= 2\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta P_{>s}\|_2^2}{(\lambda_j-\lambda_{d+1})^2} + 2\|S_{\le r}\Delta P_{\le s}\hat{P}_{>d}\|_2^2.
\end{aligned} \tag{3.13}$$

On the event $\{\|S_{\le s}\Delta S_{\le s}\|_\infty \le 1/4\}$, the last term is bounded via

$$\begin{aligned}
2\|S_{\le r}\Delta P_{\le s}\hat{P}_{>d}\|_2^2 &= 2\|S_{\le r}\Delta S_{\le s}R_{\le s}\hat{P}_{>d}\|_2^2\\
&\le 2\|S_{\le r}\Delta S_{\le s}\|_\infty^2\|R_{\le s}\hat{P}_{>d}\|_2^2\\
&\le 2\|S_{\le s}\Delta S_{\le s}\|_\infty^2\|R_{\le s}\hat{P}_{>d}\|_2^2 \le \|R_{\le s}\hat{P}_{>d}\|_2^2/8,
\end{aligned}$$

where we also used that $r \le s$. Thus, on $\{\|S_{\le s}\Delta S_{\le s}\|_\infty \le 1/4\}$, we get

$$\begin{aligned}
&\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta\hat{P}_{>d}\|_2^2}{(\lambda_j-\lambda_{d+1})^2}\\
&\quad\le 2\sum_{j\le r}(\lambda_j-\mu)\frac{\|P_j\Delta P_{>s}\|_2^2}{(\lambda_j-\lambda_{d+1})^2} + \frac{1}{8}\sum_{j\le s}(\lambda_j-\mu)\|P_j\hat{P}_{>d}\|_2^2.
\end{aligned} \tag{3.14}$$

Using also that $\|P_j \Delta \hat{P}_{>d}\|_2^2 \le \|P_j \Delta\|_2^2$, we conclude that

$$
\begin{aligned}
&4 \sum_{j \le r} (\lambda_j - \mu) \frac{\|P_j \Delta \hat{P}_{>d}\|_2^2}{(\lambda_j - \lambda_{d+1})^2} \\
&\quad \le 8 \sum_{j \le r} (\lambda_j - \mu) \frac{\|P_j \Delta P_{>s}\|_2^2}{(\lambda_j - \lambda_{d+1})^2} + \frac{1}{2} \mathcal{E}_{\le d}^{\mathrm{PCA}}(\mu) \\
&\qquad + 4 \sum_{j \le r} (\lambda_j - \mu) \frac{\|P_j \Delta\|_2^2}{(\lambda_j - \lambda_{d+1})^2} \mathbb{1}\big(\|S_{\le s} \Delta S_{\le s}\|_\infty > 1/4\big).
\end{aligned}
$$

Plugging this into (3.5), we obtain the second claim. □

Similarly, we can upper bound the second risk part $\mathcal{E}_{>d}^{\mathrm{PCA}}$. The only difference in the proof in Appendix A.1 in the Supplementary Material [28] is that an additional argument deals with the sum over all sufficiently large k.

PROPOSITION 3.8. *For $\mu \in [\lambda_{d+1}, \lambda_d]$ and $l = d+1, \dots, p+1$, we have*

$$
\begin{aligned}
\mathcal{E}_{>d}^{\mathrm{PCA}}(\mu) \le{}& 4 \sum_{k \ge l} (\mu - \lambda_k) \frac{\|P_k \Delta \hat{P}_{\le d}\|_2^2}{(\lambda_d - \lambda_k)^2} + \sum_{k=(d+1)\vee(l-d)}^{l-1} (\mu - \lambda_k) \\
&+ \sum_{\substack{k \ge l: \\ \lambda_k \ge \lambda_d/2}} (\mu - \lambda_k) \mathbb{1}\big(\hat{\lambda}_d - \lambda_d < -(\lambda_d - \lambda_k)/2\big) \\
&+ d(\mu - \lambda_p) \mathbb{1}(\hat{\lambda}_d - \lambda_d < -\lambda_d/4).
\end{aligned} \tag{3.15}
$$

Note that for $p = \infty$ the convention of Proposition 2.7 is still in force.

3.3. *Concentration inequalities.* In order to make the deterministic upper bounds of the previous section useful, one has to show that the events in the remainder terms occur with small probability. We establish concentration inequalities for the weighted sample covariance operators as well as deviation inequalities for the empirical eigenvalues $\hat{\lambda}_d$ and $\hat{\lambda}_{d+1}$, based on the concentration inequality [19], Corollary 2, for sample covariance operators which we use in the form

$$
\mathbb{P}\big(\|\Delta\|_\infty > C_3 \lambda_1 x\big) \le e^{-n(x \wedge x^2)}, \tag{3.16}
$$

whenever

$$
\operatorname{tr}(\Sigma) \le n \lambda_1 \big(x \wedge x^2\big),
$$

where $C_3 > 1$ is a constant which depends only on C_1. First, consider the weighted projector $S_{\le s}$ from (3.6) for $\mu \in [0, \lambda_s)$. Then, as in (2.9), $X' = S_{\le s} X$ satisfies Assumption 2.1 with the same constant C_1 as X and has covariance operator

$$
\Sigma' = S_{\le s} \Sigma S_{\le s} = \sum_{j \le s} \frac{\lambda_j}{\lambda_j - \mu} P_j.
$$

The eigenvalues of Σ' (in decreasing order) are $\lambda'_j = \lambda_{s+1-j}/(\lambda_{s+1-j} - \mu)$, noting that the order is reversed by the weighting. Using the sample covariance $\hat{\Sigma}' = S_{\le s} \hat{\Sigma} S_{\le s}$ and choosing $x = 1/(4C_3 \lambda'_1)$, which is smaller than 1, the concentration inequality (3.16) applied to $\Delta' = \Sigma' - \hat{\Sigma}'$ yields the following.

LEMMA 3.9. *Grant Assumption 2.1. If $\mu \in [0, \lambda_s)$ and if*

$$\frac{\lambda_s}{\lambda_s - \mu} \sum_{j \le s} \frac{\lambda_j}{\lambda_j - \mu} \le n/(16C_3^2)$$

holds with the constant C_3 from (3.16), then

$$\mathbb{P}\big(\|S_{\le s} \Delta S_{\le s}\|_\infty > 1/4\big) \le \exp\Big(-\frac{n(\lambda_s - \mu)^2}{16C_3^2 \lambda_s^2}\Big).$$

Next, we will state deviation inequalities for the empirical eigenvalues $\hat\lambda_d$ and $\hat\lambda_{d+1}$, namely right-deviation inequalities for $\hat\lambda_{d+1}$ and left-deviation inequalities for $\hat\lambda_d$.

PROPOSITION 3.10. *Grant Assumption 2.1. For all $x > 0$, satisfying*

$$\max\Big(\frac{C_3\lambda_{d+1}}{x}, 1\Big) \sum_{k>d} \frac{\lambda_k}{\lambda_{d+1} - \lambda_k + x} \le n/C_3, \tag{3.17}$$

we have

$$\mathbb{P}(\hat\lambda_{d+1} - \lambda_{d+1} > x) \le \exp\Big(-n \min\Big(\frac{x^2}{C_3^2\lambda_{d+1}^2}, \frac{x}{C_3\lambda_{d+1}}\Big)\Big),$$

where C_3 is the constant in (3.16).

PROOF. First, we apply the min-max characterisation of eigenvalues and obtain $\hat\lambda_{d+1} \le \lambda_1(P_{>d}\hat\Sigma P_{>d})$. This gives

$$\mathbb{P}(\hat\lambda_{d+1} - \lambda_{d+1} > x) \le \mathbb{P}\big(\lambda_1(P_{>d}\hat\Sigma P_{>d}) - \lambda_1(P_{>d}\Sigma P_{>d}) > x\big). \tag{3.18}$$

We now use the following lemma, proven later.

LEMMA 3.11. *Let S and T be self-adjoint, positive compact operators on $\mathcal{H}$ and $y > \lambda_1(S)$. Then*

$$\lambda_1(T) > y \quad \Longleftrightarrow \quad \lambda_1\big((y - S)^{-1/2}(T - S)(y - S)^{-1/2}\big) > 1.$$

Applying this lemma to $S = P_{>d}\Sigma P_{>d}$, $T = P_{>d}\hat\Sigma P_{>d}$, and $y = \lambda_1(S) + x = \lambda_{d+1} + x$, we get

$$\mathbb{P}\big(\lambda_1(P_{>d}\hat\Sigma P_{>d}) - \lambda_1(P_{>d}\Sigma P_{>d}) > x\big) \le \mathbb{P}\big(\|T_{>d}\Delta T_{>d}\|_\infty > 1\big) \tag{3.19}$$

with

$$T_{>d} = \sum_{k>d} \frac{1}{\sqrt{\lambda_{d+1} - \lambda_k + x}} P_k.$$

Thus, as in (2.9), we consider $X' = T_{>d}X$, satisfying Assumption 2.1 with the same constant C_1, and obtain the covariance operator

$$\Sigma' = T_{>d}\Sigma T_{>d} = \sum_{k>d} \frac{\lambda_k}{\lambda_{d+1} - \lambda_k + x} P_k.$$

Hence choosing

$$x' = \frac{1}{C_3\lambda_1'} = \frac{x}{C_3\lambda_{d+1}},$$

the concentration inequality (3.16), applied to $\Delta' = T_{>d}\Delta T_{>d}$ and x', gives

$$\mathbb{P}\big(\|T_{>d}\Delta T_{>d}\|_\infty > 1\big) \le \exp\Big(-n\min\Big(\frac{x^2}{C_3^2\lambda_{d+1}^2}, \frac{x}{C_3\lambda_{d+1}}\Big)\Big) \tag{3.20}$$

in view of Condition (3.17). Combining (3.18)–(3.20), the claim follows.

It remains to prove Lemma 3.11. We have

$$\lambda_1\big((y-S)^{-1/2}(T-S)(y-S)^{-1/2}\big) \le 1$$

if and only if (for a linear operator $L : \mathcal{H} \to \mathcal{H}$, we write $L \ge 0$ if L is positive, that is, if $\langle Lx, x\rangle \ge 0$ for all $x \in \mathcal{H}$)

$$(y-S)^{-1/2}(y-T)(y-S)^{-1/2} = I - (y-S)^{-1/2}(T-S)(y-S)^{-1/2} \ge 0.$$

Since $(y-S)^{-1/2}$ is self-adjoint and strictly positive, this is the case if and only if $y - T \ge 0$, that is, $\lambda_1(T) \le y$. A logical negation yields the assertion of the lemma. □

In view of the error decompositions (3.5) and (3.7), we want to apply Proposition 3.10 with $x = (\lambda_j - \lambda_{d+1})/2$, $j \le d$. For this, we require

$$\max\Big(\frac{2C_3\lambda_{d+1}}{\lambda_j - \lambda_{d+1}}, 1\Big)\sum_{k>d}\frac{\lambda_k}{\lambda_j - \lambda_k} \le n/(2C_3).$$

Simplifying the maximum yields the following.

COROLLARY 3.12. *Grant Assumption 2.1 and let $j \le d$. Suppose that*

$$\frac{\lambda_j}{\lambda_j - \lambda_{d+1}}\sum_{k>d}\frac{\lambda_k}{\lambda_j - \lambda_k} \le n/\big(4C_3^2\big). \tag{3.21}$$

Then

$$\mathbb{P}\Big(\hat\lambda_{d+1} - \lambda_{d+1} > \frac{\lambda_j - \lambda_{d+1}}{2}\Big) \le \exp\Big(-\frac{n(\lambda_j - \lambda_{d+1})^2}{4C_3^2\lambda_j^2}\Big). \tag{3.22}$$

The corresponding left-deviation result for $\hat\lambda_d$ is proved in Appendix A.2 in the Supplementary Material [28].

PROPOSITION 3.13. *Grant Assumption 2.1. For all $x > 0$ satisfying*

$$\max\Big(\frac{C_3\lambda_d}{x}, 1\Big)\sum_{j\le d}\frac{\lambda_j}{\lambda_j - \lambda_d + x} \le n/C_3, \tag{3.23}$$

we have

$$\mathbb{P}(\hat\lambda_d - \lambda_d < -x) \le \exp\Big(-n\min\Big(\frac{x^2}{C_3^2\lambda_d^2}, \frac{x}{C_3\lambda_d}\Big)\Big),$$

where C_3 is the constant in (3.16).

In particular, choosing $x = (\lambda_d - \lambda_k)/2$, we get the following.

COROLLARY 3.14. *Grant Assumption 2.1 and let $k > d$. Suppose that*

$$\frac{\lambda_d}{\lambda_d - \lambda_k} \sum_{j \le d} \frac{\lambda_j}{\lambda_j - \lambda_k} \le n/(4C_3^2). \tag{3.24}$$

Then

$$\mathbb{P}\big(\hat\lambda_d - \lambda_d < -(\lambda_d - \lambda_k)/2\big) \le \exp\Big(-\frac{n(\lambda_d - \lambda_k)^2}{4C_3^2 \lambda_d^2}\Big). \tag{3.25}$$

REMARK 3.15. Let us consider the important special case $j = s = d$, $k = d + 1$ and $\mu = \lambda_{d+1}$. Then all three conditions in Lemma 3.9 and Corollaries 3.12, 3.14 are implied by

$$\frac{\lambda_d}{\lambda_d - \lambda_{d+1}} \Big(\sum_{j \le d} \frac{\lambda_j}{\lambda_j - \lambda_{d+1}} + \sum_{k > d} \frac{\lambda_k}{\lambda_d - \lambda_k} \Big) \le n/(16C_3^2). \tag{3.26}$$

In particular, if (3.26) holds, then all events in the remainder terms in Propositions 3.5 and 3.8 occur with small probability.

The localised analysis of this section can be compared to the following absolute one. All events considered in Lemma 3.9 and Corollaries 3.12, 3.14 are contained in $\{\|\Delta\|_\infty > (\lambda_d - \lambda_{d+1})/4\}$ and by (3.16) this occurs with small probability if

$$\frac{\lambda_1 \operatorname{tr}(\Sigma)}{(\lambda_d - \lambda_{d+1})^2} \le n/(16C_3^2). \tag{3.27}$$

Note that the condition that $\|\Delta\|_\infty$ is small relative to certain spectral gaps, here $\lambda_d - \lambda_{d+1}$, is often encountered in perturbation theory; see, for example, [5], Theorem VII.3.1, [13], Theorems 5.1.4 and 5.1.8, and [18], Lemma 1. Many of our mathematical issues arise from showing that Condition (3.27) can be replaced by the localised version in (3.26).

REMARK 3.16. Our concentration inequalities rely on Assumption 2.1. Generalisations are possible under weaker moment assumptions, including $\sup_{j \ge 1} \mathbb{E}[|\lambda_j^{-1/2} \langle X, u_j \rangle|^k] < \infty$ for some $k > 4$. Since the latter seemingly leads to stronger eigenvalue conditions than formulated in Lemma 3.9 and Corollaries 3.12, 3.14, such generalisations are not pursued here.

4. Proofs. In this section, we provide the proofs for the results in Section 2.3, by combining the error decompositions in Section 3.2 with the concentration inequalities in Section 3.3.

4.1. *Proof of Lemma 2.6.* Inserting the spectral representation of Σ, the excess risk can be written as

$$\mathcal{E}_d^{\mathrm{PCA}} = \langle \Sigma, P_{\le d} - \hat P_{\le d} \rangle = \sum_{j \ge 1} \lambda_j \langle P_j, P_{\le d} - \hat P_{\le d} \rangle.$$

By $P_{\le d} - \hat P_{\le d} = \hat P_{> d} - P_{> d}$, we obtain

$$\begin{aligned} \mathcal{E}_d^{\mathrm{PCA}} &= \sum_{j \le d} \lambda_j \langle P_j, \hat P_{> d} - P_{> d} \rangle - \sum_{k > d} \lambda_k \langle P_k, \hat P_{\le d} - P_{\le d} \rangle \\ &= \sum_{j \le d} \lambda_j \langle P_j, \hat P_{> d} \rangle - \sum_{k > d} \lambda_k \langle P_k, \hat P_{\le d} \rangle. \end{aligned}$$

Moreover, the identity

$$\langle P_{\le d}, \hat P_{> d} \rangle = \langle P_{\le d}, \hat P_{> d} - P_{> d} \rangle = -\langle P_{> d}, P_{\le d} - \hat P_{\le d} \rangle = \langle P_{> d}, \hat P_{\le d} \rangle$$

implies $\sum_{j\le d}\mu\langle P_j, \hat{P}_{>d}\rangle = \sum_{k>d}\mu\langle P_k, \hat{P}_{\le d}\rangle$, and thus

$$\mathcal{E}_d^{\mathrm{PCA}} = \sum_{j\le d}(\lambda_j-\mu)\langle P_j, \hat{P}_{>d}\rangle + \sum_{k>d}(\mu-\lambda_k)\langle P_k, \hat{P}_{\le d}\rangle.$$

The claim now follows from inserting the identities $\langle P_j, \hat{P}_{>d}\rangle = \|P_j\hat{P}_{>d}\|_2^2$ and $\langle P_k, \hat{P}_{\le d}\rangle = \|P_k\hat{P}_{\le d}\|_2^2$. □

4.2. *Proof of Proposition 2.7.* We only prove the first inequality. The proof of the second one follows the same line of arguments and is given in Appendix A.3 in the Supplementary Material [28]. Taking expectation in (3.5) and using $\mathbb{E}[\|P_j\Delta\hat{P}_{>d}\|_2^2] \le \mathbb{E}[\|P_j\Delta\|_2^2] \le 2C_2\lambda_j\operatorname{tr}(\Sigma)/n$, we get

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}_{\le d}^{\mathrm{PCA}}(\mu)\big] \le 8C_2\sum_{j\le r}(\lambda_j-\mu)\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2} + \sum_{j=r+1}^{d\wedge(r+p-d)}(\lambda_j-\mu)\\ + \sum_{j\le r}(\lambda_j-\mu)\mathbb{P}\big(\hat{\lambda}_{d+1}-\lambda_{d+1} > (\lambda_j-\lambda_{d+1})/2\big).\end{aligned}$$

Hence, the first inequality follows from the following lemma.

LEMMA 4.1. *Let $j \le d$. Then*

$$\begin{aligned}&8C_2\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2} + \mathbb{P}\big(\hat{\lambda}_{d+1}-\lambda_{d+1} > (\lambda_j-\lambda_{d+1})/2\big)\\ &\quad\le \big(8C_2+4C_3^2\big)\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2}.\end{aligned}$$

PROOF OF LEMMA 4.1. If

$$\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2} \le 1/\big(4C_3^3\big), \tag{4.1}$$

then Condition (3.21) is satisfied and we can apply (3.22). Thus, in this case, the left-hand side can be bounded by

$$8C_2\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2} + \exp\bigg(-\frac{n(\lambda_j-\lambda_{d+1})^2}{4C_3^2\lambda_j^2}\bigg) \le \big(8C_2+2C_3^2\big)\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2},$$

where the inequality follows from $x\exp(-x) \le 1/e \le 1/2$, $x \ge 0$. On the other hand, if (4.1) is not satisfied, then the left-hand side can be bounded by

$$8C_2\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2} + 1 \le \big(8C_2+4C_3^2\big)\frac{\lambda_j\operatorname{tr}(\Sigma)}{n(\lambda_j-\lambda_{d+1})^2}.$$

Hence, we get the claim in both cases. □

4.3. *Proof of Corollary 2.8.* The claim follows from Proposition 2.7 together with the facts that for $\mu \in [\lambda_{d+1}, \lambda_d]$ the terms $\lambda_j-\mu$ (resp., $\mu-\lambda_k$) can be upper bounded by $\lambda_j-\lambda_{d+1}$ (resp., $\lambda_d-\lambda_k$) and that $\lambda \mapsto \lambda/(\lambda-\lambda_{d+1})^2$ is decreasing for $\lambda > \lambda_{d+1}$ (resp., $\lambda \mapsto \lambda/(\lambda_d-\lambda)^2$ is increasing for $\lambda < \lambda_d$). □

4.4. *Proof of Theorem 2.9.* In Corollary 2.8, only summands with $\lambda_j > \lambda_{d+1}$ and $\lambda_k < \lambda_d$, respectively, appear. Neglecting the minimum with $\lambda_j - \lambda_{d+1}$ (resp., $\lambda_d - \lambda_k$) in each summand, the local bound follows.

For the global bound use, the inequality $\min(a/x, x) \le \sqrt{a}$ for $a, x \ge 0$ to obtain from Corollary 2.8

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_{\le d}(\mu)\big] \le \sum_{j \le d} \sqrt{\frac{C\lambda_j \operatorname{tr}(\Sigma)}{n}}.$$

Considering $\mathcal{E}^{\mathrm{PCA}}_{>d}(\mu)$, the value l in Proposition 2.7 has to be chosen carefully. For $a > 0$, let $d < l = l(a) \le p+1$ be the index such that $\lambda_d - \lambda_k \ge a$ for $k \ge l$ and $\lambda_d - \lambda_k < a$ for $d < k < l$. Then the second inequality of Proposition 2.7 and the inequality $\mu \le \lambda_d$ imply

$$\mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_{>d}(\mu)\big] \le \sum_{k>d} \frac{C\lambda_k \operatorname{tr}(\Sigma)}{na} + da + \lambda_d e^{-\frac{n}{32C_3^2}}.$$

Minimizing over $a > 0$ and incorporating the remainder in the summand for $j = d$ gives the global bound in (2.10). □

4.5. *Proof of Proposition 2.10.* We begin with the following extension of Lemma 4.1, proved in Appendix A.4 in the Supplementary Material [28].

LEMMA 4.2. *Let $j \le s \le d$. Then*

$$\begin{aligned} &16C_2 \frac{\lambda_j \operatorname{tr}_{>s}(\Sigma)}{n(\lambda_j - \lambda_{d+1})^2} + 2\mathbb{P}\big(\hat{\lambda}_{d+1} - \lambda_{d+1} > (\lambda_j - \lambda_{d+1})/2\big) \\ &\quad \le \big(16C_2 + 8C_3^2\big) \frac{\lambda_j \operatorname{tr}_{>s}(\Sigma)}{n(\lambda_j - \lambda_{d+1})^2}. \end{aligned} \tag{4.2}$$

Taking expectation in (3.7) with $\mu = \lambda_{d+1}$ and using Lemma 4.2, we obtain

$$\begin{aligned} \mathbb{E}\big[\mathcal{E}^{\mathrm{PCA}}_{\le d}(\lambda_{d+1})\big] &\le \big(16C_2 + 8C_3^2\big) \sum_{j \le r} \frac{\lambda_j \operatorname{tr}_{>s}(\Sigma)}{n(\lambda_j - \lambda_{d+1})} + 2 \sum_{r < j \le d} (\lambda_j - \lambda_{d+1}) \\ &\quad + 8\mathbb{E}\bigg[\sum_{j \le r} \frac{\|P_j \Delta\|_2^2}{\lambda_j - \lambda_{d+1}} \mathbb{1}\big(\|S_{\le s} \Delta S_{\le s}\|_\infty > 1/4\big)\bigg]. \end{aligned}$$

If Condition (2.11) holds, then Lemma 3.9 with $\mu = \lambda_{d+1}$ gives

$$\mathbb{P}\big(\|S_{\le s} \Delta S_{\le s}\|_\infty > 1/4\big) \le \exp\bigg(-\frac{n(\lambda_s - \lambda_{d+1})^2}{16C_3^2 \lambda_s^2}\bigg). \tag{4.3}$$

Thus the claim follows from applying the Cauchy–Schwarz inequality, (4.3), and the following lemma.

LEMMA 4.3. *For all $r \le d$, we have*

$$\bigg(\mathbb{E}\bigg[\bigg(\sum_{j \le r} \frac{\|P_j \Delta\|_2^2}{\lambda_j - \lambda_{d+1}}\bigg)^2\bigg]\bigg)^{1/2} \le 128 C_1^4 \sum_{j \le r} \frac{\lambda_j \operatorname{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})}.$$

Lemma 4.3 follows from the Minkowski inequality and Assumption 2.1, see Appendix A.5 in the Supplementary Material [28] for the details. □

4.6. *Proof of Theorem 2.11.* By assumption, we have $\lambda_j - \lambda_p \le c_1^{-1}(\lambda_j - \lambda_{d+1})$ for all $j \le d$. Thus, Lemma 2.6 applied with $\mu = \lambda_p$ yields

$$\mathcal{E}_d^{\mathrm{PCA}} \le \sum_{j \le d} (\lambda_j - \lambda_p) \| P_j \hat{P}_{>d} \|_2^2 \le c_1^{-1} \mathcal{E}_{\le d}^{\mathrm{PCA}}(\lambda_{d+1}). \tag{4.4}$$

The local bound now follows from Proposition 2.10 applied with $r = s = d$. The proof of the global is more technical and given in Appendix A.6 in the Supplementary Material [28]. □

4.7. *Proof of Theorem 2.12.* Similarly as in (4.4), we have

$$\mathcal{E}_d^{\mathrm{PCA}} \le \sum_{j \le d} \lambda_j \| P_j \hat{P}_{>d} \|_2^2 \le \frac{\lambda_s}{\lambda_s - \lambda_{d+1}} \sum_{j \le s} (\lambda_j - \lambda_{d+1}) \| P_j \hat{P}_{>d} \|_2^2 + \mathrm{tr}_{>s}(\Sigma).$$

By (3.10) and (3.14) with $\mu = \lambda_{d+1}$ and $r = s$, we have

$$\begin{aligned}
&\sum_{j \le s} (\lambda_j - \lambda_{d+1}) \| P_j \hat{P}_{>d} \|_2^2 \\
&\quad \le 16 \sum_{j \le s} \frac{\| P_j \Delta P_{>s} \|_2^2}{\lambda_j - \lambda_{d+1}} + 2 \sum_{j \le s} (\lambda_j - \lambda_{d+1}) \mathbb{1}\big(\hat{\lambda}_{d+1} - \lambda_{d+1} > (\lambda_j - \lambda_{d+1})/2\big) \\
&\qquad + 8 \sum_{j \le s} \frac{\| P_j \Delta \|_2^2}{\lambda_j - \lambda_{d+1}} \mathbb{1}\big(\| S_{\le s} \Delta S_{\le s} \|_\infty > 1/4\big).
\end{aligned}$$

As shown in the proof of Proposition 2.10 the inequality

$$\mathbb{E}\bigg[\sum_{j \le s} (\lambda_j - \lambda_{d+1}) \| P_j \hat{P}_{>d} \|_2^2 \bigg] \le C \sum_{j \le s} \frac{\lambda_j \, \mathrm{tr}_{>s}(\Sigma)}{n(\lambda_j - \lambda_{d+1})} + R$$

holds with remainder term R given in Proposition 2.10 with $r = s$. Inserting this into the above inequality and using Condition (2.11), we get

$$\mathbb{E}[\mathcal{E}_d^{\mathrm{PCA}}] \le C \, \mathrm{tr}_{>s}(\Sigma) + C \, \mathrm{tr}(\Sigma) \exp\bigg(-\frac{n(\lambda_s - \lambda_{d+1})^2}{16 C_3^2 \lambda_s^2} \bigg)$$

with a constant $C > 0$ depending only on C_1. □

Acknowledgements. We are grateful for the helpful comments by the two anonymous referees.

The second author was supported in part by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 "Data Assimilation", Project (A4) "Nonlinear statistical inverse problems with random observations".

SUPPLEMENTARY MATERIAL

Supplement to "Nonasymptotic upper bounds for the reconstruction error of PCA" (DOI: 10.1214/19-AOS1839SUPP; .pdf). The supplement includes a section with additional proofs and a section on linear expansions.

REFERENCES

[1] ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An Introduction to Random Matrices*. *Cambridge Studies in Advanced Mathematics* **118**. Cambridge Univ. Press, Cambridge. MR2760897

[2] ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0771294

[3] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. MR2166554 https://doi.org/10.1214/009053605000000282

[4] BERCU, B., DELYON, B. and RIO, E. (2015). *Concentration Inequalities for Sums and Martingales*. *Springer Briefs in Mathematics*. Springer, Cham. MR3363542 https://doi.org/10.1007/978-3-319-22099-4

[5] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662 https://doi.org/10.1007/978-1-4612-0653-8

[6] BLANCHARD, G., BOUSQUET, O. and ZWALD, L. (2007). Statistical properties of kernel principal component analysis. *Mach. Learn.* **66** 259–294.

[7] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[8] CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

[9] DAUXOIS, J., POUSSE, A. and ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154. MR0650934 https://doi.org/10.1016/0047-259X(82)90088-4

[10] DAVIS, C. and KAHAN, W. M. (1969). Some new bounds on perturbation of subspaces. *Bull. Amer. Math. Soc.* **75** 863–868. MR0246155 https://doi.org/10.1090/S0002-9904-1969-12330-X

[11] HALL, P. and HOSSEINI-NASAB, M. (2009). Theory for high-order bounds in functional principal components analysis. *Math. Proc. Cambridge Philos. Soc.* **146** 225–256. MR2461880 https://doi.org/10.1017/S0305004108001850

[12] HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. *Springer Series in Statistics*. Springer, New York. MR2920735 https://doi.org/10.1007/978-1-4614-3655-3

[13] HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis*, *with an Introduction to Linear Operators*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

[14] JIRAK, M. (2016). Optimal eigen expansions and uniform bounds. *Probab. Theory Related Fields* **166** 753–799. MR3568039 https://doi.org/10.1007/s00440-015-0671-3

[15] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

[16] JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2036084

[17] KATO, T. (1995). *Perturbation Theory for Linear Operators*. *Classics in Mathematics*. Springer, Berlin. Reprint of the 1980 edition. MR1335452

[18] KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* **52** 1976–2013. MR3573302 https://doi.org/10.1214/15-AIHP705

[19] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

[20] KOLTCHINSKII, V. and LOUNICI, K. (2017). New asymptotic results in principal component analysis. *Sankhya A* **79** 254–297. MR3707422 https://doi.org/10.1007/s13171-017-0106-6

[21] KOLTCHINSKII, V. and LOUNICI, K. (2017). Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.* **45** 121–157. MR3611488 https://doi.org/10.1214/16-AOS1437

[22] LAX, P. D. (2002). *Functional Analysis*. Wiley, New York.

[23] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon*. *Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[24] LEDOUX, M. (2007). Deviation inequalities on largest eigenvalues. In *Geometric Aspects of Functional Analysis*. *Lecture Notes in Math.* **1910** 167–219. Springer, Berlin. MR2349607 https://doi.org/10.1007/978-3-540-72053-9_10

[25] MAS, A. and RUYMGAART, F. (2015). High-dimensional principal projections. *Complex Anal. Oper. Theory* **9** 35–63. MR3300524 https://doi.org/10.1007/s11785-014-0371-5

[26] MECKES, M. W. (2004). Concentration of norms and eigenvalues of random matrices. *J. Funct. Anal.* **211** 508–524. MR2057479 https://doi.org/10.1016/S0022-1236(03)00198-8

[27] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013 https://doi.org/10.1214/08-AOS618

[28] REISS, M. and WAHL, M. (2020). Supplement to "Nonasymptotic upper bounds for the reconstruction error of PCA." https://doi.org/10.1214/19-AOS1839SUPP.

[29] SCHÖLKOPF, B. and SMOLA, A. J. (2001). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press, Cambridge, MA.

[30] SHAWE-TAYLOR, J., WILLIAMS, C., CRISTIANINI, N. and KANDOLA, J. (2002). On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **2533** 23–40. Springer, Berlin. MR2071605 https://doi.org/10.1007/3-540-36169-3_4

[31] SHAWE-TAYLOR, J., WILLIAMS, C. K. I., CRISTIANINI, N. and KANDOLA, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans*. *Inform*. *Theory* **51** 2510–2522. MR2246374 https://doi.org/10.1109/TIT.2005.850052

[32] VAKHANIA, N. N., TARIELADZE, V. I. and CHOBANYAN, S. A. (1987). *Probability Distributions on Banach Spaces*. *Mathematics and Its Applications* (*Soviet Series*) **14**. D. Reidel Publishing Co., Dordrecht. MR1435288 https://doi.org/10.1007/978-94-009-3873-1

[33] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[34] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann*. *Statist*. **41** 2905–2947. MR3161452 https://doi.org/10.1214/13-AOS1151

[35] WAHL, M. (2018). A note on the prediction error of principal component regression. Available at https://arxiv.org/pdf/1811.02998.

[36] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006 https://doi.org/10.1093/biomet/asv008