PERTURBATION BOUNDS FOR EIGENSPACES UNDER A RELATIVE GAP CONDITION

MORITZ JIRAK AND MARTIN WAHL

ABSTRACT. A basic problem in operator theory is to estimate how a small perturbation affects the eigenspaces of a self-adjoint compact operator. In this paper, we prove upper bounds for the subspace distance, taylored for relative perturbations. As a main example, we consider the empirical covariance operator, and show that a sharp bound can be achieved under a relative gap condition. The proof is based on a novel contraction phenomenon, contrasting previous spectral perturbation approaches.

1. INTRODUCTION

Let Σ be a positive self-adjoint compact operator on a separable Hilbert space $\mathcal{H}$. 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 $\Sigma = \sum_{i \geq 1} \lambda_i u_i \otimes u_i$. For $u, v \in \mathcal{H}$, we denote by $u \otimes v$ the rank-one operator defined by $(u \otimes v)x = \langle v, x \rangle u$, $x \in \mathcal{H}$.

Let $\hat{\Sigma}$ be another positive self-adjoint compact operator on $\mathcal{H}$. We consider $\hat{\Sigma}$ as a perturbed version of Σ and write $E = \hat{\Sigma} - \Sigma$ for the perturbation, which will be thought of as small. Again, by the spectral theorem, there exists a sequence $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots > 0$ of positive eigenvalues, together with an orthonormal system of eigenvectors $\hat{u}_1, \hat{u}_2, \ldots$ such that $\hat{\Sigma} = \sum_{i \geq 1} \hat{\lambda}_i \hat{u}_i \otimes \hat{u}_i$.

Given a finite subset $\mathcal{I} \subseteq \mathbb{N}$, a basic problem is to bound the distance between the eigenspaces $U_{\mathcal{I}} = \operatorname{span}(u_i : i \in \mathcal{I})$ and $\hat{U}_{\mathcal{I}} = \operatorname{span}(\hat{u}_i : i \in \mathcal{I})$. Letting $P_{\mathcal{I}} = \sum_{i \in \mathcal{I}} u_i \otimes u_i$ and $\hat{P}_{\mathcal{I}} = \sum_{i \in \mathcal{I}} \hat{u}_i \otimes \hat{u}_i$ be the orthogonal projections onto $U_{\mathcal{I}}$ and $\hat{U}_{\mathcal{I}}$, respectively, a natural distance is given by the Hilbert-Schmidt distance $\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2$, which is equal to $\sqrt{2}$ times the Euclidean norm of the sines of the canonical angles between the corresponding subspaces, see e.g. [3, Chapter VII.1]).

A first answer to this problem is given by the Davis-Kahan $\sin\Theta$ theorem, a version of which commonly used in probability and statistics reads as

$$\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2 \leq 2\sqrt{2}\|E\|_2 / g_{\mathcal{I}}, \quad \text{with } g_{\mathcal{I}} = \min_{i \in \mathcal{I}, j \notin \mathcal{I}} |\lambda_i - \lambda_j|, \tag{1.1}$$

see e.g. [8, 3, 29], where (1.1) is proven in [29] for the case that $\mathcal{I}$ is an interval. Quantity $\|E\|_2$ is often replaced with $\sqrt{|\mathcal{I}|}$ times the operator norm $\|E\|_\infty$.

More recently, there has been increasing interest in the case where $\hat{\Sigma}$ arises from Σ by random perturbation. In this regard, one of the most prominent examples is the empirical covariance operator, a central object in high-dimensional probability due to its importance in statistics and machine learning. The stochastic nature of

2010 *Mathematics Subject Classification.* 15A42, 47A55, 62H25.

Key words and phrases. Relative perturbation bounds, eigenspace, covariance operator.

this problem leaves room for significant improvements of (1.1), see e.g. [2, 7, 10, 21, 18, 22]. Combined with tools from probability theory, a powerful machinery to derive more precise perturbation results is given by the holomorphic functional calculus for linear operators, see e.g. [17, 6, 12]. For instance, assuming that

$$\delta_{\mathcal{I}} = 2\|E\|_\infty / g_{\mathcal{I}} < 1, \tag{1.2}$$

we have, under the convention in Section 1.1, the first order perturbation expansion

$$\hat{P}_{\mathcal{I}} - P_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \sum_{j\notin\mathcal{I}} \frac{1}{\lambda_i - \lambda_j}(P_i E P_j + P_j E P_i) + S_{\mathcal{I}}(E) \tag{1.3}$$

with remainder term satisfying $\|S_{\mathcal{I}}(E)\|_\infty \le |\mathcal{I}|\delta_{\mathcal{I}}^2/(1-\delta_{\mathcal{I}})$, cf. [12] or [18]. The first term on the right-hand side of (1.3) represents a first order approximation for $\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}$. While its Hilbert-Schmidt norm is usually of smaller magnitude than the upper bound in (1.1), the main drawback of this approach is the requirement (1.2).

Let us illustrate this in the special case where Σ and $\hat{\Sigma}$ are the population and the empirical covariance operator, respectively (see Section 3 below). Then Lemma 1 below shows that under mild assumptions, the squared Hilbert-Schmidt norm of the first order approximation satisfies

$$\sum_{i\in\mathcal{I}} \sum_{j\notin\mathcal{I}} \frac{2\|P_i E P_j\|_2^2}{(\lambda_i - \lambda_j)^2} \le C\frac{\log(n)}{n} \sum_{i\in\mathcal{I}} \sum_{j\notin\mathcal{I}} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \tag{1.4}$$

with high probability, where $C > 0$ is a constant. A key feature of reproducing kernel Hilbert spaces and functional data approaches in machine learning and statistics are eigenvalues with an exponential or polynomial decay. For instance, for exponentially decaying eigenvalues and the choice $\mathcal{I} = \{1, \dots, k\}$, $k \ge 1$, the right-hand side of (1.4) is of order $\log(n)/n$, while $\delta_{\mathcal{I}}$ explodes exponentially in k, meaning that (1.2) is quickly violated and the above approach breaks down. It is thus natural to ask whether the first order approximation in (1.3) still gives accurate bounds (with high probability), if (1.2) is no longer satisfied.

The aim of this paper is to provide an affirmative answer to this question, with a view towards empirical covariance operators. Our main finding is that sharp bounds of the type (1.4) can be derived for $\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2$, replacing (1.2) with a relative gap condition. This is achieved by exploring a novel contraction phenomenon, bypassing arguments based on the holomorphic functional calculus.

The paper is organised as follows. In Section 2.1 we derive perturbation bounds in the case that certain relative coefficients (resp. blocks) are bounded. These bounds are deduced from a more general statement, given in Section 2.2. Section 3.1 presents our main applications to the empirical covariance operator. Besides, our approach can deal with a variety of other structured random perturbations. To illustrate this further, we also discuss random perturbations of low rank matrices in Section 3.2. Finally, the proof of our main result is given in Section 4.

1.1. **Further notation.** Let $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the inner product and the norm on $\mathcal{H}$, respectively. Let $p = \dim \mathcal{H}$ be the dimension of $\mathcal{H}$. Abusing notation, an index $i \in \mathbb{N}$ or a set $\mathcal{I} \subseteq \mathbb{N}$ of indices is to be understood as a subset of $\{1, \dots, p\}$ if p is finite. The set $\mathcal{I}^c$ denotes the complement of $\mathcal{I}$ (with respect to $\{1, \dots, p\}$ if p is finite). For $i \ge 1$, we write $P_i = u_i \otimes u_i$ and $\hat{P}_i = \hat{u}_i \otimes \hat{u}_i$. Hence for $\mathcal{I} \subseteq \mathbb{N}$, we have $P_{\mathcal{I}} = \sum_{i\in\mathcal{I}} P_i$ and $\hat{P}_{\mathcal{I}} = \sum_{i\in\mathcal{I}} \hat{P}_i$. If $p < \infty$, then we extend the

sequence of eigenvalues of Σ and $\hat{\Sigma}$ by adding zeros such that the corresponding eigenvectors form an orthonormal basis of $\mathcal{H}$. (We proceed similarly if $\hat{\Sigma}$ is finite-rank.) If $p = \infty$, then we assume (without loss of generality) that the eigenvectors $u_1, u_2, \dots$ form an orthonormal basis of $\mathcal{H}$. Thus we always have $\sum_{i\geq 1} P_i = I$. Given a bounded (resp. Hilbert-Schmidt) operator A on $\mathcal{H}$, we write $\|A\|_\infty$ (resp. $\|A\|_2$) for the operator norm (resp. the Hilbert-Schmidt norm). Given a trace class operator A on $\mathcal{H}$, we denote the trace of A by $\mathrm{tr}(A)$.

2. Main results

2.1. **Sharp relative perturbation bounds.**

We assume throughout Section 2.1 that the eigenvalues (λ_i) are strictly positive and summable, meaning that Σ is a strictly positive, self-adjoint trace class operator. We begin with introducing the crucial relative eigenvalue separation measure.

Definition 1. For a subset $\mathcal{I} \subseteq \mathbb{N}$, we define

$$\mathbf{r}_{\mathcal{I}}(\Sigma) = \sum_{i\in\mathcal{I}} \frac{\lambda_i}{\min_{j\notin\mathcal{I}} |\lambda_i - \lambda_j|} + \sum_{j\notin\mathcal{I}} \frac{\lambda_j}{\min_{i\in\mathcal{I}} |\lambda_j - \lambda_i|}.$$

The quantity $\mathbf{r}_{\mathcal{I}}(\Sigma)$ measures in a weighted way how well the eigenvalues in $(\lambda_i)_{i\in\mathcal{I}}$ are separated from the rest of the spectrum. Let us consider two examples. First, for $k \geq 1$, we have

$$\mathbf{r}_{\{i:\lambda_i=\lambda_k\}}(\Sigma) = \frac{m_k\lambda_k}{g_k} + \sum_{j:\lambda_j\neq\lambda_k} \frac{\lambda_j}{|\lambda_j - \lambda_k|} \tag{2.1}$$

with multiplicity $m_k = |\{i : \lambda_i = \lambda_k\}|$ and gap $g_k = \min_{i:\lambda_i\neq\lambda_k} |\lambda_i - \lambda_k|$. Second, for $k \geq 1$, we have

$$\mathbf{r}_{\{1,\dots,k\}}(\Sigma) = \sum_{i\leq k} \frac{\lambda_i}{\lambda_i - \lambda_{k+1}} + \sum_{j>k} \frac{\lambda_j}{\lambda_k - \lambda_j}. \tag{2.2}$$

The expressions in (2.1) and (2.2) can be easily evaluated if the λ_j have e.g. exponential or polynomial decay (cf. Section 3). We now state our first main result.

Theorem 1. *Let $\mathcal{I} \subseteq \mathbb{N}$ be finite. Suppose that there is a real number $x > 0$ such that for all $i, j \geq 1$,*

$$\|P_i E P_j\|_2 \leq x\sqrt{\lambda_i\lambda_j}. \tag{2.3}$$

If

$$\mathbf{r}_{\mathcal{I}}(\Sigma) \leq 1/(8x), \tag{2.4}$$

then we have

$$\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2^2 \leq 16x^2 \sum_{i\in\mathcal{I}} \sum_{j\notin\mathcal{I}} \frac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}. \tag{2.5}$$

Remark 1. The numerical constants in (2.4) and (2.5) are selected for convenience.

Remark 2. Motivated by the empirical covariance operator, Theorem 1 considers a perturbation problem where the perturbation E is related to Σ. There is, however, also a connection to numerical analysis. If p is finite, then x can be chosen as the maximum of the absolute values of the $\langle u_i, Eu_j\rangle/\sqrt{\lambda_i\lambda_j}$, $i, j \in \{1, \dots, p\}$. These quantities are the coefficients of the so called relative perturbation $\Sigma^{-1/2}E\Sigma^{-1/2}$ with respect to the eigenvectors of Σ. The latter matrix plays a prominent role in

relative perturbation theory, see e.g. [13, 14]. The novel ingredient of Theorem 1 is Condition (2.4), ensuring that sharp bounds can be derived. Indeed, (2.5) gives the size of the squared Hilbert-Schmidt norm of the first order approximation in (1.3), provided that the bounds in (2.3) are sufficiently tight.

Remark 3. An inspection of the proof shows that the inequality

$$\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2^2 \leq 8 \sum_{i\in\mathcal{I}} \sum_{j\notin\mathcal{I}} \frac{\|P_i E P_j\|_2^2}{(\lambda_i - \lambda_j)^2} + 512 x^4 \mathbf{r}_{\mathcal{I}}^2(\Sigma) \sum_{i\in\mathcal{I}} \sum_{j\notin\mathcal{I}} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \tag{2.6}$$

holds, from which (2.5) follows by inserting (2.3) and (2.4).

Next, we state the following generalization of Theorem 1, more suitable for infinite-dimensional Hilbert spaces:

Theorem 2. *Let $\mathcal{I} \subseteq \mathbb{N}$ be finite. Write $\mathcal{I} = \dot{\cup}_{r\leq m}\mathcal{I}_r$ such that for all $r \leq m$ and all $i, j \in \mathcal{I}_r$ we have $\lambda_i = \lambda_j$. Let $\mathcal{I}' \subseteq \mathbb{N}$ be another finite subset such that $|\lambda_i - \lambda_j| \geq \lambda_i/2$ for all $i \in \mathcal{I}$ and all $j \notin \mathcal{I}'$. Write $\mathcal{I}' \setminus \mathcal{I} = \dot{\cup}_{m<r\leq m+n}\mathcal{I}_r$ such that for all $m < r \leq m+n$ and all $i, j \in \mathcal{I}_r$ we have $\lambda_i = \lambda_j$. Let $\mathcal{I}_{m+n+1} = \mathcal{I}'^c$. Suppose that there is a real number $x > 0$ such that for all $r, s \leq m+n+1$,*

$$\|P_{\mathcal{I}_r} E P_{\mathcal{I}_s}\|_2 \leq x \sqrt{\sum_{i\in\mathcal{I}_r} \lambda_i} \sqrt{\sum_{j\in\mathcal{I}_s} \lambda_j}. \tag{2.7}$$

If $\mathbf{r}_{\mathcal{I}}(\Sigma) \leq 1/(8x)$, then (2.5) holds with the constant 16 replaced by 64.

Theorem 2 reveals that it actually suffices to have adequate bounds for certain blocks corresponding to the same eigenvalue, and that the far away part represented by $\mathcal{I}'^c$ can be dealt with separately. Note that (2.7) follows from (2.3), as can be seen by squaring out the Hilbert-Schmidt norm.

2.2. **A general perturbation bound.** We now present a more technical perturbation bound and show how it implies Theorems 1 and 2. In order to deal with the different assumptions on certain coefficients (resp. blocks) of E from the last section, we introduce some flexibility with respect to the structure of E.

Theorem 3. *Let $\mathcal{I} \subseteq \mathbb{N}$ be a finite subset and let $\{\mathcal{I}_1, \dots, \mathcal{I}_m\}$ be a partition of $\mathcal{I}$ (meaning that $\mathcal{I}_1, \dots, \mathcal{I}_m$ are non-empty, disjoint subsets of $\mathcal{I}$ whose union is equal to $\mathcal{I}$). Let $\{\mathcal{I}_{m+1}, \mathcal{I}_{m+2}, \dots\}$ be a (possibly finite) partition of $\mathcal{I}^c$ into intervals. Let (a_r) and (b_r) be sequences of non-negative real numbers such that for all $r, s \geq 1$,*

$$\|P_{\mathcal{I}_r} E P_{\mathcal{I}_s}\|_\infty \leq \max(\sqrt{a_r b_s}, \sqrt{b_r a_s}), \qquad \|P_{\mathcal{I}_r} E P_{\mathcal{I}_s}\|_2 \leq \sqrt{b_r b_s}. \tag{2.8}$$

Suppose that

$$\Big(\sum_{r\geq 1} \frac{a_r}{g_r}\Big)\Big(\sum_{r\geq 1} \frac{b_r}{g_r}\Big) \leq 1/64 \tag{2.9}$$

with $g_r = \min_{i\in\mathcal{I}_r, j\in\mathcal{I}^c} |\lambda_i - \lambda_j|$ for $r \leq m$ and $g_r = \min_{j\in\mathcal{I}_r, i\in\mathcal{I}} |\lambda_i - \lambda_j|$ otherwise. Then we have

$$\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2^2 \leq 12 \sum_{r>m} \sum_{s\leq m} \frac{b_r b_s}{g_{r,s}^2} + 256 \Big(\sum_{r\geq 1} \frac{b_r}{g_r}\Big)^2 \sum_{r>m} \sum_{s\leq m} \frac{a_r b_s}{g_{r,s}^2} \tag{2.10}$$

with $g_{r,s}^2 = \min_{i\in\mathcal{I}_r, j\in\mathcal{I}_s} (\lambda_i - \lambda_j)^2$.

In particular, if (2.8) *holds with* $a_r = b_r$ *and if* $\sum_{r\geq 1} b_r/g_r \leq 1/8$, *then we have*

$$\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2^2 \leq 16 \sum_{r>m} \sum_{s\leq m} \frac{b_r b_s}{g_{r,s}^2}. \tag{2.11}$$

Remark 4. From (2.9), it follows that (b_r) is summable. Combining this with Assumption (2.8), we see that E has to be Hilbert-Schmidt.

Remark 5. Theorem 3 includes a version of the Davis-Kahan $\sin\Theta$ theorem. Indeed, the simple choice $\mathcal{I}_1 = \mathcal{I}$, $\mathcal{I}_2 = \mathcal{I}^c$ and $a_r = \|E\|_\infty^2/\|E\|_2$, $b_r = \|E\|_2$, $r = 1, 2$, leads to $\|\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}\|_2 \leq 4\|E\|_2/g_{\mathcal{I}}$, provided that $\|E\|_\infty/g_{\mathcal{I}} \leq 1/16$, with $g_{\mathcal{I}} = \min_{i\in\mathcal{I}, j\notin\mathcal{I}} |\lambda_i - \lambda_j|$. One advantage of the Davis-Kahan $\sin\Theta$ theorem is that it depends only on a small number of parameters: this version, for instance, shows that the sensitivity of $\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}$ can be described by the size of the perturbation relative to the gap $g_{\mathcal{I}}$. Our main objective is to go beyond this simple worst-case scenario using only a single gap. This corresponds to choosing finer partitions. In the extreme case where both partitions consist of singletons, the bound reflects the magnitude of the first order approximation given in (1.3), and involves gaps between all relevant eigenvalues.

Remark 6. Assumption (2.8) is designed to deal with random perturbations. While it might be difficult to check (2.8) for a given fixed Σ and $\hat{\Sigma}$, we show in Section 3 that it holds with high probability for a variety of structured random perturbations. In this respect, note that Assumption (2.8) allows for some flexibility when bounding $P_{\mathcal{I}_r} E P_{\mathcal{I}_s}$. Observe that since $\|\cdot\|_\infty \leq \|\cdot\|_2$, the second condition implies the first if $a_r = b_r$ for all $r \geq 1$. If, however, significantly better bounds for the operator norm are available, see e.g. [26, 4, 25], we may select $a_r \ll b_r$, yielding much weaker conditions in (2.9).

Remark 7. If $p = \dim \mathcal{H} < \infty$, then Theorem 3 holds for all self-adjoint operators Σ and $\hat{\Sigma}$. Indeed, we can always find a real number $y > 0$ such that $\Sigma + yI$ and $\hat{\Sigma} + yI$ are positive, and the claim follows because eigenvectors, gaps, and E are invariants of this transformation. If $p = \infty$, then positiveness is also not necessary, but the statement and its proof are notationally more involved.

We conclude this section by showing how Theorems 1 and 2 can be obtained by an application of Theorem 3.

Proof of Theorems 1 and 2. In order to obtain Theorem 1, take partitions of $\mathcal{I}$ and $\mathcal{I}^c$ consisting of singletons. Choose $a_j = b_j = x\lambda_j$, with x from (2.3). Then (2.8) holds because $\|P_i E P_j\|_\infty = \|P_i E P_j\|_2$ and (2.9) coincides with (2.4). Thus (2.5) follows from (2.11). Regarding Theorem 2, the partition is already given. In addition, for $r \leq m + n + 1$, set $a_r = b_r = x \sum_{i\in\mathcal{I}_r} \lambda_i$. Then it is easy to see that (2.8) and (2.9) are implied by (2.7) and (2.4), respectively, and the claim follows from (2.11), using that by construction of $\mathcal{I}'$,

$$\frac{x^2\lambda_i \sum_{j\in\mathcal{I}'^c} \lambda_j}{\min_{j\in\mathcal{I}'^c}(\lambda_i - \lambda_j)^2} \leq 4x^2 \sum_{j\in\mathcal{I}'^c} \frac{\lambda_i\lambda_j}{\lambda_i^2} \leq 4x^2 \sum_{j\in\mathcal{I}'^c} \frac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}$$

for all $i \in \mathcal{I}$. □

3. Applications

3.1. **Empirical covariance operators.** Let us discuss applications of our main result to the empirical covariance operator. Let X be a random variable taking values in $\mathcal{H}$. We suppose that X is centered and strongly square-integrable, meaning that $\mathbb{E}X = 0$ and $\mathbb{E}\|X\|^2 < \infty$. Let $\Sigma = \mathbb{E}X \otimes X$ be the covariance operator of X, which is a positive, self-adjoint trace class operator, see e.g. [12, Theorem 7.2.5]. For $j \geq 1$, let $\eta_j = \lambda_j^{-1/2}\langle u_j, X\rangle$ be the j-th Karhunen-Loève coefficient of X. Let $X_1, \dots, X_n$ be independent copies of X and let

$$\hat{\Sigma} = \frac{1}{n}\sum_{l=1}^{n} X_l \otimes X_l$$

be the empirical covariance operator. Combining Theorem 2 with concentration inequalities, we get:

Theorem 4. *In the above setting, suppose that for some $q > 4$ and $C_\eta > 0$,*

$$\sup_{j\geq 1} \mathbb{E}|\eta_j|^q \leq C_\eta. \tag{3.1}$$

Then there are constants $c_1, C_1 > 0$ depending only on C_η and q, such that for all $k, k_0 \geq 1$ with $\lambda_{k_0} \leq \lambda_k/2$ and all $t \geq 1$ satisfying

$$\frac{t}{\sqrt{n}}\Big(\sum_{i\leq k}\frac{\lambda_i}{\lambda_i - \lambda_{k+1}} + \sum_{j>k}\frac{\lambda_j}{\lambda_k - \lambda_j}\Big) \leq c_1, \tag{3.2}$$

we have

$$\mathbb{P}\Big(\|\hat{P}_{\{1,\dots,k\}} - P_{\{1,\dots,k\}}\|_2^2 > \frac{C_1 t^2}{n}\sum_{i\leq k}\sum_{j>k}\frac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}\Big) \tag{3.3}$$
$$\leq k_0^2\Big(\frac{n^{1-q/4}}{t^{q/2}} + \exp(-t^2)\Big).$$

Remark 8. For $t = \sqrt{\log n}$, (3.3) gives a similar high probability bound as in (1.4).

Remark 9. Theorem 4 gives useful bounds for $t \geq \sqrt{\log k_0}$. Corresponding bounds for $t < \sqrt{\log k_0}$ can be obtained using Remark 3, we omit the details. In (3.3), k_0 can be replaced by the number of distinct eigenvalues with indices smaller than or equal to k_0.

Remark 10. In the literature it is often assumed that the η_j are independent and satisfy some moment growth condition, see e.g. [21], or that X is sub-Gaussian or even Gaussian, see e.g. [19, 18]. In contrast, we only need the existence of a uniform moment bound on the η_j of order $q > 4$. In fact, since our bounds are based on the Fuk-Nagaev inequality, we expect our moment assumptions to be minimal. Despite this generality, we obtain sharp results, capable of serving as a new tool in functional PCA or kernel PCA (cf. [12, 11, 24]).

Following [16], we call (3.2) a relative rank condition, in contrast to the effective rank condition introduced in [18], where the latter is based on (1.2) and the concentration inequality in [19, 1]. The relative rank condition can be easily verified for exponentially and polynomially decaying eigenvalues. For instance, for polynomially decaying eigenvalues, the eigenvalue expressions in (3.2) and (3.3) are of order $k\log(k)$ and $k^2\log(k)$, respectively (cf. [15, Lemma 7.13]).

Corollary 1. *Grant Assumption (3.1). If for some $\alpha > 0$, $\lambda_j = j^{-\alpha-1}$, $j \geq 1$, then there are constants $c_1, C_1 > 0$ depending only on α, C_η, and q such that for all $k \geq 2$ and all $t \geq 1$ satisfying $tk\log(k) \leq c_1\sqrt{n}$, we have*

$$\mathbb{P}\Big(\|\hat{P}_{\{1,\dots,k\}} - P_{\{1,\dots,k\}}\|_2^2 > \frac{C_1 t^2 k^2 \log(k)}{n}\Big) \leq k^2\Big(\frac{n^{1-q/4}}{t^{q/2}} + \exp(-t^2)\Big).$$

Moreover, if $\lambda_j = \exp(-\alpha j)$, $j \geq 1$, then for all $k \geq 1$ and all $t \geq 1$ satisfying $tk \leq c_1\sqrt{n}$, we have

$$\mathbb{P}\Big(\|\hat{P}_{\{1,\dots,k\}} - P_{\{1,\dots,k\}}\|_2^2 > \frac{C_1 t^2}{n}\Big) \leq k^2\Big(\frac{n^{1-q/4}}{t^{q/2}} + \exp(-t^2)\Big).$$

Relative perturbation bounds for the empirical covariance operator have recently attracted attention in the literature. In [21, 15, 16], using different arguments, the special case $\mathcal{I} = \{i\}$ was treated. The general case is more complicated. For instance, [21] combines the holomorphic functional calculus with a normalization argument to go beyond the standard approach outlined in the introduction. They were, however, not able to obtain the sharp leading term in (3.3) by their method of proof, and require much stronger probabilistic conditions. Several types of relative perturbations have also been investigated in the deterministic case, see e.g. [14]. However, designed for a different purpose, they give significantly inferior results when applied to the empirical covariance operator.

Let us now turn to the proof of Theorem 4. The following lemma provides the necessary concentration inequality needed to deal with Condition (2.7).

Lemma 1. *Let $\mathcal{I}, \mathcal{J} \subseteq \mathbb{N}$. Under the assumptions of Theorem 4, there is a constant $C_1 > 0$ depending only on C_η and q, such that for all $t \geq 1$,*

$$\mathbb{P}\Big(\frac{\|P_\mathcal{I} E P_\mathcal{J}\|_2}{(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\lambda_i\lambda_j)^{1/2}} > \frac{C_1 t}{\sqrt{n}}\Big) \leq \frac{n^{1-q/4}}{t^{q/2}} + \exp(-t^2).$$

Theorem 4 is now an immediate consequence of Theorem 2, the union bound, and Lemma 1. Lemma 1 itself follows from [9, Theorem 3.1], a Banach space version of the Fuk-Nagaev inequality. For the sake of completeness, we describe the relevant computations below.

Proof of Lemma 1. Observe that

$$nP_\mathcal{I} E P_\mathcal{J} = \sum_{l=1}^n \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} (\langle X_l, u_i\rangle\langle X_l, u_j\rangle - \delta_{ij}\sqrt{\lambda_i\lambda_j})u_i \otimes u_j =: \sum_{l=1}^n Z_l,$$

where the Z_l are i.i.d. random variables taking values in the Hilbert space of all Hilbert-Schmidt operators on $\mathcal{H}$, endowed with the Hilbert-Schmidt scalar product $\langle\cdot,\cdot\rangle_2$. First, for every Hilbert-Schmidt operator A on $\mathcal{H}$ with $\|A\|_2 \leq 1$, we have

$$\mathbb{E}\sum_{l=1}^n \langle A, Z_l\rangle_2^2 \leq n\mathbb{E}\|Z_1\|_2^2 = n\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\lambda_i\lambda_j\mathbb{E}(\eta_i\eta_j - \delta_{ij})^2 \leq C_2 n\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\lambda_i\lambda_j,$$

as can be seen by the Cauchy-Schwarz inequality, the definition of the η_j, Jensen's inequality, and (3.1). Similarly, by Jensen's inequality and (3.1), we have

$$\mathbb{E}\Big\|\sum_{l=1}^n Z_l\Big\|_2 \leq \Big(n\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\lambda_i\lambda_j\mathbb{E}(\eta_i\eta_j - \delta_{ij})^2\Big)^{1/2} \leq C_3\Big(n\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\lambda_i\lambda_j\Big)^{1/2}.$$

Finally, by Minkowski's inequality, Jensen's inequality, and (3.1), we have

$$(\mathbb{E}\|Z_1\|_2^{q/2})^{4/q} \le \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \lambda_i\lambda_j (\mathbb{E}(\eta_i\eta_j - \delta_{ij})^{q/2})^{4/q} \le C_4 \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \lambda_i\lambda_j.$$

Lemma 1 now follows from [9, Theorem 3.1] applied with $s = q/2$. □

3.2. **Using the operator norm.** Let us give two simple examples showing how one can benefit from having two different norms in Condition (2.8).

Random perturbation of a low rank matrix. Let $\mathcal{H} = \mathbb{R}^p$ and let $\Sigma = \sum_{i\le k} \lambda_i u_i u_i^T$ be a symmetric matrix with $\lambda_1 \ge \cdots \ge \lambda_k > 0$ and $u_1, \dots, u_k$ orthonormal system in $\mathbb{R}^p$. Let $\hat\Sigma = \Sigma + \epsilon\xi$, where $\epsilon > 0$ and $\xi = (\xi_{ij})_{1\le i,j\le p}$ is a GOE matrix, i.e. a symmetric random matrix whose upper triangular entries are independent zero mean Gaussian random variables with $\mathbb{E}\xi_{ij}^2 = 1$ for $1 \le i < j \le p$ and $\mathbb{E}\xi_{ii}^2 = 2$ for $i = 1, \dots, p$. Then Theorem 3 yields that for all $t \ge 1$, with probability at least $1 - 18\exp(-c_1 t)$,

$$\|\hat P_1 - P_1\|_2^2 \le C_1\bigg(\frac{\epsilon^2 tk}{(\lambda_1-\lambda_2)^2} + \frac{\epsilon^2 t(p-k)}{\lambda_1^2} + \frac{\epsilon^4 t^2(p-k)^2}{\lambda_1^2(\lambda_1-\lambda_2)^2}\bigg), \tag{3.4}$$

where $c_1, C_1 > 0$ are absolute constants. In comparison, the Davis-Kahan $\sin\Theta$ theorem in (1.1) with $\|E\|_2$ replaced by $\sqrt{|\mathcal{I}|}\|E\|_\infty$ yields a bound of order $\epsilon^2 p/(\lambda_1 - \lambda_2)^2$, which is inferior to (3.4) for k smaller than p and $\lambda_1 - \lambda_2$ smaller than λ_1. The bound in (3.4) can be compared to [27, Theorem 8] and [22, Remark 15], where a structurally similar third term appears.

Let us deduce (3.4) from (2.10), using also Remark 7. Since ξ is invariant under orthogonal transformations, we may assume that u_i is the i-th standard basis vector in $\mathbb{R}^p$. Hence, Condition (2.8) is dealing with submatrices of $E = \epsilon\xi$. We choose $\mathcal{I}_1 = \{1\}$, $\mathcal{I}_2 = \{2, \dots, k\}$, and $\mathcal{I}_3 = \{k+1, \dots, p\}$. Applying concentration results for the operator norm of random matrices (e.g. [4, Theorem 5.6] and [20, Theorem 1]) and the bound $\|P_{\mathcal{I}_r} E P_{\mathcal{I}_s}\|_2 \le \sqrt{|\mathcal{I}_r| \wedge |\mathcal{I}_s|}\|P_{\mathcal{I}_r} E P_{\mathcal{I}_s}\|_\infty$, we get that for all $t \ge 1$, (2.8) is satisfied with probability at least $1 - 18\exp(-c_1 t)$, provided that we choose

$$(a_1, a_2, a_3) = (\epsilon\sqrt{t}, \epsilon\sqrt{t}, \epsilon\sqrt{t}), \quad (b_1, b_2, b_3) = (C_2\epsilon\sqrt{t}, C_2(k-1)\epsilon\sqrt{t}, C_2(p-k)\epsilon\sqrt{t}).$$

By Theorem 3, we get, for all $t \ge 1$, with probability at least $1 - 18\exp(-c_1 t)$,

$$\|\hat P_1 - P_1\|_2^2 \le C_3\bigg(\frac{\epsilon^2 tk}{(\lambda_1-\lambda_2)^2} + \frac{\epsilon^2 t(p-k)}{\lambda_1^2} + \frac{\epsilon^4 t^2 k^2}{(\lambda_1-\lambda_2)^4} + \frac{\epsilon^4 t^2(p-k)^2}{\lambda_1^2(\lambda_1-\lambda_2)^2}\bigg),$$

provided that

$$\frac{\epsilon^2 tk}{(\lambda_1-\lambda_2)^2} + \frac{\epsilon^2 t(p-k)}{\lambda_1(\lambda_1-\lambda_2)} \le c_2. \tag{3.5}$$

Since always $\|\hat P_1 - P_1\|_2^2 \le 2$, Condition (3.5) can be dropped and the above bound can be rearranged into the desired form (3.4), by adjusting C_3.

Spiked covariance model. Consider the empirical covariance operator from Section 3.1. Let $\mathcal{H} = \mathbb{R}^p$ and let $\Sigma = \mu_1 P_{\mathcal{I}_1} + \mu_2 P_{\mathcal{I}_2} + \mu_3 P_{\mathcal{I}_3}$ with $\mu_1 > \mu_2 > \mu_3 > 0$ and $m_r = |\mathcal{I}_r|$, $r = 1, 2, 3$. Assume that the η_j are independent and sub-Gaussian,

meaning that for some constant $C_\eta > 0$, $\mathbb{E}^{1/q}|\eta_j|^q \le C_\eta\sqrt{q}$ for all natural numbers $q \ge 1$ and all $j = 1, \dots, p$. Then Theorem 3 yields that for all $t \ge 1$, with probability at least $1 - 9\exp(-(m_1 \wedge m_2 \wedge m_3)t)$,

$$\|\hat{P}_{\mathcal{I}_1} - P_{\mathcal{I}_1}\|_2^2 \le C_1 m_1 \bigg(\frac{\mu_1^2 k}{(\mu_1-\mu_2)^2}\frac{t}{n} + \frac{\mu_1^2(p-k)}{(\mu_1-\mu_3)^2}\frac{t}{n} + \frac{\mu_1^4(p-k)^2}{(\mu_1-\mu_2)^2(\mu_1-\mu_3)^2}\frac{t^2}{n^2}\bigg). \tag{3.6}$$

where $k = m_1 + m_2$ is the dimension of the spiked part and $C_1 > 0$ is a constant depending only on C_η. In comparison, the Davis-Kahan $\sin\Theta$ theorem in (1.1) with $\|E\|_2$ replaced by $\sqrt{|\mathcal{I}|}\|E\|_\infty$ yields a bound of order $\mu_1^2 m_1 p/(n(\mu_1-\mu_2)^2)$.

Let us deduce (3.6) from Theorem 3, by replacing Lemma 1 with the following concentration inequality for the operator norm of empirical covariance operators.

Lemma 2. *In the above setting, there is a constant $C_2 > 0$ depending only on C_η such that for all $r, s = 1, 2, 3$ and all $t \ge 1$ satisfying $t(m_r \vee m_s)/n \le 1$, we have*

$$\mathbb{P}\bigg(\|P_{\mathcal{I}_r} E P_{\mathcal{I}_s}\|_\infty > C_2\sqrt{\frac{\mu_r\mu_s(m_r \vee m_s)t}{n}} \bigg) \le \exp(-(m_r \vee m_s)t).$$

The case $r = s$ follows from [19, Theorem 1], the non-diagonal case follows from a similar standard net argument as presented therein (see also [26, 22]). Proceeding as in the proof of (3.4), we get that for all $1 \le t \le n/(k \vee (p-k))$, (2.8) is satisfied with probability at least $1 - 9\exp(-(m_1 \wedge m_2 \wedge m_3)t)$, provided that we choose

$$\begin{aligned}(a_1, a_2, a_3) &= (\mu_1\sqrt{t/n}, \mu_2\sqrt{t/n}, \mu_3\sqrt{t/n}),\\ (b_1, b_2, b_3) &= (C_2\mu_1 m_1\sqrt{t/n}, C_2\mu_2 m_2\sqrt{t/n}), C_2\mu_3 m_3\sqrt{t/n}).\end{aligned}$$

Applying Theorem 3, (3.6) follows from a similar computation leading to (3.4).

Given two groups of eigenvalues $\Sigma = \mu_1 P_{\mathcal{I}_1} + \mu_2 P_{\mathcal{I}_2}$ with $\mu_1 > \mu_2$ and $m_r = |\mathcal{I}_r|$, $r = 1, 2$, a similar computation yields a bound of order $\mu_1\mu_2 m_1 m_2/(n(\mu_1-\mu_2)^2)$. The latter is known to be optimal in a minimax sense, see e.g. Theorem 8 and 9 in [5] and also [23]. Extensions of this bound to a high-dimensional context can be found in [5, 28]. In comparison, the three-group bound (3.6) depends on two different gaps, and contains a second order perturbation term. As already mentioned in the previous example, a similar third term is present in the results in [27, 22]. Note that this term can be avoided by applying Theorem 4, yet under the additional relative gap condition in (3.2).

4. Proof of main theorem

4.1. **Separation of eigenvalues.** In this section, we show that under Condition (2.9), the perturbed eigenvalues $(\hat{\lambda}_i)_{i\in\mathcal{I}}$ are well-separated from $(\lambda_j)_{j\notin\mathcal{I}}$.

Lemma 3. *Under the assumptions of Theorem 3, we have*

$$|\hat{\lambda}_i - \lambda_j| \ge \frac{|\lambda_i - \lambda_j|}{2}$$

for all $i \in \mathcal{I}$ and $j \notin \mathcal{I}$.

The proof is based on the following result, which is an intermediate step in the proof of Propositions 3.10 and 3.13 in [23]. In fact, (4.1), for instance, follows from the min-max characterisation of eigenvalues in combination with [23, Lemma 3.11].

Proposition 1. *For all $i \geq 1$ and $y > 0$, we have the implications*

$$\Big\| \Big(\sum_{k \geq i} \frac{1}{\sqrt{\lambda_i + y - \lambda_k}} P_k \Big) E \Big(\sum_{k \geq i} \frac{1}{\sqrt{\lambda_i + y - \lambda_k}} P_k \Big) \Big\|_\infty^2 \leq 1 \Rightarrow \hat{\lambda}_i - \lambda_i \leq y \tag{4.1}$$

and

$$\Big\| \Big(\sum_{k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k \Big) E \Big(\sum_{k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k \Big) \Big\|_\infty^2 \leq 1 \Rightarrow \hat{\lambda}_i - \lambda_i \geq -y. \tag{4.2}$$

Proof of Lemma 3. It suffices to show that

$$\hat{\lambda}_i - \lambda_j \geq \frac{\lambda_i - \lambda_j}{2} \tag{4.3}$$

for all $i \in \mathcal{I}$ and $j \notin \mathcal{I}$ such that $i < j$, and that

$$\lambda_j - \hat{\lambda}_i \geq \frac{\lambda_j - \lambda_i}{2} \tag{4.4}$$

for all $i \in \mathcal{I}$ and $j \notin \mathcal{I}$ such that $j < i$. We only prove (4.3), the proof of (4.4) follows the same line of arguments. First, (4.3) is equivalent to

$$\hat{\lambda}_i - \lambda_i \geq -\frac{\lambda_i - \lambda_j}{2} \tag{4.5}$$

for all $i \in \mathcal{I}$ and $j \notin \mathcal{I}$ such that $i < j$. Thus, it suffices to show that the left-hand side in (4.2) is satisfied with $y = (\lambda_i - \lambda_j)/2$. For $r \geq 1$, set

$$T_r = \sum_{k \in \mathcal{I}_r, k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k.$$

(Set $T_r = 0$ if the summation is empty.) Using that the T_r are self-adjoint and have orthogonal ranges, we have

$$\begin{aligned} &\Big\| \Big(\sum_{k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k \Big) E \Big(\sum_{k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k \Big) \Big\|_\infty^2 \\ &= \Big\| \Big(\sum_{r \geq 1} T_r \Big) E \Big(\sum_{s \geq 1} T_s \Big) \Big\|_\infty^2 \leq \sum_{r \geq 1} \sum_{s \geq 1} \|T_r E T_s\|_\infty^2. \end{aligned} \tag{4.6}$$

Using the identities $T_r = T_r P_{\mathcal{I}_r} = P_{\mathcal{I}_r} T_r$ and (2.8), we have

$$\|T_r E T_s\|_\infty^2 \leq (a_r b_s + b_r a_s) \|T_r\|_\infty^2 \|T_s\|_\infty^2$$

for all $r, s \geq 1$. Hence,

$$\sum_{r \geq 1} \sum_{s \geq 1} \|T_r E T_s\|_\infty^2 \leq 2 \Big(\sum_{r \geq 1} a_r \|T_r\|_\infty^2 \Big) \Big(\sum_{s \geq 1} b_s \|T_s\|_\infty^2 \Big). \tag{4.7}$$

Now, using that $\min_{k \in \mathcal{I}_r, k \leq i} (\lambda_k + y - \lambda_i) \geq \min_{k \in \mathcal{I}_r} |\lambda_k - \lambda_i| \geq g_r$ for $r > m$, and $\min_{k \in \mathcal{I}_r, k \leq i} (\lambda_k + y - \lambda_i) \geq \min_{k \in \mathcal{I}_r} |\lambda_k - \lambda_j|/2 \geq g_r/2$ for $r \leq m$, we obtain that

$$\|T_r\|_\infty^2 \leq 2/g_r \tag{4.8}$$

for all $r \geq 1$. Using (4.6)-(4.8) in combination with (2.9), we conclude that

$$\Big\| \Big(\sum_{k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k \Big) E \Big(\sum_{k \leq i} \frac{1}{\sqrt{\lambda_k + y - \lambda_i}} P_k \Big) \Big\|_\infty^2 \leq 1/8 \leq 1,$$

and the claim follows from (4.2). □

4.2. **Key contraction phenomenon.** The proof of Theorem 3 is based on a first order perturbation expansion, combined with a recursive argument to get control of the remainder term. The main technical lemma is as follows:

Lemma 4. *Under the assumptions of Theorem 3, the inequality*

$$\sqrt{\sum_{i\in\mathcal{I}} \frac{\|P_{\mathcal{I}_r} E\hat{P}_i\|_2^2}{(\hat{\lambda}_i-\lambda_j)^2}} \le \left(\frac{3}{2}\sqrt{b_r} + 4\sqrt{a_r}\sum_{s\ge 1}\frac{b_s}{g_s}\right)\sqrt{\sum_{s\le m}\frac{b_s}{\min_{i\in\mathcal{I}_s}(\lambda_i-\lambda_j)^2}}$$

holds for all $r \ge 1$ and all $j \notin \mathcal{I}$.

Proof. We recall some simple properties of the Hilbert-Schmidt norm which we will use in the sequel without further comment. For Hilbert-Schmidt operators A and B on $\mathcal{H}$, we have $\|AB\|_2 \le \|A\|_\infty \|B\|_2$. Moreover, for a Hilbert-Schmidt operator A on $\mathcal{H}$ and a bounded sequence of real numbers $(x_i)_{i\ge 1}$ we have $\|\sum_{i\ge 1} x_i P_i A\|_2^2 = \sum_{i\ge 1} x_i^2 \|P_i A\|_2^2$ and the same identity holds for $P_i A$ replaced by AP_i.

Let $r \ge 1$ be arbitrary. By the identity $I = P_{\mathcal{I}} + P_{\mathcal{I}^c}$ (see the convention in Section 1.1), and the triangular inequality, we have

$$\begin{aligned}\sqrt{\sum_{i\in\mathcal{I}} \frac{\|P_{\mathcal{I}_r} E\hat{P}_i\|_2^2}{(\hat{\lambda}_i-\lambda_j)^2}} &= \left\|\sum_{i\in\mathcal{I}} \frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}_r} E\hat{P}_i\right\|_2 \\ &\le \left\|\sum_{i\in\mathcal{I}} \frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}_r} EP_{\mathcal{I}}\hat{P}_i\right\|_2 + \left\|\sum_{i\in\mathcal{I}} \frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}_r} EP_{\mathcal{I}^c}\hat{P}_i\right\|_2.\end{aligned} \tag{4.9}$$

Note that all denominators are non-zero by Lemma 3. We start with the first term on the right-hand side of (4.9). By the identity

$$(\hat{\lambda}_i-\lambda_k)P_k\hat{P}_i = P_k E\hat{P}_i, \tag{4.10}$$

valid for every $i, k \ge 1$, we have

$$\begin{aligned}\sum_{i\in\mathcal{I}} \frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}}\hat{P}_i - \sum_{k\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j} P_k\hat{P}_{\mathcal{I}} &= \sum_{k\in\mathcal{I}}\sum_{i\in\mathcal{I}} \left(\frac{1}{\hat{\lambda}_i-\lambda_j} - \frac{1}{\lambda_k-\lambda_j}\right) P_k\hat{P}_i \\ &= -\sum_{k\in\mathcal{I}}\sum_{i\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j}\frac{1}{\hat{\lambda}_i-\lambda_j} P_k E\hat{P}_i.\end{aligned}$$

Using this identity and the triangular inequality, we get

$$\begin{aligned}\left\|\sum_{i\in\mathcal{I}} \frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}_r} EP_{\mathcal{I}}\hat{P}_i\right\|_2 &\le \left\|\sum_{k\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j} P_{\mathcal{I}_r} EP_k\hat{P}_{\mathcal{I}}\right\|_2 \\ &\quad + \left\|\sum_{k\in\mathcal{I}}\sum_{i\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j}\frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}_r} EP_k E\hat{P}_i\right\|_2.\end{aligned} \tag{4.11}$$

The first term on the right-hand side of (4.11) is bounded as follows:

$$\left\|\sum_{k\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j} P_{\mathcal{I}_r} EP_k\hat{P}_{\mathcal{I}}\right\|_2 \le \left\|\sum_{k\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j} P_{\mathcal{I}_r} EP_k\right\|_2. \tag{4.12}$$

Next, consider the second term on the right-hand side of (4.11). Using the triangular inequality, we have

$$\left\|\sum_{k\in\mathcal{I}}\sum_{i\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j}\frac{1}{\hat{\lambda}_i-\lambda_j} P_{\mathcal{I}_r} EP_k E\hat{P}_i\right\|_2 \tag{4.13}$$

$$\leq \sum_{s\leq m} \Big\| \sum_{k\in\mathcal{I}_s} \sum_{i\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j}\frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2$$
$$= \sum_{s\leq m} \sqrt{\sum_{i\in\mathcal{I}} \frac{1}{(\hat\lambda_i-\lambda_j)^2} \Big\| \sum_{k\in\mathcal{I}_s} \frac{1}{\lambda_k-\lambda_j} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2^2}.$$

Now, for each $s\leq m$ and $i\in\mathcal{I}$,

$$\sum_{k\in\mathcal{I}_s} \frac{1}{\lambda_k-\lambda_j} P_{\mathcal{I}_r} E P_k E \hat P_i = P_{\mathcal{I}_r} E P_{\mathcal{I}_s} \Big(\sum_{k\in\mathcal{I}_s} \frac{1}{\lambda_k-\lambda_j} P_k \Big) P_{\mathcal{I}_s} E \hat P_i$$

and by (2.8) and the definition of the g_s, this implies that

$$\Big\| \sum_{k\in\mathcal{I}_s} \frac{1}{\lambda_k-\lambda_j} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2 \leq \Big\| P_{\mathcal{I}_r} E P_{\mathcal{I}_s} \Big\|_\infty \Big\| \sum_{k\in\mathcal{I}_s} \frac{1}{\lambda_k-\lambda_j} P_k \Big\|_\infty \Big\| P_{\mathcal{I}_s} E \hat P_i \Big\|_2$$
$$\leq \Big(\frac{\sqrt{a_r b_s}}{g_s} + \frac{\sqrt{b_r a_s}}{g_s} \Big) \Big\| P_{\mathcal{I}_s} E \hat P_i \Big\|_2 .$$

Inserting this inequality into (4.13), we get

$$\Big\| \sum_{k\in\mathcal{I}} \sum_{i\in\mathcal{I}} \frac{1}{\lambda_k-\lambda_j}\frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2 \tag{4.14}$$
$$\leq \sum_{s\leq m} \Big(\frac{\sqrt{a_r b_s}}{g_s} + \frac{\sqrt{b_r a_s}}{g_s} \Big) \Big\| \sum_{i\in\mathcal{I}} \frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_s} E \hat P_i \Big\|_2 .$$

For the second term on the right-hand side of (4.9) we proceed similarly. By (4.10), Lemma 3, and the triangular inequality, we have

$$\Big\| \sum_{i\in\mathcal{I}} \frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_r} E P_{\mathcal{I}^c} \hat P_i \Big\|_2 = \Big\| \sum_{k\notin\mathcal{I}} \sum_{i\in\mathcal{I}} \frac{1}{\hat\lambda_i-\lambda_k}\frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2$$
$$\leq \sum_{s>m} \sqrt{\sum_{i\in\mathcal{I}} \frac{1}{(\hat\lambda_i-\lambda_j)^2} \Big\| \sum_{k\in\mathcal{I}_s} \frac{1}{\hat\lambda_i-\lambda_k} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2^2}. \tag{4.15}$$

Now, for $s>m$ and $i\in\mathcal{I}$, we have

$$\sum_{k\in\mathcal{I}_s} \frac{1}{\hat\lambda_i-\lambda_k} P_{\mathcal{I}_r} E P_k E \hat P_i = P_{\mathcal{I}_r} E P_{\mathcal{I}_s} \Big(\sum_{k\in\mathcal{I}_s} \frac{1}{\hat\lambda_i-\lambda_k} P_k \Big) P_{\mathcal{I}_s} E \hat P_i$$

and by (2.8) and Lemma 3, this implies that

$$\Big\| \sum_{k\in\mathcal{I}_s} \frac{1}{\hat\lambda_i-\lambda_k} P_{\mathcal{I}_r} E P_k E \hat P_i \Big\|_2 \leq 2 \Big(\frac{\sqrt{a_r b_s}}{g_s} + \frac{\sqrt{b_r a_s}}{g_s} \Big) \Big\| P_{\mathcal{I}_s} E \hat P_i \Big\|_2 .$$

Inserting this into (4.15), we conclude that

$$\Big\| \sum_{i\in\mathcal{I}} \frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_r} E P_{\mathcal{I}^c} \hat P_i \Big\|_2 \tag{4.16}$$
$$\leq 2 \sum_{s>m} \Big(\frac{\sqrt{a_r b_s}}{g_s} + \frac{\sqrt{b_r a_s}}{g_s} \Big) \Big\| \sum_{i\in\mathcal{I}} \frac{1}{\hat\lambda_i-\lambda_j} P_{\mathcal{I}_s} E \hat P_i \Big\|_2 .$$

Collecting (4.9), (4.11)-(4.16), we conclude that

$$\begin{aligned}\Big\|\sum_{i\in\mathcal{I}}\frac{1}{\hat\lambda_i-\lambda_j}P_{\mathcal{I}_r}E\hat P_i\Big\|_2 &\le \Big\|\sum_{i\in\mathcal{I}}\frac{1}{\lambda_i-\lambda_j}P_{\mathcal{I}_r}EP_i\Big\|_2 \\ &+2\sum_{s\ge1}\Big(\frac{\sqrt{a_rb_s}}{g_s}+\frac{\sqrt{b_ra_s}}{g_s}\Big)\Big\|\sum_{i\in\mathcal{I}}\frac{1}{\hat\lambda_i-\lambda_j}P_{\mathcal{I}_s}E\hat P_i\Big\|_2 \quad \forall r\ge1.\end{aligned} \tag{4.17}$$

It remains to solve this recursive inequality. First, using (2.8), we have

$$\begin{aligned}\Big\|\sum_{i\in\mathcal{I}}\frac{1}{\lambda_i-\lambda_j}P_{\mathcal{I}_r}EP_i\Big\|_2 &= \sqrt{\sum_{i\in\mathcal{I}}\frac{\|P_{\mathcal{I}_r}EP_i\|_2^2}{(\lambda_i-\lambda_j)^2}} \\ \le \sqrt{\sum_{s\le m}\frac{\|P_{\mathcal{I}_r}EP_{\mathcal{I}_s}\|_2^2}{\min_{i\in\mathcal{I}_s}(\lambda_i-\lambda_j)^2}} &\le \sqrt{\sum_{s\le m}\frac{b_rb_s}{\min_{i\in\mathcal{I}_s}(\lambda_i-\lambda_j)^2}} =: \sqrt{b_r}B\end{aligned} \tag{4.18}$$

for all $r\ge1$. If we set

$$A_r=\Big\|\sum_{i\in\mathcal{I}}\frac{1}{\hat\lambda_i-\lambda_j}P_{\mathcal{I}_r}E\hat P_i\Big\|_2 \qquad \forall r\ge1,$$

then (4.17) implies that

$$A_r\le\sqrt{b_r}B+2\sqrt{a_r}\Big(\sum_{s\ge1}\frac{\sqrt{b_s}}{g_s}A_s\Big)+2\sqrt{b_r}\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big) \quad \forall r\ge1. \tag{4.19}$$

Multiplying both sides with $\sqrt{b_r}/g_r$ and summing over $r\ge1$, we have

$$\begin{aligned}&\sum_{r\ge1}\frac{\sqrt{b_r}}{g_r}A_r \\ &\le\Big(\sum_{r\ge1}\frac{b_r}{g_r}\Big)B+2\Big(\sum_{r\ge1}\frac{\sqrt{a_rb_r}}{g_r}\Big)\Big(\sum_{s\ge1}\frac{\sqrt{b_s}}{g_s}A_s\Big)+2\Big(\sum_{r\ge1}\frac{b_r}{g_r}\Big)\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big).\end{aligned}$$

By (2.9) and the Cauchy-Schwarz inequality, this implies

$$\sum_{r\ge1}\frac{\sqrt{b_r}}{g_r}A_r\le\frac43\Big(\sum_{r\ge1}\frac{b_r}{g_r}\Big)B+\frac83\Big(\sum_{r\ge1}\frac{b_r}{g_r}\Big)\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big).$$

Inserting this inequality into (4.19), we get

$$\begin{aligned}A_r&\le\sqrt{b_r}B+\frac83\sqrt{a_r}\Big(\sum_{s\ge1}\frac{b_s}{g_s}\Big)B \\ &+\frac{16}{3}\sqrt{a_r}\Big(\sum_{s\ge1}\frac{b_s}{g_s}\Big)\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big)+2\sqrt{b_r}\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big) \quad \forall r\ge1.\end{aligned} \tag{4.20}$$

Now, multiplying both sides with $\sqrt{a_r}/g_r$ and summing over $r\ge1$, we have

$$\begin{aligned}\sum_{r\ge1}\frac{\sqrt{a_r}}{g_r}A_r&\le\Big(\sum_{r\ge1}\frac{\sqrt{a_rb_r}}{g_r}\Big)B+\frac83\Big(\sum_{r\ge1}\frac{a_r}{g_r}\Big)\Big(\sum_{s\ge1}\frac{b_s}{g_s}\Big)B \\ &+\frac{16}{3}\Big(\sum_{r\ge1}\frac{a_r}{g_r}\Big)\Big(\sum_{s\ge1}\frac{b_s}{g_s}\Big)\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big)+2\Big(\sum_{r\ge1}\frac{\sqrt{a_rb_r}}{g_r}\Big)\Big(\sum_{s\ge1}\frac{\sqrt{a_s}}{g_s}A_s\Big).\end{aligned}$$

By (2.9) and the Cauchy-Schwarz inequality, this implies

$$\sum_{r\geq 1}\frac{\sqrt{a_r}}{g_r}A_r \leq \frac{B}{8}+\frac{B}{24}+\frac{1}{12}\bigg(\sum_{s\geq 1}\frac{\sqrt{a_s}}{g_s}A_s\bigg)+\frac{1}{4}\bigg(\sum_{s\geq 1}\frac{\sqrt{a_s}}{g_s}A_s\bigg)$$

and thus

$$\sum_{r\geq 1}\frac{\sqrt{a_r}}{g_r}A_r \leq \frac{B}{4}.$$

Inserting this into (4.20), we conclude that

$$A_r \leq \frac{3}{2}\sqrt{b_r}B + 4\bigg(\sqrt{a_r}\sum_{s\geq 1}\frac{b_s}{g_s}\bigg)B \qquad \forall r\geq 1,$$

and the claim follows from inserting the definitions of A_r and B. □

4.3. **End of proof of Theorem 3.** Using that orthogonal projections are idempotent and self-adjoint, we have $\|(I-P_{\mathcal{I}})\hat{P}_{\mathcal{I}}\|_2^2 = \mathrm{tr}((I-P_{\mathcal{I}})\hat{P}_{\mathcal{I}}) = \mathrm{tr}(P_{\mathcal{I}}(I-\hat{P}_{\mathcal{I}})) = \|P_{\mathcal{I}}(I-\hat{P}_{\mathcal{I}})\|_2^2$. Hence, by the identity $\hat{P}_{\mathcal{I}}-P_{\mathcal{I}} = (I-P_{\mathcal{I}})\hat{P}_{\mathcal{I}} - P_{\mathcal{I}}(I-\hat{P}_{\mathcal{I}})$, we get

$$\|\hat{P}_{\mathcal{I}}-P_{\mathcal{I}}\|_2^2 = 2\|(I-P_{\mathcal{I}})\hat{P}_{\mathcal{I}}\|_2^2 = 2\sum_{r>m}\|P_{\mathcal{I}_r}\hat{P}_{\mathcal{I}}\|_2^2. \tag{4.21}$$

By (4.10), we have for every $r>m$,

$$\|P_{\mathcal{I}_r}\hat{P}_{\mathcal{I}}\|_2^2 = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}_r}\frac{\|P_j E\hat{P}_i\|_2^2}{(\hat{\lambda}_i-\lambda_j)^2} \leq \sum_{i\in\mathcal{I}}\frac{\|P_{\mathcal{I}_r}E\hat{P}_i\|_2^2}{\min_{j\in\mathcal{I}_r}(\hat{\lambda}_i-\lambda_j)^2}. \tag{4.22}$$

Note that all denominators are non-zero by Lemma 3. Now, using (4.3), (4.4), and the fact that $\mathcal{I}_r$ is an interval, we get that $\min_{j\in\mathcal{I}_r}(\hat{\lambda}_i-\lambda_j)^2$ is attained at at most two points, namely at the endpoints of $\mathcal{I}_r$. Hence, there are $j_0, j_1 \in \mathcal{I}_r$ such that

$$\|P_{\mathcal{I}_r}\hat{P}_{\mathcal{I}}\|_2^2 \leq \sum_{i\in\mathcal{I}}\frac{\|P_{\mathcal{I}_r}E\hat{P}_i\|_2^2}{(\hat{\lambda}_i-\lambda_{j_0})^2} + \sum_{i\in\mathcal{I}}\frac{\|P_{\mathcal{I}_r}E\hat{P}_i\|_2^2}{(\hat{\lambda}_i-\lambda_{j_1})^2}.$$

Inserting this into (4.21) and applying Lemma 4, the claim follows from a simple computation, using the inequality $(y+z)^2 \leq 4y^2/3+4z^2$. □

Acknowledgement. We are grateful to the anonymous referees for their helpful comments. The research of Martin Wahl has been partially funded by Deutsche Forschungsgemeinschaft (DFG) via FOR 1735.

References

[1] R. Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electron. Commun. Probab.*, 20:no. 72, 13 pp, 2015.
[2] T.W. Anderson. Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34:122–148, 1963.
[3] R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997.
[4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.
[5] T. Cai, Z. Ma, and Y. Wu. Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.*, 41:3074–3110, 2013.
[6] F. Chatelin. *Spectral approximation of linear operators*. Academic Press, New York, 1983.
[7] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.*, 12:136–154, 1982.

[8] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.
[9] U. Einmahl and D. Li. Characterization of LIL behavior in Banach space. *Trans. Amer. Math. Soc.*, 360:6677–6693, 2008.
[10] P. Hall and M. Hosseini-Nasab. Theory for high-order bounds in functional principal components analysis. *Math. Proc. Cambridge Philos. Soc.*, 146:225–256, 2009.
[11] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer, New York, 2012.
[12] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, Ltd., Chichester, 2015.
[13] I. C. F. Ipsen. Relative perturbation results for matrix eigenvalues and singular values. *Acta numerica*, 7:151–201, 1998.
[14] I. C. F. Ipsen. An overview of relative $\sin\theta$ theorems for invariant subspaces of complex matrices. *J. Comput. Appl. Math.*, 123:131–153, 2000.
[15] M. Jirak. Optimal eigen expansions and uniform bounds. *Probab. Theory Related Fields*, 166:753–799, 2016.
[16] M. Jirak and M. Wahl. Relative perturbation bounds with applications to empirical covariance operators. Available at https://arxiv.org/pdf/1802.02869, 2018.
[17] T. Kato. *Perturbation theory for linear operators*. Springer-Verlag, Berlin, reprint of the 1980 edition, 1995.
[18] V. Koltchinskii and K. Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré*, 52:1976–2013, 2016.
[19] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23:110–133, 2017.
[20] R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282, 2005.
[21] A. Mas and F. Ruymgaart. High-dimensional principal projections. *Complex Anal. Oper. Theory*, 9:35–63, 2015.
[22] S. O'Rourke, V. Vu, and K. Wang. Random perturbation of low rank matrices: improving classical bounds. *Linear Algebra Appl.*, 540:26–59, 2018.
[23] M. Reiß and M. Wahl. Non-asymptotic upper bounds for the reconstruction error of PCA. Ann. Statist., to appear.
[24] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
[25] R. Van Handel. Structured random matrices. In *Convexity and Concentration*, The IMA Volumes in Mathematics and its Applications, 161, pages 107–165. Springer, New York, 2017.
[26] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.
[27] V. Vu. Singular vectors under random perturbation. *Random Structures Algorithms*, 39:526–538, 2011.
[28] V. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41:2905–2947, 2013.
[29] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

Moritz Jirak, Institut für Mathematische Stochastik, Technische Universität Braunschweig, Universitätsplatz 2, 38106 Braunschweig, Germany.
E-mail address: m.jirak@tu-braunschweig.de

Martin Wahl, Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.
E-mail address: martin.wahl@math.hu-berlin.de