Van Trees inequality, group equivariance, and estimation of principal subspaces

Martin Wahl

Abstract We establish non-asymptotic lower bounds for the estimation of principal subspaces. As applications, we obtain new results for the excess risk of principal component analysis and the matrix denoising model.

Key words: Van Trees inequality, Cramér-Rao inequality, group equivariance, orthogonal group, Haar measure, principal subspace, doubly substochastic matrix

1 Introduction

Many learning algorithms and statistical procedures rely on the spectral decomposition of some empirical matrix or operator. Leading examples are principal component analysis (PCA) and its extensions to kernel PCA or manifold learning. In modern statistics and data science, such methods are typically studied in a high-dimensional or infinite-dimensional setting; see e.g. [13, 29, 25] for an overview. Moreover, a major focus is on results that are non-asymptotic, that is, one seeks results that depend optimally on the underlying parameters (e.g. sample size and dimension); see e.g. [17, 20] for two more recent developments.

In this paper, we are concerned with non-asymptotic lower bounds for the estimation of principal subspaces, e.g. the eigenspace of the, say d, leading eigenvalues. As stated in [5], it is highly nontrivial to obtain such lower bounds which depends optimally on all underlying parameters, in particular the eigenvalues and d. In fact, in contrast to asymptotic settings in which one can e.g. apply the local asymptotic minimax theorem due to Hájek [12], it seems unavoidable to use some more specific (resp. deeper) facts on the underlying parameter space of all orthonormal bases in order to obtain non-asymptotic lower bounds. A state-of-the-art result, obtained in

Martin Wahl
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin
e-mail: martin.wahl@math.hu-berlin.de

[5] and [26], provides a non-asymptotic lower bound for a spiked covariance model with two groups of eigenvalues. To state their result, consider the statistical model defined by

$$(\mathbb{P}_U)_{U\in O(p)}, \qquad \mathbb{P}_U = \mathcal{N}(0, U\Lambda U^T)^{\otimes n}, \tag{1}$$

where $O(p)$ denotes the orthogonal group, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ is a diagonal matrix with $\lambda_1 \geq \dots \geq \lambda_p > 0$ and $\mathcal{N}(0, U\Lambda U^T)$ denotes a Gaussian distribution with expectation zero and covariance matrix $U\Lambda U^T$. This statistical model provides a decision-theoretic framework for principal component analysis (PCA). It corresponds to observing n independent $\mathcal{N}(0, U\Lambda U^T)$-distributed random variables $X_1, \dots, X_n$, and we will write $\mathbb{E}_U$ to denote expectation with respect to $X_1, \dots, X_n$ having law $\mathbb{P}_U$. Moreover, in this model, the d-th principal subspace (resp. its corresponding orthogonal projection) is given by $P_{\leq d}(U) = \sum_{i\leq d} u_i u_i^T$, where $u_1, \dots, u_p$ are the columns of $U \in O(p)$.

Theorem 1 ([5]) *Consider the statistical model* (1) *with* $\lambda_1 = \dots = \lambda_d > \lambda_{d+1} = \dots = \lambda_p > 0$. *Then there is an absolute constant* $c > 0$ *such that*

$$\inf_{\hat{P}} \sup_{U\in O(p)} \mathbb{E}_U \|\hat{P} - P_{\leq d}(U)\|_2^2 \geq c \cdot \min\Big(\frac{d(p-d)}{n}\frac{\lambda_d\lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2}, d, p-d\Big),$$

where the infimum is taken over all estimators $\hat{P} = \hat{P}(X_1, \dots, X_n)$ *with values in the class of all orthogonal projections on* $\mathbb{R}^p$ *of rank* d *and* $\|\cdot\|_2$ *denotes the Hilbert-Schmidt norm.*

The proof is based on applying lower bounds under metric entropy conditions [30, 3, 19] combined with the metric entropy of the Grassmann manifold [21]. This so-called Grassmann approach has been applied to many other principal subspaces estimation problems and spiked structures; see e.g. [6, 4, 10, 18]. In principle, this approach can also be applied to the infinite-dimensional case by considering finite-dimensional (spiked) submodels. Yet, since this leads to lower bounds of a specific multiplicative form, it seems difficult to recover the optimal weighted eigenvalue expressions $2\sum_{i\leq d}\sum_{j>d} \lambda_i\lambda_j/(\lambda_i - \lambda_j)^2$ appearing in the non-asymptotic upper bounds from [16, 22] and in the asymptotic limit [7].

To overcome this difficulty, [27] proposed a new approach based on a version of the van Trees inequality with reference measure being the Haar measure on the special orthogonal group $SO(p)$. The key ingredient was to explore the group equivariance of the model (1), allowing to derive a non-asymptotic analogue of the local asymptotic minimax theorem. For instance, using also large deviations techniques to design optimal prior densities, a main consequence of the developed theory is as follows.

Theorem 2 ([27]) *Consider the statistical model* (1) *with* $\lambda_1 \geq \dots \geq \lambda_p > 0$. *Then there are absolute constants* $c, C > 0$ *such that for every* $h \geq C$, *we have*

$$\inf_{\hat{P}} \int_{SO(p)} \mathbb{E}_U \|\hat{P} - P_{\leq d}(U)\|_2^2 \, \pi_h(\mathrm{tr}\, U) dU \geq c \cdot \sum_{i\leq d}\sum_{j>d} \min\Big(\frac{1}{n}\frac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}, \frac{1}{h^2 p}\Big),$$

where the infimum is taken over all $\mathbb{R}^{p\times p}$-valued estimators $\hat{P} = \hat{P}(X_1, \dots, X_n)$, dU denotes the Haar measure on $SO(p)$, $\operatorname{tr} U$ denotes the trace of U, and the prior probability density π_h is given by

$$\pi_h(\operatorname{tr} U) = \frac{\exp(hp \operatorname{tr} U)}{\int_{SO(p)} \exp(hp \operatorname{tr} U)\, dU}, \qquad h > 0.$$

Theorem 2 is a slight reformulation of [27, Theorem 2], where the special choice $h = C$ is considered. Obviously, for this (from a non-asymptotic point of view) optimal choice for h, Theorem 2 implies Theorem 1, as can be seen from inserting $2d(p-d)/p \geq \min(d, p-d)$. Moreover, as shown in [27, Section 1.3], Theorem 2 can be used to derive tight non-asymptotic minimax lower bounds for standard examples from functional PCA or kernel PCA, including exponentially and polynomially decaying eigenvalues.

The goal of this paper is to extend the theory of [27] in several directions. First, we provide a lower bound for the excess risk of PCA. This loss function is not covered in [27] and a variation of the approach is needed to deal with it. Second, we provide a slightly complementary (and less general) van Trees-type inequality tailored for principal subspace estimation problems and dealing solely with the uniform prior. Interestingly, such uniform prior densities lead to trivial results in the Trees inequality from [27] (as well as in previous classical van Trees approaches [24, 11]). Indeed, while the Fisher information of such uniform priors is zero, the average in the numerator is zero as well, meaning that we get the trivial lower bound. Finally, we provide lower bounds that are characterized by doubly substochastic matrices whose entries are bounded by the different Fisher information directions, confirming previous non-asymptotic upper bounds that hold for the principal subspaces of the empirical covariance operator [22, Section 2.3].

2 A van Trees inequality for the estimation of principal subspaces

In this section, we state a general van Trees-type inequality tailored for principal subspace estimation problems. Applications to more concrete settings are presented in Section 4. Let $(\mathcal{X}, \mathcal{F}, (\mathbb{P}_U)_{U\in O(p)})$ be a statistical model with parameter space being the orthogonal group $O(p)$. Let $(A, \langle\cdot,\cdot\rangle)$ be a real inner product space of dimension $m \in \mathbb{N}$ and let $\psi : O(p) \to A$ be a derived parameter. We suppose that $O(p)$ acts (from the left, measurable) on $\mathcal{X}$ and A such that

(A1) $(\mathbb{P}_U)_{U\in O(p)}$ is $O(p)$-equivariant, i.e. $\mathbb{P}_{VU}(VE) = \mathbb{P}_U(E)$ for all $U, V \in O(p)$ and all $E \in \mathcal{F}$;

(A2) ψ is $O(p)$-equivariant, i.e. $\psi(VU) = V\psi(U)$ for all $U, V \in O(p)$;

(A3) $\langle Ua, Ub\rangle = \langle a, b\rangle$ for all $a, b \in A$ and all $U \in O(p)$.

Condition (A1) says that for a random variable X with distribution $\mathbb{P}_U$ we have that VX has distribution $\mathbb{P}_{VU}$. For more background on statistical models under group action, the reader is deferred to [8, 9] and also to [27, Section 2.3].

Next, we specify the allowed loss functions. Let $(v_1, \dots, v_m) : O(p) \to A^m$ be such that for all $j = 1, \dots, m$,

(A4) $v_1(U), \dots, v_m(U)$ is an orthonormal basis of A for all $U \in O(p)$;
(A5) v_j are $O(p)$-equivariant, i.e. $v_j(VU) = Vv_j(U)$ for all $U, V \in O(p)$.

For $w \in \mathbb{R}^m_{>0}$ we now consider the loss function

$$l_w : O(p) \times A \to \mathbb{R}_{\geq 0}, \qquad l_w(U, a) = \sum_{k=1}^m w_k \langle v_k(U), a - \psi(U)\rangle^2.$$

If $w_1 = \cdots = w_m = 1$, then l_w does not depend on $v_1, \dots, v_m$ and is equal to the squared norm in A

$$l_{(1,\dots,1)}(U, a) = \|a - \psi(U)\|^2 = \langle a - \psi(U), a - \psi(U)\rangle. \tag{2}$$

For general w, the loss function l_w is itself invariant in the sense that

$$l_w(VU, Va) = l_w(U, a) \quad \text{for all } U, V \in O(p), a \in A, \tag{3}$$

as can be seen from (A2), (A3) and (A5). For an estimator $\hat{\psi}(X)$ based on an observation X from the experiment, the l_w-risk is defined as $\mathbb{E}_U l_w(U, \hat{\psi}(X))$, where $\mathbb{E}_U$ denotes expectation with respect to X having distribution $\mathbb{P}_U$.

In order to formulate our abstract main result, we also need some differentiability conditions on ψ and the v_j. We assume that ψ and v_j are differentiable at the identity matrix I_p in the sense that for all $\xi \in \mathfrak{so}(p)$, all $a \in A$ and all $j = 1, \dots, m$, we have

(A6) $\lim_{t\to 0} \left\langle \frac{\psi(\exp(t\xi)) - \psi(I_p)}{t}, a \right\rangle = \langle d\psi(I_p)\xi, a\rangle$;
(A7) $\lim_{t\to 0} \left\langle \frac{v_j(\exp(t\xi)) - v_j(I_p)}{t}, a \right\rangle = \langle dv_j(I_p)\xi, a\rangle$.

Here, $d\psi(I_p)\xi$ and $dv_j(I_p)\xi$ denote the directional derivatives at I_p defined on the Lie algebra $\mathfrak{so}(p)$ on $SO(p)$ (i.e. the tangent space of $O(p)$ at I_p). Since A is finite-dimensional, conditions (A6) and (A7) can also formulated in a norm-sense, e.g. $\lim_{t\to 0} \|(\psi(\exp(t\xi)) - \psi(I_p))/t - d\psi(I_p)\xi\| = 0$ for all $\xi \in \mathfrak{so}(p)$. For some background on the special orthogonal group $SO(p)$ and its Lie algebra $\mathfrak{so}(p)$ see e.g. [27, Section 2.1].

Proposition 1 *Assume (A1)–(A7). Let $\xi_1, \dots, \xi_m \in \mathfrak{so}(p)$ be such that $\mathbb{P}_{\exp(t\xi_j)} \ll \mathbb{P}_{I_p}$ for all $j = 1, \dots, m$ and all t small enough. Suppose that there are $a_1, \dots, a_m \in$ $(0, \infty]$ such that for all $j = 1, \dots, m$,*

$$\lim_{t\to 0} \frac{\chi^2(\mathbb{P}_{\exp(t\xi_j)}, \mathbb{P}_{I_p})}{t^2} = a_j^{-1}, \tag{4}$$

where $\chi^2(\cdot,\cdot)$ denotes the χ^2-divergence (cf. Remark 1 below). Then, for all estimators $\hat{\psi} = \hat{\psi}(X)$ with values in A, we have

$$\int_{O(p)} \mathbb{E}_U l_w(U, \hat{\psi}(X))\, dU \geq \frac{\Big(\sum_{j=1}^m \langle v_j(I_p), d\psi(I_p)\xi_j\rangle\Big)^2}{\sum_{j=1}^m w_j^{-1} a_j^{-1} + \sum_{k=1}^m w_k^{-1}\Big(\sum_{j=1}^m \langle v_k(I_p), dv_j(I_p)\xi_j\rangle\Big)^2}.$$

Remark 1 The χ^2-divergence between two probability measures $\mathbb{P} \ll \mathbb{Q}$ is defined as $\chi^2(\mathbb{P},\mathbb{Q}) = \int (\frac{d\mathbb{P}}{d\mathbb{Q}})^2\, d\mathbb{Q} - 1$.

Remark 2 In the applications, the $a_j^{-1} \in [0,\infty)$ will be the different Fisher information directions. We use the inverse notation because it will be more suitable to solve the final optimization problem in Section 5.2.

Remark 3 Let us briefly compare Proposition 1 to [27, Proposition 1 and Theorem 3], where a more general van Trees inequality is presented. In fact, the bound [27, Theorem 3] has a more classical form and involves a general prior, an average over the prior in the numerator and Fisher informations of the prior in the denominator. Yet, while these Fisher informations are zero for the uniform prior considered in Proposition 1, the averages in the denominator are zero as well. Hence, [27, Theorem 3] is trivial for the uniform prior. The reason that we can deal with the uniform prior lies in the fact that in addition to the equivariance condition (A1) for the statistical model, we also require equivariance of the derived parameter and invariance of the loss function.

3 Proof of Proposition 1

We provide a proof which manifests Proposition 1 as a version of the Cramér-Rao inequality for equivariant estimators.

3.1 Reduction to a pointwise risk

We use [27, Lemma 4] in order to reduce the Bayes risk of Proposition 1 to a pointwise risk minimized over the class of all equivariant estimators. For completeness we briefly repeat the (standard) argument. Let $\tilde{\psi}$ be an arbitrary estimator with values in A. Without loss of generality we may restrict ourselves to estimators with bounded Hilbert-Schmidt norm $\sup_{x\in\mathcal{X}} \|\tilde{\psi}(x)\|_2 < \infty$. (Indeed, by (A2) and (A3) we know that $\sup_{U\in O(p)} \|\psi(U)\| = C < \infty$. Hence, setting $\tilde{\psi}(x) = 0$ whenever $\|\tilde{\psi}(x)\|_2 > C_w = 2C(w_{\max}/w_{\min})^{1/2}$ with $w_{\max} = \max_k w_k$ and $w_{\min} = \min_k w_k$, the l_w-risk is lowered. To see this use that for such an x, we have $l_w^{1/2}(U,0) \leq w_{\max}^{1/2} C$, while, $l_w^{1/2}(U,\tilde{\psi}) > w_{\min}^{1/2} C_w - w_{\max}^{1/2} C = w_{\max}^{1/2} C$.) Hence, we can construct

$$\hat{\psi}(x) = \int_{O(p)} V^T \tilde{\psi}(Vx)\, dV, \qquad x \in \mathcal{X}.$$

By [27, Lemma 4] this defines an $O(p)$-equivariant estimator (i.e. it holds that $\hat{\psi}(Ux) = U\hat{\psi}(x)$ for all $x \in \mathcal{X}$ and all $U \in O(p)$) satisfying

$$\int_{O(p)} \mathbb{E}_U l_w(U, \tilde{\psi}(X))\, dU \geq \int_{O(p)} \mathbb{E}_U l_w(U, \hat{\psi}(X))\, dU,$$

where we used (A1) and the facts that the loss function l_w is convex in the second argument and satisfies (3). Moreover, using that $\hat{\psi}$ is $O(p)$-equivariant, it follows again from [27, Lemma 4] that the risk $\mathbb{E}_U l_w(U, \hat{\psi}(X))$ is constant over $U \in O(p)$. Hence, we arrive at

$$\inf_{\tilde{\psi}} \int_{O(p)} \mathbb{E}_U l_w(U, \tilde{\psi}(X))\, dU \geq \inf_{\hat{\psi}\ O(p)\text{-equivariant}} \mathbb{E}_{I_p} l_w(I_p, \hat{\psi}(X)),$$

and it suffices to lower bound the right-hand side.

3.2 A pointwise Cramér-Rao inequality for equivariant estimators

The classical Cramér-Rao inequality provides a lower bound for the (co-)variance of unbiased estimators. In this section, we show that in our context, a similar lower bound can be proved for the class of all equivariant estimators.

Lemma 1 *Assume (A1)–(A7). Let $\xi_1, \dots, \xi_m \in \mathfrak{so}(p)$ be such that $\mathbb{P}_{\exp(t\xi_j)} \ll \mathbb{P}_{I_p}$ for all $j = 1, \dots, m$ and all t small enough. Suppose that there are $a_1, \dots, a_m \in (0, \infty]$ such that $\lim_{t \to 0} \chi^2(\mathbb{P}_{\exp(t\xi_j)}, \mathbb{P}_{I_p})/t^2 = a_j^{-1}$ for all $j = 1, \dots, m$. Then, for any $O(p)$-equivariant estimators $\hat{\psi}(X)$ with values in A, we have*

$$\mathbb{E}_{I_p} l_w(I_p, \hat{\psi}(X)) \geq \frac{\Big(\sum_{j=1}^m \langle v_j(I_p), d\psi(I_p)\xi_j\rangle\Big)^2}{\sum_{j=1}^m w_j^{-1} a_j^{-1} + \sum_{k=1}^m w_k^{-1}\Big(\sum_{j=1}^m \langle v_k(I_p), dv_j(I_p)\xi_j\rangle\Big)^2}.$$

Proof For $U_j = \exp(t\xi_j)$, $j = 1, \dots, m$, consider the expression

$$\sum_{j=1}^m \mathbb{E}_{I_p}\langle v_j(I_p), \hat{\psi}(X) - \psi(I_p)\rangle - \sum_{j=1}^m \mathbb{E}_{I_p}\langle v_j(I_p), \hat{\psi}(X) - \psi(U_j^T)\rangle. \tag{5}$$

Clearly (5) is equal to

$$\sum_{j=1}^m \langle v_j(I_p), \psi(U_j^T) - \psi(I_p)\rangle. \tag{6}$$

On the other hand, using (A1)–(A3), (A5) and the equivariance of $\hat{\psi}$, we have

$$\begin{aligned}
&\mathbb{E}_{I_p}\langle v_j(I_p), \hat{\psi}(X) - \psi(U_j^T)\rangle \\
&= \mathbb{E}_{U_j}\langle v_j(I_p), \hat{\psi}(U_j^T X) - \psi(U_j^T)\rangle = \mathbb{E}_{U_j}\langle v_j(I_p), U_j^T\hat{\psi}(X) - U_j^T\psi(I_p)\rangle \\
&= \mathbb{E}_{U_j}\langle v_j(U_j), \hat{\psi}(X) - \psi(I_p)\rangle = \mathbb{E}_{I_p}\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X)\langle v_j(U_j), \hat{\psi}(X) - \psi(I_p)\rangle.
\end{aligned}$$

Hence, (5) is also equal to

$$\mathbb{E}_{I_p}\sum_{j=1}^m\langle v_j(I_p), \hat{\psi}(X) - \psi(I_p)\rangle - \mathbb{E}_{I_p}\sum_{j=1}^m\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X)\langle v_j(U_j), \hat{\psi}(X) - \psi(I_p)\rangle \qquad (7)$$

Using (5)–(7), Parseval's identity, (A4) and the Cauchy-Schwarz inequality (twice), we arrive at

$$\begin{aligned}
&\Big(\sum_{j=1}^m\langle v_j(I_p), \psi(U_j^T) - \psi(I_p)\rangle\Big)^2 \\
&= \Big(\mathbb{E}_{I_p}\sum_{j=1}^m\langle v_j(I_p), \hat{\psi}(X) - \psi(I_p)\rangle - \mathbb{E}_{I_p}\sum_{j=1}^m\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X)\langle v_j(U_j), \hat{\psi}(X) - \psi(I_p)\rangle\Big)^2 \\
&= \Big(\mathbb{E}_{I_p}\sum_{k=1}^m\Big\{\sum_{j=1}^m\langle v_j(I_p), v_k(I_p)\rangle - \sum_{j=1}^m\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X)\langle v_j(U_j), v_k(I_p)\rangle\Big\}\langle v_k(I_p), \hat{\psi}(X) - \psi(I_p)\rangle\Big)^2 \\
&\leq \Big(\mathbb{E}_{I_p}\sum_{k=1}^m w_k\langle v_k(I_p), \hat{\psi}(X) - \psi(I_p)\rangle^2\Big) \\
&\quad\cdot\Big(\mathbb{E}_{I_p}\sum_{k=1}^m w_k^{-1}\Big\{\sum_{j=1}^m\langle v_j(I_p), v_k(I_p)\rangle - \sum_{j=1}^m\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X)\langle v_j(U_j), v_k(I_p)\rangle\Big\}^2\Big).
\end{aligned}$$

The first term on the right-hand side is equal to the l_w-risk at I_p. Moreover, the second term can be written as

$$\sum_{k=1}^m w_k^{-1}\Big\{\Big(\sum_{j=1}^m\langle v_j(U_j) - v_j(I_p), v_k(I_p)\rangle\Big)^2 + \mathbb{E}_{I_p}\Big(\sum_{j=1}^m\Big(\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X) - 1\Big)\langle v_j(U_j), v_k(I_p)\rangle\Big)^2\Big\}.$$

In particular, we have proved that

$$\mathbb{E}_{I_p} l_w(I_p, \hat{\psi}(X)) \geq \frac{D^2}{\sum_{k=1}^m w_k^{-1}(D_k^2 + \mathbb{E}_{I_p}(B_k + C_k)^2)} \qquad (8)$$

with

$$\begin{aligned}
D &= \sum_{j=1}^m \langle v_j(I_p), \psi(U_j^T) - \psi(I_p)\rangle,\\
D_k &= \sum_{j=1}^m \langle v_j(U_j) - v_j(I_p), v_k(I_p)\rangle,\\
B_k &= \sum_{j=1}^m \Big(\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X) - 1\Big)\langle v_j(I_p), v_k(I_p)\rangle,\\
C_k &= \sum_{j=1}^m \Big(\frac{d\mathbb{P}_{U_j}}{d\mathbb{P}_{I_p}}(X) - 1\Big)\langle v_j(U_j) - v_j(I_p), v_k(I_p)\rangle.
\end{aligned}$$

We now invoke a limiting argument to deduce Lemma 1 from (8). For this, recall that $U_j = \exp(t\xi_j)$, $\xi_j \in \mathfrak{so}(p)$, multiply numerator and denominator by $1/t^2$ and let $t \to 0$. First, by (A6) and (A7), we have

$$\begin{aligned}
\frac{1}{t}D &\to -\sum_{j=1}^m \langle v_j(I_p), d\psi(I_p)\xi_j\rangle,\\
\frac{1}{t}D_k &\to \sum_{j=1}^m \langle v_k(I_p), dv_j(I_p)\xi_j\rangle
\end{aligned}$$

as $t \to 0$. Moreover, by assumption (4), we have

$$\frac{1}{t^2}\sum_{k=1}^m w_k^{-1}\mathbb{E}_{I_p}B_k^2 = \frac{1}{t^2}\sum_{j=1}^m w_j^{-1}\chi^2(\mathbb{P}_{U_j}, \mathbb{P}_{I_p}) \to \sum_{j=1}^m w_j^{-1}a_j^{-1} \quad \text{as } t \to 0.$$

On the other hand, C_k is asymptotically negligible, as can be seen from

$$\frac{1}{t^2}\mathbb{E}_{I_p}C_k^2 \le \frac{1}{t^2}\Big(\sum_{j=1}^m \chi^2(\mathbb{P}_{U_j}, \mathbb{P}_{I_p})\Big)\Big(\sum_{j=1}^m \langle v_j(U_j) - v_j(I_p), v_k(I_p)\rangle^2\Big) \to 0$$

as $t \to 0$. Here, we used (4) and (A7). Thus,

$$\begin{aligned}
&\frac{1}{t^2}\Big|\sum_{k=1}^m w_k^{-1}\mathbb{E}_{I_p}(B_k + C_k)^2 - \sum_{k=1}^m w_k^{-1}\mathbb{E}_{I_p}B_k^2\Big|\\
&\le \frac{1}{t^2}\sum_{k=1}^m w_k^{-1}(2(\mathbb{E}_{I_p}B_k^2)^{1/2}(\mathbb{E}_{I_p}C_k^2)^{1/2} + \mathbb{E}_{I_p}C_k^2) \to 0
\end{aligned}$$

as $t \to 0$. The proof now follows from inserting these limits into (8). □

4 Applications

In this section, we specialize our lower bounds in the context of principal component analysis (PCA) and a low-rank denoising model. In doing so, we will focus on the derived parameter

$$\psi(U) = P_{\leq d}(U) = \sum_{i\leq d} u_i u_i^T, \qquad U \in O(p),$$

where $1 \leq d \leq p$ and $u_1, \ldots, u_p$ are the columns of $U \in O(p)$. This will correspond to the estimation of the d-th principal subspace. We discuss several loss functions based on the Hilbert-Schmidt distance and the excess risk in the reconstruction error.

4.1 PCA and the subspace distance

In this section, we consider the statistical model given in (1)

$$(\mathbb{P}_U)_{U\in O(p)}, \qquad \mathbb{P}_U = \mathcal{N}(0, U\Lambda U^T)^{\otimes n},$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p > 0$. The following theorem proved in Section 5 applies Proposition 1 to the above model, derived parameter $P_{\leq d}$, and loss function given by the Hilbert-Schmidt distance (cf. Section 5.1 below).

Theorem 3 *Consider the statistical model* (1). *Then, for each* $\delta > 0$, *we have*

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \|\hat{P} - P_{\leq d}(U)\|_2^2 \, dU \geq I_\delta$$

with infimum taken over all $\mathbb{R}^{p\times p}$*-valued estimators* $\hat{P} = \hat{P}(X_1, \ldots, X_n)$ *and*

$$I_\delta = \frac{1}{1+2\delta} \max \Big\{ \sum_{i\leq d} \sum_{j>d} x_{ij} \ : 0 \leq x_{ij} \leq \tfrac{2}{n} \tfrac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \quad \text{for all } i \leq d, j > d,$$
$$\sum_{i\leq d} x_{ij} \leq \delta \quad \text{for all } j > d,$$
$$\sum_{j>d} x_{ij} \leq \delta \quad \text{for all } i \leq d \Big\}.$$

Remark 4 We write $i \leq d$ for $i \in \{1, \ldots, d\}$ and $j > d$ for $j \in \{d+1, \ldots, p\}$.

Remark 5 A (non-square) matrix (x_{ij}) is called doubly substochastic (cf. [2, Section 2]) if

$$\begin{aligned} x_{ij} &\geq 0 \quad \text{for all} \quad i,j,\\ \sum_i x_{ij} &\leq 1 \quad \text{for all} \quad j,\\ \sum_j x_{ij} &\leq 1 \quad \text{for all} \quad i. \end{aligned}$$

Hence, choosing $\delta = 1$, Theorem 3 holds with

$$I_1 = \frac{1}{3}\max\Big\{\sum_{i\leq d}\sum_{j>d} x_{ij} : (x_{ij}) \text{ doubly substochastic with} \\ x_{ij} \leq \tfrac{2}{n}\tfrac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2} \quad \text{for all } i \leq d, j > d\Big\}.$$

Remark 6 That doubly substochastic matrices play a role is no coincidence. Such a structure also appears in the upper bounds for the principal subspaces of the empirical covariance operator; see e.g. [22]. To explain this, let $X, X_1, \dots, X_n$ be independent random variables with expectation zero and covariance matrix Σ and let $\hat{\Sigma} = n^{-1}\sum_{i=1}^n X_i X_i^T$ be the empirical covariance operator. Moreover, let $\lambda_1 \geq \cdots \geq \lambda_p$ (resp. $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$) be the eigenvalues of Σ (resp. $\hat{\Sigma}$) and let $u_1, \dots, u_p$ (resp. $\hat{u}_1, \dots, \hat{u}_p$) be the corresponding eigenvectors of Σ (resp. $\hat{\Sigma}$). Then, for $P_{\leq d} = \sum_{i\leq d} u_i u_i^T$ and $\hat{P}_{\leq d} = \sum_{i\leq d}\hat{u}_i\hat{u}_i^T$, we have (cf. [22, 16])

$$\|\hat{P}_{\leq d} - P_{\leq d}\|_2^2 = 2\sum_{i\leq d}\sum_{j>d} x_{ij} \quad \text{with} \quad x_{ij} = \langle u_i, \hat{u}_j\rangle^2.$$

There are two completely different possibilities to bound this Hilbert-Schmidt distance. First, by Bessel's inequality, we always have the trivial bounds

$$\sum_{i\leq d} x_{ij} \leq 1 \quad \text{and} \quad \sum_{j>d} x_{ij} \leq 1.$$

On the other hand, using perturbative methods, one e.g. has

$$n x_{ij} = n\langle u_i, \hat{u}_j\rangle^2 \xrightarrow{d} \mathcal{N}\Big(0, \frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}\Big),$$

see e.g. [1], and also [15] for a non-asymptotic version of this result. Hence, the lower bound in Theorem 3 can be interpreted as the fact that we can not do better than the best mixture of trivial and perturbative bounds.

Remark 7 A simple and canonical choice of the x_{ij} in Theorem 3 is given by

$$x_{ij} = \min\Big(\frac{2}{n}\frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}, \frac{1}{p}\Big),$$

in which case we rediscover the bound [27, Theorem 1]. Yet, let us point out that the result in Theorem 2 is stronger in the sense that it allows for priors that are highly concentrated around I_p (cf. h of size $\sqrt{n}$), while Theorem 3 provides a lower bound for the uniform prior.

Remark 8 In general it seems difficult to find a simple closed form expression for the lower bound in Theorem 3. One exception is given e.g. by the case $d = 1$, in which case we have

$$I_1 = \frac{1}{3} \min\Big(\frac{2}{n} \sum_{j>1} \frac{\lambda_1\lambda_j}{(\lambda_1 - \lambda_j)^2}, 1\Big).$$

Remark 9 Using decision-theoretic arguments, the result can be extended to random variables with values in a Hilbert space; see [27, Section 1.4] for the details.

4.2 PCA and the excess risk

Theorem 3 provides a lower bound for the squared Hilbert-Schmidt distance $\|\hat{P} - P_{\le d}(U)\|_2^2$. If the estimator $\hat{P}$ is itself an orthogonal projection of rank d, then $\|\hat{P} - P_{\le d}(U)\|_2^2$ is equal to $\sqrt{2}$ times the Euclidean norm of the sines of the canonical angles between the corresponding subspaces, see e.g. [2, Chapter VII.1]. This so-called $\sin\Theta$ distance is a well-studied distance in linear algebra, numerical analysis and statistics; see e.g. [2, 14, 31]. In the context of statistical learning, another important loss function arises if one introduces PCA as an empirical risk minimization problem with respect to the reconstruction error.

For $1 \le d \le p$, let $\mathcal{P}_d$ be the set of all orthogonal projections $P : \mathbb{R}^p \to \mathbb{R}^p$ of rank d. Consider the statistical model defined by (1). Then the reconstruction error is defined by

$$R_U(P) = \mathbb{E}_U \|X - PX\|^2, \qquad P \in \mathcal{P}_d, U \in O(p)$$

and it is easy to see that (cf. [22])

$$P_{\le d}(U) \in \arg\min_{P\in\mathcal{P}_d} R_U(P).$$

Hence, the performance of $\hat{P} \in \mathcal{P}_d$ can be measured by its excess risk defined by

$$\mathcal{E}_U(\hat{P}) = R_U(\hat{P}) - \min_{P\in\mathcal{P}_d} R_U(P) = R_U(\hat{P}) - R_U(P_{\le d}(U)). \tag{9}$$

In Section 5.3, we show that $\mathcal{E}_U(\hat{P})$ can be written in the form l_w for some suitable choices for A, v and w, and Proposition 1 yields the following.

Theorem 4 *Consider the statistical model* (1) *with the excess risk loss function from* (9)*. Assume that $\lambda_d > \lambda_{d+1}$. Then, for any natural numbers r, s satisfying $1 \le r \le d < s \le p$ and $\mu \in (\lambda_{d+1}, \lambda_d)$, we have*

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \mathcal{E}_U(\hat{P})\, dU \geq J_\mu$$

with infimum taken over all $\mathcal{P}_d$-valued estimators $\hat{P} = \hat{P}(X_1, \dots, X_n)$ and

$$\begin{aligned} J_\mu = \frac{1}{3} \max \Big\{ \sum_{i \leq r} \sum_{j > s} x_{ij} \ : \ 0 \leq x_{ij} \leq \tfrac{1}{n} \tfrac{\lambda_i \lambda_j}{\lambda_i - \lambda_j} \quad & \textit{for all } i \leq r, j > s, \\ \sum_{i \leq r} x_{ij} \leq \mu - \lambda_j \quad & \textit{for all } j > s, \\ \sum_{j > s} x_{ij} \leq \lambda_i - \mu \quad & \textit{for all } i \leq r \Big\}. \end{aligned}$$

Remark 10 The lower bound has a similar structure than the mixture bounds established in [22]. In particular, as in the case of the Hilbert-Schmidt distance, the term $n^{-1}\lambda_i\lambda_j/(\lambda_i - \lambda_j)$ corresponds to the size of certain weighted projector norms, while the other two constrains correspond to trivial bounds. The lower bound strenghtens the reciprocal dependence of the excess risk on spectral gaps (the excess risk might be small in both cases, small and large gaps); see e.g. [22, Section 2.3].

An important special case is given when the last two restrictions in J_μ are satisfied for $r = d$, $s = d + 1$ and $x_{ij} = n^{-1}\lambda_i\lambda_j/(\lambda_i - \lambda_j)$. Then, letting $\mu = (\lambda_d + \lambda_{d+1})/2$, they are satisfied if and only if

$$\frac{\lambda_{d+1}}{\mu - \lambda_{d+1}} \sum_{i \leq d} \frac{\lambda_i}{\lambda_i - \lambda_{d+1}} \leq n \quad \text{and} \quad \frac{\lambda_d}{\lambda_d - \mu} \sum_{j > d} \frac{\lambda_j}{\lambda_d - \lambda_j} \leq n,$$

as can be seen from a monotonicity argument. A simple modification leads to the following corollary.

Corollary 1 *We have*

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \mathcal{E}_U(\hat{P})\, dU \geq \frac{1}{3n} \sum_{i \leq d} \sum_{j > d} \frac{\lambda_i \lambda_j}{\lambda_i - \lambda_j},$$

provided that

$$\frac{\lambda_d}{\lambda_d - \lambda_{d+1}} \Big(\sum_{i \leq d} \frac{\lambda_i}{\lambda_i - \lambda_{d+1}} + \sum_{j > d} \frac{\lambda_j}{\lambda_d - \lambda_j} \Big) \leq \frac{n}{2}. \tag{10}$$

Remark 11 Condition (10) is the main condition of [22] under which perturbation bounds for the empirical covariance operator are developed (cf. [22, Remark 3.15]). If it is not satisfied, then the accuracy of empirical spectral projectors is expected to break down; see also [28]. The quantity in the brackets is called in [15, 16] relative rank.

The involved eigenvalue expressions in Corollary 1 can be easily evaluated if the λ_j have e.g. exponential or polynomial decay (cf. [23]).

Example 1 If for some $\alpha > 0$, we have $\lambda_j = j^{-\alpha-1}$, $j = 1, \dots, p$, then there are constants $c_1, c_2 > 0$ depending only on α such that

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \mathcal{E}_U(\hat{P})\, dU \geq c_1 \frac{d^{2-\alpha}}{n}, \quad \text{provided that} \quad d^2 \log d \leq c_2 n.$$

Moreover, if for some $\alpha > 0$, we have $\lambda_j = e^{-\alpha j}$, $j = 1, \dots, p$, then there are constants $c_1, c_2 > 0$ depending only on α such that

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \mathcal{E}_U(\hat{P})\, dU \geq c_1 \frac{de^{-\alpha d}}{n}, \quad \text{provided that} \quad d \leq c_2 n.$$

4.3 Low-rank matrix denoising

For a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ with $\lambda_1 \geq \dots \geq \lambda_p \geq 0$, we consider the statistical model defined by

$$(\mathbb{P}_U)_{U \in O(p)}, \qquad \mathbb{P}_U = \mathcal{N}(\operatorname{vec}(U\Lambda U^T), \epsilon^2 I_{p^2})^{\otimes n}, \tag{11}$$

where $\operatorname{vec}(U\Lambda U^T)$ denotes the vectorization of $U\Lambda U^T$. This statistical model corresponds to observing

$$X = U\Lambda U^T + \epsilon(\xi_{ij}) \in \mathbb{R}^{p \times p},$$

with ξ_{ij} being independent Gaussian random variables with expectation 0 and variance 1. Similarly, it is also possible to consider a GOE matrix in which case one would have a symmetric perturbation. The following theorem is the analogue of Theorem 3.

Theorem 5 *Consider the statistical model* (11). *Then for each $\delta > 0$, we have*

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U \|\hat{P} - P_{\leq d}(U)\|_2^2\, dU \geq I'_\delta$$

with infimum taken over all $\mathbb{R}^{p \times p}$-valued estimators $\hat{P} = \hat{P}(X_1, \dots, X_n)$ and

$$\begin{aligned} I'_\delta = \frac{1}{1+2\delta} \max \Big\{ \sum_{i \leq d} \sum_{j > d} x_{ij} \ : 0 \leq x_{ij} \leq \tfrac{\epsilon^2}{(\lambda_i - \lambda_j)^2} \quad & \textit{for all } i \leq d, j > d, \\ \sum_{i \leq d} x_{ij} \leq \delta \quad & \textit{for all } j > d, \\ \sum_{j > d} x_{ij} \leq \delta \quad & \textit{for all } i \leq d \Big\}. \end{aligned}$$

Example 2 Suppose that $\operatorname{rank}(\Lambda) = d \leq p - d$. Then, setting

$$x_{ij} = \min\Big(\frac{\epsilon^2}{\lambda_i^2}, \frac{1}{p-d}\Big), \quad \text{we get} \quad I_1' \geq \frac{1}{3}\sum_{i\leq d} \min\Big(\frac{\epsilon^2(p-d)}{\lambda_i^2}, 1\Big).$$

5 Proofs for Section 4

In this section we show how Theorems 3–5 can be obtained by an application of Proposition 1.

5.1 Specialization to principal subspaces

We start with specializing Proposition 1 in the case where

$$\psi(U) = P_{\leq d}(U) = \sum_{i\leq d} u_i u_i^T = U \sum_{i\leq d} e_i e_i^T U^T, \qquad U \in O(p),$$

Here $u_i = Ue_i$ is the j-th column of U and $e_1, \dots, e_p$ denotes the standard basis in $\mathbb{R}^p$. We consider $A = \mathbb{R}^{p\times p}$ endowed with the trace inner product $\langle a, b\rangle_2 = \operatorname{tr}(a^T b)$, $a, b \in \mathbb{R}^{p\times p}$ and choose

$$v : O(p) \to \mathbb{R}^{p\times p}, \qquad v(U) = (u_k u_l^T)_{1\leq k,l\leq p}.$$

Hence, for $w \in \mathbb{R}^{p\times p}_{>0}$, we consider the loss function defined by

$$l_w(U, a) = \sum_{k=1}^{p}\sum_{l=1}^{p} w_{kl}\langle u_k u_l^T, a - P_{\leq d}(U)\rangle.$$

In particular, if $w_{kl} = 1$ for all k, l, then $l_w(U, a) = \|a - P_{\leq d}(U)\|_2^2 = \langle a - P_{\leq d}(U), a - P_{\leq d}(U)\rangle_2$ is the squared Hilbert-Schmidt (or Frobenius) distance. Note that, in contrast to Section 2, we consider a double index in this section. We equip A with the group action given by conjugation $U \cdot a = UaU^T$, $a \in \mathbb{R}^{p\times p}$, $U \in O(p)$. Using this definition it is easy to see that (A2), (A3), (A4) and (A5) are satisfied. Moreover, the following lemma verifies (A6) and (A7) in this case.

Lemma 2 *For $\xi \in \mathfrak{so}(p)$, we have*

(i) $dP_{\leq d}(I_p)\xi = \xi \sum_{i\leq d} e_i e_i^T - \sum_{i\leq d} e_i e_i^T \xi$,
(ii) $dv_{ij}(I_p)\xi = \xi e_i e_j^T - e_i e_j^T \xi$.

In particular, for $i \neq j$ and $L^{(ij)} = e_i e_j^T - e_j e_i^T \in \mathfrak{so}(p)$, we have

(i) $dP_{\leq d}(I_p)L^{(ij)} = -dP_{\leq d}(I_p)L^{(ji)} = -e_i e_j^T - e_j e_i^T$ *if $i \leq d$ and $j > d$,*
(ii) $dv_{ij}(I_p)L^{(ij)} = e_i e_i^T - e_j e_j^T$.

Remark 12 We have $dP_{\leq d}(I_p)L^{(kl)} = 0$ if $k, l \leq d$ or $k, l > d$.

Proof By definition, for $U \in SO(p)$ and $\xi \in \mathfrak{so}(p)$, we have $dP_{\leq d}(I_p)\xi = f'(0)$ with $f : \mathbb{R} \to \mathbb{R}^{p\times p}, t \mapsto \sum_{i\leq d}(\exp(t\xi)e_i)(\exp(t\xi)e_i)^T$. Hence, using $(d/dt)\exp(t\xi) = \xi\exp(t\xi)$, (i) follows. Claim (ii) can be shown analogously and (iii) and (iv) follow from inserting $\xi = L^{(ij)}$ into (i) and (ii), respectively. □

Corollary 2 *Consider the above setting with $\psi = P_{\leq d}$. Suppose that (A1) holds and that there is a bilinear form $\mathcal{I} : \mathfrak{so}(p) \times \mathfrak{so}(p) \to \mathbb{R}$ such that*

$$\lim_{t\to 0} \frac{\chi^2(\mathbb{P}_{\exp(t\xi)}, \mathbb{P}_{I_p})}{t^2} = \mathcal{I}(\xi, \xi) \qquad \textit{for all } \xi \in \mathfrak{so}(p). \tag{12}$$

Let $I \subseteq \{1, \ldots, d\}$ and $J \subseteq \{d+1, \ldots, p\}$. Then, for every z_{ij}, $i \in I$, $j \in J$, we have

$$\inf_{\hat{P}} \int_{O(p)} \mathbb{E}_U l_w(U, \hat{P})\, dU \tag{13}$$
$$\geq \frac{\Big(\sum_{i\in I}\sum_{j\in J} z_{ij}\Big)^2}{\sum_{i\in I}\sum_{j\in J}((w_{ij}+w_{ji})a_{ij})^{-1}z_{ij}^2 + \sum_{i\in I} w_{ii}^{-1}\Big(\sum_{j\in J} z_{ij}\Big)^2 + \sum_{j\in J} w_{jj}^{-1}\Big(\sum_{i\in I} z_{ij}\Big)^2},$$

where $a_{ij}^{-1} = \mathcal{I}(L^{(ij)}, L^{(ij)})$ and $L^{(ij)} = e_ie_j^T - e_je_i^T$, $i \in I, j \in J$.

Proof We choose $\xi^{(ij)} = y_{ij}L^{(ij)}$ and $\xi^{(ji)} = -y_{ji}L^{(ji)} = y_{ji}L^{(ij)}$ for $i \in I$ and $j \in J$ and we set $\xi^{(lk)} = 0$ in all other cases. Then, by Lemma 2, the sum in the numerator of Proposition 1 is equal to

$$\sum_{i\in I}\sum_{j\in J}\langle v_{ij}(I_p), dP_{\leq d}(I_p)\xi^{(ij)}\rangle + \sum_{j\in J}\sum_{i\in I}\langle v_{ji}(I_p), dP_{\leq d}(I_p)\xi^{(ji)}\rangle$$
$$= \sum_{i\in I}\sum_{j\in J}\big(y_{ij}\langle e_ie_j^T, -e_ie_j^T - e_je_i^T\rangle + y_{ji}\langle e_je_i^T, -e_ie_j^T - e_je_i^T\rangle\big) = -\sum_{i\in I}\sum_{j\in J} y_{ij} + y_{ji}.$$

On the other hand, for $1 \leq k, l \leq p$, the term in the squared brackets in the denominator is equal to

$$\sum_{i\in I}\sum_{j\in J}\langle v_{kl}(I_p), dv_{ij}(I_p)\xi^{(ij)}\rangle + \sum_{j\in J}\sum_{i\in I}\langle v_{kl}(I_p), dv_{ji}(I_p)\xi^{(ji)}\rangle$$
$$= \sum_{i\in I}\sum_{j\in J}(y_{ij}\langle e_ke_l^T, e_ie_i^T - e_je_j^T\rangle - y_{ji}\langle e_ke_l^T, e_je_j^T - e_ie_i^T\rangle)$$

and the latter is equal to

$$\begin{cases} \sum_{j\in J} y_{kj} + y_{jk}, & k = l \in I, \\ \sum_{i\in I} y_{ik} + y_{ki}, & k = l \in J, \\ 0, & \text{else.} \end{cases}$$

Hence the second term in the denominator is equal to

$$\sum_{i\in I} w_{ii}^{-1}\Big(\sum_{j\in J} y_{ij} + y_{ji}\Big)^2 + \sum_{j\in J} w_{jj}^{-1}\Big(\sum_{i\in I} y_{ij} + y_{ji}\Big)^2$$

Finally, the Fisher information term is equal to

$$\sum_{i\in I}\sum_{j\in J}\big(w_{ij}^{-1}\mathcal{I}(\xi^{(ij)},\xi^{(ij)}) + w_{ji}^{-1}\mathcal{I}(\xi^{(ji)},\xi^{(ji)})\big) = \sum_{i\in I}\sum_{j\in J}(w_{ij}^{-1}a_{ij}^{-1}y_{ij}^2 + w_{ji}^{-1}a_{ij}^{-1}y_{ji}^2).$$

Plugging all these formulas into Proposition 1, we get that, for every $y_{ij}, y_{ji} \in \mathbb{R}$, $i \in I$, $j \in J$, the left-hand side in (13) is lower bounded by

$$\frac{\Big(\sum_{i\in I}\sum_{j\in J} y_{ij} + y_{ji}\Big)^2}{\sum_{i\in I}\sum_{j\in J}(w_{ij}^{-1}a_{ij}^{-1}y_{ij}^2 + w_{ji}^{-1}a_{ij}^{-1}y_{ji}^2) + \sum_{i\in I} w_{ii}^{-1}\Big(\sum_{j\in J} y_{ij} + y_{ji}\Big)^2 + \sum_{j\in J} w_{jj}^{-1}\Big(\sum_{i\in I} y_{ij} + y_{ji}\Big)^2}.$$

For $a, b \geq 0$ and $z \in \mathbb{R}$, it is easy to that minimizing $a^{-1}x^2 + b^{-1}y^2$ subject to $x + y = z$ leads to the value $(a+b)^{-1}z^2$. Hence, using this with $a = w_{ij}a_{ij}$, $b = w_{ji}a_{ij}$ and $z = z_{ij}$, the claim follows. □

5.2 A simple optimization problem

We now consider the optimization problem

$$\max_{\substack{z_{ij}\in\mathbb{R} \\ i\in I, j\in J}} \frac{\Big(\sum_{i\in I}\sum_{j\in J} z_{ij}\Big)^2}{\sum_{i\in I}\sum_{j\in J} b_{ij}^{-1}z_{ij}^2 + \sum_{i\in I} w_{ii}^{-1}\Big(\sum_{j\in J} z_{ij}\Big)^2 + \sum_{j\in J} w_{jj}^{-1}\Big(\sum_{i\in I} z_{ij}\Big)^2}, \tag{14}$$

where w_{ii} and w_{jj} are positive real numbers and $b_{ij} = (w_{ij} + w_{ji})a_{ij} \in (0, \infty]$, $i \in I$, $j \in J$. If I or J is a singleton, then a solution to (14) can be given explicitly.

Lemma 3 *Suppose that* $I = \{1\}$ *and that* $w_{11} = w_{jj} = 1$ *for all* $j \in J$. *Then a solution of* (14) *is given by*

$$z_{1j} = (1 - b_{1j}^{-1})^{-1}\sum_{k\in J}(1 - b_{1k}^{-1})^{-1}, \qquad b_{1j} = 2a_{1j}$$

leading to the maximum

$$\frac{\sum_{j\in J}(1-b_{1j}^{-1})^{-1}}{1+\sum_{j\in J}(1-b_{1j}^{-1})^{-1}} \geq \frac{1}{4}\min\Big(\sum_{j\in J}\min(b_{1j},1),1\Big) = \frac{1}{4}\min\Big(\sum_{j\in J}b_{1j},1\Big). \tag{15}$$

Proof The inequality in (15) follows from the inequality $x/(1+x) \geq (1/2)\min(x,1)$. Obviously, the values for z_{1j} given in Lemma 3 lead to the expression (15). Hence, it remains to show that (14) is upper bounded by the left-hand side in (15), which can be seen by inserting the (Cauchy-Schwarz) inequality $(\sum_{j\in J}(1+b_{1j}^{-1})^{-1})^{-1}(\sum_{j\in J}z_{1j})^2 \leq \sum_{j\in J}(1+b_{1j}^{-1})z_{1j}^2$ into (14). □

In general, it seems more difficult to give an explicit formula for (14) using e.g. only b_{ij} and $\wedge$. Yet, the following lower bound is sufficient for our purposes. In the special case of Lemma 3, it gives the second bound in (15).

Lemma 4 *For each $\delta > 0$, the value defined through* (14) *is lower bounded by*

$$\begin{array}{llll} \textit{maximize} & \dfrac{1}{1+2\delta}\sum_{i\in I}\sum_{j\in J}x_{ij} & & \\ \textit{subject to} & 0 \leq x_{ij} \leq b_{ij} & \textit{for all} & i\in I, j\in J, \\ & \sum_{i\in I}x_{ij} \leq \delta w_{jj} & \textit{for all} & j\in J, \\ & \sum_{j\in J}x_{ij} \leq \delta w_{ii} & \textit{for all} & i\in I. \end{array} \tag{16}$$

Proof Let $z_{ij} = x_{ij}$ be real values satisfying the constraints in (16). Then we have

$$\sum_{i\in I}\sum_{j\in J}b_{ij}^{-1}z_{ij}^2 + \sum_{i\in I}w_{ii}^{-1}\Big(\sum_{j\in J}z_{ij}\Big)^2 + \sum_{j\in J}w_{jj}^{-1}\Big(\sum_{i\in I}z_{ij}\Big)^2 \leq (1+2\delta)\sum_{i\in I}\sum_{j\in J}z_{ij}.$$

Inserting this into (14), the claim follows □

Remark 13 If $b_{ij} = \infty$, then the first constraint in (16) can be written as $0 \leq x_{ij} < \infty$.

5.3 End of proofs of the consequences

Proof (Proof of Theorem 3) By [27, Lemma 1], Condition (12) is satisfied with

$$\mathcal{I}(\xi,\xi) = \frac{n}{2}\sum_{i,j=1}^{p}\xi_{ij}^2\frac{(\lambda_i-\lambda_j)^2}{\lambda_i\lambda_j}, \qquad \xi \in \mathfrak{so}(p).$$

Moreover, letting $O(p)$ act coordinate-wise on $\prod_{i=1}^n\mathbb{R}^p$ the statistical model in (1) satisfies (A1). Hence, applying Corollary 2 with $I = \{1,\dots,d\}$ and $J = \{d+$

$1, \dots, p\}$, $w_{kl} = w_{kl} = 1$ for all k, l (leading to the Hilbert-Schmidt distance, cf. (2)), the claim follows from Lemma 4, using that $b_{ij} = 2a_{ij} = 2\mathcal{I}(L^{(ij)}, L^{(ij)})^{-1} = (2/n)\lambda_i\lambda_j/(\lambda_i - \lambda_j)^2 \in (0, \infty]$. □

Proof (Proof of Theorem 4) The main remaining point is to show that the excess risk $\mathcal{E}_U(\hat{P})$, $\hat{P} \in \mathcal{P}_d$, is of the form $l_w(U, \hat{P})$ for some $w \in \mathbb{R}^{p\times p}_{\geq 0}$. This can be deduced from [22, Lemma 2.6].

Lemma 5 *For $\hat{P} \in \mathcal{P}_d$ and $\mu \in [\lambda_{d+1}, \lambda_d]$, we have*

$$\mathcal{E}_U(\hat{P}) = \sum_{k=1}^{p}\sum_{l=1}^{p} w_{kl}\langle u_k u_l^T, \hat{P} - P_{\leq d}(U)\rangle_2^2 = l_w(U, \hat{P})$$

with $w_{kl} = \lambda_k - \mu$ for $k \leq d$ and $w_{kl} = \mu - \lambda_k$ for $k > d$. □

Proof For brevity we write $P_{\leq d}(U) = P_{\leq d}$ and $P_k = P_k(U) = u_k u_k^T$. By [22, Lemma 2.6], we have

$$\mathcal{E}_U(\hat{P}) = \sum_{k\leq d}(\lambda_k - \mu)\|P_k(I - \hat{P})\|_2^2 + \sum_{k>d}(\mu - \lambda_k)\|P_k\hat{P}\|_2^2.$$

Inserting

$$\begin{aligned} \|P_k(I - \hat{P})\|_2^2 &= \|P_k(P_{\leq d} - \hat{P})\|_2^2, \qquad k \leq d, \\ \|P_k\hat{P}\|_2^2 = \|P_k(\hat{P} - P_{\leq d})\|_2^2 &= \|P_k(P_{\leq d} - \hat{P})\|_2^2, \qquad k > d, \end{aligned}$$

we obtain

$$\begin{aligned} \mathcal{E}_U(\hat{P}) &= \sum_{k\leq d}(\lambda_k - \mu)\|P_k(P_{\leq d} - \hat{P})\|_2^2 + \sum_{k>d}(\mu - \lambda_k)\|P_k(P_{\leq d} - \hat{P})\|_2^2 \\ &= \sum_{k\leq d}\sum_{l=1}^{p}(\lambda_k - \mu)\|P_k(P_{\leq d} - \hat{P})P_l\|_2^2 + \sum_{k>d}\sum_{l=1}^{p}(\mu - \lambda_k)\|P_k(P_{\leq d} - \hat{P})P_l\|_2^2, \end{aligned}$$

and the claim follows from inserting the identity $\|P_k B P_l\|_2^2 = \langle u_k u_l^T, B\rangle_2^2$, $B \in \mathbb{R}^{p\times p}$. □

Applying Lemma 5, we get

$$\inf_{\hat{P}\in\mathcal{P}_d}\int_{O(p)} \mathbb{E}_U\mathcal{E}_U(\hat{P})\,dU = \inf_{\hat{P}\in\mathcal{P}_d}\int_{O(p)} \mathbb{E}_U l_w(U, \hat{P})\,dU \geq \inf_{\hat{P}}\int_{O(p)} \mathbb{E}_U l_w(U, \hat{P})\,dU,$$

where the last infimum is over all estimators $\hat{P}$ with values in $\mathbb{R}^{p\times p}$. Hence, applying Corollary 2 with $w = (w_{kl})$ from Lemma 5, $I = \{1, \dots, r\}$, $J = \{s, \dots, p\}$ and

$$b_{ij} = (w_{ij} - w_{ji})a_{ij} = (\lambda_i - \mu - (\mu - \lambda_j))\mathcal{I}(L^{(ij)}, L^{(ij)})^{-1} = \frac{1}{n}\frac{\lambda_i\lambda_j}{\lambda_i - \lambda_j},$$

the claim follows from Lemma 4. □

Proof (Proof of Theorem 5) We let $O(p)$ act on $\mathbb{R}^{p\times p}$ by conjugation. Since $V(\xi_{ij})V^T \overset{d}{=} (\xi_{ij})$, we have $VXV^T \overset{d}{=} VUX(VU)^T + \epsilon(\xi_{ij})$, meaning that the statistical model in (11) satisfies (A1). Using the identity $\chi^2(\mathcal{N}(\mu_1, \epsilon I_p), \mathcal{N}(\mu_2, \epsilon I_p)) = \exp(\epsilon^{-2}\|\mu_1 - \mu_2\|_2^2) - 1$, we get

$$\chi^2(\mathbb{P}_{\exp(t\xi)}, \mathbb{P}_{I_p}) = \exp(\epsilon^{-2}\|\exp(t\xi)\Lambda\exp(-t\xi) - \Lambda\|_2^2) - 1.$$

From this, it easily follows that (12) is satisfied with

$$\mathcal{I}(\xi,\xi) = \epsilon^{-2}\|\xi\Lambda - \Lambda\xi\|_2^2 = \epsilon^{-2}\sum_{i=1}^{p}\sum_{j=1}^{p}\xi_{ij}^2(\lambda_i - \lambda_j)^2.$$

Hence, applying Corollary 2 with $w_{kl} = 1$, $I = \{1,\dots,d\}$ and $J = \{d+1,\dots,p\}$, the claim follows from Lemma 4. □

Acknowledgements This research has been supported by a Feodor Lynen Fellowship of the Alexander von Humboldt Foundation.

References

1. T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, Inc., New York, second edition, 1984.
2. R. Bhatia. *Matrix analysis*. Springer-Verlag, New York, 1997.
3. L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.
4. T. T. Cai, H. Li, and R. Ma. Optimal structured principal subspace estimation: Metric entropy and minimax rates. Available at https://arxiv.org/abs/2002.07624.
5. T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 2013.
6. T. T. Cai and A. Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.*, 46(1):60–89, 2018.
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(1):136–154, 1982.
8. M. L. Eaton. *Group invariance applications in statistics*, volume 1 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA, 1989.
9. M. L. Eaton. *Multivariate statistics: A vector space approach*. Institute of Mathematical Statistics, Beachwood, OH, 2007. Reprint of the 1983 original.
10. C. Gao, Z. Ma, Z. Ren, and H. H. Zhou. Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.*, 43(5):2168–2197, 2015.
11. R. D. Gill and Boris Y. Levit. Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.
12. J. Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I: Theory of statistics*, pages 175–194, 1972.

13. T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, Ltd., Chichester, 2015.
14. I. C. F. Ipsen. An overview of relative sin θ theorems for invariant subspaces of complex matrices. *J. Comput. Appl. Math.*, 123:131–153, 2000.
15. M. Jirak and M. Wahl. Relative perturbation bounds with applications to empirical covariance operators. Available at https://arxiv.org/pdf/1802.02869, 2018.
16. M. Jirak and M. Wahl. Perturbation bounds for eigenspaces under a relative gap condition. *Proc. Amer. Math. Soc.*, 148(2):479–494, 2020.
17. V. Koltchinskii and K. Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.*, 45(1):121–157, 2017.
18. Z. Ma and X. Li. Subspace perspective on canonical correlation analysis: dimension reduction and minimax rates. *Bernoulli*, 26(1):432–470, 2020.
19. P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.
20. A. Naumov, V. Spokoiny, and V. Ulyanov. Bootstrap confidence sets for spectral projectors of sample covariance. *Probab. Theory Related Fields*, 174(3-4):1091–1132, 2019.
21. A. Pajor. Metric entropy of the Grassmann manifold. In *Convex geometric analysis (Berkeley, CA, 1996)*, volume 34 of *Math. Sci. Res. Inst. Publ.*, pages 181–188. Cambridge Univ. Press, Cambridge, 1999.
22. M. Reiss and M. Wahl. Nonasymptotic upper bounds for the reconstruction error of PCA. *Ann. Statist.*, 48(2):1098–1123, 2020.
23. M. Reiß and M. Wahl. Supplement to "Non-asymptotic upper bounds for the reconstruction error of PCA". 2020.
24. A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009. Revised and extended from the 2004 French original.
25. R. Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.
26. V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.*, 41(6):2905–2947, 2013.
27. M. Wahl. Information inequalities for the estimation of principal components. Available at https://arxiv.org/abs/2005.06869.
28. M. Wahl. On the perturbation series for eigenvalues and eigenprojections. Available at https://arxiv.org/abs/1910.08460, 2019.
29. M. J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.
30. Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.
31. Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.