Electronic Journal of Statistics
Vol. 11 (2017) 3903–3926
ISSN: 1935-7524
DOI: 10.1214/17-EJS1343

# Robust PCA and pairs of projections in a Hilbert space

**Ilaria Giulini***

*Laboratoire de Probabilités et Modèles Aléatoires*
*Université Paris Diderot, 75013, Paris, France*
*e-mail:* giulini@math.univ-paris-diderot.fr

**Abstract:** This is a study of principal component analysis performed on a statistical sample. We assume that this data sample is made of independent copies of some random variable ranging in a separable real Hilbert space. This covers data in function spaces as well as data represented in reproducing kernel Hilbert spaces. Based on some new inequalities about the perturbation of nonnegative self-adjoint operators, we provide new bounds for the statistical fluctuations of the principal component representation with the draw of the statistical sample.

We suggest two kinds of improvements to decrease these fluctuations: the first is to use a robust estimate of the covariance operator, for which non-asymptotic bounds of the estimation error are available under weak polynomial moment assumptions. The second improvement is to use some modification of the projection on the principal components based on functional calculus applied to the covariance operator. Using this modified projection, we can obtain bounds that do not depend on the spectral gap but on some more favorable factor.

In appendix, we provide a new approach to the analysis of the relative positions of two orthogonal projections that is useful for our proofs and that has an interest of its own.

**MSC 2010 subject classifications:** 62G35, 62G05, 62H25.
**Keywords and phrases:** PAC-Bayesian learning, Principal Component Analysis, robust estimation, spectral projectors, perturbation of self-adjoint operators, principal angles.

Received May 2016.

## Contents



*The results presented in this paper were obtained while the author was preparing her PhD under the supervision of Olivier Catoni at the Département de Mathématiques et Applications, École Normale Supérieure, Paris, with the financial support of the Région Île de France.

## 1. Introduction

Principal Component Analysis (PCA) is a classical tool for dimensionality reduction that relies on the spectral properties of the covariance matrix. In this paper we consider a data set $(X_1, \dots, X_n)$ made of independent copies of a random variable $X$ taking its values in a real separable Hilbert space $\mathcal{H}$.

The basic idea of PCA is to reduce the dimensionality of $X$ by projecting it into a finite dimension linear subspace while keeping the variance as high as possible. This subspace, as is well known, is the linear span of the eigenvectors of the covariance operator associated with the largest eigenvalues and called the principal components of $X$.

Several results can be found in the literature concerning the non-asymptotic setting. These results rely on sharp non-asymptotic bounds for the approximation error of the covariance matrix (e.g. Rudelson [18], Tropp [22], Vershynin [23]).

PCA in a separable Hilbert space that we consider here includes the analysis of samples in a functional space (PCA for functional data, Ramsay and Silverman [17]) and of samples embedded in a reproducing kernel Hilbert space. The latter is for example the case of kernel PCA, that uses the kernel trick to embed the dataset in a reproducing kernel Hilbert space in order to get a representation with a simplified geometry. (e.g. Schölkopf, Smola, Müller [21], Zwald, Bousquet, Blanchard [25], Shawe-Taylor, Williams, Cristianini, Kandola [19], [20]).

We consider the covariance operator

$$\Sigma = \mathbb{E}\Big(\langle\ , X - \mathbb{E}(X)\rangle\ \big(X - \mathbb{E}(X)\big)\Big) \tag{1.1}$$

where $\mathbb{E}$ is the expectation with respect to the law of the random vector $X$, or the Gram operator

$$G = \mathbb{E}\big(\langle\ , X\rangle X\big),$$

whose principal eigenvectors provides the directions with maximum energy instead of maximum variance. Moreover, as we will show, the study of $\Sigma$ can be deduced from the study of $G$.

We assume that the law of $X$ is unknown, so that we cannot work directly with $\Sigma$ but we have to construct an estimator $\widehat{\Sigma}$ as a function of the sample $(X_1, \dots, X_n)$. Results concerning the estimation of the spectral projectors of the covariance operator by their empirical counterparts in a Hilbert space can be found in Koltchinskii, Lounici [14], [15]. The authors study the problem in the case of Gaussian centered random vectors, based on the bounds obtained in [13], and in the setting where both the sample size $n$ and the trace of the covariance operator are large.

A question that arises in standard PCA is how to determine the number of relevant components. A common choice is to maximise the gap between the lowest eigenvalue that is kept and the next one.

This type of choice is justified by the fact that the bounds available for the statistical deviations of the representation depend on the inverse of this spectral gap.

Our goal is to improve these bounds by improving on one hand the choice of the estimator $\widehat{\Sigma}$ of the covariance matrix and on the other hand the choice of the representation itself, to make principal component analysis more robust to statistical fluctuations depending on the draw of the sample $(X_1, \ldots, X_n)$.

So the kind of robustness we are after is not the same as in Candès, Li, Ma, Wright in [3] where they show that it is possible to recover the principal components of a data matrix in the case where the observations are contained in a low-dimensional space but arbitrarily corrupted by additive noise.

Our approach provides two kinds of robustness. The first idea is to replace the projection on the principal eigenvectors by an alternative using functional calculus on the covariance operator. The fluctuations of this modified projection from sample to sample can be bounded depending on a quantity that is more favorable than the spectral gap.

While this improvement is of no help if we are precisely interested in performing a projection on the span of a given number of eigenvectors, in many situations, PCA is used more loosely to shrink the dimension of the data space while keeping as much of the variance as possible. One example of such a case is $k$-means unsupervised clustering for high-dimensional data. The usual recommendation is to avoid using directly the $k$-means algorithm in a high-dimensional space, but rather to perform a PCA reduction first. In such a context, except in some restrictive models, there is no reason why there should always be a large gap between meaningful and meaningless eigenvalues. Nonetheless, we still need a stable dimension reduction method, because it is still desirable to minimise the fluctuations of the cluster boundaries when the statistical sample used to compute them is replaced by another one. To start with, we cannot hope to compute stable clusters if we base the clustering on a change of representation that is sample dependent. This is where our modified projection may help: it will remain weakly dependent on the statistical sample choice (the precise meaning of this statement being provided by a non-asymptotic bound), even when no large spectral gap is available.

The second kind of robustness consists in using an estimator of the covariance operator from a statistical sample that has sub-exponential non asymptotic deviation bounds under polynomial moment assumptions, as explained in section 3 on page 3912.

The paper is divided into two parts. One is devoted to the theory of perturbations of nonnegative self-adjoint operators and the other one to the statistical analysis of PCA. In appendix, we propose a new treatment to the analysis of the relative positions of two projections, that we need for the proofs, and that has also some interest of its own.

## 2. A contribution to the perturbation theory of nonnegative self-adjoint operators

**Proposition 2.1.** *Let $A, B : \mathcal{H} \to \mathcal{H}$ be two compact self-adjoint nonnegative operators on the separable real Hilbert space $\mathcal{H}$. According to the spectral*

*representation theorem, we can write*

$$A = \sum_{i=1}^{\infty} \lambda_i \langle\ , p_i\rangle p_i,$$
$$B = \sum_{i=1}^{\infty} \mu_i \langle\ , q_i\rangle q_i,$$

*where* $\{p_i, 1 \le i < \infty\}$ *(resp.* $\{q_i, 1 \le i < \infty\}$*) form an orthonormal basis of eigenvectors of* $A$ *(resp.* $B$*) and where* $\lambda_1 \ge \lambda_2 \ge \cdots$ *are the eigenvalues of* $A$ *(resp.* $\mu_1 \ge \mu_2 \ge \cdots$ *are the eigenvalues of* $B$*), sorted in decreasing order and satisfying*

$$\lim_{i\to\infty} \lambda_i = 0, \qquad (\textit{resp.} \qquad \lim_{i\to\infty} \mu_i = 0).$$

*Let us consider the orthogonal projectors on the span of the* $r$ *first eigenvectors of each operator, defined as*

$$\Pi_r(A) = \sum_{i=1}^{r} \langle\ , p_i\rangle p_i,$$
$$\Pi_r(B) = \sum_{i=1}^{r} \langle\ , q_i\rangle q_i.$$

*Define the spectral gaps*

$$\begin{aligned} \gamma_r(A,B) &= \max\bigl\{\lambda_r - \mu_{r+1}, \mu_r - \lambda_{r+1}\bigr\} \\ &\ge \frac{1}{2}\Bigl(\lambda_r - \lambda_{r+1} + \mu_r - \mu_{r+1}\Bigr) \\ &\ge \frac{1}{2}\max\bigl\{\lambda_r - \lambda_{r+1}, \mu_r - \mu_{r+1}\bigr\} \ge 0 \\ \textit{and} \qquad \widetilde{\gamma}_r(A,B) &= \max\bigl\{0, \min\{\lambda_r - \mu_{r+1}, \mu_r - \lambda_{r+1}\}\bigr\}. \end{aligned}$$

*The differences* $\Pi_r(A) - \Pi_r(B)$ *and* $A - B$ *are related by the relations*

$$\bigl\|\Pi_r(A) - \Pi_r(B)\bigr\|_\infty \le \frac{\sqrt{r}}{\gamma_r(A,B)}\|A - B\|_\infty,$$
$$\bigl\|\Pi_r(A) - \Pi_r(B)\bigr\|_{\mathtt{HS}} \le \frac{\sqrt{2r}}{\gamma_r(A,B)}\|A - B\|_\infty,$$
$$\bigl\|\Pi_r(A) - \Pi_r(B)\bigr\|_{\mathtt{HS}} \le \frac{\sqrt{2}\|A - B\|_{\mathtt{HS}}}{\gamma_r(A,B)},$$
$$\bigl\|\Pi_r(A) - \Pi_r(B)\bigr\|_{\mathtt{HS}} \le \frac{\|A - B\|_{\mathtt{HS}}}{\widetilde{\gamma}_r(A,B)},$$

*where* $\|\ \|_\infty$ *is the operator norm and* $\|\ \|_{\mathtt{HS}}$ *is the Hilbert-Schmidt norm.*

*Proof.* Recall that

$$\|A\|_{\mathtt{HS}}^2 = \mathbf{Tr}(A^*A) = \sum_{i=1}^{\infty}\|Ae_i\|^2 = \sum_{i=1}^{\infty}\lambda_i^2$$

for any orthonormal basis $\{e_i, 1 \le i < \infty\}$ of $\mathcal{H}$. Our proofs are essentially based on the identity

$$\begin{aligned} A - B &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(\lambda_i - \mu_j)\langle p_i, q_j\rangle\langle\ , p_i\rangle q_j, \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(\lambda_i - \mu_j)\langle p_i, q_j\rangle\langle\ , q_j\rangle p_i, \end{aligned} \tag{2.1}$$

and its consequence

$$\|A - B\|_{\mathtt{HS}}^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(\lambda_i - \mu_j)^2\langle p_i, q_j\rangle^2, \tag{2.2}$$

as well as on the relative positions of two projections, as described in the appendix. Eq. (2.1) is implied by the identities

$$p_i = \sum_{j=1}^{\infty}\langle p_i, q_j\rangle q_j \qquad \text{and} \qquad q_j = \sum_{i=1}^{\infty}\langle p_i, q_j\rangle p_i.$$

According to Lemma A.4 on page 3921,

$$\big\|\Pi_r(A) - \Pi_r(B)\big\|_{\infty}^2 = \sup_{u\in\mathcal{S}\cap\mathbf{Im}\,\Pi_r(B)}\big\|\big[\Pi_r(A) - \Pi_r(B)\big]u\big\|^2,$$

where $\mathcal{S}$ denotes the unit sphere of $\mathcal{H}$. Moreover, for any $u \in \mathcal{S}\cap\mathbf{Im}\,\Pi_r(B)$,

$$\big\|\big[\Pi_r(A) - \Pi_r(B)\big]u\big\|^2 = \big\|\Pi_r(A) - u\big\|^2 = \Big\|\sum_{i=r+1}^{\infty}\langle u, p_i\rangle p_i\Big\|^2 = \sum_{i=r+1}^{\infty}\langle u, p_i\rangle^2.$$

On the other hand, $u = \sum_{j=1}^{r}\langle u, q_j\rangle q_j$ and $\sum_{j=1}^{r}\langle u, q_j\rangle^2 = \|u\|^2 = 1$.
Using the Cauchy-Schwarz inequality and assuming without loss of generality that $\mu_r - \lambda_{r+1} \ge 0$, we get

$$\begin{aligned} \big\|\big[\Pi_r(A) - \Pi_r(B)\big]u\big\|^2 &= \sum_{i=r+1}^{\infty}\Big(\sum_{j=1}^{r}\langle u, q_j\rangle\langle q_j, p_i\rangle\Big)^2 \le \sum_{i=r+1}^{\infty}\sum_{j=1}^{r}\langle q_j, p_i\rangle^2 \\ &\le \sum_{i=r+1}^{\infty}\sum_{j=1}^{r}\frac{(\mu_j - \lambda_i)^2}{(\mu_r - \lambda_{r+1})^2}\langle p_i, q_j\rangle^2 \le \sum_{j=1}^{r}\sum_{i=1}^{\infty}\frac{(\lambda_i - \mu_j)^2}{(\mu_r - \lambda_{r+1})^2}\langle p_i, q_j\rangle^2 \\ &= \sum_{j=1}^{r}\frac{\|(A-B)q_j\|^2}{(\mu_r - \lambda_{r+1})^2} \le \frac{r}{(\mu_r - \lambda_{r+1})^2}\|A - B\|_{\infty}^2. \end{aligned} \tag{2.3}$$

Exchanging the roles of $A$ and $B$, this proves also that when $\lambda_r - \mu_{r+1} \geq 0$,

$$\big\|\Pi_r(A) - \Pi_r(B)\big\|_\infty^2 \leq \sum_{i=1}^{r} \frac{\|(A-B)p_i\|^2}{(\lambda_r - \mu_{r+1})^2} \leq \frac{r}{(\lambda_r - \mu_{r+1})^2}\|A-B\|_\infty^2,$$

so that

$$\|\Pi_r(A) - \Pi_r(B)\|_\infty \leq \frac{\sqrt{r}}{\gamma_r(A,B)}\|A-B\|_\infty.$$

According to Lemma A.4 again, assuming without loss of generality that $\mu_r - \lambda_{r+1} \geq 0$, and using eq. (2.3), we see that

$$\begin{aligned}\|\Pi_r(A) - \Pi_r(B)\|_{\mathtt{HS}}^2 &= 2\sum_{j=1}^{r}\big\|\big(\Pi_r(A) - \Pi_r(B)\big)q_j\big\|^2 \\ &= 2\sum_{j=1}^{r}\sum_{i=r+1}^{\infty}\langle p_i, q_j\rangle^2 \leq 2\sum_{j=1}^{r}\frac{\|(A-B)q_j\|^2}{(\mu_r - \lambda_{r+1})^2},\end{aligned}$$

so that

$$\begin{aligned}\big\|\Pi_r(A) - \Pi_r(B)\big\|_{\mathtt{HS}}^2 &\leq \frac{2r}{(\mu_r - \lambda_{r+1})^2}\|A-B\|_\infty^2 \\ \text{and} \qquad \big\|\Pi_r(A) - \Pi_r(B)\big\|_{\mathtt{HS}}^2 &\leq \frac{2}{(\mu_r - \lambda_{r+1})^2}\|A-B\|_{\mathtt{HS}}^2.\end{aligned}$$

Exchanging $A$ and $B$, this proves also that

$$\begin{aligned}\big\|\Pi_r(A) - \Pi_r(B)\big\|_{\mathtt{HS}}^2 &\leq \frac{2r}{(\lambda_r - \mu_{r+1})^2}\|A-B\|_\infty^2 \\ \text{and} \qquad \big\|\Pi_r(A) - \Pi_r(B)\big\|_{\mathtt{HS}}^2 &\leq \frac{2}{(\lambda_r - \mu_{r+1})^2}\|A-B\|_{\mathtt{HS}}^2,\end{aligned}$$

and therefore that

$$\big\|\Pi_r(A) - \Pi_r(B)\big\|_{\mathtt{HS}} \leq \frac{\sqrt{2}\min\big\{\sqrt{r}\|A-B\|_\infty, \|A-B\|_{\mathtt{HS}}\big\}}{\gamma_r(A,B)}.$$

Let us now assume without loss of generality that $\lambda_r - \mu_{r+1} > 0$ and $\mu_r - \lambda_{r+1} > 0$ (because otherwise $\widetilde{\gamma}_r(A,B) = 0$ and there is nothing to prove). Applying eq. (2.2) to $\Pi_r(A)$ and $\Pi_r(B)$ shows that

$$\begin{aligned}\big\|\Pi_r(A) - \Pi_r(B)\big\|_{\mathtt{HS}}^2 &= \left(\sum_{i=1}^{r}\sum_{j=r+1}^{\infty} + \sum_{i=r+1}^{\infty}\sum_{j=1}^{r}\right)\langle p_i, q_j\rangle^2 \\ &\leq \frac{1}{\min\{(\lambda_r - \mu_{r+1})_+, (\mu_r - \lambda_{r+1})_+\}^2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(\lambda_i - \mu_j)^2\langle p_i, q_j\rangle^2 \\ &= \frac{\|A-B\|_{\mathtt{HS}}^2}{\widetilde{\gamma}_r(A,B)}.\end{aligned}$$

Except maybe for the precise definition of the spectral gap, the last two results are well-known, but usually proved in a different way, applying the Cauchy integral formula to the resolvent. One advantage of our proof is to lead to a definition of the spectral gap $\gamma_r(A,B) \geq \big(\lambda_r - \lambda_{r+1}\big)/2$ that is bounded from below independently of the value of $B$, so that we do not need to assume that the perturbation $B-A$ is small in any sense and still get a meaningful bound. □

**Proposition 2.2.** *In the same setting as in Proposition 2.1 on page 3905, let us consider some arbitrary $L$-Lipschitz function $f : \mathbb{R}_+ \to \mathbb{R}$ and define*

$$f(A) = \sum_{i=1}^{\infty} f(\lambda_i)\langle\ ,p_i\rangle p_i.$$

*Define similarly $f(B)$. The transformed operators $f(A)$ and $f(B)$ are such that*

$$\begin{aligned}
\|f(A)-f(B)\|_{\mathtt{HS}} &\leq L\|A-B\|_{\mathtt{HS}} \\
&\leq L \inf_{r\in\mathbb{N}} \sqrt{2r\|A-B\|_\infty^2 + \mathbf{Tr}_{r+1}(A^2) + \mathbf{Tr}_{r+1}(B^2)} \\
&\leq 2^{3/4} L \|A-B\|_\infty^{1/2} \Big(\mathbf{Tr}(A)^2 + \mathbf{Tr}(B)^2\Big)^{1/4}, \\
\textit{and } \|f(A)-f(B)\|_\infty &\leq L\left(\|A-B\|_\infty + \inf_{r\in\mathbb{N}} \sqrt{8r\|A-B\|_\infty^2 + 2\,\mathbf{Tr}_{r+1}(A^2)}\right), \\
&\leq L\left(\|A-B\|_\infty + 2\sqrt{2}\|A-B\|_\infty^{1/2}\,\mathbf{Tr}(A)^{1/2}\right),
\end{aligned}$$

*where* $\mathbf{Tr}_{r+1}(A^2) = \sum_{i=r+1}^{\infty} \lambda_i^2$.

*Remark* 2.1. For the second bound to be finite, $A$ should be Hilbert-Schmidt, but not necessarily $B$. On the other hand, as may be expected, the first bound is finite only when both $A$ and $B$ are Hilbert-Schmidt operators. Note that the question of extending inequalities for projections to other functions of $A$ and $B$ was already identified as important in [6, Open question 10.4, page 44]. We will explain in the next section how to use functional calculus to replace the usual principal component representation by some alternative for which more favorable statistical deviation bounds can be proved.

*Proof.* First remark that

$$\begin{aligned}
\big\|f(A)-f(B)\big\|_{\mathtt{HS}}^2 &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \big(f(\lambda_i)-f(\mu_j)\big)^2 \langle p_i, q_j\rangle^2 \\
&\leq L^2 \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \big(\lambda_i-\mu_j\big)^2 \langle p_i, q_j\rangle^2 = L^2\|A-B\|_{\mathtt{HS}}^2.
\end{aligned}$$

Moreover, for any positive integer $r$,

$$\begin{aligned}\|A-B\|_{\mathtt{HS}}^2 &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\big(\lambda_i-\mu_j\big)^2\langle p_i,q_j\rangle^2\\ &\le \left(\sum_{i=1}^{r}\sum_{j=1}^{\infty}+\sum_{i=1}^{\infty}\sum_{j=1}^{r}+\sum_{i=r+1}^{\infty}\sum_{j=r+1}^{\infty}\right)\big(\lambda_i-\mu_j\big)^2\langle p_i,q_j\rangle^2\\ &= \sum_{i=1}^{r}\|(A-B)p_i\|^2+\sum_{j=1}^{r}\|(A-B)q_j\|^2+\sum_{i=r+1}^{\infty}\sum_{j=r+1}^{\infty}\big(\lambda_i-\mu_j\big)^2\langle p_i,q_j\rangle^2\\ &\le 2r\|A-B\|_\infty^2+\sum_{i=r+1}^{\infty}\sum_{j=r+1}^{\infty}\big(\lambda_i^2+\mu_j^2\big)\langle p_i,q_j\rangle^2\\ &\le 2r\|A-B\|_\infty^2+\sum_{i=r+1}^{\infty}\lambda_i^2+\sum_{j=r+1}^{\infty}\mu_j^2,\end{aligned}$$

proving that

$$\|A-B\|_{\mathtt{HS}}\le\inf_{r\in\mathbb{N}}\sqrt{2r\|A-B\|_\infty^2+\mathbf{Tr}_{r+1}(A^2)+\mathbf{Tr}_{r+1}(B^2)}.$$

Let us now remark that

$$\mathbf{Tr}_{r+1}(A^2)\le\lambda_{r+1}\,\mathbf{Tr}(A)\le(r+1)^{-1}\,\mathbf{Tr}(A)^2,$$

so that

$$\begin{aligned}&\inf_{r\in\mathbb{N}}\sqrt{2r\|A-B\|_\infty^2+\mathbf{Tr}_{r+1}(A^2)+\mathbf{Tr}_{r+1}(B^2)}\\ &\qquad\le\inf_{r\in\mathbb{N}}\sqrt{2r\|A-B\|_\infty^2+(r+1)^{-1}\big(\mathbf{Tr}(A)^2+\mathbf{Tr}(B)^2\big)}.\end{aligned}$$

Consider $r_*=\sqrt{\dfrac{\mathbf{Tr}(A)^2+\mathbf{Tr}(B)^2}{2\|A-B\|_\infty^2}}$ and choose $r$ such that $r\le r_*\le r+1$. Remark that for this choice of $r$

$$\begin{aligned}&\sqrt{2r\|A-B\|_\infty^2+(r+1)^{-1}\big(\mathbf{Tr}(A)^2+\mathbf{Tr}(B)^2\big)}\\ &\qquad\le\sqrt{2r_*\|A-B\|_\infty^2+r_*^{-1}\big(\mathbf{Tr}(A)^2+\mathbf{Tr}(B)^2\big)}\\ &\qquad=\sqrt{2\|A-B\|_\infty\sqrt{2\big(\mathbf{Tr}(A)^2+\mathbf{Tr}(B)^2\big)}},\end{aligned}$$

so that

$$\|A-B\|_{\mathtt{HS}}\le 2^{3/4}\|A-B\|_\infty^{1/2}\big(\mathbf{Tr}(A)+\mathbf{Tr}(B)\big)^{1/4}.$$

Introduce $C=\sum_{i=1}^{\infty}\lambda_i\langle\ ,q_i\rangle q_i$ the operator obtained by replacing $\mu_i$ by $\lambda_i$ in the definition of $B$, and remark that

$$\|f(A)-f(B)\|_\infty\le\|f(A)-f(C)\|_\infty+\|f(C)-f(B)\|_\infty.$$

On the one hand,

$$\|f(C)-f(B)\|_\infty^2 = \sup_{u\in\mathcal{S}}\sum_{i=1}^{\infty}\big(f(\lambda_i)-f(\mu_i)\big)^2\langle u,q_i\rangle^2$$
$$= \sup_i\big(f(\lambda_i)-f(\mu_i)\big)^2 \le L^2\sup_i\big(\lambda_i-\mu_i\big)^2 \le L^2\|A-B\|_\infty^2,$$

where the last inequality will be proved later, in Proposition 2.3. On the other hand,

$$\|f(A)-f(C)\|_\infty \le \|f(A)-f(C)\|_{\mathtt{HS}}$$
$$\le L\inf_{r\in\mathbb{N}}\sqrt{2r\|A-C\|_\infty^2+\mathbf{Tr}_{r+1}(A)+\mathbf{Tr}_{r+1}(C)}$$
$$\le 2^{3/4}L\|A-C\|_\infty^{1/2}\big(\mathbf{Tr}(A)^2+\mathbf{Tr}(B)^2\big)^{1/4}$$

We can then remark that

$$\|A-C\|_\infty \le \|A-B\|_\infty+\|B-C\|_\infty \le 2\|A-B\|_\infty,$$

and that $\mathbf{Tr}_{r+1}(C)=\mathbf{Tr}_{r+1}(A)$, to conclude that

$$\|f(A)-f(C)\|_\infty \le L\inf_{r\in\mathbb{N}}\sqrt{8r\|A-B\|_\infty^2+2\,\mathbf{Tr}_{r+1}(A^2)}$$
$$\le 2^{3/2}L\|A-B\|_\infty^{1/2}\,\mathbf{Tr}(A)^{1/2}.$$

□

**Proposition 2.3.** *Let $\mathcal{S}$ be the unit sphere of $\mathcal{H}$. In the same setting as in Proposition 2.1 on page 3905, assume that for some non-decreasing function $g:\mathbb{R}_+\to\mathbb{R}_+$, for any $u\in\mathcal{S}$,*

$$\big|\langle(A-B)u,u\rangle\big| \le g\big(\langle Au,u\rangle\big).$$

*Assume moreover that $t\mapsto t-g(t)$ is also non-decreasing on $\mathbb{R}_+$. Then for any positive integer $i$,*

$$|\lambda_i-\mu_i| \le g(\lambda_i).$$

*In particular taking for $g$ the constant function equal to $\|A-B\|_\infty$, we obtain that*

$$\sup_{i\ge1}|\lambda_i-\mu_i| \le \|A-B\|_\infty.$$

*Proof.* Let us remark that

$$\lambda_i = \inf_e\sup\big\{\langle Au,u\rangle,\ u\in\mathbf{span}\{e_1,\dots,e_{i-1}\}^\perp\cap\mathcal{S}\big\},$$

where the infimum is taken on all families of $i-1$ independent vectors $e_1,\dots,e_{i-1}$. This is similar to inequality (1.10) of [6] and is also related to eq. (6.73) on page 60 of [11] (or page 62 of [12]). Consequently,

$$\lambda_i-g(\lambda_i) = \inf_e\sup\big\{\langle Au,u\rangle-g\big(\langle Au,u\rangle\big),\ u\in\mathbf{span}\{e_1,\dots,e_{i-1}\}^\perp\cap\mathcal{S}\big\}$$
$$\le \sup\big\{\langle Bu,u\rangle,\ u\in\mathbf{span}\{q_1,\dots,q_{i-1}\}^\perp\cap\mathcal{S}\big\}=\mu_i.$$

On the other hand,

$$\begin{aligned}\mu_i &= \inf_e \sup\big\{\langle Bu,u\rangle,\ u\in \mathbf{span}\{e_1,\dots,e_{i-1}\}^\perp\cap\mathcal{S}\big\}\\ &\le \sup\big\{\langle Au,u\rangle + g\big(\langle Au,u\rangle\big),\ u\in \mathbf{span}\{p_1,\dots,p_{i-1}\}^\perp\cap\mathcal{S}\big\}\\ &= \lambda_i + g(\lambda_i).\end{aligned}$$

□

## 3. Statistical study of principal component analysis in a separable Hilbert space

In this section, we will apply the previous bounds on the perturbation of a self-adjoint operator to the estimation of the principal components of a random vector $X$ taking its values in a real separable Hilbert space $\mathcal{H}$.

In [8, Proposition 3.4] we constructed an estimator $\widehat{G}$ of the Gram operator $G=\mathbb{E}\big(\langle\ ,X\rangle X\big)$ satisfying the following properties (where we have made the choice of $\sigma_n$ explained just after [8, eq. (2.13)]).

**Proposition 3.1.** *Assume that*

$$\mathbf{Tr}(G)=\mathbb{E}\big(\|X\|^2\big)\le T<\infty$$

*and that*

$$\sup\left\{\frac{\mathbb{E}\big(\langle u,X\rangle^4\big)}{\mathbb{E}\big(\langle u,X\rangle^2\big)^2},\ u\in\mathcal{H},\mathbb{E}\big(\langle u,X\rangle^2\big)>0\right\}\le\kappa<\infty,$$

*where $\kappa$ and $T$ are known constants. Remark that as a consequence*

$$\mathbb{E}\big(\|X\|^4\big)\le\kappa T^2<\infty.$$

*Assume without loss of generality that $\kappa\ge 3/2$. Let $\delta,\epsilon>0$ and* $\sigma_n=\dfrac{100\kappa T}{n/128-4.35-\log(\delta^{-1})}$ *and define*

$$\begin{aligned}\gamma_n(t)&=\sqrt{\frac{2.4(\kappa-1)}{n}\left(\frac{0.73T}{t}+4.35+\log(\delta^{-1})\right)}+\sqrt{\frac{99\kappa T}{nt}},\\ \eta_n(t)&=\frac{2\gamma_n(t)}{1-4\gamma_n(t)},\\ g_n(t)&=\max\{t,\sigma_n\}\eta\big(\min\big\{\max\{t,\sigma_n\},\kappa^{1/2}T\big\}\big)+\sigma_n+\varepsilon\|G\|_{\mathtt{HS}}.\end{aligned}$$

*Remark that for any fixed positive value of $t$, $g_n(t)$ is of order $n^{-1/2}+\varepsilon\|G\|_{\mathtt{HS}}$, and that more precisely,*

$$g_n(t)\le\mathcal{O}\left(\sqrt{\frac{\kappa t}{n}\Big(T+t\log(\delta^{-1})\Big)}+\varepsilon\|G\|_{\mathtt{HS}}\right),$$

$$t\ge\mathcal{O}\left(\frac{\kappa T}{n}\right),\quad n\ge\mathcal{O}\Big(\log(\delta^{-1})\Big),\quad(3.1)$$

*where $\mathcal{O}$ represents numerical constants in non-asymptotic bounds. For any value of $\varepsilon > 0$, we can compute an estimator $\widehat{G}_n$ of $G$, based on a sample $(X_1, \dots, X_n)$ made of $n$ independent copies of $X$, whose computation cost depends on $\varepsilon$, such that, with probability at least $1 - 2\delta$, for any $u \in \mathcal{S}$,*

$$\begin{aligned} &\big|\langle u, Gu\rangle - \langle u, \widehat{G}_n u\rangle\big| \le g_n\big(\langle u, Gu\rangle\big) \\ \text{and} \qquad &\big|\langle u, Gu\rangle - \langle u, \widehat{G}_n u\rangle\big| \le g_n\big(\langle u, \widehat{G}_n u\rangle\big). \end{aligned}$$

*Moreover, the functions $g_n$ and $t \mapsto t - g_n(t)$ are non-decreasing on $\mathbb{R}_+$. As a consequence, on the same event of probability at least $1 - 2\delta$ as above,*

$$\|G - \widehat{G}_n\|_\infty \le \min\Big\{g_n\big(\|G\|_\infty\big), g_n\big(\|\widehat{G}_n\|_\infty\big)\Big\}.$$

*Moreover, if $\lambda_1 \ge \lambda_2 \ge \dots$ are the eigenvalues of $G$ and $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \dots$ are the eigenvalues of $\widehat{G}_n$, counted with their multiplicities, on the event of probability at least $1 - 2\delta$ mentioned above,*

$$|\lambda_i - \widehat{\lambda}_i| \le \min\big\{g_n(\lambda_i), g_n(\widehat{\lambda}_i)\big\}, \quad 1 \le i < \infty.$$

The fact that $g_n$ is non-decreasing is proved in [8, Lemma 7.7]. The fact that $t \mapsto t - g_n(t)$ is non-decreasing is a straightforward consequence of the fact that $\gamma_n$ is non-increasing.

This proposition, along with the results of the previous section concerning the perturbation of self-adjoint operators, can be used to estimate the fluctuations of $\Pi_r(\widehat{G}_n)$, the estimated principal component representation of $X$. We give in the following corollary both theoretical and empirical bounds for these fluctuations. We give also the generalization of these bounds to the estimation of Lipschitz functionals $f(G)$.

**Corollary 3.1.** *With probability at least $1 - 2\delta$, for any $r \in \mathbb{N}$,*

$$\begin{aligned} \|\Pi_r(\widehat{G}_n) - \Pi_r(G)\|_\infty &\le \frac{2\sqrt{r}}{\lambda_r - \lambda_{r+1}} g_n\big(\|G\|_\infty\big) \\ \|\Pi_r(\widehat{G}_n) - \Pi_r(G)\|_\infty &\le \frac{2\sqrt{r}}{\widehat{\lambda}_r - \widehat{\lambda}_{r+1}} g_n\big(\|\widehat{G}_n\|_\infty\big), \end{aligned}$$

*and for any $L$-Lipschitz function $f : \mathbb{R}_+ \to \mathbb{R}_+$,*

$$\begin{aligned} \|f(\widehat{G}_n) - f(G)\|_\infty &\le L\bigg(g_n(\|G\|_\infty) + \inf_{r \in \mathbb{N}} \sqrt{8r g_n(\|G\|_\infty)^2 + 2\,\mathbf{Tr}_{r+1}(G^2)}\bigg) \\ &\le L\bigg(g_n(\|G\|_\infty) + 2\sqrt{2} g_n(\|G\|_\infty)^{1/2}\,\mathbf{Tr}(G)^{1/2}\bigg), \\ \text{and } \|f(\widehat{G}_n) - f(G)\|_\infty &\le L\bigg(g_n(\|\widehat{G}_n\|_\infty) \\ &\qquad + \inf_{r \in \mathbb{N}} \sqrt{8r g_n(\|\widehat{G}_n\|_\infty)^2 + 2\,\mathbf{Tr}_{r+1}(\widehat{G}_n^2)}\ \bigg). \end{aligned}$$

Let us give an example of the use of a Lipschitz functional $f(G)$ instead of a projection $\Pi_r(G)$. For any positive real parameters $a > b \geq 0$, let us define

$$f_{a,b}(t) = \max\Big\{0, \min\Big\{1, \frac{t-b}{a-b}\Big\}\Big\}.$$

This is a Lipschitz function with Lipschitz constant $(a-b)^{-1}$. There is a unique pair of indices $r \leq s \in \mathbb{N}$ such that $\lambda_r \geq a > \lambda_{r+1}$ and $\lambda_s > b \geq \lambda_{s+1}$. We can express $f_{a,b}(G)$ as

$$f_{a,b}(G) = \underbrace{\sum_{i=1}^{r} \langle\ , p_i\rangle p_i}_{=\Pi_r(G)} + \sum_{i=r+1}^{s} \frac{\lambda_i - b}{a-b} \langle\ , p_i\rangle p_i$$

where $p_i$ is an orthonormal basis of eigenvectors of $G$ and where

$$0 < \frac{\lambda_i - b}{a-b} < 1, \qquad r < i \leq s.$$

We see therefore that $f_{a,b}(G)$ lies between $\Pi_r(G)$ and $\Pi_s(G)$, in the sense that

$$\langle u, \Pi_r(G) u\rangle \leq \langle u, f(G) u\rangle \leq \langle u, \Pi_s(G) u\rangle, \qquad u \in \mathcal{H}.$$

Consequently, the energy kept by the representation $f(G)X$ also lies in between:

$$\mathbb{E}\Big(\big\|\Pi_r(G)X\big\|^2\Big) \leq \mathbb{E}\Big(\big\|f(G)X\big\|^2\Big) \leq \mathbb{E}\Big(\big\|\Pi_s(G)X\big\|^2\Big).$$

In the same way, there is a unique pair of indices $\widehat{r} \leq \widehat{s}$ such that $\widehat{\lambda}_{\widehat{r}} \geq a > \widehat{\lambda}_{\widehat{r}+1}$ and $\widehat{\lambda}_{\widehat{s}} > b \geq \widehat{\lambda}_{\widehat{s}+1}$. The estimate $f_{a,b}(\widehat{G}_n)$ of $f_{a,b}(G)$ is also a kind of interpolation between the two projections $\Pi_{\widehat{r}}(\widehat{G}_n)$ and $\Pi_{\widehat{s}}(\widehat{G}_n)$, since it can be written as

$$f_{a,b}(\widehat{G}_n) = \sum_{i=1}^{\widehat{r}} \langle\ , q_i\rangle q_i + \sum_{i=\widehat{r}+1}^{\widehat{s}} \frac{\widehat{\lambda}_i - b}{a-b} \langle\ , q_i\rangle q_i,$$

where $q_i$ is an orthonormal basis of eigenvectors of $\widehat{G}_n$.

The benefit of using $f_{a,b}(\widehat{G}_n)$ instead of $\Pi_{\widehat{s}}(G)$ to map the data into a space of dimension $\widehat{s}$, is that the fluctuations of this representation now depend on the larger multi-step gap $a - b \simeq \widehat{\lambda}_{\widehat{r}} - \widehat{\lambda}_{\widehat{s}}$, rather than on the single-step gap $\widehat{\lambda}_{\widehat{r}} - \widehat{\lambda}_{\widehat{r}+1}$.

Another option is to use a data dependent function $\widehat{f} = f_{\widehat{\lambda}_r, \widehat{\lambda}_s}$, for some pair of indices $r < s$. Our deviation bounds being uniform on the choice of $f$, they allow for a data dependent $f$, so that we get for instance with probability at least $1 - 2\delta$ that

$$\big\|\widehat{f}(\widehat{G}_n) - \widehat{f}(G)\big\|_\infty \leq \frac{1}{\widehat{\lambda}_r - \widehat{\lambda}_s} \Bigg( g_n\big(\|\widehat{G}_n\|_\infty\big) + \inf_{t \in \mathbb{N}} \sqrt{8 t g_n\big(\|\widehat{G}_n\|_\infty\big)^2 + 2\,\mathbf{Tr}_{t+1}\big(\widehat{G}_n^2\big)} \Bigg).$$

**Proposition 3.2** (Energy estimates)**.** *We will show here that it is possible to get an empirical estimate of the energy contained in the estimated representation.*

*In the previous setting, with probability at least* $1-2\delta$*, for any* $r \in \mathbb{N}$*,*

$$0 \le \mathbb{E}\Big(\big\|\Pi_r(G)X\big\|^2\Big) - \mathbb{E}\Big(\big\|\Pi_r(\widehat{G}_n)X\big\|^2 \,|\, (X_1,\dots,X_n)\Big) \le 2\min\left\{\sum_{i=1}^r g_n(\widehat{\lambda}_i),\ \sum_{i=1}^r g_n(\lambda_i)\right\} \quad (3.2)$$

*and*

$$\left|\mathbb{E}\Big(\big\|\Pi_r(\widehat{G}_n)X\big\|^2 \,|\, (X_1,\dots,X_n)\Big) - \sum_{i=1}^r \widehat{\lambda}_i\right| \le \sum_{i=1}^r g_n(\widehat{\lambda}_i) \le \sum_{i=1}^r g_n\big(\lambda_i + g_n(\lambda_i)\big). \quad (3.3)$$

*Moreover, for any non-decreasing measurable function* $f : \mathbb{R}_+ \to \mathbb{R}_+$*,*

$$\left|\mathbb{E}\Big(\big\|f(\widehat{G}_n)X\big\|^2 \,|\, (X_1,\dots,X_n)\Big) - \sum_{i=1}^\infty \widehat{\lambda}_i f(\widehat{\lambda}_i)^2\right| \le \sum_{i=1}^\infty g_n(\widehat{\lambda}_i) f(\widehat{\lambda}_i)^2, \quad (3.4)$$

*and for any indices* $r < s$*,*

$$\mathbb{E}\Big(\big\|\Pi_r(G)X\big\|^2\Big) - 2\sum_{i=1}^r g_n(\widehat{\lambda}_i) \le \mathbb{E}\Big(\big\|f_{\widehat{\lambda}_r,\widehat{\lambda}_s}(\widehat{G}_n)X\big\|^2 \,|\, (X_1,\dots,X_n)\Big) \le \mathbb{E}\Big(\big\|\Pi_s(G)X\big\|^2\Big). \quad (3.5)$$

*Proof.* Remark that

$$\mathbb{E}\Big(\big\|\Pi_r(\widehat{G}_n)X\big\|^2 \,|\, (X_1,\dots,X_n)\Big) = \mathbb{E}\left(\sum_{i=1}^r \langle X, q_i\rangle^2 \,|\, (X_1,\dots,X_n)\right) = \sum_{i=1}^r \langle q_i, Gq_i\rangle. \quad (3.6)$$

Moreover, with probability at least $1-2\delta$,

$$\langle q_i, \widehat{G}_n q_i\rangle - \langle q_i, Gq_i\rangle \le g_n\big(\langle q_i, \widehat{G}_n q_i\rangle\big), \qquad 1 \le i < \infty,$$

that can also be written as

$$\langle q_i, Gq_i\rangle \ge \widehat{\lambda}_i - g_n(\widehat{\lambda}_i), \qquad 1 \le i < \infty.$$

Since on the same event of probability at least $1-2\delta$

$$\widehat{\lambda}_i \geq \lambda_i - \min\bigl\{g(\widehat{\lambda}_i), g(\lambda_i)\bigr\}, \qquad 1 \leq i < \infty$$

we have both

$$\langle q_i, Gq_i\rangle \geq \lambda_i - 2g_n(\widehat{\lambda}_i), \qquad 1 \leq i < \infty,$$

and, since the functions $g_n$ and $t \mapsto t - g_n(t)$ are non-decreasing,

$$\langle q_i, Gq_i\rangle \geq \lambda_i - g_n(\lambda_i) - g_n\bigl(\lambda_i - g_n(\lambda_i)\bigr) \geq \lambda_i - 2g_n(\lambda_i), \qquad 1 \leq i < \infty.$$

Since

$$\sum_{i=1}^{r} \lambda_i = \mathbb{E}\Bigl(\bigl\|\Pi_r(G)X\bigr\|^2\Bigr),$$

this implies that

$$\mathbb{E}\Bigl(\bigl\|\Pi_r(\widehat{G}_n)X\bigr\|^2 \,|\, (X_1,\ldots,X_n)\Bigr) \\ \geq \mathbb{E}\Bigl(\bigl\|\Pi_r(G)X\bigr\|^2\Bigr) - 2\min\Biggl\{\sum_{i=1}^{r} g_n(\lambda_i),\ \sum_{i=1}^{r} g_n(\widehat{\lambda}_i)\Biggr\}.$$

On the other hand,

$$\mathbb{E}\Bigl(\bigl\|\Pi_r(\widehat{G}_n)X\bigr\|^2 \,|\, (X_1,\ldots,X_n)\Bigr) \\ \leq \sup_{\substack{P \text{ projector} \\ \text{of rank } k}} \mathbb{E}\Bigl(\bigl\|PX\bigr\|^2\Bigr) \leq \mathbb{E}\Bigl(\bigl\|\Pi_r(G)X\bigr\|^2\Bigr),$$

ending the proof of eq. (3.2). From eq. (3.6) and the fact that with probability at least $1-2\delta$

$$\bigl|\langle q_i, Gq_i\rangle - \widehat{\lambda}_i\bigr| \leq g_n(\widehat{\lambda}_i), \qquad 1 \leq i < \infty, \tag{3.7}$$

we obtain eq. (3.3). Observe now that

$$\mathbb{E}\Bigl(\bigl\|f(\widehat{G}_n)X\bigr\|^2 \,|\, (X_1,\ldots,X_n)\Bigr) = \mathbb{E}\Biggl(\sum_{i=1}^{\infty} f(\widehat{\lambda}_i)^2\langle X, q_i\rangle^2 \,|\, (X_1,\ldots,X_n)\Biggr) \\ = \sum_{i=1}^{\infty} f(\widehat{\lambda}_i)^2\langle q_i, Gq_i\rangle,$$

that leads to eq. (3.4) when combined with eq. (3.7). Finally, eq. (3.5) is a consequence of eq. (3.2) and the fact that

$$\mathbb{E}\Bigl(\bigl\|\Pi_r(\widehat{G}_n)X\bigr\|^2 \,|\, (X_1,\ldots,X_n)\Bigr) \\ \leq \mathbb{E}\Bigl(\bigl\|f_{\widehat{\lambda}_r,\widehat{\lambda}_s}(\widehat{G}_n)X\bigr\|^2 \,|\, (X_1,\ldots,X_n)\Bigr) \\ \leq \mathbb{E}\Bigl(\bigl\|\Pi_s(\widehat{G}_n)X\bigr\|^2 \,|\, (X_1,\ldots,X_n)\Bigr).$$

□

If we want to exploit the inequality $\|f(A)-f(B)\|_{\mathtt{HS}} \leq L\|A-B\|_{\mathtt{HS}}$ of Proposition 2.2 on page 3909, we can use an estimator $\widehat{G}$ of $G$ such that $\|\widehat{G}-G\|_{\mathtt{HS}}$ is properly controlled. One is given by Minsker in [16, Corollary 4.3]. It is based on a multidimensional extension of the median of means estimator and is such that with probability at least $1-\delta$,

$$\|\widehat{G}-G\|_{\mathtt{HS}} \leq 11\sqrt{\frac{\left[\mathbb{E}\left(\|X\|^4\right)-\mathbf{Tr}(G^2)\right]\log\left(1.4/\delta\right)}{n}}.$$

In this setting, the only assumption on the data is that $\mathbb{E}\left(\|X\|^4\right)<\infty$. Minsker's estimator is such that with probability at least $1-\delta$, for any $a \geq b \in \mathbb{R}_+$,

$$\left\|f_{a,b}(\widehat{G})-f_{a,b}(G)\right\|_{\mathtt{HS}} \leq \frac{11}{a-b}\sqrt{\frac{\left[\mathbb{E}\left(\|X\|^4\right)-\mathbf{Tr}(G^2)\right]\log(1.4/\delta)}{n}}.$$

Remark that since $\|\widehat{G}-G\|_\infty \leq \|\widehat{G}-G\|_{\mathtt{HS}}$, Minsker's estimator can also be used in conjunction with the operator norm bounds of Propositions 2.1 on page 3905 and Proposition 2.2 on page 3909, but that it would give looser inequalities. Note that the estimator we proposed in [8] has a better proved operator bound than Minsker's, at least in some cases. Therefore, it makes sense to use our estimator in conjunction with operator norm bounds instead of Minsker's. Indeed, under the assumptions of Proposition 3.1 on page 3912, considering that $\mathbf{Tr}(G)^2 \leq \mathbb{E}\left(\|X\|^4\right) \leq \kappa\,\mathbf{Tr}(G)^2$, we do not weaken much Minsker's bound by replacing $\mathbb{E}\left(\|X\|^4\right)$ with $\kappa\,\mathbf{Tr}(G)^2$, at least when $\kappa$ is of order one. Making this substitution for the sake of comparison, we get that, for Minsker's estimator,

$$\|\widehat{G}_n-G\|_\infty \leq \|\widehat{G}_n-G\|_{\mathtt{HS}} \leq \mathcal{O}\left(\sqrt{\frac{\kappa\,\mathbf{Tr}(G)^2\log(\delta^{-1}))}{n}}\right), \quad n \geq \mathcal{O}\Big(\log\big(\delta^{-1}\big)\Big),$$

whereas for our estimator we get the bound described in Proposition 3.1, that is of order

$$\|\widehat{G}_n-G\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{\kappa\|G\|_\infty}{n}\Big(\mathbf{Tr}(G)+\|G\|_\infty\log(\delta^{-1})\Big)}\right),$$
$$\|G\|_\infty \geq \mathcal{O}\Big(\frac{\kappa\,\mathbf{Tr}(G)}{n}\Big), \quad n \geq \mathcal{O}\Big(\log\big(\delta^{-1}\big)\Big),$$

where $\mathcal{O}$ represents numerical constants in non-asymptotic bounds. We have dropped the additional term in $\varepsilon$ present in eq. (3.1), since it can be made arbitrarily small depending on the computation cost of the estimator. The difference between the two bounds is better understood while making a parallel with the finite dimension case. Here, after normalization by $\|G\|_\infty$, $\mathbf{Tr}(G)$ plays the role that would be played by the dimension $d$. Minsker's estimate uses the Hilbert-Schmidt norm. In other words, in the finite dimension case, it estimates the

matrix $G$, considering it as a vector of coefficients of dimension $d \times d$. Therefore Minsker gets a convergence rate of order $\sqrt{d^2/n}$, whereas considering $G$ as a matrix and working with the operator norm, we can get a rate in $\sqrt{d/n}$ instead (although for a different estimator). In our bound, the interplay between $\log(\delta^{-1})$ and the substitute for the dimension $\mathbf{Tr}(G)$ is also more favorable.

So far we considered the Gram operator $G = \mathbb{E}\big(\langle\ , X\rangle X\big)$, whereas principal component analysis is most often concerned with the covariance operator

$$\Sigma = \mathbb{E}\Big[\langle\ , X - \mathbb{E}(X)\rangle\big(X - \mathbb{E}(X)\big)\Big].$$

We can nevertheless use the results we presented for $G$ in order to study $\Sigma$. One way to make the link between the two settings is to consider two independent copies $(X_1, X_2)$ of $X$ and to remark that

$$\Sigma = \frac{1}{2}\mathbb{E}\Big(\langle\ , X_1 - X_2\rangle(X_1 - X_2)\Big),$$

so that $\Sigma$ turns out to be the Gram operator of $(X_1 - X_2)/\sqrt{2}$, and can be estimated as such from a sample $(X_1, \ldots, X_{2n})$ of size $2n$ by forming the reduced centered sample $\big\{(X_{2i-1} - X_{2i})/\sqrt{2}, 1 \le i \le n\big\}$ of size $n$.

## Appendix A: Pairs of Orthogonal Projections

In this appendix we introduce some results on orthogonal projectors that are interesting for their own sake.

Let $P$, $Q : \mathcal{H} \to \mathcal{H}$ be two orthogonal projectors with finite ranks, defined on some separable real Hilbert space $\mathcal{H}$. Let $\mathcal{S}$ be the unit sphere of $\mathcal{H}$. The description of the relative positions of $P$ and $Q$, or equivalently of $\mathbf{Im}(P)$ and $\mathbf{Im}(Q)$ is a classical topic that goes back at least to [10]. More recently it has been treated in [6, 12]. We would like to bring a contribution to this question based on the use of an orthonormal basis of eigenvectors of $P+Q$. In such a basis the description can be split into an orthogonal sum of problems of dimension one or two, reducing the description to the relative positions of two lines in the plane. This decomposition is related to the notion of principal angles introduced by Jordan [10]. In the following contribution, we obtain it from a straightforward construction that to our knowledge was not available in the literature, although the relation between the spectrum of $P + Q$ and principal angles is proved in [7].

Let us start by listing the properties of the eigenvectors of $P + Q$.

**Lemma A.1.** *Let $x \in \mathcal{S}$ be an eigenvector of $P + Q$ with eigenvalue $\lambda$.*

1. *In the case when $\lambda = 0$, then $Px = Qx = 0$, so that $x \in \mathbf{ker}(P) \cap \mathbf{ker}(Q)$;*
2. *in the case when $\lambda = 1$, then $PQx = QPx = 0$, so that*

$$x \in \mathbf{ker}(P) \cap \mathbf{Im}(Q) \oplus \mathbf{Im}(P) \cap \mathbf{ker}(Q);$$

3. *in the case when* $\lambda = 2$*, then* $x = Px = Qx$*, so that* $x \in \mathbf{Im}(P) \cap \mathbf{Im}(Q)$*;*
4. *otherwise,* $\lambda \in ]0,1[\cup]1,2[$*,*

$$(P-Q)^2 x = (2-\lambda)\lambda x \neq 0,$$

*so that* $(P-Q)x \neq 0$*. As*

$$(P+Q)(P-Q)x = (2-\lambda)(P-Q)x,$$

*the vector* $(P-Q)x$ *is an eigenvector of* $P+Q$ *with eigenvalue* $2-\lambda$*. Moreover*

$$0 < \|Px\| = \|Qx\| = \sqrt{\frac{\lambda}{2}} < 1, \qquad \langle Px, Qx\rangle = \frac{1}{2}\lambda(\lambda-1),$$

$x - Px \neq 0$*, and* $\big(Px, x-Px\big)$ *is an orthogonal basis of* $\mathbf{span}\big\{x, (P-Q)x\big\}$*.*

*Proof.* The operator $P+Q$ is self-adjoint, nonnegative, of finite rank, and $\|P+Q\|_\infty \leq 2$, so that there is a basis of eigenvectors and all eigenvalues are in the intervall $[0,2]$.

In case 1, $0 = \langle Px+Qx, x\rangle = \|Px\|^2 + \|Qx\|^2$, so that $Px = Qx = 0$.

In case 2, $PQx = P(x-Px) = 0$ and similarly $QPx = Q(x-Qx) = 0$, so that $x = Px + Qx$, where $Px \in \mathbf{ker}(Q) \cap \mathbf{Im}(P)$ and $Qx \in \mathbf{ker}(P) \cap \mathbf{Im}(P)$.

In case 3,

$$\begin{aligned}\|Px\|^2 + \|Qx\|^2 &= \langle (P+Q)x, x\rangle = 2\langle x, x\rangle \\ &= \|Px\|^2 + \|x-Px\|^2 + \|Qx\|^2 + \|x-Qx\|^2,\end{aligned}$$

so that $\|x-Px\| = \|x-Qx\| = 0$.

In case 4, remark that

$$PQx = P(\lambda x - Px) = (\lambda-1)Px$$

and similarly $QPx = Q(\lambda x - Qx) = (\lambda-1)Qx$. Consequently

$$(P-Q)(P-Q)x = (P-QP-PQ+Q)x = (2-\lambda)(P+Q)x = (2-\lambda)\lambda x \neq 0,$$

so that $(P-Q)x \neq 0$. Moreover

$$(P+Q)(P-Q)x = (P-PQ+QP-Q)x = (2-\lambda)(P-Q)x.$$

Therefore $(P-Q)x$ is an eigenvector of $P+Q$ with eigenvalue $2-\lambda \neq \lambda$. Remark now that

$$\langle Px, Qx\rangle = \langle x, PQx\rangle = (\lambda-1)\langle x, Px\rangle.$$

Similarly

$$\langle Px, Qx\rangle = \langle QPx, x\rangle = (\lambda-1)\langle x, Qx\rangle.$$

so that

$$\langle Px, Qx\rangle = \frac{1}{2}(\lambda-1)\langle x, (P+Q)x\rangle = \frac{1}{2}\lambda(\lambda-1).$$

Coming back to the two previous equations, we then deduce that

$$\|Px\|^2 = \langle x, Px\rangle = \frac{\lambda}{2}.$$

In the same way

$$\|Qx\|^2 = \langle x, Qx\rangle = \frac{\lambda}{2}.$$

Now $\|x - Px\|^2 = \|x\|^2 - \|Px\|^2 > 0$, proving that $x - Px \neq 0$. Similarly, $x - Qx \neq 0$.

As $P$ is an orthogonal projector, $(Px, x - Px)$ is an orthogonal pair of non-zero vectors. Moreover

$$x = x - Px + Px \in \mathbf{span}\{Px, x - Px\}$$

and

$$(P - Q)x = 2Px - \lambda x = (2 - \lambda)Px - \lambda(x - Px) \in \mathbf{span}\{Px, x - Px\}$$

therefore, $(Px, x - Px)$ is an orthogonal basis of $\mathbf{span}\{x, (P - Q)x\}$. □

**Lemma A.2.** *There is an orthonormal basis* $\{x_i, 1 \leq i < \infty\}$ *of eigenvectors of* $P + Q$ *with corresponding eigenvalues* $\{\lambda_i,\ 1 \leq i < \infty\}$ *and indices* $2m \leq p \leq q \leq s$, *such that*

1. $\lambda_i \in ]1, 2[$, *if* $1 \leq i \leq m$,
2. $\lambda_{m+i} = 2 - \lambda_i$, *if* $1 \leq i \leq m$, *and* $x_{m+i} = \|(P - Q)x_i\|^{-1}(P - Q)x_i$,
3. $\mathbf{span}\{x_{2m+1}, \ldots, x_p\} = \mathbf{Im}(P) \cap \mathbf{ker}(Q)$, *and* $\lambda_{2m+1} = \cdots = \lambda_p = 1$,
4. $\mathbf{span}\{x_{p+1}, \ldots, x_q\} = \mathbf{Im}(Q) \cap \mathbf{ker}(P)$, *and* $\lambda_{p+1} = \cdots = \lambda_q = 1$,
5. $\mathbf{span}\{x_{q+1}, \ldots, x_s\} = \mathbf{Im}(P) \cap \mathbf{Im}(Q)$, *and* $\lambda_{q+1} = \cdots = \lambda_s = 2$,
6. $\mathbf{span}\{x_i, s < i < \infty\} = \mathbf{ker}(P) \cap \mathbf{ker}(Q)$, *and* $\lambda_i = 0, i > s$.

*Proof.* As already explained in the beginning of the proof of Lemma A.1, there exists a basis of eigenvectors of $P + Q$. From the previous lemma, we have that all eigenvectors in the kernel of $P + Q$ are in $\mathbf{ker}(P) \cap \mathbf{ker}(Q)$, and on the other hand obviously $\mathbf{ker}(P) \cap \mathbf{ker}(Q) \subset \mathbf{ker}(P + Q)$ so that

$$\mathbf{ker}(P + Q) = \mathbf{ker}(P) \cap \mathbf{ker}(Q).$$

In the same way the previous lemma proves that the eigenspace corresponding to the eigenvalue 2 is equal to $\mathbf{Im}(P) \cap \mathbf{Im}(Q)$. It also proves that the eigenspace corresponding to the eigenvalue 1 is included in and consequently is equal to $\big(\mathbf{Im}(P) \cap \mathbf{ker}(Q)\big) \oplus \big(\mathbf{ker}(P) \cap \mathbf{Im}(Q)\big)$, so that we can form an orthonormal basis of this eigenspace by taking the union of an orthonormal basis of $\mathbf{Im}(P) \cap \mathbf{ker}(Q)$ and an orthonormal basis of $\mathbf{ker}(P) \cap \mathbf{Im}(Q)$.

Consider now an eigenspace corresponding to an eigenvalue $\lambda \in ]0, 1[ \cup ]1, 2[$ and let $x, y$ be two orthonormal eigenvectors in this eigenspace. From the previous lemma, remark that

$$\langle (P - Q)x, (P - Q)y\rangle = \langle (P - Q)^2 x, y\rangle = (2 - \lambda)\lambda\langle x, y\rangle = 0.$$

Therefore, if $x_1, \dots, x_k$ is an orthonormal basis of the eigenspace $V_\lambda$ corresponding to the eigenvalue $\lambda$, then $(P-Q)x_1, \dots, (P-Q)x_k$ is an orthogonal system in $V_{2-\lambda}$. If this system was not spanning $V_{2-\lambda}$, we could add to it an orthogonal unit vector $y_{k+1} \in V_{2-\lambda}$ so that $x_1, \dots, x_k, (P-Q)y_{k+1}$ would be an orthogonal set of non-zero vectors in $V_\lambda$, which would contradict the fact that $x_1, \dots, x_k$ was supposed to be an orthonormal basis of $V_\lambda$. Therefore,

$$\Big(\|(P-Q)x_i\|^{-1}(P-Q)x_i,\, 1 \le i \le k\Big)$$

is an orthonormal basis of $V_{2-\lambda}$. Doing this construction for all the eigenspaces $V_\lambda$ such that $\lambda \in ]0,1[$ achieves the construction of the orthonormal basis described in the lemma. □

**Lemma A.3.** *Consider the orthonormal basis of the previous lemma. The set of vectors*

$$\begin{gathered}\big(Px_1, \dots, Px_m, x_{2m+1}, \dots, x_p, x_{q+1}, \dots, x_s\big),\\ \big(x_1 - Px_1, \dots, x_m - Px_m, x_{p+1}, \dots x_q, x_{s+1}, \dots\big),\\ \big(Qx_1, \dots, Qx_m, x_{p+1}, \dots, x_q, x_{q+1}, \dots, x_s\big),\\ \big(x_1 - Qx_1, \dots, x_m - Qx_m, x_{2m+1}, \dots, x_p, x_{s+1}, \dots\big)\end{gathered}$$

*are respectively orthogonal bases of* $\mathbf{Im}(P)$, $\mathbf{ker}(P)$, $\mathbf{Im}(Q)$ *and* $\mathbf{ker}(Q)$.

*Proof.* According to Lemma A.1, $(Px_i, x_i - Px_i)$ is an orthogonal basis of $\mathbf{span}\{x_i, x_{m+i}\}$, so that

$$\big(Px_1, \dots, Px_m, x_1 - Px_1, \dots, x_m - Px_m, x_{2m+1}, \dots\big)$$

is another orthogonal basis of $\mathcal{H}$. Each vector of this basis is either in $\mathbf{Im}(P)$ or in $\mathbf{ker}(P)$ and more precisely

$$\begin{gathered}Px_1, \dots, Px_m, x_{2m+1}, \dots, x_p, x_{q+1}, \dots, x_s \in \mathbf{Im}(P),\\ x_1 - Px_1, \dots, x_m - Px_m, x_{p+1}, \dots, x_q, x_{s+1}, \dots \in \mathbf{ker}(P).\end{gathered}$$

This proves the claim of the lemma concerning $P$. Since $P$ and $Q$ play symmetric roles, this proves also the claim concerning $Q$, *mutatis mutandis*. □

**Corollary A.1.** *The projectors $P$ and $Q$ have the same rank if and only if*

$$p - 2m = q - p.$$

**Lemma A.4.** *Assume that* $\mathbf{rank}(P) = \mathbf{rank}(Q)$. *Then*

$$\|P-Q\|_\infty = \sup_{x \in \mathbf{Im}(P) \cap \mathcal{S}} \|(P-Q)x\|, \tag{A.1}$$

*and for any orthonormal basis* $\big(e_1, \dots, e_r\big)$ *of* $\mathbf{Im}(P)$,

$$\|P-Q\|_{\mathtt{HS}}^2 = 2\sum_{i=1}^{r} \|(P-Q)e_i\|^2. \tag{A.2}$$

*Proof.* As $P-Q$ is a self-adjoint operator, we have

$$\sup_{x\in\mathcal{S}}\|(P-Q)x\|^2 = \sup\Big\{\langle (P-Q)^2x,x\rangle \mid x\in\mathcal{S}\Big\}$$
$$= \sup\Big\{\langle (P-Q)^2x,x\rangle \mid x\in\mathcal{S} \text{ is an eigenvector of } (P-Q)^2\Big\}.$$

Remark that the basis described in Lemma A.2 is also a basis of eigenvectors of $(P-Q)^2$. More precisely, according to Lemma A.1

$$\begin{aligned}
(P-Q)^2x_i &= \lambda_i(2-\lambda_i)x_i, && 1\le i\le m,\\
(P-Q)^2x_{m+i} &= \lambda_i(2-\lambda_i)x_{m+i}, && 1\le i\le m,\\
(P-Q)^2x_i &= x_i, && 2m<i\le q,\\
(P-Q)^2x_i &= 0, && q<i\le d.
\end{aligned}$$

If $q-2m>0$, then $\|P-Q\|_\infty=1$, and $q-p>0$, according to Lemma A.1, so that $\|(P-Q)x_{p+1}\|=1$, where $x_{p+1}\in\mathbf{Im}(Q)$. If $q=2m$ and $m>0$, there is $i\in\{1,\dots,m\}$ such that $\|P-Q\|^2_\infty=\lambda_i(2-\lambda_i)$. Since $x_i$ and $x_{m+i}$ are two eigenvectors of $(P-Q)^2$ corresponding to this eigenvalue, all the non-zero vectors in $\mathbf{span}\{x_i,x_{m+i}\}$ (including $Px_i$) are also eigenvectors of the same eigenspace. Consequently $(P-Q)^2Px_i=\lambda_i(2-\lambda_i)Px_i$, proving that

$$\left\|(P-Q)\frac{Px_i}{\|Px_i\|}\right\|^2=\lambda_i(2-\lambda_i),$$

and therefore that $\sup_{x\in\mathcal{S}}\|(P-Q)x\|$ is reached in $\mathbf{Im}(P)$. Finally, if $q=0$, then $P-Q$ is the null operator, so that $\sup_{x\in\mathcal{S}}\|(P-Q)x\|$ is reached everywhere, including in $\mathbf{Im}(P)\cap\mathcal{S}$. This concludes the proof of eq. (A.1).
As for eq. (A.2), since

$$\sum_{i=1}^{r}\|(P-Q)e_i\|^2$$

is the trace of $P(P-Q)^2P$, its value is independent of the choice of the orthonormal basis $(e_1,\dots,e_r)$ of $\mathbf{Im}(P)$. Therefore it is enough to prove eq. (A.2) for any special choice of orthonormal basis for $\mathbf{Im}(P)$. Let us put

$$\begin{aligned}
e_i &= \|Px_i\|^{-1}Px_i, && 1\le i\le m,\\
e_{m+i} &= \|x_i-Px_i\|^{-1}(x_i-Px_i), && 1\le i\le m,\\
e_i &= x_i, && 2m<i<\infty.
\end{aligned}$$

According to Lemma A.3, $\{e_i, 1\le i<\infty\}$ is an orthonormal basis of $\mathcal{H}$ while $(e_1,\dots,e_m,e_{2m+1},\dots,e_p,e_{q+1},\dots,e_s)$ is an orthonormal basis of $\mathbf{Im}(P)$. Moreover, according to Lemma A.1 and Lemma A.2,

$$\begin{aligned}
\langle (P-Q)^2e_i,e_i\rangle &= \lambda_i(2-\lambda_i), && 1\le i\le m,\\
\langle (P-Q)^2e_{m+i},e_{m+i}\rangle &= \lambda_i(2-\lambda_i), && 1\le i\le m,
\end{aligned}$$

$$\begin{aligned}\langle (P-Q)^2 e_i, e_i\rangle &= 1, && 2m < i \le q,\\ \langle (P-Q)^2 e_i, e_i\rangle &= 0, && q < i < \infty.\end{aligned}$$

Accordingly, remembering that $p - 2m = q - p$, as stated in Corollary A.1, we see that

$$\begin{aligned}\|P-Q\|_{\mathtt{HS}}^2 &= \sum_{i=1}^{\infty} \langle (P-Q)^2 e_i, e_i\rangle \\ &= 2\left(\sum_{i=1}^{m} + \sum_{i=2m+1}^{p} + \sum_{i=q+1}^{s}\right)\langle (P-Q)^2 e_i, e_i\rangle \\ &= 2\left(\sum_{i=1}^{m} + \sum_{i=2m+1}^{p} + \sum_{i=q+1}^{s}\right)\|(P-Q)e_i\|^2.\end{aligned}$$

□

Our study of the eigenvectors of $P + Q$ is related to the notion of principal angles, as shown in [7]. Nevertheless, in [7], the properties of $P + Q$ and its spectrum are deduced from other results about principal angles, whereas here we show conversely that the above fairly simple and elementary study of $P+Q$ can be used to derive a description of principal angles.

**Proposition A.1.** *In the previous setting, (of finite rank orthogonal projectors in a separable real Hilbert space) the principal angles between* $\mathbf{Im}(P)$ *and* $\mathbf{Im}(Q)$ *are recursively defined as*

$$\cos(\theta_k) = \langle u_k, v_k\rangle, \qquad \theta_k \in [0, \pi/2].$$

*where* $(u_k, v_k) \in \arg\max_{(u,v)} \Big\{ \langle u, v\rangle,\ u \in \mathbf{Im}(P) \cap \mathcal{S}, v \in \mathbf{Im}(Q) \cap \mathcal{S},$
$$u \perp \mathbf{span}\{u_1, \ldots, u_{k-1}\}, v \perp \mathbf{span}\{v_1, \ldots, v_{k-1}\}\Big\},$$

*for* $1 \le k \le \min\{\mathbf{rank}(P), \mathbf{rank}(Q)\}$.

*In the previous setting, assuming without loss of generality that* $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_m$, *and that* $s - q + p - m = \mathbf{rank}(P) \le \mathbf{rank}(Q) = s - p + m$,

$$\cos(\theta_k) = \begin{cases} 1, & 1 \le k \le s-q,\\ \lambda_{k-s+q} - 1, & s-q+1 \le k \le s-q+m,\\ 0, & s-q+m+1 \le k \le s-q+p-m \end{cases}$$

*A possible choice of* $u_k$ *and* $v_k$ *is*

$$u_k = v_k = x_{q+k}, \qquad 1 \le k \le s-q,$$

$$\begin{cases} u_k = \|Px_{k-s+q}\|^{-1} Px_{k-s+q}, \\ v_k = \|Qx_{k-s+q}\|^{-1} Qx_{k-s+q}, \end{cases} \qquad s-q+1 \le k \le s-q+m,$$

$$\begin{cases} u_k = x_{m+k-s+q}, \\ v_k = x_{p+k-s+q-m}, \end{cases} \qquad s-q+m+1 \le k \le s-q+p-m.$$

*Proof.* In the case when $s-q>0$, obviously $\cos(\theta_1)=1$, $u_1=v_1 \in \mathbf{Im}(P)\cap \mathbf{Im}(Q) = \mathbf{span}\{x_{q+1},\dots,x_s\}$ and any unit vector in this set can be chosen. This reasoning can be repeated for the restriction of $P$ and $Q$ to $\mathbf{span}\{u_1,\dots,u_{k-1}\}^\perp$. After $s-q$ iterations, the restriction $\widetilde{P},\widetilde{Q}$ of $P$ and $Q$ to $\mathbf{span}\{u_1,\dots,u_{s-q}\}^\perp = \Big(\mathbf{Im}(P)\cap\mathbf{Im}(Q)\Big)^\perp$ will be such that $\mathbf{Im}(\widetilde{P})\cap\mathbf{Im}(\widetilde{Q})=\{0\}$.

So let us now assume without loss of generality that this is the case from the beginning (that is from iteration one). In other words, let us assume that $s-q=0$. Let us write, according to Lemma A.3 on page 3921,

$$u = \sum_{i=1}^{m} \alpha_i \|Px_i\|^{-1} Px_i + \sum_{i=m+1}^{p-m} \alpha_i x_{2m+i}$$
$$\text{and } v = \sum_{i=1}^{m} \beta_i \|Qx_i\|^{-1} Qx_i + \sum_{i=m+1}^{m+q-p} \beta_i x_{p+i-m}.$$

Remark that

$$\|u\|^2 = \sum_{i=1}^{p-m} \alpha_i^2 = 1, \tag{A.3}$$
$$\|v\|^2 = \sum_{i=1}^{m+q-p} \beta_i^2 = 1,$$

and that

$$\langle u,v\rangle = \sum_{i=1}^{m} \alpha_i\beta_i \frac{\langle Px_i, Qx_i\rangle}{\|Px_i\|\|Qx_i\|} = \sum_{i=1}^{m} \alpha_i\beta_i(\lambda_i-1)$$

From there, it is elementary to deduce that $\langle u,v\rangle$ is maximum if and only if

$$\{i : \alpha_i \neq 0\} \in \arg\max_i \lambda_i \quad \text{and} \quad \beta_i = \alpha_i,$$

in which case $\langle u,v\rangle = \lambda_1 - 1$. Indeed, from the Cauchy-Schwarz inequality,

$$\sum_{i=1}^{m} \alpha_i\beta_i(\lambda_i-1) \le \sqrt{\sum_{i=1}^{m} \alpha_i^2(\lambda_i-1)^2},$$

and in view of (A.3), the right-hand side of this inequality is maximal, and equal to $\lambda_1-1$, if and only if $\{i:\alpha_i\neq 0\} = \arg\max_i \lambda_i$, and equality with the left-hand side then occurs if and only if $\beta_i=\alpha_i$.

We can repeat this reasoning $m$ times, showing that

$$\cos(\theta_k) = \lambda_{k-s+q}, \qquad s-q+1 \leq k \leq s-q+m.$$

and describing the possible choices of $(u_k, v_k)$ as above. We can then assume without loss of generality that $s-q=m=0$, or equivalently that $\mathbf{Im}(P) \subset \mathbf{ker}(Q)$ and $\mathbf{Im}(Q) \subset \mathbf{ker}(P)$. In this case we can choose for $u$ any vector in $\mathbf{Im}(P) \cap \mathcal{S} = \mathbf{span}\{x_{2m+1}, \dots, x_p\} \cap \mathcal{S}$ and for $v$ any vector in $\mathbf{Im}(Q) \cap \mathcal{S} = \mathbf{span}\{x_{p+1}, \dots, x_q\} \cap \mathcal{S}$, and $\langle u, v \rangle = 0$. This proves that $\theta_k = \pi/2$, when $s-q+m+1 \leq k \leq s-q+p-m$, and shows the possible choices for $(u_k, v_k)$ in this range of values of the index $k$. □

## References


[1] Biau, G. and Mas, A. (2012). *PCA-kernel estimation*. Statistics & Risk Modeling, 29:19–46. MR2901802

[2] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press. MR2061575

[3] Candès, E., Li, X., Ma, Y. and Wright, J. (2011). *Robust principal component analysis?* J. ACM, 58(3):11:1–11:37. MR2811000

[4] Catoni, O. (2012). *Challenging the empirical mean and empirical variance: a deviation study*. Ann. Inst. Henri Poincaré Probab. Stat., 48(4):1148–1185. MR3052407

[5] Catoni, O. (2016). *PAC-Bayesian bounds for the Gram matrix and least square regression with a random design*, preprint arXiv:1603.05229.

[6] Davis, C. and Kahan, W. M. (1970). *The rotation of eigenvectors by a perturbation. III*. SIAM Journal on Numerical Analysis, 7:1–46. MR0264450

[7] Galántai, A. (2008). *Subspaces, angles and pairs of orthogonal projections*. Linear and Multilinear Algebra, 56(3):227–260. MR2384652

[8] Giulini, I. (2015). *Robust dimension-free Gram operator estimates*, Bernoulli, to appear, *preprint available at arXiv:1511.06259*.

[9] Giulini, I. (2015). *Generalization bounds for random samples in Hilbert spaces*. PhD Thesis.

[10] Jordan, C. (1875). *Essai sur la géométrie à n dimensions*. Bulletin de la S. M. F., 3:103–174. MR1503705

[11] Kato, T. (1980). *Perturbation Theory for Linear Operators*. Springer-Verlag, New York. MR0203473

[12] Kato, T. (1982). *A Short Introduction to Perturbation Theory for Linear Operators*. Springer-Verlag, New York. MR0678094

[13] Koltchinskii, V. and Lounici, K. (2014). *Asymptotics and concentration bounds for spectral projectors of sample covariance*, preprint arXiv:1408.4643. MR3573302

[14] Koltchinskii, V. and Lounici, K. (2014). *Concentration inequalities and moment bounds for sample covariance operators*, preprint arXiv:1405.2468. MR3556768

[15] Koltchinskii, V. and Lounici, K. (2015). *Normal approximation and concentration of spectral projectors of sample covariance*, preprint arXiv:1504.07333. MR3611488

[16] Minsker, S. (2015). *Geometric median and robust estimation in Banach spaces*, preprint arXiv:1308.1334. MR3378468

[17] Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer-Verlag, New York. MR2168993

[18] Rudelson, M. (1999). *Random vectors in the isotropic position*. J. Funct. Anal., 164(1):60–72. MR1694526

[19] Shawe-Taylor J., Williams, C., Cristianini, N. and Kandola, J. (2002). *On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum*, Eds.): ALT 2002, LNAI 2533, pages 23–40. Springer-Verlag. MR2071605

[20] Shawe-Taylor J., Williams, C., Cristianini, N. and Kandola, J. (2005). *On the Eigenspectrum of the Gram Matrix and the Generalisation Error of Kernel PCA*. IEEE Transactions on Information Theory, 51:2005. MR2246374

[21] Schölkopf, B., Smola, A. and Müller, K. (1998). *Nonlinear Component Analysis As a Kernel Eigenvalue Problem*. Neural Comput., 10(5):1299–1319.

[22] Tropp, J. A. (2012). *User-friendly tail bounds for sums of random matrices*. Found. Comput. Math., 12(4):389–434. MR2946459

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

[24] Zhu, P. and Knyazev, A. (2013). *Angles between subspaces and their tangents*. Journal of Numerical Mathematics, 21(4):325–340. MR3245378

[25] Zwald, L., Bousquet, O. and Blanchard, G. (2007). *Statistical properties of kernel principal component analysis*. Machine Learning, 66(2–3):259–294, 2007.

[26] Zwald, L. and Blanchard, G. (2005). *On the convergence of eigenspaces in kernel principal components analysis*. Advances in Neural Inf. Proc. Systems (NIPS 05).