A note on the prediction error of principal component regression

Martin Wahl

Abstract

We analyse the prediction error of principal component regression (PCR) and prove non-asymptotic upper bounds for the corresponding squared risk. Under mild assumptions, we conclude that PCR performs as well as the oracle method obtained by replacing empirical principal components by their population counterparts. Our approach relies on perturbation bounds for the excess risk of principal component analysis.

1 Introduction

1.1 Model

Let \((\mathcal{H}, \langle \cdot, \cdot \rangle)\) be a separable Hilbert space and let \((Y, X)\) be a pair of random variables satisfying the regression equation

\[ Y = \langle f, X \rangle + \epsilon, \]

where \(f \in \mathcal{H}\), \(X\) is a centered, strongly square-integrable random variable taking values in \(\mathcal{H}\), and \(\epsilon\) is a real-valued random variable such that \(E(\epsilon|X) = 0\) and \(E(\epsilon^2|X) = \sigma^2\). We suppose that we observe \(n\) independent copies \((Y_1, X_1), \ldots, (Y_n, X_n)\) of \((Y, X)\) and consider the problem of estimating \(f\).

Allowing for general Hilbert spaces \(\mathcal{H}\), the considered statistical model covers a variety of regression problems. It includes a special case of the functional linear model, in which the responses are scalars and the covariates are curves. This model has been extensively studied in the literature.
see e.g. Horváth and Kokoszka [10] and Hsing and Eubank [11]. Moreover, it includes several kernel learning problems. For instance, nonparametric regression with random design can be written in the form (1.1) if the regression function is contained in a reproducing kernel Hilbert space. This connection between learning theory and the theory of ill-posed inverse problems has been developed e.g. in De Vito et al. [16].

In this note, we focus on estimating $f$ by principal component regression (PCR). PCR is a widely known estimation procedure in which principal component analysis (PCA) is applied in a first step to reduce the high dimensionality of the data. Then, in a second step, the responses are regressed on the leading empirical principal components. We consider the prediction error of PCR and investigate its relation to the reconstruction error of PCA. We show that, in general, the error due to regressing on empirical principal components instead of regressing on their population counterparts can be described through the excess risk in the reconstruction error of PCA. Moreover, we show that the situation is slightly different under additional source conditions, relating $f$ to the spectral characteristics of the covariance operator of $X$. In this case, good reconstruction properties of the empirical principal components alone are not sufficient to get oracle-type bounds, for which now additional perturbation bounds related to subspace distances are required.

A general account on PCR is given in the monograph by Jolliffe [13]. For a study of the prediction error of PCR, with focus on minimax optimal rates of convergence, see Hall and Horowitz [8] and Brunel, Mas, and Roche [3] for the functional context and Blanchard and Mücke [2] for the kernel learning context. Note that the latter studies PCR (resp. spectral cut-off) within a larger class of regularisation methods. Bounds for the reconstruction error using the theory of empirical risk minimisation are derived in Blanchard, Bousquet, and Zwald [1]. The final result obtained in this paper rely on the perturbation approach to the excess risk developed in Reiβ and Wahl [15].

Further notation

Let $\|\cdot\|$ denote the norm on $\mathcal{H}$. Given a bounded (resp. Hilbert-Schmidt) operator $A$ on $\mathcal{H}$, we denote the operator norm (resp. the Hilbert-Schmidt norm) of $A$ by $\|A\|_\infty$ (resp. $\|A\|_2$). Given a trace class operator $A$ on $\mathcal{H}$, we denote the trace of $A$ by $\text{tr}(A)$. For $g \in \mathcal{H}$, the empirical and the $L^2(\mathbb{P}_X)$ norm are defined by $\|g\|_n^2 = \frac{1}{n} \sum_{i=1}^n \langle g, X_i \rangle^2$ and $\|g\|_{L^2(\mathbb{P}_X)}^2 = \mathbb{E} \langle g, X \rangle^2$, respectively. For $g, h \in \mathcal{H}$ we denote by $g \otimes h$ the rank-one operator defined by $(g \otimes h)x = \langle h, x \rangle g$, $x \in \mathcal{H}$. We write $Y = (Y_1, \ldots, Y_n)^T$ and $\epsilon = \ldots$
Throughout the paper, \(c\) and \(C\) denote constants that may change from line to line (by a numerical value).

### 1.2 Principal component analysis

The covariance operator of \(X\) is denoted by \(\Sigma = E[X \otimes X]\). By the spectral theorem there exists a sequence \(\lambda_1 \geq \lambda_2 \geq \cdots > 0\) of positive eigenvalues (which is either finite or converges to zero) together with an orthonormal system of eigenvectors \(u_1, u_2, \ldots\) such that \(\Sigma\) has the spectral representation

\[
\Sigma = \sum_{j \geq 1} \lambda_j P_j \quad \text{with} \quad P_j = u_j \otimes u_j.
\]  

(1.2)

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

\[
P_{\leq d} = \sum_{j \leq d} P_j, \quad P_{>d} = I - P_{\leq d} = \sum_{k > d} P_k
\]

for the orthogonal projections onto the linear subspace spanned by the first \(d\) eigenvectors of \(\Sigma\), and onto its orthogonal complement. Moreover, let

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

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

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

and

\[
\hat{P}_{\leq d} = \sum_{j \leq d} \hat{P}_j, \quad \hat{P}_{>d} = I - \hat{P}_{\leq d} = \sum_{k > d} \hat{P}_k.
\]

Introducing \(\mathcal{P}_d = \{ P : H \to H | P\ \text{is orthogonal projection of rank} \ d \}\), the reconstruction error of \(P \in \mathcal{P}_d\) is defined by

\[
R(P) = E[\|X - PX\|^2].
\]

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

\[
P_{\leq d} \in \arg\min_{P \in \mathcal{P}_d} R(P).
\]
Similarly, the empirical reconstruction error of \( P \in \mathcal{P}_d \) is defined by
\[
R_n(P) = \frac{1}{n} \sum_{i=1}^{n} \| X_i - PX_i \|_2^2,
\]
and we have
\[
\hat{P}_{\leq d} \in \arg\min_{P \in \mathcal{P}_d} R_n(P).
\]
The excess risk of the PCA projector \( \hat{P}_{\leq d} \) is defined by
\[
\mathcal{E}_d^{PCA} = R(\hat{P}_{\leq d}) - R(P_{\leq d}).
\]
It is easy to see and derived in [15, Equation (2.3)] that
\[
\mathcal{E}_d^{PCA} = \text{tr}(\Sigma(P_{\leq d} - \hat{P}_{\leq d})).
\]

### 1.3 Principal component regression

Let \( \hat{U}_d = \text{span}(\hat{u}_1, \ldots, \hat{u}_d) \) be the linear subspace spanned by the first \( d \) eigenvectors of \( \hat{\Sigma} \). Then the PCR estimator in dimension \( d \) is defined by
\[
\hat{f}^{PCR}_d \in \arg\min_{g \in \hat{U}_d} \| Y - S_n g \|_n^2
\]
with sampling operator \( S_n : \mathcal{H} \to \mathbb{R}^n, h \mapsto (\langle h, X_i \rangle)_{i=1}^{n} \). The PCR estimator can be defined explicitly by using the singular value decomposition of \( S_n \). In fact, writing
\[
n^{-1/2} S_n = \sum_{j=1}^{n} \lambda_j^{1/2} \hat{v}_j \otimes \hat{u}_j
\]
with orthonormal basis \( \hat{v}_1, \ldots, \hat{v}_n \) of \( \mathbb{R}^n \) (note that \( S_n^* S_n = n \hat{\Sigma} \) with adjoint operator \( S_n^* \), we have
\[
\hat{f}^{PCR}_d = n^{-1/2} \sum_{j=\leq d} \lambda_j^{-1/2} (Y, \hat{v}_j) \hat{u}_j,
\]
provided that \( \hat{\lambda}_d > 0 \). Our main goal is to analyse the squared prediction error of PCR given by
\[
\| \hat{f}^{PCR}_d - f \|_{L^2(\mathbb{P}, \mathcal{X})}^2 = \langle \hat{f}^{PCR}_d - f, \Sigma(\hat{f}^{PCR}_d - f) \rangle.
\]
To avoid technical issues related to the mean squared prediction error of PCR, we define our final estimator by
\[
\bar{f}^{PCR}_d = \hat{f}^{PCR}_d \text{ if } \hat{\lambda}_d \geq \lambda_d/2 \text{ and } \bar{f}^{PCR}_d = 0 \text{ otherwise}.
\]
Note that it is possible to replace \( \lambda_d/2 \) by a feasible threshold.
1.4 The oracle estimator

Letting \( U_d = \text{span}(u_1, \ldots, u_d) \), the oracle PCR estimator in dimension \( d \) is defined by

\[
\hat{f}_d^{\text{oracle}} = \arg \min_{g \in U_d} \| Y - S_n g \|_n^2.
\]

The oracle PCR estimator has an analogous decomposition as in (1.4), with empirical eigenvalues and singular vectors replaced by those corresponding to the projected observations \( P_{\leq d} X_i \). We define our final estimator by

\[
\tilde{f}_d^{\text{oracle}} = \hat{f}_d^{\text{oracle}} \quad \text{if the}\ d^{\text{th}} \text{ largest eigenvalue of the empirical covariance operator of the } P_{\leq d} X_i \text{ is larger than or equal to } \lambda_d/2 \text{ and } \tilde{f}_d^{\text{oracle}} = 0 \text{ otherwise.}
\]

The following bound will serve as a benchmark later on:

**Proposition 1.** Suppose that \( X \) is sub-Gaussian with factor \( C_1 > 0 \), as defined in Assumption 1 below. Then there are constants \( c, C > 0 \) depending only on \( C_1 \) such that

\[
E \| \tilde{f}_d^{\text{oracle}} - f \|^2_{L_2(\mathbb{P}^X)} \leq C \left( \sum_{k > d} \lambda_k \langle f, u_k \rangle^2 + \frac{\sigma^2 d}{n} + R \right)
\]

for all \( d \leq cn \), where \( R = \lambda_1 \lambda_d^{-1}(\sigma^2 + \|f\|^2_{L_2(\mathbb{P}^X)}) e^{-cn} \). In particular, we have

\[
E \| \tilde{f}_d^{\text{oracle}} - f \|^2_{L_2(\mathbb{P}^X)} \leq C \left( \lambda_{d+1} \|f\|^2 + \frac{\sigma^2 d}{n} + R \right)
\]

for all \( d \leq cn \).

Proposition 1 is a standard risk bound for the linear least squares estimator, see e.g. [7, Chapter 11] for similar results. For completeness, a proof is given in Appendix A.1.

The oracle bias term \( \sum_{k > d} \lambda_k \langle f, u_k \rangle^2 \) suffers from the fact that \( U_d \) does not have good approximation properties for \( f \) in general. In the extreme case, it is equal to \( \lambda_{d+1} \|f\|^2 \). Sometimes it pays off to choose a different set of principal components, see e.g. [13, Chapter 8.2]. Another widely used alternative to PCR is partial least squares, see e.g. [13, Chapter 8.4].

2 Analysis of the prediction error

2.1 The bias-variance decomposition

We start with deriving a bias-variance decomposition of the mean squared prediction error of PCR conditional on the design:
Lemma 1. If $\hat{\lambda}_d > 0$, then we have

$$
\mathbb{E}(\|\hat{f}_d^{PCR} - f\|^2_{L^2(PX)}|X_1, \ldots, X_n) = \langle \hat{P}_{>d}f, \Sigma \hat{P}_{>d}f \rangle + \frac{\sigma^2}{n} \sum_{j \leq d} \hat{\lambda}_j^{-1} \text{tr}(\hat{P}_j \Sigma).
$$

Proof. Assume that $\hat{\lambda}_d > 0$. Inserting $Y = S_n f + \epsilon$ into (1.4), we have

$$
\hat{f}_d^{PCR} = \hat{P}_{\leq df} + n^{-1/2} \sum_{j \leq df} \hat{\lambda}_j^{-1/2} \langle \epsilon, \hat{v}_j \rangle \hat{u}_j.
$$

Inserting (2.1) and the identity $f = \hat{P}_{\leq df} + \hat{P}_{>d}f$ into (1.5), we get

$$
\|f_d^{PCR} - f\|^2_{L^2(PX)} = \langle \hat{P}_{>d}f, \Sigma \hat{P}_{>d}f \rangle - 2n^{-1/2} \lambda_d^{-1/2} \langle \hat{P}_{>d}f, \Sigma \hat{u}_j \rangle \langle \epsilon, \hat{v}_j \rangle
$$

$$
+ n^{-1} \sum_{j \leq d} \sum_{k \leq d} \hat{\lambda}_j^{-1/2} \hat{\lambda}_k^{-1/2} \langle \hat{u}_j, \Sigma \hat{u}_k \rangle \langle \epsilon, \hat{v}_j \rangle \langle \epsilon, \hat{v}_k \rangle.
$$

The result now follows from the fact that, conditional on the design, the $\langle \epsilon, \hat{v}_j \rangle$ are uncorrelated, each with expectation zero and variance $\sigma^2$.

2.2 Relating the bias to the excess risk

The following lemma shows a clear connection of the bias in Lemma 1 with the bias in Proposition 1 and the excess risk:

Lemma 2. We have

$$
\langle \hat{P}_{>d}f, \Sigma \hat{P}_{>d}f \rangle = \| \left( \sum_{k > d} \lambda_k^{1/2} P_k + \sum_{j \leq d} \lambda_j^{1/2} P_j \hat{P}_{>d} - \sum_{k > d} \lambda_k^{1/2} P_k \hat{P}_{\leq d} \right) f \|^2.
$$

Proof. By (1.2) and the identity $\hat{P}_{>d} = I - \hat{P}_{\leq d}$, we have

$$
\Sigma^{1/2} \hat{P}_{>d} = \sum_{j \leq d} \lambda_j^{1/2} P_j \hat{P}_{>d} - \sum_{k > d} \lambda_k^{1/2} P_k \hat{P}_{\leq d} + \sum_{k > d} \lambda_k^{1/2} P_k.
$$

Inserting this into $\langle \hat{P}_{>d}f, \Sigma \hat{P}_{>d}f \rangle = \| \Sigma^{1/2} \hat{P}_{>d}f \|^2$, the claim follows.

The squared norm of $\sum_{k > d} \lambda_k^{1/2} P_k f$ is equal to the oracle bias term. On the other hand, the remaining part in Lemma 2 is connected to the excess risk. In fact, [15, Lemma 2.6] says that for any $\mu \in \mathbb{R}$, we have

$$
\mathcal{E}_{d \rightarrow d}^{PCA} = \mathcal{E}_{\leq d}^{PCA}(\mu) + \mathcal{E}_{>d}^{PCA}(\mu)
$$

(2.2)
with
\[ \mathcal{E}^{PCA}_{\leq d}(\mu) = \sum_{j \leq d} (\lambda_j - \mu) \| P_j \hat{P}_{>d} \|^2, \quad \mathcal{E}^{PCA}_{>d}(\mu) = \sum_{k > d} (\mu - \lambda_k) \| P_k \hat{P}_{\leq d} \|^2. \]

If the terms in the sums are non-negative, then the Hilbert-Schmidt norm can be moved outside the sum, leading to a similar structure as in Lemma 2.

Our main result of this section gives a quantitative version of this connection:

**Theorem 1.** We have
\[ \langle \hat{P}_{>d} f, \Sigma \hat{P}_{>d} f \rangle \leq \lambda_{d+1} \| f \|^2 + \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1}) \| f \|^2. \]

Moreover, for each \( \delta > 0 \), we have
\[ \langle \hat{P}_{>d} f, \Sigma \hat{P}_{>d} f \rangle \leq (1 + \delta) \sum_{k > d} \lambda_k \langle u_k, f \rangle^2 + (1 + \delta^{-1}) \lambda_{d+1} \| \hat{P}_{\leq d} - P_{\leq d} \|^2 \| f \|_\infty^2 + \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1}) \| f \|^2. \]

**Proof.** Applying (1.2), we have
\[ \langle \hat{P}_{>d} f, \Sigma \hat{P}_{>d} f \rangle = \| \Sigma^{1/2} \hat{P}_{>d} f \|^2 = \sum_{j \leq d} \lambda_j \| P_j \hat{P}_{>d} f \|^2 + \sum_{k > d} \lambda_k \| P_k \hat{P}_{>d} f \|^2. \] (2.3)
The first term on the right-hand of (2.3) is equal to
\[ \sum_{j \leq d} (\lambda_j - \lambda_{d+1}) \| P_j \hat{P}_{>d} f \|^2 + \lambda_{d+1} \| P_{\leq d} \hat{P}_{>d} f \|^2. \]

Hence, by (2.2), we get
\[ \sum_{j \leq d} \lambda_j \| P_j \hat{P}_{>d} f \|^2 \leq \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1}) \| f \|^2 + \lambda_{d+1} \| P_{\leq d} \hat{P}_{>d} f \|^2. \] (2.4)

Similarly, we have
\[ \sum_{k > d} \lambda_k \| P_k \hat{P}_{>d} f \|^2 \leq \lambda_{d+1} \| P_{>d} \hat{P}_{>d} f \|^2. \] (2.5)

The first claim now follows from (2.3)-(2.5) in combination with
\[ \| P_{\leq d} \hat{P}_{>d} f \|^2 + \| P_{>d} \hat{P}_{>d} f \|^2 = \| \hat{P}_{>d} f \|^2 \leq \| f \|^2. \] (2.6)
It remains to prove the second claim. By the identity $P_k \hat{P}_{>d} f = P_k f - P_k \hat{P}_{\leq d} f$, the triangle inequality, and the inequality $(a+b)^2 \leq (1+\delta)a^2 + (1+\delta^{-1})b^2$, $a, b, \delta \geq 0$, we get
\[
\sum_{k>d} \lambda_k \|P_k \hat{P}_{>d} f\|^2 \leq (1+\delta) \sum_{k>d} \lambda_k \langle u_k, f \rangle^2 + (1+\delta^{-1}) \sum_{k>d} \lambda_k \|P_k \hat{P}_{\leq d} f\|^2.
\]
(2.7)

We have
\[
\sum_{k>d} \lambda_k \|P_k \hat{P}_{\leq d} f\|^2 \leq \lambda_{d+1} \|P_{>d} \hat{P}_{<d} f\|^2.
\]

Moreover,
\[
\|P_{<d} \hat{P}_{>d} f\|^2 + \|P_{>d} \hat{P}_{<d} f\|^2 = \|P_{<d} \hat{P}_{>d} f - P_{>d} \hat{P}_{<d} f\|^2.
\]
Inserting $\hat{P}_{<d} - P_{<d} = P_{>d} \hat{P}_{<d} - P_{>d} \hat{P}_{>d}$, this is bounded by
\[
\|\hat{P}_{<d} - P_{<d}\|_2^2 \|f\|_2^2,
\]
and the second claim follows from combining this with (2.4) and (2.7).

**Remark 1.** The term $\lambda_{d+1}$ in the first bound of Theorem 1 is unavoidable in general. In fact, suppose that $X$ is a Gaussian random variable in $\mathbb{R}^p$ with expectation 0 and covariance $\Sigma = I_p$. Then we have $\mathcal{E}^{PCA}(\lambda_{d+1}) = 0$. On the other hand, the empirical eigenbasis (considered as an orthogonal matrix, the sign of each column chosen uniformly at random) is distributed according to the Haar measure on the orthogonal group (see e.g. [6, Theorem 5.3.1]). In particular each eigenvector is distributed according to the uniform measure $\mu$ on the $(p-1)$-sphere $S^{p-1}$ and we get
\[
\mathbb{E}\langle \hat{P}_{>d} f, \Sigma \hat{P}_{>d} f \rangle = \mathbb{E}\langle \hat{P}_{>d} f, f \rangle = (p-d) \int_{S^{p-1}} \langle u, f \rangle^2 d\mu(u) = \frac{p-d}{p}.
\]

While this example implies that the oracle bias cannot be achieved in general, the second bound in Proposition 1 shows that it can be achieved, provided that $\lambda_{d+1} \|\hat{P}_{<d} - P_{<d}\|_2^2$ is sufficiently small. Note that the operator norm distance can only be small if $\lambda_d \neq \lambda_{d+1}$, since otherwise, it is not even uniquely defined.

**Remark 2.** For every $r \leq d$, the first bound can be improved to
\[
\langle \hat{P}_{>d} f, \Sigma \hat{P}_{>d} f \rangle \leq 2\lambda_{d+1} \|P_{>d} f\|_2^2 + 2\lambda_{d+1} \|\hat{P}_{\leq d} - P_{\leq d}\|_\infty^2 \|f\|_2^2 + \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1}) \|f\|_2^2;
\]
which follows from replacing (2.6) with \( \| \hat{P}_{>d} f \|^2 < \| \hat{P}_{>r} f \|^2 < 2 \| P_{>r} f \|^2 + 2 \| (\hat{P}_{<r} - P_{<r}) f \|^2 \). While the second bound in Theorem 1 suffices to deal e.g. with exact polynomial and exponential decay of the \( \lambda_j \) (see Section 3.2 below), the above flexibility in \( r \leq d \) is important to deal with the case that polynomial (resp. exponential) decay only holds approximately.

### 2.3 Analysis of the variance

The variance in Lemma 1 contains trace scalar products with \( \Sigma \), similarly as the excess risk in (1.3). The main difference is the weighting by the \( \hat{\lambda}_j^{-1} \). Grouping together eigenvalues which are of comparable magnitude, one can bound the variance by a constant times \( \sigma^2 d n^{-1} \) plus remainder terms having a similar structure as the representation of the excess risk given in (2.2).

**Theorem 2.** Let \( 0 = r_0 < r_1 < \cdots < r_{d'} = d \) be natural numbers such that \( \lambda_{r_l}^{-1} \lambda_{r_l-1} \leq 2 \) for every \( l \leq d' \). If \( \hat{\lambda}_d \geq \lambda_d / 2 \), then we have

\[
\sum_{j \leq d} \hat{\lambda}_j^{-1} \text{tr}(\hat{P}_j \Sigma) \\
\leq 4d + 2 \sum_{2 \leq l \leq d'} \lambda_{r_l}^{-1} \sum_{k \notin J_l} |\lambda_k - \lambda_{r_l-1}| \| \hat{P}_{J_l} P_k \|_2^2 + 2 \lambda_1 \lambda_d^{-1} \sum_{j \leq d} \mathbb{1}(\hat{\lambda}_j < \lambda_j / 2)
\]

with \( J_l = \{ r_{l-1} + 1, \ldots, r_l \} \) and \( \hat{P}_{J_l} = \sum_{j \in J_l} \hat{P}_j \).

**Proof.** If \( \hat{\lambda}_d \geq \lambda_d / 2 \), then

\[
\sum_{j \leq d} \hat{\lambda}_j^{-1} \text{tr}(\hat{P}_j \Sigma) \\
= \sum_{j \leq d} \hat{\lambda}_j^{-1} \text{tr}(\hat{P}_j \Sigma) \mathbb{1}(\hat{\lambda}_j \geq \lambda_j / 2) + \sum_{j \leq d} \hat{\lambda}_j^{-1} \text{tr}(\hat{P}_j \Sigma) \mathbb{1}(\hat{\lambda}_j < \lambda_j / 2) \\
\leq 2 \sum_{j \leq d} \lambda_j^{-1} \text{tr}(\hat{P}_j \Sigma) + 2 \lambda_1 \lambda_d^{-1} \sum_{j \leq d} \mathbb{1}(\hat{\lambda}_j < \lambda_j / 2),
\]

where we applied \( \hat{\lambda}_j^{-1} \leq \lambda_d^{-1} \leq 2 \lambda_1^{-1} \) and \( \text{tr}(\hat{P}_j \Sigma) \leq \lambda_1 \) in the last inequality. It remains to bound the first term on the right-hand side. We have

\[
\sum_{j \leq d} \lambda_j^{-1} \text{tr}(\hat{P}_j \Sigma) \leq \sum_{l \leq d'} \lambda_{r_l}^{-1} \text{tr}(\hat{P}_{J_l} \Sigma) \tag{2.8}
\]

By (1.2) and the fact that orthogonal projectors are idempotent and self-adjoint, we have

\[
\text{tr}(\hat{P}_{J_l} \Sigma) = \sum_{k \geq 1} \lambda_k \text{tr}(\hat{P}_{J_l} P_k) = \sum_{k \geq 1} \lambda_k \| \hat{P}_{J_l} P_k \|_2^2.
\]
Hence, \( \text{tr}(\hat{P}_J \Sigma) \leq \lambda_1 \| \hat{P}_J \|^2 \) and
\[
\text{tr}(\hat{P}_J \Sigma) \leq \lambda_{r_l-1} \| \hat{P}_J \|^2 + \sum_{k \leq r_l-1} (\lambda_k - \lambda_{r_l-1+1}) \| \hat{P}_J P_k \|^2 \\
\leq \lambda_{r_l-1} (r_l - r_l-1) + \sum_{k \notin J_l} (\lambda_k - \lambda_{r_l-1+1}) \| \hat{P}_J P_k \|^2
\]
for every \( 2 \leq l \leq d' \). Inserting this into (2.8) and using that \( \lambda_{r_l}^{-1} \lambda_{r_l-1} \leq 2 \) for every \( l \leq d' \), the claim follows.

2.4 Bounds for the prediction error

Combining Lemma 1 with Theorems 1 and 2, we obtain an upper bound for the mean squared prediction error of PCR conditional on the design. This bound has the same leading terms appearing in Proposition 1 for the oracle PCR estimator. In Section 3.1, we will see how the remainder terms can be bounded using the excess risk bounds from [15].

**Corollary 1.** Grant the assumptions of Theorem 2. Then, on the event \( \{ \hat{\lambda}_d \geq \lambda_d/2 \} \), we have
\[
\mathbb{E}(\| \hat{f}^{PCR}_d - f \|_{L^2(\mathbb{P}|X_1, \ldots, X_n)}^2) \leq 2 \sum_{k > d} \lambda_k (f, u_k)^2 + \frac{\sigma^2}{n} (4d + R_1) + R_2
\]
with
\[
R_1 = 2 \sum_{2 \leq l \leq d'} \lambda_{r_l}^{-1} \sum_{k \notin J_l} (\lambda_k - \lambda_{r_l-1+1}) \| \hat{P}_J P_k \|^2 + 2 \lambda_1 \lambda_{r_l}^{-1} \sum_{j \leq d} 1(\hat{\lambda}_j < \lambda_j/2).
\]
and
\[
R_2 = \mathcal{E}_{d}^{PCA}(\lambda_{d+1}) \| f \|^2 + 2 \lambda_{d+1} \| \hat{P}_{\leq d} - P_{\leq d} \|_{\infty}^2 \| f \|^2.
\]

While the prediction error is standard in the learning context, estimates in \( \mathcal{H} \)-norm are standard for inverse problems. Following a similar but simpler line of arguments, we get the following upper bound in \( \mathcal{H} \)-norm. We observe that the excess risk does no longer play a role.

**Corollary 2.** For every \( r \leq d \), we have on the event \( \{ \hat{\lambda}_d \geq \lambda_d/2 \} \),
\[
\mathbb{E}(\| \hat{f}^{PCR}_d - f \|^2|X_1, \ldots, X_n) \leq 2 \| P_{>r} f \|^2 + \frac{2\sigma^2}{n} \left( \sum_{j \leq d} \lambda_j^{-1} + R_1 \right) + R_2
\]
with
\[
R_1 = \lambda_d^{-1} \sum_{j \leq d} 1(\hat{\lambda}_j < \lambda_j/2), \quad R_2 = 2 \| \hat{P}_{\leq r} - P_{\leq r} \|_{\infty}^2 \| f \|^2.
\]
3 Bounds for the mean squared prediction error

3.1 Bounds for the excess risk

In this section, we state the main perturbation bounds for the excess risk obtained in [15]. From now on, we suppose that $X$ is sub-Gaussian. In order to deal with weaker moment assumptions one can alternatively use the results obtained in [12], though under stronger eigenvalue assumptions.

**Assumption 1.** Suppose that there is a constant $C_1$ such that
\[
\sup_{k \geq 1} k^{-1/2} \mathbb{E} \left[ |\langle g, X \rangle|^k \right]^{1/k} \leq C_1 \mathbb{E} [\langle g, X \rangle^2]^{1/2}
\]
for every $g \in H$.

First, [15, Corollary 2.8] says that
\[
\mathbb{E} \mathbb{E}^{PCA}(\lambda_{d+1}) \leq C \sum_{j \leq d} \min \left( \frac{\lambda_j \text{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})}, \lambda_j - \lambda_{d+1} \right)
\]
with a constant $C$ depending only on $C_1$. For instance, if $j\lambda_j \geq (d+1)\lambda_{d+1}$ for every $j \leq d$, implying that $\lambda_j/(\lambda_j - \lambda_{d+1}) \leq (d+1)/(d+1-j)$, then
\[
\mathbb{E} \mathbb{E}^{PCA}(\lambda_{d+1}) \leq C \frac{d \log(ed)}{n}.
\]

While this bound is only up to a logarithmic term larger than the variance term $\sigma^2 dn^{-1}$, [15, Proposition 2.10] states the following improvement, valid under additional eigenvalue conditions:

**Theorem 3.** Grant Assumption 1. Then for all $r \leq d$ such that
\[
\frac{\lambda_r}{\lambda_r - \lambda_{d+1}} \sum_{j \leq r} \frac{\lambda_j}{\lambda_j - \lambda_{d+1}} \leq cn,
\]
we have
\[
\mathbb{E} \mathbb{E}^{PCA}(\lambda_{d+1}) \leq C \sum_{j \leq r} \sum_{k > r} \frac{\lambda_j \lambda_k}{n(\lambda_j - \lambda_{d+1})} + 2 \sum_{r < j \leq d} (\lambda_j - \lambda_{d+1}) + R
\]
with remainder term given by
\[
R = C \sum_{j \leq r} \frac{\lambda_j \text{tr}(\Sigma)}{n(\lambda_j - \lambda_{d+1})} e^{-cn(\lambda_r - \lambda_{d+1})^2/\lambda_r^2}.
\]

Here, $c, C > 0$ are constants depending only on $C_1$. 

11
If for some $\alpha > 1$

$$\lambda_j = j^{-\alpha}, \quad j \geq 1,$$

then Theorem 3 with $r = d$ gives

$$\mathbb{E} \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1}) \leq C \frac{d \log(ed)}{n} \left( \sum_{k > d} \lambda_k + \text{tr}(\Sigma)e^{-cn/d^2} \right) \leq C \frac{d^{2-\alpha} \log(ed)}{n},$$

(provided that $d^2 \log(ed) \leq cn$, where $c, C > 0$ are constants depending only on $C_1$ and $\alpha$). Similarly, if for some $\alpha > 0$

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

then Theorem 3 with $r = d$ gives

$$\mathbb{E} \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1}) \leq C \frac{de^{-\alpha d}}{n},$$

(provided that $d \leq cn$). A simple inequality relating the Hilbert-Schmidt distance with the excess risk is given by

$$\|\hat{P}_d - P_d\|_2^2 \leq \frac{2\mathbb{E} \mathcal{E}^{PCA}_{\leq d}(\lambda_{d+1})}{\lambda_d - \lambda_{d+1}},$$

see e.g. [15, Equation (2.21)]. This means that all excess risk bounds imply bounds on the Hilbert-Schmidt distance up to a spectral gap factor.

Using the techniques leading to Theorem 3, we prove in Appendix A.2 the following upper bound dealing with the first remainder term in Theorem 2.

**Theorem 4.** For natural numbers $s > r > 0$, let $J = \{r+1, \ldots, s\}$. Suppose that $\lambda_{r+1} - \lambda_s \leq \lambda_s - \lambda_{s+1}$ and that

$$\left( \max_{k \geq 1} \frac{\lambda_k}{g_k} \right) \left( \sum_{k \geq 1} \frac{\lambda_k}{g_k} \right) \leq cn$$

(with $g_k = \lambda_k - \lambda_{r+1}$ for $k \leq r$, $g_k = \min(\lambda_r - \lambda_k, \lambda_{s+1} - \lambda_k)$ for $k \in J$, and $g_k = \lambda_s - \lambda_k$ for $k > s$). Then we have

$$\mathbb{E} \sum_{k \notin J} |\lambda_{r+1} - \lambda_k|\|\hat{P}_J P_k\|_2^2 \leq C \sum_{j \in J} \sum_{k \notin J} \frac{\lambda_j \lambda_k}{n|\lambda_{r+1} - \lambda_k|} + R$$

with remainder term

$$R = C\lambda_1 (s - r)^2 e^{-cn \max_{k \geq 1} g_k^2/\lambda_k^2}.$$
In particular, we have

\[ \mathbb{E} \lambda^{-1} \sum_{k \not\in J} |\lambda_{r+1} - \lambda_k||\hat{P}_j P_k\|^2_2 \leq C(s-r) + \lambda^{-1} R. \] (3.8)

Here, \( c, C > 0 \) are constants depending only on \( C_1 \).

Turning to the setting of Theorem 2, suppose that (3.7) holds for each pair \((r, s) = (r_{l-1}, r_l), 2 \leq l \leq d'\). Then it follows from (3.8) that the expectation of the first remainder term in Theorem 2 is bounded by a constant times \( d \) plus an (exponentially) small remainder term. A simple choice is given by \( r_j = j \) for every \( j \leq d \). In this case, (3.7) with \( J = \{j\} \) is implied by

\[ \frac{\lambda_j}{g_j} \left( \frac{\lambda_j}{g_j} + \sum_{k \neq j} \frac{\lambda_k}{|\lambda_j - \lambda_k|} \right) \leq cn \]

with \( g_j = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}) \). If e.g. (3.2) holds, then Theorem 4 gives

\[ \mathbb{E} \sum_{j \leq d} \lambda_j^{-1} \sum_{k \neq j} |\lambda_j - \lambda_k||\hat{P}_j P_k\|^2_2 \leq Cd + C d \lambda_1 \lambda^{-1} e^{-cn/d^2} \leq Cd, \] (3.9)

provided that \( d^2 \log(ed) \leq cn \), where \( c, C > 0 \) are constants depending only on \( C_1 \) and \( \alpha \). Similarly, if (3.4) holds, then Theorem 4 gives

\[ \mathbb{E} \sum_{j \leq d} \lambda_j^{-1} \sum_{k \neq j} |\lambda_j - \lambda_k||\hat{P}_j P_k\|^2_2 \leq Cd, \] (3.10)

provided that \( d \leq cn \). We conclude this section by stating \([15, \text{Theorem 2.15}]\) (applied with \( y = 1/2 \)), needed to deal with the second remainder term in Theorem 2.

**Proposition 2.** Grant Assumption 1. Then there are constants \( c_1, c_2 > 0 \) depending only on \( C_1 \) such that

\[ \mathbb{P}(\hat{\lambda}_j < \lambda_j/2) \leq e^{1-c_1 n} \]

for all \( j \leq c_2 n \).

### 3.2 Standard model assumptions

Let us apply our results to standard model assumptions on the \( \lambda_j \) and \( f \).
Assumption 2. Suppose that there is a constant $C_2 > 0$ such that

$$
\sum_{k \neq j} |\lambda_k - \lambda_j| \leq C_2 j \log(ej), \quad \frac{\lambda_{j-1}}{\lambda_j - \lambda_j} \leq j
$$

(3.11)

for all $j \leq d + 1$.

Assumption 2 is satisfied under (3.2) and (3.4). It holds under quite general conditions, see e.g. [5, Lemma 1] and [9, Lemma 10.1].

Corollary 3. Grant Assumptions 1 and 2. Then we have

$$
\mathbb{E}\|\hat{f}_d^{PCR} - f\|_{L^2(P_X)}^2 \leq C\left(\lambda_{d+1}\|f\|^2 + \frac{\sigma^2 d}{n} + R_1\right)
$$

for all $d \geq 1$ such that $d^2 \log(ed) \leq cn$, where

$$
R_1 = \frac{d \log(ed)}{n} \left(\sum_{k > d} \lambda_k\|f\|^2 + \left(\text{tr}(\Sigma)\|f\|^2 + \lambda_1\lambda_{d+1}^{-1}\frac{\sigma^2 d}{n}\right)e^{-cn/d^2}\right).
$$

Moreover, we have

$$
\mathbb{E}\|\hat{f}_d^{PCR} - f\|_{L^2(P_X)}^2 \leq C\left(\sum_{k > d} \lambda_k\langle f, u_k \rangle^2 + \frac{\sigma^2 d}{n} + R_2\right)
$$

for all $d \geq 1$ such that $d^2 \log(ed) \leq cn$, where

$$
R_2 = \frac{d \log(ed)}{n} \left(\sum_{k > d} \lambda_k\|f\|^2 + \left(\text{tr}(\Sigma)\|f\|^2 + \lambda_1\lambda_{d+1}^{-1}\frac{\sigma^2 d}{n}\right)e^{-cn/d^2}\right).
$$

In both inequalities, $c, C > 0$ are constants depending only on $C_1$ and $C_2$.

Corollary 3 follows from Corollary 1 (applied with $r_j = j$, $j \leq d$) in combination with the results from Section 3.1. Note that the first inequalities in (3.3) and (3.9) also hold under Assumption 2. Moreover, the first term in $R_2$ comes from applying (3.6), Theorem 3, and Assumption 2 to $\lambda_{d+1}\|\tilde{P}_{\leq d} - P_{\leq d}\|_\infty$.

Example 1 (Polynomial decay). Suppose that (3.2) holds. Then, Assumption 2 holds with a constant $C_2 > 0$ depending only on $\alpha$ and the first bound in Corollary 3 implies

$$
\mathbb{E}\|\hat{f}_d^{PCR} - f\|_{L^2(P_X)}^2 \leq C\left(d^{-\alpha}\|f\|^2 + \frac{\sigma^2 d}{n}\right),
$$
provided that $d^2 \log(ed) \leq cn$, where $c, C > 0$ are constants depending only on $C_1$ and $\alpha$. Similarly, the second bound in Corollary 3 implies

$$
\mathbb{E}\|\tilde{f}^{PCR}_d - f\|_{L^2(P \mathcal{X})}^2 \leq C \left( \sum_{k > d} k^{-\alpha} \langle f, u_k \rangle^2 + \frac{\sigma^2 d}{n} + \frac{d^{1-\alpha} \log(ed)}{n} \|f\|^2 + \|f\|^2 e^{-cn/d^2} \right),
$$

(3.12)

provided that $d^2 \log(ed) \leq cn$. We see that the first remainder term can be added to the variance term, provided that $\alpha > 2$. Let us evaluate (3.12) under the source condition

$$
f \in H(r, L) = \{ g \in \mathcal{H} : g = \Sigma^r h, \|h\| \leq L \}, \quad r > 0, L > 0,$$

which is equivalent to assuming that $f$ lies in a Sobolev-type ellipsoid. In this case, assuming $\alpha > 2$ and choosing $d$ of size $n^{1/(2r\alpha + \alpha + 1)}$ leads to

$$
\sup_{f \in H(r,L)} \mathbb{E}\|\tilde{f}^{PCR}_d - f\|_{L^2(P \mathcal{X})}^2 \leq C n^{-(2r\alpha + \alpha)/(2r\alpha + \alpha + 1)}
$$

with a constant $C > 0$ depending only on $C_1$, $\alpha$, $r$, $L$, and $\sigma^2$. This rate is minimax optimal, see [8, 4] and [2] for the functional and the kernel learning context, respectively.

**Example 2 (Exponential decay).** Suppose that (3.4) holds. Then Corollary 3 can be improved, by using (3.5) and (3.10). In fact, we have

$$
\mathbb{E}\|\tilde{f}^{PCR}_d - f\|_{L^2(P \mathcal{X})}^2 \leq C (e^{-\alpha d}\|f\|^2 + \frac{\sigma^2 d}{n}),
$$

provided that $d \leq cn$, where $c, C > 0$ are constants depending only on $C_1$ and $\alpha$. In particular, choosing $d$ of size $\log(n)/\alpha$ gives

$$
\sup_{f \in \mathcal{H} : \|f\| \leq 1} \mathbb{E}_f \|\tilde{f}_d^{\text{oracle}} - f\|_{L^2(P \mathcal{X})}^2 \leq C \frac{\log(n)}{n}
$$

with a constant $C > 0$ depending only on $C_1$, $\alpha$, and $\sigma^2$. Again, the rate is minimax optimal, see e.g. [4].

**Acknowledgements**

The research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 “Data Assimilation”, Project (A4) “Non-linear statistical inverse problems with random observations”.

15
References

[1] G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Mach. Learn.*, 66:259–294, 2007.

[2] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18:971–1013, 2018.

[3] E. Brunel, A. Mas, and A. Roche. Non-asymptotic adaptive prediction in functional linear models. *J. Multivariate Anal.*, 143:208–232, 2016.

[4] H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. *J. Multivariate Anal.*, 101:395–408, 2010.

[5] H. Cardot, A. Mas, and P. Sarda. CLT in functional linear regression models. *Probab. Theory Related Fields*, 138:325–361, 2007.

[6] R. H. Farrell. *Multivariate calculation*. Springer-Verlag, New York, 1985.

[7] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002.

[8] P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35:70–91, 2007.

[9] N. Hilgert, A. Mas, and N. Verzelen. Minimax adaptive tests for the functional linear model. *Ann. Statist.*, 41:838–869, 2013.

[10] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer, New York, 2012.

[11] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, Ltd., Chichester, 2015.

[12] M. Jirak and M. Wahl. Perturbation bounds for eigenspaces under a relative rank condition. Available at https://arxiv.org/pdf/1803.03868, 2018.

[13] I. T. Jolliffe. *Principal component analysis*. Springer-Verlag, New York, second edition, 2002.
[14] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23:110–133, 2017.

[15] M. Reiß and M. Wahl. Non-asymptotic upper bounds for the reconstruction error of PCA. Available at https://arxiv.org/abs/1609.03779, 2016.

[16] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.

A Additional proofs

A.1 Proof of Proposition 1

We abbreviate \(\hat{f}_{d,\text{oracle}}\) and \(\hat{f}_{d,\text{oracle}}\) by \(\hat{f}_d\) and \(\hat{f}_d\), respectively. Consider \(X' = P_{\leq d}X, X'_j = P_{\leq d}X_j\) (defined on \(U_d\)) which lead to covariance and sample covariance \(\Sigma' = P_{\leq d}\Sigma P_{\leq d}, \hat{\Sigma}' = P_{\leq d}\hat{\Sigma} P_{\leq d}\). In the following, we use the notation introduced in Sections 1.2, 1.3 with an additional superscript \('\).

With this notation, we have

\[
\hat{f}_d = n^{-1/2} \sum_{j \leq d} \hat{\lambda}'_j^{-1/2} \langle Y, \hat{v}'_j \rangle \hat{u}'_j,
\]

provided that \(\hat{\lambda}'_d > 0\). We define the events \(A_d = \{\hat{\lambda}'_d \geq \lambda_d/2\}\) and

\[
\mathcal{E}_d = \{(1/2)\|g\|^2_{L^2(P_X)} \leq \|g\|^2_n \leq (3/2)\|g\|^2_{L^2(P_X)} \forall g \in U_d\}.
\]

Inserting \(\|g\|^2_n = \langle g, \hat{\Sigma}'^\ast g \rangle, \|g\|^2_{L^2(P_X)} = \langle g, \Sigma' g \rangle, g \in U_d,\) and applying [14, Theorem 1], there are constants \(c_1, c_2 > 0\) depending only on \(C_1\) such that

\[
\mathbb{P}(\mathcal{E}_d) = \mathbb{P}(\|\Sigma'_{-1/2}(\Sigma' - \Sigma')\Sigma'_{-1/2}\|_{\infty} > 1/2) \leq e^{-c_1 n},
\]

provided that \(d \leq c_2 n\). We now decompose the mean squared prediction error as follows:

\[
\mathbb{E}\|\hat{f}_d - f\|^2_{L^2(P_X)} \leq \mathbb{E}1_{A_d \cap \mathcal{E}_d}\|\hat{f}_d - P_{\leq d}f\|^2_{L^2(P_X)} + \mathbb{E}1_{A_d \cap \mathcal{E}_d}\|\hat{f}_d - P_{\leq d}f\|^2_{L^2(P_X)}
\]

\[
+ \mathbb{E}1_{A_d} \|P_{\leq d}f\|^2_{L^2(P_X)} + \|P_{>d}f\|^2_{L^2(P_X)} =: I_1 + \cdots + I_4.
\]

The term \(I_4\) is exactly the oracle bias term. By Proposition 2, we have

\[
I_3 \leq \|f\|^2_{L^2(P_X)}e^{1-c_1 n},
\]
provided that $d \leq c_2 n$. Moreover,

$$I_2 \leq 2 \mathbb{P}(\mathcal{E}_d^c) \|f\|_{L^2(\mathbb{P}^X)}^2 + 2 \mathbb{E} 1_{A_d \cap \mathcal{E}_d} \|\hat{f}_d\|_{L^2(\mathbb{P}^X)}^2.$$ 

Inserting that on $A_d$,

$$\|\hat{f}_d\|_{L^2(\mathbb{P}^X)}^2 \leq \lambda_1 \|\hat{f}_d\|^2 \leq 2\lambda_1 \lambda_d^{-1} \|Y\|_n^2 \leq 4\lambda_1 \lambda_d^{-1}(\|f\|_n^2 + \|\epsilon\|_n^2),$$

we get

$$I_2 \leq 2 \mathbb{P}(\mathcal{E}_d^c) \|f\|_{L^2(\mathbb{P}^X)}^2 + 8\lambda_1 \lambda_d^{-1} \mathbb{P}(\mathcal{E}_d^c) \mathbb{E}^{1/2} \|f\|_n^4 + 8\lambda_1 \lambda_d^{-1} \mathbb{P}(\mathcal{E}_d^c) \sigma^2.$$ 

By the Cauchy-Schwarz inequality and Assumption 1, we get

$$\mathbb{E}^{1/2} \|f\|_n^4 \leq \mathbb{E}^{1/2}(f, X)^4 \leq 4C_2^2 \|f\|_{L^2(\mathbb{P}^X)}^2$$

and thus

$$I_2 \leq (32C_2^2 + 10)\lambda_1 \lambda_d^{-1}(\sigma^2 + \|f\|_{L^2(\mathbb{P}^X)}^2)e^{-c_1 n}.$$

Finally,

$$(1/2)I_1 \leq \mathbb{E} 1_{A_d \cap \mathcal{E}_d} \|\hat{f}_d - P_{\leq d} f\|_n^2.$$ 

Letting $\hat{\Pi}_n : \mathbb{R}^n \to S_n U_d, y \mapsto \arg\min_{g \in U_d} \|y - S_n g\|_n^2$ be the orthogonal projection onto $S_n U_d$, we have

$$(1/2)I_1 \leq \mathbb{E} \|\hat{\Pi}_n f - P_{\leq d} f\|_n^2 + \mathbb{E} \mathbb{E}(\|\hat{\Pi}_n \epsilon\|_n^2 | X_1, \ldots, X_n)$$

$$\leq \mathbb{E} \|f - P_{\leq d} f\|_n^2 + \sigma^2 d n^{-1} = \|P_{> d} f\|_{L^2(\mathbb{P}^X)}^2 + \sigma^2 d n^{-1},$$

where we applied the projection theorem in the second inequality. The claim follows from collecting all these inequalities. \(\square\)

### A.2 Proof of Theorem 4

The following lemma is an extension of [15, Proposition 3.5]:

**Lemma 3.** Let $\Delta = \Sigma - \hat{\Sigma}$, $J = \{r + 1, \ldots, s\}$, $\hat{P}_j = \sum_{j \in J} \hat{P}_j$, and

$$S_J = \sum_{k \notin J} |\lambda_{r+1} - \lambda_k|^{-1/2} P_k.$$ 

Suppose that $\lambda_{r+1} - \lambda_s \leq \lambda_s - \lambda_{s+1}$. Then, on the event $\bigcap_{j \in J, k \notin J} \{ |\lambda_j - \lambda_k| \geq |\lambda_j - \lambda_k| / 2 \} \cap \{ ||S_J \Delta S_J反\|_\infty \leq 1/8 \}$, we have

$$\sum_{k \notin J} |\lambda_{r+1} - \lambda_k| \|\hat{P}_j P_k\|_2^2 \leq 64 \sum_{k \notin J} \frac{\|P_j \Delta P_k\|_2^2}{|\lambda_{r+1} - \lambda_k|}.$$
Proof. On the event \( \bigcap_{j \in J, \lambda \notin J} \{ |\hat{\lambda}_j - \lambda_k| \geq |\lambda_j - \lambda_k|/2 \} \) it follows from [15, Equation (3.1)] that
\[
\| \hat{P}_j P_k \|_2^2 \leq \frac{4 \| \hat{P}_j \Delta P_k \|_2^2}{(\lambda_j - \lambda_k)^2}, \tag{A.1}
\]
for every \( j \in J, \lambda \notin J \). By \( \lambda_{r+1} - \lambda_s \leq \lambda_s - \lambda_{s+1} \), the inequalities \( \lambda_j - \lambda_{s+1} \geq \lambda_s - \lambda_{s+1} \geq (\lambda_{r+1} - \lambda_{s+1})/2 \) and \( \lambda_j - \lambda_k = \lambda_j - \lambda_{s+1} + \lambda_{s+1} - \lambda_k \geq (\lambda_{r+1} - \lambda_k)/2 \) hold for every \( j \in J, \lambda \notin J \). Using this and (A.1), we get on \( \bigcap_{j \in J, \lambda \notin J} \{ |\lambda_j - \lambda_k| \geq |\lambda_j - \lambda_k|/2 \} \),
\[
\| \hat{P}_j P_k \|_2^2 \leq \frac{16 \| \hat{P}_j \Delta P_k \|_2^2}{(\lambda_{r+1} - \lambda_k)^2}
\]
for every \( j \in J, \lambda \notin J \) and thus
\[
\sum_{\lambda \notin J} |\lambda_{r+1} - \lambda_k| \| \hat{P}_j P_k \|_2 \leq 16 \sum_{\lambda \notin J} \| \hat{P}_j \Delta P_k \|_2^2 / |\lambda_{r+1} - \lambda_k| \tag{A.2}
\]
We introduce the operators
\[
P_{J^c} = \sum_{\lambda \notin J} P_k \quad \text{and} \quad T_{J^c} = \sum_{\lambda \notin J} |\lambda_{r+1} - \lambda_k|^{1/2} P_k,
\]
which satisfy the identities \( T_{J^c} S_{J^c} = P_{J^c} \) and
\[
\| \hat{P}_j T_{J^c} \|_2^2 = \sum_{\lambda \notin J} |\lambda_{r+1} - \lambda_k| \| \hat{P}_j P_k \|_2^2.
\]
Hence, (A.2) can be written as
\[
\| \hat{P}_j T_{J^c} \|_2^2 \leq 16 \| \hat{P}_j \Delta S_{J^c} \|_2^2. \tag{A.3}
\]
On the event \( \{ \| S_{J^c} \Delta S_{J^c} \|_{\infty} \leq 1/8 \} \), the last term is bounded via
\[
\| \hat{P}_j E S_{J^c} \|_2^2 \leq 2 \| \hat{P}_j P_{J^c} \Delta S_{J^c} \|_2^2 + 2 \| \hat{P}_j P_{J^c} \Delta S_{J^c} \|_2^2 \\
= 2 \| \hat{P}_j P_{J^c} \Delta S_{J^c} \|_2^2 + 2 \| \hat{P}_j T_{J^c} \Delta S_{J^c} \|_2^2 \\
\leq 2 \| P_{J^c} \Delta S_{J^c} \|_2^2 + 2 \| \hat{P}_j T_{J^c} \|_2^2 \| S_{J^c} \Delta S_{J^c} \|_{\infty}^2 \\
\leq 2 \| P_{J^c} \Delta S_{J^c} \|_2^2 + \| \hat{P}_j T_{J^c} \|_2^2 / 32. \tag{A.4}
\]
Inserting (A.4) into (A.3), we get
\[
\| \hat{P}_j T_{J^c} \|_2^2 \leq 32 \| P_{J^c} E S_{J^c} \|_2^2 + \| \hat{P}_j T_{J^c} \|_2^2 / 2,
\]
and the claim follows. \( \square \)
Lemma 4. Grant Assumption 1. Let $E = \bigcap_{j \in J, k \notin J} \{ |\hat{\lambda}_j - \lambda_k| \geq |\lambda_j - \lambda_k|/2 \} \cap \{ \|S_{J^c} \Delta S_{J^c}\|_\infty \leq 1/8 \}$. If (3.7) holds, then we have

$$P(E^c) \leq (2(s - r) + 1) \exp \left( - cn \min \left( \frac{(\lambda_r - \lambda_{r+1})^2}{\lambda_r^2}, \frac{(\lambda_s - \lambda_{s+1})^2}{\lambda_s^2} \right) \right)$$

with a constant $c > 0$ depending only on $C_1$.

Proof. Proceeding as in [15, Lemma 9], we have

$$P(\|S_{J^c} \Delta S_{J^c}\|_\infty > 1/8) \leq \exp \left( - n \min \left( \frac{(\lambda_r - \lambda_{r+1})^2}{64C_3^2 \lambda_r^2}, \frac{(\lambda_s - \lambda_{s+1})^2}{64C_3^2 \lambda_s^2} \right) \right),$$

provided that

$$\max \left( \frac{\lambda_r}{\lambda_r - \lambda_{r+1}}, \frac{\lambda_{s+1}}{\lambda_s - \lambda_{s+1}} \right) \left( \sum_{j \leq r} \frac{\lambda_j}{\lambda_j - \lambda_{r+1}} + \sum_{k > s} \frac{\lambda_k}{\lambda_s - \lambda_k} \right) \leq \frac{n}{64C_3^2},$$

where $C_3$ is the constant from [15, Equation (3.16)] depending only on $C_1$. The latter is satisfied under (3.7) if $c \leq 1/(64C_3^2)$. We turn to the empirical eigenvalues. First, the condition that $|\hat{\lambda}_j - \lambda_k| \geq |\lambda_j - \lambda_k|/2$ for every $j \in J, k \notin J$ is implied by the condition that $\hat{\lambda}_j - \lambda_k \geq (\lambda_j - \lambda_k)/2$ for every $j \in J, k > s$ and $\lambda_k - \hat{\lambda}_j \geq (\lambda_k - \lambda_j)/2$ for every $j \in J, k \leq r$. It is simple to check that the latter condition is equivalent to $-(\lambda_j - \lambda_{s+1})/2 \leq \hat{\lambda}_j - \lambda_j \leq (\lambda_r - \lambda_j)/2$ for every $j \in J$. By [15, Corollary 3.14], we have for $j \in J$,

$$P(\hat{\lambda}_j - \lambda_j < -(\lambda_j - \lambda_{s+1})/2) \leq \exp \left( - \frac{n(\lambda_j - \lambda_{s+1})^2}{4C_3^2 \lambda_j^2} \right) \leq \exp \left( - \frac{n(\lambda_s - \lambda_{s+1})^2}{4C_3^2 \lambda_s^2} \right),$$

provided that

$$\frac{\lambda_j}{\lambda_j - \lambda_{s+1}} \sum_{k \leq j} \frac{\lambda_k}{\lambda_k - \lambda_{s+1}} \leq \frac{n}{4C_3^2}.$$

The latter condition is strongest for $j = s$ and the resulting condition is implied by (3.7). Similarly, by [15, Corollary 3.12] we have for $j \in J$,

$$P(\hat{\lambda}_j - \lambda_j > (\lambda_r - \lambda_j)/2) \leq \exp \left( - \frac{n(\lambda_r - \lambda_j)^2}{4C_3^2 \lambda_r^2} \right) \leq \exp \left( - \frac{n(\lambda_r - \lambda_{r+1})^2}{4C_3^2 \lambda_r^2} \right),$$

provided that

$$\frac{\lambda_r}{\lambda_r - \lambda_j} \sum_{k > j} \frac{\lambda_k}{\lambda_r - \lambda_k} \leq \frac{n}{4C_3^2}.$
The latter condition is strongest for $j = r + 1$ and the resulting condition is implied by (3.7). The claim now follows from the above concentration inequalities and the union bound.

$\square$

End of proof of Theorem 4. By Lemma 3 and the bound $\mathbb{E}\|P_j \Delta P_k\|_2^2 \leq 16C_1^2$ (cf. [15, Section 2.1]), we get

\[
\mathbb{E} \sum_{k \notin J} |\lambda_{r+1} - \lambda_k| \|\hat{P}_j P_k\|_2^2 \leq 1024C_1^2 \sum_{j \in J} \sum_{k \notin J} \frac{\lambda_j \lambda_k}{|\lambda_{r+1} - \lambda_k|} + \lambda_1 (s - r)\mathbb{P}(\mathcal{E}^c)
\]

and the first claim follows from Lemma 4. Inserting (3.7) into the first claim gives (3.8). $\square$