ESAIM: PS 23 (2019) 662–671
https://doi.org/10.1051/ps/2019002

ESAIM: Probability and Statistics
www.esaim-ps.org

WALD STATISTICS IN HIGH-DIMENSIONAL PCA*

MATTHIAS LÖFFLER**

Abstract. In this study, we consider PCA for Gaussian observations $X_1, \ldots, X_n$ with covariance $\Sigma = \sum_i \lambda_i P_i$ in the 'effective rank' setting with model complexity governed by $\mathbf{r}(\Sigma) := \mathrm{tr}(\Sigma)/\|\Sigma\|$. We prove a Berry-Essen type bound for a Wald Statistic of the spectral projector $\hat{P}_r$. This can be used to construct non-asymptotic goodness of fit tests and confidence ellipsoids for spectral projectors P_r. Using higher order pertubation theory we are able to show that our Theorem remains valid even when $\mathbf{r}(\Sigma) \gg \sqrt{n}$.

Mathematics Subject Classification. 62H25, 62G20, 62F25.

Received September 9, 2018. Accepted January 10, 2019.

1. INTRODUCTION

Principal component analysis (PCA) is a widely used dimension-reduction technique in statistics. In the traditional Gaussian setting going back to Anderson [2] one observes n *i.i.d.* zero mean Gaussian random variables with covariance matrix Σ. In more recent years mainly three working assumptions on Σ have been considered: The *spiked covariance model*, the *spiked sparse covariance model* and the *'effective rank'* setting.

Johnstone [8] introduced the spiked covariance model in which Σ is given by

$$\Sigma = \sum_{i=1}^{l} s_i \theta_i \theta_i^T + \sigma^2 I_p. \tag{1.1}$$

Subsequent work in this model has mainly been focused on the failures of PCA in high-dimensions when the dimension $p \to \infty$ and $p/n \to const.$ [9, 15, 18, 25].

A remedy is to assume that the leading eigenvectors θ_i in (1.1) are sparse, enabling thus inference even when $p/n \to \infty$ [4, 5, 7, 22, 24].

We will consider the effective rank setting [10–12, 14, 16, 19]. Here no assumptions on the particular structure of Σ are made, except that the effective rank $\mathbf{r}(\Sigma) := \mathrm{tr}(\Sigma)/\|\Sigma\| = o(n)$ where $\mathrm{tr}(\cdot)$ denotes the trace and $\|\cdot\|$ the operator norm. This allows for a wider range of models, for example Σ with a polynomial or exponential decay of the eigenvalues [19].

*Supported by ERC grant UQMSI/647812 and EPSRC grant EP/L016516/1.

Keywords and phrases: PCA, spectral projectors, central limit theorem, confidence sets, goodness of fit tests.

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, CB3 0WB Cambridge, UK.

** Corresponding author: m.loffler@statslab.cam.ac.uk

Rates of convergence and limiting results for empirical spectral projectors and empirical eigenvectors were proven in [10]. Building upon this [14] proved that functionals of a de-biased empirical eigenvector attain the semi-parametric efficiency bound. Here we consider inference for a a whole spectral projector P_r, treating the rest of Σ as a nuisance parameter. Traditionally, in the (confirmatory) factor model literature [1, 3, 17, 21] goodness of fit tests for the whole covariance matrix Σ are performed. Inference for P_r can be seen as a complement to this, allowing for more detailed testing of the eigenspaces of Σ.

One possibility to construct such tests is to obtain the limiting distribution for $\|\hat{P}_r - P_r\|_F^2$ and related quantities where the asymptotic distribution does not depend on unknown parameters. This approach has been pursued in the two papers by Koltchinskii and Lounici [12, 13]. However, their approach requires sample splitting into three samples and the assumption that $\|\Sigma\|_F^2 \to \infty$. The latter is not necessarily fulfilled, for example in a decaying eigenvector model where the eigenvalues λ_i scale as $i^{-\alpha}$, $\alpha > 1/2$. Others approaches based on the bootstrap and Bayesian inference, respectively, were proposed by [16, 20] but require at least $\mathbf{r}(\Sigma) = o(n^{1/3})$ and do not deal with the harder case $\mathbf{r}(\Sigma) \gg \sqrt{n}$ where one needs to account for bias.

In this note we present a third possibility based on Wald Statistics for constructing goodness of fit tests and confidence ellipsoids for P_r. We show that when $p \to \infty$ a further normalized Wald statistic of $\hat{P}_r - P_r$ is asymptotically Gaussian.

Our main contribution is that we show how to deal with the critical case $\mathbf{r}(\Sigma) \gg \sqrt{n}$ by using second order perturbation theory, requiring for example in the spiked covariance model (1.1) that $\mathbf{r}(\Sigma) = o(n^{2/3})$.

1.1. Set-up & notation

For matrices A, B, C we define the Kronecker product as $(A \otimes B)C = ACB^T$ and the Frobenius inner product $\langle A, B\rangle := \operatorname{tr}(A^T B)$. $\|\cdot\|_F$ denotes the Frobenius-norm. The notation $\|\cdot\|$ will be used for the operator norm and in slight abuse of notation for the Euclidean norm of vectors with corresponding Euclidean inner product $\|v\| := \langle v, v\rangle := v^T v$.

We will frequently use the following convention throughout the paper: for nonnegative a, b $a \lesssim b$ means that there exists a constant C not depending on n or $\mathbf{r}(\Sigma)$ such that $a \le Cb$.

We assume that $X_1, \dots, X_n$ are *i.i.d* centred Gaussian vectors in $\mathbb{R}^p$ with $\mathbb{E}\|X\|^2 < \infty$. We denote by $\Sigma = \mathbb{E}X_1X_1^T$ the covariance matrix of the observations $X_1, \dots, X_n$ and we denote the empirical covariance matrix by

$$\hat{\Sigma} =: \frac{1}{n}\sum_{j=1}^{n} X_j X_j^T. \tag{1.2}$$

We define the *effective rank*: $\mathbf{r}(\Sigma) := \operatorname{tr}(\Sigma)/\|\Sigma\|$.

Since Σ is symmetric and positive semidefinite it has spectral decomposition $\Sigma = \sum_s \lambda_s P_s$ where λ_s are distinct strictly positive, descending eigenvalues and P_s are the corresponding spectral projectors. Let μ_j denote the eigenvalues of Σ arranged in a non-increasing order and repeated with their multiplicities. Define $\Delta_r := \{j : \mu_j = \lambda_r\}$. We denote by $\bar{g}_r := \min(\lambda_{r-1} - \lambda_r, \lambda_r - \lambda_{r+1})$ the spectral gap of the eigenvalue λ_r with the convention that $\lambda_0 = \infty$. By $\lambda_{\min}$ we denote the smallest eigenvalue of Σ and likewise by $\hat{\lambda}_{\min}$ the smallest eigenvalue of $\hat{\Sigma}$. Assuming that $\|\hat{\Sigma} - \Sigma\| < \min_{1\le s\le r} \bar{g}_s/4$ Lemma 2.2 in [14] proves that Δ_r can be identified on an event of high probability and hence it suffices to assume that Δ_r is known.

We thus denote by $\hat{P}_r$ the spectral projector corresponding to the eigenvalues $\{\mu_j(\hat{\Sigma}), j \in \Delta_r\}$ and by $\hat{\lambda}_r$ one arbitrary chosen eigenvalue from the same set. If $\operatorname{card}(\Delta_r) := m_r = 1$ we define $\hat{\theta}_r$ to be a sample eigenvector belonging to the eigenvalue $\hat{\lambda}_r$.

2. Main result

Wald statistics [23] are commonly used when the dimension of the parameter space is $p = const$. The Fisher information for the model $X \sim N(0, \Sigma)$ is $\mathbb{I}(\Sigma) = \frac{1}{2}(\Sigma^{-1} \otimes \Sigma^{-1})$ (*e.g.* [6]). If p is constant the maximum likelihood estimator $\hat{\Sigma}$ for n *i.i.d.* observations is asymptotically Gaussian distributed with $\sqrt{n}$-rate and limiting covariance $\mathbb{I}(\Sigma)^{-1} = 2(\Sigma \otimes \Sigma)$. Applying the delta method to $g(\Sigma) := P_r$ shows that $g(\hat{\Sigma})$ is asymptotically Gaussian, too, and has limiting covariance

$$\mathbb{I}(P_r)^{-1} := 2\sum_{s\neq r} (P_s \otimes P_r + P_r \otimes P_s) \frac{\lambda_r \lambda_s}{(\lambda_r - \lambda_s)^2} = 2(\Sigma C_r^2 \otimes P_r \Sigma + \Sigma P_r \otimes C_r^2 \Sigma), \tag{2.1}$$

where we define the resolvent operator $C_r := \sum_{s\neq r} \frac{P_s}{\lambda_r - \lambda_s}$.

If p remains fixed and $\operatorname{rank}(P_r) = m_r$ it follows that the Wald statistic below converges to a χ^2-distributed random variable,

$$n\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2 \xrightarrow{d} \chi^2_{m_r(p-m_r)},$$

where we denote

$$\mathbb{I}(P_r)^{1/2} = \frac{1}{\sqrt{2}} \left(\Sigma^{-1/2} C_r^{-1} \otimes P_r \Sigma^{-1/2} + \Sigma^{-1/2} P_r \otimes C_r^{-1} \Sigma^{-1/2} \right) \tag{2.2}$$

and, slightly abusing notation, $C_r^{-1} := \lambda_r I - \Sigma$.

In the high-dimensional regime with $p \to \infty$ the test statistic above is stochastically unbounded and thus the χ^2-approximation becomes invalid. Hence one has to further normalize, eventually obtaining a Gaussian limit instead. Moreover, higher order terms do not simply vanish anymore as $n \to \infty$. Particularly, when applying $\mathbb{I}(P_r)^{1/2}$ to $(\hat{P}_r - P_r)$ one has to multiply with $\Sigma^{-1/2}$ and thus the size of higher order error terms depends on the smallest eigenvalue of Σ which we denote by $\lambda_{\min}$.

Lemma 2.1. *Suppose that* $\mathbf{r}(\Sigma) = o(n)$ *and that* $\mathbb{E}\|\hat{\Sigma} - \Sigma\| \leq (1-\gamma)\bar{g}_r/2$ *for some* $\gamma \in (0,1)$. *For* w_i *denoting i.i.d. standard Gaussian random variables we have that,*

$$\frac{n\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2 - m_r(p-m_r)}{\sqrt{2m_r(p-m_r)}} \overset{d}{=} \frac{\sum_{i=1}^{m_r(p-m_r)}(w_i^2 - 1)}{\sqrt{2m_r(p-m_r)}} + Z_n, \tag{2.3}$$

where Z_n *fulfills with probability at least* $1 - e^{-t}$ *for every* $1 \leq t \leq n$

$$\begin{aligned} |Z_n| \leq & C(\gamma, \bar{g}_r, m_r, \|\Sigma\|, \lambda_r) \cdot \\ & \left(\frac{t}{\sqrt{n}} \bigvee \frac{(\mathbf{r}(\Sigma) \vee t)\sqrt{t}}{\lambda_{\min}\sqrt{np}} \bigvee \sqrt{\frac{p}{n}}\sqrt{t} \bigvee \frac{t^{3/2}}{\sqrt{np}} \bigvee \frac{\sqrt{p}\mathbf{r}(\Sigma)}{n} \bigvee \frac{\mathbf{r}(\Sigma)^2}{n\sqrt{p}\lambda_{\min}} \right). \end{aligned} \tag{2.4}$$

The bounds on $|\mathbb{E}\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2 - m_r(p-m_r)|$ obtained in the proof of Lemma 2.1 are sharp in their dependence on p and $\mathbf{r}(\Sigma)$. Particularly this implies that without a further de-biasing step akin to the procedure in [13] it is impossible to improve the dependence on p and $\mathbf{r}(\Sigma)$ in (2.4).

In principle Lemma 2.1 could be used to construct confidence sets and tests for P_r. However, in statistical applications $\mathbb{I}(P_r)$ is usually not known and one needs to estimate it from the data.

For this we use the plug-in estimator given by

$$\hat{\mathbb{I}}(\hat{P}_r)^{1/2} := \frac{1}{\sqrt{2}} \left[\hat{\Sigma}^{-1/2} \hat{C}_r^{-1} \otimes \hat{P}_r \hat{\Sigma}^{-1/2} + \hat{\Sigma}^{-1/2} \hat{P}_r \otimes \hat{C}_r^{-1} \hat{\Sigma}^{-1/2} \right], \tag{2.5}$$

where we denote $\hat{C}_r^{-1} = \hat{\lambda}_r I - \hat{\Sigma}$. To be able to consistently estimate $\lambda_{\min}$ we need to assume that it is of larger order than $\|\hat{\Sigma} - \Sigma\| \asymp \sqrt{\mathbf{r}(\Sigma)/n}$. Our main result is then following Berry-Essen type Theorem.

Theorem 2.2. *Suppose that $\boldsymbol{r}(\Sigma) = o(n)$, that $\mathbb{E}\|\hat{\Sigma} - \Sigma\| \leq (1-\gamma)\bar{g}_r/2$ for some $\gamma \in (0,1)$, that $\bar{g}_r > \bar{c}$ for some constant $\bar{c} > 0$ large enough and that for another large enough constant $c > 0$*

$$\lambda_{\min} \geq c\sqrt{\frac{\mathbf{r}(\Sigma) \bigvee \log(p)}{n}}. \tag{2.6}$$

Then, for Φ denoting the distribution function of a standard Gaussian random variable we have that

$$\begin{aligned}
&\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{n\|\hat{\mathbb{I}}(\hat{P}_r)^{1/2}(\hat{P}_r - P_r)\|_F^2 - m_r(p-m_r)}{\sqrt{2m_r(p-m_r)}} \leq x \right) - \Phi(x) \right| \\
&\leq C(\gamma, \bar{g}_r, m_r, \|\Sigma\|, \lambda_r)\left[\frac{1}{\sqrt{p}} + \frac{\boldsymbol{r}(\Sigma)^2 \vee \log(p)^2}{n\sqrt{p}\lambda_{\min}^3} + \sqrt{\frac{p\log(p)}{n}} \right. \\
&\left. + \frac{(\mathbf{r}(\Sigma) \vee \log(p))\sqrt{\log(p)}}{\lambda_{\min}\sqrt{np}} + \frac{\sqrt{p}\boldsymbol{r}(\Sigma)}{n} \right].
\end{aligned}$$

All quantities except P_r in the Wald statistic above are known or, as in the case of m_r, can assumed to be known (see [14], Lem. 2.2). This allows the construction of goodness of fit tests and confidence ellipsoids for P_r.

Considering the spiked covariance model (1.1) we have that $\mathbf{r}(\Sigma) \asymp p$ and $\lambda_{\min} \asymp 1$ and thus one can meaningfully apply Theorem 2.2 if $\mathbf{r}(\Sigma) = o(n^{2/3})$.

If $\lambda_{\min}$ shrinks to 0 the sample size requirements are becoming worse. For example, for models with decaying eigenvalues such that $\lambda_i \asymp i^{-\alpha}$, $0 \leq \alpha < 1$ we have that $\mathbf{r}(\Sigma) \asymp p^{1-\alpha}$ and $\lambda_{\min} \asymp p^{-\alpha}$. Therefore the application of Theorem 2.2 is feasible if $p^{3/2+\alpha} = o(n)$.

In case of the spiked covariance model the conditions of Theorem 2.2 compare favorably to the bootstrap approach used by [16] who need to assume that $\mathbf{r}(\Sigma) = o(n^{1/3})$. Moreover, for models with decaying eigenvalues with $\alpha < 3/8$ their condition is worse than our requirement $p^{3/2+\alpha} = o(n)$ whereas for $\alpha > 3/8$ their condition is better.

The construction proposed in [12, 13] requires no assumption on $\lambda_{\min}$ and allows for $\mathbf{r}(\Sigma) = o(n)$, but instead relies on sample splitting into three samples, assumes that $m_r = 1$ and that $\|\Sigma\|_F^2 \to \infty$. The last condition makes their Theorem unfeasible for application to models with quickly decaying eigenvalues $\lambda_i \asymp i^{-\alpha}$, $\alpha > 1/2$.

3. Proofs

We first collect a few results from [10, 11] which we will frequently use throughout our proof. The first Lemma is a perturbation bound for spectral projectors proven in [10].

Lemma 3.1. *Suppose that $\tilde{\Sigma} = \Sigma + E$. Let $\tilde{P}_r$ be the spectral projector corresponding to the eigenvalues $\{\mu_j(\tilde{\Sigma}),\ j \in \Delta_r\}$. Then the following bound holds*

$$\|\tilde{P}_r - P_r\| \leq 4\frac{\|E\|}{\bar{g}_r}. \tag{3.1}$$

Moreover,

$$\tilde{P}_r - P_r = L_r(E) + S_r(E) \tag{3.2}$$

where $L_r(E) = C_r E P_r + P_r E C_r$ *and where* C_r *denotes the resolvent operator*

$$C_r = \sum_{s \neq r} \frac{P_s}{\lambda_r - \lambda_s} \tag{3.3}$$

and where the remainder term can be bounded

$$\|S_r(E)\| \leq 14 \left(\frac{\|E\|}{\bar{g}_r} \right)^2 . \tag{3.4}$$

In the course of our proofs we will also need a finer analysis of the non-linear term $S_r(E)$.

Lemma 3.2. *The following bound holds*

$$S_r(E) = Z_r(E) + R_r(E) \tag{3.5}$$

where

$$Z_r(E) = P_r E C_r E C_r + C_r E C_r E P_r + C_r E P_r E C_r - P_r E P_r E C_r^2 - P_r E C_r^2 E P_r - C_r^2 E P_r E P_r \tag{3.6}$$

and where the third order remainder term is symmetric and fulfills

$$\|R_r(E)\| \leq 72 \left(\frac{\|E\|}{\bar{g}_r} \right)^3 . \tag{3.7}$$

Proof. The first part and symmetry of $R_r(E)$ follows immediately by inspecting the proof of Lemma 5 in [13]. Moreover,

$$R_r(E) = -\frac{1}{2\pi i} \oint_{\gamma_r} \sum_{k \geq 3} (-1)^k (R_\Sigma(\eta) E)^k R_\Sigma(\eta) \mathrm{d}\eta,$$

where γ_r denotes the circle of radius $\bar{g}_r/2$ centred at λ_r with counterclockwise orientation and $R_\Sigma(\eta)$ denotes the resolvent of P_r, *i.e.*

$$R_\Sigma(\eta) = \sum_{j \geq 1} \frac{P_j}{\mu_j - \eta}.$$

Assume first that $\|E\| \leq \bar{g}_r/3$. Then we have that

$$\begin{aligned} \left\| -\frac{1}{2\pi i} \oint_{\gamma_r} \sum_{k \geq 3} (-1)^k (R_\Sigma(\eta) E)^k R_\Sigma(\eta) \mathrm{d}\eta \right\| &\leq 2\pi \frac{\bar{g}_r}{2} \left(\frac{2}{\bar{g}_r} \right)^4 \|E\|^3 \sum_{k=0}^{\infty} \left(\frac{2\|E\|}{\bar{g}_r} \right)^k \\ &\leq 24 \left(\frac{\|E\|}{\bar{g}_r} \right)^3 . \end{aligned}$$

If $\|E\| \geq \bar{g}_r/3$ then by Lemma 3.1 and the explicit representation of the second order pertubation term in Lemma 5 in [13] we obtain that

$$\begin{aligned}
&\left\|-\frac{1}{2\pi i}\oint_{\gamma_r}\sum_{k\geq 3}(-1)^k(R_\Sigma(\eta)E)^k R_\Sigma(\eta)\mathrm{d}\eta\right\| \\
&\leq \|\hat{P}_r - P_r\| + \|L_r(E)\| + \left\|-\frac{1}{2\pi i}\oint_{\gamma_r}(R_\Sigma(\eta)E)^2 R_\Sigma(\eta)\mathrm{d}\eta\right\| \\
&\leq 4\frac{\|E\|}{\bar{g}_r} + \frac{2\|E\|}{\bar{g}_r} + \frac{6\|E\|^2}{\bar{g}_r} \leq 72\left(\frac{\|E\|}{\bar{g}_r}\right)^3
\end{aligned}$$

□

To bound $\|\hat{\Sigma} - \Sigma\|$ we will frequently use the following bound and concentration inequality obtained by Koltchinskii and Lounici in [11].

Theorem 3.3. *Let $X_1, \ldots, X_n$ be i.i.d. centred Gaussian random vectors with covariance matrix Σ and such that $\mathbb{E}\|X_1\|^2 < \infty$. Suppose that $\boldsymbol{r}(\Sigma) = o(n)$. Then, for some constant $C_q > 0$*

$$\left(\mathbb{E}\|\hat{\Sigma} - \Sigma\|^q\right)^{1/q} \leq C_q\|\Sigma\|\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}}. \tag{3.8}$$

Moreover, there exists another constant $C' > 0$ such that for all $t \geq 1$ with probability at least $1 - e^{-t}$ we have that,

$$\left|\|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\|\right| \leq C'\|\Sigma\|\left(\sqrt{\frac{t}{n}}\bigvee\frac{t}{n}\right). \tag{3.9}$$

In the following we denote $E = \hat{\Sigma} - \Sigma$, $L_r := L_r(E)$, $S_r := S_r(E)$, $Z_r := Z_r(E)$ and $R_r := R_r(E)$. We now turn to the proof of Lemma 2.1.

Proof of Lemma 2.1. Going line by line through the proofs of Lemma 5, Theorem 5 and the calculations leading to display (5.17) in [12] it is easy to see that one can adjust them to show

$$\frac{n\left[\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2 - \mathbb{E}\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2\right]}{\sqrt{2m_r(p-m_r)}} \overset{d}{=} \frac{\sum_{i=1}^{m_r(p-m_r)}(w_i^2 - 1)}{\sqrt{2m_r(p-m_r)}} + Z_n', \tag{3.10}$$

where Z_n' fulfills with probability at least $1 - e^{-t}$ for every $1 \leq t \leq n$

$$|Z_n'| \leq C'(\gamma, \bar{g}_r, m_r, \|\Sigma\|, \lambda_r)\left(\frac{t}{\sqrt{n}}\bigvee\frac{(\mathbf{r}(\Sigma)\vee t)\sqrt{t}}{\lambda_{\min}\sqrt{np}}\bigvee\sqrt{\frac{p}{n}}\sqrt{t}\bigvee\frac{t^{3/2}}{\sqrt{np}}\right). \tag{3.11}$$

Thus it remains to obtain a tight bound for $\mathbb{E}\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2$. Using decomposition (3.2) we obtain that

$$\mathbb{E}\|\mathbb{I}(P_r)^{1/2}(\hat{P}_r - P_r)\|_F^2 = \|\mathbb{I}(P_r)^{1/2}L_r\|_F^2 + \|\mathbb{I}(P_r)^{1/2}S_r\|_F^2 + \operatorname{tr}\left((\mathbb{I}(P_r)^{1/2}L_r)(\mathbb{I}(P_r)^{1/2}S_r)\right). \tag{3.12}$$

As in the proof of Theorem 5 in [12] we obtain that

$$n\mathbb{E}\|\mathbb{I}(P_r)^{1/2}L_r\|_F^2 = n\mathbb{E}\|P_r^{\perp}(\Sigma)^{-1/2}E(\Sigma)^{-1/2}P_r\|_F^2 = m_r(p-m_r).$$

Moreover, the second term in (3.12) can be bound by applying Lemma 3.1 and Theorem 3.3 as follows

$$\mathbb{E}\|\mathbb{I}(P_r)^{1/2}S_r\|_F^2 \le 2m_r\mathbb{E}\left\|C_r^{-1}\Sigma^{-1/2}S_rP_r\Sigma^{-1/2}\right\|^2 \lesssim \frac{2m_r\|\Sigma\|^2}{\bar{g}_r^4\lambda_r\lambda_{\min}}\mathbb{E}\|E\|^4 = O\left(\frac{\mathbf{r}(\Sigma)^2}{n^2\lambda_{\min}}\right).$$

For the last remainder term in (3.12) the naive use of Cauchy–Schwarz does not suffice and and we have to use higher-order perturbation theory to obtain good enough bounds. Applying Lemma 3.2 and using symmetry of L_r, R_r, Z_r and $\mathbb{I}(P_r)^{1/2}$ we obtain that,

$$\begin{aligned}\mathbb{E}\left\langle \mathbb{I}(P_r)^{1/2}L_r, \mathbb{I}(P_r)^{1/2}S_r\right\rangle &= \mathbb{E}\left\langle \Sigma^{-1}P_rEP_r^{\perp}\Sigma^{-1}, P_rEC_rEP_r^{\perp} - P_rEP_rEC_r\right\rangle \\ &+ \mathbb{E}\left\langle \Sigma^{-1}P_rEP_r^{\perp}\Sigma^{-1}, P_rR_r(E)C_r^{-1}\right\rangle. \end{aligned} \tag{3.13}$$

We now bound the three terms in (3.13) separately. For $\{\theta_j\}_{j\in\Delta_r}$ denoting the eigenvectors of an eigen-decomposition of P_r $\langle X_1, \theta_j\rangle$ and $P_r^{\perp}X_1$ are independent. We also decompose $\Sigma = \sum \mu_j\theta_j\theta_j^T$. Thus, we obtain that the first term in (3.13) equals

$$\begin{aligned}&\mathbb{E}\left\langle \Sigma^{-1}P_rEP_r^{\perp}\Sigma^{-1}, P_rEC_rEP_r^{\perp}\right\rangle \\ &=\frac{1}{n^2}\sum_{j\in\Delta_r}\operatorname{tr}\left(\Sigma^{-1}P_r^{\perp}X_1X_1^T\theta_j\theta_j^T\Sigma^{-1}\theta_j\theta_j^TX_1X_1^TC_r(X_1X_1^T-\Sigma)P_r^{\perp})\right) \\ &=\frac{m_r}{n^2}\sum_{i\neq l, i,l\notin\Delta_r}\frac{1}{\lambda_i(\lambda_r-\lambda_l)}\mathbb{E}\operatorname{tr}\left[\theta_i\theta_i^TX_1X_1^T\theta_l\theta_l^TX_1X_1^T\theta_i\theta_i^T\right] \\ &\quad+\frac{m_r}{n^2}\sum_{i\notin\Delta_r}\frac{1}{\lambda_i(\lambda_r-\lambda_i)}\mathbb{E}\operatorname{tr}\left[\theta_i\theta_i^TX_1X_1^T\theta_i\theta_i^T(X_1X_1^T-\Sigma)\theta_i\theta_i^T\right] \\ &=\frac{m_r}{n^2}\left[\sum_{i\neq l, i,l\notin\Delta_r}\frac{1}{\lambda_i(\lambda_r-\lambda_l)}\mathbb{E}\langle X_1,\theta_i\rangle^2\mathbb{E}\langle\theta_l, X_1\rangle^2 + \sum_{i\notin\Delta_r}\frac{1}{\lambda_i(\lambda_r-\lambda_i)}(\mathbb{E}\langle X_1,\theta_i\rangle^4 - \mathbb{E}\langle X_1,\theta_i\rangle^2\lambda_i)\right] \\ &=\frac{m_r}{n^2}\left[\sum_{i\neq l, i,l\notin\Delta_r}\frac{\lambda_l}{\lambda_r-\lambda_l} + \sum_{i\neq\Delta_r}\frac{2\lambda_i}{\lambda_r-\lambda_i}\right] \lesssim \frac{m_r}{\bar{g}_r}\frac{p\mathbf{r}(\Sigma)}{n^2} \end{aligned} \tag{3.14}$$

The second term in (3.13) can be estimated similarly,

$$\begin{aligned}&\frac{1}{n^2\lambda_r}\sum_{j\in\Delta_r}\mathbb{E}\operatorname{tr}\left(\Sigma^{-1}P_r^{\perp}X_1X_1^T\theta_j\theta_j^T(X_1X_1^T-\Sigma)\theta_j\theta_j^TX_1X_1^TC_r\right) \\ &=\frac{2m_r\lambda_r}{n^2}\sum_{i\notin\Delta_r}\frac{\mathbb{E}\langle\theta_i, X_1\rangle^2}{\lambda_i(\lambda_r-\lambda_i)} \lesssim \frac{m_r\|\Sigma\|}{\bar{g}_r}\frac{p}{n^2}. \end{aligned} \tag{3.15}$$

The last term can be bound using Cauchy Schwarz, Lemma 3.2 and 3.3,

$$
\begin{aligned}
&\mathbb{E}\left\langle \Sigma^{-1}P_rEP_r^{\perp}\Sigma^{-1/2}, P_rR_r(E)C_r^{-1}\right\rangle \\
\le &\frac{\|\Sigma\|\sqrt{m_r}}{\sqrt{\lambda_r\lambda_{\min}}}\sqrt{\mathbb{E}\|P_r\Sigma^{-1/2}E\Sigma^{-1/2}P_r^{\perp}\|_F^2}\sqrt{\mathbb{E}\|R_r(E)\|^2} \\
\lesssim &\frac{\|\Sigma\|m_r}{\bar{g}_r^3\sqrt{\lambda_r\lambda_{\min}}}\sqrt{\frac{p}{n}}(\mathbb{E}\|E\|^6)^{1/2} \lesssim \frac{\|\Sigma\|^4 m_r}{\bar{g}_r^3\sqrt{\lambda_r}}\frac{\sqrt{p}\mathbf{r}(\Sigma)^{3/2}}{n^2\sqrt{\lambda_{\min}}}.
\end{aligned} \tag{3.16}
$$

Thus, summarizing, we have that

$$
\frac{n\mathbb{E}\|\mathbb{I}(P_r)^{-1/2}(\hat{P}_r - P_r)\|_F^2}{\sqrt{2m_r(p-m_r)}} = \frac{m_r(p-m_r)}{\sqrt{2m_r(p-m_r)}} + O\left(\frac{\sqrt{p}\mathbf{r}(\Sigma)}{n} + \frac{\mathbf{r}(\Sigma)^2}{n\sqrt{p}\lambda_{\min}}\right),
$$

and the claim follows. □

We now turn to the proof of Theorem 2.2 and show first that we can replace $\mathbb{I}(P_r)^{1/2}$ by $\hat{\mathbb{I}}(\hat{P}_r)^{1/2}$.

Proof of Theorem 2.2. We have that

$$
\begin{aligned}
&\|(\mathbb{I}(P_r)^{1/2} - \hat{\mathbb{I}}(\hat{P}_r)^{1/2})(\hat{P}_r - P_r)\|_F \\
&\le 2\sqrt{m_r}\|\hat{P}_r - P_r\|\|\hat{C}_r^{-1}\|\|\hat{\Sigma}^{-1/2}\|\left(\left|\frac{1}{\sqrt{\hat{\lambda}_r}} - \frac{1}{\sqrt{\lambda_r}}\right| + \frac{\|\hat{P}_r - P_r\|}{\sqrt{\lambda_r}}\right) \\
&\quad + \frac{2\sqrt{m_r}}{\sqrt{\lambda_r}}\|\hat{P}_r - P_r\|\left(\|\hat{C}_r^{-1}\|\|\hat{\Sigma}^{-1/2} - \Sigma^{-1/2}\| + \|C_r^{-1} - \hat{C}_r^{-1}\|\|\Sigma^{-1/2}\|\right) \\
&=: I + II + III + IV.
\end{aligned} \tag{3.17}
$$

We now bound each of these four terms separately. We have that

$$
\begin{aligned}
I \le &8\sqrt{m_r}\frac{\|E\|}{\bar{g}_r}\|\Sigma + E\||\hat{\lambda}_{\min}|^{-1/2}\sum_{k=1}^{\infty}\frac{|\hat{\lambda}_r - \lambda_r|^k}{\lambda_r^{k+1/2}} \\
\le &16\sqrt{m_r}\frac{\|E\|}{\bar{g}_r}\|\Sigma\|\,|\lambda_{\min} - \|E\||^{-1/2}\sum_{k=1}^{\infty}\frac{\|E\|^k}{\lambda_r^{k+1/2}} \lesssim \frac{\sqrt{m_r}\|\Sigma\|^3}{\bar{g}_r\sqrt{\lambda_r}}\frac{\mathbf{r}(\Sigma)\vee t}{n\sqrt{\lambda_{\min}}},
\end{aligned} \tag{3.18}
$$

with probability at least $1 - e^{-t}$ for $1 \le t \le \log(p)$ and where we used Theorem 3.3 to bound $\|E\|$, Lidski's inequality to bound $|\hat{\lambda}_r - \lambda_r|$ and the $\lambda_{\min}$ condition (2.6) to bound $\lambda_{\min} - \|E\| \ge \lambda_{\min}/2$. The second term can be bounded likewise, *i.e.* on the same event as the bound for I we have with probability at least $1 - e^{-t}$ for $1 \le t \le \log(p)$ that

$$
II \le \frac{32\sqrt{m_r + m_r}}{\sqrt{\lambda_r}}\frac{\|E\|^2}{\bar{g}_r^2}\|\hat{\Sigma}\|\|\hat{\Sigma}^{-1/2}\| \lesssim \frac{\sqrt{m_r}\|\Sigma\|^3}{\bar{g}_r^2\sqrt{\lambda_r}}\frac{\mathbf{r}(\Sigma)\vee t}{n\sqrt{\lambda_{\min}}}. \tag{3.19}
$$

Using matrix series expansions of $\Sigma\hat{\Sigma}^{-1}$ and of $(\Sigma\hat{\Sigma}^{-1})^{1/2}$ around I we can bound the third term on the same event. We have with probability at least $1-e^{-t}$ for $1 \le t \le \log(p)$:

$$\begin{aligned} III &\le \frac{8\sqrt{m_r}\|\hat{\Sigma}\|\|E\|}{\bar{g}_r\sqrt{\lambda_r}}\|\hat{\Sigma}^{-1/2}-\Sigma^{-1/2}\| \\ &\lesssim \frac{\sqrt{m_r}\|\hat{\Sigma}\|\|E\|}{\bar{g}_r\sqrt{\lambda_{\min}\lambda_r}}\sum_{j=1}^{\infty}\left(\sum_{k=1}^{\infty}\|\Sigma^{-1}\|^k\|E\|^k\right)^j \\ &\lesssim \frac{\sqrt{m_r}\|\Sigma\|^3}{\bar{g}_r\sqrt{\lambda_r}}\frac{\mathbf{r}(\Sigma)\vee t}{n\lambda_{\min}^{3/2}}, \end{aligned} \tag{3.20}$$

where we used again the $\lambda_{\min}$ condition (2.6) to ensure convergence of the series. Finally, the fourth term can be bound in the same fashion on the same event. For $1 \le t \le \log(p)$ with probability at least $1-e^{-t}$ we have that

$$IV \le \frac{8\sqrt{2m_r}\|E\|}{\bar{g}_r\sqrt{\lambda_r\lambda_{\min}}}(|\hat{\lambda}_r-\lambda_r|+\|E\|) \lesssim \frac{m_r\|\Sigma\|^2}{\bar{g}_r\sqrt{\lambda_r}}\frac{\mathbf{r}(\Sigma)\vee t}{n\sqrt{\lambda_{\min}}}. \tag{3.21}$$

Thus, summarizing, and since we bounded I, II, III and IV on the same event we have, choosing $t=\log(p)$ with probability at least $1-1/p$ that

$$\frac{n\|(\mathbb{I}(P_r)^{1/2}-\hat{\mathbb{I}}(P_r)^{1/2})(\hat{P}_r-P_r)\|_F^2}{\sqrt{2m_r(p-m_r)}} \le C(\|\Sigma\|,\lambda_r,m_r,\bar{g}_r)\frac{\mathbf{r}(\Sigma)^2\vee\log(p)^2}{n\sqrt{p}\lambda_{\min}^3}. \tag{3.22}$$

Defining for random variables η and ξ $\Delta(\eta,\xi) := \sup_{x\in\mathbb{R}}|\mathbb{P}(\xi\le x)-\mathbb{P}(\eta\le x)|$ the anti-concentration bound in Lemma 4.6. from [14] combined with (3.22) and Lemma 2.1 thus implies that for $\eta := \frac{n\|\hat{\mathbb{I}}(P_r)^{1/2}(\hat{P}_r-P_r)\|_F^2-m_r(p-m_r)}{\sqrt{2m_r(p-m_r)}}$ and Z denoting a standard Gaussian random variable we have that,

$$\begin{aligned} \Delta(\eta,Z) \le &\Delta(\xi,Z)+\frac{1}{p}\bigvee C(\gamma,\bar{g}_r,m_r,\|\Sigma\|,\lambda_r)\left[\frac{\mathbf{r}(\Sigma)^2\vee\log(p)^2}{n\sqrt{p}\lambda_{\min}^3}+\sqrt{\frac{p\log(p)}{n}}\right. \\ &\left.+\frac{(\mathbf{r}(\Sigma)\vee\log(p))\sqrt{\log(p)}}{\lambda_{\min}\sqrt{np}}+\frac{\sqrt{p}\mathbf{r}(\Sigma)}{n}\right], \end{aligned}$$

where the main term ξ is defined as $\sum_{i=1}^{m_r(p-m_r)}(w_i^2-1)/\sqrt{2m_r(p-m_r)}$. Theorem 2.2 now follows from the bound above and the Berry-Essen Theorem applied to $\Delta(\xi,Z)$. □

References

[1] Y. Amemiya and T.W. Anderson, Asymptotic Chi-square tests for a large class of factor analysis models. *Ann. Statist.* **18** (1990) 1453–1463.
[2] T.W. Anderson, Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** (1963) 122–148.
[3] P. Bentler, A testing method for covariance structure analysis. In Vol. 24 of *IMS Lecture Notes Monogr. Ser.*, Inst. Math. Statist., Hayward, CA (1994) 123–136.
[4] Q. Berthet and P. Rigollet, Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** (2013) 1780–1815.
[5] T. Cai, Z. Ma and Y. Wu, Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.* **41** (2013) 3074–3110.
[6] M.L. Eaton, Multivariate Statistics: A Vector Space Approach. John Wiley & Sons, Inc., New York (1983).
[7] C. Gao and H. Zhou, Rate-optimal posterior contraction for sparse PCA. *Ann. Statist.* **43** (2015) 785–818.
[8] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** (2001) 295–327.

[9] I.M. Johnstone and A.Y. Lu, On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc.* **104** (2009) 682–693.
[10] V. Koltchinskii and K. Lounici, Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.* **52** (2016) 1976–2013.
[11] V. Koltchinskii and K. Lounici, Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** (2017) 110–133.
[12] V. Koltchinskii and K. Lounici, Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.* **45** (2017) 121–157.
[13] V. Koltchinskii and K. Lounici, New asymptotic results in principal component analysis. *Sankhya A* **79** (2017) 254–297.
[14] V. Koltchinskii, M. Löffler and R. Nickl, Efficient estimation of linear functionals of principal components. Preprint at ArXiv:1708.07642 (2017).
[15] B. Nadler, Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann. Statist.* **36** (2008) 2791–2817.
[16] A. Naumov, V. Spokoiny and V. Ulyanov, Confidence sets for Spectral projectors of covariance matrices. *Dokl. Math.* **98** (2018) 511–514.
[17] D. Passemier, Z. Li and J. Yao, On estimation of the noise variance in high dimensional probabilistic principal component analysis. *J. R. Statist. Soc. B* **79** (2017) 51–67.
[18] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sin.* **17** (2007) 1617–1642.
[19] M. Reiss and M. Wahl, Non-asymptotic upper bounds for the reconstruction error of PCA. Preprint at ArXiv:1609.03779 (2016).
[20] I. Silin and V. Spokoiny, Bayesian inference for spectral projectors of covariance matrix. *Electr. J. Stat.* **12** (2018) 1948–1987.
[21] J. Sun, Assessing goodness of fit in confirmatory factor analysis. *Meas. Eval. Counsel. Dev.* **37** (2005) 240–256.
[22] V. Vu and J. Lei, Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** (2013) 2905–2947.
[23] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **54** (1943) 426–482.
[24] T. Wang, Q. Berthet and R.J. Samworth, Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.* **44** (2016) 1896–1930.
[25] W. Wang and J. Fan, Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** (2017) 1342–1374.