Notes on "Minimax Rates of Estimation for Sparse PCA in High Dimensions"

Fan Wu

November 11, 2020

In these notes we review the paper "Minimax Rates of Estimation for Sparse PCA in High Dimensions" (Vu and Lei, 2015), which studies the problem of sparse principal component analysis (PCA) in the high-dimensional setting, where the number of variables p can be (much) larger than the number of observations n, and establishes non-asympototic on the minimax bounds on the estimation error for the leading eigenvector, which is assumed to belong to an ℓ_q ball for a $q \in [0, 1]$. Note that the case $q = 0$ corresponds to exact sparsity. Before we begin, we give a brief, non-exhaustive introduction to sparse PCA.

1 Introduction

1.1 Principal component analysis

A good overview of both methodology and theory on sparse PCA is given in (Zou and Xue, 2018). Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a given data matrix, and assume for simplicity that its columns all have mean zero, so that the empirical covariance matrix is given by $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$. The first principal component can be defined as $Z_1 = \sum_{j=1}^{p} \alpha_{1j} X_j$, where X_j denotes the jth row of the data matrix $\mathbf{X}$, and $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{1,p})^T$ maximizes the variance of Z_1, that is

$$\alpha_1 = \arg\max_{\alpha} \alpha^T \hat{\Sigma} \alpha \quad \text{subject to } \|\alpha\|_2 = 1. \tag{1}$$

The subsequent principal components can be defined sequentially by

$$\alpha_k = \arg\max_{\alpha} \alpha^T \hat{\Sigma} \alpha \quad \text{subject to } \|\alpha\|_2 = 1, \alpha^T \alpha_l = 0 \text{ for all } l < k. \tag{2}$$

This definition means that the principal components can be obtained from the singular value decomposition (SVD) of the data matrix $\mathbf{X}$. Let

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where $\mathbf{D} \in \mathbb{R}^{p\times p}$ is a diagonal matrix and $\mathbf{U} \in \mathbb{R}^{n\times p}$ and $\mathbf{V} \in \mathbb{R}^{p\times p}$ are orthonormal matrices. We can assume that the diagonal elements $d_1, \ldots, d_p$ of $\mathbf{D}$ are in descending order. Since the

columns of $\mathbf{V}$ are the eigenvectors of $\hat{\Sigma}$, we see from $\mathbf{XV} = \mathbf{UD}$ that the principal components are given by $Z_k = U_k d_k$, where U_k is the kth column of $\mathbf{U}$.

Geometrically, PCA corresponds to the best low-rank approximation of the observed data. Writing $\mathbf{V}_k = [V_1|\dots|V_k] \in \mathbb{R}^{p\times k}$ for the first k principal components, the projection onto their span is given by the projection operator $\mathbf{P}_k = \mathbf{V}_k\mathbf{V}_k^T$. Equivalently, the first k principal components can be obtained by solving the minimization problem

$$\min_{\mathbf{V}_k} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{V}_k\mathbf{V}_k^T\mathbf{x}_i\|_2^2.$$

1.2 Sparse PCA

Each principal component is a linear combination of *all* p variables, which can be undesirable in terms of interpretability. Further, in the high-dimensional limit, the PCA has been shown to be inconsistent, see e.g. (Baik and Silverstein, 2006). More precisely, assuming that the $X_i \in \mathbb{R}^p$ are drawn independently from some probability distribution with covariance matrix Σ, and $n \to \infty$ with $\frac{p}{n} \to c > 0$, then, we have under some fairly general regularity conditions,

$$\mathbb{P}[\lim_{n\to\infty} |\mathbf{v}_1^T\hat{\mathbf{v}}_1| = 0] = 1,$$

that is the leading eigenvector $\mathbf{v}_1$ of the population covariance matrix Σ and the leading eigenvector $\hat{\mathbf{v}}_1$ of the sample covariance matrix Σ are almost surely asymptotically orthogonal.

Motivated by these shortcomings of PCA, sparse variants of PCA have attracted much attention. Inspired by the Lasso for sparse linear regression, Jolliffe et al. (2003) proposed to solve the variance maximization problem (2) under an additional ℓ_1 constraint on the loading vectors,

$$\alpha_k = \arg\max_{\alpha} \alpha^T\hat{\Sigma}\alpha \quad \text{subject to } \|\alpha\|_2 = 1, \alpha^T\alpha_l = 0 \text{ for all } l < k, \|\alpha\|_1 \le t,$$

for some tuning parameter $t > 0$. Subsequently, many other approaches have been proposed, including an elastic net formulation, a semidefnite programming approach and iterative thresholding methods; see (Zou and Xue, 2018) for more details.

2 "Minimax Rates of Estimation for Sparse PCA in High Dimensions"

The paper (Vu and Lei, 2015) considers the problem of estimating the first principal component. This was later generalized to estimating the subspace spanned by the first k principal components (Vu and Lei, 2013).

2.1 Setting

Assume that the random vectors $X_i \in \mathbb{R}^p$ are independent and from the same distribution with covariance matrix $\Sigma = \mathbb{E}[X_i X_i^T] - (\mathbb{E}[X_i])(\mathbb{E}[X_i])^T$. For the minimax framework, we need to define the parameter space and loss function under consideration.

Parameter space. For $\lambda_1 > \lambda_2 \geq 0$, assume that the population covariance matrix is given by

$$\Sigma = \lambda_1 \theta_1 \theta_1^T + \lambda_2 \Sigma_0, \tag{3}$$

where $\theta_1 \in \mathbb{S}_2^{p-1}$, the unit sphere of ℓ_2, and $\Sigma_0 \in \mathbb{R}^{p \times p}$ is a symmetric matrix satisfying $\Sigma_0 \preceq 0$, $\Sigma_0 \theta_1 = 0$, and $\|\Sigma_0\|_2 = 1$. The last assumption, together with $\lambda_1 > \lambda_2$, implies that the covariance matrix Σ has a unique largest eigenvalue λ_1.

The **sparsity constraint** is modeled by assuming

$$\theta_1 \in \mathbb{B}_q^p(R_q)) \{\theta \in \mathbb{R}^p : \sum_{j=1}^p |\theta_j|^q \leq R_q\}$$

for a $q \in [0, 1]$, where we use the convention $0^0 = 0$. The case $q = 0$ corresponds to "hard" sparsity, i.e. θ can have at most R_0 non-zero entries, while $q > 0$ corresponds to "soft" sparsity, where only a few of the entries of θ are allowed to be large, while the majority are small.

Throughout these notes, we consider the class

$$\mathcal{M}_q(\lambda_1, \lambda_2, \bar{R}_q, \alpha, \kappa),$$

which consists of all probability distributions on X_i with covariance matrix satisfying (3) with $\theta_1 \in \mathbb{B}_q^p(\bar{R}_q + 1)$ and Assumption 2.2.1 below, where α and κ only depend on q.

Loss function. When estimating the leading eigenvector, we have a sign ambiguity, in the sense that we could either estimate θ_1 or $-\theta_1$. More generally, if we were considering estimating the subspace spanned by the first k principal components, say, Θ, then Θ is only unique up to orthogonal transformation, as Θ and ΘV span the same subspace for any orthogonal $k \times k$ matrix V. However, the projection $\Pi = \Theta \Theta^T$ is unique, and hence it makes sense to consider the loss function defined by

$$\|\widehat{\Pi} - \Pi\|_F.$$

In our case we have $k = 1$ and only an sign ambiguity. Still, we consider the same loss function

$$\|\hat{\theta}_1 \hat{\theta}_1^T - \theta \theta^T\|_F.$$

With these definitions in place, the goal is to establish non-asymptotic bounds on the minimax error

$$\min_{\hat{\theta}_1} \max_{P \in \mathcal{M}_q(\lambda_1, \lambda_2, \bar{R}_q, \alpha, \kappa)} \mathbb{E}_P \left[\|\hat{\theta}_1 \hat{\theta}_1^T - \theta_1 \theta_1^T\|_F \right],$$

where the minimum is taken over all estimators that only depend on the data $X_1, \ldots, X_n$.

2.2 Minimax rate - lower bound

Define the quantity

$$\sigma^2 = \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2}.$$

Intuitively, σ^2 relates to the gap between the first two eigenvalues and corresponds to the effective noise-to-signal ratio. The following assumption on $\bar{R}_q$ ensures that the eigenvector θ_1 is not too dense.

Assumption 2.2.1. *There exists a constant $\alpha \in (0, 1]$ that depends only on q satisfying*

$$\bar{R}_q \leq \kappa^q (p-1)^{1-\alpha} \bar{R}_q^{\frac{2\alpha}{2-q}} \left[\frac{\sigma^2}{n} \log \frac{p-1}{\bar{R}_q^{\frac{2}{2-q}}} \right]^{\frac{q}{2}}, \tag{4}$$

where $\kappa \leq c\alpha/16$ is a constant that only depends on q, and

$$1 \leq \bar{R}_q \leq e^{-1}(p-1)^{1-q/2}. \tag{5}$$

Assumption 2.2.1 also ensures that the effective noise-to-signal ratio σ^2 is not too small. Consider the extreme case where $\lambda_2 = 0$. Then, we have $\sigma^2 = 0$, the covariance matrix Σ from (3) becomes a symmetric rank one matrix, and the distribution of X_i lies in a one-dimensional subspace. The relationship between the parameters $n, p, \bar{R}_q$ and σ^2 required by Assumption 2.2.1 describes a regime in which the problem is neither impossible nor trivially easy.

Note that in the case $q = 0$, Assumption 2.2.1 simplifies, as (4) is satisfied for $\alpha = 1$, and we only require

$$1 \leq \bar{R}_0 \leq e^{-1}(p-1).$$

The following Theorem establishes a lower bound on the minimax estimation error.

Theorem 2.2.2. *(Lower bound for sparse PCA) Let $q \in [0, 1]$. If Assumption 2.2.1 holds, then there exists a constant $c > 0$ that only depends on q, such that every estimator $\hat{\theta}_1$ satisfies*

$$\max_{P \in \mathcal{M}_q(\lambda_1, \lambda_2, \bar{R}_q, \alpha, \kappa)} \mathbb{E}_P \left[\|\hat{\theta}_1 \hat{\theta}_1^T - \theta_1 \theta_1^T\|_F \right] \geq c \min \left\{ 1, \bar{R}_q \left[\frac{\sigma^2}{n} \log \left((p-1)/\bar{R}_q^{\frac{2}{2-q}} \right) \right]^{1-\frac{q}{2}} \right\}^{\frac{1}{2}} \tag{6}$$

This bound is easiest to interpret in the case $q = 0$, where R_0 corresponds to the number of active variables. Then, the bound becomes (up to constants)

$$\min \left\{ 1, R_0 \frac{\sigma^2}{n} \log \frac{p}{R_0} \right\}^{\frac{1}{2}}.$$

Here, the 1 means that the error cannot exceed one, as the distance $\|\hat{\theta}_1 \hat{\theta}_1^T - \theta_1 \theta_1^T\|_F$ is related to the sign of the angle between $\hat{\theta}_1$ and θ_1, which cannot exceed 1. The term $R_0 \frac{\sigma^2}{n}$ is the error

if the support of θ_1 was known, and we were able to do PCA just on the support. The term $\log \frac{p}{R_0}$ is the error due to the subset selection we need to do. Note that this bound matches the lower bound for sparse linear regression (with the effective noise level σ^2 defined as above), which means that the sparse PCA problem of estimating the first principal component is statistically as hard as sparse linear regression.

2.3 Minimax rate - upper bound

The idea to establishing the upper bound on the minimax rate consists of analyzing the following ℓ_q-constrained maximization problem,

$$\max_b b^T S b \quad \text{subject to } b \in \mathbb{S}_2^{p-1} \cap \mathbb{B}_q^p(\rho_q), \tag{7}$$

where we write $S = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ for the sample covariance matrix (recall that we assumed the columns of $\mathbf{X}$ to be zero mean, that is $\bar{\mathbf{X}}\bar{\mathbf{X}}^T = \mathbf{0}$). Note that the feasible set is non-empty if $\rho_q \geq 1$, and the ℓ_q constraint is active only when $\rho_q \leq p^{1-\frac{q}{2}}$. For $q = 2$ and $\rho_q = 1$, we recover the standard PCA. The case $q = 1$ corresponds to the Lasso estimator and coincides with the method proposed in (Jolliffe et al., 2003).

To analyze the solution to the maximization problem (7), we need to assume that the tails of the distribution of the data vector are sub-Gaussian, which can be described in terms of the Orlicz ψ_α-norm.

Definition 2.3.1. *For a random variable $Y \in \mathbb{R}$, the Orlicz ψ_α-norm for $\alpha \geq 1$ is defined as*

$$\|Y\|_{\psi_\alpha} = \inf\{c > 0 : \mathbb{E}[\exp(|Y/c|^\alpha)] \leq 2\}.$$

If a random variable has finite ψ_α-norm, then its tails are bounded by $\exp(-Cx^\alpha)$. In particular, the case $\alpha = 2$ corresponds to sub-Gaussian random variables.

Assumption 2.3.2. *There exist i.i.d. random vectors $Z_1, \ldots, Z_n \in \mathbb{R}^p$ such that $\mathbb{E}[Z_i] = \mathbf{0}$, $\mathbb{E}[Z_i Z_i^T] = \mathbf{I}_p$,*

$$X_i = \Sigma^{\frac{1}{2}} Z_i, \quad \text{and} \quad \sup_{x \in \mathbb{S}_2^{p-1}} \|\langle Z_i, x\rangle\|_{\psi_2} \leq K,$$

for a constant $K > 0$.

Under this assumption, the following upper bound holds.

Theorem 2.3.3. *(Upper bound for sparse PCA) Let $\hat{\theta}_1$ be the solution to the ℓ_q-constrained maximization problem (7) with $\rho_q = R_q := \bar{R}_q + 1$, and let $\tilde{\sigma} = \frac{\lambda_1}{\lambda_1 - \lambda_2}$. If the distribution of $(X_1, \ldots, X_n)$ belongs to $\mathcal{M}_q(\lambda_1, \lambda_2, \bar{R}_q, \alpha, \kappa)$ and satisfies Assumptions 2.2.1 and 2.3.2, then there exists a constant $c > 0$ that only depends on K such that the following holds:*

1. *If* $q \in (0,1)$, *then*

$$\mathbb{E}\left[\|\hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T\|_F^2\right] \leq c \min\left\{1, R_q^2\left[\frac{\tilde{\sigma}^2}{n}\log p\right]^{1-\frac{q}{2}}\right\}.$$

2. *If* $q = 1$, *then*

$$\mathbb{E}\left[\|\hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T\|_F^2\right] \leq c \min\left\{1, R_1\left[\frac{\tilde{\sigma}^2}{n}\log \frac{p}{R_1^2}\right]^{1-\frac{q}{2}}\right\}.$$

3. *If* $q = 0$, *then*

$$\mathbb{E}\left[\|\hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T\|_F\right]^2 \leq c \min\left\{1, R_0\frac{\tilde{\sigma}^2}{n}\log \frac{p}{R_0}\right\}.$$

The reason for the different bounds depending on q are the different tools available for controlling empirical processes in ℓ_q balls.

Comparing the bounds of Theorems 2.2.2 and 2.3.3 in the case $q = 0$, we see that the bounds agree up to a factor $\sqrt{\lambda_2/\lambda_1}$. In the cases $q \in (0,1)$ and $q = 1$, an upper bound for the squared error can be obtained using the inequality $\mathbb{E}[Y^2] \geq \mathbb{E}[Y]^2$. Thus, both bounds have the same dependence on p and n for all $q \in [0,1]$, which shows that the bounds are sharp in terms of p and n.

3 Proofs

In this section, we present the proofs of Theorems 2.2.2 and 2.3.3. We will state supporting Lemmas from (Vu and Lei, 2015) without proof. We will write $\langle A, B\rangle = \mathrm{Tr}(A^T B)$ for matrices A and B with compatible dimensions, so that $\|A\|_F^2 = \langle A, A\rangle$. For probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$, we consider the Kullback-Leibler divergence

$$D(\mathbb{P}_1\|\mathbb{P}_2) = \int \log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_2}\right) d\mathbb{P}_1.$$

3.1 Proof of Theorem 2.2.2

The main idea to prove Theorem 2.2.2 is to convert the problem from estimation to testing by discretizing the parameter space, and then to apply Fano's inequality (Lemma 3.1.1 below). Then, we need to find a sufficiently large finite subset of the parameter space, such that the parameters are α-separated under the loss function, but close under the KL divergence of the corresponding probability measures. We will provide the main ingredients and lemmas used, but we will omit the details of the calculations.

Lemma 3.1.1. *(Generalized Fano method) Let $N \geq 1$ be an integer and $\theta_1, \dots, \theta_N \subset \Theta$ index a collecion of probability measures $\mathbb{P}_{\theta_i}$ on a measurable space $(\mathcal{X}, \mathcal{A})$. Let d be a pseudometric on Θ and assume that, for all $i \neq j$,*

$$d(\theta_i, \theta_j) \geq \alpha_N \quad \text{and} \quad D(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta_j}) \leq \beta_N.$$

Then, every $\mathcal{A}$-measurable estimator $\hat{\theta}$ satisfies

$$\max_i \mathbb{E}_{\theta_i}[d(\hat{\theta}, \theta_i)] \geq \frac{\alpha_N}{2} \left(1 - \frac{\beta_N + \log 2}{\log N}\right).$$

The subset of the parameter space which is α-separated under the loss function but close under the KL divergence is given by the following lemma.

Lemma 3.1.2. *(Local packing set) Let $\bar{R}_q = R_q - 1 \geq 1$ and $p \geq 5$. There exists a finite subset $\Theta_\epsilon \subset \mathbb{S}_2^{p-1} \cap \mathbb{B}_q^p(R_q)$ and a universal constant $c > 0$ such that each pair $\theta_1 \neq \theta_2 \in \Theta_\epsilon$ satisfies*

$$\frac{\epsilon}{\sqrt{2}} < \|\theta_1 - \theta_2\|_2 \leq \sqrt{2}\epsilon \quad \text{and} \quad \log|\Theta_\epsilon| \geq \left(\frac{\bar{R}_q}{\epsilon^q}\right)^{\frac{2}{2-q}} \left[\log(p-1) - \log\left(\frac{\bar{R}_q}{\epsilon^q}\right)^{\frac{2}{2-q}}\right].$$

By Lemma A.1.2 of (Vu and Lei, 2015), the first bound in Lemma 3.1.2 implies

$$\frac{\epsilon^2}{2} \leq \|\theta_1\theta_1^T - \theta_2\theta_2^T\|_F^2 \leq 4\epsilon^2.$$

For each $\theta \in \Theta_\epsilon$, we consider the matrix

$$\Sigma_\theta = (\lambda_1 - \lambda_2)\theta\theta^T + \lambda_2 \mathbf{I}_p.$$

Then, we consider probability distributions $\mathbb{P}_\theta$ defined as the n-fold product of $\mathcal{N}(\mathbf{0}, \Sigma_\theta)$, and bound the KL divergence with the following lemma.

Lemma 3.1.3. *Let $x_1, x_2 \in \mathbb{S}_2^{p-1}$, $\lambda_1 > \lambda_2 > 0$,*

$$\Sigma_i = (\lambda_1 - \lambda_2)x_i x_i^T + \lambda_2 \mathbf{I}_p, \quad i = 1, 2,$$

and $\mathbb{P}_i$ be the n-fold product of $\mathcal{N}(\mathbf{0}, \Sigma_i)$. Then, for $\sigma^2 = \lambda_1\lambda_2/(\lambda_1 - \lambda_2)^2$,

$$D(\mathbb{P}_1 \| \mathbb{P}_2) = \frac{n}{2\sigma^2}\|x_1x_1^T - x_2x_2^T\|_F^2.$$

With Lemma 3.1.3, we can upper-bound

$$D(\mathbb{P}_{\theta_1} \| \mathbb{P}_{\theta_2}) = \frac{n}{2\sigma^2}\|\theta_1\theta_1^T - \theta_2\theta_2^T\|_F^2 \leq \frac{2n\epsilon^2}{\sigma^2},$$

and Lemma 3.1.1 implies

$$\max_{\theta \in \Theta_\epsilon} \mathbb{E}_\theta \left[\|\hat{\theta}\hat{\theta}^T - \theta\theta^T\|_F\right] \geq \frac{\epsilon}{2\sqrt{2}} \left(1 - \frac{2n\epsilon^2/\sigma^2 + \log 2}{\log|\Theta_\epsilon|}\right).$$

Finally, Theorem 2.2.2 follows from choosing ϵ on the correct order, which relies crucially on Assumption 2.2.1 and can be shown in a somewhat lengthy calculation, see (Vu and Lei, 2015).

3.2 Proof of Theorem 2.3.3

Lemma 3.2.1. *Let $\theta \in \mathbb{S}_2^{p-1}$. If $\Sigma \preceq 0$ as a unique largest eigenvalue λ_1 with eigenvector θ_1, then*

$$\frac{1}{2}(\lambda_1 - \lambda_2)\|\theta\theta^T - \theta_1\theta_1^T\|_F^2 \leq \langle \Sigma, \theta_1\theta_1^T - \theta\theta^T \rangle.$$

Now consider the solution to the maximization problem (7) $\hat{\theta}_1$. Let $\epsilon = \|\hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T\|_F$ denote the estimation error to be bounded. By Lemma 3.2.1, we can bound

$$\begin{aligned}
\frac{1}{2}(\lambda_1 - \lambda_2)\epsilon^2 &\leq \langle S, \theta_1\theta_1^T \rangle - \langle \Sigma, \hat{\theta}_1\hat{\theta}_1^T \rangle - \langle S - \Sigma, \theta_1\theta_1^T \rangle \\
&\leq \langle S - \Sigma, \hat{\theta}_1\hat{\theta}_1^T \rangle - \langle S_\Sigma, \theta_1\theta_1^T \rangle \\
&= \langle S - \Sigma, \hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T \rangle. \qquad (8)
\end{aligned}$$

We consider the three cases $q \in (0, 1)$, $q = 1$ and $q = 0$.

Case 1: $q \in (0, 1)$. By Hölder's inequality, we can rearrange (8) to

$$\frac{1}{2}\epsilon^2 \leq \frac{\|\operatorname{vec}(S - \Sigma)\|_\infty}{\lambda_1 - \lambda_2}\|\operatorname{vec}(\theta_1\theta_1^T - \hat{\theta}_1\hat{\theta}_1^T)\|_1,$$

where $\operatorname{vec}(A) \in \mathbb{R}^{p^2}$ denotes the vectorized form of a matrix $A \in \mathbb{R}^{p\times p}$. Since we have $\theta_1, \hat{\theta}_1 \in \mathbb{B}_q^p(R_q)$, we can use a truncation trick (see e.g. Lemma 5 of Raskutti et al. (2011)) to bound, for any $t > 0$,

$$\begin{aligned}
\|\operatorname{vec}(\theta_1\theta_1^T - \hat{\theta}_1\hat{\theta}_1^T)\|_1 &\leq \sqrt{2}R_q\|\operatorname{vec}(\theta_1\theta_1^T - \hat{\theta}_1\hat{\theta}_1^T)\|_2 t^{-q/2} + 2R_q^2 t^{1-q} \\
&= \sqrt{2}R_q\epsilon t^{-q/2} + 2R_q^2 t^{1-q}.
\end{aligned}$$

Choosing $t = \|\operatorname{vec}(S - \Sigma)\|_\infty/(\lambda_1 - \lambda_2)$, we have

$$\frac{1}{2}\epsilon^2 \leq \sqrt{2}t^{1-q/2}R_q\epsilon + 2t^{2-q}R_q^2.$$

Define m implicitly via $\epsilon = m\sqrt{2}t^{1-q/s}R_q$. Then, above inequality becomes $m^2/2 \leq m + 1$, which means that we must have $m < 3$. Hence, we have

$$\epsilon \leq 3\sqrt{2}t^{1-q/2}R_q = 3\sqrt{2}R_q\left(\frac{\|\operatorname{vec}(S - \Sigma)\|_\infty}{\lambda_1 - \lambda_2}\right)^{1-q/2}.$$

The next lemma allows us to bound the quantity $\|\operatorname{vec}(S - \Sigma)\|_\infty$.

Lemma 3.2.2. *If Assumption 2.3.2 holds and $\Sigma = \sum_{j=1}^p \lambda_j\theta_j\theta_j^T$ where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, then there is a universal constant $c > 0$ such that*

$$\left\|\|\operatorname{vec}(S - \Sigma)\|_\infty\right\|_{\psi_1} \leq cK^2\lambda_1 \max\left\{\sqrt{\frac{\log p}{n}}, \frac{\log p}{n}\right\}$$

Using Lemma 3.2.2, we can bound

$$\|\epsilon^{2/(2-q)}\|_{\psi_1} \le cK^2 R_q^{2/(2-q)} \tilde{\sigma} \max\left\{\sqrt{\frac{\log p}{n}}, \frac{\log p}{n}\right\}.$$

A straightforward computation shows that $\mathbb{E}[|X|^m] \le (m!)^m \|X\|_{\psi_1}^m$ for $m \ge 1$ (consider the 0th and mth terms of the series expansion of e^x). With this, we have

$$\mathbb{E}[\epsilon^2] \le c_K R_q^2 \tilde{\sigma}^{2-q} \max\left\{\sqrt{\frac{\log p}{n}}, \frac{\log p}{n}\right\}^{2-q} =: \mathbf{M},$$

so that, together with $\epsilon \le 2$, we get

$$\mathbb{E}[\epsilon^2] \le \min\{2, \mathbf{M}\}.$$

The proof for the case $q \in (0,1)$ is completed by the fact that we only need to consider the squareroot term for $\log p < n$. If $\log p > n$, then we have $\mathbb{E}[\epsilon^2] \le 2$.

Case 2: $q = 1$**.** In this case, $\theta_1, \hat{\theta}_1 \in \mathbb{B}_1^p(R_1)$ and we apply the triangle inequality to (8).

$$\begin{aligned}\frac{1}{2}(\lambda_1 - \lambda_2)\epsilon^2 &\le \langle S - \Sigma, \hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T\rangle \\ &\le |\hat{\theta}^T(S-\Sigma)\hat{\theta}_1| + |\theta_1^T(S-\Sigma)\theta_1| \\ &\le 2 \sup_{b \in \mathbb{S}_2^{p-1} \cap \mathbb{B}_1^p(R_1)} |b^T(S-\Sigma)b|\end{aligned}$$

This supremum is bounded in the following lemma.

Lemma 3.2.3. *If Assumption 2.3.2 holds and* $\Sigma = \sum_{j=1}^p \lambda_j \theta_j \theta_j^T$ *where* $\lambda_1 \ge \cdots \ge \lambda_p \ge 0$*, then there is a universal constant* $c > 0$ *such that, for all* $R_1^2 \in [1, p/e]$*,*

$$\mathbb{E}\left[\sup_{b \in \mathbb{S}_2^{p-1} \cap \mathbb{B}_1^p(R_1)} |b^T(S-\Sigma)b|\right] \le c\lambda_1 K^2 \max\left\{R_1\sqrt{\frac{\log(p/R_q^2)}{n}}, R_1^2\frac{\log(p/R_q^2)}{n}\right\}$$

Assumption 2.2.1 guarantees that $R_1^2 \in [1, p/e]$, and Lemma 3.2.3 completes the proof as in the previous case.

Case 3: $q = 0$**.** In this case, $\theta_1, \hat{\theta}_1 \in \mathbb{B}_0^p(R_0)$, so $\theta_1 - \hat{\theta}_1 \in \mathbb{B}_0^p(2R_0)$. Denote by Π the projection matrix onto the union of the supports of θ_1 and $\hat{\theta}_1$, that is Π is a diagonal matrix with a diagonal entry equal 1 where θ_1 or $\hat{\theta}_1$ are non-zero, and zero otherwise. Then, $\Pi\hat{\theta}_1 = \hat{\theta}_1$ and $\Pi\theta_1 = \theta_1$, and by the von Neumann trace inequality, we can bound

$$\begin{aligned}\frac{1}{2}(\lambda_1 - \lambda_2)\epsilon^2 &\le |\langle S - \Sigma, \Pi(\hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T)\Pi\rangle| \\ &= |\langle \Pi(S-\Sigma)\Pi, \hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T\rangle| \\ &\le \sum_{j=1}^p \alpha_j\beta_j,\end{aligned}$$

where by α_j and β_j we denote the singular values of the matrices $\Pi(S-\Sigma)\Pi$ and $\hat{\theta}_1\hat{\theta}_1^T - \theta_1\theta_1^T$ respectively, in descending order. By Lemma A.1.1 of (Vu and Lei, 2015), we can bound the sum of singular values $\sum_{j=1}^p \beta_j \leq \sqrt{2}\epsilon$, so that

$$\frac{1}{2}(\lambda_1 - \lambda_2)\epsilon^2 \leq \|\Pi(S-\Sigma)\Pi\|_2\sqrt{2}\epsilon \leq \sup_{b\in\mathbb{S}_2^{p-1}\cap\mathbb{B}_0^p(2R_0)} |b^T(S-\Sigma)b|\sqrt{2}\epsilon.$$

Similar to the $q=1$ case, the following lemma concludes the proof.

Lemma 3.2.4. *If Assumption 2.3.2 holds and $\Sigma = \sum_{j=1}^p \lambda_j\theta_j\theta_j^T$ where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, then there is a universal constant $c>0$ such that, for all $R_1^2 \in [1, p/e]$,*

$$\mathbb{E}\left[\sup_{b\in\mathbb{S}_2^{p-1}\cap\mathbb{B}_0^p(d)} |b^T(S-\Sigma)b|\right] \leq c\lambda_1 K^2 \max\left\{\sqrt{\frac{d}{n}\log\frac{p}{d}}, \frac{d}{n}\log\frac{p}{d}\right\}$$

References

J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, 97(6):1382–1408, 2006.

I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *J. Comput. Graph. Stat.*, 12(3):531–547, 2003.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

V. Vu and J. Lei. Minimax rates of estimation for sparse pca in high dimensions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1278–1286, 2015.

V. Vu and K. Lei. Minimax sparse principal subspace estimation in high dimensions. *Ann. Stat.*, 41(6):2905–2947, 2013.

H. Zou and L. Xue. A selective overview of sparse principal component analysis. *Proceedings of the IEEE*, 106(8):1311–1320, 2018.