Oracle Inequalities for Sparse Principal Component Analysis based on the Linear Manifold Approximation

Lei Shi, Haoyi Yang, and Lingzhou Xue

May 15, 2023

Abstract

Sparse principal component analysis has been widely used as a powerful method for dimension reduction and feature extraction in high-dimensional data analysis. Due to the nonconvex nature of the problem, many methods only focus on the reconstruction of the principal spaces instead of the individual principal components (PCs). This paper introduces a component-wise sparse PCA scheme based on one classical formulation that extracts PCs in a greedy style. Combined with penalization as well as a delicately designed deflation procedure, the framework adapts well to high-dimensional scenarios. Under mild assumptions like sparsity, the proposed procedure generates a sequence of estimates that proves to be asymptotically consistent with minimax convergence rates. In terms of implementation, an adaptation of the proximal gradient method is applied to tackle the step-wise penalized loss minimization, which yields accurate results with low computation expenses. Numerical results indicate that the proposed scheme is highly competitive among existing methods.

Keywords: Principal component analysis; High Dimension; Asymptotic convergence bounds

1 Introduction

Sparse principal component analysis (PCA) has witnessed a rapid development of methodology and theory as well as a lot of successful real applications in many research areas including human face recognition (Hancock et al., 1996), pandemic forecasting (Mahmoudi et al., 2021), financial marketing (Nobre and Neves, 2019), gene identification (Yano et al., 2019), and so on.

In the statistics and machine learning literature, several sparse PCA frameworks have been developed by exploring and extending different interpretations of the classical PCA (Pearson, 1901) over the past decades. Inspired by the LASSO (Tibshirani, 1996), Jolliffe et al. (2003) proposed the sequential estimation procedure called SCoTLASS to estimate sparse loading vectors by imposing an ℓ_1 constraint. After SCoTLASS, by extending the linear manifold approximation view of Pearson

(1901), Zou et al. (2006) proposed the first computational efficient method that is named SPCA and designed an alternating minimization algorithm to solve the bi-convex formulation, and the recent work by Chen et al. (2020) introduced an alternating manifold proximal gradient method with convergence guarantees to solve SPCA. Lee et al. (2010) and Lu et al. (2016) extended SPCA and proposed sparse exponential-family PCA for any type of data following exponential family distributions. On the other hand, d'Aspremont et al. (2005) proposed a semidefinite programming approach to construct a convex relaxation of the ℓ_0 penalized variance maximization program to estimate sparse loading factors, and Vu et al. (2013); Vu and Lei (2013) generalized the semidefinite programming approach and proposed the Fantope projection and selection to recover the sparse principal eigen-spaces. From the point of low rank matrix approximation, Shen and Huang (2008) proposed an iterative thresholding procedure based on singular value decomposition of the data matrix, and Witten et al. (2009) studied a penalized matrix factorization framework that includes sparse PCA as an example. Moreover, there are many works built up from other perspectives to extend PCA in high dimensions, for example, subset selection (Johnstone and Lu, 2009), matrix factorization (Chen and Wainwright, 2015), matrix decomposition and thresholding (Ma, 2013b), generalized power method (Journée et al., 2010), among others. Please see the recent review paper by Zou and Xue (2018) for more details.

On the other hand, many efforts have been devoted to understanding the theoretical property of the sparse PCA regimes. A thread of works (Baik and Silverstein, 2006; Nadler, 2008; Paul, 2007; Johnstone and Lu, 2009) investigated the statistical property of the classical PCA and critically pointed out its fundamental drawback in increasing dimensions. To this end, Johnstone and Lu (2009) proved the first consistency justification for sparse PCA with their subset selection procedure. Amini and Wainwright (2008) studied the support recovery property of the semi-definite programming approach under the k-sparse assumption for the leading eigenvector in the rank-1 spiked covariance model. Shen et al. (2013); Ma (2013b); Vu et al. (2013) also proved consistency or derived convergence bounds of their estimators under different sparse eigen-structures. Janková and van de Geer (2021) proposed a debiased SPCA scheme to performance inference in a high dimensional setting. In term of minimax optimality, Birnbaum et al. (2013); Cai et al. (2013); Vu and Lei (2013), among others, provided the minimax rates of convergence and adaptive estimation for a variety of models under high dimensional scalings.

Although significant developments have been made in the literature, there are still several missing pieces on the puzzle. One particular drawback is that, many schemes focus on establishing minimax convergence bounds upon eigenspace estimation, which is often characterized by the projection operator. This pursuit typically hides valuable information that we were supposed to draw from the data. For classical PCA people usually highlight the usage and interpretation of the loading matrix, which is formulated by the scaled eigenvectors and carries information about PC and variable importance. See for example Hastie et al. (2009); Bollen et al. (2009); Bonnier and

Byrne (2012). On the other hand, many component-based sparse PCA algorithms(like Zou et al. (2006); Mackey (2009); Shen et al. (2013) build themselves upon biconvex formulations that lead to computational efficiency but are not equipped with convergence rates that match the minimax limits. Therefore, it is of high importance to integrate practical interpretability, theoretical validity and computational efficiency for sparse PCA regimes.

In this paper, as an attempt to handle the above concerns, we propose a component-based sparse PCA estimation scheme based on a step-wise manifold approximation perspective(in the spirit of Hotelling (1933)). Concretely speaking, for a single PC, we adopt the nonconvex projection formulation of PCA and induce sparsity by adding penalization. To proceed from obtained PCs to the next potential PC, we apply a deflation strategy that can temporarily remove the extracted information from the data and avoid accumulative counting along the sequential procedure. As a brief preview of our work, we summarize the following contributions:

- From the methodology aspect, we develop a novel step-wise PC extraction scheme that borrows wisdom from classical PCA formulations as well as achieves adaptation to high dimensional setups. Within each step, it solves a penalized non-convex minimization problem to learn the best sparse one-dimensional projections that captures most information from the data. When proceeding to the next PC, it incorporates a specially designed deflation strategy to avoid the "double counting" issues due to non-orthogonality. This deflation procedure turns out to be crucial for both empirical performance as well as theoretical justification.

- From the theory aspect, we formulate general sufficient conditions that validates our estimation schemes. When these conditions hold, we presente the convergence rates for the extracted eigenvectors and show that with either exact or approximate sparsity, they are able to achieve the minimax-optimal rate (Birnbaum et al., 2013; Cai et al., 2013). Moreover, the proof techniques involved might be of separate interest. More concretely, we establish convergence results by bounding an empirical process that measures the distance between the population and sample quantities. Due to the deflation step, this bound cannot be directly transplanted directly from other popular tricks. To this end, we provide a novel solution for justifying the deflation-based estimation schemes by introducing a pseudo covariance matrix to bridge the population and sample counterparts, which disentangles the complex probabilistic relations and leads to a rigorous induction justification.

The remainder of our paper is structured as follows. In Section 3, we start from an elaboration on the motivation of the current work and introduce the proposed component-wise sparse PCA procedure. Section 4 studies the theoretical asymptotic properties of our estimates, which involves identification of sufficient conditions as well as derivation of the convergence bounds for the procedure. Section 5 presents several synthetic simulation cases to demonstrate the numerical

performance of the proposed method and make comparison with the state-of-arts. Technical proofs and relevant theoretical tools are provided in the Section 6.

Notations. For a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|_2$ are respectively its ℓ_1 and ℓ_2 norm. For a matrix $\boldsymbol{P} \in \mathbb{R}^{m_1 \times m_2}$, $\|\boldsymbol{P}\|_{\text{op}}$ stands for the spectral norm, which equals the largest singular value of $\boldsymbol{P}$. We also use $\|\boldsymbol{P}\|_F = \sqrt{\text{trace}(\boldsymbol{P}\boldsymbol{P}^\top)}$ and $\|\boldsymbol{P}\|_1$ to denote the Frobenius norm and element-wise ℓ_1 norm, respectively. We refer to Horn and Johnson (2012) for a linear algebra and matrix analysis background. Besides, we use the following standard probabilistic asymptotic notations: for two sequences of random variables a_n and b_n, we say $a_n = O(b_n)$ if a_n/b_n is bounded in probability and $a_n = o(b_n)$ if $a_n/b_n \to 0$ in probability. Let $\psi_p(x) = e^{x^p} - 1, p \geq 1$, then the ψ_p-Orlicz norm of a random variable X is defined as: $\|X\|_{\psi_p} = \inf\{t > 0 : \mathbb{E}\{\psi_p(|X|/t)\} \leq 1\}$. For a random vector $\boldsymbol{x} \in \mathbb{R}^d$, we define its ψ_p-Orlicz norm $\|\boldsymbol{x}\|_{\psi_p} := \sup_{\boldsymbol{v} \in \mathcal{D}^{d-1}} \left\|\boldsymbol{v}^\top \boldsymbol{x}\right\|_{\psi_p}$, where $\mathcal{D}^{d-1}$ is the d-dimensional unit sphere.

2 Preliminaries

Suppose we have observed a collection of n data points, $\{\boldsymbol{x}_i\}_{i=1}^n$, which are centered, independent and identically distributed (i.i.d) samples. and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ the sample matrix. We denote the population and the sample covariance respectively by $\boldsymbol{\Sigma}_1^\star$ and $\widehat{\boldsymbol{\Sigma}}$. $\lambda_1^\star \geq \cdots \geq \lambda_p^\star$ are eigenvalues of $\boldsymbol{\Sigma}_1^\star$, while $\boldsymbol{\beta}_i, i = 1 \cdots, p$ are the corresponding eigenvectors.

2.1 Revisiting Zou et al. (2006)

Our work takes off from the Sparse PCA(SPCA) proposed by Zou et al. (2006), which gives birth to the first computationally efficient algorithm for high dimensional principal component analysis. Specifically, SPCA for the first PC is based on a regression formulation in their Theorem 2, plus a ℓ_1 penalty term:

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = &\underset{\boldsymbol{\alpha},\, \boldsymbol{\beta}}{\arg\min} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\alpha}\boldsymbol{\beta}^\top \boldsymbol{x}_i\|_2^2 + \rho_1 \|\boldsymbol{\beta}\|_2^2 + \rho_2 \|\boldsymbol{\beta}\|_1 \\ &\text{subject to } \|\boldsymbol{\alpha}\|_2 = 1.\end{aligned} \tag{1}$$

Then $\widehat{\boldsymbol{\beta}}_{unit}$ is obtained by performing normalization on $\widehat{\boldsymbol{\beta}}$. The objective function in (1) gives a separable biconvex formulation, which makes it solvable via alternately updating α and β. However, the nonconvex manifold constraint $\|\boldsymbol{\alpha}\|_2 = 1$ imposes new drawbacks on the statistical property of $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$. Besides, if we instead turn to a convex relaxation to this constraint for the sake of theoretical analysis, the choice of the oracle solution would be challenging. This fact is suggested by the observation that, after expansion, the population risk of (1) is

$$R(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{tr}(\boldsymbol{\Sigma}_1^\star) - 2\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} + \|\boldsymbol{\alpha}\|_2^2 \cdot \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_2^2.$$

In order to get a local optimal solution(without the constraint), we set the first derivative of $R(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to be zero:

$$\begin{cases} -\boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} + (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\boldsymbol{\alpha} = 0 \\ -\boldsymbol{\Sigma}_1^\star \boldsymbol{\alpha} + \|\boldsymbol{\alpha}\|_2^2 \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} + \lambda \boldsymbol{\beta} = 0 \end{cases}$$

Now the dilemma is as follows: if $\lambda = 0$, by taking $\boldsymbol{\beta} = t\boldsymbol{\beta}_1^\star$ and $\boldsymbol{\alpha} = \frac{1}{t}\boldsymbol{\beta}_1^\star$, for any $t > 0$, the equations can be satisfied, which implies the oracle solution is not unique. On the other hand, if $\lambda > 0$, by taking $\boldsymbol{\beta} = t\boldsymbol{\beta}_1^\star$, the first equation holds with and only with $\boldsymbol{\alpha} = \frac{1}{t}\boldsymbol{\beta}_1^\star$; but in this case the second equality shall never be attained and we cannot convince ourselves of the existence of a potential oracle solution anymore.

In a nutshell, the SPCA proposed by Zou et al. (2006) utilized a biconvex formulation to lower the burden of computation, but also introduced some drawbacks in statistical analysis due to the nonconvex constraint $\|\boldsymbol{\alpha}\|_2 = 1$ as well as the difficulty of locating a unique oracle solution. To handle these drawbacks, one question naturally comes up: what in general could be classified as a preferable landscape, for which we can target a precise oracle point and build a statistical guarantee?

2.2 Characterization of a good nonconvex landscape

The analysis in the last section motivates us to construct a set of criteria for the identification of a "good" landscape for a nonconvex formulation. Our proposal is in the similar vein to M-estimation (Huber, 2004), where the loss forms a sample mean and the population risk is minimized at the ground truth. Specifically speaking, among the class of loss functions in the form of a sample mean, i.e.,

$$\widehat{R}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \psi(\boldsymbol{x}_i, \boldsymbol{\beta}),$$

we pursue the subfamily which is associated with a population risk $R(\boldsymbol{\beta}) = \mathbb{E}(\widehat{R}_n(\boldsymbol{\beta}))$ satisfying the following property:

(a) $\widehat{R}(\boldsymbol{\beta})$ and $R(\boldsymbol{\beta})$ are defined over a convex set $\mathcal{C}$ containing the truth $\boldsymbol{\beta}^\star$.

(b) $R(\boldsymbol{\beta})$ is minimized globally at the ground truth $\boldsymbol{\beta}^\star$;

(c) $R(\boldsymbol{\beta})$ is differentiable, and has a strongly convex structure within a neighborhood $\mathbb{B}_\delta(\boldsymbol{\beta}^\star)$ of $\boldsymbol{\beta}^\star$, i.e.,

$$\exists\ \sigma > 0,\ \text{s.t. } R(\boldsymbol{\beta}_1) \geq R(\boldsymbol{\beta}_2) + \nabla R(\boldsymbol{\beta}_2)^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + \frac{\sigma}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2,\ \forall\ \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{B}_\delta(\boldsymbol{\beta}^\star).$$

Generally it is not necessary for a function to be differentiable in order to be strongly convex(see Appendix B, Section 1.1 of Bertsekas et al. (2003)). But for our interest, it is sufficient to focus on population risks that are twice continuously differentiable, where strong convexity is guaranteed if and only if $\left(\nabla^2 R(\boldsymbol{\beta}) - \sigma I_p\right)$ is positive definite for every $\beta \in \mathbb{B}_\delta(\boldsymbol{\beta}^\star)$.

We shall briefly state why these conditions are sufficient and necessary for generating a consistent estimator. For the sufficiency part,

$$\begin{aligned}\frac{\sigma}{2}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2 &\leq R(\widehat{\boldsymbol{\beta}}) - R(\boldsymbol{\beta}^\star) \text{ using (b) and (c)} \\ &\leq \underbrace{\widehat{R}(\widehat{\boldsymbol{\beta}}) - \widehat{R}(\boldsymbol{\beta}^\star)}_{\mathbf{I}} + \underbrace{|\widehat{R}(\widehat{\boldsymbol{\beta}}) - \widehat{R}(\boldsymbol{\beta}^\star) - (R(\widehat{\boldsymbol{\beta}}) - R(\boldsymbol{\beta}^\star))|}_{\mathbf{II}}. \end{aligned} \tag{2}$$

Now if we can control **I** and **II**, the consistency under ℓ_2 measurement can be concluded. In high dimensional literature, like Lasso (Bühlmann and Van De Geer, 2011; Bickel et al., 2009) or low rank matrix recovery (Negahban and Wainwright, 2011; Elsener et al., 2018), Term **I** can often be controlled under several conditions, including assuming sparsity for the true parameter and performing regularization using a decomposable norm among others (see e.g. Negahban et al. (2012); Elsener and van de Geer (2018)). Term **II** can be controlled using an argument based on empirical process considering that $\widehat{R}$ is formulated as a sample mean.

Meanwhile, these conditions are somewhat necessary when studying a nonconvex landscape. Ideally we need to expect a local region over which the formulation has a tractable structure and is possible to generate an effective estimator. Condition (a) serves as a basic requirement for this purpose. Many nonconvex sparse PCA formulation pursues a convex relation on the constraints to achieve (a), for example, see d'Aspremont et al. (2005) and Vu et al. (2013). From an asymptotic point of view, the sample mean $\widehat{R}(\boldsymbol{\beta})$ converges to its expectation $R(\boldsymbol{\beta})$, hence intuitively the minimizer of the loss function should also approach that of the population risk. If (b) is violated, then $\widehat{\boldsymbol{\beta}}$ might deviate away from the truth and head into a wrong direction. For (c), on one hand, it guarantees that there is no other global minimizer of $R(\boldsymbol{\beta})$ than $\boldsymbol{\beta}^\star$ within $\mathbb{B}_\delta(\boldsymbol{\beta}^\star)$. On the other hand, when $R(\boldsymbol{\beta})$ approaches the global minimum, this condition specifies the quadratic rate of convergence for $\boldsymbol{\beta}$, which aligns well with the spirits of classical high dimensional M-estimation schemes (Negahban et al., 2012).

Now based on our analysis in last section, formulation (1) violates these conditions, making it difficult to compose a theoretical argument. On the contrary, a nonconvex programming scheme we notice that satisfies Condition (a)-(c) comes from Section 12.6 of **?** and Section 3.2 of Elsener and van de Geer (2018), where the authors proposes the following estimator for the first PC:

$$\widehat{\boldsymbol{\beta}} = \underset{\substack{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\sqrt{\lambda_1^\star}\boldsymbol{\beta}_1^\star\|_2\leq\eta \\ \|\boldsymbol{\beta}\|_1\leq L}}{\arg\min} \frac{1}{4}\|\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\beta}\boldsymbol{\beta}^\top\|_F^2 + \rho\|\boldsymbol{\beta}\|_1. \tag{3}$$

where $\widehat{\Sigma}_1$ is the sample covariance matrix. Their argument shows $\boldsymbol{\beta}^\star$ is the minimizer of the population risk, which has strong convexity over the convex constraint region under several assumptions.

Using these properties along with several others, they were able to establish oracle inequalities for $\widehat{\beta}$ and show the consistency under both ℓ_1 and ℓ_2 measurement, which validates the effects of the above conditions.

3 Methodology

3.1 Estimation of the first PC in the direct formulation of sparse PCA

Our arguments in the previous section motivate us to pursue another estimation scheme based on manifold approximation that has more favorable landscape than (1). Even though (3) satisfies this purpose, several limitation hinders its further application. Firstly, the center of the ℓ_2 neighborhood constraint is $\sqrt{\lambda_1^\star}\boldsymbol{\beta}_1^\star$, making it hard to find an appropriate initialization. Besides, the extraction procedure for further PCs is not included.

Starting from this section we introduce a direct formulation of sparse PCA. Recall for low dimensional PCA (Pearson, 1901), the first principal component can be extracted by projection:

$$\begin{aligned}\boldsymbol{\beta}_1^\star =& \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{x}_i\|_2^2 \\ &\text{subject to } \|\boldsymbol{\beta}\|_2 = 1.\end{aligned} \tag{4}$$

where $\boldsymbol{\beta}\boldsymbol{\beta}^\top$ serves as the projection operator. The objective function and the theoretical risk stand as nonconvex functions, whose complexity increases significantly as the dimension of the problem grows. In order to introduce sparsity in (4), we propose to append an ℓ_1 penalty and some convex constraint:

$$\tilde{\boldsymbol{\beta}}_1 = \underset{\substack{\boldsymbol{\beta}:\|\boldsymbol{\beta}-\boldsymbol{\beta}_1^\star\|_2\le\eta \\ \|\boldsymbol{\beta}\|_1\le L}}{\arg\min} \frac{1}{2(n-1)}\sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{x}_i\|_2^2 + \rho\|\boldsymbol{\beta}\|_1. \tag{5}$$

Here, $\eta > 0$ and $L > 0$ are some parameters and we will add more discussion over them later. A unit estimator $\widehat{\boldsymbol{\beta}}_1$ is obtained by further normalizing $\tilde{\boldsymbol{\beta}}_1$:

$$\widehat{\boldsymbol{\beta}}_1 = \frac{\tilde{\boldsymbol{\beta}}_1}{\|\tilde{\boldsymbol{\beta}}_1\|_2}. \tag{6}$$

Before diving into technical details of this "direct" estimator, we first get a glimpse of the landscape of the risk formulation to see how it meets the criteria we stated in Section 2.2. The first notable fact is, if the first principal component is identifiable, i.e, for the eigenvalues we have $\lambda_1^\star > \lambda_2^\star$, the global minimizer of the population risk is given by $\boldsymbol{\beta}_1^\star$, which is unique(up to a sign) over $\mathbb{R}^p$. Mathematically we have:

$$\boldsymbol{\beta}_1^\star = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ R_1(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{x}_i\|_2^2\right],$$

and no other $\boldsymbol{\beta} \in \mathbb{R}^p$ minimize this objective function. Note that we do not apply the constraint $\|\boldsymbol{\beta}\|_2 = 1$ here, thus able to circumvent the nonconvex region and meets Condition (a) and (b). The second fact is, even though we are faced with a nonconvex loss, the local landscape around the minimizer $\boldsymbol{\beta}_1^\star$ is endowed with a strongly convex structure, which is highlighted by the following lemma:

Lemma 1. *Suppose $\lambda_1^\star > \lambda_2^\star$. Denote the smallest eigenvalue of $\nabla^2 R_1(\boldsymbol{\beta})$ by $\lambda_{\min}(\nabla^2 R_1(\boldsymbol{\beta}))$. There exists an $\eta > 0$ such that for any $\boldsymbol{\beta}$, $\|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \leq \eta$, we have*

$$\lambda_{\min}\left(\nabla^2 R_1(\boldsymbol{\beta})\right) \geq (1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star > 0.$$

Furthermore, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \leq \eta\}$,

$$R_1(\boldsymbol{\beta}_1) - R_1(\boldsymbol{\beta}_2) - \nabla R_1(\boldsymbol{\beta}_2)^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \geq \frac{\left\{(1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star\right\}}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2.$$

These facts build up the foundation of several desiring statistical property for the direct formulation. The rest of this work will take advantage of Condition (a) to (c) and providing theoretical guarantees for the direct estimation scheme. Note that (5) requires locating the minimization around the truth $\boldsymbol{\beta}_1^\star$ due to the global non-convex landscape. The following proposition shows the complex nonconvex structure of the loss function as well as the necessity of this localization procedure.

Proposition 1. *The stationary points of the function*

$$R_1(\boldsymbol{\beta}) = \mathbb{E}\left[\widehat{R}_1(\boldsymbol{\beta})\right] = \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}_1^\star) + (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_1^\star\boldsymbol{\beta}\right]$$

are the eigenvectors of $\boldsymbol{\Sigma}_1^\star$. $\pm\boldsymbol{\beta}_1^\star$ are the only global minimizer of the population risk, and the other eigenvectors are all saddle points.

This fact indicates that once it fails for $\boldsymbol{\beta}$ to lie close enough to our target $\boldsymbol{\beta}_1^\star$, the optimization progress would be problematic due to the impact of saddle points. To avoid this situation we hope to introduce a good initialization point $\bar{\boldsymbol{\beta}}_1$, which is an consistent estimator with a sub-optimal convergence rate:

$$\|\bar{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = O(\xi_n),$$

where $\xi_n = o(1)$. Many papers have provided such slow-rate estimators, like in Johnstone and Lu (2009); Shen et al. (2013), etc. Therefore, together with properly chosen localization parameters η and L, our final estimator will be:

$$\tilde{\boldsymbol{\beta}}_1 = \underset{\substack{\boldsymbol{\beta}:\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_1\|_2 \leq \frac{2\eta}{3} \\ \|\boldsymbol{\beta}\|_1 \leq L}}{\arg\min} \frac{1}{2(n-1)} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{x}_i\|_2^2 + \rho\|\boldsymbol{\beta}\|_1. \tag{7}$$

3.2 Estimation of further PCs in the direct formulation of sparse PCA

Now we move forward to extract the further principal components in a sequential style. The basic idea starts from the deflation interpretation for PCA. Let $\boldsymbol{B}_0 = 0$, and $\boldsymbol{B}_k = [\boldsymbol{\beta}_1^\star, \boldsymbol{\beta}_2^\star, \ldots, \boldsymbol{\beta}_k^\star]$ be the first k eigenvectors of $\widehat{\boldsymbol{\Sigma}}$, and the residual points after projection are $\boldsymbol{y}_i^{(k+1)} = \boldsymbol{x}_i - \boldsymbol{B}_k \boldsymbol{B}_k^\top \boldsymbol{x}_i$, $i = 1, \ldots, k$. Then the $(k+1)$-th eigenvector $\boldsymbol{\beta}_{k+1}^\star$ is exactly the global minimizer of following risk function:

$$R_{k+1}(\boldsymbol{\beta}) = \mathbb{E}\left(\left\| \boldsymbol{y}_i^{(k+1)} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{y}_i^{(k+1)} \right\|_2^2 \right)$$

which can be verified by taking the $(k+1)$-th population covariance to be:

$$\boldsymbol{\Sigma}_{k+1}^\star = (\boldsymbol{I} - \boldsymbol{B}_k \boldsymbol{B}_k^\top) \boldsymbol{\Sigma}_1^\star (\boldsymbol{I} - \boldsymbol{B}_k \boldsymbol{B}_k^\top), \tag{8}$$

and using the fact that $\boldsymbol{\beta}_{k+1}^\star$ is the leading eigenvector for (8). This motivates us to extract the $k+1$-th PC by:

$$\tilde{\boldsymbol{\beta}}_{k+1} = \underset{\substack{\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_{k+1}\|_2 \le \frac{2\eta}{3} \\ \|\boldsymbol{\beta}\|_1 \le L}}{\arg\min} \frac{1}{n-1} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i^{(k+1)} - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{y}_i^{(k+1)} \right\|_2^2 + \rho \|\boldsymbol{\beta}\|_1 . \tag{9}$$

Analogous to the first PC, a unit estimator is further obtained by normalization, i.e.,

$$\widehat{\boldsymbol{\beta}}_{k+1} = \frac{\tilde{\boldsymbol{\beta}}_{k+1}}{\|\tilde{\boldsymbol{\beta}}_{k+1}\|_2}. \tag{10}$$

Mackey (2009) pointed out that it might be problematic if we simply substitute the first k eigenvectors with the first k estimators in each step, which may result in "double counting" in reducing the variance. Therefore we choose to perform the orthogonal projection deflation. Concretely speaking, we propose to wedge a step of Schmidt orthogonalization on the previous k estimators and obtain:

$$\widehat{\boldsymbol{Q}}_k = [\widehat{\boldsymbol{q}}_1, \cdots, \widehat{\boldsymbol{q}}_k], \text{ where } \widehat{\boldsymbol{q}}_j = \frac{\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{Q}}_{j-1} \widehat{\boldsymbol{Q}}_{j-1}^\top \widehat{\boldsymbol{\beta}}_j}{\left\| \widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{Q}}_{j-1} \widehat{\boldsymbol{Q}}_{j-1}^\top \widehat{\boldsymbol{\beta}}_j \right\|_2}. \tag{11}$$

Then $\widehat{\boldsymbol{Q}}_k$ will be used for further deflation. The full roadmap of our regime is summarized in Algorithm 1.

Algorithm 1 The Direct Formulation of Sparse PCA

Input: k(the number of interested PCs); $\bar{\boldsymbol{\beta}}_j$, $j = 1, \cdots, k$(initial points); ρ_j, L_j and η_j, $j = 1, \cdots, k$(parameters).

Output: $\widehat{\boldsymbol{\beta}}_j$, $j = 1, \ldots, k$.

[For the first principal component($j = 1$)]

1: Solve the optimization problem 5 to obtain $\tilde{\boldsymbol{\beta}}_1$. Then normalize $\tilde{\boldsymbol{\beta}}_1$ to obtain $\widehat{\boldsymbol{\beta}}_1$. Record $\widehat{\boldsymbol{Q}}_1 = \widehat{\boldsymbol{\beta}}_1$.

[For the further principal components($j \geq 2$)]

2: **[Deflation]** Set $\boldsymbol{y}_i^{(j)} = (\boldsymbol{I} - \widehat{\boldsymbol{Q}}_{j-1}\widehat{\boldsymbol{Q}}_{j-1}^\top)\boldsymbol{x}_i$, for $i = 1, \ldots, n$.

3: **[Minimization]** Get the j-th unnormalized estimator $\tilde{\boldsymbol{\beta}}_j$ by Formulation (7) and (9);

4: **[Normalization]** Extract the j-th unit estimator $\widehat{\boldsymbol{\beta}}_j$ by a step of normalization using (6) and (10);

5: **[Orthogonalization]** Update $\widehat{\boldsymbol{q}}_j$ and $\widehat{\boldsymbol{Q}}_j$ according to (11).

6: $j \leftarrow j + 1$. Loop Step 2-5 until $j = k + 1$.

In terms of implementation, for Step 3, we apply proximal gradient descent method proposed by Nesterov (2013). Other schemes are also possible, such as applying ADMM (Boyd et al., 2011) etc.

4 Theoretical Analysis

In this section we study the theoretical property of our estimators. Our analysis starts with the first principal component, where we conduct our study based on the discussion from Section 2.2. Concretely speaking, we seek to bound the two terms given in 2 by adapting a framework proposed in Elsener and van de Geer (2018). In the second subsection we analyze the asymptotic performance of the further estimated components, which is far more than a trivial generalization of the results for the first component.

According to the optimality condition for subgradient methods (Bertsekas et al., 2003), the minimizer for each iteration in Algorithm 1 should satisfy the following condtion:

$$\nabla \widehat{R}_i(\tilde{\boldsymbol{\beta}}_i) + \rho\tilde{v} = 0, \text{ for some } \tilde{v} \in \partial(\|\boldsymbol{\beta}\|_1)\big|_{\tilde{\boldsymbol{\beta}}_i}.$$

where $\partial(\|\cdot\|_1)$ is the subgradient of ℓ_1 norm. Let $\mathcal{C}(\bar{\boldsymbol{\beta}}_i, L_i, \eta_i)$ be the constraint region in (7) and (9) where we perform the minimization. By the definition of subgradient, We have $\|\boldsymbol{\beta}\|_1 - \|\tilde{\boldsymbol{\beta}}_i\|_1 \geq \tilde{v}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_i)$.By this inequality, times $\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_i$ on both sides of the condition, then we have that

$$\nabla \widehat{R}_i(\tilde{\boldsymbol{\beta}}_i)^\top(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_i) + \rho\|\boldsymbol{\beta}\|_1 - \rho\|\tilde{\boldsymbol{\beta}}_i\|_1 \geq 0, \forall\ \boldsymbol{\beta} \in \mathcal{C}(\bar{\boldsymbol{\beta}}_i, L_i, \eta_i), \tag{12}$$

which is called "the two point inequality" according to Elsener and van de Geer (2018) and serves as the basic property we will use in our analysis.

4.1 Asymptotic bounds for the first component

For this part we need some assumptions to carry on our theory:

Assumption 1. *We have the following four assumptions.*

1.A: (Sub-gaussianity) *The features $x_1, \ldots, x_n$ are i.i.d. copies of a sub-Gaussian random vector $X \in \mathbb{R}^{1\times p}$ with positive definite covariance matrix $\mathbf{\Sigma}_1^\star$, and parameter $K = \sup_{\|u\|_2=1} \|Xu\|_{\psi_2}$. Here $\|\cdot\|_{\psi_2}$ is the sub-gaussian norm.*

1.B: (Sparsity) *The first true PC $\boldsymbol{\beta}_1^\star$ is sparse. We say $\boldsymbol{\beta}_1^\star$ is exactly sparse if the cardinal of $\boldsymbol{\beta}_1^\star$ equals $s_0 < p$; we say $\boldsymbol{\beta}_1^\star$ is approximately sparse with respect to an ℓ_q ball, $\mathbb{B}_q(s_q)$, for some $q \in (0,1)$ and $s_q > 0$, if*

$$\sum_{k=1}^{p} |\boldsymbol{\beta}_{1k}^\star|^q \le s_q.$$

1.C: (Separation) *The two largest eigenvalues $\lambda_1^\star$ and $\lambda_2^\star$ are fixed, and satisfy*

$$\lambda_1^\star > \lambda_2^\star > \sigma \ge \lambda_3^\star \ge \cdots \ge \lambda_p^\star.$$

1.D: (Asymptotics) *p increases subexponentially as $n \to \infty$; namely, $p > n$, and $\frac{\log p}{n} \to 0$. For exact sparse case, the number of the sparse elements s_0 satisfies that $s_0\sqrt{\frac{\log p}{n}} \to 0$. The ℓ_1 bound $L > 0$ satisfies $L = O(\sqrt{s_0})$. Besides, consider the reality we assume that $p = O(n^c)$ for some constant $c > 0$. For the approximate low rank case, we assume $s_q\,(\log p/n)^{(1-q)/2} \to 0$, and $L = O\{\sqrt{s_q}(\log p/n)^{-q/4}\}$.*

We give some simple elaboration on these assumptions. With the sub-gaussianity assumption, the tail of the random vector we investigate can be well controlled with high probability. In regards of sparsity, note that when $q \to 0$,

$$\sum_{k=1}^{p} |\boldsymbol{\beta}_{1k}^\star|^q \to \#\{k : |\boldsymbol{\beta}_{1k}^\star| \neq 0\} = s_0.$$

That is, s_q degenerates to the exact sparsity s_0. Therefore, for simplicity, we can incorporate the exactly sparse case into the approximate one by allowing $q = 0$. For the separation assumption, it guarantees that, the first eigenvalue is well-separated from the others, where $\frac{\lambda_2^\star}{\lambda_1^\star} < \epsilon < 1$. The asymptotic quantification, $L \asymp \sqrt{s_1}$, is natural from the view that $\|\boldsymbol{\beta}_1^\star\|_1 \le \sqrt{s_1}\|\boldsymbol{\beta}_1^\star\|_2 = \sqrt{s_1}$. And $p = O(n^c)$ allows the dimension to grow polynomially. In general our theory applies to a sub-exponential setup, but it suffices to have a polynomial type growth in many real-world applications.

Our idea of building the oracle inequality for stationary points is inspired by Elsener and van de Geer (2018). First we will show that with our assumptions the theoretical risk for the first formulation (5) exhibits the property of strong convexity.

In the next lemma we show that the first order approximation for the difference between the sample and the population risk is controlled by a combination of the ℓ_2 norm term and an ℓ_1 norm term. This result enables us to

Lemma 2. *Let $K := \sup_{\|u\|_2=1} \|Xu\|_{\psi_2}$. We still use the notations in Lemma 1. Let σ be defined as in the Seperation assumption. Fix a $\boldsymbol{\beta}^\star \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \le \eta, \|\boldsymbol{\beta}\|_1 \le L\}$. For $t > 0$ we define*

$$
\begin{aligned}
M_0(\log 2p) &:= 2K^2\left(1 + \frac{2\lambda_1^\star}{\sigma}\right)\left\{\frac{6 + 2\log(2p)}{c_K n}(1 + (2\zeta)^{-1}) + \zeta\right\},\\
M_1(\log 2p) &:= 2K^2\left(1 + \frac{2\lambda_1^\star}{\sigma}\right)\left\{\frac{2(6 + \log(2p))}{c_K n} + \frac{6 + \log(2p)}{\zeta c_K n}\right\},\\
M_2(\log 2p) &= 2\left(K^2\frac{2\log 2p}{c_K n} + K^2\sqrt{\frac{2\log 2p}{c_K n}}\right),\\
M_3(\log 2p) &= 4K^2\frac{\log 4p}{c_K n} + 4K^2\sqrt{\frac{\log 4p}{c_K n}}.
\end{aligned}
$$

Let $M_\epsilon = 8LM_1(\log 2p) + 3M_2(\log 2p) + 2M_3(\log 2p)$, $\gamma = 4M_0(\log 2p)$, and

$$
\zeta = \left\{32K^2\left(1 + \frac{2\lambda_1^\star}{\sigma}\right)\right\}^{-1}.
$$

When the sample size n satisfies

$$
n > \zeta^{-1}(1 + \zeta^{-1})\frac{6 + 2\log 2p}{c_K},
$$

we have that, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \le \eta, \|\boldsymbol{\beta}\|_1 \le L\}$

$$
\begin{aligned}
&\left|[\nabla\widehat{R}_1(\boldsymbol{\beta}) - \nabla R_1(\boldsymbol{\beta})]^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\right|\\
\le&\frac{\left\{(1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star\right\}\gamma}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_2^2 + M_\epsilon\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1
\end{aligned}
$$

with probability at least $1 - 3\cdot(2p)^{-1}$,

Note that from an asymptotic point of view, $LM_1(\log 2p) = O(L\frac{\log p}{n}) \lesssim \sqrt{\frac{L\log p}{n}}$, $M_2(\log 2p) \asymp \sqrt{\frac{\log p}{n}}$, and $M_3(\log 2p) \asymp \sqrt{\frac{\log p}{n}}$, so asymptotically we obtain that

$$
M_\epsilon = O\left(\sqrt{\frac{L\log p}{n}}\right).
$$

With all the lemmas stated before we are now ready to prove our oracle inequality.

Theorem 1 (Oracle Inequality for the First Estimator)**.** *Let $\tilde{\boldsymbol{\beta}}_1$ be a stationary point of the optimization problem* (7)*. We inherit the conditions and notations in Lemma 1 and Lemma 2. Now pick a penalty factor $\rho = 2M_\epsilon$. With a fixed $0 < \delta < 1$ we have with probability at least $1 - 3\cdot(2p)^{-1}$*

$$
\begin{aligned}
&R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}) + \frac{\rho\delta}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1\\
\le&\frac{(3 + \delta)^2\rho^2 s}{8(1 - \gamma)\left\{(1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star\right\}} + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1.
\end{aligned}
$$

Now we can use this result to build the consistency of (5). We consider two scenarios respectively: exact and approximate sparsity for the first PC, presented respectively in the following two corollaries.

Corollary 1 (Exact Sparsity)**.** *Assume the same condition as in Theorem 1. Besides, we assume the exact sparsity assumption(Assumption 1.B) holds for $\boldsymbol{\beta}_1^\star$ with s_0. By the choice of $\rho = 2M_\epsilon$ and a fixed $0 < \delta < 1$, we have with probability at least $1 - 3 \cdot (2p)^{-1}$*

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \leq \frac{(3+\delta)^2 \rho s_0}{4\delta(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}} = O\left(s_0\sqrt{\frac{\log p}{n}}\right),$$
$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 \leq \frac{(3+\delta)\rho\sqrt{s_0}}{2\sqrt{(1-\gamma)}[(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star]} = O\left(\sqrt{\frac{s_0 \log p}{n}}\right).$$

Further, consider the normalization of $\tilde{\boldsymbol{\beta}}_1$, which we denote as $\widehat{\boldsymbol{\beta}}_1$, and the error of projection matrix $\Delta_1 = \widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_1^\top - \boldsymbol{\beta}_1^\star\boldsymbol{\beta}_1^{\star T}$, we have:

$$\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 = O\left(s_0\sqrt{\frac{\log p}{n}}\right), \|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = O\left(\sqrt{s_0\frac{\log p}{n}}\right), \|\Delta_1\|_F = O\left(\sqrt{s_0\frac{\log p}{n}}\right).$$

For the approximate sparsity case we have the following corollary:

Corollary 2 (Approximate Sparsity)**.** *Assume the same condition as in Theorem 1. Besides, we assume the exact sparsity assumption(Assumption 1.B) holds for $\boldsymbol{\beta}_1^\star$ with $\mathbb{B}_q(s_q)$. By the choice of $\rho = 2M_\epsilon$ and a fixed $0 < \delta < 1$, we have with probability at least $1 - 3 \cdot (2p)^{-1}$*

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \leq \frac{6}{\delta}\left\{\frac{(3+\delta)^2}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}\right\}^{1-q} s_q\rho^{1-q} = O\left\{s_q\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right\},$$
$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 \leq \sqrt{6}\left\{\frac{(3+\delta)^2}{8(1-\gamma)}\right\}^{\frac{1-q}{2}} \frac{1}{\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}^{1-q/2}}\sqrt{s_q}\rho^{1-q/2} = O\left\{\sqrt{s_q}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q}{4}}\right\}.$$

Further, consider the normalization of $\tilde{\boldsymbol{\beta}}_1$, which we denote as $\widehat{\boldsymbol{\beta}}_1$, and the error of projection matrix $\Delta_1 = \widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_1^\top - \boldsymbol{\beta}_1^\star\boldsymbol{\beta}_1^{\star T}$, we have:

$$\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 = O\left\{s_q\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right\}, \|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = O\left\{\sqrt{s_q}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q}{4}}\right\}, \|\Delta_1\|_F = O\left\{\sqrt{s_q}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q}{4}}\right\}.$$

4.2 Asymptotic Bounds for Further Components

Now we turn to the analysis of further components. We first state the necessary assumptions for establishing the theories.

Assumption 2. *We assume the following conditions:*

2.A: (Sub-gaussianity) *The features $X_1, \ldots, X_n$ are i.i.d. copies of a zero mean sub-Gaussian random vector $X \in \mathbb{R}^{1\times p}$ with positive definite covariance matrix $\mathbf{\Sigma}_1^\star$.*

2.B: (Separation) *The eigenvalues of $\mathbf{\Sigma}_1^\star$ are arranged in the following pattern: for some $\sigma > 0$,*

$$\lambda_1^\star > \lambda_2^\star > \cdots > \lambda_k^\star > \lambda_{k+1}^\star \geq \lambda_{k+2}^\star \geq \cdots \geq \lambda_p^\star > \sigma > 0.$$

Besides, the first k eigenvalues are fixed with no variation when n and p change.

2.C: (Sparsity) *The first k eigenvectors $\boldsymbol{B}_k = [\boldsymbol{\beta}_1^\star, \cdots, \boldsymbol{\beta}_k^\star]$ of $\mathbf{\Sigma}_1^\star$ are exactly sparse, with support size respectively $s_0^\star, \cdots, s_k^\star$.*

2.D: (Asymptotics) *$p > n$. Let s be the maximal sparsity of the first k principal components. We assume that when $n \to \infty, p \to \infty$,*

$$s^k \cdot \sqrt{\frac{\log p}{n}} \to 0.$$

We can easily show that Assumption 2 is a sufficient condition for Assumption 1 if we only focus on $k = 1$ and the exact sparsity case.

Recall at the beginning of this section, we mentioned the difficulties in dealing with further PCs. One of our key innovations is the introduction of the so-called "intermediate risk". More specifically, consider for the j-th step, after simple algebra, we obtain the expression for the empirical risk and the population risk, given respectively by

$$\widehat{R}_j(\boldsymbol{\beta}) = \frac{1}{2}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_j) + (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top \widehat{\mathbf{\Sigma}}_j \boldsymbol{\beta}. \tag{13}$$

$$R_j(\boldsymbol{\beta}) = \frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_j^\star) + (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top \mathbf{\Sigma}_j^\star \boldsymbol{\beta}. \tag{14}$$

where the sample covariance $\widehat{\mathbf{\Sigma}}_j$ and the population covariance $\mathbf{\Sigma}_j^\star$ are

$$\widehat{\mathbf{\Sigma}}_j = \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i^{(j)} \boldsymbol{y}_i^{(j)\top} = (\boldsymbol{I}_p - \widehat{\boldsymbol{Q}}_{j-1}\widehat{\boldsymbol{Q}}_{j-1}^\top)\widehat{\mathbf{\Sigma}}_1(\boldsymbol{I}_p - \widehat{\boldsymbol{Q}}_{j-1}\widehat{\boldsymbol{Q}}_{j-1}^\top),$$
$$\mathbf{\Sigma}_j^\star = (\boldsymbol{I}_p - \boldsymbol{B}_{j-1}\boldsymbol{B}_{j-1}^\top)\mathbf{\Sigma}_1^\star(\boldsymbol{I}_p - \boldsymbol{B}_{j-1}\boldsymbol{B}_{j-1}^\top).$$

Note that in (13) and (14), the expectation of $\widehat{R}_j(\beta)$ is not necessarily $R_j(\beta)$, since $\widehat{\boldsymbol{Q}}_j$ is also determined by the data points x_i. To circumvent this issue, we need to introduce *the intermediate risk*, which can be defined as:

$$\breve{R}_j(\boldsymbol{\beta}) = \frac{1}{2}\mathrm{tr}(\breve{\mathbf{\Sigma}}_j) + (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top \breve{\mathbf{\Sigma}}_j \boldsymbol{\beta}. \tag{15}$$

where the "intermediate covariance" $\breve{\mathbf{\Sigma}}_j$ is given by:

$$\breve{\mathbf{\Sigma}}_j = (\boldsymbol{I}_p - \widehat{\boldsymbol{Q}}_{j-1}\widehat{\boldsymbol{Q}}_{j-1}^\top)\mathbf{\Sigma}_1^\star(\boldsymbol{I}_p - \widehat{\boldsymbol{Q}}_{j-1}\widehat{\boldsymbol{Q}}_{j-1}^\top).$$

We will show in our proof that, while it is hard to compare (13) and (14) directly, using (15) as a bridge can circumvent the theoretical difficulty and leads to a successful induction argument. To this end, we first establish several results in a conditional sense.

Lemma 3. *Given that* $\|\Delta_{i-1}\|_F = O(\sqrt{s\log p/n})$. *Denote the smallest eigenvalue of* $\nabla^2\breve{R}_i(\boldsymbol{\beta})$ *by* $\lambda_{\min}(\nabla^2\breve{R}_i(\boldsymbol{\beta}))$. *The leading eigenvector of* $\breve{\boldsymbol{\Sigma}}_i$ *is denoted by* $\breve{\boldsymbol{\beta}}_i$. *There is an* $\eta > 0$ *such that when* $n > 0$ *is large enough, for any* $\boldsymbol{\beta}$, $\left\|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_i\right\|_2 \leq \eta$, *we have*

$$\lambda_{\min}(\nabla^2\breve{R}_i(\boldsymbol{\beta})) \geq (1 - 14\eta - 4\eta^2)\frac{2\lambda_i^\star + \lambda_{i+1}^\star}{3} - (1+2\eta)\frac{\lambda_i^\star + 2\lambda_{i+1}^\star}{3} > 0.$$

Furthermore, when n *is large enough, for any* $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_i\|_2 < \eta\}$,

$$\begin{aligned}
&\breve{R}_i(\boldsymbol{\beta}_1) - \breve{R}_i(\boldsymbol{\beta}_2) - \nabla\breve{R}_i(\boldsymbol{\beta}_2)^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \\
\geq &\frac{\left\{(1 - 14\eta - 4\eta^2)(2\lambda_i^\star + \lambda_{i+1}^\star) - (1+2\eta)(\lambda_i^\star + 2\lambda_{i+1}^\star)\right\}}{6}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2.
\end{aligned}$$

The proof of this theorem is *not* a simple generalization for Lemma 1. This is because, when considering the intermediate covariance, the eigenvalues are functions of the sample matrix X. Hence, there is no guarantee that $\lambda_1(\breve{\boldsymbol{\Sigma}}_k) > \lambda_2(\breve{\boldsymbol{\Sigma}}_k)$ holds strictly. However, under the prerequisite that $\|\Delta_{k-1}\|_F \lesssim O\left(\sqrt{\frac{s\log p}{n}}\right)$, we can set n to be large enough to ensure a desired closenss between $\lambda_i(\breve{\boldsymbol{\Sigma}}_k)$ and $\lambda_{i+k-1}^\star$, implied by Mirsky's Theorem(Lemma 5). Specifically we can guarantee that

$$|\lambda_i^\star - \lambda_1(\breve{\boldsymbol{\Sigma}}_i)| < \frac{\lambda_i^\star - \lambda_{i+1}^\star}{3}, |\lambda_{i+1}^\star - \lambda_2(\breve{\boldsymbol{\Sigma}}_i)| < \frac{\lambda_i^\star - \lambda_{i+1}^\star}{3}.$$

Then we can formally mimic the proof of Lemma 1 to obtain the above results.

Now for further components we have the counterpart for Lemma 2. Again there is a gap between these generalized counterparts and the original first-component case, which we discussed in detail in our proof section.

Lemma 4. *For some* $1 \leq i < k$, *we assume that, for* $1 \leq j \leq i$,

$$\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \lesssim O\left(\sqrt{\frac{s\log p}{n}}\right), \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O\left(s\sqrt{\frac{\log p}{n}}\right).$$

Define $K := \sup_{\|u\|_2=1}\|Xu\|_{\psi_2}$. *Fix a* $\boldsymbol{\beta}^\star$ *in* $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 < \eta\}$. *For* $t > 0$ *we also define*

$$\begin{aligned}
M_0(\log 2p) &:= 2K^2\left(1 + \frac{2\lambda_1^\star}{\sigma}\right)\left[\frac{6 + 2\log(2p)}{c_K n}(1 + (2\zeta)^{-1}) + \zeta\right], \\
M_1(\log 2p) &:= 2K^2(1 + \frac{2\lambda_1}{\sigma})\left\{\frac{2(6+\log(2p))}{c_K n} + \frac{6+\log(2p)}{\zeta c_K n}\right\}, \\
M_2(\log 2p) &= 2\left(K^2\frac{2\log 2p}{c_K n} + K^2\sqrt{\frac{2\log 2p}{c_K n}}\right), \\
M_3(\log 2p) &= 4K^2\frac{2\log 2}{c_K n} + 4K^2\sqrt{\frac{2\log 2}{c_K n}}.
\end{aligned}$$

Let $M_\epsilon = 8KM_1(\log 2p)L^3 + 3M_2(\log 2p) + 2M_3(\log 2p)$, $\gamma = 4M_0(\log 2p)$, *and*

$$\zeta = \left(32K^2(1 + \frac{2\lambda_1^\star}{\sigma})\right)^{-1}.$$

When the sample size n *satisfies*

$$n > \zeta^{-1}(1 + \zeta^{-1})\frac{6 + 2\log 2p}{c_K},$$

$\gamma < 1$, *and*

$$\begin{aligned}&\left|(\nabla \widehat{R}_{i+1}(\boldsymbol{\beta}) - \nabla \breve{R}_{i+1}(\boldsymbol{\beta}))^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\right| \le M_\epsilon\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 \\ &+ \gamma\frac{\left\{(1 - 14\eta - 4\eta^2)(2\lambda_{i+1}^\star + \lambda_{i+2}^\star) - (1 + 2\eta)(\lambda_{i+1}^\star + 2\lambda_{i+2}^\star)\right\}}{6}\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_2^2.\end{aligned}$$

In light of these lemmas, we are ready to prove an asymptotic bound for further PCs:

Theorem 2 (Oracle inequalities for further components)**.** *For some* $1 \le i < k$, *we assume that, for* $1 \le j \le i$,

$$\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \lesssim O\left(\sqrt{s\frac{\log p}{n}}\right),\ \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O\left(s\sqrt{\frac{\log p}{n}}\right).$$

Let $\tilde{\boldsymbol{\beta}}_{i+1}$ *be the* $(i+1)$*-th stationary point of the optimization problem and* $s = \max\{s_1, \cdots, s_k\}$. *Suppose the conditions and notations in Lemma 4 are inherited. Now pick a penalty factor* $\rho = 2M_\epsilon$. *With a* $0 < \delta < 1$, *with probability at least* $1 - 3\cdot(2p)^{-1}$, *we have*

$$\begin{aligned}&\frac{\delta\rho_{i+1}}{2}\|\tilde{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star\|_1 + \breve{R}_{i+1}(\tilde{\boldsymbol{\beta}}_{i+1}) \\ \le&\breve{R}_{i+1}(\boldsymbol{\beta}_{i+1}^\star) + \frac{3(3+\delta)^2 s\rho_{i+1}^2}{2(1-\gamma)\left\{(1 - 14\eta - 4\eta^2)(2\lambda_{i+1}^\star + \lambda_{i+2}^\star) - (1 + 2\eta)(\lambda_{i+1}^\star + 2\lambda_{i+2}^\star)\right\}}.\end{aligned} \tag{16}$$

Now one can easily translate this conditional oracle inequality into a joint guarantee for the first k extracted PCs:

Corollary 3. *Under Assumption 2, consider our estimated PCs, with a sequence of penalty factor* $\rho_i, i = 1, \cdots, k$ *chosen in order using Lemma 4. With probability at least* $(1 - 3\cdot(2p)^{-1})^k$, *we have*

$$\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star\|_2 = O\left(\sqrt{s\frac{\log p}{n}}\right),\ \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star\|_1 = O\left(s\sqrt{\frac{\log p}{n}}\right),\ \textit{for } i = 1, \cdot, k$$

and

$$\|\widehat{\boldsymbol{Q}}_k\widehat{\boldsymbol{Q}}_k^\top - \boldsymbol{Q}_k\boldsymbol{Q}_k^\top\|_F = O\left(\sqrt{s\frac{\log p}{n}}\right).$$

As a concluding comment, note these bounds match the minimax rates derived in Birnbaum et al. (2013), if we focus on the generic setting therein; that is, assume the population matrix is given by

$$\Sigma_1^\star = \sum_{i=1}^{k} \lambda_i \beta_i^\star \beta_i^{\star\top} + \sigma^2 I.$$

Our setting is more general than this setup. Nevertheless, this spike model has been studied widely due to its simplicity and typicality; see, for example, Ma (2013b); Birnbaum et al. (2013) among others.

5 Numerical Experiments

In this section we provide Monte Carlo simulation to demonstrate the effectiveness of our sparse PCA scheme. We consider three different experiments adapted from other authors' proposal so that we can better compare with existing methods. The candidate competing method include iterative thresholding SPCA(ITSPCA, Ma (2013b)), diagonal thresholding SPCA(DTSPCA, Johnstone and Lu (2009)), augmented SPCA(AUGSPCA, Birnbaum et al. (2013)), correlation augmented SPCA(CORSPCA, Nadler (2009)), Fantope projection & selection(Fantope, Vu et al. (2013)) and semi-definite programming(ADAL, Ma (2013a)).

Since we are considering nonconvex penalized optimization, three main issues are standing ahead: the parameter tuning, the initialization and the algorithm for solving the penalized minimization problem. For parameter tuning, it is generally not easy to propose theoretically guaranteed data driven schemes. In our simulations we try out different choices of penalty factors based on our theoretical results and show that our choices of parameter lead to satisfactory numerical performance. For other methods, we follow the choice in the original codes if available(including ITSPCA, DTSPCA, AUGSPCA, CORSPCA); if not we also finalize the tuning based on several rounds of burn-in simulation trials based on the theoretical values(including Fantope and ADAL).

Besides, for the choice of a proper initialization point, we consider two candidates in our simulation: DTSPCA estimator and the semi-definite programming(ADAL) estimator, which are both consistent. DTSPCA does not converge in optimal rates; to this end our theoretical analysis has shown that our estimation scheme can take one step forward and improve this rate. Even though semi-definite programming could guarantee optimality for the first PC, as shown by Amini and Wainwright (2008) and Vu et al. (2013), it is not clear whether further PCs obtained via deflation are also optimal, hence it is of particular interest to also include ADAL in our consideration.

Lastly, we implement a proximal gradient optimization method (Nesterov, 2013) to minimize the objective function, which is composed of a smooth part and a nonsmooth one. The intuition of this algorithm is to approximate the smooth part by a quadratic function, which then gives a

closed-form solution for ℓ_1 penalization using soft thresholding operator. This idea has also been popularized in many other penalization optimization problems (Agarwal et al., 2010; Rennie and Srebro, 2005; Ji and Ye, 2009). For more details see Nesterov (2013) and our MATLAB codes.

From the above standpoints, we consider three general models in the following sections, with details provided in each section. Our numerical results are averaged over 100 repetitions for each model and presented in Table 1-6 and Figure 1-4.

5.1 Recovery of principal subspace(projection matrix)

For the first experiment we include two synthetic models(Model 1 and Model 2), generated in the same way as Vu et al. (2013). Specifically speaking, for Model 1, we sample $n = 100$ i.i.d. observations from a normal distribution, $\mathcal{N}_p(0, \mathbf{\Sigma}^\star)$, $p = 200$. The population covariance matrix $\mathbf{\Sigma}^\star$ is constructed in the form of $\mathbf{\Sigma}^\star = \alpha\Pi + (\boldsymbol{I} - \Pi)\mathbf{\Sigma}_0(\boldsymbol{I} - \Pi)$., where $\mathbf{\Sigma}_0$ is a Wishart matrix with p degrees of freedom and $\alpha > 0$ is a constant to adjust the "effective noise level" (Vu et al., 2013), $\sigma^2 = \sqrt{\lambda_1^\star \lambda_2^\star}/(\lambda_d^\star - \lambda_{d+1}^\star) \in \{1, 10\}$. As to $\mathbf{\Pi} = \boldsymbol{B}\boldsymbol{B}^\top$, $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ is the first d sparse eigenvectors. The sparsity follows a *disjoint* pattern, i.e. the support set of the five vectors are disjointed, with support size $s \in \{10, 25\}$. The nonzero entries come from a standard normal distribution. For Model 2, it follows the same mechanism as the first one, except that, the sparsity pattern in this model is set to be *shared*, i.e., the nonzeros elements in V are aligned. To measure the performance, for each model we consider two criteria: the Frobenius error of projection matricx estimation (Vu et al., 2013) $\|\widehat{\boldsymbol{Q}}_k\widehat{\boldsymbol{Q}}_k^\top - \boldsymbol{Q}_k\boldsymbol{Q}_k^\top\|_F^2$, and the subspace distance $\|\widehat{\boldsymbol{Q}}_k\widehat{\boldsymbol{Q}}_k^\top - \boldsymbol{Q}_k\boldsymbol{Q}_k^\top\|_{\text{op}}^2$ (Ma, 2013b).

The results are presented in Table 1. We make several interesting comparisons. First, from the table, Fantope has greater advantages over other methods in dealing with these equal-eigenvalue models. Although this setting is not pursed in our theoretical analysis, our method still works well, especially for the disjoint-support sparsity pattern, which outperforms several other competing methods in certain cases. Also, when comparing the initialization methods(DTSPCA and ADAL) and the corresponding PSPCA results, we see that our method improves the original subspace and Frobenius estimation error in almost all cases, which suggests that PSPCA can serve as a powerful post augmentation step for schemes with non-optimal performance.

5.2 Single-spike settings

In this experiment and the next one we follow the factor model studied by Ma (2013b), which formulates the following data generating model:

$$\boldsymbol{x}_i = \sum_{j=1}^{d} \sqrt{\lambda_j^\star} v_{ij} \boldsymbol{\beta}_j^\star + \boldsymbol{z}_i,\ i = 1, \ldots, n.$$

Here v_{ij} are i.i.d. standard normal random variables, which are independent of the white noise vector $\boldsymbol{z}_i \sim N(0, \sigma^2 \boldsymbol{I}_p)$, and $\boldsymbol{\beta}_j^\star$ are the first d eigenvectors, with $\lambda_j^\star$ being the corresponding

eigenvalues. Note this gives a population covariance matrix

$$\boldsymbol{\Sigma}^{\star} = \sum_{j=1}^{d} \lambda_j^{\star} \boldsymbol{\beta}_j^{\star} \boldsymbol{\beta}_j^{\star T} + \sigma^2 \boldsymbol{I}_p. \tag{17}$$

In our simulation we set $p = 2048$ and $n = 1024$, and take $d = 1$ for the current single spike settings. Specifically we consider two choices for $\boldsymbol{\beta}_1^{\star}$: the "Single Peak" model for Model 3 and the "Piecewise Polynomial" model for Model 4, which are both sparse wavelet signals under Symmlet 8 bases.

In Model 3, $\boldsymbol{\beta}_1^{\star}$ is generated from a single peak(SP) function(Figure 1(a)), i.e., we have $\boldsymbol{\beta}_1 = (f(1/p), \cdots, f(p/p))$. After transforming the data to the wavelet domain, $\boldsymbol{\beta}_1$ shows strong sparse pattern. Besides, the noise level is set to be $\sigma^2 = 1$, and the first eigenvalue ranges from $\lambda_1 \in \{100, 25, 10, 5\}$. In Model 4, we continue the single spike simulation, with all the settings remaining unchanged except for the choice of a less sparse $\boldsymbol{\beta}_1^{\star}$, the piecewise polynomial(PP) function, plotted in Figure 2(a).

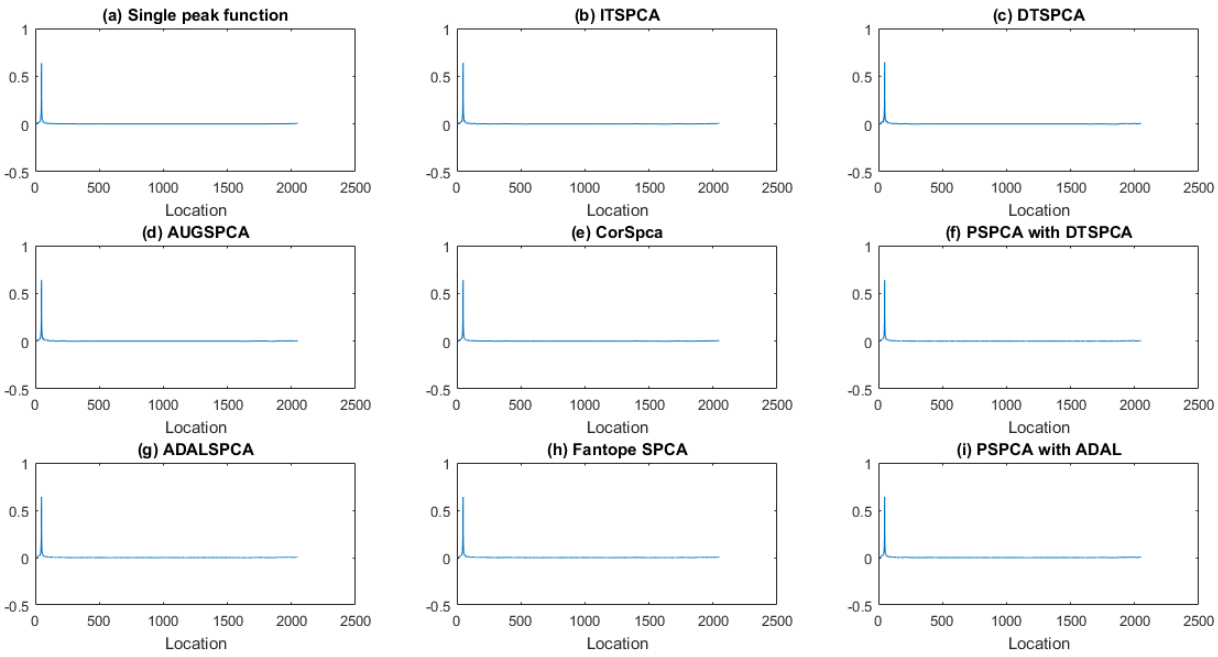

Figure 1: Single Peak Signal

In Table 2, we compare the subspace distance and the estimated support size as checked in Ma (2013b). To verify the results in the current work, the ℓ_1 and ℓ_2 loss of the single spike estimator are presented in Table 3. As to model selection results, Table 4 presents the false positive rate(FP, the proportion of true zero entries misspecified as nonzero) and the false negative rate(FN, the proportion of true nonzero entries misspecified as zero). Table 5 compares the averaged CPU time

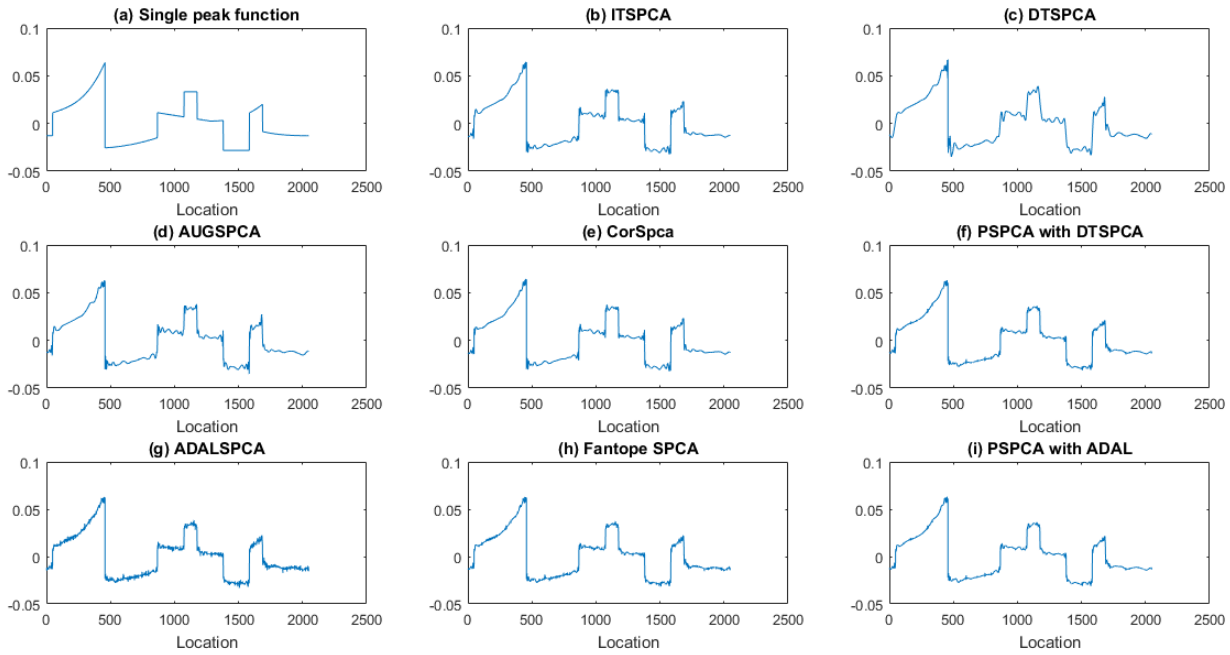

Figure 2: Piecewise Polynomial Signal

for each running. From the tables, we can see all methods achieved successful recovery. Similar to the previous simulation, one significant advantage of our method is that it improves the initialization methods significantly(like for DTSPCA under all cases, or ADAL under small eigenvalue settings). For many cases, ITSPCA and CORSPCA perform quite well, while our method also gives competing estimation. As for the sparsity specification, we see that the first four methods tend to give sparser representations. Generally ADAL are selecting too many features. PSPCA(DT) induces moderate sparsity instead.

We notice that in our algorithm implementation, the gradient methods sometimes give a quite small thresholding value due to the estimation of curvature, rendering too many small values in $\widehat{\boldsymbol{\beta}}$. From our simulation we see that this seems to be an issue for all penalized minimization methods, so in practice we suggest a post screening to filter these small remainders. Finally, from Table 5, we can see that the first four methods are more efficient generally, since they are based on straightforward matrix computation(spectral decomposition, QR decomposition, etc.) and simple thresholding, thus can be solved quickly using current well-optimized matrix computing packages. However, the rest four methods are all based on constrained optimization, which involves projection, matrix multiplication etc. and shows a slower performance. Nevertheless, our PSPCA still demonstrates much faster speed than other optimization-based methods, since we focus on solving a single vector each time while others' works try to directly estimate the projection matrices.

5.3 Multiple-spike settings

Our last two synthetic models, Model 5 and Model 6, focus on the multiple spike settings. In the factor model 17, we set $d = 4$, i.e. four sparse PCs. In Model 5, the eigenvectors $\boldsymbol{\beta}_1^\star, \cdots, \boldsymbol{\beta}_4^\star$ are generated respectively from the step function, piecewise polynomial function, three peak function and single peak function(Figure 3), with eigenvalues $(\lambda_1^\star, \lambda_2^\star, \lambda_3^\star, \lambda_4^\star) = (100, 75, 50, 25)$. In the last model, we use a different choice of eigenvalues: $(\lambda_1^\star, \lambda_2^\star, \lambda_3^\star, \lambda_4^\star) = (60, 55, 50, 45)$. Tabel 6 summarizes the subspace loss as well as the ℓ_1 , ℓ_2 error for each of the components. For the estimation error we can see that our PSPCA estimation scheme works pretty well. Especially for single-component ℓ_1 and ℓ_2 errors, PSPCA improved the initial estimate significantly and outperforming other methods in many cases. Also, we find that the accuracy of estimation is highly related to the eigenvalue gaps. Performance of individual eigenvector extraction would deteriorate if the spikes are not well separated(the second block of 6), resulting in high ℓ_1 and ℓ_2 error. That verifies the necessity of our separation assumption.

In summary, our simulation suggests the following facts:

1. PSPCA is a powerful sequential estimation scheme. It works as well as many other competing methods for both component-wise extraction and subspace approximation. This power is especially demonstrated in multiple-spike estimation. Also it can be solved via proximal gradient methods, which works more efficiently than other optimization-based estimators.

2. As a nonconvex estimation scheme, PSPCA requires an initial estimate with consistency. We could utilize PSPCA to improve the accuracy of the initializer significantly.

3. The effectiveness of PSPCA is highly related to the sparsity of the PCs as well as the gaps between eigenvalues. Sparse pattern and large gaps are required to recover individual PCs. But it is worth noting that subspace estimation is also achievable for small gaps.

5.4 Application to single-cell RNA sequence data

In this section, we applied several spca methods we have mentioned before to single-cell RNA sequence data. The scRNA-seq data matrix X($X \in \mathbf{R}^{n \times p}$) is a high-dimensional data matrix, with a dimension of about 10,000 and a sample size of only a few hundred. We consider 6 different single cell data sets from Ting et al. (2014) (Pollen et al., 2014) Treutlein et al. (2014) ,Schlitzer et al. (2015)Deng et al. (2014)Buettner et al. (2015), and each dataset includes cells with known labels. We compared the performance of spca methods in six different single-cell datasets and summarized them in the table. We do k-means clustering based on the spca results, and then compare the NMI scores of different methods based on known labels as following:

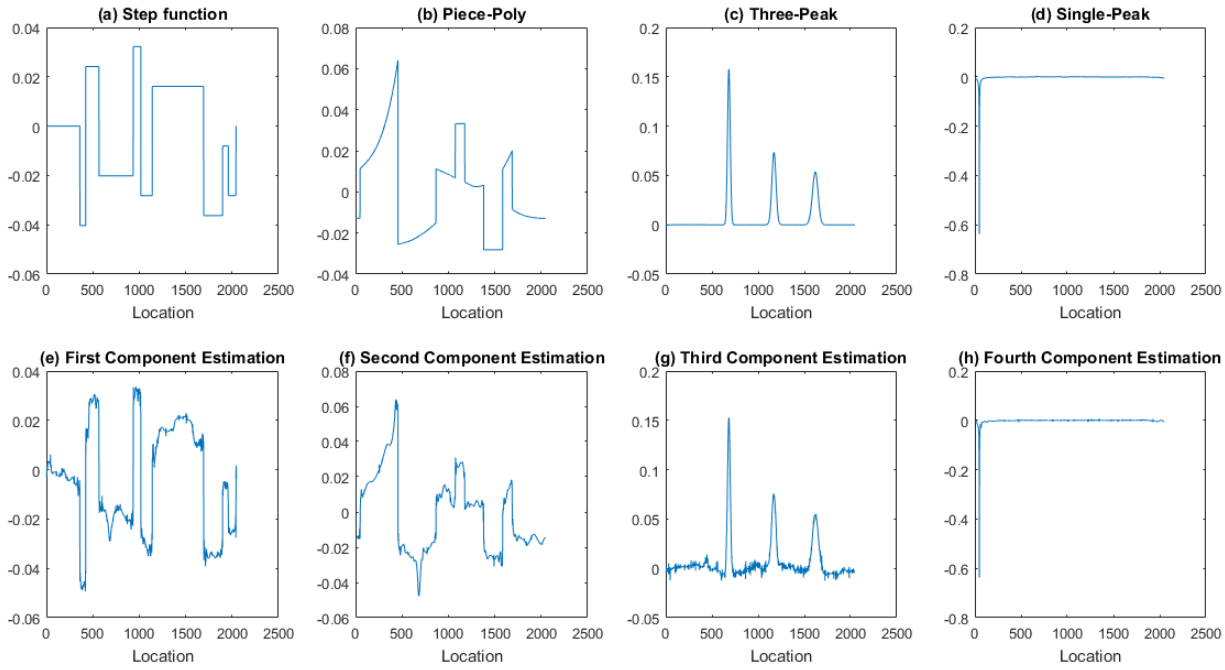

Figure 3: PSPCA with DTSPCA initialization

Table 7: NMI scores for the single-cell data sets. Higher values indicate better performance

DATASET	DENG	TREUTLEIN	TING	GINHOUX	BUENTTNER	POLLEN
DTSPCA	0.719	0.546	0.567	0.444	0.377	0.479
ASPCA	0.705	0.410	0.618	0.444	0.414	0.498
PSPCA(DT)	0.690	0.517	0.572	0.444	0.398	0.508
CORRSPCA	0.690	0.276	0.599	0.422	0.385	0.495
AUGSPCA	0.724	0.415	0.572	0.376	0.365	0.490
ITSPCA	0.727	0.191	0.625	0.373	0.391	0.500

From the table, we can summarize the following points:

- Our approach performs stably overall and performs better than all other competing methods on some datasets (such as POLLEN), with no obvious poor performance.
- However, it is worth noting that the improvement of the initial value is not as obvious as in the simulation.

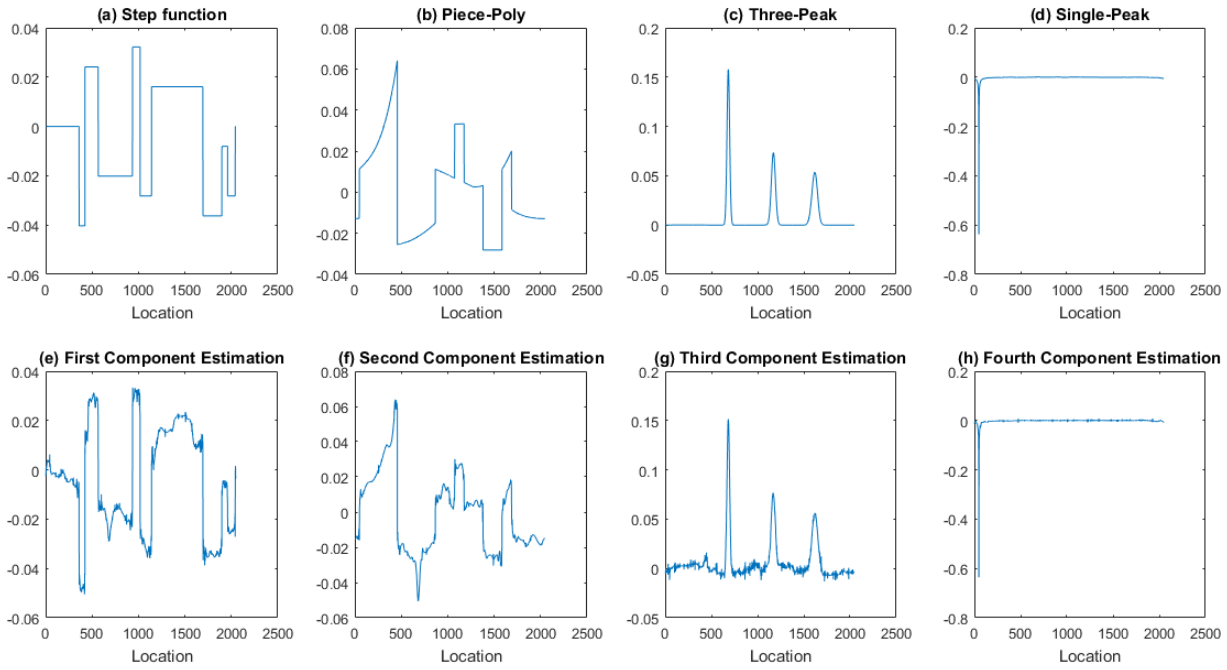

Figure 4: PSPCA with ADAL SPCA initialization

- Though all the methods can provide useful information in clustering, there is no method that can outperform in all situations consistently.

6 Collection of technical proofs

6.1 Proof of Theorem 1

At the beginning we do some simple summary and calculation based on (5). The empirical risk(loss) we study is

$$\widehat{R}_1(\boldsymbol{\beta}) = \frac{1}{2(n-1)} \sum_{i=1}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{x}_i \right\|_2^2 \tag{18}$$

$$= \frac{1}{2(n-1)} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{x}_i)^\top (\boldsymbol{x}_i - \boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{x}_i) \tag{19}$$

$$= \frac{1}{2(n-1)} \sum_{i=1}^{n} (\boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{\beta}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \|\boldsymbol{\beta}\|_2^2 \boldsymbol{\beta}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta}) \tag{20}$$

$$= \frac{1}{2} \left[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_1) + (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{\beta} \right], \tag{21}$$

which corresponds to the theoretical risk is

$$R_1(\boldsymbol{\beta}) = \mathbb{E}\left\{ \widehat{R}_1(\boldsymbol{\beta}) \right\} = \frac{1}{2}[\mathrm{tr}(\boldsymbol{\Sigma}_1^\star) + (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta}]. \tag{22}$$

The first derivative of $R(\boldsymbol{\beta})$ is

$$\nabla R_1(\boldsymbol{\beta}) = (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} + (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\boldsymbol{\beta}, \tag{23}$$

and the second derivative is

$$\nabla^2 R_1(\boldsymbol{\beta}) = (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\Sigma}_1^\star + 2\boldsymbol{\Sigma}_1^\star \boldsymbol{\beta}\boldsymbol{\beta}^\top + (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\boldsymbol{I}_p + 2\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star. \tag{24}$$

Proof. At the beginning we need to incorporate in the intialization point $\bar{\boldsymbol{\beta}}$ for asymptotic cases by adjusting the ℓ-2 radius η. Note that $\|\bar{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = O(\xi_n) = o(1)$, when n is large enough, we have $\|\bar{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 < \frac{\eta}{3}$. Now we have

$$\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_1\|_2 \le \frac{2\eta}{3}\} \subset \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \le \eta\}.$$

Suppose $\tilde{\boldsymbol{\beta}}_1$ is the unnormalized stationary point, solved from (7), satisfying (12):

$$\nabla \widehat{R}_1(\tilde{\boldsymbol{\beta}}_1)^\top (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1) + \lambda\|\boldsymbol{\beta}\|_1 - \lambda\|\tilde{\boldsymbol{\beta}}_1\|_1 \ge 0, \tag{25}$$

for all $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_1\|_2 \le \frac{2\eta}{3}, \|\boldsymbol{\beta}\|_1 \le L\}$. Let

$$\mathrm{Rem}_1(\tilde{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) = R_1(\boldsymbol{\beta}) - R_1(\tilde{\boldsymbol{\beta}}_1) - \nabla R_1(\tilde{\boldsymbol{\beta}}_1)(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1).$$

Using Lemma 1, and Taylor's expansion we have for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_2 < \eta\}$

$$\mathrm{Rem}_1(\tilde{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) \ge \frac{\{(1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star\}}{2} \left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1\right\|_2^2.$$

Now for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_1\|_2 \le \frac{2\eta}{3}, \|\boldsymbol{\beta}\|_1 \le L\}$, using (25) we have

$$\begin{aligned}
&- \nabla R_1(\tilde{\boldsymbol{\beta}}_1)^\top (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1) + \frac{\delta\rho}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 \\
\le &\big(\nabla \widehat{R}_1(\tilde{\boldsymbol{\beta}}_1) - \nabla R_1(\tilde{\boldsymbol{\beta}}_1)\big)^\top (\boldsymbol{\beta}_1^\star - \tilde{\boldsymbol{\beta}}_1) + \rho\|\boldsymbol{\beta}\|_1 - \rho\|\tilde{\boldsymbol{\beta}}_1\|_1 + \frac{\delta\rho}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1.
\end{aligned}$$

By Lemma 2 and the Decomposibility of ℓ_1 norm, we have

$$\begin{aligned}
&\big(\nabla \widehat{R}_1(\tilde{\boldsymbol{\beta}}_1) - \nabla R_1(\tilde{\boldsymbol{\beta}}_1)\big)^\top (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1) \\
\le &M_\epsilon\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 + \gamma\frac{\{(1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star\}}{2} \left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1\right\|_2^2 \\
= &M_\epsilon\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_S\|_1 + M_\epsilon\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_{S^c}\|_1 + \gamma\frac{\{(1 - 14\eta - 4\eta^2)\lambda_1^\star - (1 + 2\eta)\lambda_2^\star\}}{2} \left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1\right\|_2^2.
\end{aligned}$$

Using decomposibility of ℓ_1 norm and the triangle inequality, for an index set S with cardinal s, we can obtain

$$\begin{aligned}
&\rho\|\boldsymbol{\beta}\|_1 - \rho\|\tilde{\boldsymbol{\beta}}_1\|_1 + \delta M_\epsilon\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 \\
\le &\rho\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_S\|_1 + \rho\|\boldsymbol{\beta}_{S^c}\|_1 - \rho\|(\tilde{\boldsymbol{\beta}}_1)_{S^c}\|_1 + \delta M_\epsilon\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 \\
\le &\rho\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_S\|_1 + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1 - \rho\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_{S^c}\|_1 + \delta M_\epsilon\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1.
\end{aligned}$$

Recall our choice of penalty factor, $\rho = 2M_\epsilon$. The above results thus give

$$\begin{aligned}
&- \nabla R_1(\tilde{\boldsymbol{\beta}}_1)^\top(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1) + \delta M_\epsilon \|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 \\
&\leq \frac{3+\delta}{2}\rho\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_S\|_1 - \frac{1-\delta}{2}\rho\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_{S^c}\|_1 + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1 \\
&+ \gamma\frac{\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}{2}\left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1\right\|_2^2 \\
&\leq \frac{(3+\delta)\sqrt{s}}{2}\rho\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2 + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1 \\
&+ \gamma\frac{\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}{2}\left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1\right\|_2^2,
\end{aligned}$$

where for the last inequality we apply $\|(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})_S\|_1 \leq \sqrt{s}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2$. Following this, let c_0 be the curvature parameter:

$$c_\eta = 2^{-1}\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\},$$

by Fanchel's inequality we further have

$$\begin{aligned}
\frac{(3+\delta)\sqrt{s}}{2}\rho\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2 &= \frac{(3+\delta)\sqrt{s}\rho}{2\sqrt{2c_\eta(1-\gamma)}} \cdot \sqrt{2c_\eta(1-\gamma)}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2 \\
&\leq \frac{(3+\delta)^2\rho^2 s}{16c_\eta(1-\gamma)} + c_\eta(1-\gamma)\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2^2.
\end{aligned}$$

Hence we have

$$\begin{aligned}
&R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}) + \mathrm{Rem}_1(\tilde{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) + \delta M_\epsilon\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 \\
\leq &\frac{(3+\delta)^2\rho^2 s}{16c_\eta(1-\gamma)} + c_\eta(1-\gamma)\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2^2 + c_\eta\gamma\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|^2 + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1 \\
\leq &\frac{(3+\delta)^2\rho^2 s}{16c_\eta(1-\gamma)} + c_\eta\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_2^2 + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1.
\end{aligned}$$

Now using again Lemma 1:

$$\mathrm{Rem}_1(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq c_\eta\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2^2 = \frac{\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}{2}\left\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_1\right\|_2^2,$$

we conclude that

$$\begin{aligned}
&R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}) + \frac{\rho\delta}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_1 \\
\leq &\frac{(3+\delta)^2\rho^2 s}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}} + 2\rho\|\boldsymbol{\beta}_{S^c}\|_1.
\end{aligned}$$

□

6.2 Proof of Corollary 1

Proof. In Theorem 1, we take $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_1^\star$, and S as the support of $\boldsymbol{\beta}_1^\star$, then we have

$$R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}_1^\star) + \frac{\rho\delta}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \leq \frac{(3+\delta)^2\rho^2 s_0}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}.$$

Note the last term vanishes since $\|(\boldsymbol{\beta}_1^\star)_{S^c}\|_2 = 0$. Besides, using Lemma 1 we have

$$R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}_1^\star) \geq \frac{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2^2.$$

Note that from the expression of M_ϵ in Lemma 2, we have that $\rho = 2M_\epsilon \asymp \sqrt{\frac{\log p}{n}}$. Therefore,

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \leq \frac{(3+\delta)^2\rho s_0}{4\delta(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}} \asymp s_0\sqrt{\frac{\log p}{n}},$$
$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 \leq \frac{(3+\delta)\rho\sqrt{s_0}}{2\sqrt{(1-\gamma)}[(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star]} \asymp \sqrt{\frac{s_0\log p}{n}}.$$

Furthermore, let $\widehat{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1/\|\tilde{\boldsymbol{\beta}}_1\|_2$. For ℓ_2 norm,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = \left\|\frac{\tilde{\boldsymbol{\beta}}_1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{\boldsymbol{\beta}_1^\star}{\|\boldsymbol{\beta}_1^\star\|_2}\right\|_2 \\
=&\left\|\frac{\tilde{\boldsymbol{\beta}}_1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{\tilde{\boldsymbol{\beta}}_1}{\|\boldsymbol{\beta}_1^\star\|_2} + \frac{\tilde{\boldsymbol{\beta}}_1}{\|\boldsymbol{\beta}_1^\star\|_2} - \frac{\boldsymbol{\beta}_1^\star}{\|\boldsymbol{\beta}_1^\star\|_2}\right\|_2 \\
\leq&\frac{\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2}{\|\boldsymbol{\beta}_1^\star\|_2} + \left|\frac{1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{1}{\|\boldsymbol{\beta}_1^\star\|_2}\right|\|\tilde{\boldsymbol{\beta}}_1\|_2 \\
\leq&\frac{2\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2}{\|\boldsymbol{\beta}_1^\star\|_2} = O(\sqrt{\frac{s_0\log p}{n}}).
\end{aligned}$$

And for ℓ_1 norm,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 = \left\|\frac{\tilde{\boldsymbol{\beta}}_1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{\boldsymbol{\beta}_1^\star}{\|\boldsymbol{\beta}_1^\star\|_2}\right\|_1 \\
=&\left\|\frac{\tilde{\boldsymbol{\beta}}_1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{\boldsymbol{\beta}_1^\star}{\|\tilde{\boldsymbol{\beta}}_1\|_2} + \frac{\boldsymbol{\beta}_1^\star}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{\boldsymbol{\beta}_1^\star}{\|\boldsymbol{\beta}_1^\star\|_2}\right\|_1 \\
\leq&\frac{\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} + \left|\frac{1}{\|\tilde{\boldsymbol{\beta}}_1\|_2} - \frac{1}{\|\boldsymbol{\beta}_1^\star\|_2}\right|\|\boldsymbol{\beta}_1^\star\|_1 \\
\leq&\frac{\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 + \sqrt{s_0}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2}{\|\tilde{\boldsymbol{\beta}}_1\|_2} \\
\leq&\frac{\sqrt{s_0}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 + \|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1}{\|\boldsymbol{\beta}_1^\star\|_2 - \|\boldsymbol{\beta}_1^\star - \tilde{\boldsymbol{\beta}}_1\|_2} = O(s_0\sqrt{\frac{\log p}{n}}).
\end{aligned}$$

Last, for the projection estimator, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_1^\top - \boldsymbol{\beta}_1\boldsymbol{\beta}_1^\top\|_F &= \|\widehat{\boldsymbol{\beta}}_1(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star)^\top + (\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star)\boldsymbol{\beta}_1^{\star T}\|_F \\ &\le 2\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = O(\sqrt{s_0\frac{\log p}{n}}).\end{aligned}$$

□

6.3 Proof of Corollary 2

Proof. In Theorem 1, we take $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_1^\star$, then for some S with cardinal $s^\star$, we have

$$R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}_1^\star) + \frac{\rho\delta}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \le \frac{(3+\delta)^2\rho^2 s^\star}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}} + 2\rho\|(\boldsymbol{\beta}_1^\star)_{S^c}\|_1.$$

Now we need a trade-off between the two terms on the RHS of the above result. Recall that $\boldsymbol{\beta}_1^\star \in \mathbb{B}_q(s_q)$ for some $0 < q < 1$, i.e.,

$$\sum_{k=1}^p |\beta_{1k}^\star|^q \le s_q.$$

We pick a thresholding value τ, and set S to be the indices where $|\beta_{1k}^\star| > \tau$. Note that this gives

$$s^\star\tau^q \le s_q,$$

hence $s^\star \le s_q\tau^{-q}$. On the other hand, since $q < 1$,

$$\sum_{k\in S^c} |\beta_{1k}^\star/\tau| \le \sum_{k\in S^c} |\beta_{1k}^\star/\tau|^q \le s_q\tau^{-q},$$

which gives

$$\|(\boldsymbol{\beta}_1^\star)_{S^c}\|_1 \le s_q\tau^{1-q}.$$

Now set τ to be

$$\tau = \frac{(3+\delta)^2\rho}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}},$$

we obtain the following bound:

$$R_1(\tilde{\boldsymbol{\beta}}_1) - R_1(\boldsymbol{\beta}_1^\star) + \frac{\rho\delta}{2}\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \le 3\left[\frac{(3+\delta)^2}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}\right]^{1-q} s_q\rho^{2-q}.$$

Note that from the expression of M_ϵ in Lemma 2, we have that $\rho = 2M_\epsilon \asymp \sqrt{\frac{\log p}{n}}$.

Similar to the analysis of last proof, we can obtain

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 \le \frac{6}{\delta}\left[\frac{(3+\delta)^2}{8(1-\gamma)\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}\right]^{1-q} s_q\rho^{1-q} = O\left(s_q\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right),$$

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 \le \sqrt{6}\left[\frac{(3+\delta)^2}{8(1-\gamma)}\right]^{\frac{1-q}{2}} \frac{1}{\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}^{1-q/2}}\sqrt{s_q}\rho^{1-q/2} = O\left(\sqrt{s_q}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q}{4}}\right).$$

Furthermore, let $\widehat{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_1/\|\tilde{\boldsymbol{\beta}}_1\|_2$. Following a similar argument to the last proof, we can take one step forward and show

$$\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_1 = O\left(s_q\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right), \|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^\star\|_2 = O\left(\sqrt{s_q}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q}{4}}\right), \|\Delta_1\|_F = O\left(\sqrt{s_q}\left(\frac{\log p}{n}\right)^{\frac{1}{2}-\frac{q}{4}}\right).$$

□

6.4 Proof of Theorem 2 and Corollary 3

Now the following theorems are famous results in matrix perturbation theory (Stewart and Sun, 1990).

Lemma 5 (Mirsky's Theorem)**.** *Let X and $\tilde{X}$ be matrices of the same dimensions with singular values*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p,$$
$$\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_p.$$

Then for any unitarily invariant norm $\|\cdot\|$,

$$\|\mathrm{Diag}(\tilde{\sigma}_i - \sigma_i)\| \leq \|\tilde{X} - X\|.$$

Lemma 6 (Davis and Kahan's Second SinΘ Theorem)**.** *Let A have the spectral resolution*

$$\begin{bmatrix} X_1^H \\ \\ X_2^H \end{bmatrix} A[X_1, X_2] = \mathrm{Diag}(L_1, L_2),$$

where $[X_1\ X_2]$ is unitary with $X_1 \in \mathbb{C}^{n\times k}$. Let $Z \in \mathbb{C}^{n\times k}$ have orthonormal columns, and for any Hermitian M of order k, let

$$R = AZ - ZM.$$

Let $\mathcal{L}(P)$ be the set of eigenvalues of P, and $\mathcal{R}(\cdot)$ be the column space of P. Suppose that

$$\mathcal{L}(M) \subset [a, b].$$

and that for some $\delta > 0$,

$$\mathcal{L}(L_2) \subset \mathbb{R}\backslash[a - \delta, b + \delta].$$

Then for any unitarily invariant norm

$$\|\sin\Theta[\mathcal{R}(X_1), \mathcal{R}(Z)]\| \leq \frac{\|R\|}{\delta}.$$

Here $\sin\Theta[\mathcal{R}(X_1), \mathcal{R}(Z)]$ is the sines of canonical angles between $\mathcal{X}_\infty$ and $\mathcal{R}(Z)$, defined as the singular values of $X_1^H Z$.

Target: Let $s = \max\{s_0, \cdots, s_k^\star\}$. For $i = 1, \cdots, k$, show that for the i-th estimator, we have

1. $\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star\|_2 \lesssim O(\sqrt{s\frac{\log p}{n}})$;
2. $\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star\|_1 \lesssim O(s\sqrt{\frac{\log p}{n}})$;
3. $\|\widehat{\boldsymbol{Q}}_k\widehat{\boldsymbol{Q}}_k^\top - \boldsymbol{Q}_k\boldsymbol{Q}_k^\top\|_F \lesssim O(\sqrt{s\frac{\log p}{n}})$.

Base: We already prove the case when $k = 1$ in Theorem 1.

Hypothesis: Let $s = \max\{s_1, \cdots, s_k\}$. For some $1 \le i < k$, we assume that

1. $\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \lesssim O(\sqrt{s\frac{\log p}{n}})$, for $1 \le j \le i$.
2. $\|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O(s\sqrt{\frac{\log p}{n}})$, for $1 \le j \le i$.

We will use the second mathematical induction to deduce the target. We do this in the following steps:

I. Bounds in ℓ-2 norm: from $\widehat{\boldsymbol{\beta}}_j$ to $\widehat{\boldsymbol{q}}_j$.

1. $\|\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{\beta}_i^\star\boldsymbol{\beta}_i^{\star T}\|_2 \le 2\|\widehat{\boldsymbol{q}}_i - \boldsymbol{\beta}_i^\star\|_2$, for any $i \le k$.

Proof.

$$\begin{aligned}\|\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{\beta}_i^\star\boldsymbol{\beta}_i^{\star T}\|_2 =& \|\widehat{\boldsymbol{q}}_i(\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{\beta}_i^\star)^\top + (\widehat{\boldsymbol{q}}_i - \boldsymbol{\beta}_i^\star)\boldsymbol{\beta}_i^{\star T}\|_2\\ \le& \|\widehat{\boldsymbol{q}}_i(\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{\beta}_i^\star)^\top\|_2 + \|(\widehat{\boldsymbol{q}}_i - \boldsymbol{\beta}_i^\star)\boldsymbol{\beta}_i^{\star T}\|_2\\ \le& 2\|\widehat{\boldsymbol{q}}_i - \boldsymbol{\beta}_i^\star\|_2.\end{aligned}$$

□

2. For any i, j, $i > j$, $\|\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top\widehat{\boldsymbol{\beta}}_i\|_2 \le 2\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_2 + \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star\|_2$. Note that for any $\widehat{\boldsymbol{q}}_j$ satisfying the bound above, we have the result: for any $j < i$,

$$\begin{aligned}\|\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top\widehat{\boldsymbol{\beta}}_i\|_2 =& \|(\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top - \boldsymbol{\beta}_j^\star\boldsymbol{\beta}_j^{\star T})\widehat{\boldsymbol{\beta}}_i + \boldsymbol{\beta}_j^\star\boldsymbol{\beta}_j^{\star T}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star)\|_2\\ \le& \|(\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top - \boldsymbol{\beta}_j^\star\boldsymbol{\beta}_j^{\star T})\widehat{\boldsymbol{\beta}}_i\|_2 + \|\boldsymbol{\beta}_j^\star\boldsymbol{\beta}_j^{\star T}(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star)\|_2\\ \le& 2\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_2 + \|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^\star\|_2\end{aligned}$$

3. For all $j \le i$, $\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_2 \lesssim O(\sqrt{\frac{s\log p}{n}})$. To prove this we use a simple step of mathematical induction. Using the Base, when $j = 1$ apparently it is true. If we already obtain that for all $j < l$, $\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_2 \lesssim O(\sqrt{\frac{s\log p}{n}})$, then for l, we have

$$\|\widehat{\boldsymbol{\beta}}_l - \sum_{j=1}^{l-1}\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^\star\|_2 \le \|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l^\star\|_2 + \sum_{j=1}^{l-1}\|\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top\widehat{\boldsymbol{\beta}}_l\|_2 \lesssim O(\sqrt{\frac{s\log p}{n}}).$$

Now the normalization process will merely change the rate up to a constant. Hence we know that when n is large enough, this guarantees

$$\|\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star\|_2 \lesssim O(\sqrt{\frac{s\log p}{n}}).$$

Conclusion: Under the Hypothesis, for all $j \le i$, $\|\widehat{\boldsymbol{q}}_i - \boldsymbol{\beta}_i^\star\|_2$, $\|\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{\beta}_i^\star\boldsymbol{\beta}_i^{\star T}\|_2$, and $\|\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top\widehat{\boldsymbol{\beta}}_i\|_2$ all have the rate: $O(\sqrt{\frac{s\log p}{n}})$.

II. Bounds in ℓ_1 norm: from $\widehat{\boldsymbol{\beta}}_j$ to $\widehat{\boldsymbol{q}}_j$

1. We will use mathematical induction to prove the following results regarding ℓ_1 bound: for any $j \le i$, $\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O(s^j\sqrt{\frac{\log p}{n}})$, $\|\widehat{\boldsymbol{q}}_j\|_1 \lesssim \sqrt{s}$. Note that under Base, this is true for $j = 1$.

Now we assume that: for some $j \le i$, and all $l < j$, we have

$$\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O(s^j\sqrt{\frac{\log p}{n}}), \|\widehat{\boldsymbol{q}}_j\|_1 \lesssim \sqrt{s}.$$

2. First we show: $\|\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top - \boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T}\|_1 \le s^{l+1/2}\sqrt{\frac{\log p}{n}} = o(1)$, for any $l < j$.

$$\begin{aligned}
\|\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top - \boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T}\|_1 =& \|\widehat{\boldsymbol{q}}_l(\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star)^\top + (\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star)\boldsymbol{\beta}_l^{\star T}\|_1 \\
\le& \|\widehat{\boldsymbol{q}}_l(\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star)^\top\|_1 + \|(\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star)\boldsymbol{\beta}_l^{\star T}\|_1 \\
\le& \|\widehat{\boldsymbol{q}}_l\|_1\|\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star\|_1 + \|\widehat{\boldsymbol{q}}_l - \boldsymbol{\beta}_l^\star\|_1\|\boldsymbol{\beta}_l^\star\|_1 \\
\lesssim& s^{l+1/2}\sqrt{\frac{\log p}{n}} = o(1).
\end{aligned}$$

3. Next we show that: for all l, $l < j$, $\|\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top\widehat{\boldsymbol{\beta}}_j\|_1 \le O(s^{l+1}\sqrt{\frac{\log p}{n}})$. Note that for any $\widehat{\boldsymbol{q}}_l$ satisfying the bound above, we have the result: for any $l < j$,

$$\begin{aligned}
\|\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top\widehat{\boldsymbol{\beta}}_j\|_1 =& \|(\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top - \boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T})\widehat{\boldsymbol{\beta}}_j + \boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star)\|_1 \\
\le& \|(\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top - \boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T})\widehat{\boldsymbol{\beta}}_j\|_1 + \|\boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star)\|_1 \\
\le& \|(\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top - \boldsymbol{\beta}_l^\star\boldsymbol{\beta}_l^{\star T})\|_1\|\widehat{\boldsymbol{\beta}}_j\|_1 + \|\boldsymbol{\beta}_l^\star\|_1\|\boldsymbol{\beta}_l^\star\|_1\|(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star)\|_1 \lesssim O(s^{l+1}\sqrt{\frac{\log p}{n}}).
\end{aligned}$$

4. Finally we can move on to show that, for j, we have $\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O(s^j\sqrt{\frac{\log p}{n}})$, and $\|\widehat{\boldsymbol{q}}_j\|_1 \lesssim \sqrt{s}$, thus finishing the induction. This proves true by noting that

$$\|\widehat{\boldsymbol{\beta}}_j - \sum_{l=1}^{j-1}\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top\widehat{\boldsymbol{\beta}}_j\|_1 < \|\widehat{\boldsymbol{\beta}}_j\|_1 + o(1) \lesssim \sqrt{s}.$$

and that

$$\|\widehat{\boldsymbol{\beta}}_j - \sum_{l=1}^{j-1}\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_1 \le \|\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_1 + \sum_{l=1}^{j-1}\|\widehat{\boldsymbol{q}}_l\widehat{\boldsymbol{q}}_l^\top\widehat{\boldsymbol{\beta}}_i\|_1 \lesssim O(s^j\sqrt{\frac{\log p}{n}}).$$

Conclusion: From the above proof we see that for any $j \leq i$, $\|\widehat{\boldsymbol{q}}_j - \boldsymbol{\beta}_j^\star\|_1 \lesssim O(s^j\sqrt{\frac{\log p}{n}})$, $\|\widehat{\boldsymbol{q}}_j\|_1 \lesssim \sqrt{s}$.

III. $\|\breve{\boldsymbol{\Sigma}}_{i+1} - \boldsymbol{\Sigma}_{i+1}^\star\|_F \leq (2+2\sqrt{i}\lambda_1^\star)\|\Delta_i\|_F \lesssim O(\sqrt{\frac{s\log p}{n}})$.

1. For $\boldsymbol{\Sigma}_{i+1}^\star$, $\breve{\boldsymbol{\Sigma}}_{i+1}$, and Δ_i defined as the notations, we have the following inequality:

$$\|\boldsymbol{\Sigma}_{i+1}^\star - \breve{\boldsymbol{\Sigma}}_{i+1}\|_F \leq (2+2\sqrt{i}\lambda_1^\star)\|\Delta_i\|_F.$$

The proof is very straight forward.

$$\begin{aligned}\|\boldsymbol{\Sigma}_{i+1}^\star - \breve{\boldsymbol{\Sigma}}_{i+1}\|_F =& \|(\boldsymbol{I}-\boldsymbol{Q}_i\boldsymbol{Q}_i^\top)\boldsymbol{\Sigma}_1^\star(\boldsymbol{I}-\boldsymbol{Q}_i\boldsymbol{Q}_i^\top) - (\boldsymbol{I}-\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star(\boldsymbol{I}-\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\|_F \\ \leq& \|(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star\|_F + \|\boldsymbol{\Sigma}_1^\star(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\|_F \\ &+\|(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top)\boldsymbol{\Sigma}_1^\star(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top) - (\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star(\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\|_F.\end{aligned}$$

We know that

$$\|(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star\|_F \leq \lambda_1^\star\|\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top\|_F,$$

and it is the same for $\|\boldsymbol{\Sigma}_1^\star(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\|_F$.

Next we have

$$\begin{aligned}&\|(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top)\boldsymbol{\Sigma}_1^\star(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top) - (\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star(\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\|_F \\ \leq& \|(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top)\|_F + \|(\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star(\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{Q}_i\boldsymbol{Q}_i^\top)\|_F \\ \leq& \Big(\|(\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top)\boldsymbol{\Sigma}_1^\star\|_F + \|\boldsymbol{\Sigma}_1^\star(\boldsymbol{Q}_i\boldsymbol{Q}_i^\top)\|_F\Big)\|\widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top - \boldsymbol{Q}_i\boldsymbol{Q}_i^\top\|_F \\ \leq& 2\sqrt{i}\lambda_1^\star\|\boldsymbol{Q}_i\boldsymbol{Q}_i^\top - \widehat{\boldsymbol{q}}_i\widehat{\boldsymbol{q}}_i^\top\|_F.\end{aligned}$$

Thus the inequality follows naturally.

2. Now combining I. we can prove that

$$\|\breve{\boldsymbol{\Sigma}}_{i+1} - \boldsymbol{\Sigma}_{i+1}^\star\|_F \leq (2+2\sqrt{i}\lambda_1^\star)\|\Delta_i\|_F \lesssim O(\sqrt{\frac{s\log p}{n}}).$$

IV. A Bound for $\|\breve{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}_{i+1}^\star\|_2$

In I we have shown that $\|\Delta_i\|_F \lesssim O(\sqrt{s\log p/n})$. Next we move on to show that, if for some $i \leq k$, $\|\Delta_i\|_F \leq g_i\sqrt{s\log p/n}$ holds, then when n is large enough we have

$$\|\boldsymbol{\beta}_{i+1}^\star - \breve{\boldsymbol{\beta}}_{i+1}\|_2 \leq \frac{6(2+2\sqrt{i}\lambda_1^\star)g_i}{\lambda_{i+1}^\star - \lambda_{i+2}^\star}\sqrt{\frac{s\log p}{n}}.$$

Here g_i is a function of $\lambda_1^\star, \cdots, \lambda_{i+1}^\star$, thus a constant since we are considering fixed design for the eigenvalues.

Proof. We are going to use Davis and Kahan's Second SinΘ Theorem(Lemma 6) to complete our proof. Let $[\boldsymbol{\beta}_{i+1}^\star, \boldsymbol{X}_{i+1}]$ be any orthogonal matrix, then we have the spectral resolution

$$\begin{bmatrix} \boldsymbol{\beta}_{i+1}^\star \\ \boldsymbol{X}_{i+1} \end{bmatrix}^\top \boldsymbol{\Sigma}_{i+1}^\star \left[\boldsymbol{\beta}_{i+1}^\star, \boldsymbol{X}_{i+1}\right] = \begin{bmatrix} \lambda_{i+1}^\star & O \\ O & L_2 \end{bmatrix},$$

Besides, let $M = \lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1})$, and $\delta = \frac{\lambda_{i+1}^\star - \lambda_{i+2}^\star}{3}$. Using Mirsky's Theorem(by taking $\boldsymbol{X} = \boldsymbol{\Sigma}_{i+1}^\star$, $\tilde{\boldsymbol{X}} = \breve{\boldsymbol{\Sigma}}_{i+1}$ in Lemma 5), and the unitarily invariant norm as the Frobenius norm, we know that when n is large enough, we have

$$|\lambda_{i+1}^\star - \lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1})| < \delta,$$

and

$$|\lambda_{i+2}^\star - \lambda_2(\breve{\boldsymbol{\Sigma}}_{i+1})| < \delta.$$

Note that for $\boldsymbol{\Sigma}_{i+1}^\star$, all of its eigenvalues except for the largest one satisfies

$$\lambda_j^\star \le \lambda_{i+2}^\star < \lambda_{i+2}^\star + \delta < \lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1}) - \delta, \text{ for } j \ge i+2.$$

Using this fact we can see that, the eigenvalues of L_2 statisfy

$$\mathcal{L}_2 \subset (0, \lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1}) - \delta).$$

Now let $Z = \breve{\boldsymbol{\beta}}_{i+1}$, the leading eigenvector of $\breve{\boldsymbol{\Sigma}}_{i+1}$. $R = \boldsymbol{\Sigma}_{i+1}^\star Z - ZM = (\boldsymbol{\Sigma}_{i+1}^\star - \breve{\boldsymbol{\Sigma}}_{i+1})\breve{\boldsymbol{\beta}}_{i+1}$, and

$$\|R\|_F = \|(\boldsymbol{\Sigma}_{i+1}^\star - \breve{\boldsymbol{\Sigma}}_{i+1})\breve{\boldsymbol{\beta}}_{i+1}\|_F \le \|\boldsymbol{\Sigma}_{i+1}^\star - \breve{\boldsymbol{\Sigma}}_{i+1}\|_F \lesssim (2 + 2\sqrt{i}\lambda_1) g_i \sqrt{\frac{\log p}{n}}.$$

Finally applying Lemma 6 we get

$$\|\sin\Theta[\mathcal{R}(\boldsymbol{\beta}_{i+1}), \mathcal{R}(\breve{\boldsymbol{\beta}}_{i+1})]\|_F \le \frac{\|R\|_F}{\delta} \lesssim \frac{3(2 + 2\sqrt{i}\lambda_1) g_i}{\lambda_{i+1}^\star - \lambda_{i+2}^\star} \sqrt{\frac{s \log p}{n}}.$$

When n is large enough such that the right hand side is smaller than $\frac{\sqrt{2}}{2}$,

$$\|\boldsymbol{\beta}_{i+1}^\star - \breve{\boldsymbol{\beta}}_{i+1}\|_2 = 2\sin\frac{\Theta}{2} \le 2\sin\Theta \lesssim \frac{6(2 + 2\sqrt{i}\lambda_1) g_i}{\lambda_{i+1}^\star - \lambda_{i+2}^\star} \sqrt{\frac{s \log p}{n}}.$$

□

V. Strong Convexity for $\breve{R}_{i+1}$ over $\{\|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 \le \eta\}$, Lemma 3

1. Denote the smallest eigenvalue of $\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta})$ by $\lambda_{\min}(\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta}))$. Follow the proof of Lemma 1 we hope to show that, for some $\eta > 0$,

$$\lambda_{\min}(\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta})) \ge (1 - 14\eta - 4\eta^2)\lambda_{i+1}(\breve{\boldsymbol{\Sigma}}_{i+1}) - (1 + 2\eta)\lambda_{i+2}(\breve{\boldsymbol{\Sigma}}_{i+1}) > 0.$$

However, $\lambda_{i+1}(\breve{\boldsymbol{\Sigma}}_{i+1})$ and $\lambda_{i+2}(\breve{\boldsymbol{\Sigma}}_{i+1})$ are both subject to the sample, thus possibly equal, laying barriers for picking a valid η. Luckily, this can be solved observing the fact that under our Hypothesis, when n is large enough, eigenvalues of $\breve{\boldsymbol{\Sigma}}_{i+1}$ are located very close to those of $\boldsymbol{\Sigma}^{\star}_{i+1}$. To be specific, by Lemma 5 and III, with $\delta = \frac{\lambda^{\star}_{i+1}-\lambda^{\star}_{i+2}}{3}$ we have that

$$|\lambda^{\star}_{i+1} - \lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1})| < \delta,$$

and

$$|\lambda^{\star}_{i+2} - \lambda_2(\breve{\boldsymbol{\Sigma}}_{i+1})| < \delta.$$

Therefore, with a proper η we can have a more stable result:

$$\lambda_{\min}(\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta})) \geq (1-14\eta-4\eta^2)\frac{2\lambda^{\star}_{i+1}+\lambda^{\star}_{i+2}}{3} - (1+2\eta)\frac{\lambda^{\star}_{i+1}+2\lambda^{\star}_{i+2}}{3} > 0.$$

Furthermore, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 < \eta\}$,

$$\begin{aligned}
&\breve{R}_{i+1}(\boldsymbol{\beta}_1) - \breve{R}_{i+1}(\boldsymbol{\beta}_2) - \nabla \breve{R}_{i+1}(\boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \\
\geq &\frac{\left\{(1-14\eta-4\eta^2)(2\lambda^{\star}_{i+1}+\lambda^{\star}_{i+2}) - (1+2\eta)(\lambda^{\star}_{i+1}+2\lambda^{\star}_{i+2})\right\}}{6}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2.
\end{aligned}$$

VI. A Bound for $\left|\breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1}) - \breve{R}_{i+1}(\boldsymbol{\beta}^{\star}_{i+1})\right|$

1. We will show that under Hypothesis,

$$\left|\breve{R}_{i+1}(\boldsymbol{\beta}^{\star}_{i+1}) - \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1})\right| \leq \frac{11(4\lambda^{\star}_{i+1}-\lambda^{\star}_{i+2})}{3} \cdot (\frac{6(2+2\sqrt{i}\lambda^{\star}_1)g_i}{\lambda^{\star}_{i+1}-\lambda^{\star}_{i+2}})^2 \cdot \frac{s\log p}{n} \lesssim O(\frac{s\log p}{n}).$$

Proof. We will use Taylor's expansion to finish the proof. Note that $\breve{\boldsymbol{\beta}}_{i+1}$ is the global minimum point of $\breve{R}_{i+1}$, we have

$$\breve{R}_{i+1}(\boldsymbol{\beta}^{\star}_{i+1}) = \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1}) + \frac{1}{2}(\breve{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^{\star}_{i+1})^{\top}\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta}_{med})(\breve{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^{\star}_{i+1}),$$

where $\boldsymbol{\beta}_{med}$ is some point in $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 \leq \eta\}$.

Now we will build an upper bound for the largest eigenvalue of $\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta}_{med})$. Since

$$\begin{aligned}
\|\nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta})\|_{\text{op}} &= \left\|(\|\boldsymbol{\beta}\|_2^2 - 2)\breve{\boldsymbol{\Sigma}}_{i+1} + 2\breve{\boldsymbol{\Sigma}}_{i+1}\boldsymbol{\beta}\boldsymbol{\beta}^{\top} + (\boldsymbol{\beta}^{\top}\breve{\boldsymbol{\Sigma}}_{i+1}\boldsymbol{\beta})\boldsymbol{I}_p + 2\boldsymbol{\beta}\boldsymbol{\beta}^{\top}\breve{\boldsymbol{\Sigma}}_{i+1}\right\|_{\text{op}} \\
&\leq 2\|\breve{\boldsymbol{\Sigma}}_{i+1}\|_{\text{op}} + 4\|\breve{\boldsymbol{\Sigma}}_{i+1}\|_{\text{op}}\|\boldsymbol{\beta}\boldsymbol{\beta}^{\top}\|_{\text{op}} + \boldsymbol{\beta}^{\top}\breve{\boldsymbol{\Sigma}}_{i+1}\boldsymbol{\beta} \\
&\leq 2\lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1}) + 4\lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1})(1+\eta)^2 + \lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1})(1+\eta)^2 \\
&\leq 22\lambda_1(\breve{\boldsymbol{\Sigma}}_{i+1}) \leq 22(\lambda^{\star}_{i+1} + \frac{\lambda^{\star}_{i+1}-\lambda^{\star}_{i+2}}{3}) \\
&= \frac{22(4\lambda^{\star}_{i+1}-\lambda^{\star}_{i+2})}{3}.
\end{aligned}$$

Therefore, combining with IV, we have that

$$\begin{aligned}
&\big|\breve{R}_{i+1}(\boldsymbol{\beta}^\star_{i+1}) - \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1})\big| \\
=&\big|\frac{1}{2}(\breve{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star_{i+1})^\top \nabla^2 \breve{R}_{i+1}(\boldsymbol{\beta}_{med})(\breve{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star_{i+1})\big| \\
\leq&\frac{11(4\lambda^\star_{i+1} - \lambda^\star_{i+2})}{3}\|\breve{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star_{i+1}\|_2^2 \\
\leq&\frac{11(4\lambda^\star_{i+1} - \lambda^\star_{i+2})}{3} \cdot (\frac{6(2+2\sqrt{i}\lambda_1)g_i}{\lambda^\star_{i+1} - \lambda^\star_{i+2}})^2 \cdot \frac{s\log p}{n},
\end{aligned}$$

where in the last inequality we apply the result in IV. This shows that

$$\big|\breve{R}_{i+1}(\boldsymbol{\beta}^\star_{i+1}) - \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1})\big| \lesssim O(\frac{s\log p}{n}).$$

□

VII. Bound for the Linear Difference, Lemma 4

Define $K := \sup_{\|u\|_2=1}\|Xu\|_{\phi_2}$. Consider a $\boldsymbol{\beta}^\star$ in $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 < \eta\}$. For $t > 0$ we also define

$$\begin{aligned}
M_0(\log 2p) &:= 2K^2(1 + \frac{2\lambda^\star_1}{\lambda_{\min}})\left\{\frac{6 + 2\log(2p)}{c_K n}(1 + (2\zeta)^{-1}) + \zeta\right\}, \\
M_1(\log 2p) &:= 2K^2(1 + \frac{2\lambda_1}{\lambda_{\min}})\left\{\frac{2(6 + \log(2p))}{c_K n} + \frac{6 + \log(2p)}{\zeta c_K n}\right\}, \\
M_2(\log 2p) &= 2\bigg(K^2\frac{t + \log 2p}{c_K n} + K^2\sqrt{\frac{t + \log 2p}{c_K n}}\bigg), \\
M_3(\log 2p) &= 4K^2\frac{t + \log 2}{c_K n} + 4K^2\sqrt{\frac{t + \log 2}{c_K n}}.
\end{aligned}$$

$C > 0$ is a proper constant, and we let $M_\epsilon = 8CM_1(\log 2p))L^3 + 3M_2(\log 2p) + 2M_3(\log 2p)$. Pick a proper $\gamma < 1$. Then we have, with probablity at least $1 - 3\cdot(2p)^{-1}$, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 < \eta\}$ we have

$$\begin{aligned}
&\big|(\nabla\widehat{R}_{i+1}(\boldsymbol{\beta}) - \nabla\breve{R}_{i+1}(\boldsymbol{\beta}))^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\big| \leq M_\epsilon\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 \\
&+ \gamma\frac{\big\{(1 - 14\eta - 4\eta^2)(2\lambda^\star_{i+1} + \lambda^\star_{i+2}) - (1 + 2\eta)(\lambda^\star_{i+1} + 2\lambda^\star_{i+2})\big\}}{6}\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_2^2.
\end{aligned}$$

This proof is quite similar to the one for the first component. However, some crucial difference still exists here, which we demonstrate as follows:

1. For *Part I: Three Terms Splitting*, the process is similar since the structure of $\breve{R}_{i+1}$ and R_1 is pretty analogous, after adding projection matrices on both sides of the covariance difference. Therefore there is a counterpart three terms splitting for $\breve{R}_{i+1}$.

2. For *Part II: First Term Bounding*, we consider the coefficients before $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2$ and $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1$ respectively. For $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2$, note that by adding a projection factor the norm does not decrease; that is, for any othorgonal projection matrix P,

$$\|P(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_2^2 \leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2,$$

therefore, the coefficient before the ℓ_2 norm can be inherited from the base case. For $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1$, on the other hand, we need to consider the structure of the projection matrix we multiply. In our case,

$$\widehat{P}_i = I - \boldsymbol{\Sigma}_{j=1}^i \widehat{\boldsymbol{q}}_j \widehat{\boldsymbol{q}}_j^\top.$$

By our proof in II, we know that $\|\widehat{\boldsymbol{q}}_j\|_1 \lesssim L$, or, to be more specific, for some constant $C>0$, $\|\widehat{\boldsymbol{q}}_j\|_1 \leq C\cdot L$, thus

$$\begin{aligned}\|\widehat{P}_i(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_1 &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1 + \boldsymbol{\Sigma}_{j=1}^i \|\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_1 \\ &\leq (1+\boldsymbol{\Sigma}_{j=1}^i \|\widehat{\boldsymbol{q}}_j\widehat{\boldsymbol{q}}_j^\top\|_1)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_1 \\ &\leq (1+iC^2L^2)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_1.\end{aligned}$$

Consider that $L \asymp \sqrt{s}$, which we assume goes to infinity as n increases, we can redefine the C such that

$$\|\widehat{P}_i(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_1 \leq CL^2\|(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\|_1.$$

Now we can see that the coefficient before the ℓ_1 norm term should be refined by a factor CL^2. Meanwhile we should notice that in asymptotics this requires $L^3\sqrt{\frac{\log 2p}{n}} = o(1)$, which is also guaranteed by $s^i\sqrt{\frac{\log 2p}{n}} = o(1)$ when $i \geq 2$.

3. For *Part III: Second Term Bounding* and *Part IV: Third Term Bounding*, the concentration inequality we mainly used is the Bernstein type inequality, for which it is not hard to see that when multiplying a projection factor the bounds still hold without even the need to change the coefficient of the base case.

4. For *Part V: Combined Bounding*, the result shall be modified according to our comments above to obtain:

$$\begin{aligned}&\big|(\nabla\widehat{R}_{i+1}(\boldsymbol{\beta}) - \nabla\breve{R}_{i+1}(\boldsymbol{\beta}))^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\big| \\ \leq& 4M_0(\log 2p)\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2 \\ +&\big(8CL^3M_1(\log 2p) + 3M_2(\log 2p) + 2M_3(\log 2p)\big)\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1.\end{aligned}$$

5. For *Part VI: the EPC*, the selection of γ needs to be modified slightly since the result for strong convexity differs a little in terms of the coefficients.

After all the similar procedure the desired result can be proved.

VIII. A basic inequality for the $(i+1)$-th component, Theorem 16

Suppose our Hypothesis holds. Consider $\boldsymbol{\beta}^\star_{i+1}$ in $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 < \eta\}$ whose size of support set S is at most s. Let $\tilde{\boldsymbol{\beta}}_{i+1}$ be a stationary point of the optimization problem for the $(i+1)$-th component. Suppose the conditions in Lemma 4 are satisfied. Now let $\rho_{i+1} = 2M_\epsilon$. Then with probability at least $1 - 3 \cdot (2p)^{-1}$ we have

$$\begin{aligned}
&\frac{\delta\rho_{i+1}}{2}\|\tilde{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star\|_1 + \breve{R}_{i+1}(\tilde{\boldsymbol{\beta}}_{i+1}) \\
\leq &\breve{R}_{i+1}(\boldsymbol{\beta}^\star_{i+1}) + \frac{3(3+\delta)^2 s\rho_{i+1}^2}{2(1-\gamma)\left\{(1-14\eta-4\eta^2)(2\lambda^\star_{i+1}+\lambda^\star_{i+2}) - (1+2\eta)(\lambda^\star_{i+1}+2\lambda^\star_{i+2})\right\}}.
\end{aligned}$$

The proof again is quite similar to the one in the base case. We just need to replace R_1 by $\breve{R}_{i+1}$, and apply V and VI. we have proved to some mediate steps. The proof also involves to set n to be large enough so that $\boldsymbol{\beta}^\star_{i+1}$ and $\breve{\boldsymbol{\beta}}_{i+1}$ fall into the ball $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_{i+1}\| \leq \frac{2\eta}{3}\}$, which is possible since $\|\bar{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star_{i+1}\| = o(1)$ in a slower rate, and $\|\boldsymbol{\beta}^\star_{i+1} - \breve{\boldsymbol{\beta}}_{i+1}\|_2 = O(\sqrt{\frac{s\log p}{n}})$, implied by IV. Other changes mainly involves some modification in the coefficients which have no inherent influence on the structure of the proof.

IX. The final step towards $(i+1)$ case

Now we explore the asymptotic performance based on the conclusions we have proved. Using VII, we have with probability at least $1 - 3 \cdot (2p)^{-1}$

$$\begin{aligned}
&\frac{\delta\rho_{i+1}}{2}\|\tilde{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star\|_1 + \breve{R}_{i+1}(\tilde{\boldsymbol{\beta}}_{i+1}) \\
\leq &\breve{R}_{i+1}(\boldsymbol{\beta}^\star_{i+1}) + \frac{3(3+\delta)^2 s\rho_{i+1}^2}{2(1-\gamma)\left\{(1-14\eta-4\eta^2)(2\lambda^\star_{i+1}+\lambda^\star_{i+2}) - (1+2\eta)(\lambda^\star_{i+1}+2\lambda^\star_{i+2})\right\}}.
\end{aligned}$$

Note that from the picking of M_ϵ in VII,

$$M_\epsilon = O(\sqrt{\frac{\log p}{n}}).$$

According to VI we know that

$$\begin{aligned}
&\left|\breve{R}_{i+1}(\boldsymbol{\beta}^\star_{i+1}) - \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1})\right| \\
\leq &\frac{11(4\lambda^\star_{i+1} - \lambda^\star_{i+2})}{3} \cdot (\frac{6(2+2\sqrt{i}\lambda^\star_1)g_i}{\lambda^\star_{i+1} - \lambda^\star_{i+2}})^2 \cdot \frac{s\log p}{n},
\end{aligned}$$

By adding up the results above we obtain that

$$\frac{\delta\rho_{i+1}}{2}\|\tilde{\boldsymbol{\beta}}_{i+1} - \boldsymbol{\beta}^\star_{i+1}\|_1 + \breve{R}_{i+1}(\tilde{\boldsymbol{\beta}}_{i+1}) - \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1}) \lesssim O(\frac{s\log p}{n}).$$

$\breve{\boldsymbol{\beta}}_{i+1}$ is the global minimal point of $\breve{R}_{i+1}(\boldsymbol{\beta})$, which implies that $\nabla\breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1}) = 0$, combining the strong convexity we prove in V, it holds that

$$\breve{R}_{i+1}(\tilde{\boldsymbol{\beta}}_{i+1}) - \breve{R}_{i+1}(\breve{\boldsymbol{\beta}}_{i+1}) \geq \tau\|\tilde{\boldsymbol{\beta}}_{i+1} - \breve{\boldsymbol{\beta}}_{i+1}\|_2^2.$$

where τ is the positive curvature which can be found in V:

$$\tau = \frac{\left\{(1-14\eta-4\eta^2)(2\lambda^\star_{i+1}+\lambda^\star_{i+2})-(1+2\eta)(\lambda^\star_{i+1}+2\lambda^\star_{i+2})\right\}}{6}.$$

Summarizing we have that

$$\frac{\delta\rho_{i+1}}{2}\|\tilde{\boldsymbol{\beta}}_{i+1}-\boldsymbol{\beta}^\star_{i+1}\|_1+\tau\|\tilde{\boldsymbol{\beta}}_{i+1}-\breve{\boldsymbol{\beta}}_{i+1}\|_2^2 \lesssim O(\frac{s\log p}{n}).$$

This immediately gives two compelling results:

$$\|\tilde{\boldsymbol{\beta}}_{i+1}-\boldsymbol{\beta}^\star_{i+1}\|_1 \lesssim O\left(s\sqrt{\frac{\log p}{n}}\right),$$
$$\|\tilde{\boldsymbol{\beta}}_{i+1}-\breve{\boldsymbol{\beta}}_{i+1}\|_2 \lesssim O\left(\sqrt{\frac{s\log p}{n}}\right).$$

In IV. we have proved by perturbation that

$$\|\boldsymbol{\beta}^\star_{i+1}-\breve{\boldsymbol{\beta}}_{i+1}\|_2 \le \frac{6(2+2\sqrt{i}\lambda_1)g_i}{\lambda^\star_{i+1}-\lambda^\star_{i+2}}\sqrt{\frac{s\log p}{n}} \lesssim O\left(\sqrt{\frac{s\log p}{n}}\right).$$

Therefore, we have

$$\|\tilde{\boldsymbol{\beta}}_{i+1}-\boldsymbol{\beta}^\star_{i+1}\|_1 \lesssim O\left(s\sqrt{\frac{\log p}{n}}\right),$$
$$\|\tilde{\boldsymbol{\beta}}_{i+1}-\boldsymbol{\beta}^\star_{i+1}\|_2 \lesssim O\left(\sqrt{\frac{s\log p}{n}}\right).$$

Now the distance between our aim and what we have achieved so far has been reduced to a mere step of normalization. The procedure is again pretty similar to single PC case, mimicing which we can obtain

$$\|\widehat{\boldsymbol{\beta}}_{i+1}-\boldsymbol{\beta}^\star_{i+1}\|_1 \lesssim O\left(s\sqrt{\frac{\log p}{n}}\right),$$
$$\|\widehat{\boldsymbol{\beta}}_{i+1}-\boldsymbol{\beta}^\star_{i+1}\|_2 \lesssim O\left(\sqrt{\frac{s\log p}{n}}\right).$$

By induction our target is finally proved, with probability at least $1-3k\cdot(2p)^{-1}$.

7 Conclusion

In this article, we study sparse principal component analysis in high dimensional settings and propose a component-based PCA regime that can induce sparsity in estimation and achieve near-optimal convergence rates. The proposed scheme focuses direct PC extraction and provides more

information for practical interpretation. Besides, it enjoys high statistical accuracy and computational efficiency, which is demonstrated by our numerical experiments. Following this thread, there are also several potential problems and extensions that deserves further investigation. Firstly, the proposed formulation shares a similar inspiration of the SPCA framework by Zou et al. (2006), hence it is of great interest to see whether the mathematical techniques involved are enlightening for providing further theoretical insights for SPCA as well as other component-based algorithms. Secondly, in practice people need to handle various types of data structure; hence it is also of particular interest to generalize the proposed scheme to other eigenstruture related problems, such as factor analysis, canonical component analysis and PCA for generalized linear models.

8 Bibliography

Agarwal, A., Negahban, S., and Wainwright, M. J. (2010). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems 23(NIPS-2010)*, pages 37–45, New York, NY, USA. Curran Associates Inc.

Amini, A. A. and Wainwright, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE international symposium on information theory*, pages 2454–2458. IEEE.

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

Bertsekas, D., Nedic, A., and Ozdaglar, A. (2003). *Convex analysis and optimization*, volume 1. Athena Scientific.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Birnbaum, A., Johnstone, I. M., Nadler, B., and Paul, D. (2013). Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055.

Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022.

Bonnier, F. and Byrne, H. (2012). Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems. *Analyst*, 137(2):322–332.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155–160.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

Chen, S., Ma, S., Xue, L., and Zou, H. (2020). An alternating manifold proximal gradient method for sparse principal component analysis and sparse canonical correlation analysis. *INFORMS Journal on Optimization*, 2(3):192–208.

Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. (2005). A direct formulation for sparse pca using semidefinite programming. In *Advances in Neural Information Processing Systems*, pages 41–48.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196.

Elsener, A. and van de Geer, S. (2018). Sharp oracle inequalities for stationary points of nonconvex penalized m-estimators. *IEEE Transactions on Information Theory*, 65(3):1452–1472.

Elsener, A., van de Geer, S., et al. (2018). Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509.

Hancock, P. J., Burton, A. M., and Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & cognition*, 24(1):26–40.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, New York, USA.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.

Janková, J. and van de Geer, S. (2021). De-biased sparse pca: Inference for eigenstructure of large covariance matrices. *IEEE Transactions on Information Theory*, 67(4):2507–2527.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning(ICML-2009)*, pages 457–464, New York, NY, USA. Association for Computing Machinery.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547.

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2).

Lee, S., Huang, J. Z., and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The annals of applied statistics*, 4(3):1579.

Lu, M., Huang, J. Z., and Qian, X. (2016). Sparse exponential family principal component analysis. *Pattern recognition*, 60:681–691.

Ma, S. (2013a). Alternating direction method of multipliers for sparse principal component analysis. *Journal of the Operations Research Society of China*, 1(2):253–274.

Ma, Z. (2013b). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801.

Mackey, L. W. (2009). Deflation methods for sparse pca. In *Advances in Neural Information Processing Systems*, pages 1017–1024.

Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., and Band, S. S. (2021). Principal component analysis to study the relations between the spread rates of covid-19 in high risks countries. *Alexandria Engineering Journal*, 60(1):457–464.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791–2817.

Nadler, B. (2009). Discussion of "on consistency and sparsity for principal components analysis in high dimensions". *Journal of the American Statistical Association*, 104(486):694–697.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Nobre, J. and Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets. *Expert Systems with Applications*, 125:181–194.

Oliveira, R. I. (2013). The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *arXiv preprint arXiv:1312.2903*.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature biotechnology*, 32(10):1053–1058.

Rennie, J. D. and Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning(ICML-2005)*, pages 713–719, New York, NY, USA. Association for Computing Machinery.

Schlitzer, A., Sivakamasundari, V., Chen, J., Sumatoh, H. R. B., Schreuder, J., Lum, J., Malleret, B., Zhang, S., Larbi, A., Zolezzi, F., et al. (2015). Identification of cdc1-and cdc2-committed dc progenitors reveals early lineage priming at the common dc progenitor stage in the bone marrow. *Nature immunology*, 16(7):718–728.

Shen, D., Shen, H., and Marron, J. S. (2013). Consistency of sparse pca in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115:317–333.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034.

Stewart, G. W. and Sun, J.-g. (1990). *Matrix perturbation theory*. Academic press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Ting, D. T., Wittner, B. S., Ligorio, M., Jordan, N. V., Shah, A. M., Miyamoto, D. T., Aceto, N., Bersani, F., Brannigan, B. W., Xega, K., et al. (2014). Single-cell rna sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell reports*, 8(6):1905–1918.

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500):371–375.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*, pages 2670–2678.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Yano, K., Morinaka, Y., Wang, F., Huang, P., Takehara, S., Hirai, T., Ito, A., Koketsu, E., Kawamura, M., Kotake, K., et al. (2019). Gwas with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proceedings of the National Academy of Sciences*, 116(42):21262–21267.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

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

A Proof of Propositions and Lemmas

A.1 Proof of Lemma 4

Proof. Note that

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} R(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} \tag{26}$$

$$= \arg\min_{c\geq 0} \min_{\|\boldsymbol{\beta}\|_2=c} (c^2-2)\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} \tag{27}$$

$$= \arg\min_{0\leq c\leq 2} \left[(c^2-2)\cdot (\max_{\|\boldsymbol{\beta}\|_2=c} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\right] \tag{28}$$

$$= c_0 \boldsymbol{\beta}_1^\star, \tag{29}$$

where

$$c_0 = \arg\min_{0\leq c\leq 2} \left[\lambda_1 c^2 (c^2-2))\right] = 1. \tag{30}$$

Note that (29) holds because for any fixed $c \geq 0$,

$$\arg\max_{\|\boldsymbol{\beta}\|_2=c} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} = c\boldsymbol{\beta}_1^\star.$$

And when $\lambda_1^\star > \lambda_2^\star$, this maximum point is unique(up to a sign). Now combining (29) and (30), we finish our proof. □

B Proof of Proposition 1

Proof. We start the proof by solving the equation:

$$\nabla R_1(\boldsymbol{\beta}) = (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\Sigma}_1^\star \boldsymbol{\beta} + (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\boldsymbol{\beta} = 0.$$

Following a similar discussion to the proof of Lemma 4 we can know that the stationary points are exactly the eigenvectors of $\boldsymbol{\Sigma}_1^\star$. We check the second order derivative to give further information. Note that

$$\nabla^2 R_1(\boldsymbol{\beta}) = (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\Sigma}_1^\star + 2\boldsymbol{\Sigma}_1^\star \boldsymbol{\beta}\boldsymbol{\beta}^\top + (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\boldsymbol{I}_p + 2\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star.$$

We set $\boldsymbol{\beta} = \boldsymbol{\beta}_0 = \boldsymbol{q}_1$ in $\nabla^2 R(\boldsymbol{\beta})$ to get

$$\nabla^2 R_1(\boldsymbol{\beta}_1) = -\boldsymbol{\Sigma}_1^\star + \lambda_1^\star \boldsymbol{I}_p + 4\lambda_1^\star \boldsymbol{\beta}_1^\star \boldsymbol{\beta}_1^{\star T}.$$

Consider when $\boldsymbol{x}^\top \nabla^2 R_1(\boldsymbol{\beta}_1)\boldsymbol{x}$ is equal to 0. This happens iff.

$$\boldsymbol{x}^\top \boldsymbol{\beta}_1^\star = 0, \text{ and } \boldsymbol{x}^\top (\lambda_1^\star \boldsymbol{I}_p - \boldsymbol{\Sigma}_1^\star)x = 0.$$

From the equations, when $\lambda_1^\star > \lambda_i^\star$, for $i = 2, \cdots, p$, x must be 0,which shows that $\boldsymbol{\beta}_1^\star$ is a minimum point.

On the other hand, for $i \geq 2$,

$$\nabla^2 R_1(\boldsymbol{\beta}_i^\star) = -\boldsymbol{\Sigma}_1^\star + \lambda_i^\star \boldsymbol{I}_p + 4\lambda_i^\star \boldsymbol{\beta}_i^\star \boldsymbol{\beta}_i^{\star T}.$$

We have then

$$\boldsymbol{\beta}_i^{\star T} \nabla^2 R_1(\boldsymbol{\beta}_i^\star) \boldsymbol{\beta}_i^\star = 4\lambda_i^\star > 0, \text{ and } \boldsymbol{\beta}_1^{\star T} \nabla^2 R_1(\boldsymbol{\beta}_i^\star) \boldsymbol{\beta}_1^\star = -\lambda_1^\star + \lambda_i^\star < 0.$$

Thus we conclude $\boldsymbol{\beta}_i^\star$ is a saddle point. □

B.1 Proof of Lemma 1

Proof. We pick an $\eta > 0$ satisfying the following conditions:

- $0 < \eta < \frac{1}{2}$;
- $(1 - 2\eta)\lambda_1^\star > (1 + 2\eta)\lambda_2^\star$. Using the Separation Assumption we can pick $\eta < \frac{1-\epsilon}{2(1+\epsilon)}$;
- $1 - 14\eta - 4\eta^2 > \epsilon(1 + 2\eta)$, where $0 < \epsilon < 1$ comes from the Separation Assumption. Such $\eta > 0$ exists since the value of the function $1 - 14\eta - 4\eta^2 - \epsilon(1 + 2\eta)$ at $\eta = 0$ is $1 - \epsilon > 0$.

Suppose we have picked an $\eta > 0$. Fix a $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \leq \eta\}$. Using that $\boldsymbol{\Sigma}_1^\star = \Gamma\Lambda\Gamma^\top$ we have that

$$\begin{aligned}
\boldsymbol{x}^\top \nabla^2 R_1(\boldsymbol{\beta}) x &= \boldsymbol{x}^\top \left\{ (\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\Sigma}_1^\star + 2\boldsymbol{\Sigma}_1^\star \boldsymbol{\beta}\boldsymbol{\beta}^\top + (\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \boldsymbol{\beta})\boldsymbol{I}_p + 2\boldsymbol{\beta}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_1^\star \right\} x \\
&= \boldsymbol{x}^\top \Gamma \left\{ (\|\boldsymbol{\beta}\|_2^2 - 2)\Lambda + 2\Lambda\Gamma^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \Gamma + (\boldsymbol{\beta}^\top \Gamma\Lambda\Gamma^\top \boldsymbol{\beta})\boldsymbol{I}_p + 2\Gamma^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \Gamma\Lambda \right\} \Gamma^\top x
\end{aligned}$$

Note that

$$\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\|_2 \leq \eta\}, \text{ which means } \Gamma^\top \boldsymbol{\beta} \in \{\tilde{\boldsymbol{\beta}} : \left\|\tilde{\boldsymbol{\beta}} - \epsilon_1\right\|_2 \leq \eta\},$$

where $\epsilon_1 = (1, 0, \ldots, 0)^\top$. So now, for $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \epsilon_1\|_2 \leq \eta\}$, we need to provide a lower bound for the smallest eigenvalue of $\left\{ (\|\boldsymbol{\beta}\|_2^2 - 2)\Lambda + 2\Lambda\Gamma^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \Gamma + (\boldsymbol{\beta}^\top \Gamma\Lambda\Gamma^\top \boldsymbol{\beta})\boldsymbol{I}_p + 2\Gamma^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \Gamma\Lambda \right\}$.

Note that

$$\begin{aligned}
(\|\boldsymbol{\beta}\|^2 - 2)\Lambda &= (\|\boldsymbol{\beta}\|^2 - 2) \sum_{k=1}^p \lambda_k^\star \epsilon_k \epsilon_k^\star, \\
(\boldsymbol{\beta}^\top \Lambda \boldsymbol{\beta})\boldsymbol{I}_p &= (\boldsymbol{\beta}^\top \Lambda \boldsymbol{\beta}) \sum_{k=1}^p \epsilon_k \epsilon_k^\top,
\end{aligned}$$

which gives

$$(\|\boldsymbol{\beta}\|^2 - 2)\Lambda + (\boldsymbol{\beta}^\top \Lambda \boldsymbol{\beta})\boldsymbol{I}_p = \left[(\|\boldsymbol{\beta}\|_2^2 - 2)\lambda_1^\star + \boldsymbol{\beta}^\top \Lambda \boldsymbol{\beta} \right] \epsilon_1 \epsilon_1^\top + \sum_{i=2}^n \left[\|\boldsymbol{\beta}\|_2^2 - 2)\lambda_i^\star + \boldsymbol{\beta}^\top \Lambda \boldsymbol{\beta} \right] \epsilon_i \epsilon_i^\top.$$

Besides, we have

$$\begin{aligned}2\Lambda\boldsymbol{\beta}\boldsymbol{\beta}^\top + 2\boldsymbol{\beta}\boldsymbol{\beta}^\top\Lambda &= 2\Lambda(\boldsymbol{\beta}-\epsilon_1+\epsilon_1)(\boldsymbol{\beta}-\epsilon_1+\epsilon_1)^\top + 2(\boldsymbol{\beta}-\epsilon_1+\epsilon_1)(\boldsymbol{\beta}-\epsilon_1+\epsilon_1)^\top\Lambda \\ &= 4\lambda_1^\star\epsilon_1\epsilon_1^\top + 2\big(\Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top + (\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top\Lambda\big) \\ &\quad + 2\big(\Lambda(\boldsymbol{\beta}-\epsilon_1)\epsilon_1^\top + \epsilon_1(\boldsymbol{\beta}-\epsilon_1)^\top\Lambda\big) + 2\big(\lambda_1^\star(\boldsymbol{\beta}-\epsilon_1)\epsilon_1^\top + \lambda_1^\star\epsilon_1(\boldsymbol{\beta}-\epsilon_1)^\top\big).\end{aligned}$$

Now we do the estimation separately for the terms in the above formula.

- First we have

$$\begin{aligned}\boldsymbol{\beta}^\top\Lambda\boldsymbol{\beta} &= (\boldsymbol{\beta}+\epsilon_1-\epsilon_1)^\top\Lambda(\boldsymbol{\beta}+\epsilon_1-\epsilon_1) \\ &= \epsilon_1^\top\Lambda\epsilon_1 + (\boldsymbol{\beta}-\epsilon_1)^\top\Lambda(\boldsymbol{\beta}-\epsilon_1) + 2(\boldsymbol{\beta}-\epsilon_1)^\top\Lambda\epsilon_1 \\ &\geq \lambda_1^\star\epsilon_1^\top\epsilon_1 - 2\lambda_1^\star\left\|\epsilon_1\right\|_2\left\|\boldsymbol{\beta}-\epsilon_1\right\|_2 \\ &= (1-2\eta)\lambda_1^\star.\end{aligned}$$

  Note that from here we must pick η to be smaller than $\frac{1}{2}$.

  Then it holds that

$$(\|\boldsymbol{\beta}\|_2^2 - 2)\lambda_1^\star + \boldsymbol{\beta}^\top\Lambda\boldsymbol{\beta} \geq [(1-\eta)^2 - 2]\lambda_1^\star + (1-2\eta)\lambda_1^\star > -4\eta\lambda_1^\star.$$

  So for the first term we have

$$\left\{(\|\boldsymbol{\beta}\|_2^2 - 2)\lambda_1^\star + \boldsymbol{\beta}^\top\Lambda\boldsymbol{\beta}\right\}\boldsymbol{x}^\top\epsilon_1\epsilon_1^\top\boldsymbol{x} \geq -4\eta\lambda_1^\star\left\|\boldsymbol{x}\right\|_2^2.$$

- Next we have

$$\begin{aligned}&(\|\boldsymbol{\beta}\|_2^2 - 2)\lambda_i^\star + \boldsymbol{\beta}^\top\Lambda\boldsymbol{\beta} \\ \geq&[(1-\eta)^2 - 2]\lambda_i^\star + (1-2\eta)\lambda_1^\star \\ \geq&(-1-2\eta)\lambda_i^\star + (1-2\eta)\lambda_1^\star \\ \geq&(-1-2\eta)\lambda_2^\star + (1-2\eta)\lambda_1^\star.\end{aligned}$$

  Here we see that if and only if $\lambda_1^\star < \lambda_2^\star$ can we pick η to ensure $(-1-2\eta)\lambda_2^\star + (1-2\eta)\lambda_1^\star > 0$.

  Hence we have

$$\begin{aligned}&\sum_{i=2}^n\left\{\|\boldsymbol{\beta}\|_2^2 - 2)\lambda_1^\star + \boldsymbol{\beta}^\top\Lambda\boldsymbol{\beta}\right\}\boldsymbol{x}^\top\epsilon_i\epsilon_i^\top x \\ \geq&[(-1-2\eta)\lambda_2^\star + (1-2\eta)\lambda_1^\star]\sum_{i=2}^n\boldsymbol{x}^\top\epsilon_i\epsilon_i^\top x.\end{aligned}$$

- Then we have, trivially, $4\lambda_1 \boldsymbol{x}^\top \epsilon_1 \epsilon_1^\top x \geq [(-1-2\eta)\lambda_2^\star + (1-2\eta)\lambda_1^\star]\boldsymbol{x}^\top \epsilon_1 \epsilon_1^\top x$, then combine this estimation and the last one we get

$$\begin{aligned}
&\sum_{i=2}^{n} \left\{ \|\boldsymbol{\beta}\|_2^2 - 2)\lambda_1^\star + \boldsymbol{\beta}^\top \Lambda \boldsymbol{\beta} \right\} \boldsymbol{x}^\top \epsilon_i \epsilon_i^\top x + 4\lambda_1^\star \boldsymbol{x}^\top \epsilon_1 \epsilon_1^\top x \\
\geq & [(-1-2\eta)\lambda_2^\star + (1-2\eta)\lambda_1^\star] \sum_{i=1}^{n} \boldsymbol{x}^\top \epsilon_i \epsilon_i^\top x \\
= & [(-1-2\eta)\lambda_2^\star + (1-2\eta)\lambda_1^\star] \left\|\boldsymbol{x}\right\|_2^2 .
\end{aligned}$$

- Next we estimate the term

$$\begin{aligned}
&2\boldsymbol{x}^\top \left\{ \Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top + (\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top \Lambda \right\} x \\
= & 4\boldsymbol{x}^\top \Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top x.
\end{aligned}$$

  First using $\mathrm{rank}(\Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top)=1$, we see that its first singular value is

$$\begin{aligned}
\phi_1 &= \sqrt{\mathrm{tr}((\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top \Lambda^2 (\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top)} \\
&= \|\boldsymbol{\beta}-\epsilon_1\|_2 \sqrt{(\boldsymbol{\beta}-\epsilon_1)^\top \Lambda^2 (\boldsymbol{\beta}-\epsilon_1)} \\
&\leq \lambda_1^\star \eta^2.
\end{aligned}$$

  Now we see that

$$\begin{aligned}
&|\boldsymbol{x}^\top \Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top x| \\
\leq & \|\boldsymbol{x}\|_2 \left\| \Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top x \right\|_2 \\
\leq & \|\boldsymbol{x}\|_2 \left\| \Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top \right\|_{\mathrm{op}} \|\boldsymbol{x}\|_2 \\
\leq & \lambda_1^\star \eta^2 \left\|\boldsymbol{x}\right\|_2^2 .
\end{aligned}$$

  Here $\|\cdot\|_{\mathrm{op}}$ stands for the spectral norm. Therefore we conclude that

$$2\boldsymbol{x}^\top \big(\Lambda(\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top + (\boldsymbol{\beta}-\epsilon_1)(\boldsymbol{\beta}-\epsilon_1)^\top \Lambda\big) x \geq -4\lambda_1^\star \eta^2 \left\|\boldsymbol{x}\right\|_2^2 .$$

- Now the rest terms can be estimated in the same method as the last one. We provide the result directly:

$$\begin{aligned}
&2\boldsymbol{x}^\top \big(\Lambda(\boldsymbol{\beta}-\epsilon_1)\epsilon_1^\top + \epsilon_1(\boldsymbol{\beta}-\epsilon_1)^\top \Lambda\big) x \geq -4\eta\lambda_1^\star \left\|\boldsymbol{x}\right\|_2^2 , \\
&2\boldsymbol{x}^\top \big(\lambda_1^\star(\boldsymbol{\beta}-\epsilon_1)\epsilon_1^\top + \lambda_1^\star \epsilon_1(\boldsymbol{\beta}-\epsilon_1)^\top\big) x \geq -4\eta\lambda_1^\star \left\|\boldsymbol{x}\right\|_2^2 .
\end{aligned}$$

Finally, by integrating our term-wise estimation it is easy to see that, with an η satisfying the three conditions formulated in the beginning of proof, for any $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_1^\star\|_2 \leq \eta\}$, it holds that

$$\boldsymbol{x}^\top \nabla^2 R(\boldsymbol{\beta}) x \geq \left[(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right] \left\|\boldsymbol{x}\right\|_2^2 .$$

Now the first part of the lemma is proved. The second part follows naturally after performing Taylor's expansion over the theoretical risk:

$$\begin{aligned}
&R_1(\boldsymbol{\beta}_1) - R_1(\boldsymbol{\beta}_2) - \nabla R_1(\boldsymbol{\beta}_2)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \\
=&\frac{1}{2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^\top \nabla^2 R_1(\boldsymbol{\beta}_m)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \\
\geq&\frac{\left\{(1-14\eta-4\eta^2)\lambda_1^\star - (1+2\eta)\lambda_2^\star\right\}}{2}\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right\|_2^2.
\end{aligned}$$

□

B.2 Proof of Lemma 2

In this part we need to prove one crucial result to bridge the empirical loss and the population risk. One of the fundamental instruments is the concentration bounds for subgaussian variables and vectors, as well as quadratic forms built from them. The next two results are pretty standard on these topics.

Lemma 7 (Bernstein-type inequality)**.** . *Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be i.i.d copies of (X, Y), where $K_1 = \|X\|_{\psi_2} < \infty$ and $K_2 = \|Y\|_{\psi_2} < \infty$. Then for all $t > 0$*

$$P(|\frac{1}{n}\sum_{i=1}^n X_i Y_i - E(XY)| \geq K_1 K_2 \frac{t}{c_K n} + K_1 K_2 \sqrt{\frac{t}{c_K n}}) \leq 2\exp(-t),$$

where $c_K > 0$ is an absolute constant only related to $\|X\|_{\psi_2}$ and $\|Y\|_{\psi_2}$.

Lemma 8 (High probability bounds for random bilineary form)**.** *Consider a row vector $X \in R^s$ with $\boldsymbol{\Sigma}^\star := EX^\top X$. Let $\sup_{\|\boldsymbol{\beta}\|_2 \leq 1} \|X\boldsymbol{\beta}\|_{\psi_2} =: K < \infty$. Then for all $t > 0$, with probability at least $1 - \exp(-t)$, it holds that*

$$\sup_{\|u\|_2=1, \|v\|_2=1} |u^\top(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star)v| \leq 2K^2 \frac{t + \log 2 + 6s}{c_K n} + 2K^2 \sqrt{\frac{t + \log 2 + 6s}{c_K n}}.$$

The following lemma, called "Transfer Principle", comes from Lemma 5.1 in Oliveira (2013), whose proof is also provided therein.

Lemma 9 (Transfer Principle)**.** *Suppose $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}^\star$ are matrices with non-negative diagonal entries, and assume $\eta \in (0, 1)$, $d \in \{1, \ldots, p\}$ are such that*

$$\forall v \in \mathbb{R}^p \text{ with } |v|_0 \leq d, v^\top(\widehat{\boldsymbol{\Sigma}} - (1-\eta)\boldsymbol{\Sigma}^\star)v \geq 0.$$

Assume D is a diagonal matrix whose elements $D[j, j]$ are non-negative and satisfy $D[j, j] \geq \widehat{\boldsymbol{\Sigma}}[j, j] - (1-\eta)\boldsymbol{\Sigma}^\star[j, j]$. Then

$$\forall x \in \mathbb{R}^p, (d-1)\boldsymbol{x}^\top[\widehat{\boldsymbol{\Sigma}} - (1-\eta)\boldsymbol{\Sigma}^\star]\boldsymbol{x} \geq -|D^{1/2}\boldsymbol{x}|_1^2.$$

The basic idea of the following lemma is attributed to Sara and Elsener's Lemma D.2 in Elsener and van de Geer (2018). We adapt the methodology to our scheme and provide a self-contained proof as follows.

Lemma 10 (High probability bound on random quadratic forms)**.** *Consider a set of centered i.i.d sample* $\{\boldsymbol{x}_i \in R^s, i = 1, \cdots, n\}$*, from a zero-mean sub-gaussian random vector* X *with* $\sup_{\|\boldsymbol{\beta}\|_2 \le 1} \|X\boldsymbol{\beta}\|_{\psi_2} =: K < \infty$*. Let* $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$ *(the sample covariance),* $\boldsymbol{\Sigma}^\star := EX^\top X$*. Let We have for all* $u \in \mathbb{R}^p$*,* $\|u\|_1 \le L\|u\|_2\big(L > 0$*, and satisfies* $\frac{L^2 \log p}{n} = o(1)$ *in an asymptotics view*$\big)$ *and for all* $t > 0$*, when* n *is large enough, with probability at least* $1 - \exp(-t)$*, it holds*

$$
\begin{aligned}
&\forall u : \|u\|_1 \le L\|u\|_2, \\
&\big|u^\top(\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}})u\big| \le 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\frac{t + (6 + \log(2p))}{c_K n}\|u\|_2^2 \\
&+ 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\frac{2(6 + \log(2p))}{c_K n}\|u\|_1^2 \\
&+ 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\sqrt{\frac{t + (6 + \log(2p))}{c_K n}}\|u\|_2^2 \\
&+ 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\sqrt{\frac{2(6 + \log(2p))}{c_K n}}\|u\|_2\|u\|_1.
\end{aligned}
$$

Now we are ready to prove Lemma 2, which is crucial for bridging the gap between the empirical loss and population risk in a probabilistic sense.

Proof of Lemma 2. We divide the proof in the following parts.

PART 1. Three-term Spliting

For any $\boldsymbol{\beta}, \boldsymbol{\beta}^\star \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_1^\star\| \le \eta, \|\boldsymbol{\beta}\|_1 \le L\}$,

$$
\begin{aligned}
&\big|(\nabla R(\boldsymbol{\beta}) - \nabla R(\boldsymbol{\beta}))^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\big| \\
=&\big|(\|\boldsymbol{\beta}\|_2^2 - 2)\boldsymbol{\beta}^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}^\star - \boldsymbol{\beta}) + (\boldsymbol{\beta}^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta})\boldsymbol{\beta}^\top(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big| \\
\le&2\big|\boldsymbol{\beta}^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big| + \big|\boldsymbol{\beta}^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}\big|\big|\boldsymbol{\beta}^\top(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big| \\
\le&2\big|(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\big| + 2\big|\boldsymbol{\beta}^{*^\top}(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big| + \big|\boldsymbol{\beta}^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}\big|\big|\boldsymbol{\beta}^\top(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big| \\
\le&4\big|(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\big| + 3\big|\boldsymbol{\beta}^{*^\top}(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big| + \big|\boldsymbol{\beta}^{\star T}(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}^\star\big|\big|\boldsymbol{\beta}^\top(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\big|,
\end{aligned}
$$

which are mainly due to partitioning and triangle inequality. Now we estimate the three terms in the above formula separately.

PART 2. First Term Bounding

For the first term, we follow the beginning of the proof of Lemma 3.10 in Elsener and van de

Geer (2018). Using Lemma 10 the following event holds with probability at least $1-\exp(-t)$

$$
\begin{aligned}
&\left|(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)^\top(\boldsymbol{\Sigma}_1^\star-\widehat{\boldsymbol{\Sigma}}_1)(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\right| \\
&\leq 2K^2(1+\frac{2\lambda_1^\star}{\sigma})\frac{t+(6+\log(2p))}{c_K n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2 \\
&+2K^2(1+\frac{2\lambda_1^\star}{\sigma})\frac{2(6+\log(2p))}{c_K n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1^2 \\
&+2K^2(1+\frac{2\lambda_1^\star}{\sigma})\sqrt{\frac{t+(6+\log(2p))}{c_K n}}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2 \\
&+2K^2(1+\frac{2\lambda_1^\star}{\sigma})\sqrt{\frac{2(6+\log(2p))}{c_K n}}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1.
\end{aligned}
$$

Now using Young's Inequality with a constant $\zeta>0$ we have

$$
\begin{aligned}
\sqrt{\frac{t+(6+\log(2p))}{c_K n}}&=\sqrt{\frac{[t+(6+\log(2p))]\zeta}{c_K n\zeta}}\leq\frac{t+(6+\log(2p))}{2c_K n\zeta}+\frac{\zeta}{2}, \\
\sqrt{\frac{2(6+\log(2p))}{c_K n}}&=\sqrt{\frac{2(6+\log(2p))\zeta}{c_K n\zeta}}\leq\frac{6+\log(2p)}{c_K n\zeta}+\frac{\zeta}{2},
\end{aligned}
$$

which gives

$$
\begin{aligned}
&\left|(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)^\top(\boldsymbol{\Sigma}_1^\star-\widehat{\boldsymbol{\Sigma}}_1)(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\right| \\
&\leq 2K^2(1+\frac{2\lambda_1^\star}{\sigma})\frac{t+(6+\log(2p))}{c_K n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2 \\
&+2K^2(1+\frac{2\lambda_1^\star}{\sigma})\frac{2(6+\log(2p))}{c_K n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1^2 \\
&+2K^2(1+\frac{2\lambda_1^\star}{\sigma})(\frac{t+(6+\log(2p))}{2\zeta c_K n}+\frac{\zeta}{2})\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2 \\
&+2K^2(1+\frac{2\lambda_1^\star}{\sigma})(\frac{6+\log(2p)}{\zeta c_K n}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1^2+\frac{\zeta}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2).
\end{aligned}
$$

Now choose $t=\log(2p)$. Note that the coefficient before $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2$ is

$$
M_0(\log 2p):=2K^2(1+\frac{2\lambda_1^\star}{\sigma})\left\{\frac{6+2\log(2p)}{c_K n}(1+(2\zeta)^{-1})+\zeta\right\}, \tag{31}
$$

and that before $\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1^2$ is

$$
M_1(\log 2p):=2K^2(1+\frac{2\lambda_1^\star}{\sigma})\left\{\frac{2(6+\log(2p))}{c_K n}+\frac{6+\log(2p)}{\zeta c_K n}\right\}.
$$

We obtain

$$
\begin{aligned}
&4\left|(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)^\top(\widehat{\boldsymbol{\Sigma}}_1-\boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}-\boldsymbol{\beta}^\star)\right| \\
&\leq 4M_0(\log 2p)\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2+4M_1(\log 2p)\,\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1^2 \\
&\leq 4M_0(\log 2p)\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_2^2+8M_1(\log 2p)L\,\|\boldsymbol{\beta}-\boldsymbol{\beta}^\star\|_1\,.
\end{aligned}
$$

PART 3. Second Term Bounding

Now we begin to bound the second term. This part can also be tracked to the work in Elsener and van de Geer (2018). We will apply the dual norm inequality. We notice that

$$\left\|(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}^\star\right\|_\infty = \max_{1\le j\le p} |e_j^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}^\star|.$$

For all $j = 1, \ldots, p$ and all $t > 0$, using Lemma 7, the event

$$\left\{|e_j^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}^\star| \le \left(K^2\frac{t+\log 2p}{c_K n} + K^2\sqrt{\frac{t+\log 2p}{c_K n}}\right)(\|\boldsymbol{\beta}_1^\star\|_2 + \eta)\right\}$$

has probability at least $1 - \exp(-t)$, so does the following event:

$$\begin{aligned}
&\left|{\boldsymbol{\beta}^*}^\top(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\right| \\
\le&\left(K^2\frac{t+\log 2p}{c_K n} + K^2\sqrt{\frac{t+\log 2p}{c_K n}}\right)(\|\boldsymbol{\beta}_1^\star\|_2 + \eta)\,\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 \\
\le&2\left(K^2\frac{t+\log 2p}{c_K n} + K^2\sqrt{\frac{t+\log 2p}{c_K n}}\right)\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 .
\end{aligned}$$

Let

$$M_2(t) := 2\left(K^2\frac{t+\log 2p}{c_K n} + K^2\sqrt{\frac{t+\log 2p}{c_K n}}\right),$$

then

$$\left|\boldsymbol{\beta}^{\star T}(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\right| \le M_2(t)\,\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 .$$

PART 4. Third Term Bounding

Then comes the third term. For the quadratic coefficient we simply use Lemma 7 to get a bound with probability at least $1 - \exp(-t)$:

$$\left|\boldsymbol{\beta}^{\star T}(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}^\star\right| \le 4K^2\frac{t+\log 2}{c_K n} + 4K^2\sqrt{\frac{t+\log 2}{c_K n}} := M_3(t).$$

Then note that

$$\left|\boldsymbol{\beta}^\top(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\right| \le \|\boldsymbol{\beta}\|_2\,\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}\|_2 \le (1+\eta)\cdot\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 .$$

Combining with the quadratic coefficient bound we have

$$\left|\boldsymbol{\beta}^{\star T}(\widehat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1^\star)\boldsymbol{\beta}^\star\right|\left|\boldsymbol{\beta}^\top(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\right| \le 2M_3(t)\,\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 .$$

PART 5. Combined Bounding

Now we integrate the proof above to get the following inequality holds with probability at least $1 - 3\exp(-\log(2p))$:

$$\begin{aligned}
&\left|(\nabla R(\boldsymbol{\beta}) - \nabla R(\boldsymbol{\beta}))^\top(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\right| \\
\le&4M_0(\log 2p)\cdot\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_2^2 + \{8M_1(\log 2p))L + 3M_2(\log 2p) + 2M_3(\log 2p)\}\,\|\boldsymbol{\beta} - \boldsymbol{\beta}^\star\|_1 .
\end{aligned}$$

PART 6. Summary

Let $M_\epsilon = 8M_1(\log 2p))L + 3M_2(\log 2p) + 2M_3(\log 2p) > 0$, where M_i can be tracked in the proof above. Besides, we hope to pick ζ and the sample size n so that $\gamma = 4M_0(\log(2p)) < 1$ holds with high probability. Let

$$\zeta = \left(32K^2(1 + \frac{2\lambda_1^\star}{\sigma})\right)^{-1},$$
$$n > \zeta^{-1}(1 + \zeta^{-1})\frac{6 + 2\log 2p}{c_K}.$$

Plug these choices into (31), we get $\gamma = 4M_0(\log 2p) < 1$. Summarizing the above parts, we have proved the main result holds with probability at least $1 - 3 \cdot (2p)^{-1}$.

C Proof of Concentration Results

C.1 Proof of Lemma 7

Proof. Using the sub-gaussianity of (X, Y) we know that XY is sub-exponential, with $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$, thus a Bernstein-type tail bound exists according to Vershynin (2010):

$$P(|\frac{1}{n}\sum_{i=1}^n X_iY_i - E(XY)| \geq t) \leq 2\exp\big(-c_K n \min(\frac{t^2}{(K_1K_2)^2}, \frac{t}{K_1K_2})\big).$$

Let $s = c_K n \min(\frac{t^2}{(K_1K_2)^2}, \frac{t}{K_1K_2})$, then $t = \max\big(K_1K_2\frac{s}{c_K n}, K_1K_2\sqrt{\frac{s}{c_K n}}\big)$, which shows that

$$\begin{aligned}
&P\bigg(|\frac{1}{n}\sum_{i=1}^n X_iY_i - E(XY)| \geq K_1K_2\frac{s}{c_K n} + K_1K_2\sqrt{\frac{s}{c_K n}}\bigg) \\
\leq &P\bigg(|\frac{1}{n}\sum_{i=1}^n X_iY_i - E(XY)| \geq \max\big\{K_1K_2\frac{s}{c_K n}, K_1K_2\sqrt{\frac{s}{c_K n}}\big\}\bigg) \leq 2\exp\big(-s\big).
\end{aligned}$$

□

C.2 Proof of Lemma 8

Proof. Define

$$\widehat{A} := \sup_{\|u\|_2=1, \|v\|_2=1} \big|u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)v\big|.$$

Let $u, v, \tilde{u}, \tilde{v}$ be arbitrary. Then

$$\begin{aligned}
u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)v = &\tilde{u}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)\tilde{v} \\
&+ (u - \tilde{u})^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)(v - \tilde{v}) + \tilde{u}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)(v - \tilde{v}) + (u - \tilde{u})(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)\tilde{v}.
\end{aligned}$$

Thus for all $\epsilon > 0$ and for $\|u - \tilde{u}\|_2 \leq \epsilon$ and $\|v - \tilde{v}\|_2 \leq \epsilon$

$$|u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)v| \leq |\tilde{u}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)\tilde{v}| + 2\epsilon\widehat{A} + \epsilon^2\widehat{A}.$$

We now take S_ϵ to be a minimal ϵ-covering net of the unit sphere $\omega \in \mathbb{R}^s : \|\omega\|_2 = 1$. Then $\mathcal{S}_\epsilon \leq (1 + \frac{\epsilon}{2})^s$. It follows that

$$(1 - 2\epsilon - \epsilon^2)\widehat{A} \leq \max_{\tilde{u}\in\mathcal{S}_\epsilon, \tilde{v}\in\mathcal{S}_\epsilon} |\tilde{u}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)\tilde{v}|.$$

For each $\tilde{u}$ and $\tilde{v}$ in the unit sphere, we know that $\|X\tilde{u}\|_{\psi_2}\|X\tilde{v}\|_{\psi_2} \leq K^2$. Hence for each such $\tilde{u}, \tilde{v}$, using Lemma 7, for all $t > 0$, with probability at least $1 - 2\exp(-t)$,

$$|\tilde{u}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)\tilde{v}| \leq K^2\frac{t}{c_K n} + K^2\sqrt{\frac{t}{c_K n}}.$$

It follows that for all $t > 0$, with probability at least $1 - \exp(-t)$,

$$\max_{\tilde{u}\in\mathcal{S}_\epsilon, \tilde{v}\in\mathcal{S}_\epsilon} |\tilde{u}^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)\tilde{v}| \leq K^2\frac{t + \log(2|S_\epsilon|^2)}{c_K n} + K^2\sqrt{\frac{t + \log(2|S_\epsilon|^2)}{c_K n}}.$$

We now choose

$$\epsilon := \frac{\sqrt{6} - 2}{2}.$$

Then

$$1 - 2\epsilon - \epsilon^2 = \frac{1}{2}.$$

Moreover,

$$1 + \frac{2}{\epsilon} = 1 + \frac{4}{\sqrt{6} - 2} = 1 + \frac{4(\sqrt{6} + 2)}{2} = 2\sqrt{6} + 5.$$

Thus,

$$2|S_\epsilon|^2 \leq 2(2\sqrt{6} + 5)^{2s},$$

and

$$\log(2|S_\epsilon|^2) \leq \log 2 + 2s\log(2\sqrt{6} + 5) \leq \log 2 + 6s.$$

□

C.3 Proof of Lemma 10

Proof. Let $q \geq 2$. Define

$$t(q, p) := 2K^2\frac{t + \log 2 + 6q + q\log p}{c_K n} + 2K^2\sqrt{\frac{t + \log 2 + 6q + q\log p}{c_K n}}.$$

Consider the event

$$\mathcal{E}_q = \left\{ \sup_{\substack{\|u\|_2\leq 1, \|v\|_2\leq 1 \\ \|u\|_0\leq q, \|v\|_0\leq q}} |u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)v| \leq t(q, p) \right\}$$

Using Lemma 8, and the method of union bound we can show that $\mathbb{P}(\mathcal{E}_q) \geq 1 - \exp(-t)$. Then on $\mathcal{E}_q$, for all $u \in \mathbb{R}^p$, $\|u\|_0 \leq q$,

$$u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)u \geq -t(q,p)\|u\|_2^2 \geq -\frac{t(q,p)}{\lambda_{\min}(\mathbf{\Sigma}^\star)}u^\top\mathbf{\Sigma}^\star u,$$

which means that for all $u \in \mathbb{R}^p$, $\|u\|_0 \leq q$,

$$u^\top\left\{\widehat{\mathbf{\Sigma}} - (1 - \frac{t(q,p)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\mathbf{\Sigma}^\star\right\}u \geq 0.$$

Pick n to be large enough such that $t(q,p) < \lambda_{\min}(\mathbf{\Sigma}^\star)$, then apply the transfer principle (Lemma 9) we have on $\mathcal{E}_s$, for all $u \in \mathbb{R}^p$,

$$u^\top\left\{\widehat{\mathbf{\Sigma}} - (1 - \frac{t(q,p)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\mathbf{\Sigma}^\star\right\}u \geq -(\max_j \widehat{B}_{j,j})\|u\|_1^2/(q-1),$$

where $\widehat{B} = \widehat{\mathbf{\Sigma}} - [1 - t(q,p)/\lambda_{\min}(\mathbf{\Sigma}^\star)]\mathbf{\Sigma}^\star$.

But on $\mathcal{E}_q$,

$$\sup_{\|u\|_2=1,\|u\|_0\leq q} u^\top(\widehat{\mathbf{\Sigma}} - (1 - \frac{t(q,p)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\mathbf{\Sigma}^\star)u \leq \left\{1 + \frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)}\right\}t(q,p),$$

Therefore we find on $\mathcal{E}_q$, for all $u \in \mathbb{R}^p$,

$$u^\top\left\{\widehat{\mathbf{\Sigma}} - (1 - \frac{t(q,p)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\mathbf{\Sigma}^\star\right\}u \geq -[(1 + \frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})t(q,p)]\|u\|_1^2/(q-1).$$

Or equivalently,

$$u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)u \geq -t(q,p)\left\{\frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)}\|u\|_2^2 + (1 + \frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\|u\|_1^2/(q-1)\right\}.$$

The same exercise can be done to find that on $\mathcal{E}_q$, also for all $u \in \mathbb{R}^p$,

$$u^\top(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}})u \geq -t(q,p)\left\{\frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)}\|u\|_2^2 + (1 + \frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\|u\|_1^2/(q-1)\right\}.$$

Thus on $\mathcal{E}_q$, also for all $u \in \mathbb{R}^p$,

$$|u^\top(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}})u| \leq t(q,p)\left\{\frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)}\|u\|_2^2 + (1 + \frac{\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\|u\|_1^2/(q-1)\right\}.$$

This shows that, on $\mathcal{E}_q$ for all u, such that $\|u\|_2 \leq 1$, $\|u\|_1^2 \leq q-1$, we have

$$|u^\top(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star)u| \leq (1 + \frac{2\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})t(q,p).$$

Consider for $k \geq 1$, the event

$$\begin{aligned}\mathcal{F}_k :=& \Big\{ \sup_{\|u\|_2\leq 1,\|u\|_1^2\leq k} \left|u^\top(\mathbf{\Sigma}^\star - \widehat{\mathbf{\Sigma}})u\right| \\ &\geq 2K^2(1 + \frac{2\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\frac{t + \log(2p) + 6 + k(6 + \log p)}{c_K n} \\ &+ 2K^2(1 + \frac{2\lambda_{max}(\mathbf{\Sigma}^\star)}{\lambda_{\min}(\mathbf{\Sigma}^\star)})\sqrt{\frac{t + \log(2p) + 6 + k(6 + \log p)}{c_K n}}\Big\}.\end{aligned}$$

Then we have shown that $\mathcal{F}_k \cap \mathcal{E}_{k+1} = \emptyset$, which means $\mathbb{P}(\mathcal{F}_k) \leq \exp(-t)$.

Note that with $A = [L^2] + 1$, we have the partition:

$$\{\|u\|_1^2 \leq 1\} \cup \{1 \leq \|u\|_1^2 \leq 2\} \cup \cdots \cup \{A - 1 \leq \|u\|_1^2 \leq A\},$$

This partition reveals that we need to guarantee a sufficiently small $t(s,p)$ for $s = 1, \cdots, A+1$, and since $A \asymp L^2$, we have

$$\frac{s \log p}{n} \asymp \frac{A \log p}{n} \asymp \frac{L^2 \log p}{n} = o(1).$$

If for some $i \geq 2$ it holds that $\|u\|_1^2 > i - 1$, then the event

$$\Big\{ \exists u : \|u\|_2 \leq 1, \|u\|_1^2 > i-1,$$
$$\big|u^\top (\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}) u\big| \geq 2K^2\big(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)}\big) \frac{t + \log(2p) + 6 + 2\|u\|_1^2(6 + \log(2p))}{c_K n}$$
$$+ 2K^2\big(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)}\big) \sqrt{\frac{t + \log(2p) + 6 + 2\|u\|_1^2(6 + \log(2p))}{c_K n}} \Big\}.$$

implies

$$\Big\{ \exists u : \|u\|_2 \leq 1, \|u\|_1^2 > i-1,$$
$$\big|u^\top (\boldsymbol{\Sigma}_1^\star - \widehat{\boldsymbol{\Sigma}}) u\big| \geq 2K^2\big(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)}\big) \frac{t + \log(2p) + 6 + k(6 + \log(2p))}{c_K n}$$
$$+ 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)}) \sqrt{\frac{t + \log(2p) + 6 + k(6 + \log(2p))}{c_K n}} \Big\}.$$

Hence the event

$$\Big\{ \exists u : \|u\|_2 \leq 1, \|u\|_1 \leq L,$$
$$\big|u^\top (\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}) u\big| \geq 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)}) \frac{t + \log(2p) + 6 + 2\|u\|_1^2(6 + \log(2p))}{c_K n}$$
$$+ 2K^2(1 + \frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)}) \sqrt{\frac{t + \log(2p) + 6 + 2\|u\|_1^2(6 + \log(2p))}{c_K n}} \Big\}.$$

has probability at most

$$\sum_{i=1}^{A} \exp[-(t + i \log 2)] \leq \exp(-t) \sum_{i=1}^{\infty} 2^{-i} = \exp(-t).$$

In other words we have shown that the event

$$\Big\{ \forall u : \|u\|_1 \leq L\|u\|_2,$$
$$\big|u^\top (\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}}) u\big| \leq 2K^2(1 + \frac{2\lambda_{max}}{\lambda_{\min}}) \frac{t + (2\|u\|_1^2/(\|u\|_2^2) + 1)(6 + \log(2p))}{c_K n} \|u\|_2^2$$
$$+ 2K^2(1 + \frac{2\lambda_{max}}{\lambda_{\min}}) \sqrt{\frac{t + (2\|u\|_1^2/(\|u\|_2^2) + 1)(6 + \log(2p))}{c_K n}} \|u\|_2^2 \Big\}.$$

has probability at least $1-\exp(-t)$. It follows that with probability at least $1-\exp(-t)$

$$
\begin{aligned}
&\forall u : \|u\|_1 \leq L\|u\|_2, \\
&\left|u^\top(\boldsymbol{\Sigma}^\star - \widehat{\boldsymbol{\Sigma}})u\right| \leq 2K^2(1+\frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\frac{t+(6+\log(2p))}{c_K n}\|u\|_2^2 \\
&+2K^2(1+\frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\frac{2(6+\log(2p))}{c_K n}\|u\|_1^2 \\
&+2K^2(1+\frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\sqrt{\frac{t+(6+\log(2p))}{c_K n}}\|u\|_2^2 \\
&+2K^2(1+\frac{2\lambda_{max}(\boldsymbol{\Sigma}^\star)}{\lambda_{\min}(\boldsymbol{\Sigma}^\star)})\sqrt{\frac{2(6+\log(2p))}{c_K n}}\|u\|_2\|u\|_1.
\end{aligned}
$$

□

Table 1: Average of 100 Runs for Subspace Loss and Projection Matrix MSE(Vu and Lei's Model)

Noise	Method	s=10, disjoint		s=25, disjoint		s=10, shared		s=25, shared	
		S-Loss	F-Loss	S-Loss	F-Loss	S-Loss	F-Loss	S-Loss	F-Loss
1.0	ITSPCA	0.0508	0.2157	0.1043	0.5151	0.0539	0.2266	0.1243	0.6184
	DTSPCA	0.0558	0.2425	0.1083	0.5228	0.0487	0.2118	0.0969	0.4541
	AUGSPCA	0.0454	0.2059	0.1396	0.7508	0.0453	0.2177	0.1533	0.7814
	CORSPCA	0.0385	0.1242	0.1158	0.5203	0.0246	0.0899	0.1164	0.4809
	Fantope	0.0263	0.0684	0.0667	0.2930	0.0618	0.2459	0.1169	0.6176
	ADAL	0.0424	0.2262	0.0967	0.4999	0.0842	0.3621	0.1498	0.8628
	PSPCA(DT)	0.0438	0.2315	0.0823	0.4595	0.0419	0.1989	0.0901	0.4569
	PSPCA(ADAL)	0.0381	0.1842	0.0688	0.3586	0.0583	0.2630	0.1452	0.8205
10.0	ITSPCA	0.2097	0.6638	0.2292	1.2338	0.2396	0.8786	0.4360	2.3978
	DTSPCA	0.1861	0.6547	0.3506	1.8851	0.2406	0.9270	0.4700	2.4410
	AUGSPCA	0.1824	0.8501	0.3613	2.0446	0.2590	1.0424	0.6087	2.8975
	CORSPCA	0.1741	0.5768	0.4152	1.9447	0.1740	0.5402	0.5577	2.7051
	Fantope	0.1091	0.5659	0.1869	0.9050	0.3118	1.2209	0.6308	3.1143
	ADAL	0.1096	0.4000	0.3082	1.4138	0.3470	1.2888	0.7476	3.2423
	PSPCA(DT)	0.1441	0.5658	0.2900	1.5142	0.2059	0.8114	0.4679	2.4247
	PSPCA(ADAL)	0.1020	0.3685	0.2215	0.9895	0.3510	1.2014	0.7505	3.1670

Table 2: Average of 100 Runs for Subspace Loss and Size(Single-spike Model)

Model	Method	$\lambda_1 = 100$		$\lambda_1 = 25$		$\lambda_1 = 10$		$\lambda_1 = 5$	
		Loss	Size	Loss	Size	Loss	Size	Loss	Size
SP	ITSPCA	0.0016	38.0	0.0068	27.1	0.0161	20.3	0.0279	17.3
	DTSPCA	0.0070	26.3	0.0237	17.5	0.0377	13.6	0.0547	12.7
	AUGSPCA	0.0025	33.3	0.0095	23.1	0.0233	16.6	0.0372	13.2
	CORSPCA	0.0014	43.6	0.0060	31.8	0.0154	20.9	0.0313	15.1
	Fantope	0.0016	40.3	0.0069	40.2	0.0172	47.8	0.0320	67.3
	ADAL	0.0025	40.2	0.0083	47.9	0.0219	28.0	0.0453	20.7
	PSPCA(DT)	0.0020	42.9	0.0087	32.8	0.0162	70.0	0.0314	27.9
	PSPCA(ADAL)	0.0025	40.3	0.0082	40.6	0.0163	71.3	0.0292	30.1
PP	ITSPCA	0.0060	83.2	0.0171	52.8	0.0348	39.2	0.0594	30.3
	DTSPCA	0.0191	49.5	0.0523	29.0	0.0979	20.4	0.1796	14.0
	AUGSPCA	0.0090	66.2	0.0253	41.4	0.0527	27.4	0.0835	20.3
	CORSPCA	0.0051	92.2	0.0171	53.3	0.0402	34.3	0.0684	24.4
	Fantope	0.0049	113.5	0.0167	109.8	0.0364	98.5	0.0797	78.5
	ADAL	0.0059	111.9	0.0215	75.3	0.0564	50.0	0.1130	34.5
	PSPCA(DT)	0.0051	127.0	0.0179	196.3	0.0325	99.9	0.0715	42.1
	PSPCA(ADAL)	0.0056	108.6	0.0177	199.5	0.0320	102.0	0.0642	46.0

PS: In the "Model" column, "SP" stands for the single peak signal, and "PP" for piecewise polynomial model.

Table 3: Average of 100 Runs for l_1 and l_2 Loss(Single-spike Model)

Model	Method	$\lambda_1 = 100$		$\lambda_1 = 25$		$\lambda_1 = 10$		$\lambda_1 = 5$	
		l_1	l_2	l_1	l_2	l_1	l_2	l_1	l_2
SP	ITSPCA	1.2580	0.0400	2.0855	0.0823	2.8236	0.1264	3.4605	0.1671
	DTSPCA	2.0648	0.0831	3.2459	0.1539	3.7255	0.1947	4.0674	0.2335
	AUGSPCA	1.4991	0.0501	2.3386	0.0976	3.2627	0.1529	3.7165	0.1935
	CORSPCA	1.1415	0.0375	1.9758	0.0771	2.7714	0.1237	3.5667	0.1770
	Fantope	1.2377	0.0399	2.2378	0.0831	3.1364	0.1311	3.9503	0.1786
	ADAL	1.4471	0.0502	2.4509	0.0912	3.1321	0.1483	3.8742	0.2137
	PSPCA(DT)	1.3297	0.0451	2.2675	0.0929	3.2032	0.1271	3.5495	0.1773
	PSPCA(ADAL)	1.4480	0.0502	2.2916	0.0904	3.2320	0.1276	3.5063	0.1711
PP	ITSPCA	2.3769	0.0774	4.0275	0.1311	5.7152	0.1871	7.5565	0.2452
	DTSPCA	4.1635	0.1383	7.0143	0.2299	10.0556	0.3160	14.9033	0.4328
	AUGSPCA	2.9171	0.0948	4.8911	0.1594	7.0781	0.2309	9.3426	0.2918
	CORSPCA	2.1815	0.0717	4.0178	0.1308	6.1273	0.2012	8.3203	0.2636
	Fantope	2.2907	0.0703	4.2049	0.1295	6.1594	0.1915	9.4964	0.2843
	ADAL	2.5302	0.0767	4.6381	0.1470	7.6992	0.2389	11.8042	0.3410
	PSPCA(DT)	2.3105	0.0712	4.3828	0.1341	5.7215	0.1807	8.7849	0.2694
	PSPCA(ADAL)	2.4305	0.0750	4.3402	0.1332	5.6593	0.1794	8.1938	0.2552

PS: In the "Model" column, "SP" stands for the single peak signal, and "PP" for piecewise polynomial model.

Table 4: Average of 100 Runs for False Positive Rate and False Negative Rate(Single-spike Model)

Model	Method	$\lambda_1 = 100$		$\lambda_1 = 25$		$\lambda_1 = 10$		$\lambda_1 = 5$	
		FP	FN	FP	FN	FP	FN	FP	FN
SP	ITSPCA	0.0001	0.0059	0.0000	0.0112	0.0001	0.0146	0.0002	0.0162
	DTSPCA	0.0002	0.0117	0.0002	0.0161	0.0001	0.0179	0.0003	0.0185
	AUGSPCA	0.0000	0.0082	0.0000	0.0131	0.0000	0.0163	0.0000	0.0179
	CORSPCA	0.0011	0.0042	0.0008	0.0097	0.0001	0.0143	0.0000	0.0171
	Fantope	0.0003	0.0050	0.0037	0.0085	0.0100	0.0111	0.0215	0.0130
	ADAL	0.0003	0.0051	0.0069	0.0079	0.0019	0.0127	0.0015	0.0158
	PSPCA(DT)	0.0009	0.0044	0.0011	0.0094	0.0201	0.0103	0.0037	0.0145
	PSPCA(ADAL)	0.0003	0.0050	0.0038	0.0084	0.0207	0.0103	0.0046	0.0143
PP	ITSPCA	0.0002	0.0230	0.0001	0.0378	0.0002	0.0446	0.0005	0.0492
	DTSPCA	0.0002	0.0395	0.0001	0.0494	0.0002	0.0537	0.0002	0.0568
	AUGSPCA	0.0000	0.0311	0.0000	0.0432	0.0000	0.0501	0.0000	0.0535
	CORSPCA	0.0006	0.0191	0.0001	0.0376	0.0001	0.0468	0.0001	0.0516
	Fantope	0.0048	0.0128	0.0161	0.0260	0.0196	0.0349	0.0181	0.0433
	ADAL	0.0044	0.0132	0.0042	0.0309	0.0024	0.0414	0.0018	0.0484
	PSPCA(DT)	0.0094	0.0108	0.0526	0.0202	0.0202	0.0348	0.0033	0.0462
	PSPCA(ADAL)	0.0034	0.0138	0.0540	0.0201	0.0210	0.0347	0.0044	0.0454

PS: In the "Model" column, "SP" stands for the single peak signal, and "PP" for piecewise polynomial model.

Table 5: Average CPU time(seconds) of 100 Runs(Single-spike Model)

Model	Method	$\lambda = 100$	$\lambda = 25$	$\lambda = 10$	$\lambda = 5$
SP	ITSPCA	4.28	4.24	4.20	4.25
	DTSPCA	3.12	3.21	3.31	3.20
	AUGSPCA	4.82	4.50	4.20	4.20
	CORSPCA	3.86	3.62	3.69	3.79
	Fantope	561.94	533.60	591.01	610.20
	ADAL	345.63	678.41	1113.46	1574.37
	PSPCA(DT)	23.47	21.10	15.58	23.10
	PSPCA(ADAL)	15.57	42.64	32.64	79.14
PP	ITSPCA	4.50	4.52	4.64	4.36
	DTSPCA	3.30	3.44	3.52	3.53
	AUGSPCA	5.00	4.56	4.49	4.19
	CORSPCA	3.73	3.96	3.81	3.66
	Fantope	605.42	610.33	639.36	621.54
	ADAL	490.43	1217.48	1065.11	1513.83
	PSPCA(DT)	20.02	18.80	21.23	17.61
	PSPCA(ADAL)	25.91	24.03	48.53	67.22

Table 6: Average of 100 Runs for Subspace Distance, l_1 and l_2 Loss(Mulit-spike Model)

Model	Method	Loss	Comp-1		Comp-2		Comp-3		Comp-4	
			l_1	l_2	l_1	l_2	l_1	l_2	l_1	l_2
1	ITSPCA	0.0086	4.6878	0.1359	5.3446	0.1606	3.9250	0.1118	2.4442	0.0900
	DTSPCA	0.0223	6.0407	0.1915	6.0034	0.1911	4.2497	0.1244	3.3775	0.1442
	AUGSPCA	0.0125	5.1785	0.1557	5.4811	0.1689	3.9350	0.1133	2.5922	0.1039
	CORSPCA	0.0211	4.9592	0.1479	5.5264	0.1738	4.1696	0.1214	4.0083	0.1436
	Fantope	0.0144	5.9651	0.2834	6.2778	0.2349	7.9504	0.2558	11.2510	0.3142
	ADAL	0.0254	6.5946	0.1975	6.8540	0.2040	4.6923	0.1442	3.1897	0.1600
	PSPCA(DT)	0.0101	4.7794	0.1381	5.2343	0.1543	3.8966	0.1139	2.3674	0.0857
	PSPCA(ADAL)	0.0092	4.9683	0.1430	5.3426	0.1556	3.8470	0.1113	2.3351	0.0817
2	ITSPCA	0.0173	19.0499	0.5489	21.8101	0.6618	17.4425	0.5880	11.7352	0.3994
	DTSPCA	0.0319	18.1165	0.5275	20.1756	0.6125	15.4730	0.5415	11.9668	0.4138
	AUGSPCA	0.0210	16.8681	0.4885	19.0664	0.5780	14.2196	0.5128	11.2295	0.3980
	CORSPCA	0.0202	16.6373	0.4837	19.1245	0.5820	14.5630	0.5195	11.5600	0.4051
	Fantope	0.0208	7.4746	0.4027	7.9962	0.3073	11.1917	0.4292	17.3186	0.5399
	ADAL	0.0605	25.8973	0.7176	27.3344	0.7992	17.0537	0.6131	10.8361	0.3818
	PSPCA(DT)	0.0172	17.4207	0.4949	19.4514	0.5786	14.1997	0.5129	9.4161	0.3473
	PSPCA(ADAL)	0.0152	22.3498	0.6119	24.4374	0.7172	15.8997	0.5576	9.0777	0.3328

PS: In Block 1, $[\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [100, 75, 50, 25]$. In Block 2, $[\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [60, 55, 50, 45]$.