

Federated PCA and Estimation for Spiked Covariance Matrices: Optimal Rates and Efficient Algorithm

Jingyang Li¹, T. Tony Cai², Dong Xia³ and Anru R. Zhang⁴

¹ Department of Statistics and Department of Mathematics, University of Michigan Ann Arbor

² Department of Statistics and Data Science, University of Pennsylvania

³ Department of Mathematics, Hong Kong University of Science and Technology

⁴ Department of Biostatistics & Bioinformatics and Department of Computer Science, Duke University

(November 26, 2024)

Abstract

Federated Learning (FL) has gained significant recent attention in machine learning for its enhanced privacy and data security, making it indispensable in fields such as healthcare, finance, and personalized services. This paper investigates federated PCA and estimation for spiked covariance matrices under distributed differential privacy constraints.

We establish minimax rates of convergence, with a key finding that the central server's optimal rate is the harmonic mean of the local clients' minimax rates. This guarantees consistent estimation at the central server as long as at least one local client provides consistent results. Notably, consistency is maintained even if some local estimators are inconsistent, provided there are enough clients. These findings highlight the robustness and scalability of FL for reliable statistical inference under privacy constraints.

To establish minimax lower bounds, we derive a matrix version of van Trees' inequality, which is of independent interest. Furthermore, we propose an efficient algorithm that preserves differential privacy while achieving near-optimal rates at the central server, up to a logarithmic factor. We address significant technical challenges in analyzing this algorithm, which involves a three-layer spectral decomposition. Numerical performance of the proposed algorithm is investigated using both simulated and real data.

1 Introduction

Principal Component Analysis (PCA) and its variants are fundamental tools in statistics and machine learning, particularly valuable for dimension reduction and data visualization when high-dimensional data lie in a low-dimensional space. PCA has been widely applied in data denoising and compression, feature extraction, clustering analysis, factor models, correlation analysis, and regression analysis. Population principal components are typically defined using the leading eigenvectors of population covariance matrices. The estimation and inference of these components from sample data have been extensively studied across various fields, including probability, statistics, and machine learning (see, for example, [Vershynin \(2018\)](#); [Jolliffe and Cadima \(2016\)](#); [Silverstein and Bai \(1995\)](#); [Bickel and Levina \(2008\)](#); [Koltchinskii and Lounici \(2016\)](#); [Benaych-Georges and Nadakuditi \(2011\)](#); [Cai et al. \(2013, 2015\)](#); [Zhang et al. \(2022\)](#); [Cai and Zhang \(2016\)](#)). See [Cai et al. \(2016\)](#) for a survey on optimal estimation of high-dimensional covariance structures.

With the digital shift in human activities, such as social networking, online shopping, and healthcare, vast amounts of personal information are collected and analyzed by large information technology firms and governmental organizations. The centralization of data storage raises critical concerns about the misuse of sensitive personal information, whether through intentional abuse or unintentional leaks. Traditional privacy-preserving methods like anonymization have proven insufficient, particularly in the context of classical PCA. As shown by [Dwork et al. \(2006\)](#) and [Chaudhuri et al. \(2013\)](#), classical PCA is vulnerable to alterations in individual data points and poses a significant risk of personal information leakage.

Differential privacy (DP) provides a robust framework to ensure that published statistics do not reveal whether any individual’s data was included in the dataset. Initially introduced by [Dwork et al. \(2006\)](#), DP has become a widely accepted standard in both industrial and governmental applications ([Erlingsson et al., 2014](#); [Ding et al., 2017](#); [Apple Differential Privacy Team, 2017](#); [Abowd, 2016](#); [Abowd et al., 2020](#)). DP is typically achieved by adding random noise to statistical outputs, using mechanisms such as the Gaussian or Laplace mechanisms. However, this randomization can compromise the accuracy of statistical methods. Consequently, a growing body of literature explores the trade-offs between privacy and accuracy in fundamental statistical and machine learning problems. The minimax optimal rates for differentially private PCA and covariance matrix estimation under the spiked model are established in [Cai et al. \(2024b\)](#).

Federated Learning (FL) is a decentralized machine learning framework where local clients train their models and communicate with a central server without sharing raw data ([McMahan et al., 2017](#)). Instead, clients privatize their learned models and share them with the central server or other clients, enabling collaborative machine learning while maintaining data privacy. Federated

learning has applications in healthcare, finance, Internet of Things (IoT), and more. However, the heterogeneity of datasets, privacy constraints, and the increasing number of local clients pose significant challenges to understanding the theoretical performance of federated learning. Under the DP constraint, the special case where each client holds only one datum is referred to as the *local* differential privacy setting (Duchi et al., 2013).

This paper investigates the minimax optimal rates in federated PCA under the spiked covariance model, considering diverse privacy constraints and sample sizes at local clients. A surprising and significant finding is that the minimax optimal rates achieved by the central server are the (scaled) harmonic mean of the minimax optimal rates achieved by local clients. This indicates that federated learning is multiply robust, meaning the central server attains a consistent estimator as long as at least one local client provides a consistent estimator. We believe this phenomenon is general and applies to many other federated learning problems under DP constraints.

The lower bound is established by leveraging a matrix version of the van Trees’ inequality, inspired by a similar strategy in Cai et al. (2024a). This matrix version of van Trees’ inequality is of independent interest. Additionally, we develop a computationally efficient algorithm that preserves DP at local clients and achieves the minimax optimal rate at the central server (up to logarithmic factors). The final estimator is obtained by applying three layers of spectral decomposition, posing significant technical challenges in deriving the sharp upper bound.

When there is only one local client, federated PCA simplifies to DP-PCA, and the upper bound derived in this study aligns with the results presented in Cai et al. (2024b). However, we emphasize that the technical contributions of these two works are fundamentally distinct. The primary contribution of Cai et al. (2024b) lies in the precise characterization of the sensitivity of empirical spectral projectors and eigenvalues under the spiked covariance model, which serves as the foundation for our methodology and theoretical framework for Federated PCA presented in this paper. Specifically, their results are directly leveraged to determine the appropriate level of artificial noise to be added at each local client. In contrast, the technical challenges in Federated PCA stem from the need to perform a sharp analysis of aggregated PCA across multiple clients. Our proposed method involves not one, but *three* layers of spectral decomposition, and the precise perturbation analysis of the final estimator relies on an explicit characterization of both the stochastic error and the artificial noise introduced in the first and second layers, respectively. Each layer’s spectral decomposition is represented by a Neumann series expansion, leading to a composition of three Neumann series. Consequently, we had to develop a unified strategy to derive concentration bounds for numerous higher-order perturbation terms, which required new techniques beyond those used in single-client DP-PCA. See the proof sketch of Theorem 1 for more details on our approach.

1.1 Problem formulation

The spiked covariance model has been widely applied and extensively investigated for extracting low-dimensional covariance structure from potentially high-dimensional data. It has found applications in diverse fields such as genomics (Leek and Storey, 2007), wireless communication (Telatar, 1999), asset pricing (Chamberlain and Rothschild, 1982), econometrics (Fan et al., 2008), and population genetics (Patterson et al., 2006; Novembre and Stephens, 2008). Under the spiked model, the covariance matrix Σ is a low-rank deformation of the (scaled) identity matrix, which admits the following decomposition:

$$\Sigma = U\Lambda U^\top + \sigma^2 I_p, \tag{1}$$

where $U = (u_1, \dots, u_r) \in \mathbb{O}^{p \times r}$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ are the leading eigenvectors and eigenvalues of the low-rank deformation with $\lambda_1 \geq \dots \geq \lambda_r > 0$. Here I_p represents the $p \times p$ identity matrix and $\mathbb{O}^{p \times r}$ is the set of $p \times r$ matrices satisfying $U^\top U = I_r$.

Estimating the population covariance matrix and its leading eigenvectors from a random sample $X = (X_1, \dots, X_n) \in \mathbb{R}^{p \times n}$, where the column vectors are i.i.d. with $\text{cov}(X_i) = \Sigma$, is a fundamental problem in multivariate statistical analysis. The spiked structure often significantly facilitates the theoretical derivation of the distribution of the sample eigenvectors and eigenvalues. Minimax optimal PCA and covariance matrix estimation have been extensively studied under the spiked model. An incomplete list of representative work includes Nadler (2008); Donoho et al. (2018); Cai et al. (2010, 2016); Koltchinskii and Lounici (2017); Johnstone (2001); Fan et al. (2008); Paul (2007) and references therein.

Differential privacy (Dwork et al., 2006) is a framework designed to provide privacy guarantee when analyzing and sharing data. Let $X \in \mathbb{R}^{p \times n}$ be a data set consisting of n observations. In standard definitions, a matrix $X' \in \mathbb{R}^{p \times n}$ is called a neighboring data set of X if and only if X and X' differ by only one datum, i.e., one column of X is replaced by some other, possibly arbitrary, observation of the same dimension. In the context of PCA, since the observations in X are independently sampled from the same distribution, a neighboring data set X' is obtained by replacing one datum in X with an independent copy. For a given data set X and any $\varepsilon > 0, \delta \in [0, 1)$, a randomized algorithm A that maps X into $\mathbb{R}^{d_1 \times d_2}$ is called (ε, δ) -differentially private $((\varepsilon, \delta)$ -DP) over the data set X if

$$\mathbb{P}(A(X) \in \mathcal{Q}) \leq e^\varepsilon \cdot \mathbb{P}(A(X') \in \mathcal{Q}) + \delta,$$

for all measurable subset $\mathcal{Q} \subset \mathbb{R}^{d_1 \times d_2}$ and all neighboring data set X' .

Differentially private PCA algorithms have been proposed and investigated in Blum et al. (2005); Chaudhuri et al. (2011); Dwork et al. (2014b) by treating each datum X_i as a fixed vector. More

recently, [Liu et al. \(2022\)](#) and [Cai et al. \(2024b\)](#) studied the minimax optimal rates for differentially private PCA and covariance estimation under the spiked covariance model (1). In particular, [Cai et al. \(2024b\)](#) showed that the minimax optimal rates, up to logarithmic factors, are

$$\begin{aligned} \inf_{\widehat{U} \in \mathcal{U}_{\varepsilon, \delta}} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2 &\asymp \Psi_0^2(n, \varepsilon, \delta) := \left(\frac{\sigma^4}{\lambda^2} + \frac{\sigma^2}{\lambda} \right) \left(\frac{pr}{n} + \frac{p^2 r^2}{n^2 \varepsilon^2} \right); \\ \inf_{\widehat{\Sigma} \in \mathcal{M}_{\varepsilon, \delta}} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 &\asymp \lambda^2 \cdot \Psi_1(n, \varepsilon, \delta) + \lambda^2 \cdot \Psi_0(n, \varepsilon, \delta) \\ &:= \lambda^2 \left(\frac{r^2}{n} + \frac{r^4}{n^2 \varepsilon^2} \right) + \sigma^2 (\lambda + \sigma^2) \left(\frac{pr}{n} + \frac{p^2 r^2}{n^2 \varepsilon^2} \right), \end{aligned} \quad (2)$$

conditioned on $\max\{\Psi_0(n, \varepsilon, \delta), \Psi_1(n, \varepsilon, \delta)\} \leq \sqrt{r}$ (otherwise, a trivial estimator suffices) and under certain constraint on δ . The parameter set $\Theta(\lambda, \sigma^2)$ is defined by

$$\Theta(\lambda, \sigma^2) := \left\{ \Sigma = U\Lambda U^\top + \sigma^2 I : U \in \mathbb{O}^{p \times r}, \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r), c_0 \lambda \leq \lambda_r \leq \dots \leq \lambda_1 \leq C_0 \lambda \right\},$$

with universal constants $c_0, C_0 > 0$. Here $\mathcal{U}_{\varepsilon, \delta}$ and $\mathcal{M}_{\varepsilon, \delta}$ represent the collection of all (ε, δ) -DP estimators of U and Σ , respectively. The terms in (2) involving ε reflect the cost of privacy. The error bound of $\mathbb{E} \|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2$ consists of two terms, where $\lambda^2 \cdot \Psi_0^2(n, \varepsilon, \delta)$ and $\lambda^2 \cdot \Psi_1^2(n, \varepsilon, \delta)$ are mainly contributed from estimating the eigenvectors and eigenvalues, respectively.

We formulate the problem of differentially private federated PCA as follows. There are m local clients, where the j -th client holds data $\mathcal{D}_j := \{X_i^{(j)} \in \mathbb{R}^p, i = 1, \dots, n_j\}$ for each $j \in [m]$. Under the spiked model, we assume that $X_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} N(0, \Sigma)$ for all $j \in [m]$ and for all $i \in [n_j]$. Here n_j represents the sample size in the j -th local client. All the data share a common spiked covariance matrix, and we assume zero mean and Gaussian distribution for simplicity. There is a central server that can communicate with the local clients, whose goal is to estimate the underlying covariance matrix Σ and its principal components. Local clients have privacy constraints and cannot share raw data with the central server or other local clients. Let $\varepsilon_j > 0$ and $\delta_j \in [0, 1], j \in [m]$ be two given sequences representing the privacy budgets on all local clients. Basically, the j -th local client requires to achieve the $(\varepsilon_j, \delta_j)$ -differential privacy when communicating its local information to the central server. Denote $\boldsymbol{\varepsilon} := (\varepsilon_1, \dots, \varepsilon_m)^\top$, $\boldsymbol{\delta} := (\delta_1, \dots, \delta_m)^\top$, and $\mathbf{n} = (n_1, \dots, n_m)^\top$. An estimator is called *federated* $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ -DP if it is $(\varepsilon_j, \delta_j)$ -DP for data in the j -th local client for all $j \in [m]$. In this paper, we aim to investigate the minimax optimal rates for federated $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ -DP PCA and covariance matrix estimation under the spiked model. We also propose computationally and communication-efficient estimators that are federated $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ -DP and achieve the minimax optimality. By slightly abuse of notations, we denote $\mathcal{U}_{\mathbf{n}, \boldsymbol{\varepsilon}, \boldsymbol{\delta}}$ and $\mathcal{M}_{\mathbf{n}, \boldsymbol{\varepsilon}, \boldsymbol{\delta}}$ as the collection of all federated $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ -DP estimators of U and Σ , respectively, when the sample sizes in local clients are represented by \mathbf{n} .

1.2 Main contribution

In this paper, we establish the minimax optimal rates for federated PCA and covariance matrix estimation under the spiked model with distributed DP constraints. Recall the definitions of $\Phi_0(n_j, \varepsilon_j, \delta_j)$ and $\Phi_1(n_j, \varepsilon_j, \delta_j)$ in (2). Under mild conditions, these rates, up to logarithmic factors, are

$$\begin{aligned} \inf_{\widehat{U} \in \mathcal{U}_{n, \varepsilon, \delta}} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\widehat{U}\widehat{U}^\top - UU^\top\|_F^2 &\asymp \frac{1}{\sum_{j=1}^m \Psi_0^{-2}(n_j, \varepsilon_j, \delta_j)} \bigwedge r \\ &\asymp \left(\left(\frac{\sigma^2}{\lambda} + \frac{\sigma^4}{\lambda^2} \right) \frac{1}{\sum_{j=1}^m \left(\frac{n_j}{rp} \wedge \frac{n_j^2 \varepsilon_j^2}{r^2 p^2} \right)} \right) \bigwedge r, \end{aligned} \quad (3)$$

and

$$\begin{aligned} \inf_{\widehat{\Sigma} \in \mathcal{M}_{n, \varepsilon, \delta}} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\widehat{\Sigma} - \Sigma\|_F^2 &\asymp \left(\frac{\lambda^2}{\sum_{j=1}^m \Psi_0^{-2}(n_j, \varepsilon_j, \delta_j)} + \frac{\lambda^2}{\sum_{j=1}^m \Psi_1^{-2}(n_j, \varepsilon_j, \delta_j)} \right) \bigwedge (r\lambda^2) \\ &\asymp \left(\frac{(\lambda\sigma^2 + \sigma^4)}{\sum_{j=1}^m \left(\frac{n_j}{rp} \wedge \frac{n_j^2 \varepsilon_j^2}{r^2 p^2} \right)} + \frac{\lambda^2}{\sum_{j=1}^m \left(\frac{n_j}{r^2} \wedge \frac{n_j^2 \varepsilon_j^2}{r^4} \right)} \right) \bigwedge (r\lambda^2). \end{aligned} \quad (4)$$

The bounds in (3) and (4) show that the minimax optimal rates achievable by the central server are proportional to the *harmonic mean* of the minimax optimal rates achievable by local clients. By the harmonic mean inequality¹, we get

$$\frac{1}{\sum_{j=1}^m \Psi_0^{-2}(n_j, \varepsilon_j, \delta_j)} \leq \min \left\{ \min_{j \in [m]} \Psi_0^2(n_j, \varepsilon_j, \delta_j), \frac{\text{avg}\{\Psi_0^2(n_j, \varepsilon_j, \delta_j)\}_{j=1}^m}{m}, \frac{2\text{med}\{\Psi_0^2(n_j, \varepsilon_j, \delta_j)\}_{j=1}^m}{m} \right\},$$

where *avg* and *med* stand for the sample mean and median, respectively. Two intriguing implications can be derived from the aforementioned bound. First, federated PCA exhibits multiple robustness in the sense that the estimator produced by the central server remains consistent as long as at least one local estimator is consistent. Second, even if all local estimators are inconsistent, the central server can still deliver a consistent estimator provided the number of local clients $m \rightarrow \infty$.

Federated PCA reduces to the differentially private PCA problem when $m = 1$, in which case the bounds (3) and (4) align with the minimax optimal rates established in Cai et al. (2024b). Our results immediately imply a performance bound for (non-interactive) *local differentially private* (LDP) PCA under the spiked model. By setting $n_j \equiv r \equiv 1$ and assuming $\varepsilon_j \equiv \varepsilon = O(1)$, the bound (3) suggests that the rate of LDP PCA under the spiked model is $p^2/(m\varepsilon^2)$. Here, m represents the sample size. The minimax lower bound easily follows from Theorem 3. However, our proposed estimator from Algorithm 1 will require a strong signal-to-noise ratio condition as stated

¹Harmonic mean inequality: $\frac{m}{\sum_{i=1}^m a_i^{-1}} \leq \frac{\sum_{i=1}^m a_i}{m}$ and the fact: $\sum_{i=1}^m a_i^{-1} \geq \sum_{i=1}^{\lfloor m/2 \rfloor} a_i^{-1} \geq (m/2) \cdot \text{med}\{a_i^{-1}\}_{i=1}^m$ for positive numbers $0 < a_1 \leq a_2 \leq \dots \leq a_m$.

in Theorem 1 because spectral decomposition is implemented on a single datum. We leave this as future work.

1.3 Related work

Differentially private PCA was studied by Blum et al. (2005); Chaudhuri et al. (2011); Dwork et al. (2014b) in a deterministic setting without assuming data are independently sampled from a common distribution. Liu et al. (2022) investigated online methods and established the minimax optimal rate for rank-one PCA under the spiked model. Cai et al. (2024b) leveraged spectral tools and established the minimax optimal rates for general rank- r PCA and covariance matrix estimation. Federated PCA with homogeneous sample sizes and privacy constraints was studied by Grammenos et al. (2020), assuming data arrive sequentially and all data points are uniformly bounded. Their estimator is sub-optimal without exploiting the statistical properties of sample data under the spiked covariance model. Wang and Xu (2020) studied non-interactive local differentially private PCA assuming that the observations are sampled independently from a common distribution but are uniformly bounded.

1.4 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we introduce a federated algorithm for differentially private PCA and covariance estimation. The algorithm incorporates three layers of spectral decomposition and employs the Gaussian mechanism to ensure privacy guarantees. We demonstrate that the algorithm produces valid DP estimators of the population covariance matrix and its spectral projectors, achieving minimax optimal error rates up to logarithmic factors. Additionally, we provide a proof sketch of the main theorem, outlining the technical challenges and our proof strategy. Section 3 establishes the minimax lower bounds for differentially private federated PCA and covariance estimation. The proof leverages a matrix version of Van Tree’s inequality, which we believe is of independent interest. In Section 4, we comprehensively evaluate the performance of our algorithm through numerical experiments and real data analysis, comparing it with existing methods. All technical proofs are included in the supplementary material.

2 Optimal Federated PCA by Gaussian Mechanism

In this section, we present the federated PCA and covariance matrix estimators under distributed differential privacy constraints. Due to the different levels of sensitivity of eigenvectors and eigenvalues, our approach estimates the eigenvectors and eigenvalues separately. Based on the given

privacy budget, each local client produces its own differentially private estimator of the eigenvectors and send them to the central server. The central server aggregate these estimators with specially designed weights. Since the aggregation may break the geometric constraints of eigenvectors, an additional step of eigen-decomposition is applied, from which the spectral projector serves as the final estimator of eigenvectors. The algorithm essentially consists of *three* layers of spectral decomposition: two performed by the local clients and one by the central server. These multiple spectral decompositions are crucial for ensuring differential privacy and achieving minimax optimality. They pose significant technical challenges to the theoretical analysis, where we leverage sophisticated spectral representation tools (Xia, 2021; Cai et al., 2024b) to carefully examine the behavior of three-layer eigen-decompositions.

After the differentially private estimator of eigenvectors is determined, the central server broadcasts them back to the local clients. These are then used to produce differentially private estimators of eigenvalues at each local client according to the given privacy budget. The central server receives these estimators, aggregates them by a weighted sum, and outputs the final estimator of the covariance matrix. The details of our approach are summarized in Algorithm 1. The operation $\text{svd}_r(\cdot)$ returns the top- r left singular vectors of a matrix. For simplicity, we assume that the rank r and the nuisance noise level σ^2 are both known. The algorithmic parameters α_j and β_j represent the sensitivity levels of empirical eigenvectors and eigenvalues (up to rotations).

Lemma 1. *Suppose that $X_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} N(0, \Sigma)$ with $\Sigma \in \Theta(\lambda, \sigma^2)$ for $j \in [m]$ and $i \in [n_j]$. For any weight vectors $\mathbf{v} = (v_1, \dots, v_m)^\top$ and $\mathbf{w} = (w_1, \dots, w_m)^\top$, the output $\widehat{U}\widehat{U}^\top$ and $\widehat{\Sigma}$ by Algorithm 1 are federated (ϵ, δ) -differentially private with probability at least $1 - 20 \sum_{j=1}^m e^{-c_0(n_j \wedge p)} - \sum_{j=1}^m n_j^{-100}$ for some absolute constant $c_0 > 0$.*

By the post-processing property of differential privacy (Dwork et al., 2014a, Proposition 2.1), $\widehat{U}\widehat{U}^\top$ is federated (ϵ, δ) -DP as long as the estimator $\widehat{U}_j\widehat{U}_j^\top$ is (ϵ_j, δ_j) -DP at the j -th local client for all $j \in [m]$. The proof of Lemma 1 mainly focuses on establishing the privacy guarantee at local clients, which follows similarly to the proof of Lemma 2.2 in Cai et al. (2024b) except that we have an improved probability bound here.

The following theorem shows that the final estimator $\widehat{U}\widehat{U}^\top$ is minimax optimal, up to logarithmic factors and the dependence on δ_j 's, if the weights w_k are properly chosen. Recall that $\Psi_0(n_j, \epsilon_j, \delta_j)$, defined in (2), quantifies the error rate for $\mathbb{E}\|\widehat{U}_j\widehat{U}_j^\top - UU^\top\|_F$ achieved at the j -th local client.

Theorem 1. *Suppose $X_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} N(0, \Sigma)$ with $\Sigma \in \Theta(\lambda, \sigma^2)$, $n_j \geq C_1(r \log n_j + \log^2 n_j)$, $p \geq C_1 \log n_j$ for some large constant $C_1 > 0$, and define $\widetilde{\Psi}_0(n_j, \epsilon_j, \delta_j)$ as*

$$\widetilde{\Psi}_0(n_j, \epsilon_j, \delta_j) := \left(\frac{\sigma^2}{\lambda} + \sqrt{\frac{\sigma^2}{\lambda}} \right) \left(\sqrt{\frac{rp}{n_j}} + \frac{p\sqrt{r(r + \log n_j)}}{n_j \epsilon_j} \sqrt{\log \frac{2.5}{\delta_j}} \right) < c_1 \sqrt{r}, \quad \forall j \in [m]. \quad (5)$$

Algorithm 1 Differentially Private Federated PCA and Covariance Estimation

Input: sample data $\mathcal{D}_j := \{X_i^{(j)} : i \in [n_j]\}$ at the j -th local client and its privacy budget $(\varepsilon_j, \delta_j)$ for any $j \in [m]$; weights w_j and $v_j > 0$ satisfying $\sum_{j=1}^m w_j = \sum_{j=1}^m v_j = 1$.

▼ Part 1: PCA

for $j = 1, \dots, m$ **do** ▷ on each local client

Sample covariance matrix and eigenvectors

$$\widehat{\Sigma}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^{(j)} X_i^{(j)\top} \quad \text{and} \quad \widetilde{U}_j = \text{svd}_r(\widehat{\Sigma}_j),$$

Gaussian mechanism for ensuring $(\varepsilon_j, \delta_j)$ -DP:

$$\widehat{U}_j = \text{svd}_r(\widetilde{U}_j \widetilde{U}_j^\top + Z_j), \quad [Z_j]_{kl} = [Z_j]_{lk} \stackrel{i.i.d.}{\sim} N(0, \alpha_j^2), k > l, [Z_j]_{kk} \stackrel{i.i.d.}{\sim} N(0, 2\alpha_j^2),$$

$$\text{with } \alpha_j^2 := \frac{8}{\varepsilon_j^2} \log\left(\frac{2.5}{\delta_j}\right) \frac{\sigma^2}{\lambda} \left(\frac{\sigma^2}{\lambda} + 1\right) \frac{p(r+\log n_j)}{n_j^2}.$$

Send \widehat{U}_j to the central server.

end for

Weighted average: $\widehat{U} = \text{svd}_r(\sum_{j=1}^m w_j \widehat{U}_j \widehat{U}_j^\top)$. ▷ on central server

▼ Part 2: Covariance matrix estimation

Send \widehat{U} to local client ▷ on central server

for $j = 1, \dots, m$ **do**

$(\varepsilon_j, \delta_j)$ -DP estimator of eigenvalues: ▷ on each local client

$$\widehat{\Lambda}_j = \widehat{U}^\top (\widehat{\Sigma}_j - \sigma^2 I) \widehat{U} + E_j, \quad [E_j]_{kl} = [E_j]_{lk} \stackrel{i.i.d.}{\sim} N(0, \beta_j^2), k > l, [E_j]_{kk} \stackrel{i.i.d.}{\sim} N(0, 2\beta_j^2),$$

$$\text{with } \beta_j^2 := \frac{8}{\varepsilon_j^2} \log\left(\frac{2.5}{\delta_j}\right) \frac{\lambda^2 (r+\log n_j)^2 + \sigma^4 p^2}{n_j^2}.$$

Send $\widehat{\Lambda}_j$ to the central server.

end for

$\widehat{\Sigma} := \sum_{j=1}^m v_j \widehat{U} \widehat{\Lambda}_j \widehat{U}^\top + \sigma^2 I$. ▷ on central server

Output: \widehat{U} and $\widehat{\Sigma}$.

satisfying $\tilde{\Psi}_0(n_j, \varepsilon_j, \delta_j) < c_1\sqrt{r}$ for some small universal constant $c_1 \in (0, 1/2)$ for all $j \in [m]$. Let \hat{U} be the estimator output by Algorithm 1 with weight $w_k := \tilde{\Psi}_0^{-2}(n_k, \varepsilon_k, \delta_k) / \sum_{j=1}^m \tilde{\Psi}_0^{-2}(n_j, \varepsilon_j, \delta_j)$ for all $k \in [m]$. Then there exist absolute constants $c_2, C_2 > 0$ such that

$$\|\hat{U}\hat{U}^\top - UU^\top\|_{\text{F}}^2 \leq \frac{C_2}{\sum_{j=1}^m \tilde{\Psi}_0^{-2}(n_j, \varepsilon_j, \delta_j)} \bigwedge (2r), \quad (6)$$

which holds with probability at least $1 - 22 \sum_{j=1}^m e^{-c_2(n_j \wedge p)}$. Moreover, if $(\lambda/\sigma^2) \sum_{j=1}^m n_j \leq e^{c_2 \min_{j \in [m]}(n_j \wedge p)}$, then

$$\mathbb{E}\|\hat{U}\hat{U}^\top - UU^\top\|_{\text{F}}^2 \leq \frac{C_2}{\sum_{j=1}^m \tilde{\Psi}_0^{-2}(n_j, \varepsilon_j, \delta_j)} \bigwedge (2r). \quad (7)$$

Note that the order of $\tilde{\Psi}_0(n_j, \varepsilon_j, \delta_j)$ and $\Psi_0(n_j, \varepsilon_j, \delta_j)$ only differs by $O(\sqrt{\log(1/\delta_j)})$ and $O(\log n_j)$ factors. They represent the minimax optimal spectral norm rate of estimating UU^\top for the j -th local client. The condition $\tilde{\Psi}_0(n_j, \varepsilon_j, \delta_j) < \sqrt{r}$ requires that the differentially private estimator published by each local client is non-trivial and informative, albeit not necessarily consistent. Based on Theorem 1, the optimal weights w_k for aggregation are proportional to $\tilde{\Psi}_0^{-2}(n_k, \varepsilon_k, \delta_k)$, respectively. While the definitions of $\tilde{\Psi}_0(n_k, \varepsilon_k, \delta_k)$'s involve the unknown signal strength λ , the weight w_k only depends on known sample sizes and privacy constraints. In fact, we can simply set the following data-independent weight:

$$w_k := \frac{\sqrt{p/n_k} + (p/n_k \varepsilon_k) \sqrt{(r + \log n_k) \log(2.5/\delta_k)}}{\sum_{j=1}^m \sqrt{p/n_j} + (p/n_j \varepsilon_j) \sqrt{(r + \log n_j) \log(2.5/\delta_j)}}, \quad \forall k \in [m].$$

In the homogeneous case where $n_k \asymp n$, $\varepsilon_k \asymp \varepsilon$, and $\delta_k \asymp \delta$ for all $k \in [m]$, these weights are $w_k \asymp m^{-1}$ of the same order.

The upper bound (6) is the (scaled) harmonic mean of the error bounds for $\|\hat{U}_j \hat{U}_j^\top - UU^\top\|_{\text{F}}^2$ for all $j \in [m]$. Let us briefly elaborate on the technical challenges. Under mild conditions, the Davis-Kahan theorem (Davis and Kahan, 1970) yields

$$\mathbb{E}\|\hat{U}\hat{U}^\top - UU^\top\|_{\text{F}}^2 \lesssim \mathbb{E}\left\|\sum_{j=1}^m w_j \Delta_j\right\|_{\text{F}}^2 = \sum_{j=1}^m w_j^2 \mathbb{E}\|\Delta_j\|_{\text{F}}^2 + \sum_{1 \leq k_1 \neq k_2 \leq m} w_{k_1} w_{k_2} \mathbb{E}\langle \Delta_{k_1}, \Delta_{k_2} \rangle, \quad (8)$$

where $\Delta_j := \hat{U}_j \hat{U}_j^\top - UU^\top$. The bound (7) is primarily contributed by the first term. It remains to carefully control the expected inner product $\mathbb{E}\langle \Delta_{k_1}, \Delta_{k_2} \rangle$, where the naive approach by applying the Cauchy-Schwartz inequality delivers a sub-optimal bound. We exploit the spectral representation formula from Xia (2021) to show that the second term in (8) is dominated by the first one.

Proof sketch of Theorem 1. There exist three layers of spectral decomposition in Algorithm 1. Applying the spectral representation formula from Xia (2021) to the last eigen-decomposition, we

obtain

$$\begin{aligned} & \frac{1}{2} \|\widehat{U}\widehat{U}^\top - UU^\top\|_F^2 \\ &= \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0}} \sum_{j_1, \dots, j_l \in [m]} w_{j_1} \cdots w_{j_l} \cdot \text{tr} \left(U^\top M(s_1) \underline{M(s_1)}^\top \Delta_{j_1} M(s_2) \right. \\ & \quad \left. \cdots \underline{M(s_l)}^\top \Delta_{j_l} M(s_{l+1}) M(s_{l+1})^\top U \right), \end{aligned} \quad (9)$$

where $M(s)$ is a matrix-valued function, such that $M(0) = U_\perp$ and $M(s) = U$ for $s > 0$, and $\mathbb{S}_l := \{\mathbf{s} = (s_1, \dots, s_{l+1})^\top \in \mathbb{Z}^{l+1} : s_1, \dots, s_{l+1} \geq 0, s_1 + \dots + s_{l+1} = l\}$. We use the underline below to emphasize the recurrent terms in the pattern $M(s_l) \Delta_j M(s_{l+1})$. Essentially, three different patterns of terms appear in the summands of products in eq. (9): $U^\top \Delta_j U$, $U^\top \Delta_j U_\perp$, and $U_\perp^\top \Delta_j U_\perp$.

Recall $\Delta_j = \widehat{U}_j \widehat{U}_j^\top - UU^\top$ where \widehat{U}_j consists of the top- r eigenvectors of $UU^\top + D_j$ with $D_j := \widetilde{U}_j \widetilde{U}_j^\top - UU^\top + Z_j$. Similarly, we can write

$$\Delta_j = \sum_{l \geq 1} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} M(s_1) \cdot \underline{M(s_1)}^\top D_j M(s_2) \cdots \underline{M(s_l)}^\top D_j M(s_{l+1}) \cdot M(s_{l+1})^\top \quad (10)$$

and

$$\begin{aligned} \widetilde{U}_j \widetilde{U}_j^\top - UU^\top &= \sum_{l \geq 1} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \\ & \quad M(s_1) \Lambda^{-s_1} \underline{M(s_1)}^\top \Xi_j M(s_2) \Lambda^{-s_2} \cdots \Lambda^{-s_l} \underline{M(s_l)}^\top \Xi_j M(s_{l+1}) \Lambda^{-s_{l+1}} M(s_{l+1})^\top. \end{aligned} \quad (11)$$

The above representation formulas show that the basic building elements are the terms $U^\top (\Xi_j + Z_j) U$, $U^\top (\Xi_j + Z_j) U_\perp$, and $U_\perp^\top (\Xi_j + Z_j) U_\perp$. As a result, we will show that there is an event \mathcal{E} with $\mathbb{P}(\mathcal{E}) \geq 1 - 14 \sum_{j=1}^m e^{-c_2(p \wedge n_j)}$, in which the following bounds hold

$$\begin{aligned} \max \left\{ \|U^\top \Delta_j U\|, \|U_\perp^\top \Delta_j U_\perp\| \right\} &\lesssim \left(\frac{\sigma^2}{\lambda} + \frac{\sigma^4}{\lambda^2} \right) \left(\frac{p}{n_j} + \frac{p^2(r + \log n_j)}{n_j^2 \varepsilon_j^2} \log \frac{2.5}{\delta_j} \right), \\ \|U^\top \Delta_j U_\perp\| &\lesssim \left(\frac{\sigma}{\sqrt{\lambda}} + \frac{\sigma^2}{\lambda} \right) \left(\sqrt{\frac{p}{n_j}} + \frac{p \sqrt{r + \log n_j}}{n_j \varepsilon_j} \log^{1/2} \frac{2.5}{\delta_j} \right). \end{aligned} \quad (12)$$

For each fixed $\mathbf{s} \in \mathbb{S}_l$, we consider the upper bound for

$$\left| \mathbb{E} \sum_{j_1, \dots, j_l \in [m]} w_{j_1} \cdots w_{j_l} \text{tr} \left(U^\top M(s_1) \underline{M(s_1)}^\top \Delta_{j_1} M(s_2) \cdots \underline{M(s_l)}^\top \Delta_{j_l} M(s_{l+1}) M(s_{l+1})^\top U \right) \cdot \mathbf{1}(\mathcal{E}) \right|. \quad (13)$$

The above summand is non-zero if and only if $s_1, s_{l+1} \geq 1$. Since $s_1 + \dots + s_{l+1} = l$, there exists $1 \leq i_1 < i_2 \leq l$, such that $s_{i_1} > 0, s_{i_1+1} = 0$, and $s_{i_2} = 0, s_{i_2+1} > 0$. We define

$$\mathbb{I}_1(\mathbf{s}) = \left\{ \mathbf{j} \in [m]^l : j_{i_1} \neq j_{i_2}, \{j_{i_1}, j_{i_2}\} \cap \{j_1, \dots, \bar{j}_{i_1}, \dots, \bar{j}_{i_2}, \dots, j_l\} = \emptyset \right\}.$$

Here $\bar{\cdot}$ means \cdot is absent from the set. Then $|\mathbb{I}_1(\mathbf{s})| = m(m-1)(m-2)^{l-2}$. Denote $\mathbb{I}_2(\mathbf{s}) := [m]^l \setminus \mathbb{I}_1(\mathbf{s})$. The sum in (13) can be decomposed into two parts: over $\mathbb{I}_1(\mathbf{s})$ and $\mathbb{I}_2(\mathbf{s})$, respectively. The proof is concluded by bounding the summands in (13) for all $\mathbf{j} \in \mathbb{I}_s(\mathbf{s})$ using the facts (12). \square

We now show that the covariance matrix estimator $\widehat{\Sigma}$ output by Algorithm 1 achieves the minimax optimal rate. For each $j \in [m]$, define

$$\widetilde{\Psi}_1(n_j, \varepsilon_j, \delta_j) := \sqrt{\frac{r(r + \log n_j)}{n_j}} + \frac{\sqrt{r(r + \log n_j)^3}}{n_j \varepsilon_j} \sqrt{\log \frac{2.5}{\delta_j}}, \quad (14)$$

which satisfies $\widetilde{\Psi}_1(n_j, \varepsilon_j, \delta_j) \asymp \Psi_1(n_j, \varepsilon_j, \delta_j)$, up to $O(\sqrt{\log(1/\delta_j)})$ and $O(\log n_j)$ factors. Recall that $\lambda \cdot \widetilde{\Psi}_1(n_j, \varepsilon_j, \delta_j)$ quantifies the error rate for estimating eigenvalues under the $(\varepsilon_j, \delta_j)$ -DP constraint achieved by the j -th local client.

Theorem 2. *Suppose the conditions in Theorem 1 hold, and set the weights in Algorithm 1 such that $\sum_{j=1}^m v_j = 1$ and*

$$v_j \propto \left(\frac{\lambda^2 + \sigma^4}{n_j} + \frac{8}{\varepsilon_j^2} \log \left(\frac{2.5}{\delta_j} \right) \frac{\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2}{n_j^2} \right)^{-1}.$$

There exist absolute constants $c_2, C_2 > 0$ such that the bound

$$\|\widehat{\Sigma} - \Sigma\|_{\text{F}}^2 \leq C_2 \left(\frac{\lambda^2}{\sum_{j=1}^m \widetilde{\Psi}_1^{-2}(n_j, \varepsilon_j, \delta_j)} + \frac{\lambda^2}{\sum_{j=1}^m \widetilde{\Psi}_0^{-2}(n_j, \varepsilon_j, \delta_j)} \right) \wedge (2r\lambda^2)$$

holds with probability at least $1 - 23 \sum_{j=1}^m e^{-c_0(n_j \wedge p)} - \sum_{j=1}^m n_j^{-100}$. Moreover, if $(\lambda/\sigma^2) \sum_{j=1}^m n_j \leq e^{c_0 \min_{j \in [m]} (n_j \wedge p)}$, then we have

$$\mathbb{E} \|\widehat{\Sigma} - \Sigma\|_{\text{F}}^2 \leq C_2 \left(\frac{\lambda^2}{\sum_{j=1}^m \widetilde{\Psi}_1^{-2}(n_j, \varepsilon_j, \delta_j)} + \frac{\lambda^2}{\sum_{j=1}^m \widetilde{\Psi}_0^{-2}(n_j, \varepsilon_j, \delta_j)} \right) \wedge (2r\lambda^2). \quad (15)$$

Our proposed Algorithm 1 separately estimates the eigenvectors and eigenvalues under privacy constraints. The central server aggregates differentially private estimators of both the eigenvalues and eigenvectors sent from the local clients. Therefore, the bound (15) involves two terms, primarily contributed by the estimation of the eigenvalues and eigenvectors, respectively. The bound (8) demonstrates the (doubly) multiple robustness of the estimator $\widehat{\Sigma}$. As long as one client can provide a consistent estimator of the eigenvectors and another (can be the same client) can provide a consistent estimator of the eigenvalues, the aggregated estimator $\widehat{\Sigma}$ delivered by the central server remains consistent. The weights v_j rely on the unknown eigenvalue λ . For simplicity, the empirical eigenvalue can be used in practice. Alternatively, one can resort to random matrix theory (Benaych-Georges and Nadakuditi, 2011) to obtain a sharper estimate of λ .

In the homogeneous case when $n_j \asymp n, \varepsilon_j \asymp \varepsilon$ and $\delta_j \asymp \delta$, we have $\Psi_0(n_j, \varepsilon_j, \delta_j) \asymp \Psi_0(n, \varepsilon, \delta)$ and $\Psi_1(n_j, \varepsilon_j, \delta_j) \asymp \Psi_1(n, \varepsilon, \delta)$ for all $j \in [m]$. Theorems 1 and 2 show that the estimators \widehat{U} and $\widehat{\Sigma}$ output by Algorithm 1 achieve the rates (up to logarithmic factors):

$$\begin{aligned}\mathbb{E}\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\text{F}}^2 &\lesssim \left(\frac{\sigma^4}{\lambda^2} + \frac{\sigma^2}{\lambda}\right) \left(\frac{pr}{mn} + \frac{p^2r^2}{mn^2\varepsilon^2}\right) \wedge r; \\ \mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\text{F}}^2 &\lesssim \lambda^2 \left(\frac{r^2}{mn} + \frac{r^4}{mn^2\varepsilon^2}\right) + \sigma^2(\lambda + \sigma^2) \left(\frac{pr}{mn} + \frac{p^2r^2}{mn^2\varepsilon^2}\right) \wedge (r\lambda^2),\end{aligned}$$

which decay whenever the number of local clients m or local sample size n increases. The aggregate sample size across all local clients is mn . The statistical error, quantified by the rate $pr/(mn)$, is inversely proportional to this total sample size. Notably, this rate aligns with the minimax optimal rate achievable by estimators that utilize all observations collectively (Cai et al., 2016). This implies that distributing observations evenly among m local clients does not compromise statistical efficiency. In contrast, the privacy cost is represented by the rate $p^2r^2/(mn^2\varepsilon^2)$, which decreases as the number of local clients increases. As demonstrated in Cai et al. (2024b), the rate $p^2r^2/(n^2\varepsilon^2)$ reflects the privacy cost at each individual local client. This suggests that aggregating multiple differentially private estimators can effectively reduce the overall privacy cost. Another interpretation of the rate $p^2r^2/(mn^2\varepsilon^2)$ is to express it as $m \cdot p^2r^2/(m^2n^2\varepsilon^2)$, where mn in the denominator represents the total sample size. When the total sample size is fixed, the privacy cost increases with the number of local clients m . This is because maintaining differential privacy becomes more challenging as the number of observations per local client decreases.

3 Minimax Lower Bound

In this section, we establish the minimax lower bounds for PCA and covariance matrix estimation under the federated (ε, δ) -DP constraints. These lower bounds match, up to logarithmic factors and δ_j -terms, the upper bounds achieved by our proposed estimators derived from Algorithm 1.

Under the spiked model with a covariance matrix $\Sigma \in \Theta(\lambda, \sigma^2)$, we denote $\mathcal{U}_{\mathbf{n}, \varepsilon, \delta}$ and $\mathcal{M}_{\mathbf{n}, \varepsilon, \delta}$ the collection of all federated (ε, δ) -DP estimators of U and Σ , respectively. The vector $\mathbf{n} := (n_1, \dots, n_m)^\top$ stands for the sample sizes at local clients. Recall that the rates $\Psi_0(n_j, \varepsilon_j, \delta_j)$ and $\Psi_1(n_j, \varepsilon_j, \delta_j)$ defined in (2) characterize the minimax optimal rates for differentially private estimators achievable at the j -th local client. Moreover, $\Psi_0(n_j, \varepsilon_j, \delta_j) \asymp \widetilde{\Psi}_0(n_j, \varepsilon_j, \delta_j)$ and $\Psi_1(n_j, \varepsilon_j, \delta_j) \asymp \widetilde{\Psi}_1(n_j, \varepsilon_j, \delta_j)$, up to logarithmic factors, for all $j \in [m]$. For presentation clarity, the following theorem focuses on the case $\varepsilon_j = O(1)$ for all $j \in [m]$.

Theorem 3. *Suppose $X_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} N(0, \Sigma)$, $p \geq 2r$, and $\max_{j \in [m]} \varepsilon_j \leq C_0$ for some large absolute constant $C_0 > 0$. There exist absolute constants $c_0, c_1 > 0$ such that if $(rp + \sqrt{rpn_j})\delta_j^{0.9} \leq c_1 n_j \varepsilon_j^2$*

for all $j \in [m]$, then

$$\begin{aligned} \inf_{\hat{U} \in \mathcal{U}(\mathbf{n}, \varepsilon, \delta)} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\hat{U}\hat{U}^\top - UU^\top\|_{\text{F}}^2 &\geq \frac{c_0}{\sum_{j=1}^m \Psi_0^{-2}(n_j, \varepsilon_j, \delta_j)} \bigwedge (2r), \\ \inf_{\hat{\Sigma} \in \mathcal{M}(\mathbf{n}, \varepsilon, \delta)} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\hat{\Sigma} - \Sigma\|_{\text{F}}^2 &\geq c_0 \left(\frac{\lambda^2}{\sum_{j=1}^m \Psi_0^{-2}(n_j, \varepsilon_j, \delta_j)} + \frac{\lambda^2}{\sum_{j=1}^m \Psi_1^{-2}(n_j, \varepsilon_j, \delta_j)} \right) \bigwedge (2r\lambda^2). \end{aligned} \tag{16}$$

In the special case of $m = 1$, the bound (16) matches the lower bound for differentially private PCA established in Cai et al. (2024b). Theorem 3 shows that the minimax lower bound in federated PCA is the harmonic mean of the minimax lower bounds at each local client. The technical tool in Cai et al. (2024b) is a differentially private version of Fano’s lemma, which imposes a restricted condition on the range of allowed δ_j ’s. In contrast, Theorem 3 allows a much wider range of δ_j ’s. We remark that the exponent 0.9 can be replaced by $k/(k+1)$ for any positive integer $k \geq 1$. The minimax lower bounds in Theorem 3 hold as long as $(rp + \sqrt{rpn_j})\delta_j^{1-\zeta} \leq c_1 n_j \varepsilon_j^2$ for any $\zeta \in (0, 1)$.

Our main technical tool for proving Theorem 3 is a matrix version of Van Tree’s inequality, which quantifies a lower bound for the average error rate of estimating principal components under privacy constraints. We then establish the inequality (16) by specifying a prior distribution over the set $\mathbb{O}^{p \times r}$ and bounding the Fisher information. The detailed proof is provided in Appendix A.4 in the supplementary materials.

4 Numerical Experiments

Our proposed algorithm, Fed-DP-PCA, is easy to implement. In this section, we evaluate its numerical performance through simulations and demonstrate its practical utility by applying it to a lung cancer dataset. To provide a comprehensive evaluation, we also compare its performance against two alternative approaches: the equal-weight aggregation method and the Fed-DP-Oja algorithm (Grammenos et al., 2020; Liu et al., 2022).

4.1 Simulations

We present simulation results comparing our proposed algorithm, Fed-DP-PCA, with existing algorithms and their variations. Specifically, we evaluate the Fed-DP-Oja algorithm introduced in Grammenos et al. (2020), which addresses federated PCA under homogeneous sample sizes and privacy constraints. Additionally, we compare our approach with an alternative aggregation method that assigns equal weights to each client. We also examine a strategy where each local client transmits $\tilde{U}_j \tilde{U}_j^\top + Z_j$ to the central server. While this method ensures privacy protection, it is not an optimal estimator of principal components, as it generally fails to qualify as a valid spectral projector and

incurs additional communication costs. Nevertheless, we include the results from this approach as a reference. In all experiments, we set the covariance matrix to $\Sigma = \lambda UU^\top + I_p$, where $U \in \mathbb{R}^{p \times r}$ is an orthogonal matrix generated by extracting the left singular vectors of a randomly generated matrix with i.i.d. entries via QR decomposition. Performance is assessed using the projection distance between the estimated subspace and the true subspace, defined by $\|\widehat{U}\widehat{U}^\top - UU^\top\|_F$.

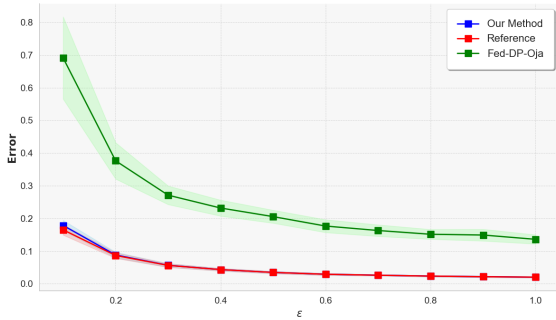
In the first simulation setting, we examine the utility-privacy trade-off under homogeneous conditions. We set the dimensionality to $p = 50$, rank to $r = 1$, and signal strength to $\lambda = 10$. The data are distributed across $m = 10$ clients, each with a privacy budget of $\varepsilon_j \equiv \varepsilon$ and $\delta_j \equiv 0.1$, and a sample size of $n_j = 10,000$. Given the homogeneous setting, the optimal choice of weights is equal weighting. Therefore, we compare our proposed method with the Fed-DP-Oja algorithm and the reference approach. The privacy budget ε varies between 0.1 and 1.0. For each choice of ε , the simulation is repeated 50 times. The results, presented in Figure 1a, demonstrate that the Fed-DP-Oja algorithm significantly underperforms compared to both our proposed method and the reference approach. In contrast, our method achieves performance nearly identical to the reference. These findings confirm that transmitting the top r left singular vectors of $\widetilde{U}_j\widetilde{U}_j^\top + Z_j$ to the server is sufficient for effective federated PCA. Additionally, larger values of ε correspond to weaker privacy guarantees but result in more accurate estimations. This behavior aligns with our theoretical predictions.

In the second experiment, we evaluate the estimation quality as the total number of total clients m varies. We use the same parameters: $p = 50, r = 1, \lambda = 10$. Each client is assigned a privacy budget of $\varepsilon_j \equiv 0.5$ and $\delta_j \equiv 0.1$. We consider a homogeneous setting where each client has a sample size of $n_j = n = 1000$ and vary the number of clients $m \in \{10, 20, \dots, 100\}$. For each value of m , the simulation is repeated 50 times. The results, depicted in Figure 1b, show that our proposed method achieves performance comparable to the reference approach while significantly outperforming the Fed-DP-Oja algorithm. Furthermore, as the number of clients increases, the estimation accuracy improves. These findings indicate that our method effectively leverages information from multiple clients, enhancing the quality of the estimated principal components as the client population grows.

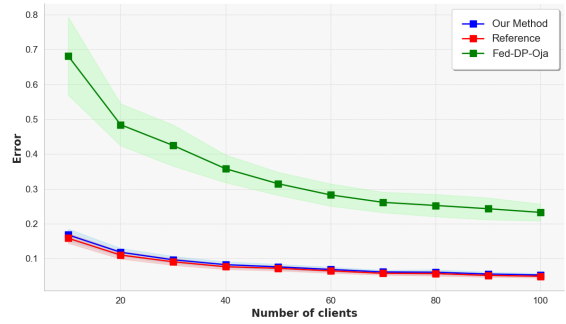
In the third experiment, we investigate the effect of varying the number of clients m on estimation quality while maintaining a fixed total number of samples N . Specifically, for each client j , the sample size is set to $n_j \equiv N/m$. We configure the parameters as $p = 50, r = 1, \lambda = 10$, with each client assigned a privacy budget of $\varepsilon_j \equiv 0.5$ and $\delta_j \equiv 0.1$. The total sample size is fixed at $N = 100,000$, and we vary the number of clients m across the values $\{10, 20, 25, 50\}$. For each configuration, we conduct 50 independent simulation runs. The results are illustrated in Figure 1c. The findings indicate that, under a fixed sample complexity, a smaller number of clients leads to more accurate estimations. This occurs because fewer clients allow for larger sample sizes

per client, thereby enhancing the quality of the local principal component estimates and facilitates easier privacy preservation. These results align with our theoretical predictions.

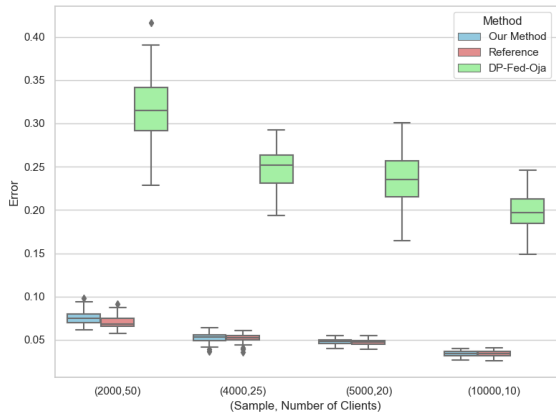
Lastly, we assess the performance of our method under heterogeneous sample sizes and privacy budgets. We set $p = 50, r = 1$, and $\lambda = 10$, with data distributed across $m = 10$ clients. For each client, the privacy parameters ε_j, δ_j are independently and uniformly drawn from $(0.1, 0.3)$ and $(0.1, 0.2)$, respectively. To introduce heterogeneity in sample sizes, we allocate a sample size of $2 * N_{\text{sample}}$ to the first five clients and $20 * N_{\text{sample}}$ to the remaining five clients, where $N_{\text{sample}} \in \{100, 200, \dots, 1000\}$. The results are presented in Figure 1d. Our proposed method outperforms the equal weight aggregation approach and even the reference method. This superior performance is attributed to our method's ability to optimally weight clients based on their individual sample sizes and privacy budgets, thereby effectively balancing the trade-offs inherent in a heterogeneous setting. In contrast, the reference method does not account for such heterogeneity in its weighting scheme, resulting in less efficient estimation.



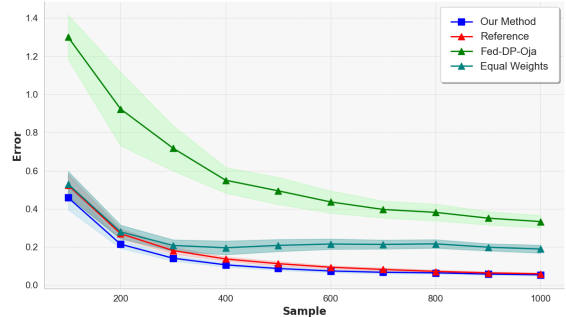
(a) Privacy-utility trade-off



(b) Estimation error versus number of clients



(c) Estimation error under fixed total sample size



(d) Heterogeneous sample sizes and privacy budgets

Figure 1: Numerical simulations comparing our method with existing methods and their variations. The performance is assessed using the projection distance $\|\widehat{U}\widehat{U}^\top - UU^\top\|_F$.

4.2 The Lung Cancer Data

In this section, we illustrate the practical utility of the proposed algorithm, Fed-DP-PCA, by applying it to a lung cancer dataset. We also compare its performance with the equal-weight aggregation approach and the Fed-DP-Oja algorithm.

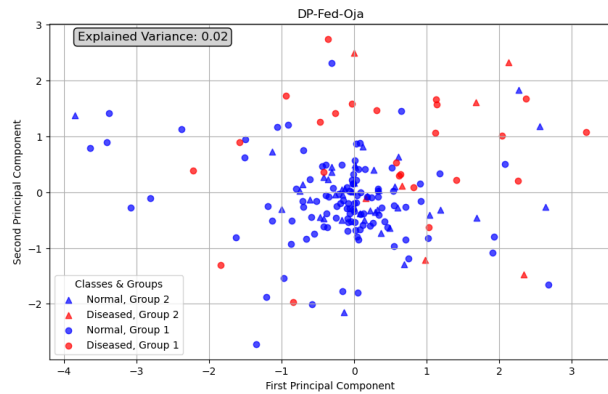
The Lung Cancer dataset, initially collected and cleaned by [Gordon et al. \(2002\)](#), comprises expression data for 12,533 genes across 181 subjects, categorized into diseased and normal groups. Following the refinement by [Jin and Wang \(2016\)](#), genes without differential expression between the groups were excluded, resulting in a curated data matrix with dimensions $p = 251$.

For our experiment, we consider a federated setting with $m = 2$ clients. We randomly shuffle the sample indices and assign the first 130 samples to Client 1, and the remaining 51 samples to Client 2. We set the target rank to $r = 5$. For each client, the signal strength λ is estimated by averaging the first three eigenvalues of the sample covariance matrix, and the noise variance σ^2 is estimated as the mean of the 51st to 251st sample eigenvalues. Both clients are allocated identical privacy budgets of $\varepsilon = 0.4$ and $\delta = 0.1$.

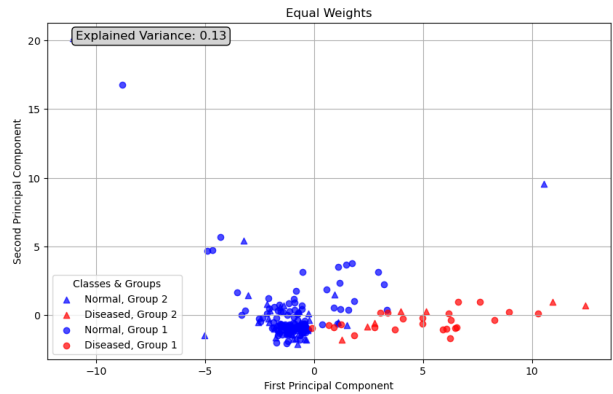
Subsequently, each client computes a differentially private subspace estimation and transmits it, along with the corresponding unnormalized weights, to a central server for aggregation. We compare the performance of our method with that of the Fed-DP-Oja algorithm and the equal-weight aggregation approach. After aggregation, we perform dimensionality reduction using the estimated subspace at the central server and report the explained variance as the evaluation metric. The results are illustrated in [Figures 2a, 2b, and 2c](#). The experimental outcomes indicate that our proposed method outperforms the Fed-DP-Oja algorithm, which requires the addition of excessively large noise, thereby degrading its performance. Moreover, when compared to the equal-weight aggregation approach, our method achieves a higher explained variance, demonstrating its superior ability to capture the underlying data structure effectively. These results underscore the efficacy of our method in balancing privacy constraints with estimation accuracy in a federated learning environment.

5 Discussions

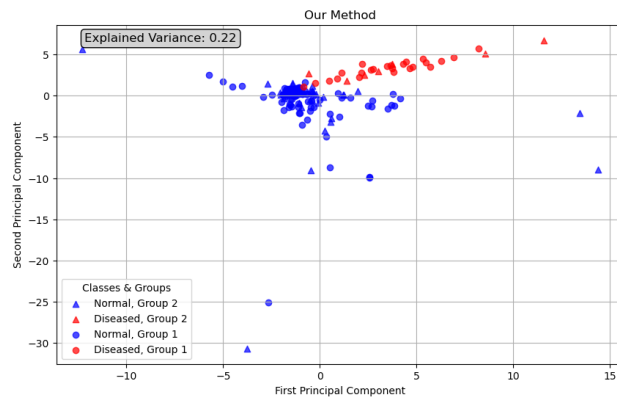
This paper establishes minimax optimal rates and demonstrates the multiple robustness and scalability of federated PCA. The central server’s estimator remains consistent as long as at least one local estimator is consistent. Moreover, even if all local estimators are inconsistent, the central estimator can still be consistent given a sufficient number of local clients. These findings highlight federated learning’s potential for reliable and robust statistical inference in a privacy-preserving manner, paving the way for further research and application in fields requiring stringent privacy



(a) Fed-DP-Oja



(b) Using equal weights



(c) Our method

Figure 2: We compare our proposed method with the equal-weight aggregation approach and the Fed-DP-Oja algorithm using the Lung Cancer dataset. For simplicity, the entire dataset is unevenly divided between two clients, each allocated a privacy budget of $\epsilon = 0.4, \delta = 0.1$.

and data security measures.

For simplicity, we assume in this paper that the mean vector of the data distribution is either zero or known. However, the approach can be readily adapted to handle cases where the mean is unknown. In such instances, we calculate the client-specific sample covariance matrix as $\widehat{\Sigma}_j = (n_j - 1)^{-1} \sum_{i=1}^{n_j} (X_i^{(j)} - \bar{X}^{(j)})(X_i^{(j)} - \bar{X}^{(j)})^\top$, where $\bar{X}^{(j)}$ denotes the sample mean vector for the j -th local client. Under the Gaussian assumption, the distribution of $(n_j - 1)\widehat{\Sigma}_j$ remains Wishart, which preserves the validity of all technical proofs presented in this work, except that the sensitivity of empirical spectral projectors and eigenvalues need to be carefully re-examined. For analytical convenience, we assume a Gaussian data distribution throughout our study. Extending these results to sub-Gaussian or more general distributions is an intriguing avenue for future research. Nevertheless, as highlighted earlier, the main technical challenges lie in developing a unified framework to bound higher-order perturbation terms that arise from the three layers of spectral decomposition.

An interesting future research direction is the study of federated SVD under the low-rank matrix denoising model. While SVD and PCA are closely related in traditional settings, they differ significantly in the context of federated learning under DP constraints due to differences in measurement units. Specifically, the covariance matrix is symmetric, whereas the low-rank signal in the matrix denoising model can have dimensions that differ drastically (Cai and Zhang, 2018). This introduces additional challenges and unique features when investigating minimax optimal rates for estimating the left and right singular subspaces under distributed differential privacy constraints. Nonetheless, we believe the multiple robustness phenomenon observed in federated PCA also applies to federated SVD, with the minimax optimal rate at the central server being the harmonic mean of the minimax optimal rates achievable at each local client.

Additionally, it is worthwhile to explore the minimax optimal rates in federated sparse PCA (Cai et al., 2013) and tensor PCA (Zhang and Xia, 2018) under privacy constraints. These problems often rely on iterative algorithms, making the development of sharp upper bounds technically challenging. Moreover, these settings are known to exhibit a statistical-to-computational gap even without privacy constraints. Understanding the interplay between privacy constraints and computational feasibility in these problems remains an open and important research problem.

6 Acknowledgment

Tony Cai’s research was supported in part by NSF grant DMS-2413106 and NIH grants R01-GM123056 and R01-GM129781. Dong Xia’s research was partially supported by Hong Kong RGC Grant GRF 16302323 and 16303224. Anru R. Zhang’s research was partially supported by NSF

Grant CAREER-2203741 and NIH Grants R01HL169347 and R01HL168940.

References

- John M Abowd. The challenge of scientific reproducibility and privacy protection for statistical agencies. *Census Scientific Advisory Committee*, 2016.
- John M Abowd, Ian M Rodriguez, William N Sexton, Phyllis E Singer, and Lars Villhuber. The modernization of statistical disclosure limitation at the us census bureau. *US Census Bureau*, 2020.
- Apple Differential Privacy Team. Learning with privacy at scale. 2017. URL <https://machinelearning.apple.com/research/learning-with-privacy-at-scale>.
- Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.
- Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604, 2008. doi: 10.1214/08-AOS600. URL <https://doi.org/10.1214/08-AOS600>.
- Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.
- T Tony Cai and Anru Zhang. Minimax estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis*, 150:55–74, 2016.
- T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60 – 89, 2018. doi: 10.1214/17-AOS1541. URL <https://doi.org/10.1214/17-AOS1541>.
- T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118 – 2144, 2010. doi: 10.1214/09-AOS752. URL <https://doi.org/10.1214/09-AOS752>.
- T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074 – 3110, 2013. doi: 10.1214/13-AOS1178. URL <https://doi.org/10.1214/13-AOS1178>.

- T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1 – 59, 2016. doi: 10.1214/15-EJS1081. URL <https://doi.org/10.1214/15-EJS1081>.
- T Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *arXiv preprint arXiv:2406.06755*, 2024a.
- T Tony Cai, Dong Xia, and Mengyue Zha. Optimal differentially private PCA and estimation for spiked covariance matrices. *arXiv preprint arXiv:2401.03820*, 2024b.
- Tony Cai, Zongming Ma, and Yihong Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3):781–815, 2015.
- Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets, 1982.
- Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.
- Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 14, 2013.
- Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.
- Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.
- Luc Devroye, Silvio Lattanzi, Gábor Lugosi, and Nikita Zhivotovskiy. On mean estimation for heteroscedastic random variables. In *Annales de l’Institut Henri Poincaré (B) Probabilités et statistiques*, volume 59, pages 1–20. Institut Henri Poincaré, 2023.
- Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.
- David L Donoho, Matan Gavish, and Iain M Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742, 2018.
- John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th annual symposium on foundations of computer science*, pages 429–438. IEEE, 2013.

- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.
- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014a.
- Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014b.
- Úlfar Erlingsson, Vasyli Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS '14*, page 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329576. doi: 10.1145/2660267.2660348. URL <https://doi.org/10.1145/2660267.2660348>.
- Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.
- Gavin J Gordon, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, Sridhar Ramaswamy, William G Richards, David J Sugarbaker, and Raphael Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62(17):4963–4967, 2002.
- Andreas Grammenos, Rodrigo Mendoza Smith, Jon Crowcroft, and Cecilia Mascolo. Federated principal component analysis. *Advances in Neural Information Processing Systems*, 33:6453–6464, 2020.
- Peisong Han and Lu Wang. Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430, 2013.
- Jiashun Jin and Wanjie Wang. Influential features PCA for high dimensional clustering. 2016.
- Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.
- Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

- Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1976 – 2013, 2016. doi: 10.1214/15-AIHP705. URL <https://doi.org/10.1214/15-AIHP705>.
- Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.
- Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.
- Xiyang Liu, Weihao Kong, Prateek Jain, and Sewoong Oh. DP-PCA: Statistically optimal and differentially private PCA. *Advances in Neural Information Processing Systems*, 35:29929–29943, 2022.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- Boaz Nadler. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics*, 36(6):2791 – 2817, 2008. doi: 10.1214/08-AOS618.
- John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.
- Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.
- Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.
- C Radhakrishna Rao. Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65(329):161–172, 1970.
- Jack W Silverstein and Zhi Dong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.
- Emre Telatar. Capacity of multi-antenna gaussian channels. *European transactions on telecommunications*, 10(6):585–595, 1999.
- Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

- Dietrich Von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, pages 97–109, 1988.
- Di Wang and Jinhui Xu. Principal component analysis in the local differential privacy model. *Theoretical Computer Science*, 809:296–312, 2020.
- Dong Xia. Non-asymptotic bounds for percentiles of independent non-identical random variables. *Statistics & Probability Letters*, 152:111–120, 2019.
- Dong Xia. Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851, 2021.
- Hui Yuan and Yingyu Liang. Learning entangled single-sample distributions via iterative trimming. In *International Conference on Artificial Intelligence and Statistics*, pages 2666–2676. PMLR, 2020.
- Anru Zhang and Dong Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.
- Anru R Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80, 2022.

A Proofs

A.1 Proof of Theorem 1

We first derive the upper bound for the expectation, and then derive the high probability upper bound. In the proof, we set $\lambda_{\max} = \lambda_1$ and $\lambda_{\min} = \lambda_r$.

Upper bound for expectation. We now derive the upper bound for $\mathbb{E}\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2$. We will first expand $\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2$. Denote $\Delta = \sum_{j=1}^m w_j \widehat{U}_j \widehat{U}_j^\top - UU^\top =: \sum_{j=1}^m w_j \Delta_j$. We define the event $\mathcal{F}_0 = \{\|\sum_{j=1}^m w_j \widehat{U}_j \widehat{U}_j^\top - UU^\top\| \leq 1/4\}$ and we will show shortly from (26), \mathcal{F}_0 holds with high probability. Notice that the columns of \widehat{U} are the top r left singular vectors of $\sum_{j=1}^m w_j \widehat{U}_j \widehat{U}_j^\top$, we can use the representation formula developed in Xia (2021) to show the following expansion holds under \mathcal{F}_0 :

$$\widehat{U}\widehat{U}^\top - UU^\top = \sum_{l \geq 1} \mathcal{S}_{UU^\top, l}(\Delta).$$

Here $\mathcal{S}_{UU^\top, l}(\Delta)$ takes the following form:

$$\mathcal{S}_{UU^\top, l}(\Delta) = \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} M(s_1) \underline{M(s_1)^\top \Delta M(s_2)} \cdots \underline{M(s_l)^\top \Delta M(s_{l+1})} M(s_{l+1})^\top,$$

where $M(s)$ is a matrix-valued function, such that $M(0) = U_\perp$ and $M(s) = U$ for $s > 0$, and

$$\mathbb{S}_l = \{(s_1, \dots, s_{l+1}) : s_1, \dots, s_{l+1} \geq 0, s_1 + \dots + s_{l+1} = l\}.$$

Under \mathcal{F}_0 , we can expand $\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2$ as

$$\begin{aligned} \frac{1}{2} \|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2 &= r - \langle \widehat{U}\widehat{U}^\top, UU^\top \rangle = -\langle \widehat{U}\widehat{U}^\top - UU^\top, UU^\top \rangle \\ &= -\sum_{l \geq 2} \langle \mathcal{S}_{UU^\top, l}(\Delta), UU^\top \rangle. \end{aligned}$$

Plug in the expression for $\mathcal{S}_{UU^\top, l}(\Delta)$, and we have the following expansion under \mathcal{F}_0 :

$$\begin{aligned} &\frac{1}{2} \|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2 \\ &= \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0}} \langle \underline{M(s_1) M(s_1)^\top \Delta M(s_2)} \cdots \underline{M(s_l)^\top \Delta M(s_{l+1})} M(s_{l+1})^\top, UU^\top \rangle \\ &= \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0}} \sum_{j_1, \dots, j_l \in [m]} w_{j_1} \cdots w_{j_l} \cdot \text{tr}(U^\top M(s_1) \underline{M(s_1)^\top \Delta_{j_1} M(s_2)} \\ &\quad \cdots \underline{M(s_l)^\top \Delta_{j_l} M(s_{l+1})} M(s_{l+1})^\top U). \end{aligned} \tag{17}$$

Here the first equality holds due to $\langle \mathcal{S}_{UU^\top,1}(\Delta), UU^\top \rangle = 0$. Recall $\Delta_j = \widehat{U}_j \widehat{U}_j^\top - UU^\top$. Notice \widehat{U}_j is the top r left singular vectors of $\widetilde{U}_j \widetilde{U}_j^\top + Z_j$, and U is the top r left singular vectors of UU^\top . We denote $D_j = \widetilde{U}_j \widetilde{U}_j^\top - UU^\top + Z_j$, then

$$\Delta_j = \widehat{U}_j \widehat{U}_j^\top - UU^\top = \sum_{l \geq 1} \mathcal{S}_{UU^\top, l}(D_j), \quad (18)$$

and

$$\mathcal{S}_{UU^\top, l}(D_j) = \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} M(s_1) \cdot \underline{M(s_1)^\top D_j M(s_2)} \cdots \underline{M(s_l)^\top D_j M(s_{l+1})} \cdot M(s_{l+1})^\top.$$

Therefore

$$\Delta_j = \sum_{l \geq 1} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} M(s_1) \cdot \underline{M(s_1)^\top D_j M(s_2)} \cdots \underline{M(s_l)^\top D_j M(s_{l+1})} \cdot M(s_{l+1})^\top. \quad (19)$$

Since $D_j = \widetilde{U}_j \widetilde{U}_j^\top - UU^\top + Z_j$, we consider the expression for $\widetilde{U}_j \widetilde{U}_j^\top - UU^\top$. Consider the event $\mathcal{E}_0^{(j)} = \{\|\widehat{\Sigma}_j - \Sigma\| \leq \lambda_{\min}/4\}$. Then from Lemma 5, we have $\mathbb{P}(\mathcal{E}_0^{(j)}) \geq 1 - e^{-p \wedge n_j}$. We denote $\Xi_j = \widehat{\Sigma}_j - \Sigma$, then under $\mathcal{E}_0^{(j)}$, we have the following expansion under the event $\mathcal{E}_0^{(j)}$:

$$\widetilde{U}_j \widetilde{U}_j^\top - UU^\top = \sum_{l \geq 1} \mathcal{S}_{U\Lambda U^\top, l}(\Xi_j),$$

where

$$\begin{aligned} & \mathcal{S}_{U\Lambda U^\top, l}(\Xi_j) \\ &= \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} M(s_1) \Lambda^{-s_1} \underline{M(s_1)^\top \Xi_j M(s_2)} \Lambda^{-s_2} \cdots \Lambda^{-s_l} \underline{M(s_l)^\top \Xi_j M(s_{l+1})} \Lambda^{-s_{l+1}} M(s_{l+1})^\top, \end{aligned} \quad (20)$$

here we denote $\Lambda^{-0} = I_{p-r}$ with slight abuse of notation. We denote $g_i^{(j)} = U^\top X_i^{(j)}$ and $h_i^{(j)} = U_\perp^\top X_i^{(j)}$. Then

$$g_i^{(j)} \sim N(0, \Lambda + \sigma^2 I_r), \quad h_i^{(j)} \sim N(0, \sigma^2 I_{p-r}).$$

We define the matrix $G^{(j)} \in \mathbb{R}^{r \times n_j}$, and $H^{(j)} \in \mathbb{R}^{(p-r) \times n_j}$ as

$$G^{(j)} = [g_1^{(j)}, \dots, g_{n_j}^{(j)}], \quad H^{(j)} = [h_1^{(j)}, \dots, h_{n_j}^{(j)}].$$

Then, $G^{(j)}$ and $H^{(j)}$ are independent. We also have

$$M(s_1)^\top \Xi_j M(s_2) = \begin{cases} \frac{1}{n_j} H^{(j)} H^{(j)\top} - \sigma^2 I_{d-r}, & \text{if } s_1 = s_2 = 0, \\ \frac{1}{n_j} H^{(j)} G^{(j)\top} & \text{if } s_1 = 0, s_2 > 0, \\ \frac{1}{n_j} G^{(j)} H^{(j)\top} & \text{if } s_1 > 0, s_2 = 0, \\ \frac{1}{n_j} G^{(j)} G^{(j)\top} - (\Lambda + \sigma^2 I_r) & \text{if } s_1, s_2 > 0. \end{cases} \quad (21)$$

Notice Z_j is a Gaussian orthogonal ensemble (GOE), and thus is invariant to orthogonal conjugation. Therefore, $U^\top Z_j U, U_\perp^\top Z_j U, U_\perp^\top Z_j U_\perp$ are independent. We will in the following denote

$$\begin{bmatrix} Z_{j,1} & Z_{j,2} \\ Z_{j,2}^\top & Z_{j,3} \end{bmatrix} = \begin{bmatrix} U^\top Z_j U & U^\top Z_j U_\perp \\ U_\perp^\top Z_j U & U_\perp^\top Z_j U_\perp \end{bmatrix}. \quad (22)$$

We recap the observations so far in the following diagram.

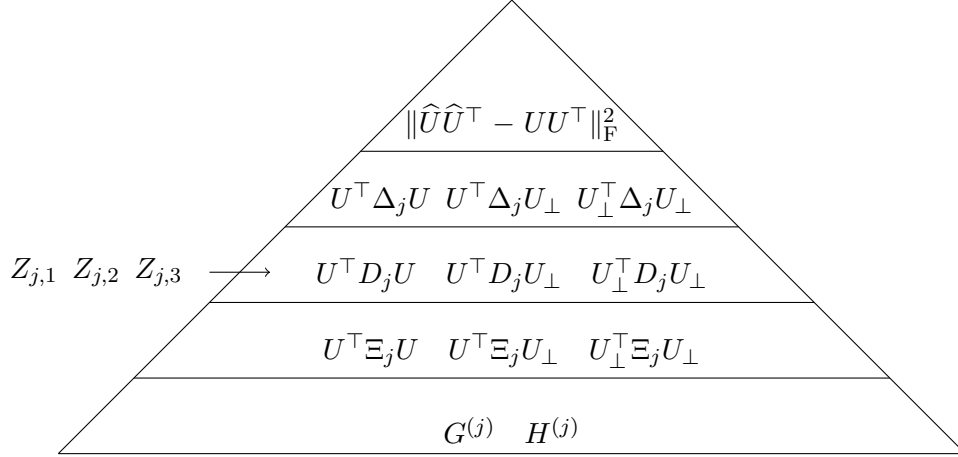


Figure 3: Layer by layer decomposition of $\|\hat{U}\hat{U}^\top - UU^\top\|_F^2$, with the building blocks $\{G^{(j)}, H^{(j)}, Z_{j,1}, Z_{j,2}, Z_{j,3}\}_{j=1}^M$

Now we analyze each terms in Figure 3 from bottom to top. We first consider $U^\top D_j U$:

$$U^\top D_j U = U^\top (\tilde{U}_j \tilde{U}_j^\top - UU^\top) U + U^\top Z_j U.$$

We have

$$\begin{aligned} & U^\top (\tilde{U}_j \tilde{U}_j^\top - UU^\top) U \\ &= \sum_{l \geq 1} \sum_{s \in \mathbb{S}_l} (-1)^{\|s\|_{\ell_0} + 1} U^\top M(s_1) \Lambda^{-s_1} \underline{M(s_1)^\top \Xi_j M(s_2)} \Lambda^{-s_2} \\ & \quad \dots \Lambda^{-s_l} \underline{M(s_l)^\top \Xi_j M(s_{l+1})} \Lambda^{-s_{l+1}} M(s_{l+1})^\top U. \end{aligned}$$

A simple fact is in each summand above, it is symmetric in both $G^{(j)}, H^{(j)}$. In details, we have

$$U^\top D_j U = f_1(G^{(j)}, H^{(j)}) + Z_{j,1},$$

where f_1 is a matrix-valued function such that $f_1(G^{(j)}, H^{(j)}) = f_1(-G^{(j)}, H^{(j)}) = f_1(G^{(j)}, -H^{(j)})$.

And we can similarly show

$$U_\perp^\top D_j U_\perp = f_3(G^{(j)}, H^{(j)}) + Z_{j,3},$$

for some f_3 such that $f_3(G^{(j)}, H^{(j)}) = f_3(-G^{(j)}, H^{(j)}) = f_3(G^{(j)}, -H^{(j)})$.

For $U^\top D_j U_\perp$, we have

$$U^\top D_j U_\perp = f_2(G^{(j)}, H^{(j)}) + Z_{j,2},$$

where $-f_2(G^{(j)}, H^{(j)}) = f_2(-G^{(j)}, H^{(j)}) = f_2(G^{(j)}, -H^{(j)})$.

Notice $U^\top D_j U$, $U^\top D_j U_\perp$, and $U_\perp^\top D_j U_\perp$ are the building blocks for Δ_j , we have

$$U^\top \Delta_j U = g_1(G^{(j)}, H^{(j)}, Z_{j,1}, Z_{j,2}, Z_{j,3}),$$

where g_1 is an even function in both $G^{(j)}, H^{(j)}, Z_{j,2}$. Similarly,

$$U^\top \Delta_j U_\perp^\top = g_2(G^{(j)}, H^{(j)}, Z_{j,1}, Z_{j,2}, Z_{j,3}), \quad (23)$$

where g_2 is an odd function in both $G^{(j)}, H^{(j)}, Z_{j,2}$, and

$$U_\perp^\top \Delta_j U_\perp^\top = g_3(G^{(j)}, H^{(j)}, Z_{j,1}, Z_{j,2}, Z_{j,3}),$$

where g_3 is an even function in both $G^{(j)}, H^{(j)}, Z_{j,2}$.

Upper bounds for $\|U^\top \Xi_j U\|, \|U^\top \Xi_j U_\perp\|, \|U_\perp^\top \Xi_j U_\perp\|$. We denote the normalized versions of $g_i^{(j)}, h_i^{(j)}$ as

$$\bar{g}_i^{(j)} = (\Lambda + \sigma^2 I_r)^{-1/2}, \quad \bar{h}_i^{(j)} = \sigma^{-1} h_i^{(j)},$$

and

$$\bar{G}^{(j)} = [\bar{g}_1^{(j)}, \dots, \bar{g}_{n_j}^{(j)}], \quad \bar{H}^{(j)} = [\bar{h}_1^{(j)}, \dots, \bar{h}_{n_j}^{(j)}].$$

Then we have

$$U^\top \Xi_j U_\perp = \frac{\sigma}{n_j} (\Lambda^{1/2} + \sigma I_r) \bar{G}^{(j)} \bar{H}^{(j)\top}.$$

Standard ε -net argument shows with probability exceeding $1 - 2e^{-p} - 2e^{-n_j}$,

$$\|U^\top \Xi_j U_\perp\| \lesssim (\lambda_{\max}^{1/2} + \sigma) \sigma \sqrt{\frac{p-r}{n_j}}.$$

For $U^\top \Xi_j U$, we have

$$U^\top \Xi_j U = \frac{1}{n_j} (\Lambda^{1/2} + \sigma I_r) \bar{G}^{(j)} \bar{G}^{(j)\top} (\Lambda^{1/2} + \sigma I_r) - (\Lambda + \sigma^2 I_r).$$

And with probability exceeding $1 - 2e^{-\eta_j}$,

$$\|U^\top \Xi_j U\| \lesssim (\lambda_{\max} + \sigma^2) \frac{\sqrt{r + \eta_j}}{\sqrt{n_j}}.$$

For $U_\perp^\top \Xi_j U_\perp$, we have

$$U_\perp^\top \Xi_j U_\perp = \frac{\sigma^2}{n_j} \bar{H}^{(j)} \bar{H}^{(j)\top} - \sigma^2 I_{d-r}.$$

And with probability exceeding $1 - 2e^{-p}$,

$$\|U_\perp^\top \Xi_j U_\perp\| \lesssim \sigma^2 \frac{\sqrt{p-r}}{\sqrt{n_j}}.$$

We define the event

$$\begin{aligned} \mathcal{E}_1^{(j)} := & \left\{ \|U^\top \Xi_j U\| \lesssim (\lambda_{\max}^{1/2} + \sigma) \sigma \sqrt{\frac{p-r}{n_j}} \right\} \cap \left\{ \|U^\top \Xi_j U\| \lesssim (\lambda_{\max} + \sigma^2) \frac{\sqrt{r + \eta_j}}{\sqrt{n_j}} \right\} \\ & \cap \left\{ \|U_\perp^\top \Xi_j U_\perp\| \lesssim \sigma^2 \sqrt{\frac{p-r}{n_j}} \right\}. \end{aligned} \quad (24)$$

Then $\mathbb{P}(\mathcal{E}_1^{(j)}) \geq 1 - 4e^{-p} - 2e^{-n_j} - 2e^{-\eta_j}$.

Upper bounds for $\|Z_{j,1}\|, \|Z_{j,2}\|$ and $\|Z_{j,3}\|$. From (22), we see that with probability exceeding $1 - 2e^{-\eta_j} - 4e^{-p}$,

$$\|Z_{j,1}\| \lesssim \alpha_j \sqrt{r + \eta_j}, \quad \|Z_{j,2}\| \lesssim \alpha_j \sqrt{p}, \quad \|Z_{j,3}\| \lesssim \alpha_j \sqrt{p}.$$

We define the event,

$$\mathcal{E}_2^{(j)} := \left\{ \|Z_{j,1}\| \lesssim \alpha_j \sqrt{r + \eta_j} \right\} \cap \left\{ \|Z_{j,2}\| \lesssim \alpha_j \sqrt{p} \right\} \cap \left\{ \|Z_{j,3}\| \lesssim \alpha_j \sqrt{p} \right\}, \quad (25)$$

$\mathcal{E}^{(j)} = \mathcal{E}_0^{(j)} \cap \mathcal{E}_1^{(j)} \cap \mathcal{E}_2^{(j)}$, and $\mathcal{E} := \bigcap_{j=1}^m \mathcal{E}^{(j)}$. Then $\mathbb{P}(\mathcal{E}^{(j)}) \geq 1 - 4e^{-\eta_j} - 10e^{-p \wedge n_j}$.

Upper bounds for $\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\mathbb{F}}^2 \cdot \mathbf{1}(\mathcal{E})$. Under the SNR condition, we have $\lambda_{\min}^{-1} \left((\lambda_{\max}^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}} \right) \lesssim 1$, and under the event \mathcal{E} , we have

$$\begin{aligned} \|U^\top D_j U\| & \leq C^2 \lambda_{\min}^{-2} \left((\lambda_{\max}^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}} \right)^2 \sum_{l \geq 2} 2^{-l+2} + C \alpha_j \sqrt{r + \eta_j} \\ & \lesssim \lambda_{\min}^{-2} (\lambda_{\max} + \sigma^2) \sigma^2 \frac{p}{n_j} + \alpha_j \sqrt{r + \eta_j}, \end{aligned}$$

and

$$\begin{aligned}\|U^\top D_j U_\perp\| &\leq C\lambda_{\min}^{-1}\left((\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}}\right)\sum_{l\geq 1}2^{-l+1} + C\alpha_j\sqrt{p} \\ &\lesssim \lambda_{\min}^{-1}\left((\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}}\right) + \alpha_j\sqrt{p},\end{aligned}$$

and similarly,

$$\begin{aligned}\|U_\perp^\top D_j U_\perp\| &\leq C^2\lambda_{\min}^{-2}\left((\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}}\right)^2\sum_{l\geq 2}2^{-l+2} + C\alpha_j\sqrt{p} \\ &\lesssim \lambda_{\min}^{-2}(\lambda_{\max} + \sigma^2)\sigma^2\frac{p}{n_j} + \alpha_j\sqrt{p}.\end{aligned}$$

As long as $\eta_j + r \leq p$ and since $\alpha_j\sqrt{p} \lesssim 1$, we have

$$\max\left\{\|U^\top D_j U\|, \|U^\top D_j U_\perp\|, \|U_\perp^\top D_j U_\perp\|\right\} \leq \lambda_{\min}^{-1}\left((\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}}\right) + \alpha_j\sqrt{p}. \quad (26)$$

We denote the right hand side bound $u_j = \lambda_{\min}^{-1}(\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}} + \alpha_j\sqrt{p}$. Notice under the given SNR, $\max_j u_j \leq \frac{1}{4}$. Together with (18), we conclude $\|\widehat{U}_j\widehat{U}_j^\top - UU^\top\| \leq \frac{1}{4}$. This also implies $\|\sum_{j=1}^m w_j\widehat{U}_j\widehat{U}_j^\top - UU^\top\| \leq 1/4$. That is, \mathcal{E} implies \mathcal{F}_0 .

For the terms related to Δ_j , we have

$$\begin{aligned}\|U^\top \Delta_j U\| &\leq \sum_{l\geq 2} C^l \left(\lambda_{\min}^{-1}\left((\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}}\right) + \alpha_j\sqrt{p}\right)^l \\ &\lesssim \lambda_{\min}^{-2}(\lambda_{\max} + \sigma^2)\sigma^2\frac{p}{n_j} + \alpha_j^2 p, \\ \|U^\top \Delta_j U_\perp\| &\lesssim \lambda_{\min}^{-1}(\lambda_{\max}^{1/2} + \sigma)\sigma\sqrt{\frac{p}{n_j}} + \alpha_j\sqrt{p}, \\ \|U_\perp^\top \Delta_j U_\perp\| &\lesssim \lambda_{\min}^{-2}(\lambda_{\max} + \sigma^2)\sigma^2\frac{p}{n_j} + \alpha_j^2 p.\end{aligned}$$

Then u_j is also the upper bound for $\|U^\top \Delta_j U\|$, $\|U^\top \Delta_j U_\perp\|$, and $\|U_\perp^\top \Delta_j U_\perp\|$ under \mathcal{E} .

Now we go back to (17). For each $\mathbf{s} \in \mathbb{S}_l$, in order for the summand in (17) to be non-zero, s_1, s_{l+1} should be strictly greater than 0. Since $s_1 + \dots + s_{l+1} = l$, there exists $1 \leq i_1 < i_2 \leq l$, such that $s_{i_1} > 0, s_{i_1+1} = 0$, and $s_{i_2} = 0, s_{i_2+1} > 0$. We define

$$\mathbb{I}_1(\mathbf{s}) = \left\{ \mathbf{j} \in [m]^l : j_{i_1} \neq j_{i_2}, \{j_{i_1}, j_{i_2}\} \cap \{j_1, \dots, \bar{j}_{i_1}, \dots, \bar{j}_{i_2}, \dots, j_l\} = \emptyset \right\}.$$

Here $\bar{\cdot}$ means \cdot is absent in the set. Then $|\mathbb{I}_1(\mathbf{s})| = m(m-1)(m-2)^{l-2}$. We define the complement of $\mathbb{I}_1(\mathbf{s})$ as $\mathbb{I}_2(\mathbf{s}) = [m]^l \setminus \mathbb{I}_1(\mathbf{s})$. Then

$$\mathbb{I}_2(\mathbf{s}) \subset \{ \mathbf{j} \in [m]^l : j_{i_1} = j_{i_2} \} \cup \left(\cup_{k \neq i_1, i_2} \{ \mathbf{j} \in [m]^l : j_{i_1} = j_k \} \right) \cup \left(\cup_{k \neq i_1, i_2} \{ \mathbf{j} \in [m]^l : j_{i_2} = j_k \} \right). \quad (27)$$

Next we consider the upper bound for

$$\left| \mathbb{E} \sum_{j_1, \dots, j_l \in [m]} w_{j_1} \cdots w_{j_l} \text{tr}(U^\top M(s_1) \underline{M(s_1)}^\top \Delta_{j_1} M(s_2) \cdots M(s_l)^\top \Delta_{j_l} M(s_{l+1}) \underline{M(s_{l+1})}^\top U) \cdot \mathbf{1}(\mathcal{E}) \right|.$$

We can split the above sum into two parts, namely

$$\sum_{j_1, \dots, j_l \in [m]} = \sum_{\mathbb{I}_1(\mathbf{s})} + \sum_{\mathbb{I}_2(\mathbf{s})}.$$

Notice from (23), we have $\mathbb{E} U^\top \Delta_{j_1} U_\perp \cdot \mathbf{1}(\mathcal{E}^{(j_{i_1})}) = 0$. Then for all $\mathbf{j} \in \mathbb{I}_1(\mathbf{s})$, we have

$$\mathbb{E} \text{tr}(U^\top M(s_1) \underline{M(s_1)}^\top \Delta_{j_1} M(s_2) \cdots M(s_l)^\top \Delta_{j_l} M(s_{l+1}) \underline{M(s_{l+1})}^\top U) \cdot \mathbf{1}(\mathcal{E}) = 0.$$

For each summand with index $\mathbf{j} \in \mathbb{I}_2(\mathbf{s})$, using Cauchy-Schwarz inequality, we have

$$|\text{tr}(U^\top M(s_1) \underline{M(s_1)}^\top \Delta_{j_1} M(s_2) \cdots M(s_l)^\top \Delta_{j_l} M(s_{l+1}) \underline{M(s_{l+1})}^\top U)| \leq r \cdot u_{j_1} \cdots u_{j_l}.$$

Using these facts, we have

$$\begin{aligned} & \left| \mathbb{E} \sum_{j_1, \dots, j_l \in [m]} w_{j_1} \cdots w_{j_l} \text{tr}(U^\top M(s_1) \underline{M(s_1)}^\top \Delta_{j_1} M(s_2) \cdots M(s_l)^\top \Delta_{j_l} M(s_{l+1}) \underline{M(s_{l+1})}^\top U) \cdot \mathbf{1}(\mathcal{E}) \right| \\ & \leq r \cdot \sum_{\mathbf{j} \in \mathbb{I}_2(\mathbf{s})} u_{j_1} \cdots u_{j_l}. \end{aligned}$$

Using the inclusion relation in (27), this is further upper bounded by

$$r \cdot \left(\sum_{j_1=j_2} + \sum_{k \neq i_1, i_2} \sum_{j_1=j_k} + \sum_{k \neq i_1, i_2} \sum_{j_2=j_k} \right) u_{j_1} \cdots u_{j_l} \leq 2lr \left(\sum_{k=1}^m w_k^2 u_k^2 \right) \left(\sum_{k=1}^m w_k u_k \right)^{l-2}. \quad (28)$$

Therefore we have

$$\mathbb{E} \|\widehat{U} \widehat{U}^\top - U U^\top\|_{\mathbb{F}}^2 \cdot \mathbf{1}(\mathcal{E}) \leq \sum_{l \geq 2} 4^l \cdot 2lr \left(\sum_{k=1}^m w_k^2 u_k^2 \right) \left(\sum_{k=1}^m w_k u_k \right)^{l-2} \leq 4r \left(\sum_{k=1}^m w_k^2 u_k^2 \right), \quad (29)$$

where the last inequality is due to $u_k \leq \frac{1}{2}$. Finally, we set w_k to be such that $\sum_{k=1}^m w_k^2 u_k^2$ is minimized, that is $w_k \propto u_k^{-2}$.

On the other hand, we can set $\eta_j = c_0(n_j \wedge p)$, then we have

$$\mathbb{E} \|\widehat{U} \widehat{U}^\top - U U^\top\|_{\mathbb{F}}^2 \cdot \mathbf{1}(\mathcal{E}^c) \leq 2r \cdot \mathbb{P}(\mathcal{E}^c) \leq 28r \sum_{j=1}^m e^{-c_0(n_j \wedge p)}.$$

In summary, we have

$$\mathbb{E} \|\widehat{U} \widehat{U}^\top - U U^\top\|_{\mathbb{F}}^2 \leq \frac{4r}{\sum_{j=1}^m u_j^{-2}} + 28r \sum_{j=1}^m e^{-c_0(n_j \wedge p)}.$$

High probability upper bound. Recall the event $\mathcal{E}_1^{(j)}, \mathcal{E}_2^{(j)}$ defined respectively in (24) and (25). Moreover, we define

$$\mathcal{E}_3^{(j)} := \left\{ \|\bar{H}^{(j)}\| \lesssim \sqrt{n_j \vee p} \right\} \cap \left\{ \|\bar{G}^{(j)}\| \lesssim \sqrt{p} \right\}.$$

Then $\mathbb{P}(\mathcal{E}_3^{(j)}) \geq 1 - 2e^{-p} - 2e^{-n_j \vee p}$. We denote $\mathcal{F}^{(j)} = \mathcal{E}_1^{(j)} \cap \mathcal{E}_2^{(j)} \cap \mathcal{E}_3^{(j)}$ and $\mathcal{F} = \cap_{j=1}^m \mathcal{F}^{(j)}$.

We define the function $\phi(s; t_0)$ for given $t_0 > 0$ as

$$\phi(s; t_0) = \begin{cases} 1, & s \leq t_0 \\ 2 - \frac{s}{t_0}, & t_0 < s \leq 2t_0 \\ 0, & s \geq 2t_0. \end{cases} \quad (30)$$

Also define

$$\begin{aligned} \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) &:= \phi(\|\bar{H}^{(j)}\|; \sqrt{n_j \vee p}) \cdot \phi(\|\bar{H}^{(j)} \bar{H}^{(j)\top} - n_j I\|; \sqrt{pn_j}) \cdot \phi(\|\bar{G}^{(j)} \bar{H}^{(j)\top}\|; \sqrt{n_j p}) \\ &\quad \cdot \mathbf{1}(\|\bar{G}^{(j)}\| \leq \sqrt{n_j}) \cdot \mathbf{1}(\|\bar{G}^{(j)} \bar{G}^{(j)\top} - n_j I\| \leq \sqrt{(r + \eta_j)n_j}). \end{aligned}$$

Then we have

$$\|U^\top \Xi_j U\|_{\mathbb{F}} \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) \leq \frac{\sqrt{r + \eta_j}}{\sqrt{n_j}} (\lambda_{\max} + \sigma^2).$$

In the following, we shall condition on $\bar{G}^{(j)}$. In order to compute the Lipschitz constant, we denote

$$\Xi'_j = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} \frac{1}{n_j} G^{(j)} G^{(j)\top} - (\Lambda + \sigma^2 I) & \frac{1}{n_j} G^{(j)} H^{(j)\top} \\ \frac{1}{n_j} H^{(j)'} G^{(j)\top} & \frac{1}{n_j} H^{(j)'} H^{(j)\top} - \sigma^2 I \end{bmatrix} \begin{bmatrix} U^\top \\ U_\perp^\top \end{bmatrix}$$

Then we have

$$\begin{aligned} & \left| \|U_\perp^\top \Xi_j U\|_{\mathbb{F}} \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) - \|U_\perp^\top \Xi'_j U\|_{\mathbb{F}} \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)'}) \right| \\ & \leq \frac{1}{\sqrt{n_j}} \cdot \sigma(\lambda_{\max}^{1/2} + \sigma) \cdot \|\bar{H}^{(j)} - \bar{H}^{(j)'}\|_{\mathbb{F}}. \end{aligned}$$

And

$$\begin{aligned} & \left| \|U_\perp^\top \Xi_j U_\perp\|_{\mathbb{F}} \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) - \|U_\perp^\top \Xi'_j U_\perp\|_{\mathbb{F}} \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)'}) \right| \\ & \leq \frac{2\sigma^2}{n_j} \cdot \sqrt{n_j \vee p} \cdot \|\bar{H}^{(j)} - \bar{H}^{(j)'}\|_{\mathbb{F}}. \end{aligned}$$

Under the given SNR condition, we have

$$\frac{1}{\sqrt{n_j}} \cdot \sigma(\lambda_{\max}^{1/2} + \sigma) \geq \frac{2\sigma^2}{n_j} \cdot \sqrt{n_j \vee p}.$$

Next we analyze $U^\top D_j U, U_\perp^\top D_j U$ and $U_\perp^\top D_j U_\perp$. Recall $D_j = \tilde{U}_j \tilde{U}_j^\top - U U^\top + Z_j$, and

$$\begin{aligned} & \tilde{U}_j \tilde{U}_j^\top - U U^\top \\ &= \sum_{l \geq 1} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \underline{M(s_1)} \Lambda^{-s_1} \underline{M(s_1)^\top \Xi_j M(s_2)} \Lambda^{-s_2} \dots \Lambda^{-s_l} \underline{M(s_l)^\top \Xi_j M(s_{l+1})} \Lambda^{-s_{l+1}} \underline{M(s_{l+1})}^\top, \end{aligned}$$

Now for each $l, \mathbf{s} \in \mathbb{S}_l$, we have

$$\begin{aligned} & \left\| \underline{M(s_1)} \Lambda^{-s_1} \underline{M(s_1)^\top \Xi_j M(s_2)} \Lambda^{-s_2} \dots \Lambda^{-s_l} \underline{M(s_l)^\top \Xi_j M(s_{l+1})} \Lambda^{-s_{l+1}} \underline{M(s_{l+1})}^\top \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) \right. \\ & \quad \left. - \underline{M(s_1)} \Lambda^{-s_1} \underline{M(s_1)^\top \Xi'_j M(s_2)} \Lambda^{-s_2} \dots \Lambda^{-s_l} \underline{M(s_l)^\top \Xi'_j M(s_{l+1})} \Lambda^{-s_{l+1}} \underline{M(s_{l+1})}^\top \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)'}) \right\|_{\mathbb{F}} \\ & \leq l \cdot \frac{1}{8^{l-1}} \cdot \lambda_{\min}^{-1} \cdot \sqrt{\frac{1}{n_j}} \sigma(\lambda_{\max}^{1/2} + \sigma) \cdot \|\bar{H}^{(j)} - \bar{H}^{(j)'}\|_{\mathbb{F}} \end{aligned}$$

Therefore we conclude

$$\begin{aligned} & \|U^\top D_j U \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) - U^\top D'_j U \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)'})\|_{\mathbb{F}} \\ & \quad \lesssim \lambda_{\min}^{-1} \cdot \sqrt{\frac{1}{n_j}} \sigma(\lambda_{\max}^{1/2} + \sigma) \cdot \|\bar{H}^{(j)} - \bar{H}^{(j)'}\|_{\mathbb{F}} + \alpha_j \|\bar{Z}_{j,1} - \bar{Z}'_{j,1}\|_{\mathbb{F}}, \\ & \|U_\perp^\top D_j U \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) - U_\perp^\top D'_j U \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)'})\|_{\mathbb{F}} \\ & \quad \lesssim \lambda_{\min}^{-1} \cdot \sqrt{\frac{1}{n_j}} \sigma(\lambda_{\max}^{1/2} + \sigma) \cdot \|\bar{H}^{(j)} - \bar{H}^{(j)'}\|_{\mathbb{F}} + \alpha_j \|\bar{Z}_{j,2} - \bar{Z}'_{j,2}\|_{\mathbb{F}}, \\ & \|U_\perp^\top D_j U_\perp \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) - U_\perp^\top D'_j U_\perp \cdot \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)'})\|_{\mathbb{F}} \\ & \quad \lesssim \lambda_{\min}^{-1} \cdot \sqrt{\frac{1}{n_j}} \sigma(\lambda_{\max}^{1/2} + \sigma) \cdot \|\bar{H}^{(j)} - \bar{H}^{(j)'}\|_{\mathbb{F}} + \alpha_j \|\bar{Z}_{j,3} - \bar{Z}'_{j,3}\|_{\mathbb{F}}. \end{aligned}$$

Recall

$$\Delta_j = \sum_{l \geq 1} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \underline{M(s_1)} \cdot \underline{M(s_1)^\top D_j M(s_2)} \dots \underline{M(s_l)^\top D_j M(s_{l+1})} \cdot \underline{M(s_{l+1})}^\top.$$

We also define

$$\tilde{\psi}_j(\bar{G}^{(j)}, \bar{H}^{(j)}, \bar{Z}_{j,1}, \bar{Z}_{j,2}, \bar{Z}_{j,3}) := \psi_j(\bar{G}^{(j)}, \bar{H}^{(j)}) \cdot \phi(\|\bar{Z}_{j,1}\|; \sqrt{r + \eta_j}) \cdot \phi(\|\bar{Z}_{j,2}\|; \sqrt{p}) \cdot \phi(\|\bar{Z}_{j,3}\|; \sqrt{p}).$$

And we have conditioning on $\bar{G}^{(j)}$, for arbitrary given $\|M\|_F \leq 1$, the function $\text{tr}(U_\perp^\top \Delta_j U M) \cdot \tilde{\psi}_j(\bar{G}^{(j)}, \bar{H}^{(j)}, \bar{Z}_{j,1}, \bar{Z}_{j,2}, \bar{Z}_{j,3})$ is

$$\lambda_{\min}^{-1} \cdot \sqrt{\frac{1}{n_j} \sigma (\lambda_{\max}^{1/2} + \sigma)} + \alpha_j$$

Lipschitz. Using Gaussian concentration theorem, this indicates

$$\|\text{vec}(U_\perp^\top \Delta_j U) \cdot \tilde{\psi}_j(\bar{G}^{(j)}, \bar{H}^{(j)}, \bar{Z}_{j,1}, \bar{Z}_{j,2}, \bar{Z}_{j,3})\|_{\psi_2} \leq \underbrace{C \lambda_{\min}^{-1} \cdot \sqrt{\frac{1}{n_j} \sigma (\lambda_{\max}^{1/2} + \sigma)} + C \alpha_j}_{=: L_j}. \quad (31)$$

For notation simplicity, we collect $\mathbf{g}_j := [\text{vec}(\bar{G}^{(j)})^\top, \text{vec}(\bar{H}^{(j)})^\top, \text{vec}(\bar{Z}_{j,1})^\top, \text{vec}(\bar{Z}_{j,2})^\top, \text{vec}(\bar{Z}_{j,3})^\top]^\top$. Then $\mathbf{g}_j \sim N(0, I)$. We shall define two matrix-valued functions f_1, f_2 as

$$f_1(\mathbf{g}_j) = U^\top \Delta_j U_\perp \cdot \tilde{\psi}_j(\mathbf{g}_j),$$

$$f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) = \underline{U_\perp^\top \Delta_{j_2} M(s_3)} \cdots \underline{M(s_{l-1})^\top \Delta_{j_{l-1}} U_\perp} \cdot \prod_{u=2}^{l-1} \tilde{\psi}_{j_u}(\mathbf{g}_{j_u}).$$

Then it boils down to estimating

$$\sum_{j \in \mathbb{I}_1} w_{j_1} \cdots w_{j_l} \cdot \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top)$$

We define a projection map: $\pi_2^{l-1} : (j_1, \dots, j_l) \mapsto (j_2, \dots, j_{l-1})$. And we denote $\pi_2^{l-1}(\mathbb{I}_1)$ the image of π_2^{l-1} applied to \mathbb{I}_1 . And then we can rewrite $\sum_{j \in \mathbb{I}_1}$ as

$$\sum_{j \in \mathbb{I}_1} = \sum_{(j_2, \dots, j_{l-1}) \in \pi_2^{l-1}(\mathbb{I}_1)} \sum_{\substack{j_1 \neq j_l \\ \{j_1, j_l\} \cap \{j_2, \dots, j_{l-1}\} = \emptyset}}.$$

Next, we shall fix (j_2, \dots, j_{l-1}) , and use the decoupling to derive the upper bound. For given (j_2, \dots, j_{l-1}) , condition on $\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}$, using the decoupling technique (e.g. [De la Pena and Giné \(2012\)](#)), we have

$$\mathbb{P}\left(\left| \sum_{\substack{j_1 \neq j_l \\ \{j_1, j_l\} \cap \{j_2, \dots, j_{l-1}\} = \emptyset}} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right| \geq t \middle| \mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}\right)$$

$$\leq C \mathbb{P}\left(C \left| \sum_{\substack{j_1 \neq j_l \\ \{j_1, j_l\} \cap \{j_2, \dots, j_{l-1}\} = \emptyset}} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}'_{j_l})^\top) \right| \geq t \middle| \mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}\right),$$

for some absolute constant $C > 0$, where \mathbf{g}'_j is an i.i.d. copy of \mathbf{g}_j . Notice

$$\begin{aligned} & \sum_{\substack{j_1 \neq j_l \\ \{j_1, j_l\} \cap \{j_2, \dots, j_{l-1}\} = \emptyset}} w_{j_1} \cdots w_{j_l} \cdot \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}'_{j_l})^\top) \\ = & \text{tr} \left(\left(\sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} f_1(\mathbf{g}_{j_1}) \right) \cdot w_{j_2} \cdots w_{j_{l-1}} f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot \left(\sum_{j_l \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_l} f_1(\mathbf{g}'_{j_l}) \right)^\top \right) \\ & - \sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1}^2 w_{j_2} \cdots w_{j_{l-1}} \text{tr} \left(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}'_{j_1})^\top \right). \end{aligned}$$

For the first term above, due to (31), we have

$$\begin{aligned} \left\| \sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} \text{vec}(f_1(\mathbf{g}_{j_1})) \right\|_{\psi_2}^2 &= \left\| \sum_{j_l \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_l} \text{vec}(f_1(\mathbf{g}'_{j_l})) \right\|_{\psi_2}^2 \\ &\leq \sum_{j \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_j^2 L_j^2 \leq \sum_{j \in [m]} w_j^2 L_j^2 \end{aligned}$$

Therefore we have

$$\begin{aligned} & \left| \text{tr} \left(\left(\sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} f_1(\mathbf{g}_{j_1}) \right) \cdot w_{j_2} \cdots w_{j_{l-1}} f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot \left(\sum_{j_l \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_l} f_1(\mathbf{g}'_{j_l}) \right)^\top \right) \right| \left\| \{\mathbf{g}_j\}_{j=1}^m \right\| \\ & \leq \left\| \left(\sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} f_1(\mathbf{g}_{j_1}) \right) \cdot w_{j_2} \cdots w_{j_{l-1}} f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \right\|_{\text{F}} \cdot \left(\sum_j w_j^2 L_j^2 \right)^{1/2} \cdot s_{l,1} \end{aligned}$$

holds with probability exceeding $1 - e^{-s_{l,1}^2}$. Using Lemma 6, we have

$$\begin{aligned} & \left\| \left(\sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} f_1(\mathbf{g}_{j_1}) \right) \cdot w_{j_2} \cdots w_{j_{l-1}} f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \right\|_{\text{F}} \\ & \leq w_{j_2} \cdots w_{j_{l-1}} \left\| f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \right\| \cdot \left\| \sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} f_1(\mathbf{g}_{j_1}) \right\|_{\text{F}} \\ & \leq w_{j_2} u_{j_2} \cdots w_{j_{l-1}} u_{j_{l-1}} \cdot \left(\sum_j w_j^2 L_j^2 \right)^{1/2} \cdot (s_{l,2} + \sqrt{pr}), \end{aligned}$$

with probability exceeding $1 - e^{-s_{l,2}^2}$. Taking union bound, we conclude

$$\begin{aligned} & \left| \text{tr} \left(\left(\sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_1} f_1(\mathbf{g}_{j_1}) \right) \cdot w_{j_2} \cdots w_{j_{l-1}} f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot \left(\sum_{j_l \in [m] \setminus \{j_2, \dots, j_{l-1}\}} w_{j_l} f_1(\mathbf{g}'_{j_l}) \right)^\top \right) \right| \\ & \leq w_{j_2} u_{j_2} \cdots w_{j_{l-1}} u_{j_{l-1}} \sum_j w_j^2 L_j^2 \cdot s_{l,1} (s_{l,2} + \sqrt{pr}) \end{aligned}$$

with probability exceeding $1 - M^{l-2}(e^{-s_{l,1}^2} + e^{-s_{l,2}^2})$. This leads to

$$\begin{aligned} & \left| \sum_{\substack{j_1 \neq j_l \\ \{j_1, j_l\} \cap \{j_2, \dots, j_{l-1}\} = \emptyset}} w_{j_1} \cdots w_{j_l} \cdot \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}'_{j_l})^\top) \right| \\ & \leq w_{j_2} u_{j_2} \cdots w_{j_{l-1}} u_{j_{l-1}} \cdot \left(\sum_j w_j^2 L_j^2 \right)^{1/2} \cdot (s_{l,2} + \sqrt{pr}) + r \cdot \sum_{j_1 \in [m] \setminus \{j_2, \dots, j_{l-1}\}} (w_{j_1} u_{j_1})^2 w_{j_2} u_{j_2} \cdots w_{j_{l-1}} u_{j_{l-1}}. \end{aligned}$$

In conclusion, we have

$$\begin{aligned} & \left| \sum_{\mathbf{j} \in \mathbb{I}_1} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right| \\ & \leq \left(\sum_{j=1}^m w_j u_j \right)^{l-2} \cdot \sum_j w_j^2 L_j^2 \cdot s_1 (s_2 + \sqrt{pr}) + r \cdot \sum_{\mathbf{j} \in \mathbb{I}_2} w_{j_1} u_{j_1} \cdots w_{j_l} u_{j_l} \\ & \leq \frac{1}{8^{l-2}} \sum_j w_j^2 L_j^2 \cdot s_{l,1} (s_{l,2} + \sqrt{pr}) + r \cdot \sum_{\mathbf{j} \in \mathbb{I}_2} w_{j_1} u_{j_1} \cdots w_{j_l} u_{j_l}. \end{aligned}$$

Now we sum up over all $\mathbf{s} \in \mathbb{S}_l$ and $l \geq 2$, and we set $s_{l,1} = s_{l,2} = \max\{\tau_0, \sqrt{2l \cdot \log(4m)}\}$ for some $\tau_0 \leq \sqrt{pr}$ to be chosen later, and we get with failure probability

$$\begin{aligned} & \sum_{l \geq 2} 2 \cdot 4^l m^{l-2} \cdot \exp(-\max\{\tau_0^2, l \log m\}) \\ & \leq \sum_{l=2}^{\lceil \tau_0^2 / 2 \log(4m) \rceil} (4m)^l \cdot e^{-\tau_0^2} + \sum_{l \geq \lceil \tau_0^2 / 2 \log(4m) \rceil + 1} (4m)^{-l} \\ & \leq 4e^{-\tau_0^2/2}, \end{aligned}$$

the following holds:

$$\begin{aligned} & \left| \sum_{\mathbf{j} \in \mathbb{I}_1} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right| \\ & \leq \frac{1}{8^{l-2}} \sum_j w_j^2 L_j^2 \cdot \max\{\tau_0, \sqrt{2l \cdot \log(4m)}\} (\max\{\tau_0, \sqrt{2l \cdot \log(4m)}\} + \sqrt{pr}) \\ & \quad + r \cdot \sum_{\mathbf{j} \in \mathbb{I}_2} w_{j_1} u_{j_1} \cdots w_{j_l} u_{j_l} \\ & \leq \frac{1}{8^{l-2}} \sum_j w_j^2 L_j^2 \cdot \max\{\tau_0, \sqrt{2l \cdot \log(4m)}\} (\max\{\tau_0, \sqrt{2l \cdot \log(4m)}\} + \sqrt{pr}) \\ & \quad + 2lr \cdot \left(\sum_{k=1}^m w_k^2 u_k^2 \right) \left(\sum_{k=1}^m w_k u_k \right)^{l-2}, \end{aligned}$$

where the last line is due to (28). Using this, we have

$$\begin{aligned}
& \left| \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \sum_{\mathbf{j} \in [m]^l} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right| \\
& \leq \left| \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \sum_{\mathbf{j} \in \mathbb{I}_1(\mathbf{s})} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right| \\
& \quad + \left| \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \sum_{\mathbf{j} \in \mathbb{I}_2(\mathbf{s})} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right|.
\end{aligned}$$

The first term above can be bounded by

$$\begin{aligned}
& \sum_j w_j^2 L_j^2 \cdot \left(\sum_{l=2}^{\lceil \tau_0^2 / 2 \log(4m) \rceil} \tau_0 \sqrt{pr} \cdot 2^{-l} + \sum_{l \geq \lceil \tau_0^2 / 2 \log(4m) \rceil + 1} 2^{-l} (2l \log(4m) + \sqrt{2lpr \log(4m)}) \right) \\
& \quad + \sum_{l \geq 2} 4^l \cdot 2lr \left(\sum_{k=1}^m w_k^2 u_k^2 \right) \left(\sum_{k=1}^m w_k u_k \right)^{l-2} \\
& \leq C \sum_j w_j^2 L_j^2 \left(\tau_0 \sqrt{pr} + \log(4m) + \sqrt{pr \log(4m)} \right) + Cr \left(\sum_{k=1}^m w_k^2 u_k^2 \right) \\
& \leq C \tau_0 \sqrt{pr} \sum_j w_j^2 L_j^2 + Cr \left(\sum_{k=1}^m w_k^2 u_k^2 \right),
\end{aligned}$$

if we set $\tau_0 = c_0 \sqrt{p} \geq \log(4m)$. Here the first inequality is due to (29).

In summary,

$$\begin{aligned}
& \left| \sum_{l \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_l} (-1)^{\|\mathbf{s}\|_{\ell_0} + 1} \sum_{\mathbf{j} \in [m]^l} w_{j_1} \cdots w_{j_l} \text{tr}(f_1(\mathbf{g}_{j_1}) \cdot f_2(\mathbf{g}_{j_2}, \dots, \mathbf{g}_{j_{l-1}}) \cdot f_1(\mathbf{g}_{j_l})^\top) \right| \\
& \leq C \tau_0 \sqrt{pr} \sum_{j=1}^m w_j^2 L_j^2 + Cr \left(\sum_{j=1}^m w_j^2 u_j^2 \right).
\end{aligned}$$

with probability $1 - 22 \sum_{j=1}^m e^{-c_0(n_j \wedge p)}$. By setting $w_j = \frac{u_j^{-2}}{\sum_j u_j^{-2}}$, we obtain

$$\begin{aligned}
\|\widehat{U}\widehat{U}^\top - UU^\top\|_{\text{F}}^2 & \leq Cr \left(\sum_j u_j^{-2} \right)^{-1} = \frac{Cpr}{\sum_j (\lambda_{\min}^{-1} (\lambda_{\max}^{1/2} + \sigma) \sigma \sqrt{\frac{1}{n_j} + \alpha_j})^{-2}} \\
& \leq \frac{Cpr}{\sum_j \left(n_j \wedge (n_j^2 \varepsilon_j^2 \log^{-1}(\frac{2.5}{\delta_j}) p^{-1} (r + \log n_j)^{-1}) \right)} \cdot \frac{\sigma^2}{\lambda_{\min}} (\lambda_{\max} / \lambda_{\min} + \frac{\sigma^2}{\lambda_{\min}}).
\end{aligned}$$

Finally since $\lambda_{\max} \asymp \lambda_{\min} \asymp \lambda$, we finish the proof.

A.2 Proof of Theorem 2

We have

$$\begin{aligned}\widehat{\Sigma} - \Sigma &= \sum_j v_j (\widehat{U}\widehat{U}^\top (\widehat{\Sigma}_j - \sigma^2 I) \widehat{U}\widehat{U}^\top + \widehat{U}E_j\widehat{U}^\top) - UU^\top (\Sigma - \sigma^2 I) UU^\top \\ &= \sum_j v_j \widehat{U}\widehat{U}^\top (\widehat{\Sigma}_j - \sigma^2 I) \widehat{U}\widehat{U}^\top - UU^\top (\Sigma - \sigma^2 I) UU^\top + \sum_j v_j \widehat{U}E_j\widehat{U}^\top.\end{aligned}\quad (32)$$

For the first term in (32), we can further decompose it as

$$\begin{aligned}&\sum_j v_j \widehat{U}\widehat{U}^\top (\widehat{\Sigma}_j - \sigma^2 I) \widehat{U}\widehat{U}^\top - UU^\top (\Sigma - \sigma^2 I) UU^\top \\ &= (\widehat{U}\widehat{U}^\top - UU^\top) \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) UU^\top + \widehat{U}\widehat{U}^\top \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U}\widehat{U}^\top - UU^\top) \\ &\quad + UU^\top (\sum_j v_j \widehat{\Sigma}_j - \Sigma) UU^\top.\end{aligned}\quad (33)$$

And notice from Lemma 5, for each j , with probability exceeding $1 - e^{-t_{1,j}}$,

$$\|\widehat{\Sigma}_j - \Sigma\| \lesssim \left(\sqrt{\frac{\tilde{r} + t_{1,j}}{n_j}} \vee \frac{\tilde{r} + t_{1,j}}{n_j} \right) (\lambda + \sigma^2),$$

where $\tilde{r} = \frac{r\lambda + p\sigma^2}{\lambda + \sigma^2}$. Under the given SNR, and by setting $t_{1,j} = p \wedge n_j$, we obtain $\|\widehat{\Sigma}_j - \Sigma\| \lesssim \lambda$, and therefore with probability exceeding $1 - e^{-p \wedge n_j}$,

$$\|\widehat{\Sigma}_j - \sigma^2 I\| \leq \|\widehat{\Sigma}_j - \Sigma\| + \|\Sigma - \sigma^2 I\| \lesssim \lambda.$$

Moreover, we have

$$U^\top (\sum_j v_j \widehat{\Sigma}_j - \Sigma) U = \sum_{j=1}^m \sum_{i=1}^{n_j} \frac{v_j}{n_j} U^\top X_i^{(j)} X_i^{(j)\top} U - (\Lambda + \sigma^2 I).$$

Now applying Lemma 5, we obtain with probability exceeding $1 - e^{-t_2}$,

$$\|U^\top (\sum_j v_j \widehat{\Sigma}_j - \Sigma) U\| \lesssim (\sigma^2 + \lambda) \sqrt{\sum_j \frac{v_j^2}{n_j}} \sqrt{r + t_2}.$$

In summary, we obtain the upper bound as follows:

$$\begin{aligned}
& \left\| \sum_j v_j \widehat{U} \widehat{U}^\top (\widehat{\Sigma}_j - \sigma^2 I) \widehat{U} \widehat{U}^\top - U U^\top (\Sigma - \sigma^2 I) U U^\top \right\|_{\mathbb{F}} \\
& \lesssim \lambda \cdot \|\widehat{U} \widehat{U}^\top - U U^\top\|_{\mathbb{F}} + (\sigma^2 + \lambda) \sqrt{\sum_j \frac{v_j^2}{n_j} \sqrt{(r + t_2)r}} \\
& \lesssim \left(\sqrt{\frac{pr}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 p^{-1} (r + \log n_j)^{-1} \log^{-1}(2.5 \delta_j^{-1})) \right)}} \cdot \sqrt{(\sigma^2 + \lambda)\sigma^2} \right) \wedge (\sqrt{2r}\lambda) \\
& \quad + (\sigma^2 + \lambda) \sqrt{\sum_j \frac{v_j^2}{n_j} \sqrt{(r + t_2)r}},
\end{aligned}$$

where the last inequality comes from Theorem 1. For the second term in (32), $[\sum_j v_j E_j]_{kl} = [\sum_j v_j E_j]_{kl} \sim N(0, \sum_j v_j^2 \beta_j^2)$, and $[\sum_j v_j E_j]_{kl} \sim N(0, 2 \sum_j v_j^2 \beta_j^2)$. And with probability exceeding $1 - e^{-t_2}$,

$$\left\| \sum_j v_j E_j \right\|_{\mathbb{F}} \lesssim \sqrt{r} \cdot \sqrt{r + t_2} \sqrt{\sum_j v_j^2 \beta_j^2}.$$

So we conclude with probability exceeding $1 - 23 \sum_j e^{-(n_j \wedge p)} - e^{-t_2}$,

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 & \lesssim \left(\frac{pr}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 d^{-1} (r + \log n_j)^{-1} \log^{-1}(2.5 \delta_j^{-1})) \right)} \cdot (\sigma^2 + \lambda)\sigma^2 \right) \wedge (2r\lambda^2) \\
& \quad + (r + t_2)r \cdot \sum_j v_j^2 \left(\frac{1}{n_j} (\lambda^2 + \sigma^4) + \beta_j^2 \right).
\end{aligned}$$

Next we consider the expectation for $\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2$. From (32) and (33), we see

$$\begin{aligned}
\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 & \lesssim \mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^2 + \mathbb{E} \left\| \sum_j v_j E_j \right\|_{\mathbb{F}}^2 \\
& \quad + \mathbb{E} \|U^\top (\sum_j v_j \widehat{\Sigma}_j - \Sigma) U\|_{\mathbb{F}}^2.
\end{aligned} \tag{34}$$

We consider the event

$$\mathcal{F}_1 = \bigcap_{j=1}^m \left\{ \|\widehat{\Sigma}_j - \sigma^2 I\| \leq C(\lambda + \sigma^2) \left(\sqrt{\frac{\widetilde{r} + t_{1,j}}{n_j}} \vee \frac{\widetilde{r} + t_{1,j}}{n_j} \right) + \lambda \right\}.$$

By setting $t_{1,j} = n_j \wedge p$, then under \mathcal{F}_1 , we have $\|\widehat{\Sigma}_j - \sigma^2 I\| \leq 2\lambda$ under the given SNR condition,

and $\mathbb{P}(\mathcal{F}_1^c) \leq \sum_j e^{-(n_j \wedge p)}$. Then

$$\begin{aligned} & \mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^2 \\ &= \mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^2 \cdot \mathbf{1}(\mathcal{F}_1) + \mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^2 \cdot \mathbf{1}(\mathcal{F}_1^c) \\ &\leq 4\lambda^2 \mathbb{E} \left\| \widehat{U} \widehat{U}^\top - U U^\top \right\|_{\mathbb{F}}^2 + \left(\mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^4 \right)^{1/2} \cdot (\mathbb{P}(\mathcal{F}_1^c))^{1/2}. \end{aligned}$$

Notice

$$\begin{aligned} \mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^4 &\leq 4r^2 \cdot \mathbb{E} \left\| \sum_j v_j \widehat{\Sigma}_j - \sigma^2 I \right\|^4 \\ &\leq 4r^2 \cdot m^3 \sum_j v_j^4 \mathbb{E} \left\| \widehat{\Sigma}_j - \sigma^2 I \right\|^4. \end{aligned}$$

And

$$\mathbb{E} \left\| \widehat{\Sigma}_j - \sigma^2 I \right\|^4 \leq 8 \mathbb{E} \left\| \widehat{\Sigma}_j - \Sigma \right\|^4 + 8\lambda^4.$$

From Lemma 5, we see

$$\mathbb{P} \left(\left\| \widehat{\Sigma}_j - \Sigma \right\| \leq C(\lambda + \sigma^2) \left(\sqrt{\frac{\widetilde{r} + t}{n_j}} \vee \frac{\widetilde{r} + t}{n_j} \right) \right) \leq e^{-t}.$$

Then from Lemma 7, we see

$$\mathbb{E} \left\| \widehat{\Sigma}_j - \Sigma \right\|^4 \leq C(\lambda + \sigma^2)^4 \frac{\widetilde{r}^2}{n_j^2} \leq \lambda^4,$$

where the last inequality comes from the SNR condition. And thus

$$\mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^4 \leq 64r^2 m^3 \lambda^4$$

Therefore

$$\left(\mathbb{E} \left\| \sum_j v_j (\widehat{\Sigma}_j - \sigma^2 I) (\widehat{U} \widehat{U}^\top - U U^\top) \right\|_{\mathbb{F}}^4 \right)^{1/2} \cdot (\mathbb{P}(\mathcal{F}_1^c))^{1/2} \leq 8rm^{3/2} \lambda^2 \cdot \left(\sum_j e^{-p \wedge n_j} \right)^{1/2}.$$

This term is dominated by the first term as long as $\lambda/\sigma^2 \lesssim \frac{1}{m^{3/2}(\sum_j n_j)(\sum_j e^{-(n_j \wedge p)})}$. For the second term in (34), we have

$$\mathbb{E} \left\| \sum_j v_j E_j \right\|_{\mathbb{F}}^2 = r^2 \sum_j v_j^2 \beta_j^2.$$

We now consider the last term in (34). From Lemma 5, we have

$$\mathbb{P}\left(\|U^\top\left(\sum_j v_j \widehat{\Sigma}_j - \Sigma\right)U\| \geq C(\sigma^2 + \lambda) \sqrt{\sum_j \frac{v_j^2}{n_j} \sqrt{r+t}}\right) \leq e^{-t}.$$

Which gives

$$\mathbb{E}\|U^\top\left(\sum_j v_j \widehat{\Sigma}_j - \Sigma\right)U\|_{\mathbb{F}}^2 \leq r\mathbb{E}\|U^\top\left(\sum_j v_j \widehat{\Sigma}_j - \Sigma\right)U\|^2 \leq Cr^2(\lambda + \sigma)^2 \sum_j \frac{v_j^2}{n_j}.$$

In conclusion, we have

$$\begin{aligned} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 &\leq \left(\frac{Cpr}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 p^{-1} (r + \log n_j)^{-1} \log^{-1}(2.5\delta_j^{-1})) \right)} (\sigma^2 \lambda + \sigma^4) \right) \wedge (2r\lambda^2) \\ &\quad + Cr^2 \sum_j v_j^2 \left(\frac{\lambda^2 + \sigma^4}{n_j} + \frac{8}{\varepsilon_j^2} \log\left(\frac{2.5}{\delta_j}\right) \frac{\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2}{n_j^2} \right) \end{aligned}$$

Now by setting $v_j \propto \left(\frac{\lambda^2 + \sigma^4}{n_j} + \frac{8}{\varepsilon_j^2} \log\left(\frac{2.5}{\delta_j}\right) \frac{\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2}{n_j^2} \right)^{-1}$, we obtain

$$\begin{aligned} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 &\leq \left(\frac{Cpr}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 p^{-1} (r + \log n_j)^{-1} \log^{-1}(2.5\delta_j^{-1})) \right)} (\sigma^2 \lambda + \sigma^4) \right) \wedge (2r\lambda^2) \\ &\quad + \frac{Cr^2}{\sum_{j=1}^m \left((n_j (\lambda^2 + \sigma^4)^{-1}) \wedge (n_j^2 \varepsilon_j^2 \log^{-1}\left(\frac{2.5}{\delta_j}\right) (\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2)^{-1}) \right)}. \end{aligned}$$

Finally we show

$$\begin{aligned} &\frac{Cr^2}{\sum_{j=1}^m \left((n_j (\lambda^2 + \sigma^4)^{-1}) \wedge (n_j^2 \varepsilon_j^2 \log^{-1}\left(\frac{2.5}{\delta_j}\right) (\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2)^{-1}) \right)} \\ &\leq \frac{Cpr}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 p^{-1} (r + \log n_j)^{-1} \log^{-1}(2.5\delta_j^{-1})) \right)} (\sigma^2 \lambda + \sigma^4) \\ &\quad + \frac{Cr^2}{\sum_{j=1}^m \left((n_j \lambda^{-2}) \wedge (n_j^2 \varepsilon_j^2 \log^{-1}\left(\frac{2.5}{\delta_j}\right) \lambda^{-2} (r + \log n_j)^{-2}) \right)}. \end{aligned}$$

We consider the different cases for λ/σ^2 . When $\lambda/\sigma^2 \leq 1$, the left hand side is bounded by

$$\frac{Cr^2 \sigma^4}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 \log^{-1}\left(\frac{2.5}{\delta_j}\right)) \right)},$$

which is bounded by the first term on the right hand side. Next if $\lambda/\sigma^2 \geq \frac{p}{r}$, we have the left hand side is bounded by

$$\frac{Cr^2\lambda^2}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 \log^{-1} \left(\frac{2.5}{\delta_j} \right) (r + \log n_j)^{-2}) \right)},$$

which is bounded by the second term on the right hand side. Finally we consider if $1 \leq \lambda/\sigma^2 \leq \frac{p}{r}$. Then the left hand side is bounded by

$$\frac{Cr^2}{\sum_{j=1}^m \left((n_j \lambda^{-2}) \wedge (n_j^2 \varepsilon_j^2 \log^{-1} \left(\frac{2.5}{\delta_j} \right) (\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2)^{-1}) \right)}$$

Notice the first term on the right hand side is lower bounded by

$$\frac{Cpr\sigma^2\lambda}{\sum_{j=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 p^{-1} (r + \log n_j)^{-1} \log^{-1} (2.5 \delta_j^{-1})) \right)}.$$

Therefore it is equivalent to showing

$$\begin{aligned} & \frac{Cr^2}{\sum_{j=1}^m \left(\frac{\lambda^2}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2} \log \left(\frac{2.5}{\delta_j} \right) (\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2) \right)^{-1}} \\ & \leq \frac{Cpr\sigma^2\lambda}{\sum_{j=1}^M \left(\frac{1}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2} p (r + \log n_j) \log (2.5 \delta_j^{-1}) \right)^{-1}}, \end{aligned}$$

which is true if and only if

$$r \sum_{j=1}^m \left(\frac{1}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2} p (r + \log n_j) \log (2.5 \delta_j^{-1}) \right)^{-1} \leq p \sigma^2 \lambda \sum_{j=1}^m \left(\frac{\lambda^2}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2} \log \left(\frac{2.5}{\delta_j} \right) (\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2) \right)^{-1}.$$

This can be implied by

$$r \left(\frac{\lambda^2}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2} \log \left(\frac{2.5}{\delta_j} \right) (\lambda^2 (r + \log n_j)^2 + \sigma^4 p^2) \right) \leq d \sigma^2 \lambda \left(\frac{1}{n_j} + \frac{1}{n_j^2 \varepsilon_j^2} p (r + \log n_j) \log (2.5 \delta_j^{-1}) \right),$$

which is true if $1 \leq \frac{\lambda}{\sigma^2} \leq \frac{p}{r}$.

A.3 Proof of Lemma 1

We first state the following lemma, which will be helpful.

Lemma 2. *For any $j \in [m]$, suppose $\lambda/\sigma^2 \geq C_1(p/n_j + \sqrt{p/n_j})$, and $p \geq \log n_j$. Then with probability exceeding $1 - n_j^{-100} - 12e^{-c_0(p \wedge n_j)}$,*

$$\max_{i \in [n_j]} \|\tilde{U}_j \tilde{U}_j^\top - \tilde{U}_j^{(i)} \tilde{U}_j^{(i)\top}\|_{\mathbb{F}} \leq C \frac{1}{n_j} \sqrt{\frac{\lambda + \sigma^2}{\lambda}} \frac{\sigma^2}{\lambda} \sqrt{p(r + \log n_j)}.$$

Proof. Most of the proof is the same as the proof in Lemma 3 in Cai et al. (2024b), we only aim at improving the probability. Notice we have from Lemma 5, $\|\widehat{\Sigma}_j - \Sigma\|, \|\widehat{\Sigma}_j^{(i)} - \Sigma\| \leq c_1\lambda$ with probability exceeding $1 - 2e^{-c_0(n_j \wedge p)}$, where $\widehat{\Sigma}_j^{(i)} = \frac{1}{n_j} \sum_{i' \neq i} X_{i'}^{(j)} X_{i'}^{(j)\top} + \frac{1}{n_j} \widetilde{X}_i^{(j)} \widetilde{X}_i^{(j)\top}$, where $\widetilde{X}_i^{(j)}$ is an i.i.d. copy of $X_i^{(j)}$. Therefore we have

$$\begin{aligned}\widetilde{U}_j \widetilde{U}_j^\top - UU^\top &= \sum_{k \geq 1} \mathcal{S}_{\Sigma, k}(\Xi), \\ \widetilde{U}_j^{(i)} \widetilde{U}_j^{(i)\top} - UU^\top &= \sum_{k \geq 1} \mathcal{S}_{\Sigma, k}(\Xi^{(i)}),\end{aligned}$$

where $\Xi = \widehat{\Sigma}_j - \Sigma$, and $\Xi^{(i)} = \widehat{\Sigma}_j^{(i)} - \Sigma$. This implies

$$\widetilde{U}_j \widetilde{U}_j^\top - \widetilde{U}_j^{(i)} \widetilde{U}_j^{(i)\top} = \mathcal{S}_{\Sigma, 1}(\Xi) - \mathcal{S}_{\Sigma, 1}(\Xi^{(i)}) + \sum_{k \geq 2} (\mathcal{S}_{\Sigma, k}(\Xi) - \mathcal{S}_{\Sigma, k}(\Xi^{(i)})).$$

Notice

$$\begin{aligned}\mathcal{S}_{\Sigma, 1}(\Xi) - \mathcal{S}_{\Sigma, 1}(\Xi^{(i)}) &= U\Lambda^{-1}U^\top(\Xi - \Xi^{(i)})U_\perp U_\perp^\top + U_\perp U_\perp^\top(\Xi - \Xi^{(i)})U\Lambda^{-1}U^\top \\ &= \frac{1}{n_j} U\Lambda^{-1}U^\top (X_i^{(j)} X_i^{(j)\top} - \widetilde{X}_i^{(j)} \widetilde{X}_i^{(j)\top}) U_\perp U_\perp^\top \\ &\quad + \frac{1}{n_j} U_\perp U_\perp^\top (X_i^{(j)} X_i^{(j)\top} - \widetilde{X}_i^{(j)} \widetilde{X}_i^{(j)\top}) U\Lambda^{-1}U^\top.\end{aligned}$$

We consider the event

$$\begin{aligned}\mathcal{E}_1 &= \left\{ \|U^\top X_i^{(j)}\|_{\ell_2}, \|U^\top \widetilde{X}_i^{(j)}\|_{\ell_2} \lesssim \sqrt{\lambda + \sigma^2} \sqrt{r + \log n_j} : \forall i \in [n_j] \right\} \\ &\quad \cap \left\{ \|U_\perp^\top X_i^{(j)}\|_{\ell_2}, \|U_\perp^\top \widetilde{X}_i^{(j)}\|_{\ell_2} \lesssim \sigma\sqrt{p} : \forall i \in [n_j] \right\}.\end{aligned}$$

Then $\mathbb{P}(\mathcal{E}_1) \geq 1 - n_j^{-100}$. Therefore under \mathcal{E}_1 ,

$$\begin{aligned}\|\mathcal{S}_{\Sigma, 1}(\Xi) - \mathcal{S}_{\Sigma, 1}(\Xi^{(i)})\|_{\text{F}} &\leq \frac{2}{n_j} \|U\Lambda^{-1}U^\top X_i^{(j)} X_i^{(j)\top} U_\perp U_\perp^\top\|_{\text{F}} + \frac{2}{n_j} \|U\Lambda^{-1}U^\top \widetilde{X}_i^{(j)} \widetilde{X}_i^{(j)\top} U_\perp U_\perp^\top\|_{\text{F}} \\ &\lesssim \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j}.\end{aligned}$$

Now we consider for $k \geq 2$. We denote the index set

$$\mathbb{S}_k = \{\mathbf{s} = (s_1, \dots, s_{k+1}) : s_1, \dots, s_{k+1} \geq 0, s_1 + \dots + s_{k+1} = k\},$$

whose cardinality is bounded by $|\mathbb{S}_k| = \binom{2k}{k} \leq 4^k$. We define

$$\begin{aligned}\mathcal{T}_{\Sigma, k, \mathbf{s}, l} &= M(s_1) \Lambda^{-s_1} \underline{M(s_1)^\top \Xi^{(i)} M(s_2)} \cdots \underline{M(s_l)^\top (\Xi - \Xi^{(i)}) M(s_{l+1})} \\ &\quad \cdots \underline{M(s_k)^\top \Xi M(s_{k+1})} \Lambda^{-s_{k+1}} M(s_{k+1})^\top\end{aligned}$$

for $k \geq 2$, $\mathbf{s} \in \mathbb{S}_k$, and $l \in [k]$ and $M(0) := U_\perp$, $M(s) = U$ for $s > 0$. With slight abuse of notation, $\Lambda^{-0} = I_{p-r}$. Then we have

$$\sum_{k \geq 2} (\mathcal{S}_{\Sigma, k}(\Xi) - \mathcal{S}_{\Sigma, k}(\Xi^{(i)})) = \sum_{k \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_k} \sum_{l \in [k]} \mathcal{T}_{\Sigma, k, \mathbf{s}, l}. \quad (35)$$

We consider the event

$$\begin{aligned} \mathcal{E}_2 = & \left\{ \|U^\top \Xi U\|, \|U^\top \Xi^{(i)} U\| \lesssim (\lambda + \sigma^2) \sqrt{\frac{r + \eta}{n_j}} : \forall i \in [n_j] \right\} \\ & \cap \left\{ \|U_\perp^\top \Xi U\|, \|U_\perp^\top \Xi^{(i)} U\| \lesssim (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}} : \forall i \in [n_j] \right\} \\ & \cap \left\{ \|U_\perp^\top \Xi U_\perp\|, \|U_\perp^\top \Xi^{(i)} U_\perp\| \lesssim \sigma^2 \sqrt{\frac{p}{n_j}} : \forall i \in [n_j] \right\}. \end{aligned}$$

Then $\mathbb{P}(\mathcal{E}_2) \geq 1 - 4(e^{-c_0 p} + e^{-c_0 n_j}) - n_j e^{-\eta}$ for some $\eta > 0$ to be specified. Then as long as $\eta + r \leq p \wedge n_j$, and under the given SNR condition, we have

$$\lambda^{-1} \max \left\{ (\lambda + \sigma^2) \sqrt{\frac{r + \eta}{n_j}}, (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}}, \sigma^2 \sqrt{\frac{p}{n_j}} \right\} \leq \frac{1}{10}.$$

Now we bound $\|\mathcal{T}_{\Sigma, k, \mathbf{s}, l}\|_F$ under $\mathcal{E}_1 \cap \mathcal{E}_2$. We discuss different choices of s_l, s_{l+1} .

Case 1: $s_l, s_{l+1} > 0$. In this case, we have

$$M(s_l)^\top (\Xi - \Xi^{(i)}) M(s_{l+1}) = \frac{1}{n_j} U^\top (X_i^{(j)} X_i^{(j)\top} - \tilde{X}_i^{(j)} \tilde{X}_i^{(j)\top}) U.$$

Therefore under \mathcal{E}_1 , we have

$$\|M(s_l)^\top (\Xi - \Xi^{(i)}) M(s_{l+1})\|_F \lesssim \frac{(\lambda + \sigma^2)(r + \log n_j)}{n_j}.$$

Now we consider the rest terms in $\mathcal{T}_{\Sigma, k, \mathbf{s}, l}$. Since $s_l, s_{l+1} > 0$, there exists $l' \neq l \in [k]$, $s_{l'} = 0, s_{l'+1} > 0$ or $s_{l'} > 0, s_{l'+1} = 0$. Therefore we have

$$\begin{aligned} \|\mathcal{T}_{\Sigma, k, \mathbf{s}, l}\|_F & \leq \frac{C_1}{10^{k-2}} \lambda^{-2} \frac{(\lambda + \sigma^2)(r + \log n_j)}{n_j} (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}} \\ & \leq \frac{1}{10^{k-1}} \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j}, \end{aligned}$$

where the last line holds given the SNR condition.

Case 2: $s_l = 0, s_{l+1} > 0$ or $s_l > 0, s_{l+1} = 0$. In this case, we have

$$\begin{aligned} \|M(s_l)^\top (\Xi - \Xi^{(i)}) M(s_{l+1})\|_F &= \frac{1}{n_j} \|U_\perp^\top (X_i^{(j)} X_i^{(j)\top} - \tilde{X}_i^{(j)} \tilde{X}_i^{(j)\top}) U_\perp\|_F \\ &\lesssim \frac{1}{n_j} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j}. \end{aligned}$$

And under \mathcal{E}_2 , we have

$$\|\mathcal{T}_{\Sigma, k, s, l}\|_F \leq \frac{1}{10^{k-1}} \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j}.$$

Case 3: $s_l = s_{l+1} = 0$. In order to derive a tight upper bound, we need to use the leave-one-out technique. Notice

$$\begin{aligned} \mathcal{T}_{\Sigma, k, s, l} &= M(s_1) \Lambda^{-s_1} \underline{M(s_1)^\top \Xi^{(i)} M(s_2)} \cdots \underline{U_\perp^\top (\Xi - \Xi^{(i)}) U_\perp} \\ &\quad \cdots \underline{M(s_k)^\top \Xi M(s_{k+1}) \Lambda^{-s_{k+1}} M(s_{k+1})^\top} \\ &= \frac{1}{n_j} M(s_1) \Lambda^{-s_1} \underline{M(s_1)^\top \Xi^{(i)} M(s_2)} \cdots \underline{U_\perp^\top (X_i^{(j)} X_i^{(j)\top} - \tilde{X}_i^{(j)} \tilde{X}_i^{(j)\top}) U_\perp} \\ &\quad \cdots \underline{M(s_k)^\top \Xi M(s_{k+1}) \Lambda^{-s_{k+1}} M(s_{k+1})^\top}. \end{aligned}$$

We only consider the bound for

$$\|M(s_1) \Lambda^{-s_1} \underline{M(s_1)^\top \Xi^{(i)} M(s_2)} \cdots \underline{U_\perp^\top X_i^{(j)} X_i^{(j)\top} U_\perp} \cdots \underline{M(s_k)^\top \Xi M(s_{k+1}) \Lambda^{-s_{k+1}} M(s_{k+1})^\top}\|_F,$$

and the other term can be bounded similarly. Since $s_l = s_{l+1} = 0$. There exists some $l_0 \in [k+1]$, $s_{l_0} > 0$. We assume wlog $l_0 > l+1$ and that l_0 is the smallest integer that $s_{l_0} > 0$. In fact, if $l_0 < l$, then the term can be easier to bound due to the independence between $X_i^{(j)}$ and $\Xi^{(i)}$. Now we consider the term

$$\| \underbrace{X_i^{(j)\top} U_\perp}_{1\text{st}} \cdot \underbrace{U_\perp^\top \Xi U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi U}_F \|_F.$$

We now decompose $\Xi = \Xi_1 + \Xi_2$, with $\Xi_1 = \frac{1}{n_j} (X_i^{(j)} X_i^{(j)\top} - \Sigma)$, and $\Xi_2 = \frac{1}{n_j} (\sum_{i' \neq i} X_j^{(i')} X_j^{(i')\top} - \Sigma)$.

Then

$$\begin{aligned} \underbrace{X_i^{(j)\top} U_\perp}_{1\text{st}} \cdot \underbrace{U_\perp^\top \Xi U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi U}_F &= \underbrace{X_i^{(j)\top} U_\perp}_{1\text{st}} \cdot \underbrace{U_\perp^\top \Xi U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi_1 U}_F \\ &\quad + \underbrace{X_i^{(j)\top} U_\perp}_{1\text{st}} \cdot \underbrace{U_\perp^\top \Xi U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi_1 U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi_2 U}_F \\ &\quad + \cdots \\ &\quad + \underbrace{X_i^{(j)\top} U_\perp}_{1\text{st}} \cdot \underbrace{U_\perp^\top \Xi_2 U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi_2 U_\perp}_{(l_0-l)\text{-th}} \cdots \underbrace{U_\perp^\top \Xi_2 U}_F \\ &=: g_1^\top + \cdots + g_{l_0-l}^\top. \end{aligned}$$

Notice Ξ_2 is independent of $X_i^{(j)}$. Therefore condition on Ξ_2 ,

$$g_{l_0-l} \sim N(0, \sigma^2 \underline{U^\top \Xi_2 U_\perp} \cdot \underline{U_\perp^\top \Xi_2 U_\perp} \cdots \underline{U_\perp^\top \Xi_2 U_\perp} \cdot \underline{U_\perp^\top \Xi_2 U_\perp} \cdots \underline{U_\perp^\top \Xi_2 U_\perp} \cdot \underline{U_\perp^\top \Xi_2 U_\perp} \cdots \underline{U_\perp^\top \Xi_2 U_\perp}).$$

We define

$$\mathcal{E}_3 = \left\{ \|U_\perp^\top \Xi_2 U\| \lesssim (\lambda^{1/2} + \sigma)\sigma \sqrt{\frac{p}{n_j}} : \forall i \in [n_j] \right\} \cap \left\{ \|U_\perp^\top \Xi_2 U_\perp\| \lesssim \sigma^2 \sqrt{\frac{p}{n_j}} : \forall i \in [n_j] \right\}.$$

Then $\mathbb{P}(\mathcal{E}_3) \geq 1 - 4(e^{-c_0 p} + e^{-c_0 n_j})$. Then under $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$,

$$\|g_{l_0-l}\|_{\ell_2} \lesssim \frac{1}{10^{l_0-l-2}} \sqrt{r} \sigma (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}}.$$

And similarly for all $l' = 1, \dots, l_0 - l - 1$, we have

$$\|g_{l'}\|_{\ell_2} \lesssim \frac{1}{10^{l_0-l-2}} \frac{p}{n_j} \sigma^3 \sqrt{r} \lesssim \frac{1}{10^{l_0-l-2}} \sqrt{r} \sigma (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}}.$$

In summary,

$$\| \underbrace{X_i^{(j)\top U_\perp}_{1\text{st}}} \cdot \underbrace{U_\perp^\top \Xi U_\perp}_{(l_0-l)\text{-th}} \|_{\text{F}} \lesssim \frac{l_0}{10^{l_0-l-2}} \sqrt{r} \sigma (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}}.$$

Now we use the event $\mathcal{E}_1, \mathcal{E}_2$ to bound the rest of the terms, which give

$$\begin{aligned} & \|M(s_1) \Lambda^{-s_1} M(s_1)^\top \Xi^{(i)} M(s_2) \cdots U_\perp^\top X_i^{(j)} X_i^{(j)\top} U_\perp \cdots M(s_k)^\top \Xi M(s_{k+1}) \Lambda^{-s_{k+1}} M(s_{k+1})^\top \|_{\text{F}} \\ & \lesssim \lambda^{-2} \sqrt{p} \sigma \frac{1}{10^{k-2}} \sqrt{r} \sigma (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}}. \end{aligned}$$

This implies

$$\|\mathcal{T}_{\Sigma, k, s, l}\|_{\text{F}} \lesssim \frac{1}{n_j} \lambda^{-2} \sqrt{p} \sigma \frac{1}{10^{k-2}} \sqrt{r} \sigma (\lambda^{1/2} + \sigma) \sigma \sqrt{\frac{p}{n_j}}.$$

In other words, under the given SNR, we have

$$\|\mathcal{T}_{\Sigma, k, s, l}\|_{\text{F}} \leq \frac{1}{10^{k-1}} \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j}.$$

Finally from (35), we conclude

$$\begin{aligned} \left\| \sum_{k \geq 2} (\mathcal{S}_{\Sigma, k}(\Xi) - \mathcal{S}_{\Sigma, k}(\Xi^{(i)})) \right\|_{\text{F}} & \leq \sum_{k \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_k} \sum_{l \in [k]} \|\mathcal{T}_{\Sigma, k, s, l}\|_{\text{F}} \\ & \leq \sum_{k \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_k} \sum_{l \in [k]} \frac{1}{10^{k-1}} \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j} \\ & \lesssim \sum_{k \geq 2} \sum_{\mathbf{s} \in \mathbb{S}_k} \frac{1}{8^k} \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j} \\ & \lesssim \sum_{k \geq 2} \frac{1}{2^k} \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j} \\ & \leq \frac{1}{n_j} \lambda^{-1} \sqrt{\lambda + \sigma^2} \sigma \sqrt{p} \sqrt{r + \log n_j}. \end{aligned}$$

In summary, by setting $\eta = n_j \wedge p$ and taking union bound over all $i \in [n_j]$, we conclude with probability exceeding $1 - n_j^{-100} - 12e^{-c_0(p \wedge n_j)}$,

$$\max_{i \in [n_j]} \|\tilde{U}_j \tilde{U}_j^\top - \tilde{U}_j^{(i)} \tilde{U}_j^{(i)\top}\|_F \leq C \frac{1}{n_j} \sqrt{\frac{\lambda + \sigma^2}{\lambda}} \frac{\sigma^2}{\lambda} \sqrt{p(r + \log n_j)}.$$

□

The proof for the sensitivity of singular values is a direct result of Lemma 4 in [Cai et al. \(2024b\)](#). The claim of Lemma 1 then follows the sensitivity of Gaussian mechanism (see e.g. Lemma 1 in [Cai et al. \(2024b\)](#)).

A.4 Proof of Theorem 3

We first show the lower bound for subspace estimation and then the lower bound for covariance matrix estimation.

Lower bound for subspace estimation. Let Θ be a random matrix of size $p \times r$ with its entries i.i.d. $N(0, 1)$. The density function of Θ is $p(\Theta) = (2\pi)^{-pr/2} \cdot \exp(-\|\Theta\|_F^2/2)$. Let $W := \Theta^\top \Theta$ has the Wishart distribution $\mathcal{W}_r(I_r, p)$. Define a map $\psi : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{p \times p}$ as $\psi(\Theta) = \Theta(\Theta^\top \Theta)^{-1} \Theta^\top$. Denote $\psi_{k_1 k_2}(\Theta) := e_{k_1}^\top \psi(\Theta) e_{k_2}$ be the (k_1, k_2) -th component function of $\psi(\Theta)$ for all $k_1, k_2 \in [p]$. Basically, ψ maps a given $p \times r$ matrix to a $p \times p$ rank- r projection matrix. Moreover, denote $\bar{\Theta} := \Theta(\Theta^\top \Theta)^{-1/2} \in \mathbb{O}^{p \times r}$ the left singular vectors of Θ . It is clear by definition that $\psi(\Theta) = \bar{\Theta} \bar{\Theta}^\top$.

Suppose that $X_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} N(0, \lambda \bar{\Theta} \bar{\Theta}^\top + \sigma^2 I_p)$ for all $j \in [m]$ and $\forall i \in [n_j]$. Denote $\bar{\Theta}_\perp \in \mathbb{O}^{p \times (p-r)}$ such that $(\bar{\Theta}, \bar{\Theta}_\perp)$ is a $p \times p$ orthogonal matrix. We denote

$$\begin{bmatrix} Y_i^{(j)} \\ Z_i^{(j)} \end{bmatrix} := \begin{bmatrix} \bar{\Theta}^\top \\ \bar{\Theta}_\perp^\top \end{bmatrix} X_i^{(j)} \stackrel{\text{i.i.d.}}{\sim} \begin{bmatrix} N(0, (\lambda + \sigma^2) I_r) \\ N(0, \sigma^2 I_{p-r}) \end{bmatrix}, \quad \forall j \in [m], i \in [n_j] \quad (36)$$

We define the score corresponding to $X_i^{(j)}$ as

$$S_{j,i} := \nabla \log p(X_i^{(j)}; \Theta) = ((\lambda + \sigma^2)^{-1} - \sigma^{-2}) \bar{\Theta}_\perp Z_i^{(j)} Y_i^{(j)\top} W^{-1/2} \in \mathbb{R}^{p \times r}. \quad (37)$$

Denote $\mathcal{D}_j := \{X_i^{(j)} : i \in [n_j]\}$ the data set stored at j -th local client. We define

$$S_j := \nabla \log p(\mathcal{D}_j; \Theta) = \sum_{i=1}^{n_j} \nabla \log p(X_i^{(j)}; \Theta) = \sum_{i=1}^{n_j} S_{j,i}.$$

This induces a linear operator $\mathbb{R}^{p \times r} \mapsto \mathbb{R}^{p \times r}$ for all $j \in [m], i \in [n_j]$ defined by

$$\mathcal{C}_{j,i}(V) := \mathbb{E} \langle S_{j,i}, V \rangle S_{j,i} = \frac{\lambda^2}{(\lambda + \sigma^2) \sigma^2} \bar{\Theta}_\perp \bar{\Theta}_\perp^\top V W^{-1}, \quad (38)$$

where the expectation is taken w.r.t. $X_i^{(j)}$. We denote the sum as

$$\mathcal{C}_j(V) := \sum_{i=1}^{n_j} \mathcal{C}_{j,i}(V) = \frac{n_j \lambda^2}{(\lambda + \sigma^2) \sigma^2} \bar{\Theta}_\perp \bar{\Theta}_\perp^\top V W^{-1}. \quad (39)$$

The following lemma states a matrix version of the Van Trees' inequality. We first clarify some useful notations. In the following, we view the gradient $\nabla\psi(\Theta)$ as an operator maps from $\mathbb{R}^{p \times r}$ to $\mathbb{R}^{p \times p}$, i.e., $\nabla\psi(\Theta)(Y) \in \mathbb{R}^{p \times p}$ for all $Y \in \mathbb{R}^{p \times r}$ as a directional derivative. See more details in Appendix C.1. Similarly, the gradient $\nabla \log p(\mathcal{D}_j; \Theta | \hat{U}_j) \in \mathbb{R}^{p \times r}$ can be identified as an operator maps from $\mathbb{R}^{p \times r} \rightarrow \mathbb{R}$ such that $\cdot \mapsto \langle \nabla \log p(\mathcal{D}_j; \Theta | \hat{U}_j), \cdot \rangle$. Let $\nabla\psi(\Theta)^* : \mathbb{R}^{p \times p} \mapsto \mathbb{R}^{p \times r}$ the adjoint operator satisfying

$$\langle \nabla\psi(\Theta)(Y), M \rangle = \langle \nabla\psi(\Theta)^*(M), Y \rangle, \quad \text{for } \forall Y \in \mathbb{R}^{p \times r} \quad \text{and} \quad \forall M \in \mathbb{R}^{p \times p}.$$

Let \circ denote the composition of operators. The trace of a self-adjoint operator \mathcal{L} that maps from $\mathbb{R}^{p \times p}$ to itself is defined by

$$\text{tr}(\mathcal{L}) := \sum_{i,j \in [p]} \langle \mathcal{L}(e_i e_j^\top), e_i e_j^\top \rangle,$$

where e_i denotes the i -th canonical basis vector of \mathbb{R}^p .

Lemma 3. *For any estimator $\hat{U} \in \mathbb{O}^{p \times r}$ of $\psi(\Theta)$, its average-case error rate is lower bounded by*

$$\int \mathbb{E} \|\hat{U} \hat{U}^\top - \psi(\Theta)\|_F^2 \cdot p(\Theta) d\Theta \geq \frac{(\int \text{tr}(\nabla\psi(\Theta) \circ \nabla\psi(\Theta)^*) \cdot p(\Theta) d\Theta)^2}{\sum_{j=1}^M \mathbb{E} \int \text{tr}(\nabla\psi(\Theta) \circ \mathcal{I}(\Theta | \hat{U}_j) \circ \nabla\psi(\Theta)^*) \cdot p(\Theta) d\Theta + \mathcal{J}(p)},$$

where \hat{U}_j denotes any $(\varepsilon_j, \delta_j)$ -DP estimator based on dataset \mathcal{D}_j at j -th local client and

$$\begin{aligned} \mathcal{J}(p) &= \sum_{k_1, k_2 \in [p]} \int \Delta \psi_{k_1 k_2}^2(\Theta) p(\Theta) d\Theta, \\ \Delta \psi_{k_1 k_2}^2(\Theta) &= \sum_{(i,j) \in [p] \times [r]} \frac{\partial^2 (\psi_{k_1 k_2}^2)}{\partial \Theta_{ij}^2}(\Theta), \quad \forall k_1, k_2 \in [p] \\ \mathcal{I}(\Theta | \hat{U}_j) &= \mathbb{E} [(\nabla \log p(\mathcal{D}_j; \Theta | \hat{U}_j))^* \circ \nabla \log p(\mathcal{D}_j; \Theta | \hat{U}_j)]. \end{aligned}$$

It suffices to control the three terms involved in the right hand side of Lemma 3. We will show (see Appendix C.1 for more details) $\nabla\psi(\Theta) : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{p \times p}$ is the following linear map:

$$\nabla\psi(\Theta)(Y) = \bar{\Theta}_\perp \bar{\Theta}_\perp^\top Y W^{-1/2} \bar{\Theta}^\top + \bar{\Theta} W^{-1/2} Y^\top \bar{\Theta}_\perp \bar{\Theta}_\perp^\top. \quad (40)$$

Meanwhile (see Appendix C.2), $\nabla\psi(\Theta)^* : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p \times r}$ is given by

$$\nabla\psi(\Theta)^*(M) = \bar{\Theta}_\perp \bar{\Theta}_\perp^\top (M + M^\top) \bar{\Theta} W^{-1/2}. \quad (41)$$

Lower bound for $\int \text{tr}(\nabla\psi(\Theta) \circ \nabla\psi(\Theta)^*)p(\Theta)d\Theta$. Based on (40) and (41), we have for all $M \in \mathbb{R}^{p \times p}$ that

$$\begin{aligned} & \nabla\psi(\Theta) \circ \nabla\psi(\Theta)^*(M) \\ &= \bar{\Theta}_\perp \bar{\Theta}_\perp^\top (M + M^\top) \bar{\Theta} W^{-1} \bar{\Theta}^\top + \bar{\Theta} W^{-1} \bar{\Theta}^\top (M + M^\top) \bar{\Theta}_\perp \bar{\Theta}_\perp^\top. \end{aligned}$$

So, by definition, we get

$$\begin{aligned} & \text{tr}(\nabla\psi(\Theta) \circ \nabla\psi(\Theta)^*) \\ &= \sum_{i,j \in [p]} e_i^\top \left(\bar{\Theta}_\perp \bar{\Theta}_\perp^\top (e_i e_j^\top + e_j e_i^\top) \bar{\Theta} W^{-1} \bar{\Theta}^\top + \bar{\Theta} W^{-1} \bar{\Theta}^\top (e_i e_j^\top + e_j e_i^\top) \bar{\Theta}_\perp \bar{\Theta}_\perp^\top \right) e_j \\ &= 2\text{tr}(\bar{\Theta}_\perp \bar{\Theta}_\perp^\top) \cdot \text{tr}(\bar{\Theta} W^{-1} \bar{\Theta}^\top) + 2\langle \bar{\Theta} W^{-1} \bar{\Theta}^\top, \bar{\Theta}_\perp \bar{\Theta}_\perp^\top \rangle \\ &= 2(p-r)\text{tr}(W^{-1}). \end{aligned}$$

Therefore,

$$\int \text{tr}(\nabla\psi(\Theta) \circ \nabla\psi(\Theta)^*)p(\Theta)d\Theta = 2(p-r)\mathbb{E}\text{tr}(W^{-1}),$$

where $W \sim W_r(I_r, p)$ follows the Wishart distribution. Following the Theorem 3.1 of [Von Rosen \(1988\)](#), we have $\mathbb{E}W^{-1} = (p-r-1)^{-1}I_r$ if $p-r-1 \geq 1$. So we have

$$\int \text{tr}(\nabla\psi(\Theta) \circ \nabla\psi(\Theta)^*)p(\Theta)d\Theta = \frac{2(p-r)r}{p-r-1} \geq 2r.$$

Upper bound for $\mathcal{J}(p)$. Simple calculations show, for all $k_1, k_2 \in [p]$, that

$$\begin{aligned} \nabla\psi_{k_1 k_2}(\Theta) &= \bar{\Theta}_\perp \bar{\Theta}_\perp^\top (e_{k_1} e_{k_2}^\top + e_{k_2} e_{k_1}^\top) \bar{\Theta} W^{-1/2}, \\ \Delta\psi_{k_1 k_2} &= 2[\bar{\Theta}_\perp \bar{\Theta}_\perp^\top]_{k_1 k_2} \cdot \text{tr}(W^{-1}) - 2(p-r)[\bar{\Theta} W^{-1} \bar{\Theta}^\top]_{k_1 k_2}. \end{aligned}$$

Since $\nabla p(\Theta) = (2\pi)^{-pr/2} \exp(-\|\Theta\|_{\mathbb{F}}^2/2)(-\Theta)$, we have $\langle \nabla p, \nabla\psi_{k_1 k_2} \rangle = 0$ for all $k_1, k_2 \in [p]$. As a result,

$$\begin{aligned} \int \sum_{k_1, k_2 \in [p]} (\Delta\psi_{k_1 k_2})^2 p(\Theta) d\Theta &= \int \left(4(p-r)^2 \|W^{-1}\|_{\mathbb{F}}^2 + 4(p-r)(\text{tr}(W^{-1}))^2 \right) p(\Theta) d\Theta \\ &= 4(p-r)^2 \mathbb{E} \|W^{-1}\|_{\mathbb{F}}^2 + 4(p-r) \mathbb{E} (\text{tr}(W^{-1}))^2 \\ &\leq 4(p-r)^2 \mathbb{E} \|W^{-1}\|_{\mathbb{F}}^2 + 4(p-r)r \mathbb{E} \|W^{-1}\|_{\mathbb{F}}^2 \\ &\leq 8(p-r)^2 \mathbb{E} \|W^{-1}\|_{\mathbb{F}}^2 \end{aligned}$$

as long as $p \geq 2r$.

Following the Corollary 3.1 of [Von Rosen \(1988\)](#), we have ²

$$\mathbb{E}W^{-2} = (c_1 + c_2 + c_2r)I_r,$$

where $c_1 = (p - r - 2)c_2$ and $c_2 = [(p - r)(p - r - 1)(p - r - 3)]^{-1}$. As a result,

$$\int \sum_{k_1, k_2 \in [p]} (\Delta\psi_{k_1 k_2})^2 p(\Theta) d\Theta \lesssim r. \quad (42)$$

Upper bound for $\mathbb{E} \int \text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta)) \cdot p(\Theta) d\Theta$. In fact, for all $j \in [m]$, we have

$$\begin{aligned} & \mathbb{E} \int \text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta)) \\ &= \mathbb{E}[\mathbb{E}\|\nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j))\|_{\mathbb{F}}^2] \\ &= \mathbb{E}\left[\sum_{i=1}^{n_j} \underbrace{\mathbb{E}\langle \nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j)), \nabla\psi(\Theta)(\nabla \log p(X_i^{(j)}; \Theta|\widehat{U}_j)) \rangle}_{=: G_i^{(j)}}\right]. \end{aligned} \quad (43)$$

Meanwhile, for all $j \in [m], i \in [n_j]$, we also define

$$\widetilde{G}_i^{(j)} := \mathbb{E}\langle \nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j)), \nabla\psi(\Theta)(\nabla \log p(\widetilde{X}_i^{(j)}; \Theta|\widehat{U}_j)) \rangle,$$

where $\widetilde{X}_i^{(j)}$ is an i.i.d. copy of $X_i^{(j)}$. Note that the expectation is taken conditional on \widehat{U}_j , implying that $\mathbb{E}\widetilde{G}_i^{(j)} = 0$.

Denote $(G_i^{(j)})^+ := 0 \vee G_i^{(j)}$ and $(G_i^{(j)})^- := -0 \wedge G_i^{(j)}$. By slightly abuse of notation, we denote $X^{(j)} := [X_1^{(j)}, \dots, X_{n_j}^{(j)}]$ and $p_{X^{(j)}}$ the corresponding density function. Since \widehat{U}_j is $(\varepsilon_j, \delta_j)$ -DP, by definition, we have

$$\begin{aligned} & \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) = \int \int \mathbb{P}\left((G_i^{(j)})^+ \geq t \mid X^{(j)} = x^{(j)}\right) p_{X^{(j)}}(x^{(j)}) p_{\widetilde{X}_i^{(j)}}(\widetilde{x}_i^{(j)}) dx^{(j)} d\widetilde{x}_i^{(j)} \\ & \leq \int \int \left(e^{\varepsilon_j} \mathbb{P}\left((\widetilde{G}_i^{(j)})^+ \geq t \mid X^{(j)} = x^{(j)}, \widetilde{X}_i^{(j)} = \widetilde{x}_i^{(j)}\right) + \delta_j\right) p_{X^{(j)}}(x^{(j)}) p_{\widetilde{X}_i^{(j)}}(\widetilde{x}_i^{(j)}) dx^{(j)} d\widetilde{x}_i^{(j)} \\ & = e^{\varepsilon_j} \mathbb{P}\left((\widetilde{G}_i^{(j)})^+ \geq t\right) + \delta_j. \end{aligned}$$

Therefore, for an arbitrary $\tau > 0$ to be determined later, we have

$$\begin{aligned} \int_0^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt &= \int_0^\tau \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt + \int_\tau^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt \\ &\leq e^{\varepsilon_j} \int_0^\tau \mathbb{P}\left((\widetilde{G}_i^{(j)})^+ \geq t\right) dt + \tau\delta_j + \int_\tau^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt \\ &\leq (1 + C_1\varepsilon_j) \int_0^\tau \mathbb{P}\left((\widetilde{G}_i^{(j)})^+ \geq t\right) dt + \tau\delta_j + \int_\tau^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt, \end{aligned}$$

²There appears to be a typo in Corollary 3.1 (i), where the coefficient of the second term on the right hand side should be c_2 instead of c_1 .

where in the last inequality we used the fact that $\max_{j \in [m]} \varepsilon_j = O(1)$. And similarly we can show

$$\begin{aligned} \int_0^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt &\geq \int_0^{+\infty} \mathbb{P}\left((\tilde{G}_i^{(j)})^- \geq t\right) dt - C_1 \varepsilon_j \int_0^{+\infty} \mathbb{P}\left((\tilde{G}_i^{(j)})^- \geq t\right) dt \\ &\quad - \tau \delta_j - \int_\tau^{+\infty} \mathbb{P}\left((\tilde{G}_i^{(j)})^+ \geq t\right) dt. \end{aligned}$$

Combine these two inequalities and we get

$$\mathbb{E}G_i^{(j)} \leq \mathbb{E}\tilde{G}_i^{(j)} + 2C_1\varepsilon_j\mathbb{E}|\tilde{G}_i^{(j)}| + 2\tau\delta_j + \int_\tau^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt + \int_\tau^{+\infty} \mathbb{P}\left((\tilde{G}_i^{(j)})^- \geq t\right) dt. \quad (44)$$

The first term in above right hand side vanishes. We now bound $\mathbb{E}|\tilde{G}_i^{(j)}|$. By Cauchy-Schwarz inequality, we get

$$\begin{aligned} \mathbb{E}|\tilde{G}_i^{(j)}| &\leq \sqrt{\mathbb{E}|\tilde{G}_i^{(j)}|^2} \\ &= \sqrt{\mathbb{E}\left(\mathbb{E}\langle \nabla\psi(\Theta)(\nabla\log p(\mathcal{D}_j; \Theta|\hat{U}_j)), \nabla\psi(\Theta)(\nabla\log p(\tilde{X}_i^{(j)}; \Theta|\hat{U}_j)) \rangle^2\right)} \\ &\leq \sqrt{\mathbb{E}\left[\mathbb{E}\|\nabla\psi(\Theta)(\nabla\log p(\mathcal{D}_j; \Theta|\hat{U}_j))\|_{\mathbb{F}}^2\right]} \cdot \sqrt{\mathbb{E}\left[\mathbb{E}\|\nabla\psi(\Theta)(\nabla\log p(\tilde{X}_i^{(j)}; \Theta|\hat{U}_j))\|_{\mathbb{F}}^2\right]}. \end{aligned}$$

Using the data processing inequality, we have

$$\mathbb{E}\left[\mathbb{E}\|\nabla\psi(\Theta)(\nabla\log p(\tilde{X}_i^{(j)}; \Theta|\hat{U}_j))\|_{\mathbb{F}}^2\right] \leq \mathbb{E}\|\nabla\psi(\Theta)(\nabla\log p(\tilde{X}_i^{(j)}; \Theta))\|_{\mathbb{F}}^2.$$

From (37) and (40), we obtain

$$\mathbb{E}\left[\mathbb{E}\|\nabla\psi(\Theta)(\nabla\log p(\tilde{X}_i^{(j)}; \Theta|\hat{U}_j))\|_{\mathbb{F}}^2\right] \leq \frac{2\lambda^2}{(\lambda + \sigma^2)\sigma^2} \|W^{-2}\|.$$

In summary, we have

$$\mathbb{E}|\tilde{G}_i^{(j)}| \leq \sqrt{\frac{2\lambda^2}{(\lambda + \sigma^2)\sigma^2} \|W^{-2}\|} \cdot \sqrt{\mathbb{E}\left[\mathbb{E}\|\nabla\psi(\Theta)(\nabla\log p(\mathcal{D}_j; \Theta|\hat{U}_j))\|_{\mathbb{F}}^2\right]}. \quad (45)$$

It remains to bound the tail probabilities $\mathbb{P}\left((G_i^{(j)})^+ \geq t\right)$ and $\mathbb{P}\left((\tilde{G}_i^{(j)})^- \geq t\right)$. Without loss of generality, we take $i = 1$. We shall first consider $\mathbb{E}|G_1^{(j)}|^k$ for some large and absolute integer $k > 0$. Recall $G_1^{(j)} = \mathbb{E}\langle \nabla\psi(\Theta)(\nabla\log p(\mathcal{D}_j; \Theta|\hat{U}_j)), \nabla\psi(\Theta)(\nabla\log p(X_1^{(j)}; \Theta|\hat{U}_j)) \rangle$. By definition, we get

$$\begin{aligned} \mathbb{E}|G_1^{(j)}|^k &= \mathbb{E}\left|\mathbb{E}\langle \nabla\psi(\Theta)(\nabla\log p(\mathcal{D}_j; \Theta|\hat{U}_j)), \nabla\psi(\Theta)(\nabla\log p(X_1^{(j)}; \Theta|\hat{U}_j)) \rangle\right|^k \\ &\leq \mathbb{E}\left|\langle \nabla\psi(\Theta)(S_j), \nabla\psi(\Theta)(S_{j,1}) \rangle\right|^k. \end{aligned}$$

where the inequality is due to Jensen's inequality and recall

$$\begin{aligned} S_j &:= \nabla\psi(\Theta)(\nabla\log p(\mathcal{D}_j; \Theta|\widehat{U}_j)), \\ S_{j,1} &:= \nabla\psi(\Theta)(\nabla\log p(X_1^{(j)}; \Theta|\widehat{U}_j)) \end{aligned}$$

Observe that $\langle \nabla\psi(\Theta)(S_j), \nabla\psi(\Theta)(S_{j,1}) \rangle = \sum_{i=2}^{n_j} \langle \nabla\psi(\Theta)(S_{j,i}), \nabla\psi(\Theta)(S_{j,1}) \rangle$. Therefore

$$\begin{aligned} &\mathbb{E}|\langle \nabla\psi(\Theta)(S_j), \nabla\psi(\Theta)(S_{j,1}) \rangle|^k \\ &= \mathbb{E}\left| \|\nabla\psi(\Theta)(S_{j,1})\|_{\mathbb{F}}^2 + \left\langle \sum_{i=2}^{n_j} \nabla\psi(\Theta)(S_{j,i}), \nabla\psi(\Theta)(S_{j,1}) \right\rangle \right|^k \\ &\leq 2^k \mathbb{E}\|\nabla\psi(\Theta)(S_{j,1})\|_{\mathbb{F}}^{2k} + 2^k \mathbb{E}\left| \left\langle \sum_{i=2}^{n_j} \nabla\psi(\Theta)(S_{j,i}), \nabla\psi(\Theta)(S_{j,1}) \right\rangle \right|^k. \end{aligned}$$

Denote

$$Y_{2:n_j}^{(j)} := [Y_2^{(j)}, \dots, Y_{n_j}^{(j)}] \quad \text{and} \quad Z_{2:n_j}^{(j)} := [Z_2^{(j)}, \dots, Z_{n_j}^{(j)}].$$

Then we can write

$$\begin{aligned} \sum_{i=2}^{n_j} \nabla\psi(\Theta)(S_{j,i}) &= ((\lambda + \sigma^2)^{-1} - \sigma^{-2}) \left(\bar{\Theta}_{\perp} \sum_{i=2}^{n_j} Z_i^{(j)} Y_i^{(j)\top} W^{-1} \bar{\Theta}^{\top} + \bar{\Theta} W^{-1} \sum_{i=2}^{n_j} Y_i^{(j)} Z_i^{(j)\top} \bar{\Theta}_{\perp}^{\top} \right) \\ &= ((\lambda + \sigma^2)^{-1} - \sigma^{-2}) \left(\bar{\Theta}_{\perp} Z_{2:n_j}^{(j)} Y_{2:n_j}^{(j)\top} W^{-1} \bar{\Theta}^{\top} + \bar{\Theta} W^{-1} Y_{2:n_j}^{(j)} Z_{2:n_j}^{(j)\top} \bar{\Theta}_{\perp}^{\top} \right). \end{aligned}$$

By the definitions in eq. (36), we know that all entries of $Y_{2:n_j}^{(j)}$ are i.i.d. obeying distribution $N(0, \lambda + \sigma^2)$. Similarly, all entries of $Z_{2:n_j}^{(j)}$ are i.i.d. obeying $N(0, \sigma^2)$. Based on these facts, we get

$$\sum_{i=2}^{n_j} \langle \nabla\psi(\Theta)(S_{j,i}), \nabla\psi(\Theta)(S_{j,1}) \rangle = 2((\lambda + \sigma^2)^{-1} - \sigma^{-2})^2 \langle Z_{2:n_j}^{(j)} Y_{2:n_j}^{(j)\top} W^{-1}, Z_1^{(j)} Y_1^{(j)\top} W^{-1} \rangle.$$

By denoting $\mu := ((\lambda + \sigma^2)^{-1} - \sigma^{-2})^2 (\lambda + \sigma^2) \sigma^2$, we can write

$$\sum_{i=2}^{n_j} \langle \nabla\psi(\Theta)(S_{j,i}), \nabla\psi(\Theta)(S_{j,1}) \rangle = 2\mu \langle \bar{Z}_{2:n_j}^{(j)} \bar{Y}_{2:n_j}^{(j)\top} W^{-1}, \bar{Z}_1^{(j)} \bar{Y}_1^{(j)\top} W^{-1} \rangle,$$

where $\bar{\cdot}$ are the normalized version, i.e., the entries of $\bar{Z}^{(j)}$ and $\bar{Y}^{(j)}$ are i.i.d. standard normal random variables.

Using the tower rule, we get

$$\begin{aligned}
& \mathbb{E}|\langle \bar{Z}_{2:n_j}^{(j)} \bar{Y}_{2:n_j}^{(j)\top} W^{-1}, \bar{Z}_1^{(j)} \bar{Y}_1^{(j)\top} W^{-1} \rangle|^k = \mathbb{E}|\langle \bar{Z}_{2:n_j}^{(j)}, \bar{Z}_1^{(j)} \bar{Y}_1^{(j)\top} W^{-2} \bar{Y}_{2:n_j}^{(j)} \rangle|^k \\
& \leq k^{k/2} \cdot \mathbb{E}\|\bar{Z}_1^{(j)}\|_{\ell_2}^k \cdot \mathbb{E}\|\bar{Y}_{2:n_j}^{(j)\top} W^{-2} \bar{Y}_1^{(j)}\|_{\ell_2}^k \\
& = k^{k/2} \cdot \prod_{i=0}^{(k/2)-1} (p-r+2i) \cdot \prod_{l=0}^{(k/2)-1} (n_j-1+2l) \cdot \mathbb{E}\|W^{-2} \bar{Y}_1^{(j)}\|_{\ell_2}^k \\
& \leq C^k k^k \cdot \prod_{i=0}^{(k/2)-1} (p-r+2i) \cdot \prod_{l=0}^{(k/2)-1} (n_j-1+2l) \cdot \|W^{-2}\|_{\mathbb{F}}^k,
\end{aligned}$$

where, in the first and last inequalities, we used Lemma 4 to show that $\mathbb{E}\|W^{-2} \bar{Y}_1^{(j)}\|_{\ell_2}^k \leq (Ck^{1/2}\|W^{-2}\|_{\mathbb{F}})^k$. Here $C > 0$ is an absolute constant.

Similarly, we get

$$\begin{aligned}
\mathbb{E}\|\nabla\psi(\Theta)(S_{j,1})\|_{\mathbb{F}}^{2k} &= (2\mu)^k \cdot \mathbb{E}\|\bar{Z}_1^{(j)} \bar{Y}_1^{(j)\top} W^{-1}\|_{\mathbb{F}}^{2k} \\
&= (2\mu)^k \cdot \mathbb{E}\|\bar{Z}_1^{(j)}\|_{\ell_2}^{2k} \cdot \mathbb{E}\|W^{-1} \bar{Y}_1^{(j)}\|_{\ell_2}^{2k} \\
&\leq C^k \cdot \prod_{i=0}^{k-1} (p-r+2i) \cdot k^k \|W^{-1}\|_{\mathbb{F}}^{2k}.
\end{aligned}$$

In summary, we have

$$\begin{aligned}
\mathbb{E}|G_1^{(j)}|^k &\leq \mathbb{E}|\langle \nabla\psi(\Theta)(S_j), \nabla\psi(\Theta)(S_{j,1}) \rangle|^k \\
&\leq C^k k^k \mu^k \left((p-r)^k \|W^{-1}\|_{\mathbb{F}}^{2k} + (p-r)^{k/2} n_j^{k/2} \|W^{-2}\|_{\mathbb{F}}^k \right).
\end{aligned}$$

We can similarly show the upper bound for $\mathbb{E}|\tilde{G}_i^{(j)}|^k$ as

$$\mathbb{E}|\tilde{G}_i^{(j)}|^k \leq C^k k^k \mu^k \left((d-r)^k \|W^{-1}\|_{\mathbb{F}}^{2k} + (d-r)^{k/2} n_j^{k/2} \|W^{-2}\|_{\mathbb{F}}^k \right).$$

By Markov's inequality, we get

$$\mathbb{P}\left((G_i^{(j)})^+ \geq t \right) \leq \mathbb{P}\left(|G_i^{(j)}| \geq t \right) = \mathbb{P}\left(|G_i^{(j)}|^k \geq t^k \right) \leq \frac{\mathbb{E}|G_i^{(j)}|^k}{t^k}.$$

Therefore

$$\int_{\tau}^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t \right) dt \leq \int_{\tau}^{+\infty} \frac{\mathbb{E}|G_i^{(j)}|^k}{t^k} dt = \frac{1}{k-1} \tau^{-k+1} \mathbb{E}|G_i^{(j)}|^k.$$

Observe that, by setting $\tau = (\delta_j^{-1} \mathbb{E}|G_i^{(j)}|^k)^{1/k}$, we get

$$\begin{aligned}
& \tau \delta_j + \int_{\tau}^{+\infty} \mathbb{P}\left((G_i^{(j)})^+ \geq t\right) dt + \int_{\tau}^{+\infty} \mathbb{P}\left((\tilde{G}_i^{(j)})^- \geq t\right) dt \\
& \leq \underbrace{\frac{1}{k-1} \delta_j \tau + \dots + \frac{1}{k-1} \delta_j \tau + \frac{2}{k-1} \tau^{-k+1} \mathbb{E}|G_i^{(j)}|^k}_{k-1 \text{ terms}} \\
& \leq 2(\mathbb{E}|G_i^{(j)}|^k)^{1/k} \delta_j^{\frac{k-1}{k}} \\
& \leq Ck \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} \left((d-r) \|W^{-1}\|_{\mathbb{F}}^2 + (d-r)^{1/2} n_j^{1/2} \|W^{-2}\|_{\mathbb{F}} \right) \delta_j^{\frac{k-1}{k}}.
\end{aligned}$$

By plugging the above bound into (44), we get

$$\begin{aligned}
\mathbb{E}G_i^{(j)} & \leq 2C_1 \varepsilon_j \sqrt{\frac{2\lambda^2}{(\lambda + \sigma^2)\sigma^2} \|W^{-2}\|} \sqrt{\mathbb{E}[\mathbb{E}\|\nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j))\|_{\mathbb{F}}^2]} \\
& \quad + Ck \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} \left((d-r) \|W^{-1}\|_{\mathbb{F}}^2 + (d-r)^{1/2} n_j^{1/2} \|W^{-2}\|_{\mathbb{F}} \right) \delta_j^{\frac{k-1}{k}}.
\end{aligned}$$

Together with (43), we get

$$\begin{aligned}
& \mathbb{E}[\mathbb{E}\|\nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j))\|_{\mathbb{F}}^2] \\
& \leq 2C_1 n_j \varepsilon_j \sqrt{\frac{2\lambda^2}{(\lambda + \sigma^2)\sigma^2} \|W^{-2}\|} \sqrt{\mathbb{E}[\mathbb{E}\|\nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j))\|_{\mathbb{F}}^2]} \\
& \quad + Ckn_j \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} \left((p-r) \|W^{-1}\|_{\mathbb{F}}^2 + (p-r)^{1/2} n_j^{1/2} \|W^{-2}\|_{\mathbb{F}} \right) \delta_j^{\frac{k-1}{k}}.
\end{aligned}$$

Therefore, as long as

$$Ck \cdot n_j \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} \left((p-r)r + (p-r)^{1/2} r^{1/2} n_j^{1/2} \right) \delta_j^{\frac{k-1}{k}} \leq C_1^2 n_j^2 \varepsilon_j^2 \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2},$$

we have

$$\mathbb{E}[\mathbb{E}\|\nabla\psi(\Theta)(\nabla \log p(\mathcal{D}_j; \Theta|\widehat{U}_j))\|_{\mathbb{F}}^2] \leq C_1^2 n_j^2 \varepsilon_j^2 \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} \|W^{-2}\|.$$

As a result, we get

$$\mathbb{E} \int \text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta)) \cdot p(\Theta) d\Theta \leq C_1^2 n_j^2 \varepsilon_j^2 \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} \cdot \mathbb{E}\|W^{-2}\|, \quad (46)$$

where recall that $W \sim W_r(I_r, p)$ follows the Wishart distribution.

Using the data processing inequality, we have another upper bound for $\mathbb{E}\text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta))$ as

$$\mathbb{E}\text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta)) \leq \text{tr}(\nabla\psi(\Theta) \circ \mathcal{C}_j \circ \nabla\psi(\Theta)^*),$$

where \mathcal{C}_j is defined in (39). From (38) and $\mathcal{C}_j = \sum_{i=1}^{n_j} \mathcal{C}_{j,i}$, we see

$$\text{tr}(\nabla\psi(\Theta) \circ \mathcal{C}_j \circ \nabla\psi(\Theta)^*) = \frac{2n_j\lambda^2}{(\lambda + \sigma^2)\sigma^2}(p-r)\text{tr}(W^{-2}).$$

Therefore,

$$\int \text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta)) \cdot p(\Theta)d\Theta \leq \frac{4n_j\lambda^2}{(\lambda + \sigma^2)\sigma^2} \frac{r}{p-r}.$$

In summary, we have

$$\int \text{tr}(\nabla\psi(\Theta)^* \circ \mathcal{I}(\Theta|\widehat{U}_j) \circ \nabla\psi(\Theta)) \cdot p(\Theta)d\Theta \leq \min \left\{ C_1^2 n_j^2 \varepsilon_j^2 \mathbb{E}\|W^{-2}\|, \frac{4n_j r}{p-r} \right\} \cdot \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2}.$$

Finally, we plug these bounds into the right hand side of the inequality in Lemma 3, we obtain (recall that we focus on the regime $\max_{j \in [m]} \varepsilon_j = O(1)$)

$$\int \|\widehat{U}\widehat{U}^\top - \psi(\Theta)\|_{\mathbb{F}}^2 \cdot p(\Theta)d\Theta \gtrsim \frac{r^2}{\sum_{i=1}^m \min \left\{ n_j^2 \varepsilon_j^2 \cdot \mathbb{E}\|W^{-2}\|, \frac{n_j r}{p-r} \right\} \cdot \frac{\lambda^2}{(\lambda + \sigma^2)\sigma^2} + r}$$

Finally we bound $\mathbb{E}\|W^{-2}\|$. Denote the event $\mathcal{F}_0 := \{\|W - pI_r\| \leq p/2\}$. From the basic concentration inequality of sample covariance matrix (Koltchinskii and Lounici, 2017), we have $\mathbb{P}(\mathcal{F}_0) \geq 1 - e^{-c_1 p}$. Under \mathcal{F}_0 , we have $\lambda_{\min}(W) \geq p/2$. So we have

$$\begin{aligned} \mathbb{E}\|W^{-2}\| &= \mathbb{E}\|W^{-2}\| \cdot \mathbf{1}(\mathcal{F}_0) + \mathbb{E}\|W^{-2}\| \cdot \mathbf{1}(\mathcal{F}_0^c) \\ &\leq 4p^{-2} + (\mathbb{E}\|W^{-2}\|^2)^{1/2} \cdot e^{-c_1 p/2} \\ &\leq 4p^{-2} + (\mathbb{E}\|W^{-2}\|_{\mathbb{F}}^2)^{1/2} \cdot e^{-c_1 p/2} \\ &= 4p^{-2} + (\mathbb{E}\text{tr}(W^{-4}))^{1/2} \cdot e^{-c_1 p/2}. \end{aligned}$$

The term $\mathbb{E}\text{tr}(W^{-4})$ can be computed using the Theorem 4.1 of Von Rosen (1988), which implies $\mathbb{E}\text{tr}(W^{-4}) \cdot e^{-c_1 p/2} \leq p^{-2}$.

Lower bound for covariance matrix estimation. We consider a subset Θ_1 of $\Theta(\lambda, \sigma^2)$:

$$\Theta_1 = \left\{ \Sigma = \lambda U U^\top + \sigma^2 I : U \in \mathbb{O}_{p,r} \right\}$$

In this set, both λ and σ^2 are known to us, and it boils down to estimating U . Therefore

$$\begin{aligned} \inf_{\widehat{\Sigma}} \sup_{\Sigma \in \Theta_1} \mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 &= \inf_{\widehat{U} \in \mathcal{M}(\varepsilon, \delta)} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \lambda^2 \cdot \mathbb{E}\|\widehat{U}\widehat{U}^\top - U U^\top\|_{\mathbb{F}}^2 \\ &\geq \frac{c_0 p r}{\sum_{i=1}^m \left(n_j \wedge (n_j^2 \varepsilon_j^2 \cdot p^{-1} r^{-1}) \right)} (\lambda \sigma^2 + \sigma^4) \bigwedge (r \lambda^2). \end{aligned} \quad (47)$$

Now if $\lambda/\sigma^2 \geq 1$, then in addition to (47), we consider another set

$$\Theta_2 = \left\{ \Sigma = \begin{bmatrix} (\lambda + \sigma^2)VV^\top + (\lambda + \sigma^2)I_r & 0 \\ 0 & \sigma^2 I_{p-r} \end{bmatrix} : V \in \mathbb{O}_{r, \frac{r}{2}} \right\}.$$

For any $\Sigma \in \Theta_2$, it admits the following decomposition:

$$\Sigma = \begin{bmatrix} [V & V_\perp] \\ 0 & \end{bmatrix} \text{diag}(\underbrace{2\lambda + \sigma^2, \dots, 2\lambda + \sigma^2}_{\frac{r}{2} \text{ times}}, \lambda, \dots, \lambda) \begin{bmatrix} [V & V_\perp] \\ 0 & \end{bmatrix}^\top + \sigma^2 I,$$

where $V_\perp \in \mathbb{O}_{r, \frac{r}{2}}$ is the orthogonal complement of V . Since $\lambda/\sigma^2 \geq 1$, we can conclude $\Theta_2 \subset \Theta(\lambda, \sigma^2)$. Now the original problem reduces to a smaller one. Define

$$\tilde{\Theta}(\lambda, \sigma^2) = \left\{ \Sigma = V\Lambda V^\top + \sigma^2 I : \right. \\ \left. V \in \mathbb{O}_{r, \frac{r}{2}}, \Lambda = \text{diag}(\lambda_1, \dots, \lambda_r), c_0\lambda \leq \lambda_r \leq \dots \leq C_0\lambda \right\}.$$

Then from (47), we have

$$\begin{aligned} \inf_{\tilde{\Sigma}} \sup_{\Sigma \in \tilde{\Theta}(\lambda + \sigma^2, \lambda + \sigma^2)} \mathbb{E} \|\tilde{\Sigma} - \Sigma\|_{\text{F}}^2 &\geq \left(\frac{c_0 r^2}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot r^{-2}))} (\lambda + \sigma^2)^2 \right) \wedge (r(\lambda + \sigma)^2) \\ &\geq \left(\frac{c_0 r^2 \lambda^2}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot r^{-2}))} \right) \wedge (r\lambda^2) \end{aligned}$$

Note that the estimation of V in $\tilde{\Theta}_2$ is a sub-problem of estimating $\begin{bmatrix} [V & V_\perp] \\ 0 & \end{bmatrix}$ in Θ_2 , we have

$$\begin{aligned} \inf_{\hat{\Sigma}} \sup_{\Sigma \in \Theta_2} \mathbb{E} \|\hat{\Sigma} - \Sigma\|_{\text{F}}^2 &\geq \inf_{\tilde{\Sigma}} \sup_{\Sigma \in \tilde{\Theta}(\lambda + \sigma^2, \lambda + \sigma^2)} \mathbb{E} \|\tilde{\Sigma} - \Sigma\|_{\text{F}}^2 \\ &\geq \left(\frac{c_0 r^2 \lambda^2}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot r^{-2}))} \right) \wedge (r\lambda^2). \end{aligned}$$

Together with the bound in (47), we conclude when $\lambda/\sigma^2 \geq 1$,

$$\begin{aligned} \inf_{\hat{\Sigma}} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\hat{\Sigma} - \Sigma\|_{\text{F}}^2 &\geq \inf_{\hat{\Sigma}} \sup_{\Sigma \in \Theta_1 \cup \Theta_2} \mathbb{E} \|\hat{\Sigma} - \Sigma\|_{\text{F}}^2 \\ &\geq \left(\left(\frac{c_0 p r}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot d^{-1} r^{-1}))} (\lambda \sigma^2 + \sigma^4) \right) \wedge (r\lambda^2) \right) \vee \left(\left(\frac{c_0 r^2 \lambda^2}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot r^{-2}))} \right) \wedge (r\lambda^2) \right). \end{aligned}$$

Notice when $\lambda/\sigma^2 \leq 1$,

$$\left(\frac{c_0 dr}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot p^{-1} r^{-1}))} (\lambda \sigma^2 + \sigma^4) \right) \wedge (r \lambda^2) \geq \left(\frac{c_0 r^2 \lambda^2}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot r^{-2}))} \right) \wedge (r \lambda^2).$$

Therefore we conclude for any λ, σ^2 satisfy the condition in the theorem,

$$\begin{aligned} & \inf_{\widehat{\Sigma}} \sup_{\Sigma \in \Theta(\lambda, \sigma^2)} \mathbb{E} \|\widehat{\Sigma} - \Sigma\|_{\mathbb{F}}^2 \\ & \geq \left(\left(\frac{c_0 pr}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot p^{-1} r^{-1}))} (\lambda \sigma^2 + \sigma^4) \right) \wedge (r \lambda^2) \right) \vee \left(\left(\frac{c_0 r^2 \lambda^2}{\sum_{i=1}^m (n_j \wedge (n_j^2 \varepsilon_j^2 \cdot r^{-2}))} \right) \wedge (r \lambda^2) \right). \end{aligned}$$

This finishes the proof of Theorem 3.

A.5 Proof of Lemma 3

We use $x = \{X_i^{(j)}, i = 1, \dots, n_j\}_{j=1}^m$ to represent the collection of all data, $X^{(j)} = \{X_i^{(j)}, i = 1, \dots, n_j\}$, and Θ to be the parameter. Condition on $\{\widehat{U}_j\}_{j=1}^m$, we define the random matrices

$$\begin{aligned} A &= \widehat{U} \widehat{U}^\top - \psi(\Theta), \\ B_{ij} &= \sum_{(k,l) \in [p] \times [r]} \frac{\partial}{\partial \Theta_{kl}} ([\nabla \psi_{ij}(\Theta)]_{kl} \cdot p(x, \Theta | \{\widehat{U}_j\}_{j=1}^m) \cdot p(\Theta)) \frac{1}{p(x, \Theta | \{\widehat{U}_j\}_{j=1}^m) p(\Theta)}, \end{aligned}$$

where $p(x, \Theta | \{\widehat{U}_j\})$ is the conditional density with parameter Θ and we have

$$p(x, \Theta | \{\widehat{U}_j\}) = \prod_{j=1}^m p(X^{(j)}, \Theta | \widehat{U}_j).$$

Now we define the conditional expectation $\mathbb{E}[\langle A, B \rangle | \{\widehat{U}_j\}] = \int \int \langle A, B \rangle p(x, \Theta | \{\widehat{U}_j\}) p(\Theta) d\Theta dx$.

Then using Cauchy-Schwarz inequality, we see

$$\mathbb{E}[\|A\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}] \geq \frac{(\mathbb{E}[\langle A, B \rangle | \{\widehat{U}_j\}])^2}{\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}]} \quad (48)$$

Simple calculation shows

$$\begin{aligned} \mathbb{E}[\langle A, B \rangle | \{\widehat{U}_j\}] &= \int \int \sum_{ij} [\widehat{U} \widehat{U}^\top - \psi(\Theta)]_{ij} \cdot \sum_{k,l} \frac{\partial}{\partial \Theta_{kl}} ([\nabla \psi_{ij}(\Theta)]_{kl} p(x, \Theta | \{\widehat{U}_j\}) p(\Theta)) d\Theta dx \\ &= \int \int \sum_{ij,kl} [\nabla \psi_{ij}(\Theta)]_{kl}^2 \cdot p(\Theta) d\Theta dx \\ &= \int \text{tr}(\nabla \psi(\Theta) \circ \nabla \psi(\Theta)^*) p(\Theta) d\Theta, \end{aligned}$$

where the second equality holds from integration by parts, $\int p(x, \Theta | \{\widehat{U}_j\}) dx = 1$ and $\sum_{ij,kl} [\nabla \psi_{ij}(\Theta)]_{kl}^2 = \text{tr}(\nabla \psi(\Theta) \circ \nabla \psi(\Theta)^*)$. Meanwhile,

$$\mathbb{E}[\|A\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}] = \int \|\widehat{U} \widehat{U}^\top - \psi(\Theta)\|_{\mathbb{F}}^2 \cdot p(\Theta) d\Theta.$$

Notice the right hand side is still a function of $\{\widehat{U}_j\}$. Next we consider the expectation $\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}]$:

$$\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}] = \mathbb{E} \sum_{ij} \left(\Delta \psi_{ij}(\Theta) + \langle \nabla \psi_{ij}(\Theta), \nabla \log p(x, \Theta | \{\widehat{U}_j\}) + \nabla \log p(\Theta) \rangle \right)^2.$$

Since $\psi(\Theta)$ is independent of $\{\widehat{U}_j\}$,

$$\mathbb{E} \sum_{ij} \Delta \psi_{ij}^2(\Theta) = \int \sum_{ij} \Delta \psi_{ij}^2(\Theta) p(\Theta) d\Theta.$$

Notice

$$\int \nabla \log p(X^{(j)}, \Theta | \widehat{U}_j) \cdot p(X^{(j)}, \Theta | \widehat{U}_j) dx = \nabla \left[\int p(X^{(j)}, \Theta | \widehat{U}_j) dx \right] = 0, \quad (49)$$

where the last equality is due to $\int p(X^{(j)}, \Theta | \widehat{U}_j) dx = 1$. Thus, $\mathbb{E} \nabla \log p(x, \Theta | \{\widehat{U}_j\}) = 0$. Also notice $\langle \nabla \psi_{ij}(\Theta), \nabla \log p(\Theta) \rangle = 0$. Therefore

$$\begin{aligned} & \mathbb{E} \sum_{ij} \langle \nabla \psi_{ij}(\Theta), \nabla \log p(x, \Theta | \{\widehat{U}_j\}) + \nabla \log p(\Theta) \rangle^2 \\ &= \mathbb{E} \sum_{ij} \langle \nabla \psi_{ij}(\Theta), \nabla \log p(x, \Theta | \{\widehat{U}_j\}) \rangle^2 \\ &= \mathbb{E} \sum_{ij} \langle \nabla \psi_{ij}(\Theta), \sum_{k=1}^m \nabla \log p(X^{(j)}, \Theta | \widehat{U}_k) \rangle^2 \\ &= \sum_{k=1}^m \mathbb{E} \sum_{ij} \langle \nabla \psi_{ij}(\Theta), \nabla \log p(X^{(j)}, \Theta | \widehat{U}_k) \rangle^2, \end{aligned}$$

where in the last line the cross terms vanish due to (49). Recall

$$\mathcal{I}(\Theta | \widehat{U}_j) = \mathbb{E}[(\nabla \log p(X^{(j)}; \Theta | \widehat{U}_j))^* \circ \nabla \log p(X^{(j)}; \Theta | \widehat{U}_j)].$$

Therefore

$$\mathbb{E} \sum_{ij} \langle \nabla \psi_{ij}(\Theta), \nabla \log p(X^{(j)}, \Theta | \widehat{U}_k) \rangle^2 = \int \text{tr}(\nabla \psi(\Theta)^* \circ \mathcal{I}(\Theta | \widehat{U}_k) \circ \nabla \psi(\Theta)) \cdot p(\Theta) d\Theta$$

Using (49) again, we obtain

$$\mathbb{E} \sum_{ij} \Delta \psi_{ij}(\Theta) \cdot \langle \nabla \psi_{ij}(\Theta), \nabla \log p(x, \Theta | \{\widehat{U}_j\}) \rangle = 0.$$

So we conclude

$$\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}] = \int \sum_{ij} \Delta \psi_{ij}^2(\Theta) p(\Theta) d\Theta + \sum_{j=1}^m \int \text{tr}(\nabla \psi(\Theta) \circ \mathcal{I}(\Theta | \widehat{U}_j) \circ \nabla \psi(\Theta)^*) \cdot p(\Theta) d\Theta.$$

Now taking expectation w.r.t. \widehat{U}_j in (48), and using Jensen's inequality yield the desired result

$$\begin{aligned} \int \mathbb{E}[\|\widehat{U}\widehat{U}^\top - \psi(\Theta)\|_{\mathbb{F}}^2 \cdot p(\Theta) d\Theta] &\geq \mathbb{E} \frac{(\mathbb{E}[\langle A, B \rangle | \{\widehat{U}_j\}])^2}{\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}]} \\ &= \mathbb{E} \frac{\left(\int \text{tr}(\nabla \psi(\Theta) \circ \nabla \psi(\Theta)^*) p(\Theta) d\Theta \right)^2}{\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}]} \\ &\geq \frac{\left(\int \text{tr}(\nabla \psi(\Theta) \circ \nabla \psi(\Theta)^*) p(\Theta) d\Theta \right)^2}{\mathbb{E}[\mathbb{E}[\|B\|_{\mathbb{F}}^2 | \{\widehat{U}_j\}]]}. \end{aligned}$$

B Technical Lemma

Lemma 4. Let $g_1, \dots, g_m \stackrel{i.i.d.}{\sim} N(0, 1)$, and $c_1, \dots, c_m \geq 0$. Then for any integer $l \geq 0$, we have

$$\mathbb{E}\left(\sum_{i=1}^m c_i g_i^2\right)^l \leq (Cl \cdot \sum_{i=1}^m c_i)^l$$

for some absolute constant $C > 0$.

Proof. We show this by expanding $(\sum_{i=1}^m c_i g_i^2)^l$. In fact, we have

$$\begin{aligned} \mathbb{E}\left(\sum_{i=1}^m c_i g_i^2\right)^l &= \mathbb{E} \sum_{i_1, \dots, i_l=1}^m c_{i_1} \cdots c_{i_l} g_{i_1}^2 \cdots g_{i_l}^2 \\ &\leq \left(\sum_i c_i\right)^l \cdot \mathbb{E} g^{2l} \leq (Cl)^l \cdot \left(\sum_i c_i\right)^l, \end{aligned}$$

where $g \sim N(0, 1)$ and we use the moment bound for Gaussian in the last inequality. \square

Lemma 5 (Koltchinskii and Lounici (2017)). Let X_1, \dots, X_n be i.i.d. samples from $N(0, \Sigma)$, and $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$. Then

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\frac{\widetilde{r}}{n} \vee \sqrt{\frac{\widetilde{r}}{n}} \right),$$

where $\widetilde{r} = \frac{\text{tr}(\Sigma)}{\|\Sigma\|}$ is the effective rank of Σ . Moreover, there exists an absolute constant $C_1 > 0$, such that for all $t \geq 1$, with probability exceeding $1 - e^{-t}$,

$$\left| \|\widehat{\Sigma} - \Sigma\| - \mathbb{E}\|\widehat{\Sigma} - \Sigma\| \right| \leq C_1 \|\Sigma\| \left(\frac{t}{n} \vee \sqrt{\frac{t}{n}} \right).$$

Lemma 6. Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with $\mathbb{E}X = 0$, and denote $\|X\|_{\psi_2}$ its ψ_2 norm. Then we have for any $t > 0$,

$$\mathbb{P}\left(\|X\|_{\ell_2} \geq t\right) \leq 4^d \exp\left(-\frac{Ct^2}{\|X\|_{\psi_2}^2}\right)$$

for some absolute constant $C > 0$.

Proof. Let $\{x_i\}_{i=1}^N$ be an $1/2$ cover of the unit sphere \mathbb{S}^{d-1} , then $N \leq 4^d$. Notice $\|X\|_{\ell_2} = \langle X, \frac{X}{\|X\|_{\ell_2}} \rangle$. Then there exists some $x_0 \in \{x_i\}_{i=1}^N$, such that $\|x_0 - X/\|X\|_{\ell_2}\|_{\ell_2} \leq 1/2$. Now

$$\|X\|_{\ell_2} = \langle X, \frac{X}{\|X\|_{\ell_2}} \rangle = \langle X, \frac{X}{\|X\|_{\ell_2}} - x_0 \rangle + \langle X, x_0 \rangle \leq \frac{1}{2} \|X\|_{\ell_2} + \langle X, x_0 \rangle.$$

This implies $\|X\|_{\ell_2} \leq 2\langle X, x_0 \rangle$. We conclude

$$\mathbb{P}(\|X\|_{\ell_2} \geq t) \leq 4^d \cdot \mathbb{P}(\langle X, x_0 \rangle \geq t/2) \leq 4^d \exp\left(-\frac{Ct^2}{\|X\|_{\psi_2}^2}\right).$$

\square

Lemma 7. Let X be random variable such that

$$\mathbb{P}(|X| \geq \max\{a + bt, \sqrt{a + bt}\}) \leq e^{-t}$$

for some $1 > a > 0, b > 0$ and all $t > 0$, then we have

$$\begin{aligned}\mathbb{E}|X|^2 &\leq a + b + 2ab + 2b^2, \\ \mathbb{E}|X|^4 &\leq a^2 + 2ab + 2b^2 + 16a^3b + 96b^4.\end{aligned}$$

Proof. We have

$$\mathbb{E}|X|^2 = \int_0^{+\infty} \mathbb{P}(|X|^2 \geq s) ds = \int_0^{+\infty} \mathbb{P}(|X| \geq \sqrt{s}) 2s ds.$$

We then decompose the integral into three parts:

$$\int_0^{\infty} = \int_0^{\sqrt{a}} + \int_{\sqrt{a}}^1 + \int_1^{\infty}. \quad (50)$$

For the first part, we have

$$\int_0^{\sqrt{a}} \mathbb{P}(|X| \geq s) 2s ds \leq a.$$

For the second part, we have

$$\begin{aligned}\int_{\sqrt{a}}^1 \mathbb{P}(|X| \geq s) 2s ds &= \int_0^{\frac{1-a}{b}} \mathbb{P}(|X| \geq \sqrt{a + bt}) 2(a + bt)^{1/2} \frac{1}{2} (a + bt)^{-1/2} b dt \\ &\leq b \int_0^{\frac{1-a}{b}} e^{-t} dt \leq b.\end{aligned}$$

For the third part, we have

$$\begin{aligned}\int_1^{\infty} \mathbb{P}(|X| \geq s) 2s ds &= \int_{\frac{1-a}{b}}^{\infty} \mathbb{P}(|X| \geq a + bt) 2(a + bt) b dt \\ &\leq 2b \int_{\frac{1-a}{b}}^{\infty} e^{-t} (a + bt) dt \\ &\leq 2ab + 2b^2\end{aligned}$$

For the fourth order moment, we have similarly

$$\mathbb{E}|X|^4 = \int_0^{+\infty} \mathbb{P}(|X|^4 \geq s) ds = \int_0^{+\infty} \mathbb{P}(|X| \geq \sqrt[4]{s}) 4s^3 ds.$$

Using the decomposition as in (50), we have for the first part,

$$\int_0^{\sqrt{a}} \mathbb{P}(|X| \geq s) 4s^3 ds \leq a^2.$$

For the second part, we have

$$\begin{aligned}
\int_{\sqrt{a}}^1 \mathbb{P}(|X| \geq s) 4s^3 ds &= \int_0^{\frac{1-a}{b}} \mathbb{P}(|X| \geq \sqrt{a+bt}) 4(a+bt)^{3/2} \frac{1}{2} (a+bt)^{-1/2} b dt \\
&\leq 2b \int_0^{\frac{1-a}{b}} e^{-t} (a+bt) dt \\
&\leq 2ab + 2b^2.
\end{aligned}$$

For the third part, we have

$$\begin{aligned}
\int_1^\infty \mathbb{P}(|X| \geq s) 4s^3 ds &= \int_{\frac{1-a}{b}}^\infty \mathbb{P}(|X| \geq a+bt) 4(a+bt)^3 b dt \\
&\leq 4b \int_{\frac{1-a}{b}}^\infty e^{-t} (a+bt)^3 dt \\
&\leq 16a^3 b + 96b^4.
\end{aligned}$$

□

C Some Linear Algebras

C.1 Derivation for $\nabla\psi(\Theta)$

Let the map $\psi : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{p \times p}$ be defined as $\psi(\Theta) = \Theta(\Theta^\top \Theta)^{-1} \Theta^\top$. Then the gradient of ψ evaluated at Θ is a linear map: $\nabla\psi(\Theta) : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{p \times p}$. We set

$$\begin{aligned}
\psi_1 : \mathbb{R}^{p \times r} &\rightarrow \mathbb{R}^{p \times r} \text{ as } \psi_1(\Theta) = \Theta, \\
\psi_2 : \mathbb{R}^{p \times r} &\rightarrow \mathbb{R}^{r \times r} \text{ as } \psi_2(\Theta) = (\Theta^\top \Theta)^{-1}, \\
\psi_3 : \mathbb{R}^{p \times r} &\rightarrow \mathbb{R}^{r \times p} \text{ as } \psi_3(\Theta) = \Theta^\top.
\end{aligned}$$

Then using product rule, we have for any $Y \in \mathbb{R}^{p \times r}$,

$$\begin{aligned}
\nabla\psi(\Theta)(Y) &= \nabla\psi_1(\Theta)(Y) \cdot \psi_2(\Theta) \cdot \psi_3(\Theta) + \psi_1(\Theta) \cdot \nabla\psi_2(\Theta)(Y) \cdot \psi_3(\Theta) \\
&\quad + \psi_1(\Theta) \cdot \psi_2(\Theta) \cdot \nabla\psi_3(\Theta)(Y).
\end{aligned}$$

Notice here $\nabla\psi_1(\Theta) : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{p \times r}$ is defined as $\nabla\psi_1(\Theta)(Y) = Y$ and $\nabla\psi_3(\Theta) : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{r \times p}$ is defined as $\nabla\psi_3(\Theta)(Y) = Y^\top$. Now we compute $\nabla\psi_2(\Theta)$. Following definition of gradient, $\nabla\psi_2(\Theta) : \mathbb{R}^{p \times r} \rightarrow \mathbb{R}^{r \times r}$. We set $\psi_{2,1}(\Theta) = \Theta^\top \Theta$, and $\psi_{2,2}(M) = M^{-1}$. Then using product rule,

$$\nabla\psi_{2,1}(\Theta)(Y) = Y^\top \Theta + \Theta^\top Y.$$

We also define $\psi_{2,3}(M) = M$. $\nabla\psi_{2,2}(M) : \mathbb{R}^{r \times r} \rightarrow \mathbb{R}^{r \times r}$ can be calculated using product rule. Notice for any $N \in \mathbb{R}^{r \times r}$,

$$0 = \nabla(\psi_{2,2} \cdot \psi_{2,3})(M)(N) = \nabla\psi_{2,2}(M)(N) \cdot M + M^{-1} \cdot N,$$

which implies $\nabla\psi_{2,2}(M)(N) = -M^{-1}NM^{-1}$. Notice $\psi_2(\Theta) = \psi_{2,2} \circ \psi_{2,1}(\Theta)$. Using chain rule, we have

$$\begin{aligned} \nabla\psi_2(\Theta)(Y) &= \nabla(\psi_{2,2} \circ \psi_{2,1})(\Theta)(Y) = \nabla\psi_{2,2}(\psi_{2,1}(\Theta))(\nabla\psi_{2,1}(\Theta)(Y)) \\ &= \nabla\psi_{2,2}(\Theta^\top \Theta)(Y^\top \Theta + \Theta^\top Y) \\ &= -(\Theta^\top \Theta)^{-1}(Y^\top \Theta + \Theta^\top Y)(\Theta^\top \Theta)^{-1}. \end{aligned}$$

We have

$$\psi_1(\Theta) \cdot \nabla\psi_2(\Theta)(Y) \cdot \psi_3(\Theta) = -\Theta(\Theta^\top \Theta)^{-1}(Y^\top \Theta + \Theta^\top Y)(\Theta^\top \Theta)^{-1}\Theta^\top.$$

In summary, we have

$$\begin{aligned} \nabla\psi(\Theta)(Y) &= Y(\Theta^\top \Theta)^{-1}\Theta^\top - \Theta(\Theta^\top \Theta)^{-1}(Y^\top \Theta + \Theta^\top Y)(\Theta^\top \Theta)^{-1}\Theta^\top + \Theta(\Theta^\top \Theta)^{-1}Y^\top \\ &= \bar{\Theta}_\perp \bar{\Theta}_\perp^\top Y(\Theta^\top \Theta)^{-1}\Theta^\top + \Theta(\Theta^\top \Theta)^{-1}Y^\top \bar{\Theta}_\perp \bar{\Theta}_\perp^\top, \end{aligned}$$

where $\bar{\Theta}_\perp \bar{\Theta}_\perp^\top = I_p - \Theta(\Theta^\top \Theta)^{-1}\Theta^\top$.

C.2 Derivation for $\nabla\psi(\Theta)^*$

Once we obtain the closed-form for $\nabla\psi(\Theta)$, we can compute its adjoint $\nabla\psi(\Theta)^* : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p \times r}$.

For any $Y \in \mathbb{R}^{p \times r}$, $M \in \mathbb{R}^{p \times p}$, we have

$$\begin{aligned} \langle \nabla\psi(\Theta)^*(M), Y \rangle &= \langle M, \nabla\psi(\Theta)(Y) \rangle \\ &= \langle M, \bar{\Theta}_\perp \bar{\Theta}_\perp^\top Y(\Theta^\top \Theta)^{-1}\Theta^\top + \Theta(\Theta^\top \Theta)^{-1}Y^\top \bar{\Theta}_\perp \bar{\Theta}_\perp^\top \rangle \\ &= \langle \bar{\Theta}_\perp \bar{\Theta}_\perp^\top (M + M^\top) \Theta(\Theta^\top \Theta)^{-1}, Y \rangle. \end{aligned}$$

So we conclude $\nabla\psi(\Theta)^*(M) = \bar{\Theta}_\perp \bar{\Theta}_\perp^\top (M + M^\top) \Theta(\Theta^\top \Theta)^{-1}$.