Data-dependent Confidence Regions of Singular Subspaces

Dong Xia
Department of Mathematics
Hong Kong University of Science and Technology

(January 3, 2019)

Abstract

Matrix singular value decomposition (SVD) is popular in statistical data analysis which shows superior efficiency of extracting the unknown low-rank singular subspaces embedded in noisy matrix observations. This article is about the statistical inference for the singular subspaces when the noise matrix has i.i.d. entries and our goal is to construct data-dependent confidence regions of the unknown singular subspaces from one single noisy matrix observation.

We derive an explicit representation formula for the empirical spectral projectors. The formula is neat and holds for deterministic matrix perturbations. Based on this representation formula, we calculate, up to the fourth-order approximations, the expected joint projection distance between the empirical singular subspace and the true singular subspace. Then, we prove the normal approximation of the joint projection distance with an explicit normalization factor and different levels of bias corrections. In particular, with the fourth-order bias corrections, we show that the asymptotical normality holds under the signal-to-noise ration (SNR) condition $O((d_1+d_2)^{9/16})$ where $d_1$ and $d_2$ denote the matrix sizes. An intriguing observation is that the higher order bias terms are negligible when $d_1 \approx d_2$ so that $|d_1 - d_2| = O((d_1 + d_2)^{1/2})$. We will propose
a shrinkage estimator of the singular values by borrowing recent results from random matrix theory. Equipped with these estimators, we introduce data-dependent centering and normalization factors for the joint projection distance between the empirical singular subspace and the true singular subspace. We will prove such data-dependent asymptotical normality under SNR condition $O((d_1 + d_2)^{9/16})$. Therefore, we are able to construct data-dependent confidence regions for the true singular subspaces which attains the pre-determined confidence levels asymptotically. The convergence rates of the asymptotical normality are also presented. Finally, we provide comprehensive simulation results to illustrate our theoretical discoveries.

1 Introduction

Matrix singular value decomposition (SVD) is a powerful tool for various purposes across diverse fields. In numerical linear algebra, SVD has been successfully applied for solving linear inverse problems, low-rank matrix approximation, to name a few. See, e.g., Golub and Van Loan (2012), for more examples. For numerous machine learning tasks, SVD is crucial for designing computationally efficient algorithms, such as matrix and tensor completion (Cai et al. (2010), Keshavan et al. (2010), Candes and Tao (2010), Xia and Yuan (2018+), Xia et al. (2017)), and phase retrieval (Ma et al. (2017), Candès et al. (2015)), where SVD is also employed for generating warm initial points for non-convex optimization algorithms. In statistical data analysis, SVD is superior for the matrix denoising and dimension reduction. For instance in Kim et al. (2005), as an intermediate step for dimension reduction, SVD was successfully applied for text classification. See also Li and Wang (2007). In Shabalin and Nobel (2013), SVD delivered appealing performances in low rank matrix denoising. More specifically, in Donoho and Gavish (2014), they proved the precise singular value thresholds for the statistically optimal denoising of noisy matrix data. Recently, matrix SVD is generalized to tensor SVD for low rank tensor denoising, see Xia and Zhou (2017) and Zhang and Xia (2018).

Characterizing its perturbation is critical for advancing the further developments of SVD.
driven methodologies for analyzing noisy matrix observations. In this case, the observed data matrix often equals a low-rank information matrix plus a noise matrix. The deterministic perturbation bounds of matrix SVD have been well established by Davis-Kahan (Davis and Kahan (1970)) and Wedin (Wedin (1972)) many decades ago. Generally speaking, those deterministic perturbation bounds demonstrated that the perturbation of singular subspaces is governed by the so-called signal-to-noise ratio (SNR) where ”signal” refers to the smallest non-zero singular value of the information matrix and the ”noise” refers to the spectral norm of the noise matrix. The result is quite general because these bounds do not rely on the wellness of alignments between the singular subspaces of the information and of the noise matrices. These general perturbation bounds turn out to be often satisfactorily sharp in statistical data analysis where the additive noise matrix is usually assumed to contain i.i.d. random entries. However, more refined characterizations of the empirical singular subspaces are demanded on the frontiers of statistical analysis of SVD for noisy matrix observations. Indeed, the Davis-Kahan Theorem and Wedin’s perturbation bounds are described by the non-zero smallest singular value of the information matrix, whereas the effects of those larger singular values are missing. Moreover, the exact numerical factor is also not well recognized so far. These important questions will be answered in this article with the ingredients from random matrix theory (RMT).

The behaviour of singular values and singular subspaces of low rank perturbations of large rectangular random matrices is popular in the recent years among the RMT community. The asymptotic limits of singular values and singular subspaces were initially established by Benaych-Georges and Nadakuditi (2012), where the convergence rate of the largest singular value is also provided. Recently, in Ding (2017), the more precise non-asymptotical concentration bounds for those empirical singular values were developed. Meanwhile, Ding (2017) and Xia and Zhou (2017) established the non-asymptotic perturbation bounds of the empirical singular vectors for each simple singular values with multiplicity 1. In a recent work Bao et al. (2018), the authors investigated the asymptotical limit distributions of the empirical singular subspaces. Specifically, they showed that if the noise matrix has i.i.d.
Gaussian entries, then the limit distribution of the projection distance between the empirical singular subspaces and the true singular subspaces is also Gaussian. In [Xia, 2018], the low rank matrix regression model was studied where the author proposed a de-biased estimator built on the nuclear norm penalized least squares estimator. The de-biased estimator ends up with an analogous form of the low rank perturbation of rectangular random matrices. Then, the non-asymptotic normal approximation theories for the projection distance between the empirical singular subspace and the true singular subspace are developed, under nearly optimal sample size requirements. The paramount observation is that, in the limit normal distributions, the mean values are significantly larger than the corresponding standard deviations. As a result, in order to construct data-dependent confidence regions of the singular subspaces, a much larger than regular sample size requirement is necessary to tradeoff the errors of estimating the mean values. Most recently, [Chen et al., 2018] revealed an interesting phenomenon of the eigenvalues and eigenvectors in the similar random perturbation models, showing that the perturbation of eigen structures is much smaller than the perturbation of singular structures. In addition, non-asymptotic perturbation bounds of the linear forms of empirical singular vectors can be found in [Koltchinskii and Xia, 2016]. General entry-wise perturbation bounds when the true singular subspaces are incoherent are proved in [Abbe et al., 2017]. The one-sided minimax optimal perturbation bounds of the singular subspaces for low rank perturbations of large rectangular random matrices were also established in [Cai and Zhang, 2018] and [Zheng and Tomioka, 2015].

Our goal in this article is to construct data-dependent confidence regions of the singular subspaces in the low rank perturbation model of high-dimensional random matrices. Confidence regions are crucial for the statistical inference. Unlike the nicely established perturbation bounds of the empirical singular subspaces in the literature, the methods of designing data-dependent confidence regions of the singular subspaces is much less known. As elucidated in [Xia, 2018], the major difficulty arises from how to precisely determine the mean value of the projection distance between the empirical singular subspaces and true singular subspaces. One conclusive contribution of this article is providing an explicit repre-
sentation formula of the empirical spectral projectors corresponding to the empirical singular subspaces. The representation formula is neat and holds also for deterministic matrix perturbations. It allows us to reach much more precise calculations of the expected projection distance between the empirical singular subspaces and the true singular subspaces. As a consequence, we are able to prove the asymptotical normality with higher order bias corrections under nearly minimal requirements of SNR. Based on the asymptotical normality, we can construct data-dependent confidence regions of the true singular subspaces from one single noisy matrix observation.

In order to better present our results and highlight our contributions, let’s begin with the standard notations which shall appear throughout the article. We denote $M = U \Lambda V^T$ the unknown $d_1 \times d_2$ matrix where $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$ are its left and right singular vectors. The rank of $M$ is denoted by $r$ which is much smaller than $d_1$ and $d_2$. The diagonal matrix $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r)$ consists of $M$’s non-increasing positive singular values. The observed data matrix $\hat{M} \in \mathbb{R}^{d_1 \times d_2}$ is contaminated with additive Gaussian noise:

$$\hat{M} = M + Z \quad \text{where} \quad Z_{j_1 j_2} \sim \mathcal{N}(0, 1) \quad \text{for} \quad 1 \leq j_1 \leq d_1, 2 \leq j_2 \leq d_2. \quad (1.1)$$

Here in (1.1), we fix the noise variance to be 1, just for simplicity. As a result, the signal-to-noise ratio (SNR) is completely determined by the signal strength $\lambda_r$. Let $\hat{U} \in \mathbb{R}^{d_1 \times r}$ and $\hat{V} \in \mathbb{R}^{d_2 \times r}$ be the top-$r$ left and right singular vectors of $\hat{M}$. Let $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ denotes the top-$r$ singular values of $\hat{M}$. Given $\hat{M}$, our ultimate goal is to construct data-dependent confidence regions of $U$ and $V$. The loss function we will focus on is the joint projection distance between the empirical singular subspaces and the true singular subspaces. To be exact, the joint projection distance is defined by

$$\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] := \|\hat{U}U^T - UU^T\|_F^2 + \|\hat{V}V^T - VV^T\|_F^2, \quad (1.2)$$

where $\| \cdot \|_F$ denotes the matrix Frobenius norm. By Davis-Kahan Theorem (Davis and Kahan (1970)) or Wedin’s sin $\Theta$ theorem (Wedin (1972)), $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ is non-trivial only on the event $\{\lambda_r > 2\|Z\|\}$ where $\| \cdot \|$ denotes the matrix spectral norm. It is well-known by RMT that $\|Z\| = O_P(d_{\text{max}}^{1/2})$ where $d_{\text{max}} = \max\{d_1, d_2\}$. As a result, the lower bound of
SNR requirement is $\lambda_r \gg d_{\text{max}}^{1/2}$, which is the minimal requirement on the signal strength. See also, e.g., Tao (2012), Koltchinskii and Xia (2016), Cai and Zhang (2018), Vershynin (2010) and references therein.

To construct confidence regions of the true singular subspaces, our strategy is to investigate the distribution of the joint projection distance $\text{dist}^2((\hat{U}, \hat{V}), (U, V))$ defined as in (1.2). In this article, our contributions can be summarized as follows.

1. We derive an explicit representation formula of the empirical spectral projectors $\hat{U}\hat{U}^T$ and $\hat{V}\hat{V}^T$. In particular, $\hat{U}\hat{U}^T$ and $\hat{V}\hat{V}^T$ are completely determined by a sum of a series of matrix products involving only $\Lambda, UU^T, U_\perp U_\perp^T, VV^T, V_\perp V_\perp^T$ and $Z$, where $U_\perp \in \mathbb{R}^{d_1 \times (d_1 - r)}$ and $V_\perp \in \mathbb{R}^{d_2 \times (d_2 - r)}$ are chosen so that $(U, U_\perp)$ and $(V, V_\perp)$ are orthonormal matrices, respectively. To derive this useful representation formula, we will apply the Reisz formula, combinatorics formulas, contour integrals, the residue theorem and the generalized Leibniz rule. It worths to point out that the representation formula is deterministic which holds for any noise matrix $Z$ as long as $\|Z\| < \lambda_r/2$. We believe that this representation formula of the empirical spectral projectors should be of independent interest for various of purposes.

2. By the representation formula of the empirical spectral projectors $\hat{U}\hat{U}^T$ and $\hat{V}\hat{V}^T$, we will study the normal approximation of $\hat{\varepsilon}_1 := \frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - \text{dist}^2((\hat{U}, \hat{V}), (U, V))}{\sqrt{8d_\star \|\Lambda^{-2}\|_F}}$ where $d_\star = d_1 + d_2 - 2r$. In particular, we show that $\hat{\varepsilon}_1$ converges to a standard normal distribution as long as $\frac{(rd_{\text{max}})^{1/2}}{\lambda_r} \to 0$ when $d_{\text{max}} \to \infty$. Here, we denote by $d_{\text{max}} = \max\{d_1, d_2\}$. In the case that $r = O(1)$, the aforementioned SNR requirement for the asymptotical normality of $\hat{\varepsilon}_1$ is optimal. The proof strategy is based on the Gaussian isoperimetric inequality and the Berry-Esseen theorem.

3. Note that the centering term in $\hat{\varepsilon}_1$ is the unknown expected joint projection distance $\mathbb{E}\text{dist}^2((\hat{U}, \hat{V}), (U, V))$, which can be viewed as the bias of $\text{dist}^2((\hat{U}, \hat{V}), (U, V))$. To provide with user-friendly normal approximations of $\text{dist}^2((\hat{U}, \hat{V}), (U, V))$, it is important to derive more transparent descriptions of the expected projection distance...
\( \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] \). Thanks to the representation formula of \( \hat{U} \hat{U}^\top \) and \( \hat{V} \hat{V}^\top \), we are able to obtain approximations of \( \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] \). In this article, we present four different levels of approximating \( \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] \) which ends up with different levels of bias corrections. These four levels of approximating \( \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] \) are as follows.

(a) First-order approximation: \( B_1 = 2d_\ast \|\Lambda^{-1}\|_F^2 \) where \( d_\ast = d_{1-} + d_{2-} \) with \( d_{1-} = d_1 - r \) and \( d_{2-} = d_2 - r \). Then, the approximation error is

\[
\left| \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] - B_1 \right| = O\left( \frac{rd_{\text{max}}^2}{\lambda_r^4} \right).
\]

(b) Second-order approximation: \( B_2 = 2(d_\ast \|\Lambda^{-1}\|_F^2 - \Delta_d^2 \|\Lambda^{-2}\|_F^2) \) where \( \Delta_d = d_{1-} - d_{2-} \). Then, the approximation error is

\[
\left| \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] - B_2 \right| = O\left( \frac{rd_{\text{max}}^3}{\lambda_r^8} \right).
\]

(c) Third-order approximation: \( B_3 = 2(d_\ast \|\Lambda^{-1}\|_F^2 - \Delta_d^2 \|\Lambda^{-2}\|_F^2 + d_\ast \Delta_d^2 \|\Lambda^{-3}\|_F^2) \). Then, the approximation error is

\[
\left| \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] - B_3 \right| = O\left( \frac{rd_{\text{max}}^4}{\lambda_r^{10}} \right).
\]

(d) Fourth-order approximation:

\[
B_4 = 2(d_\ast \|\Lambda^{-1}\|_F^2 - \Delta_d^2 \|\Lambda^{-2}\|_F^2 + d_\ast \Delta_d^2 \|\Lambda^{-3}\|_F^2 - (d_{1-} + d_{1-}d_{2-} + d_{2-}^2) \Delta_d^2 \|\Lambda^{-4}\|_F^2).
\]

Then, the approximation error is

\[
\left| \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)] - B_4 \right| = O\left( \frac{rd_{\text{max}}^5}{\lambda_r^{10}} \right).
\]

All the above approximation errors are established under the minimal SNR condition \( \lambda_r \gg d_{\text{max}}^{1/2} \). An intriguing phenomenon is that if \( |d_1 - d_2| = O(d_{\text{max}}^{1/2}) \), i.e., the two dimensions of \( M \) are comparable, then the second-order, third-order and fourth-order approximations have the same effects as the first-order approximation with respect to the SNR conditions. Simulation results in Section 5 indeed confirm that the first-order approximation by \( B_1 \) is satisfactorily accurate when \( d_1 = d_2 \).
4. We then replace the expected joint projection distance \( \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) with the approximations \( B_1, B_2, B_3 \) or \( B_4 \) in the definition of \( \hat{\varepsilon}_1 \), and prove the normal approximation of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). Different levels of bias corrections will end up with different levels of SNR requirements to ensure the asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). For instance, we will investigate the non-asymptotic normal approximation of \( \hat{\varepsilon}_2 := \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_4}{\sqrt{8d_\varepsilon \| \Lambda^{-2} \|_F}} \) where the fourth-order approximation \( B_4 \) is utilized. Consequently, if \( r = O(1) \), we show the asymptotical normality of \( \hat{\varepsilon}_2 \) as long as \( \frac{d_{\text{max}}^{9/16}}{\lambda_\varepsilon} \to 0 \) when \( d_{\text{max}} \to \infty \). The rate \( d_{\text{max}}^{9/16} \) is slightly larger than the minimal SNR requirement \( \lambda_\varepsilon \gg d_{\text{max}}^{3/2} \). Moreover, when \( \Delta_d = O(d_{\text{max}}^{1/2}) \), the asymptotical normality still holds if we replace \( B_4 \) with the first-order bias correction \( B_1 \). In principle, the rate \( d_{\text{max}}^{9/16} \) is improvable if even higher order bias corrections are implemented. However, the calculations of higher order approximations of \( \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) are intricate. In addition, to construct data-dependent confidence regions of \((U, V)\), we will need efficient estimators of \( \| \Lambda^{-1} \|_F^2 \). Since the rate \( \lambda_\varepsilon \gg d_{\text{max}}^{9/16} \) also appears in the efficiency of a shrinkage estimator of the singular values which still exists even if we apply higher-order approximations of \( \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \), we give up (at least in this article) calculating those even higher order approximations of \( \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). Therefore, in this article, \( \lambda_\varepsilon \gg d_{\text{max}}^{9/16} \) is the sharpest SNR condition we could obtain.

5. Finally, by RMT, we define a shrinkage estimator of the singular values \( \lambda_j \) which is more effective than the empirical singular values \( \hat{\lambda}_j \). More specifically, we define the diagonal matrix \( \hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_r) \) where

\[
\hat{\lambda}_j^2 = \lambda_j^2 - (d_1 + d_2) - \frac{d_1 d_2}{\lambda_j^2 - (d_1 + d_2)}, \quad \text{for all } 1 \leq j \leq r.
\]

Together with \( \hat{\Lambda} \), we define the plug-in estimator of \( B_4 \) as

\[
\hat{B}_4 = 2(d_* \| \hat{\Lambda}^{-1} \|_F^2 - \Delta_d^2 \| \hat{\Lambda}^{-2} \|_F^2 + d_* \Delta_d^2 \| \hat{\Lambda}^{-3} \|_F^2 - (d_1^2 + d_1 - d_2 - d_2^2) \Delta_d^2 \| \hat{\Lambda}^{-4} \|_F^2).
\]

In the end, we prove that \( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \hat{B}_4}{\sqrt{8d_\varepsilon \| \Lambda^{-2} \|_F}} \) converges to the standard normal distribution as long as \( \frac{d_{\text{max}}^{9/16}}{\lambda_\varepsilon} \to 0 \), based on which, confidence regions of the true singular
subspaces are constructed. The result is neat and the convergence rates to the standard normal distributions will also be presented.

The rest of the paper is organized as follows. In Section 2, we derive the explicit representation formula for the empirical spectral projector. The representation formula will be established for deterministic symmetric matrix perturbations. By applying the representation formula, we prove the normal approximation of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) in Section 3. Especially, we show that \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) is asymptotically normal under nearly optimal requirement of SNR. In Section 3, we also develop the four different levels of approximating the expected joint projection distance \( \mathbb{E}\text{dist}^2[\hat{(U, V)}, (U, V)] \). Corresponding normal approximation theorems are specifically established with different requirements for SNR. In Section 4, we propose a shrinkage estimator of the unknown singular values and prove the asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) based on data-dependent centering and normalizing factors. We then display comprehensive simulation results in Section 5, where, for instance, we show the importance of those higher order approximations of \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) when the matrix has unbalanced sizes and we show the effectiveness of the shrinkage estimator of the singular values in the normal approximation of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). The proof is collected in Section 7.

## 2 Representation formula of spectral projectors

In this section, we derive the representation formula of the empirical spectral projectors when a low rank symmetric matrix is perturbed by a symmetric noise matrix. To be exact, let \( A \) and \( X \) be \( d \times d \) symmetric matrices. The matrix \( A \) has low rank \( r = \text{rank}(A) \ll d \). Denote the eigen-decomposition of \( A \),

\[
A = \Theta \Lambda \Theta^T = \sum_{j=1}^{r} \lambda_j \theta_j \theta_j^T
\]

where the diagonal matrix \( \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r) \) consists of the non-zero and non-increasing eigenvalues of \( A \). The \( d \times r \) matrix \( \Theta = (\theta_1, \cdots, \theta_r) \) contains the eigenvectors of \( A \). The
matrix $X$ is a noise matrix and $\|X\| < \min_{1 \leq i \leq r} \frac{|\lambda_i|}{2}$ where $\|\cdot\|$ denotes the matrix operator norm. Given the perturbed version $\hat{A} = A + X$ where $A$ and $X$ are both unknown, the goal is to estimate $\Theta$. We denote by $\hat{\Theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_r)$ the $d \times r$ matrices containing the eigenvectors of $\hat{A}$ with the largest $r$ eigenvalues in absolute values. Therefore, $\hat{\Theta}$ represents the empirical version of $\Theta$. In this following, we derive the representation formula for the empirical spectral projector $\hat{\Theta}\hat{\Theta}^T$ when $X$ is a deterministic perturbation matrix. Generally speaking, we will represent $\hat{\Theta}\hat{\Theta}^T - \Theta\Theta^T$ by using only $\Theta, X, \Theta_\perp$ and $\Lambda$. The formula should be useful for various of purposes.

To this end, define $\Theta_\perp = (\theta_{r+1}, \cdots, \theta_d)$ the $d \times (d-r)$ matrix such that $(\Theta, \Theta_\perp)$ is an orthonormal matrix. Define the spectral projector corresponding to the orthogonal projection of $\Theta\Theta^T$,

$$P_\perp = \sum_{j=r+1}^{d} \theta_j \theta_j^T = \Theta_\perp \Theta_\perp^T.$$  

Note that $\Theta_\perp$ may not be uniquely definable whereas the spectral projector $\Theta_\perp \Theta_\perp^T$ is unique. Also, we define

$$P^{-1} := \sum_{j=1}^{r} \lambda_j^{-1} \theta_j \theta_j^T = \Theta \Lambda^{-1} \Theta^T.$$  

As a result, we can write $P^{-k} = \Theta \Lambda^{-k} \Theta^T$ for all $k \geq 1$. Theorem 1 provides an explicit representation formula for $\hat{\Theta}\hat{\Theta}^T - \Theta\Theta^T$ which is an infinite sum of the matrix products. For notational simplicity, we will denote $P^0 = P_\perp$ and denote the $k$-th order perturbation term of $\hat{\Theta}\hat{\Theta}^T - \Theta\Theta^T$ by

$$S_{A,k}(X) = \sum_{s; s_1 + \cdots + s_{k+1} = k} (-1)^{1+\tau(s)} \cdot P^{-s_1}XP^{-s_2}X \cdots X P^{-s_{k+1}} \quad (2.1)$$

where $s = (s_1, \cdots, s_{k+1})$ contains non-negative integer indices and

$$\tau(s) = \sum_{j=1}^{k+1} I(s_j > 0)$$

denotes the number of positive indices in $s$. For instance, if $k = 1$, we will have

$$S_{A,1}(X) = P^{-1}XP_\perp + P_\perp XP^{-1}.$$
If \( k = 2 \), then we consider \( s_1 + s_2 + s_3 = 2 \) for \( s_1, s_2, s_3 \geq 0 \). As a result of (2.1), we have
\[
S_{A,2}(X) = (\mathcal{P}^{-2}X\mathcal{P}^\perp X\mathcal{P}^\perp + \mathcal{P}^\perp X\mathcal{P}^{-2}X\mathcal{P}^\perp + \mathcal{P}^\perp X\mathcal{P}^\perp X\mathcal{P}^{-2})
- (\mathcal{P}^\perp X\mathcal{P}^{-1}X\mathcal{P}^{-1} + \mathcal{P}^{-1}X\mathcal{P}^\perp X\mathcal{P}^{-1} + \mathcal{P}^{-1}X\mathcal{P}^{-1}X\mathcal{P}^\perp).
\]

**Theorem 1.** If \( \|X\| < \min_{1 \leq i \leq r} |\lambda_i|/2 \), then
\[
\hat{\Theta}^T\hat{\Theta} - \Theta^T\Theta = \sum_{k \geq 1} S_{A,k}(X)
\]
where \( S_{A,k}(X) \) is defined in (2.1) and we set \( \mathcal{P}^\perp = \Theta^\perp \Theta^T \) for notational simplicity.

Apparently, by eq. (2.1), a simple fact is
\[
\|S_{A,k}(X)\| \leq \left(\frac{2^k}{k}\right) \cdot \frac{\|X\|^k}{\min_{1 \leq i \leq r} |\lambda_i|^k} \leq \left(\frac{4\|X\|}{\min_{1 \leq i \leq r} |\lambda_i|}\right)^k, \quad \forall \ k \geq 1.
\]

### 3 Normal approximation of spectral projectors

Recall from (1.1) that \( \hat{M} = M + Z \in \mathbb{R}^{d_1 \times d_2} \) with SVD of \( M \) denoted by \( M = U\Lambda V^T \) where \( U \in \mathbb{R}^{d_1 \times r} \) and \( V \in \mathbb{R}^{d_2 \times r} \) satisfying \( U^TU = I_r \) and \( V^TV = I_r \). Here, the \( r \times r \) identity matrix is denoted by \( I_r \). The diagonal matrix \( \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r) \) contains non-increasing positive singular values of \( M \). As in model (1.1), the noise matrix \( Z \) has i.i.d. standard normal entries, i.e., \( Z_{j_1j_2} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \) for \( 1 \leq j_1 \leq d_1 \) and \( 1 \leq j_2 \leq d_2 \). Let \( \hat{U} \) and \( \hat{V} \) be \( \hat{M} \)'s top-\( r \) left and right singular vectors. In this section, we will investigate the normal approximation of
\[
\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] = \|\hat{U}\hat{U}^T - UU^T\|^2_F + \|\hat{V}\hat{V}^T - VV^T\|^2_F,
\]
which is the (squared) joint projection distance between the empirical singular subspace and the true singular subspace. It is also called the (squared) projection distance on Grassmannians. The notation \( \|\cdot\|_F \) denotes the matrix Frobenius norm. To this end, we clarify several important notations which will appear frequently throughout the article.

In order to apply the representation formula developed in Theorem 1, we shall turn \( \hat{M}, M \) and \( Z \) into symmetric matrices. To be consistent with the notations in Section 2, we create
\((d_1 + d_2) \times (d_1 + d_2)\) symmetric matrices as 
\[
\hat{A} = \begin{pmatrix}
0 & \hat{M} \\
\hat{M}^T & 0
\end{pmatrix}, \quad A = \begin{pmatrix}
0 & M \\
M^T & 0
\end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix}
0 & Z \\
Z^T & 0
\end{pmatrix}.
\]

The model (1.1) is thus translated into \(\hat{A} = A + X\) which are the notations used in Section 2.

Now, the symmetric matrix \(A\) has eigenvalues \(\lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{-r} \geq \cdots \geq \lambda_{-1}\) where \(\lambda_{-i} = -\lambda_i\) for \(1 \leq i \leq r\). The eigenvectors corresponding to \(\lambda_i\) and \(\lambda_{-i}\) are, respectively,
\[
\theta_i = \frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ v_i \end{pmatrix} \quad \text{and} \quad \theta_{-i} = \frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ -v_i \end{pmatrix}
\]
for \(1 \leq i \leq r\), where \(\{u_i\}_{i=1}^{r}\) and \(\{v_i\}_{i=1}^{r}\) are the columns of \(U\) and \(V\). Note that the eigenvectors \(\{\theta_i\}_{i=1}^{r}\) may not be uniquely definable if some singular value \(\lambda_i\) has multiplicity larger than 1. However, the spectral projectors \(UU^T\) and \(VV^T\) are still uniquely definable regardless of the multiplicities of \(M\)'s singular values.

Following the same routine of notations, we denote by \(\Theta = (\theta_1, \cdots, \theta_r, \theta_{-r}, \cdots, \theta_{-1}) \in \mathbb{R}^{(d_1 + d_2) \times 2r}\) and \(\Theta_\perp \in \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2 - 2r)}\) such that \((\Theta, \Theta_\perp)\) is an orthonormal matrix. Then, we have
\[
\Theta \Theta^T = \sum_{1 \leq |j| \leq r} \theta_j \theta_j^T = \begin{pmatrix}
UU^T & 0 \\
0 & VV^T
\end{pmatrix}
\]
and similarly,
\[
\hat{\Theta} \hat{\Theta}^T = \sum_{1 \leq |j| \leq r} \hat{\theta}_j \hat{\theta}_j^T = \begin{pmatrix}
\hat{U} \hat{U}^T & 0 \\
0 & \hat{V} \hat{V}^T
\end{pmatrix}
\]
where \(\hat{U}\) and \(\hat{V}\) represent \(\hat{M}\)'s top-\(r\) left and right singular vectors. Similarly, for all \(k \geq 1\), we denote by
\[
\mathcal{Y}^{-k} = \sum_{1 \leq |j| \leq r} \frac{1}{\lambda_j^k} \theta_j \theta_j^T = \begin{cases}
\begin{pmatrix}
0 & U \Lambda^{-k} V^T \\
V \Lambda^{-k} U^T & 0
\end{pmatrix} & \text{if } k \text{ is odd} \\
\begin{pmatrix}
U \Lambda^{-k} U^T & 0 \\
0 & V \Lambda^{-k} V^T
\end{pmatrix} & \text{if } k \text{ is even}
\end{cases}
\]
12
The orthogonal spectral projector can therefore be written as

$$
\mathcal{P}^\perp = \Theta_\perp \Theta_\perp^T = \begin{pmatrix}
U_\perp U_\perp^T & 0 \\
0 & V_\perp V_\perp^T
\end{pmatrix}
$$

where \((U, U_\perp)\) and \((V, V_\perp)\) are orthonormal matrices. Actually, the columns of \(\Theta_\perp\) can be explicitly expressed by the columns of \(U_\perp\) and \(V_\perp\). Indeed, if we denote the columns of \(\Theta_\perp \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2-2r)}\) by \(\Theta_\perp = (\theta_{r+1}, \cdots, \theta_{d_1}, \theta_{d_1-1}, \cdots, \theta_{d_1-2r})\), then we can write

$$
\theta_{j_1} = \begin{pmatrix} u_{j_1} \\ 0 \end{pmatrix} \quad \text{and} \quad \theta_{-j_2} = \begin{pmatrix} 0 \\ v_{j_2} \end{pmatrix}
$$

for \(r + 1 \leq j_1 \leq d_1\) and \(r + 1 \leq j_2 \leq d_2\).

### 3.1 Normal approximation of the joint projection distance

By the above notations, it is clear that we have

$$
\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] = \|\hat{\Theta} \Theta^T - \Theta \Theta^T\|^2_F.
$$

Therefore, it suffices to prove the normal approximation of \(\|\hat{\Theta} \Theta^T - \Theta \Theta^T\|^2_F\). Observe that

$$
\|\hat{\Theta} \Theta^T - \Theta \Theta^T\|^2_F = 4r - 2\langle \Theta \Theta^T, \hat{\Theta} \Theta^T \rangle = -2\langle \Theta \Theta^T, \hat{\Theta} \Theta^T - \Theta \Theta^T \rangle.
$$

By Theorem 1 and the fact \(\Theta \Theta^T \mathcal{P}^\perp = 0\), we can write

$$
\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] = -2 \sum_{k \geq 2} \langle \Theta \Theta^T, S_{A,k} (X) \rangle = -2\|\mathcal{P}^\perp X \mathcal{P}^{-1}\|^2_F - 2 \sum_{k \geq 3} \langle \Theta \Theta^T, S_{A,k} (X) \rangle.
$$

(3.1)

where we also used the fact \(\mathcal{P}^\perp \mathcal{P}^\perp = \mathcal{P}^\perp\) so that

$$
-2\langle \Theta \Theta^T, S_{A,2} (X) \rangle = 2\langle \Theta \Theta^T, \mathcal{P}^{-1} X \mathcal{P}^\perp X \mathcal{P}^{-1} \rangle = 2 \text{tr} (\mathcal{P}^{-1} X \mathcal{P}^\perp X \mathcal{P}^{-1}) = 2\|\mathcal{P}^\perp X \mathcal{P}^{-1}\|^2_F.
$$

In Theorem 2, we establish the central limit theorem of \(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\) with theoretic bias corrections where the explicit normalization factor is derived.
**Theorem 2.** Suppose that \( d_{\max} \geq 3r \) where \( d_{\max} = \max\{d_1, d_2\} \). There exist absolute constants \( C_1, C_2, c_1 > 0 \) such that if \( \lambda_r \geq C_1 d_{\max}^{1/2} \), then for any \( t \geq 1 \),

\[
\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V))}{\sqrt{8d_*}||\Lambda^{-2}||_F} \leq x \right) - \Phi(x) \right| \leq C_2 t^{1/2} \left( \frac{\sqrt{r}}{||\Lambda^{-2}||_F} \right) \left( \frac{rd_{\max}}{\lambda_r} \right)^{1/2} + e^{-t} + e^{-c_1 d_{\max}} + C_2 \left( \frac{||\Lambda^{-1}||^4_F}{||\Lambda^{-2}||^2_F} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} ,
\]

where \( d_* = d_1 + d_2 - 2r \) and \( \Phi(x) \) denotes the cumulative distribution function of the standard normal distributions.

By setting \( t = \left( \frac{||\Lambda^{-2}||_F \lambda_r^2}{\sqrt{r}} \right) \cdot \frac{\lambda_r}{\left( \frac{rd_{\max}}{\lambda_r} \right)^{1/2}} \) in Theorem 2, we can conclude that the asymptotical normality of \( \text{dist}^2((\hat{U}, \hat{V}), (U, V)) - \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) holds as long as

\[
\left( \frac{\sqrt{r}}{||\Lambda^{-2}||_F \lambda_r^2} \right) \left( \frac{rd_{\max}}{\lambda_r} \right)^{1/2} \to 0 \quad \text{and} \quad \left( \frac{||\Lambda^{-1}||^4_F}{||\Lambda^{-2}||^2_F} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} \to 0
\]

when \( d_1, d_2 \to \infty \). Apparently, such SNR condition is optimal up to the rank parameter \( r \).

**Remark 1.** The normalization factor \( \sqrt{8d_*}||\Lambda^{-2}||_F \) comes from the fact that

\[
\text{Var}(2||\mathcal{P}^{-1}X\mathcal{P}^\perp||_F^2) = 8d_*||\Lambda^{-2}||_F^2.
\]

Actually, it turns out that, under the conditions in Theorem 2,

\[
\text{Var}(\text{dist}^2((\hat{U}, \hat{V}), (U, V))) = [8 + o(1)] \cdot d_*||\Lambda^{-2}||_F^2.
\]

However, the centering term \( \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) in Theorem 2 is not explicitly determined. Approximating \( \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) requires much more delicate treatments. Under the conditions of Theorem 2 if we approximate \( \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) by its dominating term \( 2\mathbb{E}||\mathcal{P}^{-1}X\mathcal{P}^\perp||_F^2 \), we will obtain the first order approximation

\[
\mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) = [2 + o(1)] \cdot d_*||\Lambda^{-1}||_F^2.
\]

That is, the bias \( \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) is approximately \( d_*^{1/2} \) factor of the normalization factor \( \sqrt{8d_*}||\Lambda^{-2}||_F \). It implies that we shall approximate \( \mathbb{E} \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) to a much higher accuracy to maintain the asymptotical normality. It is the primary subject in section 3.2.
3.2 Approximating the bias

To obtain more explicit normal approximations of \(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\), we need a precise characterization of \(\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\). Recall from (3.1), we have

\[
\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] = 2\mathbb{E}\|\mathcal{P}^\perp X\mathcal{P}^{-1}\|_F^2 - 2\sum_{k\geq 2}\mathbb{E}\langle \Theta\Theta^T, S_{A,2k}(X) \rangle,
\]

where we used the fact \(\mathbb{E}S_{A,2k+1}(X) = 0\) for any positive integer \(k \geq 1\). Our goal is therefore to determine the expectations \(\mathbb{E}\|\mathcal{P}^\perp X\mathcal{P}^{-1}\|_F^2\) and \(\mathbb{E}\langle \Theta\Theta^T, S_{A,2k}(X) \rangle\) for all \(k \geq 2\). Clearly, if we can obtain explicit expressions of \(\mathbb{E}\langle \Theta\Theta^T, S_{A,2k}(X) \rangle\) for larger \(k\)’s, we will end up with more precise characterization of \(\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\). In the following lemmas, we provide four different levels of approximating the expected joint projection distance \(\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\).

**Lemma 1** (First order approximation). The following equation holds

\[
\mathbb{E}\|\mathcal{P}^\perp X\mathcal{P}^{-1}\|_F^2 = d_\star\|\Lambda^{-1}\|_F^2
\]

where \(d_\star = d_1 + d_2 - 2r\). Moreover, if \(\lambda_r \geq C_1 d_\star^{1/2}\) for some large enough constant \(C_1 > 0\), then

\[
\left| \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star\|\Lambda^{-1}\|_F^2 \right| \leq C_2 \frac{r d_\star^2 d_{\max}}{\lambda_r^4}
\]

where \(C_2 > 0\) is an absolute constant (depending on the constant \(C_1\)).

In Lemma 2 and Lemma 3 we calculate the leading terms of \(-\mathbb{E}\langle \Theta\Theta^T, S_{A,4}(X) \rangle\) and \(-\mathbb{E}\langle \Theta\Theta^T, S_{A,6}(X) \rangle\). They enable us to reach the second and third order approximations of \(\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\). Moreover, Lemma 2 and 3 suggest that the first order approximation \(2d_\star\|\Lambda^{-1}\|_F^2\) over-estimates \(\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\), whereas the second order approximation \(2(d_\star\|\Lambda^{-1}\|_F^2 - \Delta_\star^2\|\Lambda^{-2}\|_F^2)\) under-estimates \(\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\).

**Lemma 2** (Second order approximation). The following fact holds

\[
\left| -\mathbb{E}\langle \Theta\Theta^T, S_{A,4}(X) \rangle + \Delta_\star^2\|\Lambda^{-2}\|_F^2 \right| \leq C_2 \frac{r^2 d_{\max}}{\lambda_r^4}
\]
where \( d_* = d_1 + d_2 - 2r \) and \( \Delta_d = d_1 - d_2 \) and \( C_2 > 0 \) is an absolute constant. Moreover, if \( \lambda_r \geq C_1 d_{\max}^{1/2} \) for some large enough constant \( C_1 > 0 \), then

\[
\left| \mathbb{E} \text{ dist}^2[\hat{U}, \hat{V}], (U, V)] - 2(d_* \| \Lambda^{-1} \|_F^2 - \Delta_d^2 \| \Lambda^{-2} \|_F^2) \right| \leq C_3 \frac{r^2 d_{\max}^2}{\lambda_r^4} + C_4 \frac{r d_{\max}^3}{\lambda_r^6}
\]

where \( C_3, C_4 > 0 \) are absolute constants (depending on \( C_1 \) only).

Note that in both the second and third order approximations, the dimension difference \( \Delta_d = d_1 - d_2 \) is involved in the higher order bias terms. If \( d_1 \approx d_2 \) so that \( |d_1 - d_2| = O(d_{\max}^{1/2}) \), then these higher order approximations essentially have the same effects as the first order approximation.

**Lemma 3** (Third order approximation). The following fact holds

\[
\left| - \mathbb{E} \langle \Theta \Theta^T, S_{A,6}(X) \rangle - d_* \Delta_d^2 \| \Lambda^{-2} \|_F^2 \right| \leq C_2 \frac{r^2 d_{\max}^2 + r^3 d_{\max}}{\lambda_r^6}
\]

where \( \Delta_d = d_1 - d_2, d_* = d_1 + d_2 \) and \( C_2 > 0 \) is an absolute constant. Moreover, if \( \lambda_r \geq C_1 d_{\max}^{1/2} \) for some large enough constant \( C_1 > 0 \), then

\[
\left| \mathbb{E} \text{ dist}^2[\hat{U}, \hat{V}], (U, V)] - 2d_* \| \Lambda^{-1} \|_F^2 + 2\Delta_d^2 \| \Lambda^{-2} \|_F^2 - 2d_* \Delta_d^2 \| \Lambda^{-3} \|_F^2 \right| \leq C_3 \frac{r^2 d_{\max}^2 + r^3 d_{\max}}{\lambda_r^6} + C_4 \frac{r d_{\max}^4}{\lambda_r^8}
\]

for some absolute constants \( C_3, C_4 > 0 \) (depending on \( C_1 \) only).

**Lemma 4** (Fourth order approximation). The following fact holds

\[
\left| - \mathbb{E} \langle \Theta \Theta^T, S_{A,8}(X) \rangle + (d_{1-}^2 + d_{2-}^2 + d_1 - d_2) \Delta_d^2 \| \Lambda^{-4} \|_F^2 \right| \leq C_2 \frac{r^2 d_{\max}^3}{\lambda_r^8}
\]

where \( d_* = d_1 + d_2 \) and \( \Delta_d = d_1 - d_2 \) and \( C_2 > 0 \) is an absolute constant. Moreover, if \( \lambda_r \geq C_1 d_{\max}^{1/2} \) for some large enough constant \( C_1 > 0 \), then

\[
\left| \mathbb{E} \text{ dist}^2[\hat{U}, \hat{V}], (U, V)] - B_4 \right| \leq C_3 \frac{r^2 d_{\max}^3}{\lambda_r^8} + C_4 \frac{r d_{\max}^5}{\lambda_r^{10}}
\]

for some absolute constants \( C_3, C_4 > 0 \) (depending on \( C_1 \) only) where

\[
B_4 = 2(d_* \| \Lambda^{-1} \|_F^2 - \Delta_d^2 \| \Lambda^{-3} \|_F^2 + d_* \Delta_d^2 \| \Lambda^{-3} \|_F^2 - (d_{1-}^2 + d_{2-}^2 + d_1 - d_2) \Delta_d^2 \| \Lambda^{-4} \|_F^2).
\]

16
Remark 2. In this article, we only approximate the expected joint projection distance up to the fourth order as in Lemma 4. If $r \ll d_{\text{max}}$, then the remainder error by the fourth order approximation as in Lemma 4 is of the order $O(rd_{\text{max}}^5/\lambda r^{10})$. In principle, we can approximate the bias $E \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ to even higher orders. However, the calculations of even higher order terms are intricate and thus will not be pursued here. Moreover, in Section 4, we will see that the fourth order approximation is sufficient for constructing the data-dependent confidence regions of the true singular subspaces. The reason is that the data-dependent estimates of the singular values related parameters, e.g., $\|\Lambda^{-1}\|_F^2$ and $\|\Lambda^{-2}\|_F$, will dominate the fourth-order approximation errors of $E\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$.

3.3 Normal approximation after bias corrections

In this section, we will display the normal approximation of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ with explicit centering and normalizing factors. In view of Theorem 2, it now suffices to replace $E \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ with the explicit approximation formulas from Lemma 1-4 which ends up with more transparent central limit theorem for $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$.

Similarly as in Section 3.2, we will provide four levels of normal approximation of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$, each of which corresponds to different levels of bias corrections for $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Clearly, higher order bias corrections, while having more complicate bias reduction terms, require lower levels of SNR to guarantee the asymptotical normality. For instance, the first order bias correction in Theorem 3 requires SNR as $\lambda r \gg d_{\text{max}}^3$ for asymptotical normality, while the fourth order bias correction in Theorem 6 only requires SNR as $\lambda r \gg d_{\text{max}}^{9/16}$ for asymptotical normality.

**Theorem 3** (First order CLT). Suppose that $d_{\text{max}} \geq 3r$ where $d_{\text{max}} = \max\{d_1, d_2\}$. There exist absolute constants $C_1, C_2, C_3, c_1 > 0$ such that if $\lambda r \geq C_1 d_{\text{max}}^{1/2}$, then, for all $t \geq 1$,

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_* \|\Lambda^{-1}\|_F^2}{\sqrt{8d_* \|\Lambda^{-2}\|_F}} \leq x \right) - \Phi(x) \right| \leq C_2 \left( \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F \lambda r^2} \right) \cdot \left( \frac{rd_{\text{max}}^t}{\lambda r} \right)^{1/2} + e^{-c_1d_{\text{max}}} + C_2 \left( \frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{d_{\text{max}}^{1/2}} + C_3 \frac{rd_{\text{max}}^{3/2}}{\lambda r^2} + 2e^{-t},$$

where $d_* = d_1 + d_2 - 2r$. 

17
By Theorem 3, we conclude with the asymptotical normality
\[ \frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2d_*\|\Lambda^{-1}\|_F^2}{\sqrt{8d_*\|\Lambda^{-2}\|_F}} \xrightarrow{d} \mathcal{N}(0, 1) \]
as \(d_1, d_2 \to \infty\)
if
\[ \frac{rd_{\text{max}}^{1/2} + r^{1/2}d_{\text{max}}^{3/4}}{\lambda_r} \to 0 \quad \text{and} \quad \left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{d_{\text{max}}^{1/2}} \to 0. \]

In the case \(r = O(1)\), the above conditions require SNR condition \(\lambda_r \gg d_{\text{max}}^{3/4}\). The SNR requirement \(\lambda_r \gg d_{\text{max}}^{3/4}\) is larger than the standard SNR requirement \(\lambda_r \gg d_{\text{max}}^{1/2}\). The exponent \(\frac{3}{4}\) can be improved if we utilize the higher order approximations of \(\mathbb{E}\) dist\(^2\)[(\(\hat{U}, \hat{V}), (U, V)]\].

**Theorem 4** (Second order CLT). Suppose that \(d_{\text{max}} \geq 3r\) where \(d_{\text{max}} = \max\{d_1, d_2\}\). There exist absolute constants \(C_1, C_2, C_3, c_1 > 0\) such that if \(\lambda_r \geq C_1d_{\text{max}}^{1/2}\), then, for all \(t \geq 1\),
\[
\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2(d_*\|\Lambda^{-1}\|_F^2 - \Delta_d^2\|\Lambda^{-2}\|_F^2)}{\sqrt{8d_*\|\Lambda^{-2}\|_F}} \leq x \right) - \Phi(x) \right| \\
\leq C_2\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_F \lambda_r^2}\right) \cdot \left(\frac{rd_{\text{max}}t}{\lambda_r}\right)^{1/2} + e^{-c_1d_{\text{max}}} + C_2\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{d_{\text{max}}^{1/2}} + 2e^{-t} \\
+ C_3\frac{r^2d_{\text{max}}^{1/2}}{\lambda_r^2} + C_4\frac{r^2d_{\text{max}}^{5/2}}{\lambda_r^4},
\]
where \(d_* = d_1 + d_2 - 2r\) and \(\Delta_d = d_1 - d_2\).

By Theorem 4, the CLT of \(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\) with second order bias correction holds as long as the SNR satisfies \(\lambda_r \gg d_{\text{max}}^{5/8}\) when \(r = O(1)\). Apparently, the SNR condition is much weaker than SNR required in Theorem 3. Moreover, we can conclude that when \(\Delta_d = O(d_{\text{max}}^{1/2})\), then the CLT of \(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\) with the first order bias correction holds also under the SNR requirement \(\lambda_r \gg d_{\text{max}}^{5/8}\). Similarly, we have the following CLT of \(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\) with even higher order bias corrections. The proofs are identical to the proof of Theorem 3 and will be omitted.

**Theorem 5** (Third order CLT). Suppose that \(d_{\text{max}} \geq 3r\) where \(d_{\text{max}} = \max\{d_1, d_2\}\). There exist absolute constants \(C_1, C_2, C_3, C_4, c_1 > 0\) such that if \(\lambda_r \geq C_1d_{\text{max}}^{1/2}\), then, for all \(t \geq 1\),
\[
\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2(d_*\|\Lambda^{-1}\|_F^2 - \Delta_d^2\|\Lambda^{-2}\|_F^2 + d_*\Delta_d^2\|\Lambda^{-3}\|_F^2)}{\sqrt{8d_*\|\Lambda^{-2}\|_F}} \leq x \right) - \Phi(x) \right| \\
\leq C_2\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_F \lambda_r^2}\right) \cdot \left(\frac{rd_{\text{max}}t}{\lambda_r}\right)^{1/2} + e^{-c_1d_{\text{max}}} + C_2\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{d_{\text{max}}^{1/2}} + 2e^{-t} \\
+ C_3\frac{r^2d_{\text{max}}^{1/2}}{\lambda_r^2} + C_4\frac{r^2d_{\text{max}}^{5/2}}{\lambda_r^4},
\]
\[ \leq C_2 \left( \frac{\sqrt{r}}{\| A^{-2} \|_F \lambda_r^2} \right) \cdot \left( \frac{(r \lambda_{\max})^{1/2}}{\lambda_r} \right) + e^{-c_1 \lambda_{\max}} + C_2 \left( \frac{\| A^{-1} \|_F^4}{\| A^{-2} \|_F^2} \right)^{3/2} \cdot \frac{1}{\lambda_{\max}^{1/2}} + 2e^{-t} + C_3 \frac{r^2 \lambda_{\max}^{3/2}}{\lambda_r^4} + C_4 \frac{r \lambda_{\max}^{7/2}}{\lambda_r^6}, \]

where \( d_* = d_1 + d_2 - 2r \) and \( \Delta_d = d_1 - d_2 \).

**Theorem 6** (Fourth order CLT). Suppose that \( \lambda_{\max} \geq 3r \) where \( \lambda_{\max} = \max\{d_1, d_2\} \). There exist absolute constants \( C_1, C_2, C_3, C_4, c_1 > 0 \) such that if \( \lambda_r \geq C_1 \lambda_{\max}^2 \), then, for all \( t \geq 1 \),

\[
\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_4}{\sqrt{8d_* \| A^{-2} \|_F^2}} \leq x \right) - \Phi(x) \right| \\
\leq C_2 \left( \frac{\sqrt{r}}{\| A^{-2} \|_F \lambda_r^2} \right) \cdot \left( \frac{(r \lambda_{\max})^{1/2}}{\lambda_r} \right) + e^{-c_1 \lambda_{\max}} + C_2 \left( \frac{\| A^{-1} \|_F^4}{\| A^{-2} \|_F^2} \right)^{3/2} \cdot \frac{1}{\lambda_{\max}^{1/2}} + 2e^{-t} + C_3 \frac{r^2 \lambda_{\max}^{5/2}}{\lambda_r^4} + C_4 \frac{r \lambda_{\max}^{9/2}}{\lambda_r^6},
\]

where \( d_* = d_1 - d_2 \), \( \Delta_d = d_1 - d_2 \) with \( d_1 = d_1 - r \) and \( d_2 = d_2 - r \) and

\[
B_4 = 2(d_* \| A^{-1} \|_F^2 - \Delta_d^2 \| A^{-2} \|_F^2 + d_1 \Delta_d^2 \| A^{-3} \|_F^2 - (d_1 + d_2 + d_1 - d_2) \Delta_d^2 \| A^{-4} \|_F^2).
\]

By Theorem 5, the asymptotical normality holds under the SNR requirement \( \lambda_r \gg \lambda_{\max}^{7/12} \) when \( r = O(1) \). Similarly, the asymptotical normality in Theorem 6 holds under the SNR condition \( \lambda_r \gg \lambda_{\max}^{1/16} \) when \( r = O(1) \). Moreover, if \( \Delta_d = O(\lambda_{\max}^{1/2}) \), then Theorem 6 implies that the asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) with the first order bias correction \( 2d_* \| A^{-1} \|_F^2 \) also holds under the SNR condition \( \lambda_r \gg \lambda_{\max}^{9/16} \).

Note that the asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) in Theorem 3 holds on the unknown parameters \( \Lambda \) which should be estimated from the data matrix \( \hat{M} \). In Section 4, we propose efficient estimators of \( \{\lambda_j \}_{j=1}^r \) based on recent developments of random matrix theory in the literature. As a result, we will obtain asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) with data-dependent centering and normalizing factors, based on which, the data-dependent confidence regions of the true singular subspaces will be constructed under the SNR condition \( \lambda_r \gg \lambda_{\max}^{9/16} \).
4 Confidence regions of singular subspaces

4.1 Shrinkage estimation of singular values

To replace the unknown singular values $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r)$ in Theorem 3 with data-dependent estimates, the immediate yet naive estimates are the empirical singular values $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ which consists of the top-$r$ singular values of $\hat{M}$. In the noisy matrix SVD model (1.1), the data matrix $\hat{M} = M + Z$ where the noise matrix $Z$ has i.i.d. standard normal entries. Recall that the SVD of $\hat{M}$ is written as

$$\hat{M} = \sum_{j=1}^{\min(d_1, d_2)} \hat{\lambda}_j \cdot (\hat{u}_j \hat{v}_j^T).$$

It is well known that the empirical singular values of $\{\hat{\lambda}_j\}_{j=1}^{r}$ are biased estimators of the true singular values $\{\lambda_j\}_{j=1}^{r}$. See Benaych-Georges and Nadakuditi (2012) and Ding (2017) for more details. By the non-asymptotic results from Ding (2017), if $\lambda_r > 2d_1^{1/2}$, then with probability at least $1 - d_3/\max$, for all $1 \leq j \leq r$,

$$\left| \hat{\lambda}_j^2 - \left( \lambda_j^2 + (d_1 + d_2) + \frac{d_1 d_2}{\lambda_j^2} \right) \right| \leq C_1 d_3^{1/4} \lambda_j^{1/2}$$

for some absolute constant $C_1 > 0$. Therefore, by the fact $|\hat{\lambda}_j^2 - \lambda_j^2| \geq c_0(d_1 + d_2)$, we get

$$\frac{d_3 \|\Lambda^{-1}\|_F^2 - \|\hat{\Lambda}^{-1}\|_F^2}{\sqrt{8d_3 \|\Lambda^{-2}\|_F}} \geq c_1 \frac{d_3^{3/2}}{\lambda_r^2}$$

for some absolute constants $c_0, c_1 > 0$. As a result, if we directly apply $\|\hat{\Lambda}^{-1}\|_F^2$ as the plug-in estimates of $\|\Lambda^{-1}\|_F^2$ in Theorem 3, we will require $\frac{d_3^{3/4}}{\lambda_r} \to 0$ to guarantee the asymptotical normality. This SNR condition is the same as in Theorem 3. However, it is significantly stronger than the SNR conditions in Theorem 4-6 when higher order bias corrections are applied. Therefore, we need more effective estimators of the true singular values.

To this end, based on (4.1), we define the shrinkage estimator of $\lambda_j^2$ as

$$\tilde{\lambda}_j^2 = \hat{\lambda}_j^2 - (d_1 + d_2) - \frac{d_1 d_2}{\lambda_j^2} - (d_1 + d_2)$$

for all $1 \leq j \leq r$. (4.2)

We have the following lemma to characterize the effectiveness of the shrinkage estimator of singular values.
Lemma 5. If $\lambda_r \geq C_1 d_{\text{max}}^{1/2}$ for a large enough constant $C_1 > 0$, then with probability at least $1 - d_{\text{max}}^{-2}$, for all $1 \leq j \leq r$:

$$|\tilde{\lambda}_j^2 - \lambda_j^2| \leq C_2 \frac{d_1^2 d_2^2}{\lambda_j^6} + C_3 d_{\text{max}}^{1/4} \lambda_j^{1/2}$$

where $C_2, C_3$ are some absolute positive constants.

Proof of Lemma 5. By definition of $\tilde{\lambda}_j^2$ in (4.2), we get

$$|\tilde{\lambda}_j^2 - \lambda_j^2| = \left| \frac{\tilde{\lambda}_j^2 - \lambda_j^2 - (d_1 + d_2)}{\lambda_j^2 - (d_1 + d_2)} \right|$$

$$\leq \left| \frac{d_1 d_2}{\lambda_j^2} - \frac{d_1 d_2}{\lambda_j^2 - (d_1 + d_2)} \right| + C_1 d_{\text{max}}^{1/4} \lambda_j^{1/2}$$

$$\leq C_1 d_1 d_2 \left( \frac{\lambda_j^2 - (d_1 + d_2) - \lambda_j^2}{\lambda_j^2} \right) + C_2 d_{\text{max}}^{1/4} \lambda_j^{1/2}$$

$$\leq C_1 \frac{d_1^2 d_2^2}{\lambda_j^6} + C_2 \frac{d_1 d_2 d_{\text{max}}^{1/4}}{\lambda_j^{7/2}} + C_3 d_{\text{max}}^{1/4} \lambda_j^{1/2}$$

$$\leq C_1 \frac{d_1^2 d_2^2}{\lambda_j^6} + C_2 d_{\text{max}}^{1/4} \lambda_j^{1/2}$$

where we used the fact $\tilde{\lambda}_j^2 - (d_1 + d_2) \geq c_0 \lambda_j^2$ for some absolute constant $c_0 \in (0, 1)$ and the last inequality is due to the fact $\lambda_r \geq C_1 d_{\text{max}}^{1/2}$ for some large enough constant $C_1 > 0$.

4.2 Normal approximation with data-dependent bias corrections

We denote $\tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_r)$ which contains the thresholding estimators of singular values $\{\tilde{\lambda}_j\}_{j=1}^r$ defined as (4.2). We will use $\|\tilde{\Lambda}^{-1}\|_F^2$, $\|\tilde{\Lambda}^{-2}\|_F^2$, $\|\tilde{\Lambda}^{-3}\|_F^2$ and $\|\tilde{\Lambda}^{-4}\|_F^2$ as estimators for $\|\Lambda^{-1}\|_F^2$, $\|\Lambda^{-2}\|_F^2$, $\|\Lambda^{-3}\|_F^2$ and $\|\Lambda^{-4}\|_F^2$, respectively, which are needed as the bias corrections in Theorem 3-6. In this section, we prove the CLT of $\text{dist}^2([\hat{U}, \hat{V}), (U, V)]$ with the data-dependent centering and normalization factors. Only the proof of Theorem 7 will be provided in Section 7 because the idea of proving the other theorems are essentially identical.

Theorem 7 (First order data-dependent CLT). Suppose that $d_{\text{max}} \geq 3r$ where $d_{\text{max}} = \max\{d_1, d_2\}$. There exist absolute constants $C_0, C_1, C_2, C_3, C_4, c_1 > 0$ such that if $\lambda_r \geq$
$C_0 d_{\max}^{1/2}$, then for all $t \geq 1$,

$$
\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left( \frac{\text{dist}^2(\hat{U}, \hat{V}), (U, V) - 2d_\star \|\Lambda^{-1}\|^2_F}{\sqrt{8d_\star \|\Lambda^{-2}\|^2_F}} \leq x \right) - \Phi(x) \right| \leq C_1 \left( \frac{r d_{\max} t}{\lambda_r} \right)^{1/2} + 2e^{-c_1 d_{\max}} + C_2 \left( \frac{\|\Lambda^{-1}\|^4_F}{\|\Lambda^{-2}\|^2_F} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + 2e^{-t} + C_3 \frac{r d_{\max}^3}{\lambda_r^2} + C_4 \frac{r d_{\max}^2}{\lambda_r^3},
$$

where $d_\star = d_1 + d_2 - 2r$ and $\Delta_d = d_1 - d_2$.

**Remark 3.** The normal approximation in Theorem 3 has two dominating rates $\frac{r d_{\max}^{3/2}}{\lambda_r^2}$ and $\frac{r d_{\max}^{3/2}}{\lambda_r^3}$. The rate $\frac{r d_{\max}^{3/2}}{\lambda_r^2}$ is due to the first order approximation of the bias $\mathbb{E} \text{dist}^2(\hat{U}, \hat{V}), (U, V)$ by $2d_\star \|\Lambda^{-1}\|^2_F$ (see Lemma 1), while the rate $\frac{r d_{\max}^{3/2}}{\lambda_r^3}$ is due to the error of estimating $d_\star \|\Lambda^{-1}\|^2_F$ by the thresholding estimators $d_\star \|\tilde{\Lambda}^{-1}\|^2_F$ (see Lemma 5) which can not be improved with just higher order approximation of $\mathbb{E} \text{dist}^2(\hat{U}, \hat{V}), (U, V)$. Note that the rate $\frac{r d_{\max}^{3/2}}{\lambda_r^3}$ coincides with the error rate of fourth order approximation of $\mathbb{E} \text{dist}^2(\hat{U}, \hat{V}), (U, V)$ (see Theorem 6).

**Theorem 8** (Second order data-dependent CLT). Suppose that $d_{\max} \geq 3r$ where $d_{\max} = \max\{d_1, d_2\}$. There exist absolute constants $C_0, C_1, C_2, C_3, C_4, c_1 > 0$ such that if $\lambda_r \geq C_0 d_{\max}^{1/2}$, then for all $t \geq 1$,

$$
\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left( \frac{\text{dist}^2(\hat{U}, \hat{V}), (U, V) - 2d_\star \|\tilde{\Lambda}^{-1}\|^2_F - \Delta_d^2 \|\tilde{\Lambda}^{-2}\|^2_F}{\sqrt{8d_\star \|\tilde{\Lambda}^{-2}\|^2_F}} \leq x \right) - \Phi(x) \right| \leq C_1 \left( \frac{r d_{\max} t}{\lambda_r} \right)^{1/2} + 2e^{-c_1 d_{\max}} + C_2 \left( \frac{\|\tilde{\Lambda}^{-1}\|^4_F}{\|\tilde{\Lambda}^{-2}\|^2_F} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + 2e^{-t} + C_3 \frac{r d_{\max}^5}{\lambda_r^4} + C_4 \frac{r d_{\max}^4}{\lambda_r^5},
$$

where $d_\star = d_1 + d_2 - 2r$ and $\Delta_d = d_1 - d_2$.

**Theorem 9** (Third order data-dependent CLT). Suppose that $d_{\max} \geq 3r$ where $d_{\max} = \max\{d_1, d_2\}$. There exist absolute constants $C_0, C_1, C_2, C_3, C_4, c_1 > 0$ such that if $\lambda_r \geq C_0 d_{\max}^{1/2}$, then for all $t \geq 1$,

$$
\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left( \frac{\text{dist}^2(\hat{U}, \hat{V}), (U, V) - 2(d_\star \|\tilde{\Lambda}^{-1}\|^2_F - \Delta_d^2 \|\tilde{\Lambda}^{-2}\|^2_F + d_\star \Delta_d^2 \|\tilde{\Lambda}^{-3}\|^2_F)}{\sqrt{8d_\star \|\tilde{\Lambda}^{-2}\|^2_F}} \leq x \right) - \Phi(x) \right|
$$
\[
\leq C_1 \frac{\sqrt{r}}{\lambda_r^2 \| \Lambda^{-2} \|_F} \cdot \left( \frac{r d_{\max} t}{\lambda_r} \right)^{1/2} + 2e^{-c_1 d_{\max}} + C_2 \left( \frac{\| \Lambda^{-1} \|_F^4}{\| \Lambda^{-2} \|_F^2} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + 2e^{-t} + C_3 \frac{r d_{\max}^{7/2}}{\lambda_r^6} + C_4 \frac{r d_{\max}^{9/2}}{\lambda_r^8},
\]

where \( d_* = d_1 + d_2 - 2r \) and \( \Delta_d = d_1 - d_2 \).

Clearly by Theorem 9 after the third-order bias corrections with the data-dependent centering and normalization, the asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) holds with SNR requirement \( \lambda_r \gg d_{\max}^{7/12} \). Moreover, if \( \Delta_d = O(d_{\max}^{1/2}) \), then the second and third bias corrections can be ignorable.

**Theorem 10** (Fourth order data-dependent CLT). Suppose that \( d_{\max} \geq 3r \) where \( d_{\max} = \max\{d_1, d_2\} \). There exist absolute constants \( C_0, C_1, C_2, C_3, c_1 > 0 \) such that if \( \lambda_r \geq C_0 d_{\max}^{1/2} \), then for all \( t \geq 1 \),

\[
\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \tilde{B}_4}{\sqrt{8d_r} \| \Lambda^{-2} \|_F} \leq x \right) - \Phi(x) \right|
\leq C_1 \frac{\sqrt{r}}{\lambda_r^2 \| \Lambda^{-2} \|_F} \cdot \left( \frac{r d_{\max} t}{\lambda_r} \right)^{1/2} + 2e^{-c_1 d_{\max}} + C_2 \left( \frac{\| \Lambda^{-1} \|_F^4}{\| \Lambda^{-2} \|_F^2} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + C_3 \frac{r d_{\max}^{7/2}}{\lambda_r^6} + 2e^{-t},
\]

where \( \tilde{B}_4 \) is defined by

\[
\tilde{B}_4 = 2 \left( d_* \| \Lambda^{-1} \|_F^2 - \Delta_d^2 \| \Lambda^{-2} \|_F^2 + d_* \Delta_d^2 \| \Lambda^{-3} \|_F^2 - (d_{1-}^2 + d_{2-}^2 + d_{1-} d_{2-}) \Delta_d^2 \| \Lambda^{-4} \|_F^2 \right)
\]

with \( d_{1-} = d_1 - r, d_{2-} = d_2 - r \) and \( d_* = d_{1-} + d_{2-} \) and \( \Delta_d = d_{1-} - d_{2-} \).

**Remark 4.** By Theorem 10, after the fourth order data-dependent bias corrections, the asymptotical normality of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) holds as long as \( \lambda_r \gg d_{\max}^{9/16} \) when \( r = O(1) \). Similarly, if \( d_1 \approx d_2 \) so that \( |d_1 - d_2| = O(d_{\max}^{1/2}) \), then there is no need to implement the higher order bias corrections. Put it differently, when \( \Delta_d = O(d_{\max}^{1/2}) \), then the asymptotical normality of \( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_* \| \Lambda^{-1} \|_F}{\sqrt{8d_r} \| \Lambda^{-2} \|_F} \) holds by Theorem 10 as long as \( \lambda_r \gg d_{\max}^{9/16} \).

### 4.3 Data-dependent confidence regions of singular subspaces

From the data-dependent central limit theorems of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) established in Section 4.2 we can construct data-dependent confidence regions of the true singular subspaces.
In this section, we will construct confidence regions of the true singular subspaces which attain the pre-determined confidence levels asymptotically when \(d_1, d_2 \to +\infty\). Note that in the asymptotic scheme where the dimension \(d_1, d_2 \to \infty\), the underlying parameters \(r(\mathbf{d}_1, \mathbf{d}_2)\), \(U(\mathbf{d}_1, \mathbf{d}_2)\), \(V(\mathbf{d}_1, \mathbf{d}_2)\) and \(\Lambda(\mathbf{d}_1, \mathbf{d}_2)\) also depend on the dimension parameters \(d_1, d_2\). For notational simplicity, we will omit the superscripts \((\mathbf{d}_1, \mathbf{d}_2)\) for the parameters of interest without causing confusions.

Recall that we denote by

\[
\tilde{B}_4 = 2(d_\bullet \|\tilde{\Lambda}^{-1}\|_F^2 - \Delta^2 \|\tilde{\Lambda}^{-2}\|_F^2 + d_\bullet \Delta^2 \|\tilde{\Lambda}^{-3}\|_F^2 - (d_1^+ + d_2^+ + d_1^- d_2^-) \Delta^2 \|\tilde{\Lambda}^{-4}\|_F^2),
\]

where \(d_1^- = d_1 - r\) and \(d_2^- = d_2 - r\). By Theorem 10, we have that

\[
\frac{\text{dist}^2([\hat{U}, \hat{V}], (U, V)) - \tilde{B}_4}{\sqrt{8d_\bullet \|\tilde{\Lambda}^{-2}\|_F}} \to_{d_1, d_2} N(0, 1)
\]

as \(d_1, d_2 \to +\infty\) when

\[
\lim_{d_1, d_2 \to \infty} \max \left\{ \frac{r d_{\max}^{1/2} + r^{1/8} d_{\max}^{9/16} \lambda_r}{\lambda_r} + \left( \frac{\|\Lambda^{-1}\|_F}{\|\Lambda^{-2}\|_F} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} \right\} = 0. \tag{4.4}
\]

In order to construct \(100(1 - \alpha)\%\) confidence regions of the joint columns spaces for \(U\) and \(V\), we define a confidence set based on \((\hat{U}, \hat{V})\) and \(\tilde{\Lambda}\)

\[
\mathcal{M}_\alpha(\hat{U}, \hat{V}) := \left\{ (L, R) : L \in \mathbb{R}^{d_1 \times r}, R \in \mathbb{R}^{d_2 \times r}, L^T L = R^T R = I_r, \text{ and } \left| \text{dist}^2([L, R], (\hat{U}, \hat{V})) - \tilde{B}_4 \right| \leq \sqrt{8d_\bullet z_{\alpha/2}} \|\tilde{\Lambda}^{-2}\|_F \right\}
\]

where \(\tilde{B}_4\) is defined in (4.3) and \(d_\bullet = d_1^- + d_2^-\) and \(\Delta = d_1^+ - d_2^-\) and \(z_\alpha\) denotes the critical value of the standard normal distribution, i.e., \(z_\alpha = \Phi^{-1}(1 - \alpha)\). The following theorem follows immediately from Theorem 10.

**Theorem 11.** Suppose that conditions in Theorem 6 hold. Then, for any \(\alpha \in (0, 1)\) and \(t \geq 0\), we get

\[
\mathbb{P}((U, V) \in \mathcal{M}_\alpha(\hat{U}, \hat{V})) - (1 - \alpha) \leq C_1 \frac{\sqrt{r}}{\lambda_r^2 \|\Lambda^{-2}\|_F} \cdot (r d_{\max} t)^{1/2} + 2e^{-c_1 d_{\max}} + C_2 \left( \frac{\|\Lambda^{-1}\|_F^3}{\|\Lambda^{-2}\|_F^3} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + C_3 \frac{r d_{\max}^{9/2}}{\lambda_r^5} + 2e^{-t}
\]

24
for some absolute constants $C_1, C_2, C_3, c_1 > 0$. Therefore, if condition (4.4) holds, then

$$\lim_{d_1, d_2 \to \infty} \mathbb{P}\left( (U, V) \in \mathcal{M}_\alpha(\hat{U}, \hat{V}) \right) = 1 - \alpha.$$  

By Theorem [11] we conclude that the claimed confidence region $\mathcal{M}_\alpha(\hat{U}, \hat{V})$ attains any pre-determined confidence level $1 - \alpha$ asymptotically. The required SNR condition is that $\lambda_r \gg d_9^{9/16}$ when $r = O(1)$. The confidence region $\mathcal{M}_\alpha(\hat{U}, \hat{V})$ applies the fourth order approximation of $\mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$.

5 Numerical experiments

For all the simulation cases considered below, we choose the rank $r = 6$ and the singular values are set as $\lambda_i = 2^{r-i} \cdot \lambda$ for $i = 1, \cdots, r$ for some positive number $\lambda$. As a result, the signal strength is determined by $\lambda$. The true singular vectors $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$ are computed from the left and right singular subspaces of a $d_1 \times d_2$ Gaussian random matrix.

5.1 Effectiveness of higher order approximations of the bias.

In this subsection, we display the simulating results of approximating $\mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. In Simulation 1, we show the effectiveness of approximating $\mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ by the first order approximations $2d_* \| \Lambda^{-1} \|_F^2$ where $d_* = d_1 + d_2 - 2r$. In comparison, we also show the inefficiency of approximating $\mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ by the first order approximations when $|d_1 - d_2| \gg \min(d_1, d_2)$. In Simulation 2, we illustrate the appealing effectiveness of the higher order approximations for $\mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ when $|d_1 - d_2| \gg \min(d_1, d_2)$.

Simulation 1. In this simulation, we study the accuracy of the first order approximation and its dependence on $\Delta_d = d_1 - d_2$. First, we set $d_1 = d_2 = d$ where $d = 100, 200, 300$ and denote $d_* = d_1 + d_2 - 2r$. The signal strength $\lambda$ is chosen as $30, 30.5, \cdots, 39.5, 40$. For each given $\lambda$, the first order approximation $2d_* \| \Lambda^{-1} \|_F^2$ is recorded. Since we don’t have the complete explicit expressions for $\mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$, we repeat the experiments for 500 times for each $\lambda$. Then, for each $\lambda$, the average of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ based on those
500 repetitions is recorded, which is called the simulated mean value of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \).
We compare the simulated mean of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) and the first order approximation \( 2d_\ast \| \Lambda^{-1} \|_F^2 \), which is displayed in Figure 1(a). Since \( d_1 = d_2 = d \), we know from Remark 4 that the first order approximation has the same efficiency as the fourth order approximation of \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). Therefore, we should expect that the first order approximation is close to the simulated mean, which is indeed confirmed in Figure 1(a). Second, we set \( d_1 = \frac{d_2}{2} = d \) for \( d = 100, 200, 300 \). As a result, \( \Delta_d = d_2 - d_1 = d \) which is significantly large. Similar experiments are conducted for the same signal strength \( \lambda \) as in the previous settings.

We also compare the simulated mean \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \), based on 500 realizations for each \( \lambda \), with the first order approximations \( 2d_\ast \| \Lambda^{-1} \|_F^2 \). The results are displayed in Figure 1(b) which clearly shows that the first order approximation is insufficient to estimate \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). Therefore, if \( |d_1 - d_2| \gg 0 \), we need the higher order bias corrections to better estimate \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \).

**Simulation 2.** In this simulation, we study the effects of the higher order bias corrections when \( |d_1 - d_2| \gg 0 \). More specifically, we choose \( d_1 = 500 \) and \( d_2 = 1000 \). The signal strength \( \lambda = 50, 51, \cdots, 60 \). For each \( \lambda \), we repeat the experiments for 500 times producing 500 realizations of \( \text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \) whose average is recorded as the simulated mean \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). Meanwhile, for each \( \lambda \), we also record the 1st order approximation \( 2d_\ast \| \Lambda^{-1} \|_F^2 \), the 2nd order approximation \( 2(d_\ast \| \Lambda^{-1} \|_F^2 - \Delta_1 \| \Lambda^{-2} \|_F^2) \), the 3rd order approximation \( 2(d_\ast \| \Lambda^{-1} \|_F^2 - \Delta_2 \| \Lambda^{-2} \|_F^2 + d_\ast \Delta_2 \| \Lambda^{-3} \|_F^2) \) and the 4th order approximation \( 2(d_\ast \| \Lambda^{-1} \|_F^2 - \Delta_3 \| \Lambda^{-2} \|_F^2 + d_\ast \Delta_3 \| \Lambda^{-3} \|_F^2 - (d_{1\_}^2 + d_{2\_}^2 + d_{1\_d_2\_}) \Delta_4 \| \Lambda^{-4} \|_F^2) \) where \( \Delta_d = d_1 - d_2, d_{1\_} = d_1 - r \) and \( d_{2\_} = d_2 - r \). All the results are displayed in Figure 2. The figure shows that the higher order bias corrections indeed improve the accuracy of estimating \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \). It also illustrates that the 1st and 3rd order approximations over-estimate \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \); while, the 2nd and 4th order approximations underestimate \( \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \).
(a) First order approximation \(2 d_\ast \| \Lambda^{-1} \|_F^2 \) is accurate when \( \Delta_d = d_1 - d_2 = 0 \). Here \( d_\ast = d_1 + d_2 - 2r \) and the rank \( r = 6 \). There is no need for the higher order bias corrections.

(b) First order approximation \(2 d_\ast \| \Lambda^{-1} \|_F^2 \) is not sufficiently accurate when \( |d_1 - d_2| \gg 0 \). Here \( d_\ast = d_1 + d_2 - 2r \) and rank \( r = 6 \). The higher order bias corrections are indeed necessary.

Figure 1: Comparison between \( \mathbb{E} \text{dist}^2([\hat{U}, \hat{V}]), ([U, V]) \) and the first order approximation \(2 d_\ast \| \Lambda^{-1} \|_F^2 \). It shows that the accuracy of the first order approximation depends on the dimension difference \( \Delta_d = d_1 - d_2 \). Here the red curves represent the simulated mean \( \mathbb{E} \text{dist}^2([\hat{U}, \hat{V}]), ([U, V]) \) based on 500 realizations of \( \text{dist}^2([\hat{U}, \hat{V}]), ([U, V]) \). The blue curve are the theoretical first order approximations \(2 d_\ast \| \Lambda^{-1} \|_F^2 \) based on Lemma \([1]\). The above left figure shows that the first order approximation is accurate if \( d_1 = d_2 \) which is consistent with Lemma \([2]\) because the higher order bias corrections depend on \( \Delta_d = d_1 - d_2 \).
Figure 2: The higher order approximations of $E_{dist^2}[(\hat{U}, \hat{V}), (U, V)]$. The simulated mean represents the $E_{dist^2}[(\hat{U}, \hat{V}), (U, V)]$ estimated by the average of 500 realizations of $dist^2((\hat{U}, \hat{V}), (U, V))$. The 1st order approximation is $2d_\star \|\Lambda^{-1}\|_F^2$; the 2nd order approximation is $2(d_\star \|\Lambda^{-1}\|_F^2 - \Delta d^2 \|\Lambda^{-2}\|_F^2 + d_1 \Delta d^2 \|\Lambda^{-3}\|_F^2)$ and the 4th order approximation is $2(d_\star \|\Lambda^{-1}\|_F^2 - \Delta d^2 \|\Lambda^{-2}\|_F^2 + d_1 \Delta d^2 \|\Lambda^{-3}\|_F^2 - (d_2^2 - d_1^2) \Delta d^2 \|\Lambda^{-3}\|_F^2)$ where $\Delta d = d_1 - d_2$, $d_{1-} = d_1 - r$, $d_{2-} = d_2 - r$ and $d_\star = d_{1-} + d_{2-}$. Here the rank $r = 6$. It shows that the 3-rd and 4th order approximations are already close to the simulated mean. It also shows that the 1st and 3rd order approximations over-estimate $E_{dist^2}[(\hat{U}, \hat{V}), (U, V)]$; whereas the 2nd and 4th order approximations underestimate $E_{dist^2}[(\hat{U}, \hat{V}), (U, V)]$. 

28
5.2 Normal approximation with data-dependent bias corrections

In this subsection, we present the simulation results for the normal approximations of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ with various bias corrections and normalization factors. In Simulation 3 and Simulation 4, the goal is to show that when $d_1 \approx d_2$, the first order bias correction is enough for the normal approximation. Meanwhile, they also show the importance of applying the shrinkage estimators of the singular values as defined in eq. (4.2), because the empirical singular values $\hat{\lambda}_j$ over-estimates the true singular values $\lambda_j$. In Simulation 5, we demonstrate the necessity of applying higher order bias corrections to guarantee the normal approximations when $|d_1 - d_2| \gg 0$. Indeed, when $|d_1 - d_2| \gg \min(d_1, d_2)$, the results illustrate that the first and second order bias corrections are not sufficient.

Simulation 3. We now show the normal approximation of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star \|\hat{\Lambda}^{-1}\|_F^2$ in the case $d_1 = d_2 = 100$ and $r = 6$. Since $d_1 = d_2$, we only apply the 1st order approximation of $\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Here, $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ denotes the top-$r$ empirical singular values of $\hat{M}$. The signal strength $\lambda = 25, 50, 65, 75$. For each $\lambda$, we record $(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star \|\hat{\Lambda}^{-1}\|_F^2) / \sqrt{8d_\star \|\hat{\Lambda}^{-2}\|_F}$ from 5000 independent experiments and the density histogram is displayed. The density histogram is compared with the probability density function of the standard normal distribution shown as the red curve. The results are presented in Figure 3. Since we are using $\hat{\Lambda}$ to estimate $\Lambda$ where, as shown in Lemma 5, each $\hat{\lambda}_j$ over-estimates the true $\lambda_j$, the bias correction $2d_\star \|\hat{\Lambda}^{-1}\|_F^2$ is not sufficiently significant. Therefore, we observe that the density histograms shift rightward compared with the standard normal curve, especially when signal strength $\lambda$ is moderately strong.

Simulation 4. We show the normal approximation of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star \|\hat{\Lambda}^{-1}\|_F^2$ in the case $d_1 = d_2 = 100$ and $r = 6$. Again, we only apply the 1st order approximation of $\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Here, $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ denotes the top-$r$ shrinkage estimators of $\lambda_j$s as in (4.2). The signal strength $\lambda = 25, 50, 65, 75$. For each $\lambda$, we record $(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star \|\hat{\Lambda}^{-1}\|_F^2) / \sqrt{8d_\star \|\hat{\Lambda}^{-2}\|_F}$ from 5000 independent experiments and the density histogram is displayed. The density histogram is compared with the probability density function of the standard normal distribution as the red curve. The results are
Figure 3: Normal approximation of \( \frac{\text{dist}^2[(\hat{U}, \hat{V}),(U,V)] - 2d_1\lambda^{-1}\|\hat{\Lambda}\|_F}{\sqrt{8d_1}\|\Lambda^{-1}\|_F} \) with \( d_1 = d_2 = 100 \) and \( r = 6 \). The density histogram is based on 5000 realizations from independent experiments. The empirical singular values \( \hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r) \) are calculated from the data matrix \( \hat{M} \). The red curve presents the p.d.f. of standard normal distributions. Since \( d_1 = d_2 \), we apply only first order bias corrections to \( \text{dist}^2[(\hat{U}, \hat{V}),(U,V)] \). As shown in Lemma 5, \( \hat{\lambda}_j \) over-estimates \( \lambda_j \) which explains why the density histogram shifts rightward compared with the standard normal curve, especially when signal strength \( \lambda \) is not significantly strong.
shown in Figure 4. In comparison with Simulation 3 and Figure 3 where \( \hat{\Lambda} \) is used instead of \( \tilde{\Lambda} \), we conclude that \( 2d_\star \|\tilde{\Lambda}^{-1}\|_F^2 \) is more accurate than \( 2d_\star \|\hat{\Lambda}^{-1}\|_F^2 \) for the bias corrections. Indeed, we see that the normal approximation of 
\[
\frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2d_\star \|\hat{\Lambda}^{-1}\|_F^2}{\sqrt{8d_\star \|\hat{\Lambda}^{-2}\|_F^2}}
\]
is already satisfactory when the signal strength \( \lambda = 35 \). On the other hand, the normal approximation of 
\[
\frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2d_\star \|\tilde{\Lambda}^{-1}\|_F^2}{\sqrt{8d_\star \|\tilde{\Lambda}^{-2}\|_F^2}}
\]
is satisfactory only when \( \lambda \geq 75 \).

**Simulation 5.** In this simulation, we show the normal approximation of 
\[
\frac{\text{dist}^2((\hat{U}, \hat{V}), (U, V)) - \hat{B}}{\sqrt{8d_\star \|\hat{\Lambda}^{-2}\|_F^2}}
\]
in the case \( d_1 = 100, d_2 = 600 \) and rank \( r = 6 \). Since \( |d_1 - d_2| \gg \min(d_1, d_2) \), the higher order approximations of \( \mathbb{E}\text{dist}^2((\hat{U}, \hat{V}), (U, V)) \) are crucial for the normal approximations. We fixed the signal strength \( \lambda = 50 \). Similarly as in Simulation 4, the density histogram is based on 5000 realizations from independent experiments. We consider all the four different levels of bias corrections where \( \hat{B} \) can denote the 1st, 2nd, 3rd or 4th order approximations of \( \mathbb{E}\text{dist}^2((\hat{U}, \hat{V}), (U, V)) \), denoted by \( \{\hat{B}_k\}_{k=1}^4 \). More specifically,
\[
\hat{B}_1 = 2d_\star \|\tilde{\Lambda}^{-1}\|_F^2, \quad \text{and} \quad \hat{B}_2 = 2(d_\star \|\tilde{\Lambda}^{-1}\|_F^2 - \Delta^2_d \|\tilde{\Lambda}^{-2}\|_F^2)
\]
and
\[
\hat{B}_3 = 2(d_\star \|\tilde{\Lambda}^{-1}\|_F^2 - \Delta^2_d \|\tilde{\Lambda}^{-2}\|_F^2 + d_\star \Delta^2_d \|\tilde{\Lambda}^{-3}\|_F^2)
\]
and
\[
\hat{B}_4 = 2(d_\star \|\tilde{\Lambda}^{-1}\|_F^2 - \Delta^2_d \|\tilde{\Lambda}^{-2}\|_F^2 + d_\star \Delta^2_d \|\tilde{\Lambda}^{-3}\|_F^2 - (d_1^2 + d_2^2 + d_1d_2) \Delta^2_d \|\tilde{\Lambda}^{-4}\|_F^2).
\]
The results are summarized in Figure 5. This experiment aims to demonstrate the necessity of those higher order bias corrections. Indeed, from the density histograms in Figure 5, we observe that the first and second order bias corrections are not sufficiently strong to guarantee the normal approximations, at least when \( \lambda \leq 50 \), where the density histograms either shift leftward or rightward compared with the standard normal curve. On the other hand, after third or fourth order bias corrections, the normal approximation is very satisfactory at the same level of signal strength \( \lambda = 50 \).
Figure 4: Normal approximation of \( \text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2d_{\star} \| \hat{\Lambda}^{-1} \|_F^2 \) \( \frac{\sqrt{8d_{\star} \| \hat{\Lambda}^{-1} \|_F}}{\sqrt{8d_{\star} \| \hat{\Lambda}^{-1} \|_F}} \) with \( d_1 = d_2 = 100 \) and \( r = 6 \). The density histogram is based on 5000 realizations from independent experiments. The shrinkage estimators of singular values \( \hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r) \) are calculated as eq. (4.2). The red curve presents the p.d.f. of standard normal distributions. Since \( d_1 = d_2 \), we apply only the first order bias corrections to \( \text{dist}^2((\hat{U}, \hat{V}), (U, V)) \). In comparison with Simulation 3 and Figure 3 where \( \hat{\Lambda} \) is used instead of \( \hat{\Lambda} \), we conclude that \( 2d_{\star} \| \hat{\Lambda}^{-1} \|_F^2 \) is more accurate than \( 2d_{\star} \| \hat{\Lambda}^{-1} \|_F^2 \) for the bias corrections. Indeed, we observe that the normal approximation of \( \text{dist}^2((\hat{U}, \hat{V}), (U, V)) - 2d_{\star} \| \hat{\Lambda}^{-1} \|_F^2 \) \( \frac{\sqrt{8d_{\star} \| \hat{\Lambda}^{-1} \|_F}}{\sqrt{8d_{\star} \| \hat{\Lambda}^{-1} \|_F}} \) is already satisfactory when the signal strength \( \lambda = 35 \).
\[ B_1 = 2 d \| \tilde{\Lambda}^{-1} \|_F^2 \]

\[ B_2 = 2 \left( d \| \tilde{\Lambda}^{-1} \|_F^2 - \Delta_2^2 \| \tilde{\Lambda}^{-2} \|_F^2 \right) \]

\[ B_3 = 2 \left( d_1 \| \tilde{\Lambda}^{-1} \|_F^2 - \Delta_3^2 \| \tilde{\Lambda}^{-3} \|_F^2 \right) \]

\[ B_4 = 2 \left( d_1 \| \tilde{\Lambda}^{-1} \|_F^2 - \Delta_4^2 \| \tilde{\Lambda}^{-4} \|_F^2 \right) - \left( d_{1-}^2 + d_{2-}^2 + d_{1-} d_{2-} \right) \Delta_4^2 \| \tilde{\Lambda}^{-2} \|_F^2 \]

Figure 5: Normal approximation of \( \frac{\text{dist}_2[\hat{U}, \hat{V}; (U, V)] - \hat{B}}{\sqrt{8d_e \| \tilde{\Lambda}^{-2} \|_F^2}} \) with higher order bias corrections when \( d_1 = 100, d_2 = 600 \) and \( r = 6 \). The density histogram is based on 5000 realizations from independent experiments. The shrinkage estimators of singular values \( \tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r) \) are calculated as eq. (4.2). The red curve presents the p.d.f. of standard normal distributions. Since \( |d_1 - d_2| \gg \min(d_1, d_2) \), this experiment aims to demonstrate the necessity of those higher order bias corrections. The bias correction \( \hat{B} \) can be the 1st, 2nd, 3rd and 4th order bias corrections. The above figures clearly affirm that the first and second order bias corrections are not sufficiently strong for ensuring the normal approximations.
6 Acknowledgement

The author would like to thank Yik-Man Chiang for insightful recommendations on applying Residue theorem, and Jeff Yao for encouragement on improving the former results.

7 Proofs

7.1 Proof of Theorem 1

For notational simplicity, we assume that all $\lambda_i > 0$ for $1 \leq i \leq r$, i.e., the matrix $A$ is positive semidefinite. The proof is almost identical when $A$ has negative eigenvalues.

Since $A$ is positively semidefinite, we have $\min_{1 \leq i \leq r} |\lambda_i| = \lambda_r$. Therefore, the condition in Theorem 1 is equivalent to $\lambda_r > 2\|X\|$. Recall that we denote by $\{\hat{\lambda}_i, \hat{\theta}_i\}_{i=1}^d$ the singular value and singular vectors of $\hat{A}$. Define the following contour plot $\gamma_A$ on the complex plane (shown as in Figure 6):

![Contour Plot](image)

Figure 6: The contour plot $\gamma_A$ which includes $\{\hat{\lambda}_i, \lambda_i\}_{i=1}^r$ leaving out 0 and $\{\hat{\lambda}_i\}_{i=r+1}^d$.

, where the contour $\gamma_A$ is chosen such that $\min_{\eta \in \gamma_A} \min_{1 \leq i \leq r} |\eta - \lambda_i| = \frac{\lambda_r}{2}$.

Recall from Weyl’s lemma that $\max_{1 \leq i \leq r} |\hat{\lambda}_i - \lambda_i| \leq \|X\|$. We observe that, when $\|X\| < \frac{\lambda_r}{2}$, all $\{\hat{\lambda}_i\}_{i=1}^r$ are inside the contour $\gamma_A$ while 0 and $\{\hat{\lambda}_i\}_{i=r+1}^d$ are outside of the contour $\gamma_A$. By Cauchy’s integral formula, we get

$$
\frac{1}{2\pi i} \oint_{\gamma_A} (\eta I - \hat{A})d\eta = \sum_{i=1}^r \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{\eta - \hat{\lambda}_i}(\hat{\theta}_i\hat{\theta}_i^T) + \sum_{i=r+1}^d \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{\eta - \lambda_i}(\hat{\theta}_i\hat{\theta}_i^T)
= \sum_{i=1}^r \hat{\theta}_i\hat{\theta}_i^T = \hat{\Theta}\hat{\Theta}^T.
$$

34
As a result, we could write

$$\hat{\Theta} \hat{\Theta}^T = \frac{1}{2\pi i} \oint_{\gamma_A} (\eta I - \hat{A})^{-1} d\eta.$$ 

Note that

$$(\eta I - \hat{A})^{-1} = (\eta I - A - X)^{-1} = [(\eta I - A)(I - R_A(\eta) X)]^{-1} = (I - R_A(\eta) X)^{-1} R_A(\eta)$$

where $R_A(\eta) := (\eta I - A)^{-1}$. Observe that

$$\|R_A(\eta) X\| \leq \|R_A(\eta)\| \|X\| < 1.$$ 

Therefore, we write the Neumann series:

$$(I - R_A(\eta) X)^{-1} = I + \sum_{j \geq 1} [R_A(\eta) X]^j. \quad (7.1)$$

By (7.1), we could write $\hat{\Theta} \hat{\Theta}^T$ more explicitly as

$$\hat{\Theta} \hat{\Theta}^T = \frac{1}{2\pi i} \oint_{\gamma_A} (\eta I - A)^{-1} d\eta$$

$$= \frac{1}{2\pi i} \oint_{\gamma_A} R_A(\eta) d\eta + \sum_{j \geq 1} \frac{1}{2\pi i} \oint_{\gamma_A} [R_A(\eta) X]^j R_A(\eta) d\eta.$$ 

Clearly, $\frac{1}{2\pi i} \oint_{\gamma_A} R_A(\eta) d\eta = \Theta \Theta^T$, we end up with

$$\hat{\Theta} \hat{\Theta}^T - \Theta \Theta^T = S_A(X) := \sum_{j \geq 1} \frac{1}{2\pi i} \oint_{\gamma_A} [R_A(\eta) X]^j R_A(\eta) d\eta.$$ 

Now, for all $k \geq 1$, we define

$$S_{A,k}(X) = \frac{1}{2\pi i} \oint_{\gamma_A} [R_A(\eta) X]^k R_A(\eta) d\eta \quad (7.2)$$

which essentially corresponds to the $k$-th order perturbation. Therefore, we obtain

$$\hat{\Theta} \hat{\Theta}^T - \Theta \Theta^T = \sum_{k \geq 1} S_{A,k}(X). \quad (7.3)$$
By (7.3), it suffices to derive explicit expression formulas for all \( \{S_{A,k}(X)\}_{k \geq 1} \). Before proving the formulas for general \( k \), let us derive \( S_{A,k}(X) \) for \( k = 1, 2, 3 \) to interpret the shared styles where we simply apply the Cauchy integral formula.

To this end, we denote \( I_r \) the \( r \times r \) identity matrix and write

\[
R_A(\eta) = \Theta(\eta \cdot I_r - \Lambda)^{-1}\Theta^T - \eta^{-1}\Theta\Theta^T = \sum_{j=1}^{d} \frac{1}{\eta - \lambda_j}\theta_j\theta_j^T
\]

where we set \( \lambda_j = 0 \) for all \( r + 1 \leq j \leq d \). For notational simplicity, we denote \( P_j = \theta_j\theta_j^T \) for all \( 1 \leq j \leq d \). Therefore, \( P_j \) represents the spectral projector onto \( \theta_j \).

**Derivation of \( S_{A,1}(X) \).** By definitions of \( S_{A,1}(X) \), we write

\[
S_{A,1}(X) = \frac{1}{2\pi i} \oint_{\gamma_A} R_A(\eta)X\partial R_A(\eta)d\eta
\]

\[
= \sum_{j_1=1}^{d} \sum_{j_2=1}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1}XP_{j_2}.
\]

(7.4)

**Case 1:** both \( j_1 \) and \( j_2 \) are greater than \( r \). In this case, the contour integral in (7.4) is zero by Cauchy integral formula.

**Case 2:** only one of \( j_1 \) and \( j_2 \) is greater than \( r \). W.L.O.G, let \( j_2 > r \), we get

\[
\sum_{j_1=1}^{r} \sum_{j_2>r} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{\eta^{-1}d\eta}{\eta - \lambda_{j_1}} P_{j_1}XP_{j_2}
\]

\[
= \sum_{j_1=1}^{r} \sum_{j_2>r} \lambda_{j_1}^{-1} P_{j_1}XP_{j_2} = P^{-1}XP^{-1}.
\]

**Case 3:** none of \( j_1 \) and \( j_2 \) is greater than \( r \). In this case, the contour integral in (7.4) is zero by Cauchy integral formula.

To sum up, we conclude that \( S_{A,1}(X) = P^{-1}XP^{-1} + P^{\perp}XP^{\perp} \).

**Derivation of \( S_{A,2}(X) \).** By definition of \( S_{A,2}(X) \), we write

\[
S_{A,2}(X) = \frac{1}{2\pi i} \oint_{\gamma_A} R_A(\eta)X\partial R_A(\eta)X\partial R_A(\eta)d\eta
\]

36
Derivation of Case 3 by Cauchy integral formula.

Case 1: all \( j_1, j_2, j_3 \) are greater than \( r \). In this case, the contour integral in (7.5) is zero by Cauchy integral formula.

Case 2: two of \( j_1, j_2, j_3 \) are greater than \( r \). W.L.O.G., let \( j_1 \leq r \) and \( j_2, j_3 > r \), we get

\[
\begin{align*}
\sum_{j_1=1}^{r} \sum_{j_2 \neq 1, j_3 > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2} X P_{j_3} \\
= \sum_{j_1=1}^{r} \sum_{j_2 \neq 1, j_3 > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{\lambda_{j_1}^2} P_{j_1} X P_{j_2} X P_{j_3} = \mathbb{P}^{-2} X \mathbb{P}^\perp X \mathbb{P}^\perp.
\end{align*}
\]

Case 3: one of \( j_1, j_2, j_3 \) is greater than \( r \). W.L.O.G., let \( j_1, j_2 \leq r \) and \( j_3 > r \), we get

\[
\begin{align*}
\sum_{j_1=1}^{r} \sum_{j_2 \neq 1, j_3 > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2} X P_{j_3} \\
= \sum_{j_1=1}^{r} \sum_{j_2 \neq 1, j_3 > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{\lambda_{j_1}^2} P_{j_1} X P_{j_2} X P_{j_3} \\
+ \sum_{j_1 \neq j_2 \neq 1, j_3 > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2} X P_{j_3} \\
= - \sum_{j_1=1}^{r} \sum_{j_2 \neq 1, j_3 > r}^{d} \lambda_{j_1}^{-2} P_{j_1} X P_{j_2} X P_{j_3} - \sum_{j_1 \neq j_2 \neq 1, j_3 > r}^{d} (\lambda_{j_1} \lambda_{j_2})^{-1} P_{j_1} X P_{j_2} X P_{j_3} \\
= - \mathbb{P}^{-1} X \mathbb{P}^{-1} X \mathbb{P}^\perp.
\end{align*}
\]

Case 4: none of \( j_1, j_2, j_3 \) is greater than \( r \). In this case, the contour integral in (7.5) is zero by Cauchy integral formula.

To sum up, we conclude that

\[
S_{A,2}(X) = (\mathbb{P}^{-2} X \mathbb{P}^\perp X \mathbb{P}^\perp + \mathbb{P}^\perp X \mathbb{P}^{-2} X \mathbb{P}^\perp + \mathbb{P}^\perp X \mathbb{P}^\perp X \mathbb{P}^{-2}) \\
- (\mathbb{P}^\perp X \mathbb{P}^{-1} X \mathbb{P}^{-1} + \mathbb{P}^{-1} X \mathbb{P}^\perp X \mathbb{P}^{-1} + \mathbb{P}^{-1} X \mathbb{P}^{-1} X \mathbb{P}^\perp).
\]

Derivation of \( S_{A,3}(X) \). By definition of \( S_{A,3} \), we get

\[
S_{A,3} = \frac{1}{2\pi i} \oint_{\gamma_A} [\mathcal{R}_A(\eta)X]^3 \mathcal{R}_A(\eta) d\eta
\]
\[
= \sum_{j_1=1}^{d} \sum_{j_2=1}^{d} \sum_{j_3=1}^{d} \sum_{j_4=1}^{d} \frac{1}{2\pi i} \oint_{\gamma M} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})(\eta - \lambda_{j_3})(\eta - \lambda_{j_4})} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4}.
\]

(7.6)

**Case 1:** all \( j_1, j_2, j_3, j_4 \) are greater than \( r \). The contour integral in (7.6) is zero by Cauchy integral formula.

**Case 2:** three of \( j_1, j_2, j_3, j_4 \) are greater than \( r \). W.L.O.G., let \( j_1 \leq r \) and \( j_2, j_3, j_4 > r \), we get

\[
\sum_{j_1=1}^{r} \sum_{j_2,j_3,j_4>r} \frac{1}{2\pi i} \oint_{\gamma A} \frac{\eta^{-3} d\eta}{\eta - \lambda_{j_1}} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4}
\]

\[
= \sum_{j_1=1}^{r} \sum_{j_2,j_3,j_4>r} \frac{1}{\lambda_{j_1}^3} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4} = \Psi^{-3}\varphi^T \varphi^T \varphi^T \varphi^T.
\]

**Case 3:** two of \( j_1, j_2, j_3, j_4 \) are greater than \( r \). W.L.O.G., let \( j_1, j_2 \leq r \) and \( j_3, j_4 > r \), we get

\[
\sum_{j_1,j_2 \leq r} \sum_{j_3,j_4 > r} \frac{1}{2\pi i} \oint_{\gamma A} \frac{\eta^{-2} d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4}
\]

\[
= \sum_{j_1,j_2 \leq r} \sum_{j_3,j_4 > r} \frac{1}{\lambda_{j_1}^2 \lambda_{j_2}} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4} + \sum_{j_1,j_2 \leq r} \sum_{j_3,j_4 > r} \frac{1}{\lambda_{j_1} \lambda_{j_2}^2} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4}
\]

\[
= \left( \Psi^{-2} \Xi^{-1} \Xi \Xi \Xi + \Psi^{-1} \Xi^{-2} \Xi \Xi \Xi \right).
\]

**Case 4:** one of \( j_1, j_2, j_3, j_4 \) is greater than \( r \). W.L.O.G, let \( j_1, j_2, j_3 \leq r \) and \( j_4 > r \), we get

\[
\sum_{j_1,j_2 \geq r} \sum_{j_3 \geq r, j_4 > r} \frac{1}{2\pi i} \oint_{\gamma A} \frac{\eta^{-1} d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})(\eta - \lambda_{j_3})} P_{j_1} X P_{j_2} X P_{j_3} X P_{j_4}
\]

\[
= \sum_{j_1=1}^{r} \frac{1}{\lambda_{j_1}^3} P_{j_1} X P_{j_1} X \Xi + \sum_{j_1=1}^{r} \sum_{j_2 \neq j_3} \sum_{j_4 > r} \frac{1}{\lambda_{j_1} \lambda_{j_2}^2} P_{j_1} X P_{j_2} X P_{j_2} X \Xi
\]

38
+ \sum_{j_2=1}^{r} \sum_{j_1=j_2}^{r} \frac{1}{\lambda_j^{j_2}} \sum_{j_3=1}^{r} \sum_{j_2=j_1 \neq j_3}^{r} \frac{1}{\lambda_j^{j_3}} P_{j_1}XP_{j_2}XP_{j_3}X\mathbb{P}^\perp
+ \sum_{j_1 \neq j_3 \neq 1}^{r} \frac{1}{\lambda_j^{j_1} \lambda_j^{j_2} \lambda_j^{j_3}} P_{j_1}XP_{j_2}XP_{j_3}X\mathbb{P}^\perp
= \mathbb{P}^{-1}X\mathbb{P}^{-1}X\mathbb{P}^{-1}X\mathbb{P}^\perp.

Case 5: none of $j_1, j_2, j_3, j_4$ is greater than $r$. In this case, the contour integral in (7.6) is zero by Cauchy integral formula.

To sum up, we conclude that

$$ S_{A,3}(X) = \sum_{s:s_1 + \cdots + s_4 = 3} (-1)^{1+r(s)} \cdot \mathbb{P}^{-s_1}X\mathbb{P}^{-s_2}X\mathbb{P}^{-s_3}X\mathbb{P}^{-s_4} $$

where we set $\mathbb{P}^0 = \mathbb{P}^\perp$.

**Derivation of $S_{A,k}(X)$ for general $k$.** Recall the definition of $S_{A,k}(X)$, we write

$$ S_{A,k}(X) = \sum_{j_1, \ldots, j_{k+1} \geq 1}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \left( \prod_{i=1}^{k+1} \frac{1}{\eta - \lambda_{j_i}} \right) d\eta P_{j_1}XP_{j_2}X \cdots P_{j_{k+1}}X \mathbb{P}^\perp. \quad (7.7) $$

We consider components of summations in (7.7). For instance, consider the cases that some $k_1$ indices from $\{j_1, \ldots, j_{k+1}\}$ are not larger than $r$. W.L.O.G., let $j_1, \ldots, j_{k_1} \leq r$ and $j_{k_1+1}, \ldots, j_{k+1} > r$. By Cauchy integral formula, the integral in (7.7) is zero if $k_1 = 0$ or $k_1 = k + 1$. Therefore, we only focus on the cases that $1 \leq k_1 \leq k$. Then,

$$ \sum_{j_1, \ldots, j_{k_1} \geq 1}^{r} \sum_{j_{k_1+1} > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \left( \prod_{i=1}^{k_1} \frac{1}{\eta - \lambda_{j_i}} \right) \eta^{k_1-k-1} d\eta P_{j_1}XP_{j_2}X \cdots P_{j_{k_1}}X \mathbb{P}^\perp
= \sum_{j_1, \ldots, j_{k_1} \geq 1}^{r} \frac{1}{2\pi i} \oint_{\gamma_A} \left( \prod_{i=1}^{k_1} \frac{1}{\eta - \lambda_{j_i}} \right) \eta^{k_1-k-1} d\eta P_{j_1}XP_{j_2}X \cdots P_{j_{k_1}}X \mathbb{P}^\perp$$

Recall that our goal is to prove

$$ S_{A,k}(X) = \sum_{s:s_1 + \cdots + s_{k+1} = k} (-1)^{1+r(s)} \cdot \mathbb{P}^{-s_1}X\mathbb{P}^{-s_2}X \cdots \mathbb{P}^{-s_{k+1}}. $$

39
Accordingly, in the above summations, we consider the components, where \( s_1, \ldots, s_k \geq 1 \) and \( s_{k+1} = \cdots = s_{k+1} = 0 \), namely,

\[
\sum_{s_1 + \cdots + s_k = k \atop s_j \geq 1} (-1)^{k+1} \frac{1}{s_1 \cdots s_k} X \cdots X^{s_k} X^{\perp} \cdots X^{\perp}.
\]

It turns out that we need to prove

\[
\sum_{j_1, \ldots, j_k} \frac{1}{2\pi i} \oint_{\gamma_A} \left( \prod_{i=1} \frac{1}{\eta - \lambda_{j_i}} \right) \eta^{k-1} \, d\eta P_{j_1} X P_{j_2} X \cdots P_{j_k} = \sum_{j_1, \ldots, j_k} (-1)^{k+1} \frac{1}{\lambda_{j_1}^{s_1} \cdots \lambda_{j_k}^{s_k}} P_{j_1} X P_{j_2} X \cdots P_{j_k}.
\]

It suffices to prove that for all \( j = (j_1, \ldots, j_k) \in \{1, \ldots, r\}^k \),

\[
\frac{1}{2\pi i} \oint_{\gamma_A} \left( \prod_{i=1} \frac{1}{\eta - \lambda_{j_i}} \right) \frac{d\eta}{\eta^{k+1}} = \sum_{s_1 + \cdots + s_k = k \atop s_j \geq 1} (-1)^{k+1} \frac{1}{\lambda_{j_1}^{s_1} \cdots \lambda_{j_k}^{s_k}}. \tag{7.8}
\]

To prove (7.8), we rewrite its right hand side. Given any \( j = (j_1, \ldots, j_k) \in \{1, \ldots, r\}^k \), define

\[
v_i(j) := \{1 \leq t \leq k : j_t = i\} \quad \text{for } 1 \leq i \leq r
\]

, that is, \( v_i(j) \) contains the location \( s \) such that \( \lambda_{j_s} = \lambda_i \). Meanwhile, denote \( v_i(j) = \text{Card}(v_i(j)) \). Then, the right hand side of (7.8) is written as

\[
\sum_{s_1 + \cdots + s_k = k \atop s_j \geq 1} (-1)^{k+1} \frac{1}{\lambda_{j_1}^{s_1} \cdots \lambda_{j_k}^{s_k}} = (-1)^{k+1} \sum_{s_1 + \cdots + s_k = k \atop s_j \geq 1} \lambda_i^{-\sum_{p \in v_i(j)} s_p} \cdots \lambda_r^{-\sum_{p \in v_r(j)} s_p}.
\]

Now, we denote \( t_i(j) = \sum_{p \in v_i(j)} s_p \) for \( 1 \leq i \leq r \), we can write the above equation as

\[
\sum_{s_1 + \cdots + s_k = k \atop s_j \geq 1} (-1)^{k+1} \frac{1}{\lambda_{j_1}^{s_1} \cdots \lambda_{j_k}^{s_k}} = (-1)^{k+1} \sum_{s_1 + \cdots + s_k = k \atop s_j \geq 1} \prod_{t_i(j) \geq v_i(j)} \left( \frac{t_i(j) - 1}{v_i(j) - 1} \right) \lambda_i^{t_i(j)}
\]

\[
\quad \quad \quad = (-1)^{k+1} \sum_{t_1(j) + \cdots + t_r(j) = k-k \atop t_i(j) \geq v_i(j)} \prod_{t_i(j) \geq v_i(j)} \left( \frac{t_i(j) + v_i(j) - 1}{v_i(j) - 1} \right) \lambda_i^{t_i(j)-v_i(j)}
\]

40
where the last equality is due to the fact $v_1(j) + \cdots + v_r(j) = k_1$. Similarly, the left hand side of (7.8) can be written as
\[
\frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_1) \cdots (\eta - \lambda_{j_k}) \eta^{k+1-k_1}} = \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_1)^{v_1(j)} \cdots (\eta - \lambda_{j_r})^{v_r(j)} \eta^{k+1-k_1}}.
\]
Therefore, in order to prove (7.8), it suffices to prove that for any $j = (j_1, \cdots, j_{k_1})$ the following equality holds
\[
\frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_1)^{v_1} \cdots (\eta - \lambda_{j_r})^{v_r} \eta^{k+1-k_1}} = (-1)^{k_1+1} \sum_{t_1+\cdots+t_r=k_1 \atop t_i=0 \; \text{if} \; v_i=0} \prod_{i:v_i \geq 1} (t_i + v_i - 1) \lambda_i^{t_i-v_i}
\]
where we omitted the index $j$ in definitions of $v_i(j)$ and $t_i(j)$ without causing any confusions.

The non-negative numbers $v_1 + \cdots + v_r = k_1$. We define the function
\[
\varphi(\eta) = \frac{1}{(\eta - \lambda_1)^{v_1} \cdots (\eta - \lambda_{j_r})^{v_r} \eta^{k+1-k_1}}
\]
and we will calculate $\frac{1}{2\pi i} \oint_{\gamma_A} \varphi(\eta)d\eta$ by Residue theorem. Indeed, by Residue theorem,
\[
\frac{1}{2\pi i} \oint_{\gamma_A} \varphi(\eta)d\eta = -\text{Res}(\varphi, \eta = \infty) - \text{Res}(\varphi, \eta = 0).
\]
Clearly, $\text{Res}(\varphi, \eta = \infty) = 0$ and it suffices to calculate $\text{Res}(\varphi, \eta = 0)$. To this end, let $\gamma_0$ be a contour plot around 0 where none of $\{\lambda_k\}_{k=1}^r$ is inside it. Then,
\[
\text{Res}(\varphi, \eta = 0) = \frac{1}{2\pi i} \oint_{\gamma_0} \varphi(\eta)d\eta.
\]
By Cauchy integral formula, we obtain
\[
\text{Res}(\varphi, \eta = 0) = \frac{1}{(k-k_1)!} \left[ \prod_{i:v_i \geq 1} (\eta - \lambda_i)^{-v_i} \right]_{\eta=0}^{(k-k_1)}
\]
where we denote by $f(x)^{(k-k_1)}$ the $k - k_1$-th order differentiation of $f(x)$. Then, we use general Leibniz rule and get
\[
\text{Res}(\varphi, \eta = 0) = \frac{1}{(k-k_1)!} \sum_{t_1+\cdots+t_r=k-k_1 \atop t_i=0 \; \text{if} \; v_i=0} \frac{(k-k_1)!}{t_1!t_2!\cdots t_r!} \left[ \prod_{i:v_i \geq 1} (\eta - \lambda_i)^{-v_i} \right]_{\eta=0}^{(t_i)}
\]

\[
= (-1)^{k-k_1} \sum_{t_1+\ldots+t_r=k-k_1 \ i:v_i \geq 1 \ \ t_i=0 \ if \ v_i=0} \prod_{i} \frac{v_i(v_i+1)\cdots(v_i+t_i-1)}{t_i!} (-\lambda_i)^{-v_i-t_i}
\]
\[
= (-1)^{k-k_1} \sum_{t_1+\ldots+t_r=k-k_1 \ i:v_i \geq 1 \ \ t_i=0 \ if \ v_i=0} \prod_{i} \left( \frac{t_i+v_i-1}{v_i-1} \right) (-\lambda_i)^{-v_i-t_i}
\]
\[
= (-1)^{2k-k_1} \sum_{t_1+\ldots+t_r=k-k_1 \ i:v_i \geq 1 \ \ t_i=0 \ if \ v_i=0} \prod_{i} \left( \frac{t_i+v_i-1}{v_i-1} \right) \lambda_i^{-v_i-t_i}.
\]
Therefore,
\[
\frac{1}{2\pi i} \oint_{\gamma_a} \varphi(\eta)d\eta = (-1)^{k_1+1} \sum_{t_1+\ldots+t_r=k-k_1 \ i:v_i \geq 1 \ \ t_i=0 \ if \ v_i=0} \prod_{i} \left( \frac{t_i+v_i-1}{v_i-1} \right) \lambda_i^{-v_i-t_i}
\]
which proves (7.9). We conclude the proof of Theorem 1.

7.2 Proof of Theorem 2

By rank(\(\hat{\Theta}\)) = rank(\(\Theta\)) = 2r, we can write
\[
\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] = \|\hat{\Theta}^T - \Theta^T\|_F^2 = 4r - 2\langle \hat{\Theta}^T, \Theta^T \rangle.
\]
We will apply the representation formula of \(\hat{\Theta}^T\) as in Theorem 1. Since \(X\) is random, we shall take care of the operator norm of \(X\). Observe that \(\|X\| = \|Z\|\) where the operator norm of \(Z\) is well-known in random matrix theory (see, e.g., Tao (2012) and Vershynin (2010)). Indeed, there exist some absolute constants \(C_1, C_2, c_1 > 0\) such that
\[
\mathbb{E}\|X\| \leq C_1 d_{\text{max}}^{1/2} \quad \text{and} \quad \mathbb{P}\left(\|X\| \geq C_2 d_{\text{max}}^{1/2}\right) \leq e^{-c_1 d_{\text{max}}}
\]
where \(d_{\text{max}} = \max\{d_1, d_2\}\). Denote the event \(\mathcal{E}_1 := \{\|X\| \leq C_2 d_{\text{max}}^{1/2}\}\) so that \(\mathbb{P}(\mathcal{E}_1) \geq 1 - e^{-c_1 d_{\text{max}}}\). Suppose that \(\lambda_r > 2C_2 d_{\text{max}}^{1/2}\), our analysis is conditioned on the event \(\mathcal{E}_1\). By Theorem 1 on event \(\mathcal{E}_1\), we have
\[
\hat{\Theta}^T = \Theta^T + S_{A,1}(X) + S_{A,2}(X) + \sum_{k \geq 3} S_{A,3}(X)
\]
where $S_{A,1}(X) = \mathcal{P}^{-1}X\mathcal{P}^\perp + \mathcal{P}^\perp X\mathcal{P}^{-1}$ and

$$
S_{A,2}(X) = (\mathcal{P}^{-2}X\mathcal{P}^\perp X\mathcal{P}^\perp + \mathcal{P}^\perp X\mathcal{P}^{-2}X\mathcal{P}^\perp + \mathcal{P}^\perp X\mathcal{P}^\perp X\mathcal{P}^{-2})
$$

$$
- (\mathcal{P}^\perp X\mathcal{P}^{-1}X\mathcal{P}^\perp + \mathcal{P}^{-1}X\mathcal{P}^\perp X\mathcal{P}^{-1} + \mathcal{P}^{-1}X\mathcal{P}^{-1}X\mathcal{P}^\perp).
$$

Therefore, conditioned on $\mathcal{E}_1$, we get

$$
\|\hat{\Theta}\hat{\Theta}^T - \Theta\Theta^T\|_F^2 = 2 \text{tr} (\mathcal{P}^{-1}X\mathcal{P}^\perp X\mathcal{P}^{-1}) - 2 \sum_{k \geq 3} \langle \Theta\Theta^T, S_{A,k}(X) \rangle
$$

$$
= 2\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2 - 2 \sum_{k \geq 3} \langle \Theta\Theta^T, S_{A,k}(X) \rangle.
$$

As a result, we have

$$
\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]
$$

$$
= 2\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2 - 2 \mathbb{E}\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2
$$

$$
- 2 \sum_{k \geq 3} \langle \Theta\Theta^T, S_{A,k}(X) - \mathbb{E}S_{A,k}(X) \rangle.
$$

We will investigate the normal approximation of

$$
\frac{2\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2 - 2 \mathbb{E}\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2}{\sqrt{8(d_1 + d_2 - 2r)} \cdot \|\Lambda^{-2}\|_F}
$$

and show that

$$
\frac{2 \sum_{k \geq 3} \langle \Theta\Theta^T, S_{A,k}(X) - \mathbb{E}S_{A,k}(X) \rangle}{\sqrt{8(d_1 + d_2 - 2r)} \cdot \|\Lambda^{-2}\|_F}
$$

is ignorable when signal strength $\lambda_r$ is sufficiently strong. Now define a function for some $t > 0$ determined later

$$
f_t(X) = 2 \sum_{k \geq 3} \langle \Theta\Theta^T, S_{A,k}(X) \rangle \cdot \phi\left(\frac{\|X\|}{t \cdot d_{max}^{1/2}}\right)
$$

(7.11)

where we view $X$ as a variable in $\mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)}$ and the function $\phi(\cdot): \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined as

$$
\phi(s) :=
\begin{cases}
1 & \text{if } s \leq 1, \\
2 - s & \text{if } 1 < s \leq 2, \\
0 & \text{if } s > 2.
\end{cases}
$$
Clearly, \( \phi(s) \) is Lipschitz with constant 1. The following lemma shows that \( f(\cdot) \) is a Lipschitz function as long as the signal strength is sufficiently strong. The proof of Lemma 6 is in Appendix, Section 8.2.

**Lemma 6.** There exist absolute constants \( C_3, C_4 > 0 \) so that if \( \lambda_r \geq C_3 t^2 d_{\text{max}}^{1/2} \), then

\[
|f_t(X_1) - f_t(X_2)| \leq C_4 t^2 \frac{rd_{\text{max}}}{\lambda_r^2} \cdot \|X_1 - X_2\|_F
\]

where \( f_t(X) \) is defined as in (7.11).

By Lemma 6, we apply Gaussian isoperimetric inequality (see, e.g., Koltchinskii and Lounici (2016)), with probability at least \( 1 - e^{-s} \) for any \( s \geq 1 \),

\[
\left| 2 \sum_{k \geq 3} \langle \Theta^T, S_{A,k}(X) \rangle \cdot \phi\left( \frac{\|X\|}{t \cdot d_{\text{max}}^{1/2}} \right) - 2 \sum_{k \geq 3} \langle \Theta^T, S_{A,k}(X) \rangle \cdot \phi\left( \frac{\|X\|}{C_2 \cdot t \cdot d_{\text{max}}^{1/2}} \right) \right| \leq C_5 \sqrt{st^2 \frac{rd_{\text{max}}}{\lambda_r^4}} \tag{7.12}
\]

for some absolute constant \( C_5 > 0 \). Now, set \( t = C_2 \) where \( C_2 \) is defined in (7.10). Therefore, on event \( E_1 \), we have \( \phi\left( \frac{\|X\|}{C_2 \sqrt{d_{\text{max}}}} \right) = 1 \). Meanwhile, the following fact holds

\[
\left| 2 \sum_{k \geq 3} \langle \Theta^T, S_{A,k}(X) \rangle \cdot \phi\left( \frac{\|X\|}{C_2 \cdot t \cdot d_{\text{max}}^{1/2}} \right) - 2 \sum_{k \geq 3} \langle \Theta^T, S_{A,k}(X) \rangle \cdot \phi\left( \frac{\|X\|}{C_2 \cdot t \cdot d_{\text{max}}^{1/2}} \right) \|_{E_1^c} \right|
\]

\[
\leq \left| 2 \sum_{k \geq 3} \langle \Theta^T, S_{A,k}(X) \rangle \cdot \phi\left( \frac{\|X\|}{C_2 \cdot t \cdot d_{\text{max}}^{1/2}} \right) \|_{E_1^c} \right| + \left| 2 \sum_{k \geq 3} \langle \Theta^T, S_{A,k}(X) \rangle \|_{E_1^c} \right|
\]

\[
\leq 4 \sum_{k \geq 3} \mathbb{E}\|\Theta^T, S_{A,k}(X)\|_{E_1^c} \leq 8r \sum_{k \geq 3} \mathbb{E}^{1/2} \|S_{A,k}(X)\| \cdot e^{-c_1 d_{\text{max}}/2}
\]

\[
\leq e^{-c_1 d_{\text{max}}/2} \cdot 8r \sum_{k \geq 3} \mathbb{E}^{1/2} \frac{16^k \|X\|^{2k}}{\lambda_r^{2k}} \leq e^{-c_1 d_{\text{max}}/2} \cdot 8r \sum_{r \geq 3} \left( \frac{C_6 \cdot d_{\text{max}}^{1/2}}{\lambda_r} \right)^k
\]

\[
\leq e^{-c_1 d_{\text{max}}/2} \cdot \frac{C_6 r d_{\text{max}}^{3/2}}{\lambda_r^2} \leq C_6 \frac{r d_{\text{max}}}{\lambda_r^2}
\]

where the last inequality holds as long as \( e^{-c_1 d_{\text{max}}/2} \leq \frac{1}{\sqrt{d_{\text{max}}}} \) and we used the fact \( \mathbb{E}^{1/p} \|X\|^{p} \leq C_6 d_{\text{max}}^{1/2} \) for some absolute constant \( C_6 > 0 \) and any positive integer \( p \). (See, e.g., Koltchinskii
and Xia (2016), Vershynin (2010) and Tao (2012). Together with (7.12), we conclude that with probability at least $1 - e^{-s} - e^{-c_1 d_{\max}}$ for any $s \geq 1$,

$$
\left| 2 \sum_{k \geq 3} \left\langle \Theta \Theta^T, S_{A,k}(X) \right\rangle - \mathbb{E} 2 \sum_{k \geq 3} \left\langle \Theta \Theta^T, S_{A,k}(X) \right\rangle \right| \leq C_6 s^{1/2} \frac{rd_{\max}}{\lambda^3 r}
$$

for some absolute constant $C_6 > 0$. Therefore, for any $s \geq 1$, with probability at least $1 - e^{-s} - e^{-c_1 d_{\max}}$,

$$
\left| 2 \sum_{k \geq 3} \left\langle \Theta \Theta^T, S_{A,k}(X) \right\rangle - \mathbb{E} 2 \sum_{k \geq 3} \left\langle \Theta \Theta^T, S_{A,k}(X) \right\rangle \right| \leq C_6 s^{1/2} \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F} \cdot \frac{r^{1/2} d_{\max}^{1/2}}{\lambda_r}
$$

where we assumed $d_{\max} \geq 3r$. Note that the bound in (7.13) tends to be ignorable if $\frac{r d_{\max}^{1/2}}{\lambda_r} = o(1)$. We next prove the normal approximation of $2\|\mathcal{P}^{-1} X \mathcal{P}^\perp\|_F^2$. Similarly as in Xia (2018), by definitions of $\mathcal{P}^{-1}, X$ and $\mathcal{P}^\perp$, we could write

$$
\mathcal{P}^{-1} X \mathcal{P}^\perp = \begin{pmatrix} U \Lambda^{-1} V^T Z^T U \perp U^T \perp & 0 \\ 0 & V \Lambda^{-1} U^T Z V \perp V^T \perp \end{pmatrix}.
$$

As a result, we obtain

$$
\|\mathcal{P}^{-1} X \mathcal{P}^\perp\|_F^2 = \|U \Lambda^{-1} V^T Z^T U \perp U^T \perp\|_F^2 + \|V \Lambda^{-1} U^T Z V \perp V^T \perp\|_F^2.
$$

Denote by $z_j \in \mathbb{R}^{d_1}$ the $j$-th column of $Z$ for $1 \leq j \leq d_2$. Then, $z_1, \cdots, z_{d_2}$ are independent Gaussian random vectors and $\mathbb{E} z_j z_j^T = I_{d_1}$ for all $j$. Therefore, we can write

$$
U^T Z = \sum_{j=1}^{d_2} (U^T z_j) e_j^T
$$

where $\{e_j\}_{j=1}^{d_2}$ represent the standard basis vectors in $\mathbb{R}^{d_2}$. Similarly,

$$
U_\perp^T Z = \sum_{j=1}^{d_2} (U_\perp^T z_j) e_j^T.
$$

Observe that $U^T z_j$ and $U_\perp^T z_j$ are Gaussian random vectors and

$$
\mathbb{E} U^T z_j (U_\perp^T z_j)^T = U^T U_\perp = 0
$$

45
implying that \( \{ U^T z_j \}_{j=1}^{d_2} \) are independent with \( \{ U^T z_j \}_{j=1}^{d_2} \). Therefore, \( \| U \Lambda^{-1} V^T Z U \|_F^2 \) is independent with \( \| V \Lambda^{-1} U \|_F^2 \). Denote by \( \tilde{Z} \) an independent copy of \( Z \), we conclude that \( (Y_1 \overset{d}{=} Y_2 \text{ denotes equivalence of } Y_1 \text{ and } Y_2 \text{ in distribution}) \)

\[
\| \mathcal{P}^{-1} X \mathcal{P} \|_F^2 \overset{d}{=} \| U \Lambda^{-1} V^T Z U \|_F^2 + \| V \Lambda^{-1} U \|_F^2 \\
= \sum_{j=r+1}^{d_1} \| U \Lambda^{-1} V^T u_j \|_2^2 + \sum_{j=r+1}^{d_2} \| V \Lambda^{-1} U \tilde{v}_j \|_2^2 \\
= \sum_{j=r+1}^{d_1} \| \Lambda^{-1} V^T u_j \|_2^2 + \sum_{j=r+1}^{d_2} \| \Lambda^{-1} U \tilde{v}_j \|_2^2
\]

where \( \{ u_j \}_{j=r+1}^{d_1} \) and \( \{ v_j \}_{j=r+1}^{d_2} \) represent the columns of \( U \) and \( V \), respectively. Observe that \( Z^T u_j \sim \mathcal{N}(0, I_{d_1}) \) for all \( r+1 \leq j \leq d_1 \) and

\[
\mathbb{E}(Z^T u_{j_1})(Z^T u_{j_2})^T = 0 \text{ for all } r+1 \leq j_1 \neq j_2 \leq d_1.
\]

Therefore, \( \{ Z^T u_j \}_{j=r+1}^{d_1} \) are independent normal random vectors. Similarly, \( \tilde{v}_j \sim \mathcal{N}(0, I_{d_1}) \) are independent for all \( r+1 \leq j \leq d_2 \). Similarly, it is also easy to show that \( V^T Z u_j \sim \mathcal{N}(0, I_r) \) and \( U \tilde{v}_j \sim \mathcal{N}(0, I_r) \) are all independent for \( r+1 \leq j_1 \leq d_1 \) and \( r+1 \leq j_2 \leq d_2 \).

As a result, let \( d_* = d_1 + d_2 - 2r \), we conclude that

\[
\| \mathcal{P}^{-1} X \mathcal{P} \|_F^2 \overset{d}{=} \sum_{j=1}^{d_*} \| \Lambda^{-1} z_j \|_2^2 \quad \text{(7.14)}
\]

where we abuse the notations and denote \( \{ z_j \}_{j=1}^{d_*} \) the i.i.d. Gaussian random vector with \( z_j \sim \mathcal{N}(0, I_r) \). By Berry-Esseen theorem (\cite{Berry1941} and \cite{Esseen1942}), we get, for some absolute constant \( C_7 > 0 \),

\[
\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{2\| \mathcal{P}^{-1} X \mathcal{P} \|_F^2 - 2\mathbb{E}\| \mathcal{P}^{-1} X \mathcal{P} \|_F^2}{\sqrt{8(d_1+d_2-2r)\| \Lambda^{-2} \|_F^2}} \leq x \right) - \Phi(x) \right| \leq C_7 \left( \frac{\| \Lambda^{-1} \|_F^4}{\| \Lambda^{-2} \|_F^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_*}} \quad \text{(7.15)}
\]

where we used the fact \( \text{Var}(\| \Lambda^{-1} z_j \|_2^2) = 2\| \Lambda^{-2} \|_F^2 \) and for some absolute constant \( C_7 > 0 \),

\[
\mathbb{E}\| \Lambda^{-1} z_j \|_2^6 \leq C_7 \sum_{j_1,j_2,j_3 \geq 1} \frac{1}{\lambda_{j_1}^2 \lambda_{j_2}^2 \lambda_{j_3}^2} \leq C_7 \| \Lambda^{-1} \|_F^6.
\]

46
In (7.15), the function $\Phi(x)$ denotes the cumulative distribution function of a standard normal random variable. Recall that, on event $\mathcal{E}_1$,

$$\frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F} = \frac{2\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2 - 2\mathbb{E}\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F} + \frac{2\sum_{k \geq 3} \langle \Theta \Theta^T, S_{A,k}(X) - \mathbb{E}S_{A,k}(X) \rangle}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F}$$

where normal approximation of the first term is given in (7.15) and upper bound of the second term is given in (7.13). Based on (7.13), we get for any $x \in \mathbb{R}$ and any $s \geq 1$,

$$\mathbb{P}\left(\frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F} \leq x\right) \leq \mathbb{P}\left(\frac{2\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2 - 2\mathbb{E}\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F} \leq x + C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2 \lambda_r^2} \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r}\right) + e^{-s} + e^{-c_1 d_{\max}}$$

$$\leq \Phi\left(x + C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2 \lambda_r^2} \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r}\right) + e^{-s} + e^{-c_1 d_{\max}} + C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_*}}$$

$$\leq \Phi(x) + C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2 \lambda_r^2} \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r} + e^{-s} + e^{-c_1 d_{\max}} + C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_*}}$$

where the last inequality is due to (7.15) and the Lipschitz property of function $\Phi(x)$. Similarly, for any $x \in \mathbb{R}$ and any $s \geq 1$,

$$\mathbb{P}\left(\frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E} \text{dist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F} \leq x\right) \geq \mathbb{P}\left(\frac{2\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2 - 2\mathbb{E}\|\mathcal{P}^{-1}X\mathcal{P}^\perp\|_F^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_F} \leq x - C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2 \lambda_r^2} \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r}\right) - e^{-s} - e^{-c_1 d_{\max}}$$

$$\geq \Phi\left(x - C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2 \lambda_r^2} \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r}\right) - e^{-s} - e^{-c_1 d_{\max}} - C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_*}}$$

$$\geq \Phi(x) - C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2 \lambda_r^2} \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r} - e^{-s} - e^{-c_1 d_{\max}} - C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_*}}.$$
\[ \leq C_6 s^{1/2} \left( \frac{\sqrt{r}}{\|A^{-2}\|_F^2 r} \right)^{1/2} \left( \frac{rd_{\text{max}}}{\lambda_r} \right)^{1/2} + e^{-s} + e^{-c_1d_{\text{max}}} + C_7 \left( \frac{\|A^{-1}\|_F^4}{\|A^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_*}} \]

where \( d_* = d_1 + d_2 - 2r \) and \( C_6, C_7, c_1 \) are absolute positive constants.

### 7.3 Proof of lemmas in Section 3.2

Observe that \( S_{A,k}(X) \) involves the product of \( X \) for \( k \) times. If \( k \) is an odd number, we immediately get \( \mathbb{E}S_{A,k}(X) = 0 \) since \( Z \) has i.i.d. standard normal entries. Therefore, it suffices to consider \( \mathbb{E}\langle \Theta \Theta^T, S_{A,k}(X) \rangle \) for even numbers \( k \).

**Proof of Lemma 1.** Recall the definitions of \( \Psi^\perp \), \( X \) and \( \Psi^{-1} \), we have

\[
\mathbb{E}\|\Psi^\perp X \Psi^{-1}\|_F^2 = \mathbb{E}\left\| U A^{-1} V^T Z U \|_F^2 + \mathbb{E}\| V A^{-1} U^T Z U \|_F^2 \right.
\]

\[
= \mathbb{E}\| A^{-1} V^T Z U \|_F^2 + \mathbb{E}\| A^{-1} U^T Z U \|_F^2.
\]

Similar as in the proof of Theorem 2, we obtain \( \mathbb{E}\|\Psi^\perp X \Psi^{-1}\|_F^2 = (d_1 + d_2 - 2r)\|A^{-1}\|_F^2 \) which obtains the first claim. To prove the second claim, recall the representation formulas of \( S_{A,2k}(X) \) from Theorem 1, we get

\[
\left| \mathbb{E}\|\tilde{\Theta} \Theta^T - \Theta \Theta^T\|_F^2 - 2d_* \|A^{-1}\|_F^2 \right| \leq 2 \sum_{k \geq 2} \mathbb{E}\langle \Theta \Theta^T, S_{A,2k}(X) \rangle
\]

\[
\leq 2 \sum_{k \geq 2} \left| \mathbb{E}\langle \Theta \Theta^T, \sum_{s, s_1 + \cdots + s_{2k+1} = 2k} (-1)^{1+\tau(s)} \Psi^{-s_1} \Psi^{-s_2} X \cdots X \Psi^{-s_{2k}} X \Psi^{-s_{2k+1}} \rangle \right|
\]

\[
= 2 \sum_{k \geq 2} \left| \mathbb{E}\langle \Theta \Theta^T, \sum_{s, s_1 + \cdots + s_{2k+1} = 2k} (-1)^{1+\tau(s)} \Psi^{-s_1} \Psi^{-s_2} X \cdots X \Psi^{-s_{2k}} X \Psi^{-s_{2k+1}} \rangle \right|
\]

where we used the fact \( \Theta \Theta^T \Psi^0 = \Psi^0 \Theta \Theta = 0 \). As a result, we can write

\[
\left| \mathbb{E}\|\tilde{\Theta} \Theta^T - \Theta \Theta^T\|_F^2 - 2d_* \|A^{-1}\|_F^2 \right| \leq 4r \sum_{k \geq 2} \sum_{s, s_1 + \cdots + s_{2k+1} = 2k} \mathbb{E}\| \Psi^{-s_1} \Psi^{-s_2} X \cdots X \Psi^{-s_{2k}} X \Psi^{-s_{2k+1}} \|
\]

\[
\leq 4r \sum_{k \geq 2} \sum_{s, s_1 + \cdots + s_{2k+1} = 2k} \frac{\mathbb{E}\|X\|_F^{2k}}{\lambda_r^{2k}}
\]
\[ \leq 4r \sum_{k \geq 2} \left( \frac{4k}{2k} \right) \frac{E \| X \|_{2k}^2}{\lambda_r^2} \leq C_2 r \sum_{k \geq 2} \frac{4^{2k} E \| X \|_{2k}^2}{\lambda_r^2}. \]

for some absolute constant $C_2 > 0$. As in the proof of Theorem 1 (see also Koltchinskii and Xia, 2016, Lemma 3), there exists an absolute constant $C_1 > 0$ such that

\[ E \| X \|_p \leq C_1 d_{\max}^{p/2} \]

for all positive integer $p \geq 1$. Therefore, we conclude that

\[ \left| E \| \hat{\Theta} \hat{\Theta}^T - \Theta \Theta^T \|_F^2 - 2d_s \| \Lambda^{-1} \|_F^2 \right| \leq C_2 r \sum_{k \geq 2} \left( \frac{16C_1^2 d_{\max}}{\lambda_r^2} \right)^k \leq C_2 r d_{\max}^{2r} \]

where the last inequality holds as long as $\lambda_r \geq 5C_1 d_{\max}^{1/2}$.

**An important property of** $\langle \Theta \Theta^T, ES_{A,2k}(X) \rangle$. In order to calculate higher order approximations such as Lemma 2, we need the following useful property of $ES_{2k}(X)$ in this section. Note that, in general, the leading terms of $\left| \langle \Theta \Theta^T, ES_{A,2k}(X) \rangle \right|$ are of the order $O\left( \frac{d_{\max}^{2}}{\lambda_r^{2k}} \right)$. Basically, we will determine which terms in $\langle \Theta \Theta^T, ES_{A,2k}(X) \rangle$ attain the magnitude of such an order.

Recall from Theorem 1 we can write

\[ \langle \Theta \Theta^T, S_{A,2k}(X) \rangle = \sum_{s_1,s_2,\cdots,s_{2k+1}=2k} (-1)^{1+\tau(s)} \cdot \text{tr} \left( \mathcal{P}^{-s_1} X \cdots X \mathcal{P}^{-s_{2k+1}} \right). \]

Clearly, for any $\tau(s) = \tau \geq 2$, there exists positive integers $s_{j_1}, s_{j_2}, \cdots, s_{j_r}$ and positive integers $t_1, t_2, \cdots, t_{\tau-1}$ so that we can write

\[ \mathcal{P}^{-s_1} X \cdots X \mathcal{P}^{-s_{2k+1}} = \mathcal{P}^{-s_{j_1}} \underbrace{X \mathcal{P}^{-1} X \cdots X \mathcal{P}^{-1}}_{t_1 \text{ of } X} \mathcal{P}^{-s_{j_2}} \cdots \mathcal{P}^{-s_{j_{\tau-1}}} \underbrace{X \mathcal{P}^{-1} \cdots \mathcal{P}^{-1}}_{t_{\tau-1} \text{ of } X} X \mathcal{P}^{-s_{j_{\tau}}}, \]

where

\[ s_{j_1} + \cdots + s_{j_\tau} = 2k \quad \text{and} \quad t_1 + \cdots + t_{\tau-1} = 2k. \]
Therefore, we write for positive integers \(s_1, \cdots, s_{2k+1}, t_1, \cdots, t_{2k} \geq 1\),

\[
\langle \Theta^T, \mathbb{E} S_{4,2k}(X) \rangle = \sum_{\tau \geq 2} (-1)^{1+\tau} \sum_{s_1+\cdots+s_\tau=2k} \sum_{t_1+\cdots+t_{\tau-1}=2k} \mathbb{E} \operatorname{tr} (Q_{t_1t_2^* \cdots t_{\tau-1}^*}^{(s_1s_2^* \cdots s_\tau)})
\]

where the matrix \(Q_{t_1t_2^* \cdots t_{\tau-1}^*}^{(s_1s_2^* \cdots s_\tau)}\) is defined as

\[
Q_{t_1t_2^* \cdots t_{\tau-1}^*}^{(s_1s_2^* \cdots s_\tau)} = \mathcal{P}^{-s_1}X\mathcal{P}^{-s_2} \cdots X\mathcal{P}^{-s_{\tau-1}}X\mathcal{P}^{-s_\tau}.
\]

**Case 1:** if any of \(t_1, t_2, \cdots, t_{\tau-1}\) is one. W.L.O.G., let \(t_1 = 1\). Then, there involves the product of the matrix \(\mathcal{P}^{-s_1}X\mathcal{P}^{-s_2}\). Then,

\[
|\mathbb{E} \operatorname{tr} (Q_{t_1t_2^* \cdots t_{\tau-1}^*}^{(s_1s_2^* \cdots s_\tau)})| \leq 2r \cdot \mathbb{E} \|\mathcal{P}^{-s_1}X\mathcal{P}^{-s_2}\| \cdot \|X\|^{2k-1} \lambda_{\tau}^{2k}
\]

\[
\leq 2r \cdot \mathbb{E} \|\Theta^T X \Theta^T\| \cdot \|X\|^{2k-1} \lambda_{\tau}^{2k}
\]

\[
\leq 2r \cdot \mathbb{E}^{1/2} \|\Theta^T X \Theta^T\|^2 \mathbb{E}^{1/2} \|X\|^{4k-2}
\]

\[
\leq C_1 \frac{r^{3/2} \sigma_{\max}^{k-\frac{1}{2}}}{\lambda_{\tau}^{2k}}
\]

where we used the fact \(\Theta^T X \Theta^T = \begin{pmatrix} 0 & UU^T ZV^T \\ VV^T ZU^T & 0 \end{pmatrix}\) which is of rank at most 2r and \(\mathbb{E}^{1/2} \|U^T Z V\|^2 \leq C_2 \sqrt{r}\). We also used the fact \(\mathbb{E} \|X\|^p \leq \sigma_{\max}^p \sigma_{\max}^{p/2}\) for some absolute constant \(C_2 > 0\) and all positive integers \(p \geq 1\). Therefore, if any of \(t_1, \cdots, t_{\tau-1}\) is one, then the magnitude of \(|\mathbb{E} \operatorname{tr} (Q_{t_1t_2^* \cdots t_{\tau-1}^*}^{(s_1s_2^* \cdots s_\tau)})|\) is of the order \(O\left(\frac{\sigma_{\max}^{k-\frac{1}{2}}}{\lambda_{\tau}^{2k}}\right)\) which is much smaller than the leading order \(O\left(\frac{\sigma_{\max}^{k-\frac{1}{2}}}{\lambda_{\tau}^{2k}}\right)\) of \(|\langle \Theta^T, \mathbb{E} S_{4,2k}(X) \rangle|\).

**Case 2:** if any of \(t_1, \cdots, t_{\tau-1}\) is an odd number greater than 1. W.L.O.G., let \(t_1\) be an odd number and \(t_1 \geq 3\). More specifically, let \(t_1 = 2p + 3\) for some non-negative integer \(p \geq 0\). Then,

\[
|\mathbb{E} \langle \Theta^T, Q_{t_1t_2^* \cdots t_{\tau-1}^*}^{(s_1s_2^* \cdots s_\tau)} \rangle| \leq \operatorname{tr} \left( X(\mathcal{P} \perp X)^{t_1-1} \mathcal{P}^{-s_2} X(\mathcal{P} \perp X)^{t_2-1} \mathcal{P}^{-s_3} X \cdots \mathcal{P}^{-s_{\tau-1}} X(\mathcal{P} \perp X)^{t_{\tau-1}-1} \mathcal{P}^{-s_{\tau}} \right)
\]

50
$$\leq \mathbb{E} \left\| \mathcal{P}^{-1} X (\mathcal{P} \perp X \mathcal{P} \perp) 2^{p+1} X \mathcal{P}^{-1}\right\|_F \cdot \frac{\sqrt{2r} \left\| X \right\|^{2k-t_1}}{\lambda_r^2}$$

$$\leq \mathbb{E} \left\| \mathcal{P}^{-1} X (\mathcal{P} \perp X \mathcal{P} \perp) 2^{p+1} X \mathcal{P}^{-1}\right\|_F \cdot \frac{\sqrt{2r} \left\| X \right\|^{2k-t_1}}{\lambda_r^2} \mathbb{P}_{\mathcal{E}_1}$$

$$+ \mathbb{E} \left\| \mathcal{P}^{-1} X (\mathcal{P} \perp X \mathcal{P} \perp) 2^{p+1} X \mathcal{P}^{-1}\right\|_F \cdot \frac{\sqrt{2r} \left\| X \right\|^{2k-t_1}}{\lambda_r^2} \mathbb{P}_{\mathcal{E}_1}$$

where, as in the proof of Theorem 2, define the event $\mathcal{E}_1 = \{ \| X \| \leq C_2 \cdot \frac{d_1}{\lambda_r} \}$ for some absolute constant $C_2 > 0$ such that $\mathbb{P}(\mathcal{E}_1) \geq 1 - e^{-c_1 d_{\text{max}}}$. As a result, we get

$$\left\| \mathbb{E}(\Theta \Theta^T, Q^{(s_{i_1^{t_2^{-t_r, \ldots, t_{r-1}}})}) \right\| \leq \mathbb{E} \left\| \mathcal{P}^{-1} X (\mathcal{P} \perp X \mathcal{P} \perp) 2^{p+1} X \mathcal{P}^{-1}\right\|_F \cdot \frac{\sqrt{2r} \left\| X \right\|^{2k-t_1}}{\lambda_r^2}$$

$$+ C_{2k} \cdot \frac{r d_{\text{max}}^{2k}}{\lambda_r^2} \cdot e^{-c_1 d_{\text{max}}}$$

where $\Theta = (\theta_1, \ldots, \theta_r, \theta_{-r}, \ldots, \theta_{-1}) \in \mathbb{R}^{(d_1 + d_2) \times (2r)}$. Since we are only interested in the case $2k \leq 4 \log d_{\text{max}}$, we can conclude that $C_{2k} \cdot \frac{r d_{\text{max}}^{2k}}{\lambda_r^2} \cdot e^{-c_1 d_{\text{max}}} = O\left(\frac{r d_{\text{max}}^{2k}}{\lambda_r^2}\right)$ as long as $d_{\text{max}}$ is large enough. In addition, we can write

$$\mathbb{E} \left\| \Theta^T X (\mathcal{P} \perp X \mathcal{P} \perp) 2^{p+1} X \Theta \right\|^2 = \sum_{1 \leq |j_1|, |j_2| \leq r} \mathbb{E}(\theta_{j_1}^T X (\mathcal{P} \perp X \mathcal{P} \perp) 2^{p+1} X \theta_{j_2})^2.$$

Observe that, for any integer $p \geq 0$,

$$(\mathcal{P} \perp X \mathcal{P} \perp)^{2p} = \begin{pmatrix} (U \perp U \perp^T V \perp V \perp^T Z^T U \perp U \perp^T)^p & 0 \\ 0 & (V \perp V \perp^T Z^T U \perp U \perp^T V \perp V \perp^T)^p \end{pmatrix}.$$
Observe that $Z_{v_1}$ is independent with $ZV_\perp$ and $Z^T u_j$ is independent with $Z^T U_\perp$. Therefore,

$$\mathbb{E}(\theta_{j_1}^T X (P^\perp X P^\perp)^{2p+1} X \theta_{j_2})^2 \leq 2^{-1} \mathbb{E}\left\| \left( U_\perp U^\perp ZV_\perp V_\perp^T U_\perp U^\perp \right)^p U_\perp U^\perp ZV_\perp V_\perp^T u_j \right\|_2^2 + 2^{-1} \mathbb{E}\left\| \left( V_\perp V^\perp ZT U_\perp U^\perp ZV_\perp V_\perp^T \right)^p V_\perp V^\perp ZT U_\perp U^\perp Z v_j \right\|_2^2 \leq 2^{-1} \mathbb{E}\left\| \left( U_\perp U^\perp ZV_\perp V_\perp^T U_\perp U^\perp \right)^p U_\perp U^\perp ZV_\perp \right\|_2^2 + 2^{-1} \mathbb{E}\left\| \left( V_\perp V^\perp ZT U_\perp U^\perp ZV_\perp V_\perp^T \right)^p V_\perp V^\perp ZT U_\perp U^\perp \right\|_2^2 \leq C_2^{4p+2} d_{\max}^{2p+1} = C_2^{4p+2} d_{\max}^{t_1-2},$$

where the last inequality is due to the independence between $Z^T u_j$ and $Z^T U_\perp$, the independence between $Z_{v_1}$ and $ZV_\perp$. We conclude that

$$\left| \mathbb{E}\langle \Theta \Theta^T, Q^{(s_1 s_2 \cdots s_r)}_{t_1 t_2 \cdots t_{\tau-1}} \rangle \right| \leq C_2 \cdot \frac{r^{3/2} d_{\max}^{k-1}}{\lambda_{\min}^{2k}}$$

where $C_2 > 0$ is some absolute constant. Based on Case 1 and Case 2, we conclude that, whenever any of $t_1, \cdots, t_{\tau-1}$ is an odd number, the term

$$\left| \mathbb{E}\langle \Theta \Theta^T, Q^{(s_1 s_2 \cdots s_r)}_{t_1 t_2 \cdots t_{\tau-1}} \rangle \right| \leq C_2 \cdot \frac{r^{3/2} d_{\max}^{k-1}}{\lambda_{\min}^{2k}}.$$

Recall that the leading term of $\mathbb{E}\langle \Theta \Theta^T, S_{A,2k}(X) \rangle$ is of the order $O\left( \frac{d_{\max}^k}{\lambda_{\min}^{2k}} \right)$. Therefore, based on Case 1 and Case 2, it suffices to consider the cases that all of $t_1, \cdots, t_{\tau-1}$ are even numbers.

**Proof of Lemma 2** From the above analysis, to calculate $\mathbb{E}\langle \Theta \Theta^T, S_{A,4}(X) \rangle$, it suffices to calculate

$$\sum_{\tau=2}^{3} (-1)^{1+\tau} \sum_{s_1+\cdots+s_\tau=4} \sum_{t_1+\cdots+t_{\tau-1}=4} \mathbb{E}\langle \Theta \Theta^T, Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}} \rangle$$
where \( t_1, \ldots, t_{\tau-1} \) are positive even numbers and \( s_1, \ldots, s_\tau \) are positive numbers.

**Case 1:** \( \tau = 2 \). In this case, \( t_1 = 4 \) and \( s_1 + s_2 = 4 \). Therefore, for any \( s_1, s_2 \) such that \( s_1 + s_2 = 4 \), we shall calculate

\[
Q_4^{(s_1s_2)} = \mathbf{\Psi}^{-s_1} X (\mathbf{\Psi}^\perp X^{\perp})^2 X \mathbf{\Psi}^{-s_2} = \mathbb{E} \operatorname{tr} (Q_4^{(s_1s_2)}) = \mathbb{E} \operatorname{tr} (\Theta \Theta^T X (\mathbf{\Psi}^\perp X^{\perp})^2 X \Theta \Theta^T \mathbf{\Psi}^{-4}).
\]

Clearly, we have

\[
\Theta \Theta^T X (\mathbf{\Psi}^\perp X^{\perp})^2 X \Theta \Theta^T = \begin{pmatrix}
UU^T ZV_1 Z^T U_1 U_1^T ZV_1 V_1^T Z^T U U^T & 0 \\
0 & VV^T Z^T U_1 U_1^T ZV_1 V_1^T Z^T U_1 U_1^T ZV V^T
\end{pmatrix}.
\]

By the independence between \( U^T Z \) and \( U_1^T Z \), independence between \( V^T Z \) and \( V_1^T Z \), we immediately obtain

\[
\mathbb{E} \Theta \Theta^T X (\mathbf{\Psi}^\perp X^{\perp})^2 X \Theta \Theta^T = \mathbb{E} \begin{pmatrix}
d_1-UU^T ZV_1 V_1^T Z^T U U^T & 0 \\
0 & d_2-VV^T Z^T U_1 U_1^T ZV V^T
\end{pmatrix}
= d_1 - d_2 - \Theta \Theta^T
\]

where \( d_1 = d_1 - r \) and \( d_2 = d_2 - r \). Then,

\[
\mathbb{E} \langle \Theta \Theta^T, Q_4^{(s_1s_2)} \rangle = 2d_1 - d_2 - \| \Lambda^{-2} \|_F^2
\]

for all \((s_1, s_2) = (1, 3), (s_1, s_2) = (2, 2) \) and \((s_1, s_2) = (3, 1)\).

**Case 2:** \( \tau = 3 \). In this case, the only possible even numbers are \( t_1 = 2 \) and \( t_2 = 2 \). There are three pairs of \((s_1, s_2, s_3) \in \{(1, 1, 2), (1, 2, 1), (2, 1, 1) \}\). W.L.O.G., consider \( s_1 = 1, s_2 = 1, s_3 = 2 \), we have

\[
Q_2^{(112)} = \mathbf{\Psi}^{-1} X \mathbf{\Psi}^\perp X \mathbf{\Psi}^{-1} X \mathbf{\Psi}^\perp X \mathbf{\Psi}^{-2}.
\]

Similarly, we can write

\[
\mathbb{E} \operatorname{tr} (Q_2^{(112)}) = \mathbb{E} \operatorname{tr} (U \Lambda^{-1} V^T Z U_1 U_1^T Z V \Lambda^{-1} U^T Z V_1 V_1^T Z^T U U^T \Lambda^{-2} U^T) \\
+ \mathbb{E} \operatorname{tr} (V \Lambda^{-1} U^T Z V_1 V_1^T Z^T U \Lambda^{-1} V^T Z^T U_1 U_1^T Z V \Lambda^{-2} V^T)
\]

53
\[ d_2 \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z U \Lambda^{-3} U^T \right) + d_1 \mathbb{E} \text{tr} \left( V \Lambda^{-1} Z V U \Lambda^{-3} U^T \right) = 2d_1 d_2 \| \Lambda^{-2} \|_F^2. \]

By symmetricity, the same equation holds for \( \mathbb{E} \text{tr}(Q_{22}^{(121)}) \). Next, we consider \((s_1, s_2, s_3) = (1, 2, 1)\). We will write

\[
\mathbb{E} \text{tr}(Q_{22}^{(121)}) = \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z U \Lambda^{-2} V^T Z U \Lambda^{-1} U^T \right) + \mathbb{E} \text{tr} \left( V \Lambda^{-1} Z V U \Lambda^{-2} U^T Z U \Lambda^{-1} U^T \right)
\]

\[
- \mathbb{E} \| \Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-1} \|_F^2 - \mathbb{E} \| \Lambda^{-1} \tilde{Z}_2 \tilde{Z}_2^T \Lambda^{-1} \|_F^2
\]

where \( \tilde{Z}_1 \in \mathbb{R}^{r \times d_1} \) and \( \tilde{Z}_2 \in \mathbb{R}^{r \times d_2} \) contain i.i.d. standard normal entries. By Lemma 11 in the Appendix, we obtain

\[
\mathbb{E} \text{tr} \left( Q_{22}^{(121)} \right) = (d_1^2 + d_2^2) \| \Lambda^{-2} \|_F^2 + (d_1 - d_2)(\| \Lambda^{-2} \|_F^2 + \| \Lambda^{-1} \|_F^2).
\]

Therefore, we conclude that

\[
\left| - \mathbb{E} \langle \Theta \Theta^T, S_{A,4}(X) \rangle + (d_1 - d_2)^2 \| \Lambda^{-2} \|_F^2 \right| \leq C_2 \cdot \frac{r^2 d_{\max} + r^2 d_{\max}}{\lambda_r^4}
\]

for some absolute constant \( C_1 > 0 \) where the term \( \frac{r^3 d_{\max}}{\lambda_r^4} \) comes from those smaller terms when some \( t_i \) is an odd number as discussed in the previous section. Together with the proof of Lemma 1 we conclude that

\[
\left| \mathbb{E} \| \tilde{\Theta} \tilde{\Theta}^T - \Theta \Theta^T \|_F^2 - 2(d_1 \| \Lambda^{-1} \|_F^2 - \Delta_d^2 \| \Lambda^{-2} \|_F^2) \right| \leq C_1 \cdot \frac{r^2 d_{\max}}{\lambda_r^4} + C_2 \cdot \frac{r d_{\max}^2}{\lambda_r^6}
\]

where \( \Delta_d = d_1 - d_2 \) and \( C_1, C_2 > 0 \) are absolute constants.

**Proof of Lemma 3** Similar as the proof of Lemma 2, it suffices to consider \( \tau = 2, 3, 4 \).

**Case 1: \( \tau = 2 \)**. Then, \( t_1 = 6 \) and \( s_1 + s_2 = 6 \). There are 5 possible pairs \((s_1, s_2) \in \{(k, 6-k)\}_{k=1}^5\). To this end, we can write

\[
Q_6^{(s_1, s_2)} = \mathcal{P}^{-s_1} X (\mathcal{P}^\perp X \mathcal{P}^\perp)^4 X \mathcal{P}^{-s_2}.
\]
Clearly, we have $\mathbb{E}\langle \Theta \Theta^T, Q^{(s_1 s_2)}_6 \rangle = \mathbb{E} \text{tr} \left( \mathbf{P}^{-3} X (\mathbf{P}^\perp X \mathbf{P}^\perp)^4 X \mathbf{P}^{-3} \right)$. We can immediately check that

$$
(\mathbf{P}^\perp X \mathbf{P}^\perp)^4 = \begin{pmatrix} T_U^2 & 0 \\ 0 & T_V^2 \end{pmatrix}
$$

where $T_U = U_{\perp} U_{\perp}^T Z V_{\perp} V_{\perp}^T Z^T U_{\perp} U_{\perp}^T$ and $T_V = V_{\perp} V_{\perp}^T Z^T U_{\perp} U_{\perp}^T Z V_{\perp} V_{\perp}^T$. Therefore, we obtain

$$
\mathbb{E}\langle \Theta \Theta^T, Q^{(s_1 s_2)}_6 \rangle = \mathbb{E} \text{tr} (U \Lambda^{-3} V^T Z^T T_U^2 Z V \Lambda^{-3} U^T) + \mathbb{E} \text{tr} (V \Lambda^{-3} U^T Z T_V^2 Z^T U \Lambda^{-3} V^T) = \|\Lambda^{-3}\|^2_F \mathbb{E} \text{tr} (T_U^2) + \|\Lambda^{-3}\|^2_F \mathbb{E} \text{tr} (T_V^2),
$$

where we used the independence between $Z V$ and $T_U$, and the independence between $U^T Z$ and $U_{\perp} Z$. By Lemma [8] in the Appendix, we obtain $\mathbb{E}\|T_U\|^2_F = \mathbb{E}\|T_V\|^2_F = d_1 - d_2 - (d_r + 1)$. Therefore, for all $s_1 + s_2 = 6$, we have

$$
\mathbb{E}\langle \Theta \Theta^T, Q^{(s_1 s_2)}_6 \rangle = 2d_1 - d_2 - (d_r + 1) \|\Lambda^{-3}\|^2_F.
$$

**Case 2**: $\tau = 3$. There are two possible pairs $(t_1, t_2) = (2, 4)$ or $(t_1, t_2) = (4, 2)$. W.L.O.G., let $(s_1, s_2, s_3) = (1, 1, 4)$ and $t_1 = 2$, $t_2 = 4$. Then, we denote

$$
Q^{(114)}_{24} = \mathbf{P}^{-1} X \mathbf{P}^\perp X \mathbf{P}^{-1} X (\mathbf{P}^\perp X \mathbf{P}^\perp)^2 X \mathbf{P}^{-4}.
$$

For simplicity, we will denote $L^{(114)}_{24} = \mathbb{E}\langle \Theta \Theta^T, Q^{(114)}_{24} \rangle$. Then, we get

$$
L^{(114)}_{24} = \mathbb{E} \text{tr} (Q^{(114)}_{24}) = \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z^T U_{\perp} U_{\perp}^T Z V \Lambda^{-1} U^T Z T_V Z^T U \Lambda^{-4} U^T \right)
+ \mathbb{E} \text{tr} \left( V \Lambda^{-1} U^T Z V_{\perp} V_{\perp}^T Z^T U \Lambda^{-1} V^T Z^T T_U Z V \Lambda^{-4} V^T \right)
= d_1 \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z^T U_{\perp} U_{\perp}^T Z V \Lambda^{-1} U^T Z V_{\perp} V_{\perp}^T Z^T U \Lambda^{-4} U^T \right)
+ d_2 \mathbb{E} \text{tr} \left( V \Lambda^{-1} U^T Z V_{\perp} V_{\perp}^T Z^T U \Lambda^{-1} V^T Z^T T_U Z V \Lambda^{-4} V^T \right)
= d_1 \mathbb{E} \|\Lambda^{-5/2} V^T Z^T U_{\perp} U_{\perp}^T Z V \Lambda^{-1/2}\|^2_F + d_2 \mathbb{E} \|\Lambda^{-5/2} U^T Z V_{\perp} V_{\perp}^T Z^T U \Lambda^{-1/2}\|^2_F
= d_1 d_2 - d_r \|\Lambda^{-3}\|^2_F + O \left( \frac{r^2 d_{\max}^2}{\lambda_p^2} \right),
$$

where the last inequality is due to Lemma [11]. Similarly, we can get the following results (some details are provided in Section [8.1]):

$$
L^{(114)}_{24} = L^{(213)}_{24} = L^{(312)}_{24} = L^{(411)}_{24}
$$
\[ L^{(114)}_{42} = L^{(213)}_{42} = L^{(312)}_{42} = L^{(411)}_{42} = d_1 d_2 d_* \Lambda^{-3} \|_{F}^2 + O \left( \frac{r^2 d^2_{\max}}{\lambda^6_r} \right) \]

\[ L^{(123)}_{24} = L^{(222)}_{24} = L^{(321)}_{24} = L^{(412)}_{24} = L^{(312)}_{24} = L^{(421)}_{24} = d_1 d_2 d_* \Lambda^{-3} \|_{F}^2 + O \left( \frac{r^2 d^2_{\max}}{\lambda^6_r} \right) \]

\[ L^{(132)}_{24} = L^{(231)}_{24} = d_1 d_2 d_* \Lambda^{-3} \|_{F}^2 + O \left( \frac{r^2 d^2_{\max}}{\lambda^6_r} \right) \]

\[ L^{(141)}_{24} = d_1 d_2 d_* \Lambda^{-3} \|_{F}^2 + O \left( \frac{r^2 d^2_{\max}}{\lambda^6_r} \right) \]

Case 3: \( \tau = 4 \). There is only one possible \((t_1, t_2, t_3) = (2, 2, 2)\). W.L.O.G., let \( s_1 = s_2 = s_3 = 1 \) and \( s_4 = 3 \). Then, we consider

\[ Q^{(113)}_{222} = \Psi^{-1}(X \Psi^\perp X) \Psi^{-1}(X \Psi^\perp X) \Psi^{-1}(X \Psi^\perp X) \Psi^{-3}. \]

Similarly, we denote \( L^{(113)}_{222} = \mathbb{E} \langle \Theta \Theta^T, Q^{(113)}_{222} \rangle \) and obtain

\[ L^{(113)}_{222} = \mathbb{E} \text{tr} \left( Q^{(113)}_{222} \right) \]

\[ = \mathbb{E} \text{tr} \left( \Lambda^{-1} V^T U_1^T U_1 U_1^T Z V \Lambda^{-1} U_1^T Z V \Lambda^{-1} U_1^T Z V \Lambda^{-3} U_1^T \right) + \mathbb{E} \text{tr} \left( V \Lambda^{-1} U_1^T Z V \Lambda^{-3} V U_1^T Z V \Lambda^{-1} U_1^T Z V \Lambda^{-3} V \right) \]

\[ = d_2 \mathbb{E} \text{tr} \left( \Lambda^{-1} V^T U_1^T U_1^T Z V \Lambda^{-3} U_1^T \right) + d_1 \mathbb{E} \text{tr} \left( V \Lambda^{-1} U_1^T Z V \Lambda^{-3} V \right) \]

\[ = d_1 d_2 d_* \| \Lambda^{-3} \|_{F}^2 + 2 d_1 d_2 d_* \left( \| \Lambda^{-3} \|_{F}^2 + \| \Lambda^{-3} \|_{F}^2 \| \Lambda^{-3} \|_{F}^2 \right), \]

where the last equality is due to Lemma 11. Considering all the pairs \((s_1, s_2, s_3, s_4)\), we obtain the following results (some details are provided in Section 8.1):

\[ L^{(113)}_{222} = L^{(212)}_{222} = L^{(311)}_{222} \]

\[ = L^{(112)}_{222} = L^{(211)}_{222} \]

\[ = L^{(121)}_{222} = L^{(221)}_{222} \]

\[ = d_1 d_2 d_* \| \Lambda^{-3} \|_{F}^2 + 2 d_1 d_2 d_* \left( \| \Lambda^{-3} \|_{F}^2 + \| \Lambda^{-3} \|_{F}^2 \| \Lambda^{-3} \|_{F}^2 \right) \]

\[ L^{(131)}_{222} = L^{(111)}_{222} = d_1 d_2 d_* \| \Lambda^{-3} \|_{F}^2 + 2 d_1 d_2 d_* \left( \| \Lambda^{-3} \|_{F}^2 + \| \Lambda^{-3} \|_{F}^2 \| \Lambda^{-3} \|_{F}^2 \right) \]

56
\( L_{222}^{(1221)} = (d_1^3 + d_2^3 + 3d_1^2 + 3d_2^2 + 4d_\tau) \| \Lambda^{-3} \|_F^2 + 3(d_1^2 + d_2^2 + d_\tau) \| \Lambda^{-1} \|_F^2 \| \Lambda^{-2} \|_F^2 + d_\tau \| \Lambda^{-1} \|_F^2. \)

The equations of \( L_{122221} \) is due to Lemma \[12\]. By combining all the above cases, we conclude that

\[
-2\mathbb{E}\langle \Theta \Theta^T, S_{A,6}(X) \rangle - 2d_\tau (d_1 - d_2)^2 \| \Lambda^{-6} \| \leq C_4 \frac{r^2 d_\tau^2 \Lambda_{max}^2 + r^2 d_\tau^2 \Delta_{max}^2 + r^3 d_\tau^3}{\lambda_{max}^6}
\]

for some absolute constant \( C_4 > 0 \). Following similar proof as in Lemma \[1\] we conclude that

\[
\left| \mathbb{E}\| \Theta \Theta^T - \Theta \Theta^T \|_F^2 - 2d_\tau \| \Lambda^{-1} \|_F^2 + 2\Delta_{max}^2 \| \Lambda^{-2} \|_F^2 - 2d_\tau \Delta_{max}^2 \| \Lambda^{-3} \|_F^2 \right| \leq C_4 \frac{r^3 d_\tau^2 \Lambda_{max}^2 + r^2 d_\tau^2 \Delta_{max}^2 + r^3 d_\tau^3}{\lambda_{max}^6} + C_5 \frac{rd_{max}}{\lambda_{max}^8}
\]

for some absolute constants \( C_4, C_5 > 0 \).

**Proof of Lemma \[4\]** Similarly, as in the proof of Lemma \[2\] we need to choose positive even numbers \( t_1, \cdots, t_\tau \) so that \( t_1 + \cdots + t_{\tau-1} = 8 \). Clearly, it suffices to consider \( \tau = 2, 3, 4, 5 \).

**Case 1: \( \tau = 2 \)**. Then, we consider \( Q_{s_1t_1s_2} \) where \( s_1 + s_2 = 8 \) and \( t_1 = 8 \). There are 7 possible pairs of \( (s_1, s_2) \). For every pair \( (s_1, s_2) \) such that \( s_1 + s_2 = 8 \), we have

\[
Q_{s_1s_2}^{(s_1s_2)} = \mathcal{P}^{s_1X(\mathcal{P}^{\perp}X\mathcal{P}^{\perp})^6X\mathcal{P}^{-s_2}}.
\]

Similarly, we denote by \( L_{s_1s_2}^{(s_1s_2)} = \mathbb{E} \text{tr} \left( Q_{s_1s_2}^{(s_1s_2)} \Theta \Theta^T \right) \). Since \( Z \) has i.i.d. normal entries, we can easily check that \( \Theta \Theta^T X \) is independent with \( (\mathcal{P}^{\perp}X\mathcal{P}^{\perp})^2 \) and so that

\[
L_{s_1s_2}^{(s_1s_2)} = \mathbb{E} \Theta \Theta^T X (\mathcal{P}^{\perp}X\mathcal{P}^{\perp})^6X\Theta \Theta^T = \Theta \Theta^T \mathbb{E} \text{tr} (\mathcal{P}^{\perp}X\mathcal{P}^{\perp})^6.
\]

Therefore, for all \( (s_1, s_2) \) with \( s_1 + s_2 = 8 \), we obtain

\[
L_{s_1s_2}^{(s_1s_2)} = \| \Lambda^{-4} \|_F^2 \cdot \mathbb{E} \text{tr} (\mathcal{P}^{\perp}X\mathcal{P}^{\perp})^6.
\]

After simple algebra, we can get \( \mathbb{E} \text{tr} (\mathcal{P}^{\perp}X\mathcal{P}^{\perp})^6 = \mathbb{E} \text{tr} (\tilde{Z}_1\tilde{Z}_1^T)^3 + \mathbb{E} \text{tr} (\tilde{Z}_2\tilde{Z}_2^T)^3 \) where \( \tilde{Z}_1 \in \mathbb{R}^{(d_1-r)\times(d_2-r)} \) and \( \tilde{Z}_2 \in \mathbb{R}^{(d_2-r)\times(d_1-r)} \) contain i.i.d. standard normal entries. By Lemma \[12\] we conclude that

\[
L_{s_1s_2}^{(s_1s_2)} = d_{1-}d_{2-}(2d_1^2 + 2d_2^2 + 6d_1d_2 - E) + O \left( \frac{d_\tau^3}{\lambda_{max}^8} \right) \quad \text{for all } s_1 + s_2 = 8.
\]

57
Case 2: $\tau = 3$. We shall consider $Q_{t_1t_2}^{(s_1s_2s_3)}$ where $t_1 + t_2 = 8$ and $s_1 + s_2 + s_3 = 8$. W.L.O.G., let $t_1 = t_2 = 4$ and $s_1 = s_3 = 2$ and $s_2 = 4$. Then, we get

\[
\langle \Theta \Theta^T, Q_{44}^{(242)} \rangle = \text{tr}(V^TZ^TUZV\Lambda^{-4}V^TZ^TUZV\Lambda^{-4}) + \text{tr}(U^TZVT\Lambda^{-4}U^TZVT\Lambda^{-4})
\]

\[
= \|\Lambda^{-2}V^TZ^TUZV\Lambda^{-2}\|^2_F + \|\Lambda^{-2}U^TZVT\Lambda^{-2}\|^2_F
\]

where

\[
T_U = U_\perp U_\perp^TZV_\perp U_\perp^TZU_\perp^T \quad \text{and} \quad T_V = V_\perp V_\perp^TZU_\perp^TZV_\perp^T.
\]

Therefore, we get, conditioned on $T_U$,

\[
E_{ZV}\langle \Theta \Theta^T, Q_{44}^{(242)} \rangle = \sum_{i=1}^{r} \frac{1}{\lambda_i} \left[ 2\|T_U\|^2_F + (\text{tr}(T_U))^2 \right] + \sum_{1 \leq i_1 \neq i_2 \leq r} \frac{1}{\lambda_{i_1}^{2} \lambda_{i_2}^{2}} \|T_U\|^2_F
\]

\[
+ \sum_{i_1=1}^{r} \frac{1}{\lambda_{i_1}^{2}} \left[ 2\|T_V\|^2_F + (\text{tr}(T_V))^2 \right] + \sum_{1 \leq i_1 \neq i_2 \leq r} \frac{1}{\lambda_{i_1}^{2} \lambda_{i_2}^{2}} \|T_V\|^2_F
\]

\[
= \|\Lambda^{-4}\|^2_F \|T_U\|^2_F + (\text{tr}(\Lambda^{-4}))^2 \|T_U\|^2_F + \|\Lambda^{-4}\|^2_F (\text{tr}(T_U))^2
\]

\[
+ \|\Lambda^{-4}\|^2_F \|T_V\|^2_F + (\text{tr}(\Lambda^{-4}))^2 \|T_V\|^2_F + \|\Lambda^{-4}\|^2_F (\text{tr}(T_V))^2
\]

By Lemma 8 and Lemma 9, we obtain $E\|T_U\|^2_F = E\|T_V\|^2_F = d_1d_2(d_\ast + 1)$ and $E(\text{tr}(T_U))^2 = E(\text{tr}(T_V))^2 = d_1^2d_2^2 + 2d_1d_2d_\ast$. Therefore, we obtain

\[
L_{44}^{(242)} = 2d_1^2d_2^2\|\Lambda^{-4}\|^2_F + O\left(\frac{r^2d_\ast^3}{\lambda_r^8}\right)
\]

where $d_1 = d_1 - r$ and $d_2 = d_2 - r$ so that $d_\ast = d_1 + d_2$. Similarly, we get the following results (some details are provided in Section 8.1):

\[
L_{44}^{(116)} = L_{44}^{(215)} = L_{44}^{(314)} = L_{44}^{(413)} = L_{44}^{(512)} = L_{44}^{(611)} = 2d_1^2d_2^2\|\Lambda^{-4}\|^2_F + O\left(\frac{r^2d_\ast^3}{\lambda_r^8}\right)
\]

\[
L_{44}^{(125)} = L_{44}^{(224)} = L_{44}^{(323)} = L_{44}^{(422)} = L_{44}^{(521)} = 2d_1^2d_2^2\|\Lambda^{-4}\|^2_F + O\left(\frac{r^2d_\ast^3}{\lambda_r^8}\right)
\]

\[
L_{44}^{(134)} = L_{44}^{(233)} = L_{44}^{(332)} = L_{44}^{(431)} = 2d_1^2d_2^2\|\Lambda^{-4}\|^2_F + O\left(\frac{r^2d_\ast^3}{\lambda_r^8}\right)
\]

\[
L_{44}^{(143)} = L_{44}^{(242)} = L_{44}^{(341)} = 2d_1^2d_2^2\|\Lambda^{-4}\|^2_F + O\left(\frac{r^2d_\ast^3}{\lambda_r^8}\right)
\]

58
Now, set $t_1 = 2$ and $t_2 = 6$ and we obtain the following results (some details are provided in Section 8.1):

$$L_{44}^{(152)} = L_{44}^{(251)} = 2d_1^2 d_2^2 \| \Lambda^{-4} \|_F^2 + O \left( \frac{r \sigma_{\max}^2}{\lambda_r^8} \right)$$

$$L_{44}^{(161)} = 2d_1^2 d_2^2 \| \Lambda^{-4} \|_F^2 + O \left( \frac{r \sigma_{\max}^3}{\lambda_r^8} \right).$$

In the case $t_1 = 6$ and $t_2 = 2$, the terms are the same as above. After summing over all the above terms, we conclude that

$$| S_{8,3} - 42d_1 d_2 (d_1 d_2 + d_2^2) \| \Lambda^{-4} \|_F^2 | \leq C_1 \cdot \frac{r \sigma_{\max}^3}{\lambda_r^8}$$

where $C_1 > 0$ is an absolute constant and

$$S_{8,3} = \sum_{s_1 + s_2 + s_3 = 8 \atop t_1 + t_2 = 8} L_{t_1 t_2 t_3}^{(s_1 s_2 s_3)}.$$

Case 2: $\tau = 4$. Similarly, we denote by

$$S_{8,4} = \sum_{s_1 + s_2 + s_3 + s_4 = 8 \atop t_1 + t_2 + t_3 = 8} L_{t_1 t_2 t_3 t_4}^{(s_1 s_2 s_3 s_4)}$$

where $s_1, s_2, s_3, s_4 \geq 1$ and $t_1, t_2, t_3$ are positive even numbers. First, we consider $(t_1, t_2, t_3) = (2, 4, 2)$. Then, we have the following results. (Some detailed calculates are provided in Section 8.1)

$$L_{242}^{(1115)} = L_{242}^{(2114)} = L_{242}^{(3113)} = L_{242}^{(4112)} = L_{242}^{(5111)} = d_1 d_2 (d_1^2 + d_2^2) \| \Lambda^{-4} \|_F^2 + O \left( \frac{r \sigma_{\max}^3}{\lambda_r^8} \right)$$

59
\[ L_{242}^{(1124)} = L_{242}^{(2123)} = L_{242}^{(3122)} = L_{242}^{(4121)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O\left( \frac{rd_{\text{max}}^3}{\lambda_r^8} \right) \]
\[ L_{242}^{(1133)} = L_{242}^{(2132)} = L_{242}^{(3131)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O\left( \frac{rd_{\text{max}}^3}{\lambda_r^8} \right) \]
\[ L_{242}^{(1142)} = L_{242}^{(2141)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O\left( \frac{rd_{\text{max}}^3}{\lambda_r^8} \right) \]
\[ L_{242}^{(1151)} = d_1 - d_2 \| \Lambda^{-4} \|_F^2 + O\left( \frac{rd_{\text{max}}^3}{\lambda_r^8} \right) \]
\[ L_{242}^{(1214)} = L_{242}^{(2213)} = L_{242}^{(3212)} = L_{242}^{(4211)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1223)} = P_{242}^{(2222)} = L_{242}^{(3221)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1232)} = L_{242}^{(2331)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1312)} = L_{242}^{(3131)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1322)} = L_{242}^{(2321)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1412)} = L_{242}^{(4111)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1421)} = L_{242}^{(4211)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{242}^{(1511)} = d_1 - d_2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]

Now, consider the tuple \((t_1, t_2, t_3) = (2, 2, 4)\). Then, we obtain the following results. (Some detailed calculations are provided in Section 8.1)
\[ L_{224}^{(1115)} = L_{224}^{(2114)} = L_{224}^{(3113)} = L_{224}^{(4112)} = L_{224}^{(5111)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1124)} = L_{224}^{(2123)} = L_{224}^{(3122)} = L_{224}^{(4121)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1133)} = L_{224}^{(2132)} = L_{224}^{(3131)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1142)} = L_{224}^{(2141)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1151)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1214)} = L_{224}^{(2213)} = L_{224}^{(3212)} = L_{224}^{(4211)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1223)} = L_{224}^{(2222)} = L_{224}^{(3221)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1232)} = L_{224}^{(2231)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1312)} = L_{224}^{(3113)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1322)} = L_{224}^{(3221)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1412)} = L_{224}^{(4111)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1421)} = L_{224}^{(4211)} = 2d_1^2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]
\[ L_{224}^{(1511)} = d_1 - d_2 \| \Lambda^{-4} \|_F^2 + O(rd_{\text{max}}^3/\lambda_r^8) \]

60
Case 4: $\tau$ for some absolute constant we consider

We obtain the following results. (Again, we provide some detailed calculations in Section 8.1).

The terms for $(t_1, t_2, t_3) = (4, 2, 2)$ are similar as above. We conclude that

$$
|S_{8,4} - \left( 45d_1 - d_2/d_1 + d_2^2 \right)||\Lambda^{-4}/||F| + O(rd_{\text{max}}^3/\lambda_r^8)
$$

for some absolute constant $C_1 > 0$.

Case 4: $\tau = 5$. Then, $t_1 = t_2 = t_3 = t_4 = 2$ is the only possible pair for $(t_1, t_2, t_3, t_4)$. Now, we consider

$$
S_{8,5} = \sum_{s_8 s_1 + \cdots + s_5 = 8} L^{(s_1 \cdots s_8)}_{2222}.
$$

We obtain the following results. (Again, we provide some detailed calculations in Section 8.1).
We conclude that

\[ |S_{8,5} - (16d_1d_2 - (d_1^2 + d_2^2) + 36d_1^2d_2 + (d_1^4 + d_2^4))\|\Lambda^{-4}\|_F^2| \leq C_2 \frac{rd_{\max}^3}{\lambda_r^8}. \]

By summing over the terms in \( S_{8,2}, S_{8,3}, S_{8,4}, S_{8,5} \) and taking into consideration of the signs, we conclude that

\[ | -E\langle \Theta^T S_{A,8}(X) \rangle + (d_1^2 + d_2^2 + d_1d_2)\Delta_d^2\Lambda^{-4}\|_F^2| \leq C_2 \cdot \frac{r^2d_{\max}^3}{\lambda_r^8} \]

for some absolute constant \( C_2 > 0 \). Together with Lemma 1, Lemma 2, Lemma 3, we conclude that

\[ \left| E\|\hat{\Theta}^T - \Theta^T\|_F^2 - 2B_4 \right| \leq C_1 \frac{r^2d_{\max}^3}{\lambda_r^8} + C_2 \frac{rd_{\max}^5}{\lambda_r^{10}} \]

for some absolute constant \( C_1, C_2 > 0 \) where

\[ B_4 = d_\star\|\Lambda^{-1}\|_F^2 - \Delta_d^2\|\Lambda^{-2}\|_F^2 + d_\star\Delta_d^2\|\Lambda^{-3}\|_F^2 - (d_1^2 + d_2d_2 + d_2^2)\Delta_d^2\|\Lambda^{-4}\|_F^2. \]

### 7.4 Proof of CLT theorems in Section 3.3

**Proof of Theorem 3** Recall Theorem 2, we end up with for all \( t \geq 1 \),

\[ \sup_{x \in \mathbb{R}} \left| P\left( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \text{Edist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)\|\Lambda^{-2}\|_F}} \leq x \right) - \Phi(x) \right| \]

62
\[ \leq C_2 \left( \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2} \right) \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r} + e^{-c_1 d_{\max}} + C_2 \left( \frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + e^{-t}. \]

By Lemma 1, we get

\[ |E_{\text{dist}}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_*\|\Lambda^{-1}\|_F^2| \leq C_2 \frac{r d_{\max}^2}{\lambda_r^4}. \]

Therefore,

\[ \left| \frac{E_{\text{dist}}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_*\|\Lambda^{-1}\|_F^2}{\sqrt{8d_*\|\Lambda^{-2}\|_F}} \right| \leq C_2 \frac{r d_{\max}^3}{\lambda_r^2}. \]

Now, by the Lipschitz property of \( \Phi(x) \) and applying similar technical as in proof of Theorem 2, we can get

\[ \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_*\|\Lambda^{-1}\|_F^2}{\sqrt{8d_*\|\Lambda^{-2}\|_F}} \leq x \right) - \Phi(x) \right| \leq C_2 \left( \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F^2} \right) \cdot \frac{(rd_{\max})^{1/2}}{\lambda_r} + e^{-c_1 d_{\max}} + C_2 \left( \frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + C_3 \frac{r d_{\max}^3}{\lambda_r^6} + 2e^{-t}. \]

**Proof of Theorem 4** By Lemma 2 we have

\[ \left| E_{\text{dist}}^2[(\hat{U}, \hat{V}), (U, V)] - 2\left( d_*\|\Lambda^{-1}\|_F^2 - \Delta_d^2\|\Lambda^{-2}\|_F^2 \right) \right| \leq C_1 \frac{r^2 d_{\max}}{\lambda_r^4} + C_2 \frac{r d_{\max}^3}{\lambda_r^6}. \]

The rest of the proof is the same as in the proof of Theorem 3

### 7.5 Proof of Theorem 7

We denote

\[ \bar{T} := \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_*\|\Lambda^{-1}\|_F^2}{\sqrt{8d_*\|\Lambda^{-2}\|_F}} \]

and denote

\[ T := \frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_*\|\Lambda^{-1}\|_F^2}{\sqrt{8d_*\|\Lambda^{-2}\|_F}}. \]

Then, we can write

\[ \bar{T} = T + \frac{2d_*\|\Lambda^{-1}\|_F^2 - \|\Lambda^{-1}\|_F^2}{\sqrt{8d_*\|\Lambda^{-2}\|_F}}. \]

63
By Lemma 5 if \( \lambda_r \geq C_1 d_{\max}^{1/2} \) for a large enough constant \( C_1 > 0 \), then we have \( \frac{\sqrt{2}}{2} \lambda_j \leq \tilde{\lambda}_j \leq \sqrt{2} \lambda_j \) for all \( 1 \leq j \leq r \). Then, by Lemma 5 we get

\[
\frac{2d_* (\|A^{-1}\|_F^2 - \|\tilde{A}^{-1}\|_F^2)}{\sqrt{8d_* \|A^{-2}\|_F}} \leq C_1 \frac{r d_{\max}^{3/2}}{\lambda_r^8} + C_2 \frac{r d_{\max}^{3/2}}{\lambda_r^{3/2}}
\]

for some absolute constants \( C_1, C_2 > 0 \). Therefore, if \( \lambda_r \geq C_1 d_{\max}^{1/2} \), we conclude that

\[
|\tilde{T} - T| \leq C_1 \frac{r d_{\max}^{3/2}}{\lambda_r^8} + C_2 \frac{r d_{\max}^{3/2}}{\lambda_r^{3/2}}
\]

which holds with probability at least \( 1 - \frac{1}{d_{\max}} \) by Lemma 5. Moreover, we define

\[
T_1 := \frac{\text{dist}^2((\tilde{U}, \tilde{V}), (U, V)) - 2d_* \|A^{-1}\|_F^2}{\sqrt{8d_* \|A^{-2}\|_F}}.
\]

Then, together with Lemma 5 we have

\[
|T - T_1| = \text{dist}^2((\tilde{U}, \tilde{V}), (U, V)) - 2d_* \|A^{-1}\|_F^2 \cdot \frac{\|A^{-2}\|_F - \|\tilde{A}^{-2}\|_F}{\sqrt{8d_* \|A^{-2}\|_F \|\tilde{A}^{-2}\|_F}} \leq \text{dist}^2((\tilde{U}, \tilde{V}), (U, V)) - 2d_* \|A^{-1}\|_F^2 \cdot \left( C_1 \frac{r d_{\max}^{3/2}}{\lambda_r^8} + C_2 \frac{r d_{\max}^{3/2}}{\lambda_r^{3/2}} \right)
\]

for some absolute constants \( C_1, C_2 > 0 \) and the last inequality holds with probability at least \( 1 - d_{\max}^{-2} \). By the CLT in Theorem 3 we conclude that with the probability at least \( C_1 \sqrt{\gamma} / (\lambda_r^2 \|A^{-2}\|_F) \cdot \left( \frac{r d_{\max} t^{1/2}}{\lambda_r} \right) + e^{-c_1 d_{\max}} + C_2 (\|A^{-1}\|_F^4 / \|A^{-2}\|_F^2)^{3/2} \cdot d_{\max}^{1/2} + C_3 r d_{\max}^{3/2} / \lambda_r^2 + 2e^{-t} \),

\[
|\text{dist}^2((\tilde{U}, \tilde{V}), (U, V)) - 2d_* \|A^{-1}\|_F^2 | \leq C_4 \sqrt{d_* \|A^{-2}\|_F} \cdot t^{1/2}
\]

for all \( t \geq 0 \) and some absolute constant \( C_4 > 0 \). As a result, we get, with the same probability,

\[
|T - T_1| \leq C_1 \frac{r d_{\max} t^{1/2}}{\lambda_r^9} + C_2 \frac{r d_{\max}^{3/4} \cdot t^{1/2}}{\lambda_r^{11/2}}.
\]

By the asymptotical normality of \( T_1 \) in Theorem 3 and the Lipschitz property of \( \Phi(x) \) as in the proof of Theorem 2 we conclude that for all \( t \geq 0 \),

\[
\sup_{x \in \mathbb{R}} \mathbb{P}\left( \frac{\text{dist}^2((\tilde{U}, \tilde{V}), (U, V)) - 2d_* \|A^{-1}\|_F^2}{\sqrt{8d_* \|A^{-2}\|_F}} \leq x \right) - \Phi(x)
\]

64
$$\leq C_1 \frac{\sqrt{r}}{\lambda_r^2 \|A^{-2}\|_F} \cdot \left( \frac{rd_{\max}}{\lambda_r} \right)^{1/2} + 2e^{-c_1 d_{\max}} + C_2 \left( \frac{\|A^{-1}\|_F}{\|A^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{d_{\max}^{1/2}} + 2e^{-t} + C_3 \frac{r d_{\max}^{3/2}}{\lambda_r^2} + C_4 \frac{r d_{\max}^{9/2}}{\lambda_r^8}$$

where we used the facts $\lambda_r \geq C_0 d_{\max}^{1/2}$ for a large enough constant $C_0 > 0$.

8 Appendix

8.1 Details of calculating $L$ values

Calculate $L_{24}^{(123)}$ in the proof of Lemma 3. By definition, we could write

$$L_{24}^{(123)} = \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z \Lambda U \Lambda V^{-2} V^T Z \Lambda V \Lambda^{-3} U^T \right) + \mathbb{E} \text{tr} \left( V \Lambda^{-1} U^T Z \Lambda V^{-2} U^T Z \Lambda V^{-3} V^T \right)$$

$$= d_2 \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z \Lambda U \Lambda V^{-2} V^T Z \Lambda V \Lambda^{-3} U^T \right) + d_1 \mathbb{E} \text{tr} \left( V \Lambda^{-1} U^T Z \Lambda V^{-2} U^T Z \Lambda V^{-3} V^T \right)$$

$$= d_1 d_2 d_4 \|A^{-3}\|_F^2 + O(\frac{r^2 d_{\max}^2}{\lambda_r^6}),$$

where the last equality is due to Lemma 11.

Calculate $L_{222}^{(1122)}$ in the proof of Lemma 3. By definition, we write

$$L_{222}^{(1122)} = \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z \Lambda U \Lambda V^{-2} V^T Z \Lambda V \Lambda^{-3} U^T \right) + \mathbb{E} \text{tr} \left( V \Lambda^{-1} U^T Z \Lambda V^{-2} U^T Z \Lambda V^{-3} V^T \right)$$

$$= d_1 \mathbb{E} \text{tr} \left( U \Lambda^{-1} V^T Z \Lambda U \Lambda V^{-2} V^T Z \Lambda V \Lambda^{-3} U^T \right) + d_2 \mathbb{E} \text{tr} \left( V \Lambda^{-2} U^T Z \Lambda V^{-2} U^T Z \Lambda V^{-3} V^T \right)$$

$$= d_1 d_2 d_4 \|A^{-3}\|_F^2 + 2d_1 d_2 \left( \|A^{-3}\|_F^2 + \|A^{-1}\|_F^2 \|A^{-2}\|_F^2 \right),$$

where the last equality is due to Lemma 11.
Calculate $L_{222}^{(1221)}$ in the proof of Lemma 3. By definition, we write

$$L_{222}^{(1221)} = \mathbb{E} \tr (UA^{-1}V^TZ^TU_{\perp}ZV\Lambda^{-2}V^TZ^TU_{\perp}ZV\Lambda^{-1}U^T)$$

$$+ \mathbb{E} \tr (VA^{-1}U^TZV_{\perp}V^TZU_{\perp}ZV\Lambda^{-2}U^TZV_{\perp}ZV^TU\Lambda^{-1}V^T)$$

$$=(d_1^2 + d_2^2)\|\Lambda^{-3}\|_F^2 + O \left( \frac{r^2 d_{\max}^2 + r^3 d_{\max}}{\lambda^6} \right)$$

where the last equality is due to Lemma 12. Note that the smaller terms are omitted here.

Calculate $L_{44}^{(116)}$ in the proof of Lemma 4. By definition, we write

$$L_{44}^{(116)} = \mathbb{E} \tr (UA^{-1}V^TZ^TU_{\perp}ZV\Lambda^{-1}U^TZV^TU\Lambda^{-6}V^T)$$

$$+ \mathbb{E} \tr (VA^{-1}U^TZV_{\perp}V^TZU_{\perp}ZV\Lambda^{-1}V^TZV_{\perp}ZV^TU\Lambda^{-6}V^T)$$

$$=2\|\Lambda^{-4}\|_F^2 \cdot \mathbb{E} \tr(T_U) \tr(T_V) = 2\|\Lambda^{-4}\|_F^2 \cdot \mathbb{E} (\tr(T_U))^2$$

$$=2d_1^2 - d_2^2 \|\Lambda^{-4}\|_F^2 + 4d_1 - d_2 \|\Lambda^{-4}\|_F^2$$

where the last equality is due to Lemma 9.

Calculate $L_{26}^{(116)}$ in the proof of Lemma 4. By definition, we can write

$$L_{26}^{(116)} = \mathbb{E} \tr (UA^{-1}V^TZ^TU_{\perp}ZV\Lambda^{-1}U^TZV^TU\Lambda^{-6}V^T)$$

$$+ \mathbb{E} \tr (VA^{-1}U^TZV_{\perp}V^TZU_{\perp}ZV\Lambda^{-1}V^TZV_{\perp}ZV^TU\Lambda^{-6}V^T)$$

$$=d_1 \mathbb{E} \tr (\Lambda^{-4}U^TZV_{\perp}) + d_2 \mathbb{E} \tr (\Lambda^{-4}V^TU\Lambda^{-4})$$

$$=d_1 \mathbb{E} \tr(T_U^2)\|\Lambda^{-4}\|_F^2 + d_2 \mathbb{E} \tr(T_V^2)\|\Lambda^{-4}\|_F^2$$

$$=d_1 d_2 + (d_1 + 1)\|\Lambda^{-4}\|_F^2$$

where the last equality is due to Lemma 9.

Calculate $L_{26}^{(125)}$ in the proof of Lemma 4. By definition, we write

$$L_{26}^{(125)} = \mathbb{E} \tr (UA^{-1}V^TZ^TU_{\perp}ZV\Lambda^{-2}V^TZ^TU_{\perp}ZV\Lambda^{-5}U^T)$$

$$+ \mathbb{E} \tr (VA^{-1}U^TZV_{\perp}V^TZU_{\perp}ZV\Lambda^{-2}U^TZV_{\perp}ZV^TU\Lambda^{-5}V^T)$$

66
\[
E(V^T Z U^\perp Z V A^{-2} V^T Z U^\perp E(U^\perp T U^\perp Z V A^{-6})
+ E \text{ tr} (U^T Z V U^\perp Z V A^{-2} V^T Z V U^\perp E(V^\perp T V U^\perp Z V A^{-6})
= E(V^T Z U^\perp Z V A^{-2} V^T Z U^\perp Z V A^{-6})(d_1 - d_2 - d_2^2 + d_2 - d_2^2)
+ E \text{ tr} (U^T Z V U^\perp Z V A^{-2} V^T Z V U^\perp Z V A^{-6})(d_1 - d_2 - d_2^2 + d_2 - d_2^2),
\]
where the last equality is due to Lemma [13]. Now, apply Lemma [9], we immediately obtain
\[
L_26^{(125)} = d_1 - d_2 - d_2^2 \| \Lambda^{-4} \|_F^2 + O \left( \frac{r d_2^2}{\lambda_2^8} \right).
\]

**Calculate \( L_242^{(1115)} \) in the proof of Lemma 4.** By definition, we have
\[
L_242^{(1115)} = E \text{ tr} (U A^{-1} V^T Z U^\perp Z V A^{-1} V^T Z U A^{-1} V^T Z U A^{-1} V^T Z U A^{-5} U T)
+ E \text{ tr} (V A^{-1} U T Z U^\perp Z V U A^{-1} V^T Z U A^{-1} V^T Z U A^{-5} V T)
= E \text{ tr} (V^T Z U^\perp Z V A^{-2} V^T Z U^\perp Z V A^{-6})E \text{ tr}(U T)
+ E \text{ tr} (U^T Z V U^\perp Z V A^{-2} U^T Z V U^\perp Z V A^{-6})E \text{ tr}(U T)
= d_1 - d_2 - E \text{ tr} (V^T Z U^\perp Z V A^{-2} V^T Z U^\perp Z V A^{-6})
+ d_1 - d_2 - E \text{ tr} (U^T Z V U^\perp Z V A^{-2} U^T Z V U^\perp Z V A^{-6})
= d_1 - d_2 - (d_1^2 + d_2^2) \| \Lambda^{-4} \|_F^2 + O \left( \frac{r d_2^2}{\lambda_2^8} \right),
\]
where the last equality is due to Lemma [9].

**Calculate \( L_242^{(1214)} \) in the proof of Lemma 4.** By definition, we write
\[
L_242^{(1214)} = E \text{ tr} (U A^{-1} V^T Z U^\perp Z V A^{-2} V^T Z U^\perp Z V A^{-4} U T)
+ E \text{ tr} (V A^{-1} U T Z U^\perp Z V U A^{-1} V^T Z U A^{-1} V^T Z U A^{-4} V T)
= d_2 - E \text{ tr} (U A^{-1} V^T Z U^\perp Z V A^{-2} V^T Z U^\perp Z V A^{-4} U T)
+ d_1 - E \text{ tr} (V A^{-1} U T Z U^\perp Z V A^{-2} U^T Z U A^{-4} V T)
= d_2^2 - E \text{ tr} (U A^{-1} V^T Z U^\perp Z V A^{-2} V^T Z U^\perp Z V A^{-4} U T)
+ d_1^2 - E \text{ tr} (V A^{-1} U T Z U^\perp Z V A^{-2} U^T Z U A^{-4} V T)
= 2 d_1^2 - d_2^2 - \| \Lambda^{-4} \|_F^2 + O \left( \frac{r d_2^2}{\lambda_2^8} \right),
\]
Calculate $L^{(1223)}_{242}$ in the proof of Lemma 4. By definition, we write

$$L^{(1223)}_{242} = E \text{ tr } (UA^{-1}VTZTU \perp V \perp ZVA - 2V^T ZTU \perp ZVA - 2V^T ZTU \perp ZVA - 3U^T)$$

$$+ E \text{ tr } (VA^{-1}UTZV \perp V \perp ZUA - 2UTZTV \perp ZUA - 2UTZV \perp ZUA - 3V^T)$$

$$= d_2 - E \text{ tr } (UA^{-1}VTZTU \perp V \perp ZVA - 2V^T ZTU \perp ZVA - 3U^T)$$

$$+ d_1 - E \text{ tr } (VA^{-1}UTZV \perp V \perp ZUA - 2UTZTV \perp ZUA - 3V^T)$$

$$= d_1 - d_2 - (d_1^2 + d_2^2) ||\Lambda^-4||_F^2 + O \left( \frac{r \delta_{\max}^3}{\lambda^4} \right)$$

where the last equality is due to Lemma 12.

Calculate $L^{(1115)}_{224}$ in the proof of Lemma 4. By definition, we have

$$L^{(1115)}_{224} = E \text{ tr } (UA^{-1}VTZTU \perp V \perp ZVA - 1VTZV \perp V \perp ZTA - 1VTZTU \perp V \perp ZVA - 5U^T)$$

$$+ E \text{ tr } (VA^{-1}UTZV \perp V \perp ZUA - 1VTZTU \perp V \perp ZVA - 5V^T)$$

$$= d_2 - E \text{ tr } (UA^{-1}VTZTU \perp V \perp ZVA - 2V^T ZTU \perp ZVA - 5U^T)$$

$$+ d_1 - E \text{ tr } (VA^{-1}UTZV \perp V \perp ZUA - 2UTZTV \perp ZUA - 5V^T)$$

$$= d_2 - d_1^2 - d_2^2 ||\Lambda^-4||_F^2 + O \left( \frac{r \delta_{\max}^3}{\lambda^4} \right),$$

where the last equality is due to Lemma 9.

Calculate $L^{(1124)}_{224}$ in the proof of Lemma 4. By definition, we have

$$L^{(1124)}_{224} = E \text{ tr } (UA^{-1}VTZTU \perp V \perp ZVA - 1VTZV \perp V \perp ZTA - 2VTZV \perp V \perp ZVA - 4U^T)$$

$$+ E \text{ tr } (VA^{-1}UTZV \perp V \perp ZUA - 2VTZTV \perp ZUA - 4V^T)$$

$$= d_1 - E \text{ tr } (UA^{-1}UTZV \perp V \perp ZUA - 2UTZTU \perp V \perp ZVA - 4U^T)$$

$$+ d_2 - E \text{ tr } (VA^{-1}UTZV \perp V \perp ZUA - 2VTZTU \perp V \perp ZVA - 4V^T)$$

$$= d_1 - d_2^2 - d_1^2 ||\Lambda^-4||_F^2 + O \left( \frac{r \delta_{\max}^3}{\lambda^4} \right).$$

68
Calculate $L_{224}^{(1214)}$ in the proof of Lemma 4. By definition, we have
\[
L_{224}^{(1214)} = \mathbb{E} \left( \operatorname{tr} \left( UA^{-1}V^T Z^T U \perp U \perp \perp Z \Lambda^{-2} V^T Z^T U \perp U \perp \perp Z \Lambda^{-4} U^T \right) \right) \\
+ \mathbb{E} \left( \operatorname{tr} \left( VA^{-1} U^T Z V \perp V \perp Z U \Lambda^{-2} U^T Z V \perp V \perp Z U \Lambda^{-4} U^T \right) \right) \\
= d_1 d_2 \mathbb{E} \left( \operatorname{tr} \left( UA^{-1}V^T Z^T U \perp U \perp \perp Z \Lambda^{-2} V^T Z^T U \perp U \perp \perp Z \Lambda^{-4} U^T \right) \right) \\
+ d_1 d_2 \mathbb{E} \left( \operatorname{tr} \left( VA^{-1} U^T Z V \perp V \perp Z U \Lambda^{-2} U^T Z V \perp V \perp Z U \Lambda^{-4} U^T \right) \right) \\
= d_1 d_2 \left( d_1^2 + d_2^2 \right) \mathbb{E} \left( \operatorname{tr} \left( \Lambda^{-2} \right) \right) \\
+ O \left( \frac{r d_{\max}^3}{\lambda_{\perp}^8} \right),
\]
where the last equality is due to Lemma 11.

Calculate $L_{224}^{(1223)}$ in the proof of Lemma 4. By definition, we have
\[
L_{224}^{(1223)} = \mathbb{E} \left( \operatorname{tr} \left( UA^{-1}V^T Z^T U \perp U \perp \perp Z \Lambda^{-2} V^T Z^T U \perp U \perp \perp Z \Lambda^{-4} U^T \right) \right) \\
+ \mathbb{E} \left( \operatorname{tr} \left( VA^{-1} U^T Z V \perp V \perp Z U \Lambda^{-2} U^T Z V \perp V \perp Z U \Lambda^{-4} U^T \right) \right) \\
= d_1 d_2 \mathbb{E} \left( \operatorname{tr} \left( UA^{-1}V^T Z^T U \perp U \perp \perp Z \Lambda^{-2} V^T Z^T U \perp U \perp \perp Z \Lambda^{-4} U^T \right) \right) \\
+ d_1 d_2 \mathbb{E} \left( \operatorname{tr} \left( VA^{-1} U^T Z V \perp V \perp Z U \Lambda^{-2} U^T Z V \perp V \perp Z U \Lambda^{-4} U^T \right) \right) \\
= d_1 d_2 \left( d_1^2 + d_2^2 \right) \mathbb{E} \left( \operatorname{tr} \left( \Lambda^{-2} \right) \right) \\
+ O \left( \frac{r d_{\max}^3}{\lambda_{\perp}^8} \right),
\]
where the last equality is due to Lemma 12.

Calculate $L_{2222}^{(11114)}$ in the proof of Lemma 4. By definition, we write
\[
L_{2222}^{(11114)} = \mathbb{E} \left( \operatorname{tr} \left( UA^{-1}V^T Z^T U \perp U \perp \perp Z \Lambda^{-2} V^T Z^T U \perp U \perp \perp Z \Lambda^{-4} U^T \right) \right)
\]
\[
+E \operatorname{tr} (V \Lambda^{-1} U^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-1} U^T Z U \perp U \perp^T Z V \Lambda^{-4} V^T ) \\
= d_2 \cdot E \operatorname{tr} (\Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-5} ) \operatorname{tr} (\Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-1} ) + (d_2^2 + d_2^-) \operatorname{diag} (\Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-5} ) \operatorname{diag} (\Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-1} ) \\
+ d_1 \cdot E \operatorname{tr} (\Lambda^{-1} \tilde{Z}_2 \tilde{Z}_2^T \Lambda^{-5} ) \operatorname{tr} (\Lambda^{-1} \tilde{Z}_2 \tilde{Z}_2^T \Lambda^{-1} ) + (d_1^2 + d_1^-) \operatorname{diag} (\Lambda^{-1} \tilde{Z}_2 \tilde{Z}_2^T \Lambda^{-5} ) \operatorname{diag} (\Lambda^{-1} \tilde{Z}_2 \tilde{Z}_2^T \Lambda^{-1} ) \\
\]
where the last equality is due to Lemma [13] and \( \tilde{Z}_1 \in \mathbb{R}^{r \times d_1^-} \), \( \tilde{Z}_2 \in \mathbb{R}^{r \times d_2^-} \) have i.i.d. standard normal entries. Since \( E \operatorname{tr} (\Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-5} ) \operatorname{tr} (\Lambda^{-1} \tilde{Z}_1 \tilde{Z}_1^T \Lambda^{-1} ) = O (r^2 d_2^0) \), we end up with
\[
L_{2222}^{(1114)} = 2d_1^2 d_2^- \| \Lambda^{-4} \|_F^2 + O \left( \frac{r^2 d_2^0}{\lambda_p^8} \right).
\]

**Calculate \( L_{2222}^{(11123)} \) in the proof of Lemma 4.**

By definition, we write
\[
L_{2222}^{(11123)} = E \operatorname{tr} (U \Lambda^{-1} V^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
+ E \operatorname{tr} (V \Lambda^{-1} U^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-1} V^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
= d_2 \cdot E \operatorname{tr} (U \Lambda^{-1} V^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
+ d_1 \cdot E \operatorname{tr} (V \Lambda^{-1} U^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
= d_1 d_2^- (d_1^- + d_2^-) \| \Lambda^{-1} \|_F^2 + O \left( \frac{r d_2^0}{\lambda_p^8} \right),
\]
where the last equality is due to Lemma [12]

**Calculate \( L_{2222}^{(11123)} \) in the proof of Lemma 4.**

By definition, we write
\[
L_{2222}^{(11123)} = E \operatorname{tr} (U \Lambda^{-1} V^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
+ E \operatorname{tr} (V \Lambda^{-1} U^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-1} V^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
= E \operatorname{tr} (U \Lambda^{-1} V^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-1} V^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
+ E \operatorname{tr} (V \Lambda^{-1} U^T Z V \perp V \perp^T Z U \perp U \perp^T Z V \Lambda^{-1} V^T Z U \perp U \perp^T Z V \Lambda^{-2} V^T Z U \perp U \perp^T Z V \Lambda^{-3} U^T ) \\
\]
where the last equality is due to Lemma [13] Therefore, by Lemma [11] we get
\[
L_{2222}^{(11123)} = 2d_1^- d_2^- \| \Lambda^{-1} \|_F^2 + O \left( \frac{r^2 d_2^0}{\lambda_p^8} \right).
\]
Calculate $L^{(1224)}_{2222}$ in the proof of Lemma 4. By definition, we write

$$L^{(1224)}_{2222} = \mathbb{E} \text{tr} \left( UA^{-1} V^T Z^T U_1 U_1^T Z V A^{-1} U^T Z V_1 V_1^T Z U A^{-2} U^T Z V_1 V_1^T Z U A^{-4} U^T \right)$$

$$+ \mathbb{E} \text{tr} \left( VA^{-1} U^T Z V_1 V_1^T Z U A^{-1} V^T Z U_1 U_1^T Z V A^{-2} V^T Z U_1 U_1^T Z U A^{-4} V^T \right)$$

$$= d_1 \mathbb{E} \text{tr} \left( U A^{-2} U^T Z V_1 V_1^T Z U A^{-2} U^T Z V_1 V_1^T Z U A^{-4} U^T \right)$$

$$+ d_2 \mathbb{E} \text{tr} \left( VA^{-2} V^T Z U_1 U_1^T Z V A^{-2} V^T Z U_1 U_1^T Z V A^{-4} V^T \right)$$

$$= d_1 d_2 (d_1^2 + d_2^2) \| \Lambda^{-4} \|_F^2 + O \left( \frac{r^2 d_{\text{max}}^3}{\lambda_r^8} \right),$$

where the last equality is due to Lemma 12.

Calculate $L^{(1222)}_{2222}$ in the proof of Lemma 4. By definition, we write

$$L^{(1222)}_{2222} = \mathbb{E} \text{tr} \left( UA^{-1} V^T Z^T U_1 U_1^T Z V A^{-2} V^T Z^T U_1 U_1^T Z V A^{-4} U^T Z V_1 V_1^T Z U A^{-2} U^T \right)$$

$$+ \mathbb{E} \text{tr} \left( VA^{-1} U^T Z V_1 V_1^T Z U A^{-1} V^T Z U_1 U_1^T Z V A^{-2} V^T Z U_1 U_1^T Z V A^{-4} V^T \right)$$

$$= \mathbb{E} \text{tr} \left( U A^{-1} V^T Z^T U_1 U_1^T Z V A^{-2} V^T Z^T U_1 U_1^T Z V A^{-4} U^T \right)$$

$$+ \mathbb{E} \text{tr} \left( VA^{-1} U^T Z V_1 V_1^T Z U A^{-2} V^T Z U_1 U_1^T Z V A^{-4} V^T \right)$$

$$= (d_1^4 + d_2^4) \| \Lambda^{-4} \|_F^2 + O \left( \frac{r^2 d_{\text{max}}^3}{\lambda_r^8} \right),$$

where the last equality is due to Lemma 13. Therefore, by Lemma 11, we end up with

$$L^{(1222)}_{2222} = 2d_1^2 d_2^2 \| \Lambda^{-4} \|_F^2 + O \left( \frac{r^2 d_{\text{max}}^3}{\lambda_r^8} \right).$$

Calculate $L^{(1221)}_{2222}$ in the proof of Lemma 4. By definition, we write

$$L^{(1221)}_{2222} = \mathbb{E} \text{tr} \left( UA^{-1} V^T Z^T U_1 U_1^T Z V A^{-2} V^T Z^T U_1 U_1^T Z V A^{-1} U^T Z V_1 V_1^T Z U A^{-2} U^T \right)$$

$$+ \mathbb{E} \text{tr} \left( VA^{-1} U^T Z V_1 V_1^T Z U A^{-1} V^T Z U_1 U_1^T Z V A^{-2} V^T Z U_1 U_1^T Z V A^{-1} U^T \right)$$

$$= (d_1^4 + d_2^4) \| \Lambda^{-4} \|_F^2 + O \left( \frac{r^2 d_{\text{max}}^3}{\lambda_r^8} \right)$$

where the last equality is due to Lemma 14.

8.2 Supporting lemmas

Proof of Lemma 4. Recall that

$$f_t(X_1) = \sum_{k \geq 3} \langle \Theta \Theta^T, S_{A,k}(X_1) \rangle \phi \left( \frac{\| X_1 \|}{t \cdot d_{\text{max}}^{1/2}} \right).$$
Case 1: if \( \|X_1\| > 2td_{\text{max}}^{1/2} \) and \( \|X_2\| > 2td_{\text{max}}^{1/2} \), then \( f_t(X_1) = f_t(X_2) = 0 \) by definition of \( \phi(\cdot) \) where the claimed inequality holds trivially.

Case 2: if \( \|X_1\| \leq 2td_{\text{max}}^{1/2} \) and \( \|X_2\| > 2td_{\text{max}}^{1/2} \), then \( f_t(X_2) = 0 \). We get, by Lipschitz property of \( \phi(\cdot) \), that

\[
\left| f_t(X_1) - f_t(X_2) \right| = \left| \sum_{k \geq 3} \left( \Theta \Theta^T, S_{A,k}(X_1) \right) \cdot \left( \phi\left( \frac{\|X_1\|}{t \cdot d_{\text{max}}^{1/2}} \right) - \phi\left( \frac{\|X_2\|}{t \cdot d_{\text{max}}^{1/2}} \right) \right) \right|
\]

\[
\leq \sum_{k \geq 3} 2r \left\| S_{A,k}(X_1) \right\| \cdot \frac{\|X_1 - X_2\|_F}{t \cdot d_{\text{max}}^{1/2}}
\]

\[
\leq \frac{2r}{t \cdot d_{\text{max}}^{1/2}} \sum_{k \geq 3} \sum_{s: s_1 + \cdots + s_{k+1} = k} \left\| \mathcal{P}^{-s_1} X_1 \mathcal{P}^{-s_2} X_1 \cdots X_1 \mathcal{P}^{-s_{k+1}} \right\|
\]

\[
\leq \frac{2r}{t \cdot d_{\text{max}}^{1/2}} \sum_{k \geq 3} \sum_{s: s_1 + \cdots + s_{k+1} = k} \|X_1\|^k \lambda_r^k
\]

\[
\leq \frac{2r}{t \cdot d_{\text{max}}^{1/2}} \sum_{k \geq 3} \left( \frac{4}{\lambda_r} \|X_1\| \right)^k \lambda_r^{3/2}
\]

\[
\leq C_4 t^2 r \|X_1 - X_2\|_F \frac{d_{\text{max}}^{1/2}}{d_{\text{max}}^{1/2}} \lambda_r^{3/2}
\]

where the last inequality holds as long as \( \lambda_r \geq 9td_{\text{max}}^{1/2} \).

Case 3: if \( \|X_1\| \leq 2td_{\text{max}}^{1/2} \) and \( \|X_2\| \leq 2td_{\text{max}}^{1/2} \). Then,

\[
\left| f_t(X_1) - f_t(X_2) \right| \leq 2r \sum_{k \geq 3} \left\| S_{A,k}(X_1) \phi\left( \frac{\|X_1\|}{t \cdot d_{\text{max}}^{1/2}} \right) - S_{A,k}(X_2) \phi\left( \frac{\|X_2\|}{t \cdot d_{\text{max}}^{1/2}} \right) \right\|
\]

\[
\leq 2r \sum_{k \geq 3} \sum_{s: s_1 + \cdots + s_{k+1} = k} \left\| \mathcal{P}^{-s_1} X_1 \cdots X_1 \mathcal{P}^{-s_{k+1}} \phi\left( \frac{\|X_1\|}{t \cdot d_{\text{max}}^{1/2}} \right) - \mathcal{P}^{-s_1} X_2 \cdots X_2 \mathcal{P}^{-s_{k+1}} \phi\left( \frac{\|X_2\|}{t \cdot d_{\text{max}}^{1/2}} \right) \right\|
\]

\[
\leq 2r \sum_{k \geq 3} \sum_{s: s_1 + \cdots + s_{k+1} = k} \left( k + 2 \right) \frac{3}{(2td_{\text{max}}^{1/2})^{k-1}} \|X_1 - X_2\|_F \leq C_4 t^2 \cdot \frac{v d_{\text{max}}}{\lambda_r^3} \|X_1 - X_2\|_F
\]

where the last inequality holds as long as \( \lambda_r \geq 9td_{\text{max}}^{1/2} \). Therefore, we conclude the proof of Lemma 6.

\[
\square
\]

**Lemma 7.** Let \( A \in \mathbb{R}^{d \times d} \) be a deterministic symmetric matrix and \( z \sim \mathcal{N}(0, I_d) \). Then,

\[
\mathbb{E}(z^T A z)^2 = 2\|A\|_F^2 + (\text{tr}(A))^2.
\]

72
Proof of Lemma 7. Denote by $A = \sum_{i=1}^{d} \lambda_i(\theta_i\theta_i^T)$ the eigen-decomposition of $A$. Then,

$$z^T A z = \sum_{i=1}^{d} \lambda_i (z^T \theta_i)^2$$

where we conclude that $\{(z^T \theta_i)\}_{i=1}^{d}$ are i.i.d. standard normal random variables, since $\{\theta_i\}_{i=1}^{d}$ are orthonormal vectors. Therefore,

$$\mathbb{E}(z^T A z)^2 = \sum_{i=1}^{d} \lambda_i^2 \cdot \mathbb{E}(z^T \theta_i)^4 + \sum_{1 \leq i_1 \neq i_2 \leq d} \lambda_{i_1} \lambda_{i_2} \cdot \mathbb{E}(z^T \theta_{i_1})^2 \mathbb{E}(z^T \theta_{i_2})^2$$

$$= 2 \|A\|_F^2 + \left( \sum_{i=1}^{d} \lambda_i \right)^2 = 2 \|A\|_F^2 + \left( \text{tr}(A) \right)^2.$$

Lemma 8. Let $Z \in \mathbb{R}^{d_1 \times d_2}$ be a random matrix containing i.i.d. standard normal entries. Then,

$$\mathbb{E}\|ZZ^T\|_F^2 = d_1 d_2 (d_1 + d_2 + 1).$$

Proof of Lemma 8. Let $z_i^T \in \mathbb{R}^{d_2}$ be the $i$-th row of $Z$. Then,

$$\mathbb{E}\|ZZ^T\|_F^2 = \sum_{i_1=1}^{d_1} \mathbb{E}(z_{i_1}z_{i_1})^2 + \sum_{1 \leq i_1 \neq i_2 \leq d_1} \mathbb{E}(z_{i_1}z_{i_2})^2$$

$$= d_1 d_2 (d_1 + d_2 + 1).$$

Lemma 9. Let $Z \in \mathbb{R}^{d_1 \times d_2}$ be a random matrix containing i.i.d. standard normal entries. Then,

$$\mathbb{E}\|Z\|_F^4 = d_1^2 d_2^2 + 2d_1 d_2.$$

Proof of Lemma 9. Denote by $\{z_i\}_{i=1}^{d_1 d_2}$ entries of $Z$. Then,

$$\mathbb{E}\|Z\|_F^4 = \sum_{i=1}^{d_1 d_2} \mathbb{E}z_i^4 + \sum_{1 \leq i_1 \neq i_2 \leq d_1 d_2} \mathbb{E}z_{i_1}^2 \mathbb{E}z_{i_2}^2$$

$$= d_1^2 d_2^2 + 2d_1 d_2.$$
Lemma 10. Let \( A, B \) be deterministic \( d \times d \) symmetric matrix and \( z \sim \mathcal{N}(0, I_d) \). Then,
\[
\mathbb{E}(z^T A z)(z^T B z) = \text{tr}(A) \text{tr}(B) + 2 \text{tr}(AB)
\]

Proof of Lemma 10. Write
\[
z^T A z = \sum_{i_1, i_2} A_{i_1 i_2} z_{i_1} z_{i_2} \quad \text{and} \quad z^T B z = \sum_{j_1, j_2} B_{j_1 j_2} z_{j_1} z_{j_2}
\]
where \( \{z_j\}_{j=1}^d \) denotes the entries of \( z \). Then,
\[
\mathbb{E}(z^T A z)(z^T B z) = \mathbb{E} \left( \sum_{i_1} A_{i_1 i_1} z_{i_1}^2 + \sum_{1 \leq i_1 \neq i_2} A_{i_1 i_2} z_{i_1} z_{i_2} \right) \left( \sum_{i_1} B_{i_1 i_1} z_{i_1}^2 + \sum_{1 \leq i_1 \neq i_2} B_{i_1 i_2} z_{i_1} z_{i_2} \right)
\]
\[
= \text{tr}(A) \text{tr}(B) + 2 \text{tr}(AB).
\]

\[\square\]

Lemma 11. Let \( \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r) \) and \( Z \in \mathbb{R}^{r \times d} \) be a random matrix containing i.i.d. standard normal entries. Then, for any positive numbers \( j_1, j_2 \), we have
\[
\mathbb{E}\|\Lambda^{-j_1} Z Z^T \Lambda^{-j_2}\|_F^2 = d^2 \|\Lambda^{-j_1-j_2}\|_F^2 + d \left( \|\Lambda^{-j_1-j_2}\|_F^2 + \|\Lambda^{-j_1}\|_F^2 \right).
\]

Proof of Lemma 11. Let \( z_1, \cdots, z_r \in \mathbb{R}^d \) denote the columns of \( Z^T \). Therefore, we can write
\[
\|\Lambda^{-j_1} Z Z^T \Lambda^{-j_2}\|_F^2 = \sum_{i=1}^r \frac{1}{\lambda_i^{2(j_1+j_2)}} (z_i^T z_i)^2 + \sum_{1 \leq i_1 \neq i_2 \leq r} \frac{1}{\lambda_{i_1}^{2j_1} \lambda_{i_2}^{2j_2}} (z_{i_1}^T z_{i_2})^2.
\]
Then, we get
\[
\mathbb{E}\|\Lambda^{-1} Z Z^T \Lambda^{-1}\|_F^2 = \sum_{i=1}^r \frac{d^2 + 2d}{\lambda_i^{2(j_1+j_2)}} + \sum_{1 \leq i_1 \neq i_2 \leq r} \frac{d}{\lambda_{i_1}^{2j_1} \lambda_{i_2}^{2j_2}}
\]
\[
= d^2 \|\Lambda^{-j_1-j_2}\|_F^2 + d \left( \|\Lambda^{-j_1-j_2}\|_F^2 + \|\Lambda^{-j_1}\|_F^2 \right).
\]

\[\square\]

Lemma 12. Let \( Z \in \mathbb{R}^{r \times d} \) contain i.i.d. standard normal entries and \( \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r) \). Then, for positive integers \( j_1, j_2, j_3, j_4 \), we have
\[
\mathbb{E} \text{tr} \left( \Lambda^{-j_1} Z Z^T \Lambda^{-j_2} Z Z^T \Lambda^{-j_3} Z Z^T \Lambda^{-j_4} \right) = (d^6 + 3d^2 + 4d) \text{tr}(\Lambda^{-j_1-j_2-j_3-j_4})
\]
\[
+ (d^2 + d) \cdot \left[ \text{tr}(\Lambda^{-j_1-j_3}) \text{tr}(\Lambda^{-j_2-j_4}) + \text{tr}(\Lambda^{-j_1-j_2-j_4}) \text{tr}(\Lambda^{-j_3}) + \text{tr}(\Lambda^{-j_1-j_3-j_4}) \text{tr}(\Lambda^{-j_2}) \right]
\]
\[
+ d \cdot \text{tr}(\Lambda^{-j_2}) \text{tr}(\Lambda^{-j_3}) \text{tr}(\Lambda^{-j_1-j_4})
\]

74
Proof of Lemma 12. For integer $1 \leq k \leq r$,

\[
\left(\Lambda^{-j_1} ZZ^T \Lambda^{-j_2} ZZ^T \Lambda^{-j_3} ZZ^T \Lambda^{-j_4}\right)_{kk} = \frac{1}{\lambda_k^{j_1+j_4}} z_k^T \Lambda^{-j_2} ZZ^T \Lambda^{-j_3} Z z_k
\]

where $\{z_k\}_{k=1}^r$ denote the columns of $Z$. We write

\[
\frac{1}{\lambda_k^{j_1+j_4}} z_k^T \Lambda^{-j_2} ZZ^T \Lambda^{-j_3} Z z_k = \frac{1}{\lambda_k^{j_1+j_4+j_2}} \|z_k\|_{\ell_2}^2 \cdot z_k^T \left( \sum_{i_1 \neq k} \frac{1}{\lambda_{i_1}} z_{i_1} z_{i_1}^T \right) z_k
\]

\[
+ \frac{1}{\lambda_k^{j_1+j_4+j_3}} \|z_k\|_{\ell_2}^6 \cdot z_k^T \left( \sum_{i_1 \neq k} \frac{1}{\lambda_{i_1}} z_{i_1} z_{i_1}^T \right) \left( \sum_{i_2 \neq k} \frac{1}{\lambda_{i_2}} z_{i_2} z_{i_2}^T \right) z_k.
\]

Taking expectation and simplify the expressions, we obtain the claimed equation.

Lemma 13. Let $B \in \mathbb{R}^{r \times r}$ be a deterministic symmetric matrix and $Z \in \mathbb{R}^{r \times d}$ contain i.i.d. standard normal entries. Then,

\[
\mathbb{E}(ZZ^T B ZZ^T) = d \cdot \text{tr}(B) \cdot I_r + (d^2 + d) \cdot \text{diag}(B).
\]

Proof of Lemma 13. Since $Z$ has i.i.d. standard normal entries, we can assume $B$ is a diagonal matrix. Then,

\[
(ZZ^T B ZZ^T)_{ii} = z_i^T \left( \sum_{j \neq i} B_{jj} z_j z_j^T \right) z_i + B_{ii} \|z_i\|_{\ell_2}^4.
\]

The claim immediately follows by taking expectation.

Lemma 14. Let $Z \in \mathbb{R}^{r \times d}$ contain i.i.d. standard normal entries and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$. Then, for any non-negative numbers $j_1, j_2, j_3, j_5$,

\[
\mathbb{E} \text{tr}(\Lambda^{-j_1} ZZ^T \Lambda^{-j_2} ZZ^T \Lambda^{-j_3} ZZ^T \Lambda^{-j_4} ZZ^T) = d^4 \cdot \text{tr}(\Lambda^{-j_1-j_2-j_3-j_4}) + O(r^2 d^3 / \lambda_1^{j_1+j_2+j_3+j_4}).
\]

Proof of Lemma 14. We denote by $z_i^T \in \mathbb{R}^d$ the $i$-th row of $Z$. Then, we write

\[
\text{tr}(\Lambda^{-j_1} ZZ^T \Lambda^{-j_2} ZZ^T \Lambda^{-j_3} ZZ^T \Lambda^{-j_4} ZZ^T)
\]

\[
= \sum_{i_1=1}^r \frac{1}{\lambda_i^{j_1}} z_i^T \left( \sum_{i_2=1}^r \frac{1}{\lambda_{i_2}} z_{i_2} z_{i_2}^T \right) \left( \sum_{i_3=1}^r \frac{1}{\lambda_{i_3}} z_{i_3} z_{i_3}^T \right) \left( \sum_{i_4=1}^r \frac{1}{\lambda_{i_4}} z_{i_4} z_{i_4}^T \right) z_i.
\]

By the moment of Chi squared random variables, the claim immediately follows by taking expectation.
References

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Zhigang Bao, Xiucai Ding, and Ke Wang. Singular vector and singular subspace distribution for the matrix denoising model. *arXiv preprint arXiv:1809.10476*, 2018.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Andrew C Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

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.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Yuxin Chen, Chen Cheng, and Jianqing Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*, 2018.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.
Xiucai Ding. High dimensional deformed rectangular matrices with applications in matrix denoising. *arXiv:1702.06975*, 2017.

David Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

Carl-Gustav Esseen. On the liapunoff limit of error in the theory of probability. *Arkiv for matematik, astronomi och fysik*, A28:1–19, 1942.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

Hyunsoo Kim, Peg Howland, and Haesun Park. Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6(Jan):37–53, 2005.

Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1976–2013. Institut Henri Poincaré, 2016.

Vladimir Koltchinskii and Dong Xia. Perturbation of linear forms of singular vectors under gaussian noise. In *High Dimensional Probability VII*, pages 397–423. Springer, 2016.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Andrey A Shabalin and Andrew B Nobel. Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.
Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

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

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Dong Xia. Confidence interval of singular vectors for high-dimensional and low-rank matrix regression. *arXiv preprint arXiv:1805.09871*, 2018.

Dong Xia and Ming Yuan. On polynomial time methods for exact low rank tensor completion. *Foundations of Computational Mathematics*, 2018+.

Dong Xia and Fan Zhou. The $\ell_\infty$ perturbation of hosvd and low rank tensor denoising. *arXiv preprint arXiv:1707.01207*, 2017.

Dong Xia, Ming Yuan, and Cun-Hui Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 2018.

Qinqing Zheng and Ryota Tomioka. Interpolating convex and non-convex tensor decompositions via the subspace norm. In *Advances in Neural Information Processing Systems*, pages 3106–3113, 2015.