# Normal approximation and confidence region of singular subspaces[*]

## Dong Xia

*Department of Mathematics,*
*Hong Kong University of Science and Technology,*
*Clear Water Bay, Kowloon, Hong Kong*
*e-mail:* madxia@ust.hk

**Abstract:** This paper is on the normal approximation of singular subspaces when the noise matrix has i.i.d. entries. Our contributions are threefold. First, we derive an explicit representation formula of the empirical spectral projectors. The formula is neat and holds for deterministic matrix perturbations. Second, we calculate the expected projection distance between the empirical singular subspaces and true singular subspaces. Our method allows obtaining arbitrary $k$-th order approximation of the expected projection distance. Third, we prove the non-asymptotical normal approximation of the projection distance with different levels of bias corrections. By the $\lceil \log(d_1 + d_2) \rceil$-th order bias corrections, the asymptotical normality holds under optimal signal-to-noise ratio (SNR) condition where $d_1$ and $d_2$ denote the matrix sizes. In addition, it shows that higher order approximations are unnecessary when $|d_1 - d_2| = O((d_1 + d_2)^{1/2})$. Finally, we provide comprehensive simulation results to merit our theoretic discoveries.

Unlike the existing results, our approach is non-asymptotical and the convergence rates are established. Our method allows the rank $r$ to diverge as fast as $o((d_1 + d_2)^{1/3})$. Moreover, our method requires no eigen-gap condition (except the SNR) and no constraints between $d_1$ and $d_2$.

**MSC2020 subject classifications:** Primary 62H10, 62H25; secondary 62G20.

**Keywords and phrases:** Singular value decomposition, projection distance, normal approximation, random matrix theory, spectral perturbation.

Received November 2020.

## 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 and etc. See, e.g., Golub and Van Loan (2012), for more examples. In many 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), Candès and Tao (2010), Xia and Yuan (2018), Xia et al. (2017)), and phase retrieval (Ma et al. (2017), Candes et al. (2015)), where SVD is

often applied for generating a warm initial point for non-convex optimization algorithms. In statistical data analysis, SVD is superior for denoising and dimension reduction. For instance, SVD, as a dimension reduction tool, is used for text classification in Kim et al. (2005). See also Li and Wang (2007). In Shabalin and Nobel (2013), SVD shows appealing performances in low rank matrix denoising. More specifically, in Donoho and Gavish (2014), they proved that statistically minimax optimal matrix denoising can be attained via precise singular value thresholding. Recently, matrix SVD is generalized to tensor SVD for tensor denoising, see Xia and Zhou (2019) and Zhang and Xia (2018).

The perturbation analysis is critical for advancing the theoretical developments of SVD for low-rank matrix denoising where 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), Yu et al. (2014)) and Wedin (Wedin (1972)) many years ago. Among those deterministic perturbation bounds, one simple yet useful bound shows that the perturbation of singular vectors 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. It is a quite general result since the bound does not rely on the wellness of alignments between the singular subspaces of the information and of the noise matrices. Such a general bound turns out to be somewhat satisfactorily sharp when the noise matrix contains i.i.d. random entries. However, more refined characterizations of singular vectors are needed on the frontiers of statistical inference for matrix SVD. The Davis-Kahan Theorem and Wedin's perturbation bounds are illustrated by the non-zero smallest singular value of the information matrix, where the effects of those large singular values are usually missing. Moreover, the exact numerical factor is also not well recognized.

The behavior of singular values and singular vectors of low rank perturbations of large rectangular random matrices is popular in recent years. They play a key role in statistical inference with diverse applications. See Li and Li (2018), Naumov et al. (2017), Tang et al. (2018) for some examples in network testing. The asymptotic limits of singular values and singular vectors were firstly developed by Benaych-Georges and Nadakuditi (2012), where the convergence rate of the largest singular value was also established. Recently, by Ding (2017), more precise non-asymptotic concentration bounds for empirical singular values were obtained. Meanwhile, Ding (2017) also proved non-asymptotic perturbation bounds of empirical singular vector when the associated singular value has multiplicity 1. In a recent work (Bao et al., 2018), the authors studied the asymptotic limit distributions of the empirical singular subspaces when (scaled) singular values are bounded. Specifically, they showed that if the noise matrix has Gaussian distribution, then the limit distribution of the projection distance is also Gaussian. Unlike these prior arts (Ding (2017), Bao et al. (2018)), we focus on the non-asymptotical normal approximations of the joint singular subspaces in a different regime. Our approach allows the rank to diverge, and imposes no

constraints between $d_1$ and $d_2$. In addition, we establish the convergence rates and impose no eigen-gap conditions (except SNR).

In Xia (2019), the low rank matrix regression model is investigated where the author proposed a de-biased estimator built on nuclear normal penalized least squares estimator. The de-biased estimator ends up with an analogous form of the low rank perturbation of rectangular random matrices. Then, non-asymptotical normal approximation theory of the projection distance is proved, under near optimal sample size requirement. The paramount observation is that the mean value in the limit normal distribution is significantly larger than its standard deviation. As a result, a much larger than regular sample size requirement is necessary to tradeoff the estimation error of the expected projection distance. Most recently, Chen et al. (2018) revealed an interesting phenomenon of the perturbation of eigenvalues and eigenvectors of such non-asymmetric random perturbations, showing that the perturbation of eigen structures is much smaller than the singular structures. In addition, some non-asymptotic perturbation bounds of empirical singular vectors can be found in Koltchinskii and Xia (2016),Bloemendal et al. (2016) and Abbe et al. (2017). The minimax optimal bounds of singular subspace estimation for low rank perturbations of large rectangular random matrices are established in Cai and Zhang (2018).

Our goal is to investigate the central limit theorems of singular subspaces in the low rank perturbation model of large rectangular random matrices. As illustrated in Xia (2019), the major difficulty arises from how to precisely determine the expected projection distance. One conclusive contribution of this paper is an explicit representation formula of the empirical spectral projector. This explicit representation formula allows us to obtain precise characterization of the (non-asymptotical) expected projection distance. After those higher order bias corrections, we prove normal approximation of the singular subspaces with optimal (in the consistency regime) SNR requirement. For better presenting the results and highlighting the contributions, let's begin with introducing the standard notations. We denote $M = U\Lambda V^{\mathsf{T}}$ 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 diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_r)$ contains $M$'s non-increasing positive singular values. The observed data matrix $\hat{M} \in \mathbb{R}^{d_1 \times d_2}$ satisfies the additive model:

$$\hat{M} = M + Z \quad \text{where} \quad Z_{j_1 j_2} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \quad \text{for } 1 \le j_1 \le d_1, 1 \le j_2 \le d_2. \quad (1)$$

Here, we fix the noise variance to be 1, just for simplicity. For ease of exposition, let $d_1 \le d_2$. 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 $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ denote the top-$r$ singular values of $\hat{M}$. We focus on the projection distance between the empirical and true singular subspaces which is defined by

$$\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] := \|\hat{U}\hat{U}^{\mathsf{T}} - UU^{\mathsf{T}}\|_{\mathrm{F}}^2 + \|\hat{V}\hat{V}^{\mathsf{T}} - VV^{\mathsf{T}}\|_{\mathrm{F}}^2. \quad (2)$$

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 on the event $\{\lambda_r > 2\|Z\|\}$. It is well-known that $\|Z\| = O_P(\sqrt{d_2})$ where $\|\cdot\|$ denotes the spectral norm and $d_2 = \max\{d_1, d_2\}$. Therefore, it is convenient to consider $\lambda_r \gtrsim \sqrt{d_2}$. In this paper, we focus on the consistency regime[1] so that the empirical singular subspaces are consistent which requires $\lambda_r \gg \sqrt{rd_2}$. See, e.g., Tao ([2012](#)), Koltchinskii and Xia ([2016](#)), Cai and Zhang ([2018](#)) and Vershynin ([2010](#)).

Our contributions are summarized as follows.

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

2. By the representation formula, we prove the normal approximation of $\hat{\varepsilon}_1 := \big(\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]\big)/\big(\sqrt{8d_\star}\|\Lambda^{-2}\|_\mathrm{F}\big)$ 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 $\sqrt{rd_2}/\lambda_r \to 0$ and $r^3/d_2 \to 0$ as $d_1, d_2 \to \infty$. The required SNR is optimal in the consistency regime. Note that our result allows $r$ to diverge as fast as $o((d_1 + d_2)^{1/3})$. In addition, no conditions on the eigen-gaps (except $\lambda_r$) are required. The convergence rate is also established. The proof strategy is based on the Gaussian isoperimetric inequality and Berry-Esseen theorem.

3. The unknown $\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ plays the role of centering in $\hat{\varepsilon}_1$. To derive user-friendly normal approximations of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$, it suffices to explicitly calculate its expectation (non-asymptotically). By the representation formula of $\hat{U}\hat{U}^\mathsf{T}$ and $\hat{V}\hat{V}^\mathsf{T}$, we obtain approximations of $\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Different levels of approximating the expectation ends up with different levels of bias corrections. These levels of approximations are

   (a) Level-1 approximation: $B_1 = 2d_\star\|\Lambda^{-1}\|_\mathrm{F}^2$. The approximation error is
   $$\left|\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_1\right| = O\Big(\frac{rd_2^2}{\lambda_r^4}\Big).$$

   (b) Level-2 approximation: $B_2 = 2(d_\star\|\Lambda^{-1}\|_\mathrm{F}^2 - \Delta_d^2\|\Lambda^{-2}\|_\mathrm{F}^2)$ where $\Delta_d = d_1 - d_2$. Then,
   $$\left|\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_2\right| = O\Big(\frac{rd_2^3}{\lambda_r^6}\Big).$$

---

[1]We note that, in RMT literature (see, e.g., Bao et al. ([2018](#)),Ding ([2017](#))), many works studied the problem when $\lambda_r = O(\sqrt{d_2})$ and $\lambda_r \gtrsim (d_1 d_2)^{1/4}$. In this paper, we focus on the regime when empirical singular subspaces are consistent, i.e., $\mathbb{E}\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] \to 0$ when $d_2 \to \infty$. As shown in Cai and Zhang ([2018](#)), such consistency requires $\sqrt{rd_2}/\lambda_r \to 0$.

(c) Level-$k$ approximation: $B_k = 2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2 - 2\sum_{k_0=2}^{k}(-1)^{k_0}\Delta_d(d_{1-}^{k_0-1} - d_{2-}^{k_0-1})\|\Lambda^{-k_0}\|_{\mathrm{F}}^2$ where $d_{1-} = d_1 - r$ and $d_{2-} = d_2 - r$. Then, for all $k \geq 2$,

$$\left|\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] - B_k\right|$$
$$= O\Big(\frac{r^2 d_2}{\lambda_r^4} + \frac{r^2}{\sqrt{d_2}}\cdot\Big(\frac{d_2}{\lambda_r^2}\Big)^3 + r\Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^{k+1}\Big)$$

where $C_2 > 0$ is some absolute constant.

The aforementioned approximation errors hold whenever $C_2 d_2/\lambda_r^2 < 1$. Explicit formula for $B_\infty$ is also derived. An intriguing fact is that if $|d_1 - d_2| = O(\sqrt{d_2})$, i.e., the two dimensions of $M$ are comparable, then higher level approximations have similar effects as the Level-1 approximation. Simulation results show that Level-1 approximation by $B_1$ is indeed satisfactorily accurate when $d_1 = d_2$.

4. By replacing $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ with $B_k$, we prove the normal approximation of $\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$. Different levels of bias corrections require different levels of SNR conditions for the asymptotical normality. For instance, we prove the normal approximation of $\hat{\varepsilon}_2 := \big(\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] - B_{\lceil\log d_2\rceil}\big)/\big(\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}\big)$ with the $\lceil\log d_2\rceil$-th order bias correction. More exactly, we show the asymptotical normality of $\hat{\varepsilon}_2$ when $\sqrt{rd_2}/\lambda_r \to 0$ and $r^3/d_2 \to 0$ as $d_1, d_2 \to \infty$. As far as we know, this is the first result about the limiting distribution of singular subspaces which allows the rank $r$ to diverge. Meanwhile, no eigen-gap conditions (except SNR) are needed. Since our normal approximation is non-asymptotical, we impose no constraints on the relation between $d_1$ and $d_2$.

The rest of the paper is organized as follows. In Section 2, we derive the explicit representation formula of empirical spectral projector. The representation formula is established under deterministic perturbation. We prove normal approximation of $\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ in Section 3. Especially, we show that $\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ is asymptotically normal under optimal SNR conditions. In Section 4 and 5, we develop the arbitrarily $k$-th level approximations of $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ and its corresponding normal approximation, where requirements for SNR are specifically developed. In Section 6, we propose confidence regions and discuss about data-adaptive shrinkage estimator of singular values. We then display comprehensive simulation results in Section 7, where, for instance, we show the importance of higher order approximations of $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ when the matrix has unbalanced sizes and the effectiveness of shrinkage estimation of singular values. The proofs are collected in Section 8 and Appendix 8.

## 2. Representation formula of spectral projectors

Let $A$ and $X$ be $d \times d$ symmetric matrices. The matrix $A$ has rank $r = \text{rank}(A) \leq d$. Denote the eigen-decomposition of $A$,

$$A = \Theta \Lambda \Theta^{\mathsf{T}} = \sum_{j=1}^{r} \lambda_j \theta_j \theta_j^{\mathsf{T}}$$

where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r)$ contains the non-zero non-increasing eigenvalues of $A$. The $d \times r$ matrix $\Theta = (\theta_1, \cdots, \theta_r)$ consists of $A$'s eigenvectors. The noise matrix $X$ satisfies $\|X\| < \min_{1 \leq i \leq r} \frac{|\lambda_i|}{2}$ where $\|\cdot\|$ denotes the matrix operator norm. Given $\hat{A} = A + X$ where $A$ and $X$ are unknown, our goal is to estimate $\Theta$. We denote $\hat{\Theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_r)$ the $d \times r$ matrix containing the eigenvectors of $\hat{A}$ with largest $r$ eigenvalues in absolute values. Therefore, $\hat{\Theta}$ represents the empirical version of $\Theta$. We derive the representation formula of $\hat{\Theta}\hat{\Theta}^{\mathsf{T}}$ for deterministic $X$. The formula is 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 orthonormal. Define the spectral projector,

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

Also, define

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

Meanwhile, we write $\mathfrak{P}^{-k} = \Theta \Lambda^{-k} \Theta^{\mathsf{T}}$ for all $k \geq 1$. For notational simplicity, we denote $\mathfrak{P}^0 = \mathfrak{P}^\perp$ and denote the $k$-th order perturbation term

$$\mathcal{S}_{A,k}(X) = \sum_{\mathbf{s}: s_1 + \cdots + s_{k+1} = k} (-1)^{1+\tau(\mathbf{s})} \cdot \mathfrak{P}^{-s_1} X \mathfrak{P}^{-s_2} X \cdots X \mathfrak{P}^{-s_{k+1}} \qquad (3)$$

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

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

denotes the number of positive indices in $\mathbf{s}$. For instance, if $k = 1$, we have

$$\mathcal{S}_{A,1}(X) = \mathfrak{P}^{-1} X \mathfrak{P}^\perp + \mathfrak{P}^\perp X \mathfrak{P}^{-1}.$$

If $k = 2$, by considering $s_1 + s_2 + s_3 = 2$ for $s_1, s_2, s_3 \geq 0$ in (3), we have

$$\mathcal{S}_{A,2}(X) = \left( \mathfrak{P}^{-2} X \mathfrak{P}^\perp X \mathfrak{P}^\perp + \mathfrak{P}^\perp X \mathfrak{P}^{-2} X \mathfrak{P}^\perp + \mathfrak{P}^\perp X \mathfrak{P}^\perp X \mathfrak{P}^{-2} \right)$$
$$- \left( \mathfrak{P}^\perp X \mathfrak{P}^{-1} X \mathfrak{P}^{-1} + \mathfrak{P}^{-1} X \mathfrak{P}^\perp X \mathfrak{P}^{-1} + \mathfrak{P}^{-1} X \mathfrak{P}^{-1} X \mathfrak{P}^\perp \right).$$

**Theorem 1.** *If* $\|X\| < \min_{1 \le i \le r} \frac{|\lambda_i|}{2}$, *then*

$$\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}} = \sum_{k \ge 1} \mathcal{S}_{A,k}(X)$$

*where* $\mathcal{S}_{A,k}(X)$ *is defined in* (3) *and we set* $\mathfrak{P}^0 = \mathfrak{P}^\perp = \Theta_\perp \Theta_\perp^{\mathsf{T}}$ *for notational simplicity.*

Apparently, by eq. (3), a simple fact is

$$\|\mathcal{S}_{A,k}(X)\| \le \binom{2k}{k} \cdot \frac{\|X\|^k}{\lambda_r^k} \le \left(\frac{4\|X\|}{\lambda_r}\right)^k, \quad \forall \, k \ge 1.$$

Compared with the famous Wedin's and Davis-Kahan's first-order (w.r.t. $\|\mathbf{X}\|$) perturbation bound (Davis and Kahan, 1970; Wedin, 1972), Theorem 1 provides a precise formula for the empirical spectral projector. For instance, we can obtain the second-order approximation $\hat{\Theta}\hat{\Theta}^{\top} - \Theta\Theta^{\top} - \mathcal{S}_{A,1}(X)$ and even higher order approximations. The proof of Theorem 1 is based on complex analysis of the resolvent, a technique has been used in Koltchinskii and Lounici (2016); Xia (2019); Löffler et al. (2019). We note that our representation formula is similar, in spirit, to the perturbation series of the spectral projector for a single eigenvalue developed in Kato (2013). However, our formula is to investigate the spectral projector for all eigenvalues jointly, which has recently become more useful in low-rank methods.

## 3. Normal approximation of spectral projectors

Recall from (1) that $\hat{M} = M + Z \in \mathbb{R}^{d_1 \times d_2}$ with $M = U\Lambda V^{\mathsf{T}}$ where $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$ satisfying $U^{\mathsf{T}}U = I_r$ and $V^{\mathsf{T}}V = I_r$. The diagonal matrix $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_r)$ contains non-increasing positive singular values of $M$. Let $\hat{U}$ and $\hat{V}$ be $\hat{M}$'s top-$r$ left and right singular vectors. We derive the normal approximation of

$$\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] = \|\hat{U}\hat{U}^{\mathsf{T}} - UU^{\mathsf{T}}\|_{\mathrm{F}}^2 + \|\hat{V}\hat{V}^{\mathsf{T}} - VV^{\mathsf{T}}\|_{\mathrm{F}}^2,$$

which is often called the (squared) projection distance on Grassmannians. To this end, we clarify important notations which shall appear frequently throughout the paper.

To apply the representation formula from Theorem 1, we turn $\hat{M}, M$ and $Z$ into symmetric matrices. For notational consistency, we create $(d_1 + d_2) \times (d_1 + d_2)$ symmetric matrices as

$$\hat{A} = \begin{pmatrix} 0 & \hat{M} \\ \hat{M}^{\mathsf{T}} & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & M \\ M^{\mathsf{T}} & 0 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 0 & Z \\ Z^{\mathsf{T}} & 0 \end{pmatrix}.$$

The model (1) is thus translated into $\hat{A} = A + X$. The symmetric matrix $A$ has eigenvalues $\lambda_1 \ge \cdots \ge \lambda_r \ge \lambda_{-r} \ge \cdots \ge \lambda_{-1}$ where $\lambda_{-i} = -\lambda_i$ for $1 \le i \le r$.

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 $V$. Here, $\{\theta_i\}_{i=1}^r$ may not be uniquely defined if the singular value $\lambda_i$ has multiplicity larger than 1. However, the spectral projector $UU^\mathsf{T}$ and $VV^\mathsf{T}$ are unique regardless of the multiplicities of $M$'s singular values.

Following the same routine of notations, we denote

$$\Theta = (\theta_1, \cdots, \theta_r, \theta_{-r}, \cdots, \theta_{-1}) \in \mathbb{R}^{(d_1+d_2)\times 2r}$$

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,

$$\Theta\Theta^\mathsf{T} = \sum_{1\leq|j|\leq r} \theta_j\theta_j^\mathsf{T} = \begin{pmatrix} UU^\mathsf{T} & 0 \\ 0 & VV^\mathsf{T} \end{pmatrix}$$

and

$$\hat{\Theta}\hat{\Theta}^\mathsf{T} = \sum_{1\leq|j|\leq r} \hat{\theta}_j\hat{\theta}_j^\mathsf{T} = \begin{pmatrix} \hat{U}\hat{U}^\mathsf{T} & 0 \\ 0 & \hat{V}\hat{V}^\mathsf{T} \end{pmatrix}$$

where $\hat{U}$ and $\hat{V}$ represent $\hat{M}$'s top-$r$ left and right singular vectors. Similarly, for all $k \geq 1$, denote

$$\mathfrak{P}^{-k} = \sum_{1\leq|j|\leq r} \frac{1}{\lambda_j^k}\theta_j\theta_j^\mathsf{T} = \begin{cases} \begin{pmatrix} 0 & U\Lambda^{-k}V^\mathsf{T} \\ V\Lambda^{-k}U^\mathsf{T} & 0 \end{pmatrix} & \text{if } k \text{ is odd} \\ \begin{pmatrix} U\Lambda^{-k}U^\mathsf{T} & 0 \\ 0 & V\Lambda^{-k}V^\mathsf{T} \end{pmatrix} & \text{if } k \text{ is even.} \end{cases}$$

The orthogonal spectral projector is written as

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

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

$$\Theta_\perp = (\theta_{r+1}, \cdots, \theta_{d_1}, \theta_{-r-1}, \cdots, \theta_{-d_2})$$

, 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 $r+1 \leq j_2 \leq d_2$.

By the above notations, it is clear that

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

It suffices to prove the normal approximation of $\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2$. Observe that

$$\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 = 4r - 2\langle\Theta\Theta^{\mathsf{T}}, \hat{\Theta}\hat{\Theta}^{\mathsf{T}}\rangle = -2\langle\Theta\Theta^{\mathsf{T}}, \hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\rangle.$$

By Theorem 1 and $\Theta\Theta^{\mathsf{T}}\mathfrak{P}^{\perp} = 0$, we can write

$$\text{dist}^2[(\hat{U},\hat{V}),(U,V)] = -2\sum_{k\geq 2}\langle\Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X)\rangle$$

$$= 2\|\mathfrak{P}^{\perp}X\mathfrak{P}^{-1}\|_{\mathrm{F}}^2 - 2\sum_{k\geq 3}\langle\Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X)\rangle. \qquad (4)$$

where we used the fact $\mathfrak{P}^{\perp}\mathfrak{P}^{\perp} = \mathfrak{P}^{\perp}$ so that

$$-2\langle\Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2}\rangle = 2\langle\Theta\Theta^{\mathsf{T}}, \mathfrak{P}^{-1}X\mathfrak{P}^{\perp}X\mathfrak{P}^{-1}\rangle$$

$$= 2\text{tr}(\mathfrak{P}^{-1}X\mathfrak{P}^{\perp}X\mathfrak{P}^{-1}) = 2\|\mathfrak{P}^{\perp}X\mathfrak{P}^{-1}\|_{\mathrm{F}}^2.$$

We prove CLT of $\text{dist}^2[(\hat{U},\hat{V}),(U,V)]$ with an explicit normalizing factor. Without loss of generality, we assume $d_1 \leq d_2$ hereafter.

**Theorem 2.** *Suppose $d_2 \geq 3r$ where $d_2 = \max\{d_1, d_2\}$. There exist absolute constants $C_1, C_2 > 0$ such that if $\lambda_r \geq C_1\sqrt{d_2}$, then for any $s \geq 1$,*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\text{dist}^2[(\hat{U},\hat{V}),(U,V)] - \mathbb{E}\,\text{dist}^2[(\hat{U},\hat{V}),(U,V)]}{\sqrt{8d_{\star}}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x\right) - \Phi(x)\right|$$

$$\leq C_2 s^{1/2}\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\right)\cdot\frac{(rd_2)^{1/2}}{\lambda_r} + e^{-s} + C_2\left(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\right)^{3/2}\cdot\frac{1}{\sqrt{d_2}},$$

*where $d_{\star} = d_1 + d_2 - 2r$ and $\Phi(x)$ denotes the c.d.f. of standard normal distributions. By setting $s = \frac{\lambda_r}{\sqrt{rd_2}}$, we conclude that*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\text{dist}^2[(\hat{U},\hat{V}),(U,V)] - \mathbb{E}\,\text{dist}^2[(\hat{U},\hat{V}),(U,V)]}{\sqrt{8d_{\star}}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x\right) - \Phi(x)\right|$$

$$\leq C_2\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\right)\cdot\sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}} + C_2\left(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\right)^{3/2}\frac{1}{\sqrt{d_2}}.$$

By Theorem 2, the asymptotical normality holds as long as

$$\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\right)\cdot\sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}} \to 0 \quad\text{and}\quad \left(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\right)^{3/2}\cdot\frac{1}{\sqrt{d_2}} \to 0 \qquad (5)$$

as $d_1, d_2 \to \infty$. If $\sqrt{r} = O(\lambda_r^2\|\Lambda^{-2}\|_{\mathrm{F}})$, then the first condition in (5) is equivalent to $\frac{\sqrt{rd_2}}{\lambda_r} \to 0$. Such SNR condition is optimal in the consistency regime.

In addition, Cauchy-Schwartz inequality implies that $\|\Lambda^{-1}\|_F^4 \leq r \cdot \|\Lambda^{-2}\|_F^2$. Therefore, the second condition in (5) holds when

$$\frac{r^3}{d_2} \to 0 \quad \text{as } d_1, d_2 \to \infty.$$

Therefore, $r$ is allowed to grow as fast as $o\big((d_1 + d_2)^{1/3}\big)$.

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

$$\operatorname{Var}\big(2\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_F^2\big) = 8d_\star\|\Lambda^{-2}\|_F^2.$$

*Clearly, this conclusion relies on the Gaussian assumption. If the entries of $Z$ are not Gaussian, this variance should involve the kurtosis of the unknown distribution. The unknown kurtosis makes the data-driven statistical inference even more challenging. Finally, we remark by the proof of Theorem 2 that no constraints between $d_1$ and $d_2$ are needed.*

Note that $\mathbb{E}\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ in Theorem 2 is not transparent yet. Calculating $\mathbb{E}\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ needs delicate analysis. If we approximate this expectation by its leading term $2\mathbb{E}\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_F^2$, we obtain

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

The primary subject of section 4 is to approximate $\mathbb{E}\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ to a higher accuracy.

## 4. Approximating the bias

Recall (4), we have

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

where we used the fact $\mathbb{E}\,\mathcal{S}_{A,2k+1}(X) = 0$ for any positive integer $k \geq 1$. We aim to determine $\mathbb{E}\|\mathfrak{P}^\perp X\mathfrak{P}^{-1}\|_F^2$ and $\mathbb{E}\langle\Theta\Theta^\mathsf{T}, \mathcal{S}_{A,2k}(X)\rangle$ for all $k \geq 2$. Apparently, by obtaining explicit formulas of $\mathbb{E}\langle\Theta\Theta^\mathsf{T}, \mathcal{S}_{A,2k}(X)\rangle$ for larger $k$s, we end up with more precise approximation of $\mathbb{E}\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. In Lemma 1-3, we provide arbitrarily $k$-th order approximation of the bias.

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

$$\mathbb{E}\|\mathfrak{P}^\perp X\mathfrak{P}^\perp\|_F^2 = d_\star\|\Lambda^{-1}\|_F^2$$

*where $d_\star = d_1 + d_2 - 2r$. Moreover, if $\lambda_r \geq C_1\sqrt{d_2}$ for some large enough constant $C_1 > 0$, then*

$$\left|\mathbb{E}\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star\|\Lambda^{-1}\|_F^2\right| \leq C_2 r\left(\frac{d_2}{\lambda_r^2}\right)^2$$

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

In Lemma 2, we calculate $\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,4}(X)\rangle$. It yields the second order approximation of $\mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$.

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

$$\left|\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,4}(X)\rangle - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2\right| \leq C_2 \frac{r^2 d_2}{\lambda_r^4}$$

*where $d_\star = d_1 + d_2 - 2r$ and $\Delta_d = d_1 - d_2$ and $C_2$ is an absolute constant. Moreover, if $\lambda_r \geq C_1\sqrt{d_2}$ for some large enough constant $C_1 > 0$, then*

$$\left|\mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2\big(d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2\big)\right|$$

$$\leq C_1 \frac{r^2 d_2}{\lambda_r^4} + C_2 r \Big(\frac{d_2}{\lambda_r^2}\Big)^3$$

*where $C_2, C_3 > 0$ are absolute constants (depending on $C_1$).*

In general, we calculate the arbitrary $k$-th order approximation in Lemma 3. Recall that $d_{1-} = d_1 - r$ and $d_{2-} = d_2 - r$.

**Lemma 3** (Arbitrary $k$-th order approximation)**.** *For a positive integer $k \geq 2$ and $\sqrt{d_2} \geq \log^2 d_2$ and $e^{-c_1 d_2} \leq \frac{1}{\sqrt{d_2}}$, the following fact holds*

$$\left|\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X)\rangle - (-1)^k(d_{1-}^{k-1} - d_{2-}^{k-1})(d_{1-} - d_{2-})\|\Lambda^{-k}\|_{\mathrm{F}}^2\right|$$

$$\leq \frac{C_1(r^2 + k)}{\sqrt{d_2}} \cdot \Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k$$

*where $c_1, C_1, C_2 > 0$ are some absolute constants. Then, the following bound holds*

$$\left|\mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_k\right|$$

$$\leq C_4 \frac{r^2 d_2}{\lambda_r^4} + \frac{C_5 r^2}{\sqrt{d_2}} \cdot \Big(\frac{d_2}{\lambda_r^2}\Big)^3 + C_6 r \Big(\frac{C_3 d_2}{\lambda_r^2}\Big)^{k+1}$$

*where $C_3, C_4, C_5, C_6$ are some absolute constants and $B_k$ is defined by*

$$B_k = 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - 2\sum_{k_0=2}^{k}(-1)^{k_0}(d_{1-}^{k_0-1} - d_{2-}^{k_0-1})(d_{1-} - d_{2-})\|\Lambda^{-k_0}\|_{\mathrm{F}}^2. \quad (6)$$

The second and higher order terms involve the dimension difference $\Delta_d = d_1 - d_2$. If $d_1 = d_2$, these higher order approximations essentially have similar effects as the first order approximation.

**Remark 2.** *By choosing $k = \lceil \log d_2 \rceil$ so that $(C_3 d_2/\lambda_r^2)^{k+1} \lesssim (d_2/\lambda_r^2)^3/\sqrt{d_2}$, we get*

$$\left|\mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_{\lceil \log d_2 \rceil}\right| \leq C_4 \frac{r^2 d_2}{\lambda_r^4} + C_5 \frac{r^2}{\sqrt{d_2}} \cdot \Big(\frac{d_2}{\lambda_r^2}\Big)^3$$

*for some absolute constants $C_4, C_5 > 0$. In addition, for each $1 \leq j \leq r$, we*

*have*

$$2d_{1-}\lambda_j^{-2} - 2\sum_{k=2}^{\infty}(-1)^k(d_{1-} - d_{2-})d_{1-}^{k-1}\lambda_j^{-2k} = \frac{2d_{1-}(\lambda_j^2 + d_{2-})}{\lambda_j^2(\lambda_j^2 + d_{1-})}$$

*which matches* $\mathbb{E}\|\hat{u}_j\hat{u}_j^\mathsf{T} - u_ju_j^\mathsf{T}\|_{\mathrm{F}}^2$ *developed in (Bao et al., 2018, Theorem 2.9) if* $\min\{\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j\}$ *is bounded away from 0 and r is fixed. Similarly, we have*

$$2d_{2-}\lambda_j^{-2} - 2\sum_{k=2}^{\infty}(-1)^k(d_{2-} - d_{1-})d_{2-}^{k-1}\lambda_j^{-2k} = \frac{2d_{2-}(\lambda_j^2 + d_{1-})}{\lambda_j^2(\lambda_j^2 + d_{2-})}$$

*which matches* $\mathbb{E}\|\hat{v}_j\hat{v}_j^\mathsf{T} - v_jv_j^\mathsf{T}\|_{\mathrm{F}}^2$ *developed in (Bao et al., 2018, Theorem 2.3). Compared with Bao et al. (2018), our results are non-asymptotical. We impose no eigen-gap conditions and no upper bounds on r.*

**Remark 3.** *The proof of Lemma 3 imply that if* $\lambda_r \geq C_1\sqrt{d_2}$, *then*

$$\mathbb{E}\|\hat{U}\hat{U}^\mathsf{T} - UU^\mathsf{T}\|_{\mathrm{F}}^2 = 2\sum_{j=1}^{r}\frac{d_{1-}(\lambda_j^2 + d_{2-})}{\lambda_j^2(\lambda_j^2 + d_{1-})} + O\Big(\frac{r^2d_2}{\lambda_r^4} + \frac{r^2}{\sqrt{d_2}}\cdot\frac{d_2^3}{\lambda_r^6}\Big)$$

*and*

$$\mathbb{E}\|\hat{V}\hat{V}^\mathsf{T} - VV^\mathsf{T}\|_{\mathrm{F}}^2 = 2\sum_{j=1}^{r}\frac{d_{2-}(\lambda_j^2 + d_{1-})}{\lambda_j^2(\lambda_j^2 + d_{2-})} + O\Big(\frac{r^2d_2}{\lambda_r^4} + \frac{r^2}{\sqrt{d_2}}\cdot\frac{d_2^3}{\lambda_r^6}\Big).$$

## 5. Normal approximation after bias corrections

In this section, we prove the normal approximation of $\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ with explicit centering and normalizing terms. By Theorem 2, it suffices to substitute $\mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ with the explicit formulas from Lemma 1-3.

Similarly as in Section 4, we consider arbitrarily $k$-th levels of bias corrections for $\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Higher order bias corrections, while involving 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 $\lambda_r \gg \sqrt{rd_2^{3/2}}$ for asymptotical normality, while the $\lceil\log d_2\rceil$-th order bias correction in Theorem 4 only requires optimal $\lambda_r \gg \sqrt{rd_2}$ for asymptotical normality. Again, the rank $r$ is allowed to diverge as fast as $o\big((d_1 + d_2)^{1/3}\big)$.

**Theorem 3** (First order CLT). *Suppose* $d_2 \geq 3r$. *There exist absolute constants* $C_1, C_2, C_3 > 0$ *such that if* $\lambda_r \geq C_1\sqrt{d_2}$, *then,*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_1}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x\right) - \Phi(x)\right|$$

$$\leq C_2\Big(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\Big)\cdot\sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}} + C_2\Big(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\Big)^{3/2}\cdot\frac{1}{\sqrt{d_2}} + C_3\frac{rd_2^{3/2}}{\lambda_r^2},$$

*where* $d_\star = d_1 + d_2 - 2r$ *and* $B_1$ *is defined by (6).*

By Theorem 3, we conclude that

$$\frac{\operatorname{dist}^2[(\hat{U},\hat{V}),(U,V)] - 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

as $d_1, d_2 \to \infty$ if $\sqrt{r} = O(\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2)$ and

$$\frac{\sqrt{rd_2} + \sqrt{rd_2^{3/2}}}{\lambda_r} \to 0 \quad \text{and} \quad \Big(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\Big)^{3/2} \cdot \frac{1}{\sqrt{d_2}} \to 0.$$

The above conditions require $\lambda_r \gg \sqrt{rd_2^{3/2}}$ and $r^3 \ll d_2$. The order $d_2^{3/4}$ is larger than the optimal rate $\sqrt{d_2}$. It is improvable if we apply higher order bias corrections.

**Theorem 4** (Arbitrary $k$-th order CLT). *Suppose that $d_2 \geq 3r$ and $k \geq 3$. There exist absolute constants $C_0, C_1, C_2, C_3 > 0$ such that if $\lambda_r \geq C_1\sqrt{d_2}$, then,*

$$\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{\operatorname{dist}^2[(\hat{U},\hat{V}),(U,V)] - B_k}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x\Big) - \Phi(x)\Big| \leq C_2\Big(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\Big)\cdot\sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}}$$

$$+ C_0\frac{r^2\sqrt{d_2}}{\lambda_r^2} + C_2\Big(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\Big)^{3/2}\cdot\frac{1}{\sqrt{d_2}} + C_1\frac{r^2d_2^2}{\lambda_r^4} + C_2 r\sqrt{d_2}\cdot\Big(\frac{C_3 d_2}{\lambda_r^2}\Big)^k,$$

*where $B_k$ is defined by (6).*

Tthe asymptotical normality of $\big(\operatorname{dist}^2[(\hat{U},\hat{V}),(U,V)] - B_k\big)/\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}$ requires

$$\frac{\sqrt{rd_2} + rd_2^{1/4} + \sqrt{d_2}\cdot(r^2d_2)^{1/4k}}{\lambda_r} \to 0$$

as $d_1, d_2 \to \infty$ when $\sqrt{r} = O(\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2)$. By choosing $k = \lceil\log d_2\rceil$, it boils down to $\sqrt{rd_2}/\lambda_r \to 0$ which is optimal in the consistency regime. Similarly as in Theorem 2, the condition $(\|\Lambda^{-1}\|_{\mathrm{F}}^4/\|\Lambda^{-2}\|_{\mathrm{F}}^2)^{3/2}/\sqrt{d_2} \to 0$ requires that $r^3/d_2 \to 0$ as $d_1, d_2 \to \infty$.

**Remark 4.** *To avoid computing the sum of $k$ terms in $B_k$ (6), it suffices to apply $B_\infty$ which by Remark 2 is*

$$B_\infty = 2\sum_{j=1}^{r}\frac{1}{\lambda_j^2}\Big(d_{1-}\cdot\frac{\lambda_j^2 + d_{2-}}{\lambda_j^2 + d_{1-}} + d_{2-}\cdot\frac{\lambda_j^2 + d_{1-}}{\lambda_j^2 + d_{2-}}\Big).$$

*By setting $k = \infty$ in Theorem 4, we obtain*

$$\frac{\operatorname{dist}^2[(\hat{U},\hat{V}),(U,V)] - B_\infty}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}} \to \mathcal{N}(0,1)$$

*as long as $\sqrt{rd_2}/\lambda_r \to 0$ and $r^3/d_2 \to 0$ when $d_1, d_2 \to \infty$.*

We note that it is also possible to develop the asymptotic distribution for the one-sided singular vectors $\|\hat{U}\hat{U}^\top - UU^\top\|_F^2$ and $\|\hat{V}\hat{V}^\top - VV^\top\|_F^2$. For that purpose, some treatments should be adjusted and the proof has to be modified accordingly. We leave it as a future work.

## 6. Data-driven Confidence regions of singular subspaces

By the normal approximation of $\text{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ in Theorem 4, we construct confidence regions of $U$ and $V$. The confidence regions of $(U, V)$ attain the pre-determined confidence level asymptotically. In the asymptotic scheme, we shall consider $d_1, d_2 \to \infty$. Therefore, the parameters $r^{(d_1, d_2)}$, $U^{(d_1, d_2)}$, $V^{(d_1, d_2)}$ and $\Lambda^{(d_1, d_2)}$ also depend on $d_1, d_2$. For notational simplicity, we omit the superscripts $(d_1, d_2)$ without causing confusions.

In particular, we set $k = \lceil \log d_2 \rceil$ in Theorem 4 and get

$$\frac{\text{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_{\lceil \log d_2 \rceil}}{\sqrt{8d_\star}\|\Lambda^{-2}\|_F} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

as $d_1, d_2 \to +\infty$ when $\sqrt{r} = O(\lambda_r^2 \|\Lambda^{-2}\|_F)$ and

$$\lim_{d_1, d_2 \to \infty} \max\left\{ \frac{\sqrt{rd_2} + rd_2^{1/4}}{\lambda_r} + \left( \frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_2}} \right\} = 0. \tag{7}$$

We define the confidence region based on $(\hat{U}, \hat{V})$ by

$$\mathcal{M}_\alpha(\hat{U}, \hat{V}) := \left\{ (L, R) : L \in \mathbb{R}^{d_1 \times r}, R \in \mathbb{R}^{d_2 \times r}, L^\mathsf{T} L = R^\mathsf{T} R = I_r \right.$$

$$\left. , \left| \text{dist}^2[(L, R), (\hat{U}, \hat{V})] - B_{\lceil \log d_2 \rceil} \right| \leq \sqrt{8d_\star} z_{\alpha/2} \|\Lambda^{-2}\|_F \right\}$$

where $z_\alpha$ denotes the critical value of standard normal distribution, i.e., $z_\alpha = \Phi^{-1}(1 - \alpha)$. Theorem 5 follows immediately from Theorem 4.

**Theorem 5.** *Suppose that conditions in Theorem 4 hold. Then, for any $\alpha \in (0, 1)$, we get*

$$\left| \mathbb{P}\big((U, V) \in \mathcal{M}_\alpha(\hat{U}, \hat{V})\big) - (1 - \alpha) \right| \leq C_1 \frac{\sqrt{r}}{\lambda_r^2 \|\Lambda^{-2}\|_F} \cdot \sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}}$$

$$+ C_2 \left( \frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_2}} + C_3 \frac{r^2 \sqrt{d_2}}{\lambda_r^2} + C_4 \frac{r^2 d_2^2}{\lambda_r^4}$$

*for some absolute constants $C_1, C_2, C_3, C_4 > 0$. If condition (7) holds, then*

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

**Remark 5.** *We can also simply replace $B_{\lceil \log d_2 \rceil}$ with $B_\infty$ and Theorem 5 still holds under the same conditions.*

Note that $\Lambda$ is usually unknown. An immediate choice is the empirical singular values $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$, i.e., top-$r$ singular values of $\hat{M}$. It is well known that $\{\hat{\lambda}_j\}_{j=1}^r$ are biased estimators of $\{\lambda_j\}_{j=1}^r$. See (Benaych-Georges and Nadakuditi, 2012) and (Ding, 2017) for more details. The results of Section 3.1 in (Benaych-Georges and Nadakuditi, 2012) show, under the condition $d_1/d_2$ converges to $\alpha \in (0, 1]$, that

$$\frac{\hat{\lambda}_j^2}{d_2} \xrightarrow{\text{a.s.}} \frac{(1 + \lambda_j^2/d_2)(d_1/d_2 + \lambda_j^2/d_2)}{\lambda_j^2/d_2}, \quad \text{as } d_2 \to \infty \tag{8}$$

for all $j = 1, \cdots, r$. The square root of RHS of (8) is called the class location of the empirical singular value. Bound (8) inspires the following shrinkage estimator of $\lambda_j^2$:

$$\tilde{\lambda}_j^2 = \frac{\hat{\lambda}_j^2 - (d_1 + d_2)}{2} + \frac{\sqrt{(\hat{\lambda}_j^2 - (d_1 + d_2))^2 - 4d_1 d_2}}{2} \quad \text{for all } 1 \leq j \leq r. \tag{9}$$

By replacing $\Lambda$ with data-dependent estimates $\tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_r)$, we define

$$\tilde{B}_k = 2d_* \|\tilde{\Lambda}^{-1}\|_{\mathrm{F}}^2 - 2 \sum_{k_0=2}^k (-1)^{k_0} (d_{1-}^{k_0-1} - d_{2-}^{k_0-1})(d_{1-} - d_{2-}) \|\tilde{\Lambda}^{-k_0}\|_{\mathrm{F}}^2$$

and similarly

$$\tilde{B}_\infty = 2 \sum_{j=1}^r \frac{1}{\lambda_j^2} \left( d_{1-} \cdot \frac{\lambda_j^2 + d_{2-}}{\lambda_j^2 + d_{1-}} + d_{2-} \cdot \frac{\lambda_j^2 + d_{1-}}{\lambda_j^2 + d_{2-}} \right).$$

To this end, we define the data-driven confidence region based on $(\hat{U}, \hat{V})$ by

$$\widehat{\mathcal{M}}_\alpha(\hat{U}, \hat{V}) := \Big\{ (L, R) : L \in \mathbb{R}^{d_1 \times r}, R \in \mathbb{R}^{d_2 \times r}, L^{\mathsf{T}} L = R^{\mathsf{T}} R = I_r$$

$$, \big| \mathrm{dist}^2[(L, R), (\hat{U}, \hat{V})] - \tilde{B}_{\lceil \log d_2 \rceil} \big| \leq \sqrt{8d_\star} z_{\alpha/2} \|\tilde{\Lambda}^{-2}\|_{\mathrm{F}} \Big\}.$$

**Theorem 6.** *Suppose that conditions in Theorem 4 and (7) hold, and $r = O(1)$. Then, for any $\alpha \in (0, 1)$, we have*

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

By Theorem 6, the data-driven confidence region $\widehat{\mathcal{M}}_\alpha(\hat{U}, \hat{V})$ is a valid confidence region asymptotically. For simplicity, we only consider the case of fixed ranks. We remark that Theorem 6 still holds if we replace $\tilde{B}_{\lceil \log d_2 \rceil}$ with $\tilde{B}_\infty$.

## 7. 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.

### 7.1. Higher order approximations of bias and normal approximation

In *Simulation* 1, we show the effectiveness of approximating $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ by the first order approximation $2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2$ where $d_\star = d_1 + d_2 - 2r$. Meanwhile, we show the inefficiency of first order approximation when $|d_1 - d_2| \gtrsim \min(d_1, d_2)$. In *Simulation* 2, we demonstrate the benefits of higher order approximations when $|d_1 - d_2| \gtrsim \min(d_1, d_2)$.

*Simulation* 1. In this simulation, we study the accuracy of first order approximation and its relevance with $\Delta_d = d_1 - d_2$. First, we set $d_1 = d_2 = d$ where $d = 100, 200, 300$. The signal strength $\lambda$ is chosen as $30, 30.5, \cdots, 40$. For each given $\lambda$, the first order approximation $2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2$ is recorded. To obtain $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$, we repeat the experiments for 500 times for each $\lambda$ and the average of $\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ is recorded, which denotes the simulated value of $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. We compare the simulated $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ with $2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2$, which is displayed in Figure 1(a). Since $d_1 = d_2 = d$, the first order approximation has similar effect as higher order approximation which is verified by 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 and the results are displayed in Figure 1(b), which clearly shows that first order approximation is insufficient to estimate $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Therefore, if $|d_1 - d_2| \gg 0$, we need higher order approximation of $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$.

*Simulation* 2. In this simulation, we study the effects of higher order approximations 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 $\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ whose average is recorded as the simulated $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. Meanwhile, for each $\lambda$, we record the 1st-4th order approximations $B_1, B_2, B_3$ and $B_4$ which are defined by (6). All the results are displayed in Figure 2. It verifies that higher order bias corrections indeed improve the accuracy of approximating $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. It also shows that the 1st and 3rd order approximations over-estimate $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$; while, the 2nd and 4th order approximations under-estimate $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$.

*Simulation* 3. We apply higher order approximations and show the normal approximation of $\left(\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_k\right) / \sqrt{8d_\star} \|\Lambda^{-2}\|_{\mathrm{F}}$ when $d_1 = 100, d_2 = 600$ and rank $r = 6$. We fixed the signal strength $\lambda = 50$. The density histogram is based on 5000 realizations from independent experiments. We consider 1st-4th order approximations, denoted by $\{B_k\}_{k=1}^4$. More specifically,

$$B_1 = 2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2, \quad \text{and} \quad B_2 = 2(d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2 \|\Lambda^{-2}\|_{\mathrm{F}}^2)$$
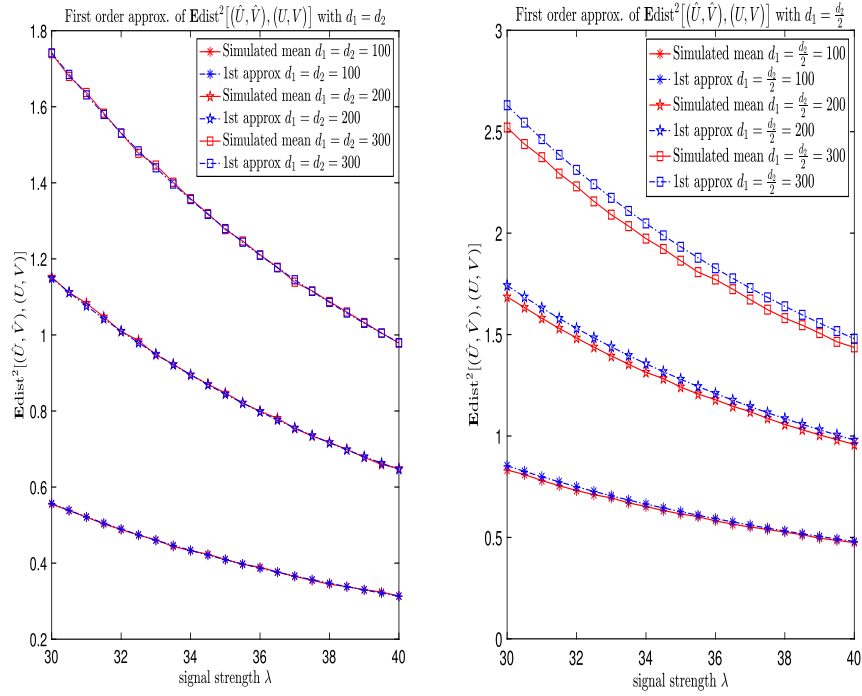
(a) First order approximation $2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2$ is accurate when $\Delta_d = d_1 - d_2 = 0$ and rank $r = 6$. Here $d_\star = d_1 + d_2 - 2r$. There is no need for higher order approximations.

(b) First order approximation $2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2$ is not sufficiently accurate when $|d_1 - d_2| \gg 0$. Here $d_\star = d_1 + d_2 - 2r$ and rank $r = 6$. The higher order approximations are indeed necessary.

FIG 1. *Comparison between* $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ *and the first order approximation:* $2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2$. *It verifies that the accuracy of first order approximation depends on the dimension difference* $\Delta_d = d_1 - d_2$. *Here the red curves represent the simulated mean* $\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$ *based on* 500 *realizations of* $\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. *The blue curves are the theoretical first order approximations* $2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2$ *based on Lemma* 1. *The above left figure clearly shows that first order approximation is accurate if* $d_1 = d_2$.

and

$$B_3 = 2\big(d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2 + d_\star\Delta_d^2\|\Lambda^{-3}\|_{\mathrm{F}}^2\big)$$

and

$$B_4 = 2\big(d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2 + d_\star\Delta_d^2\|\Lambda^{-3}\|_{\mathrm{F}}^2 - (d_{1-}^3 - d_{2-}^3)\Delta_d\|\Lambda^{-4}\|_{\mathrm{F}}^2\big).$$

The results are shown in Figure 3. This experiment aims to demonstrate the necessity of higher order bias corrections. Indeed, by the density histograms in Figure 3, 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
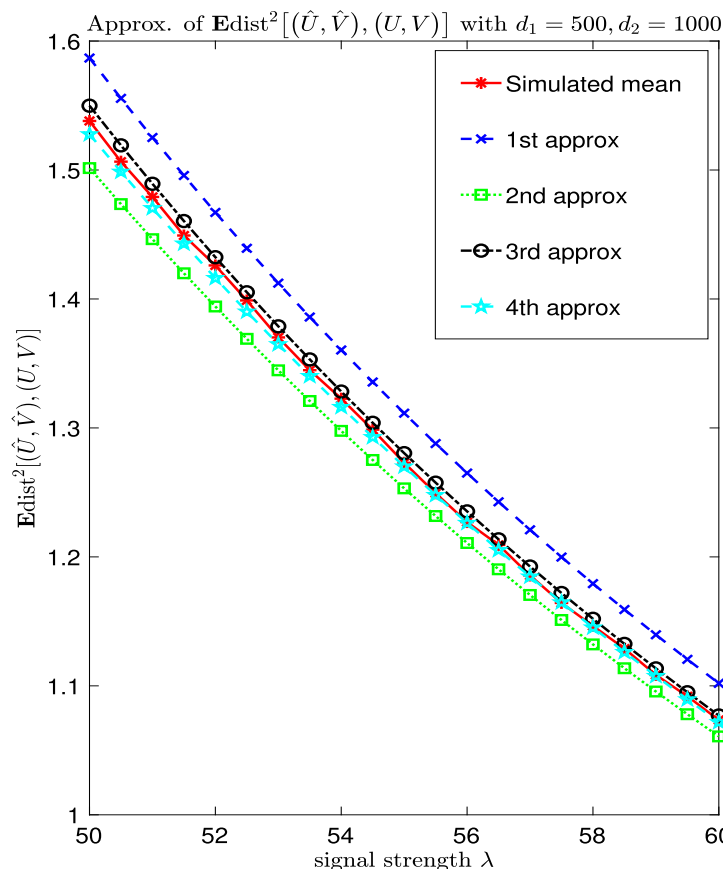
Approx. of $\mathbf{E}\mathrm{dist}^2\big[(\hat{U},\hat{V}),(U,V)\big]$ with $d_1 = 500, d_2 = 1000$



FIG 2. *The higher order approximations of $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$. The simulated mean represents $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ calculated by the average of 500 realizations of $\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$. The 1st order approximation is $2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2$; 2nd order approximation is $2(d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2)$, 3rd order approximation is $2(d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2 + d_\star\Delta_d^2\|\Lambda^{-3}\|_{\mathrm{F}}^2)$ and 4th order approximation is $2(d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2 + d_\star\Delta_d^2\|\Lambda^{-3}\|_{\mathrm{F}}^2 - (d_\star^2 - d_{1-}d_{2-})\Delta_d^2\|\Lambda^{-4}\|_{\mathrm{F}}^2)$ where $\Delta_d = d_1 - d_2$, $d_{1-} = d_1 - r$, $d_{2-} = d_2 - r$ and $d_\star = d_{1-} + d_{2-}$ with $r = 6$. Clearly, the 3rd and 4th order approximations are already close to the simulated mean. We observe that the 1st and 3rd order approximations over-estimate $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$; while, the 2nd and 4th order approximations under-estimate $\mathbb{E}\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$.*

curve. On the other hand, after third or fourth order corrections, the normal approximation is very satisfactory at the same level of signal strength $\lambda = 50$.

### 7.2. Normal approximation with data-dependent bias corrections

Next, we show normal approximations of $\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]$ with data-dependent bias corrections and normalization factors.
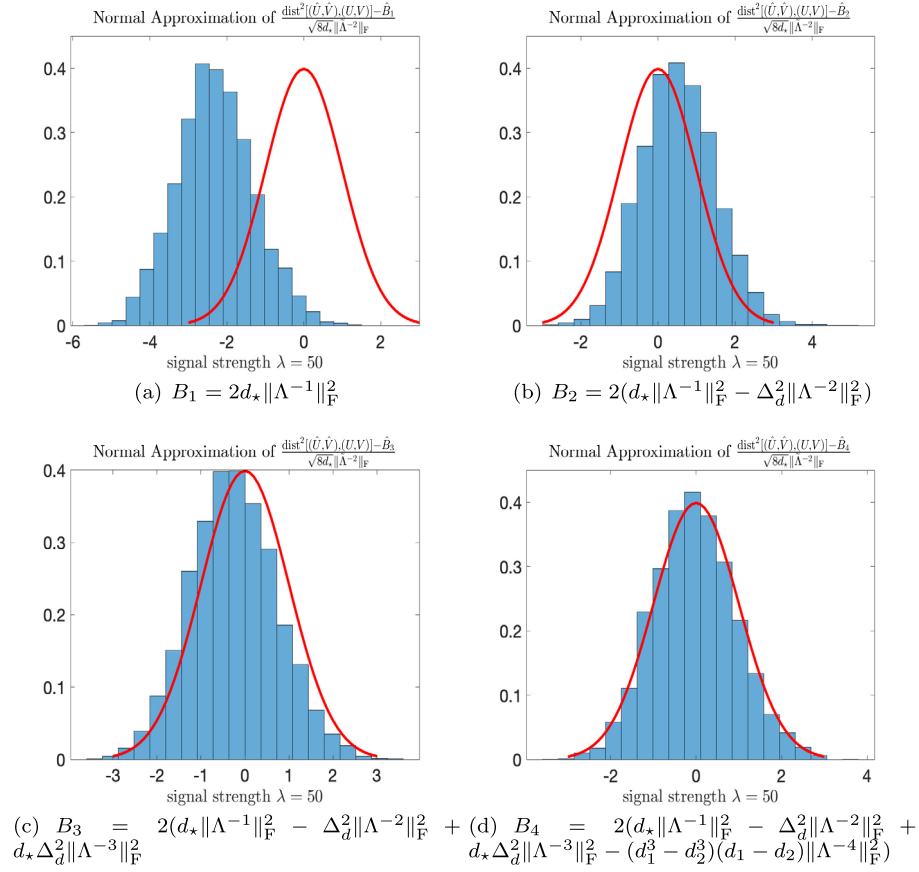
(a) $B_1 = 2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2$

(b) $B_2 = 2(d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2 \|\Lambda^{-2}\|_{\mathrm{F}}^2)$

(c) $B_3 = 2(d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2 \|\Lambda^{-2}\|_{\mathrm{F}}^2 + d_\star \Delta_d^2 \|\Lambda^{-3}\|_{\mathrm{F}}^2$

(d) $B_4 = 2(d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2 \|\Lambda^{-2}\|_{\mathrm{F}}^2 + d_\star \Delta_d^2 \|\Lambda^{-3}\|_{\mathrm{F}}^2 - (d_1^3 - d_2^3)(d_1 - d_2)\|\Lambda^{-4}\|_{\mathrm{F}}^2)$

FIG 3. *Normal approximation of $\frac{\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] - B_k}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}}$ 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 red curve presents p.d.f. of standard normal distributions. Since $|d_1 - d_2| \gg 0$, this experiment demonstrates the necessity of higher order bias corrections. The bias correction $\hat{B}_k$ can be 1st -4th order bias corrections.*

*Simulation* 4. We apply the 1st order approximation and show normal approximation of $\left(\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] - 2d_\star\|\hat{\Lambda}^{-1}\|_{\mathrm{F}}^2\right)/\sqrt{8d_\star}\|\hat{\Lambda}^{-2}\|_{\mathrm{F}}$ when $d_1 = d_2 = 100$ and $r = 6$. Here, $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ denotes the top-$r$ empirical singular values of $\hat{M}$. The signal strength $\lambda = 25, 50, 65, 75$. For each $\lambda$, we record $\left(\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] - 2d_\star\|\hat{\Lambda}^{-1}\|_{\mathrm{F}}^2\right)/\sqrt{8d_\star}\|\hat{\Lambda}^{-2}\|_{\mathrm{F}}$ from 5000 thousand independent experiments and draw the density histogram. The p.d.f. of standard normal distribution is displayed by the red curve. The results are shown in Figure 4. Since each $\hat{\lambda}_j$ over-estimates the true $\lambda_j$, the bias correction $2d_\star\|\hat{\Lambda}^{-1}\|_{\mathrm{F}}^2$ is not sufficiently significant. It explains why the density histograms shift rightward compared with the standard normal curve, especially when signal strength $\lambda$ is moderately strong.
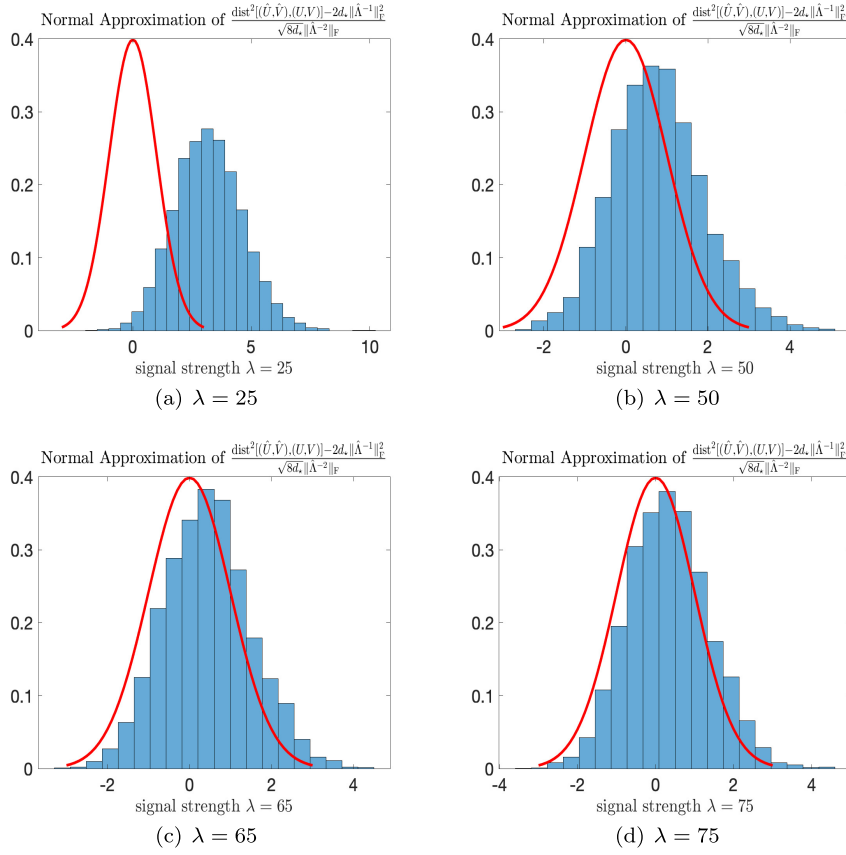
FIG 4. *Normal approximation of* $\frac{\text{dist}^2[(\hat{U},\hat{V}),(U,V)]-2d_\star\|\hat{\Lambda}^{-1}\|_F^2}{\sqrt{8d_\star}\|\hat{\Lambda}^{-2}\|_F}$ *with* $d_1 = d_2 = 100$ *and* $r = 6$. *The density histogram is based on* 5000 *realizations from independent experiments. The empirical singular values* $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_r)$ *are calculated from* $\hat{M}$. *The red curve presents p.d.f. of standard normal distributions. Since* $\hat{\lambda}_j$ *over-estimates* $\lambda_j$, *it explains why the density histogram shifts to the right compared with the standard normal curve, especially when signal strength* $\lambda$ *is not significantly strong.*

*Simulation* 5. We apply the 1st order approximation and show normal approximation of $\left(\text{dist}^2[(\hat{U},\hat{V}),(U,V)] - 2d_\star\|\tilde{\Lambda}^{-1}\|_F^2\right)/\sqrt{8d_\star}\|\tilde{\Lambda}^{-2}\|_F$ when $d_1 = d_2 = 100$ and $r = 6$. Here, $\tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_r)$ denotes the top-$r$ shrinkage estimators of $\lambda_j$s as in (9). The signal strength $\lambda = 25, 50, 65, 75$. For each $\lambda$, we record $\left(\text{dist}^2[(\hat{U},\hat{V}),(U,V)]-2d_\star\|\tilde{\Lambda}^{-1}\|_F^2\right)/\sqrt{8d_\star}\|\tilde{\Lambda}^{-2}\|_F$ from 5000 thousand independent experiments and draw the density histogram. The results are shown in Figure 5. In comparison with *Simulation* 4 and Figure 4, we conclude that $2d_\star\|\tilde{\Lambda}^{-1}\|_F^2$ works better than $2d_\star\|\hat{\Lambda}^{-1}\|_F^2$ for bias corrections. Indeed, normal approximation of $\left(\text{dist}^2[(\hat{U},\hat{V}),(U,V)] - 2d_\star\|\tilde{\Lambda}^{-1}\|_F^2\right)/\sqrt{8d_\star}\|\tilde{\Lambda}^{-2}\|_F$ is already satisfactory when signal strength $\lambda = 35$, compared with $\lambda \geq 75$ when $\hat{\Lambda}$ is used.
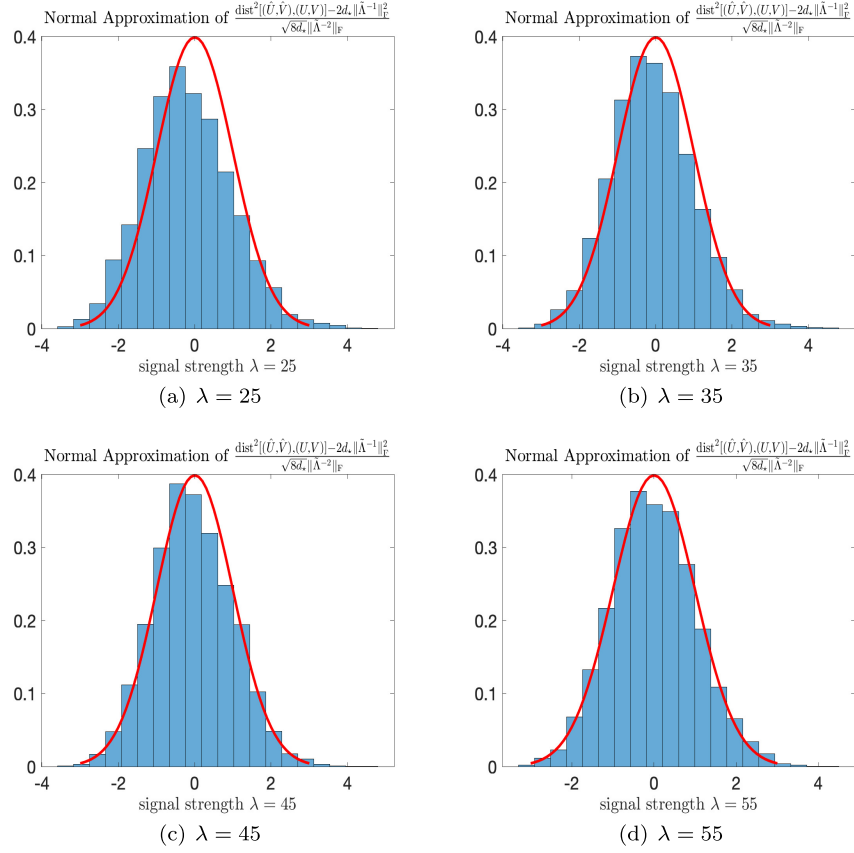
(a) $\lambda = 25$



(b) $\lambda = 35$



(c) $\lambda = 45$



(d) $\lambda = 55$

FIG 5. *Normal approximation of* $\frac{\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]-2d_\star\|\tilde{\Lambda}^{-1}\|_{\mathrm{F}}^2}{\sqrt{8d_\star}\|\tilde{\Lambda}^{-2}\|_{\mathrm{F}}}$ *with* $d_1 = d_2 = 100$ *and* $r = 6$. *The density histogram is based on* $5000$ *realizations from independent experiments. The shrinkage estimators* $\tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_r)$ *are calculated as eq.* (9). *The red curve presents p.d.f. of standard normal distributions. Since* $d_1 = d_2$, *we apply first order bias corrections to* $\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]$. *In comparison with Simulation* 4 *and Figure* 4 *where* $\hat{\Lambda}$ *is used instead of* $\tilde{\Lambda}$, *we conclude that* $2d_\star\|\tilde{\Lambda}^{-1}\|_{\mathrm{F}}^2$ *is more accurate than* $2d_\star\|\hat{\Lambda}^{-1}\|_{\mathrm{F}}^2$ *for bias corrections. Indeed, we see that normal approximation of* $\frac{\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)]-2d_\star\|\tilde{\Lambda}^{-1}\|_{\mathrm{F}}^2}{\sqrt{8d_\star}\|\tilde{\Lambda}^{-2}\|_{\mathrm{F}}}$ *is already satisfactory when signal strength* $\lambda = 35$.

# 8. Proofs

We only provide the proof of Theorem 1 in this section. Proofs of other theorems are collected in the supplementary file.

## 8.1. Proof of Theorem 1

For notational simplicity., we assume $\lambda_i > 0$ for $1 \le i \le r$, i.e., the matrix $A$ is positively semidefinite. The proof is almost identical if $A$ has negative

eigenvalues. Indeed, if there exist negative eigenvalues, we should also construct a contour plot which includes those negative eigenvalues.

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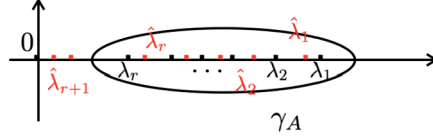


FIG 6. *The contour plot $\gamma_A$ which includes $\{\hat{\lambda}_i, \lambda_i\}_{i=1}^r$ leaving out 0 and $\{\hat{\lambda}_i\}_{i=r+1}^d$.*

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

Weyl's lemma implies that $\max_{1 \le i \le r} |\hat{\lambda}_i - \lambda_i| \le \|X\|$. 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

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

$$= \sum_{i=1}^r \hat{\theta}_i \hat{\theta}_i^{\mathsf{T}} = \hat{\Theta} \hat{\Theta}^{\mathsf{T}}.$$

As a result, we have

$$\hat{\Theta} \hat{\Theta}^{\mathsf{T}} = \frac{1}{2\pi i} \oint_{\gamma_A} (\eta I - \hat{A})^{-1} d\eta. \tag{10}$$

Note that

$$(\eta I - \hat{A})^{-1} = (\eta I - A - X)^{-1} = \left[ (\eta I - A)(I - \mathcal{R}_A(\eta) X) \right]^{-1}$$
$$= (I - \mathcal{R}_A(\eta) X)^{-1} \mathcal{R}_A(\eta)$$

where $\mathcal{R}_A(\eta) := (\eta I - A)^{-1}$. clearly

$$\left\| \mathcal{R}_A(\eta) X \right\| \le \|\mathcal{R}_A(\eta)\| \|X\| \le \frac{2\|X\|}{\lambda_r} < 1.$$

Therefore, we write the Neumann series:

$$(I - \mathcal{R}_A(\eta) X)^{-1} = I + \sum_{j \ge 1} [\mathcal{R}_A(\eta) X]^j. \tag{11}$$

By (11) and (10), we get

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

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

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

$$\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}} = \mathcal{S}_A(X) := \sum_{j \geq 1} \frac{1}{2\pi i} \oint_{\gamma_A} \left[\mathcal{R}_A(\eta)X\right]^j \mathcal{R}_A(\eta) d\eta.$$

For $k \geq 1$, we define

$$\mathcal{S}_{A,k}(X) = \frac{1}{2\pi i} \oint_{\gamma_A} \left[\mathcal{R}_A(\eta)X\right]^k \mathcal{R}_A(\eta) d\eta \tag{12}$$

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

$$\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}} = \sum_{k \geq 1} \mathcal{S}_{A,k}(X). \tag{13}$$

By (13), it suffices to derive explicit expression formulas for $\{\mathcal{S}_{A,k}(X)\}_{k \geq 1}$. Before dealing with general $k$, let us derive $\mathcal{S}_{A,k}(X)$ for $k = 1, 2$ to interpret the shared styles.

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

$$\mathcal{R}_A(\eta) = \Theta(\eta \cdot I_r - \Lambda)^{-1}\Theta^{\mathsf{T}} + \eta^{-1}\Theta_{\perp}\Theta_{\perp}^{\mathsf{T}} = \sum_{j=1}^{d} \frac{1}{\eta - \lambda_j}\theta_j\theta_j^{\mathsf{T}}$$

where we set $\lambda_j = 0$ for all $r + 1 \leq j \leq d$. Denote $P_j = \theta_j\theta_j^{\mathsf{T}}$ for all $1 \leq j \leq d$ which represents the spectral projector onto $\theta_j$.

**Derivation of $\mathcal{S}_{A,1}(X)$.** By the definition of $\mathcal{S}_{A,1}(X)$,

$$\mathcal{S}_{A,1}(X) = \frac{1}{2\pi i} \oint_{\gamma_A} \mathcal{R}_A(\eta) X \mathcal{R}_A(\eta) d\eta$$

$$= \sum_{j_1=1}^{d} \sum_{j_2=1}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2}. \tag{14}$$

*Case 1:* both $j_1$ and $j_2$ are greater than $r$. In this case, the contour integral in (14) 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

$$\sum_{j_1=1}^{r} \sum_{j_2 > r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{\eta^{-1} d\eta}{\eta - \lambda_{j_1}} P_{j_1} X P_{j_2} = \sum_{j_1=1}^{r} \sum_{j_2 > r} \lambda_{j_1}^{-1} P_{j_1} X P_{j_2} = \mathfrak{P}^{-1} X \mathfrak{P}^{\perp}.$$

*Case 3:* none of $j_1$ and $j_2$ is greater than $r$. Clearly, the contour integral in (14) is zero.

To sum up, we conclude with $\mathcal{S}_{A,1}(X) = \mathfrak{P}^{-1} X \mathfrak{P}^{\perp} + \mathfrak{P}^{\perp} X \mathfrak{P}^{-1}$.

**Derivation of $\mathcal{S}_{A,2}(X)$.** By the definition of $\mathcal{S}_{A,2}(X)$,

$$
\begin{aligned}
\mathcal{S}_{A,2}(X) =& \frac{1}{2\pi i} \oint_{\gamma_A} \mathcal{R}_A(\eta) X \mathcal{R}_A(\eta) X \mathcal{R}_A(\eta) d\eta \\
=& \sum_{j_1=1}^{d} \sum_{j_2=1}^{d} \sum_{j_3=1}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})(\eta - \lambda_{j_3})} P_{j_1} X P_{j_2} X P_{j_3}.
\end{aligned}
$$

$$(15)$$

*Case 1*: all $j_1, j_2, j_3$ are greater than $r$. The contour integral in (15) 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

$$
\sum_{j_1=1}^{r} \sum_{j_2,j_3>r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{\eta^{-2} d\eta}{\eta - \lambda_{j_1}} P_{j_1} X P_{j_2} X P_{j_3}
$$

$$
= \sum_{j_1=1}^{r} \sum_{j_2,j_3>r}^{d} \frac{1}{\lambda_{j_1}^2} P_{j_1} X P_{j_2} X P_{j_3} = \mathfrak{P}^{-2} X \mathfrak{P}^{\perp} X \mathfrak{P}^{\perp}.
$$

*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

$$
\sum_{j_1,j_2=1}^{r} \sum_{j_3>r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{\eta^{-1} d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2} X P_{j_3}
$$

$$
= \sum_{j_1=j_2=1}^{r} \sum_{j_3>r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{\eta^{-1} d\eta}{(\eta - \lambda_{j_1})^2} P_{j_1} X P_{j_1} X P_{j_3}
$$

$$
+ \sum_{j_1 \neq j_2 \geq 1}^{r} \sum_{j_3>r}^{d} \frac{1}{2\pi i} \oint_{\gamma_A} \frac{\eta^{-1} d\eta}{(\eta - \lambda_{j_1})(\eta - \lambda_{j_2})} P_{j_1} X P_{j_2} X P_{j_3}
$$

$$
= - \sum_{j_1=1}^{r} \lambda_{j_1}^{-2} P_{j_1} X P_{j_1} X \mathfrak{P}^{\perp} - \sum_{j_1 \neq j_2 \geq 1}^{r} (\lambda_{j_1} \lambda_{j_2})^{-1} P_{j_1} X P_{j_2} X \mathfrak{P}^{\perp}
$$

$$
= - \mathfrak{P}^{-1} X \mathfrak{P}^{-1} X \mathfrak{P}^{\perp}.
$$

*Case 4:* none of $j_1, j_2, j_3$ is greater than $r$. Clearly, the contour integral in (15) is zero.

To sum up, we obtain

$$
\begin{aligned}
\mathcal{S}_{A,2}(X) =& \left( \mathfrak{P}^{-2} X \mathfrak{P}^{\perp} X \mathfrak{P}^{\perp} + \mathfrak{P}^{\perp} X \mathfrak{P}^{-2} X \mathfrak{P}^{\perp} + \mathfrak{P}^{\perp} X \mathfrak{P}^{\perp} X \mathfrak{P}^{-2} \right) \\
& - \left( \mathfrak{P}^{\perp} X \mathfrak{P}^{-1} X \mathfrak{P}^{-1} + \mathfrak{P}^{-1} X \mathfrak{P}^{\perp} X \mathfrak{P}^{-1} + \mathfrak{P}^{-1} X \mathfrak{P}^{-1} X \mathfrak{P}^{\perp} \right).
\end{aligned}
$$

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

$$\mathcal{S}_{A,k}(X) = \sum_{j_1,\cdots,j_{k+1}\geq 1}^{d} \frac{1}{2\pi i}\oint_{\gamma_A} \Big(\prod_{i=1}^{k+1}\frac{1}{\eta-\lambda_{j_i}}\Big)d\eta\, P_{j_1}XP_{j_2}X\cdots P_{j_k}XP_{j_{k+1}}.$$

(16)

We consider components of summations in (16). For instance, consider the cases that some $k_1$ indices from $\{j_1,\cdots,j_{k+1}\}$ are not larger than $r$. W.L.O.G., let $j_1,\cdots,j_{k_1}\leq r$ and $j_{k_1+1},\cdots,j_{k+1} > r$. By Cauchy integral formula, the integral in (16) 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,

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

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

Recall that our goal is to prove

$$\mathcal{S}_{A,k}(X) = \sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k}(-1)^{1+\tau(\mathbf{s})}\cdot\mathfrak{P}^{-s_1}X\mathfrak{P}^{-s_2}X\cdots X\mathfrak{P}^{-s_{k+1}}.$$

Accordingly, in the above summations, we consider the components, where $s_1,\cdots,s_{k_1}\geq 1$ and $s_{k_1+1} = \cdots = s_{k+1} = 0$, namely,

$$\sum_{\substack{s_1+\cdots+s_{k_1}=k\\ s_j\geq 1}}(-1)^{k_1+1}\mathfrak{P}^{-s_1}X\cdots X\mathfrak{P}^{-s_{k_1}}X\mathfrak{P}^{\perp}\cdots X\mathfrak{P}^{\perp}.$$

It turns out that we need to prove

$$\sum_{j_1,\cdots,j_{k_1}\geq 1}^{r}\frac{1}{2\pi i}\oint_{\gamma_A}\Big(\prod_{i=1}^{k_1}\frac{1}{\eta-\lambda_{j_i}}\Big)\eta^{k_1-k-1}d\eta\, P_{j_1}XP_{j_2}X\cdots P_{j_{k_1}}$$

$$= \sum_{j_1,\cdots,j_{k_1}\geq 1}^{r}\sum_{\substack{s_1+\cdots+s_{k_1}=k\\ s_j\geq 1}}(-1)^{k_1+1}\frac{1}{\lambda_{j_1}^{s_1}\cdots\lambda_{j_{k_1}}^{s_{k_1}}}P_{j_1}XP_{j_2}X\cdots XP_{j_{k_1}}.$$

It suffices to prove that for all $\mathbf{j} = (j_1,\ldots,j_{k_1})\in\{1,\cdots,r\}^{k_1}$,

$$\frac{1}{2\pi i}\oint_{\gamma_A}\frac{d\eta}{(\eta-\lambda_{j_1})\cdots(\eta-\lambda_{j_{k_1}})\eta^{k+1-k_1}} = \sum_{\substack{s_1+\cdots+s_{k_1}=k\\ s_j\geq 1}}(-1)^{k_1+1}\frac{1}{\lambda_{j_1}^{s_1}\cdots\lambda_{j_{k_1}}^{s_{k_1}}}.$$

(17)

To prove (17), we rewrite its right hand side. Given any $\mathbf{j} = (j_1, \cdots, j_{k_1}) \in \{1, \cdots, r\}^{k_1}$, define

$$\mathbf{v}_i(\mathbf{j}) := \{1 \le t \le k_1 : j_t = i\} \quad \text{for } 1 \le i \le r$$

, that is, $\mathbf{v}_i(\mathbf{j})$ contains the location $s$ such that $\lambda_{j_s} = \lambda_i$. Meanwhile, denote $v_i(\mathbf{j}) = \mathrm{Card}(\mathbf{v}_i(\mathbf{j}))$. Then, the right hand side of (17) is written as

$$\sum_{\substack{s_1+\cdots+s_{k_1}=k \\ s_j \ge 1}} (-1)^{k_1+1} \frac{1}{\lambda_{j_1}^{s_1} \cdots \lambda_{j_{k_1}}^{s_{k_1}}}$$

$$= (-1)^{k_1+1} \sum_{\substack{s_1+\cdots+s_{k_1}=k \\ s_j \ge 1}} \lambda_1^{-\sum_{p \in \mathbf{v}_1(\mathbf{j})} s_p} \cdots \lambda_r^{-\sum_{p \in \mathbf{v}_r(\mathbf{j})} s_p}.$$

Now, we denote $t_i(\mathbf{j}) = \sum_{p \in \mathbf{v}_i(\mathbf{j})} s_p$ for $1 \le i \le r$, we can write the above equation as

$$\sum_{\substack{s_1+\cdots+s_{k_1}=k \\ s_j \ge 1}} (-1)^{k_1+1} \frac{1}{\lambda_{j_1}^{s_1} \cdots \lambda_{j_{k_1}}^{s_{k_1}}}$$

$$= (-1)^{k_1+1} \sum_{\substack{t_1(\mathbf{j})+\cdots+t_r(\mathbf{j})=k \;\; i:v_i(\mathbf{j}) \ge 1 \\ t_i(\mathbf{j}) \ge v_i(\mathbf{j}) \\ t_i(\mathbf{j})=0 \text{ if } v_i(\mathbf{j})=0}} \prod \binom{t_i(\mathbf{j})-1}{v_i(\mathbf{j})-1} \lambda_i^{-t_i(\mathbf{j})}$$

$$= (-1)^{k_1+1} \sum_{\substack{t_1(\mathbf{j})+\cdots+t_r(\mathbf{j})=k-k_1 \;\; i:v_i(\mathbf{j}) \ge 1 \\ t_i(\mathbf{j})=0 \text{ if } v_i(\mathbf{j})=0}} \prod \binom{t_i(\mathbf{j})+v_i(\mathbf{j})-1}{v_i(\mathbf{j})-1} \lambda_i^{-t_i(\mathbf{j})-v_i(\mathbf{j})}$$

where the last equality is due to the fact $v_1(\mathbf{j}) + \cdots + v_r(\mathbf{j}) = k_1$. Similarly, the left hand side of (17) can be written as

$$\frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_{j_1}) \cdots (\eta - \lambda_{j_{k_1}}) \eta^{k+1-k_1}}$$

$$= \frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_1)^{v_1(\mathbf{j})} \cdots (\eta - \lambda_{j_r})^{v_r(\mathbf{j})} \eta^{k+1-k_1}}.$$

Therefore, in order to prove (17), it suffices to prove that for any $\mathbf{j} = (j_1, \cdots, j_{k_1})$ the following equality holds

$$\frac{1}{2\pi i} \oint_{\gamma_A} \frac{d\eta}{(\eta - \lambda_1)^{v_1} \cdots (\eta - \lambda_{j_r})^{v_r} \eta^{k+1-k_1}}$$

$$= (-1)^{k_1+1} \sum_{\substack{t_1+\cdots+t_r=k-k_1 \;\; i:v_i \ge 1 \\ t_i=0 \text{ if } v_i=0}} \prod_{i=1}^{r} \binom{t_i+v_i-1}{v_i-1} \lambda_i^{-t_i-v_i} \qquad (18)$$

where we omitted the index $\mathbf{j}$ in definitions of $v_i(\mathbf{j})$ and $t_i(\mathbf{j})$ without causing any confusions. The non-negative numbers $v_1 + \cdots + v_r = k_1$. We define the function

$$\varphi(\eta) = \frac{1}{(\eta - \lambda_1)^{v_1} \cdots (\eta - \lambda_r)^{v_r} \eta^{k+1-k_1}}$$

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

$$\text{Res}(\varphi, \eta = 0) = \frac{1}{(k-k_1)!} \left[ \prod_{i:v_i \geq 1}^r (\eta - \lambda_i)^{-v_i} \right]^{(k-k_1)} \Bigg|_{\eta=0}$$

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

$$\text{Res}(\varphi, \eta = 0) = \frac{1}{(k-k_1)!} \sum_{\substack{t_1+\cdots+t_r=k-k_1 \\ t_i=0 \text{ if } v_i=0}} \frac{(k-k_1)!}{t_1! t_2! \cdots t_r!} \prod_{i:v_i \geq 1}^r \left[ (\eta - \lambda_i)^{-v_i} \right]^{(t_i)} \Bigg|_{\eta=0}$$

$$= (-1)^{k-k_1} \sum_{\substack{t_1+\cdots+t_r=k-k_1 \\ t_i=0 \text{ if } v_i=0}} \prod_{i:v_i \geq 1}^r \frac{v_i(v_i+1)\cdots(v_i+t_i-1)}{t_i!} (-\lambda_i)^{-v_i-t_i}$$

$$= (-1)^{k-k_1} \sum_{\substack{t_1+\cdots+t_r=k-k_1 \\ t_i=0 \text{ if } v_i=0}} \prod_{i:v_i \geq 1}^r \binom{t_i+v_i-1}{v_i-1} (-\lambda_i)^{-v_i-t_i}$$

$$= (-1)^{2k-k_1} \sum_{\substack{t_1+\cdots+t_r=k-k_1 \\ t_i=0 \text{ if } v_i=0}} \prod_{i:v_i \geq 1}^r \binom{t_i+v_i-1}{v_i-1} \lambda_i^{-v_i-t_i}.$$

Therefore,

$$\frac{1}{2\pi i} \oint_{\gamma_A} \varphi(\eta) d\eta = (-1)^{k_1+1} \sum_{\substack{t_1+\cdots+t_r=k-k_1 \\ t_i=0 \text{ if } v_i=0}} \prod_{i:v_i \geq 1}^r \binom{t_i+v_i-1}{v_i-1} \lambda_i^{-v_i-t_i}$$

which proves (18). We conclude the proof of Theorem 1.

# References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*. MR4124330

Bao, Z., Ding, X., and Wang, K. (2018). Singular vector and singular subspace distribution for the matrix denoising model. *arXiv preprint arXiv:1809.10476*. MR4206682

Benaych-Georges, F. and Nadakuditi, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135. MR2944410

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136. MR0003498

Bloemendal, A., Knowles, A., Yau, H.-T., and Yin, J. (2016). On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1-2):459–552. MR3449395

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982. MR2600248

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89. MR3766946

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007. MR3332993

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080. MR2723472

Chen, Y., Cheng, C., and Fan, J. (2018). Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *arXiv preprint arXiv:1811.12804*. MR4206685

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46. MR0264450

Ding, X. (2017). High dimensional deformed rectangular matrices with applications in matrix denoising. *arXiv:1702.06975*. MR4036038

Ding, X. and Yang, F. (2020). Edge statistics of large dimensional deformed rectangular matrices. *arXiv preprint arXiv:2009.00389*.

Donoho, D. and Gavish, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440. MR3269984

Esseen, C.-G. (1942). On the liapunoff limit of error in the theory of probability. *Arkiv for matematik, astronomi och fysik*, A28:1–19. MR0011909

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press. MR3024913

Götze, F. and Tikhomirov, A. (2011). On the rate of convergence to the marchenko–pastur distribution. *arXiv preprint arXiv:1110.1284*. MR2655937

Kato, T. (2013). *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media. MR0203473

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998. MR2683452

Kim, H., Howland, P., and Park, H. (2005). Dimension reduction in text classification with support vector machines. *Journal of Machine Learning Research*, 6(Jan):37–53. MR2249814

Koltchinskii, V. and Lounici, K. (2016). 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. MR3573302

Koltchinskii, V. and Lounici, K. (2017). Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157. MR3611488

Koltchinskii, V. and Xia, D. (2016). Perturbation of linear forms of singular vectors under gaussian noise. In *High Dimensional Probability VII*, pages 397–423. Springer. MR3565274

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008. MR2354409

Li, Y. and Li, H. (2018). Two-sample test of community memberships of weighted stochastic block models. *arXiv preprint arXiv:1811.12593*.

Löffler, M., Zhang, A. Y., and Zhou, H. H. (2019). Optimality of spectral clustering in the gaussian mixture model. *arXiv preprint arXiv:1911.00538*.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2017). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*. MR4099988

Mingo, J. A. and Speicher, R. (2017). *Free probability and random matrices*, volume 35. Springer. MR3585560

Naumov, A., Spokoiny, V., and Ulyanov, V. (2017). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields*, pages 1–42. MR3980312

Shabalin, A. A. and Nobel, A. B. (2013). Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76. MR3054091

Tang, M., Priebe, C. E., et al. (2018). Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415. MR3845021

Tao, T. (2012). *Topics in random matrix theory*, volume 132. American Mathematical Soc. MR2906465

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*. MR2963170

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111. MR0309968

Xia, D. (2019). Confidence interval of singular subspaces for high-dimensional and low-rank matrix regression. *IEEE Transactions on Information Theory*. MR4030894

Xia, D. and Yuan, M. (2018+). On polynomial time methods for exact low rank tensor completion. *Foundations of Computational Mathematics*. MR4029842

Xia, D., Yuan, M., and Zhang, C.-H. (2017). Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*. MR4206670

Xia, D. and Zhou, F. (2019). The sup-norm perturbation of hosvd and low rank tensor denoising. *Journal of Machine Learning Research*, 20(61):1–42. MR3960915

Yu, Y., Wang, T., and Samworth, R. J. (2014). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323. MR3371006

Zhang, A. and Xia, D. (2018). Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*. MR3876445

## Appendix A: Proofs

### *A.1. Proof of Theorem 2*

By $\mathrm{rank}(\hat{\Theta}) = \mathrm{rank}(\Theta) = 2r$, we get

$$\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] = \|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 = 4r - 2\langle \hat{\Theta}\hat{\Theta}^{\mathsf{T}}, \Theta\Theta^{\mathsf{T}}\rangle.$$

Since $X$ is random, we shall take care of the "size" of $X$. Observe that $\|X\| = \|Z\|$ and the operator norm of $Z$ is well-known (see, e.g., (Tao, 2012) and (Vershynin, 2010)). Indeed, there exist some absolute constants $C_1, C_2, c_1 > 0$ such that

$$\mathbb{E}\|X\| \leq C_1\sqrt{d_2} \quad \text{and} \quad \mathbb{P}\big(\|X\| \geq C_2\sqrt{d_2}\big) \leq e^{-c_1 d_2} \tag{19}$$

where $d_2 = \max\{d_1, d_2\}$. Meanwhile, $\mathbb{E}^{1/p}\|X\|^p \leq C_3\sqrt{d_2}$ for all integer $p \geq 1$. See (Koltchinskii and Xia, 2016, Lemma 3).

Denote the event $\mathcal{E}_1 := \{\|X\| \leq C_2\sqrt{d_{\max}}\}$ so that $\mathbb{P}(\mathcal{E}_1) \geq 1 - e^{-c_1 d_2}$. Assume that $\lambda_r > 2C_2\sqrt{d_2}$, our analysis is conditioned on $\mathcal{E}_1$. By Theorem 1, on event $\mathcal{E}_1$, we have

$$\hat{\Theta}\hat{\Theta}^{\mathsf{T}} = \Theta\Theta^{\mathsf{T}} + \mathcal{S}_{A,1}(X) + \mathcal{S}_{A,2}(X) + \sum_{k \geq 3} \mathcal{S}_{A,3}(X)$$

where $\mathcal{S}_{A,1}(X) = \mathfrak{P}^{-1}X\mathfrak{P}^{\perp} + \mathfrak{P}^{\perp}X\mathfrak{P}^{-1}$ and

$$\begin{aligned}
\mathcal{S}_{A,2}(X) =& \big(\mathfrak{P}^{-2}X\mathfrak{P}^{\perp}X\mathfrak{P}^{\perp} + \mathfrak{P}^{\perp}X\mathfrak{P}^{-2}X\mathfrak{P}^{\perp} + \mathfrak{P}^{\perp}X\mathfrak{P}^{\perp}X\mathfrak{P}^{-2}\big) \\
&- \big(\mathfrak{P}^{\perp}X\mathfrak{P}^{-1}X\mathfrak{P}^{-1} + \mathfrak{P}^{-1}X\mathfrak{P}^{\perp}X\mathfrak{P}^{-1} + \mathfrak{P}^{-1}X\mathfrak{P}^{-1}X\mathfrak{P}^{\perp}\big).
\end{aligned}$$

Therefore, we get

$$\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 = 2\mathrm{tr}\big(\mathfrak{P}^{-1}X\mathfrak{P}^{\perp}X\mathfrak{P}^{-1}\big) - 2\sum_{k \geq 3}\big\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X)\big\rangle$$

$$=2\|\mathfrak{P}^{-1}X\mathfrak{P}^{\perp}\|_{\mathrm{F}}^2 - 2\sum_{k\geq 3}\big\langle\Theta\Theta^{\mathsf{T}},\mathcal{S}_{A,k}(X)\big\rangle.$$

Then,

$$\begin{aligned}
\mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] &- \mathbb{E}\ \mathrm{dist}^2[(\hat{U},\hat{V}),(U,V)] \\
&= \Big(2\|\mathfrak{P}^{-1}X\mathfrak{P}^{\perp}\|_{\mathrm{F}}^2 - 2\mathbb{E}\|\mathfrak{P}^{-1}X\mathfrak{P}^{\perp}\|_{\mathrm{F}}^2\Big) - 2\sum_{k\geq 3}\big\langle\Theta\Theta^{\mathsf{T}},\mathcal{S}_{A,k}(X) - \mathbb{E}\mathcal{S}_{A,k}(X)\big\rangle.
\end{aligned}$$

We investigate the normal approximation of

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

and show that

$$\frac{2\sum_{k\geq 3}\big\langle\Theta\Theta^{\mathsf{T}},\mathcal{S}_{A,k}(X) - \mathbb{E}\mathcal{S}_{A,k}(X)\big\rangle}{\sqrt{8(d_1 + d_2 - 2r)}\cdot\|\Lambda^{-2}\|_{\mathrm{F}}}$$

is ignorable when signal strength $\lambda_r$ is sufficiently strong. For some $t > 0$ which shall be determined later, define a function

$$f_t(X) = 2\sum_{k\geq 3}\big\langle\Theta\Theta^{\mathsf{T}},\mathcal{S}_{A,k}(X)\big\rangle\cdot\phi\Big(\frac{\|X\|}{t\cdot\sqrt{d_2}}\Big) \tag{20}$$

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 by

$$\phi(s) := \begin{cases} 1 & \text{if } s \leq 1, \\ 2 - s & \text{if } 1 < s \leq 2, \\ 0 & \text{if } s > 2. \end{cases}$$

Clearly, $\phi(s)$ is Lipschitz with constant 1. Lemma 4 shows that $f(\cdot)$ is Lipschitz when $\lambda_r \geq C_4\sqrt{d_2}$. The proof of Lemma 4 is in Appendix, Section B.1.

**Lemma 4.** *There exist absolute constants $C_3, C_4 > 0$ so that if $\lambda_r \geq C_3 t^2\sqrt{d_2}$, then*

$$\big|f_t(X_1) - f_t(X_2)\big| \leq C_4 t^2\frac{rd_2}{\lambda_r^3}\cdot\|X_1 - X_2\|_{\mathrm{F}}$$

*where $f_t(X)$ is defined by (20).*

By Lemma 4 and Gaussian isoperimetric inequality (see, e.g., (Koltchinskii and Lounici, 2016, 2017)), it holds with probability at least $1 - e^{-s}$ for any $s \geq 1$ that

$$\left|2\sum_{k\geq 3}\big\langle\Theta\Theta^{\mathsf{T}},\mathcal{S}_{A,k}(X)\big\rangle\cdot\phi\Big(\frac{\|X\|}{t\cdot\sqrt{d_2}}\Big) - \mathbb{E}2\sum_{k\geq 3}\big\langle\Theta\Theta^{\mathsf{T}},\mathcal{S}_{A,k}(X)\big\rangle\cdot\phi\Big(\frac{\|X\|}{t\cdot\sqrt{d_2}}\Big)\right|$$

$$\leq C_5\sqrt{s}t^2\frac{rd_2}{\lambda_r^3} \tag{21}$$

for some absolute constant $C_5 > 0$. Now, set $t = C_2$ where $C_2$ is defined in (19). Therefore, $\phi\left(\frac{\|X\|}{C_2 \cdot \sqrt{d_2}}\right) = 1$ on event $\mathcal{E}_1$. Meanwhile, the following fact holds

$$
\left| \mathbb{E}2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \cdot \phi\left(\frac{\|X\|}{C_2 \cdot \sqrt{d_2}}\right) - \mathbb{E}2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \right|
$$

$$
\leq \left| \mathbb{E}2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \cdot \phi\left(\frac{\|X\|}{C_2 \cdot \sqrt{d_2}}\right) \mathbb{I}_{\mathcal{E}_1^c} \right|
$$

$$
+ \left| \mathbb{E}2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \mathbb{I}_{\mathcal{E}_1^c} \right|
$$

$$
\leq 4 \sum_{k \geq 3} \mathbb{E}\left| \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \right| \mathbb{I}_{\mathcal{E}_1^c} \leq 8r \sum_{k \geq 3} \mathbb{E}^{1/2} \|\mathcal{S}_{A,k}(X)\|^2 \cdot e^{-c_1 d_2/2}
$$

$$
\leq e^{-c_1 d_2/2} \cdot 8r \sum_{k \geq 3} \mathbb{E}^{1/2} \frac{16^k \|X\|^{2k}}{\lambda_r^{2k}} \leq e^{-c_1 d_2/2} \cdot 8r \sum_{k \geq 3} \left(\frac{C_6 \cdot d_2^{1/2}}{\lambda_r}\right)^k
$$

$$
\leq e^{-c_1 d_2/2} \cdot \frac{C_6 r d_2^{3/2}}{\lambda_r^3} \leq C_6 \frac{r d_2}{\lambda_r^3}
$$

where the last inequality holds as long as $e^{-c_1 d_2/2} \leq \frac{1}{\sqrt{d_2}}$ and we used the fact $\mathbb{E}^{1/p} \|X\|^p \leq C_6 \sqrt{d_2}$ 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 (21), it holds with probability at least $1 - e^{-s} - e^{-c_1 d_2}$ for any $s \geq 1$ that

$$
\left| 2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle - \mathbb{E}2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \right| \leq C_6 s^{1/2} \cdot \frac{r d_2}{\lambda_r^3}
$$

for some absolute constant $C_6 > 0$. Therefore, for any $s \geq 1$, with probability at least $1 - e^{-s} - e^{-c_1 d_2}$,

$$
\frac{\left| 2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle - \mathbb{E}2 \sum_{k \geq 3} \left\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \right\rangle \right|}{\sqrt{8(d_1 + d_2 - 2r)} \|\Lambda^{-2}\|_{\mathrm{F}}}
$$

$$
\leq C_6 s^{1/2} \left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}} \lambda_r^2}\right) \cdot \frac{\sqrt{r d_2}}{\lambda_r} \tag{22}
$$

where we assumed $d_2 \geq 3r$.

We next prove the normal approximation of $2\|\mathfrak{P}^{-1} X \mathfrak{P}^{\perp}\|_{\mathrm{F}}^2$. Similar as in (Xia, 2019), by the definition of $\mathfrak{P}^{-1}, X$ and $\mathfrak{P}^{\perp}$, we could write

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

Then,

$$
\|\mathfrak{P}^{-1} X \mathfrak{P}^{\perp}\|_{\mathrm{F}}^2 = \|U\Lambda^{-1}V^{\mathsf{T}} Z^{\mathsf{T}} U_{\perp} U_{\perp}^{\mathsf{T}}\|_{\mathrm{F}}^2 + \|V\Lambda^{-1}U^{\mathsf{T}} Z V_{\perp} V_{\perp}^{\mathsf{T}}\|_{\mathrm{F}}^2.
$$

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

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

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

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

Sincet $U^{\mathsf{T}} z_j$ and $U_\perp^{\mathsf{T}} z_j$ are Gaussian random vectors and

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

, we know that $\{U^{\mathsf{T}} z_j\}_{j=1}^{d_2}$ are independent with $\{U_\perp^{\mathsf{T}} z_j\}_{j=1}^{d_2}$.

Therefore, $\|U \Lambda^{-1} V^{\mathsf{T}} Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}}\|_{\mathrm{F}}^2$ is independent with $\|V \Lambda^{-1} U^{\mathsf{T}} Z V_\perp V_\perp^{\mathsf{T}}\|_{\mathrm{F}}^2$. Denote by $\tilde{Z}$ an independent copy of $Z$, we conclude that ($Y_1 \overset{\mathrm{d}}{=} Y_2$ denotes equivalence of $Y_1$ and $Y_2$ in distribution)

$$\|\mathfrak{P}^{-1} X \mathfrak{P}^\perp\|_{\mathrm{F}}^2 \overset{\mathrm{d}}{=} \|U \Lambda^{-1} V^{\mathsf{T}} Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}}\|_{\mathrm{F}}^2 + \|V \Lambda^{-1} U^{\mathsf{T}} \tilde{Z} V_\perp V_\perp^{\mathsf{T}}\|_{\mathrm{F}}^2$$

$$= \sum_{j=r+1}^{d_1} \|U \Lambda^{-1} V^{\mathsf{T}} Z^{\mathsf{T}} u_j\|_{\ell_2}^2 + \sum_{j=r+1}^{d_2} \|V \Lambda^{-1} U^{\mathsf{T}} \tilde{Z} v_j\|_{\ell_2}^2$$

$$= \sum_{j=r+1}^{d_1} \|\Lambda^{-1} V^{\mathsf{T}} Z^{\mathsf{T}} u_j\|_{\ell_2}^2 + \sum_{j=r+1}^{d_2} \|\Lambda^{-1} U^{\mathsf{T}} \tilde{Z} v_j\|_{\ell_2}^2$$

where $\{u_j\}_{j=r+1}^{d_1}$ and $\{v_j\}_{j=r+1}^{d_2}$ denote the columns of $U_\perp$ and $V_\perp$, respectively. Observe that $Z^{\mathsf{T}} u_j \sim \mathcal{N}(0, I_{d_2})$ for all $r + 1 \le j \le d_1$ and

$$\mathbb{E} (Z^{\mathsf{T}} u_{j_1})(Z^{\mathsf{T}} u_{j_2})^{\mathsf{T}} = 0 \quad \text{for all } r + 1 \le j_1 \ne j_2 \le d_1.$$

Therefore, $\{Z^{\mathsf{T}} u_j\}_{j=r+1}^{d_1}$ are independent normal random vectors. Similarly, $\tilde{Z} v_j \sim \mathcal{N}(0, I_{d_1})$ are independent for all $r + 1 \le j \le d_2$. Clearly, $V^{\mathsf{T}} Z^{\mathsf{T}} u_{j_1} \sim \mathcal{N}(0, I_r)$ and $U^{\mathsf{T}} \tilde{Z} v_{j_2} \sim \mathcal{N}(0, I_r)$ are all independent for $r + 1 \le j_1 \le d_1$ and $r + 1 \le j_2 \le d_2$.

As a result, let $d_\star = d_1 + d_2 - 2r$, we conclude that

$$\|\mathfrak{P}^{-1} X \mathfrak{P}^\perp\|_{\mathrm{F}}^2 \overset{\mathrm{d}}{=} \sum_{j=1}^{d_\star} \|\Lambda^{-1} z_j\|_{\ell_2}^2 \tag{23}$$

where we abuse the notations and denote $\{z_j\}_{j=1}^{d_\star}$ where $z_j \overset{i.i.d.}{\sim} \mathcal{N}(0, I_r)$. By Berry-Esseen theorem ((Berry, 1941) and (Esseen, 1942)), it holds for some

absolute constant $C_7 > 0$ that

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{2\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2 - 2\mathbb{E}\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x \right) - \Phi(x) \right|$$

$$\leq C_7 \left( \frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_\star}} \tag{24}$$

where we used the fact $\mathrm{Var}\big(\|\Lambda^{-1}z_j\|_{\ell_2}^2\big) = 2\|\Lambda^{-2}\|_{\mathrm{F}}^2$ and

$$\mathbb{E}\big\|\Lambda^{-1}z_j\big\|_{\ell_2}^6 \leq C_7 \sum_{j_1, j_2, j_3 \geq 1}^{r} \frac{1}{\lambda_{j_1}^2 \lambda_{j_2}^2 \lambda_{j_3}^2} \leq C_7 \|\Lambda^{-1}\|_{\mathrm{F}}^6.$$

In (24), the function $\Phi(x)$ denotes the c.d.f. of standard normal distributions. Recall that, on event $\mathcal{E}_1$,

$$\frac{\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}}$$

$$= \frac{2\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2 - 2\mathbb{E}\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}} + \frac{2\sum_{k \geq 3}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) - \mathbb{E}\mathcal{S}_{A,k}(X)\rangle}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}}$$

where normal approximation of the first term is given in (24) and upper bound of the second term is given in (22). Based on (22), we get for any $x \in \mathbb{R}$ and any $s \geq 1$,

$$\mathbb{P}\left( \frac{\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x \right)$$

$$\leq \mathbb{P}\left( \frac{2\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2 - 2\mathbb{E}\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x + C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2} \cdot \frac{\sqrt{rd_2}}{\lambda_r} \right)$$

$$\qquad + e^{-s} + e^{-c_1 d_2}$$

$$\leq \Phi\left( x + C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2} \cdot \frac{\sqrt{rd_2}}{\lambda_r} \right) + e^{-s} + e^{-c_1 d_2} + C_7 \left( \frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_\star}}$$

$$\leq \Phi(x) + C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2} \cdot \frac{\sqrt{rd_2}}{\lambda_r} + e^{-s} + e^{-c_1 d_2} + C_7 \left( \frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2} \right)^{3/2} \cdot \frac{1}{\sqrt{d_\star}}$$

where the last inequality is due to (24) and the Lipschitz property of $\Phi(x)$. Similarly, for any $x \in \mathbb{R}$ and any $s \geq 1$,

$$\mathbb{P}\left( \frac{\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \mathbb{E}\,\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)]}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x \right)$$

$$\geq \mathbb{P}\left( \frac{2\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2 - 2\mathbb{E}\|\mathfrak{P}^{-1}X\mathfrak{P}^\perp\|_{\mathrm{F}}^2}{\sqrt{8(d_1 + d_2 - 2r)}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x - C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2} \cdot \frac{\sqrt{rd_2}}{\lambda_r} \right)$$

$$\qquad - e^{-s} - e^{-c_1 d_2}$$

$$\geq \Phi\left(x - C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F \lambda_r^2} \cdot \frac{\sqrt{r d_2}}{\lambda_r}\right) - e^{-s} - e^{-c_1 d_2} - C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_\star}}$$

$$\geq \Phi(x) - C_6 s^{1/2} \cdot \frac{\sqrt{r}}{\|\Lambda^{-2}\|_F \lambda_r^2} \cdot \frac{\sqrt{r d_2}}{\lambda_r} - e^{-s} - e^{-c_1 d_2} - C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_\star}}.$$

Finally, we conclude that for any $s \geq 1$,

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

$$\leq C_6 s^{1/2}\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_F \lambda_r^2}\right) \cdot \frac{\sqrt{r d_2}}{\lambda_r} + e^{-s} + e^{-c_1 d_2} + C_7\left(\frac{\|\Lambda^{-1}\|_F^4}{\|\Lambda^{-2}\|_F^2}\right)^{3/2} \frac{1}{\sqrt{d_\star}}$$

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

### A.2.  Proof of Theorem 6

It suffices to prove

$$\frac{\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \tilde{B}_{\lceil \log d_2 \rceil}}{\sqrt{8 d_*} \|\tilde{\Lambda}^{-2}\|_F} \xrightarrow{\mathrm{d.}} \mathcal{N}(0, 1)$$

as $d_1, d_2 \to \infty$. Toward that end, we write

$$\frac{\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)] - \tilde{B}_{\lceil \log d_2 \rceil}}{\sqrt{8 d_*} \|\tilde{\Lambda}^{-2}\|_F} = \frac{\operatorname{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_{\lceil \log d_2 \rceil}}{\sqrt{8 d_*} \|\Lambda^{-2}\|_F} \cdot \frac{\|\Lambda^{-2}\|_F}{\|\tilde{\Lambda}^{-2}\|_F}$$

$$+ \frac{B_{\lceil \log d_2 \rceil} - \tilde{B}_{\lceil \log d_2 \rceil}}{\sqrt{8 d_*} \|\tilde{\Lambda}^{-2}\|_F}, \tag{25}$$

where the first term in the RHS of (25) converges to $\mathcal{N}(0, 1)$ in view of Theorem 4 and (9). It suffices to show that the second term in the RHS of (25) converges to 0 in distribution. We prove it in two cases.

 *Case 1*: if $d_2^{3/4} \log(d_2) \lambda_r^{-1} = O(1)$. In this case, we write

$$B_{\lceil \log d_2 \rceil} - \tilde{B}_{\lceil \log d_2 \rceil} = 2 d_*(\|\Lambda^{-1}\|_F^2 - \|\tilde{\Lambda}^{-1}\|_F^2)$$

$$- 2 \sum_{k_0=2}^{\lceil \log d_2 \rceil} (-1)^{k_0} (d_{1-}^{k_0-1} - d_{2-}^{k_0-1})(d_{1-} - d_{2-})(\|\tilde{\Lambda}^{-k_0}\|_F^2 - \|\Lambda^{-k_0}\|_F^2).$$

Note that (8) and (9) implies that $\tilde{\lambda}_j^2 - \lambda_j^2 = d_2 \cdot o_p(1)$, where $o_p(1)$ denotes a random variable converges to 0 almost surely. Then,

$$\frac{B_{\lceil \log d_2 \rceil} - \tilde{B}_{\lceil \log d_2 \rceil}}{\sqrt{8 d_*} \|\tilde{\Lambda}^{-2}\|_F} = \frac{\sqrt{d_2^3 r} \log d_2}{\lambda_r^2} \cdot o_p(1),$$

which converges to 0 almost surely.

*Case 2*: if $\lambda_r = O(d_2^{3/4} \log d_2)$ and $\sqrt{rd_2}/\lambda_r = o(1)$. In this case, we apply a recent result from Ding and Yang (2020) which holds when $\lambda_r \ll d_2^{5/6}$. By Theorem 2.13[2] of Ding and Yang (2020), we have

$$\hat{\lambda}_j - \sqrt{\lambda_j^2 + d_1 + d_2 + d_1 d_2 \lambda_j^{-2}} = O_p(\log^{-1} d_2),$$

where $O_p(\log^{-1} d_2)$ stands for "asymptotically bounded by $O(\log^{-1} d_2)$ in probability". Then, by (9), we get $\tilde{\lambda}_j^2 - \lambda_j^2 = \lambda_j \cdot O_p(\log^{-1} d_2)$. As a result,

$$\frac{B_{\lceil \log d_2 \rceil} - \tilde{B}_{\lceil \log d_2 \rceil}}{\sqrt{8d_*}\|\tilde{\Lambda}^{-2}\|_{\mathrm{F}}} = \frac{\sqrt{d_2 r} \log d_2}{\lambda_r} \cdot O_p(\log^{-1} d_2),$$

which converges to 0 in probability.

The proof is concluded by combining *Case 1* and *Case 2*.

### A.3. Proof of lemmas in Section 4

Observe that $\mathcal{S}_{A,k}(X)$ involves the product of $X$ for $k$ times. If $k$ is odd, we immediately get $\mathbb{E}\mathcal{S}_{A,k}(X) = 0$ since $Z$ has i.i.d. standard normal entries. Therefore, it suffices to investigate $\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,k}(X) \rangle$ when $k$ is even.

**Proof of Lemma 1.** By the definitions of $\mathfrak{P}^\perp, X$ and $\mathfrak{P}^{-1}$,

$$\mathbb{E}\|\mathfrak{P}^\perp X \mathfrak{P}^{-1}\|_{\mathrm{F}}^2 = \mathbb{E}\|U\Lambda^{-1}V^{\mathsf{T}}Z^{\mathsf{T}}U_\perp U_\perp^{\mathsf{T}}\|_{\mathrm{F}}^2 + \mathbb{E}\|V\Lambda^{-1}U^{\mathsf{T}}ZV_\perp V_\perp^{\mathsf{T}}\|_{\mathrm{F}}^2$$
$$= \mathbb{E}\|\Lambda^{-1}V^{\mathsf{T}}Z^{\mathsf{T}}U_\perp\|_{\mathrm{F}}^2 + \mathbb{E}\|\Lambda^{-1}U^{\mathsf{T}}ZV_\perp\|_{\mathrm{F}}^2.$$

By the proof of Theorem 2, we obtain $\mathbb{E}\|\mathfrak{P}^\perp X \mathfrak{P}^{-1}\|_{\mathrm{F}}^2 = (d_1 + d_2 - 2r)\|\Lambda^{-1}\|_{\mathrm{F}}^2$ which is the first claim. To prove the second claim, it holds by Theorem 1 that

$$\left|\mathbb{E}\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 - 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2\right| \leq 2\left|\sum_{k\geq 2}\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X)\rangle\right|$$

$$\leq 2\sum_{k\geq 2}\left|\mathbb{E}\left\langle \Theta\Theta^{\mathsf{T}}, \sum_{\mathbf{s}:s_1+\cdots+s_{2k+1}=2k}(-1)^{1+\tau(\mathbf{s})}\cdot\mathfrak{P}^{-s_1}X\mathfrak{P}^{-s_2}X\cdots X\mathfrak{P}^{-s_{2k}}X\mathfrak{P}^{-s_{2k+1}}\right\rangle\right|$$

$$= 2\sum_{k\geq 2}\left|\mathbb{E}\left\langle \Theta\Theta^{\mathsf{T}}, \sum_{\substack{\mathbf{s}:s_1+\cdots+s_{2k+1}=2k \\ s_1, s_{2k+1}\geq 1}}(-1)^{1+\tau(\mathbf{s})}\cdot\mathfrak{P}^{-s_1}X\mathfrak{P}^{-s_2}X\cdots X\mathfrak{P}^{-s_{2k}}X\mathfrak{P}^{-s_{2k+1}}\right\rangle\right|$$

where we used the fact $\Theta\Theta^{\mathsf{T}}\mathfrak{P}^0 = \mathfrak{P}^0\Theta\Theta^{\mathsf{T}} = 0$. Then,

$$\left|\mathbb{E}\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 - 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2\right|$$

$$\leq 4r\sum_{k\geq 2}\sum_{\substack{\mathbf{s}:s_1+\cdots+s_{2k+1}=2k \\ s_1, s_{2k+1}\geq 1}}\mathbb{E}\left\|\mathfrak{P}^{-s_1}X\mathfrak{P}^{-s_2}X\cdots X\mathfrak{P}^{-s_{2k}}X\mathfrak{P}^{-s_{2k+1}}\right\|$$

---

[2]Note that the result in Ding and Yang (2020) is even stronger. We only use a weaker version of their results.

$$\leq 4r \sum_{k \geq 2} \sum_{\substack{\mathbf{s}: s_1 + \cdots + s_{2k+1} = 2k \\ s_1, s_{2k+1} \geq 1}} \frac{\mathbb{E}\|X\|^{2k}}{\lambda_r^{2k}}$$

$$\leq 4r \sum_{k \geq 2} \binom{4k}{2k} \frac{\mathbb{E}\|X\|^{2k}}{\lambda_r^{2k}} \leq C_2 r \sum_{k \geq 2} \frac{4^{2k} \mathbb{E}\|X\|^{2k}}{\lambda_r^{2k}}.$$

for some absolute constant $C_2 > 0$. Therefore,

$$\left| \mathbb{E}\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 - 2d_\star \|\Lambda^{-1}\|_{\mathrm{F}}^2 \right| \leq C_2 r \sum_{k \geq 2} \left( \frac{16 C_1^2 d_2}{\lambda_r^2} \right)^k \leq C_2 \frac{r d_2^2}{\lambda_r^4}$$

where the last inequality holds as long as $\lambda_r \geq 5 C_1 \sqrt{d_2}$.

**Property 1: only even order terms matter.** In order to calculate higher order approximations, we need the following useful property of $\mathbb{E}\mathcal{S}_{2k}(X)$.

By Theorem 1,

$$\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X) \rangle = \sum_{\substack{\mathbf{s}: s_1 + \cdots + s_{2k+1} = 2k \\ s_1, s_{2k+1} \geq 1}} (-1)^{1+\tau(\mathbf{s})} \cdot \mathrm{tr}\left( \mathfrak{P}^{-s_1} X \cdots X \mathfrak{P}^{-s_{2k+1}} \right).$$

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

$$\mathfrak{P}^{-s_1} X \cdots X \mathfrak{P}^{-s_{2k+1}}$$
$$= \mathfrak{P}^{-s_{j_1}} \underbrace{X \mathfrak{P}^{\perp} \cdots \mathfrak{P}^{\perp} X}_{t_1 \text{ of } X} \mathfrak{P}^{-s_{j_2}} \cdots \mathfrak{P}^{-s_{j_{\tau-1}}} \underbrace{X \mathfrak{P}^{\perp} \cdots \mathfrak{P}^{\perp} X}_{t_{\tau-1} \text{ of } X} \mathfrak{P}^{-s_{j_\tau}}$$

where

$$s_{j_1} + \cdots + s_{j_\tau} = 2k \quad \text{and} \quad t_1 + \cdots + t_{\tau-1} = 2k.$$

Therefore, for positive integers $s_1, \cdots, s_{2k+1}, t_1, \cdots, t_{2k} \geq 1$,

$$\langle \Theta\Theta^{\mathsf{T}}, \mathbb{E}\mathcal{S}_{A,2k}(X) \rangle = \sum_{\tau \geq 2} (-1)^{1+\tau} \sum_{s_1 + \cdots + s_\tau = 2k} \sum_{t_1 + \cdots + t_{\tau-1} = 2k} \mathbb{E}\mathrm{tr}\left( Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}} \right)$$

where the matrix $Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}}$ is defined by

$$Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}} = \mathfrak{P}^{-s_1} \underbrace{X \mathfrak{P}^{\perp} \cdots \mathfrak{P}^{\perp} X}_{t_1 \text{ of } X} \mathfrak{P}^{-s_2} \cdots \mathfrak{P}^{-s_{\tau-1}} \underbrace{X \mathfrak{P}^{\perp} \cdots \mathfrak{P}^{\perp} X}_{t_{\tau-1} \text{ of } X} \mathfrak{P}^{-s_\tau}. \quad (26)$$

*Case 1*: if any of $t_1, t_2, \cdots, t_{\tau-1}$ equals one. W.L.O.G., let $t_1 = 1$. Then, $Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}}$ involves the product of $\mathfrak{P}^{-s_1} X \mathfrak{P}^{-s_2}$. Then,

$$\left| \mathbb{E}\mathrm{tr}\left( Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}} \right) \right| \leq \sqrt{2r} \cdot \mathbb{E}\|\mathfrak{P}^{-s_1} X \mathfrak{P}^{-s_2}\|_{\mathrm{F}} \cdot \frac{\|X\|^{2k-1}}{\lambda_r^{2k-s_1-s_2}}$$

$$\leq \sqrt{2r} \cdot \mathbb{E}\|\Theta\Theta^\mathsf{T} X\Theta\Theta^\mathsf{T}\|_\mathrm{F} \cdot \frac{\|X\|^{2k-1}}{\lambda_r^{2k}}$$

$$\leq \frac{\sqrt{2r}}{\lambda_r^{2k}} \cdot \mathbb{E}\|\Theta\Theta^\mathsf{T} X\Theta\Theta^\mathsf{T}\|_\mathrm{F}\|X\|^{2k-1}$$

$$\leq \frac{\sqrt{2r}}{\lambda_r^{2k}} \cdot \mathbb{E}^{1/2}\|\Theta\Theta^\mathsf{T} X\Theta\Theta^\mathsf{T}\|_\mathrm{F}^2 \mathbb{E}^{1/2}\|X\|^{4k-2}$$

$$\leq C_1 \frac{C_2^{k-1} r^{3/2} d_2^{k-\frac{1}{2}}}{\lambda_r^{2k}}$$

where we used the fact $\Theta\Theta^\mathsf{T} X\Theta\Theta^\mathsf{T} = \begin{pmatrix} 0 & UU^\mathsf{T} ZVV^\mathsf{T} \\ VV^\mathsf{T} Z^\mathsf{T} UU^\mathsf{T} & 0 \end{pmatrix}$ which is of rank at most $2r$ and $\mathbb{E}^{1/2}\|U^\mathsf{T} ZV\|_\mathrm{F}^2 = O(r)$. We also used the fact $\mathbb{E}^{1/p}\|X\|^p \leq C_2\sqrt{d_2}$ for some absolute constant $C_2 > 0$ and all positive integers $p \geq 1$. Therefore, if any of $t_1, \cdots, t_{\tau-1}$ equals one, then the magnitude of $\left|\mathbb{E}\mathrm{tr}\big(Q_{t_1 t_2 \cdots t_{\tau-1}}^{(s_1 s_2 \cdots s_\tau)}\big)\right|$ is of the order $O\left(\frac{r^{3/2}}{\sqrt{d_2}} \cdot \frac{C_2^k d_2^k}{\lambda_r^{2k}}\right)$.

*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,

$$\left|\mathbb{E}\big\langle \Theta\Theta^\mathsf{T}, Q_{t_1 t_2 \cdots t_{\tau-1}}^{(s_1 s_2 \cdots s_\tau)}\big\rangle\right|$$

$$\leq \left|\mathrm{tr}\Big(\mathfrak{P}^{-s_1} X\big(\mathfrak{P}^\perp X\big)^{t_1-1}\mathfrak{P}^{-s_2} X\big(\mathfrak{P}^\perp X\big)^{t_2-1}\mathfrak{P}^{-s_3} X\right.$$

$$\left.\cdots \mathfrak{P}^{-s_{\tau-1}} X\big(\mathfrak{P}^\perp X\big)^{t_{\tau-1}-1}\mathfrak{P}^{-s_\tau}\Big)\right|$$

$$\leq \mathbb{E}\left\|\mathfrak{P}^{-s_1} X(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{2p+1} X\mathfrak{P}^{-s_2}\right\|_\mathrm{F} \cdot \frac{\sqrt{2r}\|X\|^{2k-t_1}}{\lambda_r^{2k-s_1-s_2}}$$

$$\leq \mathbb{E}\left\|\mathfrak{P}^{-s_1} X(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{2p+1} X\mathfrak{P}^{-s_2}\right\|_\mathrm{F} \cdot \frac{\sqrt{2r}\|X\|^{2k-t_1}}{\lambda_r^{2k-s_1-s_2}}\mathbb{I}_{\mathcal{E}_1}$$

$$+ \mathbb{E}\left\|\mathfrak{P}^{-s_1} X(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{2p+1} X\mathfrak{P}^{-s_2}\right\|_\mathrm{F} \cdot \frac{\sqrt{2r}\|X\|^{2k-t_1}}{\lambda_r^{2k-s_1-s_2}}\mathbb{I}_{\mathcal{E}_1^c}$$

where, as in the proof of Theorem 2, define the event $\mathcal{E}_1 = \{\|X\| \leq C_2 \cdot \sqrt{d_2}\}$ for some absolute constant $C_2 > 0$ such that $\mathbb{P}(\mathcal{E}_1) \geq 1 - e^{-c_1 d_2}$. As a result, we get

$$\left|\mathbb{E}\big\langle \Theta\Theta^\mathsf{T}, Q_{t_1 t_2 \cdots t_{\tau-1}}^{(s_1 s_2 \cdots s_\tau)}\big\rangle\right|$$

$$\leq \mathbb{E}\left\|\mathfrak{P}^{-s_1} X(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{2p+1} X\mathfrak{P}^{-s_2}\right\|_\mathrm{F} \cdot \frac{\sqrt{2r}d_2^{(2k-t_1)/2}}{\lambda_r^{2k-s_1-s_2}}\mathbb{I}_{\mathcal{E}_1} + C_2^{2k} \cdot \frac{rd_2^k}{\lambda_r^{2k}} \cdot e^{-c_1 d_2}$$

$$\leq \mathbb{E}^{1/2}\left\|\mathfrak{P}^{-s_1} X(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{2p+1} X\mathfrak{P}^{-s_2}\right\|_\mathrm{F}^2 \cdot \frac{\sqrt{2r}d_2^{(2k-t_1)/2}}{\lambda_r^{2k-s_1-s_2}} + C_2^{2k} \cdot \frac{rd_2^k}{\lambda_r^{2k}} \cdot e^{-c_1 d_2}$$

$$\leq \frac{C\sqrt{r}d_2^{(2k-t_1)/2}}{\lambda_r^{2k}} \cdot \mathbb{E}^{1/2}\left\|\Theta^\mathsf{T} X(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{2p+1} X\Theta\right\|_\mathrm{F}^2 + C_2^{2k} \cdot \frac{rd_2^k}{\lambda_r^{2k}} \cdot e^{-c_1 d_2}$$

where $\Theta = (\theta_1, \cdots, \theta_r, \theta_{-r}, \cdots, \theta_{-1}) \in \mathbb{R}^{(d_1+d_2)\times(2r)}$. In addition, we can write

$$\mathbb{E}\left\|\Theta^\mathsf{T} X(\mathfrak{P}^\perp X \mathfrak{P}^\perp)^{2p+1} X\Theta\right\|_\mathrm{F}^2 = \sum_{1\le|j_1|,|j_2|\le r} \mathbb{E}\big(\theta_{j_1}^\mathsf{T} X(\mathfrak{P}^\perp X \mathfrak{P}^\perp)^{2p+1} X\theta_{j_2}\big)^2.$$

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

$$(\mathfrak{P}^\perp X \mathfrak{P}^\perp)^{2p} = \begin{pmatrix} \big(U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T}\big)^p & 0 \\ 0 & \big(V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T}\big)^p \end{pmatrix}.$$

W.L.O.G, let $j_1, j_2 \ge 1$. Then, we write

$$\theta_{j_1}^\mathsf{T} X(\mathfrak{P}^\perp X \mathfrak{P}^\perp)^{2p+1} X\theta_{j_2}$$
$$= \frac{1}{2} v_{j_1}^\mathsf{T} Z^\mathsf{T} \big(U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T}\big)^p U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} u_{j_2}$$
$$+ \frac{1}{2} u_{j_1}^\mathsf{T} Z\big(V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T}\big)^p V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Zv_{j_2}$$

and get the simple bound

$$\mathbb{E}\big(\theta_{j_1}^\mathsf{T} X(\mathfrak{P}^\perp X \mathfrak{P}^\perp)^{2p+1} X\theta_{j_2}\big)^2$$
$$\le 2^{-1}\mathbb{E}\Big(v_{j_1}^\mathsf{T} Z^\mathsf{T} \big(U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T}\big)^p U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} u_{j_2}\Big)^2$$
$$+ 2^{-1}\mathbb{E}\Big(u_{j_1}^\mathsf{T} Z\big(V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T}\big)^p V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Zv_{j_2}\Big)^2.$$

Observe that $Zv_{j_1}$ is independent with $ZV_\perp$ and $Z^\mathsf{T} u_{j_1}$ is independent with $Z^\mathsf{T} U_\perp$. Therefore,

$$\mathbb{E}\big(\theta_{j_1}^\mathsf{T} X(\mathfrak{P}^\perp X \mathfrak{P}^\perp)^{2p+1} X\theta_{j_2}\big)^2$$
$$\le 2^{-1}\mathbb{E}\left\|\big(U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T}\big)^p U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} u_{j_2}\right\|_{\ell_2}^2$$
$$+ 2^{-1}\mathbb{E}\left\|\big(V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T}\big)^p V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Zv_{j_2}\right\|_{\ell_2}^2$$
$$\le 2^{-1}\mathbb{E}\left\|\big(U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T}\big)^p U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T}\right\|_{\ell_2}^2$$
$$+ 2^{-1}\mathbb{E}\left\|\big(V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} ZV_\perp V_\perp^\mathsf{T}\big)^p V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T}\right\|_{\ell_2}^2$$
$$\le C_2^{4p+2} d_2^{2p+1} = C_2^{4p+2} d_2^{t_1-2},$$

where the last inequality is due to the independence between $Z^\mathsf{T} u_{j_2}$ and $Z^\mathsf{T} U_\perp$, the independence between $Zv_{j_2}$ and $ZV_\perp$. We conclude that

$$\big|\mathbb{E}\langle \Theta\Theta^\mathsf{T}, Q_{t_1 t_2 \cdots t_{\tau-1}}^{(s_1 s_2 \cdots s_\tau)}\rangle\big| \le C_2^{2k} \cdot \frac{r^{3/2} d_2^{k-1}}{\lambda_r^{2k}} + r\Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k \cdot e^{-c_1 d_2} \le \frac{r^{3/2}}{d_2} \cdot \Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k$$

where $C_2 > 0$ is some absolute constant and the last inequality is due to $e^{-c_1 d_2} \leq d_2^{-1}$.

We now finalize the proof. If there exists one odd $t_i$, then there exists at least another $t_j$ which is also odd since the sum of $t_i$s is even. Following the same analysis, we conclude

$$\left| \mathbb{E} \langle \Theta\Theta^\mathsf{T}, Q_{t_1 t_2 \cdots t_{\tau-1}}^{(s_1 s_2 \cdots s_\tau)} \rangle \right| \leq \frac{r^2}{d_2} \cdot \left( \frac{C_2 d_2}{\lambda_r^2} \right)^k$$

whenever any of $t_1, \cdots, t_{\tau-1}$ is an odd number. Therefore, 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 $\mathbb{E} \langle \Theta\Theta^\mathsf{T}, \mathcal{S}_{A,4}(X) \rangle$, it suffices to calculate

$$\sum_{\tau=2}^{3} (-1)^{1+\tau} \sum_{s_1 + \cdots + s_\tau = 4} \sum_{t_1 + \cdots + t_{\tau-1} = 4} \mathbb{E} \langle \Theta\Theta^\mathsf{T}, Q_{t_1 t_2 \cdots t_{\tau-1}}^{(s_1 s_2 \cdots s_\tau)} \rangle$$

where $t_1, \cdots, t_{\tau-1}$ are positive even numbers and $s_1, \cdots, s_\tau$ 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

$$Q_4^{(s_1 s_2)} = \mathfrak{P}^{-s_1} X (\mathfrak{P}^\perp X \mathfrak{P}^\perp)^2 X \mathfrak{P}^{-s_2}$$
$$= \mathbb{E}\text{tr}\big(Q_4^{(s_1 s_2)}\big) = \mathbb{E}\text{tr}\big(\Theta\Theta^\mathsf{T} X (\mathfrak{P}^\perp X \mathfrak{P}^\perp)^2 X \Theta\Theta^\mathsf{T} \mathfrak{P}^{-4}\big).$$

Clearly, we have

$$\Theta\Theta^\mathsf{T} X (\mathfrak{P}^\perp X \mathfrak{P}^\perp)^2 X \Theta\Theta^\mathsf{T}$$
$$= \begin{pmatrix} U U^\mathsf{T} Z V_\perp V_\perp Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Z V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U U^\mathsf{T} & 0 \\ 0 & V V^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Z V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Z V V^\mathsf{T} \end{pmatrix}.$$

By the independence between $U^\mathsf{T} Z$ and $U_\perp^\mathsf{T} Z$, independence between $V^\mathsf{T} Z^\mathsf{T}$ and $V_\perp^\mathsf{T} Z^\mathsf{T}$, we immediately obtain

$$\mathbb{E}\Theta\Theta^\mathsf{T} X (\mathfrak{P}^\perp X \mathfrak{P}^\perp)^2 X \Theta\Theta^\mathsf{T}$$
$$= \mathbb{E} \begin{pmatrix} d_{1-} U U^\mathsf{T} Z V_\perp V_\perp^\mathsf{T} Z^\mathsf{T} U U^\mathsf{T} & 0 \\ 0 & d_{2-} V V^\mathsf{T} Z^\mathsf{T} U_\perp U_\perp^\mathsf{T} Z V V^\mathsf{T} \end{pmatrix} = d_{1-} d_{2-} \Theta\Theta^\mathsf{T}$$

where $d_{1-} = d_1 - r$ and $d_{2-} = d_2 - r$. Then,

$$\mathbb{E} \langle \Theta\Theta^\mathsf{T}, Q_4^{(s_1 s_2)} \rangle = 2 d_{1-} d_{2-} \|\Lambda^{-2}\|_{\mathrm{F}}^2$$

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 $(s_1, s_2, s_3) \in \big\{(1,1,2),(1,2,1),(2,1,1)\big\}$. W.L.O.G., consider $s_1 = 1, s_2 = 1, s_3 = 2$, we have

$$Q_{22}^{(112)} = \mathfrak{P}^{-1} X \mathfrak{P}^{\perp} X \mathfrak{P}^{-1} X \mathfrak{P}^{\perp} X \mathfrak{P}^{-2}.$$

Similarly, we can write

$$\mathbb{E}\mathrm{tr}(Q_{22}^{(112)}) = \mathbb{E}\mathrm{tr}\big(U\Lambda^{-1}V^{\mathsf{T}}Z^{\mathsf{T}}U_{\perp}U_{\perp}^{\mathsf{T}}ZV\Lambda^{-1}U^{\mathsf{T}}ZV_{\perp}V_{\perp}^{\mathsf{T}}Z^{\mathsf{T}}U\Lambda^{-2}U^{\mathsf{T}}\big)$$
$$+\mathbb{E}\mathrm{tr}\big(V\Lambda^{-1}U^{\mathsf{T}}ZV_{\perp}V_{\perp}^{\mathsf{T}}Z^{\mathsf{T}}U\Lambda^{-1}V^{\mathsf{T}}Z^{\mathsf{T}}U_{\perp}U_{\perp}^{\mathsf{T}}ZV\Lambda^{-2}V^{\mathsf{T}}\big)$$
$$=d_{2-}\mathbb{E}\mathrm{tr}\big(U\Lambda^{-1}V^{\mathsf{T}}Z^{\mathsf{T}}U_{\perp}U_{\perp}^{\mathsf{T}}ZV\Lambda^{-3}U^{\mathsf{T}}\big)+d_{1-}\mathbb{E}\mathrm{tr}\big(V\Lambda^{-1}U^{\mathsf{T}}ZV_{\perp}V_{\perp}^{\mathsf{T}}Z^{\mathsf{T}}U\Lambda^{-3}V^{\mathsf{T}}\big)$$
$$=2d_{1-}d_{2-}\|\Lambda^{-2}\|_{\mathrm{F}}^2.$$

By symmetricity, the same equation holds for $\mathbb{E}\mathrm{tr}(Q_{22}^{(211)})$. Next, we consider $(s_1, s_2, s_3) = (1,2,1)$. We will write

$$\mathbb{E}\mathrm{tr}(Q_{22}^{(121)}) = \mathbb{E}\mathrm{tr}\big(U\Lambda^{-1}V^{\mathsf{T}}Z^{\mathsf{T}}U_{\perp}U_{\perp}^{\mathsf{T}}ZV\Lambda^{-2}V^{\mathsf{T}}Z^{\mathsf{T}}U_{\perp}U_{\perp}^{\mathsf{T}}ZV\Lambda^{-1}U^{\mathsf{T}}\big)$$
$$+\mathbb{E}\mathrm{tr}\big(V\Lambda^{-1}U^{\mathsf{T}}ZV_{\perp}V_{\perp}^{\mathsf{T}}Z^{\mathsf{T}}U\Lambda^{-2}U^{\mathsf{T}}ZV_{\perp}V_{\perp}^{\mathsf{T}}Z^{\mathsf{T}}U\Lambda^{-1}V^{\mathsf{T}}\big)$$
$$=\mathbb{E}\|\Lambda^{-1}\tilde{Z}_1\tilde{Z}_1^{\mathsf{T}}\Lambda^{-1}\|_{\mathrm{F}}^2 + \mathbb{E}\|\Lambda^{-1}\tilde{Z}_2\tilde{Z}_2^{\mathsf{T}}\Lambda^{-1}\|_{\mathrm{F}}^2$$

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 6 in the Appendix, we obtain

$$\mathbb{E}\mathrm{tr}\big(Q_{22}^{(121)}\big) = (d_{1-}^2 + d_{2-}^2)\|\Lambda^{-2}\|_{\mathrm{F}}^2 + (d_{1-} + d_{2-})\big(\|\Lambda^{-2}\|_{\mathrm{F}}^2 + \|\Lambda^{-1}\|_{\mathrm{F}}^4\big).$$

Therefore, we conclude that

$$\left| -\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,4}(X) \rangle + (d_{1-} - d_{2-})^2 \|\Lambda^{-2}\|_{\mathrm{F}}^2 \right| \le C_1 \cdot \frac{r^2 d_2}{\lambda_r^4}$$

for some absolute constant $C_1 > 0$ where we also include those smaller terms when some $t_i$ is odd as discussed in **Property 1**. Together with the proof of Lemma 1, we conclude that

$$\left| \mathbb{E}\|\hat{\Theta}\hat{\Theta}^{\mathsf{T}} - \Theta\Theta^{\mathsf{T}}\|_{\mathrm{F}}^2 - 2\big(d_{\star}\|\Lambda^{-1}\|_{\mathrm{F}}^2 - \Delta_d^2\|\Lambda^{-2}\|_{\mathrm{F}}^2\big) \right| \le C_1 \cdot \frac{r^2 d_2}{\lambda_r^4} + C_2 \cdot \frac{r d_2^3}{\lambda_r^6}$$

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

### A.4. Proof of Lemma 3.

To characterize $\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X) \rangle$ more easily, we observe the following property.

**Property 2: effect from distinct singular values are negligible.** Recall that

$$\mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X)\rangle = \sum_{\mathbf{s}: s_1 + \cdots + s_{2k+1} = 2k} (-1)^{1+\tau(\mathbf{s})} \cdot \mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathfrak{P}^{-s_1} X \cdots X \mathfrak{P}^{-s_{2k+1}}\rangle.$$

As proved in **Property 1**, we have

$$\left| \mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X)\rangle - \sum_{\tau \geq 2} (-1)^{1+\tau} \sum_{s_1 + \cdots + s_\tau = 2k} \sum_{t_1 + \cdots + t_{\tau-1} = 2k} \mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}}\rangle \right|$$

$$\leq \frac{r^2}{d_2} \cdot \left(\frac{C_2 d_2}{\lambda_r^2}\right)^k$$

where the matrix $Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}}$ is defined as in (26) and $t_1, \cdots, t_{\tau-1}$ are positive even numbers. Recall that $\Theta\Theta^{\mathsf{T}} = \sum_{j=1}^r (P_j + P_{-j})$ and $\mathfrak{P}^{-s_1} = \sum_{j=1}^r \left[\lambda_j^{-s_1} P_j + (\lambda_{-j})^{-s_j} P_{-j}\right]$ where $\lambda_{-j} = -\lambda_j$. For each fixed $(s_1, \cdots, s_\tau)$ and $(t_1, \cdots, t_{\tau-1})$ where $t_j$s are even numbers, we write

$$\langle \Theta\Theta^{\mathsf{T}}, Q^{(s_1 s_2 \cdots s_\tau)}_{t_1 t_2 \cdots t_{\tau-1}}\rangle$$

$$= \sum_{|j_1|, |j_2|, \cdots, |j_{\tau-1}| \geq 1}^r \lambda_{j_1}^{-(s_1 + s_\tau)} \lambda_{j_2}^{-s_2} \cdots \lambda_{j_{\tau-1}}^{-s_{\tau-1}} (\theta_{j_1}^{\mathsf{T}} W_{t_1} \theta_{j_2})(\theta_{j_2}^{\mathsf{T}} W_{t_2} \theta_{j_3})$$

$$\cdots (\theta_{j_{\tau-1}}^{\mathsf{T}} W_{t_{\tau-1}} \theta_{j_1})$$

where the matrix $W_{t_1} = \underbrace{X \mathfrak{P}^{\perp} X \mathfrak{P}^{\perp} \cdots \mathfrak{P}^{\perp} X}_{t_1 \text{ of } X}$ for positive even numbers $t_1$. Observe that

$$\theta_{j_1}^{\mathsf{T}} W_{t_1} \theta_{j_2} = \theta_{j_1}^{\mathsf{T}} W_{t_1} \theta_{j_2} = \theta_{j_1}^{\mathsf{T}} X (\mathfrak{P}^{\perp} X \mathfrak{P}^{\perp})^{t_1 - 2} X \theta_{j_2}.$$

We show that if there exists $1 \leq k_0 \leq \tau - 1$ so that $|j_{k_0}| \neq |j_{k_0+1}|$, then $|\theta_{j_{k_0}}^{\mathsf{T}} W_{t_{k_0}} \theta_{j_{k_0+1}}|$ is a negligibly smaller term. W.L.O.G., assume $|j_1| \neq |j_2|$ and then

$$\left| \mathbb{E} \sum_{\substack{|j_1|, |j_2|, \cdots, |j_{\tau-1}| \geq 1 \\ |j_1| \neq |j_2|}}^r \lambda_{j_1}^{-(s_1 + s_\tau)} \lambda_{j_2}^{-s_2} \right.$$

$$\left. \cdots \lambda_{j_{\tau-1}}^{-s_{\tau-1}} (\theta_{j_1}^{\mathsf{T}} W_{t_1} \theta_{j_2})(\theta_{j_2}^{\mathsf{T}} W_{t_2} \theta_{j_3}) \cdots (\theta_{j_{\tau-1}}^{\mathsf{T}} W_{t_{\tau-1}} \theta_{j_1}) \right|$$

$$= \left| \mathbb{E} \sum_{|j_1| \neq |j_2|} \lambda_{j_1}^{-(s_1 + s_\tau)} \lambda_{j_2}^{-s_2} (\theta_{j_1}^{\mathsf{T}} W_{t_1} \theta_{j_2}) \theta_{j_2}^{\mathsf{T}} W_{t_2} \mathfrak{P}^{-s_3} W_{t_3} \mathfrak{P}^{-s_4} \cdots \mathfrak{P}^{-s_{\tau-1}} W_{t_{\tau-1}} \theta_{j_1} \right|$$

$$\leq \frac{1}{\lambda_r^{2k}} \sum_{|j_1| \neq |j_2|} \mathbb{E} \left| \theta_{j_1}^{\mathsf{T}} W_{t_1} \theta_{j_2} \right| \|X\|^{2k - t_1}$$

Since $\theta_{j_1}$ and $\theta_{j_2}$ are orthogonal, we conclude that $X\theta_{j_1}$ and $X\theta_{j_2}$ are independent normal vectors from which we get that $\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2}|\mathfrak{P}^\perp X\mathfrak{P}^\perp$ is subexponential and $\mathbb{E}|\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2}| = O\big(\|(\mathfrak{P}^\perp X\mathfrak{P}^\perp)^{t_1-2}\|_\mathrm{F}\big)$. Therefore, we get

$$
\begin{aligned}
&\mathbb{E}\big|\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2}\big|\|X\|^{2k-t_1}\\
={}&\mathbb{E}\big|\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2}\big|\|X\|^{2k-t_1}\mathbf{1}(\|X\| \le C_1\sqrt{d_2})\\
&+\mathbb{E}\big|\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2}\big|\|X\|^{2k-t_1}\mathbf{1}(\|X\| \ge C_1\sqrt{d_2})\\
\le{}&\mathbb{E}^{1/2}\big|\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2}\big|^2 \cdot (C_1^2 d_2)^{k-t_1/2}\mathbf{1}(\|X\| \le C_1 d_2^{1/2}) + e^{-d_2/2}(C_1 d_2)^k\\
\lesssim{}&\frac{1}{\sqrt{d_2}}\cdot(C_2 d_2)^k + e^{-d_2/2}(C_2 d_2)^k.
\end{aligned}
$$

As a result, we conclude that

$$
\begin{aligned}
\Bigg|\mathbb{E}\sum_{|j_1|,|j_2|,\cdots,|j_{\tau-1}|\ge 1}^{r} &\lambda_{j_1}^{-(s_1+s_\tau)}\lambda_{j_2}^{-s_2}\\
&\cdots\lambda_{j_{\tau-1}}^{-s_{\tau-1}}(\theta_{j_1}^\mathsf{T} W_{t_1}\theta_{j_2})(\theta_{j_2}^\mathsf{T} W_{t_2}\theta_{j_3})\cdots(\theta_{j_{\tau-1}}^\mathsf{T} W_{t_{\tau-1}}\theta_{j_1})\Bigg|
\end{aligned}
$$

$$
\le \frac{C_1 r^2}{\sqrt{d_2}}\cdot\Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k + C_3 e^{-d_2/2}\cdot\Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k \le \frac{C_1 r^2}{\sqrt{d_2}}\cdot\Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k
$$

for some absolute constants $C_1, C_2 > 0$.

It suggests that the dominating terms come from those tuples $(j_1, j_2, \cdots, j_{\tau-1})$ such that $|j_1| = |j_2| = \cdots = |j_{\tau-1}|$. Now, we define $\mathfrak{P}_j = \lambda_j P_j + \lambda_{-j} P_{-j}$. To this end, we conclude

$$
\begin{aligned}
\Bigg|\mathbb{E}\big\langle\Theta\Theta^\mathsf{T}, \mathcal{S}_{A,2k}(X)\big\rangle\\
-\sum_{j=1}^{r}\sum_{\tau\ge 2}(-1)^{1+\tau}\sum_{\substack{\mathbf{s}:s_1+\cdots+s_\tau=2k,s_1,s_\tau>0\\\mathbf{t}:t_1+\cdots+t_{\tau-1}=2k}}\mathbb{E}\mathrm{tr}\big(\mathfrak{P}_j^{-s_1} W_{t_1}\mathfrak{P}_j^{-s_2} W_{t_2}\cdots\mathfrak{P}_j^{-s_\tau}\big)\Bigg|
\end{aligned}
$$

$$
\le\frac{C_1 r^2}{\sqrt{d_2}}\cdot\Big(\frac{C_2 d_2}{\lambda_r^2}\Big)^k \tag{27}
$$

for some absolute constants $C_1, C_2 > 0$. The above fact suggests that it suffices to focus on the effect from individual singular values (i.e., for any fixed $1 \le j \le r$). Moreover, it is easy to check that

$$
\mathfrak{P}_j^{-s_1} W_{t_1}\mathfrak{P}_j^{-s_2} W_{t_2}\cdots\mathfrak{P}_j^{-s_\tau} = \frac{1}{\lambda_j^{2k}}\cdot\tilde{\mathfrak{P}}_j^{-s_1} W_{t_1}\tilde{\mathfrak{P}}_j^{-s_2} W_{t_2}\cdots\tilde{\mathfrak{P}}_j^{-s_\tau}
$$

where $\tilde{\mathfrak{P}}_j^{-s} = P_j + (-1)^s P_{-j}$ implying that the $k$-th order error term has dominator $\lambda_j^{2k}$. To this end, we prove the following lemma in the Appendix.

**Lemma 5.** *For any $1 \leq j \leq r$ and $k \geq 2$, we obtain*

$$\left| \sum_{\tau \geq 2} (-1)^{1+\tau} \sum_{\substack{\mathbf{s}: s_1 + \cdots + s_\tau = 2k, s_1, s_\tau > 0 \\ \mathbf{t}: t_1 + \cdots + t_{\tau-1} = 2k}} \mathbb{E}\mathrm{tr}\left(\mathfrak{P}_j^{-s_1} W_{t_1} \mathfrak{P}_j^{-s_2} \cdots \mathfrak{P}_j^{-s_\tau}\right) \right.$$

$$\left. - \frac{(-1)^k (d_{1-}^{k-1} - d_{2-}^{k-1})(d_{1-} - d_{2-})}{\lambda_j^{2k}} \right|$$

$$\leq \frac{C_1 k}{\sqrt{d_2}} \cdot \left(\frac{C_2 d_2}{\lambda_r^2}\right)^k$$

*for some absolute constants $C_1, C_2 > 0$.*

By Lemma 5 and (27), it holds for all $k \geq 2$ that

$$\left| \mathbb{E}\langle \Theta\Theta^{\mathsf{T}}, \mathcal{S}_{A,2k}(X) \rangle - (-1)^k (d_{1-}^{k-1} - d_{2-}^{k-1})(d_{1-} - d_{2-})\|\Lambda^{-k}\|_{\mathrm{F}}^2 \right|$$

$$\leq \frac{C_1(r^2 + k)}{\sqrt{d_2}} \cdot \left(\frac{C_2 d_2}{\lambda_r^2}\right)^k$$

for some absolute constants $C_1, C_2 > 0$, which concludes the proof.

### A.5. Proof of CLT theorems in Section 5

**Proof of Theorem 3** Recall Theorem 2, we end up with

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

$$\leq C_2\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\right) \cdot \sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}} + e^{-c_1 d_2} + C_2\left(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_2}} + e^{-\lambda_r/\sqrt{rd_2}}.$$

By Lemma 1, we get

$$\left| \mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2 \right| \leq C_2 \frac{rd_2^2}{\lambda_r^4}.$$

Therefore,

$$\left| \frac{\mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}} \right| \leq C_2 \frac{rd_2^{3/2}}{\lambda_r^2}.$$

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

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left( \frac{\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - 2d_\star\|\Lambda^{-1}\|_{\mathrm{F}}^2}{\sqrt{8d_\star}\|\Lambda^{-2}\|_{\mathrm{F}}} \leq x \right) - \Phi(x) \right|$$

$$\leq C_2\left(\frac{\sqrt{r}}{\|\Lambda^{-2}\|_{\mathrm{F}}\lambda_r^2}\right) \cdot \sqrt{\frac{(rd_2)^{1/2}}{\lambda_r}}$$

$$+ e^{-c_1 d_2} + C_2\left(\frac{\|\Lambda^{-1}\|_{\mathrm{F}}^4}{\|\Lambda^{-2}\|_{\mathrm{F}}^2}\right)^{3/2} \cdot \frac{1}{\sqrt{d_2}} + C_3 \frac{rd_2^{3/2}}{\lambda_r^2} + e^{-\lambda_r/\sqrt{rd_2}}.$$

**Proof of Theorem 4**  By Lemma 3, we have

$$\left| \mathbb{E}\mathrm{dist}^2[(\hat{U}, \hat{V}), (U, V)] - B_k \right| \leq C_4 \frac{r^2 d_2}{\lambda_r^4} + \frac{C_5 r^2}{\sqrt{d_2}} \cdot \left( \frac{d_2}{\lambda_r^2} \right)^3 + C_6 r \left( \frac{C_3 d_2}{\lambda_r^2} \right)^{k+1}.$$

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

## Appendix B: Appendix

### B.1. Supporting lemmas

*Proof of Lemma 4.* Recall that

$$f_t(X_1) = \sum_{k \geq 3} \langle \Theta\Theta^\mathsf{T}, \mathcal{S}_{A,k}(X_1) \rangle \phi \left( \frac{\|X_1\|}{t \cdot \sqrt{d_2}} \right).$$

*Case 1*: if $\|X_1\| > 2t\sqrt{d_2}$ and $\|X_2\| > 2t\sqrt{d_2}$, then $f_t(X_1) = f_t(X_2) = 0$ by definition of $\phi(\cdot)$ where the claimed inequality holds trivially.
*Case 2*: if $\|X_1\| \leq 2t\sqrt{d_2}$ and $\|X_2\| > 2t\sqrt{d_2}$, then $f_t(X_2) = 0$. We get, by Lipschitz property of $\phi(\cdot)$, that

$$\left| f_t(X_1) - f_t(X_2) \right| = \left| \sum_{k \geq 3} \langle \Theta\Theta^\mathsf{T}, \mathcal{S}_{A,k}(X_1) \rangle \cdot \left( \phi \left( \frac{\|X_1\|}{t \cdot \sqrt{d_2}} \right) - \phi \left( \frac{\|X_2\|}{t \cdot \sqrt{d_2}} \right) \right) \right|$$

$$\leq \sum_{k \geq 3} 2r \left\| \mathcal{S}_{A,k}(X_1) \right\| \cdot \frac{\|X_1 - X_2\|_\mathrm{F}}{t \cdot \sqrt{d_2}}$$

$$\leq \frac{2r\|X_1 - X_2\|_\mathrm{F}}{t \cdot \sqrt{d_2}} \cdot \sum_{k \geq 3} \sum_{\mathbf{s}: s_1 + \cdots + s_{k+1} = k} \left\| \mathfrak{P}^{-s_1} X_1 \mathfrak{P}^{-s_2} X_1 \cdots X_1 \mathfrak{P}^{-s_{k+1}} \right\|$$

$$\leq \frac{2r\|X_1 - X_2\|_\mathrm{F}}{t \cdot \sqrt{d_2}} \cdot \sum_{k \geq 3} \sum_{\mathbf{s}: s_1 + \cdots + s_{k+1} = k} \frac{\|X_1\|^k}{\lambda_r^k}$$

$$\leq \frac{2r\|X_1 - X_2\|_\mathrm{F}}{t \cdot \sqrt{d_2}} \cdot \sum_{k \geq 3} \left( \frac{4\|X_1\|}{\lambda_r} \right)^k$$

$$\leq C_4 t^2 \frac{r\|X_1 - X_2\|_\mathrm{F}}{\sqrt{d_2}} \cdot \frac{d_2^{3/2}}{\lambda_r^3}$$

where the last inequality holds as long as $\lambda_r \geq 9t\sqrt{d_2}$.
*Case 3*: if $\|X_1\| \leq 2t d_2^{1/2}$ and $\|X_2\| \leq 2t d_2^{1/2}$. Then,

$$\left| f_t(X_1) - f_t(X_2) \right| \leq 2r \sum_{k \geq 3} \left\| \mathcal{S}_{A,k}(X_1) \phi \left( \frac{\|X_1\|}{t \cdot \sqrt{d_2}} \right) - \mathcal{S}_{A,k}(X_2) \phi \left( \frac{\|X_2\|}{t \cdot \sqrt{d_2}} \right) \right\|$$

$$\leq 2r \sum_{k \geq 3} \sum_{\mathbf{s}: s_1 + \cdots + s_{k+1} = k} \left\| \mathfrak{P}^{-s_1} X_1 \cdots X_1 \mathfrak{P}^{-s_{k+1}} \phi \left( \frac{\|X_1\|}{t \cdot \sqrt{d_2}} \right) - \mathfrak{P}^{-s_1} X_2 \right.$$

$$\cdots X_2 \mathfrak{P}^{-s_{k+1}} \phi\Big(\frac{\|X_2\|}{t \cdot \sqrt{d_2}}\Big)\Big\|$$

$$\leq 2r \sum_{k \geq 3} \sum_{\mathbf{s}: s_1 + \cdots + s_{k+1} = k} (k+2) \cdot \frac{(2td_2^{1/2})^{k-1}}{\lambda_r^k} \|X_1 - X_2\|_{\mathrm{F}}$$

$$\leq C_4 t^2 \cdot \frac{rd_2}{\lambda_r^3} \|X_1 - X_2\|_{\mathrm{F}}$$

where the last inequality holds as long as $\lambda_r \geq 9t\sqrt{d_2}$. Therefore, we conclude the proof of Lemma 4. □

*Proof of Lemma 5.* Based on **Property 2** and eq. (27), it suffices to calculate the quantities $\mathbb{E}\mathrm{tr}\big(\mathfrak{P}_j^{-s_1} W_{t_1} \mathfrak{P}_j^{-s_2} W_{t_2} \cdots \mathfrak{P}_j^{-s_\tau}\big)$ which relies on singular values $\lambda_j$ and singular vectors $u_j, v_j$ only. Moreover, the actual forms of $u_j, v_j$ does not affect the values. By choosing $\{u_j\}_{j=1}^r$ and $\{v_j\}_{j=1}^r$ as the first $r$ canonical basis vectors in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, it is easy to check that we can reduce the calculations to the rank-one spiked model with singular value $\lambda_j$. To leverage the dimensionality effect where $U_\perp^{\mathsf{T}} Z V_\perp \in \mathbb{R}^{d_{1-} \times d_{2-}}$ has i.i.d. standard normal entries, we consider the rank-one spiked model with

$$\hat{M} = \lambda(u \otimes v) + Z \in \mathbb{R}^{(d_{1-}+1) \times (d_{2-}+1)} \tag{28}$$

where $Z$ has i.i.d. standard normal entries and $d_{1-} = d_1 - r, d_{2-} = d_2 - r$. Let $\hat{u}$ and $\hat{v}$ denote the leading left and right singular vectors of $\hat{M}$. By fact (27), it suffices to calculate the $k$-th order approximation of $\|\hat{u}\hat{u}^{\mathsf{T}} - uu^{\mathsf{T}}\|_{\mathrm{F}}^2 + \|\hat{v}\hat{v}^{\mathsf{T}} - vv^{\mathsf{T}}\|_{\mathrm{F}}^2$. In the proof, we calculate the errors $\|\hat{u}\hat{u}^{\mathsf{T}} - uu^{\mathsf{T}}\|_{\mathrm{F}}^2$ and $\|\hat{v}\hat{v}^{\mathsf{T}} - vv^{\mathsf{T}}\|_{\mathrm{F}}^2$ separately. W.L.O.G., we just deal with $\|\hat{u}\hat{u}^{\mathsf{T}} - uu^{\mathsf{T}}\|_{\mathrm{F}}^2$ and consider $d_1 \leq d_2$[3].

Recall that we aim to calculate the $k$-th order error term in $\|uu^{\mathsf{T}} - \hat{u}\hat{u}^{\mathsf{T}}\|_{\mathrm{F}}^2$. To this end, we write the error terms as

$$\mathbb{E}\|\hat{u}\hat{u}^{\mathsf{T}} - uu^{\mathsf{T}}\|_{\mathrm{F}}^2 = 2 \sum_{k=1}^{\infty} \frac{E_{2k}}{\lambda^{2k}}. \tag{29}$$

We show that $E_{2k} = (-1)^k d_{1-}^{k-1}(d_{1-} - d_{2-}) \cdot \Big[1 + O\Big(\frac{C_1^k}{\sqrt{d_2}}\Big)\Big]$ for some absolute constant $C_1 > 0$. To this end, we consider the second-order (see (Xia and Zhou, 2019)) moment trick (denote $T = \lambda^2(u \otimes u)$)

$$\hat{M}\hat{M}^{\mathsf{T}} = \lambda^2(u \otimes u) + \Delta \in \mathbb{R}^{(d_{1-}+1) \times (d_{1-}+1)} \tag{30}$$

where $\Delta = \lambda uv^{\mathsf{T}} Z^{\mathsf{T}} + \lambda Z vu^{\mathsf{T}} + Z Z^{\mathsf{T}}$. By eq. (4), we can write

$$\|\hat{u}\hat{u}^{\mathsf{T}} - uu^{\mathsf{T}}\|_{\mathrm{F}}^2 = -2 \sum_{k \geq 2} \big\langle uu^\top, \mathcal{S}_{T,k}(\Delta) \big\rangle$$

---

[3]This condition just simplifies our calculation when dealing with the Marchenko Pastur law. Our results do not rely on the condition $d_1 \leq d_2$

where we define $\mathfrak{P}_u = \lambda(u \otimes u)$ and $\mathfrak{P}_u^0 = \mathfrak{P}_u^\perp = U_\perp U_\perp^\top \in \mathbb{R}^{(d_{1-}+1)\times d_{1-}}$ and

$$\mathcal{S}_{T,k}(\Delta) = \sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k} (-1)^{\tau(\mathbf{s})+1} \cdot \mathfrak{P}_u^{-s_1} \Delta \mathfrak{P}_u^{s_2} \Delta \cdots \Delta \mathfrak{P}_u^{-s_{k+1}}.$$

Now, we investigate $\langle uu^\top, \mathcal{S}_{T,k}(\Delta) \rangle$ for all $k \geq 2$. Denote $W_{t_1} = \underbrace{\Delta \mathfrak{P}_u^\perp \Delta \cdots \mathfrak{P}_u^\perp \Delta}_{t_1 \text{ of } \Delta}$

and we can write

$$\langle uu^\top, \mathcal{S}_{T,k}(\Delta) \rangle$$

$$= \sum_{\tau=2}^{k} (-1)^{\tau+1} \sum_{\substack{t_1+\cdots+t_{\tau-1}=k,t_j\geq 1 \\ s_1+\cdots+s_\tau=k,s_j\geq 1}} \mathrm{tr}\big(\mathfrak{P}_u^{-s_1} W_{t_1} \mathfrak{P}_u^{-s_2} W_{t_2} \cdots \mathfrak{P}_u^{-s_{\tau-1}} W_{t_{\tau-1}} \mathfrak{P}_u^{-s_\tau}\big)$$

$$= \frac{1}{\lambda^{2k}} \sum_{\tau=2}^{k} (-1)^{1+\tau} \binom{k-1}{\tau-1} \sum_{t_1+\cdots+t_{\tau-1}=k,t_j\geq 1} (u^\top W_{t_1} u)(u^\top W_{t_2} u) \cdots (u^\top W_{t_{\tau-1}} u).$$

Denote $\beta_{t_1}^\Delta = u^\top W_{t_1} u$, we can write concisely

$$\mathbb{E}\langle uu^\top, \mathcal{S}_{T,k}(\Delta) \rangle = \frac{1}{\lambda^{2k}} \sum_{\tau=2}^{k} (-1)^{1+\tau} \binom{k-1}{\tau-1} \sum_{t_1+\cdots+t_{\tau-1}=k,t_j\geq 1} \mathbb{E}\big(\beta_{t_1}^\Delta \beta_{t_2}^\Delta \cdots \beta_{t_{\tau-1}}^\Delta\big).$$

$$(31)$$

Now, we investigate the concentration property of $\beta_t^\Delta = u^\top W_t u$. Clearly, we can write

$$\beta_1^\Delta = \underbrace{2\lambda \cdot (u^\top Z v)}_{\beta_{1,1}^\Delta} + \underbrace{u^\top ZZ^\top u}_{\beta_{1,0}^\Delta}$$

and for all $t \geq 2$, we write $\beta_t^\Delta = \beta_{t,1}^\Delta + \beta_{t,0}^\Delta$ where

$$\beta_{t,0}^\Delta = u^\top ZZ^\top U_\perp (U_\perp^\top ZZ^\top U_\perp)^{t-2} U_\perp^\top ZZ^\top u$$
$$\quad + \lambda^2 uv^\top Z^\top U_\perp (U_\perp^\top ZZ^\top U_\perp)^{t-2} U_\perp^\top Zvu^\top$$
$$\beta_{t,1}^\Delta = 2\lambda uv^\top Z^\top U_\perp (U_\perp^\top ZZ^\top U_\perp)^{t-2} U_\perp^\top ZZ^\top u.$$

As a result, we can calculate

$$\mathbb{E}(\beta_{t_1}^\Delta \beta_{t_2}^\Delta \cdots \beta_{t_{\tau-1}}^\Delta) = \mathbb{E}\big((\beta_{t_1,0}^\Delta + \beta_{t_1,1}^\Delta)(\beta_{t_2,0}^\Delta + \beta_{t_2,1}^\Delta) \cdots (\beta_{t_{\tau-1},0}^\Delta + \beta_{t_{\tau-1},1}^\Delta)\big).$$

It is easy to check that $\mathbb{E}\beta_1^\Delta = d_{2-} + 1$ and for $t \geq 2$

$$\mathbb{E}\beta_t^\Delta = \lambda^2 \mathbb{E}\big(v^\top Z^\top U_\perp (U_\perp^\top ZZ^\top U_\perp)^{t-2} U_\perp^\top Zv\big) + \mathbb{E}\mathrm{tr}\big(Z^\top U_\perp (U_\perp^\top ZZ^\top U_\perp)^{t-2} U_\perp^\top Z\big)$$

$$= \Big(1 + \frac{\lambda^2}{d_{2-}+1}\Big) \cdot \mathbb{E}\mathrm{tr}\big(Z^\top U_\perp (U_\perp^\top ZZ^\top U_\perp)^{t-2} U_\perp^\top Z\big)$$

$$= \Big(1 + \frac{\lambda^2}{d_{2-}+1}\Big) \cdot \mathbb{E}\mathrm{tr}\big((U_\perp^\top ZZ^\top U_\perp)^{t-1}\big)$$

where the second equality can be checked by choosing $v = e_1 \in \mathbb{R}^{d_{2^-}+1}$. Since $Z^\mathsf{T}u$ and $Z^\mathsf{T}U_\perp$ are independent, it is easy to check that

$$\mathbb{E}\beta^\Delta_{t_1,i_1}\beta^\Delta_{t_2,i_2}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}} = 0, \quad \text{if } \sum_{j=1}^{\tau-1} i_j \text{ is an odd number}$$

for all $i_1, i_2, \cdots, i_{\tau-1} \in \{0,1\}$. As a result, we observe that $\mathbb{E}\langle uu^\mathsf{T}, \mathcal{S}_{T,k}\rangle$ has contributions to $E_{2k}, E_{2k-2}, E_{2k-4}, \cdots, E_{2\lceil k/2\rceil}$. (Recall that $E_{2k}$ is the coefficient for $\frac{1}{\lambda^{2k}}$.)

Moreover, since $Z^\mathsf{T}u$ and $Z^\mathsf{T}U_\perp$ are independent, we can conclude that

$$\beta^\Delta_{1,1} \sim \mathcal{N}(0, 4\lambda^2)$$

and for all $t \geq 2$,

$$\beta^\Delta_{t,1}|U_\perp^\mathsf{T}Z \sim \mathcal{N}\big(0, 4\lambda^2\|Z^\mathsf{T}U_\perp(U_\perp^\mathsf{T}ZZ^\mathsf{T}U_\perp)^{t-2}U_\perp^\mathsf{T}Zv\|_{\ell_2}^2\big).$$

We can get, for all $t \geq 2$, that

$$\mathbb{E}^{1/2}\big[(\beta^\Delta_{t,1})^2\big|U_\perp^\mathsf{T}Z\big] \lesssim \mathbb{E}^{1/4}\big[(\beta^\Delta_{t,1})^4\big|U_\perp^\mathsf{T}Z\big] \lesssim \lambda\|U_\perp^\mathsf{T}Z\|^{2(t-1)}$$

Therefore, it is easy to check that for any $(i_1, i_2, \cdots, i_{\tau-1}) \in \{0,1\}^{\tau-1}$ where there exists some $i_j \geq 1$, then $\mathbb{E}\beta^\Delta_{t_1,i_1}\beta^\Delta_{t_2,i_2}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}}$'s contribution to any $E_{2k_1}$ is bounded by $\frac{1}{d_2}\cdot\left(\frac{C_1 d_2}{\lambda^2}\right)^{k_1}$ for some absolute constant $C_1 > 0$ and $2\lceil k/2\rceil \leq 2k_1 \leq 2k$. To show this, w.l.o.g, let $i_1 = i_2 = 1$ and observe that

$$\begin{aligned}
\mathbb{E}\beta^\Delta_{t_1,1}\beta^\Delta_{t_2,1}\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}} &= \mathbb{E}^{1/2}(\beta^\Delta_{t_1,1}\beta^\Delta_{t_2,1})^2\mathbb{E}^{1/2}\big(\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}}\big)^2 \\
&\leq \mathbb{E}^{1/4}(\beta^\Delta_{t_1,1})^4\mathbb{E}^{1/4}(\beta^\Delta_{t_2,1})^4\mathbb{E}^{1/2}\big(\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}}\big)^2 \\
&\leq \lambda^2 d_2^{t_1+t_2-2}\mathbb{E}^{1/2}\big(\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}}\big)^2
\end{aligned} \tag{32}$$

and then we get

$$\frac{1}{\lambda^{2k}}\mathbb{E}\beta^\Delta_{t_1,1}\beta^\Delta_{t_2,1}\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}} \leq \frac{1}{d_2}\cdot\left(\frac{d_2}{\lambda^2}\right)^{t_1+t_2-1}\cdot\frac{\mathbb{E}^{1/2}\big(\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}}\big)^2}{\lambda^{2(k-t_1-t_2)}}.$$

The claim follows immediately since

$$\begin{aligned}
\frac{\mathbb{E}^{1/2}\big(\beta^\Delta_{t_3,i_3}\cdots\beta^\Delta_{t_{\tau-1},i_{\tau-1}}\big)^2}{\lambda^{2(k-t_1-t_2)}} &\leq \sum_{k_1=\lceil(k-t_1-t_2)/2\rceil}^{k_1}\frac{C_1^{k_1}\mathbb{E}^{1/2}\|Z\|^{4k_1}}{\lambda^{2k_1}} \\
&\leq \sum_{k_1=\lceil(k-t_1-t_2)/2\rceil}^{k_1}\left(\frac{C_2 d_2}{\lambda^2}\right)^{k_1}
\end{aligned}$$

for some absolute constant $C_1, C_2 > 0$ and where the last inequality is due to $\mathbb{E}\|Z\|^{4k_1} \leq C_3^{4k_1}d_2^{2k_1}$ for some absolute constant $C_3 > 0$.

As a result, in order to calculate eq. (31), it suffices to calculate

$$\frac{1}{\lambda^{2k}} \sum_{\tau=2}^{k} (-1)^{1+\tau} \binom{k-1}{\tau-1} \sum_{t_1+\cdots+t_{\tau-1}=k, t_j \geq 1} \mathbb{E}\big(\beta_{t_1,0}^{\Delta} \beta_{t_2,0}^{\Delta} \cdots \beta_{t_{\tau-1},0}^{\Delta}\big). \tag{33}$$

Next, we will replace $\mathbb{E}\big(\beta_{t_1,0}^{\Delta} \beta_{t_2,0}^{\Delta} \cdots \beta_{t_{\tau-1},0}^{\Delta}\big)$ with $\mathbb{E}\beta_{t_1,0}^{\Delta} \mathbb{E}\beta_{t_2,0}^{\Delta} \cdots \mathbb{E}\beta_{t_{\tau-1},0}^{\Delta}$ for which we shall investigate the concentrations of $\beta_{t,0}^{\Delta}$. To this end, we have the sub-exponential inequality

$$\mathbb{P}\big(\big|u^{\mathsf{T}} Z Z^{\mathsf{T}} u - d_{2-}\big| \geq C_3 \sqrt{\alpha d_{2-}} + C_4 \alpha\big) \leq C_5 e^{-\alpha}, \quad \forall \alpha > 0$$

for some constants $C_3, C_4 > 0$. Again, by Gaussian isoperimetric inequality and the proof of Theorem $3^{[4]}$, we can show, for all $\alpha > 0$

$$\begin{aligned}
\mathbb{P}\big(\big|u^{\mathsf{T}} Z(Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}} Z)^t Z^{\mathsf{T}} u &- \mathbb{E}u^{\mathsf{T}} Z(Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}} Z)^t Z^{\mathsf{T}} u\big| \\
&\geq C_3 \alpha d_2^{t+1/2} + C_4 e^{-c_1 d_2} d_2^{t+1}\big) \\
&\leq C_5 e^{-\alpha^2} + C_6 e^{-c_2 d_2}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}\big(\big|v^{\mathsf{T}}(Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}} Z)^{t-1} v - \mathbb{E}v^{\mathsf{T}}(Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}} Z)^{t-1} v\big| &\geq C_3 t\alpha d_2^{t-3/2} + C_4 e^{-c_1 d_2} d_2^{t-1}\big) \\
&\leq C_5 e^{-\alpha^2} + C_6 e^{-c_2 d_2}.
\end{aligned}$$

Therefore, we can show that $\big|\mathbb{E}(\beta_{t_1,0}^{\Delta} \beta_{t_2,0}^{\Delta} \cdots \beta_{t_{\tau-1},0}^{\Delta}) - (\mathbb{E}\beta_{t_1,0}^{\Delta})(\mathbb{E}\beta_{t_2,0}^{\Delta}) \cdots$ $(\mathbb{E}\beta_{t_{\tau-1},0}^{\Delta})\big|$'s contribution to any $E_{2k_1}$ is bounded by $\frac{1}{\sqrt{d_2}} \cdot (C_1 d_2/\lambda^2)^{k_1}$ for some constant $C_1 > 0$ and $2\lceil k/2 \rceil \leq 2k_1 \leq 2k$. Indeed, the above concentration inequalities of $\beta_{t,0}^{\Delta}$ imply

$$\mathbb{E}^{1/2}(\beta_{t,0}^{\Delta} - \mathbb{E}\beta_{t,0}^{\Delta})^2 \lesssim d_2^{t-1/2} + \lambda^2 d_2^{t-3/2}, \quad \forall t \geq 1.$$

The claim can be proved as in eq. (32). Indeed, we can write

$$\begin{aligned}
&\frac{1}{\lambda^{2k}} \Big|\mathbb{E}\beta_{t_1,0}^{\Delta} \beta_{t_2,0}^{\Delta} \cdots \beta_{t_{\tau-1},0}^{\Delta} - (\mathbb{E}\beta_{t_1,0}^{\Delta})(\mathbb{E}\beta_{t_2,0}^{\Delta}) \cdots \mathbb{E}(\beta_{t_{\tau-1},0}^{\Delta})\Big| \\
&\leq \frac{1}{\lambda^{2k}} \sum_{i=1}^{\tau-1} \Big(\prod_{j=1}^{i-1} \mathbb{E}\beta_{t_j,0}^{\Delta}\Big) \Big|\mathbb{E}\big(\beta_{t_i,0}^{\Delta} - \mathbb{E}\beta_{t_i,0}^{\Delta}\big)\Big(\prod_{j=i+1}^{\tau-1} \beta_{t_j,0}^{\Delta}\Big)\Big| \\
&\leq \sum_{i=1}^{\tau-1} \frac{\mathbb{E}^{1/2}\big(\beta_{t_i,0}^{\Delta} - \mathbb{E}\beta_{t_i,0}^{\Delta}\big)^2}{\lambda^{2t_i}} \cdot \frac{1}{\lambda^{2(k-t_i)}} \prod_{j=1}^{i-1} \Big(\mathbb{E}\beta_{t_j,0}^{\Delta}\Big) \mathbb{E}^{1/2}\Big(\prod_{j=i+1}^{\tau-1} \beta_{t_j,0}^{\Delta}\Big)^2 \\
&\leq \sum_{i=1}^{\tau-1} \frac{1}{\sqrt{d_2}} \Big((d_2/\lambda^2)^{t_i} + (d_2/\lambda^2)^{t_i-1}\Big) \cdot \frac{1}{\lambda^{2(k-t_i)}} \prod_{j=1}^{i-1} \Big(\mathbb{E}\beta_{t_j,0}^{\Delta}\Big) \mathbb{E}^{1/2}\Big(\prod_{j=i+1}^{\tau-1} \beta_{t_j,0}^{\Delta}\Big)^2
\end{aligned}$$

---

[4] We just need to study the Lipschitz property of the function $f(Z) = u^{\mathsf{T}} Z(Z^{\mathsf{T}} U_\perp U_\perp^{\mathsf{T}} Z)^t Z^{\mathsf{T}} u \cdot \mathbf{1}(\|Z\| \leq C_1 \sqrt{d_2})$

which concludes the proof since $\frac{1}{\lambda^{2(k-t_i)}} \prod_{j=1}^{i-1} \left( \mathbb{E}\beta_{t_j,0}^{\Delta} \right) \mathbb{E}^{1/2} \left( \prod_{j=i+1}^{\tau-1} \beta_{t_j,0}^{\Delta} \right)^2 \leq \left( \frac{C_2 d_2}{\lambda^2} \right)^{k-t_i}$.

To this end, to calculate eq. (31), it suffices to calculate

$$\frac{1}{\lambda^{2k}} \sum_{\tau=2}^{k} (-1)^{1+\tau} \binom{k-1}{\tau-1} \sum_{t_1+\cdots+t_{\tau-1}=k, t_j \geq 1} \mathbb{E}\beta_{t_1,0}^{\Delta} \mathbb{E}\beta_{t_2,0}^{\Delta} \cdots \mathbb{E}\beta_{t_{\tau-1},0}^{\Delta}$$

Now, we compute $\mathbb{E}\beta_{t,0}^{\Delta} = \left( 1 + \lambda^2/(d_{2-}+1) \right) \cdot \mathbb{E}\mathrm{tr}\left( (U_{\perp}^{\mathsf{T}} ZZ^{\mathsf{T}} U_{\perp})^{t-1} \right)$. Note that the matrix $U_{\perp}^{\mathsf{T}} Z \in \mathbb{R}^{d_{1-} \times (d_{2-}+1)}$ has i.i.d. standard normal entries. By the moment of Marchenko-Pastur law ((Mingo and Speicher, 2017)), for all $t \geq 2$, we define (additionally, $\beta_1 = d_{2-} + 1$)

$$\frac{\beta_t}{1 + \lambda^2/(1+d_{2-})} = \frac{1}{t-1} \sum_{r=0}^{t-2} d_{1-}^{r+1} (d_{2-}+1)^{t-1-r} \binom{t-1}{r+1}\binom{t-1}{r}. \qquad (34)$$

Note that $\mathbb{E}\mathrm{tr}\left( (U_{\perp}^{\mathsf{T}} ZZ^{\mathsf{T}} U_{\perp})^{t-1} \right) = \mathbb{E}\mathrm{tr}\left( (Z^{\mathsf{T}} U_{\perp} U_{\perp}^{\mathsf{T}} Z)^{t-1} \right)$ for all $t \geq 2$. By the rate of convergence of Marchenko Pastur law ((Götze and Tikhomirov, 2011, Theorem 1.1)), we have (as long as $\sqrt{d_2} \geq \log^2 d_2$)

$$\frac{\left| \beta_t - \mathbb{E}\beta_{t,0}^{\Delta} \right|}{1 + \lambda^2/(d_{2-}+1)} \leq \frac{1}{\sqrt{d_2}} \cdot \left( C_1 d_2 \right)^{t-1}$$

for all $t \geq 2$ where $C_1 > 0$ is an absolute constant. As a result, we get that for all $t_1 + \cdots + t_{\tau-1} = k$, the contribution to $E_{2k_1}$ from $\left| \mathbb{E}\beta_{t_1,0}^{\Delta} \mathbb{E}\beta_{t_2,0}^{\Delta} \cdots \mathbb{E}\beta_{t_{\tau-1},0}^{\Delta} - \beta_{t_1}\beta_{t_2}\cdots\beta_{t_{\tau-1}} \right|$ is bounded by $\frac{1}{\sqrt{d_2}} \cdot \left( \frac{C_1 d_2}{\lambda^2} \right)^{k_1}$.

Therefore, by eq. (31), to calculate $\mathbb{E}\langle uu^{\mathsf{T}}, \mathcal{S}_{T,k}(\Delta) \rangle$, we consider the following term

$$\frac{1}{\lambda^{2k}} \sum_{\tau=2}^{k} (-1)^{1+\tau} \binom{k-1}{\tau-1} \sum_{t_1+\cdots+t_{\tau-1}=k, t_j \geq 1} \beta_{t_1}\beta_{t_2}\cdots\beta_{t_{\tau-1}}$$

which is the $k$-th order derivative of the function $\frac{1}{\lambda^{2k} \cdot (k!)} (1 - g(\alpha))^{k-1}$ at $\alpha = 0$ where

$$g(\alpha) = \beta_1 \alpha + \alpha^2 \beta_2 + \alpha^3 \beta_3 + \cdots = \sum_{k \geq 1} \beta_k \alpha^k. \qquad (35)$$

Now, we calculate the explicit form of the function $g(\alpha)$. Denote $\gamma = \frac{d_{1-}}{d_{2-}+1}$ and $Y$ the random variable obeying the Marchenko-Pastur distribution, i.e., its pdf is given by

$$f_Y(y) = \frac{1}{2\pi} \frac{\sqrt{(\gamma_+ - y)(y - \gamma_-)}}{\gamma y} \cdot \mathbf{1}(y \in [\gamma_-, \gamma_+])$$

where $\gamma_+ = (1 + \sqrt{\gamma})^2$ and $\gamma_- = (1 - \sqrt{\gamma})^2$. It is easy to check that ((Mingo and Speicher, 2017))

$$\beta_t = \left( 1 + \frac{\lambda^2}{1+d_{2-}} \right) d_{1-} (d_{2-}+1)^{t-1} \mathbb{E}Y^{t-1}, \quad \forall t \geq 2.$$

For notational simplicity, we just write $d_{2-}$ instead of $1 + d_{2-}$. As a result, we get for $\alpha \ll \frac{1}{d_2^2}$,

$$
\begin{aligned}
g(\alpha) &= \beta_1 \alpha + \left(1 + \frac{\lambda^2}{d_{2-}}\right) d_{1-} \alpha \mathbb{E} \sum_{t \geq 1} d_{2-}^t (\alpha Y)^t \\
&= \beta_1 \alpha + \left(1 + \frac{\lambda^2}{d_{2-}}\right) \mathbb{E} \frac{d_{1-} \alpha d_{2-} \alpha Y}{1 - d_{2-} \alpha Y} \\
&= \alpha d_{2-} + \left(1 + \frac{\lambda^2}{d_{2-}}\right) \cdot \frac{\left(\sqrt{1 - \alpha d_{2-} \gamma_-} - \sqrt{1 - \alpha d_{2-} \gamma_+}\right)^2}{4}
\end{aligned}
$$

where the last equality comes up by integrating $Y$ according to the p.d.f. $F_Y(y)$. Therefore, we get

$$
1 - g(\alpha) = \frac{1}{2}\left[g_+(\alpha) - \frac{\lambda^2}{d_{2-}} g_-(\alpha)\right]
$$

where

$$
g_-(\alpha) = 1 - (d_{1-} + d_{2-})\alpha - \sqrt{(1 - \alpha d_{2-} \gamma_-)(1 - \alpha d_{2-} \gamma_+)}
$$

and

$$
g_+(\alpha) = 1 - (d_{2-} - d_{1-})\alpha + \sqrt{(1 - \alpha d_{2-} \gamma_-)(1 - \alpha d_{2-} \gamma_+)}.
$$

Therefore, in order to calculate $\mathbb{E}\langle uu^\top, \mathcal{S}_{T,k}(\Delta)\rangle$, it suffices to calculate the $k$-th order derivative of function $\frac{(1 - g(\alpha))^{k-1}}{\lambda^{2k} \cdot (k!)}$ at $\alpha = 0$. Write

$$
\left.\frac{\left[\left(1 - g(\alpha)\right)^{k-1}\right]^{(k)}}{\lambda^{2k}(k!)}\right|_{\alpha=0}
$$

$$
= \frac{1}{\lambda^{2k} \cdot 2^{k-1} \cdot (k!)} \sum_{t=0}^{k-1} \binom{k-1}{t} \left(-\frac{\lambda^2}{d_{2-}}\right)^t \left.\left[g_-^t(\alpha) g_+^{k-1-t}(\alpha)\right]^{(k)}\right|_{\alpha=0}. \tag{36}
$$

Note that $g_-(\alpha) = O(\alpha^2)$. The terms in eq. (36) with $t > \frac{k}{2}$ are all 0. Recall that we are interested in the $k_0$-th order term in the error $\|\hat{u}\hat{u}^\top - uu^\top\|_F^2$ whose denominator is $\lambda^{2k_0}$. By eq. (36), the $k_0$-th order error term $\frac{1}{\lambda^{2k_0}}$ can be contributed from $\mathbb{E}\langle uu^\top, \mathcal{S}_{T,k}(\Delta)\rangle$ for $k = k_0, k = k_0 + 1, \cdots, k = 2k_0$.

By the above analysis, we conclude that the $k_0$-th error term (except the negligible error terms from translating $\mathbb{E}(\beta_{t_1}^\Delta \beta_{t_2}^\Delta \cdots \beta_{t_{\tau-1}}^\Delta)$ into $\beta_{t_1} \beta_{t_2} \cdots \beta_{t_{\tau-1}}$) of $\|\hat{u}\hat{u}^\top - uu^\top\|_F^2$ is given by $E_{2k_0} = \sum_{t=0}^{k_0} E_{2k_0,t}$ where (we change $k$ in (36) to $k_0 + t$)

$$
E_{2k_0,t} = \frac{1}{\lambda^{2k_0}} \frac{1}{2^{k_0+t-1}} \frac{1}{(k_0 + t)!} \binom{k_0 + t - 1}{t} \left(-\frac{1}{d_{2-}}\right)^t \left.\left[g_-^t(\alpha) g_+^{k_0-1}(\alpha)\right]^{(k_0+t)}\right|_{\alpha=0}
$$

When $t = k_0$, we have

$$
g_-^{k_0}(\alpha) = \frac{(4d_{1-}d_{2-})^{k_0} \alpha^{2k_0}}{\left[1 - \alpha(d_{1-} + d_{2-}) + \sqrt{(1 - \alpha d_{2-} \gamma_-)(1 - \alpha d_{2-} \gamma_+)}\right]^{k_0}}
$$

implying that

$$\left[g_-^{k_0}(\alpha)g_+^{k_0-1}(\alpha)\right]^{(2k_0)}\Big|_{\alpha=0} = (2k_0)!\frac{(4d_{1-}d_{2-})^{k_0}}{2}.$$

Therefore, we get $E_{2k_0,k_0} = (-1)^{k_0}d_{1-}^{k_0}\binom{2k_0-1}{k_0}$. Now, we consider $t \le k_0 - 1$ and we observe

$$1 - \alpha(d_{1-} + d_{2-}) + \sqrt{(1 - \alpha d_{2-}\gamma_-)(1 - \alpha d_{2-}\gamma_+)} = g_+(\alpha) - 2d_{1-}\alpha$$

so that $g_-(\alpha) = \frac{4d_{1-}d_{2-}\alpha^2}{g_+(\alpha)-2\alpha d_{1-}}$, Then, we get

$$\left[g_-^t(\alpha)g_+^{k_0-1}(\alpha)\right]^{(k_0+t)}\Big|_{\alpha=0} = \left[\frac{(4d_{1-}d_{2-}\alpha^2)^t}{(g_+(\alpha)-2\alpha d_{1-})^t}\cdot g_+^{k_0-1}(\alpha)\right]^{(k_0+t)}\Big|_{\alpha=0}$$

$$= \binom{k_0+t}{2t}(2t)!(4d_{1-}d_{2-})^t\left[\frac{g_+^{k_0-1}(\alpha)}{(g_+(\alpha)-2\alpha d_{1-})^t}\right]^{(k_0-t)}\Big|_{\alpha=0}.$$

It suffices to calculate the $(k_0 - t)$-th derivative of function $g_+^{k_0-1}(\alpha)/(g_+(\alpha) - 2\alpha d_{1-})^t$ at $\alpha = 0$. We write

$$\left[\frac{g_+^{k_0-1}(\alpha)}{(g_+(\alpha)-2\alpha d_{1-})^t}\right]^{(k_0-t)}\Big|_{\alpha=0}$$

$$= \left[\sum_{t_1=0}^{k_0-1}\binom{k_0-1}{t_1}(2\alpha d_{1-})^{k_0-1-t_1}(g_+(\alpha)-2\alpha d_{1-})^{t_1-t}\right]^{(k_0-t)}\Big|_{\alpha=0}.$$

Observe that $\left[(2\alpha d_{1-})^{k_0-1-t_1}\right]^{(k_0-t)}\Big|_{\alpha=0} = 0$ for all $t_1 < t - 1$. Then, we get

$$\left[\frac{g_+^{k_0-1}(\alpha)}{(g_+(\alpha)-2\alpha d_{1-})^t}\right]^{(k_0-t)}\Big|_{\alpha=0}$$

$$= \left[\sum_{t_1=t-1}^{k_0-1}\binom{k_0-1}{t_1}(2\alpha d_{1-})^{k_0-1-t_1}(g_+(\alpha)-2\alpha d_{1-})^{t_1-t}\right]^{(k_0-t)}\Big|_{\alpha=0}.$$

If $t_1 = t - 1$, then

$$\left[\binom{k_0-1}{t_1}(2\alpha d_{1-})^{k_0-1-t_1}(g_+(\alpha)-2\alpha d_{1-})^{t_1-t}\right]^{(k_0-t)}\Big|_{\alpha=0}$$

$$= \binom{k_0-1}{t-1}(2d_1)^{k_0-t}(k_0-t)!\cdot\frac{1}{2}$$

If $t_1 \ge t$, we have

$$\left[(2\alpha d_{1-})^{k_0-1-t_1}(g_+(\alpha)-2\alpha d_{1-})^{t_1-t}\right]^{(k_0-t)}\Big|_{\alpha=0}$$

$$= \binom{k_0-t}{k_0-1-t_1}(2d_1)^{k_0-1-t_1}(k_0-1-t_1)!\left[(g_+(\alpha)-2\alpha d_{1-})^{t_1-t}\right]^{(t_1+1-t)}\Big|_{\alpha=0}.$$

Clearly, if $t_1 = t$, then $\left[\left(g_+(\alpha) - 2\alpha d_{1-}\right)^{t_1-t}\right]^{(t_1+1-t)}\big|_{\alpha=0} = 0$. For $t_1 \geq t+1$, recall that

$$g_+(\alpha) - 2\alpha d_1 = 1 - (d_{1-} + d_{2-})\alpha + \sqrt{(1 - \alpha d_{2-}\gamma_-)(1 - \alpha d_{2-}\gamma_+)}.$$

It is easy to check that

$$\left[\left(g_+(\alpha) - 2\alpha d_{1-}\right)^{t_1-t}\right]^{(t_1+1-t)}\Big|_{\alpha=0}$$
$$= -\left[\left(1 - (d_{1-} + d_{2-})\alpha - \sqrt{(1 - \alpha d_{2-}\gamma_-)(1 - \alpha d_{2-}\gamma_+)}\right)^{t_1-t}\right]^{(t_1+1-t)}\Big|_{\alpha=0}$$
$$= -\left[\left(\frac{4d_{1-}d_{2-}\alpha^2}{g_+(\alpha) - 2\alpha d_{1-}}\right)^{t_1-t}\right]^{(t_1+1-t)}\Big|_{\alpha=0}$$

which is non-zero only when $t_1 = t+1$. In fact, when $t_1 = t+1$, we get

$$\left[\left(g_+(\alpha) - 2\alpha d_{1-}\right)^{t_1-t}\right]^{(t_1+1-t)}\Big|_{\alpha=0} = -4d_{1-}d_{2-}.$$

Therefore, we conclude that

$$\left[\frac{g_+^{k_0-1}(\alpha)}{(g_+(\alpha) - 2\alpha d_{1-})^t}\right]^{(k_0-t)}\Big|_{\alpha=0} = \binom{k_0-1}{t-1}(2d_{1-})^{k_0-t}(k_0-t)! \cdot \frac{1}{2} - \binom{k_0-1}{t+1}$$
$$\times \binom{k_0-t}{2}(2d_{1-})^{k_0-t-2}(k_0-2-t)!(4d_{1-}d_{2-}).$$

As a result, for $t \leq k_0 - 1$, we get

$$E_{2k_0,t} = d_{1-}^{k_0} \cdot (-1)^t \binom{k_0+t-1}{t}\binom{k_0-1}{t-1}$$
$$- d_{1-}^{k_0-1}d_{2-} \cdot (-1)^t \binom{k_0+t-1}{t}\binom{k_0-1}{t+1}.$$

Clearly, it also holds for $t = k_0$. Therefore, we have

$$E_{2k_0} = \sum_{t=0}^{k_0} E_{2k_0,t} = d_{1-}^{k_0}\sum_{t=0}^{k_0}(-1)^t\binom{k_0+t-1}{t}\binom{k_0-1}{t-1}$$
$$- d_{1-}^{k_0-1}d_{2-}\sum_{t=0}^{k_0-2}(-1)^t\binom{k_0+t-1}{t}\binom{k_0-1}{t+1}.$$

It is easy to check that

$$\sum_{t=0}^{k_0}(-1)^t\binom{k_0+t-1}{t}\binom{k_0-1}{t-1} = \sum_{t=1}^{k_0}(-1)^t\binom{k_0+t-1}{t}\binom{k_0-1}{t-1}$$
$$= (-1)\sum_{t=0}^{k_0-1}(-1)^t\binom{k_0+t}{t+1}\binom{k_0-1}{t}.$$

It is interesting to observe that $\sum_{t=0}^{k_0-1}(-1)^t\binom{k_0+t}{t+1}\binom{k_0-1}{t}$ equals the coefficient of $x^{k_0-1}$ in the polynomial $(1+x)^{k_0}\left[1-(1+x)\right]^{k_0-1}$. Then, it is easy to check that $\sum_{t=0}^{k_0-1}(-1)^t\binom{k_0+t}{t+1}\binom{k_0-1}{t}=(-1)^{k_0-1}$. Similarly, we can observe that

$$\sum_{t=0}^{k_0-2}(-1)^t\binom{k_0+t-1}{t}\binom{k_0-1}{t+1}=\sum_{t=1}^{k_0-1}(-1)^{t-1}\binom{k_0+t-2}{t-1}\binom{k_0-1}{t}$$
$$=(-1)\sum_{t=1}^{k_0-1}(-1)^t\binom{k_0+t-2}{t-1}\binom{k_0-1}{t}.$$

Again, it is easy to check that $\sum_{t=1}^{k_0-1}(-1)^t\binom{k_0+t-2}{t-1}\binom{k_0-1}{t}$ equals the coefficient of $x^{k_0-1}$ in the polynomial $(1+x)^{k_0-2}[1-(1+x)]^{k_0-1}$. As a result, we get $\sum_{t=1}^{k_0-1}(-1)^t\binom{k_0+t-2}{t-1}\binom{k_0-1}{t}=(-1)^{k_0-1}$. To this end, we conclude that

$$E_{2k_0}=(-1)^{k_0}d_{1-}^{k_0-1}(d_{1-}-d_{2-})$$

, i.e., the $k_0$-th error term in $\mathbb{E}\|\hat{u}\hat{u}^\mathsf{T}-uu^\mathsf{T}\|_\mathrm{F}^2$ is given by $\frac{(-1)^{k_0}d_{1-}^{k_0-1}(d_{1-}-d_{2-})}{\lambda^{2k_0}}$ (except the negligible error terms). In a similar fashion, we can show that the $k_0$-th error term in $\mathbb{E}\|\hat{v}\hat{v}^\mathsf{T}-vv^\mathsf{T}\|_\mathrm{F}^2$ is given by $\frac{(-1)^{k_0}d_{2-}^{k_0-1}(d_{2-}-d_{1-})}{\lambda^{2k_0}}$. Meanwhile, the negligible error terms from translating $\mathbb{E}(\beta_{t_1}^\Delta\beta_{t_2}^\Delta\cdots\beta_{t_{\tau-1}}^\Delta)$ into $\beta_{t_1}\beta_{t_2}\cdots\beta_{t_{\tau-1}}$ are upper bounded by $\frac{k_0}{\sqrt{d_2}}\cdot\left(\frac{C_2d_2}{\lambda_r^2}\right)^{k_0}$ which concludes the proof. $\qquad\square$

**Lemma 6.** *Let $\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_r)$ 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*

$$\mathbb{E}\|\Lambda^{-j_1}ZZ^\mathsf{T}\Lambda^{-j_2}\|_\mathrm{F}^2=d^2\|\Lambda^{-j_1-j_2}\|_\mathrm{F}^2+d\big(\|\Lambda^{-j_1-j_2}\|_\mathrm{F}^2+\|\Lambda^{-j_1}\|_\mathrm{F}^2\|\Lambda^{-j_2}\|_\mathrm{F}^2\big).$$

*Proof of Lemma 6.* Let $z_1,\cdots,z_r\in\mathbb{R}^d$ denote the columns of $Z^\mathsf{T}$. Therefore, we can write

$$\|\Lambda^{-j_1}ZZ^\mathsf{T}\Lambda^{-j_2}\|_\mathrm{F}^2=\sum_{i=1}^r\frac{1}{\lambda_i^{2(j_1+j_2)}}(z_i^\mathsf{T}z_i)^2+\sum_{1\leq i_1\neq i_2\leq r}\frac{1}{\lambda_{i_1}^{2j_1}\lambda_{i_2}^{2j_2}}(z_{i_1}^\mathsf{T}z_{i_2})^2.$$

Then, we get

$$\mathbb{E}\|\Lambda^{-1}ZZ^\mathsf{T}\Lambda^{-1}\|_\mathrm{F}^2=\sum_{i=1}^r\frac{d^2+2d}{\lambda_i^{2(j_1+j_2)}}+\sum_{1\leq i_1\neq i_2\leq r}\frac{d}{\lambda_{i_1}^{2j_1}\lambda_{i_2}^{2j_2}}$$
$$=d^2\|\Lambda^{-j_1-j_2}\|_\mathrm{F}^2+d\big(\|\Lambda^{-j_1-j_2}\|_\mathrm{F}^2+\|\Lambda^{-j_1}\|_\mathrm{F}^2\|\Lambda^{-j_2}\|_\mathrm{F}^2\big). \quad\square$$

## Acknowledgments