Multiple Testing of Linear Forms for Noisy Matrix Completion

Wanteng Ma[1], Lilun Du[2], Dong Xia[3] and Ming Yuan[4]

[1] Department of Statistics and Data Science, University of Pennsylvania

[2]Department of Decision Analytics and Operations, City University of Hong Kong

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

[4]Department of Statistics, Columbia University

March 8, 2025

Abstract

Many important tasks of large-scale recommender systems can be naturally cast as testing multiple linear forms for noisy matrix completion model. These problems, however, present unique challenges because of the subtle bias-and-variance tradeoff and the intricate dependence among the estimated entries induced by the low-rank structure. In this paper, we develop a general approach to overcome these difficulties by introducing new statistics for individual tests with sharp asymptotics both marginally and jointly, and utilizing them to control the false discovery rate (FDR) via a data splitting and symmetric aggregation scheme. We show that valid FDR control can be achieved with guaranteed power under nearly optimal sample size requirements using the proposed methodology. Extensive numerical simulations and real data examples are also presented to further illustrate its practical merits.

[3]Dong Xia's research was partially supported by Hong Kong RGC Grant GRF 16301622 and GRF 16303224.

[4]Ming Yuan's research was supported in part by NSF Grants DMS-2015285 and DMS-2052955.

1 Introduction

Popularized by the Netflix prize (Bennett and Lanning, 2007), matrix completion techniques have emerged as an essential tool for large-scale collaborative-filtering-based recommender systems. See, e.g., Resnick and Varian (1997); Schafer et al. (2007); Koren et al. (2009); Davidson et al. (2010); McAuley and Leskovec (2013); Das et al. (2017). Consider, more specifically, representing the ratings of d_1 users on d_2 products/items by a $d_1 \times d_2$ matrix. For all practical purposes, both d_1 and d_2 can be very large yet only a rather small number of the entries can be observed. The idea is that if the interaction between users and products can be approximately captured by a handful of latent user-specific and product-specific characteristics, then it is possible to infer the whole user-item rating matrix from these sparsely observed entries, and hence recommend products to users who may be genuinely interested in them. Since the pioneering works of Candès and Tao (2009); Candes and Plan (2010); Candes and Recht (2012), a lot of impressive progress has been made to make these techniques more accurate and scalable, and to better understand the statistical and computational underpinnings of the problem. See, e.g., Cai et al. (2010); Keshavan et al. (2010a); Recht et al. (2010); Gross (2011); Koltchinskii et al. (2011); Liu (2011); Negahban and Wainwright (2011); Rohde and Tsybakov (2011); Tsybakov et al. (2011); Negahban and Wainwright (2012); Sun and Zhang (2012); Klopp (2014); Cai and Zhang (2015); Cai and Zhou (2016); Gao et al. (2016), among numerous others.

Most of these existing works study recommender systems from an estimation perspective and investigate how well the user-item matrix can be estimated or reconstructed collectively. These are clearly relevant metrics for evaluating recommender systems. For example, the Netflix prize uses root mean squared error as the gold standard for the competition. Yet they do not account for the fact that only a subset of the products can be recommended to a user and as such estimation accuracy may not be directly translated into the quality of these recommendations. Instead, various classical notions for binary classification such as precision and recall are often adopted in practice to evaluate the quality of top recommendations. See, e.g., Herlocker et al. (2004). This subtlety has significant statistical implications. First of all, making quality recommendations requires a more careful uncertainty quantification. Consider recommending between a blockbuster movie and an independent film to a user. Even if both estimated ratings are similar and favorable, the uncertainty associated with the estimated rating for the former is likely to be much smaller as it has been viewed by a much greater number of people. It could therefore be more prudent to recommend it over the latter. On the other hand, as each recommendation incurs uncertainty, when making a list of recommendations, it is more helpful to assess their quality collectively rather than individually. For example, the percentage of relevant recommendations among all recommended products could be a more meaningful measure than the chance of a specific recommendation being relevant. Both aspects draw immediate comparison with multiple testing problems, for example, in high-throughput gene expression studies where, among thousands of genes, a small subset that are likely to behave differently between control and treatment groups are sought. See, e.g., Storey and Tibshirani (2003); Efron (2007, 2012). Our work is inspired by this analogy and examines the problem of item recommendations from a multiple testing perspective.

For the sake of generality, we shall adopt the framework of trace regression where each observation is a random pair (X, Y) with $X \in \mathbb{R}^{d_1 \times d_2}$ and $Y \in \mathbb{R}$. The random matrix X is sampled uniformly from the orthonormal basis $\mathfrak{E} = \{e_i e_j^\top : 1 \leq i \leq d_1, 1 \leq j \leq d_2\}$ where $\{e_i\}$ is the canonical basis vectors of an Euclidean space of conformable dimensions. The response variable Y is related to X via

$$Y = \langle M, X \rangle + \xi \tag{1}$$

where $\langle M, X \rangle = \mathrm{tr}(M^\top X)$, and the independent measurement error ξ is assumed to be a 0-mean sub-Gaussian random variable. Our goal is to infer the true user-product preference matrix M from i.i.d. copies of (X, Y) when M is of (approximately) low rank and the observations are incomplete. Specifically, the task of deciding if product j should be recommended to user i can be cast as testing the null hypothesis, denoted by $H_{0,ij}$, about the (i, j) entry of the true user-product matrix M, e.g., product j is irrelevant to user i, against the alternative, denoted by $H_{a,ij}$, that user i is interested in product j. Likewise, item recommendations in general amount to testing collectively all null hypotheses $H_{0,ij}$, $1 \leq i \leq d_1$ and $1 \leq j \leq d_2$. More broadly, one may consider testing about multiple linear forms, $\langle M, T \rangle$ for a family of $T \in \mathcal{H} \subset \mathbb{R}^{d_1 \times d_2}$. For example, one may consider T of the form $e_i e_{j_1}^\top - e_i e_{j_2}^\top$ to determine between two products (j_1 and j_2) which one to recommend to a user (i). This multiple testing framework allows us to address, among others, two most pertinent questions for recommender systems: which items should we recommend so that we can ensure a certain percentage of recommendations are relevant, or click-through rate; given a list of recommendations, what percentage of recommendations are relevant. Both questions can be naturally rephrased in terms of the false discovery rate (FDR), commonly used in the context of multiple testing.

Since its introduction in the seminal paper by Benjamini and Hochberg (1995), FDR has proven to be an extremely useful notion in a wide variety of areas including bioinformatics (Jung, 2005; Roeder and Wasserman, 2009; Brzyski et al., 2017), neuroimaging (Perone Pacifico et al., 2004; Chumbley et al., 2010), and finance (Barras et al., 2010; Bajgrowicz and Scaillet, 2012), to name a few. Numerous methodologies have also been developed to control FDR in multiple testing. Notable examples includes Benjamini and Yekutieli (2001); Sarkar (2002); Wu (2008); Clarke and Hall (2009); Barber and Candès (2015); Candes et al. (2018); Barber and Candès (2019), among many others. There are, however, considerable new challenges when considering multiple testing in the context of item recommendations or matrix completion, both in defining test statistics for individual hypothesis and in how to utilize them effectively to improve the overall performance.

In most if not all of the existing literature of multiple testing, the individual test statistics are either given or naturally defined. For matrix completion, however, finding the right test statistics is arguably one of the most difficult steps for statistical inferences. Common estimators for entries of the underlying matrix do not admit an explicit expression, which creates technical obstacles to characterize their bias and variance. This challenge is already in full display when testing a single hypothesis which occurs, for example, when deciding on whether to recommend a specific product to a particular user. See, e.g., Chen et al. (2019); Xia and Yuan (2021); Farias et al. (2022); Chen et al. (2023); Gui et al. (2023); Shao and Zhang (2023). The problem is exacerbated when dealing with multiple hypotheses where more refined bounds for the convergence of test statistics are needed both for controlling the FDR and to ensure power without unnecessary sample size and signal-to-noise ratio restriction. We shall introduce a new test statistic especially suitable for such purposes. It builds upon recent developments (e.g., Chen et al., 2019; Xia and Yuan, 2021) for inferring a single entry and is based upon a more precise characterization of variance than earlier works. In particular, it can be shown that, with the improved variance estimate, the new statistic converges to normal distribution at a faster rate, both marginally and jointly, and is thus more suitable for use in multiple testing.

Most procedures for FDR control were developed, at least initially, assuming that the individual test statistics are independent of each other. How to handle complicated dependency structure, as is the case for matrix completion, remains a critical issue and an actively researched subject in multiple testing. See, e.g., Efron (2007); Leek and Storey (2008); Fan and Han (2017); Li and Zhong (2017); Du et al. (2023); Fithian and Lei (2022). A common strategy to deal with dependence is data splitting. See, e.g., Roeder and Wasserman (2009); Song and Liang (2015); Barber and Candès (2019); Zou et al. (2020); Du et al. (2023); Dai et al. (2022, 2023), for a number of recent examples and applications of data splitting schemes. In particular, Du et al.

(2023) showed that the FDR can be properly controlled as long as the individual test statistics have nearly symmetric null distribution and the dependence among them is sufficiently weak. To make use of this insight, we derive the asymptotic correlation of our proposed individual test statistics. Interestingly, for many item recommendation tasks, these statistics are only weakly correlated and hence, the FDR can be controlled accordingly. In other settings where the test statistics can be strongly correlated, our explicit characterization of their dependence structure also suggests ways to "whitening" and "screening" so that FDR can still be controlled under minimal sample size requirement.

The rest of the paper is organized as follows. In the next section, we shall introduce our test statistics for a single linear form and study its asymptotic properties. Section 3 discusses how these individual test statistics can be aggregated to test multiple linear forms. Section 4 introduces a whitening and screening scheme to address situations where the test statistics could be strongly correlated. Numerical experiments, both simulated and real-world data examples, are presented in Section 5. We conclude with a few remarks in Section 6. Due to space limitation, all proofs, as well as further examples and discussions, are relegated to the Supplement.

Throughout the paper, let $\|\cdot\|$ denote the spectral norm of a matrix and the ℓ_2-norm of a vector, and denote $\|M\|_{2,\max} := \max_{i\in d_1}\|e_i^\top M\|$. Define $\|M\|_{\max} = \max_{i,j}|M_{ij}|$ and $\|M\|_\infty := \max_{i\in[q]}\|e_i^\top M\|_{\ell_1}$ for a matrix M. Note that $\|\cdot\|_{\max}$ and $\|\cdot\|_\infty$ are equivalent for a vector. Also, we use $\|\cdot\|_{\mathrm{F}}$ to denote the Frobenius norm.

2 An Approach for Individual Test

We begin with testing a single hypothesis:

$$H_{0T}:\langle M,T\rangle=\theta_T \qquad \text{vs} \qquad H_{aT}:\langle M,T\rangle\neq\theta_T \tag{2}$$

for some fixed $T\in\mathbb{R}^{d_1\times d_2}$ and pre-specified value $\theta_T\in\mathbb{R}$, based on n independent observations $\mathcal{D}:=\{(X_i,Y_i)\}_{i=1}^n$ following the trace regression model (1). Recall that ξ in (1) is sub-Gaussian noise with mean 0 and variance σ_ξ such that $\mathbb{E}\exp(\lambda\xi)\le\exp(c^2\sigma_\xi^2\lambda^2/2)$ for some constant $c>0$. Following the convention, we shall assume that the singular vectors of M are incoherent:

$$\max\left\{\sqrt{\frac{d_1}{r}}\|U\|_{2,\max},\sqrt{\frac{d_2}{r}}\|V\|_{2,\max}\right\}\le\sqrt{\mu}, \tag{3}$$

where r is the rank of M, and $M=U\Lambda V^\top$ its singular value decomposition. This ensures that the entries of M are delocalized so that it can be recovered even if some entries are not observed. In what follows, we shall denote by $\lambda_{\max}$ and $\lambda_{\min}$ the largest and smallest nonzero singular values of M, and κ_0 the ratio $\kappa_0 := \lambda_{\max}/\lambda_{\min}$, i.e., its condition number.

For brevity, we focus on two-sided tests here, though our discussion extends straightforwardly to one-sided tests. Without loss of generality, we assume that $d_1 \geq d_2$ throughout. The goal of this section is to develop a test statistic for (2) that can be effectively applied to testing a large number of hypotheses. The problem of testing a single linear form (2) has been previously studied by Xia and Yuan (2021); see also Chen et al. (2019); Farias et al. (2022); Chen et al. (2023), among others. However, the tests proposed in these works are neither sufficiently sharp nor directly applicable to multiple testing. For example, Xia and Yuan (2021)'s debiasing approach relies on independent initialization through data splitting, which may lead to potential power loss. Although Chen et al. (2019); Farias et al. (2022) avoids data splitting by employing a leave-one-out analysis, their methods are restricted to the entrywise case where $T = e_i e_j^\top$, and they uses Bernoulli sampling with a constrained sample size $n \leq d_1 d_2$. Moreover, in both Xia and Yuan (2021), and Chen et al. (2019); Farias et al. (2022), the rate of asymptotic normal convergence is no faster than $\sqrt{\log d_1 / d_2}$. This convergence rate is too slow for our purposes, as it unnecessarily restricts the number of hypotheses that can be tested, regardless of how large the sample size n is.

We now propose a novel approach to address these issues mentioned above. Our approach for estimating $\langle M, T \rangle$ consists of three steps: gradient-descent initialization, bias-correction and an improved low-rank projection. For the initialization, we apply the gradient descent (Chen et al., 2020) to obtain an entry-wise consistent rank-r estimator $\widehat{M}^{\mathsf{init}}$ such that for any $\tau \geq 1$,

$$\left\|\widehat{M}^{\mathsf{init}} - M\right\|_{\max} \leq C_0 \sigma_\xi \sqrt{\frac{\tau d_1 \log d_1}{n}}, \tag{4}$$

with probability at least $1 - d_1^{-\tau}$, for some parameter $C_0 > 0$ that is only $\mathrm{Poly}(\tau, \kappa_0, \mu, r, \log d_1)$. Denote the left and right r singular vectors of $\widehat{M}^{\mathsf{init}}$ as $\widehat{U}^{\mathsf{init}}$, $\widehat{V}^{\mathsf{init}}$. A key observation is that $\widehat{U}^{\mathsf{init}}$, $\widehat{V}^{\mathsf{init}}$ are also incoherent. For brevity, in what follows, we shall assume τ is a large enough constant to ensure that $n \leq O(d_1^{2\tau})$. To correct the bias of initial estimates, we then define

$$\widehat{M}^{\mathsf{unbs}} = \widehat{M}^{\mathsf{init}} + \frac{d_1 d_2}{n} \sum_{i=1}^{n} \left(Y_i - \left\langle \widehat{M}^{\mathsf{init}}, X_i \right\rangle \right) X_i. \tag{5}$$

Unfortunately this debiasing may lead to a significant increase in variance, we shall trade off between bias and variance by low-rank projection. However, different from the popular simple SVD used in the literature (Chen et al., 2019; Xia and Yuan, 2021), we use an incoherence-assisted low-rank projection. More exactly, we compute

$$\widehat{U} = \mathrm{SVD}_r\big(\widehat{M}^{\mathsf{unbs}}\widehat{V}^{\mathsf{init}}\big) \quad \text{and} \quad \widehat{V} \ = \mathrm{SVD}_r\big((\widehat{M}^{\mathsf{unbs}})^\top \widehat{U}^{\mathsf{init}}\big), \tag{6}$$

where $\mathrm{SVD}_r(\cdot)$ returns the top-r left singular vectors. This yields a low-rank estimate:

$$\widehat{M} = \widehat{U}\widehat{U}^\top \widehat{M}^{\mathsf{unbs}} \widehat{V}\widehat{V}^\top. \tag{7}$$

Finally we shall estimate $\langle M, T\rangle$ by $\langle \widehat{M}, T\rangle$. Under certain regularity conditions, we can show that

$$\frac{\langle \widehat{M}, T\rangle - \langle M, T\rangle}{\sigma_\xi \left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}} \sqrt{d_1 d_2/n}} \to_d N(0,1), \tag{8}$$

where

$$\mathcal{P}_M(A) = UU^\top AVV^\top + UU^\top AV_\perp V_\perp^\top + U_\perp U_\perp^\top AVV^\top \tag{9}$$

represents the projection onto the tangent space of low-rank matrix manifold $\mathcal{M}_r$ at point M, and $U_\perp$ and $V_\perp$ are orthonormal matrices whose columns span the orthogonal complements of the left and right singular spaces of M, respectively.

We define the alignment ratio β_T (which may goes to 0 asymptotically) as

$$\beta_T := \frac{\left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}}{\left\|T\right\|_{\mathrm{F}}} \sqrt{\frac{d_2}{r}}, \tag{10}$$

where $\left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}$ is proportional to the (asymptotic) variance of the test statistic with respect to a linear form T. Here we assume $\left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}} > 0$ throughout the discussion. When $\left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}} = 0$, the linear form $\langle M, T\rangle = 0$ and estimates with faster rate of convergence can be obtained. This condition avoids such pathological situations. Similar assumptions are also made in earlier works. See, e.g., Chen et al. (2019); Xia and Yuan (2021).

We can use this result to test (2) for a fixed T. The theoretical guarantee of our approach is presented in Theorem 1:

Theorem 1. *Given $\widehat{M}$ from (7) with dimension ratio $\alpha_d = d_1/d_2$, suppose that the sample size and and SNR condition satisfies:*

$$n \geq C_1 C_0^2 \kappa_0^2 \frac{\left\|T\right\|_{\ell_1}^2}{\beta_T^2 \left\|T\right\|_{\mathrm{F}}^2} \mu^5 r^5 d_1 \log^2 d_1, \ \frac{\lambda_{\min}}{\sigma_\xi} \geq C_2 C_0^2 \frac{\kappa_0 \left\|T\right\|_{\ell_1}}{\beta_T \left\|T\right\|_{\mathrm{F}}} \sqrt{\frac{\alpha_d \mu^6 r^5 d_1^2 d_2 \log^2 d_1}{n}} \tag{11}$$

for some constants $C_1, C_2 > 0$. Then there exists a constant $C_3 > 0$ such that for any $T \in \mathbb{R}^{d_1 \times d_2}$

satisfying (10),

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\langle\widehat{M},T\rangle-\langle M,T\rangle}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\cdot\sqrt{d_1d_2/n}}\le t\right)-\Phi(t)\right|$$
$$\le C_3\left(C_0\frac{\kappa_0\left\|T\right\|_{\ell_1}}{\beta_T\left\|T\right\|_{\mathrm{F}}}\sqrt{\frac{\mu^5r^5d_1\log^2 d_1}{n}}+\frac{C_0^2\sigma_\xi\kappa_0\left\|T\right\|_{\ell_1}}{\beta_T\lambda_{\min}\left\|T\right\|_{\mathrm{F}}}\sqrt{\frac{\alpha_d\mu^6r^5d_1^2d_2\log^2 d_1}{n}}\right).$$

We highlight the advantage of our methods compared to existing literature: First, we apply the gradient descent directly to the observations for initialization, which avoids data splitting and potential power loss (Xia and Yuan, 2021). Second, we use a new low-rank projection method akin to the subspace iteration (Bathe and Wilson, 1973; Bathe, 2013), and tensor power iteration (Richard and Montanari, 2014). The benefit of this new projection method can be explained from a theoretical point of view: recall that $\widehat{M}^{\mathsf{unbs}}-M=\frac{d_1d_2}{n}\sum_{i=1}^n\xi_iX_i+\frac{d_1d_2}{n}\sum_{i=1}^n\left\langle\widehat{M}^{\mathsf{init}}-M,X_i\right\rangle X_i-(\widehat{M}^{\mathsf{init}}-M)$. When multiplied by incoherent $\widehat{U}^{\mathsf{init}}$ or $\widehat{V}^{\mathsf{init}}$ on one side, the size of each summand (i.e., $\widehat{U}^{\mathsf{init}\top}X_i$, or $X_i\widehat{V}^{\mathsf{init}}$) can be effectively reduced, leading to sharper concentrations. The leave-one-out argument enables further bounding the negligible bias terms in $\langle\widehat{M}-M,T\rangle$, making our theory more general than those in Chen et al. (2019).

Moreover, the variance that we characterize is sharper than previous methods (Chen et al., 2019; Xia and Yuan, 2021; Farias et al., 2022) in the sense that (i) as is shown in Ma and Xia (2024), it reaches the Cramér-Rao lower bound for the regression-based noisy matrix completion, indicating its efficiency; (ii) it returns a confidence interval with minimax optimal length, as is given in our following Theorem 2; and (iii) it allows the convergence rate to be totally controlled by the sample size n and SNR ($\lambda_{\min}/\sigma_\xi$), whenever d_1, d_2 is large or not. This convergence rate is especially helpful for multiple testing since the number of hypotheses we are interested may grow with respect to d_1, d_2.

Theorem 2 (Minimax optimal length of confidence interval)**.** *Define the parameter space as*

$$\mathbf{\Theta}=\Big\{M'\in\mathbb{R}^{d_1\times d_2}:\mathrm{rank}(M')\le r,(d_1\left\|U'\right\|_{2,\max}^2)\vee(d_2\left\|V'\right\|_{2,\max}^2)\le\mu r,$$
$$\lambda_{\min}(M')\ge\lambda_{\min},\kappa(M')\le\kappa_0,\left\|\mathcal{P}_{M'}(T)\right\|_{\mathrm{F}}\ge\left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}\Big\}.$$

Here $\mathcal{P}_{M'}(\cdot)$ means the projection onto the tangent space at M'. Consider the set of any valid $1-\alpha$ confidence interval with $\alpha<\frac{1}{4}$ as:

$$\mathcal{I}_\alpha(\mathbf{\Theta},T):=\left\{\mathrm{CI}_T^\alpha\left(M,\{(X_i,Y_i)\}_{i=1}^n\right)=[l,u]:\inf_{M\in\mathbf{\Theta}}\mathbb{P}(l\le\langle M,T\rangle\le u)\ge1-\alpha\right\},$$

where l, u are any functions of observations $\{(X_i, Y_i)\}_{i=1}^n$. Then, when the SNR satisfies

$$\frac{\lambda_{\min}}{\sigma_\xi} \geq C_{\mathsf{gap}} \kappa_0 \frac{\|T\|_{\ell_1}}{\beta_T \|T\|_{\mathrm{F}}} \sqrt{\frac{\mu^6 r^3 d_1^2 d_2}{n}}$$

for a numeric constant C_{gap}, the length of the confidence interval has the minimax lower bound:

$$\inf_{\mathrm{CI}_T^\alpha\left(M, \{(X_i,Y_i)\}_{i=1}^n\right) \in \mathcal{I}_\alpha(\boldsymbol{\Theta}, T)} \sup_{M \in \boldsymbol{\Theta}} \mathbb{E} L\left(\mathrm{CI}_T^\alpha\left(M, \{(X_i, Y_i)\}_{i=1}^n\right)\right) \geq c\sigma_\xi \|\mathcal{P}_M(T)\|_{\mathrm{F}} \sqrt{\frac{d_1 d_2}{n}}.$$

To pursue a fully data driven approach for hypothesis testing, we provide the following estimates for the variance term $\sigma_\xi^2 \|\mathcal{P}_M(T)\|_{\mathrm{F}}^2$:

$$\widehat{\sigma}_\xi^2 = \frac{1}{n} \sum_{i=1}^n \left(Y_1 - \left\langle \widehat{M}^{\mathsf{init}}, X_i \right\rangle\right)^2, \widehat{s}_T^2 = \left\|\mathcal{P}_{\widehat{M}^{\mathsf{init}}}(T)\right\|_{\mathrm{F}}^2,$$

where $\mathcal{P}_{\widehat{M}^{\mathsf{init}}}(\cdot)$ follows (9) by replacing U, V with $\widehat{U}^{\mathsf{init}}$, $\widehat{V}^{\mathsf{init}}$. Now, we define our test statistic formally as

$$W_T\left(\{(X_i, Y_i)\}_{i=1}^n\right) = \frac{\langle \widehat{M}, T\rangle - \theta_T}{\widehat{\sigma}_\xi \widehat{s}_T \cdot \sqrt{d_1 d_2 / n}}. \tag{12}$$

The following result shows that the asymptotic normality continues to hold using these variance estimates.

Theorem 3. *Under the condition that* (11) *holds, if H_{0T} is true, then*

$$\begin{aligned} &\sup_{t \in \mathbb{R}} \left|\mathbb{P}\left(W_T \leq t\right) - \Phi\left(t\right)\right| \\ &\leq C_3 \left(C_0 \frac{\kappa_0 \|T\|_{\ell_1}}{\beta_T \|T\|_{\mathrm{F}}} \sqrt{\frac{\mu^5 r^5 d_1 \log^2 d_1}{n}} + \frac{C_0^2 \sigma_\xi \kappa_0 \|T\|_{\ell_1}}{\beta_T \lambda_{\min} \|T\|_{\mathrm{F}}} \sqrt{\frac{\alpha_d \mu^6 r^5 d_1^2 d_2 \log^2 d_1}{n}} \right). \end{aligned}$$

3 Multiple Tests

We now turn our attention to testing a family of hypotheses $\{H_{0T} : T \in \mathcal{H}\}$ for a subset $\mathcal{H} \subset \mathbb{R}^{d_1 \times d_2}$. In particular, we can take $\mathcal{H} = \{e_i e_j^\top : 1 \leq i \leq d_1, 1 \leq j \leq d_2\}$ for testing preferences of all user-item pairs. Denote the number of tests $|\mathcal{H}| = q$. Without loss of generality, assume that the linear forms are linearly independent so that the q is no larger than $d_1 d_2$. Denote the null set by $\mathcal{H}_0$, i.e., $\mathcal{H}_0 = \{T \in \mathcal{H} : \langle M, T\rangle = \theta_T\}$ and the non–null set $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$, with cardinality q_0 and q_1 respectively. To establish the test statistics for all $\{H_{0T} : T \in \mathcal{H}\}$, we denote the smallest alignment parameter as $\beta_0 = \min_{T \in \mathcal{H}_0} \beta_T$, and write

$$h_n := C_0 \max_{T \in \mathcal{H}_0} \left\{ \frac{\kappa_0 \|T\|_{\ell_1}}{\beta_0 \|T\|_{\mathrm{F}}} \sqrt{\frac{\mu^5 r^5 d_1 \log^2 d_1}{n}} + \frac{C_0^2 \sigma_\xi \kappa_0 \|T\|_{\ell_1}}{\beta_0 \lambda_{\min} \|T\|_{\mathrm{F}}} \sqrt{\frac{\alpha_d \mu^6 r^5 d_1^2 d_2 \log^2 d_1}{n}} \right\}, \tag{13}$$

where, for brevity, we omit the dependence of h_n on d_1. In light of Theorem 3, with appropriate initial estimates, we have

$$\left|\mathbb{P}\left(W_T \leq t\right) - \Phi\left(t\right)\right| \lesssim h_n$$

for all $T \in \mathcal{H}_0$. In what follows, we assume $h_n \to 0$ to ensure proper inference.

3.1 Symmetric Data Aggregation

With the asymptotic normality of W_T, it looks possible to directly apply Benjamini and Hochberg (1995) style of methods to control the FDR in an asymptotic sense. However, doing so may put an unreasonable limit on the number (q) of tests under consideration. This is due to the fact that the test statistic W_T has much heavier tail than that in classic multivariate normal mean problems. As a result, while W_T converges to $\mathcal{N}(0,1)$ in distribution for any linear form T (as long as signal strength is large enough), it does not necessarily converge in fourth-order or higher-order moments. Indeed, it can be shown that the $2k$-th order moment ($k \geq 2$) of W_T for a properly chosen linear form T is lower bounded by

$$\sqrt[2k]{\mathbb{E}\left|W_T\right|^{2k}} \gtrsim \left(\frac{d_1 d_2}{n}\right)^{1/4}. \tag{14}$$

If $d_1 \asymp d_2 \asymp d$, and $n \asymp d^{1+\epsilon}$ for some $\epsilon \in (0,1)$, then we have $\mathbb{E}\left|W_T\right|^{2k} \gtrsim d^{(1-\epsilon)k/2}$. See supplement for proof of (14).

Thankfully, much more powerful approaches can be developed by exploiting other salient features of W_T entailed by its asymptotic normality. In particular, we shall adopt a general strategy introduced by Du et al. (2023) by leveraging symmetricity and data aggregation. Assume that, without loss of generality, n is even with $n = 2n_0$. We split $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ into two sub-samples:

$$\mathcal{D}_1 = \left\{(X_i, Y_i)\right\}_{i=1}^{n_0} \quad \text{and} \quad \mathcal{D}_2 = \left\{(X_i, Y_i)\right\}_{i=n_0+1}^{n}.$$

We describe our approach as following: we first construct two groups of independent asymptotic symmetric statistics $\{W_T^{(1)} : T \in \mathcal{H}\}$ from $\mathcal{D}_1$ and $\{W_T^{(2)} : T \in \mathcal{H}\}$ from $\mathcal{D}_2$ by data splitting. After that, we aggregate them by multiplication: $W_T^{\mathsf{rank}} = W_T^{(1)} \cdot W_T^{(2)}$. Finally, we rank each W_T^{rank} and choose a data-driven threshold by taking advantage of symmetricity:

$$L := \inf\left\{t > 0 : \frac{\#\left\{T : W_T^{\mathsf{rank}} < -t\right\}}{\#\left\{T : W_T^{\mathsf{rank}} > t\right\} \vee 1} \leq \alpha\right\} \cup \{+\infty\}, \tag{15}$$

given any FDR level $\alpha \in (0,1)$, and reject H_{0T} if $W_T^{\mathsf{rank}} > L$. Details are given in Algorithm 1. Hereafter, we denote $M_T := \langle M, T\rangle$ for simplicity.

```
Algorithm 1 Matrix FDR Control
Require: Hypotheses $\{H_{0T} : M_T = \theta_T, T \in \mathcal{H}\}$, data splits $\mathcal{D}_1$, $\mathcal{D}_2$, rank $r$, FDR level $\alpha$.
1: Apply gradient descent on $\mathcal{D}_1$, $\mathcal{D}_2$ to construct initial estimates $\widehat{M}^{(1)}_{\mathsf{init}}$, $\widehat{M}^{(2)}_{\mathsf{init}}$, respectively.
2: Apply (7), (12) to construct two independent test statistics: $W_T^{(1)} := W_T(\mathcal{D}_1)$, $W_T^{(2)} = W_T(\mathcal{D}_2)$.
3: Compute the final ranking statistics by $W_T^{\mathsf{rank}} = W_T^{(1)} W_T^{(2)}$, and then choose a data-driven threshold $L$ by (15).
4: Reject $H_{0T}$ if $W_T^{\mathsf{rank}} > L$.
```

By the definition of L, we have

$$\mathrm{FDP} = \frac{\sum_{T\in\mathcal{H}} \mathbb{I}(W_T^{\mathsf{rank}} < -L)}{\left(\sum_{T\in\mathcal{H}} \mathbb{I}(W_T^{\mathsf{rank}} > L)\right) \vee 1} \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{\sum_{T\in\mathcal{H}} \mathbb{I}(W_T^{\mathsf{rank}} < -L)} \le \alpha \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} < -L)}.$$

The crux of our argument is that the ratio on the rightmost hand side is approximately 1 by virtue of the symmetry of W_T^{rank} under H_{0T}. To do so, we need to investigate the dependence among multiple test statistics, which will be elaborated later.

We remark that, our data aggregation scheme resembles the "mirror-statistic" in the literature (Xing et al., 2021; Dai et al., 2022), which choose the new sign to be $\mathrm{sign}(W_T^{\mathsf{new}}) = \mathrm{sign}(W_T^{(1)}) \cdot \mathrm{sign}(W_T^{(2)})$, and get absolute value by combination. In addition to multiplying the two values $|W_T^{\mathsf{new}}| = \left|W_T^{(1)} W_T^{(2)}\right|$, other ways of getting new absolute value, including $\min\left\{\left|W_T^{(1)}\right|, \left|W_T^{(2)}\right|\right\}$ and $\left|W_T^{(1)}\right| + \left|W_T^{(2)}\right|$, have also been studied earlier by Xing et al. (2021); Dai et al. (2022, 2023). Our choice of the multiplicative data aggregation is motivated by an observation that for testing about the multivariate normal mean, it can be more powerful than the other two choices. See supplement for detailed discussion.

3.2 Dependence among Test Statistics

One of the main challenges for multiple testing is how to account for the dependence structure among test statistics. To this end, we shall first derive the asymptotic distribution for the joint distribution of two estimated linear forms. In particular, for two matrices $T_1, T_2 \in \mathbb{R}^{d_1\times d_2}$, it can be shown that

$$\mathtt{corr}(\langle \widehat{M}, T_1\rangle, \langle \widehat{M}, T_2\rangle) \approx \frac{\langle \mathcal{P}_M(T_1), \mathcal{P}_M(T_2)\rangle}{\|\mathcal{P}_M(T_1)\|_{\mathrm{F}} \|\mathcal{P}_M(T_2)\|_{\mathrm{F}}} =: \rho_{T_1,T_2}. \tag{16}$$

More specifically, we have

Theorem 4. *Suppose* (11) *holds for* $T_1, T_2 \in \mathcal{H}$, *and* $|\rho_{T_1,T_2}| < 1$. *Define* $\Phi_\rho(\cdot,\cdot)$ *as the cumulative distribution function of bivariate normal distribution* $N(0,((1,\rho)^\top,(\rho,1)^\top))$. *If both* H_{0T_1} *and* H_{0T_2} *hold, then*

$$\sup_{t_1,t_2\in\mathbb{R}} \left|\mathbb{P}\left(W_{T_1} \le t_1, W_{T_2} \le t_2\right) - \Phi_{\rho_{T_1,T_2}}\left(t_1,t_2\right)\right| \le C_3(1-\rho_{T_1,T_2})^{-\frac{3}{2}} h_n.$$

This result explicitly characterizes the dependence between two test statistics which is critical for the FDR control in multiple testing. In particular, we shall separate pairs of linear forms in the null hypotheses into strongly correlated:

$$\mathcal{H}^2_{0,\text{strong}} := \left\{(T_1,T_2) \in \mathcal{H}_0 \times \mathcal{H}_0 : \rho_{T_1,T_2} \ge c q_0^{-\nu}\right\}, \tag{17}$$

where $\nu > 0$ can be any pre-specified non-vanishing number (e.g., $\nu \ge 0.1$) and $c > 0$ is some universal constant, and weakly correlated $\mathcal{H}^2_{0,\text{weak}} := (\mathcal{H}_0 \times \mathcal{H}_0) \setminus \mathcal{H}^2_{0,\text{strong}}$. The proportion of all linear form pairs that are strongly correlated is therefore

$$\beta_{\mathsf{s}} := \frac{\left|\mathcal{H}^2_{0,\text{strong}}\right|}{|\mathcal{H}^2_0|}. \tag{18}$$

Here, β_{s} measures how dependent the test statistics are. When $\beta_{\mathsf{s}} \to 0$, most of the $\{W_T\}_{T\in\mathcal{H}_0}$ are weakly correlated, leading to weak dependency among all test statistics.

We demonstrate that the measure of dependency in (18) is particularly useful for matrix linear form inference as the incoherent structure naturally ensures small β_{s} for many practical instances. Under the incoherent assumption, we have

$$\rho_{T_1,T_2} \le \frac{\mu^4 r \left\|T_1\right\|_{\ell_1} \left\|T_2\right\|_{\ell_1}}{\beta_0^2 \left\|T_1\right\|_{\mathrm{F}} \left\|T_2\right\|_{\mathrm{F}}} \frac{1}{d_2} + \frac{\left|\left\langle T_1 T_2^\top, UU^\top\right\rangle\right| + \left|\left\langle T_1^\top T_2, VV^\top\right\rangle\right|}{\left\|\mathcal{P}_M(T_1)\right\|_{\mathrm{F}} \left\|\mathcal{P}_M(T_2)\right\|_{\mathrm{F}}}.$$

Thus, two linear forms (T_1, T_2) are weakly correlated if $T_1^\top T_2 = \mathbf{0}$, $T_1 T_2^\top = \mathbf{0}$ and

$$\frac{\mu^2 \left\|T_1\right\|_{\ell_1} \left\|T_2\right\|_{\ell_1}}{\beta_0^2 \left\|T_1\right\|_{\mathrm{F}} \left\|T_2\right\|_{\mathrm{F}}} \le C. \tag{19}$$

Condition (19) holds when T_1, T_2 are sparse, i.e., the number, s_0, of nonzero entries in T_1 and T_2 is of the order $O(\beta_0^2)$. Note that these conditions concern the linear forms only and do not depend on M. We can use this to show that in the following practical examples related to item recommendations, the linear forms are weakly correlated (given $\beta_0 \gtrsim 1$), regardless of the underlying matrix M:

Inference of a submatrix. Consider the inference problem with indexing matrices $\mathcal{H} = \{e_i e_j^\top : l_1 \le i \le l_2, l_3 \le j \le l_4\}$, where $l_2 - l_1 \asymp d_1$, $l_4 - l_3 \asymp d_2$. This can represent recommendation tasks in problems including Netflix prize (Bennett and Lanning, 2007), or gene-disease association discovery (Natarajan and Dhillon, 2014), among others. Here we have the number of tests of order $O(d_1 d_2)$. Since $\|T\|_{\ell_1}/\|T\|_{\mathrm{F}} = 1$ for any $T \in \mathcal{H}$, condition (19) is easily satisfied. Therefore, at most $O(d_1)$ pairs are strongly correlated (share the same row/column) for each linear form so that $\beta_{\mathsf{s}} \lesssim 1/d_2$ for any $\nu < 0.5$.

Inference of entrywise comparisons. We can also consider comparison between two entries $M_{i,j}$ and $M_{i+1,j}$: $\mathcal{H} = \{e_i e_j^\top - e_{i+1} e_j^\top : l_1 \le i \le l_2, l_3 \le j \le l_4\}$. If $l_2 - l_1 \asymp d_1$, $l_4 - l_3 \asymp d_2$, then the total number of tests is of the order $O(d_1 d_2)$. Similar to before, $\|T\|_{\ell_1}/\|T\|_{\mathrm{F}} = \sqrt{2}$ for any $T \in \mathcal{H}$ so that there are at most $O(d_1^2 d_2)$ pairs that can be strongly correlated (share the same row/column) for each linear form. This again yields $\beta_{\mathsf{s}} \lesssim 1/d_2$ for any $\nu < 0.5$.

Inference of several user/feature groups. For many applications, groupwise recommendation (Bi et al., 2018) is of interest. This can be formulated as testing $H_{0T} : \sum_{i \in G_k} M_{ij} \le \theta_{kj}$ vs $H_{1T} : \sum_{i \in G_k} M_{ij} > \theta_{kj}$, where $(G_1, \ldots, G_K)$ is a partition of the $[d_1]$. In other words $\mathcal{H} = \{\sum_{i \in G_k} e_i e_j^\top : 1 \le k \le K, 1 \le j \le d_2\}$. Note that $\|T\|_{\ell_1}/\|T\|_{\mathrm{F}} = \sqrt{|G_k|}$ for all $T \in \mathcal{H}$. If $K = \Omega(d_2)$, then for any $\nu < 0.5$,

$$\beta_{\mathsf{s}} \lesssim \frac{d_2 K (K + d_2)}{d_2^2 K^2} \lesssim \frac{1}{d_2}.$$

3.3 Theoretical Guarantees

A crucial aspect to understand the efficacy of a multiple testing procedure is the signal strength of the non-null set, i.e., $|\langle M, T \rangle - \theta_T|$ for $T \in \mathcal{H}_1$. Recall that for any matrix completion estimator $\widetilde{M}$, the best entrywise error rate we can attain is $\|\widetilde{M} - M\|_{\max} \lesssim_p \sigma_\xi \sqrt{d_1 \log(d_1)/n}$ (Koltchinskii et al., 2011). In the case of gradient descent estimator $\widehat{M}^{\mathsf{init}}$, with high probability, one can expect

$$\left| \left\langle \widehat{M}^{\mathsf{init}} - M, T \right\rangle \right| \le \left\| \widehat{M}^{\mathsf{init}} - M \right\|_{\max} \|T\|_{\ell_1} \le C_0 \sigma_\xi \sqrt{\frac{d_1 \log d_1}{n}} \|T\|_{\ell_1}.$$

Thus, we say that a signal can be consistently identified if

$$\frac{|\langle M, T \rangle - \theta_T|}{\|T\|_{\ell_1} \sqrt{\log d_1}} \ge C_{\mathsf{gap}} \cdot C_0 \sigma_\xi \sqrt{\frac{d_1 \log d_1}{n}} \tag{20}$$

for a sufficiently large constant $C_{\mathsf{gap}} > 0$. Denote by $\mathcal{S}$ the set of all linear forms $T \in \mathcal{H}$ such that (20) holds, with its cardinality as $\eta_n := |\mathcal{S}|$. Note that β_s and $\eta_n = |\mathcal{S}|$ are the most essential quantities in characterizing the effectiveness of FDR control and power guarantee for multiple testing. We are now in position to state our main result.

Theorem 5. *Suppose that $h_n \to 0$ and*

$$\left(\sqrt{\beta_{\mathsf{s}}} \vee h_n\right) \frac{q_0}{\eta_n} \to 0.$$

Then, we have

$$\text{FDP} := \frac{\sum_{T \in \mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{\left(\sum_{T \in \mathcal{H}} \mathbb{I}(W_T^{\mathsf{rank}} > L)\right) \vee 1} \leq \alpha(1 + o_p(1))$$

and

$$\text{POWER} := \frac{\sum_{T \in \mathcal{H}_1} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{q_1} \geq \frac{\eta_n}{q_1}(1 - o_p(1)).$$

The first claim implies that

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \alpha(1 + o(1)), \tag{21}$$

which can be used for control of FDR. On the other hand, if nearly all signals are strong in that $\eta_n / q_1 \to 1$, then the second claim indicates that POWER $\to_p 1$.

For clarity, we stated the asymptotic bounds for FDP and POWER in Theorem 5. Our proof actually establishes stronger results in a nonasymptotic form. Theorem 5 is a direct consequence of these nonasymptotic results that will be presented in the supplement. It is also worth noting that both the sample size and signal-to-noise ratio (implied by the condition on h_n) requirements of Theorem 5 are comparable to those for estimation (Keshavan et al., 2010b; Ma et al., 2018; Xia and Yuan, 2021). This immediately suggests that we can effectively control FDR under conditions of weak correlation, provided the underlying matrix can be consistently recovered.

We remark that (i) the condition $\left(\sqrt{\beta_{\mathsf{s}}} \vee h_n\right) \frac{q_0}{\eta_n} \to 0$ is easily satisfied by the three practical examples discussed above, provided that the proportion of identified signals is not too small (i.e., $\frac{\eta_n}{q_0} \gg \frac{1}{d_2} \vee h_n$). Moreover, (ii) the threshold (15) ensures FDR control only when W_T^{rank} is highly likely to be positive for $T \in \mathcal{H}_1$. Indeed, if W_T^{rank} are negative for many $T \in \mathcal{H}_1$, we will yield an overly large L, which leads to poor tail symmetry of $\{W_T\}_{T \in \mathcal{H}_0}$, leaving the FDR unbounded. For instance, simple asymptotic normal-based methods may yield negative test statistics for non-null hypotheses when $\langle M, T \rangle - \theta_T < 0$, for $T \in \mathcal{H}_1$. In contrast, we favor symmetric data aggregation, which tends to keep W_T^{rank} positive because $W_T^{(1)}$, $W_T^{(2)}$ usually share the same sign when $T \in \mathcal{H}_1$. This preference for symmetric data aggregation, along with its power boost, is also highlighted and justified in Dai et al. (2022); Du et al. (2023).

4 Whitening and Screening

Theorem 5 shows that the symmetric data aggregation method can control FDR effectively if the number of strongly correlated linear form pairs is sufficiently small relative to the number of strong signals, i.e., $\sqrt{\beta_{\mathsf{s}}}q_0/\eta_n \to 0$. While this is plausible in many applications, as we have argued, there are also situations in which this may not be the case. We now discuss how this condition can be further relaxed thanks to the explicit characterization of the correlation among test statistics. In particular, as advocated by Du et al. (2023), we proceed to apply symmetric data aggregation after appropriate screening and whitening. Interestingly, by exploiting the explicit characterization of the dependence among W_Ts, we can develop a more general and intuitive theoretical framework to study the power and FDR control for matrix completion.

More specifically, denote the collection of test statistics obtained from Algorithm 1 as $Z^{(i)} = \left[W_{T_1}^{(i)}, W_{T_2}^{(i)}, \ldots, W_{T_q}^{(i)}\right]^\top \in \mathbb{R}^q$, for $i = 1, 2$. By Theorem 4, $Z^{(i)} \approx_d N(\mathsf{w}, R)$ where $\mathsf{w} \in \mathbb{R}^q$ with the i-th entry $\mathsf{w}_i = \big(\langle M, T_i\rangle - \theta_{T_i}\big)/\big(\sigma_\xi \|\mathcal{P}_M(T_i)\|_{\mathrm{F}} \sqrt{d_1 d_2/n}\big)$ and $R = (\rho_{T_j,T_k})_{1\le j,k\le q}$. If R is known, then $R^{-1/2}Z^{(i)} \approx_d N(R^{-1/2}\mathsf{w}, I_q)$ has asymptotically independent coordinates and thus allows for better FDR control. However, such a whitening step can also mask the nonzero coordinates of w, which, as suggested by Du et al. (2023), can be estimated by Lasso. Of course, ρ_{T_j,T_k} is unknown, but it can nonetheless be estimated by

$$\widehat{\rho}_{T_j,T_k} = \frac{\left\langle \mathcal{P}_{\widehat{M}_{\mathrm{init}}}(T_j), \mathcal{P}_{\widehat{M}_{\mathrm{init}}}(T_k)\right\rangle}{\left\|\mathcal{P}_{\widehat{M}_{\mathrm{init}}}(T_j)\right\|_{\mathrm{F}} \left\|\mathcal{P}_{\widehat{M}_{\mathrm{init}}}(T_k)\right\|_{\mathrm{F}}}.$$

In summary, we shall consider the following algorithm detailed in Algorithm 2. Here, to ensure valid inversion of the population correlation matrix R, we confine q as $q \le (d_1 + d_2)r - r^2$, and assume $T \in \mathcal{H}$ are linearly independent.

Here $\widehat{R}_{\mathcal{A}}^{-1/2}$ is the submatrix of $\widehat{R}^{-1/2}$ with only columns indexed by $\mathcal{A}$. Similarly, $\widehat{\mathsf{w}}_{\mathcal{A}}$ is the subvector of $\widehat{\mathsf{w}}$ with only coordinates indexed by $\mathcal{A}$. Note that Algorithm 1 can be treated as a special case of Algorithm 2 by choosing the regularization parameter $\lambda = 0$. However, as we argue below, with an appropriate choice of $\lambda > 0$, the whitening and screening may lead to a more effective multiple testing procedure. In addition, a more concrete example of testing entries of submatrix of M is given in the supplement to demonstrate the impact of whitening and screening.

It is clear that the efficacy of Algorithm 2 hinges upon the reduction of dependence among test statistics with Lasso screening. We can show that, under mild regularity conditions, the

Algorithm 2 Matrix FDR Control with Whitening and Screening

```
Require: Hypotheses $\{H_{0T_i} : M_{T_i} = \theta_{T_i}, i \in [q]\}$, data splits $\mathcal{D}_1$, $\mathcal{D}_2$, rank $r$, FDR level $\alpha$, regularization parameter $\lambda \geq 0$.
Apply Algorithm 1 to get $Z^{(1)} \in \mathbb{R}^q$, $Z^{(2)} \in \mathbb{R}^q$ from $\{\mathcal{D}_1\}$ and $\{\mathcal{D}_2\}$ respectively
From $\mathcal{D}_1$, obtain a covariance estimate $\widehat{R} = (\widehat{\rho}_{T_i,T_j})_{i,j=1}^q$ using $\widehat{M}_{\mathsf{init}}^{(1)}$ estimated from Algorithm 1, that is
    $$\widehat{\rho}_{T_i,T_j} = \frac{\left\langle \mathcal{P}_{\widehat{M}_{\mathsf{init}}^{(1)}}(T_i), \mathcal{P}_{\widehat{M}_{\mathsf{init}}^{(1)}}(T_j) \right\rangle}{\left\| \mathcal{P}_{\widehat{M}_{\mathsf{init}}^{(1)}}(T_i) \right\|_{\mathrm{F}} \left\| \mathcal{P}_{\widehat{M}_{\mathsf{init}}^{(1)}}(T_j) \right\|_{\mathrm{F}}}.$$
    And solve Lasso estimator
    $$\widehat{\mathsf{w}}^{(1)} := \underset{\mathsf{w} \in \mathbb{R}^q}{\arg\min} \left\{ \frac{1}{2} \left\| \widehat{R}^{-1/2}(Z^{(1)} - \mathsf{w}) \right\|^2 + \lambda \left\| \mathsf{w} \right\|_{\ell_1} \right\}.$$
Denote $\mathcal{A} := \mathrm{supp}(\widehat{\mathsf{w}}^{(1)})$ the support of $\widehat{\mathsf{w}}^{(1)}$. Run linear regression on $\mathcal{A}$ with new design matrix $\widehat{R}_{\mathcal{A}}^{-1/2}$ and response $\widehat{R}^{-1/2} Z^{(2)}$ to get asymptotically symmetric statistics $\widehat{\mathsf{w}}^{(2)}$, where
    $$\widehat{\mathsf{w}}_{\mathcal{A}}^{(2)} := \left( \widehat{R}_{\mathcal{A}}^{-1/2\top} \widehat{R}_{\mathcal{A}}^{-1/2} \right)^{-1} \widehat{R}_{\mathcal{A}}^{-1/2\top} \widehat{R}^{-1/2} Z^{(2)} \quad \text{and} \quad \widehat{\mathsf{w}}_{\mathcal{A}^c}^{(2)} = 0$$
    with variance estimate $\widehat{\sigma}_{\mathsf{w}i}^2 := e_i^\top \left( \widehat{R}_{\mathcal{A}}^{-1/2\top} \widehat{R}_{\mathcal{A}}^{-1/2} \right)^{-1} e_i$ for $i \in \mathcal{A}$.
Compute the final ranking statistics of each $T_i$ by $\mathsf{w}_{T_i}^{\mathsf{rank}} = \widehat{\mathsf{w}}_i^{(1)} \widehat{\mathsf{w}}_i^{(2)} / \widehat{\sigma}_{\mathsf{w}i}$, and then choose a data-driven threshold $L$ by
    $$L := \inf \left\{ t > 0 : \frac{\sum_{i=1}^q \mathbb{I}\left( \mathsf{w}_{T_i}^{\mathsf{rank}} < -t \right)}{\sum_{i=1}^q \mathbb{I}\left( \mathsf{w}_{T_i}^{\mathsf{rank}} > t \right) \vee 1} \leq \alpha \right\}.$$
Reject $H_{0T_i}$ if $\mathsf{w}_{T_i}^{\mathsf{rank}} > L$
```

asymptotic covariance matrix of $\widehat{\mathsf{w}}_{\mathcal{A}}^{(2)}$ is given by

$$Q^* := \left(R_{\mathcal{A}}^{-1/2\top} R_{\mathcal{A}}^{-1/2}\right)^{-1}.$$

Similar to before, write

$$\mathcal{H}^2_{0\mathcal{A},\text{strong}} = \left\{(T_i, T_j) \in \mathcal{A}_0 \times \mathcal{A}_0 : \left|Q^*_{jk}\right| / \sqrt{Q^*_{kk} Q^*_{jj}} \geq c|\mathcal{A}|^{-\nu}\right\},$$

where $\mathcal{A}_0 = \mathcal{A} \cap \mathcal{H}_0$. Denote by

$$\beta'_{\mathsf{s}} := \frac{\left|\mathcal{H}^2_{0\mathcal{A},\text{strong}}\right|}{|\mathcal{A}_0|^2}.$$

In other words, β'_{s} represents the proportion of strongly correlated pairs after whitening and screening. Likewise, we shall write $\eta'_n = |\mathcal{S}'|$ where $\mathcal{S}'$ is the set of strong signals to be defined. To define strong signal, write

$$T_{\mathcal{H}} = \begin{bmatrix} \text{Vec}(T_1)^\top \\ \text{Vec}(T_2)^\top \\ \vdots \\ \text{Vec}(T_q)^\top \end{bmatrix} \in \mathbb{R}^{q \times d_1 d_2} \tag{22}$$

Then the limiting covariance matrix of W_Ts is given by

$$\Sigma := \left(\langle \mathcal{P}_M(T_j), \mathcal{P}_M(T_k)\rangle\right)_{1 \leq j,k \leq q} = T_{\mathcal{H}}(I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top) T_{\mathcal{H}}^\top.$$

We assume that Σ is invertible with $\lambda_{\min}(\Sigma) \geq c$ for some small constant c, and denote the condition number $\kappa_1 = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$. We can then define new strong signals as:

$$\mathcal{S}' = \left\{T \in \mathcal{H} : \frac{|\langle M, T\rangle - \theta_T|}{\|T\|_{\ell_1} \sqrt{q_1 \log d_1}} \geq C_{\mathsf{gap}} \cdot C_0 \kappa_1^{3/2} \sqrt{\frac{d_1 \log d_1}{n}}\right\}, \tag{23}$$

with its cardinality as $\eta'_n = |\mathcal{S}'|$. We have the following theoretical guarantee for Algorithm 2.

Theorem 6. *Let $T_{\mathcal{H}}$ be a $q \times d_1 d_2$ matrix with i-th row being $\text{vec}(T_i)$ and define $\text{supp}(T_{\mathcal{H}}) := \cup_{i=1}^q \text{supp}(T_i)$. Suppose that q'_0 a uniform upper bound for $|\mathcal{A}_0|$ and*

$$\left(\sqrt{\beta'_{\mathsf{s}}} \vee \left(h_n + \left\|\mathsf{w}_{\mathcal{A}^c}\right\|_\infty\right)\right) \frac{q'_0}{\eta'_n} \xrightarrow{p} 0,$$

and

$$\lambda_{\min} \gg C_0 \left(\left\|R^{-1}\right\|_\infty + \frac{\|T_{\mathcal{H}}\|}{\|T_{\mathcal{H}}\|_{2,\max}} \left(|\operatorname{supp}(T_{\mathcal{H}})| \wedge \sqrt{d_2}\right)\right) \max_{T \in \mathcal{H}} \left\{\frac{\|T\|_{\ell_1}}{\|T\|_{\mathrm{F}}}\right\} \sigma_\xi \sqrt{\frac{q d_1^3 \log d_1}{n}}.$$

Then there exists universal constant $C_4 > 0$ such that if regularization parameter $\lambda = C_4\sqrt{\log d_1}$ in Algorithm 2, then

$$\mathrm{FDP} = \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(\mathsf{w}_T^{\mathsf{rank}} > L)}{\left(\sum_{T\in\mathcal{H}} \mathbb{I}(\mathsf{w}_T^{\mathsf{rank}} > L)\right) \vee 1} \leq \alpha(1 + o_p(1))$$

and

$$\mathrm{POWER} = \frac{\sum_{T\in\mathcal{H}_1} \mathbb{I}(\mathsf{w}_T^{\mathsf{rank}} > L)}{q_1} \geq \frac{\eta_n'}{q_1}(1 - o_p(1)).$$

Note that the covariance matrix $\Sigma = \big(\langle \mathcal{P}_M(T_i), \mathcal{P}_M(T_j)\rangle\big)_{i,j\in[q]}$ is not known and our whitening procedure uses an estimate in its place. The additional lower bound of $\lambda_{\min}$ in Theorem 6 is in place to ensure that the estimated covariance matrix indeed can be used to "whiten" the test statistics. It is also worth pointing out that we do not require the sure-screening condition of Lasso. Such conditions are common in the literature. See , e.g., Roeder and Wasserman (2009); Barber and Candès (2019); Du et al. (2023); Dai et al. (2023). For our purpose, weak signals can be entertained as long as $\|\mathsf{w}_{\mathcal{A}^c}\|_\infty$ is sufficiently small. We also note that the sample size n in our matrix completion problem has a fundamentally different meaning compared to the classical regression problem. Due to incomplete observations, each sample point provides only limited information for inferring a matrix. Even for a single linear form, constructing a test statistic requires at least $n \gg d_1 \log^2 d_1$ samples from Theorem 3. Therefore, given $q \leq d_1 d_2$, one cannot expect n to be logarithmically dependent on q, as in the classical regression problem.

5 Numerical Experiments

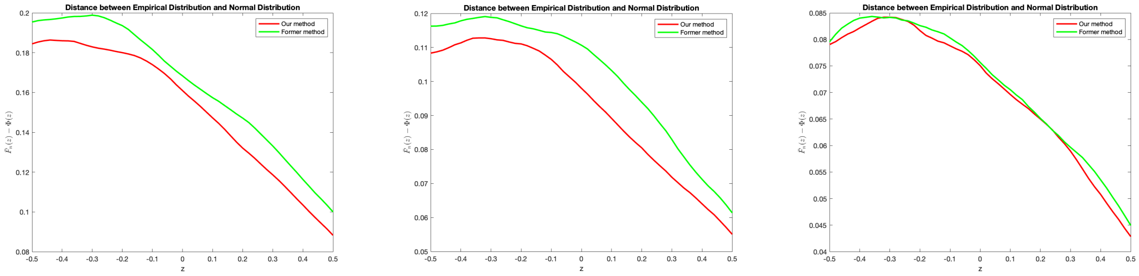
To complement our theoretical development, we also conducted several sets of numerical experiments to further demonstrate the practical merits of the proposed methodology.

5.1 Simulation Studies

We begin with a series of simulation studies aimed at illustrating the impact of several key aspects of our approach. All the simulations in this section display the averaged performance of multiple independent runs.

5.1.1 Variance of linear forms

In Section 2, we have presented the asymptotic normal test statistics for linear forms with a more accurate characterization of its variance. To justify the accuracy of our variance $\|\mathcal{P}_M(T)\|_{\mathrm{F}}$, we show the simulation of empirical distribution functions of our test statistics W_T in Theorem 1 against former test statistic in (8) whose variance is characterized by $(\|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2)^{1/2}$ in Xia and Yuan (2021). We plot the difference between empirical distribution functions $\bar{F}_n(z)$ and standard normal distribution function $\Phi(z)$ by sampling 10,000 independent realizations of test statistics. The result is shown in Figure 1. It is clear that our methods share a more precise asymptotic normal rate given smaller errors of $\bar{F}_n(z) - \Phi(z)$, especially for small sample size N.



(a) $n = 2400$

(b) $n = 3000$

(c) $n = 3600$

Figure 1: The difference between empirical distribution functions and $\Phi(z)$. Here, we compare our W_T with the former method (Xia and Yuan, 2021). We set the matrix with $d_1 = d_2 = \lambda_{\min} = 400$, and $r = 3$, and vary the number of random samples n in noisy matrix completion.

5.1.2 Data aggregation under weak dependency

We first evaluate our Algorithm 1 by simulations to corroborate two important properties of the proposed method: (1) the validity of FDR control for multiple testing of linear forms; (2) the power boost by data splitting and data aggregation; we randomly sample a low-rank matrix of dimension $d_1 = d_2 = 1000$, rank $r = 3$, with signal strength $\lambda_{\min} = 1000$. The number of observations used for Algorithm 1 is $n = 50rd_1$, and the noises $\xi \sim N(0, 1^2)$. We use the gradient descent (Wei et al., 2016; Chen et al., 2020; Cai et al., 2022) as initialization. We first verify the FDR control in weak dependency by performing blockwise matrix tests: we test each entry in $M(1:200, 1:200)$ by $H_{0,ij} : M_{ij} - m_{ij} = 0$ versus $H_{1,ij} : M_{ij} - m_{ij} \neq 0$. We randomly assign non-null hypotheses to these $200 \times 200 = 40,000$ entries with probability $p = 0.2$, which leads to the following settings of m_{ij}:

$$M_{ij} - m_{ij} = \begin{cases} \mu_{ij}, & \text{with probability } p = 0.2; \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

Here μ_{ij} are randomly-generated signals with fixed absolute mean: $\mathbb{E}\,|\mu_{ij}| = \mu$. We run Algorithm 1 and compare different methods of data aggregation (see Section A.5 for more details): I. multiplication; II. minimum absolute value with sign multiplication; III. adding absolute values with sign multiplication; IV. BHq with no data splitting. Here BHq with no data splitting means that we use data $\mathcal{D}_1$ and $\mathcal{D}_2$ together to construct asymptotic normal test statistics and then compute their p-values by the normal distribution. More specifically, we describe the BHq selection for linear forms as follows:

1. Use $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2\}$ to construct an initial estimate $\widehat{M}_{\mathsf{init}}$
2. Following the construction of W_T, but use both $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2\}$ to de-bias $\widehat{M}_{\mathsf{init}}$
3. Project the debiased matrix on the low-rank structure and get test statistics W_T^{all} for each linear form T.
4. Computing two-sided p-value $P_i = 2(1 - \Phi(\left|W_{T_i}^{\mathsf{all}}\right|))$
5. Feature selection by BHq method: finding the largest k such that $P_{(k)} \leq \frac{k}{q}\alpha$, and rejecting null hypothesis H_{0,T_i} with $P_i \leq P_{(k)}$.

This BHq selection relies on the asymptotic normality of high-dimensional features and serves as a counterpart to our methods. The result presented in Figure 2 clearly shows the excellent performance of multiplication in data aggregation with respect to both FDR control and power. By Section 3.2, the blockwise matrix entries tests here can be treated as the weakly correlated case. Although the BHq method Benjamini and Hochberg (1995) is guaranteed to be effective in the classical regression model, it fails to control the FDR at level α in our matrix completion problem. The reason this happens might be due to the large number of tests q and the heavy-tail property of W_T described in 14.

5.1.3 Whitening and screening

We now evaluate Algorithm 1 and Algorithm 2 and show the advantages of de-correlation. To this end, we still adopt the data generation mechanism in the previous section, but apply our methods to the entry comparisons between rows: we test $q = 400$ differences between first row $M(1, 1:400)$ and second row $M(2, 1:400)$, with H_{0,T_i}: $M_{1,i} - M_{2,i} = 0$. The linear forms are in the same rows, meaning that they are correlated. Since the complicated correlation structure of

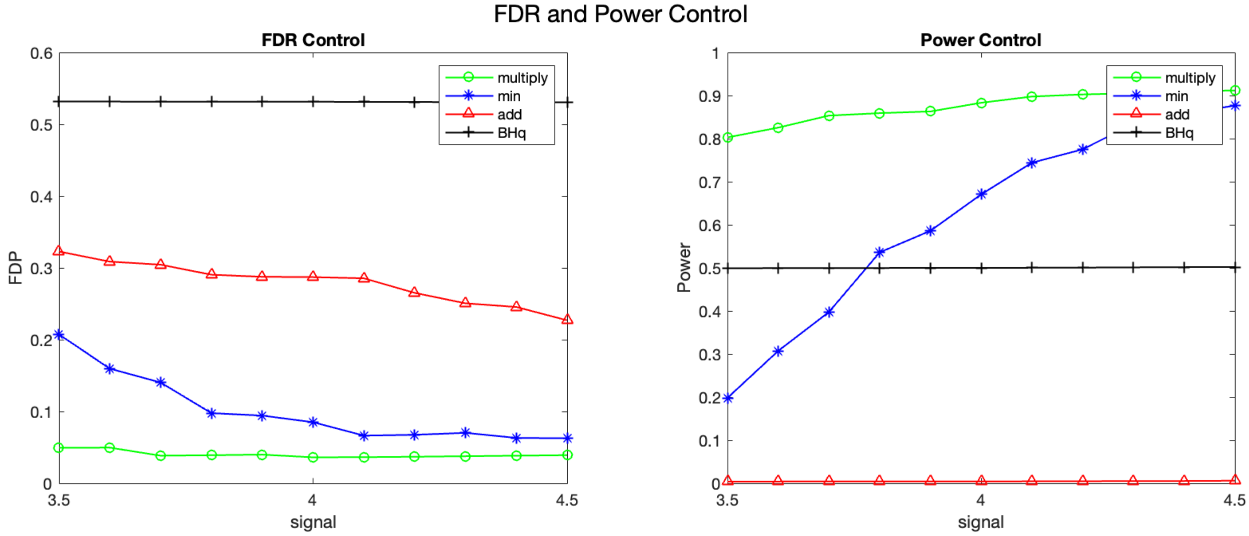


Figure 2: FDR control & Power of different data aggregation schemes in blockwise matrix tests with $\alpha = 0.1$. Here the signal is defined by μ in eq. (24).

features, here we measure the overall correlation of our case by the proportion of related pairs:

$$\varrho^*(z) = \frac{\sum_{i,j\in[q]} \mathbb{I}\left(\left|\rho_{T_i,T_j}\right| > z\right)}{q^2},$$

where ρ_{T_i,T_j} indicates the correlation of two linear form M_{T_i} and M_{T_j} and is given by (16). Here $\varrho^*(z)$ can be treated as a proxy of the strength of correlation β_{s}. In this entry comparison problem, we have $\varrho^*(0.2) = 0.3838$, which means that an indispensable proportion of feature pairs are correlated. For the SDA method, we use a known correlation matrix. The performance of Algorithm 1 and Algorithm 2 with different data aggregation methods are summarized in Figure 3. We also plot the ROC curves of different methods given two different signal levels. The result is presented in Figure 4.

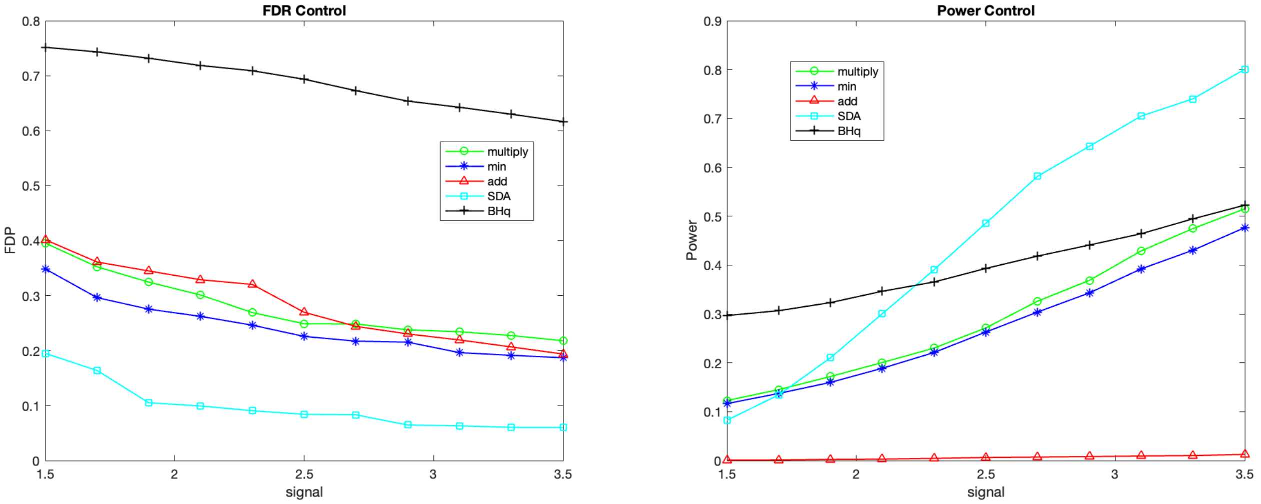
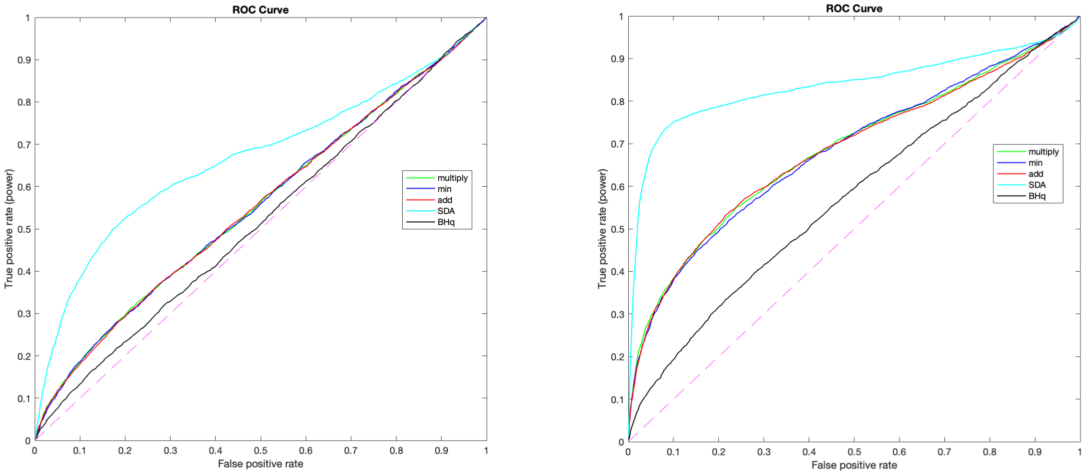


Figure 3: FDR control & Power of different data aggregation schemes in row tests with $\alpha = 0.1$. Here the signal is defined by μ in eq. (24).



(a) ROC curve with signal = 1

(b) ROC curve with signal = 2

Figure 4: ROC curve for different test statistics. Here the signal is defined by μ in eq. (24).

In Figure 3, the SDA method can effectively control the FDR level at $\alpha = 0.1$, with a notable power enhancement, while the BHq method on the other hand, fails to control the FDR given the strong correlation between features. Moreover, without de-correlation and screening, simple data aggregation methods also fail to control the FDR due to dependency. We can thereby draw the conclusion that our algorithm based on SDA outperforms others in the highly correlated case with the help of screening and de-correlation. The ROC curves in Figure 4 also clearly show the advantages of our data aggregation methods in feature selections.

5.1.4 Heavy-tailed noises

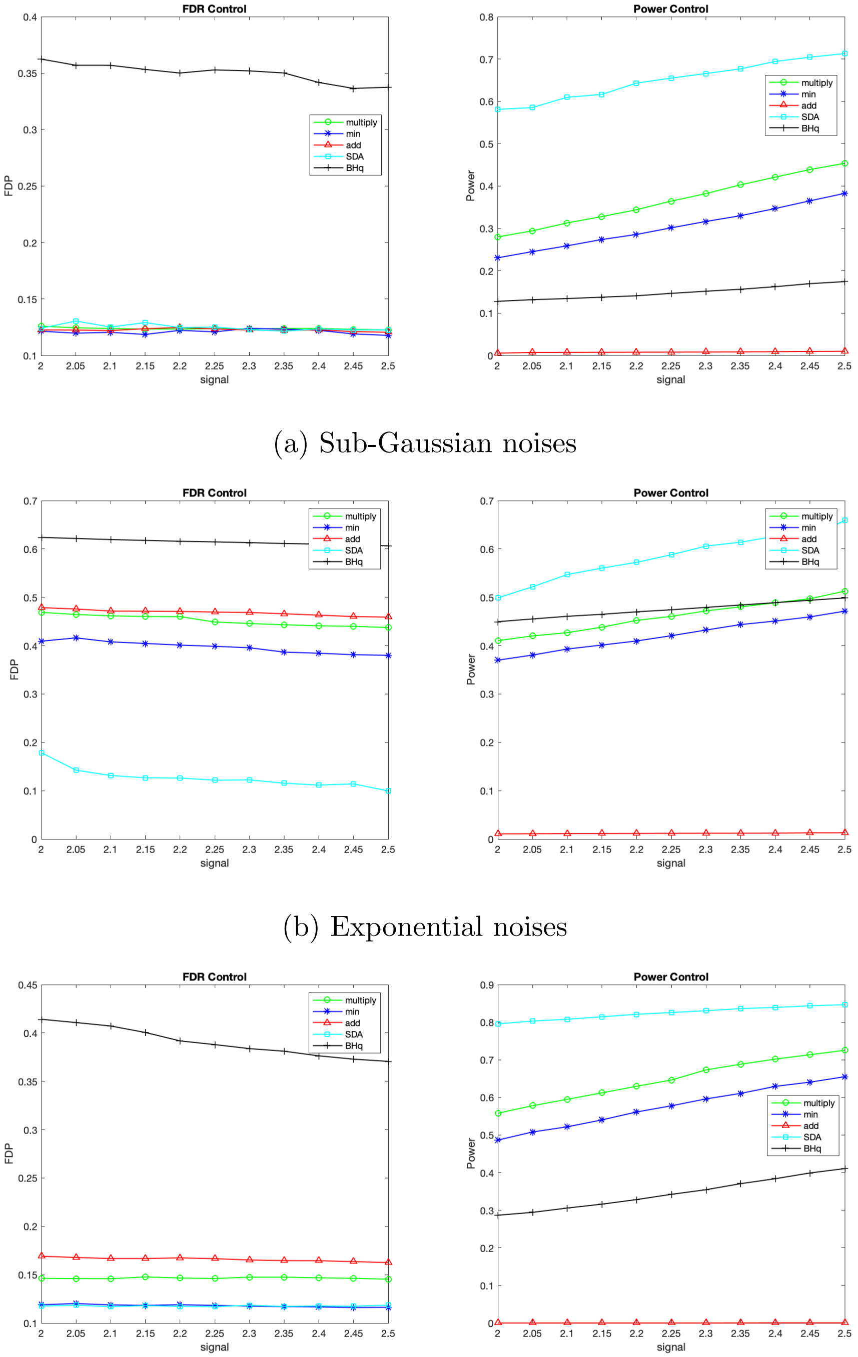
While our theories are established for sub-Gaussian noise, we observe that the proposed methods are very robust to heavy-tailed noise. This section showcases the performance of our algorithms in the existence of heavy-tailed noises, e.g., t-distribution and exponential distribution, and compares the performances of different methods. We consider moderate and strong correlations, respectively. Here M is randomly generated with dimensions $d_1 = d_2 = 400$, rank $r = 3$, $\lambda_{\min} = 400$, and the noise is fixed with a standard deviation $\sigma_\xi = 0.4$. The sample size is set by $n = 3000$. We focus on the following tasks: (i) entry comparisons between rows; (ii) entry comparisons within a block. More specifically, in the entry comparison task between rows, we compare $H_{0,T}$: $M_{i,j} - M_{i+1,1} = 0$ for every $1 \leq i \leq 4$ and $j \geq 2$. That is, we compare each entry with the first entry of the next row; in the entry comparison task within a block, we compare $H_{0,T}$: $M_{i,j} - M_{1,1} = 0$ for every $1 \leq i \leq 4$ and $j \geq 2$. For these two tasks, we all have $q = 1596$, but the correlation structures and levels are different. That is, (i) entry comparisons between rows, $\varrho^*(0.2) = 0.4541$; (ii) entry comparisons within a block, $\varrho^*(0.2) = 0.9514$. Here, (i) and (ii) can be viewed as examples of moderate and strong correlations.

We report all the results in Figure 5 and Figure 6. In both moderate and strong correlation cases, the BHq method shows unstable FDR control, while our proposed SDA method always performs well even under strong correlation. The SDA method is also robust with respect to heavy-tailed noises.

5.2 Real Data Examples

5.2.1 MovieLens

This section applies our methods to the MovieLens dataset for multiple testing and FDR control. MovieLens (Harper and Konstan, 2015), as a commonly used dataset in matrix completion problems, records millions of people's expressed preferences for movies (rated from 1-5). The dataset can be viewed as a huge, sparse matrix with heavily incomplete observations. MovieLens dataset is broadly used in matrix completion (Hastie et al., 2015; Monti et al., 2017; Xia and Yuan, 2021) and other machine learning tasks. The dataset is available on `https://grouplens.org/datasets/movielens/`. To demonstrate the reliability of the performance, we removed users with ratings less than 20 movies, resulting in 100,000 ratings (0-5) from 943 users on 1682

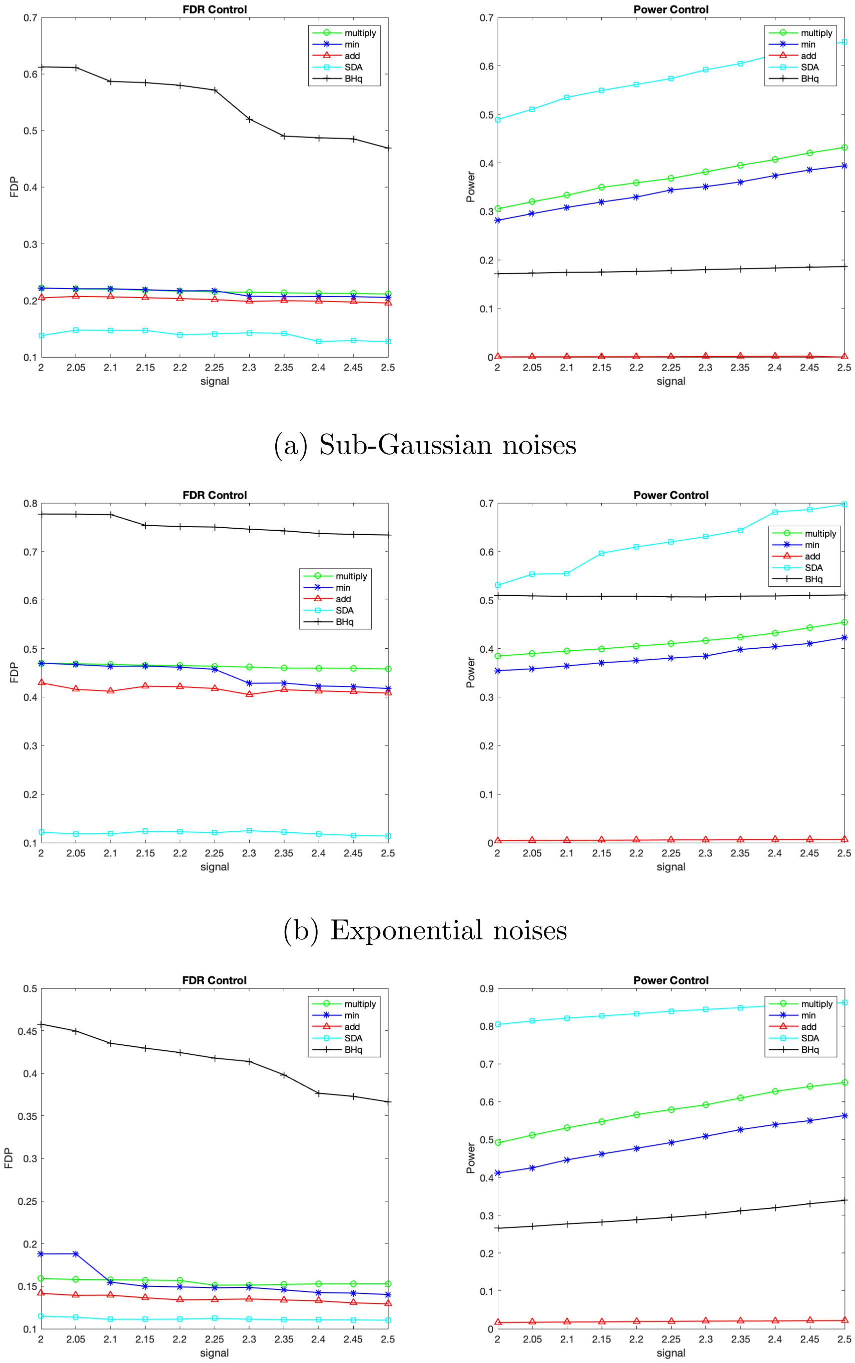


(a) Sub-Gaussian noises

(b) Exponential noises

(c) Student-t noises

Figure 5: FDR control & Power of different data aggregation schemes for entry comparisons between rows with $\alpha = 0.1$ when the noises are heavy-tailed distributed



(a) Sub-Gaussian noises

(b) Exponential noises

(c) Student-t noises

Figure 6: FDR control & Power of different data aggregation schemes for entry comparisons within a block with $\alpha = 0.1$ when the noises are heavy-tailed distributed

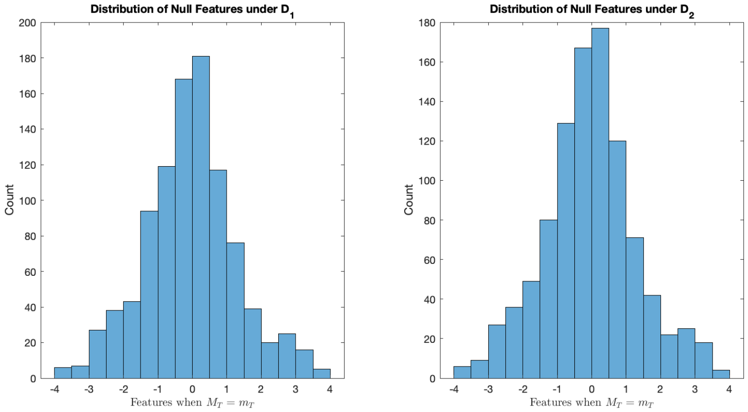


Figure 7: Symmetric distribution of all the test statistics under the null set given $\langle M, T\rangle = m_T$ for all T. The test statistics are observed to preserve a good symmetric property both on $\mathcal{D}_1$ and $\mathcal{D}_2$.

movies (where 0 stands for unrated movies). We assume the latent low-rank structure of this user-rating matrix with $r = 10$. We select $q = 1000$ adjacent and observed entry pairs, aiming to compare

$$H_{0,ij} : M(i,j) - M(i,j+1) = 0 \text{ versus } H_{1,ij} : M(i,j) - M(i,j+1) > 0,$$

for a group of suitable entries (i, j). Notice that since in the noisy matrix completion problem, we have the observation $Y(i,j) = M(i,j) + \xi(i,j)$, which means that the ground truth $M(i,j)$ is always unknown, we adopt the process in Xia and Yuan (2021) that treats $\mathbb{I}\left(Y\left(i,j+1\right) > Y\left(i,j\right)\right)$ as a proxy to differentiate H_1 from H_0.

We first randomly split data into two parts $\mathcal{D}_1$, $\mathcal{D}_2$, and then use gradient descent (Wei et al., 2016; Chen et al., 2020; Cai et al., 2022) on the two parts for initializing. Then, we run Algorithm 1, 2 with data splitting. Still, we consider 3 types of data aggregation on $\mathcal{D}_1$, $\mathcal{D}_2$. We first verify the symmetric property of our test statistics on MovieLens Data. Towards that end, we first set our hypotheses $m_{ij} = Y(i,j) - Y(i,j+1)$ and construct asymptotic statistics on $\mathcal{D}_1$, $\mathcal{D}_2$ to mimic null test statistics. Here we still use $Y(i,j) - Y(i,j+1)$ as a proxy of $M(i,j) - M(i,j+1)$. The distribution of the corresponding $W_T^{(1)}$, $W_T^{(2)}$ can be found in Figure 7, showing clearly the symmetric properties of null hypotheses.

We then apply our methods to the entrywise comparison task. Given a total of $q = 1000$, the number of instances for $\mathbb{I}\left(Y\left(i,j\right) > Y\left(i,j+1\right)\right)$ is $q_1 = 262$. We perform the tests for this one-sided hypothesis testing by dropping out hypotheses with negative test statistics on both $\mathcal{D}_1$ and $\mathcal{D}_2$. The p-values for BHq are also adjusted correspondingly. The outcomes are concisely presented in Table 1. The result table clearly shows that the SDA method outperforms other data aggregation methods and the BHq method in terms of false discovery rate control. The ineffectiveness of the first three simple data aggregation methods can be attributed to the high correlation of entry pairs, as adjacent entry pairs within a row are selected. When α is significantly small, SDA tends to be more conservative, which leads to good FDR control, while other methods remain to keep large FDRs. The result also shows business implications: instead of excessively recommending movies to users, the SDA can better select target users that are truly interested in the movies to increase the accuracy of the recommendation. By adopting our method for recommendation, the movie company can increase its profit while avoiding losing potential customers.

5.2.2 Rossmann sales dataset

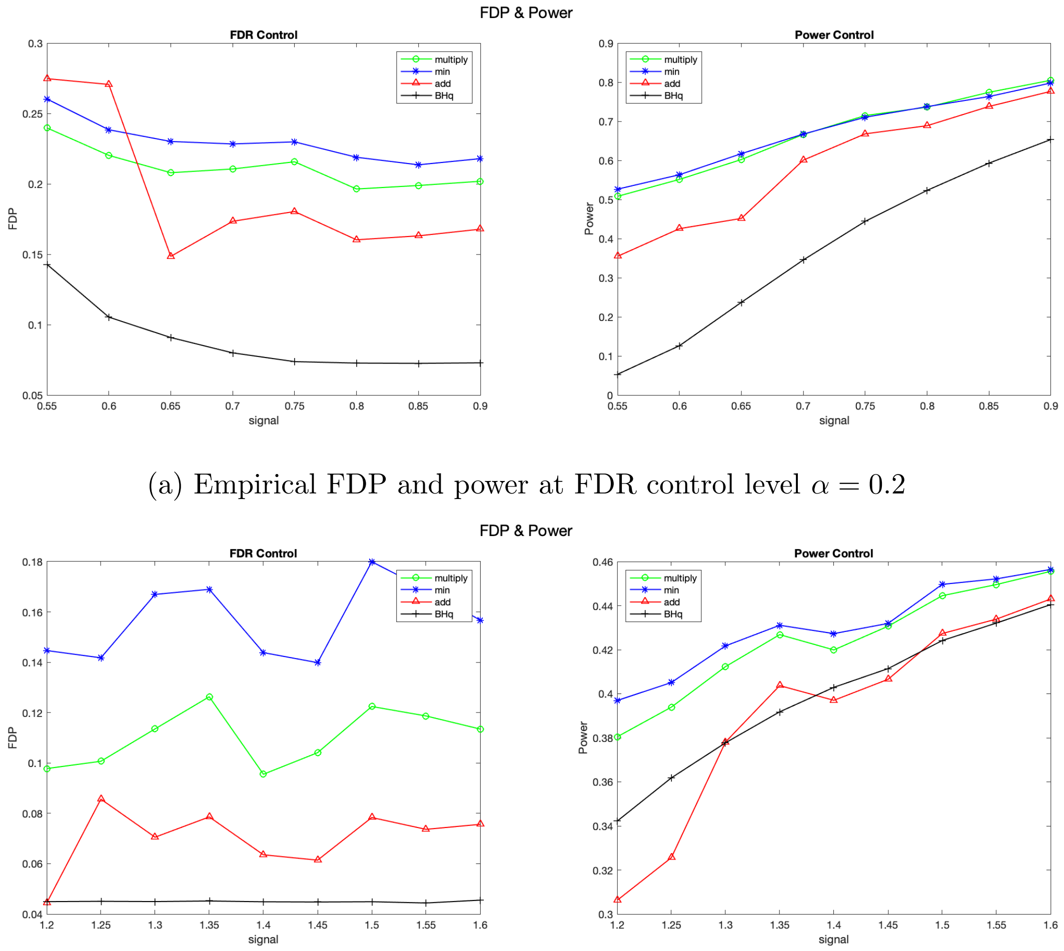
We use the Rossmann sales dataset that has recently been studied for uncertainty quantification in matrix completion (Farias et al., 2022; Gui et al., 2023). The Rossmann sales dataset records over 3,000 drug stores run by Rossmann in 7 European countries. The training set contains daily sales of 1115 drug stores on workdays from Jan 1, 2013, to July 31, 2015. The data matrix is thus of dimension 1115 × 780, where two dimensions represent drug stores and workdays, respectively. The unit of sales data is 1K. The dataset is very dense with about 80% valid (non-zero sells) observations of the full matrix; thus, we apply random masking to get sparse observations and use other data only to initialize the algorithm. In this example, we use 20% of the total records as each one split and apply Algorithm 1 on the two splits of the data that are properly processed. Noticing that most observed entries are given, we use the observations as true M_{ij} and perform multiple entrywise tests (25). We select the first $q = 20,000$ entries sorted by workdays with records in the whole dataset as our target $\mathcal{H}$. Since we consider the inference of a submatrix with a relatively large q, according to Section 3.2, the problem can be approximately treated as weakly correlated, which means simple data aggregation is enough to control FDR. We randomly assign null and non-null features by (24) but only consider positive signals. In this case, the ratio of non-null is $p = 0.3$, and we assume the latent low-rank $r = 30$. Specifically, we simultaneously test

$$H_{0,ij} : M_{ij} = m_{ij} \text{ vs } H_{1,ij} : M_{ij} > m_{ij}, \text{ for all } (i,j) \in \mathcal{H}. \tag{25}$$

We present the result in Figure 8 and the ROC curves in Figure 9. The Rossmann sales dataset is available at `https://www.kaggle.com/c/rossmann-store-sales`.

Level α	Method	False discoveries	True discoveries	FDP
$\alpha = 0.01$	Multiplication	13	59	0.1806
	Minimum	13	58	0.1831
	Addition	13	60	0.1781
	SDA	**0**	18	**0**
	BHq	1	26	0.0370
$\alpha = 0.05$	Multiplication	20	84	0.1923
	Minimum	20	83	0.1942
	Addition	20	84	0.1923
	SDA	**2**	25	**0.0741**
	BHq	10	53	0.1587
$\alpha = 0.1$	Multiplication	24	**95**	**0.2017**
	Minimum	24	94	0.2034
	Addition	25	95	0.2083
	SDA	**8**	49	**0.1404**
	BHq	22	76	0.2245
$\alpha = 0.2$	Multiplication	33	108	0.2340
	Minimum	33	108	0.2340
	Addition	33	108	0.2340
	SDA	**23**	89	**0.2054**
	BHq	36	115	0.2384

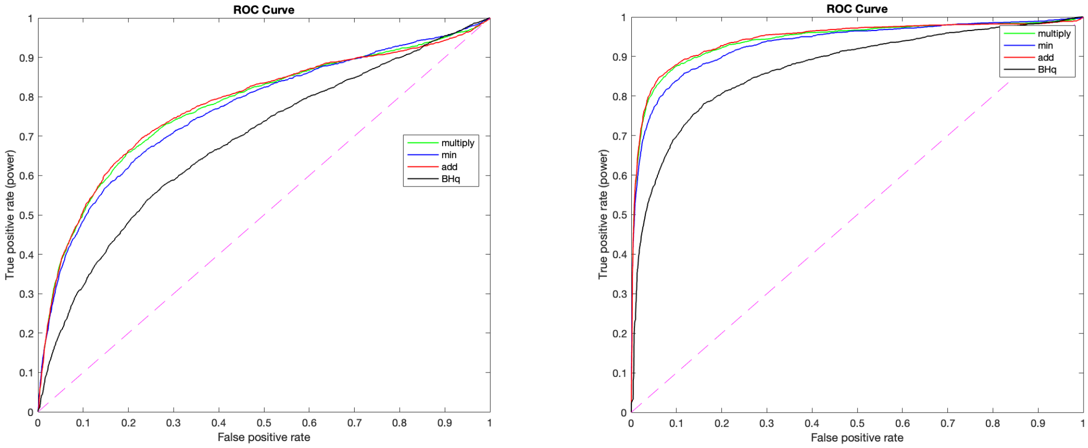
Table 1: Numbers of the discovered entry pairs with FDP by different data aggregation methods under various levels on MovieLens data.



(a) Empirical FDP and power at FDR control level $\alpha = 0.2$

(b) Empirical FDP and power at FDR control level $\alpha = 0.1$

Figure 8: FDR control & Power of different data aggregation schemes for Rossmann sales testing. Here the signals indicate the sizes of $|M_{ij} - m_{ij}|$ which are scaled by 10^3



(a) ROC curve with signal = 1

(b) ROC curve with signal = 2

Figure 9: ROC curve for different test statistics in Rossmann sales dataset

Here, three different data aggregation methods, together with BHq method, are compared. For this one-sided problem, we also drop out features that have negative statistics on $\mathcal{D}_1$ and $\mathcal{D}_2$. From Figure 8, it is clear that the data aggregation method with multiplication performs better regarding both FDR control and power. Data aggregation by taking minimum absolute values performs close to our aggregation method with multiplication in power, but it has larger FDPs. Data aggregation by adding absolute values behaves conservatively in the problem. The BHq method appears to be more conservative compared to the data aggregation methods, particularly at the FDR control level of $\alpha = 0.2$. Moreover, from the ROC curves in Figure 9, we can observe the obvious advantage of our data aggregation methods against the BHq method.

6 Concluding Remarks

In this paper, motivated by large-scale recommender systems, we study the problem of multiple testing for linear forms in noisy matrix completion and develop a general framework to control the FDR. Our approach is based upon a new test statistic for testing linear forms that enjoy sharper asymptotics than existing ones in the literature and an effective data splitting and symmetric aggregation scheme that can be shown to be especially suitable in the context of matrix completion.

Our approach can potentially be extended to many other problems with structural high-dimensional features. For example, one possible direction is the FDR control for tensor completion. Indeed, multiple testing in multilinear arrays presents a number of additional technical challenges as it requires much-involved analysis of singular subspace perturbations. As such, inferences in general for low-rank multilinear arrays are largely unexplored. We shall leave these intriguing problems for future investigation.

References

Al-Mohy, A. H. and Higham, N. J. (2009). Computing the fréchet derivative of the matrix exponential, with an application to condition number estimation. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1639–1657.

Bajgrowicz, P. and Scaillet, O. (2012). Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, 106(3):473–491.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, pages 2055–2085.

Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537.

Barras, L., Scaillet, O., and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The journal of finance*, 65(1):179–216.

Bathe, K.-J. (2013). The subspace iteration method–revisited. *Computers & Structures*, 126:177–183.

Bathe, K.-J. and Wilson, E. L. (1973). Solution methods for eigenvalue problems in structural mechanics. *International Journal for Numerical Methods in Engineering*, 6(2):213–226.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York.

Bi, X., Qu, A., and Shen, X. (2018). Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics*, 46(6B):3308–3333.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2017). Controlling the rate of gwas false discoveries. *Genetics*, 205(1):61–75.

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

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.

Cai, J.-F., Li, J., and Xia, D. (2022). Generalized low-rank plus sparse tensor estimation by fast riemannian optimization. *Journal of the American Statistical Association*, pages 1–17.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity1. *Annals of Statistics*, 45(2):615–646.

Cai, T. T. and Zhang, A. (2015). Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138.

Cai, T. T. and Zhou, W.-X. (2016). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493–1525.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(3):551–577.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *arXiv preprint arXiv:0903.1476*.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937.

Chen, Y., Li, C., Ouyang, J., and Xu, G. (2023). Statistical inference for noisy incomplete binary matrix. *Journal of Machine Learning Research*, 24(95):1–66.

Chumbley, J., Worsley, K., Flandin, G., and Friston, K. (2010). Topological fdr for neuroimaging. *Neuroimage*, 49(4):3057–3064.

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, 37(1):332–358.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association*, 0(0):1–18.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023). A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, pages 1–31.

Das, D., Sahoo, L., and Datta, S. (2017). A survey on recommendation system. *International Journal of Computer Applications*, 160(7).

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., and Livingston, B. (2010). The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296.

Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE transactions on information theory*, 47(7):2845–2862.

Du, L., Guo, X., Sun, W., and Zou, C. (2023). False discovery rate control under general dependence by symmetrized data aggregation. *Journal of the American Statistical Association*, 118(541):607–621.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Fan, J. and Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79(4):1143.

Farias, V., Li, A. A., and Peng, T. (2022). Uncertainty quantification for low-rank matrix completion with heterogeneous and sub-exponential noise. In *International Conference on Artificial Intelligence and Statistics*, pages 1179–1189. PMLR.

Fithian, W. and Lei, L. (2022). Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118.

Gao, C., Lu, Y., Ma, Z., and Zhou, H. H. (2016). Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17(1):5602–5630.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566.

Gui, Y., Barber, R., and Ma, C. (2023). Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36:4820–4844.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. SIAM.

Jung, S.-H. (2005). Sample size for fdr-control in microarray data analysis. *Bioinformatics*, 21(14):3097–3104.

Keshavan, R. H., Montanari, A., and Oh, S. (2010a). Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 11:2057–2078.

Keshavan, R. H., Montanari, A., and Oh, S. (2010b). Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

Li, J., Cai, J.-F., Chen, Y., and Xia, D. (2023). Online tensor learning: Computational and statistical trade-offs, adaptivity and optimal regret. *arXiv preprint arXiv:2306.03372*.

Li, J. and Zhong, P.-S. (2017). A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence. *The Annals of Statistics*, 45(2):557–590.

Liu, J. and Rigollet, P. (2019). Power analysis of knockoff filters for correlated designs. *Advances in Neural Information Processing Systems*, 32.

Liu, Y.-K. (2011). Universal low-rank matrix recovery from pauli measurements. In *Advances in Neural Information Processing Systems*, pages 1638–1646.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2018). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR.

Ma, W. and Xia, D. (2024). Statistical inference in tensor completion: Optimal uncertainty quantification and statistical-to-computational gaps. *arXiv preprint arXiv:2410.11225*.

McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Meyer, C. (2013). The bivariate normal copula. *Communications in Statistics-Theory and Methods*, 42(13):2402–2422.

Monti, F., Bronstein, M., and Bresson, X. (2017). Geometric matrix completion with recurrent multi-graph neural networks. *Advances in neural information processing systems*, 30.

Natarajan, N. and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697.

Perone Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014.

Raič, M. (2019). A multivariate berry–esseen theorem with explicit constants. *Bernoulli*, 25(4A):2824–2853.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Resnick, P. and Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Richard, E. and Montanari, A. (2014). A statistical model for tensor pca. *Advances in neural information processing systems*, 27.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30(1):239–257.

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. *The adaptive web: methods and strategies of web personalization*, pages 291–324.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471.

Shao, M. and Zhang, Y. (2023). Distribution-free matrix prediction under arbitrary missing pattern. *arXiv preprint arXiv:2305.11640*.

Song, Q. and Liang, F. (2015). A split-and-merge bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 947–972.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, pages 583–602.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Sun, T. and Zhang, C.-H. (2012). Calibrated elastic regularization in matrix completion. In *Advances in Neural Information Processing Systems*, pages 863–871.

Tsybakov, A., Koltchinskii, V., and Lounici, K. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. (2016). Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222.

Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candes, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*, 2(7).

Wu, W. B. (2008). On false discovery control under dependence. *The Annals of Statistics*, 36(1):364–380.

Xia, D. (2021). Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851.

Xia, D. and Yuan, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society Series B*, 83(1):58–77.

Xing, X., Zhao, Z., and Liu, J. S. (2021). Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, pages 1–20.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, C., Ren, H., Guo, X., and Li, R. (2020). A new procedure for controlling false discovery rate in large-scale t-tests. *arXiv preprint arXiv:2002.12548*.

Supplement to “Multiple Testing of Linear Forms in Noisy Matrix Completion”

To ease the understanding, here we list several important notations frequently encountered while reading our main text and proofs in Table 2.

Notation	Meaning
q, q_1, q_0	number of all, non-null, and null tests respectively
$W_T^{(i)}$	test statistic of linear form $M_T := \langle M, T\rangle$ constructed from the ith data split
μ	parameter for incoherence condition
β_0	minimum alignment parameter among all $T \in \mathcal{H}$
α_d	dimension ratio of matrix M: $\alpha_d := d_1/d_2$
ρ_T	$\rho_T = \|T\|_{\ell_1} / \|T\|_{\mathrm{F}}$
κ_0	condition number of matrix M
γ_n	accuracy of initial estimation $\left\|\widehat{M}^{\mathsf{init}} - M\right\|_{\max} \le C\sigma_\xi\gamma_n$, which may take $\gamma_n = C_0\sqrt{\frac{d_1 \log d_1}{n}}$
β_T	alignment parameter for T defined in (10)
$\mathcal{P}_M(\cdot)$	projection operators $\mathcal{P}_M(T) := T - \mathcal{P}_M^{\perp}(T) = T - U_{\perp}U_{\perp}^{\top}TV_{\perp}V_{\perp}^{\top}$
s_T	variance of testing M_T induced by random sampling: $s_T = \|\mathcal{P}_M(T)\|_{\mathrm{F}}$
h_n	asymptotic normal rate defined in (13)
β_{s}	proportion of strongly correlated linear form pairs defined in (20)
η_n	number of strong signals
κ_1	condition number of covariance matrix $\Sigma = \left(\langle \mathcal{P}_M(T_j), \mathcal{P}_M(T_k)\rangle\right)_{1\le j,k\le q}$
κ_T	shrinkage of variances caused by low-rank projection $\kappa_T = \|T_{\mathcal{H}}\| / \|\Sigma\|^{1/2}$
κ_∞	maximum row-wise ℓ_1-norm of inverse correlation matrix: $\kappa_\infty = \|R^{-1}\|_\infty := \max_i \|e_i^{\top}R\|_{\ell_1}$
q_n, q_{0n}	cardinality of support after screening $q_n = \lvert\mathcal{A}\rvert$, and $q_{0n} = \lvert\mathcal{A}\cap\mathcal{H}_0\rvert$
β_{s}', η_n'	proportion β_{s} and number of strong signals after screening

Table 2: Important notations used in the main text

A Additional Results

A.1 Covariance Matrix, Effect of Screening, and Whitening

Given $T_{\mathcal{H}} \in \mathbb{R}^{q \times d_1 d_2}$, we have the unnormalized covariance matrix for W_T as

$$\Sigma := \big(\langle \mathcal{P}_M(T_j), \mathcal{P}_M(T_k)\rangle\big)_{1\le j,k\le q} = T_{\mathcal{H}}(I_{d_1d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top)T_{\mathcal{H}}^\top.$$

Here, we have $\operatorname{rank}(I_{d_1d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top) = r(d_1+d_2) - r^2$, and $T_{\mathcal{H}}$ is of rank q. Therefore, to make sure that Σ is of full rank, we must have $q \le r(d_1+d_2) - r^2$.

Based on this covariance matrix representation, we shall discuss an example of testing about a submatrix of M to further illustrate the effect of screening and whitening. In particular, we shall show how whitening can weaken the dependence in Q^*, compared with the un-whitened $R_{\mathcal{A},\mathcal{A}}$, where

$$Q^* := \big(R_{\mathcal{A}}^{-1/2\top} R_{\mathcal{A}}^{-1/2}\big)^{-1} = R_{\mathcal{A},\mathcal{A}} - R_{\mathcal{A},\mathcal{A}^c} R_{\mathcal{A}^c,\mathcal{A}^c}^{-1} R_{\mathcal{A},\mathcal{A}^c}^\top.$$

To this end, we define the total test matrix $T_{\mathcal{H}} = [P_{d\times d}, \mathbf{0}_{d\times(d^2-d)}]$, where we set $d_1 = d_2 = q = d$. Thus, the covariance matrix of our un-standardized test statistics is

$$\begin{aligned}\Sigma =& T_{\mathcal{H}}\left(I_{d^2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right) T_{\mathcal{H}}^\top = P\left(I_d - \left(U_\perp U_\perp^\top\right)_{11} V_\perp V_\perp^\top\right)P^\top \\ =& P\left(VV^\top + u_{11} V_\perp V_\perp^\top\right)P^\top = P\,[V, V_\perp]\begin{bmatrix} I_r & 0 \\ 0 & u_{11} I_{d-r}\end{bmatrix}\left[V^\top; V_\perp^\top\right]P^\top.\end{aligned}$$

Here $u_{11} = \left(UU^\top\right)_{11}$. Without loss of generality, let $P = I_d$ be a diagonal matrix, i.e., testing multiple entries in the first row. The $q \times q$ covariance matrix $\Sigma = \left[u_{11} I_q + (1-u_{11})VV^\top\right]$, showing that the test statistics under noisy matrix completion are always correlated, due to the low-rank projection. This underscores the difficulties of multiple testing in matrix completion problems. Nevertheless, it is clear from textbook results of Multivariate Statistics that the total variance $\operatorname{tr}\big(\Sigma_{\mathcal{A},\mathcal{A}} - \Sigma_{\mathcal{A},\mathcal{A}^c}\Sigma_{\mathcal{A}^c,\mathcal{A}^c}^{-1}\Sigma_{\mathcal{A}^c,\mathcal{A}}\big) \le \operatorname{tr}(\Sigma_{\mathcal{A},\mathcal{A}})$, which is smaller than the total variance of the unscreened statistics $\operatorname{tr}(\Sigma)$.

A special case of multiple testing is defined by making

$$P = \begin{bmatrix} I_{\mathcal{A}} & B \\ 0 & I_{\mathcal{A}^c}\end{bmatrix}[V, V_\perp]^\top,$$

where for simplicity, we assume $\mathcal{A}$ is just the index set from the first $|\mathcal{A}|$ dimensions. This gives us the covariance matrix

$$\Sigma = \begin{bmatrix} \Lambda + u_{11}BB^\top & u_{11}B \\ u_{11}B^\top & u_{11}I_{\mathcal{A}^c}\end{bmatrix}.$$

Here Λ is a diagonal matrix of the size $|\mathcal{A}| \times |\mathcal{A}|$ with the first r diagonals equal to 1, and others equal to u_{11}. Obviously, this covariance matrix shows that the test statistics can be highly correlated since $R_{\mathcal{A},\mathcal{A}} = D_{\mathcal{A}}^{-\frac{1}{2}} \left(\Lambda + u_{11}BB^{\top}\right) D_{\mathcal{A}}^{-\frac{1}{2}}$ contains off-diagonal elements determined by B. Here $D_{\mathcal{A}}$ represents the the first $|\mathcal{A}|$ diagonal elements of Σ. However, the screening shows that

$$\begin{aligned} Q^* &= \left(R_{\mathcal{A}}^{-1/2\top} R_{\mathcal{A}}^{-1/2}\right)^{-1} = R_{\mathcal{A},\mathcal{A}} - R_{\mathcal{A},\mathcal{A}^c} R_{\mathcal{A}^c,\mathcal{A}^c}^{-1} R_{\mathcal{A},\mathcal{A}^c}^{\top} \\ &= D_{\mathcal{A}}^{-\frac{1}{2}} \left(\Lambda + u_{11}BB^{\top}\right) D_{\mathcal{A}}^{-\frac{1}{2}} - D_{\mathcal{A}}^{-\frac{1}{2}} u_{11}^{\frac{1}{2}} BB^{\top} u_{11}^{\frac{1}{2}} D_{\mathcal{A}}^{-\frac{1}{2}} \\ &= D_{\mathcal{A}}^{-\frac{1}{2}} \Lambda D_{\mathcal{A}}^{-\frac{1}{2}}, \end{aligned}$$

with no off-diagonal elements. This indicates that our screening and whitening procedure in the noisy matrix completion model can reduce the correlation of test statistics.

A.2 Non-asymptotic Bounds for FDR and Power

Here, we present a specific non-asymptotic version of our theoretical guarantees.

A.2.1 Weak dependence

Theorem 7. *Under the conditions of Theorem 5,*

(a) with probability at least

$$1 - C_2 \varepsilon^{-2} \log(\frac{q_0}{\alpha \eta_n}) \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\alpha^2 \eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_n q_0}{\alpha \eta_n} + (\alpha \eta_n q_0)^{-\nu/2}\right)^{\frac{1}{2}} \right) - C_2 h_n,$$

Algorithm 1 achieves false discovery proportion

$$\mathrm{FDP} := \frac{\sum_{T \in \mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{\left(\sum_{T \in \mathcal{H}} \mathbb{I}(W_T^{\mathsf{rank}} > L)\right) \vee 1} \le \alpha(1+\varepsilon), \tag{26}$$

for any $\varepsilon \in (0,1)$.

(b) with probability at least

$$1 - C_2 \log(\frac{q_0}{\alpha \eta_n}) \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\alpha^2 \eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_n q_0}{\alpha \eta_n} + (\alpha \eta_n q_0)^{-\nu/2}\right)^{\frac{1}{2}} \right) - C_2 \varepsilon^{-1} h_n,$$

Algorithm 1 can select the strong signals with power

$$\mathrm{POWER} := \frac{\sum_{T \in \mathcal{H}_1} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{q_1} \ge (1-\varepsilon)\frac{\eta_n}{q_1}. \tag{27}$$

Note that Part (a) also implies that

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}) \le \alpha + C_2 h_n + C_2 \alpha^{\frac{2}{3}} \log(\frac{q_0}{\alpha \eta_n}) \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\alpha^2 \eta_n^2}\right)^{\frac{1}{6}} + \left(\frac{h_n q_0}{\alpha \eta_n}\right)^{\frac{1}{6}} + (\alpha \eta_n q_0)^{-\frac{\nu}{12}} \right). \tag{28}$$

A.2.2 Whitening and screening

Theorem 8. *Under the settings of Theorem 6, suppose that*

$$\left(\left\|R^{-1}\right\|_{\infty}+\kappa_1\frac{\|T_{\mathcal{H}}\|}{\|\Sigma\|^{1/2}}\left(\frac{\operatorname{supp}(T_{\mathcal{H}})}{\sqrt{d_2}}\wedge 1\right)\right)\frac{\beta_T\mu\sigma_{\xi}}{\beta_0\lambda_{\min}}\sqrt{\frac{\kappa_1\alpha_d q d_1^2 d_2\log d_1}{n}}=o(1).$$

With the regularization level $\lambda = C\sqrt{\log d_1}$, the Algorithm 2 attains an FDP

$$\mathrm{FDP}=\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(\mathsf{w}_T^{\mathsf{rank}}>L)}{\left(\sum_{T\in\mathcal{H}}\mathbb{I}(\mathsf{w}_T^{\mathsf{rank}}>L)\right)\vee 1}\leq\alpha(1+\varepsilon),$$

for any $\varepsilon\in(0,1)$ with probability at least

$$1-C_1\varepsilon^{-2}\log(\frac{q_0'}{\alpha\eta_n'})\left(\left(\frac{\beta_{\mathsf{s}}'q_0'^2}{\alpha^2\eta_n'^2}\right)^{\frac{1}{2}}+\left(\frac{C_{\infty}\left(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_{\infty}\right)q_0'}{\alpha\eta_n'}+(\alpha\eta_n'q_0')^{-\nu/2}\right)^{\frac{1}{2}}\right)-C_{\infty}\left(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_{\infty}\right),$$

where C_{∞} is a constant involving $\widehat{R}$ and $\mathcal{A}$ only, defined later in Proposition 3. Moreover, the power is guaranteed to be lower bounded by:

$$\mathrm{POWER}=\frac{\sum_{T\in\mathcal{H}_1}\mathbb{I}(\mathsf{w}_T^{\mathsf{rank}}>L)}{q_1}\geq(1-\varepsilon)\frac{\eta_n'}{q_1},$$

with a probability at least

$$1-C_1\log(\frac{q_0'}{\alpha\eta_n'})\left(\left(\frac{\beta_{\mathsf{s}}'q_0'^2}{\alpha^2\eta_n'^2}\right)^{\frac{1}{2}}+\left(\frac{C_{\infty}\left(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_{\infty}\right)q_0'}{\alpha\eta_n'}+(\alpha\eta_n'q_0')^{-\nu/2}\right)^{\frac{1}{2}}\right)-C_1C_{\infty}\varepsilon^{-1}\left(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_{\infty}\right).$$

Since we further have

$$\|\Sigma\|^{\frac{1}{2}}\geq\left\|e_i^{\top}T_{\mathcal{H}}\left(I_{d_1d_2}-U_{\perp}U_{\perp}^{\top}\otimes V_{\perp}V_{\perp}^{\top}\right)\right\|_2=\|\mathcal{P}_M(T_i)\|_{\mathrm{F}}\geq\beta_0\sqrt{\frac{r}{d_1}}\|T_i\|_{\mathrm{F}},$$

i.e., $\|\Sigma\|^{\frac{1}{2}}\geq\beta_0\sqrt{\frac{r}{d_1}}\max_i\|T_i\|_{\mathrm{F}}=\beta_0\sqrt{\frac{r}{d_1}}\|T_{\mathcal{H}}\|_{2,\max}$, and

$$\frac{\|T_{\mathcal{H}}\|}{\|\Sigma\|^{\frac{1}{2}}}\left(\frac{\operatorname{supp}(T_{\mathcal{H}})}{\sqrt{d_2}}\wedge 1\right)\leq\frac{\sqrt{\alpha_d}}{\sqrt{r}\beta_0}\cdot\frac{\|T_{\mathcal{H}}\|}{\|T_{\mathcal{H}}\|_{2,\max}}\left(\operatorname{supp}(T_{\mathcal{H}})\wedge\sqrt{d_2}\right),$$

we can convert this signal requirement to a stronger but clearer one presented in Theorem 6. In the subsequent proofs, we shall prove the non-asymptotic versions of Theorem 5 and 6.

A.3 Finite-sample Guarantees for Whitening and Screening

Notice that, in our method of FDR control with whitening and screening, the condition of the correlation structure is defined on the asymptotic correlation matrix $Q^*:=\left(R_{\mathcal{A}}^{-1/2\top}R_{\mathcal{A}}^{-1/2}\right)^{-1}$.

However, conditional on $\mathcal{D}_1$, the covariance of our test statistics is determined by $\widehat{\mathsf{w}}_i^{(2)}$ and is sample-related, which is $Q := (\widehat{R}_{\mathcal{A}}^{-1/2\top}\widehat{R}_{\mathcal{A}}^{-1/2})^{-1}\widehat{R}_{\mathcal{A}}^{-1/2\top}\widehat{R}^{-1/2}R\widehat{R}^{-1/2}\widehat{R}_{\mathcal{A}}^{-1/2}(\widehat{R}_{\mathcal{A}}^{-1/2\top}\widehat{R}_{\mathcal{A}}^{-1/2})^{-1}$. The following Proposition 1 will show that, as long as the signal strength of M is strong enough, the estimation of R will be accurate enough such that the data-driven Q is also weakly correlated.

Proposition 1 (Finite-sample guarantee of weak correlation after screening)**.** *If the matrix signal strength satisfies*

$$\frac{\kappa_1^{1.5}\left\|T_{\mathcal{H}}\right\|\sigma_\xi}{\lambda_{\min}\left\|\Sigma\right\|^{\frac{1}{2}}}\left(\frac{\operatorname{supp}(T_{\mathcal{H}})}{\sqrt{d_2}}\wedge 1\right)\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}\lesssim\frac{1}{q^{\nu}},$$

then the weak correlation condition also holds for finite-sample covariance matrix Q, i.e., β_{s}' is defined as the proportion of strongly correlated pairs using Q instead of Q^*.

Proposition 2 (LASSO screening)**.** *By choosing regularization level $\lambda = C\sqrt{\log d_1}$, LASSO can recover the signal with precision*

$$\left|\widehat{\mathsf{w}}_i^{(1)}-\mathsf{w}_i\right|\leq C\kappa_1^{1.5}\sqrt{q_1\log d_1}+h_n\left|\mathsf{w}_i\right|,$$

uniformly for all $i\in[q]$ with probability at least $1-Cd_1^{-2}\log d_1$ for some universal constant $C>0$, as long as the sample requirement of SDA holds. Moreover, under this condition, if $T_i\in\mathcal{S}'$, then LASSO can surely select feature i.

In our method, LASSO is used for pre-selection. In fact, we always deliberately choose a weak regularization level so that most true signals and many false positives are included in $\mathcal{A}$, at the cost of power loss. Here, we do not require the sure-screening condition of LASSO that is commonly used in Roeder and Wasserman (2009); Barber and Candès (2019); Du et al. (2023); Dai et al. (2023). We emphasize that our theory can hold with non-identified signals as long as $\left\|\mathsf{w}_{\mathcal{A}^c}\right\|_\infty$ is small enough.

We exploit the symmetricity of $\widehat{\mathsf{w}}^{(2)}$ obtained by linear regression after LASSO. This symmetricity, described in the following Proposition 3, serves as a counterpart of Theorem 1 in the weakly correlated case.

Proposition 3 (Linear regression after screening)**.** *Suppose $T_i\in\mathcal{A}\cap\mathcal{H}_0$. Denote an upper bound of variance shrinkage effect of screening on $\mathcal{A}$ as*

$$C_\infty := \sup_{i\in\mathcal{A}}\frac{1\vee\left\|e_i^\top\left(\widehat{R}_{\mathcal{A}}^{-1/2\top}\widehat{R}_{\mathcal{A}}^{-1/2}\right)^{-1}\widehat{R}_{\mathcal{A}}^{-1/2\top}\widehat{R}_{\mathcal{A}^c}^{-1/2}\right\|_{\ell_1}}{\sqrt{Q_{ii}}}.$$

Here, we slightly abuse the notation by treating $\mathcal{A}$ as an index set of numbers. Conditional on $\mathcal{D}_1$, we have

$$\left|\mathbb{P}\left(\frac{\widehat{\mathsf{w}}_i^{(2)}}{\sqrt{Q_{ii}}} \leq t \middle| \mathcal{D}_1\right) - \Phi(t)\right| \leq C \cdot C_\infty \left(h_n + \|\mathsf{w}_{\mathcal{A}^c}\|_\infty\right).$$

Here, C_∞ can be viewed as a special kind of coherence condition that has been broadly used in LASSO selection (Donoho and Huo, 2001; Zhao and Yu, 2006; Wainwright, 2009). In this propostion, $\|\mathsf{w}_{\mathcal{A}^c}\|_\infty$ measures the error caused by inconsistent screening.

A.4 An Equivalent Version of Algorithm 2

Note that Algorithm 2 involves the computation of the correlation coefficient matrix. To ease the analysis, we rewrite Algorithm 2 as the following version that avoids computing the inverse of diagonal elements. Notice that, this is just a change of notation for mathematical analysis, rather than a new algorithm.

Algorithm 3 Matrix FDR Control with Whitening and Screening

```
Require: Hypotheses $\{H_{0T} : \langle M, T\rangle = \theta_T, T \in \mathcal{H}\}$, data splits $\mathcal{D}_1$, $\mathcal{D}_2$, rank $r$, FDR level $\alpha$.
1: Use $\mathcal{D}_0$ to construct an initial estimate $\widehat{M}_{\mathsf{init}}$
2: Apply proposed asymptotic test statistics to the second part of data $\mathcal{D}_1$ and the third part of data $\mathcal{D}_2$ respectively to get un-normalized test statistics $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, where
       $\mathbf{W}_i^{(k)} = \widehat{s}_{T_i}^{(k)} W_{T_i}^{(k)} = \frac{\widehat{M}_{T_i}^{(k)} - \theta_{T_i}}{\widehat{\sigma}_{\xi}^{(k)} \sqrt{d_1 d_2 / n}}, \ k = 1, 2, \ \text{and} \ \widehat{D} = \mathrm{diag}\left(\widehat{s}_{T_1}^{(1)}, \ldots, \widehat{s}_{T_p}^{(1)}\right).$
   Here $\widehat{s}_{T_i}^{(k)} = \left\|\mathcal{P}_{\widehat{M}_{\mathsf{init}}^{(k)}}(T_i)\right\|_{\mathrm{F}}$ is an estimate of $s_{T_i} = \left\|\mathcal{P}_M(T_i)\right\|_{\mathrm{F}}$.
3: Obtain a covariance matrix estimate $\widehat{\Sigma}$ by $\widehat{U}^{\mathsf{init}}$, $\widehat{V}^{\mathsf{init}}$ from $\mathcal{D}_0$ and $\mathcal{D}_1$:
       $\widehat{\Sigma} = T_{\mathcal{H}}(I_{d_1 d_2} - \widehat{U}_{\perp}\widehat{U}_{\perp}^{\top} \otimes \widehat{V}_{\perp}\widehat{V}_{\perp}^{\top}) T_{\mathcal{H}}^{\top}, \qquad (29)$
   and write $\mathbf{X} = \widehat{\Sigma}^{-\frac{1}{2}}$. Construct response $\mathbf{y}_1 = \mathbf{X}\mathbf{W}^{(1)}$, and solve LASSO
       $\widehat{\mathsf{w}}^{(1)} = \arg\min_{\mathsf{w} \in \mathbb{R}^q} \left\{ \frac{1}{2} \left\| \mathbf{y}_1 - \mathbf{X}\widehat{D}\mathsf{w} \right\|^2 + \lambda \left\| \mathsf{w} \right\|_{\ell_1} \right\}.$
4: Denote $\mathcal{A}$ as the support set of LASSO solution $\widehat{\mathsf{w}}^{(1)}$. We then have the separation $\mathbf{X} = [\mathbf{X}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}^c}]$. We run linear regression on $\mathcal{A}$ with new loading matrix $\mathbf{X}_{\mathcal{A}}$ and response $\mathbf{y}_2 = \mathbf{X}\mathbf{W}^{(2)}$ from $\mathcal{D}_2$ to get asymptotic symmetric statistics $\widehat{\mathsf{w}}^{(2)}$, where
       $\widehat{\mathsf{w}}_i^{(2)} = \begin{cases} e_i^{\top} \left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} \mathbf{X}_{\mathcal{A}}^{\top}\mathbf{y}_2, & i \in \mathcal{A} \\ 0, & i \in \mathcal{A}^c \end{cases}$
   with variance estimate $\widehat{\sigma}_{wi}^2 = e_i^{\top} \left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i$.
5: Compute the final ranking statistics of each $T_i$ by $\widehat{\mathsf{w}}_{T_i}^{\mathsf{rank}} = \widehat{\mathsf{w}}_i^{(1)}\widehat{\mathsf{w}}_i^{(2)}/\widehat{\sigma}_{wi}$, and then choose a data-driven threshold $L$ by
       $L = \inf\left\{ t > 0 : \frac{\sum_{T \in \mathcal{H}} \mathbb{I}\left(\widehat{\mathsf{w}}_T^{\mathsf{rank}} < -t\right)}{\sum_{T \in \mathcal{H}} \mathbb{I}\left(\widehat{\mathsf{w}}_T^{\mathsf{rank}} > t\right) \vee 1} \leq \alpha \right\}.$
6: Reject $H_{0T_i}$ if $\widehat{\mathsf{w}}_{T_i}^{\mathsf{rank}} > L$
```

It can be easily checked that $\widehat{\mathsf{w}}_i^{(1)}$ and $\widehat{\mathsf{w}}_i^{(2)}/\widehat{\sigma}_{wi}$ share the same representation as in Algorithm 2, and Algorithm 2, 3 are essentially identical. For brevity of notations, our following proofs (presented in Sections B.4-B.7) of theories in Section 4 will be based on the quantities and notations in Algorithm 3 rather than that in Algorithm 2.

A.5 Comparison of Data Aggregation Methods

The empirical success of data splitting in multiple testing leads to the problem of how to choose data aggregation methods for split data and what the theoretical explanations are behind them. In this section, we probe into the power behavior of different data aggregation methods to answer this question. Indeed, existing literature have scarcely compared the power of different FDR control procedures. Here we list some notable attempts: Genovese et al. (2006) found that the p-value weighting can improve the power compared with the original BHq method; Scott et al. (2015) showed simulation evidence that FDR regression improves the power upon traditional FDR control methods; for knockoff procedure, Liu and Rigollet (2019); Weinstein et al. (2020) focused on explaining the power behavior of knockoff under special designs. However, all these attempts have been unsuccessful in transferring to the case of data aggregation methods and in comparing the power enhancement achieved through data splitting. We compare our methods with other combination schemes in a simple mean-testing problem. Actually, several data aggregation methods have been proposed in Dai et al. (2022) by the so-called "mirror statistic" design. Namely, for any dimension $i \in [q]$, we derive two independent test statistics X_i^1, X_i^2 from two groups of data. Then we combine each pair of X_i^1, X_i^2 by

$$W(X_i^1, X_i^2) = \operatorname{sign}(X_i^1 X_i^2) f(\left|X_i^1\right|, \left|X_i^2\right|). \tag{30}$$

Possible candidates of $f(u,v)$ are

$$f_1(u,v) = uv, \quad f_2(u,v) = \min(u,v), \quad f_3(u,v) = u + v. \tag{31}$$

Among these choices, f_2 and f_3 have been discussed in Xing et al. (2021); Dai et al. (2023, 2022) and f_3 is said to be nearly optimal with respect to power under certain conditions (Dai et al., 2022). Our method can be viewed as a special kind of mirror statistic design by choosing $f_1(u,v) = uv$. This amounts to computing the Hadamard product of two test statistic vectors X^1, $X^2 \in \mathbb{R}^p$. Interestingly, in practice, it is sometimes observed that f_1 can outperform other methods; see Dai et al. (2023); Du et al. (2023) for examples. Here, we explain this empirical finding from a Bayesian perspective by mixture model. Consider the multiple testing problem that we observe q-dimensional vector X sampled from the model

$$X = \boldsymbol{\delta} + \boldsymbol{\xi}, \tag{32}$$

where noise $\boldsymbol{\xi}$ is independent multivariate gaussian with $\Sigma = I_q$ (or weakly dependent). The signals $\boldsymbol{\delta}$ are sparse and independent from an unknown non-zero prior $\boldsymbol{\Theta}$ in the sense that in each dimension $i \in [q]$, $\delta_i = 0$ or $\delta_i \sim \boldsymbol{\Theta}$, and $\pi_1 := \#\{\mu_i \sim \boldsymbol{\Theta}\}/q \to 0$. Our tests are

$$\mathcal{H}_{0i} : \delta_i = 0 \text{ versus } \mathcal{H}_{1i} : \delta_i \neq 0, \text{ for every } i \in [q].$$

To examine the impact of data aggregation, suppose two observations X^1, X^2 are given, and we aim to control the FDR by data aggregation in (30) with a certain threshold L_α. When $q \to \infty$, the performance of this data-splitting-based method can actually be explained by a mixture model. Consider a prior $H_0 : \delta = 0$, and $H_1 : \delta \sim \boldsymbol{\Theta}$, with $\mathbb{P}(H_0) = 1 - \pi_1$, $\mathbb{P}(H_1) = \pi_1$ and a variable Y with mixture distribution $Y|H_0 \sim W(\xi_1, \xi_2)$, and $Y|H_1 \sim W(\delta + \xi_1, \delta + \xi_2)$. Here ξ_1, ξ_2 are independent standard normal variables. When all the dimensions of X are independent or weakly correlated, we have

$$\frac{\#\{i : W_i > t\}}{q} \to \mathbb{P}(Y > t)$$

uniformly for any t. The limiting behavior of data aggregation method W given any threshold L is summarized as follows:

$$\begin{aligned} \mathrm{FDR}_W(L) &= \frac{\mathbb{P}(Y > L, H_0)}{\mathbb{P}(Y > L)} = \frac{(1-\pi_1)\mathbb{P}(Y > L|H_0)}{(1-\pi_1)\mathbb{P}(Y > L|H_0) + \pi_1 \mathbb{P}(Y > L|H_1)}, \\ \mathrm{Power}_W(L) &= \mathbb{P}(Y > L|H_1), \end{aligned} \tag{33}$$

where the limiting power is the expectation with respect to $\boldsymbol{\Theta}$: $\mathbb{P}(Y > L|H_1) = \mathbb{E}_{\boldsymbol{\Theta}}\mathbb{P}(Y > L|\delta, H_1)$. Suppose we can specify a threshold L_α that controls the limiting FDR at exact level α, i.e.,

$$L_\alpha = \min\{L > 0 : \mathrm{FDR}_W(L) = \alpha\}, \tag{34}$$

where L_α is determined by both FDR level α and aggregation function W. Then, at the same FDR level α, the power of different data aggregation methods is only decided by the mixture distribution Y induced by aggregation function W. To compare the limiting power of different aggregation method $W_j(u, v) = \mathrm{sign}(uv) f_j(|u|, |v|)$, $j = 1, 2, 3$, we denote $L_{\alpha j}$ as the corresponding threshold by (34). It suffices to compare $\mathrm{Power}_{W_j}(L_{\alpha j})$. This is equivalent to comparing the quantities $\mathrm{Power}_{W_j}(L_{pj})$ where L_{pj} is the p-th quantile of null distribution $Y_j|H_0 \sim W_j(\xi_1, \xi_2)$. The rationale is as follows. For the same quantile p, if the $\mathrm{Power}_{W_j}(L_{pj})$ is larger, then in order to achieve the same FDR level, one must have a smaller threshold, thus the corresponding $\mathbb{P}(Y_i > L|H_0)$ tends to be larger. It is clear that given the threshold

L_α that controls the limiting FDR at exact level α, we have $\mathbb{P}(Y > L_\alpha | H_0)$ proportional to $\mathbb{P}(Y > L_\alpha | H_1)$, which implies that larger $\mathbb{P}(Y > L | H_0)$ leads to a larger power.

If $\text{FDR}_{W_j}(L_{\alpha j}) = \alpha$, then we have

$$\mathbb{P}(Y > L_\alpha | H_0) = \frac{\pi_1}{1-\pi_1} \mathbb{P}(Y > L_\alpha | H_1) \frac{\alpha}{1-\alpha} \leq c\pi_1,$$

which indicates that to reach any fixed FDR level α, the quantity $\mathbb{P}(Y > L_\alpha | H_0)$ will decrease at the rate $O(\pi_1)$. We thus choose $p = O(\pi_1)$ and L_{pj} satisfying $\mathbb{P}(Y_j > L_{pj} | H_0) = p$ for $j = 1, 2, 3$. Let $z = \sqrt{p}\delta$. We will use Talyor expansion and compare the derivatives of $\mathbb{P}(Y_j > L_{pj} | z, H_0) = p$ with respect to z.

Theorem 9. *Consider the limiting behaviors* (33) *of different data aggregation methods in* (31) *characterized by the mixture model stated above. When achieving the same FDR level α and $\pi_1 \to 0$, we have the following asymptotic power relation:*

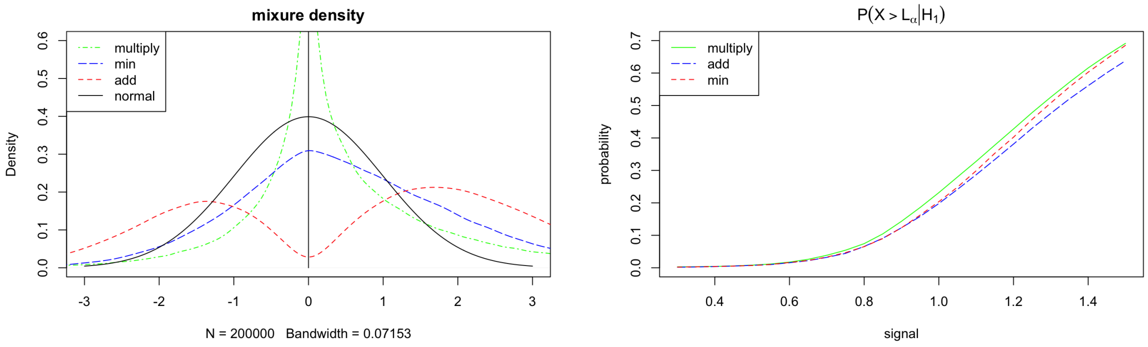
$$\text{Power}_{W_1}(L_{\alpha 1}) \geq \text{Power}_{W_2}(L_{\alpha 2}) \geq \text{Power}_{W_3}(L_{\alpha 3}),$$

for any bounded prior $\mathbf{\Theta}$: $\mathbb{P}(|\delta| \leq \delta_0 | \mathbf{\Theta}) \to 1$ where $\delta_0 = o(\sqrt{\frac{1}{\pi_1}})$.

Here, we allow the bound δ_0 to go to infinity as long as its order is of $o(\sqrt{\frac{1}{\pi_1}})$. This theorem offers a theoretical justification for the superiority of our data aggregation method over other common alternatives, a conclusion that aligns with our empirical findings in Dai et al. (2023); Du et al. (2023).

Intuitively, when the two-sided tails of mixture distribution are more unbalanced (left-skewed) and $\mathbb{P}(Y > t)$ decreases slower, the threshold L_α tends to be smaller and thus the null and non-null distributions can be well-separated. In Figure 10, we present a simulation of the density of mixture distributions and $\mathbb{P}(Y > L_\alpha | H_1)$ given different data aggregation methods.

It is observed that f_1 generates a narrower mixture distribution with unbalanced tails starting to decrease slowly when t is moderate, and the limiting power of f_1 is the highest among the three combinations.



(a) Density of mixture distribution

(b) $\mathbb{P}(Y > L_\alpha | H_1)$

Figure 10: Simulation of mixture distribution Y under different constructions

B Proofs of Main Results

B.1 Proof of Theorems 1, 3

We first verify the initialization condition (4). Let $\widehat{M}^{\mathsf{init}}$ be the output of gradient descent till convergence from Chen et al. (2020). According to the leave-one-out analysis in Chen et al. (2020) (eq 93, and Lemma 10-15), we have,

$$\begin{aligned}\left\|\widehat{M}^{\mathsf{init}} - M\right\|_{\max} \le{}& \left\|V\Lambda^{\frac{1}{2}}\right\|_{2,\max}\left\|\widehat{U}^{\mathsf{init}}(\widehat{\Lambda}^{\mathsf{init}})^{\frac{1}{2}}\widetilde{O} - U\Lambda^{\frac{1}{2}}\right\|_{2,\max} + \left\|U\Lambda^{\frac{1}{2}}\right\|_{2,\max}\left\|\widehat{V}^{\mathsf{init}}(\widehat{\Lambda}^{\mathsf{init}})^{\frac{1}{2}}\widetilde{O} - V\Lambda^{\frac{1}{2}}\right\|_{2,\max}\\ &+ \left\|\widehat{U}^{\mathsf{init}}(\widehat{\Lambda}^{\mathsf{init}})^{\frac{1}{2}}\widetilde{O} - U\Lambda^{\frac{1}{2}}\right\|_{2,\max}\left\|\widehat{V}^{\mathsf{init}}(\widehat{\Lambda}^{\mathsf{init}})^{\frac{1}{2}}\widetilde{O} - V\Lambda^{\frac{1}{2}}\right\|_{2,\max}\\ &\le C\frac{\mu r}{\sqrt{d_1 d_2}}\lambda_{\max}\cdot\frac{\kappa_0\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{d_1^2 d_2\log d_1}{n}} \le C\mu r\kappa_0^2\sqrt{\frac{d_1\log d_1}{n}}\end{aligned} \tag{35}$$

where $\widetilde{O}$ is a rotation matrix with $\widetilde{O} \in \mathbb{R}^{r\times r}$ and $\widetilde{O}^\top\widetilde{O} = I_r$. Moreover, for the leave-one-out perturbation series $\widehat{M}_{(l)}^{\mathsf{init}}$ which is constructed by knocking off observations from the l-th row, we also have

$$\left\|\widehat{M}_{(l)}^{\mathsf{init}} - M\right\|_{\max} \le \left\|\widehat{M}^{\mathsf{init}} - \widehat{M}_{(l)}^{\mathsf{init}}\right\|_{\max} + \left\|\widehat{M}^{\mathsf{init}} - M\right\|_{\max} \le C\mu r\kappa_0^2\sqrt{\frac{d_1\log d_1}{n}}. \tag{36}$$

Specifically, Chen et al. (2020) (eq 93c) suggests the leave-one-out error on the singular subspace by Wedin's $\sin\Theta$ Theorem:

$$\left\|\widehat{U}_{(l)}^{\mathsf{init}}\widehat{U}_{(l)}^{\mathsf{init}\top} - \widehat{U}^{\mathsf{init}}\widehat{U}^{\mathsf{init}\top}\right\|_{\mathrm{F}} \le \frac{\kappa_0^{\frac{1}{2}}\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{d_1^2 d_2\log d_1}{n}}\sqrt{\frac{\mu r}{d_1}}. \tag{37}$$

A corresponding bound for $\widehat{V}_{(l)}^{\mathsf{init}}$ can also be established by studying the GD iterations on the first d_1 rows and last d_2 rows of $F = \begin{bmatrix} U\Lambda^{\frac{1}{2}} \\ V\Lambda^{\frac{1}{2}} \end{bmatrix}$ respectively.

Combining (35), (36), (37), we know that the assumptions of Theorem 4 in Ma and Xia (2024) are satisfied, with the initialization error constant at most $C_0 \lesssim \mu r \kappa_0^2$. Therefore, we have Theorems 1, 3.

B.2 Proof of Theorem 4

Proof. We write the debiasing estimator as

$$\widehat{M}^{\mathsf{unbs}} = M + \underbrace{\frac{d_1 d_2}{n} \sum_{i=1}^{n} \xi_i X_i}_{\widehat{Z}_1\text{: i.i.d. noises}} + \underbrace{\left(\frac{d_1 d_2}{n} \sum_{i=1}^{n} \left\langle \widehat{M}^{\mathsf{init}} - M, X_i \right\rangle X_i - (\widehat{M}^{\mathsf{init}} - M) \right)}_{\widehat{Z}_2\text{: initialization error}}.$$

By the decomposition of SVD operators (Xia and Yuan, 2021; Ma and Xia, 2024), we have the inequalities uniformly hold for all T, we can write the couple (W_{T_1}, W_{T_2}) as

$$(W_{T_1}, W_{T_2}) = \left(\frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_1) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T_1) \right\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}} + \Delta_{T_1}, \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_2) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T_2) \right\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}} + \Delta_{T_2} \right),$$

where with probability at least $1 - C \log d_1 d_1^{-\tau}$,

$$\max \left\{ \left| \Delta_{T_1} \right|, \left| \Delta_{T_2} \right| \right\} \leq C_4 h_n \tag{38}$$

by the same argument in the proof of Theorem 4 in Ma and Xia (2024). Notice that $\langle \widehat{Z}_1, \mathcal{P}_M(T_i) \rangle$, $i = 1, 2$ are the sum of i.i.d. random variables, with covariance matrix:

$$\begin{aligned} \mathbb{E} \langle \widehat{Z}_1, \mathcal{P}_M(T_1) \rangle \langle \widehat{Z}_1, \mathcal{P}_M(T_2) \rangle &= \frac{d_1^2 d_2^2}{n^2} \sum_{i \in I_2} \xi_i^2 \left\langle X_i, \mathcal{P}_M(T_1) \right\rangle \left\langle X_i, \mathcal{P}_M(T_2) \right\rangle \\ &= \frac{d_1 d_2}{n} \sigma_\xi^2 \sum_{i \in [d_1]} \sum_{j \in [d_2]} e_i^\top \mathcal{P}_M(T_1) e_j e_i^\top \mathcal{P}_M(T_2) e_j \\ &= \frac{d_1 d_2}{n} \sigma_\xi^2 \left\langle \mathcal{P}_M(T_1), \mathcal{P}_M(T_2) \right\rangle. \end{aligned}$$

Then we have

$$\operatorname{cov} \left(\frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_1) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T_1) \right\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}}, \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_2) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T_2) \right\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}} \right) =: \rho_{T_1, T_2}.$$

We jointly control the c.d.f. of them by multivariate Berry-Essen theorem (Stein, 1972; Raič, 2019)

$$\begin{aligned}&\sup_{t_1,t_2\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\left\langle \widehat{Z}_1,\mathcal{P}_M(T_1)\right\rangle}{\sigma_\xi\left\|\mathcal{P}_M(T_1)\right\|_{\mathrm{F}}\sqrt{d_1d_2/n}}\le t_1,\frac{\left\langle \widehat{Z}_1,\mathcal{P}_M(T_2)\right\rangle}{\sigma_\xi\left\|\mathcal{P}_M(T_2)\right\|_{\mathrm{F}}\sqrt{d_1d_2/n}}t_2\right)-\Phi_{\rho_{T_1,T_2}}(t_1,t_2)\right|\\&\le C\mu\left(\frac{1}{1-\rho}\right)^{\frac{3}{2}}\sqrt{\frac{rd_1}{n}}.\end{aligned}\tag{39}$$

The gradient bound $\|\nabla\Phi_\rho(t_1,t_2)\|\le C$ indicates the Lipschitz property of $\Phi_\rho(t_1,t_2)$, which suggests the desired probability bound following (38), (39). □

B.3 Proof of Theorem 5

We remark that our proof of Theorem 5 will give a non-asymptotic bound on the FDR control (Section A.2) with novel techniques, which is totally different from the previous asymptotic analysis in Du et al. (2023); Dai et al. (2022). We proceed to prove the FDR control in the sequel by three steps: we first show that $\mathbb{I}(W_T^{(i)}>t)$ follows weak dependency and asymptotic symmetricity for $T\in\mathcal{H}_0$ when t is in a certain region $[0,L_n]$; then we show that with high probability, the data-driven threshold is in the region $[0,L_n]$; finally we control the power when strong signals dominate signals in the non-null set. Since $n\le O(d_1^{2\tau})$ in general, we treat $h_n=\Omega(\sqrt{\log d_1}/d_1^{\tau/2}\vee(rd_1/n)^{1/4})$ in the proof for simplicity. Here h_n can be chosen smaller as long as n is large. Moreover, to make the FDR control nontrivial, we assume that $q_0\ge cq$ for some $0<c<1$, and we consider the case when q is large as $n\to\infty$. We first start with the asymptotic property of $W_T^{(i)}$.

B.3.1 Weak dependence and symmetricity

From Theorem 1, 3, and definition of h_n, we have the following claim of the asymptotic normality of $W_T^{(1)}$:

Proposition 4. *There exits a constant C_2 such that $W_T^{(1)}$ follows the asymptotic normality rate:*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(W_T^{(1)}>t)-\Phi(-t)\right|\le C_2h_n,\tag{40}$$

for any $T\in\mathcal{H}_0$.

This proposition implies the asymptotic symmetricity of $W_T^{(1)}$, $W_T^{(2)}$, which is crucial for the following analysis. Since $W_T^{(1)}$, $W_T^{(2)}$ are asymptotically normal, the c.d.f. of their product W_T^{rank} will converge to an asymptotically conditional symmetric random variable. We will show that, conditional on the splits $\mathcal{D}_1$, W_T^{rank} is asymptotically symmetric for $T \in \mathcal{H}_0$. Define

$$G(t) = \frac{\sum_{T\in\mathcal{H}_0} \mathbb{P}(W_T^{(1)} Z > t|\mathcal{D}_1)}{q_0},$$

where Z is a standard Gaussian variable. Here since $W_T^{(1)} \in \sigma(\mathcal{D}_1)$, $W_T^{(1)}$ is fixed conditional on $\mathcal{D}_1$.

Denote $L_n = G^{-1}(\frac{\epsilon_n \eta_n}{q_0}) = \inf \left\{t : G(t) \leq \frac{\epsilon_n \eta_n}{q_0}\right\}$, where ϵ_n is a rate to be specified later. We can exploit the following asymptotic symmetric property of $W_T^{(1)}$ to investigate the population version of the following ratio:

$$\mathsf{R}_0 = \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > L)}{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} < -L)}.$$

Here, we introduce a weaker characterization of strong signals, that is

$$\mathcal{S} = \left\{T \in \mathcal{H} : \frac{\sqrt{n}\,|M_T - \theta_T|}{\sigma_\xi \sqrt{d_1 d_2}\,\|\mathcal{P}_M(T)\|_{\mathrm{F}} \sqrt{\log d}} \geq C_{\mathsf{gap}}\right\}, \tag{41}$$

with $\eta_n = |\mathcal{S}|$ for some large constant C_{gap}. In the following proof, we will actually focus on this definition of strong signals. This condition is actually weaker than in our main text, because $\|\mathcal{P}_M(T)\|_{\mathrm{F}} \leq \|T\|_{\ell_1} \max_{i,j} \left\|\mathcal{P}_M(e_i e_j^\top)\right\|_{\mathrm{F}} \leq 3\mu \|T\|_{\ell_1} \sqrt{\frac{r}{d_2}}$. Thus, all the signals that satisfy condition (20) can also satisfy condition (41), meaning that the η_n defined here is always larger than that defined in (20).

Lemma 1. *Conditional on $\mathcal{D}_1$, we have*

$$\sup_{0\leq t\leq L_n} \left|\frac{\sum_{T\in\mathcal{H}_0} \mathbb{P}(W_T^{\mathsf{rank}} > t)}{q_0 G(t)} - 1\right| \leq C_3 \frac{h_n q_0}{\epsilon_n \eta_n}.$$

Proof. We only focus on small h_n. For each $T \in \mathcal{H}_0$, conditional on $\mathcal{D}_1$, Proposition 4 implies that

$$\left|\mathbb{P}(W_T^{\mathsf{rank}} > t) - \mathbb{P}(W_T^{(1)} Z > t|\mathcal{D}_1)\right| \leq C_2 h_n.$$

The definition of L_n also implies $G(t) \geq \frac{\epsilon_n \eta_n}{q_0}$. Then, we can derive the following uniform bound of convergence:

$$\sup_{0\leq t\leq L_n} \left|\frac{\sum_{T\in\mathcal{H}_0} \mathbb{P}(W_T^{\mathsf{rank}} > t)}{q_0 G(t)} - 1\right| \leq \sup_{0\leq t\leq L_n} \left|\frac{\sum_{T\in\mathcal{H}_0} \mathbb{P}(W_T^{\mathsf{rank}} > t) - G(t)}{q_0 G(L_n)}\right| \leq C_3 \frac{h_n q_0}{\epsilon_n \eta_n}.$$

□

Then, we explore the weak dependency of linear forms under signals and correlations assumptions. We will show that, with high probability, the R_0 can be very close to its population version described in Lemma 1. Although we already have the intuition of dependency between different $W_T^{(1)}$ by Theorem 4, the rate provided is not enough for FDR control based on W_T^{rank}. Here, we study the correlation of W_T^{rank} between different T with a more delicate analysis. Let T_1, and T_2 be two different indexing matrices in $\mathcal{H}_0$. To this end, we introduce the following Lemma:

Lemma 2 (Weak dependency of null statistics)**.** *Conditional on $\mathcal{D}_1$,*

$$\sup_{0\le t\le L_n} \frac{\sum_{(T_i,T_j)\in\mathcal{H}^2_{0,weak}} \left|\operatorname{cov}(\mathbb{I}(W_{T_i}^{\mathsf{rank}} > t), \mathbb{I}(W_{T_j}^{\mathsf{rank}} > t))\right|}{q_0^2 G^2(t)} \le C_1 \frac{h_n q_0}{\epsilon_n \eta_n} + C_2 \frac{1}{(\epsilon_n \eta_n q_0)^{\nu/2}}, \tag{42}$$

where ν is the weak correlation parameter defined in (17)*.*

Proof. Suppose we have a pair $(T_1, T_2) \in \mathcal{H}^2_{0,\text{weak}}$. Here we adopt the notation in the proof of Theorem 3: denote

$$\begin{aligned}(W_{T_1}^{(i)}, W_{T_2}^{(i)}) &= \left(\frac{\left\langle \widehat{Z}_1^{(i)}, \mathcal{P}_M(T_1) \right\rangle}{\sigma_\xi \left\|\mathcal{P}_M(T_1)\right\|_{\mathrm{F}} \sqrt{d_1 d_2/n}} + \Delta_{T_1}^{(i)}, \frac{\left\langle \widehat{Z}_1^{(i)}, \mathcal{P}_M(T_2) \right\rangle}{\sigma_\xi \left\|\mathcal{P}_M(T_2)\right\|_{\mathrm{F}} \sqrt{d_1 d_2/n}} + \Delta_{T_2}^{(i)} \right) \\ &:= \left(\widetilde{W}_{T_1}^{(i)} + \Delta_{T_1}^{(i)}, \widetilde{W}_{T_2}^{(i)} + \Delta_{T_2}^{(i)} \right),\ i = 1, 2,\end{aligned}$$

where $\widehat{Z}_1^{(i)}$, $\Delta_T^{(i)}$ are defined analogously as in the proof of Theorem 3. We have $\mathbb{E}(\widetilde{W}_T^{(1)})^2 = \mathbb{E}(\widetilde{W}_T^{(2)})^2 = 1$. Here $\widetilde{W}_T^{(1)}$ and $\widetilde{W}_T^{(2)}$ are standardized averages of n i.i.d. samples and can be regarded as the cores which lead to asymptotic normality of $W_T^{(1)}$, $W_T^{(2)}$. By the proof of Theorem 3, the remainder term $\Delta_T^{(i)}$ is controlled by:

$$\mathbb{P}\left(\left|\Delta_T^{(i)}\right| > c_1 h_n \right) \le C_2 \frac{\log d_1}{d_1^\tau}. \tag{43}$$

For $i = 1$, as is shown in the proof of Theorem 3, by multivariate Berry–Esseen theorem, $(\widetilde{W}_{T_1}^{(2)}, \widetilde{W}_{T_2}^{(2)})$ converges to normal variable $\omega_1 \sim \mathcal{N}(0, R)$ conditional on E_0 where $R_{11} = R_{22} = 1$ and $R_{12} = R_{21} = \operatorname{cov}(\widetilde{W}_{T_1}^1, \widetilde{W}_{T_2}^1) = \rho_{T_1,T_2}$ with the error bound:

$$\left| \mathbb{P}\left((\widetilde{W}_{T_1}^{(2)}, \widetilde{W}_{T_2}^{(2)}) \in A \middle| E_0 \right) - \mathbb{P}(\omega_1 \in A) \right| \le C(1/(1-\rho))^{\frac{3}{2}} \mu \sqrt{\frac{r d_1}{n}}, \tag{44}$$

for any convex set $A \subseteq \mathbb{R}^2$. Here the $\rho := \rho_{T_1,T_2}$ is the correlation between $W_T^{(1)}$, $W_T^{(2)}$ defined in (16), and it follows $|\rho| \le c q_0^{-\nu} \le \frac{1}{2}$ due to the weak dependency in $\mathcal{H}^2_{0,\text{weak}}$. By the following calculation of the covariance between $\mathbb{I}(W_{T_1}^{\mathsf{rank}} > t)$ and $\mathbb{I}(W_{T_2}^{\mathsf{rank}} > t)$ conditional on $\mathcal{D}_1$, we have:

$$\begin{aligned}
&\left|\operatorname{cov}(\mathbb{I}(W_{T_1}^{\mathsf{rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{rank}} > t))\right| \\
&= \left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t, W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t)\mathbb{P}(W_{T_2}^{(1)}W_{T_2}^{(2)} > t)\right| \\
&\leq \left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t, W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}w_{11} > t, W_{T_2}^{(1)}w_{12} > t)\right| \text{ (1)} \\
&+ \left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t)\mathbb{P}(W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}w_{11} > t)\mathbb{P}(W_{T_2}^{(1)}w_{12} > t)\right| \text{ (2)} \\
&+ \left|\mathbb{P}(W_{T_1}^{(1)}w_{11} > t, W_{T_2}^{(1)}w_{12} > t) - \mathbb{P}(W_{T_1}^{(1)}w_{11} > t)\mathbb{P}(W_{T_2}^{(1)}w_{12} > t)\right| \text{ (3).}
\end{aligned}$$

Term (1), (2), (3) can be controlled separately. For (1), conditional on $\mathcal{D}_1$, we invoke multivariate Berry–Esseen theorem (44) to bound the joint c.d.f. of $(W_{T_1}^{(2)}, W_{T_2}^{(2)})$ by

$$\begin{aligned}
&\mathbb{P}(W_{T_1}^{(2)} > t_1, W_{T_2}^{(2)} > t_2) \\
&\leq \mathbb{P}(W_{T_1}^{(2)} > t_1, W_{T_2}^{(2)} > t_2, \left|\Delta_{T_1}^{(2)}\right| \leq c_1 h_n, \left|\Delta_{T_2}^{(2)}\right| \leq c_1 h_n) + \frac{2c_2 \log d_1}{d_1^{\tau}} \\
&\leq \mathbb{P}(\widetilde{W}_{T_1}^{(2)} > t_1 - c_1 h_n, \widetilde{W}_{T_2}^{(2)} > t_2 - c_1 h_n) + \frac{2c_2 \log d_1}{d_1^{\tau}} \\
&\leq \mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2) + c_1\left[\phi(t_1)\mathbb{P}(\omega_{12} > t_2 | \omega_{11} = t_1) + \phi(t_2)\mathbb{P}(\omega_{11} > t_1 | \omega_{12} = t_2)\right] h_n + C_2 h_n^2,
\end{aligned}$$

where we apply Taylor expansion to the c.d.f. of normal distribution $\omega_1 \sim \mathcal{N}(0, R)$ and apply the upper bound $\log d_1 \cdot d_1^{-\tau} \leq h_n^2$. Analogously, it also holds that

$$\begin{aligned}
&\mathbb{P}(W_{T_1}^{(2)} > t_1, W_{T_2}^{(2)} > t_2) \geq \mathbb{P}(\widetilde{W}_{T_1}^{(2)} > t_1 + c_1 h_n, \widetilde{W}_{T_2}^{(2)} > t_2 + c_1 h_n) - \frac{2c_2 \log d_1}{d_1^{\tau}} \\
&\geq \mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2) - c_1\left[\phi(t_1)\mathbb{P}(\omega_{12} > t_2 | \omega_{11} = t_1) + \phi(t_2)\mathbb{P}(\omega_{11} > t_1 | \omega_{12} = t_2)\right] h_n - C_2 h_n^2.
\end{aligned}$$

We conclude that, conditional on $\mathcal{D}_1$

$$\begin{aligned}
&\left|\mathbb{P}(W_{T_1}^{(2)} > t_1, W_{T_2}^{(2)} > t_2) - \mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2)\right| \\
&\leq c_1\left[\phi(t_1)\mathbb{P}(\omega_{12} > t_2 | \omega_{11} = t_1) + \phi(t_2)\mathbb{P}(\omega_{11} > t_1 | \omega_{12} = t_2)\right] h_n + C_2 h_n^2.
\end{aligned}$$

Using the Lipschitz property of $\Phi(t)$, we have

$$\begin{aligned}
&\left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t, W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}\omega_{11} > t, W_{T_2}^{(1)}\omega_{12} > t)\right| \\
&\leq 2c_1 h_n \left(\mathbb{P}(W_{T_1}^{(1)}\omega_{11} > t) + \mathbb{P}(W_{T_2}^{(1)}\omega_{12} > t)\right) + C h_n^2.
\end{aligned} \tag{45}$$

For (2), the proof of Lemma 1 also implies the following bound

$$\begin{aligned}
&\left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t)\mathbb{P}(W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}w_{11} > t)\mathbb{P}(W_{T_2}^{(1)}w_{12} > t)\right| \\
&\leq 2c_1 h_n \left(\mathbb{P}(W_{T_1}^{(1)}\omega_{11} > t) + \mathbb{P}(W_{T_2}^{(1)}\omega_{12} > t)\right) + C h_n^2,
\end{aligned} \tag{46}$$

by the same argument conditional on E_0 and $\mathcal{D}_1$. Our next step is to compare the c.d.f. of (ω_1, ω_2) with standard Gaussian (Z_1, Z_2) to control the term (3), the difference between $\mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2)$ and $\Phi(-t_1)\Phi(-t_2)$. Since $(T_1, T_2) \in \mathcal{H}^2_{0,\text{weak}}$, the covariance between w_{11}, w_{12} is thus bounded by: $|\rho| \le cq_0^{-\nu}$.

We invoke the property of bivariate Gaussian copula (Meyer, 2013):

$$|\mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2) - \Phi(-t_1)\Phi(-t_2)| = \left| \int_0^\rho \phi_2(-t_1, -t_2, z) dz \right|,$$

where $\phi_2(x, y, z)$ is the p.d.f of bivariate normal distribution with correlation coefficient z. Without loss of generality, assume $t_1, t_2 > 0$ are away from 0. Thus, it is clear that

$$\begin{aligned} |\mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2) - \Phi(-t_1)\Phi(-t_2)| &\le \int_0^\rho \phi_2(-t_1, -t_2, z) dz, \\ &\le \frac{\rho}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{t_1^2 + t_2^2}{2} + \frac{\rho t_1 t_2}{(1-\rho^2)} \right) \\ &\le \frac{2\rho}{2\pi} \exp\left(-\frac{t_1^2 + t_2^2}{2}(1 - c\rho) \right) \\ &= 2\rho \left[\phi(-t_1)\phi(-t_2) \right]^{1-c\rho}. \end{aligned}$$

For any $\nu > 0$, there exist $C_\nu > 0$ such that $\Phi(-t)^\nu \le C_\nu / t$ for all $t > 0$. Because by Mill's ratio, we have:

$$\Phi(-t)^\nu \le \frac{\phi(-t)^\nu}{t^\nu} \le C_\nu \frac{1}{t^{1-\nu}} \frac{1}{t^\nu} = C_\nu \frac{1}{t},$$

where we use the fact that $\phi(-t)^\nu \le C_\nu t^{-(1-\nu)}$. Now combine this with the upper bound of $\phi(-t)$: $\phi(-t) \le C(t+1)\Phi(-t)$, we have:

$$\begin{aligned} |\mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2) - \Phi(-t_1)\Phi(-t_2)| &\le 2\rho \left[\phi(-t_1)\phi(-t_2) \right]^{1-c\rho} \\ &\le 2\rho \left[C(\Phi(-t_1)^{-\nu} + 1)\Phi(-t_1)(\Phi(-t_2)^{-\nu} + 1)\Phi(-t_2) \right]^{1-c\rho} \\ &\le C\rho \left[\Phi(-t_1)\Phi(-t_2) \right]^{(1-\nu)(1-c\rho)}, \end{aligned} \tag{47}$$

for the term (3). Together with (45), (46), we can show that

$$
\begin{aligned}
&\sup_{0\le t\le L_n}\sum_{(T_i,T_j)\in\mathcal{H}_{0,\text{weak}}^2}\frac{\left|\operatorname{cov}(\mathbb{I}(W_{T_i}^{\mathsf{rank}}>t),\mathbb{I}(W_{T_j}^{\mathsf{rank}}>t))\right|}{q_0^2G^2(t)}\le\frac{8c_1h_nq_0G(t)}{q_0^2G(t)^2}\\
&+\sup_{0\le t\le L_n}\frac{\sum_{(T_i,T_j)\in\mathcal{H}_{0,\text{weak}}^2}C\rho\left[\mathbb{P}(W_{T_i}^{(1)}Z>t)\mathbb{P}(W_{T_j}^{(1)}Z>t)\right]^{(1-\nu)(1-c\rho)}}{q_0^2G(t)^2}\\
&\le C\frac{h_nq_0}{\epsilon_n\eta_n}+\sup_{0\le t\le L_n}\frac{C\rho\left(\sum_{T\in\mathcal{H}_0}\mathbb{P}(W_T^{(1)}Z>t)^{(1-\nu)(1-c\rho)}\right)^2}{q_0^2G(t)^2}\\
&\le C\frac{h_nq_0}{\epsilon_n\eta_n}+\sup_{0\le t\le L_n}C\rho\frac{(G(t))^{2(1-\nu)(1-c\rho)}}{G(t)^2}\\
&\le C\frac{h_nq_0}{\epsilon_n\eta_n}+C\rho(\frac{q_0}{\epsilon_n\eta_n})^{3\nu}.
\end{aligned}
$$

The argument above is valid for any ν, thus, we choose 3ν to be the $\frac{\nu}{2}$, where ν is defined in (17). It thus finishes the proof. □

We now apply the weak dependency yielded in Lemma 2 to derive a uniform bound between R and its population version:

Lemma 3. *For any $\varepsilon>0$, conditional on $\mathcal{D}_1$, it holds that*

$$
\mathbb{P}\left(\sup_{0\le t\le L_n}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>t)}{q_0G(t)}-1\right|\ge\varepsilon\right)\le\frac{C}{\varepsilon^2}\log(\frac{q_0}{\epsilon_n\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\epsilon_n^2\eta_n^2}\right)^{\frac{1}{2}}+\left(\frac{h_nq_0}{\epsilon_n\eta_n}+\frac{1}{(\epsilon_n\eta_nq_0)^{v/2}}\right)^{\frac{1}{2}}\right).
$$

Proof. To prove the uniform convergence in probability, we define a grid on $[0,L_n]$:

$$
\left\{t_k=G^{-1}\left(\frac{1}{2}(2G(L_n))^{\frac{k}{K}}\right)\right\}_{k=0}^{K},
$$

which equates each $G(t_k)$ with $\frac{1}{2}(2G(L_n))^{\frac{k}{K}}$. Then for each $t\in[t_{k-1},t_k)$, the ratio can be bounded by:

$$
\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>t_k)}{q_0G(t_{k-1})}\le\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>t)}{q_0G(t)}\le\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>t_{k-1})}{q_0G(t_k)}.
$$

Define $(2G(L_n))^{\frac{1}{K}}=r_K$, we have $G(t_k)/G(t_{k-1})=r_K$, and

$$
\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>t)}{q_0G(t)}-1\right|\le\sup_{i=k-1,k}\frac{1}{r_K}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>t_i)}{q_0G(t_i)}-1\right|+|r_K-1|\vee\left|\frac{1}{r_K}-1\right|,
$$

for each $t \in [t_{k-1}, t_k)$. Then for any $t \in [0, L_n]$, it suffices to control the quantities

$$\sup_{k=0,\ldots,K} \frac{1}{r_K} \left| \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > t_k)}{q_0 G(t_k)} - 1 \right| \leq \sup_{k=0,\ldots,K} \frac{1}{r_K} \left| \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > t_k) - \mathbb{P}(W_T^{\mathsf{rank}} > t_k)}{q_0 G(t_k)} \right| + C_3 \frac{h_n q_0}{\epsilon_n \eta_n}$$

and $|r_K - 1| \vee \left|\frac{1}{r_K} - 1\right|$ by Proposition 4. Denote

$$D_n = \sup_{k=0,\ldots,K} \frac{1}{r_K} \left| \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > t_k) - \mathbb{P}(W_T^{\mathsf{rank}} > t_k)}{q_0 G(t_k)} \right|.$$

It follows that

$$\begin{aligned} \mathbb{E}D_n^2 &\leq \frac{K}{r_K^2} \mathbb{E} \left| \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{rank}} > t_k) - \mathbb{P}(W_T^{\mathsf{rank}} > t_k)}{q_0 G(t_k)} \right|^2 \\ &\leq \frac{K}{r_K^2} \frac{\sum\limits_{(T_1,T_2)\in\mathcal{H}_{0,\text{weak}}^2} \left|\text{cov}(\mathbb{I}(W_{T_1}^{\mathsf{rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{rank}} > t))\right| + \sum\limits_{T_1,T_2\in\mathcal{H}_{0,\text{strong}}^2} \left|\text{cov}(\mathbb{I}(W_{T_1}^{\mathsf{rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{rank}} > t))\right|}{q_0^2 G^2(t)}, \end{aligned} \tag{48}$$

for any $t \in \{t_k\}$. Since the number of strong dependency pairs $\left|\mathcal{H}_{0,\text{strong}}^2\right| \leq \beta_{\mathsf{s}} q_0^2$, we have

$$\frac{\sum\limits_{T_1,T_2\in\mathcal{H}_{0,\text{strong}}^2} \left|\text{cov}(\mathbb{I}(W_{T_1}^{\mathsf{rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{rank}} > t))\right|}{q_0^2 G^2(t)} \leq \frac{\beta_{\mathsf{s}} q_0^2}{\epsilon_n^2 \eta_n^2},$$

for any $t \in [0, L_n]$. For the weak dependency pair,

$$\frac{\sum\limits_{T_1,T_2\in\mathcal{H}_{0,\text{weak}}^2} \left|\text{cov}(\mathbb{I}(W_{T_1}^{\mathsf{rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{rank}} > t))\right|}{q_0^2 G^2(t)} \leq C_1 \frac{h_n q_0}{\epsilon_n \eta_n} + C_2 \frac{1}{(\eta_n q_0)^{v/2}},$$

where we apply our previous results in Lemma 2. What remains for us is to specify the density of grid K. Choose a constant ς and we set

$$K = \log(\frac{q_0}{\epsilon_n \eta_n}) \min\left\{ \left(\frac{q_0^2 \beta_{\mathsf{s}}}{\eta_n^2 \epsilon_n}\right)^{-\varsigma}, \left(\frac{q_0 h_n}{\eta_n \epsilon_n} + \frac{1}{(\epsilon_n \eta_n q_0)^{v/2}}\right)^{-\varsigma} \right\},$$

then it is clear that $1 \leq \frac{1}{r_k} \leq \left[\frac{q_0}{\epsilon_n \eta_n}\right]^{1/K} \to 1$, and $K(\frac{\beta_{\mathsf{s}} q_0^2}{\epsilon_n^2 \eta_n^2} + \frac{h_n q_0}{\epsilon_n \eta_n}) \to 0$. Therefore

$$\begin{aligned} |r_K - 1| \vee \left|\frac{1}{r_K} - 1\right| &\leq C \frac{1}{K} \log(\frac{q_0}{\epsilon_n \eta_n}) \leq \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\epsilon_n^2 \eta_n^2}\right)^{\varsigma} + \left(\frac{h_n q_0}{\epsilon_n \eta_n} + \frac{1}{(\epsilon_n \eta_n q_0)^{v/2}}\right)^{\varsigma} \right) \\ \mathbb{E}D_n^2 \leq CK \left(\frac{\beta_{\mathsf{s}} q_0^2}{\eta_n^2} + \frac{h_n q_0}{\epsilon_n \eta_n}\right) &\leq C \log(\frac{q_0}{\epsilon_n \eta_n}) \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\epsilon_n^2 \eta_n^2} + \frac{1}{(\epsilon_n \eta_n q_0)^{v/2}}\right)^{1-\varsigma} + \left(\frac{h_n q_0}{\epsilon_n \eta_n}\right)^{1-\varsigma} \right). \end{aligned}$$

We can finish the proof of uniform convergence by using Markov's inequality with $\varsigma = \frac{1}{2}$. □

Recall the main theorem. For the ratio $\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>L)}{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}<-L)}$, we have

$$\mathsf{R}_0 = \frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>L)}{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}<-L)} = \frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}>L)}{q_0G(t)}\cdot\frac{q_0G(t)}{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}}<-L)}.$$

Then, it's clear that, under the event that $L \le L_n$, if Lemma 3 holds for a ε, then we have

$$\mathsf{R}_0 \le \frac{1+\varepsilon}{1-\varepsilon} \le 1+\frac{2\varepsilon}{1-\varepsilon} \le 1+3\varepsilon,$$

with probability at least $1-\frac{C}{\varepsilon^2}\log(\frac{q_0}{\epsilon_n\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\epsilon_n^2\eta_n^2}\right)^{\frac{1}{2}}+\left(\frac{h_nq_0}{\epsilon_n\eta_n}+(\epsilon_n\eta_nq_0)^{-\nu/2}\right)^{\frac{1}{2}}\right)$. By Lemma 3, we now successfully reduce our problem to proving our data-driven threshold $L\le L_n$ with high probability.

B.3.2 Threshold control

The gist of asymptotic threshold control is that when we choose L_n as the threshold and d_1, d_2, n go large, entries with strong signals in $\mathcal{S}$ can always pass the test, and other entries with weak signals or no signal can pass the test will little possibility. We first focus on the entries with strong signals. Denote the standardized signal $\delta_T = (M_T-\theta_T)/(\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\sqrt{d_1d_2/n})$, and $\widehat{W_T^1} = W_T^{(1)}-\delta_T$, $\widehat{W_T^2} = W_T^{(2)}-\delta_T$. Given any $T\in\mathcal{H}_1$, following the argument that is similar to the proof in Lemma 1, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(W_T^{\mathsf{rank}}>t)-\mathbb{P}\left((Z_1+\delta_T)(Z_2+\delta_T)>t\right)\right| \le Ch_n.$$

Here (Z_1, Z_2) are independent standard Gaussian. Without loss of generality, assume $M_T-\theta_T>0$. Then,

$$\begin{aligned}\mathbb{P}(W_T^{\mathsf{rank}}<L_n) \le& \mathbb{P}\left((Z_1+\delta_T)(Z_2+\delta_T)<L_n\right)+Ch_n\\ \le& 1-\mathbb{P}\left((Z_1+\delta_T)(Z_2+\delta_T)\ge L_n\right)+Ch_n\\ \le& 1-\mathbb{P}\left(Z_1\ge-\delta_T+\sqrt{L_n}\right)^2+Ch_n.\end{aligned}$$

Here, we use the fact that

$$\left\{Z_1\ge-\delta_T+\sqrt{L_n}\right\}\cap\left\{Z_2\ge-\delta_T+\sqrt{L_n}\right\}\subseteq\left\{(Z_1+\delta_T)(Z_2+\delta_T)\ge L_n\right\}.$$

An upper bound of $G(t)$ is given by

$$G(t) = \frac{\sum_{T\in\mathcal{H}_0}\mathbb{P}(W_T^{(1)}Z>t|\mathcal{D}_0,\mathcal{D}_1)}{q_0} \le \frac{\sqrt{2}}{\sqrt{\pi}}\exp\left(-\frac{t^2}{2\max_{T\in\mathcal{H}_0}\left|W_T^{(1)}\right|^2}\right).$$

From Theorem 1, an uniform upper bound of $\left|W_T^{(1)}\right|$ is given by:

$$\mathbb{P}\left(\max_{T\in\mathcal{H}_0}\left|W_T^{(1)}\right| \geq C(h_n+\sqrt{\log d_1})\right) \leq \frac{1}{d_1^{\tau-1}}.$$

If $T \in \mathcal{S}$, then $\delta_T \geq C_{\mathsf{gap}}\sqrt{\log d_1}$ by the definition of $\mathcal{S}$. The definition of L_n implies that $L_n \leq C\sqrt{\log(\frac{q_0}{\epsilon_n\eta_n})}\cdot\sqrt{\log d_1} \ll \log(\frac{1}{h_n}) \vee (\log d_1)$. Generally, we have $d^{-10} \leq h_n$, thus the term $\log(\frac{1}{h_n})$ can be omitted. Assume C_{gap} is large. It is clear that

$$\mathbb{P}\left(Z_1 \geq -\delta_T + \sqrt{L_n}\right)^2 \geq \mathbb{P}\left(Z_1 \geq -C\sqrt{(\log d_1)}\right)^2 \geq (1-ch_n)^2,$$

i.e., $\mathbb{P}(W_T^{\mathsf{rank}} < L_n) \leq Ch_n$. For any $\varepsilon > 0$, we compute the probability that $\sum\limits_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) = \eta_n$ by finding its complement:

$$\mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \leq (1-\varepsilon)\eta_n) = \mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} < L_n) > \varepsilon\eta_n) \leq \frac{\sum\limits_{T\in\mathcal{S}}\mathbb{P}(W_T^{\mathsf{rank}} < L_n)}{\varepsilon\eta_n} \leq Ch_n/\varepsilon,$$

i.e., $\mathbb{P}(\sum\limits_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \leq (1-\varepsilon)\eta_n) \to 0$, $\mathbb{P}(\sum\limits_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \geq \eta_n) \to 1$. This indicates that, all the signals in $\mathcal{S}$ can pass our test. For our data-driven threshold (15), we have

$$\sum_{T\in\mathcal{H}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \geq \sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \geq \frac{3}{4}\eta_n, \tag{49}$$

with probability at least $1 - Ch_n$

Consider the probability $\mathbb{P}(\sum\limits_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}} < -L_n) \geq \frac{\alpha}{4}\eta_n)$ for the no-signal linear forms $T \in \mathcal{H}_0$. As we have shown in the proof of Lemma 3, we have

$$\mathbb{P}(\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}} < -L_n) \geq 2\epsilon_n\eta_n) \leq \log(\frac{q_0}{\epsilon_n\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\epsilon_n^2\eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_nq_0}{\epsilon_n\eta_n} + (\epsilon_n\eta_nq_0)^{-\nu/2}\right)^{\frac{1}{2}}\right) \to 0,$$

and consequently, by taking $\epsilon_n = \alpha/8$,

$$\begin{aligned}\mathbb{P}(\sum_{T\in\mathcal{H}}\mathbb{I}(W_T^{\mathsf{rank}} < -L_n) \geq \frac{3}{4}\alpha\eta_n) &\leq \mathbb{P}(2\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{rank}} < -L_n) \geq \frac{\alpha}{2}\eta_n) + \mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} < -L_n) \geq \frac{\alpha}{4}\eta_n)\\ &\leq \log(\frac{q_0}{\alpha\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\alpha^2\eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_nq_0}{\alpha\eta_n} + (\epsilon_n\eta_nq_0)^{-\nu/2}\right)^{\frac{1}{2}}\right) + Ch_n.\end{aligned} \tag{50}$$

Combining (49) and (50), it is sufficient to conclude that

$$\mathbb{P}\left(\frac{\sum\limits_{T\in\mathcal{H}}\mathbb{I}\left(T : W_T^{\mathsf{rank}} < -L_n\right)}{\left(\sum\limits_{T\in\mathcal{H}}\mathbb{I}\left(T : W_T^{\mathsf{rank}} > L_n\right)\right)\vee 1} \geq \alpha\right) \leq \log(\frac{q_0}{\alpha\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\alpha^2\eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_nq_0}{\alpha\eta_n} + (\alpha\eta_nq_0)^{-\nu/2}\right)^{\frac{1}{2}}\right) + Ch_n,$$

i.e., $\mathbb{P}(L \leq L_n) \to 1$.

B.3.3 Power control

From the discussion on the threshold control, it is clear that for any ε,

$$\mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \le (1-\varepsilon)\eta_n) \le \frac{\sum_{T\in\mathcal{S}}\mathbb{P}(W_T^{\mathsf{rank}} < L_n)}{\varepsilon\eta_n} \le Ch_n/\varepsilon.$$

Under the event that $L \le L_n$, this also implies that with probability at least $1 - Ch_n/\varepsilon$,

$$(1-\varepsilon)\eta_n \le \sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L_n) \le \sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L).$$

The probability of $\{L \le L_n\}$ is lower bounded in Section B.3.2. We can, therefore, get the power:

$$\text{POWER} = \frac{\sum_{T\in\mathcal{H}_1}\mathbb{I}(W_T^{\mathsf{rank}} > L)}{q_1} \ge \frac{\sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{rank}} > L)}{\eta_n}\cdot\frac{\eta_n}{q_1} \ge (1-\varepsilon)\frac{\eta_n}{q_1},$$

with probability at least:

$$1 - C\log(\frac{q_0}{\alpha\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\alpha^2\eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_nq_0}{\alpha\eta_n} + (\alpha\eta_nq_0)^{-\nu/2}\right)^{\frac{1}{2}}\right) - C\varepsilon^{-1}h_n.$$

B.4 Proof of Proposition 1

Proof. By definition, we can equally use the covariance matrix $\mathbf{Q}^* = (\mathbf{X}_{\mathcal{A}}^{*\top}\mathbf{X}_{\mathcal{A}}^*)^{-1} = \left(\Sigma_{\mathcal{A}}^{-\frac{1}{2}\top}\Sigma_{\mathcal{A}}^{-\frac{1}{2}}\right)^{-1}$ to derive the correlation coefficient matrix. Here in the proof, we use bold symbols like $\mathbf{Q}$ to distinguish our analysis from the Q in the Algorithm 2 of the main text, although they lead to the same correlation structure. We will show that, if two linear forms indexed by T_i, T_j are weakly correlated in $\mathbf{Q}^*$, i.e.,

$$\left|\frac{\mathbf{Q}_{ij}^*}{\sqrt{\mathbf{Q}_{ii}^*\mathbf{Q}_{jj}^*}}\right| = \frac{\left|e_i^\top\left(\mathbf{X}_{\mathcal{A}}^{*\top}\mathbf{X}_{\mathcal{A}}^*\right)^{-1}e_j\right|}{\sqrt{\mathbf{Q}_{ii}^*\mathbf{Q}_{jj}^*}} \le C_1q_n^{-\nu},$$

then, in the data-driven covariance matrix $\mathbf{Q}$, they are also weakly correlated:

$$\left|\frac{\mathbf{Q}_{ij}}{\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}}\right| \le C_2q_n^{-\nu},$$

with probability at least $1 - Cd_1^{-2}\log d_1$. By definition, the covariance matrix of $\widehat{\mathsf{w}}^{(2)}$ without normalization is

$$\begin{aligned}\mathbf{Q} &= \left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}\Sigma\mathbf{X}^\top\mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\\ &= \left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1} + \left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^\top\Delta\Sigma\mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1},\end{aligned}$$

where we define $\Delta\Sigma = \mathbf{X}\Sigma\mathbf{X}^\top - I = \widehat{\Sigma}^{-\frac{1}{2}}(\Sigma - \widehat{\Sigma})\widehat{\Sigma}^{-\frac{1}{2}}$. The following Lemma characterizes the precision of our covariance estimation:

Lemma 4. *Suppose that we use $\widehat{U} = \widehat{U}^{\mathsf{init}}$, $\widehat{V} = \widehat{V}^{\mathsf{init}}$ obtained from $\mathcal{D}_1$ to estimate Σ:*

$$\widehat{\Sigma} = T_{\mathcal{H}}(I_{d_1 d_2} - \widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top) T_{\mathcal{H}}^\top.$$

Then with probability at least $1 - Cd_1^{-\tau}\log d_1$, we have

$$\left\|\Sigma^{-\frac{1}{2}}(\Sigma - \widehat{\Sigma})\Sigma^{-\frac{1}{2}}\right\| \le CC_0 \frac{\kappa_T \sigma_\xi}{\lambda_{\min}} \left(\frac{\mathrm{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1\right) \sqrt{\frac{\kappa_1 d_1^2 d_2 \log d_1}{n}}. \tag{51}$$

For simplicity, we denote $\kappa_T' = \kappa_T \left(\frac{\mathrm{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1\right)$. Lemma 4 implies the bound of eigenvalue : $\left|\lambda_i(\Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}}) - 1\right| = o_p(1)$ for all eigenvalues. Thus, the eigenvalues of its inverse can also be bounded by the rate in (51), i.e.,

$$\|\Delta\Sigma\| \le CC_0 \frac{\kappa_T' \sigma_\xi}{\lambda_{\min}} \sqrt{\frac{\kappa_1 d_1^2 d_2 \log d_1}{n}}.$$

We then have

$$\left|\mathbf{Q}_{ij} - \mathbf{Q}_{ij}^*\right| \le \left|e_i^\top \left(\left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1} - \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)^{-1}\right) e_j\right| + \left|e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1} \mathbf{X}_{\mathcal{A}}^\top \Delta\Sigma \mathbf{X}_{\mathcal{A}} \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1} e_j\right|. \tag{52}$$

Denote $\mathbf{Q}' = \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1}$. The first term in (52) can be controlled by:

$$\begin{aligned}
&\left|e_i^\top \left(\left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1} - \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)^{-1}\right) e_j\right| \\
&= \left|e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1} \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* - \mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right) \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)^{-1} e_j\right| \\
&\le CC_0 \frac{\kappa_1^{1.5} \kappa_T' \sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}} \sqrt{\mathbf{Q}_{ii}' \mathbf{Q}_{jj}^*},
\end{aligned} \tag{53}$$

where we use the fact that

$$\begin{aligned}
\left\|\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* - \mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right\| &\le \left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\| \le \left\|\Sigma^{-1}(\widehat{\Sigma} - \Sigma)\Sigma^{-1}\right\| + O\left(\left\|\widehat{\Sigma} - \Sigma\right\|^2\right) \\
&\le \frac{1}{\lambda_{\min}(\Sigma)} \left\|\Sigma^{-\frac{1}{2}}(\Sigma - \widehat{\Sigma})\Sigma^{-\frac{1}{2}}\right\| + o\left(\left\|\widehat{\Sigma} - \Sigma\right\|\right) \\
&\le CC_0 \frac{\kappa_T' \sigma_\xi}{\lambda_{\min}(\Sigma)\lambda_{\min}} \sqrt{\frac{\kappa_1 d_1^2 d_2 \log d_1}{n}},
\end{aligned}$$

by Fréchet derivative (Higham, 2008; Al-Mohy and Higham, 2009) and Lemma 4, and also we have

$$\begin{aligned}
\left\|\left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)^{-1} e_j\right\|^2 &\le \frac{1}{\lambda_{\min}\left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)} \left\|e_j^\top \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)^{-1} \mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^*\right)^{-1} e_j\right\| \\
&\le \lambda_{\max}(\Sigma) \mathbf{Q}_{jj}^*,
\end{aligned}$$

$$
\begin{aligned}
\left\|\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i\right\|^2 \leq{} & \frac{1}{\lambda_{\min}\left(\mathbf{X}_{\mathcal{A}}^{*\top}\mathbf{X}_{\mathcal{A}}^{*}\right)}\left|e_i^{\top}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i\right| \\
& + \frac{1}{\lambda_{\min}\left(\mathbf{X}_{\mathcal{A}}^{*\top}\mathbf{X}_{\mathcal{A}}^{*}\right)}\left|e_i^{\top}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1}\left(\mathbf{X}_{\mathcal{A}}^{*\top}\mathbf{X}_{\mathcal{A}}^{*} - \mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i\right| \\
\leq{} & \lambda_{\max}(\Sigma)\mathbf{Q}_{ii}' + CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}\left\|\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i\right\|^2,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
\left\|\left(\mathbf{X}_{\mathcal{A}}^{*\top}\mathbf{X}_{\mathcal{A}}^{*}\right)^{-1} e_j\right\| &\leq \sqrt{\lambda_{\max}(\Sigma)\mathbf{Q}_{jj}^{*}} \\
\left\|\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i\right\| &\leq (1+c)\sqrt{\lambda_{\max}(\Sigma)\mathbf{Q}_{ii}'}.
\end{aligned}
$$

The second term in (52) can be bounded given that

$$
\begin{aligned}
\left|e_i^{\top}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^{\top}\Delta\Sigma\mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_j\right| &\leq \left\|\mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_j\right\|\left\|\mathbf{X}_{\mathcal{A}}\left(\mathbf{X}_{\mathcal{A}}^{\top}\mathbf{X}_{\mathcal{A}}\right)^{-1} e_i\right\|\|\Delta\Sigma\| \\
&= \sqrt{\mathbf{Q}_{ii}'\mathbf{Q}_{jj}'}\,\|\Delta\Sigma\| \\
&\leq CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}\sqrt{\mathbf{Q}_{ii}'\mathbf{Q}_{jj}'}.
\end{aligned}
$$

However, notice that,

$$
\left|\frac{\mathbf{Q}_{ii}-\mathbf{Q}_{ii}'}{\mathbf{Q}_{ii}'}\right| \leq CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}.
$$

We can conclude that

$$
\left|\mathbf{Q}_{ij}-\mathbf{Q}_{ij}^{*}\right| \leq CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}\left(\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}^{*}}+\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}\right).
$$

Setting $i=j$, we also have

$$
\frac{\left|\mathbf{Q}_{jj}-\mathbf{Q}_{ij}^{*}\right|}{\mathbf{Q}_{jj}} \leq CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}\left(\sqrt{1+\frac{\left|\mathbf{Q}_{jj}-\mathbf{Q}_{jj}^{*}\right|}{\mathbf{Q}_{jj}}}+1\right),
$$

i.e.,

$$
\frac{\left|\mathbf{Q}_{jj}-\mathbf{Q}_{jj}^{*}\right|}{\mathbf{Q}_{jj}} \leq CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}.
$$

We now compare the difference of correlation coefficients:

$$
\begin{aligned}
\left|\frac{\mathbf{Q}_{ij}}{\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}}-\frac{\mathbf{Q}_{ij}^{*}}{\sqrt{\mathbf{Q}_{ii}^{*}\mathbf{Q}_{jj}^{*}}}\right| \leq{} & \frac{\left|\mathbf{Q}_{ij}-\mathbf{Q}_{ij}^{*}\right|}{\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}}+\left|\mathbf{Q}_{ij}^{*}\right|\frac{\left|\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}-\sqrt{\mathbf{Q}_{ii}^{*}\mathbf{Q}_{jj}^{*}}\right|}{\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}\sqrt{\mathbf{Q}_{ii}^{*}\mathbf{Q}_{jj}^{*}}} \\
& + \left|\mathbf{Q}_{ij}-\mathbf{Q}_{ij}^{*}\right|\frac{\left|\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}-\sqrt{\mathbf{Q}_{ii}^{*}\mathbf{Q}_{jj}^{*}}\right|}{\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}}\sqrt{\mathbf{Q}_{ii}^{*}\mathbf{Q}_{jj}^{*}}} \\
\leq{} & CC_0\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{d_1^2 d_2\log d_1}{n}}.
\end{aligned}
$$

If the assumption on the signal strength, i.e.,

$$\frac{\kappa_1^{1.5}\kappa_T'\sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}} \lesssim \frac{1}{q^\nu},$$

is satisfied, we also have $|\mathbf{Q}_{ij}|/\sqrt{\mathbf{Q}_{ii}\mathbf{Q}_{jj}} \lesssim q^{-\nu}$, which indicates that these two linear forms are also weakly correlated in data-driven covariance matrix $\mathbf{Q}$. □

B.5 Proof of Proposition 2

Proof. We start with the decomposition of LASSO response $\mathbf{y}_1 = \mathbf{X}\mathbf{W}^{(1)}$:

$$\mathbf{y}_1 = \widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}\widehat{\mathsf{w}} + \widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}\widetilde{\mathbf{W}},$$

where $\widehat{\mathsf{w}}_i = \frac{M_{T_i}-\theta_{T_i}}{\widehat{\sigma}_\xi^{(1)}\sqrt{d_1 d_2}\widehat{s}_{T_i}^{(1)}}\sqrt{n}$ is the standardized signals with variance estimation with respect to T_i, $\widetilde{\mathbf{W}}_i = \mathbf{W}_i^{(1)}/\widehat{s}_{T_i}^{(1)} - \mathsf{w}_i$ is the asymptotic normal noise. Here recall that $M_{T_i} := \langle M, T_i\rangle$ and $\widehat{s}_{T_i}^{(1)} = \left\|\mathcal{P}_{\widehat{M}_{\mathsf{init}}^{(1)}}(T_i)\right\|_{\mathrm{F}}$.

Our loading matrix is $\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}$, with

$$\lambda_{\min}(\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}) = \frac{1}{\left\|\widehat{\Sigma}^{\frac{1}{2}}\widehat{D}^{-1}\right\|} = \frac{1}{\sqrt{\left\|\widehat{D}^{-1}\widehat{\Sigma}\widehat{D}^{-1}\right\|}} \geq \frac{1}{\sqrt{\left\|\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\right\| + \left\|\widehat{D}^{-1}\left(\Sigma - \widehat{\Sigma}\right)\widehat{D}^{-1}\right\|}}$$

We now denote $\rho_T = \|T\|_{\ell_1}/\|T\|_{\mathrm{F}}$. By Xia and Yuan (2021); Ma and Xia (2024), we have $\left|1 - \widehat{s}_T^{(1)}/s_T\right| \leq C_2\frac{\mu\rho_T}{\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2 d_2 \log d_1}{n}}$ with probability at least $1 - Cd_1^{-\tau}\log d_1$. Here $D :=$ $\mathrm{diag}(s_{T_1}, \cdots, s_{T_q})$. Thus, the absolute value of the diagonal matrix can be controlled by:

$$\left|D^{-1} - \widehat{D}^{-1}\right| \preceq C_2\frac{\mu\rho_T}{\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2 d_2 \log d_1}{n}}D^{-1}. \tag{54}$$

This indicates that

$$\left\|\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\right\| \leq (1+c)\left\|D^{-1}\Sigma D^{-1}\right\| \leq \frac{3}{2}\kappa_1,$$

for a small c as long as $\frac{\mu\rho_T}{\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2 d_2 \log d_1}{n}} \to 0$; and also

$$\left\|\widehat{D}^{-1}\left(\Sigma - \widehat{\Sigma}\right)\widehat{D}^{-1}\right\| \leq (1+c)\left\|D^{-1}\left(\Sigma - \widehat{\Sigma}\right)D^{-1}\right\| \leq CC_0\frac{\rho_T\mu\sigma_\xi}{\beta_0\lambda_{\min}}\sqrt{\frac{\alpha_d\kappa_1 q d_1^2 d_2 \log d_1}{n}},$$

which can be derived following the same steps as Lemma 4. It thus gives the well-conditioning of our loading matrix in LASSO:

$$\lambda_{\min}\left(\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}\right) \geq \frac{1}{\sqrt{2\kappa_1}}.$$

Following a classic argument on the LASSO precision analysis (van de Geer and Bühlmann, 2009; Bühlmann and Van De Geer, 2011), we have

$$\begin{aligned}\left\|\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}\left(\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right)\right\|^2 \leq & 2\left\langle \widehat{D}\widehat{\Sigma}^{-1}\widehat{D}\widetilde{\mathbf{W}}, \widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right\rangle + 2\lambda\left(\left\|\widehat{\mathsf{w}}\right\|_{\ell_1} - \left\|\widehat{\mathsf{w}}^{(1)}\right\|_{\ell_1}\right) \\ \leq & \lambda\left\|\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right\|_{\ell_1} + 2\lambda\left(\left\|\widehat{\mathsf{w}}\right\|_{\ell_1} - \left\|\widehat{\mathsf{w}}^{(1)}\right\|_{\ell_1}\right),\end{aligned}$$

where we define λ as the value that $\mathbb{P}\left(2\left\|\widehat{D}\widehat{\Sigma}^{-1}\widehat{D}\widetilde{\mathbf{W}}\right\|_\infty \geq \lambda\right) \leq d_1^{-2}$. It is thus clear that

$$\left\|\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}\left(\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right)\right\|^2 \leq 3\lambda\left\|\widehat{\mathsf{w}}_s^{(1)}-\widehat{\mathsf{w}}_s\right\|_{\ell_1} \leq 3\lambda\sqrt{q_1}\left\|\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right\|.$$

Here, we use the subscript s to denote the support set of w. Combined with the well-conditioning property of $\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}$, we have

$$\frac{1}{2\kappa_1}\left\|\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right\|^2 \leq \left\|\widehat{\Sigma}^{-\frac{1}{2}}\widehat{D}\left(\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right)\right\|^2 \leq 3\lambda\sqrt{q_1}\left\|\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right\|,$$

i.e., $\left\|\widehat{\mathsf{w}}^{(1)}-\widehat{\mathsf{w}}\right\| \leq 6\lambda\kappa_1\sqrt{q_1}$. Then, it amounts to determining the regularization level λ. Notice that $\widehat{D}\widetilde{\mathbf{W}} = D\widehat{\mathbf{W}}$, where $\widehat{\mathbf{W}}_i = \mathbf{W}_i^{(1)}/s_{T_i} - \frac{M_{T_i}-\theta_{T_i}}{\widehat{\sigma}_\xi s_{T_i}\sqrt{d_1 d_2}}\sqrt{n}$. Here $\widehat{\mathbf{W}}_i$ and $\widetilde{\mathbf{W}}_i$ only differ in the sampling variance s_{T_i}. We adopt the notation in the proof of Theorem 1: we define an average of i.i.d. matrix as $\widehat{Z}_1 = \frac{d_1 d_2}{n}\sum_{i\in I_2}\xi_i X_i$, and split the noise $\widehat{\mathbf{W}} = \widehat{\mathbf{W}}_1 + \widehat{\mathbf{W}}_2$, where

$$\widehat{\mathbf{W}}_{1i} = \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_i)\right\rangle}{\sigma_\xi s_{T_i}\sqrt{d_1 d_2/n}}, \text{ for each } i \in [q]. \tag{55}$$

By Theorem 1, we have $\left\|\widehat{\mathbf{W}}_2\right\|_\infty \leq Ch_n$, with probability at least $1 - Cd_1^2$. Therefore,

$$\begin{aligned}2\left\|\widehat{D}\widehat{\Sigma}^{-1}\widehat{D}\widetilde{\mathbf{W}}\right\|_\infty &= 2\left\|\widehat{D}\widehat{\Sigma}^{-1}D\widehat{\mathbf{W}}\right\|_\infty \leq 2(1+ch_n)\left\|D\widehat{\Sigma}^{-1}D\widehat{\mathbf{W}}\right\|_\infty \\ &\leq 3\left(\left\|D\Sigma^{-1}D\widehat{\mathbf{W}}\right\|_\infty + \left\|D(\widehat{\Sigma}^{-1}-\Sigma^{-1})D\widehat{\mathbf{W}}\right\|_\infty\right) \\ &\leq 3\left(\left\|D\Sigma^{-1}D\widehat{\mathbf{W}}_1\right\|_\infty + \left\|D\Sigma^{-1}D\widehat{\mathbf{W}}_2\right\|_\infty + \left\|D(\widehat{\Sigma}^{-1}-\Sigma^{-1})D\widehat{\mathbf{W}}\right\|_\infty\right).\end{aligned}$$

For any i, it is clear that

$$e_i^\top D\Sigma^{-1}D\widehat{\mathbf{W}}_1 = \frac{\left\langle \operatorname{Vec}(\widehat{Z}_1)^\top, e_i^\top D\Sigma^{-1}T_{\mathcal{H}}(I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top)\right\rangle}{\sigma_\xi\sqrt{d_1 d_2/n}},$$

with

$$\mathbb{E}\left(e_i^\top D\Sigma^{-1}D\widehat{\mathbf{W}}_1\right)^2 = e_i^\top D\Sigma^{-1}De_i \leq \kappa_1.$$

According to Bernstein inequality, we have

$$\frac{1}{\sqrt{n}}\left|\frac{e_i^\top D\Sigma^{-1}D\widehat{\mathbf{W}}_1}{\left(e_i^\top D\Sigma^{-1}De_i\right)^{\frac{1}{2}}}\right| \leq C_1\sqrt{\frac{(\log d_1)}{n}} + C_2\frac{\sqrt{rd_1}(\log d_1)}{n},$$

with probability at least $1 - q^{-1}d_1^{-\tau}$. This indicates that

$$\mathbb{P}\left(\left\|D\Sigma^{-1}D\widehat{\mathbf{W}}_1\right\|_\infty \geq C\sqrt{\kappa_1(\log d_1)}\right) \leq d_1^{-\tau}.$$

If we use $\widehat{U}$, $\widehat{V}$ to estimate Σ, then a corresponding accuracy in $\|\cdot\|_\infty$-norm is given by:

Lemma 5. *If we use $\widehat{\Sigma}$ to approximate Σ, then*

$$\left\|D(\widehat{\Sigma}^{-1} - \Sigma^{-1})D\right\|_\infty \leq CC_0\left(\kappa_\infty\sqrt{\kappa_1} + \kappa_1^{1.5}\kappa_T\left(\frac{\text{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1\right)\right)\frac{\rho_T\mu\sigma_\xi}{\beta_0\lambda_{\min}}\sqrt{\frac{\alpha_d q d_1^2 d_2\log d_1}{n}}.$$

Here $\|M\|_\infty := \max_i \|e_i^\top M\|_{\ell_1}$ and $\kappa_\infty := \|R^{-1}\|_\infty$ where $R = D^{-1}\Sigma D^{-1}$.

Notice that, since $\widehat{\mathbf{W}}_{1i}$ is standardized, Bernstein inequality also gives the bound:

$$\left\|\widehat{\mathbf{W}}_1\right\|_\infty \leq C\sqrt{\log d_1},$$

with probability at least $1 - d_1^{-\tau}$. This indicates that, with probability at least $1 - Cd_1^{-\tau}$, we have

$$2\left\|\widehat{D}\widehat{\Sigma}^{-1}\widehat{D}\widetilde{\mathbf{W}}\right\|_\infty \leq C\left(\sqrt{\kappa_1(\log d_1)} + \kappa_\infty h_n\right) \leq C\sqrt{\kappa_1(\log d_1)},$$

as long as $(\kappa_\infty h_n) \vee \left(\left(\kappa_\infty\sqrt{\kappa_1} + \kappa_1^{1.5}\kappa_T\left(\frac{\text{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1\right)\right)\frac{\rho_T\mu\sigma_\xi}{\beta_0\lambda_{\min}}\sqrt{\frac{\alpha_d q d_1^2 d_2\log d_1}{n}}\right) \leq c\sqrt{\kappa_1}$ for some small constant c. Here we use the fact $\left\|D\Sigma^{-1}D\widehat{\mathbf{W}}_2\right\|_\infty \leq C\kappa_\infty h_n$.This leads to the error bound of $\widehat{\mathsf{w}}^{(1)}$:

$$\left\|\widehat{\mathsf{w}}^{(1)} - \widehat{\mathsf{w}}\right\|_\infty \leq \left\|\widehat{\mathsf{w}}^{(1)} - \widehat{\mathsf{w}}\right\| \leq 6\lambda\kappa_1\sqrt{q_1} \leq C\kappa_1^{1.5}\sqrt{q_1(\log d_1)}.$$

Since for each i,

$$|\widehat{\mathsf{w}}_i - \mathsf{w}_i| \leq \left(\frac{C_1\tau\log d_1}{\sqrt{n}} + C_2\gamma_n^2 + C_3\mu\frac{\|T\|_{\ell_1}}{\|T\|_{\mathrm{F}}\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\tau\alpha_d d_1^2 d_2\log d_1}{n}}\right)|\mathsf{w}_i| \leq Ch_n\,|\mathsf{w}_i|\,,$$

we finish the proof. □

B.6 Proof of Proposition 3

Proof. We proceed to discuss the asymptotic normality of each $e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{y}_2$: since $\mathbf{y}_2 = \mathbf{X}\mathbf{W}^{(2)}$, with

$$\mathbf{y}_2 = \widehat{\Sigma}^{-\frac{1}{2}} D\widehat{\mathsf{w}} + \widehat{\Sigma}^{-\frac{1}{2}} D\widehat{\mathbf{W}},$$

where, with a slight abuse of notation, we define $\widehat{\mathsf{w}}_i = \frac{M_T - \theta_T}{\widehat{\sigma}_\xi \sqrt{d_1 d_2 s_T}} \sqrt{n}$ is the standardized signals with variance estimation, $\widehat{\mathbf{W}}_i = \mathbf{W}_i / s_{T_i} - \widehat{\mathsf{w}}_i$ is the asymptotic normal noise. From the proof of Theorem 1, it is clear that $\widehat{\mathsf{w}}_i$ is close enough to w_i. Notice that, here, we do not assume $\mathcal{H}_1 \subseteq \mathcal{A}$. For any $i \in \mathcal{A}$, we have

$$\begin{aligned}
e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{y}_2 &= e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \left[\mathbf{X}_\mathcal{A}, \mathbf{X}_{\mathcal{A}^c}\right] D(\widehat{\mathsf{w}} + \widehat{\mathbf{W}}) \\
&= s_{T_i}\widehat{\mathsf{w}}_i + e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} D\widehat{\mathbf{W}} + e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{X}_{\mathcal{A}^c} D_{\mathcal{A}^c} \widehat{\mathsf{w}}_{\mathcal{A}^c} \\
&= s_{T_i}\widehat{\mathsf{w}}_i + e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} D \left(\widehat{\mathbf{W}}_1 + \widehat{\mathbf{W}}_2\right) + e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{X}_{\mathcal{A}^c} D_{\mathcal{A}^c} \widehat{\mathsf{w}}_{\mathcal{A}^c},
\end{aligned}$$

where the noise decomposition $\widehat{\mathbf{W}} = \widehat{\mathbf{W}}_1 + \widehat{\mathbf{W}}_2$ is the same as (55). If $T_i \in \mathcal{A} \cap \mathcal{H}_0$, we have $\mathsf{w}_i = 0$, thus $e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{y}_2 = e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} D \left(\widehat{\mathbf{W}}_1 + \widehat{\mathbf{W}}_2\right)$. We investigate the following terms: (i) the asymptotic normality of $e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} D\widehat{\mathbf{W}}_1$, (ii) the vanishing of $e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} D\widehat{\mathbf{W}}_2$, and (iii) the bias introduced by inconsistent screening $e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{X}_{\mathcal{A}^c} D_{\mathcal{A}^c} \widehat{\mathsf{w}}_{\mathcal{A}^c}$.
(i) the asymptotic normality of $\widehat{\beta}_i := e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} D\widehat{\mathbf{W}}_1$. Conditional on $\mathcal{D}_0$ and $\mathcal{D}_1$, $\widehat{\beta}_i$ can be viewed as sum of i.i.d. independent random variables:

$$\widehat{\beta}_i = \frac{\left\langle \mathrm{Vec}(\widehat{Z}_1), e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} T_\mathcal{H} \left(I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right) \right\rangle}{\sigma_\xi \sqrt{d_1 d_2 / n}}. \tag{56}$$

The variance of $\widehat{\beta}_i$ is given by

$$\begin{aligned}
\mathbb{E}\widehat{\beta}_i^2 &= \left\| e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} T_\mathcal{H} \left(I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right) \right\|^2 \\
&= e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X}_\mathcal{A}^\top \mathbf{X} \Sigma \mathbf{X}^\top \mathbf{X}_\mathcal{A} \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} e_i = \mathbf{Q}_{ii}.
\end{aligned}$$

The third-order moment of each component is also derived by

$$
\begin{aligned}
&\mathbb{E}\left|\sqrt{d_1d_2/n}\frac{\left\langle \mathrm{Vec}(\xi_i X_i), e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}T_{\mathcal{H}}\left(I_{d_1d_2}-U_\perp U_\perp^\top\otimes V_\perp V_\perp^\top\right)\right\rangle}{\sigma_\xi \mathbf{Q}_{ii}^{\frac{1}{2}}}\right|^3\\
&\quad\le C\frac{\sqrt{d_1d_2}}{n^{1.5}}\frac{\left|\left\langle \mathrm{Vec}(X_i), e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}T_{\mathcal{H}}\left(I_{d_1d_2}-U_\perp U_\perp^\top\otimes V_\perp V_\perp^\top\right)\right\rangle\right|}{\mathbf{Q}_{ii}^{\frac{1}{2}}}\\
&\quad= C\frac{\sqrt{d_1d_2}}{n^{1.5}}\frac{\left|\left\langle \mathrm{Vec}(X_i)\left(I_{d_1d_2}-U_\perp U_\perp^\top\otimes V_\perp V_\perp^\top\right), e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}T_{\mathcal{H}}\left(I_{d_1d_2}-U_\perp U_\perp^\top\otimes V_\perp V_\perp^\top\right)\right\rangle\right|}{\mathbf{Q}_{ii}^{\frac{1}{2}}}\\
&\quad\le C\frac{\sqrt{d_1d_2}}{n^{1.5}}\left\|\mathrm{Vec}(X_i)\left(I_{d_1d_2}-U_\perp U_\perp^\top\otimes V_\perp V_\perp^\top\right)\right\|_{\mathrm{F}}\\
&\quad\le C\frac{\mu\sqrt{rd_1}}{n^{1.5}},
\end{aligned}
$$

where we use the incoherence condition in the last inequality. It is thus suggested that:

$$
\left|\mathbb{P}\left(\frac{\widehat{\beta}_i}{\sqrt{\mathbf{Q}_{ii}}}\le t\middle|\mathcal{D}_0,\mathcal{D}_1\right)-\Phi(t)\right|\le C\mu\sqrt{\frac{rd_1}{n}}. \tag{57}
$$

(ii) the vanishing of $\Delta\beta_i := e_i^\top\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}D\widehat{\mathbf{W}}_2$. By the proof of Theorem 1, we have $\left\|\widehat{\mathbf{W}}_2\right\|_\infty\le Ch_n$, with probability at least $1-Cd_1^{-\tau}\log d_1$. Thus, by writing $\mathbf{X}=[\mathbf{X}_{\mathcal{A}},\mathbf{X}_{\mathcal{A}^c}]$, we have

$$
\frac{|\Delta\beta_i|}{\sqrt{\mathbf{Q}_{ii}}}=\frac{\left|e_i^\top\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}D\widehat{\mathbf{W}}_2\right|}{\sqrt{\mathbf{Q}_{ii}}}\le\frac{\left|s_{T_i}\widehat{\mathbf{W}}_{2i}\right|+\left|e_i^\top\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}^c}D_{\mathcal{A}^c}\widehat{\mathbf{W}}_{2,\mathcal{A}^c}\right|}{\sqrt{\mathbf{Q}_{ii}}}.
$$

Using the definition of C_∞, it follows that

$$
\frac{|\Delta\beta_i|}{\sqrt{\mathbf{Q}_{ii}}}\le CC_\infty h_n,
$$

uniformly for all i with probability at least $1-C\log d_1d_1^{-\tau}$.
(iii) the bias $e_i^\top\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}^c}D_{\mathcal{A}^c}\widehat{\mathsf{w}}_{\mathcal{A}^c}$ can be surely controlled by

$$
\frac{\left|e_i^\top\left(\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}}\right)^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}^c}D_{\mathcal{A}^c}\widehat{\mathsf{w}}_{\mathcal{A}^c}\right|}{\sqrt{\mathbf{Q}_{ii}}}\le C\cdot C_\infty(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_\infty).
$$

Then, combing (i), (ii), and (iii) by the Lipschiz property of $\Phi(t)$, we have

$$
\left|\mathbb{P}\left(\frac{\widehat{\mathsf{w}}_i^{(2)}}{\sqrt{\mathbf{Q}_{ii}}}\le t\middle|\mathcal{D}_0,\mathcal{D}_1\right)-\Phi(t)\right|\le C\cdot C_\infty(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_\infty)+C\mu\sqrt{\frac{rd_1}{n}}\le C\cdot C_\infty(h_n+\|\mathsf{w}_{\mathcal{A}^c}\|_\infty).
$$

□

B.7 Proof of Theorem 6

Proof. In the following proof, we write $h_n + \|\mathsf{w}_{\mathcal{A}^c}\|_\infty$ as h_n for notational simplicity. The proof essentially follows the proof of Theorem 5. Define the expected false rejection:

$$\widetilde{G}(t) = \frac{\sum_{T_i \in \mathcal{H}_0 \cap \mathcal{A}} \mathbb{P}(\widehat{\mathsf{w}}_i^{(1)} \frac{\sqrt{\mathbf{Q}_{ii}}}{\widehat{\sigma}_{wi}} Z > t | \mathcal{D}_1)}{q_{0n}},$$

where $\widehat{\sigma}_{wi}^2 = e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}}\right)^{-1} e_i$ is defined in Algorithm 3. Denote

$$L_n' = \widetilde{G}^{-1}\left(\frac{\epsilon_n \eta_n'}{q_{0n}}\right) = \inf\left\{t : \widetilde{G}(t) \leq \frac{\epsilon_n \eta_n'}{q_{0n}}\right\},$$

where ϵ_n is a rate to be specified later, and $q_{0n} = |\mathcal{A} \cap \mathcal{H}_0|$. We can rewrite Lemma 1, 2, and 3 as:

Lemma 6. *Conditional on E_0 and $\mathcal{D}_1$, we have*

$$\sup_{0 \leq t \leq L_n} \left| \frac{\sum_{T_i \in \mathcal{H}_0 \cap \mathcal{A}} \mathbb{P}(\widehat{\mathsf{w}}_i^{\mathsf{rank}} > t)}{q_{0n} \widetilde{G}(t)} - 1 \right| \leq C_3 \frac{C_\infty h_n q_{0n}}{\epsilon_n \eta_n'}.$$

Here we use $\widehat{\mathsf{w}}_i^{\mathsf{rank}}$ to indicate the combined statistics $\widehat{\mathsf{w}}_{T_i}^{\mathsf{rank}}$

Lemma 7 (Weak dependency of null features)**.** *Conditional on $\mathcal{D}_1$,*

$$\sup_{0 \leq t \leq L_n'} \frac{\sum_{(T_i, T_j) \in \mathcal{H}_{0\mathcal{A},weak}^2} \left|\mathrm{cov}(\mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{rank}} > t), \mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{rank}} > t))\right|}{q_{0n}^2 \widetilde{G}^2(t)} \leq C_1 \frac{C_\infty h_n q_{0n}}{\epsilon_n \eta_n'} + C_2 \frac{1}{(\epsilon_n \eta_n' q_{0n})^{v/2}}. \tag{58}$$

Lemma 8. *For any $\varepsilon > 0$, conditional on $\mathcal{D}_1$, it holds that*

$$\mathbb{P}\left(\sup_{0 \leq t \leq L_n'} \left| \frac{\sum_{T_i \in \mathcal{H}_0 \cap \mathcal{A}} \mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{rank}} > t))}{q_{0n} \widetilde{G}(t)} - 1 \right| \geq \varepsilon\right)$$
$$\leq \frac{C}{\varepsilon^2} \log(\frac{q_{0n}}{\epsilon_n \eta_n'}) \left(\left(\frac{\beta_{\mathsf{s}}' q_{0n}^2}{\epsilon_n^2 \eta_n'^2}\right)^{\frac{1}{2}} + \left(\frac{C_\infty h_n q_{0n}}{\epsilon_n \eta_n'} + \frac{1}{(\epsilon_n \eta_n' q_{0n})^{v/2}}\right)^{\frac{1}{2}}\right).$$

The proof of Lemma 6, 7, and 8 is same as that in Lemma 1, 2, and 3, and thus omitted. These lemmas imply that, if $L \leq L_n'$, then we have $\mathsf{R}_0 \leq 1 + 3\varepsilon$ with probability at least

$$1 - \frac{C}{\varepsilon^2} \log(\frac{q_{0n}}{\epsilon_n \eta_n'}) \left(\left(\frac{\beta_{\mathsf{s}}' q_{0n}^2}{\epsilon_n^2 \eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{C_\infty h_n q_{0n}}{\epsilon_n \eta_n'} + \frac{1}{(\epsilon_n \eta_n' q_{0n})^{v/2}}\right)^{\frac{1}{2}}\right).$$

We then prove the probability of $\mathbb{P}(L \leq L_n')$ can be very large. A matching upper bound of $\widetilde{G}(t)$ is given by, similarly as in the proof Theorem 5,

$$\widetilde{G}(t) = \frac{\sum_{T_i \in \mathcal{H}_0 \cap \mathcal{A}} \mathbb{P}(\widehat{\mathsf{w}}_i^{(1)} \frac{\sqrt{\mathbf{Q}_{ii}}}{\widehat{\sigma}_{wi}} Z > t | \mathcal{D}_0, \mathcal{D}_1)}{q_{0n}} \leq \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left(-\frac{t^2}{2 \max_{T \in \mathcal{H}_0 \cap \mathcal{A}} \left|\widehat{\mathsf{w}}_i^{(1)} \frac{\sqrt{\mathbf{Q}_{ii}}}{\widehat{\sigma}_{wi}}\right|^2}\right).$$

The LASSO results presented in Proposition 2 show that, the $\left|\widehat{\mathsf{w}}_i^{(1)}\right|$ can be uniformly bounded by:

$$\begin{aligned}
\max_{T_i\in\mathcal{H}_0\cap\mathcal{A}}\left|\widehat{\mathsf{w}}_i^{(1)}\frac{\sqrt{\mathbf{Q}_{ii}}}{\widehat{\sigma}_{wi}}\right| &\leq \max_{T\in\mathcal{H}_0\cap\mathcal{A}}\left|\widehat{\mathsf{w}}_i^{(1)}\right|\frac{\sqrt{e_i^\top\left(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A}\right)^{-1}\mathbf{X}_\mathcal{A}^\top\mathbf{X}\Sigma\mathbf{X}^\top\mathbf{X}_\mathcal{A}\left(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A}\right)^{-1}e_i}}{\sqrt{e_i^\top\left(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A}\right)^{-1}e_i}}\\
&\leq \max_{T\in\mathcal{H}_0\cap\mathcal{A}}\left|\widehat{\mathsf{w}}_i^{(1)}\right|\left(1+\frac{\left\|e_i^\top\left(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A}\right)^{-1}\mathbf{X}_\mathcal{A}^\top\right\|\sqrt{\|\mathbf{X}\Sigma\mathbf{X}^\top-I\|}}{\sqrt{e_i^\top\left(\mathbf{X}_\mathcal{A}^\top\mathbf{X}_\mathcal{A}\right)^{-1}e_i}}\right)\\
&\leq (1+c)\max_{T\in\mathcal{H}_0\cap\mathcal{A}}\left|\widehat{\mathsf{w}}_i^{(1)}\right|\\
&\leq C\kappa_1^{1.5}\sqrt{q_1(\log d_1)},
\end{aligned}$$

with probability at least $1-Cd_1^{-\tau}$. Here we use the fact that $\left\|\mathbf{X}\Sigma\mathbf{X}^\top-I\right\|\leq\frac{1}{1-c}$ if we have its inverse $\left\|\Sigma^{-\frac{1}{2}}\widehat{\Sigma}\Sigma^{-\frac{1}{2}}-I\right\|\leq c$. The definition of L_n' implies that

$$L_n'\leq C\sqrt{\log(\frac{q_{0n}}{\epsilon_n\eta_n'})}\cdot C\kappa_1^{1.5}\sqrt{q_1(\log d_1)}\ll\sqrt{\log(\frac{1}{h_n})}\cdot\kappa_1^{1.5}\sqrt{q_1(\log d_1)}.$$

If $T_i\in\mathcal{S}$, then $|\delta_{T_i}|\geq C_{\mathsf{gap}}\sqrt{\log\frac{1}{h_n}}\vee\kappa_1^{1.5}\sqrt{q_1(\log d_1)}$ by the definition of $\mathcal{S}$, and also the LASSO estimation:

$$\left|\widehat{\mathsf{w}}_i^{(1)}\right|\geq C\kappa_1^{1.5}\sqrt{q_1(\log d_1)},$$

by our assumption. Assume C_{gap} is large enough, and $\delta_{T_i}>0$. Then we have

$$\begin{aligned}
\mathbb{P}(\widehat{\mathsf{w}}_i^{\mathsf{rank}}<L_n')\leq&\mathbb{P}\left(\widehat{\mathsf{w}}_i^{(1)}(Z_2+\delta_{T_i}\frac{s_{T_i}}{\sqrt{\mathbf{Q}_{ii}}})<L_n'\right)+C_\infty h_n\\
\leq&1-\mathbb{P}\left((Z_2+\delta_{T_i}\frac{s_{T_i}}{\sqrt{\mathbf{Q}_{ii}}})\geq L_n'/\widehat{\mathsf{w}}_i^{(1)}\right)+C_\infty h_n\\
\leq&1-\mathbb{P}\left(Z_2\geq-\delta_{T_i}+\sqrt{\log\frac{1}{h_n}}\right)+C_\infty h_n\\
\leq&\mathbb{P}\left(Z_2\leq-2\sqrt{\log\frac{1}{h_n}}\right)+C_\infty h_n\\
\leq&2C_\infty h_n.
\end{aligned}$$

We compute the probability:

$$\begin{aligned}
&\mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{rank}}>L_n')\leq(1-\varepsilon)\eta_n')=\mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{rank}}<L_n)>\varepsilon\eta_n')\\
\leq&\frac{\sum_{T\in\mathcal{S}}\mathbb{P}(\widehat{\mathsf{w}}_i^{\mathsf{rank}}<L_n')}{\varepsilon\eta_n'}\leq CC_\infty h_n/\varepsilon,
\end{aligned}$$

i.e., $\mathbb{P}(\sum_{T\in\mathcal{S}} \mathbb{I}(\widehat{\mathsf{w}}_T^{\mathsf{rank}} > L_n') \leq (1-\varepsilon)\eta_n') \to 0$, $\mathbb{P}(\sum_{T\in\mathcal{S}} \mathbb{I}(\widehat{\mathsf{w}}_T^{\mathsf{rank}} > L_n') \geq \eta_n') \to 1$. By taking $\epsilon_n = \alpha/8$, other steps essentially follow the proof of Theorem 5.

□

C Additional Technical Lemmas

We now gives an error bound on the generalized SVD for any mode-m low-rank tensor, which is instantly applicable to the low-rank matrix with $m = 2$.

Lemma 9 (Perturbation of general HOSVD)**.** *Given a mode-m Tucker low-rank tensor with $\mathcal{M} = \mathcal{S} \times_{j=1}^m U_j \in \mathbb{R}^{d_1\times\cdots\times d_j}$, where $\mathcal{S} \in \mathbb{R}^{r_1\times r_2\cdots\times r_m}$, and $U_j \in \mathbb{R}^{d_j\times r_j}$ are incoherent singular subspaces, i.e., $U_j^\top U_j = I_{d_j}$, $\|U_j\|_{2,\max} \leq \sqrt{\mu r_j/d_j}$. We denote its maximum and minimum singular values as $\lambda_{\max} = \max_{j\in[m]}\{\lambda_1(\mathrm{Mat}_j(\mathcal{M}))\}$, $\lambda_{\min} = \min_{j\in[m]}\{\lambda_{r_j}(\mathrm{Mat}_j(\mathcal{M}))\}$, with condition number $\kappa_0 = \lambda_{\max}/\lambda_{\min}$. Here, $\mathrm{Mat}_j(\cdot)$ means the mode-j unfolding. Denote $d^* = \prod_{j=1}^m d_j$, $r^* = \prod_{j=1}^m r_j$, $\bar{d} = \max\{d_j\}$. Write the tangent space of the $r_1 \times r_2 \cdots \times r_m$ low-rank manifold at point $\mathcal{M}$ as $\mathbb{T}$, with the projection onto it as $\mathcal{P}_{\mathbb{T}}(\cdot)$. For any perturbation $\mathcal{E}$, we denote its higher-order SVD as*

$$\widehat{\mathcal{M}} = \mathrm{HOSVD}(\mathcal{M}+\mathcal{E}, r_1\times r_2\cdots\times r_m) = (\mathcal{M}+\mathcal{E})\times_{j=1}^m \mathcal{P}_{\widehat{U}_j},\ \textit{where}\ \widehat{U}_j = \mathrm{SVD}_{r_j}(\mathrm{Mat}_j(\mathcal{M}+\mathcal{E})).$$

Then, when

$$\|\mathcal{E}\|_{\ell_\infty} = \varepsilon_\infty \leq \frac{\lambda_{\min}}{48\kappa_0 m\sqrt{d^*}},$$

for any tensor $\mathcal{I} \in \mathbb{R}^{d_1\times\cdots\times d_j}$, we have

$$\left|\left\langle \widehat{\mathcal{M}}, \mathcal{I}\right\rangle - \langle \mathcal{M} + \mathcal{P}_{\mathbb{T}}(\mathcal{E}), \mathcal{I}\rangle\right| \leq \frac{37e^2m^2d^*\varepsilon_\infty^2}{\lambda_{\min}}\sqrt{\frac{\mu^m r^*}{d^*}}\,\|\mathcal{I}\|_{\ell_1}$$

For more introduction on the Tucker low-rank tensor and the related definitions/notations, please refer to Ma and Xia (2024).

Proof. The proof relies on the spectral decomposition that has been studied in Ma and Xia (2024). Define the mode-j unfolding of $\mathcal{M}$, $\mathcal{E}$ as M_j, E_j, correspondingly. Without loss of generality, we write each unfolding of $\mathcal{S}$ as $S_j = \Lambda_j V_j^\top$. Define $\|\mathcal{E}\|_{\mathrm{F}} = \varepsilon_{\mathrm{F}}$. According to the proof of Lemma 1 in Ma and Xia (2024), for any mode j, we have

$$\mathcal{P}_{\widehat{U}_j} - \mathcal{P}_{U_j} = \sum_{k\geq 1}\mathcal{S}_j^{(k)} = \sum_{k\geq 1}\sum_{\mathbf{s}:s_1+\cdots+s_{k+1}=k}(-1)^{1+\tau(\mathbf{s})}\cdot\mathfrak{P}_j^{-s_1}\Delta_j\mathfrak{P}_j^{-s_2}\Delta_j\cdots\Delta_j\mathfrak{P}_1^{-s_{k+1}},$$

where $s_1 \geq 0, \cdots, s_{k+1} \geq 0$ are non-negative integers with $\tau(\mathbf{s}) = \sum_{j=1}^{k+1} \mathbb{I}\left(s_j > 0\right)$, and

$$\Delta_j = (M_j + E_j)(M_j + E_j)^\top - \mathcal{M}_j \mathcal{M}_j^\top = \mathcal{M}_j E_j^\top + E_j \mathcal{M}_j^\top + E_j E_j^\top,$$

and $\mathfrak{P}_j$ are the power series whose the definition can be found in Xia (2021); Ma and Xia (2024). For the series $\mathcal{S}_j^{(k)}$, its first order is

$$\begin{aligned}
\mathcal{S}_j^{(1)} &= U_j \Lambda_j^{-2} U_j^\top \Delta_j \mathcal{P}_{U_j}^\perp + \mathcal{P}_{U_j}^\perp \Delta_j U_j \Lambda_j^{-2} U_j^\top \\
&= \underbrace{U_j \Lambda_j^{-1} V_j^\top \left(\otimes_{k \neq j} \mathcal{P}_{U_k}\right) E_j^\top \mathcal{P}_{U_j}^\perp + \mathcal{P}_{U_j}^\perp E_j \left(\otimes_{k \neq j} \mathcal{P}_{U_k}\right) V_j \Lambda_j^{-1} U_j^\top}_{:= \mathfrak{A}_j} \\
&\quad + \underbrace{U_j \Lambda_j^{-2} U_j^\top (E_j E_j^\top) \mathcal{P}_{U_j}^\perp + \mathcal{P}_{U_j}^\perp (E_j E_j^\top) U_j \Lambda_j^{-2} U_j^\top}_{:= \mathfrak{B}_j}
\end{aligned}$$

From Lemma 1 of Ma and Xia (2024), we can extract that:

$$\left\| \mathcal{P}_{\widehat{U}_j} - \mathcal{P}_{U_j} \right\|_{2,\infty} \leq \frac{8\sqrt{d^*}\varepsilon_\infty}{\lambda_{\min}} \sqrt{\frac{\mu r_j}{d_j}} \leq \frac{1}{2}\sqrt{\frac{\mu r_j}{d_j}}, \tag{59}$$

and for any e_j as the canonical basis in $\mathbb{R}^{d_j}$,

$$\begin{aligned}
\left\| e_j^\top \mathfrak{B}_j \right\|_2 \leq 4\frac{\sqrt{d^*}\varepsilon_\infty}{\lambda_{\min}} \sqrt{\frac{\mu r_j}{d_j}} \left(\frac{\varepsilon_{\mathrm{F}}}{\lambda_{\min}}\right), \ \left\| e_j^\top \mathfrak{B}_j M_j \right\|_2 \leq 4\sqrt{d^*}\varepsilon_\infty \sqrt{\frac{\mu r_j}{d_j}} \left(\frac{\varepsilon_{\mathrm{F}}}{\lambda_{\min}}\right), \\
\left\| e_j^\top \sum_{k \geq 2} \mathcal{S}_j^{(k)} \right\|_2 \leq \sqrt{\frac{\mu r_j}{d_j}} \cdot \frac{8\sqrt{d^*}\varepsilon_\infty}{\lambda_{\min}} \left(\frac{8\varepsilon_{\mathrm{F}}}{\lambda_{\min}}\right)^{k-1}.
\end{aligned} \tag{60}$$

We then have

$$\begin{aligned}
\widehat{\mathcal{M}} - \mathcal{M} &= (\mathcal{M} + \mathcal{E}) \times_{j=1}^m \mathcal{P}_{\widehat{U}_j} - \mathcal{M} \times_{j=1}^m \mathcal{P}_{U_j} \\
&= \mathcal{E} \times_{j=1}^m \mathcal{P}_{U_j} + \sum_{k=1}^m \mathcal{M} \times_k \left(\mathcal{P}_{\widehat{U}_k} - \mathcal{P}_{U_k}\right) \times_{j \neq k} \mathcal{P}_{U_j} \\
&\quad \underbrace{+ \sum_{\mathbb{Q} \subseteq [m], |\mathbb{Q}| \geq 1} \mathcal{E} \times_{k \in \mathbb{Q}} \left(\mathcal{P}_{\widehat{U}_k} - \mathcal{P}_{U_k}\right) \times_{j \notin \mathbb{Q}} \mathcal{P}_{U_j} + \sum_{\mathbb{Q} \subseteq [m], |\mathbb{Q}| \geq 2} \mathcal{M} \times_{k \in \mathbb{Q}} \left(\mathcal{P}_{\widehat{U}_k} - \mathcal{P}_{U_k}\right) \times_{j \notin \mathbb{Q}} \mathcal{P}_{U_j}}_{:= \mathfrak{C}_1} \\
&= \mathcal{E} \times_{j=1}^m \mathcal{P}_{U_j} + \sum_{k=1}^m \mathcal{M} \times_k \mathfrak{A}_k \times_{j \neq k} \mathcal{P}_{U_j} + \sum_{k=1}^m \mathcal{M} \times_k \mathfrak{B}_k \times_{j \neq k} \mathcal{P}_{U_j} + \mathfrak{C}_1 \\
&= \mathcal{P}_{\mathbb{T}}(\mathcal{E}) + \sum_{k=1}^m \mathcal{M} \times_k \mathfrak{B}_k \times_{j \neq k} \mathcal{P}_{U_j} + \mathfrak{C}_1.
\end{aligned} \tag{61}$$

Therefore, for any $\mathcal{I}$, we have

$$\left| \left\langle \widehat{\mathcal{M}}, \mathcal{I} \right\rangle - \left\langle \mathcal{M} + \mathcal{P}_{\mathbb{T}}(\mathcal{E}), \mathcal{I} \right\rangle \right| = \left| \left\langle \sum_{k=1}^m \mathcal{M} \times_k \mathfrak{B}_k \times_{j \neq k} \mathcal{P}_{U_j} + \mathfrak{C}_1, \mathcal{I} \right\rangle \right|.$$

According to (59), (60), given any single entry $\mathcal{W}$, we can control the term above by:

$$\left|\left\langle \sum_{k=1}^{m} \mathcal{M} \times_k \mathfrak{B}_k \times_{j\neq k} \mathcal{P}_{U_j}, \mathcal{W} \right\rangle\right| \leq 4m\sqrt{d^*}\varepsilon_\infty \sqrt{\frac{\mu^m r^*}{d^*}} \left(\frac{\varepsilon_{\mathrm{F}}}{\lambda_{\min}}\right), \tag{62}$$

and

$$\begin{aligned} |\langle \mathfrak{C}_1, \mathcal{W}\rangle| &\leq \sum_{k\geq 1} (em)^k \left(\frac{8\sqrt{d^*}\varepsilon_\infty}{\lambda_{\min}}\right)^k \sqrt{\frac{\mu^m r^*}{d^*}} \varepsilon_{\mathrm{F}} + \sum_{k\geq 2} \left(\frac{em}{2}\right)^k \left(\frac{8\sqrt{d^*}\varepsilon_\infty}{\lambda_{\min}}\right)^{k-1} \sqrt{\frac{\mu^m r^*}{d^*}} 8\sqrt{d^*}\varepsilon_\infty \\ &\leq \left(\frac{16emd^*\varepsilon_\infty^2}{\lambda_{\min}} + \frac{32e^2m^2d^*\varepsilon_\infty^2}{\lambda_{\min}}\right)\sqrt{\frac{\mu^m r^*}{d^*}} \leq \frac{36e^2m^2d^*\varepsilon_\infty^2}{\lambda_{\min}} \sqrt{\frac{\mu^m r^*}{d^*}}. \end{aligned} \tag{63}$$

Combining (61), (62), (63), we know that for $\widehat{\mathcal{M}}$,

$$\left|\left\langle \widehat{\mathcal{M}}, \mathcal{I}\right\rangle - \langle \mathcal{M} + \mathcal{P}_{\mathbb{T}}(\mathcal{E}), \mathcal{I}\rangle\right| \leq \frac{37e^2m^2d^*\varepsilon_\infty^2}{\lambda_{\min}} \sqrt{\frac{\mu^m r^*}{d^*}} \left\|\mathcal{I}\right\|_{\ell_1}.$$

□

D Proof of Minimax CI Length

Theorem 10 (Minimax optimal CI length for tensor completion)**.** *Consider the tensor completion model:*

$$Y_i = \langle \mathcal{X}_i, \mathcal{M}\rangle + \xi_i, \quad i \in [n],$$

where $\xi_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d., and $\mathcal{X}_i$ are independent and uniformly distributed over all the canonical bases in $\mathbb{R}^{d_1\times\cdots d_j}$, which are independent of $\{\xi_i\}_{i=1}^n$. We use the notation in Lemma 9, with $\underline{r} = \min\{r_j\}$ Denote $s_0 = \|\mathcal{P}_{\mathbb{T}}(\mathcal{I})\|_{\mathrm{F}}$. Define the parameter space as

$$\begin{aligned} \boldsymbol{\Theta} =& \big\{\mathcal{M} \in \mathbb{R}^{d_1\times\cdots d_j} : \mathrm{rank}(\mathrm{Mat}_j(\mathcal{M})) \leq r_j, \|U_j\|_{2,\max} \leq \sqrt{\mu r_j / d_j},. \\ & \lambda_{\min}(\mathcal{M}) \geq \lambda_{\min}, \kappa(\mathcal{M}) \leq \kappa_0, \left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}} \geq s_0\big\}. \end{aligned}$$

Here $\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\cdot)$ means the projection onto the tangent space at any given $\mathcal{M}$. Consider the set of any valid $1-\alpha$ confidence interval with $\alpha < \frac{1}{4}$ as:

$$\mathcal{I}_\alpha(\boldsymbol{\Theta}, \mathcal{I}) := \left\{\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M}, \{(\mathcal{X}_i, Y_i)\}_{i=1}^n\right) = [l, u] : \inf_{\mathcal{M}\in\boldsymbol{\Theta}} \mathbb{P}(l \leq \langle \mathcal{M}, \mathcal{I}\rangle \leq u) \geq 1-\alpha\right\},$$

where l, u are any functions of observations $\{(\mathcal{X}_i, Y_i)\}_{i=1}^n$. Then, when the SNR satisfies

$$\frac{\lambda_{\min}}{\sigma} \geq C_{\mathsf{gap}}\kappa_0 \left(\frac{\|\mathcal{I}\|_{\ell_1}}{\|\mathcal{P}_{\mathbb{T}}(\mathcal{I})\|_{\mathrm{F}}\sqrt{d^*/\bar{d}}} \bigvee 1\right) \sqrt{\frac{m^5(2\mu)^{3m}(r^*)^3\bar{d}d^*}{\underline{r}^2 n}}$$

for a numeric constant C_{gap}, the length of the confidence interval has the minimax lower bound:

$$\inf_{\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M},\{(\mathcal{X}_i,Y_i)\}_{i=1}^n\right)\in\mathcal{I}_\alpha(\boldsymbol{\Theta},\mathcal{I})}\ \sup_{\mathcal{M}\in\boldsymbol{\Theta}} \mathbb{E}L\left(\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M},\{(\mathcal{X}_i,Y_i)\}_{i=1}^n\right)\right) \ge c\sigma\sqrt{\frac{d^*}{n}}s_0 = c\sigma\sqrt{\frac{d^*}{n}}\left\|\mathcal{P}_{\mathbb{T}}(\mathcal{I})\right\|_{\mathrm{F}}$$

Proof. Invoking the Lemma 1 of Cai and Guo (2017), we have

$$\begin{aligned}
&\inf_{\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M},\{(\mathcal{X}_i,Y_i)\}_{i=1}^n\right)\in\mathcal{I}_\alpha(\boldsymbol{\Theta},\mathcal{I})}\ \sup_{\mathcal{M}\in\boldsymbol{\Theta}} \mathbb{E}L\left(\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M},\{(\mathcal{X}_i,Y_i)\}_{i=1}^n\right)\right)\\
&\ge \inf_{\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M},\{(\mathcal{X}_i,Y_i)\}_{i=1}^n\right)\in\mathcal{I}_\alpha(\boldsymbol{\Theta},\mathcal{I})}\ \sup_{\mathcal{M}\in\{\mathcal{M}_1,\mathcal{M}_2\}} \mathbb{E}L\left(\mathrm{CI}_{\mathcal{I}}^{\alpha}\left(\mathcal{M},\{(\mathcal{X}_i,Y_i)\}_{i=1}^n\right)\right)\\
&\ge \left|\langle\mathcal{M}_1-\mathcal{M}_2,\mathcal{I}\rangle\right|\left(1-2\alpha-\mathrm{TV}(\pi(\mathcal{M}_1),\pi(\mathcal{M}_2))\right)\\
&\ge \left|\langle\mathcal{M}_1-\mathcal{M}_2,\mathcal{I}\rangle\right|\left(1-2\alpha-\sqrt{2\,\mathrm{KL}(\pi(\mathcal{M}_1),\pi(\mathcal{M}_2))}\right)
\end{aligned}$$

where we use Pinsker's inequality for the last step. Now, with a slightly abuse of notation, we choose a $\mathcal{M}\in\boldsymbol{\Theta}$ such that $\lambda_{\min}(\mathcal{M}) = 2\lambda_{\min}$, $\kappa(\mathcal{M}) = \frac{1}{2}\kappa_0$, $\|U_j\|_{2,\max} \le \sqrt{\mu r_j/d_j}/2$, and $\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}} \ge 2s_0$. Let a new $\widehat{\mathcal{M}}$ be

$$\widehat{\mathcal{M}} = \mathrm{HOSVD}\left(\mathcal{M} + \varepsilon\frac{\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}}, r_1,\dots,r_m\right).$$

For each entry $\mathcal{W}$, we have

$$\varepsilon\left|\left\langle\frac{\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}},\mathcal{W}\right\rangle\right| = \varepsilon\left|\left\langle\frac{\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}},\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{W})\right\rangle\right| \le \varepsilon\sqrt{\frac{mr^*\mu^{m-1}\bar{d}}{\underline{r}d^*}},$$

i.e.,

$$\left\|\varepsilon\frac{\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}}\right\|_{\ell_\infty} \le \varepsilon\sqrt{\frac{mr^*\mu^{m-1}\bar{d}}{\underline{r}d^*}}.$$

Now, we set

$$\varepsilon\sqrt{\frac{mr^*\mu^{m-1}\bar{d}}{\underline{r}d^*}} \le \frac{\lambda_{\min}}{48\kappa_0 m\sqrt{d^*}},\ \text{i.e., } \varepsilon \le \frac{\lambda_{\min}\sqrt{\underline{r}}}{48\kappa_0 m^{1.5}\sqrt{r^*\bar{d}}}, \tag{64}$$

and apply (59), Lemma 9. This tells us that $\widehat{\mathcal{M}}$ is also with rank $r_1,\cdots,r_m$, and μ-incoherence. Moreover, from Lemma 1 of Li et al. (2023), we know that

$$\left\|\widehat{\mathcal{M}}-\mathcal{M}\right\|_{\mathrm{F}} \le \varepsilon + \frac{59m\varepsilon^2}{\lambda_{\min}} \le 2\varepsilon \le \frac{1}{2}\lambda_{\min}.$$

Also, we need to check the variance condition. By Lemma 14 of Ma and Xia (2024), we have

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathbb{T}(\widehat{\mathcal{M}})}(\mathcal{I})\right\|_{\mathrm{F}} \ge \left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\mathbb{T}(\widehat{\mathcal{M}})}(\mathcal{I}) - \mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}\\
&\ge 2s_0 - \frac{Cm^2\kappa_0}{\lambda_{\min}}\sqrt{\frac{(2\mu)^m r^*\bar{d}}{n}}\cdot\varepsilon\sqrt{\frac{mr^*\mu^{m-1}\bar{d}}{\underline{r}d^*}}\left\|\mathcal{I}\right\|_{\ell_1}\\
&\ge s_0,
\end{aligned}$$

where we set

$$\frac{Cm^2\kappa_0}{\lambda_{\min}}\sqrt{\mu^m r^* \bar{d}} \cdot \varepsilon\sqrt{\frac{mr^*\mu^{m-1}\bar{d}}{\underline{r}d^*}}\left\|\mathcal{I}\right\|_{\ell_1} \leq s_0, \text{ i.e., } \varepsilon \leq c\frac{s_0}{\left\|\mathcal{I}\right\|_{\ell_1}}\lambda_{\min}\frac{\sqrt{\underline{r}d^*}}{m^{2.5}\kappa_0\mu^m r^*\bar{d}}. \tag{65}$$

Therefore, $\widehat{\mathcal{M}} \in \boldsymbol{\Theta}$. Set $\mathcal{M}_1 = \mathcal{M}$, $\mathcal{M}_2 = \widehat{\mathcal{M}}$. Now, we compute $|\langle\mathcal{M}_1 - \mathcal{M}_2, \mathcal{I}\rangle|$ and $\mathrm{KL}(\pi(\mathcal{M}_1), \pi(\mathcal{M}_2))$. For the KL-divergence, clearly we have the following chain rule:

$$\begin{aligned}\mathrm{KL}(\pi(\mathcal{M}_1), \pi(\mathcal{M}_2)) &= \mathbb{E}_{\mathcal{X}}[\mathrm{KL}(\pi(\mathcal{M}_1)|\{\mathcal{X}_i\}_{i=1}^n, \pi(\mathcal{M}_2)|\{\mathcal{X}_i\}_{i=1}^n)] + \mathrm{KL}(\{\mathcal{X}_i\}_{i=1}^n, \{\mathcal{X}_i\}_{i=1}^n)\\ &= \mathbb{E}_{\mathcal{X}}\frac{\sum_{i=1}^n\left(\left\langle\widehat{\mathcal{M}} - \mathcal{M}, \mathcal{X}_i\right\rangle\right)^2}{2\sigma^2} = \frac{n\left\|\widehat{\mathcal{M}} - \mathcal{M}\right\|_{\mathrm{F}}^2}{2\sigma^2 d^*}.\end{aligned}$$

Thus, $\sqrt{2\,\mathrm{KL}(\pi(\mathcal{M}_1), \pi(\mathcal{M}_2))} \leq 2\sqrt{n/d^*}/\sigma \cdot \varepsilon$. We select

$$\varepsilon = \frac{\sigma\sqrt{d^*/n}}{8},$$

which gives $\left(1 - 2\alpha - \sqrt{2\,\mathrm{KL}(\pi(\mathcal{M}_1), \pi(\mathcal{M}_2))}\right) \geq \frac{1}{4}$.

For the term $|\langle\mathcal{M}_1 - \mathcal{M}_2, \mathcal{I}\rangle|$, we use Lemma 9, which gives

$$\begin{aligned}\left|\left\langle\widehat{\mathcal{M}}, \mathcal{I}\right\rangle - \langle\mathcal{M}, \mathcal{I}\rangle\right| &\geq \left|\left\langle\mathcal{P}_{\mathbb{T}(\mathcal{M})}\left(\varepsilon\frac{\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}}\right), \mathcal{I}\right\rangle\right| - \left|\left\langle\widehat{\mathcal{M}}, \mathcal{I}\right\rangle - \left\langle\mathcal{M} + \mathcal{P}_{\mathbb{T}(\mathcal{M})}\left(\varepsilon\frac{\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}}\right), \mathcal{I}\right\rangle\right.\\ &\geq \varepsilon\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}} - \frac{37e^2m^2d^*\varepsilon_\infty^2}{2\lambda_{\min}}\sqrt{\frac{\mu^m r^*}{2^m d^*}}\left\|\mathcal{I}\right\|_{\ell_1}\\ &\geq \varepsilon\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}} - \varepsilon^2 37e^2\frac{m^3r^*\mu^{m-1}\bar{d}}{2^m\lambda_{\min}\underline{r}}\sqrt{\frac{\mu^m r^*}{2^m d^*}}\left\|\mathcal{I}\right\|_{\ell_1}.\end{aligned}$$

Set

$$\frac{2\cdot 37e^2m^3r^*\mu^{1.5m-1}\sqrt{r^*}\bar{d}}{2^{1.5m}\underline{r}\lambda_{\min}\sqrt{d^*}}\frac{\left\|\mathcal{I}\right\|_{\ell_1}}{\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}}} \leq \frac{74e^2(r^*)^{1.5}\mu^{1.5m-1}\bar{d}}{\underline{r}\lambda_{\min}\sqrt{d^*}}\frac{\left\|\mathcal{I}\right\|_{\ell_1}}{2s_0} \leq \frac{1}{\varepsilon},$$

i.e.,

$$\varepsilon \leq c\frac{\sqrt{d^*}s_0}{\bar{d}\left\|\mathcal{I}\right\|_{\ell_1}}\frac{\underline{r}\lambda_{\min}}{(r^*)^{1.5}\mu^{1.5m-1}}, \tag{66}$$

we have

$$\left|\left\langle\widehat{\mathcal{M}}, \mathcal{I}\right\rangle - \langle\mathcal{M}, \mathcal{I}\rangle\right| \geq \frac{1}{2}\varepsilon\left\|\mathcal{P}_{\mathbb{T}(\mathcal{M})}(\mathcal{I})\right\|_{\mathrm{F}} \geq \varepsilon s_0.$$

Combining (64), (65), (66), and $\varepsilon = \frac{\sigma\sqrt{d^*/n}}{8}$, we know that when

$$\frac{\lambda_{\min}}{\sigma} \geq C_{\mathsf{gap}}\kappa_0\left(\frac{\left\|\mathcal{I}\right\|_{\ell_1}}{\left\|\mathcal{P}_{\mathbb{T}}(\mathcal{I})\right\|_{\mathrm{F}}\sqrt{d^*/\bar{d}}}\bigvee 1\right)\sqrt{\frac{m^5(2\mu)^{3m}(r^*)^3\bar{d}d^*}{\underline{r}^2 n}},$$

the minimax lower bound is given by

$$|\langle\mathcal{M}_1 - \mathcal{M}_2, \mathcal{I}\rangle|\left(1 - 2\alpha - \sqrt{2\,\mathrm{KL}(\pi(\mathcal{M}_1), \pi(\mathcal{M}_2))}\right) \geq \frac{1}{4}\frac{\sigma\sqrt{d^*/n}}{8}s_0 = \frac{\sigma\left\|\mathcal{P}_{\mathbb{T}}(\mathcal{I})\right\|_{\mathrm{F}}}{32}\sqrt{\frac{d^*}{n}}.$$

This finishes the proof. □

E Proofs of Auxiliary Results

E.1 Verification of 14

It has been shown in the proof of Theorem 1 that the test statistics W_T can be decomposed as

$$W_T = \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}} + \Delta_T,$$

where Δ_T is a vanishing term with the rate of convergence described in (38). Suppose also the distribution of ξ is symmetric. Denote I_1 the index set of observations in sample $\mathcal{D}_1$. Therefore, for any integer $k \geq 2$, we have

$$\begin{aligned}
\mathbb{E}\left|W_T\right|^{2k} &\gtrsim \mathbb{E}\left| \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}} \right|^{2k} = \mathbb{E}\left| \frac{\sqrt{d_1 d_2 / n} \sum_{i \in I_1} \xi_i \left\langle X_i, \mathcal{P}_M(T) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}} \right|^{2k} \\
&\geq \left(\mathbb{E}\left| \frac{\sqrt{d_1 d_2 / n} \sum_{i \in I_1} \xi_i \left\langle X_i, \mathcal{P}_M(T) \right\rangle}{\sigma_\xi \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}} \right|^4 \right)^{k/2} \\
&\geq \left(\frac{d_1^2 d_2^2 \left(\sum_{i \in I_1} \mathbb{E}\xi_i^4 \left\langle X_i, \mathcal{P}_M(T) \right\rangle^4 + \sum_{i,j \in I_1, i \neq j} \mathbb{E}\xi_i^2 \left\langle X_i, \mathcal{P}_M(T) \right\rangle^2 \xi_j^2 \left\langle X_j, \mathcal{P}_M(T) \right\rangle^2 \right)}{n^2 \sigma_\xi^4 \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}^4} \right)^{k/2} \\
&\gtrsim \left(\frac{d_1^2 d_2^2 \mathbb{E} \left\langle X_i, \mathcal{P}_M(T) \right\rangle^4}{n \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}^4} + 1 \right)^{k/2} \\
&= \left(\frac{d_1 d_2 \sum_{i \in [d_1], j \in [d_2]} \mathcal{P}_M(T)_{i,j}^4}{n \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}^4} + 1 \right)^{k/2}.
\end{aligned}$$

If the energy of $\mathcal{P}_M(T)$ is concentrated in a few entries, e.g., there exists an index set J such that the entries in J can dominate other entries, i.e.,

$$\sum_{(i,j) \in J} \mathcal{P}_M(T)_{i,j}^2 \geq \sum_{(i,j) \notin J} \mathcal{P}_M(T)_{i,j}^2,$$

with $s_0 := |J| = O(1)$, then we have

$$\frac{\sum_{i \in [d_1], j \in [d_2]} \mathcal{P}_M(T)_{i,j}^4}{\left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}^4} \geq \frac{\sum_{(i,j) \in J} \mathcal{P}_M(T)_{i,j}^4}{4 \left(\sum_{(i,j) \in J} \mathcal{P}_M(T)_{i,j}^2 \right)^2} \geq \frac{1}{4 s_0} \geq \Omega(1),$$

and thus, we have

$$\sqrt[2k]{\mathbb{E}\left|W_T\right|^{2k}} \gtrsim \left(\frac{d_1 d_2}{n} \right)^{1/4}.$$

E.2 Proof of Theorem 9

Proof. Define the c.d.f. of the product of two standard normal random variables as $\Psi(t)$, also $\tilde{\Psi}(t) := 1 - \Psi(t)$. The c.d.f. of standard normal distribution is denoted by $\Phi(t)$ by convention, with $\tilde{\Phi}(t) := 1 - \Phi(t)$. For $j = 1$, we have

$$\mathbb{P}(Y_1 > t|H_0) = \Psi(-t) = \tilde{\Psi}(t)$$

$$F_1(z,t) := \mathbb{P}(Y_1 > t|z, H_1) = \mathbb{P}\left(\frac{(\xi_1+\xi_2+2\delta)^2}{4} - \frac{(\xi_1-\xi_2)^2}{4} > t\right) = \mathbb{P}\left(\frac{(Z_1+\sqrt{2}\delta)^2}{2} - \frac{Z_2^2}{2} > t\right)$$
$$= \int_{\mathbb{R}} \left[\tilde{\Phi}(\sqrt{2t+y_2^2} - \sqrt{\frac{2}{p}}z) + \tilde{\Phi}(\sqrt{2t+y_2^2} + \sqrt{\frac{2}{p}}z)\right] dy_2.$$

Here ξ_1, ξ_2, and Z_1, Z_2 are all standard normal random variables. Thus $L_{p1} = \tilde{\Psi}^{-1}(p)$. Calculate the first order and second order derivative of $F_1(z,t)$ with respect to z when $t = L_{p1}$:

$$\partial_z F_1(0, L_{p1}) = 0$$
$$\partial_z^2 F_1(0, L_{p1}) = \frac{8}{q}\int_0^{+\infty} -f'(\sqrt{2L_{p1}+y_2^2})dy_2 = \frac{8}{p}f(-\sqrt{2L_{p1}}).$$

Since $\tilde{\Psi}(t) < \sqrt{2}\tilde{\Phi}(\sqrt{2t})$, we have $L_{p1} < \frac{1}{2}\tilde{\Phi}^{-1}(p/\sqrt{2})^2$. When $x \to 0$, we have

$$\sqrt{2(\log(\frac{1-r_1}{x}) - \frac{1}{2}\log\log(\frac{1-r_1}{x}))} \le \tilde{\Phi}^{-1}(x) \le \sqrt{2(\log(\frac{1}{x}) - \frac{1}{2+r_2}\log\log(\frac{1}{x}))}.$$

for any small $r_1, r_2 > 0$. Thus we have $L_{p1} < \frac{1}{2}\tilde{\Phi}^{-1}(p/\sqrt{2})^2 \le \log(\frac{\sqrt{2}}{p}) - \frac{1}{2+r_2}\log\log(\frac{\sqrt{2}}{p})$, and the second order derivative

$$\partial_z^2 F_1(0, L_{p1}) = \frac{8}{p}f(-\sqrt{2L_{p1}}) \ge c(\log(\frac{\sqrt{2}}{p}))^{1/(2+r_2)},$$

is non-vanishing.

For $j = 2$, we have

$$\mathbb{P}(Y_2 > t|H_0) = \mathbb{P}(\xi_1 > t, \xi_2 > t) + \mathbb{P}(\xi_1 < -t, \xi_2 < -t) = 2\tilde{\Phi}^2(t)$$
$$F_2(z,t) := \mathbb{P}(Y_2 > t|z, H_1) = \mathbb{P}(\xi_1 + \mu > t, \xi_2 + \mu > t) + \mathbb{P}(\xi_1 + \mu < -t, \xi_2 + \mu < -t)$$
$$= \tilde{\Phi}^2(t + \sqrt{\frac{1}{p}}z) + \tilde{\Phi}^2(t - \sqrt{\frac{1}{p}}z).$$

In this case, the threshold $L_{p2} = \tilde{\Phi}^{-1}(\sqrt{\frac{p}{2}})$. Compute the derivatives of F_2:

$$\partial_z F_2(0, L_{p2}) = 0$$
$$\partial_z^2 F_2(0, L_{p2}) = \frac{4}{p}(f^2(-L_{p2}) + \tilde{\Phi}(L_{p2})f'(-L_{p2})) \ge c((\log(\sqrt{\frac{2}{p}}))^{1/(2+r_2)} + 1),$$

which also has a non-vanishing second-order derivative.

For $j = 3$, we have

$$\begin{aligned}
\mathbb{P}(Y_3 > t|H_0) &= \mathbb{P}(\xi_1+\xi_2 > t, \xi_1 > 0, \xi_2 > 0) + \mathbb{P}(\xi_1+\xi_2 < -t, \xi_1 < 0, \xi_2 < 0)\\
&= 2\mathbb{P}(Y_1 > \frac{t}{\sqrt{2}}, -Y_1 < Y_2 < Y_1) = 2\tilde{\Phi}(\frac{t}{\sqrt{2}})(1-\tilde{\Phi}(\frac{t}{\sqrt{2}}))\\
\mathbb{P}(Y_3 > t|z, H_1) &= \mathbb{P}(Z_1 > \frac{t-2\mu}{\sqrt{2}}, -Z_1-\sqrt{2}\mu < Z_2 < Z_1+\sqrt{2}\mu)\\
&+ \mathbb{P}(Z_1 < \frac{-t-2\mu}{\sqrt{2}}, Z_1+\sqrt{2}\mu < Z_2 < -Z_1-\sqrt{2}\mu)\\
&\le \tilde{\Phi}(\frac{t}{\sqrt{2}})\left(\phi(\frac{t}{\sqrt{2}}+\sqrt{2}\mu)+\phi(\frac{t}{\sqrt{2}}-\sqrt{2}\mu)\right)\\
F_3(z,t) :&= \tilde{\Phi}(\frac{t}{\sqrt{2}})\left(\phi(\frac{t}{\sqrt{2}}+\sqrt{\frac{2}{p}}z)+\phi(\frac{t}{\sqrt{2}}-\sqrt{\frac{2}{p}}z)\right).
\end{aligned}$$

Compute the derivatives of F_3:

$$\begin{aligned}
\partial_z F_3(0, L_{p3}) &= 0\\
\partial_z^2 F_3(0, L_{p3}) &= \frac{4}{p}\tilde{\Phi}(\frac{L_{p3}}{\sqrt{2}})f'(\frac{L_{p3}}{\sqrt{2}}) \le 0.
\end{aligned}$$

If $\delta_0 = o(\sqrt{\frac{1}{\pi}})$, we have $z = \sqrt{p}\delta \to 0$. By Taylor's theorem, we have

$$\text{Power}_{W_j}(L_{pj}) = p + \mathbb{E}_{\mathbf{\Theta}}\partial_z F_j(0, L_{pj})z + \mathbb{E}_{\mathbf{\Theta}}\frac{1}{2}\partial_z^2 F_j(0, L_{pj})z^2 + o(\mathbb{E}_{\mathbf{\Theta}} z^2),$$

(or $\le$ for $j = 3$). Plugging in the derivatives of $j = 1, 2, 3$, clearly we have $\text{Power}_{W_1}(L_{p1}) \ge \text{Power}_{W_3}(L_{p3})$, and $\text{Power}_{W_2}(L_{p2}) \ge \text{Power}_{W_3}(L_{p3})$; for the second order derivative of F_1 and F_2, we also have

$$\begin{aligned}
\partial_z^2 F_1(0, L_{p1}) - \partial_z^2 F_2(0, L_{p2}) &= \frac{4}{p}(2f(-\sqrt{2L_{p1}}) - f^2(-L_{p2}) - \tilde{\Phi}(L_{p2})f'(-L_{p2}))\\
&\ge c\frac{1}{p}\exp\left(-\frac{1}{2}\tilde{\Phi}^{-1}(p/\sqrt{2})^2\right)\left(1-\exp\left(\frac{1}{2}\tilde{\Phi}^{-1}(p/\sqrt{2})^2 - \tilde{\Phi}^{-1}(\sqrt{\frac{p}{2}})^2\right)\right).
\end{aligned}$$

Since

$$\begin{aligned}
&\frac{1}{2}\tilde{\Phi}^{-1}(p/\sqrt{2})^2 - \tilde{\Phi}^{-1}(\sqrt{\frac{p}{2}})^2 = (\frac{1}{\sqrt{2}}\tilde{\Phi}^{-1}(p/\sqrt{2}) + \tilde{\Phi}^{-1}(\sqrt{\frac{p}{2}}))(\frac{1}{\sqrt{2}}\tilde{\Phi}^{-1}(p/\sqrt{2}) - \tilde{\Phi}^{-1}(\sqrt{\frac{p}{2}}))\\
&\le (\frac{1}{\sqrt{2}}\tilde{\Phi}^{-1}(p/\sqrt{2}) + \tilde{\Phi}^{-1}(\sqrt{\frac{p}{2}}))\\
&\cdot (\sqrt{\log(\frac{\sqrt{2}}{p}) - \frac{1}{1+r_2}\log\log(\frac{\sqrt{2}}{p})} - \sqrt{\log(\frac{2(1-r_1)^2}{p}) - \log\log((1-r_1)\sqrt{\frac{2}{p}})})\\
&\to -\infty,
\end{aligned}$$

we have $\partial_z^2 F_1(0, L_{p1}) - \partial_z^2 F_2(0, L_{p2}) \geq 0$, thus $\text{Power}_{W_1}(L_{p1}) \geq \text{Power}_{W_2}(L_{p2})$. Translating the $\text{Power}_{W_j}(L_{pj})$ to $\text{Power}_{W_j}(L_{\alpha j})$, we finish our proof. □

E.3 Proof of Lemma 4

Proof. For simplicity, we omit C_0 in the proof. Notice that both $\left(I_{d_1 d_2} - \widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top\right)$ and $\left(I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)$ are projection matrices with $P = P^2$. We thus have

$$\begin{aligned}\widehat{\Sigma} - \Sigma &= T_{\mathcal{H}} \left((I_{d_1 d_2} - \widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top) - (I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top) \right) T_{\mathcal{H}}^\top \\ &= T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \\ &+ T_{\mathcal{H}} \left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \\ &+ T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top .\end{aligned} \tag{67}$$

We apply (67) to the error $\Sigma^{-\frac{1}{2}}(\widehat{\Sigma} - \Sigma)\Sigma^{-\frac{1}{2}}$:

$$\begin{aligned}&\left\| \Sigma^{-\frac{1}{2}}(\widehat{\Sigma} - \Sigma)\Sigma^{-\frac{1}{2}} \right\| \leq 2 \left\| \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \Sigma^{-\frac{1}{2}} \right\| \\ &+ \left\| \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \Sigma^{-\frac{1}{2}} \right\| .\end{aligned}$$

Notice that $\left\| \left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \Sigma^{-\frac{1}{2}} \right\| \leq 1$. We only need to focus on the term

$$\begin{aligned}&\left\| \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \right\| \\ &\leq \left\| \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \right\| \left\| \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \right\| \\ &\leq \sqrt{\kappa_1} \kappa_T \left(\left\| \left(\widehat{U}_\perp \widehat{U}_\perp^\top - U_\perp U_\perp^\top \right) \otimes V_\perp V_\perp^\top \right\| + \left\| U_\perp U_\perp^\top \otimes \left(\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top \right) \right\| \right. \\ &+ \left. \left\| \left(\widehat{U}_\perp \widehat{U}_\perp^\top - U_\perp U_\perp^\top \right) \otimes \left(\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top \right) \right\| \right) \\ &\leq C_2 \sqrt{\kappa_1} \kappa_T \frac{\sqrt{\tau}\,(1 + \gamma_n)\,\sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}},\end{aligned}$$

where we use the definition of κ_T and the perturbation of singular subspaces in Xia and Yuan (2021). Moreover, when $T_{\mathcal{H}}$ is sparse, we use $e_{T,k} \in \mathbb{R}^{d_1 \times d_2}$, $k \in [\text{supp}(T_{\mathcal{H}})]$ to indicate the collective supports of all the $\mathrm{vec}(T_i)$. We then have

$$\begin{aligned}
&\left\|\Sigma^{-\frac{1}{2}} T_{\mathcal{H}}\left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\| \\
&= \left\|\Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \sum_{k=1}^{\mathrm{supp}(T_{\mathcal{H}})} e_{T,k} e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\| \\
&\leq \sqrt{\kappa_1}\kappa_T \,\mathrm{supp}(T_{\mathcal{H}}) \max_k \left\| e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\| \\
&\leq \sqrt{\kappa_1}\kappa_T \,\mathrm{supp}(T_{\mathcal{H}}) \max_k \left(\left\| e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top - U_\perp U_\perp^\top\right) \otimes V_\perp V_\perp^\top\right\|\right. \\
&\left. + \left\| e_{T,k}^\top U_\perp U_\perp^\top \otimes \left(\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top\right)\right\| + \left\| e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top - U_\perp U_\perp^\top\right) \otimes \left(\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top\right)\right\|\right).
\end{aligned} \tag{68}$$

Since each $e_{T,k}$ can also be represented as $e_{T,k} = e_{T,k}^1 \otimes e_{T,k}^2$, where $e_{T,k}^1 \in \mathbb{R}^{d_1}$ and $e_{T,k}^2 \in \mathbb{R}^{d_2}$ are also canonical bases, we then have

$$\begin{aligned}
&\left\|\Sigma^{-\frac{1}{2}} T_{\mathcal{H}}\left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\| \\
&\lesssim \sqrt{\kappa_1}\kappa_T \,\mathrm{supp}(T_{\mathcal{H}}) \left(\left\|UU^\top - \widehat{U}\widehat{U}^\top\right\|_{2,\max} + \left\|VV^\top - \widehat{V}\widehat{V}^\top\right\|_{2,\max}\right) \\
&\leq C\sqrt{\kappa_1}\kappa_T \frac{\mathrm{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \frac{\sqrt{\tau}\left(1+\gamma_n\right)\sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}},
\end{aligned}$$

because the higher-order error can be dominated. The rate γ_n converges to 0, which means that the whole error can be controlled by:

$$\left\|\Sigma^{-\frac{1}{2}}(\widehat{\Sigma} - \Sigma)\Sigma^{-\frac{1}{2}}\right\| \leq C\frac{\kappa_T \sigma_\xi}{\lambda_{\min}} \cdot \left(\frac{\mathrm{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1\right)\sqrt{\frac{\kappa_1 d_1^2 d_2 \log d_1}{n}}.$$

□

E.4 Proof of Lemma 5

Proof. For simplicity, we omit C_0 in the proof. Denote $E = \Sigma - \widehat{\Sigma}$. By Fréchet derivative, as long as $\|E\| = \left\|\Sigma - \widehat{\Sigma}\right\|$ is small for any operator norm, $\widehat{\Sigma}^{-1} - \Sigma^{-1}$ can be dominated by its Fréchet derivative $\Sigma^{-1}E\Sigma^{-1}$. Therefore, We have

$$\left\|D(\widehat{\Sigma}^{-1} - \Sigma^{-1})D\right\|_\infty \leq \left\|D\Sigma^{-1}E\Sigma^{-1}D\right\|_\infty + o(\|E\|_\infty).$$

We only need to study the convergence rate of $\|D\Sigma^{-1}E\Sigma^{-1}D\|_\infty$ as E is small. This term, however, can be decomposed following (67), i.e.,

$$\begin{aligned}
&\left\|D\Sigma^{-1}E\Sigma^{-1}D\right\|_\infty \\
&\leq \left\|D\Sigma^{-1}D\right\|_\infty \left\|D^{-1}T_{\mathcal{H}}\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)T_{\mathcal{H}}^\top \Sigma^{-1}D\right\|_\infty \\
&+ \left\|D\Sigma^{-1}T_{\mathcal{H}}\left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)T_{\mathcal{H}}^\top\Sigma^{-1}D\right\|_\infty \\
&+ \left\|D\Sigma^{-1}T_{\mathcal{H}}\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)T_{\mathcal{H}}^\top\Sigma^{-1}D\right\|_\infty \\
&\leq \kappa_\infty\sqrt{q}\left\|D^{-1}T_{\mathcal{H}}\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\|_{2,\max}\sqrt{\kappa_1} \\
&+ \sqrt{\kappa_1}\sqrt{q}\left\|\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)T_{\mathcal{H}}^\top\Sigma^{-1}D\right\| \\
&+ q\kappa_1^2\left\|D^{-1}T_{\mathcal{H}}\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\|_{2,\max}^2 \\
&\leq C\left(\kappa_\infty\sqrt{\kappa_1} + \|T_{\mathcal{H}}\|_2\,\kappa_1/\sqrt{\lambda_{\min}(\Sigma)}\right)\frac{\rho_T\mu\sigma_\xi}{\beta_0\lambda_{\min}}\sqrt{\frac{\alpha_d q d_1^2 d_2\log d_1}{n}} \\
&\leq C\left(\kappa_\infty\sqrt{\kappa_1} + \kappa_1^{1.5}\kappa_T\right)\frac{\rho_T\mu\sigma_\xi}{\beta_0\lambda_{\min}}\sqrt{\frac{\alpha_d q d_1^2 d_2\log d_1}{n}},
\end{aligned}$$

where the 2-max norm here can be bounded by:

$$\begin{aligned}
&\left\|D^{-1}T_{\mathcal{H}}\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)\right\|_{2,\max} \\
&\leq \max_{T_i\in\mathcal{H}}\frac{\sum_{j\in[d_1]}\sum_{k\in[d_2]}\left|T_i(j,k)\left[\left(UU^\top - \widehat{U}\widehat{U}^\top\right)\otimes V_\perp V_\perp^\top + U_\perp U_\perp^\top \otimes \left(\widehat{V}\widehat{V}^\top - VV^\top\right)\right]\cdot e_j\otimes e_k\right|}{s_{T_i}} \\
&\quad + \max_{T_i\in\mathcal{H}}\frac{\sum_{j\in[d_1]}\sum_{k\in[d_2]}\left|T_i(j,k)\left(UU^\top - \widehat{U}\widehat{U}^\top\right)\otimes\left(\widehat{V}\widehat{V}^\top - VV^\top\right)\cdot e_j\otimes e_k\right|}{s_{T_i}} \\
&\leq C\max_{T_i\in\mathcal{H}}\frac{\|T\|_{\ell_1}}{\|T\|_{\mathrm{F}}\,\beta_0\sqrt{r/d_1}}\frac{\mu\left(1+\gamma_n\right)\sigma_\xi}{\lambda_{\min}}\cdot\sqrt{\frac{rd_1^2\log d_1}{n}} \\
&\leq C\frac{\rho_T\mu\sigma_\xi}{\beta_0\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2 d_2\log d_1}{n}}.
\end{aligned} \tag{69}$$

Here, we use the 2-max norm bound in Xia and Yuan (2021), the alignment assumption, and the definition of κ_T. Moreover, the norm $\left\|\left(\widehat{U}_\perp\widehat{U}_\perp^\top \otimes \widehat{V}_\perp\widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)T_{\mathcal{H}}^\top\Sigma^{-1}D\right\|$ can also bounded by

$$
\begin{aligned}
&\left\| D\Sigma^{-1} T_{\mathcal{H}} \left(\widehat{U}_{\perp}\widehat{U}_{\perp}^{\top} \otimes \widehat{V}_{\perp}\widehat{V}_{\perp}^{\top} - U_{\perp}U_{\perp}^{\top} \otimes V_{\perp}V_{\perp}^{\top} \right) \left(I - U_{\perp}U_{\perp}^{\top} \otimes V_{\perp}V_{\perp}^{\top} \right) \right\| \\
&\leq \left\| D\Sigma^{-\frac{1}{2}} \cdot \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \left(\widehat{U}_{\perp}\widehat{U}_{\perp}^{\top} \otimes \widehat{V}_{\perp}\widehat{V}_{\perp}^{\top} - U_{\perp}U_{\perp}^{\top} \otimes V_{\perp}V_{\perp}^{\top} \right) \right\| \\
&\lesssim \kappa_1 \kappa_T \operatorname{supp}(T_{\mathcal{H}}) \left(\left\| UU^{\top} - \widehat{U}\widehat{U}^{\top} \right\|_{2,\max} + \left\| VV^{\top} - \widehat{V}\widehat{V}^{\top} \right\|_{2,\max} \right) \\
&\leq C\kappa_1 \kappa_T \frac{\operatorname{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \frac{\sqrt{\tau}\,(1+\gamma_n)\,\sigma_{\xi}}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}},
\end{aligned}
$$

where we use the sparsity of $T_{\mathcal{H}}$ following (68). This gives the desired bound

$$
\left\| D\Sigma^{-1} E \Sigma^{-1} D \right\|_{\infty} \leq C \left(\kappa_{\infty}\sqrt{\kappa_1} + \kappa_1^{1.5}\kappa_T \left(\frac{\operatorname{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1 \right) \right) \frac{\rho_T \mu \sigma_{\xi}}{\beta_0 \lambda_{\min}} \sqrt{\frac{\alpha_d q d_1^2 d_2 \log d_1}{n}}.
$$

□