Multiple Testing of Linear Forms for Noisy Matrix Completion

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

[1,3]Department of Mathematics, Hong Kong University of Science and Technology

[2]Department of Management Sciences, City University of Hong Kong

[4]Department of Statistics, Columbia University

(December 1, 2023)

Abstract

Many important tasks of large-scale recommender systems can be naturally cast as testing multiple linear forms for noisy matrix completion. These problems, however, present unique challenges because of the subtle bias-and-variance tradeoff of and an 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.

[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 et al. (2011); Rohde et al. (2011); Tsybakov et al. (2011); Negahban and Wainwright (2012); Sun and Zhang (2012); Klopp et al. (2014); Cai et al. (2015, 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 $\mathfrak{E} = \{e_i e_j^\top : 1 \le i \le d_1, 1 \le j \le d_2\}$ 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 centered 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 number of observations is much smaller than $d_1 d_2$. 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 $H_{0,ij}$, $1 \le i \le d_1$ and $1 \le j \le d_2$. 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 so-called 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 $\|R\|_{\max} = \max_{i,j}|R_{ij}|$ and $\|R\|_\infty := \max_{i\in[q]}\|e_i^\top R\|_{\ell_1}$ for a matrix R. Note that $\|\cdot\|_{\max}$ and $\|\cdot\|_\infty$ are equivalent for a vector.

2 Individual Tests

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 $\mathcal{D} := \{(X_i, Y_i) : 1 \le i \le n\}$ 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:

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

where r is the rank of M, $\|\cdot\|_{2,\max}$ denotes the maximum row-wise ℓ_2-norm, 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, respectively, of M, and κ_0 the ratio between the two, i.e., its condition number.

For brevity, we consider two-sided tests here but our discussion can be applied to one-sided tests straightforwardly. We shall also assume, without loss of generality, that $d_1 \geq d_2$, in what follows. Our goal of this section is to develop a test statistic for (2) that is readily applicable for testing a large number of hypotheses. The problem of testing a single linear form (2) has been previously investigated by Xia and Yuan (2021). See also Chen et al. (2019); Farias et al. (2022); Chen et al. (2023), among others, for treatment of the special case when $T = e_i e_j^\top$. The tests proposed in these earlier works however cannot be directly used for multiple testing. For example, the test statistics from Xia and Yuan (2021), and similarly others, converge to normal distribution at a rate no faster than $\sqrt{\log d_1/d_2}$. This is too slow for our purpose because it puts an unnecessary limit on the number of hypotheses we can test, regardless of how large the sample size (n) is.

We start by estimating $\langle M, T\rangle$. A general approach consists of three steps: initialization, bias-correction and low rank projection. More specifically, assume that, without loss of generality, n is an even number with $n = 2n_0$. We split $\mathcal{D}$ into two sub-samples:

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

Assume that an initial estimating procedure is available so that there exists an initial estimate $\widehat{M}_1^{\mathsf{init}}$ (or $\widehat{M}_2^{\mathsf{init}}$) from $\mathcal{D}_1$ (or $\mathcal{D}_2$) such that for any $\tau \geq 1$,

$$\left\|\widehat{M}_1^{\mathsf{init}} - M\right\|_{\max} \leq C\sigma_\xi \mu\kappa_0 \sqrt{\frac{\tau r^2 d_1 \log^2 d_1}{n}}, \tag{4}$$

with probability at least $1 - d_1^{-\tau}$, for some constant $C > 0$. Here $\|\cdot\|_{\max}$ represents the maximum entry-wise magnitude. This requirement for initialization is fairly weak and satisfied, in particular, by several recently developed matrix completion techniques, including those from Wei et al. (2016); Ma et al. (2018); Chen et al. (2020); Xia and Yuan (2021); Cai et al. (2022b) among others. For brevity, in what follows, we shall assume τ is large enough to ensure that $n = O(d_1^{2\tau})$.

To correct the bias of initial estimates, we then define

$$\widehat{M}_1^{\mathsf{unbs}} = \widehat{M}_1^{\mathsf{init}} + \frac{d_1 d_2}{n} \sum_{i=n_0+1}^{n} \left(Y_i - \left\langle \widehat{M}_1^{\mathsf{init}}, X_i\right\rangle\right) X_i$$

and similarly

$$\widehat{M}_2^{\mathsf{unbs}} = \widehat{M}_2^{\mathsf{init}} + \frac{d_1 d_2}{n} \sum_{i=1}^{n_0} \left(Y_i - \left\langle \widehat{M}_2^{\mathsf{init}}, X_i \right\rangle \right) X_i$$

Unfortunately this debiasing may lead to a significant increase in variance, we shall again trade off between bias and variance by low-rank projection, yielding an estimate

$$\widehat{M} = \frac{1}{2}[\mathcal{P}_r(\widehat{M}_1^{\mathsf{unbs}}) + \mathcal{P}_r(\widehat{M}_2^{\mathsf{unbs}})],$$

where $\mathcal{P}_r(\cdot)$ is the best rank-r approximation of a matrix, i.e., the projection of a matrix to row and column spaces spanned by its first r singular vectors. Finally we shall estimate $\langle M, T \rangle$ by $\langle \widehat{M}, T \rangle$. Inferences about $\langle M, T \rangle$ can naturally be made by studying the distribution of $\langle \widehat{M}, T \rangle$. Under certain regularity conditions, one can show that

$$\frac{\langle \widehat{M}, T \rangle - \langle M, T \rangle}{\sigma_\xi (\|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2)^{1/2} \sqrt{d_1 d_2 / n}} \to_d N(0,1), \tag{5}$$

as $n, d_1, d_2 \to \infty$, where $\|\cdot\|_{\mathrm{F}}$ denotes Frobenius norm. See, e.g., Chen et al. (2019); Xia and Yuan (2021); Cai et al. (2022a). We can use this result to test (2) for a fixed T. But if we want to test for a family of hypothesis $\{H_{0T} : T \in \mathcal{H}\}$, then more refined bounds on the convergence rates of (5) are needed.

Our key insight is that the slow convergence rates obtained in earlier works can be attributed to the fact that the variance of $\langle \widehat{M}, T \rangle$ in (5) is not sufficiently precise. More specifically, (5) uses the following variance approximation:

$$\frac{n}{d_1 d_2} \cdot \mathtt{var}(\langle \widehat{M}, T \rangle) \approx \sigma_\xi^2 (\|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2). \tag{6}$$

While this is a good first order approximation, one can derive an improved approximation. Specifically, we shall show that

$$\frac{n}{d_1 d_2} \cdot \mathtt{var}(\langle \widehat{M}, T \rangle) \approx \sigma_\xi^2 \|\mathcal{P}_M(T)\|_{\mathrm{F}}^2, \tag{7}$$

where

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

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. Note that

$$\|\mathcal{P}_M(T)\|_{\mathrm{F}}^2 = \|U^\top T\|_{\mathrm{F}}^2 + \|TV\|_{\mathrm{F}}^2 - \|U^\top TV\|_{\mathrm{F}}^2$$

so that the difference between the two variance approximations (6) and (7) is the term $\sigma_\xi^2\|U^\top TV\|_{\rm F}^2$. It is instructive to consider the special case of estimating one entry of M, i.e., $T=e_ie_j^\top$. Denote $\|\cdot\|$ the ℓ_2-norm of a vector. Then the difference becomes $\sigma_\xi^2\|e_i^\top U\|^2\|e_j^\top V\|^2$ which is of smaller order than the approximation (6): $\sigma_\xi^2(\|e_i^\top U\|^2+\|e_j^\top V\|^2)$, in light of the incoherence condition (3). Indeed, (5) immediately yields:

$$\frac{\langle\widehat{M},T\rangle-\langle M,T\rangle}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\rm F}\sqrt{d_1d_2/n}}\to_d N(0,1).$$

However, the enhanced variance approximation can significantly improve the rate of convergence as the following theorem shows.

Theorem 1. *Suppose that the sample size $n\geq C_1\mu^2rd_1\log d_1$, and*

$$\lambda_{\min}\geq C_2\mu\sigma_\xi\kappa_0^2\sqrt{\frac{rd_1^3\log^2 d_1}{n}},$$

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

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\langle\widehat{M},T\rangle-\langle M,T\rangle}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\rm F}\cdot\sqrt{d_1d_2/n}}\leq t\right)-\Phi(t)\right|$$
$$\leq C_3\left(\frac{\kappa_0\mu^2\|T\|_{\ell_1}}{\|\mathcal{P}_M(T)\|_{\rm F}}\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{r^2d_1^2\log^2 d_1}{n}}+\mu\kappa_0\sqrt{\frac{r^2d_1\log^3 d_1}{n}}\right).$$

Of course, to use the asymptotic normality established above for testing (2), we need to estimate the variance. An intuitive choice is to estimate σ_ξ^2 by

$$\widehat{\sigma}_\xi^2=\frac{1}{2n_0}\sum_{i=n_0+1}^{n}\big(Y_i-\langle\widehat{M}_1^{\rm init},X_i\rangle\big)^2+\frac{1}{2n_0}\sum_{i=1}^{n_0}\big(Y_i-\langle\widehat{M}_2^{\rm init},X_i\rangle\big)^2,$$

and $\mathcal{P}_M(T)$ by $\mathcal{P}_{\widehat{M}}(T)$. We shall therefore consider the following test statistic:

$$W_T:=\frac{\langle\widehat{M},T\rangle-\theta_T}{\widehat{\sigma}_\xi\|\mathcal{P}_{\widehat{M}}(T)\|_{\rm F}\cdot\sqrt{d_1d_2/n}}.$$

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

Theorem 2. *Under the assumptions of Theorem 1, if H_{0T} holds, then*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(W_T\leq t\right)-\Phi\left(t\right)\right|\leq C_3\left(\frac{\kappa_0\mu^2\|T\|_{\ell_1}}{\|\mathcal{P}_M(T)\|_{\rm F}}\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{r^2d_1^2\log^2 d_1}{n}}+\mu\kappa_0\sqrt{\frac{r^2d_1\log^3 d_1}{n}}\right).$$

3 Multiple Tests

We now turn our attention to testing a family of hypothesis $\{H_{0T} : T \in \mathcal{H}\}$ for a subset $\mathcal{H} \subset \mathbb{R}^{d_1 \times d_2}$. In particular, we can take $\mathcal{H} = \{e_i e_j^\top : 1 \le i \le d_1, 1 \le j \le d_2\}$ 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. In addition, we shall also assume that there exists a constant $\beta_0 > 0$ such that for all $T \in \mathcal{H}$,

$$\|\mathcal{P}_M(T)\|_{\mathrm{F}} \ge \beta_0 \|T\|_{\mathrm{F}} \sqrt{\frac{r}{d_1}}. \tag{8}$$

Recall that from the previous section, $\|\mathcal{P}_M(T)\|_{\mathrm{F}}$ is proportional to the (asymptotic) variance of the test statistic with respect to a linear form T. When $\|\mathcal{P}_M(T)\|_{\mathrm{F}} = 0$, 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., Xia and Yuan (2021). Write

$$h_n := \kappa_0 \mu^2 \sup_{T \in \mathcal{H}} \left\{\frac{\|T\|_{\ell_1}}{\|T\|_{\mathrm{F}}}\right\} \frac{\sigma_\xi}{\lambda_{\min}\beta_0} \sqrt{\frac{r d_1^3 \log^2 d_1}{n}} + \mu\kappa_0 \sqrt{\frac{r^2 d_1 \log^3 d_1}{n}}, \tag{9}$$

where, for brevity, we omit the dependence of h_n on d_1. In light of Theorem 2, with appropriate initial estimates, we have

$$|\mathbb{P}(W_T \le t) - \Phi(t)| \lesssim h_n$$

for all $T \in \mathcal{H}$.

3.1 Symmetric Data Aggregation

With the asymptotic normality of W_T, it is 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 $N(0,1)$ 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 ($k \ge 2$) of W_T for a properly chosen linear form T is lower bounded by

$$\sqrt[2k]{\mathbb{E}\,|W_T|^{2k}} \gtrsim \left(\frac{d_1 d_2}{n}\right)^{1/4}. \tag{10}$$

If $d_1 \asymp d_2 \asymp d$, and $n \asymp d^{1+\epsilon}$ for some $\epsilon \in (0,1)$, then we have $\mathbb{E}\,|W_T|^{2k} \gtrsim d^{(1-\epsilon)k/2}$. See supplement for proof of (10).

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). More specifically, we first construct two groups of (conditionally) independent asymptotic symmetric statistics $\{W_T^{(1)} : T \in \mathcal{H}\}$ and $\{W_T^{(2)} : T \in \mathcal{H}\}$ 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:

$$L := \inf\left\{t > 0 : \frac{\#\left\{T : W_T^{\mathsf{Rank}} < -t\right\}}{\#\left\{T : W_T^{\mathsf{Rank}} > t\right\} \vee 1} \leq \alpha\right\}, \tag{11}$$

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 $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, rank r, FDR level α.

1: Use $\mathcal{D}_0$ to construct an initial estimate $\widehat{M}_{\mathsf{init}}$

2: Apply bias-correction and low-rank projection using the second part of data $\mathcal{D}_1$ and the third part of data $\mathcal{D}_2$, respectively, and obtain two groups of test statistics:

$$W_T^{(1)} := \frac{\langle \widehat{M}^{(1)}, T\rangle - \theta_T}{\widehat{\sigma}_\xi^{(1)} \|\mathcal{P}_{\widehat{M}^{(1)}} T\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}}, \quad W_T^{(2)} := \frac{\langle \widehat{M}^{(2)}, T\rangle - \theta_T}{\widehat{\sigma}_\xi^{(2)} \|\mathcal{P}_{\widehat{M}^{(2)}}(T)\|_{\mathrm{F}} \sqrt{d_1 d_2 / n}}, \quad T \in \mathcal{H}$$

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 (11).

4: Reject H_{0T} if $W_T^{\mathsf{Rank}} > L$.

Here we split the data such that $|\mathcal{D}_0| \asymp |\mathcal{D}_1| = |\mathcal{D}_2| = n$ in general. Note that $\widehat{M}^{(1)}, \widehat{\sigma}_\xi^{(1)}$ are computed from $\widehat{M}_{\mathsf{init}}$ and $\mathcal{D}_1$; while $\widehat{M}^{(2)}, \widehat{\sigma}_\xi^{(2)}$ are computed from $\widehat{M}_{\mathsf{init}}$ and $\mathcal{D}_2$. Clearly, conditional on $\mathcal{D}_0$, $W_T^{(1)}$ and $W_T^{(2)}$ are independent. By the definition of L, we have

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

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}. To do so we need to first investigate the dependence among multiple test statistics.

We remark that, in addition to multiplying the two tests statistics $W_T^{\mathsf{Rank}} = W_T^{(1)} W_T^{(2)}$, other ways of aggregating the two test statistics are also possible. See, e.g., Dai et al. (2022). Notable examples including $\min\{W_T^{(1)}, W_T^{(2)}\}$ and $W_T^{(1)} + W_T^{(2)}$ that have 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

$$\texttt{corr}(\langle \widehat{M}, T_1 \rangle, \langle \widehat{M}, T_2 \rangle) \approx \frac{\langle \mathcal{P}_M(T_1), \mathcal{P}_M(T_2) \rangle}{\|\mathcal{P}_M(T_1)\|_{\mathrm{F}} \|\mathcal{P}_M(T_2)\|_{\mathrm{F}}} =: \rho_{T_1, T_2}. \tag{12}$$

More specifically, we have

Theorem 3. *Suppose that the sample size $n \geq C_1 \mu^2 r d_1 \log d_1$, and*

$$\lambda_{\min} \geq C_2 \mu \sigma_\xi \kappa_0^2 \sqrt{\frac{r d_1^3 \log^2 d_1}{n}},$$

for some constants $C_1, C_2 > 0$. For any two matrices $T_1, T_2 \in \mathbb{R}^{d_1 \times d_2}$ such that $|\rho_{T_1,T_2}| < 1$, we have

$$\begin{aligned}
&\sup_{t_1, t_2 \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\langle \widehat{M}, T_1 \rangle - \langle M, T_1 \rangle}{\sigma_\xi \|\mathcal{P}_M(T_1)\|_{\mathrm{F}} \cdot \sqrt{d_1 d_2 / n}} \leq t_1, \frac{\langle \widehat{M}, T_2 \rangle - \langle M, T_2 \rangle}{\sigma_\xi \|\mathcal{P}_M(T_2)\|_{\mathrm{F}} \cdot \sqrt{d_1 d_2 / n}} \leq t_2 \right) - \Phi_{\rho_{T_1,T_2}}(t_1, t_2) \right| \\
&\leq C_3 \left[\frac{\kappa_0 \mu^2 \sigma_\xi}{\lambda_{\min}} \left(\frac{\|T_1\|_{\ell_1}}{\|\mathcal{P}_M(T_1)\|_{\mathrm{F}}} + \frac{\|T_2\|_{\ell_1}}{\|\mathcal{P}_M(T_2)\|_{\mathrm{F}}} \right) \sqrt{\frac{r^2 d_1^2 \log^2 d_1}{n}} + \mu \kappa_0 \sqrt{\frac{r^2 d_1 \log^3 d_1}{n}} \right],
\end{aligned}$$

where $\Phi_\rho(\cdot, \cdot)$ is the cumulative distribution function of bivariate normal distribution $N(0, ((1, \rho)^\top, (\rho, 1)^\top))$. Moreover, if both H_{0T_1} and H_{0T_2} hold, then

$$\begin{aligned}
&\sup_{t_1, t_2 \in \mathbb{R}} \left| \mathbb{P}(W_{T_1} \leq t_1, W_{T_2} \leq t_2) - \Phi_{\rho_{T_1,T_2}}(t_1, t_2) \right| \\
&\leq C_3 \left[\frac{\kappa_0 \mu^2 \sigma_\xi}{\lambda_{\min}} \left(\frac{\|T_1\|_{\ell_1}}{\|\mathcal{P}_M(T_1)\|_{\mathrm{F}}} + \frac{\|T_2\|_{\ell_1}}{\|\mathcal{P}_M(T_2)\|_{\mathrm{F}}} \right) \sqrt{\frac{r^2 d_1^2 \log^2 d_1}{n}} + \mu \kappa_0 \sqrt{\frac{r^2 d_1 \log^3 d_1}{n}} \right].
\end{aligned}$$

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:

$$\mathcal{H}_{0,\text{strong}}^2 := \left\{(T_1, T_2) \in \mathcal{H}_0 \times \mathcal{H}_0 : \rho_{T_1,T_2} \geq c q_0^{-\nu}\right\}, \tag{13}$$

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

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

Under the incoherent assumptions,

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

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

$$\frac{\mu^2 \left\|T_1\right\|_{\ell_1} \left\|T_2\right\|_{\ell_1}}{\beta_0^2 \left\|T_1\right\|_{\mathrm{F}} \left\|T_2\right\|_{\mathrm{F}}} \leq C. \tag{14}$$

(14) 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. In fact, we can use this to show that in many practical examples related to item recommendations, the linear forms are weakly correlated, regardless of the underlying matrix M.

Inference of a submatrix. Consider the inference problem with indexing matrices $\mathcal{H} = \{e_i e_j^\top : l_1 \leq i \leq l_2, l_3 \leq j \leq l_4\}$, 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 (14) is easily satisfied. Therefore, at most $O(d_1)$ pairs are strongly correlated (share the same row/column) for each linear form so that $\beta_{\mathsf{s}} \lesssim 1/d_2$.

Inference of entrywise comparisons. We can also consider comparison between two entries M_{i_1,j_1} and M_{i_2,j_2}: $\mathcal{H} = \{e_{i_1} e_{j_1}^\top - e_{i_2} e_{j_2}^\top : l_1 \leq i_1, i_2 \leq l_2, l_3 \leq j_1, j_2 \leq l_4\}$. 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 $O(d_1^2 d_2^2)$. 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 $\beta_{\mathsf{s}} \lesssim 1/d_2$.

Inference of several user/feature groups. For many applications, groupwise recommendation (Bi et al., 2018) is of interest. This can be formulated as testing $H_{0T} : \sum_{i \in G_k} M_{ij} \leq \theta_{kj}$ 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 $\mathcal{H} = \{\sum_{i \in G_k} e_i e_j^\top : 1 \leq k \leq K, 1 \leq j \leq d_2\}$. Note that $\|T\|_{\ell_1} / \|T\|_{\mathrm{F}} = \sqrt{|G_k|}$ for all $T \in \mathcal{H}$. If $K = \Omega(d_2)$, then

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

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 $\|\widehat{M} - M\|_{\max} \leq \mu \kappa_0 \sigma_\xi \sqrt{r^2 d_1 \log^2(d_1)/n}$, which implies that

$$\left|\left\langle M - \widehat{M}, T \right\rangle\right| \leq \left\|\widehat{M} - M\right\|_{\max} \|T\|_{\ell_1} \leq \mu \kappa_0 \sigma_\xi \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}} \, \|T\|_{\ell_1}.$$

Thus, a signal can be consistently identified if

$$\frac{|\langle M, T\rangle - \theta_T|}{\|T\|_{\ell_1} \sqrt{\log(q \vee d_1)}} \geq C_{\mathsf{gap}} \cdot \mu \kappa_0 \sigma_\xi \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}} \tag{15}$$

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 (15) holds. 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 4. *Suppose that*

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

There exists a universal constants $C > 0$ such that if the sample size $n \geq C \mu^2 r d_1 \log d_1$, then

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

and

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

The first claim implies that

$$\mathrm{FDR} = \mathbb{E}(\mathrm{FDP}) \leq \alpha(1 + o(1)), \tag{16}$$

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 4. Our proof actually establishes stronger results in a nonasymptotic form. Theorem 4 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 4 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.

4 Whitening and Screening

Theorem 4 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 $Z^{(i)} = \left[W^{(i)}_{T_1}, W^{(i)}_{T_2}, \dots, W^{(i)}_{T_q}\right]^\top \in \mathbb{R}^q$, for $i = 1, 2$. By Theorem 3, $Z^{(i)} \approx_d N(\mathsf{w}, R)$ 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 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 $\widehat{R}_{\mathcal{A}}^{-1/2}$ is the submatrix of $\widehat{R}^{-1/2}$ with only columns indexed by $\mathcal{A}$. Similarly, $\widehat{\mathsf{w}}_{\mathcal{A}}$ is the

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 $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, rank r, FDR level α, regularization parameter $\lambda \geq 0$.

1: Apply Algorithm 1 to get $Z^{(1)}$, $Z^{(2)}$ from $\{\mathcal{D}_0, \mathcal{D}_1\}$ and $\{\mathcal{D}_0, \mathcal{D}_2\}$ respectively

2: From $\mathcal{D}_1$, obtain a covariance estimate $\widehat{R} = (\widehat{\rho}_{T_i,T_j})_{i,j=1}^q$ using $\widehat{M}^{(1)}$ estimated from Algorithm 1, that is

$$\widehat{\rho}_{T_i,T_j} = \frac{\left\langle \mathcal{P}_{\widehat{M}^{(1)}}(T_i), \mathcal{P}_{\widehat{M}^{(1)}}(T_j) \right\rangle}{\left\| \mathcal{P}_{\widehat{M}^{(1)}}(T_i) \right\|_{\mathrm{F}} \left\| \mathcal{P}_{\widehat{M}^{(1)}}(T_j) \right\|_{\mathrm{F}}},$$

Solve LASSO estimator

$$\widehat{\mathsf{w}}^{(1)} := \underset{\mathsf{w} \in \mathbb{R}^q}{\arg\min} \left\{ \frac{1}{2} \left\| \widehat{R}^{-1/2}(Z^{(1)} - \mathsf{w}) \right\|^2 + \lambda \left\| \mathsf{w} \right\|_{\ell_1} \right\}.$$

3: 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 $i \in \mathcal{A}$.

4: 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

$$L := \inf \left\{ t > 0 : \frac{\sum_{i=1}^q \mathbb{I}\left(\mathsf{w}_{T_i}^{\mathsf{Rank}} < -t \right)}{\sum_{i=1}^q \mathbb{I}\left(\mathsf{w}_{T_i}^{\mathsf{Rank}} > t \right) \vee 1} \leq \alpha \right\}.$$

5: Reject H_{0T_i} if $\mathsf{w}_{T_i}^{\mathsf{Rank}} > L$

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 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

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

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

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

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 define strong signal, write

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

Then the limiting covariance matrix of W_Ts is given by

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

Define

$$\mathcal{S}' = \left\{T \in \mathcal{H} : \frac{|\langle M, T\rangle - \theta_T|}{\|T\|_{\ell_1} \sqrt{q_1 \log(q \vee d)}} \geq C_{\mathsf{gap}} \cdot \mu \kappa_1^{3/2} \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}}\right\}, \tag{17}$$

where $\kappa_1 = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ is the condition number of Σ.

Let $T_{\mathcal{H}}$ be a $q \times d_1 d_2$ matrix with i-th row being $\mathrm{vec}(T_i)$ and define $\mathrm{supp}(T_{\mathcal{H}}) := \cup_{i=1}^q \mathrm{supp}(T_i)$. Let $\|\cdot\|$ denote the spectral norm of a matrix and the ℓ_2-norm of a vector, and denote $\|M\|_{2,\max} :=$

$\max_{i \in d_1} \|e_i^\top M\|$. By definition, we have $\|T_{\mathcal{H}}\|_{2,\max} = \max_{i \in [q]} \|T_i\|_{\mathrm{F}}$. Here define $\|R\|_\infty := \max_{i \in [q]} \|e_i^\top R\|_{\ell_1}$ for a matrix R. Note that $\|\cdot\|_{\max}$ and $\|\cdot\|_\infty$ are equivalent for a vector. We have the following theoretical guarantee for Algorithm 2.

Theorem 5. *Let $T_{\mathcal{H}}$ be a $q \times d_1 d_2$ matrix with i-th row being $\mathrm{vec}(T_i)$ and define $\mathrm{supp}(T_{\mathcal{H}}) := \cup_{i=1}^q \mathrm{supp}(T_i)$. Suppose that q_0' a uniform upper bound for $|\mathcal{A}_0|$ and*

$$\left(\sqrt{\beta_{\mathsf{s}}'} \vee \left(h_n + \left\|\mathsf{w}_{\mathcal{A}^c}\right\|_\infty\right)\right) \frac{q_0'}{\eta_n'} \to 0,$$

and

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

Then there are universal constants $C_1, C_2 > 0$ such that if $n \geq C_1 \mu^2 r d_1 \log d_1$ and regularization parameter $\lambda = C_2 \sqrt{\log d_1 + \log q}$ 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 $\Sigma = \left(\langle \mathcal{P}_M(T_i), \mathcal{P}_M(T_j) \rangle\right)_{i,j \in [q]}$ is not known and our whitening procedure uses an estimate in its place. The additional lower bound of $\lambda_{\min}$ in Theorem 5 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.

5 Numerical Experiments

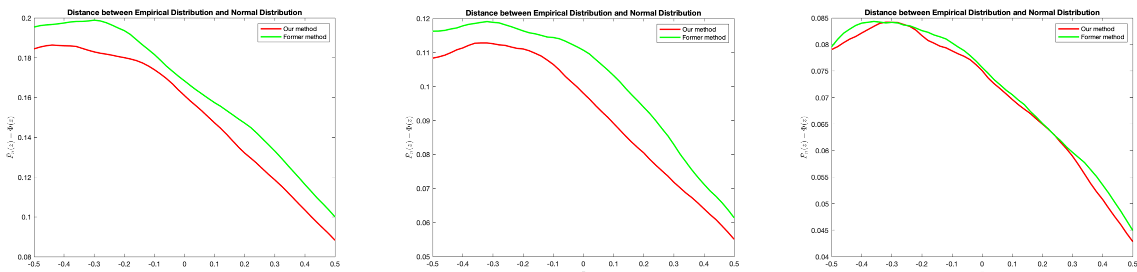
To complement our theoretical development, we also conducted several sets of numerical experiments to further demonstrate the practical merits of the proposed methodology.

5.1 Simulation Studies

We begin with a series of simulation studies aimed at illustrating the impact of several key aspects of our approach.

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 (5) 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 debiasing $n = 10^4$, and the noises $\xi \sim N(0, 1^2)$. We use the known true matrix as initialization for clearer comparisons. We first verify the FDR control in weak dependency by performing blockwise matrix tests: we test each entry in $M(1:300, 1:200)$ 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 $300 \times 200 = 60,000$ entries with probability $p = 0.2$, which leads to the following settings of m_{ij}:

$$M_{ij} - m_{ij} = \begin{cases} \mu_{ij}, & \text{with probability } p = 0.2; \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

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 $\mathcal{D}_0$ to construct an initial estimate $\widehat{M}_{\mathsf{init}}$
2. Following the construction of W_T, but use both the second and third part of data $\mathcal{D}_1$, $\mathcal{D}_2$ 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 good 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 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 linear model, it cannot exactly control the FDR at level α in our matrix completion problem. The reason this happens might be due to the heavy tail property of W_T described in Proposition 10 and the correlation of test statistics caused by low-rank projection.

5.1.3 Whitening and screening

We now evaluate Algorithm 1 and Algorithm 2 and show the advantages of de-correlation. For computational concerns, we slightly modify the implementation of Algorithm 2, which does not

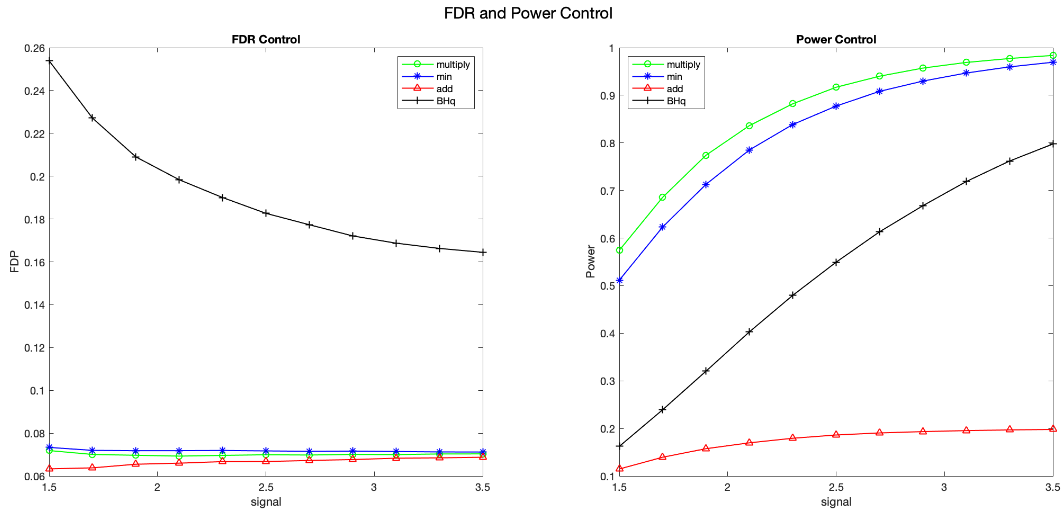


Figure 2: FDR control & Power of different data aggregation schemes in blockwise matrix tests with $\alpha = 0.1$. Here the signal is defined by μ in eq. (18).

affect the theoretical guarantees. See Algorithm 3 in the Supplement. To this end, we 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 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 $\rho_M(T_i, T_j)$ indicates the correlation of two linear form M_{T_i} and M_{T_j} and is given by (12). Here $\varrho^*(z)$ can be treated as a measure 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.

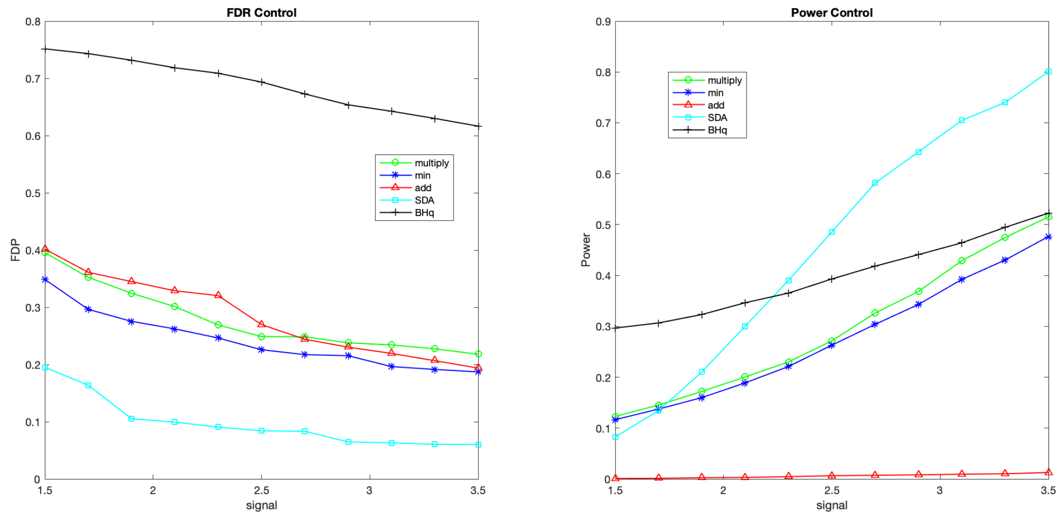
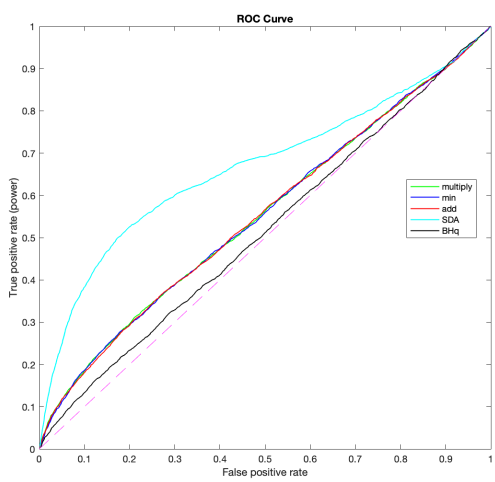
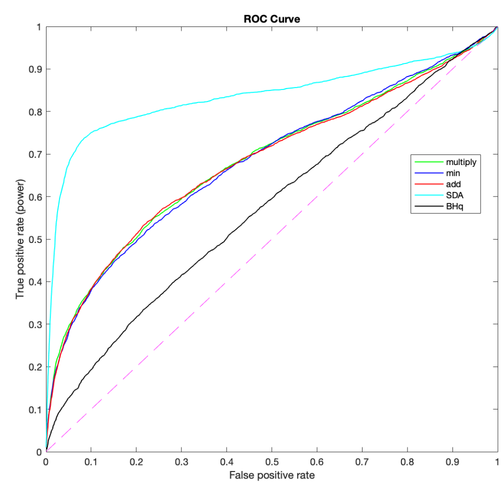


Figure 3: FDR control & Power of different data aggregation schemes in row tests with $\alpha = 0.1$. Here the signal is defined by μ in eq. (18).



(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. (18).

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

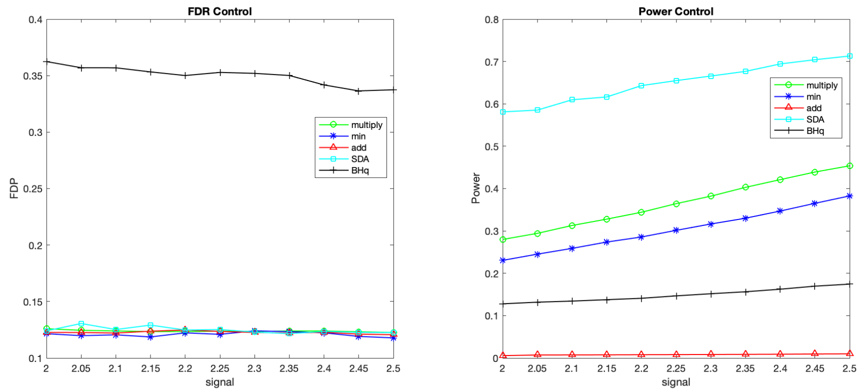
While our theories are established for sub-Gaussian noise, we observe that the proposed methods are very robust to heavy-tailed noise. This section showcases the performance of our algorithms in the existence of heavy-tailed noises, e.g., t-distribution and exponential distribution, and compares the 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. All the simulations in this section display the averaged performance of multiple independent runs.

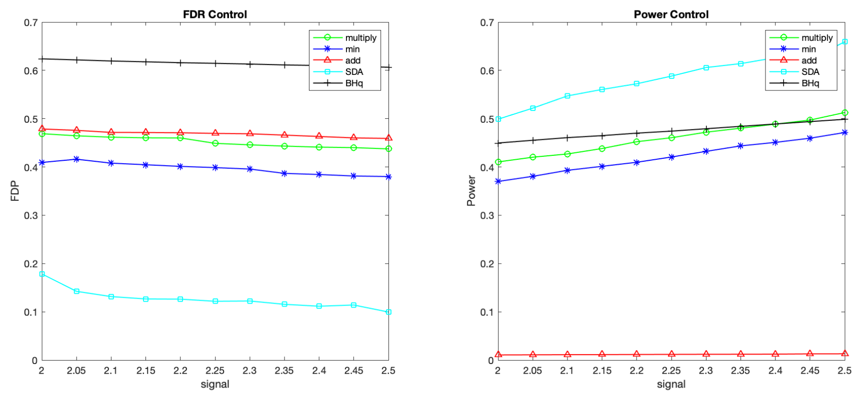
5.2 Real Data Examples

5.2.1 MovieLens

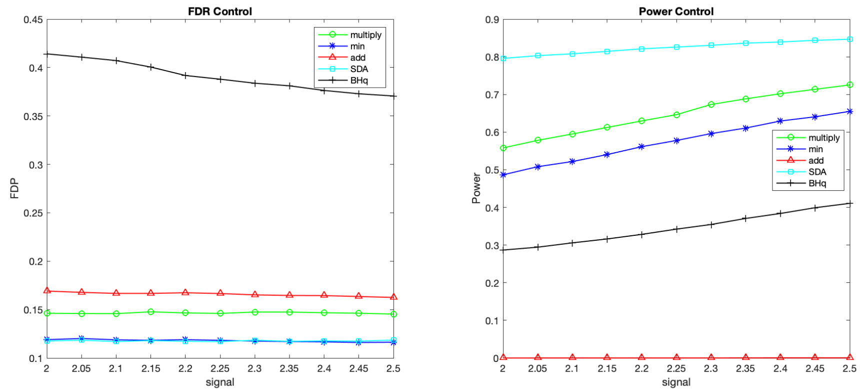
This section applies our methods to the MovieLens dataset for multiple testing and FDR control. MovieLens (Harper and Konstan, 2015), as a commonly used dataset in matrix completion problems, records millions of people's expressed preferences for movies (rated from 1-5). The dataset can be viewed as 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



(a) Sub-Gaussian noises

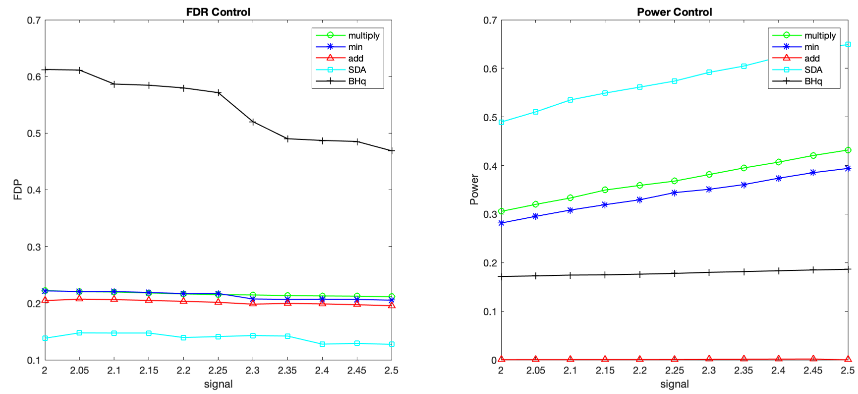


(b) Exponential noises

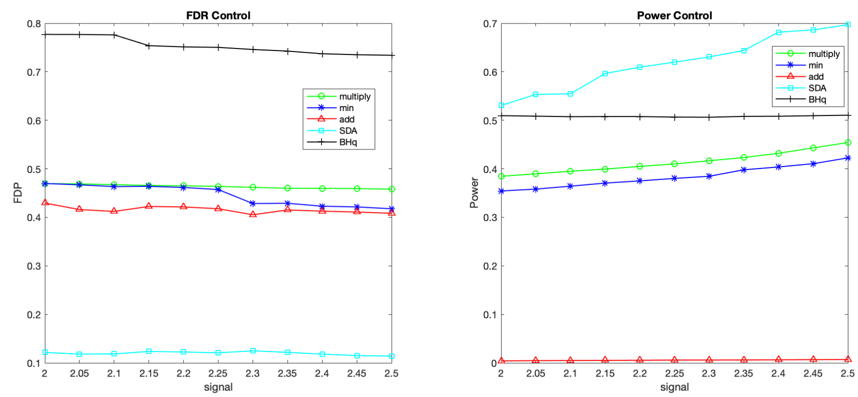


(c) Student-t noises

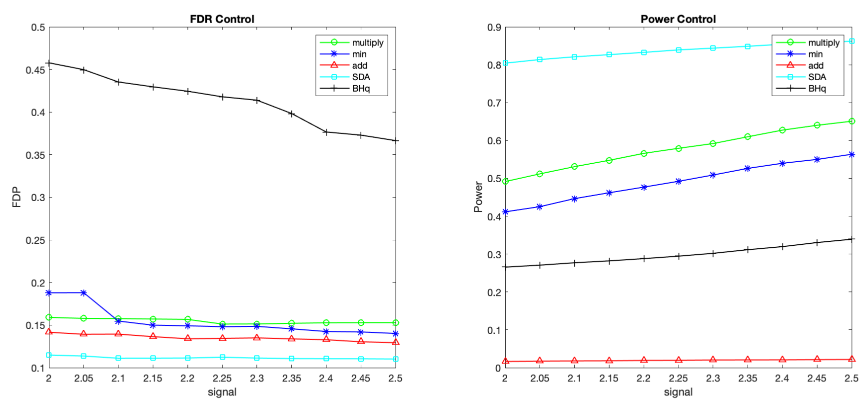
Figure 5: FDR control & Power of different data aggregation schemes for entry comparisons between rows with $\alpha = 0.1$ when the noises are heavy-tailed distributed



(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

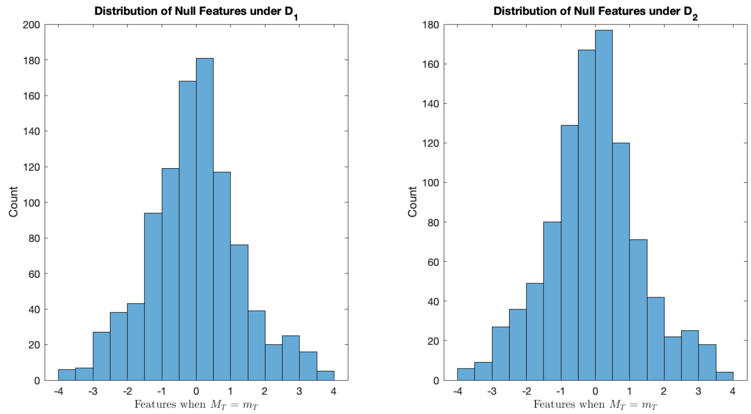


Figure 7: Symmetric distribution of all the test statistics under the null set given $\langle M, T\rangle = m_T$ for all T. The test statistics are observed to preserve a good symmetric property both on $\mathcal{D}_1$ and $\mathcal{D}_2$.

users with ratings less than 20 movies, resulting in 100,000 ratings (0-5) from 943 users on 1682 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.

Empirically, instead of initializing our algorithm by data splitting, for all the methods, we use fast Riemannian gradient descent (Wei et al., 2016; Cai et al., 2022b) on the whole data set to initialize our algorithms and then randomly split data into two parts $\mathcal{D}_1$, $\mathcal{D}_2$ to perform debiasing and 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

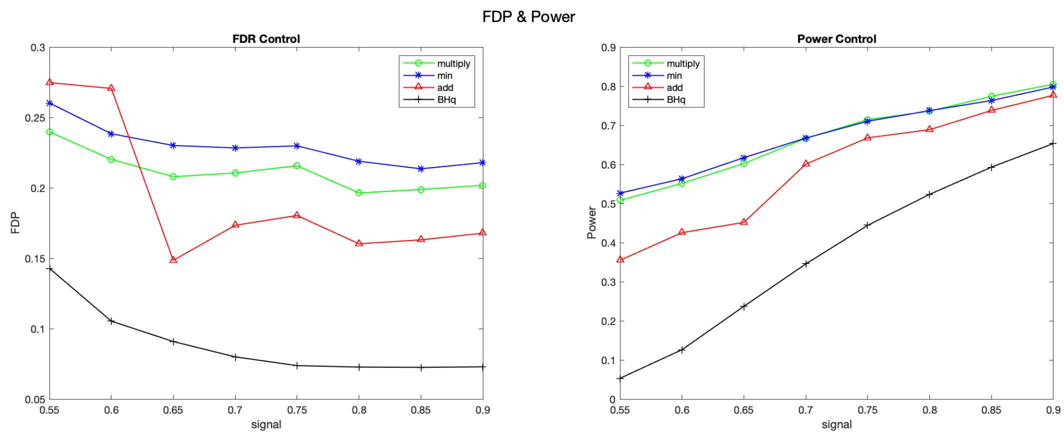
We use the Rossmann sales dataset that has recently been studied for uncertainty quantification in matrix completion (Farias et al., 2022; Gui et al., 2023). The Rossmann sales dataset records over 3,000 drug stores run by Rossmann in 7 European countries. The training set contains daily sales of 1115 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 (19). We select the first $q = 20,000$ entries sorted by rows with records in the whole dataset as our target $\mathcal{H}$. Since we select a relatively large q, according to Section 3.2, the problem is weakly correlated, which means simple data aggregation is enough to control FDR. We randomly assign null and non-null features by (18) 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

$$H_{0,ij} : M_{ij} = m_{ij} \text{ vs } H_{1,ij} : M_{ij} > m_{ij}, \text{ for all } (i,j) \in \mathcal{H}. \tag{19}$$

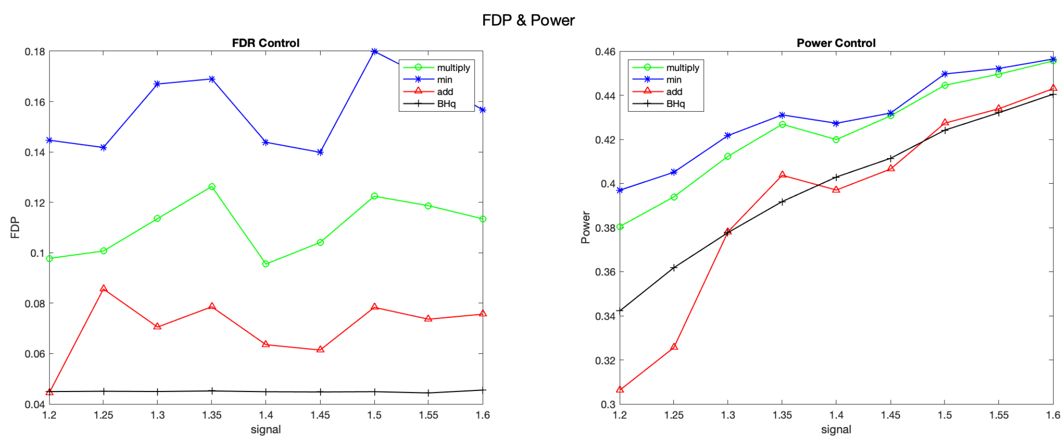
We present the result in Figure 8 and the ROC curves in Figure 9. The Rossmann sales dataset is available at `https://www.kaggle.com/c/rossmann-store-sales`.

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.

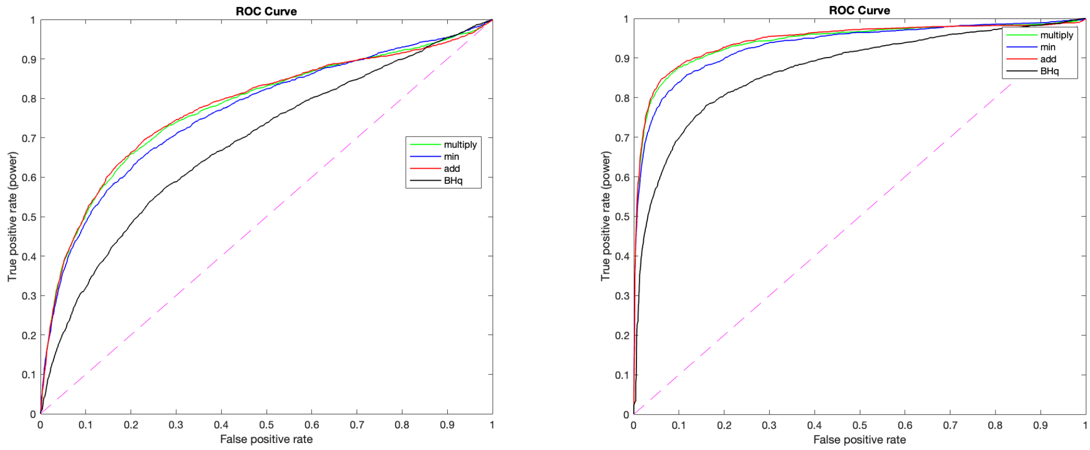


(a) Empirical FDP and power at FDR control level $\alpha = 0.2$



(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.

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, C., Poor, H. V., and Chen, Y. (2022a). Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *IEEE Transactions on Information Theory*, 69(1):407–452.

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. (2022b). Generalized low-rank plus sparse tensor estimation by fast riemannian optimization. *Journal of the American Statistical Association*, pages 1–17.

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

Cai, T. T., Zhou, W.-X., et al. (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., Livingston, B., et al. (2010). The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296.

Donoho, D. L., Huo, X., et al. (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. F., and Ma, C. (2023). Conformalized matrix completion. *arXiv preprint arXiv:2305.10637*.

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

Koltchinskii, V., Lounici, K., Tsybakov, A. B., et al. (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. 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.

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

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

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.

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., Tsybakov, A. B., et al. (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	alignment parameter
α_d	dimension ratio of matrix M: $\alpha_d := d_1/d_2$
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 $\gamma_n = \sqrt{\frac{r^2 d_1 \log^2 d_1}{n}}$
β_T	sparsity level of all indexing matrices: $\beta_T := \max_{T\in\mathcal{H}} \|T\|_{\ell_1}/\|T\|_{\mathrm{F}}$
$\mathcal{P}_M(\cdot)$	projection operators $\mathcal{P}_M(T) := T - \mathcal{P}_M^{\perp}(T) = T - U_\perp U_\perp^\top T V_\perp V_\perp^\top$
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 (9)
β_{s}	proportion of strongly correlated linear form pairs defined in (15)
η_n	number of strong signals
κ_1	condition number of covariance matrix $\Sigma = \big(\langle \mathcal{P}_M(T_j), \mathcal{P}_M(T_k)\rangle\big)_{1\le j,k\le q}$
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 $q_n = |\mathcal{A}|$, and $q_{0n} = |\mathcal{A}\cap\mathcal{H}_0|$
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 Effect of Screening and Whitening

In this subsection, 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

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

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

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

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 $\operatorname{tr}\left(\Sigma_{\mathcal{A},\mathcal{A}} - \Sigma_{\mathcal{A},\mathcal{A}^c} \Sigma_{\mathcal{A}^c,\mathcal{A}^c}^{-1} \Sigma_{\mathcal{A}^c,\mathcal{A}}\right) \le \operatorname{tr}(\Sigma_{\mathcal{A},\mathcal{A}})$, which is smaller than the total variance of the unscreened statistics $\operatorname{tr}(\Sigma)$.

A special case of multiple testing is defined by making

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

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 general non-asymptotic version of our theoretical guarantees.

A.2.1 Weak dependence

Theorem 6. *Under the conditions of Theorem 4,*

(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

$$\text{FDP} := \frac{\sum_{T \in \mathcal{H}_0} \mathbb{I}(W_T^{\mathsf{Rank}} > L)}{\left(\sum_{T \in \mathcal{H}} \mathbb{I}(W_T^{\mathsf{Rank}} > L)\right) \vee 1} \leq \alpha(1+\varepsilon), \tag{20}$$

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

$$\text{POWER} := \frac{\sum_{T \in \mathcal{H}_1} \mathbb{I}(W_T^{\mathsf{Rank}} > L)}{q_1} \geq (1-\varepsilon)\frac{\eta_n}{q_1}. \tag{21}$$

Note that Part (a) also implies that

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \alpha + C_2 h_n + C_2 \alpha^{\frac{2}{3}} \log(\frac{q_0}{\alpha\eta_n}) \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\alpha^2 \eta_n^2} \right)^{\frac{1}{6}} + \left(\frac{h_n q_0}{\alpha \eta_n} \right)^{\frac{1}{6}} + (\alpha\eta_n q_0)^{-\frac{\nu}{12}} \right). \tag{22}$$

A.2.2 Whitening and screening

Theorem 7. *Under the settings of Theorem 5, suppose that*

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

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

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

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:

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

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

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

we can convert this signal requirement to a stronger but clearer one presented in Theorem 5. In the subsequent proofs, we shall prove the non-asymptotic versions of Theorem 4 and 5.

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 $\mathcal{D}_0$ and $\mathcal{D}_1$, 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 there exists a large absolute constant $C_0 > 0$ such that the matrix signal strength satisfies*

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

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* $\lambda = C\sqrt{\log d_1 + \log q}$*, LASSO can recover the signal with precision*

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

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 stress that our theory can hold with non-identified signals as long as $\|\mathsf{w}_{\mathcal{A}^c}\|_\infty$ 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* $\mathcal{D}_0$ *and* $\mathcal{D}_1$*, we have*

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

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

A.4 A Practical Alternative to Algorithm 2

Note that Algorithm 2 involves the computation of the correlation coefficient matrix. In practice, one could also use the following alternative to Algorithm 2 that avoids computing the inverse of diagonal elements of the covariance matrix but at the same time enjoys the same theoretical guarantees.

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 $\mathcal{D}_0$, $\mathcal{D}_1$, $\mathcal{D}_2$, rank r, FDR level α.

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

$$\mathbf{W}_i^{(k)} = \widehat{s}_{T_i}^{(k)} W_{T_i}^{(k)} = \frac{\widehat{M}_{T_i}^{(k)} - \theta_{T_i}}{\widehat{\sigma}_{\xi}^{(k)} \sqrt{d_1 d_2 / n}}, \ k = 1, 2, \text{ and } \widehat{D} = \operatorname{diag}\left(\widehat{s}_{T_1}^{(1)}, \dots, \widehat{s}_{T_p}^{(1)}\right).$$

Here $\widehat{s}_{T_i}^{(k)} = \left\|\mathcal{P}_{\widehat{M}^{(k)}}(T_i)\right\|_{\mathrm{F}}$ 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 $\widehat{U}$, $\widehat{V}$ from $\mathcal{D}_0$ and $\mathcal{D}_1$:

$$\widehat{\Sigma} = T_{\mathcal{H}}(I_{d_1 d_2} - \widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top) T_{\mathcal{H}}^\top, \tag{23}$$

and write $\mathbf{X} = \widehat{\Sigma}^{-\frac{1}{2}}$. Construct response $\mathbf{y}_1 = \mathbf{X}\mathbf{W}^{(1)}$, and solve LASSO

$$\widehat{\mathsf{w}}^{(1)} = \underset{\mathsf{w} \in \mathbb{R}^q}{\arg\min} \left\{\frac{1}{2}\left\|\mathbf{y}_1 - \mathbf{X}\widehat{D}\mathsf{w}\right\|^2 + \lambda \left\|\mathsf{w}\right\|_{\ell_1}\right\}.$$

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

$$L = \inf\left\{t > 0 : \frac{\sum_{T \in \mathcal{H}} \mathbb{I}\left(\widehat{\mathsf{w}}_T^{\mathsf{Rank}} < -t\right)}{\sum_{T \in \mathcal{H}} \mathbb{I}\left(\widehat{\mathsf{w}}_T^{\mathsf{Rank}} > t\right) \vee 1} \le \alpha\right\}.$$

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. 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

$$W(X_i^1, X_i^2) = \operatorname{sign}(X_i^1 X_i^2) f(\left|X_i^1\right|, \left|X_i^2\right|). \tag{24}$$

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

$$f_1(u,v) = uv, \quad f_2(u,v) = \min(u,v), \quad f_3(u,v) = u + v. \tag{25}$$

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

$$X = \boldsymbol{\delta} + \boldsymbol{\xi}, \tag{26}$$

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 (24) 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:

$$\begin{aligned} \mathrm{FDR}_W(L) &= \frac{\mathbb{P}(Y > L, H_0)}{\mathbb{P}(Y > L)} = \frac{(1-\pi_1)\mathbb{P}(Y > L|H_0)}{(1-\pi_1)\mathbb{P}(Y > L|H_0) + \pi_1 \mathbb{P}(Y > L|H_1)}, \\ \mathrm{Power}_W(L) &= \mathbb{P}(Y > L|H_1), \end{aligned} \tag{27}$$

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.,

$$L_\alpha = \min\{L > 0 : \mathrm{FDR}_W(L) = \alpha\}, \tag{28}$$

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 (28). It suffices to compare $\mathrm{Power}_{W_j}(L_{\alpha j})$. This is equivalent to comparing the quantities $\text{Power}_{W_j}(L_{pj})$ 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 $\text{Power}_{W_j}(L_{pj})$ 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 8. *Consider the limiting behaviors* (27) *of different data aggregation methods in* (25) *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:*

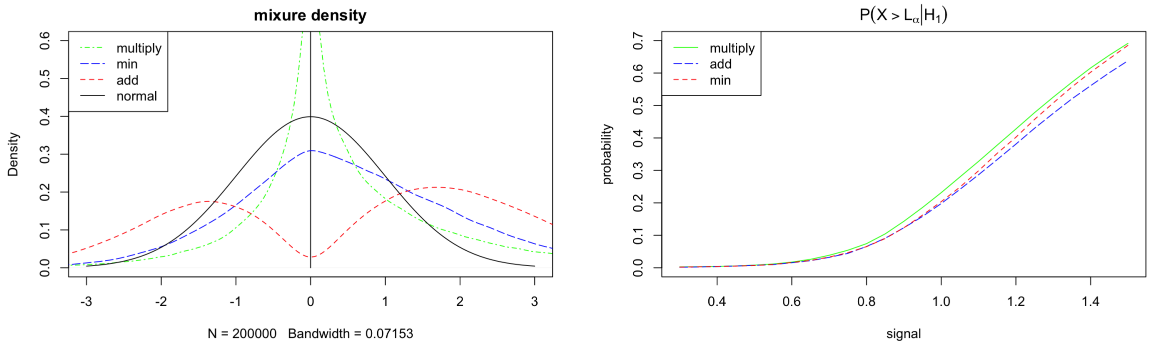
$$\text{Power}_{W_1}(L_{\alpha 1}) \geq \text{Power}_{W_2}(L_{\alpha 2}) \geq \text{Power}_{W_3}(L_{\alpha 3}),$$

for any bounded prior $\boldsymbol{\Theta}$: $\mathbb{P}(|\delta| \leq \delta_0|\boldsymbol{\Theta}) \to 1$ 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, 2

Proof. The general proof strategy is similar to that of Xia and Yuan (2021) but with a more involved discussion at the core step: separating the negligible part of the statistic out of the sum of the i.i.d part, followed by controlling the sum of the i.i.d. part by the Berry-Essen theorem. We start with the decomposition of our statistics. Denote $\widehat{\Delta} = M - \widehat{M}_{\mathsf{init}}$, and

$$\widehat{M}^{\mathsf{unbs}} = M + \underbrace{\overbrace{\frac{d_1 d_2}{n} \sum_{i \in I_2} \xi_i X_i}^{\widehat{Z}_1} + \overbrace{\left(\frac{d_1 d_2}{n} \sum_{i \in I_2} \left\langle \widehat{\Delta}, X_i \right\rangle X_i - \widehat{\Delta} \right)}^{\widehat{Z}_2}}_{\widehat{Z}}.$$

Here we denote I_2 the index set of observations in the sample $\mathcal{D}_2$. To ease the notation, we denote the initialization $\gamma(n, d_1, d_2, \tau) = \gamma_n$ such that $\left\| \widehat{M}^{\mathsf{init}} - M \right\|_{\max} \leq \gamma_n$, which holds with probability at least $1 - d_1^{\tau}$. We separate the vanishing part out of the asymptotic normal part:

$$\widehat{M}_T - M_T = \left\langle \widehat{U}\widehat{U}^\top \widehat{Z} \widehat{V}\widehat{V}^\top - UU^\top \widehat{Z} VV^\top, T \right\rangle + \left\langle \widehat{U}\widehat{U}^\top M \widehat{V}\widehat{V}^\top - M, T \right\rangle + \left\langle UU^\top \widehat{Z} VV^\top, T \right\rangle.$$

We shall show that, after proper rescaling: (1) $\left\langle \widehat{U}\widehat{U}^\top \widehat{Z} \widehat{V}\widehat{V}^\top - UU^\top \widehat{Z} VV^\top, T \right\rangle$ is vanishing; (2) $\left\langle \widehat{U}\widehat{U}^\top M \widehat{V}\widehat{V}^\top - M, T \right\rangle + \left\langle UU^\top \widehat{Z} VV^\top, T \right\rangle$ is asymptotic normal. The scale we require is exactly $\sqrt{d_1 d_2 / n} \sigma_\xi \left\| \mathcal{P}_M(T) \right\|_{\mathrm{F}}$. Define E_0 as the event that the initial estimate $\widehat{M}_{\mathsf{init}}$ constructed from $\mathcal{D}_0$ follows the error bound $\gamma(n, d_1, d_2)$. Then it holds that $E_0 \in \sigma(\mathcal{D}_0)$ and $\mathbb{P}(E_0) \geq 1 - d_1^{-\tau}$.

The first part $\left\langle \widehat{U}\widehat{U}^\top \widehat{Z} \widehat{V}\widehat{V}^\top - UU^\top \widehat{Z} VV^\top, T \right\rangle$ is vanishing with the rate described in Lemma 1.

Lemma 1. *Under the assumptions of incoherence and sufficient signal strength, there exists an absolute constant $C > 0$ such that conditional on E_0, if $n \geq C d_1 \log d_1$, then with probability at least $1 - 6\log d_1 \cdot d_1^{-\tau}$, the inequality*

$$\left|\left\langle \widehat{U}\widehat{U}^\top \widehat{Z}\widehat{V}\widehat{V}^\top - UU^\top \widehat{Z}VV^\top, T\right\rangle\right| \leq C_2 \tau \|T\|_{\ell_1} \mu^2 \frac{\sigma_\xi}{\lambda_{\min}} \sqrt{\frac{r d_1^2 d_2 \log d_1}{n}} \cdot \sigma_\xi \sqrt{\frac{r d_1 \log d_1}{n}}$$

uniformly holds for every T.

We turn to prove the asymptotic normality of part $\left\langle \widehat{U}\widehat{U}^\top M \widehat{V}\widehat{V}^\top - M, T\right\rangle + \left\langle UU^\top \widehat{Z}VV^\top, T\right\rangle$. The core i.i.d. part that contributes to asymptotic normality is the combination of the first order term of representation formula of the empirical spectral projectors (Xia, 2021), and $\left\langle UU^\top \widehat{Z}_1 VV^\top, T\right\rangle$. Define the $(d_1+d_2)\times(d_1+d_2)$ matrix A with the corresponding perturbation $\widehat{E}$ as

$$A = \begin{pmatrix} 0 & M \\ M^\top & 0 \end{pmatrix}, \quad \widehat{E} = \begin{pmatrix} 0 & \widehat{Z} \\ \widehat{Z}^\top & 0 \end{pmatrix}.$$

Also, define the singular subspace with its estimate as

$$\Theta = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}, \quad \widehat{\Theta} = \begin{pmatrix} \widehat{U} & 0 \\ 0 & \widehat{V} \end{pmatrix}.$$

By the representation formula (Xia, 2021), when $\lambda_{\min} \geq 2\left\|\widehat{Z}\right\|$, it follows that:

$$\widehat{\Theta}\widehat{\Theta}^\top - \Theta\Theta^\top = \sum_{k=1}^{\infty} \underbrace{\sum_{\Sigma: s_1+\cdots+s_{k+1}=k} (-1)^{1+\tau(\Sigma)} \cdot \mathfrak{P}^{-s_1} \widehat{E} \mathfrak{P}^{-s_2} \cdots \mathfrak{P}^{-s_k} \widehat{E} \mathfrak{P}^{-s_{k+1}}}_{\mathcal{S}_{A,k}(\widehat{E})}, \tag{29}$$

where $s_1, \cdots, s_{k+1} \geq 0$ are integers and $\tau(\Sigma) = \sum_{i=1}^{k+1} \mathbf{1}(s_i > 0)$. And $\mathfrak{P}^{-s}$ is defined by

$$\mathfrak{P}^{-s} = \begin{cases} \begin{pmatrix} U\Lambda^{-s}U^\top & 0 \\ 0 & V\Lambda^{-s}V^\top \end{pmatrix}, & \text{if } s \text{ is even;} \\ \begin{pmatrix} 0 & U\Lambda^{-s}V^\top \\ V\Lambda^{-s}U^\top & 0 \end{pmatrix}, & \text{if } s \text{ is odd.} \end{cases}$$

with the order 0 term:

$$\mathfrak{P}^0 = \mathfrak{P}^\perp = \begin{pmatrix} U_\perp U_\perp^\top & 0 \\ 0 & V_\perp V_\perp^\top \end{pmatrix}.$$

By 29, we have

$$
\begin{aligned}
&\widehat{\Theta}\widehat{\Theta}^\top A\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top A\Theta\Theta^\top\\
&=\left(\mathcal{S}_{A,1}\left(\widehat{E}\right)A\Theta\Theta^\top+\Theta\Theta^\top A\mathcal{S}_{A,1}\left(\widehat{E}\right)\right)+\sum_{k=2}^{\infty}\left(\mathcal{S}_{A,k}\left(\widehat{E}\right)A\Theta\Theta^\top+\Theta\Theta^\top A\mathcal{S}_{A,k}\left(\widehat{E}\right)\right)\\
&+\left(\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top\right)A\left(\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top\right).
\end{aligned}
$$

Define $\widetilde{T}=\begin{pmatrix}0 & T\\ 0 & 0\end{pmatrix}$. We can then write

$$
\begin{aligned}
\left\langle\widehat{U}\widehat{U}^\top M\widehat{V}\widehat{V}^\top-M,T\right\rangle=&\left\langle\widehat{\Theta}\widehat{\Theta}^\top A\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top A\Theta\Theta^\top,\widetilde{T}\right\rangle\\
=&\left\langle UU^\top\widehat{Z}_1V_\perp V_\perp^\top+U_\perp U_\perp^\top\widehat{Z}_1VV^\top,T\right\rangle+\left\langle UU^\top\widehat{Z}_2V_\perp V_\perp^\top+U_\perp U_\perp^\top\widehat{Z}_2VV^\top,T\right\rangle\\
&+\sum_{k=2}^{\infty}\left\langle\mathcal{S}_{A,k}\left(\widehat{E}\right)A\Theta\Theta^\top+\Theta\Theta^\top A\mathcal{S}_{A,k}\left(\widehat{E}\right),\widetilde{T}\right\rangle\\
&+\left\langle\left(\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top\right)A\left(\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top\right),\widetilde{T}\right\rangle.
\end{aligned}
$$

Combined with $\left\langle UU^\top\widehat{Z}VV^\top,T\right\rangle$, we have the decomposition

$$
\begin{aligned}
&\left\langle\widehat{U}\widehat{U}^\top M\widehat{V}\widehat{V}^\top-M,T\right\rangle+\left\langle UU^\top\widehat{Z}VV^\top,T\right\rangle\\
=&\left\langle UU^\top\widehat{Z}_1V_\perp V_\perp^\top+U_\perp U_\perp^\top\widehat{Z}_1VV^\top,T\right\rangle+\left\langle UU^\top\widehat{Z}_1VV^\top,T\right\rangle\\
&+\left\langle UU^\top\widehat{Z}_2V_\perp V_\perp^\top+U_\perp U_\perp^\top\widehat{Z}_2VV^\top,T\right\rangle+\left\langle UU^\top\widehat{Z}_2VV^\top,T\right\rangle\\
&+\sum_{k=2}^{\infty}\left\langle\mathcal{S}_{A,k}\left(\widehat{E}\right)A\Theta\Theta^\top+\Theta\Theta^\top A\mathcal{S}_{A,k}\left(\widehat{E}\right),\widetilde{T}\right\rangle\\
&+\left\langle\left(\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top\right)A\left(\widehat{\Theta}\widehat{\Theta}^\top-\Theta\Theta^\top\right),\widetilde{T}\right\rangle.
\end{aligned}
\tag{30}
$$

Denote the sum of i.i.d. part as $\bar{Z}=\left\langle UU^\top\widehat{Z}_1V_\perp V_\perp^\top+U_\perp U_\perp^\top\widehat{Z}_1VV^\top,T\right\rangle+\left\langle UU^\top\widehat{Z}_1VV^\top,T\right\rangle$. Compute the second-order moment and third-order moment, respectively:

$$
\begin{aligned}
\mathbb{E}\bar{Z}^2=&\frac{d_1^2d_2^2}{n}\mathbb{E}\xi^2\left(\left\langle U_\perp U_\perp^\top XVV^\top,T\right\rangle+\left\langle UUXV_\perp V_\perp^\top,T\right\rangle+\left\langle UU^\top XVV^\top,T\right\rangle\right)^2\\
=&\frac{d_1^2d_2^2}{n}\sigma_\xi^2\mathbb{E}\left[\left\langle UU^\top X,T\right\rangle^2+\left\langle U_\perp U_\perp^\top XVV^\top,T\right\rangle^2+2\left\langle UU^\top X,T\right\rangle\left\langle U_\perp U_\perp^\top XVV^\top,T\right\rangle\right]\\
=&\frac{d_1d_2}{n}\sigma_\xi^2\sum_{i_1\in[d_1]}\sum_{i_2\in[d_2]}(e_{i_1}^\top UU^\top Te_{i_2})^2+(e_{i_1}^\top U_\perp U_\perp^\top TVV^\top e_{i_2})^2+2e_{i_2}^\top VV^\top T^\top U_\perp U_\perp^\top e_{i_1}e_{i_1}^\top UU^\top Te_{i_2}\\
=&\frac{d_1d_2}{n}\sigma_\xi^2\left(\left\|U^\top T\right\|_{\mathrm{F}}^2+\left\|U_\perp^\top TV\right\|_{\mathrm{F}}^2\right)=\frac{d_1d_2}{n}\sigma_\xi^2\left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}^2.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\bar{Z}^3 \leq & C\frac{d_1^3 d_2^3}{n^2}\sigma_\xi^3 \mathbb{E}\left(\langle U_\perp U_\perp^\top X V V^\top, T\rangle + \langle U U X V_\perp V_\perp^\top, T\rangle + \langle U U^\top X V V^\top, T\rangle\right)^3 \\
\leq & C\frac{d_1^3 d_2^3}{n^2}\sigma_\xi^3 \cdot 4\mathbb{E}\left[\langle U U^\top X, T\rangle^3 + \langle U_\perp U_\perp^\top X V V^\top, T\rangle^3\right] \\
\leq & C\frac{d_1^3 d_2^3}{n^2}\sigma_\xi^3 \mathbb{E}\left[\langle U U^\top X, T\rangle^2 \left|\langle U U^\top X, T\rangle\right| + \langle U_\perp U_\perp^\top X V V^\top, T\rangle^2 \left|\langle U_\perp U_\perp^\top X V V^\top, T\rangle\right|\right].
\end{aligned}
$$

Since, by incoherence assumption,

$$
\begin{aligned}
\left|\langle U U^\top X, T\rangle\right| = \left|\langle U^\top e_{i_1} e_{i_2}^\top, U^\top T\rangle\right| \leq \left\|U^\top T\right\|_{\mathrm{F}} \mu\sqrt{\frac{r}{d_1}}, \\
\left|\langle U_\perp U_\perp^\top X V V^\top, T\rangle\right| = \left|\langle U_\perp^\top e_{i_1} e_{i_2}^\top V, U_\perp^\top T V\rangle\right| \leq \left\|U_\perp^\top T V\right\|_{\mathrm{F}} \mu\sqrt{\frac{r}{d_2}},
\end{aligned}
$$

the third-order moment is thus bounded by:

$$
\mathbb{E}\bar{Z}^3 \leq C\frac{d_1^2 d_2^{1.5}\mu\sqrt{r}}{n^2}\sigma_\xi^3\left(\left\|U^\top T\right\|_{\mathrm{F}}^3 + \left\|U_\perp^\top T V\right\|_{\mathrm{F}}^3\right) \leq C\frac{d_1^2 d_2^{1.5}\mu\sqrt{r}}{n^2}\sigma_\xi^3 \left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}^3.
$$

By invoking the Berry-Essen theorem to get the error bound of normal approximation, we have

$$
\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\left\langle U U^\top \widehat{Z}_1 + U_\perp U_\perp^\top \widehat{Z}_1 V V^\top, T\right\rangle}{\sigma_\xi \left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}\sqrt{d_1 d_2/n}} \leq t\right) - \Phi(t)\right| \leq C\mu\sqrt{\frac{r d_1}{n}}. \tag{31}
$$

We now return to (30) to control other terms. We show that, apart from $\bar{Z}$, all the other terms in (30) will vanish uniformly for all $T \in \mathcal{H}_0$ after scaling.

Lemma 2. *Under the assumptions of incoherence and sufficient signal strength, there exists an absolute constant $C > 0$ such that conditional on E_0, if $n \geq C d_1 \log d_1$, then with probability at least $1 - 2\cdot d_1^{-\tau}$, the inequality*

$$
\begin{aligned}
&\left|\left\langle U U^\top \widehat{Z}_2 V_\perp V_\perp^\top + U_\perp U_\perp^\top \widehat{Z}_2 V V^\top, T\right\rangle + \left\langle U U^\top \widehat{Z}_2 V V^\top, T\right\rangle\right| \\
&\qquad \leq C\sqrt{\tau}\sigma_\xi\gamma_n \left\|\mathcal{P}_M(T)\right\|_{\mathrm{F}}\sqrt{\frac{d_1 d_2 \log d_1}{n}}
\end{aligned}
$$

uniformly holds for every T.

Lemma 3 (Xia and Yuan (2021))**.** *Under the assumptions of incoherence and sufficient signal strength, there exists an absolute constant $C > 0$ such that conditional on E_0, if $n \geq C d_1 \log d_1$, then with probability at least $1 - 6\log d_1 \cdot d_1^{-\tau}$, inequalities*

$$
\begin{aligned}
&\left|\sum_{k=2}^{\infty}\left\langle \mathcal{S}_{A,k}\left(\widehat{E}\right) A\Theta\Theta^\top + \Theta\Theta^\top A\mathcal{S}_{A,k}\left(\widehat{E}\right), \widetilde{T}\right\rangle\right| \leq C\tau\|T\|_{\ell_1}\mu^2\sigma_\xi\sqrt{\frac{r d_1 \log d_1}{n}}\cdot\left(\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{r d_1^2 d_2 \log d_1}{n}}\right), \\
&\left|\left\langle\left(\widehat{\Theta}\widehat{\Theta}^\top - \Theta\Theta^\top\right) A\left(\widehat{\Theta}\widehat{\Theta}^\top - \Theta\Theta^\top\right), \widetilde{T}\right\rangle\right| \leq C\tau\kappa_0\|T\|_{\ell_1}\mu^2\sigma_\xi\sqrt{\frac{r d_1 \log d_1}{n}}\cdot\left(\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{r d_1^2 d_2 \log d_1}{n}}\right).
\end{aligned}
$$

uniformly hold for every T.

Now combine (31) with Lemma 1, 2, and 3. By Lipschitz property of $\Phi(t)$, we conclude that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\widehat{M}_T - M_T}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\cdot\sqrt{d_1d_2/n}}\leq t\right)-\Phi(t)\right|\leq C_2\log d_1\cdot d_1^{-\tau}+C_3\mu\sqrt{\frac{rd_1}{n}}$$
$$+C_4\tau\kappa_0\mu^2\frac{\sigma_\xi}{\lambda_{\min}}\frac{\|T\|_{\ell_1}}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}}\sqrt{\frac{\alpha_d r^2 d_1d_2\log^2 d_1}{n}}+C_5\tau\gamma_n\sqrt{\log d_1},$$

which proves the first statement.

Since

$$\frac{\widehat{M}_T - M_T}{\widehat{\sigma}_\xi\widehat{s}_T\cdot\sqrt{d_1d_2/n}}-\frac{\widehat{M}_T - M_T}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\cdot\sqrt{d_1d_2/n}}=\frac{\widehat{M}_T - M_T}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\cdot\sqrt{d_1d_2/n}}\left(\frac{\sigma_\xi\,\|\mathcal{P}_M(T)\|_{\mathrm{F}}}{\widehat{\sigma}_\xi\widehat{s}_T}-1\right), \tag{32}$$

and

$$\left|\frac{\widehat{M}_T - M_T}{\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\cdot\sqrt{d_1d_2/n}}\right|\leq C\sqrt{\tau\log d_1},$$

with probability at least $1-C\log d_1\cdot d_1^{-\tau}-C\tau h_n$, we focus on the term $\frac{\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}}{\widehat{\sigma}_\xi\widehat{s}_T}-1$ and discuss the estimation accuracy of σ_ξ and s_T respectively.

The estimation of σ_ξ shares the accuracy

$$\left|1-\frac{\widehat{\sigma}_\xi}{\sigma_\xi}\right|\leq\left|1-\frac{\widehat{\sigma}_\xi^2}{\sigma_\xi^2}\right|\leq\frac{C_1\tau\log d_1}{\sqrt{n}}+C_2\gamma_n^2,$$

with probability at least $1-2d^{-\tau}$ by Bernstein inequality; under the event that Lemma 1 holds, it is also clear that

$$\left|1-\frac{\widehat{s}_T}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}}\right|\leq\left|1-\frac{\widehat{s}_T^2}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}^2}\right|\leq\left|\frac{\|U^\top T\|_{\mathrm{F}}^2-\left\|\widehat{U}^\top T\right\|_{\mathrm{F}}^2+\|U_\perp^\top TV\|_{\mathrm{F}}^2-\left\|\widehat{U}_\perp^\top T\widehat{V}\right\|_{\mathrm{F}}^2}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}^2}\right|,$$

and

$$\left|\|U^\top T\|_{\mathrm{F}}^2-\left\|\widehat{U}^\top T\right\|_{\mathrm{F}}^2\right|\leq C_1\mu^2\tau\frac{\|T\|_{\ell_1}^2}{d_1}\cdot\left(\frac{\sigma_\xi^2}{\lambda_{\min}^2}\right)\frac{rd_1^2d_2\log d_1}{n}+C_2\|U^\top T\|_{\mathrm{F}}\|T\|_{\ell_1}\mu\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{r\tau d_1d_2\log d_1}{n}}.$$

We compute the difference of T's projections onto both singular subspaces and their orthogonal complements:

$$\left|\|U_\perp^\top TV\|_{\mathrm{F}}^2-\left\|\widehat{U}_\perp^\top T\widehat{V}\right\|_{\mathrm{F}}^2\right|\leq\left|\|U_\perp^\top TV\|_{\mathrm{F}}^2-\left\|U_\perp^\top T\widehat{V}\right\|_{\mathrm{F}}^2\right|+\left|\left\|\widehat{U}_\perp^\top TV\right\|_{\mathrm{F}}^2-\left\|\widehat{U}_\perp^\top T\widehat{V}\right\|_{\mathrm{F}}^2\right|.$$

It follows that

$$\left| \left\| U_\perp^\top T V \right\|_{\mathrm{F}}^2 - \left\| U_\perp^\top T \widehat{V} \right\|_{\mathrm{F}}^2 \right| \leq \left\| U_\perp^\top T (VV^\top - \widehat{V}\widehat{V}^\top) \right\|_{\mathrm{F}}^2 + 2 \left| \langle U_\perp^\top T (VV^\top - \widehat{V}\widehat{V}^\top), U_\perp^\top T V \rangle \right|$$
$$\leq C_1 \mu^2 \tau \frac{\|T\|_{\ell_1}^2}{d_2} \cdot \left(\frac{\sigma_\xi^2}{\lambda_{\min}^2} \right) \frac{r d_1^2 d_2 \log d_1}{n} + C_2 \|U_\perp^\top T V\|_{\mathrm{F}} \|T\|_{\ell_1} \mu \cdot \frac{\sigma_\xi}{\lambda_{\min}} \sqrt{\frac{r\tau d_1^2 \log d_1}{n}},$$

and

$$\left| \left\| U_\perp^\top T \widehat{V} \right\|_{\mathrm{F}}^2 - \left\| \widehat{U}_\perp^\top T \widehat{V} \right\|_{\mathrm{F}}^2 \right| \leq \left\| (\widehat{U}\widehat{U}^\top - UU^\top) T \widehat{V}\widehat{V}^\top \right\|_{\mathrm{F}}^2 + 2 \left| \langle U_\perp^\top T (VV^\top - \widehat{V}\widehat{V}^\top), U_\perp^\top T \widehat{V} \rangle \right|$$
$$\leq C_1 \mu^2 \tau \frac{\|T\|_{\ell_1}^2}{d_1} \cdot \left(\frac{\sigma_\xi^2}{\lambda_{\min}^2} \right) \frac{r d_1^2 d_2 \log d_1}{n} + C_2 (\|U_\perp^\top T V\|_{\mathrm{F}}$$
$$+ \|U_\perp^\top T (\widehat{V}\widehat{V}^\top - VV^\top)\|_{\mathrm{F}}) \|T\|_{\ell_1} \mu \cdot \frac{\sigma_\xi}{\lambda_{\min}} \sqrt{\frac{r\tau d_1^2 \log d_1}{n}}$$
$$\leq C_1 \mu^2 \tau \frac{\|T\|_{\ell_1}^2}{d_1} \cdot \left(\frac{\sigma_\xi^2}{\lambda_{\min}^2} \right) \frac{r d_1^2 d_2 \log d_1}{n} + C_2 \|U_\perp^\top T V\|_{\mathrm{F}} \|T\|_{\ell_1} \mu \cdot \frac{\sigma_\xi}{\lambda_{\min}} \sqrt{\frac{r\tau d_1^2 \log d_1}{n}}$$
$$+ C_3 \mu^2 \tau \frac{\|T\|_{\ell_1}^2}{d_2} \cdot \left(\frac{\sigma_\xi^2}{\lambda_{\min}^2} \right) \frac{r d_1^2 d_2 \log d_1}{n}.$$

We then can show that

$$\left| 1 - \frac{\widehat{s}_T}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}} \right| \leq \left| 1 - \frac{\widehat{s}_T^2}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}^2} \right| \leq C_2 \mu \frac{\|T\|_{\ell_1}}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}} \cdot \frac{\sigma_\xi}{\lambda_{\min}} \sqrt{\frac{\tau r \alpha_d d_1 d_2 \log d_1}{n}}, \tag{33}$$

which indicates the fact that, with probability at least $1 - C \log d_1 d_1^{-\tau}$,

$$\left| \frac{\widehat{\sigma}_\xi \widehat{s}_T}{\sigma_\xi \|\mathcal{P}_M(T)\|_{\mathrm{F}}} - 1 \right| \leq \left| \left(\frac{\widehat{\sigma}_\xi}{\sigma_\xi} - 1 \right) \left(\frac{\widehat{s}_T}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}} - 1 \right) \right| + \left| \frac{\widehat{\sigma}_\xi}{\sigma_\xi} - 1 \right| + \left| \frac{\widehat{s}_T}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}} - 1 \right|$$
$$\leq \frac{C_1 \tau \log d_1}{\sqrt{n}} + C_2 \gamma_n^2 + C_3 \mu \frac{\|T\|_{\ell_1}}{\|\mathcal{P}_M(T)\|_{\mathrm{F}}} \cdot \frac{\sigma_\xi}{\lambda_{\min}} \sqrt{\frac{\tau r \alpha_d d_1 d_2 \log d_1}{n}}.$$

Applying again the Lipschitz property of $\Phi(t)$ to the decomposition (32), we can obtain the second statement.

□

B.2 Proof of Theorem 3

Proof. Since in Lemma 1, 2, and 3, the inequalities uniformly hold for all T, we can write the couple (W_{T_1}, W_{T_2}) as

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

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

$$\begin{aligned}\max\{|\Delta_{T_1}|, |\Delta_{T_2}|\} \le & C_4\tau\kappa_0\mu^2\frac{\sigma_\xi}{\lambda_{\min}}\left(\frac{\|T_1\|_{\ell_1}}{\|\mathcal{P}_M(T_1)\|_{\mathrm{F}}} + \frac{\|T_2\|_{\ell_1}}{\|\mathcal{P}_M(T_2)\|_{\mathrm{F}}}\right)\sqrt{\frac{\alpha_d r d_1 d_2 \log^2 d_1}{n}} \\ & + C_5\tau\gamma_n\sqrt{\log d_1},\end{aligned} \tag{34}$$

by the same argument in the proof of Theorem 1. 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:

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

Then we have

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

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

$$\sup_{t_1,t_2\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_1)\right\rangle}{\sigma_\xi\|\mathcal{P}_M(T_1)\|_{\mathrm{F}}\sqrt{d_1 d_2/n}} \le t_1, \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_2)\right\rangle}{\sigma_\xi\|\mathcal{P}_M(T_2)\|_{\mathrm{F}}\sqrt{d_1 d_2/n}} \le t_2\right) - \Phi_{\rho_{T_1,T_2}}(t_1, t_2)\right| \le C\mu\sqrt{\frac{rd_1}{n}}. \tag{35}$$

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. $\square$

B.3 Proof of Theorem 4

We proceed to prove Theorem 4 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 $n = O(d_1 d_2)$ in general, we take $h_n = \Omega(\frac{\sqrt{\log d_2}}{d_2} \vee (rd_1/n)^{1/4})$ in the proof for simplicity. Here h_n can be smaller as long as n is large. Denote $E_0 \in \sigma(\mathcal{D}_0)$ as the event that we initialize our algorithm with the required accuracy using $\mathcal{D}_0$. We first start with the asymptotic property of $W_T^{(i)}$.

B.3.1 Weak dependence and symmetricity

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

Proposition 4. *Conditional on E_0, there exits a constant C_2 such that $W_T^{(1)}$ follows the asymptotic normality rate:*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(W_T^{(1)} > t|E_0) - \Phi(-t)\right| \leq C_2 h_n, \tag{36}$$

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

This proposition implies the asymptotic symmetricity of $W_T^{(1)}$, $W_T^{(2)}$ conditioned on E_0, 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 $\mathcal{D}_0$ and $\mathcal{D}_1$, W_T^{Rank} is asymptotically symmetric for $T \in \mathcal{H}_0$. Define

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

where Z is a standard Gaussian variable. Here since $W_T^{(1)} \in \sigma(\mathcal{D}_0, \mathcal{D}_1)$, $W_T^{(1)}$ is fixed conditional on $\mathcal{D}_0, \mathcal{D}_1$.

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*:

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

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

$$\mathcal{S} = \left\{T \in \mathcal{H} : \frac{\sqrt{n}\,|M_T - \theta_T|}{\sigma_\xi\sqrt{d_1d_2}\,\|\mathcal{P}_M(T)\|_{\mathrm{F}}\sqrt{(\log q + \log d)}} \geq C_{\mathsf{gap}}\right\}, \tag{37}$$

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 $\|\mathcal{P}_M(T)\|_{\mathrm{F}} \leq \|T\|_{\ell_1}\max_{i,j}\left\|\mathcal{P}_M(e_ie_j^\top)\right\|_{\mathrm{F}} \leq 3\mu\,\|T\|_{\ell_1}\sqrt{\frac{r}{d_2}}$. Thus, all the signals that satisfy condition (15) can also satisfy condition (37), meaning that the η_n defined here is always larger than that defined in (15).

Lemma 4. *Conditional on E_0 and $\mathcal{D}_1$, we have*

$$\sup_{0\le t\le L_n}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{P}(W_T^{\mathsf{Rank}}>t)}{q_0G(t)}-1\right|\le C_3\frac{h_nq_0}{\epsilon_n\eta_n}.$$

Proof. We only focus on small h_n. For each $T\in\mathcal{H}_0$, conditional on E_0 and $\mathcal{D}_1$, Proposition 4 implies that

$$\left|\mathbb{P}(W_T^{\mathsf{Rank}}>t)-\mathbb{P}(W_T^{(1)}Z>t|E_0,\mathcal{D}_1)\right|\le C_2h_n.$$

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

$$\sup_{0\le t\le L_n}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{P}(W_T^{\mathsf{Rank}}>t)}{q_0G(t)}-1\right|\le\sup_{0\le t\le L_n}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{P}(W_T^{\mathsf{Rank}}>t)-G(t)}{q_0G(L_n)}\right|\le C_3\frac{h_nq_0}{\epsilon_n\eta_n}.$$

□

Then, we explore the weak dependency of linear forms under signals and correlations assumptions. We will show that, with high probability, the *Ratio* can be very close to its population version described in Lemma 4. Although we already have the intuition of dependency between different $W_T^{(1)}$ by Theorem 3, 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 5 (Weak dependency of null statistics)**.** *Conditional on E_0, $\mathcal{D}_1$,*

$$\sup_{0\le t\le L_n}\frac{\sum_{(T_i,T_j)\in\mathcal{H}^2_{0,weak}}\left|\operatorname{cov}(\mathbb{I}(W_{T_i}^{\mathsf{Rank}}>t),\mathbb{I}(W_{T_j}^{\mathsf{Rank}}>t))\right|}{q_0^2G^2(t)}\le C_1\frac{h_nq_0}{\epsilon_n\eta_n}+C_2\frac{1}{(\epsilon_n\eta_nq_0)^{\nu/2}},\tag{38}$$

where ν is the weak correlation parameter defined in (13)*.*

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 2: denote

$$\begin{aligned}(W_{T_1}^{(i)},W_{T_2}^{(i)})&=\left(\frac{\left\langle\widehat{Z}_1^{(i)},\mathcal{P}_M(T_1)\right\rangle}{\sigma_\xi\left\|\mathcal{P}_M(T_1)\right\|_{\mathrm{F}}\sqrt{d_1d_2/n}}+\Delta_{T_1}^{(i)},\frac{\left\langle\widehat{Z}_1^{(i)},\mathcal{P}_M(T_2)\right\rangle}{\sigma_\xi\left\|\mathcal{P}_M(T_2)\right\|_{\mathrm{F}}\sqrt{d_1d_2/n}}+\Delta_{T_2}^{(i)}\right)\\&:=\left(\widetilde{W}_{T_1}^{(i)}+\Delta_{T_1}^{(i)},\widetilde{W}_{T_2}^{(i)}+\Delta_{T_2}^{(i)}\right),\ i=1,2,\end{aligned}$$

where $\widehat{Z}_1^{(i)}$, $\Delta_T^{(i)}$ are defined analogously as in the proof of Theorem 2. 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 1, the remainder term $\Delta_T^{(i)}$ is controlled by:

$$\mathbb{P}\left(\left|\Delta_T^{(i)}\right| > c_1 h_n | E_0\right) \le C_2 \frac{\log d_1}{d_1^2}. \tag{39}$$

For $i = 1$, as is shown in the proof of Theorem 2, 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 $R_{12} = R_{21} = \mathrm{cov}(\widetilde{W}_{T_1}^1, \widetilde{W}_{T_2}^1) = \rho_{T_1,T_2}$ with the error bound:

$$\left|\mathbb{P}\left((\widetilde{W}_{T_1}^{(2)}, \widetilde{W}_{T_2}^{(2)}) \in A \middle| E_0\right) - \mathbb{P}(\omega_1 \in A)\right| \le C\mu\sqrt{\frac{rd_1}{n}}, \tag{40}$$

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 (12). 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 E_0 and $\mathcal{D}_1$, we have:

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

Term (1), (2), (3) can be controlled separately. For (1), conditional on E_0 and $\mathcal{D}_1$, we invoke multivariate Berry–Esseen theorem (40) to bound the joint c.d.f. of $(W_{T_1}^{(2)}, W_{T_2}^{(2)})$ by

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

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 $\log d_1 \cdot d_1^{-2} \le h_n^2$. Analogously, it also holds that

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

We conclude that, conditional on E_0 and $\mathcal{D}_1$

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

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

$$\begin{aligned}&\left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t, W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}\omega_{11} > t, W_{T_2}^{(1)}\omega_{12} > t)\right| \\ &\le 2c_1 h_n \left(\mathbb{P}(W_{T_1}^{(1)}\omega_{11} > t) + \mathbb{P}(W_{T_2}^{(1)}\omega_{12} > t)\right) + Ch_n^2.\end{aligned} \tag{41}$$

For (2), the proof of Lemma 4 also implies the following bound

$$\begin{aligned}&\left|\mathbb{P}(W_{T_1}^{(1)}W_{T_1}^{(2)} > t)\mathbb{P}(W_{T_2}^{(1)}W_{T_2}^{(2)} > t) - \mathbb{P}(W_{T_1}^{(1)}w_{11} > t)\mathbb{P}(W_{T_2}^{(1)}w_{12} > t)\right| \\ &\le 2c_1 h_n \left(\mathbb{P}(W_{T_1}^{(1)}\omega_{11} > t) + \mathbb{P}(W_{T_2}^{(1)}\omega_{12} > t)\right) + Ch_n^2,\end{aligned} \tag{42}$$

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 $(T_1, T_2) \in \mathcal{H}_{0,\text{weak}}^2$, 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 $\phi(-t)^{\nu} \leq C_{\nu} t^{-(1-\nu)}$. Now combine this with the upper bound of $\phi(-t)$: $\phi(-t) \leq C(t+1)\Phi(-t)$, we have:

$$
\begin{aligned}
|\mathbb{P}(\omega_{11} > t_1, \omega_{12} > t_2) - \Phi(-t_1)\Phi(-t_2)| &\leq 2\rho \left[\phi(-t_1)\phi(-t_2)\right]^{1-c\rho} \\
&\leq 2\rho \left[C(\Phi(-t_1)^{-\nu}+1)\Phi(-t_1)(\Phi(-t_2)^{-\nu}+1)\Phi(-t_2)\right]^{1-c\rho} \\
&\leq C\rho \left[\Phi(-t_1)\Phi(-t_2)\right]^{(1-\nu)(1-c\rho)},
\end{aligned} \tag{43}
$$

for the term (3). Together with (41), (42), we can show that

$$
\begin{aligned}
&\sup_{0\leq t\leq L_n} \sum_{(T_i,T_j)\in\mathcal{H}^2_{0,\mathrm{weak}}} \frac{\left|\mathrm{cov}(\mathbb{I}(W^{\mathsf{Rank}}_{T_i} > t), \mathbb{I}(W^{\mathsf{Rank}}_{T_j} > t))\right|}{q_0^2 G^2(t)} \leq \frac{8c_1 h_n q_0 G(t)}{q_0^2 G(t)^2} \\
&+ \sup_{0\leq t\leq L_n} \frac{\sum_{(T_i,T_j)\in\mathcal{H}^2_{0,\mathrm{weak}}} C\rho \left[\mathbb{P}(W^{(1)}_{T_i} Z > t)\mathbb{P}(W^{(1)}_{T_j} Z > t)\right]^{(1-\nu)(1-c\rho)}}{q_0^2 G(t)^2} \\
&\leq C\frac{h_n q_0}{\epsilon_n \eta_n} + \sup_{0\leq t\leq L_n} \frac{C\rho \left(\sum_{T\in\mathcal{H}_0} \mathbb{P}(W^{(1)}_T Z > t)^{(1-\nu)(1-c\rho)}\right)^2}{q_0^2 G(t)^2} \\
&\leq C\frac{h_n q_0}{\epsilon_n \eta_n} + \sup_{0\leq t\leq L_n} C\rho \frac{(G(t))^{2(1-\nu)(1-c\rho)}}{G(t)^2} \\
&\leq C\frac{h_n q_0}{\epsilon_n \eta_n} + C\rho(\frac{q_0}{\epsilon_n \eta_n})^{3\nu}.
\end{aligned}
$$

The argument above is valid for any ν, thus, we choose 3ν to be the $\frac{\nu}{2}$, where ν is defined in (13). It thus finishes the proof. □

We now apply the weak dependency yielded in Lemma 5 to derive a uniform bound between R and its population version:

Lemma 6. *For any $\varepsilon > 0$, conditional on E_0 and $\mathcal{D}_1$, it holds that*

$$
\mathbb{P}\left(\sup_{0\leq t\leq L_n} \left|\frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W^{\mathsf{Rank}}_T > t)}{q_0 G(t)} - 1\right| \geq \varepsilon\right) \leq \frac{C}{\varepsilon^2}\log(\frac{q_0}{\epsilon_n\eta_n})\left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\epsilon_n^2\eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_n q_0}{\epsilon_n\eta_n} + \frac{1}{(\epsilon_n\eta_n q_0)^{v/2}}\right)^{\frac{1}{2}}\right).
$$

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:

$$
\frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W^{\mathsf{Rank}}_T > t_k)}{q_0 G(t_{k-1})} \leq \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W^{\mathsf{Rank}}_T > t)}{q_0 G(t)} \leq \frac{\sum_{T\in\mathcal{H}_0} \mathbb{I}(W^{\mathsf{Rank}}_T > t_{k-1})}{q_0 G(t_k)}.
$$

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

$$\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} > t)}{q_0 G(t)} - 1\right| \leq \sup_{i=k-1,k}\frac{1}{r_K}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} > t_i)}{q_0 G(t_i)} - 1\right| + |r_K - 1| \vee \left|\frac{1}{r_K} - 1\right|,$$

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

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

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

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

It follows that

$$\begin{aligned}
\mathbb{E}D_n^2 &\leq \frac{K}{r_K^2}\mathbb{E}\left|\frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} > t_k) - \mathbb{P}(W_T^{\mathsf{Rank}} > t_k)}{q_0 G(t_k)}\right|^2 \\
&\leq \frac{K}{r_K^2}\frac{\sum_{(T_1,T_2)\in\mathcal{H}_{0,\text{weak}}^2}\left|\operatorname{cov}(\mathbb{I}(W_{T_1}^{\mathsf{Rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{Rank}} > t))\right| + \sum_{T_1,T_2\in\mathcal{H}_{0,\text{strong}}^2}\left|\operatorname{cov}(\mathbb{I}(W_{T_1}^{\mathsf{Rank}} > t), \mathbb{I}(W_{T_2}^{\mathsf{Rank}} > t))\right|}{q_0^2 G^2(t)},
\end{aligned} \tag{44}$$

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

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

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

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

where we apply our previous results in Lemma 5. 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

$$|r_K - 1| \vee \left|\frac{1}{r_K} - 1\right| \leq C\frac{1}{K}\log(\frac{q_0}{\epsilon_n \eta_n}) \leq \left(\left(\frac{\beta_{\mathsf{s}} q_0^2}{\epsilon_n^2 \eta_n^2}\right)^{\varsigma} + \left(\frac{h_n q_0}{\epsilon_n \eta_n} + \frac{1}{(\epsilon_n \eta_n q_0)^{v/2}}\right)^{\varsigma}\right)$$

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

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

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

$$Ratio = \frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} > L)}{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} < -L)} = \frac{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} > L)}{q_0 G(t)} \cdot \frac{q_0 G(t)}{\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} < -L)}.$$

Then, it's clear that, under the event that $L \le L_n$, if Lemma 6 holds for a ε, then we have

$$Ratio \le \frac{1+\varepsilon}{1-\varepsilon} \le 1 + \frac{2\varepsilon}{1-\varepsilon} \le 1 + 3\varepsilon,$$

with probability at least $1 - \frac{C}{\varepsilon^2}\log(\frac{q_0}{\epsilon_n\eta_n})\left(\left(\frac{\beta_s q_0^2}{\epsilon_n^2\eta_n^2}\right)^{\frac{1}{2}} + \left(\frac{h_n q_0}{\epsilon_n\eta_n} + (\epsilon_n\eta_n q_0)^{-\nu/2}\right)^{\frac{1}{2}}\right)$. By Lemma 6, 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 4, conditional on E_0, 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) is 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:

$$\mathbb{P}\left(\max_{T\in\mathcal{H}_0}\left|W_T^{(1)}\right| \geq C(h_n+\sqrt{\log d_1+\log q})\middle|\mathcal{D}_0\right) \leq \frac{1}{d_1^2}.$$

If $T\in\mathcal{S}$, then $\delta_T \geq C_{\mathsf{gap}}\sqrt{\log d_1+\log q}$ by the definition of $\mathcal{S}$. The definition of L_n implies that $L_n \leq C\sqrt{\log(\frac{q_0}{\epsilon_n\eta_n})}\cdot\sqrt{\log d_1+\log q} \ll \log(\frac{1}{h_n})\vee(\log d_1+\log q)$. 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

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

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 (11), we have

$$\sum_{T\in\mathcal{H}}\mathbb{I}(W_T^{\mathsf{Rank}} > L_n) \geq \sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{Rank}} > L_n) \geq \frac{3}{4}\eta_n, \tag{45}$$

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 6, 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 $\epsilon_n = \alpha/8$,

$$\begin{aligned}\mathbb{P}(\sum_{T\in\mathcal{H}}\mathbb{I}(W_T^{\mathsf{Rank}} < -L_n) \geq \frac{3}{4}\alpha\eta_n) &\leq \mathbb{P}(2\sum_{T\in\mathcal{H}_0}\mathbb{I}(W_T^{\mathsf{Rank}} < -L_n) \geq \frac{\alpha}{2}\eta_n)+\mathbb{P}(\sum_{T\in\mathcal{S}}\mathbb{I}(W_T^{\mathsf{Rank}} < -L_n) \geq \frac{\alpha}{4}\eta_n)\\ &\leq \log(\frac{q_0}{\alpha\eta_n})\left(\left(\frac{\beta_{\mathsf{s}}q_0^2}{\alpha^2\eta_n^2}\right)^{\frac{1}{2}}+\left(\frac{h_nq_0}{\alpha\eta_n}+(\epsilon_n\eta_nq_0)^{-\nu/2}\right)^{\frac{1}{2}}\right)+Ch_n.\end{aligned} \tag{46}$$

Combining (45) and (46), 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 7. *Suppose that we use $\widehat{U}$, $\widehat{V}$ obtained from $\mathcal{D}_0$, $\mathcal{D}_1$ 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 $1 - Cd_1^{-2}\log d_1$, we have

$$\left\| \Sigma^{-\frac{1}{2}}(\Sigma - \widehat{\Sigma})\Sigma^{-\frac{1}{2}} \right\| \leq C \frac{\kappa_T \sigma_\xi}{\lambda_{\min}} \left(\frac{\text{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1 \right) \sqrt{\frac{\kappa_1 d_1^2 d_2 \log d_1}{n}}. \tag{47}$$

For simplicity, we denote $\kappa_T' = \kappa_T \left(\frac{\text{supp}(T_{\mathcal{H}})}{\sqrt{d_2}} \wedge 1 \right)$. Lemma 7 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 (47), i.e.,

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

We then have

$$\left| \mathbf{Q}_{ij} - \mathbf{Q}_{ij}^* \right| \leq \left| e_i^\top \left(\left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} - \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* \right)^{-1} \right) e_j \right| + \left| e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \mathbf{X}_{\mathcal{A}}^\top \Delta\Sigma \mathbf{X}_{\mathcal{A}} \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} e_j \right|. \tag{48}$$

Denote $\mathbf{Q}' = \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1}$. The first term in (48) can be controlled by:

$$\begin{aligned}
&\left| e_i^\top \left(\left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} - \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* \right)^{-1} \right) e_j \right| \\
&= \left| e_i^\top \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* - \mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right) \left(\mathbf{X}_{\mathcal{A}}^{*\top} \mathbf{X}_{\mathcal{A}}^* \right)^{-1} e_j \right| \\
&\leq C \frac{\kappa_1^{1.5} \kappa_T' \sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}} \sqrt{\mathbf{Q}_{ii}' \mathbf{Q}_{jj}^*},
\end{aligned} \tag{49}$$

where we use the fact that

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

by Fréchet derivative (Higham, 2008; Al-Mohy and Higham, 2009) and Lemma 7, and also we have

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

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

which is equivalent to

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

The second term in (48) can be bounded given that

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

However, notice that,

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

We can conclude that

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

Setting $i = j$, we also have

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

i.e.,

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

We now compare the difference of correlation coefficients:

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

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 $\mathbf{Q}$. $\square$

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 $\widehat{\mathsf{w}}_i=\frac{M_{T_i}-\theta_{T_i}}{\widehat{\sigma}_\xi^{(1)}\sqrt{d_1d_2}\widehat{s}_{T_i}^{(1)}}\sqrt{n}$ 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 $\widehat{s}_{T_i}^{(1)}=\left\|\mathcal{P}_{\widehat{M}^{(1)}}(T_i)\right\|_{\mathrm{F}}$.

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

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

By (33) in the proof of Theorem 1, we have $\left|1-\widehat{s}_T^{(1)}/s_T\right|\leq C_2\frac{\mu\beta_T}{\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2d_2\log d_1}{n}}$ with probability at least $1-Cd_1^{-2}\log d_1$. Here $D:=\operatorname{diag}(s_{T_1},\cdots,s_{T_q})$. Thus, the absolute value of the diagonal matrix can be controlled by:

$$\left|D^{-1}-\widehat{D}^{-1}\right|\preceq C_2\frac{\mu\beta_T}{\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2d_2\log d_1}{n}}D^{-1}. \tag{50}$$

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 $\frac{\mu\beta_T}{\beta_0}\cdot\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{\alpha_d d_1^2d_2\log d_1}{n}}\to 0$; and also

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

which can be derived following the same steps as Lemma 7. 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 $\widehat{\mathbf{W}}_i = \mathbf{W}_i^{(1)}/s_{T_i} - \frac{M_{T_i}-\theta_{T_i}}{\widehat{\sigma}_\xi s_{T_i}\sqrt{d_1d_2}}\sqrt{n}$. 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 $\widehat{Z}_1 = \frac{d_1d_2}{n}\sum_{i\in I_2}\xi_iX_i$, and split the noise $\widehat{\mathbf{W}} = \widehat{\mathbf{W}}_1 + \widehat{\mathbf{W}}_2$, where

$$\widehat{\mathbf{W}}_{1i} = \frac{\left\langle \widehat{Z}_1, \mathcal{P}_M(T_i)\right\rangle}{\sigma_\xi s_{T_i}\sqrt{d_1d_2/n}}, \text{ for each } i\in[q]. \tag{51}$$

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

$$e_i^\top D\Sigma^{-1}D\widehat{\mathbf{W}}_1 = \frac{\left\langle \mathrm{Vec}(\widehat{Z}_1)^\top, e_i^\top D\Sigma^{-1}T_{\mathcal{H}}(I_{d_1d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top)\right\rangle}{\sigma_\xi\sqrt{d_1d_2/n}},$$

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

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

with probability at least $1-q^{-1}d_1^{-2}$. This indicates that

$$\mathbb{P}\left(\left\|D\Sigma^{-1}D\widehat{\mathbf{W}}_1\right\|_\infty \ge C\sqrt{\kappa_1(\log d_1+\log q)}\right) \le d_1^{-2}.$$

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

Lemma 8. *If we use $\widehat{\Sigma}$ to approximate Σ, then*

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

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:

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

with probability at least $1-d_1^{-2}$. This indicates that, with probability at least $1-Cd_1^{-2}$, we have

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

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

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

Since for each i,

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

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 (51). 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:

$$\widehat{\beta}_i = \frac{\left\langle \mathrm{Vec}(\widehat{Z}_1), e_i^\top \left(\mathbf{X}_\mathcal{A}^\top \mathbf{X}_\mathcal{A}\right)^{-1} \mathbf{X} T_\mathcal{H} \left(I_{d_1 d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right) \right\rangle}{\sigma_\xi \sqrt{d_1 d_2 / n}}. \tag{52}$$

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

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

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

$$
\left|\mathbb{P}\left(\frac{\widehat{\beta}_i}{\sqrt{\mathbf{Q}_{ii}}} \leq t \middle| \mathcal{D}_0, \mathcal{D}_1\right) - \Phi(t)\right| \leq C\mu\sqrt{\frac{rd_1}{n}}. \tag{53}
$$

(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 $\left\|\widehat{\mathbf{W}}_2\right\|_\infty \leq Ch_n$, with probability at least $1 - Cd_1^{-2}\log d_1$. Thus, by writing $\mathbf{X} = [\mathbf{X}_\mathcal{A}, \mathbf{X}_{\mathcal{A}^c}]$, we have

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

Using the definition of C_∞, it follows that

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

uniformly for all i with probability at least $1 - C\log d_1 d_1^{-2}$.

(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

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

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

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

□

B.7 Proof of Theorem 5

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 4. Define the expected false rejection:

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

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 4, 5, and 6 as:

Lemma 9. *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 10 (Weak dependency of null features)**.** *Conditional on E_0, $\mathcal{D}_1$,*

$$\sup_{0 \leq t \leq L_n'} \frac{\sum_{(T_i, T_j) \in \mathcal{H}^2_{0\mathcal{A}, weak}} \left|\mathrm{cov}(\mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{Rank}} > t), \mathbb{I}(\widehat{\mathsf{w}}_i^{\mathsf{Rank}} > t))\right|}{q_{0n}^2 \widetilde{G}^2(t)} \leq C_1 \frac{C_\infty h_n q_{0n}}{\epsilon_n \eta_n'} + C_2 \frac{1}{(\epsilon_n \eta_n' q_{0n})^{v/2}}. \quad (54)$$

Lemma 11. *For any $\varepsilon > 0$, conditional on E_0 and $\mathcal{D}_1$, 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 9, 10, and 11 is same as that in Lemma 4, 5, and 6, and thus omitted. These lemmas imply that, if $L \leq L_n'$, then we have *Ratio* $\leq 1 + 3\varepsilon$ 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 4,

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

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

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

with probability at least $1-Cd_1^{-2}$. 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

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

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

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

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., $\mathbb{P}(\sum_{T\in\mathcal{S}} \mathbb{I}(\widehat{\mathsf{w}}_T^{\mathsf{Rank}} > L_n') \le (1-\varepsilon)\eta_n') \to 0$, $\mathbb{P}(\sum_{T\in\mathcal{S}} \mathbb{I}(\widehat{\mathsf{w}}_T^{\mathsf{Rank}} > L_n') \ge \eta_n') \to 1$. By taking $\epsilon_n = \alpha/8$, other steps essentially follow the proof of Theorem 4.

□

C Proofs of Auxiliary Results

C.1 Verification of (10)

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 (34). 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 $k \ge 2$, we have

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

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.,

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

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

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

and thus, we have

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

C.2 Proof of Theorem 8

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

$$\begin{aligned}
&\mathbb{P}(Y_1 > t | H_0) = \Psi(-t) = \tilde{\Psi}(t) \\
&F_1(z,t) := \mathbb{P}(Y_1 > t | z, H_1) = \mathbb{P}\left(\frac{(\xi_1 + \xi_2 + 2\delta)^2}{4} - \frac{(\xi_1 - \xi_2)^2}{4} > t\right) = \mathbb{P}\left(\frac{(Z_1 + \sqrt{2}\delta)^2}{2} - \frac{Z_2^2}{2} > t\right) \\
&\qquad = \int_{\mathbb{R}} \left[\tilde{\Phi}(\sqrt{2t + y_2^2} - \sqrt{\frac{2}{p}} z) + \tilde{\Phi}(\sqrt{2t + y_2^2} + \sqrt{\frac{2}{p}} z)\right] dy_2.
\end{aligned}$$

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 $t = L_{p1}$:

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

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

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

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

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

is non-vanishing.

For $j = 2$, we have

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

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

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

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

For $j = 3$, we have

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

Compute the derivatives of F_3:

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

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

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

Since

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

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. □

C.3 Proof of Lemma 1

Proof. To show this, we first state the perturbation of singular subspaces with respect to different norms:

Lemma 12 (Xia and Yuan (2021), Lemma 2)**.** *Under good initialization and signal requirements, there exists an absolute constant $C > 0$ such that conditional on E_0, if $n \geq Cd_1 \log d_1$, then with probability at least $1 - d_1^{-\tau}$,*

$$\left\|\widehat{Z}\right\| \leq C_2\sqrt{\tau}(1+\gamma_n)\sigma_\xi\sqrt{\frac{d_1^2 d_2 \log d_1}{n}},$$

and

$$\max\left\{\left\|UU^\top - \widehat{U}\widehat{U}^\top\right\|, \left\|VV^\top - \widehat{V}\widehat{V}^\top\right\|\right\} \leq C_2\frac{\sqrt{\tau}\,(1+\gamma_n)\,\sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{d_1^2 d_2 \log d_1}{n}}.$$

Lemma 13 (Xia and Yuan (2021), Theorem 4)**.** *Under good initialization and signal requirements, there exists an absolute constant $C > 0$ such that conditional on E_0, if $n \geq C\mu^2 r d_1 \log d_1$, then with probability at least $1 - 5\log d_1 \cdot d_1^{-\tau}$,*

$$\left\|UU^\top - \widehat{U}\widehat{U}^\top\right\|_{2,\max} \leq C_2\mu\frac{\sqrt{\tau}\,(1+\gamma_n)\,\sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{r d_1 d_2 \log d_1}{n}},$$

and

$$\left\|VV^\top - \widehat{V}\widehat{V}^\top\right\|_{2,\max} \leq C_2\mu\frac{\sqrt{\tau}\,(1+\gamma_n)\,\sigma_\xi}{\lambda_{\min}} \cdot \sqrt{\frac{r d_1^2 \log d_1}{n}}.$$

Then under the event when Lemma 12 and Lemma 13 both hold, it follows that

$$\left|\left\langle \widehat{U}\widehat{U}^\top\widehat{Z}\widehat{V}\widehat{V}^\top - UU^\top\widehat{Z}VV^\top, T\right\rangle\right| \leq \|T\|_{\ell_1}\left\|\widehat{U}\widehat{U}^\top\widehat{Z}\widehat{V}\widehat{V}^\top - UU^\top\widehat{Z}VV^\top\right\|_{\max},$$

and by triangular inequality,

$$
\begin{aligned}
\left\|\widehat{U}\widehat{U}^\top\widehat{Z}\widehat{V}\widehat{V}^\top - UU^\top\widehat{Z}VV^\top\right\|_{\max} \leq& \left\|(\widehat{U}\widehat{U}^\top - UU^\top)\widehat{Z}VV^\top\right\|_{\max} + \left\|UU^\top\widehat{Z}(\widehat{V}\widehat{V}^\top - VV^\top)\right\|_{\max} \\
&+ \left\|(\widehat{U}\widehat{U}^\top - UU^\top)\widehat{Z}(\widehat{V}\widehat{V}^\top - VV^\top)\right\|_{\max} \\
\leq& \left\|\widehat{Z}\right\|\left(\left\|\widehat{U}\widehat{U}^\top - UU^\top\right\|_{2,\max}\|V\|_{2,\max} + \left\|\widehat{V}\widehat{V}^\top - VV^\top\right\|_{2,\max}\|U\|_{2,\max}\right) \\
&+ \left\|\widehat{Z}\right\|\left\|\widehat{U}\widehat{U}^\top - UU^\top\right\|_{2,\max}\left\|\widehat{V}\widehat{V}^\top - VV^\top\right\|_{2,\max} \\
\leq& C_2\tau\mu^2\frac{\sigma_\xi}{\lambda_{\min}}\sqrt{\frac{rd_1^2d_2\log d_1}{n}}\cdot\sigma_\xi\sqrt{\frac{rd_1\log d_1}{n}}.
\end{aligned}
$$

Thus, we conclude that under the event with probability larger than $1 - d^{-\tau} - 5\log d_1 \cdot d_1^{-\tau} \geq 1 - 6d_1^{-\tau}\log d_1$, the desired bound holds. $\square$

C.4 Proof of Lemma 2

Proof. Denote I_1 the index set of observations in the sample $\mathcal{D}_1$. Since

$$
\begin{aligned}
&\left\langle UU^\top\widehat{Z}_2V_\perp V_\perp^\top + U_\perp U_\perp^\top\widehat{Z}_2VV^\top, T\right\rangle + \left\langle UU^\top\widehat{Z}_2VV^\top, T\right\rangle = \left\langle\widehat{Z}_2, \mathcal{P}_M(T)\right\rangle \\
&= \frac{d_1d_2}{n}\sum_{i\in I_1}\left\langle\widehat{\Delta}, X_i\right\rangle\langle X_i, \mathcal{P}_M(T)\rangle - \left\langle\widehat{\Delta}, \mathcal{P}_M(T)\right\rangle,
\end{aligned}
$$

and

$$
\begin{aligned}
&\left|\frac{d_1d_2}{n}\left\langle\widehat{\Delta}, X_i\right\rangle\langle X_i, \mathcal{P}_M(T)\rangle - \frac{1}{n}\left\langle\widehat{\Delta}, \mathcal{P}_M(T)\right\rangle\right| \\
&\leq \sqrt{3}\frac{\left\|\widehat{\Delta}\right\|_{\max}\mu\sqrt{rd_1^2d_2}\,\|\mathcal{P}_M(T)\|_{\mathrm{F}} + \sqrt{d_1d_2}\left\|\widehat{\Delta}\right\|_{\max}\|\mathcal{P}_M(T)\|_{\mathrm{F}}}{n}, \\
&\mathbb{E}\left|\frac{d_1d_2}{n}\left\langle\widehat{\Delta}, X_i\right\rangle\langle X_i, \mathcal{P}_M(T)\rangle - \frac{1}{n}\left\langle\widehat{\Delta}, \mathcal{P}_M(T)\right\rangle\right|^2 \leq \frac{d_1d_2\left\|\widehat{\Delta}\right\|_{\max}^2\|\mathcal{P}_M(T)\|_{\mathrm{F}}^2}{n^2},
\end{aligned}
$$

combined with initialization assumption and, by Bernstein inequality, we have

$$
\left|\frac{d_1d_2}{n}\sum_{i\in I_1}\left\langle\widehat{\Delta}, X_i\right\rangle\langle X_i, \mathcal{P}_M(T)\rangle - \left\langle\widehat{\Delta}, \mathcal{P}_M(T)\right\rangle\right| \leq C\gamma_n\sigma_\xi\|\mathcal{P}_M(T)\|_{\mathrm{F}}\sqrt{\frac{\tau d_1d_2\log d_1}{n}},
$$

with probability at least $1 - 2d^{-\tau}$. Here $n \geq C\mu^2rd_1\log d_1$. $\square$

C.5 Proof of Lemma 7

Proof. Notice that both $\left(I_{d_1d_2} - \widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top\right)$ and $\left(I_{d_1d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top\right)$ are projection matrices with $P = P^2$. We thus have

$$
\begin{aligned}
\widehat{\Sigma} - \Sigma &= T_{\mathcal{H}} \left((I_{d_1d_2} - \widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top) - (I_{d_1d_2} - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top) \right) T_{\mathcal{H}}^\top \\
&= T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \\
&+ T_{\mathcal{H}} \left(I - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top \\
&+ T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) T_{\mathcal{H}}^\top.
\end{aligned} \tag{55}
$$

We apply (55) to the error $\Sigma^{-\frac{1}{2}}(\widehat{\Sigma} - \Sigma)\Sigma^{-\frac{1}{2}}$:

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

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

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

where we use the definition of κ_T and the perturbation of singular subspaces in Lemma 12. Moreover, when $T_{\mathcal{H}}$ is sparse, we use $e_{T,k} \in \mathbb{R}^{d_1 \times d_2}$, $k \in [\mathrm{supp}(T_{\mathcal{H}})]$ to indicate the collective supports of all the $\mathrm{vec}(T_i)$. We then have

$$
\begin{aligned}
&\left\| \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \right\| \\
&= \left\| \Sigma^{-\frac{1}{2}} T_{\mathcal{H}} \sum_{k=1}^{\mathrm{supp}(T_{\mathcal{H}})} e_{T,k} e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \right\| \\
&\le \sqrt{\kappa_1} \kappa_T \, \mathrm{supp}(T_{\mathcal{H}}) \max_k \left\| e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top \otimes \widehat{V}_\perp \widehat{V}_\perp^\top - U_\perp U_\perp^\top \otimes V_\perp V_\perp^\top \right) \right\| \\
&\le \sqrt{\kappa_1} \kappa_T \, \mathrm{supp}(T_{\mathcal{H}}) \max_k \left(\left\| e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top - U_\perp U_\perp^\top \right) \otimes V_\perp V_\perp^\top \right\| \right. \\
&\left. + \left\| e_{T,k}^\top U_\perp U_\perp^\top \otimes \left(\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top \right) \right\| + \left\| e_{T,k}^\top \left(\widehat{U}_\perp \widehat{U}_\perp^\top - U_\perp U_\perp^\top \right) \otimes \left(\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top \right) \right\| \right).
\end{aligned} \tag{56}
$$

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

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

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:

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

□

C.6 Proof of Lemma 8

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 (55), 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{\beta_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{\beta_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:

$$
\begin{aligned}
&\left\| D^{-1} T_{\mathcal{H}} \left(\widehat{U}_{\perp} \widehat{U}_{\perp}^{\top} \otimes \widehat{V}_{\perp} \widehat{V}_{\perp}^{\top} - U_{\perp} U_{\perp}^{\top} \otimes V_{\perp} V_{\perp}^{\top} \right) \right\|_{2,\max} \\
&\leq \max_{T_i \in \mathcal{H}} \frac{\sum_{j \in [d_1]} \sum_{k \in [d_2]} \left| T_i(j,k) \left[\left(UU^{\top} - \widehat{U}\widehat{U}^{\top} \right) \otimes V_{\perp} V_{\perp}^{\top} + U_{\perp} U_{\perp}^{\top} \otimes \left(\widehat{V}\widehat{V}^{\top} - VV^{\top} \right) \right] \cdot e_j \otimes e_k \right|}{s_{T_i}} \\
&\quad + \max_{T_i \in \mathcal{H}} \frac{\sum_{j \in [d_1]} \sum_{k \in [d_2]} \left| T_i(j,k) \left(UU^{\top} - \widehat{U}\widehat{U}^{\top} \right) \otimes \left(\widehat{V}\widehat{V}^{\top} - VV^{\top} \right) \cdot e_j \otimes e_k \right|}{s_{T_i}} \\
&\leq C \max_{T_i \in \mathcal{H}} \frac{\|T\|_{\ell_1}}{\|T\|_{\mathrm{F}} \beta_0 \sqrt{r/d_1}} \frac{\mu (1+\gamma_n) \sigma_{\xi}}{\lambda_{\min}} \cdot \sqrt{\frac{r d_1^2 \log d_1}{n}} \\
&\leq C \frac{\beta_T \mu \sigma_{\xi}}{\beta_0 \lambda_{\min}} \sqrt{\frac{\alpha_d d_1^2 d_2 \log d_1}{n}}.
\end{aligned}
\tag{57}
$$

Here, we use the 2-max norm bound in Lemma 13, 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 (56). This gives the desired bound

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

$\square$