Statistica Sinica Preprint No: SS-2021-0406	
Title	Large-Scale Multiple Testing for Matrix-Valued Data under Cross Dependency
Manuscript ID	SS-2021-0406
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0406
Complete List of Authors	Shiyu Zhang, Xu Han and Sanat K. Sarkar
Corresponding Authors	Sanat K. Sarkar
E-mails	sanat@temple.edu
Notice: Accepted version subject to English editing.	

Large-Scale Multiple Testing for Matrix-Valued Data under Cross-Dependency

Shiyu Zhang, Xu Han and Sanat K. Sarkar*

Temple University

Abstract:

High-dimensional inference based on matrix-valued data has drawn increasing attention in modern statistical research, yet not much progress has been made in large-scale multiple testing specifically designed for analyzing such data sets. Motivated by this, we consider in this article an electroencephalography (EEG) experiment that produces matrix-valued data and presents a scope of developing novel matrix-valued data-based multiple testing methods that are of importance in such an experiment. The row-column cross-dependency of observations appearing in a matrix form, referred to as cross-dependency, is one of the main challenges in the development of such methods. We address this challenge by assuming a matrix normal distribution for the observations at each of the independent matrix data points. This allows us to capture the underlying cross-dependency informed through the row- and column-covariance matrices and develop methods that are potentially better than the corresponding one obtained by vectorizing each data point and thus ignoring the cross-dependency. Given a fixed thresholding proce-

*Corresponding author. Email: sanat@temple.edu

dure with unknown cross covariance matrices, we consider approximating the false discovery proportion capturing the underlying cross-dependency with statistical accuracy, and propose two methods of doing so. While one of these methods is a general approach under cross-dependency, the other one provides more computational efficiency for higher dimensionality. Extensive numerical studies illustrate the superior performance of the proposed methods over the principal factor approximation method of Fan and Han (2017). The proposed methods have been further applied to the aforementioned EEG data.

Key words and phrases: matrix-valued data, large-scale multiple testing, false discovery proportion, cross dependency, electroencephalogram

1. Introduction

Large-scale multiple testing is an integral part of statistical investigations in the modern era of Big Data-driven scientific research with statisticians/data scientists frequently encountering simultaneous testing of tens of thousands or even hundreds of thousands of hypotheses in such research. Despite substantial growth of research in multiple testing over the past few decades (see, for instance, Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Sarkar (2002), Storey (2002), Efron (2007), Fan, Han and Gu (2012), Fan, Ke, Sun and Zhou (2019), etc), development of multiple testing methods specifically designed for matrix-valued data under row-

column cross dependency has not yet received much attention, even though such data have been increasingly seen to occur in various applications, for instance, in brain imaging, electroencephalography (EEG), environmental science, economics and many others.

The row-column cross-dependency of observations appearing in a matrix-structured form at each data point, which we refer to as cross-dependency in this article, is a newer challenge in developing a multiple testing method specifically designed for matrix-valued data. One can, of course, mitigate this challenge by vectorizing each data point and considering to use an appropriately chosen method, depending on the problem under consideration, from the abundant literature on vector-valued data-based multiple testing methods. However, such a method does not utilize the original matrix structure of the data, and so would be less desirable and potentially less powerful than the one that can be developed by capturing the underlying cross-dependency. So, there seems to be an urgent need to develop such matrix-valued data-based multiple-testing methods.

Driven by the aforementioned need, we revisit the matrix-valued data set from an EEG experiment, used by Li, Kim and Altman (2010) and many other researchers while developing newer statistical theories and methodologies for such data sets (see also Nandi, Sarkar and Chen (2021)). This data

set presents an opportunity for us to develop our desired novel multiple-testing tools, at least in the context of such an important scientific investigation. The EEG experiment involved a control group and a treatment group comprising alcoholic subjects. Ten trials were performed on each subject and a picture was presented to the subject during each trial, while EEG activity in the form of voltage fluctuations (in microvolts) was recorded at 256 time points from 64 electrodes placed on the subject's scalp. Figure 1 in Li, Kim and Altman (2010) shows an example of the EEG pattern, averaged over measurements obtained from ten trials, for two subjects, one each from the control and alcoholic groups. This figure indicates a difference in the voltage fluctuation patterns for the two subjects over time and electrodes, sparking our interest in developing novel multiple testing methods for comparing two groups that can potentially be applied to gain a deeper understanding of brain dysfunction and regions impacted by alcoholism. We develop such methods in the framework of approximating false discovery proportion (FDP) given a fixed threshold-based multiple testing procedure.

Substantial challenges do arise in developing multiple testing methods for the aforementioned type of matrix-valued data. First of all, the number of hypotheses is often excessively large, relative to the sample size, even

when each dimension of the matrix is not very high, since the product of the two dimensions can be "quadratically high". In the EEG data, for example, there are $64 \times 256 = 16384$ hypotheses to be tested when comparing the two groups. One naive way to handle this challenge would be to vectorize the matrix data by stacking the columns and apply the Principal Factor Approximation (PFA) method in Fan and Han (2017) to the vectorized data assuming vector-variate multivariate normal. That is, suppose we have a matrix format data $\mathbf{Y}_{p\times q}$, vectorizing the matrix data and assuming the multivariate normal distribution on $vec(\mathbf{Y})_{pq\times 1} \sim \mathcal{N}(vec(\boldsymbol{\mu}), \boldsymbol{\Sigma})$. The PFA relies on an estimate of the unknown covariance matrix $\boldsymbol{\Sigma}$, particularly through the eigenvalues and eigenvectors of this matrix. When the dimensionality is quadratically high and the sample size is relatively small, the performance of PFA in approximating the FDP will deteriorate. We will illustrate this issue in the numerical studies. The second challenge, as noted above, is the cross-dependency and its effective full utilization of our methods.

We will handle this challenge by assuming that the underlying random matrix of voltage observations, say $\mathbf{X}$, follows a matrix normal distribution; i.e., $\mathbf{X} \sim \mathcal{MN}(\boldsymbol{\mu}_{p\times q}, \mathbf{U}_{p\times p}, \mathbf{V}_{q\times q})$, where $\boldsymbol{\mu}_{p\times q}$ is the location matrix parameter, $\mathbf{U}$ and $\mathbf{V}$ are, respectively, the common covariance matrices of

the column and row vectors. The matrix normal is theoretically amenable to newer methodological developments specifically for matrix-valued data and so has been a commonly used distribution for analyzing such data sets; see, i.e., Li, Kim and Altman (2010) and Xia and Li (2017), respectively, for the development of multivariate regression with dimension folding and for brain connection testing. The matrix normal has also been used for analyzing microarray data (Allen and Tibshirani , 2012) and mRNA expression data (Horenstein, Fan, Shedden and Zhou , 2019). More importantly for our research, the underlying cross-dependency can be effectively parameterized through the row- and column-covariance matrices in this distribution.

In this paper, we propose two methods of approximating FDP for a fixed thresholding procedure based on matrix normal data assuming that the row- and column-covariance matrices are completely unknown with general dependence structures. In particular, we will present two types of extension of the work of Fan and Han (2017) from vector-valued to matrix-valued data. The first method, called the noodle method, utilizes the property of matrix normal distribution through the Kronecker product. More specifically, the vectorized matrix normal, having a multivariate normal distribution with the covariance matrix as the Kronecker product of the row- and column-covariance matrices, provides structural information about the

underlying cross-dependency and thus provides a dimension reduction advantage. This not only allows full capture of the cross-dependency informed through these covariance matrices but also facilitates the estimation of FDP in a large-scale multiple-testing setup. Instead of estimating a $(pq) \times (pq)$ dimensional covariance matrix for the vectorized data, we are estimating two smaller matrices: $p \times p$ dimensional column correlation matrix of $\mathbf{U}$ and $q \times q$ dimensional row correlation matrix of $\mathbf{V}$. Kronecker product has been considered an effective tool for dimension reduction. See other applications of Kronecker product in Liu et al. (2019) and Chen et al. (2023)

Although the noodle method shows superior performance for matrix data in comparison with the PFA procedure, it suffers from some computational complexity issues. More specifically, in the first method, we need to calculate any pair of the eigenvalues and eigenvectors from the two estimated correlation matrices of $\mathbf{U}$ and $\mathbf{V}$. When p and q are large, the noodle method is computationally intensive. To circumvent this issue, we propose the second method, the sandwich method, which involves the first few principal components from the correlation matrices of $\mathbf{U}$ and $\mathbf{V}$, respectively, mostly capturing the underlying dependence structure. The sandwich method is developed to handle a large number of tests, much larger than when the noodle method can be used. Our simulation studies

will show that the sandwich method can be applied to a more ambitious setting where $p \times q = 500 \times 500 = 250000$, where the noodle method fails.

The rest of the paper is organized as follows. In Section 2, we describe the two proposed methods and their theoretical underpinnings. Section 3 provides the results of simulation studies we conducted to compare these methods with the PFA method in Fan and Han (2017) under various scenarios. Section 4 presents the results obtained from the analysis of the EEG data using these methods. All technical proofs are relegated to the Supplementary Materials.

2. Main Results

Our proposed methods will be presented in this section. First, let us introduce below some of the notations to be used throughout this paper.

- $a_n \asymp b_n : 0 < a_n/b_n + b_n/a_n = O(1)$.
- For a vector $\mathbf{x} = (x_1, \cdots, x_p)'$, $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^p x_i^2}$, $\|\mathbf{x}\|_1 = \sum_{i=1}^p |x_i|$.
- For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$, Frobenius norm: $\|\mathbf{A}\|_F = \sqrt{\text{trace}(\mathbf{A}'\mathbf{A})}$; Operator norm: $\|\mathbf{A}\| = \lambda_{\max}^{1/2}(\mathbf{A}^T\mathbf{A})$; l_1 norm: $\|\mathbf{A}\|_1 = \max_{1\le j\le n}\sum_{i=1}^m |a_{i,j}|$; and l_∞ norm: $\|\mathbf{A}\|_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$.
- For a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, vec$(A)$ is the mn-dimensional column vector

obtained by stacking its columns.

- For two matrices $\mathbf{A}_{n\times m} = (a_{ij})$ and $\mathbf{B}_{p\times q} = (b_{ij})$,

  Kronecker product of $\mathbf{A}$ and $\mathbf{B}$: $\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1m}\mathbf{B} \\ \vdots & & \vdots \\ a_{n1}\mathbf{B} & \cdots & a_{nm}\mathbf{B} \end{pmatrix}$.

2.1 Basic Setup

With $\mathbf{Y}$ and $\mathbf{Z}$ representing the $p \times q$ dimensional random matrices associated with the treatment and control groups, let $\mathbf{Y} \sim \mathcal{MN}(\boldsymbol{\mu}_y, \mathbf{U}, \mathbf{V})$ and $\mathbf{Z} \sim \mathcal{MN}(\boldsymbol{\mu}_z, \mathbf{U}, \mathbf{V})$, with unknown mean matrices $\boldsymbol{\mu}_y = (\boldsymbol{\mu}_{y,ij})$ and $\boldsymbol{\mu}_z = (\boldsymbol{\mu}_{z,ij})$ and same but unknown common column- and row-covariance matrices $\mathbf{U}$ and $\mathbf{V}$, respectively, in each of $\mathbf{Y}$ and $\mathbf{Z}$. Our problem is to test

$$H_{0,ij} : \boldsymbol{\mu}_{y,ij} - \boldsymbol{\mu}_{z,ij} = 0 \quad \text{against} \quad H_{1,ij} : \boldsymbol{\mu}_{y,ij} - \boldsymbol{\mu}_{z,ij} \neq 0, \tag{2.1}$$

simultaneously for $(i, j) = (1, 1), \ldots, (p, q)$, based on independent samples, $(\mathbf{Y}_1 = (\mathbf{Y}_{1,ij}), \ldots, \mathbf{Y}_n = (\mathbf{Y}_{n,ij}))$ and $(\mathbf{Z}_1 = (\mathbf{Z}_{1,ij}), \ldots, \mathbf{Z}_m = (\mathbf{Z}_{m,ij}))$, of matrix-valued observations on $\mathbf{Y}$ and $\mathbf{Z}$, respectively.

Let $\overline{\mathbf{Y}} = n^{-1} \sum_{l=1}^{n} \mathbf{Y}_l$ and $\overline{\mathbf{Z}} = m^{-1} \sum_{k=1}^{m} \mathbf{Z}_k$ be the sample mean matrices corresponding to the treatment and control groups. If $\mathbf{U}$ and $\mathbf{V}$ were known, $H_{0,ij}$ would have been tested marginally against $H_{1,ij}$ using the

(i, j)th entry of the following matrix-valued statistic:

$$\begin{aligned}\widetilde{\mathbf{X}} &= \sqrt{\frac{nm}{n+m}}[\mathrm{Diag}(\mathbf{U})]^{-\frac{1}{2}}\left(\overline{\mathbf{Y}}-\overline{\mathbf{Z}}\right)[\mathrm{Diag}(\mathbf{V})]^{-\frac{1}{2}} \\ \sim\ & \mathcal{MN}\left(\sqrt{\frac{nm}{n+m}}[\mathrm{Diag}(\mathbf{U})]^{-\frac{1}{2}}\left(\boldsymbol{\mu}_y-\boldsymbol{\mu}_z\right)[\mathrm{Diag}(\mathbf{V})]^{-\frac{1}{2}}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2\right) \quad (2.2)\end{aligned}$$

where $\mathrm{Diag}(\mathbf{U})$ and $\mathrm{Diag}(\mathbf{V})$ are the diagonal matrices based on the main diagonals of, and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the correlation matrices obtained from, $\mathbf{U}$ and $\mathbf{V}$, respectively.

Suppose the standard deviation of $\mathbf{Y}_{l,ij}$ or $\mathbf{Z}_{l,ij}$ is σ_{ij}, denote $\boldsymbol{\Sigma} = (\sigma_{ij}^{-1})$ as a matrix with the (i, j)th element as σ_{ij}^{-1}, then equivalently we can express $\widetilde{\mathbf{X}}$ as

$$\widetilde{\mathbf{X}} = \sqrt{\frac{nm}{n+m}}(\overline{\mathbf{Y}}-\overline{\mathbf{Z}})\circ\boldsymbol{\Sigma} \sim \mathcal{MN}(\sqrt{\frac{nm}{n+m}}(\boldsymbol{\mu}_y-\boldsymbol{\mu}_z)\circ\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2). \quad (2.3)$$

In (2.3), the notation "$\circ$" is Hadamard product, which means element wise product for the matrices. However, in practice σ_{ij} is unknown. Correspondingly, we can use pooled estimator constructed based on the two groups. More specifically, denote the sample mean $\overline{\mathbf{Y}}_{ij} = n^{-1}\sum_{l=1}^{n}\mathbf{Y}_{l,ij}$, and $\overline{\mathbf{Z}}_{ij} = m^{-1}\sum_{k=1}^{m}\mathbf{Z}_{k,ij}$, then

$$\widehat{\sigma}_{ij}^2 = \frac{1}{n+m-2}\Big\{\sum_{l=1}^{n}(\mathbf{Y}_{l,ij}-\overline{\mathbf{Y}}_{ij})^2 + \sum_{k=1}^{m}(\mathbf{Z}_{k,ij}-\overline{\mathbf{Z}}_{ij})^2\Big\}. \quad (2.4)$$

For the unknown marginal variances σ_{ij}, we denote $\widehat{\boldsymbol{\Sigma}}$ as a matrix with the (i, j)th element as $\widehat{\sigma}_{ij}^{-1}$, where each $\widehat{\sigma}_{ij}$ is defined in (2.4). We will

consider the $p \times q$ dimensional matrix

$$\mathbf{X} \equiv \sqrt{\frac{nm}{n+m}}(\overline{\mathbf{Y}} - \overline{\mathbf{Z}}) \circ \widehat{\mathbf{\Sigma}}$$

for the test statistics in this paper.

The matrix $\mathbf{X}$ is the basic ingredient for the development of our proposed methods. Let $P_{ij} = 2\Phi(-|X_{ij}|)$ be the p-value corresponding to X_{ij}, the (i,j)th entry of $\mathbf{X}$, where Φ is the cdf of $\mathcal{N}(0,1)$, and $\{P_{ij} \leq t\}$ be the rejection region for testing $H_{0,ij} : \mu_{ij} = 0$ against $H_{1,ij} : \mu_{ij} \neq 0$, given a fixed threshold t. Then, our methods are aimed at approximating the realized value of

$$\mathrm{FDP}(t) = \frac{\sum_{\{(i,j):H_{0,ij} \text{ is true}\}} \mathbf{I}(P_{ij} \leq t)}{\max(R(t), 1)},$$

where $\mathbf{I}$ is the indicator function, $R(t) = \sum_{ij} \mathbf{I}(P_{ij} \leq t)$ is the total number of rejections, in a manner that extends the work of Fan and Han (2017) from vector- to matrix-valued data. More specifically, the methods provide approximation expressions for $\mathrm{FDP}(t)$ extracting out the strong cross-dependency exhibited through $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, before estimating the unknown quantities involving these correlation matrices in such approximations from their estimates $\widehat{\mathbf{\Sigma}}_1$ and $\widehat{\mathbf{\Sigma}}_2$, respectively. We will estimate $\mathbf{\Sigma}_1$ by $\widehat{\mathbf{\Sigma}}_1$ and

$\boldsymbol{\Sigma}_2$ by $\widehat{\boldsymbol{\Sigma}}_2$ as follows:

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_1 &= (n+m-2)^{-1}q^{-1}\{\sum_{l=1}^{n}((\mathbf{Y}_l-\overline{\mathbf{Y}})\circ\widehat{\boldsymbol{\Sigma}})((\mathbf{Y}_l-\overline{\mathbf{Y}})\circ\widehat{\boldsymbol{\Sigma}})^T \\
&\quad +\sum_{k=1}^{m}((\mathbf{Z}_k-\overline{\mathbf{Z}})\circ\widehat{\boldsymbol{\Sigma}})((\mathbf{Z}_k-\overline{\mathbf{Z}})\circ\widehat{\boldsymbol{\Sigma}})^T\}, \\
\widehat{\boldsymbol{\Sigma}}_2 &= (n+m-2)^{-1}p^{-1}\{\sum_{l=1}^{n}((\mathbf{Y}_l-\overline{\mathbf{Y}})\circ\widehat{\boldsymbol{\Sigma}})^T((\mathbf{Y}_l-\overline{\mathbf{Y}})\circ\widehat{\boldsymbol{\Sigma}}) \\
&\quad +\sum_{k=1}^{n}((\mathbf{Z}_k-\overline{\mathbf{Z}})\circ\widehat{\boldsymbol{\Sigma}})^T((\mathbf{Z}_k-\overline{\mathbf{Z}})\circ\widehat{\boldsymbol{\Sigma}})\}.
\end{aligned}$$

These are pooled sample correlation estimators. For diverging p and q, $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ are not necessarily consistent estimates of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. However, we will show that for FDP approximation, $\widehat{\boldsymbol{\Sigma}}_1$ and $\widehat{\boldsymbol{\Sigma}}_2$ can still lead to good approximation results. The following two sub-sections elaborate on the development of these two methods.

It is worth noting that in (2.2), if we have $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$, estimates of $\mathbf{U}$ and $\mathbf{V}$, respectively, then we can also consider

$$\mathbf{X}^{\star} = \sqrt{\frac{nm}{n+m}}[\mathrm{Diag}(\widehat{\mathbf{U}})]^{-\frac{1}{2}}\left(\overline{\mathbf{Y}}-\overline{\mathbf{Z}}\right)[\mathrm{Diag}(\widehat{\mathbf{V}})]^{-\frac{1}{2}}$$

as the test statistics. The conventional procedure in multivariate analysis is to consider maximum likelihood estimators for $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ through iterative algorithms. However, it is difficult to derive convergence results of such $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$, where the convergence results will be important for our theoretical analysis of the FDP approximation. The representation in (2.3) and (2.4)

circumvents this issue. Instead of estimating the matrix $\mathbf{U}$ and $\mathbf{V}$, we are estimating σ_{ij}^2, the product of any pair of U_{ii} and V_{jj}, which are the diagonal elements of $\mathbf{U}$ and $\mathbf{V}$, respectively. For ease of presentation, we assume that the corresponding covariance matrices for the treatment group and the control group are the same. We will pursue the more general situation in our future research.

2.2 Noodle Method

If we vectorize $\mathbf{X}$ by stacking the column vectors denoted as $vec(\mathbf{X})$, then

$$vec(\mathbf{X}) = \mathbf{T}^{1/2} vec(\widetilde{\mathbf{X}}),$$

where $\mathbf{T}^{1/2}$ is a $(pq) \times (pq)$ dimensional diagonal matrix. The diagonal elements of $\mathbf{T}^{1/2}$ are a vectorized matrix with the (i, j)th element as $\sigma_{ij}/\widehat{\sigma}_{ij}$. By the property of matrix normal distribution, based on expression (2.3),

$$vec(\widetilde{\mathbf{X}}) \sim \mathcal{N}(\sqrt{\frac{nm}{n+m}} vec((\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \boldsymbol{\Sigma}), \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1), \tag{2.5}$$

where $vec(\widetilde{\mathbf{X}})$ is a pq dimensional column vector, the notation "$\otimes$" denotes the Kronecker product, and $\boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1$ is a $(pq) \times (pq)$ dimensional covariance matrix.

For the $vec(\widetilde{\mathbf{X}})$, since it follows a multivariate normal distribution, it can be connected with a factor model structure where the random errors

are weakly dependent. More specifically, applying eigenvalue decomposition to $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$, let $\theta_1, \cdots, \theta_{pq}$ be the non-increasing eigenvalues of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$, and $\boldsymbol{\rho}_1, \cdots, \boldsymbol{\rho}_{pq}$ be the corresponding eigenvectors. We further define $\mathbf{F} = (\sqrt{\theta}_1\boldsymbol{\rho}_1, \cdots, \sqrt{\theta}_h\boldsymbol{\rho}_h)$ for some appropriate positive integer value h, then $vec(\mathbf{X})$ can be expressed as

$$vec(\mathbf{X}) = \mathbf{T}^{1/2}(vec(\boldsymbol{\mu}^\star) + \mathbf{F}\mathbf{W} + \boldsymbol{\epsilon}), \tag{2.6}$$

where $\mathbf{W} \sim \mathcal{N}(0, \mathbf{I}_h)$, $\boldsymbol{\mu}^\star = \sqrt{\frac{nm}{n+m}}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \mathbf{\Sigma}$ for simplification, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sum_{i=h+1}^{pq} \theta_i \boldsymbol{\rho}_i \boldsymbol{\rho}_i^T)$. As shown in Fan, Han and Gu (2012), when h satisfies some regularity condition, $\boldsymbol{\epsilon}$ are weakly dependent. Due to the independence between the sample variance and sample mean, conditional on $\mathbf{T}^{1/2}$ and $\mathbf{W}$, $vec(\mathbf{X})$ are weakly dependent because of the covariance structure in $\boldsymbol{\epsilon}$. Therefore, we expect the proportion of falsely rejected hypothesis among all tests can be approximated by $(pq)^{-1}\sum_{l\in\{\text{true null}\}} P(P_l \le t|\mathbf{T}, \mathbf{W})$. Denote the diagonal elements of $\mathbf{T}^{1/2}$ as $\{\sqrt{T_l}\}_{l=1}^{pq}$. By plugging the definition of p-values and note that T_l concentrates on 1 with $var(T_l) \to 0$ as $n \to \infty$, we propose an approximation formula for FDP(t) by

$$\text{FDP}_{\text{oracle},1}(t) = \frac{1}{\max(R(t), 1)} \sum_{l\in\{\text{true null}\}} [\Phi(a_l(z_{t/2} + \zeta_l)) + \Phi(a_l(z_{t/2} - \zeta_l))],$$

where $a_l = (1 - \|\mathbf{f}_l\|^2)^{-1/2}$, $\zeta_l = \mathbf{f}_l^T\mathbf{W}$ and $\mathbf{f}_l^T$ is the lth row of $\mathbf{F}$. The following Proposition 1 shows that FDP$_{oracle,1}(t)$ is a good approximation

to the true FDP(t).

Proposition 1. *If* $(pq)^{-1}\sqrt{\theta_{h+1}^2+\cdots+\theta_{pq}^2} = O((pq)^{-\delta})$ *for some* $\delta > 0$, $R(t)^{-1} = O_p((pq)^{-(1-\zeta)})$ *for some* $\zeta \geq 0$, *then* $|FDP_{oracle,1}(t) - FDP(t)| = O_p((pq)^{\zeta}((pq)^{-\delta/2} + (n+m)^{-1/2})$.

When the number of tests increases, the number of total rejections tends to increase. Suppose $\zeta = 0$, the condition $R(t)^{-1} = O_p((pq)^{-(1-\zeta)})$ is simplified to be $R(t)^{-1} = O_p((pq)^{-1})$, i.e., the total rejection number increases at the order of pq as the total number of hypotheses increases. The ζ is introduced to allow more flexibility in the growth rate, so that $R(t)$ does not need to grow in the order of pq. However, the value of ζ should not be too large, as it will reduce the convergence rate in the FDP approximation.

Since we do not know which hypotheses are true nulls, $\text{FDP}_{\text{oracle},1}(t)$ can be approximated by

$$\text{FDP}_{A,1}(t) = \frac{1}{\max(R(t),1)}\sum_{l=1}^{pq}[\Phi(a_l(z_{t/2}+\zeta_l)) + \Phi(a_l(z_{t/2}-\zeta_l))].$$

Here $\text{FDP}_{A,1}(t)$ is an upper bound of $\text{FDP}_{\text{oracle},1}(t)$. When we assume sparse signals, these two quantities will be close.

Applying eigenvalue decomposition directly to a $(pq)\times(pq)$ dimensional matrix will be challenging. Fortunately, due to the properties of Kronecker

product, the eigenvalues and eigenvectors of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ in (2.3) can be constructed based on those of $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_1$. Let $\lambda_1, \cdots, \lambda_p$ be the non-increasing eigenvalues of $\mathbf{\Sigma}_1$, and $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_p$ be the corresponding eigenvectors. Let $\xi_1, \cdots, \xi_q$ be the non-increasing eigenvalues of $\mathbf{\Sigma}_2$ and $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_q$ be the corresponding eigenvectors. Then the eigenvalues of $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ are $\xi_j \times \lambda_i$ for $1 \le i \le p$ and $1 \le j \le q$, and the corresponding eigenvectors are $\boldsymbol{\gamma}_j \otimes \boldsymbol{\nu}_i$.

However, in practice, the correlation matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ in (2.3) are both unknown. Let $\widehat{\lambda}_1, \cdots, \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{\Sigma}}_1$, and $\widehat{\boldsymbol{\nu}}_1, \cdots, \widehat{\boldsymbol{\nu}}_p$ be the corresponding eigenvectors. Let $\widehat{\xi}_1, \cdots, \widehat{\xi}_q$ be the eigenvalues of $\widehat{\mathbf{\Sigma}}_2$, and $\widehat{\boldsymbol{\gamma}}_1, \cdots, \widehat{\boldsymbol{\gamma}}_q$ be the corresponding eigenvectors. For the eigenvalues $\{\widehat{\lambda}_i\}$ and $\{\widehat{\xi}_j\}$, we calculate the product of each possible pair to obtain the eigenvalues of $\widehat{\mathbf{\Sigma}}_2 \otimes \widehat{\mathbf{\Sigma}}_1$, and arrange these values in a non-increasing order, written as $\{\widehat{\theta}_l\}$. Correspondingly, the eigenvectors of $\widehat{\mathbf{\Sigma}}_2 \otimes \widehat{\mathbf{\Sigma}}_1$ will be written as $\{\widehat{\boldsymbol{\rho}}_l\}$. For a given integer value h, we define $(pq) \times h$ dimensional matrix $\widehat{\mathbf{F}} = (\sqrt{\widehat{\theta}_1}\widehat{\boldsymbol{\rho}}_1, \cdots, \sqrt{\widehat{\theta}_h}\widehat{\boldsymbol{\rho}}_h)$. Given an experiment, $\mathbf{W}$ is a realized but unobserved vector. We will consider a least squares estimator of $\mathbf{W}$:

$$\widehat{\mathbf{W}} = (\widehat{\mathbf{F}}^T\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}^T vec(\mathbf{X}).$$

Then we can approximate the $\mathrm{FDP}_{A,1}(t)$ by

$$\widehat{\mathrm{FDP}}_1(t) = \frac{1}{\max(R(t), 1)} \sum_{l=1}^{p\times q} [\Phi(\widehat{a}_l(z_{t/2} + \widehat{\zeta}_l)) + \Phi(\widehat{a}_l(z_{t/2} - \widehat{\zeta}_l))],$$

where $\widehat{a}_l = (1 - \|\widehat{\mathbf{f}}_l\|^2)^{-1/2}$, $\widehat{\zeta}_l = \widehat{\mathbf{f}}_l^T (\widehat{\mathbf{F}}^T\widehat{\mathbf{F}})^{-1}\widehat{\mathbf{F}}^T vec(\mathbf{X})$ and $\widehat{\mathbf{f}}_l^T$ is the lth row of $\widehat{\mathbf{F}}$.

Theorem 1. *Under the conditions in Proposition 1, in addition,* $\theta_i - \theta_{i+1} \geq g_{pq}$ *with positive* $g_{pq} \asymp pq$ *for* $i = 1, \cdots, h$, $\{\widehat{a}_l\}_{l=1}^{pq}$ *and* $\{a_l\}_{l=1}^{pq}$ *are upper bounded, then* $|\widehat{FDP}_1(t) - FDP_{A,1}(t)| = O_p\big\{(pq)^{\zeta}\big(h(n+m)^{-1/2} + (pq)^{-1/2}h(n+m)^{-1/2}\|vec(\boldsymbol{\mu}^*)\|\big)\big\}$.

In Theorem 1, we require an eigengap condition for the largest h eigenvalues. Fan, Liao and Mincheva (2013) has shown that such condition can be satisfied for factor model structures. In the FDP approximation, the convergence rate also depends on the magnitude of signals, $\|vec(\boldsymbol{\mu}^\star)\|$. When we consider sparse signals in the mean matrix for the two group comparison, we expect that $(pq)^{-1/2}\|vec(\boldsymbol{\mu}^\star)\|$ converges to zero since ζ is a very small positive number.

To determine h, we can use the eigenvalue ratio estimator in Ahn and Horenstein (2013). The estimator is $\widehat{h} = \text{argmax}_{1\leq l\leq l_{\max}}(\widehat{\theta}_l/\widehat{\theta}_{l+1})$, where $l_{\max}$ is a pre-determined maximum possible number of factors. Ahn and Horenstein (2013) has shown that under mild regularity conditions, the eigenvalue ratio (ER) estimator is consistent for the true number of factors. However, this ER estimator may end up with selecting $\hat{h} > 0$. If the true $h = 0$, we may consider to test $H_0 : \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1 = \mathbf{I}_{pq\times pq}$ versus

$H_a : \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1 \neq \mathbf{I}_{pq \times pq}$. If $\mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ adopts independent or weakly dependent structure, we can simply assign $h = 0$ instead of using the estimator $\hat{h}$ from Ahn and Horenstein (2013). Correspondingly, the condition in Proposition 1 becomes $(pq)^{-1}\sqrt{\theta_{h+1}^2 + \cdots + \theta_{pq}^2} = (pq)^{-1/2}$. Our methods would be simplified as Storey procedure, where the estimator of FDP is $\hat{\pi}_0 t / R(t)$.

In practice, if we have a priori knowledge for the two correlation matrices, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, the regularity condition in Theorem 1 can be substantially relaxed. For example, if we know that $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are sparse matrices (Bickel and Levina , 2008), we can propose consistent thresholding estimators for $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Correspondingly, the eigengap condition in Theorem 1 can be relaxed to $d_{pq} \asymp d$ for $i = 1, \cdots, h$ where d is a constant. Under such scenario, we can still achieve the FDP approximation results in Theorem 1. Nevertheless, the FDP approximation that we proposed here is more general, especially designed for strong dependence scenarios.

We call the above described procedure as noodle method, as we cut the "dough" (matrix valued data) into slices (column vectors) and stick into a long "noodle".

2.3 Sandwich Method

In the noodle method, the approximation of FDP relies on the eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}_2 \otimes \widehat{\mathbf{\Sigma}}_1$. Note that we need to calculate the Kronecker product of each possible pair from $\{\widehat{\boldsymbol{\gamma}}_j\}$ and $\{\widehat{\boldsymbol{\nu}}_i\}$. When pq is large, this step can be very computationally intensive. The question is whether we can provide an alternative approximation procedure for FDP which is more computationally efficient. The key idea in the PFA method in Fan, Han and Gu (2012) is to use the first few principal components to capture the majority dependence among the test statistics. In our paper, the problem is somewhat different, because there are two covariance matrices, $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ for modeling the column and row dependence, respectively. If we can use the first few principal components from $\widehat{\mathbf{\Sigma}}_1$ and from $\widehat{\mathbf{\Sigma}}_2$ respectively to capture the majority dependence among the test statistics, the computation will be substantially simplified, and the corresponding procedure will be very appealing. This motivates us to propose the following method.

Note that in expression (2.3), $\widetilde{\mathbf{X}} \sim \mathcal{MN}(\boldsymbol{\mu}^\star, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$, where $\boldsymbol{\mu}^\star = \sqrt{\frac{nm}{n+m}}(\boldsymbol{\mu}_y - \boldsymbol{\mu}_z) \circ \mathbf{\Sigma}$ is defined to simplify the notation. Then for some positive integer values k_1 and k_2, we can rewrite $\widetilde{\mathbf{X}}$ as

$$\widetilde{\mathbf{X}} = \boldsymbol{\mu}^\star + \mathbf{C}\widetilde{\mathbf{W}}\mathbf{D} + \boldsymbol{\epsilon} \tag{2.7}$$

where $\mathbf{C} = (\sqrt{\lambda_1}\boldsymbol{\nu}_1, \cdots, \sqrt{\lambda_{k_1}}\boldsymbol{\nu}_{k_1})$, $\mathbf{D} = (\sqrt{\xi_1}\boldsymbol{\gamma}_1, \cdots, \sqrt{\xi_{k_2}}\boldsymbol{\gamma}_{k_2})^T$, $\widetilde{\mathbf{W}} \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_{k_1}, \mathbf{I}_{k_2})$, and $\boldsymbol{\epsilon} \sim \mathcal{MN}(\mathbf{0}, \sum_{i=k_1+1}^{p} \lambda_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^T, \sum_{j=k_2+1}^{q} \xi_j \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j^T)$.

Note that $\mathbf{C}$ contains the first k_1 principal components from $\boldsymbol{\Sigma}_1$, and $\mathbf{D}$ contains the first k_2 principal components from $\boldsymbol{\Sigma}_2$. Compared with the factor model structure in Fan, Han and Gu (2012), we have $\mathbf{C}$ and $\mathbf{D}$ here to capture the column and the row dependences, respectively. When k_1 and k_2 are appropriately chosen, the covariance matrices in $\boldsymbol{\epsilon}$ are both weakly dependent.

By the properties of Kronecker product, vectorizing expression (2.7) leads to

$$vec(\widetilde{\mathbf{X}}) = vec(\boldsymbol{\mu}) + (\mathbf{D}^T \otimes \mathbf{C}) vec(\widetilde{\mathbf{W}}) + vec(\boldsymbol{\epsilon}). \tag{2.8}$$

Similar to the discussion in section 2.3, we can propose an approximation for FDP(t) as

$$\text{FDP}_{\text{oracle},2}(t) = \frac{1}{\max(R(t), 1)} \sum_{l \in \{\text{true null}\}} [\Phi(d_l(z_{t/2} + \eta_l)) + \Phi(d_l(z_{t/2} - \eta_l))]$$

where $\eta_l = \mathbf{b}_l^T vec(\widetilde{\mathbf{W}})$, $\mathbf{b}_l$ is the lth row of $\mathbf{D}^T \otimes \mathbf{C}$, and $d_l = (1 - \|\mathbf{b}_l\|^2)^{-1/2}$. The following Proposition 2 shows that $\text{FDP}_{\text{oracle},2}$ is also a good approximation for the true FDP.

Proposition 2. *If* $p^{-1}\sqrt{\lambda_{k_1+1}^2 + \cdots + \lambda_p^2} = O(p^{-\delta_1})$ *for some* $\delta_1 > 0$ *and* $q^{-1}\sqrt{\xi_{k_2+1}^2 + \cdots + \xi_q^2} = O(q^{-\delta_2})$ *for some* $\delta_2 > 0$, $R(t)^{-1} = O_p((pq)^{-(1-\zeta)})$

for some $\zeta \geq 0$, *then* $|FDP_{oracle,2}(t) - FDP(t)| = O_p((pq)^{\zeta}(p^{-\delta_1/2}q^{-\delta_2/2} + (n+m)^{-1/2}))$.

Replacing the summation over true nulls in $\mathrm{FDP}_{\mathrm{oracle},2}(t)$ by all the tests, we have an upper bound as

$$\mathrm{FDP}_{A,2}(t) = \frac{1}{\max(R(t),1)} \sum_{l=1}^{pq} [\Phi(d_l(z_{t/2} + \zeta_l)) + \Phi(d_l(z_{t/2} - \zeta_l))].$$

If $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are known, we can estimate the realized $vec(\widetilde{\mathbf{W}})$ by least squares estimator

$$[(\mathbf{D}^T \otimes \mathbf{C})^T(\mathbf{D}^T \otimes \mathbf{C})]^{-1}(\mathbf{D}^T \otimes \mathbf{C})^T vec(\mathbf{X}).$$

For unknown $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, we use $\widehat{\mathbf{\Sigma}}_1$ and $\widehat{\mathbf{\Sigma}}_2$ for the estimation. Correspondingly, we replace $\mathbf{C}$ and $\mathbf{D}$ by $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{D}}$ respectively, where eigenvalues and eigenvectors are replaced by their estimates. Furthermore, we consider the FDP approximation formula:

$$\widehat{\mathrm{FDP}}_2(t) = \sum_{l=1}^{pq} [\Phi(\widehat{d}_l(z_{t/2} + \widehat{\eta}_l)) + \Phi(\widehat{d}_l(z_{t/2} - \widehat{\eta}_l))] / \max(R(t),1)$$

where $\widehat{d}_l = (1 - \|\widehat{\mathbf{b}}_l\|^2)^{-1/2}$, $\widehat{\mathbf{b}}_l$ is the lth row of $\widehat{\mathbf{D}}^T \otimes \widehat{\mathbf{C}}$, and $\widehat{\eta}_l$ is the lth element of $[(\sum_{i=1}^{k_2} \widehat{\boldsymbol{\gamma}}_i \widehat{\boldsymbol{\gamma}}_i^T) \otimes (\sum_{j=1}^{k_1} \widehat{\boldsymbol{\nu}}_j \widehat{\boldsymbol{\nu}}_j^T)] vec(\mathbf{X})$.

It is worth mentioning in $\widehat{\mathrm{FDP}}_2(t)$, we only need to calculate the Kronecker product of the first few eigenvectors from $\widehat{\mathbf{\Sigma}}_1$ and $\widehat{\mathbf{\Sigma}}_2$. This can avoid the computational issue in noodle method for large values of p and q, where Kronecker product has to be calculated for all possible pairs.

Theorem 2. *Under the conditions in Proposition 2, in addition,* $\lambda_i - \lambda_{i+1} \geq g_p$ *with* $g_p \asymp p$ *for* $i = 1, \cdots, k_1$ *and* $\xi_j - \xi_{j+1} \geq g_q$ *with* $g_q \asymp q$ *for* $j = 1, \cdots, k_2$, $\{\widehat{d}\}_{l=1}^{pq}$ *and* $\{d_l\}_{l=1}^{pq}$ *are upper bounded, then*

$$|\widehat{FDP}_2(t) - FDP_{A,2}(t)| = O_p\Big((pq)^{\zeta}\big(k_1k_2(n+m)^{-1} + (k_1+k_2)(n+m)^{-1/2} + (pq)^{-1/2}\|vec(\boldsymbol{\mu}^{\star})\|\big)\Big).$$

To determine k_1 and k_2, we consider the eigenvalue ratio estimator:

$$\widehat{k}_1 = \text{argmax}_{1\leq l\leq l_{\max}}(\widehat{\lambda}_l/\widehat{\lambda}_{l+1}), \qquad \widehat{k}_2 = \text{argmax}_{1\leq l\leq l_{\max}}(\widehat{\xi}_l/\widehat{\xi}_{l+1}).$$

As discussed in section 2.2, if we have a priori knowledge for the two correlation matrices, the eigengap condition in Theorem 2 can be relaxed. Also, both methods proposed above simply adopt the pooled sample correlation estimators, other estimators such as MLE-type estimators could also been used in practice, even though the theoretical results of convergence rate are difficult to be derived. The MLE algorithm for the matrix normal distribution has been well described in Dutilleuel (1999), we will show in simulations that how those two proposed methods perform by using different estimators of the correlation matrices in Supplementary Materials.

The key idea in the above procedure is to express the matrix data in terms of a "sandwich" formula (2.7), where the two matrices $\mathbf{C}$ and $\mathbf{D}$ "wrap" the common factor matrix $\mathbf{W}$. Thus, we call the proposed procedure

in section 2.3 "sandwich method".

3. Simulation Studies

In our simulation studies, the treatment group data are generated from $\mathbf{Y}_i \sim \mathcal{MN}(\boldsymbol{\mu}, \mathbf{\Sigma_1}, \mathbf{\Sigma_2}), i = 1, ..., n$ and the control group data are generated from $\mathbf{Z}_j \sim \mathcal{MN}(\mathbf{0}, \mathbf{\Sigma_1}, \mathbf{\Sigma_2}), j = 1, ..., m$. The signal strength $\boldsymbol{\mu}$ equals 1 for the first 8 rows out of p rows, the first 25 columns out of q columns, and 0 otherwise. We consider sample size $n = m = 50$, dimensionality $p = q = 100$ for both the noodle method and sandwich method unless stated otherwise, threshold value $t = 0.001$, and the number of simulation rounds to be 500. We estimate the unknown number of factors by the data-driven eigenvalue ratio method for the noodle method and sandwich method, both with $k_{max} = \lfloor 0.2(n+m) \rfloor$. We now examine the performance of our two methods on simulated data sets, which are constructed under the framework of three models.

- [**Model 1: strict factor model**] Let $\mathbf{B}_1$ be a $p \times l_1$ dimensional matrix with each element generated from a distribution $\mathcal{F}_1$, Σ_{u1} be a $p \times p$ dimensional diagonal matrix with all diagonal values being 0.5, then $\mathbf{\Sigma}_1$ is the correlation matrix of $\mathbf{B}_1\mathbf{B}_1^T + \Sigma_{u1}$; Similarly, let $\mathbf{B}_2$ be a $q \times l_2$ dimensional matrix with each element generated from

a distribution $\mathcal{F}_2$, Σ_{u2} be a $q \times q$ dimensional diagonal matrix with all diagonal values being 0.5, then $\mathbf{\Sigma}_2$ is the correlation matrix of $\mathbf{B}_2\mathbf{B}_2^T + \Sigma_{u2}$. We consider two cases: $l_1 = l_2 = 3$, $\mathcal{F}_1$ and $\mathcal{F}_2$ are both $\mathcal{N}(0,1)$; $l_1 = 2, l_2 = 4$, $\mathcal{F}_1$ and $\mathcal{F}_2$ are both $U(-1,1)$. In Model 1, both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ possess some strict factor model structures.

- [**Model 2: approximate factor model**] We keep the similar setting in Model 1, but consider $\mathbf{\Sigma}_{u1}$ to be a $p \times p$ dimensional power decay matrix with ρ_1, where the (i,j)th element of $\mathbf{\Sigma}_{u1}$ is defined as $\rho_1^{|i-j|}$. Similarly, let $\mathbf{\Sigma}_{u2}$ be a $q \times q$ dimensional power decay matrix with ρ_2. In Model 2, we consider $l_1 = l_2 = 3$, and $\mathcal{F}_1$, $\mathcal{F}_2$ are both $U(-1,1)$. We examine two settings: $(\rho_1, \rho_2) = (0.5, 0.3)$ and $(\rho_1, \rho_2) = (0.5, 0.8)$. In Model 2, both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ possess some approximate factor model structures, which can be used for testing the robustness of the eigenvalue ratio estimator for the unknown number of factors under the matrix normal settings.

- [**Model 3: normality violated model**] We keep the similar setting in model 1, where both $\mathcal{F}_1$ and $\mathcal{F}_2$ are $U(-1,1)$. Let $(\lambda_1, ..., \lambda_p)$ and $(\nu_1, ...\nu_p)$ be the eigenvalues and the corresponding eigenvectors of $\mathbf{\Sigma}_1$, respectively. Let $(\xi_1, ..., \xi_q)$ and $(\gamma_1, ...\gamma_q)$ be the eigenval-

ues and the corresponding eigenvectors of $\mathbf{\Sigma_2}$, respectively. Define $\widetilde{\mathbf{C}} = (\sqrt{\lambda_1}\nu_1, ..., \sqrt{\lambda_p}\nu_p)$, and $\widetilde{\mathbf{D}} = (\sqrt{\xi_1}\gamma_1, ..., \sqrt{\xi_q}\gamma_q)$. Then the data matrix $\mathbf{X}$ is generated from $\mathbf{X} = \mu + \widetilde{\mathbf{C}}\mathbf{W}\widetilde{\mathbf{D}}$. In our simulation studies, we consider the following settings: $(l_1, l_2) = (2, 2), (3, 3), (4, 4), (2, 4)$ as the choices for $\mathbf{B}_1$ and $\mathbf{B}_2$; $\mathbf{W}$ is a $p \times q$ dimensional matrix with each element randomly generated from $\sqrt{\frac{2}{3}}t_6$ distribution, or exponential distribution with $\lambda = 1$. Model 3 is designed to test the performance of our proposed methods when the matrix normality assumption is violated.

We will compare our newly proposed methods with the PFA method in Fan and Han (2017). The PFA method was originally designed for vector data from multivariate normal distribution. In the current paper, we are dealing with matrix variated data $\mathbf{Y}$ and $\mathbf{Z}$. To apply the PFA method, we vectorize $\mathbf{Y}$ and $\mathbf{Z}$ to obtain $vec(\mathbf{Y})$ and $vec(\mathbf{Z})$, and assume $vec(\mathbf{Y}) \sim \mathcal{N}(vec(\boldsymbol{\mu}), \mathbf{\Sigma})$, $vec(\mathbf{Z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. We need to estimate the $(pq) \times (pq)$ dimensional covariance matrix $\mathbf{\Sigma}$ based on the sample data $\{vec(\mathbf{Y}_i)\}_{i=1}^n$ and $\{vec(\mathbf{Z}_j)\}_{j=1}^m$. Suppose we consider the pooled sample covariance matrix as our estimator,

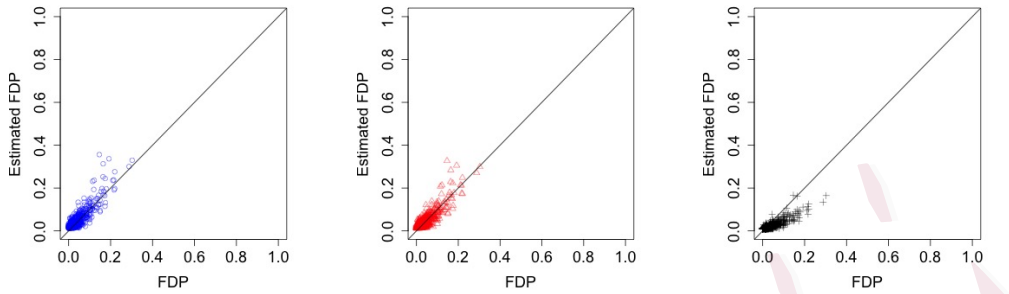
then

$$S = \frac{1}{n+m-2}\Big\{\sum_{i=1}^{n}(vec(\mathbf{Y}_i)-\overline{vec(\mathbf{Y})})(vec(\mathbf{Y}_i)-\overline{vec(\mathbf{Y})})^T + \sum_{j=1}^{m}(vec(\mathbf{Z}_j)-\overline{vec(\mathbf{Z})})(vec(\mathbf{Z}_j)-\overline{vec(\mathbf{Z})})^T\Big\}$$

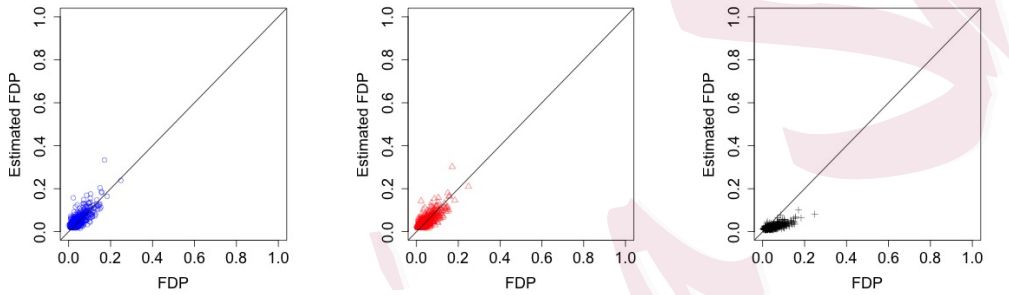
where $\overline{vec(\mathbf{Y})}$ and $\overline{vec(\mathbf{Z})}$ are the sample means of $vec(\mathbf{Y}_1), ..., vec(\mathbf{Y}_n)$ and $vec(\mathbf{Z}_1), ..., vec(\mathbf{Z}_m)$, respectively. To apply the PFA method, we need to get the eigenvalues and eigenvectors of S. However, since $p \times q$ is large, it will be time-consuming to directly apply eigenvalue decomposition to S. Instead, we consider

$$F = \sqrt{\frac{1}{n+m-2}}(\mathbf{Y}_1^{\nu}, ..., \mathbf{Y}_n^{\nu}, \mathbf{Z}_1^{\nu}, ..., \mathbf{Z}_m^{\nu}) \tag{3.9}$$

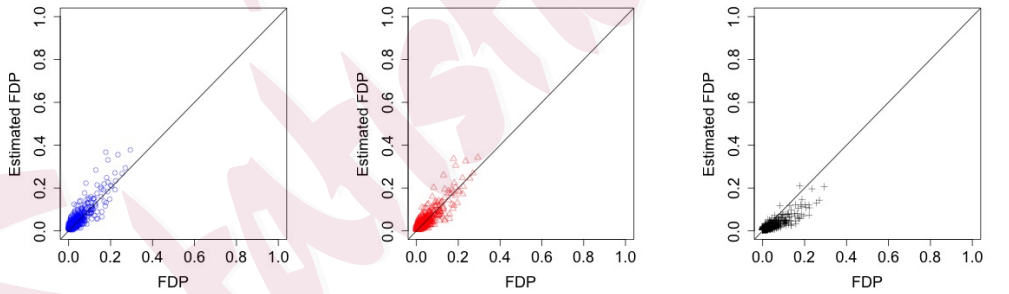
where $\mathbf{Y}_i^{\nu} = vec(\mathbf{Y}_i) - \overline{vec(\mathbf{Y})}$ and $\mathbf{Z}_j^{\nu} = vec(\mathbf{Z}_j) - \overline{vec(\mathbf{Z})}$ for $i = 1, ..., n, j = 1, ..., m$. Then clearly, $S = FF^T$. Eigenvalue decomposition of F will provide the eigenvalues and eigenvectors of S. Such construction would reduce the computation complexity.



(a) $f(2,4), \mathbf{B} \sim U(-1,1)$ (b) $f(2,4), \mathbf{B} \sim U(-1,1)$ (c) $f(2,4), \mathbf{B} \sim U(-1,1)$



(e) $(\rho_1, \rho_2) = (0.5, 0.3)$ (f) $(\rho_1, \rho_2) = (0.5, 0.3)$ (g) $(\rho_1, \rho_2) = (0.5, 0.3)$



(h) $f(3,3), \mathbf{W} \sim Exp(1)$ (i) $f(3,3), \mathbf{W} \sim Exp(1)$ (j) $f(3,3), \mathbf{W} \sim Exp(1)$

Figure 1: The estimated values of FDP obtained by the noodle method (blue circle), sandwich method (red triangle), and PFA (black crossover) are compared with the true value of FDP. From top to bottom, each row corresponds to Model 1, Model 2, and Model 3. Here, $n = m = 50, p = q = 100$, and $t = 0.001$.

We compare the estimated false discovery proportion from both of our proposed methods and PFA with the true value of the false discovery proportion. The results are summarized in Figure 1, Supplementary Materials Figures S1-S6, and Table 1. In Figure 1, points closer to the diagonal line suggest a good approximation. Under various settings, both of our proposed methods produce points slightly above the diagonal line, while the points from PFA are generally under the diagonal one. This phenomenon is further confirmed by the results in Table 1, where we calculate the mean difference between the estimated FDP and the true FDP over the 500 simulation rounds. Table 1 shows that our new estimator performs better than the PFA estimator in the sense that, the PFA estimator dramatically underestimates the true FDP, while our new method consistently overestimates the true FDP a little bit. That means, our new estimator can provide an upper bound for estimating FDP, which is meaningful in practice. It is worth mentioning that both the noodle method and the sandwich method perform roughly the same, but the sandwich method is much more computationally efficient. For a much more challenging setting: $p = q = 500$, which tests $500 \times 500 = 250,000$ hypotheses simultaneously, the results are summarized in Supplementary Materials Figures S7-S12 and Table S1. Note that the noodle method fails under this setting due to the computational

complexity. Our sandwich method still performs well for approximating the true FDP, while PFA underestimates the true value.

Table 1: Mean and standard deviation of $\widehat{\text{FDP}}(t) - \text{FDP}(t)$ are presented in percent, with $n = m = 50$, $p = q = 100$ and $t = 0.001$.

Results (%) for the following methods:						
$n = m = 50$	Noodle Method		Sandwich Method		PFA	
$p = q = 100$	Bias	Sd	Bias	Sd	Bias	Sd
Model 1						
$f(2,4), B \sim U(-1,1)$	0.981	2.640	0.437	2.415	-1.724	2.844
$f(3,3), B \sim N(0,1)$	1.049	3.732	0.653	3.428	-0.689	2.795
Model 2						
$(\rho_1, \rho_2) = (0.5, 0.3)$	1.008	2.537	0.481	2.366	-2.064	2.691
$(\rho_1, \rho_2) = (0.5, 0.8)$	0.914	3.044	0.372	2.949	-2.150	3.703
Model 3						
$f(2,2), W \sim Exp(1)$	0.986	2.556	0.515	2.396	-0.829	2.666
$f(2,2), W \sim \sqrt{\frac{2}{3}} t_6$	0.734	2.585	0.292	2.462	-0.863	2.603
$f(2,4), W \sim Exp(1)$	1.041	2.950	0.488	2.719	-1.669	2.912
$f(2,4), W \sim \sqrt{\frac{2}{3}} t_6$	0.808	2.558	0.293	2.456	-1.660	3.015
$f(3,3), W \sim Exp(1)$	0.924	2.861	0.345	2.636	-1.956	2.994
$f(3,3), W \sim \sqrt{\frac{2}{3}} t_6$	0.910	2.515	0.368	2.344	-1.715	2.712
$f(4,4), W \sim Exp(1)$	1.077	2.639	0.457	2.459	-2.740	3.271
$f(4,4), W \sim \sqrt{\frac{2}{3}} t_6$	0.995	2.465	0.423	2.285	-2.526	2.963

4. Data Analysis

Electroencephalogram (EEG) has been widely considered as an effective approach for detecting spontaneous fluctuations in brain activity. We will

illustrate our newly proposed multiple testing procedures on an EEG data set from a study to examine EEG correlating of genetic predisposition to alcoholism. We will compare the alcoholic group and control group, to study the influence of alcohol on the brain. For each subject, the data contains measurements from 64 electrodes placed on the scalp sampled at 256 Hz for 1 second, while the subjects were performing a visual object recognition task. The data set and the more detailed description can be accessed via *kdd.ics.uci.edu/databases/eeg*.

In our study, let $\mathbf{Y}_1, ..., \mathbf{Y}_n, n = 77$, denote the voltage (in microvolts) for the group of alcoholic subjects, and $\mathbf{Z}_1, ..., \mathbf{Z}_m, m = 45$, denote the voltage for the group of control subjects. Thus, each sample contains $p \times q = 64 \times 256 = 16384$ values of measurements. We further assume that the voltage of the two groups on each subject is from two matrix normal distributions with possibly different mean matrix but the same column covariance matrix $\mathbf{U}$ and row covariance matrix $\mathbf{V}$. More specifically, $\mathbf{Y}_i \sim \mathcal{MN}(\mu^y, \mathbf{U}, \mathbf{V})$ for $i = 1, ..., 77$ and $\mathbf{Z}_j \sim \mathcal{MN}(\mu^z, \mathbf{U}, \mathbf{V})$ for $j = 1, ..., 45$. We applied the empirical bootstrap procedure of Aston et al. (2017) to check the plausibility of the Kronecker product dependence decomposition for the covariance structure, the p-values of the test for alcoholic group $\{\mathbf{Y}_l\}_{l=1}^{77}$ and control group $\{\mathbf{Z}_l\}_{k=1}^{45}$ are 0.872 and 0.394,

respectively, thus we could not reject the null hypothesis that the data conforms with a Kronecker product structure. We also test the equality of the Kronecker product of dependencies across two groups using the method proposed by Cai, Liu and Xia (2013). The details are presented in Supplementary Materials. Furthermore, there are other procedures designed for matrix-valued data, to test the equality of dependency across two groups, see Xia and Li (2017), Xia and Li (2019), and Chen et al. (2023).

In our problem, testing whether EEG correlating of genetic predisposition to alcoholism can be formulated as a multiple hypothesis testing problem on $H_{0,ij} : \mu^y_{ij} = \mu^z_{ij}$ versus $H_{1,ij} : \mu^y_{ij} \neq \mu^z_{ij}$ for $i = 1, ..., 64, j = 1, ..., 256$. The EEG data reflects the brain's electric activity in a spatial-temporal pattern, where the dependence from either direction should not be simply ignored. The temporal correlation is easier to understand, as the activities of the same brain regions are recorded through time. The spatial correlation has a deeper scientific foundation, reflecting the brain functional connectivity (Fox and Raichle , 2007). Figure 2 shows the results of selected signals, by rejecting the hypotheses where the corresponding p-values are no greater than a threshold value t. Both the noodle method and sandwich method return the same results here. By choosing a larger threshold value, more signals will be detected, whereas a smaller threshold value will lead to

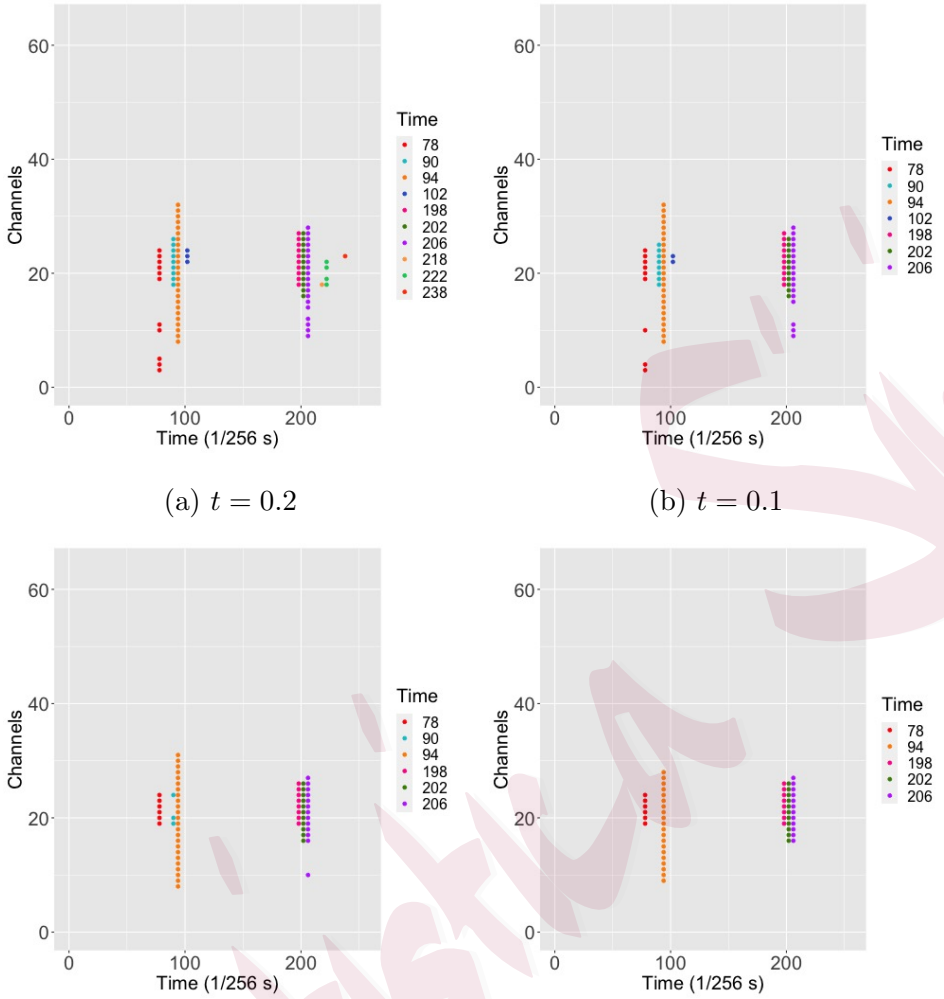
fewer discoveries. Although channels discovered are different according to different threshold values, the time when a signal is discovered is relatively stable. It is also interesting that the time lag between signals may reflect the causal effect or the direction of influence of the regional brain activities in response to the task.

5. Concluding Remarks

Our methods have shed some new light on the significance detection for the large-scale two-sample comparison. Extensions in multiple directions are possible for future research.

In the current paper, the methods are based on a sample covariance matrix estimator. This approach applies to a general scenario. When a priori knowledge of the brain function connectivity is available, we can choose a better estimator for the covariance matrices, and the corresponding FDP approximation can be further improved. For the more general non-normal data, we can also consider the robust covariance estimation in Fan, Wang and Zhong (2019).

For matrix normal assumption, it may be violated in practice, even if each element in the matrix data follows a normal distribution. One possibility is to detect significance by rows, and then by columns. The final



(a) $t = 0.2$ (b) $t = 0.1$

(c) $t = 0.01$ (d) $t = 0.001$

Figure 2: Plots of the selected hypotheses with different threshold values t: (a) 0.2 (b) 0.1 (c) 0.01 (d) 0.001

selection can be the intersection of row selection and column selection.

For the distribution of the matrix data, we could also consider relaxing the normality assumption but keeping the Kronecker product prop-

erty. For example, we could consider the setting that the matrix data $\mathbf{Y}_i \overset{iid}{\sim} \mathcal{F}(\mathbf{M}, \mathbf{U}_{p\times p}, \mathbf{V}_{q\times q})$ for some distribution $\mathcal{F}$ where $\mathbf{U}, \mathbf{V}$ characterize the column-wise and row-wise dependence, respectively. Also, we require the distribution $\mathcal{F}$ possesses Kronecker product property, i.e., $vec(\mathbf{Y}_i) \overset{iid}{\sim} \mathcal{F}'(vec(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$, then we could consider a multivariate CLT under the condition that $E(vec(\mathbf{Y}_i)_j^2) < \infty, j = 1, ..., p \times q$:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (vec(\mathbf{Y}_i) - E(vec(\mathbf{Y}_i))) \xrightarrow{d} \mathcal{N}_{pq}(\mathbf{0}, \mathbf{V} \otimes \mathbf{U}) \quad \text{as} \quad n \to \infty.$$

The simulation results of Model 3 in Section 4 have shown the robustness of our methods when the normality assumption is violated.

Furthermore, our methods can be extended to two-sample comparison for large tensor (multi-dimensional array) data, such as MRI/fMRI data (Lindquist, 2008). The ideas of extending current methods to tensors are as follows:

Let $\mathcal{A} \in \mathbb{R}^{h\times p\times q}$ be a 3-order tensor that $\mathcal{A} \sim \mathcal{N}_I(\mathcal{M}_A, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3)$ where $I := h \times p \times q$. Based on the property of tensor normal distribution, without loss of generality, Let $\mathbf{Y} = \mathcal{A}[1]$, which unfolds $\mathcal{A}$ along the mode-1, it's clear that $\mathbf{Y} \sim \mathcal{N}_{h,pq}(\mathbf{M}_A, \mathbf{\Sigma}_1, \mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2)$, where $\mathbf{M}_A$ is the unfolding of $\mathcal{M}_A$ along mode-1, $\mathbf{\Sigma}_1$ is a $h \times h$ matrix, $\mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2$ is a $pq \times pq$ matrix. Furthermore, we could vectorize the matrix, and by the property of matrix normal distribution, we have $vec(\mathbf{Y}) \sim \mathcal{N}(vec(\mathbf{M}_A), \mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1)$,

where $vec(\mathbf{Y})$ is a hpq dimensional column vector, and $\mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2 \otimes \mathbf{\Sigma}_1$ is a $(hpq) \times (hpq)$ dimensional covariance matrix. So to generalize the proposed methods to tensors (take the 3-order tensor data as an example), suppose we have n samples $\mathcal{A}_l \overset{iid}{\sim} \mathcal{N}_I(\mathcal{M}_A, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3)$ for $l = 1, ..., n$ from treatment group, m samples $\mathcal{B}_r \overset{iid}{\sim} \mathcal{N}_I(\mathcal{M}_B, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3)$ for $r = 1, ..., m$ from control group, where $I := h \times p \times q$, $\mathcal{A}$ and $\mathcal{B}$ have unknown mean tensors $\mathcal{M}_A = (\boldsymbol{\mu}_{A,ijk})$ and $\mathcal{M}_B = (\boldsymbol{\mu}_{B,ijk}), i = 1, ..., h, j = 1, ..., p, k = 1, ..., q$, but the same covariances $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \mathbf{\Sigma}_3$. Our problem would be extended to test

$$H_{0,ijk} : \boldsymbol{\mu}_{A,ijk} - \boldsymbol{\mu}_{B,ijk} = 0 \quad \text{against} \quad H_{1,ijk} : \boldsymbol{\mu}_{A,ijk} - \boldsymbol{\mu}_{B,ijk} \neq 0.$$

To apply our proposed methods, we first unfold those two groups of tensor data into matrix format. It turns out we have n samples $\mathbf{Y}_l \overset{iid}{\sim} \mathcal{MN}(\boldsymbol{\mu}_y, \mathbf{\Sigma}_1, \mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2)$ for $l = 1, ..., n$ from treatment group, m samples $\mathbf{Z}_r \overset{iid}{\sim} \mathcal{MN}(\boldsymbol{\mu}_z, \mathbf{\Sigma}_1, \mathbf{\Sigma}_3 \otimes \mathbf{\Sigma}_2)$ for $r = 1, ..., m$ from control group, where $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_z$ are the unfolding of $\mathcal{M}_A$ and $\mathcal{M}_B$ along mode-1, respectively. This will connect the tensor data to our proposed methods for matrix data.

We would like to pursue such generalizations in our future research. By and large, the current paper provides an effective step for the significance detection in the large-scale matrix-valued data, which will be useful for brain related research.

Supplementary Materials

Proofs of Theorems 1 & 2 and Propositions 1 & 2 as well as additional figures displaying results of numerical studies are relegated to the supplementary materials.

Acknowledgement

We would like to thank Dr. Yanhui Xu for her early assistance on some of the numerical studies. Our thanks also go to the associate editor and three referees for taking the time to review our work. Their valuable comments and suggestions have greatly improved the quality of the manuscript. Sarkar's research has been partially supported by NSF grant DMS 2210687,

REFERENCES

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203-1227.

Allen, G. I. and Tibshirani, R. (2012). Inference with transposable data: modeling the effects of row and column correlations. *Journal of the Royal Statistical Society, Series B* **74**, 721-743.

Aston, J. A., Pigoli, D. and Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics* **45**, 1431-1461.

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* **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165-1188.

Bickel, P. and Levina, L. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577-2604.

Cai, T., Liu, W. and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association* **108**, 265-277.

Chen, X., Yang, D., Xu, Y., Xia, Y., Wang, D. and Shen, H. (2023). Testing and support recovery of correlation structures for matrix-valued observations with an application to stock market data. *Journal of Econometrics* **232**, 544-564.

Dutilleuel, P. (1999). The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* **64**, 105-123.

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

Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association* **107**, 1019-1035.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal

orthogonal complements. *Journal of the Royal Statistical Society, Series B* **75**, 603-680.

Fan, J. and Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society, Series B* **79**, 1143-1164.

Fan, J., Ke, Y., Sun, Q. and Zhou, W. (2019). FarmTest: Factor-adjusted robust multiple testing with false discovery control. *Journal of the American Statistical Association* **114**, 1880-1893.

Fan, J., Wang, W. and Zhong, Y. (2019). Robust covariance estimation for approximate factor models. *Journal of Econometrics* **208**, 5-22.

Fox, M. D. and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neurosciences* **8**, 700-711.

Horenstein, M., Fan, R., Shedden, K. and Zhou, S. (2019). Joint mean and covariance estimation with unreplicated matrix-variate data. *Journal of the American Statistical Association* **114**, 682-696.

Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics* **38**, 1094-1121.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science* **23**, 439-464.

Liu, J., Psarakis, E., Feng, Y. and Stamos, I. (2019). A Kronecker product model for repeated pattern detection on 2D urban images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 2266-2272.

REFERENCES

Nandi, S., Sarkar, S. K. (2021). Adapting to one- and two-way classified structures of hypotheses while controlling the false discovery rate. *Journal of Statistical Planning and Inference* **215**, 95-108.

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

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479-498.

Xia, Y. and Li, L. (2017). Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics* **73**, 780-791.

Xia, Y. and Li, L. (2019). Matrix graph hypothesis testing and application in brain connectivity alternation detection. *Statistica Sinica* **29**, 303-328.

Department of Statistics, Operations, and Data Science

Fox School of Business, Temple University

Philadelphia, PA 19122, USA

E-mail: (shiyuz@temple.edu)

E-mail: (xhanprinceton@gmail.com)

E-mail: (sanat@temple.edu)