Statistica Sinica **29** (2019), 479-504
doi:https://doi.org/10.5705/ss.202017.0076

GRAPH ESTIMATION FOR MATRIX-VARIATE GAUSSIAN DATA

Xi Chen and Weidong Liu

New York University and Shanghai Jiao Tong University

Abstract: Matrix-variate Gaussian graphical models (GGM) have been widely used for modeling matrix-variate data. Since the support of sparse precision matrix represents the conditional independence graph among matrix entries, conducting support recovery yields valuable information. A commonly used approach is the penalized log-likelihood method. However, due to the complicated structure of precision matrices in the form of Kronecker products, the log-likelihood is non-convex, which presents challenges for both computation and theoretical analysis. In this paper, we propose an alternative approach by formulating the support recovery problem as a multiple testing problem. A new test statistic is developed and, based on that, we use the popular Benjamini and Hochberg's procedure to control false discovery rate (FDR) asymptotically. Our method involves only convex optimization, making it computationally attractive. Theoretically, our method allows very weak conditions and, even when the sample size is finite and the dimensions go to infinity, the asymptotic normality of the test statistics and FDR control can still be guaranteed. We further provide the power analysis result. The finite sample performance of the proposed method is illustrated with simulations and real data analysis.

Key words and phrases: Correlated samples, false discovery rate, matrix-variate Gaussian graphical models, multiple tests, support recovery.

1. Introduction

In the era of big data, matrix-variate observations are becoming prevalent in such domains as biomedical imaging, genomics, financial markets, spatio-temporal environmental data analysis, and more. A typical example is the gene expression data in genomics, in which each observation contains expression levels of p genes on q microarrays of the same subject (see, e.g., Efron (2009); Yin and Li (2012)). Another example of such data is the multi-channel electroencephalography (EEG) data for brain imaging studies (see, e.g., Bijma, De Munck and Heethaar (2005)), in which each measurement can be expressed as a matrix with rows corresponding to p different channels and columns to q time points. Leng and Tang (2012) provided more interesting examples of matrix-variate data. Due to the prevalence of matrix-variate observations (especially high-dimensional observations), it is important for us to understand the structural information encoded in these observations.

To study matrix-variate data where each observation $\mathbf{X}$ is a $p \times q$ matrix, it is commonly assumed that $\mathbf{X}$ follows a matrix-variate Gaussian distribution, e.g., Efron (2009); Allen and Tibshirani (2010); Leng and Tang (2012); Yin and Li (2012); Zhou (2014). The matrix-variate Gaussian distribution is a generalization of the familiar multivariate normal distribution for vector-variate data. In particular, let $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{pq\times 1}$ be the vectorization of matrix $\mathbf{X}$ obtained by stacking the columns of $\mathbf{X}$ on top of each other. We say that $\mathbf{X}$ follows a matrix-variate Gaussian distribution $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$ with mean matrix $\boldsymbol{\mu} \in \mathbb{R}^{p\times q}$, row covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ and column covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$ if and only if $\mathrm{vec}(\mathbf{X}') \sim N(\mathrm{vec}(\boldsymbol{\mu}'), \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$, where $\mathbf{X}'$ denotes the transpose of $\mathbf{X}$ and $\otimes$ is the Kronecker product.

Readers can refer to Dawid (1981) and Gupta and Nagar (1999) for more properties of matrix-variate Gaussian distribution. Similar to the vector-variate Gaussian graphical models (GGMs) in which the conditional independence graph is encoded in the support of the precision matrix, one can analogously define matrix-variate Gaussian graphical models (MGGM) (a.k.a. Gaussian bigraphical models). Let us denote a *conditional independence graph* by the undirected graph $G = (V, E)$, where $V = \{V_{ij}\}_{1\le i\le p, 1\le j\le q}$ contains $p \times q$ nodes and each node corresponds to an entry in the random matrix $\mathbf{X}$. We can regard the edge set E as a $pq \times pq$ matrix where there is no edge between X_{ij} and X_{kl} if and only if X_{ij} and X_{kl} are conditionally independent given the rest of the entries. The goal of the *graph estimation* is to estimate the edge set E, which unveils important structural information on the conditional independence relationship.

The estimation of the conditional independence graph is equivalent to the estimation of *the support of the precision matrix*. In particular, let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{ik})_{p\times p}$, $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1} = (\gamma_{jl})_{q\times q}$. The precision matrix of the MGGM is a $pq \times pq$ matrix $\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma} = (\boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})^{-1}$, where $(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma})_{q(i-1)+j,\, q(k-1)+l} = \omega_{ik} \cdot \gamma_{jl}$. The conditional independence among entries of $\mathbf{X}$ can be presented by the support of $\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}$ (denoted by $\mathrm{supp}(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma})$), which is equivalent to $\mathrm{supp}(\boldsymbol{\Omega}) \otimes \mathrm{supp}(\boldsymbol{\Gamma})$. To see this, we recall that X_{ij} and X_{kl} are conditionally independent given the rest of the entries, if and only if $\varrho_{ij,kl} = 0$, where $\varrho_{ij,kl}$ is the partial correlation between X_{ij} and X_{kl},

$$\varrho_{ij,kl} = -\frac{\omega_{ik}}{\sqrt{\omega_{ii}\omega_{kk}}} \cdot \frac{\gamma_{jl}}{\sqrt{\omega_{jj}\omega_{ll}}}. \tag{1.1}$$

Thus, X_{ij} and X_{kl} are conditionally independent if and only if there is at least one zero in ω_{ik} or γ_{jl}. Therefore, to estimate the conditional independence graph, one only needs to estimate $\mathrm{supp}(\mathbf{\Omega})$ and $\mathrm{supp}(\mathbf{\Gamma})$. Their Kronecker product $\mathrm{supp}(\mathbf{\Omega}) \otimes \mathrm{supp}(\mathbf{\Gamma})$ gives the edge set E. For a given matrix-variate Gaussian distribution, multiplying a constant to $\mathbf{\Omega}$ and dividing $\mathbf{\Gamma}$ by the same constant leads to the same distribution. The existing literature usually assumes that $\omega_{11} = 1$ to make the model identifiable (see, e.g., Leng and Tang (2012)). However, if we are interested in support recovery rather than values of ω_{ik} or γ_{jl}, then there is no identifiability issue.

Due to the complicated structure in the precision matrices of MGGMs, research on matrix-variate GGMs (MGGMs) is scarce compared to the large body of literature on vector-variate GGMs. The vector-variate GGM with random vector observations can be viewed as a special case of MGGM with $p = 1$ or $q = 1$, and readers can refer to Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Rothman et al. (2008); d'Aspremont, Banerjee and El Ghaoui (2008); Friedman, Hastie and Tibshirani (2008); Yuan (2010); Ravikumar et al. (2011); Cai, Liu and Luo (2011); Liu et al. (2012); Xue and Zou (2012); Liu (2013); Zhu, Shen and Pan (2014); Fan and Lv (2016); Ren et al. (2016) for the recent development in vector-variate GGMs. Our work is closely related to that of Liu (2013), which conducts graph estimation via false discovery rate (FDR) control for vector-variate GGMs. Due to the complicated structure of MGGMs, the proposed test statistics are fundamentally different from the ones in Liu (2013) and the theoretical analysis is more challenging. The details on the comparisons to Liu (2013) are deferred to Section 6.

For estimating sparse precision matrices of matrix-variate Gaussian data, one approach is based on the penalized likelihood method. However, since the precision matrices are in the form of a Kronecker product, the negative log-likelihood function is no longer convex, which makes both computation and theoretical analysis significantly more challenging than in the case of classical vector-variate GGMs. A few recent works (Allen and Tibshirani (2010); Leng and Tang (2012); Yin and Li (2012); Kalaitzis et al. (2013); Tsiligkaridis, Hero and Zhou (2013); Ying and Liu (2013); Huang and Chen (2015)) have focused on developing various penalized likelihood approaches for estimating MGGMs or extensions of MGGMs (e.g., multiple MGGMs in Huang and Chen (2015) and semiparametric extension in Ying and Liu (2013)). In particular, Leng and Tang (2012) provided theoretical guarantees on the estimated precision matrices, e.g., rates of convergence under the Frobenius norm and sparsistency. One limitation is that these results are stated in terms that there *exists* a local minimizer that enjoys good properties. In practice, it might be difficult to determine whether the obtained local minimizer from an optimization solver is a desired local minimizer. In addition, most convergence results require certain conditions on the sample size n and dimensionality p and q, e.g., p and q cannot differ too much from each other and n should go to infinity at a certain rate. We will show later that $n \to \infty$ is not necessary for the control of the false discovery rate (FDR) in support recovery for MGGMs. Zhou (2014) developed new penalized methods for estimating $\mathbf{\Sigma} \otimes \mathbf{\Psi}$ and $\mathbf{\Omega} \otimes \mathbf{\Gamma}$ and established the convergence rates under the spectral norm and the Frobenius norm. Our goal of accurate FDR control cannot be achieved by the method in Zhou (2014) or other penalized optimization approaches.

The main goal of this work is to infer the support of the precision matrix for an MGGM in high-dimensional settings, which fully characterizes the conditional independence relationship. Our method differs from the common approaches that turn the problem into a joint optimization over $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ with penalized likelihood methods. We utilize the large-scale testing framework and formulate the problem into multiple testing problems for $\mathbf{\Omega}$ and $\mathbf{\Gamma}$:

$$H_{0ij}^{\mathbf{\Omega}} : \omega_{ij} = 0 \quad \text{vs} \quad H_{0ij}^{\mathbf{\Omega}} : \omega_{ij} \neq 0, \quad 1 \leq i < j \leq p, \tag{1.2}$$

$$H_{0ij}^{\mathbf{\Gamma}} : \gamma_{ij} = 0 \quad \text{vs} \quad H_{0ij}^{\mathbf{\Gamma}} : \gamma_{ij} \neq 0, \quad 1 \leq i < j \leq q. \tag{1.3}$$

By conducting the multiple testing for (1.2) and (1.3), we obtain the estimates for the support of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$, denoted by $\widehat{\text{supp}(\mathbf{\Omega})}$ and $\widehat{\text{supp}(\mathbf{\Gamma})}$, respectively. Then, the support of $\mathbf{\Omega} \otimes \mathbf{\Gamma}$ can be naturally estimated by $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$. Instead of aiming for perfect support recovery, which requires strong conditions, our goal is to asymptotically control the false discovery rate (FDR). The FDR, originally introduced for multiple testing (Benjamini and Hochberg (1995)), has been considered one of the most important criterion for evaluating the quality of estimated networks (e.g., in the application of genetic data analysis, Schafer and Strimmer (2005); Ma, Gong and Bohnert (2007)). Refer to (4.2) and (4.5) in Section 4.2 for the definition of FDR in our graph estimation problems.

Although conducting variable selection via multiple testing is not a new idea, how to implement such a high-level idea for MGGMs with a complicated covariance structure requires several innovations in the methodology development. In particular, to conduct the multiple testing in (1.2) and (1.3), it is critical to construct a test statistic with the explicit asymptotic null distribution for each edge. To this end, we propose a new approach that fully utilizes the correlation structures among rows and columns of $\mathbf{X}$. In particular, suppose that there are n *i.i.d.* $p \times q$ matrix-variate samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$. To conduct the testing in (1.3) and estimate the support of $q \times q$ matrix $\mathbf{\Gamma}$, we treat each row of $\mathbf{X}^{(k)}$ as a q-dimensional sample. In such a way, we construct $n \cdot p$ *correlated* vector-variate samples for the testing problem in (1.3), where the correlation among these "row samples" is characterized by the covariance matrix $\mathbf{\Sigma}$. One important advantage of this approach is that it only requires number of row samples $np \to \infty$ to control FDR asymptotically, and thus allows a finite sample size n even when p and q go to infinity. On the other hand, the correlation structure among row samples also presents a significant challenge to the development of the FDR control approach, and most existing inference techniques for vector-variate GGMs rely heavily on the independence assumption (see, e.g., Liu (2013); Van de Geer et al. (2014); Ren et al. (2015)). To address this challenge, we summarize the effect of correlation among "row samples" in a simple quantity depending on $\mathbf{\Sigma}$ and, based on that, introduce a variance correction technique into the construction of the test statistics (see Section 3.1). The testing of the $\text{supp}(\mathbf{\Omega})$ in (1.2) can be performed in a completely symmetric way with nq correlated "column samples" from the data.

More specifically, the high-level description of the proposed large-scale testing approach is as follows. Given the "row samples" from the data, the first step is to construct an asymptotically normal test statistic for each $H^{\mathbf{\Gamma}}_{0ij}$ in (1.3). We utilize a fundamental result from multivariate analysis which relates the partial correlation coefficient to the correlation coefficient of residuals from linear regression. To compute the sample version of the correlation coefficient of the residuals, we first construct an initial estimator for the regression coefficients. With the initial estimator in place, we can directly show that the sample correlation coefficient of the residuals is asymptotically normal under the null. We further apply the aforementioned variance correction technique and obtain the final test statistic for each $H^{\mathbf{\Gamma}}_{0ij}$. Combining the developed test statistics with the Benjamini and Hochberg approach (Benjamini and Hochberg (1995)), we show that the resulting procedure asymptotically controls the FDR for both $\widehat{\text{supp}(\mathbf{\Gamma})}$ and $\widehat{\text{supp}(\mathbf{\Omega})}$ (and thus $\widehat{\text{supp}(\mathbf{\Gamma})} \otimes \widehat{\text{supp}(\mathbf{\Omega})}$) under some sparsity conditions of $\mathbf{\Gamma}$ and $\mathbf{\Omega}$.

This proposed method is the first to investigate FDR control in MGGMs, greatly generalizing the method on FDR control for *i.i.d.* vector-variate GGMs in Liu (2013), and it improves on optimization-based approaches. The main contribution and difference between our results and the existing ones for vector-variate GGMs (e.g., Liu (2013)) are summarized as follows,

1. We propose a novel *test statistic* in Section 3.1. By introducing a new construction of the initial regression coefficients (i.e., setting a particular element in each initial Lasso estimator to zero), our testing approach no longer requires a complicated bias-correlation step as in Eq. (6) in Liu (2013). Furthermore, the limiting null distribution of the sample covariance coefficient between residuals (see $\widehat{r}_{ij}$ in (3.7)) can be easily obtained. In fact, this idea can be used to provide simpler testing procedure for ordinary vector-variate high-dimensional graphical models.

2. Instead of relying on the *i.i.d.* assumption in GGM literature, we propose to extract *np correlated* vector-variate "rows samples" (as well *nq* correlated "column samples") from matrix-variate observations. By utilizing correlation structure among rows and columns, our approach allows for finite sample size, which is a very attractive property from both theoretical and practical perspectives. More specifically, even in the case that n is a constant and $p \to \infty$ and $q \to \infty$, our method still guarantees the asymptotic normality of the test statistics and FDR control. This is fundamentally different from the case of vector-variate GGMs, which always requires $n \to \infty$ for the support recovery. Therefore, this work greatly generalizes Liu (2013), which only deals with *i.i.d.* vector-variate Gaussian samples, to correlated data.

   In this paper, we develop new techniques and theoretical analysis to address the challenges arising from correlated samples. For example, the proposed *variance correlation technique* can be used as a general technique for high-dimensional inference with correlated samples. Moreover, the initial estimator is now based on the Lasso with correlated samples. To this end, we establish the consistency result for the Lasso with correlated samples, which itself is of independent interest for high-dimensional linear regression.

3. Theoretically, by utilizing the Kronecker product structure of the covariance matrix of $\mathbf{X}$, the proposed method allows the partial correlation between X_{ij} and X_{kl} (i.e., $\varrho_{ij,kl}$ in (1.1)) to be of the order of $\{1/(n-1)\}\sqrt{(\log p \log q)/(pq)}$ so that the corresponding edge can be detected (please see the power analysis in Section 4.3 and Theorem 4 for details). This is

essentially different from any vector-variate GGM estimator (e.g., the one from Liu (2013)) that requires the partial correlation to be at least $C(1/\sqrt{n})$.

In terms of support recovery and computational cost, our method has several advantages as compared to popular penalized likelihood approaches: First, it provides an accurate control of FDR (see Theorem 3). Existing support recovery results only guarantee that the estimated positions of zeros are supersets of the positions of true zeros in $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ with probability tending to one when n, p, q all go to infinity at certain rates. (see, e.g., Leng and Tang (2012); Yin and Li (2012)). Second, in terms of computation, as compared to existing penalized likelihood methods whose objective functions are non-convex, our approach is based on convex optimization and thus computationally more attractive. The main computational cost of our approach is the construction of initial estimates for $p+q$ regression coefficient vectors that directly lead to our test statistics. The corresponding computational problems are completely independent allowing an efficient parallel implementation. Theoretically, our approach allows a wider range of p and q. In particular, our result on FDR control holds for (p, q) such that $q^{r_2} \leq p \leq q^{r_1}$ for some $0 < r_2 \leq r_1$, while Leng and Tang (2012) require that $p\log(p) = o(nq)$ and $q\log(q) = o(np)$.

2. Notations and Organization of the Paper

We introduce some necessary notation. Let $\mathbf{X}^{(k)} = (X_{ij}^{(k)})_{p\times q}$ for $k = 1,\ldots,n$ be the n *i.i.d.* matrix-variate observations from $N_{p,q}(\boldsymbol{\mu}, \mathbf{\Sigma}\otimes\mathbf{\Psi})$ and let $\bar{\mathbf{X}} = (1/n)\sum_{k=1}^{n}\mathbf{X}^{(k)}$. Put $\mathbf{\Sigma} = (\sigma_{ij})$ and $\mathbf{\Psi} = (\psi_{ij})$. For any vector $\mathbf{x} = (x_1,\ldots,x_p)'$, let $\mathbf{x}_{-i}$ denote $p-1$ dimensional vector by removing x_i from $\mathbf{x}$. Let $\langle\mathbf{x},\mathbf{y}\rangle$ be the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$. For any $p\times q$ matrix $\mathbf{A}$, let $\mathbf{A}_{i,\cdot}$ denote the i-th row of $\mathbf{A}$ (or $\mathbf{A}_i$ when it is clear from the context) and $\mathbf{A}_{\cdot,j}$ denote the j-th column of $\mathbf{A}$ (or $\mathbf{A}_j$ when it is clear from the context). Further, let $\mathbf{A}_{i,-j}$ denote the i-th row of $\mathbf{A}$ with its j-th entry removed and $\mathbf{A}_{-i,j}$ denotes the j-th column of $\mathbf{A}$ with its i-th entry removed. $\mathbf{A}_{-i,-j}$ denotes a $(p-1)\times(q-1)$ matrix by removing the i-th row and j-th column of $\mathbf{A}$. Take $[n] = \{1,\ldots,n\}$, $[p] = \{1,\ldots,p\}$ and $[q] = \{1,\ldots,q\}$. For a p-dimensional vector $\mathbf{x}$, let $|\mathbf{x}|_0 = \sum_{j=1}^{p} I(x_j \neq 0)$, $|\mathbf{x}|_1 = \sum_{j=1}^{p}|x_j|$ and $|\mathbf{x}|_2 = \sqrt{\sum_{j=1}^{p}x_j^2}$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, let $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{i\in[p],j\in[q]}a_{ij}^2}$ be the Frobenius norm of $\mathbf{A}$, $|\mathbf{A}|_\infty = \max_{i\in[p],j\in[q]}|a_{ij}|$ be the element-wise ℓ_∞-norm of $\mathbf{A}$ and $\|\mathbf{A}\|_2 = \sup_{|\mathbf{x}|_2\leq 1}|\mathbf{A}\mathbf{x}|_2$ be the spectral norm of $\mathbf{A}$. For a square matrix $\mathbf{A}$, let $\mathrm{tr}(\mathbf{A})$ denote the trace of $\mathbf{A}$. For a given set $\mathcal{H}$, let $Card(\mathcal{H})$ be the cardinality of $\mathcal{H}$. Throughout the paper, we use $\mathbf{I}_p$ to denote the $p \times p$ identity matrix, and we use C, c, etc. to denote generic constants whose values might change from place to place.

The rest of the paper is organized as follows. In Section 3, we introduce our test statistics and then describe the FDR control procedure for MGGM estimation. Theoretical results on asymptotic normality of our test statistic and FDR control are reported in Section 4. Simulations and real data analysis are given in Section 5. In Section 6, we provide further discussions on the proposed method and also point out some future work. The proofs and some additional experimental results are relegated to the supplementary material.

3. Methodology

The false discover proportion (FDP) is the proportion of false discoveries among total rejections. If $\widehat{\text{supp}(\boldsymbol{\Omega})}$ and $\widehat{\text{supp}(\boldsymbol{\Gamma})}$ are the estimators of $\text{supp}(\boldsymbol{\Omega})$ and $\text{supp}(\boldsymbol{\Gamma})$, respectively, under the control of FDP at level $\alpha \in [0, 1]$, it is clear that the FDP of $\widehat{\text{supp}(\boldsymbol{\Omega})} \otimes \widehat{\text{supp}(\boldsymbol{\Gamma})}$ as an estimator of $\text{supp}(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma})$ is controlled at some level α'. Here, the level α' (explicitly given at (4.9)) is a monotonically increasing function in α. Therefore, we reduce our task to design an estimator of $\text{supp}(\boldsymbol{\Gamma})$ under the FDP level α and the estimator of $\text{supp}(\boldsymbol{\Omega})$ can be constructed similarly. We propose to estimate $\text{supp}(\boldsymbol{\Gamma})$ by implementing the large-scale multiple tests:

$$H_{0ij} : \gamma_{ij} = 0 \quad \text{vs.} \quad H_{1ij} : \gamma_{ij} \neq 0, \quad 1 \leq i < j \leq q. \tag{3.1}$$

3.1. Construction of test statistics

In this section, we propose our test statistic for each H_{0ij} in (3.1), constructed from n *i.i.d.* $p \times q$ matrix-variate samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ with the population distribution $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi})$. Denote the partial correlation matrix associated with $\boldsymbol{\Gamma}$ by $\boldsymbol{\rho}^{\boldsymbol{\Gamma}} = \left(\rho^{\boldsymbol{\Gamma}}_{ij\cdot}\right)_{q\times q}$, where $\rho^{\boldsymbol{\Gamma}}_{ij\cdot} = -(\gamma_{ij}/\sqrt{\gamma_{ii}\gamma_{jj}})$ is the partial correlation coefficient between X_{li} and X_{lj} for any $1 \leq l \leq p$. For $1 \leq i < j \leq q$ and any $1 \leq l \leq p$, consider the population regression coefficients,

$$\begin{aligned}(\alpha_i, \boldsymbol{\beta}_i) &= \underset{a\in\mathbb{R},\mathbf{b}\in\mathbb{R}^{q-1}}{\arg\min} \ \mathbb{E}(X_{li} - a - \mathbf{X}_{l,-i}\mathbf{b})^2, \\ (\alpha_j, \boldsymbol{\beta}_j) &= \underset{a\in\mathbb{R},\mathbf{b}\in\mathbb{R}^{q-1},}{\arg\min} \ \mathbb{E}(X_{lj} - a - \mathbf{X}_{l,-j}\mathbf{b})^2. \end{aligned} \tag{3.2}$$

The standard linear regression result shows that

$$\boldsymbol{\beta}_i = -\gamma_{ii}^{-1}\boldsymbol{\Gamma}_{-i,i} \, ; \qquad \boldsymbol{\beta}_j = -\gamma_{jj}^{-1}\boldsymbol{\Gamma}_{-j,j}. \tag{3.3}$$

For such $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$, the corresponding residuals ϵ_{li} and ϵ_{lj} take the form,

$$\varepsilon_{li} = X_{li} - \alpha_i - \mathbf{X}_{l,-i}\boldsymbol{\beta}_i \ ; \quad \varepsilon_{lj} = X_{lj} - \alpha_j - \mathbf{X}_{l,-j}\boldsymbol{\beta}_j. \tag{3.4}$$

It is known that $\mathbb{E}(\varepsilon_{li}) = \mathbb{E}(\varepsilon_{lj}) = 0$. Moreover, the correlation between ε_{li} and ε_{lj} is $\mathrm{Corr}(\varepsilon_{li}, \varepsilon_{lj}) = -\rho^{\mathbf{\Gamma}}_{ij\cdot} = \gamma_{ij}/\sqrt{\gamma_{ii}\gamma_{jj}}$. To see this, let $\boldsymbol{\gamma}_i$ be the i-th column of $\mathbf{\Gamma}$ for $1 \le i \le q$. By (3.3) and (3.4), we can equivalently write $\epsilon_{li} = -\alpha_i + \gamma_{ii}^{-1}\mathbf{X}_{l,\cdot}\boldsymbol{\gamma}_i$ and $\epsilon_{lj} = -\alpha_j + \gamma_{jj}^{-1}\mathbf{X}_{l,\cdot}\boldsymbol{\gamma}_j$. Since the covariance of $\mathbf{X}_{l,\cdot}$, $\mathrm{Cov}(\mathbf{X}_{l,\cdot})$, is $\sigma_{ll}\mathbf{\Psi} = \sigma_{ll}\mathbf{\Gamma}^{-1}$, we have

$$\mathrm{Cov}(\epsilon_{li}, \epsilon_{lj}) = \gamma_{ii}^{-1}\gamma_{jj}^{-1}\boldsymbol{\gamma}_i^T\mathrm{Cov}(\mathbf{X}_{l,\cdot})\boldsymbol{\gamma}_j = \sigma_{ll}\gamma_{ii}^{-1}\gamma_{jj}^{-1}\left(\boldsymbol{\gamma}_i^T\mathbf{\Gamma}^{-1}\right)\boldsymbol{\gamma}_j = \sigma_{ll}\gamma_{ii}^{-1}\gamma_{jj}^{-1}\gamma_{ij}, \tag{3.5}$$

where the last equality holds because $\boldsymbol{\gamma}_i^T\mathbf{\Gamma}^{-1} = \mathbf{e}_i^T$ where $\mathbf{e}_i$ denotes the i-th canonical vector. Similarly, $\mathrm{Var}(\epsilon_{li}) = \sigma_{ll}\gamma_{ii}^{-1}$ and $\mathrm{Var}(\epsilon_{lj}) = \sigma_{ll}\gamma_{jj}^{-1}$ which, together with (3.5), imply that $\mathrm{Corr}(\varepsilon_{li}, \varepsilon_{lj}) = \gamma_{ij}/\sqrt{\gamma_{ii}\gamma_{jj}}$. Therefore, testing whether $\gamma_{ij} = 0$ (or $\rho^{\mathbf{\Gamma}}_{ij\cdot} = 0$) is equivalent to testing whether $\mathrm{Corr}(\varepsilon_{li}, \varepsilon_{lj}) = 0$. We build our test statistics based on this equivalence relationship.

To implement this, one needs to construct an initial estimator for each $\boldsymbol{\beta}_j$ so that the distribution of sample correlation coefficient of residuals can be deduced easily. To do so, we first let $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{1,j}, \ldots, \widehat{\beta}_{q-1,j})'$ be any estimator for $\boldsymbol{\beta}_j$ that satisfies

$$\max_{1\le j\le q} |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j|_1 = O_{\mathbb{P}}(a_{n1}), \quad \text{and} \quad \max_{1\le j\le q} |\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j|_2 = O_{\mathbb{P}}(a_{n2}), \tag{3.6}$$

where $a_{n1} \to 0$ and $a_{n2} \to 0$ at some rate to be specified later. The Lasso (Tibshirani (1996)), Dantzig selector (Candès and Tao (2007)), or other sparse regression approaches can be adopted provided that (3.6) is satisfied (see Section 3.2 for details). Under the null H_{0ij}: $\gamma_{ij} = 0$, according to (3.3), the i-th element in $\boldsymbol{\beta}_j$ that corresponds to the covariate X_{li} in (3.4), is zero, and the $(j-1)$-th element in $\boldsymbol{\beta}_i$ that corresponds to the covariate X_{lj} in (3.4), is also zero. Hence, for each pair $i < j$, we take the initial estimators $\widehat{\boldsymbol{\beta}}_{j,\setminus i} = (\widehat{\boldsymbol{\beta}}_{1,j}, \ldots, \widehat{\boldsymbol{\beta}}_{i-1,j}, 0, \widehat{\boldsymbol{\beta}}_{i+1,j}, \ldots, \widehat{\boldsymbol{\beta}}_{q-1,j})'$ and $\widehat{\boldsymbol{\beta}}_{i,\setminus j} = (\widehat{\boldsymbol{\beta}}_{1,i}, \ldots, \widehat{\boldsymbol{\beta}}_{j-2,i}, 0, \widehat{\boldsymbol{\beta}}_{j,i}, \ldots, \widehat{\boldsymbol{\beta}}_{q-1,i})'$ for $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_i$, respectively. For convenience, we let $\widehat{\boldsymbol{\beta}}_{j,\setminus j} = \widehat{\boldsymbol{\beta}}_j$ so that the "sample residual" $\widehat{\varepsilon}^{(k)}_{ljj}$ introduced in the below is also well-defined (see (3.7)).

Given the initial estimators under the null, we construct the "sample residuals" by treating each row $l \in [p]$ of a matrix-variate observation $\mathbf{X}^{(k)}$ for $k \in [n]$ as a "row sample", i.e., $\mathbf{X}^{(k)}_{l,\cdot} = (X^{(k)}_{l1}, \ldots X^{(k)}_{lq})$. The "sample residuals" corresponding to ϵ_{li} and ϵ_{lj} in (3.4) are:

$$\widehat{\varepsilon}^{(k)}_{lij} = X^{(k)}_{li} - \bar{X}_{li} - (\mathbf{X}^{(k)}_{l,-i} - \bar{\mathbf{X}}_{l,-i})\widehat{\boldsymbol{\beta}}_{i,\setminus j},$$

$$\widehat{\varepsilon}_{lji}^{(k)} = X_{lj}^{(k)} - \bar{X}_{lj} - (\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})\widehat{\boldsymbol{\beta}}_{j,\backslash i}, \tag{3.7}$$

where $\bar{X}_{li} = (1/n)\sum_{k=1}^{n} X_{li}^{(k)}$ and $\bar{\mathbf{X}}_{l,-i} = (1/n)\sum_{k=1}^{n} \mathbf{X}_{l,-i}^{(k)}$. Let $\widehat{r}_{ij}$ be the sample covariance coefficient between constructed residuals,

$$\widehat{r}_{ij} = \frac{1}{(n-1)p}\sum_{k=1}^{n}\sum_{l=1}^{p} \widehat{\varepsilon}_{lij}^{(k)}\widehat{\varepsilon}_{lji}^{(k)}; \tag{3.8}$$

and $\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}$ be the corresponding sample correlation coefficient of residuals.

The proposed construction of the sample correlation of residuals has advantages. Under the null H_{0ij}, incorporating the fact that $\gamma_{ij} = 0$ into the regression coefficients enables us to derive the asymptotic null distribution of the sample correlation coefficients. In particular, Proposition 1 show that under the null,

$$\sqrt{\frac{(n-1)p}{A_p}}\frac{\widehat{r}_{ij}}{\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}} \Rightarrow N(0,1), \tag{3.9}$$

where

$$A_p = \frac{p\|\boldsymbol{\Sigma}\|_{\mathrm{F}}^2}{\{\mathrm{tr}(\boldsymbol{\Sigma})\}^2}. \tag{3.10}$$

Here, A_p is the asymptotic variance of $\sqrt{(n-1)p}(\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}})$ under the null. It is noteworthy that the term A_p is critical since it plays the role of *variance correction* when treating rows of matrix-variate data as correlated samples. For the proposed construction, although the sample correlation coefficients are constructed under the null, we can show that $\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}$ converges to $(1-\gamma_{ij}\psi_{ij})\rho_{ij\cdot}^{\boldsymbol{\Gamma}}$ in probability as $np \to \infty$ under both the null and alternative (see Proposition 2). This result indicates that the test statistic based on $\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}$ can properly reject the null when the magnitude of partial correlation coefficient $|\rho_{ij\cdot}^{\boldsymbol{\Gamma}}|$ is away from zero. For many widely studied covariance structures of $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma} = \boldsymbol{\Psi}^{-1}$, $\psi_{ij}\gamma_{ij}$ is often non-positive, which makes the signal strength $(1-\gamma_{ij}\psi_{ij})\rho_{ij\cdot}^{\boldsymbol{\Gamma}}$ even larger than the partial correlation coefficient and thus leads to good statistical power. For example, two variables that are directly positively/negatively correlated, they will often be positively/negatively conditionally correlated.

It is worthwhile to note that the variance correction quantity $A_p \in \mathbb{R}$ in (3.9) is unknown as it involves $\boldsymbol{\Sigma}$. In the next subsection, we will propose a ratio consistent estimator $\widehat{A}_p$ such that $\widehat{A}_p/A_p \to 1$ in probability as $nq \to \infty$. Using $\widehat{A}_p$, we construct the final test statistic for H_{0ij}:

$$\widehat{T}_{ij} = \sqrt{\frac{(n-1)p}{\widehat{A}_p}}\frac{\widehat{r}_{ij}}{\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}}, \tag{3.11}$$

which is asymptotically normal given (3.9) and the ratio consistency of $\widehat{A}_p$.

3.1.1. Estimator for A_p

We propose an estimator $\widehat{A}_p$ of A_p based on a thresholding estimator of $\mathbf{\Sigma}$. We first construct an initial estimator of $\mathbf{\Sigma}$ based on nq "column samples", where each column of $\mathbf{X}^{(k)}$ for $k \in [n]$ is treated as a p-dimensional sample. In particular, let $\bar{\mathbf{X}} = (1/n)\sum_{k=1}^{n}\mathbf{X}^{(k)}$, each centered column sample $\mathbf{Y}_{kj} = \mathbf{X}^{(k)}_{\cdot,j} - \bar{\mathbf{X}}_{\cdot,j} \in \mathbb{R}^{p\times 1}$ for $k \in [n]$ and $j \in [q]$. Take $\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p} := [1/\{(n-1)q\}]\sum_{k=1}^{n}\sum_{j=1}^{q}(\mathbf{Y}_{kj})(\mathbf{Y}_{kj})'$. In a more succinct notation, let $\mathbf{Y} = \{\mathbf{X}^{(1)} - \bar{\mathbf{X}}, \ldots, \mathbf{X}^{(n)} - \bar{\mathbf{X}}\} \in \mathbb{R}^{p\times(nq)}$ and $\widehat{\mathbf{\Sigma}} = [1/\{(n-1)q\}]\,\mathbf{Y}\mathbf{Y}'$. Then, we threshold the elements of $\widehat{\mathbf{\Sigma}}$ as follows:

$$\widehat{\sigma}_{ij,\lambda} = \widehat{\sigma}_{ij} I\left\{|\widehat{\sigma}_{ij}| \geq \lambda\sqrt{\frac{\log\max(p,nq)}{nq}}\right\} \quad \text{for } i \neq j; \tag{3.12}$$

and $\widehat{\sigma}_{ii,\lambda} = \widehat{\sigma}_{ii}$ for $i \in [p]$. Set $\widehat{\mathbf{\Sigma}}_\lambda = (\widehat{\sigma}_{ij,\lambda})_{p\times p}$ and define the plug-in estimator of A_p:

$$\widehat{A}_p = \frac{p\|\widehat{\mathbf{\Sigma}}_\lambda\|_{\mathrm{F}}^2}{\{\mathrm{tr}(\widehat{\mathbf{\Sigma}}_\lambda)\}^2}. \tag{3.13}$$

In Section 4, we show that $\widehat{A}_p/A_p \to 1$ in probability as $nq \to \infty$ for a properly chosen λ (a data-driven approach for the choice of λ will be discussed later in Section 3.2). Therefore, by (3.9), we have the desired asymptotic normality of $\widehat{T}_{ij}$ under the null: $\widehat{T}_{ij} \Rightarrow N(0,1)$ as $np \to \infty$.

We further note that when the columns of $\mathbf{X}$ are *i.i.d.* (i.e., $\mathbf{\Psi} = \mathbf{I}_q$), consistency under the spectral norm $\|\widehat{\mathbf{\Sigma}}_\lambda - \mathbf{\Sigma}\|_2$ has been established if the sparsity condition of $\mathbf{\Sigma}$ satisfies the row sparsity level $s_0(p) = O(\sqrt{nq/\log p})$ (see, e.g., Bickle and Levina (2008); Cai and Liu (2011) and references therein). We do not need such a strong consistency result for $\widehat{\mathbf{\Sigma}}_\lambda$ in the spectral norm to establish the consistency of $\widehat{A}_p$. In fact, since A_p only involves $\|\mathbf{\Sigma}\|_{\mathrm{F}}$ rather than $\mathbf{\Sigma}$ itself, the sparsity condition on $s_0(p)$ is no longer necessary; see Proposition 3 and its proof for more details. Other estimators of $\|\mathbf{\Sigma}\|_{\mathrm{F}}$ have been proposed (e.g,. by Chen and Qin (2010)). But those approaches rely heavily on the *i.i.d.* assumption on the columns of $\mathbf{X}$.

3.2. Initial estimators of regression coefficients

In the construction of the test statistic $\widehat{T}_{ij}$, we need an estimate $\widehat{\boldsymbol{\beta}}_j$ that satisfies the condition in (3.6). Here, we choose to construct $\widehat{\boldsymbol{\beta}}_j$ using Lasso for simplicity, but other approaches such as the Dantzig selector can also be used. In particular, let $\mathbf{Z} = \left\{(\mathbf{X}^{(1)})' - \bar{\mathbf{X}}', \ldots, (\mathbf{X}^{(n)})' - \bar{\mathbf{X}}'\right\} \in \mathbb{R}^{q\times(np)}$ be np q-dimensional samples extracted from the data, and $\widehat{\boldsymbol{\Psi}} = [1/\{(n-1)p\}]\,\mathbf{Z}\mathbf{Z}' =: (\widehat{\psi}_{ij})_{q\times q}$. For $1 \le j \le q$, define the scaling/normalizing vector $\mathbf{D}_j = \mathrm{diag}(\widehat{\boldsymbol{\Psi}}_{-j,-j}) \in \mathbb{R}^{(q-1)\times(q-1)}$. The coefficients $\boldsymbol{\beta}_j$ can be estimated by the Lasso as follows:

$$\widehat{\boldsymbol{\beta}}_j(\delta) = \mathbf{D}_j^{-1/2}\widehat{\boldsymbol{\alpha}}_j(\delta), \tag{3.14}$$

where

$$\widehat{\boldsymbol{\alpha}}_j(\delta) = \underset{\boldsymbol{\alpha}\in\mathbb{R}^{q-1}}{\arg\min} \left(\frac{1}{2np}\sum_{k=1}^{n}\sum_{l=1}^{p} \left\{ X_{lj}^{(k)} - \bar{X}_{lj} - (\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j})\mathbf{D}_j^{-1/2}\boldsymbol{\alpha} \right\}^2 + \theta_{nj}(\delta)|\boldsymbol{\alpha}|_1 \right) \tag{3.15}$$

and

$$\theta_{nj}(\delta) = \delta\sqrt{\frac{\widehat{\psi}_{jj}\log\max(q,np)}{np}}. \tag{3.16}$$

In the Lasso estimate in (3.15), the np covariates-response pairs $(\mathbf{X}_{l,-j}^{(k)} - \bar{\mathbf{X}}_{l,-j}, X_{lj}^{(k)} - \bar{X}_{lj})$ for $k \in [n]$ and $j \in [p]$ are not *i.i.d.* and thus the standard consistency result of Lasso cannot be applied here. By exploring the correlation structure among rows of $\mathbf{X}$, we managed to derive the rate of convergence of the Lasso estimator in the ℓ_1 and the ℓ_2 norms. This result may be of independent interest for dealing with high-dimensional correlated data. For the choice of tuning parameters, our theoretical results hold for any large enough constants λ in (3.12) for estimating $\widehat{A}_p$ (see Proposition 3) and $\delta > 0$ in (3.16) for $\widehat{\boldsymbol{\beta}}_j(\delta)$ (see Proposition 4). In our experiment, we adopt a data-driven parameter-tuning strategy from Liu (2013).

3.3. FDR control procedure

Given the constructed test statistic $\widehat{T}_{ij}$, we can carry out $(q^2-q)/2$ tests in (3.1) simultaneously using the popular Benjamini and Hochberg (BH) method Benjamini and Hochberg (1995). Let the p-values $p_{ij} = 2 - 2\Phi(|\widehat{T}_{ij}|)$ for $1 \le i < j \le q$. We sort these $m = (q^2-q)/2$ p-values such that $p_{(1)} < \cdots < p_{(m)}$. For a given $0 < \alpha < 1$, take

$$\widehat{k} = \max\left\{0 \le k \le m : p_{(k)} \le \frac{\alpha k}{m}\right\}.$$

For $1 \le i \ne j \le q$, we reject H_{0ij} if $p_{ij} \le p_{(\widehat{k})}$ and the estimated support of $\boldsymbol{\Gamma}$ is

$$\widehat{\mathrm{supp}(\boldsymbol{\Gamma})} = \{(i,j) : p_{ij} \le p_{(\widehat{k})}, 1 \le i \ne j \le q\} \cup \{(i,i) : 1 \le i \le q\}. \tag{3.17}$$

We set $\widehat{T}_{ji} = \widehat{T}_{ij}$ for $1 \le i < j \le q$ in (3.17).

Note that the original results from Benjamini and Hochberg (1995) cannot be directly applied to obtain the guarantee of FDR control since the test statistics (and thus the p-values) are correlated with each other. By utilizing some techniques of Liu (2013), we can prove that this procedure controls the FDR/FDP asymptotically in $\widehat{\mathrm{supp}(\boldsymbol{\Gamma})}$ (see Section 4.2 for details).

Estimation of supp($\boldsymbol{\Omega}$). The estimation of $\mathrm{supp}(\boldsymbol{\Omega})$ can be done in the same way. In particular, we only need to consider the transpose of each matrix-variate observation $\mathbf{X}^{(k)}$, $(\mathbf{X}^{(1)})', \ldots, (\mathbf{X}^{(n)})'$ and change some notation (e.g., p to q and q to p).

Estimation of supp($\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}$). Let $\widehat{\mathrm{supp}(\boldsymbol{\Omega})}$ and $\widehat{\mathrm{supp}(\boldsymbol{\Gamma})}$ be the estimators of $\mathrm{supp}(\boldsymbol{\Omega})$ and $\mathrm{supp}(\boldsymbol{\Gamma})$, respectively, under the control of FDR at level α. The support of $\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}$ can then be estimated by $\widehat{\mathrm{supp}(\boldsymbol{\Omega})} \otimes \widehat{\mathrm{supp}(\boldsymbol{\Gamma})}$. In Theorem 3, we show that the FDR/FDP of this estimator is controlled at level

$$\alpha' = \frac{\alpha\{(2-\alpha)ab + aq + bp\}}{\max(ab + aq + pb, 1)} \tag{3.18}$$

asymptotically, where a and b are the numbers of total discoveries in $\widehat{\mathrm{supp}(\boldsymbol{\Omega})}$ and $\widehat{\mathrm{supp}(\boldsymbol{\Gamma})}$, respectively, excluding the diagonal entries. When the FDR/FDP level α' for the joint support estimation is given, to determine the FDR level α for the estimation of $\mathrm{supp}(\boldsymbol{\Omega})$ and $\mathrm{supp}(\boldsymbol{\Gamma})$, one can try a sequence of α's from small to large. For each candidate α, we obtain the value a and b by estimating $\mathrm{supp}(\boldsymbol{\Omega})$ and $\mathrm{supp}(\boldsymbol{\Gamma})$ and plug the obtained values into (3.18). Finally, we choose the value α for the separate estimation that leads to the closest value to the pre-given α'.

Remark 1. One natural approach of solving our problem is the de-correlation method. More precisely, if $\boldsymbol{\Sigma}$ is known, the data matrix can be transformed as $\boldsymbol{\Sigma}^{-1/2}\mathbf{X} \sim N(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I}_{p\times p} \otimes \boldsymbol{\Psi})$, based on which the method from Liu (2013) can be applied. Therefore, a natural two-stage approach is to first obtain an consistent estimator of $\boldsymbol{\Sigma}^{-1/2}$ (e.g., Cai, Liu and Luo (2011)) and then apply the FDR control approach of Liu (2013) to $\widehat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{X}^{(i)}$ for $1 \le i \le n$. However, this "de-correlation" approach is not applicable here. To ensure the estimation error between $\widehat{\boldsymbol{\Sigma}}^{-1/2}$ and $\boldsymbol{\Sigma}^{-1/2}$ is negligible for the FDR control, we find that we need the condition

$$\frac{nq}{\{ps^2(p)\}^{1/(1-\tau)}} \to \infty \tag{3.19}$$

to replace $\boldsymbol{\Sigma}^{-1/2}$ by $\widehat{\boldsymbol{\Sigma}}^{-1/2}$. (This condition means that we need a large sample number nq to estimate $\boldsymbol{\Sigma}^{-1/2}$ accurately). Similarly, to replace $\boldsymbol{\Psi}^{-1/2}$ by $\widehat{\boldsymbol{\Psi}}^{-1/2}$, we need the condition

$$\frac{np}{\{qs^2(q)\}^{1/(1-\tau)}} \to \infty. \tag{3.20}$$

Hence, to get $\widehat{\operatorname{supp}(\mathbf{\Sigma}^{-1})} \otimes \widehat{\operatorname{supp}(\mathbf{\Psi}^{-1})}$, we need (3.19) and (3.20) simultaneously. However, when n is fixed or small, (3.19) and (3.20) are contrary. Hence, it is impossible to do the de-correlation for rows and columns of $\mathbf{X}$ simultaneously.

4. Theoretical Results

In this section, we provide the properties of the test statistic at (3.11), the guarantee of FDR control, power analysis and convergence rate of the initial estimator. Proofs are relegated to the supplement.

Let $\lambda_{\min}(\mathbf{\Sigma}) = \lambda_1^{(1)} \le \cdots \le \lambda_p^{(1)} = \lambda_{\max}(\mathbf{\Sigma})$ be the eigenvalues of $\mathbf{\Sigma}$ and $\lambda_{\min}(\mathbf{\Psi}) = \lambda_1^{(2)} \le \cdots \le \lambda_q^{(2)} = \lambda_{\max}(\mathbf{\Psi})$ be eigenvalues of $\mathbf{\Psi}$. We make an assumption on the eigenvalues throughout this section,

(C1). $c^{-1} \le \lambda_1^{(1)} \le \cdots \le \lambda_p^{(1)} \le c$ and $c^{-1} \le \lambda_1^{(2)} \le \cdots \le \lambda_q^{(2)} \le c$ for some constant $c > 0$.

The condition (C1) is a standard eigenvalue assumption in high-dimensional covariance estimation literature (see the survey Cai, Ren and Zhou (2016) and references therein). It is natural for many important classes of covariance matrices, e.g., bandable, Toeplitz, and sparse covariance matrices. The assumption (C1) implies that $1/c' \le A_p \le c'$ (see A_p in (3.10)) for some constant $c' > 0$. We first provide some key results on the properties of the test statistic and the estimator $\widehat{A}_p$ of A_p in the next subsection.

4.1. Asymptotic normality and convergence results of the proposed test statistics

The first result gives the asymptotic normality for the test statistic $\sqrt{\{(n-1)p\}/A_p}(\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}})$ in (3.9) under the null.

Proposition 1. *Assume that, as $np \to \infty$, $\log\max(q, np) = o(np)$, and the estimator $\widehat{\boldsymbol{\beta}}_j$ for $j \in [q]$ satisfies* (3.6) *with*

$$a_{n1} = o\left\{\frac{1}{\sqrt{\log\max(q, np)}}\right\} \quad \textit{and} \quad a_{n2} = o\{(np)^{-1/4}\}. \tag{4.1}$$

Under the null $H_{0ij} : \gamma_{ij} = 0$, we have, as $np \to \infty$,

$$\sqrt{\frac{(n-1)p}{A_p}}\frac{\widehat{r}_{ij}}{\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}} \Rightarrow N(0,1)$$

in distribution, where $\widehat{r}_{ii}$ *and* A_p *are defined in* (3.8) *and* (3.10)*, respectively.*

The next proposition shows that under alternatives, $\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}$ converge to a nonzero number, which indicates that our test statistics lead to a non-trivial power. Here $\rho^{\mathbf{\Gamma}}_{ij\cdot}$ is the partial correlation coefficient between X_{li} and X_{lj} (for any $1 \le l \le p$).

Proposition 2. *Suppose that conditions in Proposition* 1 *hold. We have, for* $1 \le i < j \le q$,

$$\frac{\widehat{r}_{ij}}{\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}}} - (1-\gamma_{ij}\psi_{ij})\rho^{\mathbf{\Gamma}}_{ij\cdot} \to 0$$

in probability as $np \to \infty$.

The condition in (4.1) will be established later in Proposition 4. It is interesting to see that in Propositions 1 and 2, we only require $np \to \infty$, which means that the sample size n can be a constant. This is a significant difference between the estimation of MGGMs and that of vector-variate GGMs. In the latter problem, to establish the asymptotic consistency or normality, the sample size is usually required to go to infinity in the existing literature (see, e.g., Rothman et al. (2008); Lam and Fan (2009); Liu (2013); Ren et al. (2015)).

We next establish the convergence rate for the estimator of A_p. To this end, we need an additional condition on $\mathbf{\Sigma}$.

(C2). For some $0 < \tau < 2$, $\sum_{j=1}^{p} |\sigma_{ij}|^{\tau} \le Cs(p)$ with $s(p) = 1/(\log q)^2 \left\{\sqrt{nq/\log\max(p,nq)}\right\}^{(2-\tau)\wedge 1}$ uniformly in $1 \le i \le p$.

Note that when $0 < \tau < 1$, this assumption becomes the typical weak sparsity assumption in high-dimensional covariance estimation (Bühlmann and van de Geer (2011)).

Proposition 3. *Let* $\widetilde{\lambda} = \lambda\sqrt{\log\max(p,nq)/nq}$ *with* λ *being sufficiently large. Suppose that* (C2) *holds. We have* $\widehat{A}_p/A_p = 1 + O_{\mathbb{P}}(\widetilde{\lambda}^{(2-\tau)\wedge 1}s(p))$ *as* $nq \to \infty$.

Combining Propositions 1 and 3, we have the asymptotic normality under the null for our final test statistic $\widehat{T}_{ij} = \sqrt{(n-1)p/\widehat{A}_p}(\widehat{r}_{ij}/\sqrt{\widehat{r}_{ii}\widehat{r}_{jj}})$ in (3.11).

4.2. Guarantees on FDP/FDR control

We show that the FDP and FDR of $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$ can be controlled asymptotically. To this end, we discuss the FDP and FDR of the estimation of $\mathrm{supp}(\mathbf{\Gamma})$ and $\mathrm{supp}(\mathbf{\Omega})$ separately. For the estimation of $\mathrm{supp}(\mathbf{\Gamma})$, take

$$\mathrm{FDP}_1 = \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{(i,j) \in \widehat{\mathrm{supp}(\mathbf{\Gamma})}\}}{\max\left(\sum_{1\le i<j\le q} I\{(i,j) \in \widehat{\mathrm{supp}(\mathbf{\Gamma})}\}, 1\right)}, \quad \mathrm{FDR}_1 = \mathbb{E}(\mathrm{FDP}_1), \quad (4.2)$$

where $\mathcal{H}_0 = \{(i,j) : \gamma_{ij} = 0, \quad 1 \le i < j \le q\}$. Let $\mathcal{H}_1 = \{(i,j) : \gamma_{ij} \neq 0, \quad 1 \le i < j \le q\}$. Write $\varpi_0 = Card(\mathcal{H}_0)$ as the total number of true nulls, $\varpi_1 = card(\mathcal{H}_1)$ as the number of true alternatives, and $\varpi = (q^2 - q)/2$ as the total number of hypotheses. For a constant $\gamma > 0$ and $1 \le i \le q$, let $\mathcal{A}_i(\gamma) = \{j : 1 \le j \le q, \ j \neq i, \ |\gamma_{ij}| \ge (\log q)^{-2-\gamma}\}$. Theorem 1 shows that our procedure controls FDP and FDR at a given level α asymptotically.

Theorem 1. *Let the dimension (p, q) satisfy $q \le (np)^r$ for some $r > 0$. Suppose that*

$$Card\left((i,j) : 1 \le i < j \le q, \quad |(1 - \gamma_{ij}\psi_{ij})\rho^{\mathbf{\Gamma}}_{ij\cdot}| \ge 4\sqrt{\frac{A_p \log q}{(n-1)p}}\right) \ge \sqrt{\log\log q}, (4.3)$$

where A_p is defined in (3.10)*. Assume that $\varpi_1 \le c\varpi$ for some $c < 1$ and that $\{\widehat{\boldsymbol{\beta}}_i\}_{i\in[q]}$ satisfy* (3.6) *with*

$$a_{n1} = o\left(\frac{1}{\log\max(q, np)}\right) \quad \text{and} \quad a_{n2} = o\left((np\log q)^{-1/4}\right). \qquad (4.4)$$

Under (C1), (C2), *and* $\max_{1\le i\le q} Card(\mathcal{A}_i(\gamma)) = O(q^{\vartheta})$ *for some $\vartheta < 1/2$ and $\gamma > 0$, we have* $\lim_{np,q\to\infty}\{\mathrm{FDR}_1/(\alpha\varpi_0/\varpi)\} = 1$ *and* $\{\mathrm{FDP}_1/(\alpha\varpi_0/\varpi)\} \to 1$ *in probability as $np, q \to \infty$.*

Condition (4.3) requires the number of true alternatives be at least $\sqrt{\log\log q}$. This condition is very mild and, in fact, is a nearly necessary condition for FDP control. Proposition 2.1 in Liu and Shao (2014) shows that, in large-scale multiple testing problems, if the number of true alternatives is fixed, then it is impossible for the Benjamini and Hochberg method (Benjamini and Hochberg (1995)) to control the FDP with probability tending to one at any desired level. Here (4.3) is only slightly stronger than the condition that the number of true alternatives goes to infinity. The condition on $Card(\mathcal{A}_i(\gamma))$ is a sparsity condition for $\mathbf{\Gamma}$. This condition is also quite weak. For the estimation of vector-variate GGMs, the existing literature often requires the row sparsity level of precision matrix to be less than $O(\sqrt{n})$. When the dimension q is much larger than n, our condition on $Card(\mathcal{A}_i(\gamma))$ in Theorem 1 is clearly much weaker. In Theorem 1, the sample size n can be a fixed constant as long as the dimension $p, q \to \infty$.

As in Theorem 1, we have the similar FDP and FDR control result for the estimation of $\text{supp}(\mathbf{\Omega})$. Let $\mathcal{H}_0' = \{(i,j) : \omega_{ij} = 0, \quad 1 \leq i < j \leq p\}$ and $\mathcal{H}_1' = \{(i,j) : \omega_{ij} \neq 0, \quad 1 \leq i < j \leq p\}$. Further, let $\kappa_0 = Card(\mathcal{H}_0')$, $\kappa_1 = Card(\mathcal{H}_1')$ and $\kappa = (p^2 - p)/2$. Start with

$$\text{FDP}_2 = \frac{\sum_{(i,j)\in\mathcal{H}_0'} I\left((i,j) \in \widehat{\text{supp}(\mathbf{\Omega})}\right)}{\max\left(\sum_{1\leq i<j\leq q} I\left((i,j) \in \widehat{\text{supp}(\mathbf{\Omega})}\right), 1\right)}, \quad \text{FDR}_2 = \mathbb{E}(\text{FDP}_2). \tag{4.5}$$

For a constant $\gamma > 0$ and $1 \leq i \leq p$, take $\mathcal{B}_i(\gamma) = \{j : 1 \leq j \leq p,\ j \neq i,\ |\omega_{ij}| \geq (\log p)^{-2-\gamma}\}$. Let $B_q = q\|\mathbf{\Psi}\|_{\text{F}}^2/\{\text{tr}(\mathbf{\Psi})\}^2$ and the partial correlation associated with $\mathbf{\Omega}$ be $\rho_{ij\cdot}^{\mathbf{\Omega}} = -(\omega_{ij}/\sqrt{\omega_{jj}\omega_{jj}})$ for $1 \leq i < j \leq p$. As (C2), we assume the following condition on $\mathbf{\Psi} = (\psi_{ij})_{p\times p}$,

(C3). For some $0 < \tau < 2$, assume that $\sum_{j=1}^{q} |\psi_{ij}|^\tau \leq Cs(q)$ with $s(q) = \{1/(\log p)^2\}\left\{\sqrt{np/\log\max(q, np)}\right\}^{(2-\tau)\wedge 1}$ uniformly in $1 \leq i \leq q$.

Theorem 2. *Let the dimension (p, q) satisfy $p \leq (nq)^r$ for some $r > 0$, and that*

$$Card\left\{(i,j) : 1 \leq i < j \leq p, |(1 - \omega_{ij}\sigma_{ij})\rho_{ij\cdot}^{\mathbf{\Omega}}| \geq 4\sqrt{\frac{B_q \log p}{(n-1)q}}\right\} \geq \sqrt{\log\log p}. \tag{4.6}$$

Assume that $\kappa_1 \leq c\kappa$ for some $c < 1$ and $\{\widehat{\boldsymbol{\beta}}_i\}_{i\in[p]}$ satisfies (3.6) *with*

$$a_{n1} = o\left(\frac{1}{\log\max(p, nq)}\right) \quad \textit{and} \quad a_{n2} = o((nq\log p)^{-1/4}). \tag{4.7}$$

Under (C1), (C3), *and* $\max_{1\leq i\leq p} Card(\mathcal{B}_i(\gamma)) = O(p^\vartheta)$ *for some* $\vartheta < 1/2$ *and* $\gamma > 0$, *we have* $\lim_{nq,p\to\infty}\{\text{FDR}_2/(\alpha\kappa_0/\kappa)\} = 1$ *and* $\{\text{FDP}_2/(\alpha\kappa_0/\kappa)\} \to 1$ *in probability as* $nq, p \to \infty$.

By Theorems 1 and 2, we can obtain the FDP and FDR result of the estimator $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$. In particular, let a_0 and a be the number of false discoveries and total discoveries in $\widehat{\text{supp}(\mathbf{\Omega})}$, excluding the diagonal entries, b_0 and b be the number of false discoveries and total discoveries in $\widehat{\text{supp}(\mathbf{\Gamma})}$, excluding the diagonal entries. It is then easy to calculate that the number of false discoveries in $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$ is $a_0(q + b) + (a - a_0)b_0 + pb_0$, and the number of total discoveries in $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$ is $pb + a(q + b)$ (excluding the diagonal entries). We have the FDP and FDR of $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$:

$$\text{FDP} = \frac{a_0(q + b) + (a - a_0)b_0 + pb_0}{\max(pb + a(q + b), 1)}, \quad \text{FDR} = \mathbb{E}(\text{FDP}). \tag{4.8}$$

The true FDP in (4.8) cannot be computed in practice since the number of false discoveries a_0 and b_0 are unknown. One straightforward estimator for FDP is to replace the unknown quantities a_0 and b_0 with αa and αb, respectively, which leads to the following FDP estimator α':

$$\alpha' = \frac{\alpha\{(2-\alpha)ab + aq + bp\}}{\max(ab + aq + bp, 1)}. \tag{4.9}$$

Note the values of a and b in (4.9) are known, which represent the number of total discoveries in $\widehat{\text{supp}(\boldsymbol{\Omega})}$ and $\widehat{\text{supp}(\boldsymbol{\Gamma})}$, respectively. The FDP estimator α' takes the value in $[0, 1]$ and is monotonically increasing as a function of α. In the next theorem, we show that $\text{FDP}/\alpha' \to 1$ in probability as $p, q \to \infty$.

Theorem 3. *Under the conditions of Theorems* 1 *and* 2 *with the sparsity condition* $\omega_1 = o(\omega)$ *and* $\kappa_1 = o(\kappa)$*, we have* $\text{FDP}/(\alpha') \to 1$ *in probability as* $p, q \to \infty$.

Theorem 3 shows that the FDP of the proposed estimator $\widehat{\text{supp}(\boldsymbol{\Omega})} \otimes \widehat{\text{supp}(\boldsymbol{\Gamma})}$ can be estimated consistently by α'. By Theorems 1 and 2, the sparsity conditions $\omega_1 = o(\omega)$ and $\kappa_1 = o(\kappa)$ imply that $\text{FDP}_1/\alpha \to 1$ and $\text{FDP}_2/\alpha \to 1$ in probability when $p, q \to \infty$. Therefore, one can replace a_0 and b_0 in FDP in (4.8) by αa and αb, respectively, and achieve the result in Theorem 3. In fact, we can still obtain the guarantee of FDP of $\widehat{\text{supp}(\boldsymbol{\Omega})} \otimes \widehat{\text{supp}(\boldsymbol{\Gamma})}$ even without the sparsity conditions $\omega_1 = o(\omega)$ and $\kappa_1 = o(\kappa)$. For any $\varepsilon > 0$, Theorems 1 and 2 show that $\mathbb{P}(\text{FDP}_1 \le \alpha(1+\varepsilon)) \to 1$ and $\mathbb{P}(\text{FDP}_2 \le \alpha(1+\varepsilon)) \to 1$ as $p, q \to \infty$ regardless of the sparsity conditions. This further implies the guarantee on the FDP of $\widehat{\text{supp}(\boldsymbol{\Omega})} \otimes \widehat{\text{supp}(\boldsymbol{\Gamma})}$ for any $\varepsilon > 0$:

$$\mathbb{P}\left(\frac{\text{FDP}}{\alpha\{(2ab + aq + bp)/\max(ab + aq + pb, 1)\}} \le 1 + \varepsilon\right) \to 1, \text{ as } p, q \to \infty.$$

4.3. Power analysis

We study the statistical power of the proposed method by considering the following class of alternatives. We assume that for some $c > 4$,

$$|\rho^{\boldsymbol{\Gamma}}_{ij\cdot}| = c\sqrt{\frac{A_p \log q}{(n-1)p}} \text{ and } |\rho^{\boldsymbol{\Omega}}_{kl\cdot}| = c\sqrt{\frac{B_q \log p}{(n-1)q}}, \quad (i,j) \in \mathcal{H}_1, \ (k,l) \in \mathcal{H}'_1. \tag{4.10}$$

We will show in the next theorem that the power of the support estimators will converge to 1 as $p, q \to \infty$.

Theorem 4. *Let the dimension* (p, q) *satisfy* $p \le (nq)^r$ *and* $q \le (np)^r$ *for some* $r > 0$. *Assume that* (C1)-(C3), (4.4) *and* (4.7) *hold. We have* $\text{supp}(\boldsymbol{\Gamma}) \subseteq \widehat{\text{supp}(\boldsymbol{\Gamma})}$, $\text{supp}(\boldsymbol{\Omega}) \subseteq \widehat{\text{supp}(\boldsymbol{\Omega})}$ *and* $\text{supp}(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}) \subseteq \widehat{\text{supp}(\boldsymbol{\Omega})} \otimes \widehat{\text{supp}(\boldsymbol{\Gamma})}$ *with probability tending to one as* $p, q \to \infty$.

Recall that the power is defined by the ratio between the number of true discoveries in $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$ and the total number of non-zero off-diagonals in $\text{supp}(\mathbf{\Omega} \otimes \mathbf{\Gamma})$. Thus, Theorem 4 shows that the power converges to 1 as $p, q \to \infty$. In addition, Theorem 4 shows that to detect the edge between X_{ij} and X_{kl}, the corresponding partial correlation $\varrho_{ij,kl} = \rho^{\mathbf{\Gamma}}_{ij\cdot} \cdot \rho^{\mathbf{\Omega}}_{kl\cdot}$ can be as small as $C\{1/(n-1)\}\sqrt{\log p \log q/pq}$ (A_p and B_q are bounded, see assumption (C1)). This is essentially different from the estimation of vector-variate GGMs. If we apply the method of estimation of vector-variate GGMs to $\text{vec}(\mathbf{X})$ directly (e.g., the method from Ren et al. (2015)), even for an individual test (detecting a single edge), the magnitude of the partial correlation $\varrho_{ij,kl}$ needs to be $C(1/\sqrt{n})$.

4.4. Convergence rate of the initial estimators of regression coefficients

Finally, the next proposition shows that the convergence rate condition of $\widehat{\boldsymbol{\beta}}_j$ in (4.1) and (4.4) can be satisfied under some regular conditions. The convergence rate condition in (4.7) can be established similarly. This result establishes the consistency of Lasso for correlated samples, which in itself is interesting.

Proposition 4. *Let δ in* (3.16) *be large enough. Suppose that* (C1) *holds and* $\max_{1\le j\le q} |\boldsymbol{\beta}_j|_0 = o(\sqrt{np}/\{\log \max(q, np)\}^{3/2})$. *We have $\widehat{\boldsymbol{\beta}}_j(\delta)$ for $1 \le j \le q$ are consistent in both ℓ_1 and ℓ_2 norms with the rate in* (4.4).

5. Numerical Results

In this section, we present numerical results on simulations and real data to investigate the performance of the proposed method on support recovery of matrix-variate data. In our experiment, we adopted the data-driven parameter-tuning approach from Liu (2013) to tune the parameters (see Section E in the supplement for details). Due to space constraints, some simulated experimental results and real data analysis are provided in Section E of the supplement.

5.1. Simulated experiments

In the simulations, we constructed $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ based on combinations of following graph structures used in Liu (2013).

1. Hub graph ("hub" for short). There are $p/10$ rows with sparsity 11. The rest of the rows have sparsity 2. We took $\mathbf{\Omega}_1 = (\omega_{ij}), \omega_{ij} = \omega_{ji} = 0.5$ for $i = 10(k-1)+1$ and $10(k-1)+2 \le j \le 10(k-1)+10, 1 \le k \le p/10$. The diagonal $\omega_{ii} = 1$ and other entries are zero. We also took $\mathbf{\Omega} = \mathbf{\Omega}_1 + (|\min(\lambda_{\min})| + 0.05)\mathbf{I}_p$ to make the matrix positive definite.

Table 1. Averaged empirical FDP, the estimated FDP/FDR level α' in (4.9) and power.

p	q	$\mathbf{\Omega}$	$\mathbf{\Gamma}$	$n = 20$		$n = 100$	
				FDP (α')	Power	FDP (α')	Power
100	100	hub	hub	0.192 (0.146)	1.000	0.155 (0.145)	1.000
		hub	band	0.158 (0.152)	1.000	0.146 (0.152)	1.000
		hub	random	0.188 (0.154)	0.916	0.156 (0.154)	1.000
		band	band	0.138 (0.161)	1.000	0.154 (0.162)	1.000
		band	random	0.152 (0.164)	0.998	0.127 (0.163)	1.000
		random	random	0.161 (0.164)	0.834	0.104 (0.165)	0.999
200	200	hub	hub	0.183 (0.146)	1.000	0.145 (0.145)	1.000
		hub	band	0.149 (0.152)	1.000	0.144 (0.152)	1.000
		hub	random	0.167 (0.154)	0.981	0.153 (0.154)	1.000
		band	band	0.138 (0.162)	1.000	0.148 (0.162)	1.000
		band	random	0.158 (0.164)	1.000	0.134 (0.163)	1.000
		random	random	0.171 (0.166)	0.980	0.134 (0.166)	1.000
200	50	hub	hub	0.166 (0.145)	1.000	0.154 (0.145)	1.000
		hub	band	0.159 (0.152)	1.000	0.151 (0.152)	1.000
		hub	random	0.138 (0.154)	0.991	0.134 (0.153)	1.000
		band	band	0.127 (0.161)	1.000	0.146 (0.161)	1.000
		band	random	0.200 (0.163)	0.894	0.120 (0.163)	0.992
		random	random	0.194 (0.162)	0.714	0.141 (0.165)	0.980
400	400	hub	hub	0.160 (0.145)	1.000	0.141 (0.145)	1.000
		hub	band	0.146 (0.152)	1.000	0.144 (0.152)	1.000
		hub	random	0.172 (0.154)	0.999	0.151 (0.154)	1.000
		band	band	0.169 (0.162)	1.000	0.148 (0.162)	1.000
		band	random	0.159 (0.164)	1.000	0.142 (0.164)	1.000
		random	random	0.180 (0.166)	1.000	0.147 (0.166)	1.000

2. Band graph ("band" for short). $\mathbf{\Omega} = (\omega_{ij})$, where $\omega_{ii} = 1$, $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$, $\omega_{ij} = 0$ for $|i-j| \geq 3$.

3. Erdös-Rényi random graph ("random" for short). There is an edge between each pair of nodes with probability $\min(0.05, 5/p)$ independently. We took $\mathbf{\Omega}_1 = (\omega_{ij})$, $\omega_{ii} = 1$ and $\omega_{ij} = u_{ij} * \delta_{ij}$ for $i \neq j$, where $u_{ij} \sim U(0.4, 0.8)$ is a uniform random variable and δ_{ij} is a Bernoulli random variable with success probability $\min(0.05, 5/p)$; u_{ij} and δ_{ij} are independent. We also let $\mathbf{\Omega} = \mathbf{\Omega}_1 + (|\min(\lambda_{\min})| + 0.05)\mathbf{I}_p$ so that the matrix was positive definite.

The matrix $\mathbf{\Gamma}$ was also constructed from one of these graph structures.

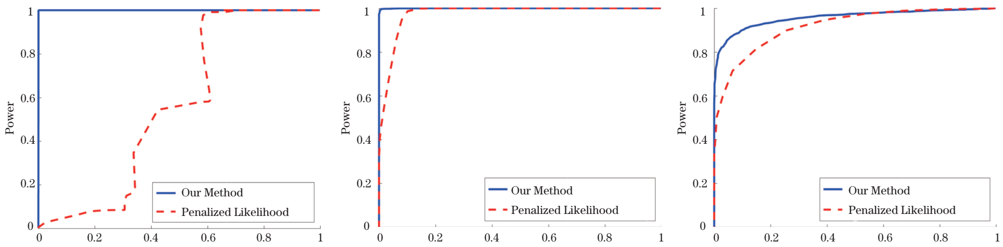
For each combination of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$, we generated n ($n = 20$ or $n = 100$) samples $(\mathbf{X}^{(k)})_{k=1}^{n}$, where each $\mathbf{X}^{(k)} \sim N_{p,q}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$ with $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ and $\mathbf{\Psi} = \mathbf{\Gamma}^{-1}$. We considered different settings of p and q: $(p,q) = (100, 100)$, $(p,q) = (200, 50)$, $(p,q) = (200, 200)$ and $(p,q) = (400, 400)$. The FDR level α for the support recovery of $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ is set to 0.1 (the observations for other α's are similar and thus omitted for space considerations). The parameters λ and δ were tuned using the data-driven approach in (23). All simulations are based on 100 independent replications.

In Table 1, we report the averaged true FDP in (4.8), the FDP estimator α' in (4.9) and the power for estimating $\text{supp}(\mathbf{\Omega} \otimes \mathbf{\Gamma})$ over 100 replications. On the one hand, according to Theorem 3, it is desirable that the true FDP be close to α'. On the other hand, we aim for a large power. In particular, the power of the estimator $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$ can be calculated as follows. Let A and B be the number of nonzero off-diagonals in $\mathbf{\Omega}$ and $\mathbf{\Gamma}$. From the definition of a_0, a, b_0, b in (4.8), we have

$$\begin{aligned} \text{Power} &= \frac{\{pb + a(q+b)\} - \{a_0(q+b) + (a-a_0)b_0 + pb_0\}}{pB + A(q+B)} \\ &= \frac{p(b-b_0) + (a-a_0)(q+b-b_0)}{pB + A(q+B)}, \end{aligned} \tag{5.1}$$

where the numerator is the number of true discoveries in $\widehat{\text{supp}(\mathbf{\Omega})} \otimes \widehat{\text{supp}(\mathbf{\Gamma})}$ and the denominator is the total number of nonzero off-diagonals in $\text{supp}(\mathbf{\Omega} \otimes \mathbf{\Gamma})$. From Table 1, in all settings with $n = 100$, the true FDPs are close to their estimates α' and the powers are all very close to 1. For $n = 20$, a more challenging case due to the small sample size, the true FDPs are still controlled by their estimates α' for most graphs. When $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ are both hub or random graphs, the true FDPs are slightly larger than the corresponding estimates. In terms of power with $n = 20$, when $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ are both generated from random graphs and either p or q is small (e.g., $p = q = 100$ or $p = 200, q = 50$) the powers could be away from 1 (but still above 0.7); for all other cases, the powers are close to 1. We examined the cases in which the power is much less than one and found that our FDP procedure generates overly sparse estimators, which leads to lower powers. In fact, a lower power for a small n and p (or q) is expected since we essentially use np correlated samples to estimate $\text{supp}(\mathbf{\Gamma})$ and nq correlated samples to estimate $\text{supp}(\mathbf{\Omega})$.

Due to space constraints, we relegate further simulation studies to the supplement. In Section E.1 of the supplement, we present the boxplots of FDPs over 100 replications. The plots show that FDPs are well concentrated, which suggests that the performance of the proposed estimator is quite stable. In Section E.2, we provide experimental results on the estimation of $\widehat{A}_p$ that empirically verify our theoretical result in Proposition 3. In Section E.3, we compare our procedure with the penalized likelihood approach in Leng and Tang (2012). When p, q are



(a) f = 1 (b) f = 2 (c) f = 3

Figure 1. ROC curves for different signal strength when $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ are bandable matrices.

small as compared to n, the penalized likelihood approach still achieves good support recovery performance (e.g., the case $n = 100, p = q = 20$ as reported in Leng and Tang (2012)). When p, q are comparable to or larger than n, our testing based method achieves better support recovery performance. In Section E.4, we provide some empirical evidences to show that the de-correlation approach in Remark 1 cannot control FDP well. In Section E.5, we further present simulation studies when the covariance matrix does not follow the form of a Kronecker product.

5.1.1. ROC curves

We make comparisons between our method and the penalized likelihood approach in Leng and Tang (2012) (with the SCAD penalty) in terms of the ROC curve. We constructed the precision matrices $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ using the graph structures in Section 5.1 while introducing an additional factor f to tune the signal strength. In particular, for the hub graph, we set $\mathbf{\Omega}_1 = (\omega_{ij})$ with $\omega_{ij} = \omega_{ji} = 0.5/f$, for the band graph, we set $\omega_{i,i+1} = \omega_{i+1,i} = 0.6/f$ and $\omega_{i,i+2} = \omega_{i+2,i} = 0.3/f$, and for the random graph, we chose $u_{ij} \sim U(0.4/f, 0.8/f)$. When $f = 1$, the construction of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ is the same as that in Section 5.1. The higher the value of f, the weaker the signal strength. Due to space constraints, we only report the comparison when $n = 20$, $p = q = 100$ and observations are similar for other settings of n, p, and q.

We compared the support recovery performance for different signal strengths by varying $f = 1, 2, 3$ and fixed $\mathbf{\Omega}$ and $\mathbf{\Gamma}$ to be bandable matrices. The usual ROC curve for binary classification is based on false positive rate (a.k.a. 1$-$ specificity) vs true positive rate (a.k.a. sensitivity or power). In our high-dimensional setting, $\mathbf{\Omega} \otimes \mathbf{\Gamma}$ is a highly sparse matrix and thus the false positive rate is extremely small for any reasonable choice of α or regularization parameter that gives a small number of discoveries. Therefore, we choose to report the ROC curve in terms of FDP vs power, from which one can easily compare powers for different methods under the same level of FDP. As seen in Figure 1, our method achieves better performance than the method in Leng and Tang (2012) for different signal strengths. When the factor $f = 3$, the ROC curve of our method is still almost vertical. In Section E.6 in the supplement, we fix the factor $f = 3$ and consider different types of $\mathbf{\Omega}$ and $\mathbf{\Gamma}$. For most cases, our method still achieves better performance.

5.2. Real data analysis

For real data analysis, we investigated the performance of the proposed method on two datasets: the U.S. agricultural export data from Leng and Tang (2012) and the climatological data from Lozano et al. (2009). Due to space constraints, the details of the data analysis are provided in Section E.7 in the supplement.

6. Discussions and future work

In this paper, we propose new test statistics with FDR control guarantees for graph estimation from matrix-variate Gaussian data. To handle the correlation structure among “row samples” and “column samples”, we develop the *variance correlation technique*. This variance correlation technique can be directly extended to address the problem of learning high-dimensional GGMs with correlated samples, which has not been studied in the existing literature but with many important applications in practice. We leave this extension as future work.

To establish the FDR control result, the correlation among “row samples” makes the theoretical analysis significantly more challenging than the *i.i.d.* case and all the analysis in the *i.i.d.* case must be carefully tailored. For example, we need to establish the consistency for Lasso estimators from correlated samples. We also need a few new large deviation bounds on sample covariance matrices with correlated samples (see Section A in the supplement).

There are several future directions to pursue. Although our paper mainly focuses on the support recovery and graph estimation, it is also interesting to estimate the Kronecker product precision matrix based on the multiple testing framework. While our work relies on the Kronecker product structure, it is interesting to consider other forms of covariance matrices, e.g., the true covariance matrix does not exactly follow Kronecker product structure but is close to that structure.

Supplementary Materials

The supplementary material consists of several technical lemmas and the proofs of our propositions and theorems. Moreover, it includes additional simulation studies and real data analysis.

Acknowledgment

The authors are very grateful to two anonymous referees and the associate editor for their detailed and constructive comments that considerably improved the quality of this paper. We would also like to thank Yichen Zhang for helpful discussions and Chenlei Leng and Cheng Yong Tang for sharing the code of Leng and Tang (2012) and U.S. export data. Xi Chen's research was supported by Alibaba innovation research award and Bloomberg data science research award. Weidong Liu's research was supported by NSFC Grant No.11431006 and 11690013, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shanghai Shuguang Program, Youth Talent Support Program, 973 Program (2015CB856004) and a grant from Australian Research Council. Weidong Liu is the corresponding author of the paper.

References

Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics* **4**, 764–790.

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 (Statistical Methodology)* **57**, 389–300.

Bickle, P. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.

Bijma, F., De Munck, J. and Heethaar, R. (2005). The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. *NeuroImage* **27**, 402–415.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data – Methods, Theory and Applications.* Springer.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.

Cai, T., Ren, Z. and Zhou, H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics* **10**, 1–89.

Cai, T. T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313–2351.

Chen, S. X. and Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

d'Aspremont, A., Banerjee, O. and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications* **30**, 56–66.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika* **68**, 265–274.

Efron, B. (2009). Are a set of microarrays independent of each other? *Annals of Applied Statistics* **3**, 922–942.

Fan, Y. and Lv, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics* **44**.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix Variate Distributions*. Chapman Hall.

Huang, F. and Chen, S. (2015). Joint learning of multiple sparse matrix Gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems* **26**, 2606–2620.

Kalaitzis, A., Lafferty, J., Lawrence, N. D. and Zhou, S. (2013). The bigraphical lasso. In *Proceedings of the 30th International Conference on Machine Learning*.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* **37**, 4254–4278.

Leng, C. and Tang, C. Y. (2012). Sparse matrix graphical models. *Journal of the American Statistical Association* **107**, 1187–1200.

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40**, 2293–2326.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics* **41**, 2948–2978.

Liu, W. and Shao, Q. M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics* **42**, 2003–2025.

Lozano, A. C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J. and Abe, N. (2009). Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ma, S., Gong, Q. and Bohnert, H. J. (2007). An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research* **17**, 1614–1625.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5**, 935–980.

Ren, Z., Kang, Y., Fan, Y. and Lv, J. (2016). Tuning-free heterogeneity pursuit in massive networks. ArXiv preprint arXiv:1606.03803.

Ren, Z., Sun, T., Zhang, C. H. and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *The Annals of Statistics* **43**, 991–1026.

Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

Schafer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754–764.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**, 267–288.

Tsiligkaridis, T., Hero, A. O. and Zhou, S. (2013). Convergence properties of kronecker graphical Lasso algorithms. *IEEE Transactions on Signal Processing* **61**, 1743–1755.

Van de Geer, S., Bhlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42**, 1166–1202.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics* **40**, 2541–2571.

Yin, J. and Li, H. (2012). Model selection and estimation in matrix normal graphical model. *Journal of Multivariate Analysis* **107**, 119–140.

Ying, Y. and Liu, H. (2013). High-dimensional semiparametric bigraphical models. *Biometrika* **100**, 655–670.

Yuan, M. (2010). Sparse inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research* **11**, 2261–2286.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics* **42**, 532–562.

Zhu, Y., Shen, X. T. and Pan, W. (2014). Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association* **109**, 1683–1696.

Department of Information, Operations & Management Sciences, Stern School of Business, New York University, New York, NY 10003, USA.

E-mail: xchen3@stern.nyu.edu

Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, Shanghai, 200240, China.

E-mail: weidongl@sjtu.edu.cn.

(Received February 2017; accepted August 2017)