Hypothesis Testing of Matrix Graph Model with Application to Brain Connectivity Analysis

Yin Xia\textsuperscript{1,2,*} and Lexin Li\textsuperscript{3,**}

\textsuperscript{1}Department of Statistics, School of Management, Fudan University, Shanghai 200433, China
\textsuperscript{2}Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.
\textsuperscript{3}Division of Biostatistics, University of California at Berkeley, Berkeley, California 94720, U.S.A.

\textit{Summary.} Brain connectivity analysis is now at the forefront of neuroscience research. A connectivity network is characterized by a graph, where nodes represent neural elements such as neurons and brain regions, and links represent statistical dependence that is often encoded in terms of partial correlation. Such a graph is inferred from the matrix-valued neuroimaging data such as electroencephalography and functional magnetic resonance imaging. There have been a good number of successful proposals for sparse precision matrix estimation under normal or matrix normal distribution; however, this family of solutions does not offer a direct statistical significance quantification for the estimated links. In this article, we adopt a matrix normal distribution framework and formulate the brain connectivity analysis as a precision matrix hypothesis testing problem. Based on the separable spatial-temporal dependence structure, we develop oracle and data-driven procedures to test both the global hypothesis that all spatial locations are conditionally independent, and simultaneous tests for identifying conditional dependent spatial locations with false discovery rate control. Our theoretical results show that the data-driven procedures perform asymptotically as well as the oracle procedures and enjoy certain optimality properties. The empirical finite-sample performance of the proposed tests is studied via intensive simulations, and the new tests are applied on a real electroencephalography data analysis.

\textit{Key words:} Brain connectivity analysis; False discovery rate; Gaussian graphical model; Matrix-variate normal distribution; Multiple testing.

1. Introduction

Matrix-valued data are recently becoming ubiquitous in a wide range of scientific and business applications (Allen and Tibshirani, 2010, 2012; Reiss and Ogden, 2010; Aston and Kirch, 2012; Leng and Tang, 2012; Yin and Li, 2012; Tsiligkaridis et al., 2013; Zhou and Li, 2014, among others). Accordingly, the matrix normal distribution is becoming increasingly popular in modeling the matrix-variate observations (Zhou, 2014). Our motivating example is an electroencephalography (EEG) study of 77 alcoholic individuals and 45 controls. For each subject, we recorded the voltage values from 61 electrodes placed at various scalp locations at a rate of 256 Hz for 1 second, resulting a $61 \times 256$ matrix. The scientific objective is to infer the connectivity patterns among 61 spatial locations for both alcoholic and control groups. This study embodies a more general class of applications of mapping brain connectivity networks, which is now at the forefront of neuroscience research. The overarching goal is to infer the brain network, characterized by a graph, where the nodes represent neural elements such as neurons and brain regions, and the links encode statistical dependence between those elements. Partial correlations, conveyed by a precision matrix, are frequently employed to describe such statistical dependence (Fornito et al., 2013), and this precision matrix, in turn, is to be derived from matrix-valued imaging data, such as EEG and functional magnetic resonance imaging.

Adopting matrix normal distribution framework, we formulate brain connectivity network analysis as a precision matrix inference problem. Specifically, let $X \in \mathbb{R}^{p \times q}$ denote the spatial-temporal matrix data from an image modality, for example, EEG. It is assumed to follow a matrix normal distribution with the Kronecker product covariance structure, $\text{Cov}(\text{vec}(X)) = \Sigma_s \otimes \Sigma_t$, where $\text{vec}(X)$ stacks the columns of the matrix $X$ to a vector, $\otimes$ is the Kronecker product, $\Sigma_s \in \mathbb{R}^{p \times p}$ denotes the covariance matrix of $p$ spatial locations, $\Sigma_t \in \mathbb{R}^{q \times q}$ denotes the covariance matrix for $q$ time points. Correspondingly,

$$\text{Cov}^{-1}(\text{vec}(X)) = \Sigma_s^{-1} \otimes \Sigma_t^{-1} = \Omega_s \otimes \Omega_t,$$

where $\Omega_s \in \mathbb{R}^{p \times p}$ and $\Omega_t \in \mathbb{R}^{q \times q}$ represent the spatial and temporal precision matrix, respectively. In brain connectivity analysis, our primary interest is to infer the connectivity network characterized by the spatial precision matrix $\Omega_s$. By contrast, the temporal precision matrix $\Omega_t$ is of little interest here and is to be treated as a nuisance parameter. We also make some remarks regarding the assumptions of our adopted framework. First, the matrix normal assumption has
been frequently adopted in numerous applications (Leng and Tang, 2012; Yin and Li, 2012), and is scientifically plausible in neuroimaging analysis. For instance, the majority standard neuroimaging processing software, such as SPM (Friston et al., 2007) and FSL (Smith et al., 2004), adopt a framework that assumes the data are normally distributed per voxel (location) with a noise factor and an autoregressive structure, which shares a similar spirit as the matrix normal formulation. Second, it is commonly assumed that the precision matrix is sparse, which we also adopt for our inferential procedure. Again, this sparsity assumption is scientifically justifiable, as it is known that brain region connections are energy consuming (Raichle and Gusnard, 2002; Fox and Raichle, 2007), and biological units tend to minimize energy-consuming activities (Bullmore and Sporns, 2009).

In this article, we aim to address the following two hypothesis testing questions. The first is to test if all spatial locations are conditionally independent, namely we test

\[ H_0: \Omega_S \text{ is diagonal versus } H_1: \Omega_S \text{ is not diagonal.} \quad (1) \]

The second is to identify those conditionally dependent pairs of locations with false discovery rate (FDR) and false discovery proportion (FDP) control; that is, we simultaneously test

\[ H_{0,i,j} : \omega_{S,i,j} = 0 \text{ versus } H_{1,i,j} : \omega_{S,i,j} \neq 0, \quad \text{for } 1 \leq i < j \leq p, \quad (2) \]

where \( \omega_{S,i,j} \) is the \((i,j)\)th element of \( \Omega_S \).

In the literature, there have been a good number of methods proposed to estimate a sparse precision matrix under normal distribution Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Friedman et al. (2008); Ravikumar et al. (2011); Cai et al. (2011). There are extensions to multiple precision matrices (Danaher et al., 2014; Zhu et al., 2014), nonparanormal distribution (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Cai et al., 2011; Ravikumar et al., 2011), and matrix-valued graph model (Leng and Tang, 2012; Yin and Li, 2012; Zhou, 2014). However, all those methods tackle the estimation aspect of the problem and induce a network from an estimated precision matrix. Only recently, hypothesis testing procedures are emerging (Drton and Perlman, 2007; Liu, 2013; Chen and Liu, 2015; Narayan et al., 2015; Xia et al., 2015).

We aim at hypothesis testing for the spatial precision matrix \( \Omega_S \) under the matrix normal framework. We separate the spatial and temporal dependence structures, and consider two scenarios. One is to assume the temporal covariance \( \Sigma_T \) is known, and we term the method as an oracle procedure. The other is to use a data-driven approach to estimate and plug in \( \Sigma_T \), and accordingly, we term it a data-driven procedure. We construct test statistics based on the covariances between the residuals from the inverse regression models. We develop a global test for (1) based on the derived limiting null distribution, and show it is particularly powerful against sparse alternatives. We then develop a multiple testing procedure for (2), and show the FDR and FDP are controlled asymptotically as well as the oracle procedures, and enjoy certain optimality under some regularity conditions. Our intensive numerical analysis supports such findings.

Our contributions and novelty are multi-fold. First, brain connectivity analysis is now becoming a central goal of neuroscience research (Fornito et al., 2013), and it constantly calls for statistical significance quantification of the inferred connection between neural elements. However, there is a paucity of systematic hypothesis testing solutions developed for this type of problems, and our proposal offers a timely response. Second, there have been some successful proposals of matrix data network estimation, most notably, Leng and Tang (2012); Yin and Li (2012); Zhou (2014). That class of solutions and our proposal can both produce, in effect, a sparse representation of the network structure. However, the former tackles network estimation, whereas our method targets network hypothesis testing, which is an utterly different problem than estimation. The key of estimation is to seek a bias-variance tradeoff, and many common sparse graph estimators are actually biased. Such methods do not produce a direct quantification of statistical significance for individual network edges. Although they often enjoy a high true positive discovery rate (power), there is no explicit control of false positive rate (significance level). By contrast, our hypothesis testing solution starts with an nearly unbiased estimator, produces an explicit significance quantification, and achieves a high power under an explicit significance level. Third, our proposal is also distinctly different from the few existing graph model hypothesis testing solutions. In particular, Liu (2013) developed one-sample testing procedures under a vector normal distribution. Directly applying Liu (2013) in our context is equivalent to assuming the columns of the spatial-temporal matrix \( X \in \mathbb{R}^{n \times q} \) are i.i.d. vector normal. This is clearly not true as the measurements at different time points are temporally highly correlated. A preprocessing step of whitening as suggested in Narayan and Allen (2015) can help induce independent columns, which basically uses \( n \) samples to estimate a \( q \times q \) temporal covariance matrix at each spatial location. By contrast, our testing procedures are built upon a linear transformation of the data \( X \Sigma_T^{-1/2} \), and pool the \( np \) correlated samples to estimate \( \Sigma_T \). Crucial to our solution is that the pooled estimator of \( \Sigma_T \) guarantees the required convergence rate, which in turn ensures the data-driven procedures perform asymptotically as well as the oracle procedures as if \( \Sigma_T \) were known. Conventional whitening, however, does not guarantee this convergence rate, and furthermore, is computationally more expensive. See also Remark 1 in Section 2.1.

In Section 4, we numerically compare with Liu (2013), with and without whitening, and demonstrate the clear advantage of our solution. In a more recent parallel work, Chen and Liu (2015) studied a similar one-sample testing problem of matrix-variate graphs. However, their work differs from ours in several ways. We study both the global test and entry-wise test and focus on the spatial precision matrix inference, whereas Chen and Liu (2015) only considered the entry-wise test and targeted both spatial and temporal precision matrices. Our solution is invariant with respect to a constant factor of the estimation of \( \Sigma_T \), and thus we do not require a strict estimate as in Chen and Liu (2015). Moreover, in their construction of test statistics, Chen and Liu (2015) required both
a bias correction and a variance correction. The variance correction involves estimation of the Frobenius norm and the trace of $\Sigma_T$, and can be challenging when the spatial dimension $p$ is large. Our method, however, separates the spatial and temporal structures and treats $\Sigma_T$ as a nuisance. The resulting test statistics do not require any variance correction. See also Remark 4 in Section 2.2. Finally, both Xia et al. (2015) and Narayan et al. (2015) studied the two-sample testing, whereas Xia et al. (2015) only considered the vector normal case, and Narayan et al. (2015) turned the matrix data into vector normal by whitening, and used bootstrap resampling method for inference.

The rest of the article is organized as follows. Section 2 develops the testing procedures, and Section 3 studies their asymptotic properties. Section 4 presents intensive simulation studies and Section 5 analyzes the motivating EEG data. All technical conditions and proofs, along with additional simulations, are relegated to the online supplement.

2. Methodology

2.1. Separation of Spatial-Temporal Dependency

Let $X_k \in \mathbb{R}^{p \times q}$, $k = 1, \ldots, n$, denote $n$ i.i.d. random samples from a matrix normal distribution with mean zero and covariance matrix $\Sigma = \Sigma_S \otimes \Sigma_T$. Note that $\Sigma_S$ and $\Sigma_T$ are not identifiable; however, our aim is to identify the set of nonzero entries of the spatial graph, and this set is invariant up to a constant. Motivated by brain functional connectivity analysis, our goal is to infer about $\Omega_S = \Sigma_S^{-1}$, while treating $\Sigma_T$ or equivalently $\Sigma_T$ as a nuisance. Our first step is to separate the spatial and temporal dependence structures. We build the test statistics based on the linear transformation of the original samples, $\{X_k, \Sigma_T^{-1/2}, k = 1, \ldots, n\}$, first assuming the temporal covariance $\Sigma_T$ is known, then plugging in a legitimate estimator $\hat{\Sigma}_T$. There are multiple ways to estimate $\Sigma_T$, or $\Sigma_T^{-1}$, as long as the resulting estimator satisfies the regularity condition (C4) as given in Web Appendix A. Examples include the usual sample covariance estimator, the banded estimator (Bickel and Levina, 2008), and the adaptive thresholding estimator (Cai and Liu, 2011).

Remark 1. Our separation step through a pooled estimator of $\Sigma_T$ can be viewed, conceptually, as matrix-variate whitening. However, it notably differs from the conventional whitening. Crucial to our proposed solution is that our test statistics are built upon a nearly unbiased estimator of $\omega_{k,i,j}$, with a convergence rate $o_p((nq \log p)^{-1/2})$ as specified in Lemma 2 in Web Appendix B. For the data-driven testing procedures, to guarantee the same convergence rate and to ensure nearly the same asymptotic properties as the oracle procedures, we require the estimator of $\Sigma_T$ to satisfy the regularity condition (C4), in terms of certain convergence rate on some norm of $\Sigma_T^{-1/2} - c\Sigma_T^{-1/2}$, with $c > 0$ a constant. Our estimator by pooling $np$ samples satisfies this requirement. Due to the correlations among the $np$ samples, the pooled estimator of $\Sigma_T$ is unbiased only up to a constant. However, our test statistics, by construction, are not affected by this constant. By contrast, the conventional whitening procedure seeks an unbiased estimator of the temporal covariance based on the $n$ samples, which does not satisfy the estimation rate of (C4). Consequently, it can not guarantee the asymptotic performance of the data-driven testing procedures.

2.2. Test Statistics

We first develop test statistics in the oracle case when $\Sigma_T$ is known. The development of the data-driven statistics with an estimated $\Sigma_T$ is similar, so we omit the details but point out the difference between the oracle and data-driven cases. For simplicity, we also use the same set of notations for the two scenarios, and only differentiate them when we study their respective asymptotic properties in Section 3.

It is well established that, under the normal distribution, the precision matrix can be described in terms of regression models (Anderson, 2003, Section 2.5). Specifically, let $Y_k = X_k \Sigma_T^{-1/2}$, $k = 1, \ldots, n$, denote the transformed samples, we have,

$$Y_{k,i,j} = \tilde{Y}_{k,i,j} + \epsilon_{k,i,j}, \quad 1 \leq i \leq p, 1 \leq j \leq q.$$  

where $\epsilon_{k,i,j} \sim N(0, \sigma_{S,i,j} - \Sigma_{k,i,j}^{-1} \Sigma_{k,i,j}^{-1} \Sigma_{k,i,j}^{-1} \Sigma_{k,i,j}^{-1})$ is independent of $Y_{k,i,j}$, and the subscript ”$\cdot$” means the $i$th entry is removed for a vector, or the $i$th row/column removed for a matrix. The regression coefficient $\beta_i = (\beta_{i,1}, \ldots, \beta_{p-1})$ and the error $\epsilon_{k,i,j}$ satisfy

$$\beta_i = -\omega_{S,i,i}^{-1} \Omega_{S,i,j}, \quad \text{and} \quad r_{ij} = \text{Cov}(\epsilon_{k,i,j}, \epsilon_{k,j,l}) = \frac{\omega_{S,i,j}}{\omega_{S,i,i} \omega_{S,j,j}}.$$  

As such, the elements $\omega_{S,i,j}$ of $\Omega_S$ can be represented in terms of $r_{ij}$. Next, we construct an estimator of $r_{ij}$ and its bias-corrected version. We then build on this estimator to obtain an estimator of $\omega_{S,i,j}$, upon which we further develop our test statistics.

A natural estimator of $r_{ij}$ is the sample covariance between the residuals $\tilde{e}_{k,i,j} = Y_{k,i,j} - \tilde{\beta}_i Y_{k,i,j} - \tilde{\beta}_j Y_{k,j,l}$, i.e., $\tilde{r}_{ij} = \frac{\sum_{k=1}^n \sum_{i=1}^p \sum_{j=1}^q \tilde{e}_{k,i,j} \tilde{e}_{k,j,l}}{np}$. Note that $\tilde{\beta}_i = \sum_{k=1}^n \sum_{i=1}^p Y_{k,i,j} / n$, $\bar{Y}_{k,j,l} = \sum_{k=1}^n Y_{k,i,j} / n \in \mathbb{R}^{(p-1) \times 1}$, and $\beta_i, i = 1, \ldots, p$, are estimators of $\beta_i$ that satisfy the regularity conditions (C1) or (C1') given in Web Appendix A. Such estimators can be obtained via standard estimation methods such as the Lasso and Dantzig selector. When $i \neq j$, however, $\tilde{r}_{ij}$ tends to be biased, due to the correlation induced by the estimated parameters. To correct such bias, following Lemma 2 in Web Appendix B, we have

$$\hat{r}_{ij} = \tilde{r}_{ij} - \tilde{r}_{i \cdot} (\tilde{\beta}_{i \cdot} - \beta_{i \cdot}) - \tilde{r}_{j \cdot} (\tilde{\beta}_{j \cdot} - \beta_{j \cdot}) + \omega_{S,i,j}((nq \log p)^{-1/2}),$$

where $\tilde{\beta}_{i \cdot} = (\tilde{\beta}_{i,1}, \ldots, \tilde{\beta}_{i,p-1})$, $\omega_{S,i,j} = (\omega_{S,i,j}^2)^{1/2}$, and $\omega_{S,i,j} = (\omega_{S,i,j}^2)^{1/2}$, for $1 \leq i < j \leq p$. For $i = j$, we let $\hat{r}_{ii} = \tilde{r}_{ii}$, which is a nearly unbiased estimator of $r_{ii}$. Consequently, an estimator of the entry $\omega_{S,i,j}$ of the spatial precision matrix $\Omega_S$ can be constructed as

$$T_{ij} = \frac{\hat{r}_{ij}}{r_{ij} \hat{r}_{ij}}, \quad 1 \leq i < j \leq p.$$
To further estimate the variance of $T_{ij}$, note that
\[
\theta_{ij} = \text{Var} \left( \varepsilon_{ik} \varepsilon_{kj} / (r_{ij} r_{kj}) \right) / \nu q = (1 + p \rho_{ij}^2) / (n q r_{ij} r_{kj}),
\]
(4)
where $\rho_{ij}^2 = \beta_{ij}^2 r_{ij} r_{kj}$. Then, $\theta_{ij}$ can be estimated by $\hat{\theta}_{ij} = (1 + \hat{\beta}_{ij}^2 r_{ij} r_{kj}) / (n q r_{ij} r_{kj})$. Given $\{T_{ij}; 1 \leq i < j \leq p\}$ are heteroscedastic and can possibly have a wide variability, we standardize $T_{ij}$ by its standard error, which leads to the standardized statistics,
\[
W_{ij} = T_{ij} / (\hat{\theta}_{ij})^{1/2}, \quad 1 \leq i < j \leq p.
\]
In the next section, we test hypotheses (1) and (2) based on $\{W_{ij}\}_{i,j=1}^p$.

**Remark 2.** Construction of the test statistics for the data-driven procedure is almost the same as that for the oracle procedure, except that the data-driven one replaces the transformed data $Y_k = X_k / \sqrt{n \nu q}$ in (3) with a plug-in estimator for $\Sigma_T$. Furthermore, the regression coefficients slightly vary at different time points in the data-driven scenario, and we replace (3) by $Y_{k,ij} = Y_k - \hat{\beta}_{ij} + \hat{\varepsilon}_{ij}$, for $1 \leq i \leq p, 1 \leq l \leq q$.

**Remark 3.** When $\Sigma_T$ is unknown, $E(\Sigma_T) = (\text{trace}(\Sigma_T)) / p \Sigma_T$. If $\text{trace}(\Sigma_T) = c$, an unbiased estimator of $\Sigma_T$ becomes $\Sigma_T / c$. Accordingly, we shall define the transformed data $Y_k = X_k \Sigma_T^{-1/2}$, for $k = 1, \ldots, n$. Then, we have the bias-corrected estimator $\hat{r}_{ij} = c r_{ij}$, which in turn leads to $T_{ij}^{\text{new}} = T_{ij} / c$, and $\hat{\theta}_{ij}^{\text{new}} = \hat{\theta}_{ij} / c^2$. Thus, the standardized statistic $W_{ij}$ remains the same, as the constant $c$ is canceled. Therefore, $c$ does not affect our final test statistics, and for simplicity, we set $c = 1$ from the beginning.

**Remark 4.** Thanks to the separation of spatial and temporal covariance structures, the errors $\{\varepsilon_{ij} \}_{i=1}^n$ are independent with each other for $k = 1, \ldots, n$ and $l = 1, \ldots, q$. As such, the variance of the estimator $\hat{T}_{ij}$ can be approximated by the variance of the products of residuals as in (4). On the other hand, Chen and Liu (2015) did not separate the spatial and temporal structures, and performed the inverse regression directly. As a result, for each spatial location, the errors of the corresponding $n q$ linear models are correlated, and the variance of $T_{ij}$ can no longer be approximated by (4). Therefore, their test statistics required an additional variance correction, whereas ours do not.

### 2.3. Global Testing Procedure

We propose the following test statistic for the global null hypothesis $H_0 : \Omega_S$ is diagonal,
\[
M_{\nu q} = \max_{1 \leq i < j \leq p} W_{ij}^2.
\]
Furthermore, let $q_* = - \log(8 \pi) - 2 \log \log(1 - \alpha)^{-1}$, we define the global test $\Psi_0$ by
\[
\Psi_0 = I(\nu q \geq q_*) + 4 \log p - \log \log p,
\]
where $I(\cdot)$ is the indicator function. The hypothesis $H_0$ is rejected whenever $\Psi_0 = 1$.

The above test is developed based on the asymptotic properties of $M_{\nu q}$, which will be studied in detail in Section 3.1. Intuitively, $(\{W_{ij}\}_{i,j=1}^p)$ are approximately standard normal variables under $H_0$, and are only weakly dependent under suitable conditions. Thus, $M_{\nu q}$ is the maximum of the squares of $\nu q (p - 1/2)$ such variables, and its value should be close to $2 \log \nu q (p - 1/2) \approx 4 \log \nu q$. We will later show that, under certain regularity conditions, $M_{\nu q} - 4 \log \nu q - \log \log \nu q$ converges to a type I extreme value distribution under $H_0$.

### 2.4. Multiple Testing Procedure

Next, we develop a multiple testing procedure, based on $W_{ij}$, for $H_{0,i,j} : \omega_{S,i,j} = 0$, to identify spatial locations that are conditionally dependent. Let $t$ be the threshold level such that $H_{0,i,j}$ is rejected if $|W_{ij}| \geq t$. Let $\mathcal{H}_0 = \{(i, j) : \Omega_{S,i,j} = 0, 1 \leq i < j \leq p\}$ be the set of true nulls. Denote by $R_0(t) = \sum_{(i, j) \in \mathcal{H}_0} I(|W_{ij}| \geq t)$ and $R(t) = \sum_{1 \leq i < j \leq p} I(|W_{ij}| \geq t)$ the total number of false positives and rejections, respectively. The FDP and FDR are defined as
\[
\text{FDP}(t) = \frac{R_0(t)}{R(t)} \vee 1, \quad \text{FDR}(t) = \mathbb{E}[\text{FDP}(t)].
\]
An ideal choice of $t$ would reject as many true positives as possible while controlling the FDP at the pre-specified level $\alpha$. That is, we select $t_0 = \inf \left\{ 0 \leq t \leq 2(\log p)^{1/2} : \text{FDP}(t) \leq \alpha \right\}$. We shall estimate $\sum_{(i, j) \in \mathcal{H}_0} I(|W_{ij}| \geq t)$ by $2[1 - \Phi(t)] |\mathcal{H}_0|$, where $\Phi(t)$ is the standard normal cumulative distribution function. Note that $|\mathcal{H}_0|$ is at most $(p^2 - p)/2$, and is close to $(p^2 - p)/2$ due to the sparsity of $\Omega_S$. This leads to the following multiple testing procedure.

Step 1. Calculate the test statistics $W_{ij}$.
Step 2. For given $0 \leq \alpha \leq 1$, calculate
\[
i = \inf \left\{ 0 \leq t \leq 2(\log p)^{1/2} : \frac{2[1 - \Phi(t)](p^2 - p)/2}{R(t) \vee 1} \leq \alpha \right\}.
\]
If $i$ does not exist, set $i = 2(\log p)^{1/2}$.
Step 3. For $1 \leq i < j \leq p$, reject $H_{0,i,j}$ if and only if $|W_{ij}| \geq i$.

### 3. Theoretical Properties

In this section, we analyze the theoretical properties of the global and multiple testing procedures for both the oracle and data-driven scenarios. We show that the data-driven procedures perform asymptotically as well as the oracle procedures and enjoy certain optimality under the regularity conditions. For separate treatment of the oracle and data-driven procedures, we now distinguish the notations of the two, and add the superscript “o” to denote the statistics and tests of the oracle procedures, e.g., $\hat{\beta}_o$, $M_{\nu q}^o$, $\Psi_o$, $\hat{p}_o$, and the superscript “d” to denote those of the data-driven procedures, e.g., $\hat{\beta}_d$, $M_{\nu q}^d$, $\Psi_d$, and $\hat{p}_d$. In the interest of space, we present all the regularity conditions and their discussion in Web Appendix A.
3.1. Oracle Global Testing Procedure

We first analyze the limiting null distribution of the oracle global test statistic $M_{\nu}^{o}$ and the power of the corresponding test $\Psi^{o}$. We are particularly interested in the power of the test under the alternatives when $\Omega_{5}$ is sparse, and show that the power is minimax rate optimal.

The following theorem states the asymptotic null distribution for $M_{\nu}^{o}$, and indicates that, under $H_{0}$, $M_{\nu}^{o} = 4 \log p + \log \log p$ converges weakly to a Gumbel random variable with distribution function $\exp\{-(8\pi)^{-1/2}e^{-t/2}\}$.

**Theorem 1.** Assuming the regularity conditions (C1)-(C3), under $H_{0}$, for any $t \in \mathbb{R}$,$$
P(M_{\nu}^{o} < 4 \log p + \log \log p \leq t) \rightarrow \exp\{-(8\pi)^{-1/2} \exp(-t/2)\},$$as $nq, p \rightarrow \infty$.

The above convergence is uniform for all $\{X_{k}\}_{k=1}^{\infty}$ satisfying (C1)-(C3).

We next study the power of the corresponding test $\Psi^{o}$. We define the following class of precision matrices for spatial locations:

$$\mathcal{U}(c) = \left\{ \Omega_{5} : \max_{1 \leq i < j \leq p} |\omega_{i,j}|/\theta_{i,j}^{1/2} \geq c(\log p)^{1/2} \right\}. \quad (5)$$

This class of matrices include all precision matrices such that there exists one standardized off-diagonal entry having the magnitude exceeding $c(\log p)^{1/2}$. By the definition in (4), $\theta_{i,j}$ is of the order $1/(nq)$, and, thus, we only require one of the off-diagonal entries of $\Omega_{5}$ to be larger than $C(\log p/(nq))^{1/2}$ with $C > 0$ fully determined by $c$ in (5) and $c_{0}$ and $c_{1}$ in the regularity condition (C2). Then, if we choose the constant $c = 4$, the next theorem shows that the null parameter set in which $\Omega_{5}$ is diagonal is asymptotically distinguishable from $\mathcal{U}(4)$ by the test $\Psi^{o}$. That is, $H_{0}$ is rejected by the test $\Psi^{o}$ with high probability if $\Omega_{5} \in \mathcal{U}(4)$.

**Theorem 2.** Assuming the regularity conditions (C1) and (C2), then,$$
\inf_{\Omega_{5} \in \mathcal{U}(4)} P(\Psi^{o} = 1) \rightarrow 1, \quad \text{as } nq, p \rightarrow \infty.
$$

The next theorem further shows that this lower bound $4(\log p)^{1/2}$ is rate-optimal. Let $\mathcal{T}_{\alpha}$ be the set of all $\alpha$-level tests, that is, $P(T_{u} = 1) \leq \alpha$ under $H_{0}$ for all $T_{u} \in \mathcal{T}_{\alpha}$.

**Theorem 3.** Suppose that $\log p = o(nq)$. Let $\alpha, \beta > 0$ and $\alpha + \beta < 1$. Then there exists a constant $c_{2} > 0$ such that for all sufficiently large $nq$ and $p$,$$
\inf_{\Omega_{5} \in \mathcal{U}(c_{2})} \sup_{T_{u} \in \mathcal{T}_{\alpha}} P(T_{u} = 1) \leq 1 - \beta.
$$

Therefore, if $c_{2}$ is sufficiently small, then any $\alpha$ level test is unable to reject the null hypothesis correctly uniformly over $\Omega_{5} \in \mathcal{U}(c_{2})$ with probability tending to one. So the order $(\log p)^{1/2}$ in the lower bound of $\max_{1 \leq i < j \leq p} |\omega_{i,j}|/\theta_{i,j}^{1/2}$ in (5) cannot be further improved.

3.2. Oracle Multiple Testing Procedure

We next investigate the properties of the oracle multiple testing procedure. The following theorem shows that it controls the FDP and FDR at the pre-specified level $\alpha$ asymptotically.

**Theorem 4.** Assuming the regularity conditions (C1) and (C2), and letting$$S_{\alpha} = \left\{(i, j) : 1 \leq i < j \leq p, |\omega_{i,j}|/\theta_{i,j}^{1/2} \geq (\log p)^{1/2+\rho} \right\}.$$

Suppose for some $\rho, \delta > 0$, $|S_{\alpha}| \geq 1/(18\pi^{1/2}a) + \delta(\log p)^{1/2}$.

We make a few remarks. First, the condition $|S_{\alpha}| \geq 1/(8\pi^{1/2}a)$ in Theorem 4 is mild, because we have $(p^{2} - p)/2$ hypotheses in total and this condition only requires a few entries of $\Omega_{5}$ having standardized magnitude exceeding $\{(\log p)^{1/2+\rho}/(nq)\}$ for some $\rho > 0$. Second, the condition $\alpha_{0} = |H_{0}| \geq c_{0}p^{2}$ for some $c_{0} > 0$, and $p \leq c(nq)^{1}$ for some $c, r > 0$. Let $l = (p^{2} - p)/2$, then,$$
\lim_{(nq,p) \rightarrow \infty} \frac{\text{FDR}(\hat{\beta})}{\alpha_{0}/l} = 1, \quad \frac{\text{FDP}(\hat{\beta})}{\alpha_{0}/l} \rightarrow 1 \quad \text{in probability, as } (nq, p) \rightarrow \infty.
$$

3.3. Data-Driven Procedures

We next turn to the data-driven procedures for both the global and multiple testing. We show that they perform as well as the oracle testing procedures asymptotically.

**Theorem 5.** Assume the regularity conditions (C1’, (C3’), (C2’)-(C4’).

(i) Under $H_{0}$, for any $t \in \mathbb{R}$,$$
P(M^{d} - 4(\log p + \log \log p \leq t) \rightarrow \exp\{-(8\pi)^{-1/2} \exp(-t/2)\},$$as $nq, p \rightarrow \infty$.

and this convergence is uniform for all $\{X_{k}\}_{k=1}^{\infty}$ satisfying (C1’), (C2’)-(C4’).

(ii) Furthermore, $\inf_{\Omega_{5} \in \mathcal{U}(c_{2})} P(\Psi^{d} = 1) \rightarrow 1, \quad \text{as } nq, p \rightarrow \infty.$

This theorem shows that $M^{d}$ has the same limiting null distribution as the oracle test statistics $M_{\nu}^{o}$, and the power of the corresponding test $\Psi^{d}$ performs as well as the oracle test and is thus minimax rate optimal. The same observation...
applies to Theorem 6 below, which shows that the data-driven multiple procedure also performs as well as the oracle case, and controls the FDP and FDR at the pre-specified level \( \alpha \) asymptotically.

**Theorem 6.** Assuming (C1), (C4), and the same conditions as in Theorem 4,

\[
\lim_{(nq,p) \to \infty} \frac{\text{FDP}(\hat{\rho})}{\alpha_0/1} = 1, \quad \lim_{(nq,p) \to \infty} \frac{\text{FDP}(\hat{\rho})}{\alpha_0/1} \to 1 \quad \text{in probability,}
\]

\[
as(nq, p) \to \infty,\]

4. **Simulation Studies**

We study in this section the finite-sample performance of the proposed testing procedures. We compare the following methods: the oracle procedure (denoted as “oracle”), the data-driven procedure with \( \Sigma_2 \) estimated by the usual sample covariance (denoted as “data-driven-S”), the one with the banded estimator (denoted as “data-driven-B”), the testing method of Liu (2013) with whitening (denoted as “whitening”), and the one without whitening (denoted as “vector normal”). We examine a range of spatial and temporal dimensions, as well as the sample sizes. Specifically, \( p = \{50, 200, 800\} \), \( q = \{20, 50, 200\} \), and \( n = \{2, 10, 50\} \).

We consider two temporal covariance structures: (1) autoregressive model \( \Sigma_t = \{\sigma_t(i,j)\} \) with elements \( \sigma_t(i,j) = 0.4^{i-j}, 1 \leq i, j \leq p \); and (2) moving average model: \( \Sigma_t = \{\sigma_t(i,j)\} \) with nonzero elements \( \sigma_t(i,j) = 1/(|i-j| + 1) \), for \( |i-j| \leq 3 \). We also examine two additional temporal structures and report the corresponding results in Web Appendix D.

In all simulations, we use Lasso to estimate \( \hat{\beta}_k \),

\[
\hat{\beta}_k = D_k^{-1/2} \arg \min_u \left\{ \frac{1}{2nq} \left( \|Y_{(i)} - \bar{Y}_{(i)}\|_2^2 + \lambda_0 u_{11} \right) \right\},
\]

where \( Y \) is the \( nq \times p \) data matrix by stacking the transformed samples \( \{(Y_{k,j}, k = 1, \ldots, n, j = 1, \ldots, q) \}, \ Y_k = X_k \Sigma_t^{-1/2} \) for the oracle procedure and \( Y_k = X_k \Sigma_t^{-1/2} \) for the data-driven procedure, \( k = 1, \ldots, n \), \( Y_{(i)} = (Y_{1,i}, \ldots, Y_{nq,i}) \in \mathbb{R}^{nq \times 1} \), \( Y_{(i)} = (\bar{Y}_{1,i}, \ldots, \bar{Y}_{nq,i}) \in \mathbb{R}^{nq \times 1} \), \( Y_{(i)} = (\bar{Y}_{1,i}, \ldots, \bar{Y}_{nq,i}) \in \mathbb{R}^{nq \times (p-1)} \) with \( Y_{(i)} = \frac{1}{nq} \sum_{k=1}^{nq} Y_{k,i}, \ \bar{Y}_{(i)} = (\bar{Y}_{1,i}, \ldots, \bar{Y}_{nq,i}) \in \mathbb{R}^{nq \times (p-1)} \), \( D_k = \text{diag}(\Sigma_k S_{k,k}) \), and \( \Sigma_k \) is the sample covariance matrix of \( \Sigma_k \) with \( nq \) transformed samples.

4.1. **Global Testing Simulation**

For global testing, the data are generated from a matrix normal distribution with precision matrix \( \Omega_2 \otimes \Omega_T \). To evaluate the size of the tests under the null, we set \( \Omega_S = I \). To evaluate the power under the alternative, we set \( \Omega_S = (I + U + \delta I)/(1 + \delta) \), where \( \delta = |2\alpha_{0.05}(I + U)| + 0.05 \), and \( U \) is a matrix with eight random nonzero entries. The locations of four nonzero entries are selected randomly from the upper triangle of \( U \), each with a magnitude generated randomly and uniformly from the set \([-4\log p/(nq)]^{1/2}, -2\log p/(nq)]^{1/2} \cup [2\log p/(nq)]^{1/2}, 4\log p/(nq)]^{1/2} \). The other four entries are determined by symmetry. We set the tuning parameters in (6) as \( \lambda_{0.05} = 2\Sigma_{S,k} \log p/(nq)]^{1/2} \), following Xia et al. (2015).

Table 1 summarizes the size and power, in percentages, of the proposed global testing based on 1000 data replications with the significance level \( \alpha_0 = 5\% \). We see from the table that the empirical sizes of both oracle and data-driven procedures are well controlled under \( \alpha_0 \) for all settings. By contrast, the vector normal based testing shows a serious distortion in size, whereas the whitening aided testing shows a clear increase in size too, especially when \( n \) is small and \( q \) is large. We also observe that, for our proposed tests, the empirical sizes are slightly below the nominal level when \( p \) is large, and this is due to the correlation among the variables. A similar phenomenon has also been observed and justified in Cai et al. (2013,Proposition 1). Moreover, we see that the proposed test based on the banded covariance estimator is powerful in all settings, even though the two spatial precision matrices differ only in eight entries with the magnitude of difference of the order \( \log p/(nq)]^{1/2} \). The new test is also more powerful than the whitening procedure. In addition, the data-driven procedure based on the banded covariance estimator is seen to perform similarly as the oracle procedure, and it clearly outperforms the one based on the sample covariance estimator. As such we recommend to use the banded estimator in our data-driven testing. The results for \( p = 800 \) are reported in Web Appendix D.

4.2. **Multiple Testing Simulation**

For multiple testing, the data \( \{X_1, \ldots, X_n\} \) are generated from the matrix normal distribution with the above two temporal covariance matrices, and five spatial graph structures. They include: (1) banded graph: a banded graph generated by the \texttt{huge} package (Zhao et al., 2012) with bandwidth equal to 3; (2) sparse graph: a random graph by \texttt{huge} with default probability 3/\( p \) that a pair of nodes has an edge; (3) relatively dense graph: a random graph by \texttt{huge} with probability 0.01 that a pair of nodes has an edge; (4) hub graph: a hub graph by \texttt{huge} with row and columns evenly partitioned into 10 disjoint groups; and (5) small-world graph: a small-world graph generated by the \texttt{huge} package \texttt{rags2ridges} (van Wieringen and Peeters, 2014) with three starting neighbors and 5\% probability of rewiring.

We select the tuning parameters \( \lambda_{0.05} \) in (6) adaptively given the data. The idea is to make \( \sum_{(i,j) \in E_{0.05}} I(|W_{i,j}| \geq t) \) and \( (2 - 2\Phi(t))(p^2/p)/2 \) close. That is, a good choice of the tuning parameter \( \lambda_{0.05} = b/20 \sqrt{\delta_{S,k} \log p/(nq)} \) should minimize the error

\[
\int_c^1 \left( \sum_{(i,j) \in E_{0.05}} I(|W_{i,j}| \geq t) \right) \left( \Phi^{-1}(1 - \alpha/2) - 1 \right)^2 dt \]

with \( c > 0 \) and \( W_{i,j}^{(0)} \) is the statistic of the corresponding tuning parameter. Toward that end, we employ the following procedure for parameter tuning. Step 2 is a discretization of the above integral, where the sum in the numerator is over \( H \), as
Table 1

Global testing: empirical size (with standard error in parentheses) and empirical power (in percentage) based on 1000 data replications. 
$n = (10, 50), p = (50, 200), q = (20, 50, 200), \alpha = 5\%$. Methods under comparison are: the oracle procedure with the true temporal covariance (“oracle”), the data-driven procedure with a sample covariance estimator (“data-driven-S”), the data-driven procedure with a banded covariance estimator (“data-driven-B”), and the procedure of Liu (2013) with (“whitening”) and without (“vector normal”) conventional whitening.

| $p$ | $q$ | $n = 10, AR$ |  | $n = 50, AR$ |  | $n = 10, MA$ |  | $n = 50, MA$ |  |
|-----|-----|-------------|---|-------------|---|-------------|---|-------------|---|
|     |     | Empirical size (in %) |  | Empirical power (in %) |  | Empirical size (in %) |  | Empirical power (in %) |  |
|     |     | Oracle | Data-driven-S | Data-driven-B | Whitening | Vector normal | Oracle | Data-driven-S | Data-driven-B | Whitening | Vector normal | Oracle | Data-driven-S | Data-driven-B | Whitening | Vector normal |
|     |     | 3.4 (0.6) | 1.7 (0.4) | 3.2 (0.6) | 5.4 (0.7) | 34.8 (1.5) | 3.8 (0.6) | 3.4 (0.6) | 3.5 (0.6) | 7.1 (0.8) | 49.3 (1.6) | 84.8 | 80.8 | 82.4 | 76.8 | 89.8 | 92.6 | 91.8 | 92.2 | 88.2 | 94.4 |
| 20  | 20  | 5.0 (0.7) | 1.4 (0.4) | 4.1 (0.6) | 7.3 (0.8) | 40.4 (1.6) | 4.8 (0.7) | 3.0 (0.5) | 4.1 (0.6) | 7.2 (0.8) | 60.6 (1.5) | 73.6 | 54.6 | 70.6 | 68.0 | 80.1 | 81.1 | 76.6 | 79.2 | 71.6 | 91.2 |
| 50  | 50  | 3.9 (0.6) | 3.3 (0.6) | 3.7 (0.6) | 7.1 (0.8) | 41.2 (1.6) | 3.9 (0.6) | 0.8 (0.3) | 4.7 (0.6) | 6.7 (0.8) | 67.4 (1.5) | 64.8 | 63.2 | 60.1 | 63.8 | 79.6 | 61.8 | 63.8 | 63.8 | 58.9 | 88.9 |
| 200 | 200 | 4.3 (0.7) | 1.4 (0.4) | 3.6 (0.6) | 4.4 (0.7) | 39.5 (1.5) | 4.9 (0.7) | 0.8 (0.3) | 4.7 (0.6) | 6.7 (0.8) | 68.4 (1.5) | 61.6 | 58.2 | 62.4 | 59.6 | 74.4 | 63.6 | 63.6 | 62.8 | 62.8 | 87.6 |
Hypothesis Testing of Matrix Graph Model with Application to Brain Connectivity Analysis

5. Real Data Analysis

We revisit the motivating electroencephalography (EEG) data and illustrate our testing methods. The data was collected in a study examining EEG correlates of genetic predisposition to alcoholism and is available at http://kdd.ics.uci.edu/datasets/eeg/eeg.data.html. It consists of 77 alcoholic individuals and 45 controls, and each subject was fitted with a 61-lead electrode cap and was recorded at 256 Hz for 1 second. There were in addition a ground and two bipolar deviation electrodes, which are excluded from the analysis. The electrode positions were located at standard sites (Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990), and were organized into frontal, central, parietal, occipital, left temporal, and right temporal regions. Each subject performed 120 trials under three types of stimuli. More details of data collection can be found in Zhang et al. (1995). We preprocessed the data by first averaging all trials under a single stimulus condition following Li et al. (2010). We then performed an α-band filtering on the signals following Hayden et al. (2006). The resulting data is a 61 × 256 matrix for each subject, and our goal is to infer the 61 × 61 connectivity network of the brain spatial locations.

We applied our testing procedures for the alcoholic and control groups separately. We employed the banded estimator of Bickel and Levina (2008) to estimate the temporal covariance. By splitting the data of each group 10 times, we obtained the best bandwidth equaling 3 for both the alcoholic and control groups, which are close to our simulation settings. We applied the global test and obtained the p-value 0 for both groups, clearly indicating that some brain regions are connected in the two groups. We then applied the data-driven multiple testing procedure, with a pre-specified FDR significance level α = 1%. For graphical illustration, we report the top 100 most significant FDRs between the left and right frontal regions in the alcoholic group compared to the control. This finding agrees with that in the literature (Hayden et al., 2006).

6. Discussion

Motivated by brain connectivity analysis, we have proposed in this article hypothesis testing procedures for detecting conditional dependence between spatial locations. Our work explicitly exploits the special covariance structure of matrix normal distribution, and our results suggest that using such information could improve the inferential capability. Our testing methods can handle a small sample size, as well as an adequately large network; in our simulations, n is as small as 2 or 10, whereas the spatial dimension p could reach 800. Meanwhile, we have treated the temporal covariance as a nuisance. We make a few remarks regarding the temporal dimension q and the temporal covariance estimator $\Sigma_T$. First, we do not require $q < n$, as we in effect pool np samples to estimate a $q \times q$ temporal covariance $\Sigma_T$. Second, there are multiple ways to estimate $\Sigma_T$. Our numerical study suggests the usual sample covariance estimator does not work well, but favors the banded estimator (Bickel and Levina, 2008). Meanwhile, several other existing covariance estimators can be employed as well (Cai and Liu, 2011; Cai et al., 2011). Third, the estimator of $\Sigma_T$ needs to satisfy the estimation rate condition (C4). We view this condition reasonable, because it holds under some sparsity requirement, which is scientifically plausible, and it allows the temporal dimension $q$ to be of the same
Table 2

*Multiple testing: empirical FDR (with standard error in parenthesis) based on 100 data replications. n = 10, p = (50, 200), q = (20, 50, 200), \( \alpha_2 = 1\% \).*

| \( p \)  | \( q = 20 \) | \( q = 50 \) | \( q = 200 \) |
|-------|-------------|-------------|-------------|
|       | spatial structure | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Oracle | Empirical FDR (SE) (in %), autoregressive temporal structure | 1.3 (1.4) | 1.1 (1.6) | 0.8 (2.3) | 0.5 (1.1) | 1.1 (2.3) | 0.5 (0.6) | 0.6 (0.8) | 0.8 (2.6) | 0.8 (1.2) | 0.8 (0.7) | 0.4 (0.5) | 1.0 (1.2) | 0.6 (2.0) | 0.9 (1.3) | 0.6 (0.6) |
| Data-driven-S | 1.5 (1.7) | 1.0 (1.6) | 0.5 (1.9) | 0.6 (1.3) | 1.4 (2.9) | 0.5 (0.5) | 0.5 (0.7) | 0.3 (1.3) | 0.7 (1.2) | 0.6 (0.7) | 0.2 (0.3) | 0.1 (0.4) | 0.0 (0.0) | 0.1 (0.5) | 0.1 (0.1) |
| Whitening | 2.4 (1.8) | 1.3 (2.2) | 1.4 (3.2) | 1.0 (1.6) | 1.8 (3.3) | 1.0 (0.8) | 0.9 (1.1) | 1.7 (3.9) | 1.2 (1.7) | 1.2 (1.1) | 1.7 (1.1) | 1.0 (1.2) | 0.7 (2.1) | 1.2 (1.6) | 1.0 (0.9) |
| Vector normal | 4.5 (2.2) | 4.4 (3.4) | 4.9 (6.2) | 3.8 (3.3) | 4.8 (3.8) | 2.5 (1.5) | 4.3 (2.3) | 9.1 (7.8) | 4.9 (3.3) | 3.2 (1.5) | 2.7 (1.4) | 5.3 (2.8) | 7.4 (6.3) | 5.1 (3.3) | 3.3 (1.6) |
| Oracle | Empirical FDR (SE) (in %), moving average temporal structure | 1.3 (0.7) | 0.7 (1.0) | 0.7 (0.7) | 1.0 (1.7) | 1.1 (1.2) | 0.5 (0.2) | 0.7 (0.5) | 0.7 (0.5) | 0.8 (0.6) | 0.6 (0.3) | 0.6 (0.3) | 0.8 (0.4) | 0.9 (0.6) | 0.8 (0.6) | 0.6 (0.3) |
| Data-driven-S | 1.3 (0.7) | 0.8 (1.0) | 0.6 (0.7) | 1.0 (1.6) | 1.0 (1.2) | 0.5 (0.2) | 0.7 (0.5) | 0.6 (0.5) | 0.9 (0.7) | 0.6 (0.3) | 0.6 (0.3) | 0.4 (0.4) | 0.4 (0.4) | 0.7 (0.5) | 0.5 (0.3) |
| Whitening | 2.0 (1.0) | 1.1 (1.9) | 1.2 (1.2) | 1.9 (2.9) | 1.5 (2.4) | 0.8 (0.4) | 1.2 (0.7) | 1.2 (0.8) | 1.4 (0.9) | 1.1 (0.5) | 1.2 (0.4) | 1.3 (0.7) | 1.3 (0.7) | 1.4 (0.8) | 1.1 (0.4) |
| Vector normal | 4.1 (1.6) | 5.4 (2.3) | 5.2 (2.4) | 6.3 (3.5) | 5.2 (2.3) | 2.9 (0.7) | 4.9 (1.5) | 5.6 (1.5) | 6.7 (1.7) | 3.8 (0.8) | 4.4 (0.9) | 6.5 (1.3) | 7.3 (2.0) | 8.0 (1.7) | 4.7 (0.9) |
| Oracle | Empirical FDR (SE) (in %), autoregressive temporal structure | 1.3 (1.4) | 0.8 (1.4) | 0.6 (2.2) | 0.7 (1.4) | 1.0 (2.2) | 0.5 (0.5) | 0.9 (1.0) | 0.8 (2.4) | 0.9 (1.5) | 0.7 (0.7) | 0.7 (0.7) | 0.7 (0.8) | 1.3 (3.5) | 0.7 (1.6) | 0.7 (0.6) |
| Data-driven-S | 1.4 (1.3) | 0.8 (1.5) | 0.6 (2.2) | 0.5 (1.1) | 1.3 (2.8) | 0.5 (0.6) | 0.6 (1.0) | 0.3 (1.3) | 0.7 (1.2) | 0.5 (0.7) | 0.1 (0.3) | 0.0 (0.0) | 0.0 (0.0) | 0.2 (0.6) | 0.0 (0.1) |
| Whitening | 2.9 (2.1) | 1.7 (3.1) | 1.2 (2.9) | 1.5 (2.0) | 1.5 (2.9) | 1.3 (0.9) | 1.4 (1.4) | 1.5 (3.3) | 1.7 (1.9) | 1.5 (1.3) | 2.8 (1.3) | 2.0 (1.5) | 2.0 (3.9) | 2.3 (2.5) | 1.6 (0.9) |
| Vector normal | 4.7 (2.6) | 4.7 (3.5) | 6.8 (7.9) | 4.5 (3.4) | 5.2 (3.9) | 3.7 (1.5) | 6.0 (2.9) | 11.6 (9.3) | 7.8 (4.2) | 4.9 (2.1) | 5.3 (1.9) | 7.9 (3.2) | 16.5 (10.4) | 10.3 (4.1) | 5.9 (2.1) |
| Oracle | Empirical FDR (SE) (in %), moving average temporal structure | 1.3 (0.7) | 0.8 (0.8) | 0.7 (0.7) | 1.0 (1.3) | 1.0 (1.1) | 0.4 (0.2) | 0.7 (0.5) | 0.7 (0.5) | 0.8 (0.6) | 0.7 (0.4) | 0.6 (0.3) | 0.8 (0.5) | 0.7 (0.5) | 0.8 (0.5) | 0.6 (0.3) |
| Data-driven-S | 1.3 (0.7) | 0.8 (0.8) | 0.6 (0.8) | 1.0 (1.2) | 1.0 (1.2) | 0.4 (0.2) | 0.7 (0.5) | 0.6 (0.5) | 0.7 (0.6) | 0.7 (0.4) | 0.5 (0.3) | 0.4 (0.4) | 0.4 (0.5) | 0.5 (0.4) | 0.4 (0.2) |
| Whitening | 2.0 (1.2) | 1.7 (1.6) | 1.3 (1.6) | 2.5 (4.5) | 1.8 (2.4) | 1.2 (0.5) | 1.4 (0.7) | 1.5 (0.8) | 1.9 (1.1) | 1.7 (0.7) | 2.0 (0.6) | 1.9 (0.9) | 1.9 (0.8) | 2.5 (1.2) | 1.4 (0.5) |
| Vector normal | 5.5 (1.3) | 5.2 (2.5) | 5.4 (2.9) | 6.9 (4.2) | 5.3 (2.5) | 4.3 (0.7) | 7.6 (1.5) | 9.3 (2.6) | 10.8 (2.3) | 6.1 (1.1) | 8.1 (1.1) | 12.1 (2.0) | 15.8 (2.9) | 14.4 (2.1) | 8.3 (1.3) |
Table 3

Multiple testing: empirical power (in percentage) based on 100 data replications. $n = 10$, $p = \{50, 200\}$, $q = \{20, 50, 200\}$, $\alpha_2 = 1\%$.

| $p$   | spatial structure | $q = 20$         | $q = 50$         | $q = 200$        |
|-------|-------------------|------------------|------------------|------------------|
|       |                   | 1    | 2    | 3    | 4    | 5    | 1    | 2    | 3    | 4    | 5    | 1    | 2    | 3    | 4    | 5    |
|       |                   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|       |                   | Oracle       | 62.7 | 53.2 | 98.9 | 98.7 | 19.1 | 100.0| 98.9 | 100.0| 100.0| 88.5 | 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Data-driven-S| 57.7 | 47.9 | 97.7 | 98.3 | 14.5 | 100.0| 97.4 | 100.0| 100.0| 77.9 | 100.0| 100.0| 100.0| 100.0| 99.9 |
| 50    |                   | Data-driven-B| 61.2 | 51.9 | 98.3 | 98.6 | 17.2 | 100.0| 98.8 | 100.0| 100.0| 87.2 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Whitening         | 47.9 | 39.1 | 96.4 | 95.1 | 11.8 | 99.6 | 96.7 | 100.0| 100.0| 74.9 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Vector normal     | 62.7 | 56.1 | 98.1 | 96.9 | 21.9 | 99.9 | 98.0 | 100.0| 100.0| 85.1 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Oracle       | 44.5 | 21.2 | 46.7 | 19.0 | 12.4 | 99.9 | 95.6 | 98.1 | 92.6 | 92.6 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Data-driven-S| 43.3 | 20.4 | 45.0 | 18.3 | 11.6 | 99.9 | 94.7 | 97.6 | 91.4 | 89.3 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
| 200   |                   | Data-driven-B| 44.1 | 20.8 | 46.1 | 18.5 | 12.2 | 99.9 | 95.5 | 98.1 | 92.4 | 92.1 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Whitening         | 32.3 | 13.7 | 32.9 | 10.1 | 5.9  | 97.8 | 89.3 | 94.6 | 76.6 | 73.7 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Vector normal     | 48.9 | 26.0 | 45.6 | 23.2 | 15.8 | 99.7 | 92.7 | 96.2 | 86.7 | 90.6 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Oracle       | 62.2 | 50.7 | 98.5 | 98.6 | 17.3 | 100.0| 98.9 | 100.0| 100.0| 88.3 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Data-driven-S| 58.1 | 45.7 | 97.9 | 98.2 | 13.6 | 99.9 | 97.6 | 100.0| 100.0| 76.9 | 100.0| 100.0| 100.0| 100.0| 99.8 |
|       |                   | Data-driven-B| 60.7 | 49.1 | 98.2 | 98.6 | 16.1 | 100.0| 98.9 | 100.0| 100.0| 87.2 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Whitening         | 44.7 | 35.7 | 93.7 | 94.1 | 10.1 | 98.7 | 95.8 | 99.9 | 100.0| 69.9 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Vector normal     | 63.5 | 53.9 | 97.2 | 96.7 | 22.2 | 99.8 | 97.8 | 100.0| 100.0| 83.9 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Oracle       | 44.5 | 27.9 | 33.6 | 19.7 | 11.7 | 99.9 | 96.7 | 97.9 | 92.9 | 80.1 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       |                   | Data-driven-S| 43.4 | 26.8 | 32.3 | 19.4 | 10.9 | 99.8 | 95.9 | 97.3 | 91.9 | 76.3 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
| 200   |                   | Data-driven-B| 44.2 | 27.5 | 33.2 | 19.1 | 11.5 | 99.9 | 96.6 | 97.8 | 92.7 | 79.4 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Whitening         | 30.1 | 17.3 | 21.3 | 9.9  | 5.8  | 96.2 | 89.2 | 92.4 | 70.1 | 59.8 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
|       | Vector normal     | 50.5 | 31.2 | 34.7 | 22.4 | 15.3 | 99.6 | 93.2 | 95.1 | 83.7 | 80.8 | 100.0| 100.0| 100.0| 100.0| 100.0| 100.0| 100.0|
order as a polynomial of $np$, which again generally holds in the neuroimaging context. Under such situations, we feel our proposed testing procedure is advisable. Finally, the empirical performance of our testing procedure depends on how good the plug-in estimator $\hat{\Sigma}_T$ is. In cases where the sparsity does not hold or $q$ far exceeds $np$, the quality of $\hat{\Sigma}_T$ can deteriorate, which would in turn adversely affect our testing procedure.

7. Supplementary Materials
Web Appendices, Tables, and Figures referenced in Sections 2–4, the EEG data and the computer code are available with this article at the Biometrics website on Wiley Online Library.

ACKNOWLEDGEMENTS
We thank the editor, the associate editor, and two reviewers for their constructive and insightful comments. This research was partially supported by NSF grants DMS-1310319 (Li), DMS-1613137 (Li), and DMS-1612906 (Xia), The Recruitment Program of Global Experts Youth Project (Xia), and the startup fund from Fudan University (Xia).

REFERENCES
Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics* 4, 764–790.
Allen, G. I. and Tibshirani, R. (2012). Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, 721–743.
Anderson, T. W. (2003). *An Introduction To Multivariate Statistical Analysis*. Third edition. New York: Wiley.
Aston, J. A. and Kirch, C. (2012). Estimation of the distribution of change-points with application to fmri data. *Annals of Applied Statistics* 6, 1906–1948.
Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* 199–227.
Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience* 10, 186–198.
Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106, 672–684.
Cai, T. T., Liu, W., and Luo, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106, 594–607.
Cai, T. 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. and Liu, W. (2015). Statistical inference for matrix-variate gaussian graphical models and false discovery rate control. *arXiv preprint arXiv:1509.05453*.
Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 76, 373–397.
Drton, M. and Perlman, M. D. (2007). Multiple testing and error control in gaussian graphical model selection. *Statistical Science* 22, 430–449.
Fornito, A., Zalesky, A., and Breakspear, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *Neuroimage* 80, 426–444.
Fox, M. D. and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience* **8**, 700–711.

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

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Academic Press.

Hayden, E. P., Wiegand, R. E., Meyer, E. T., Bauer, L. O., O’Connor, S. J., Nurnberger, J. I., et al. (2006). Patterns of regional brain activity in alcohol-dependent subjects. *Alcoholism: Clinical and Experimental Research* **30**, 1986–1991.

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

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.

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.

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

Narayan, M., Allen, G., and Tomson, S. (2015). Two sample inference for populations of graphical models with applications to functional connectivity. *arXiv preprint arXiv:1502.03853*.

Narayan, M. and Allen, G. I. (2016). Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. *Frontiers in Neuroscience* 10. doi: 10.3389/fnins.2016.00108.

Raichle, M. E. and Gusnard, D. A. (2002). Appraising the brains energy budget. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 10237–10239.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5**, 935–980.

Reiss, P. and Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66**, 61–69.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23**, Supplement 1, S208–S219.

Tsiligkaridis, T., Hero, III, A. O., and Zhou, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Transactions on Signal Processing* **61**, 1743–1755.

van Wieringen, W. N. and Peeters, C. F. (2014). Ridge estimation of inverse covariance matrices from high-dimensional data. *arXiv preprint arXiv:1403.0904*.

Xia, Y., Cai, T., and Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102**, 247–266.

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

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

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research* **13**, 1059–1096.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society, Series B* **76**, 463–483.

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

Zhu, Y., Shen, X., and Pan, W. (2014). Structural Pursuit Over Multiple Undirected Graphs. *Journal of the American Statistical Association* **109**, 1683–1696.

Received November 2015. Revised September 2016. Accepted October 2016.