Matrix Graph Hypothesis Testing and Application in Brain Connectivity Alternation Detection

Yin Xia and Lexin Li

Fudan University

University of California at Berkeley

Abstract: Brain connectivity alternation analysis reveals important insights of pathologies for a wide range of neurological disorders. It calls for development of rigorous statistical inferential tools, which can both provide an explicit statistical significance quantification as well as a rigid false discovery control. We formulate the problem as partial correlation hypothesis testing under a matrix normal distribution. We develop inferential procedures for testing equality of individual entries of partial correlation matrices across multiple groups. We derive the asymptotic properties and show the procedures can control the false discovery at the pre-specified level. We also compare our proposal with alternative testing procedures, both analytically and numerically, and demonstrate clear advantages of the new method. We illustrate with a functional connectivity analysis of an attention deficit hyperactivity disorder dataset.

Key words and phrases: Brain connectivity analysis; Functional magnetic resonance imaging; Gaussian graphical model; Matrix-variate normal distribution; Multiple testing; Partial correlation.

1 Introduction

Brain connectivity analysis is now in the foreground of neuroscience research (Bullmore and Sporns (2009); Fornito et al. (2013)), and is drawing ever-increasing attention in the statistics field as well (Kim et al. (2014); Ahn et al. (2015); Chen et al. (2015); Narayan et al. (2015); Zhang et al. (2015); Han et al. (2016); Kang et al. (2016); Qiu et al. (2016); Wang et al. (2016); Xia and Li (2017); among others). Brain functional connectivity reveals synchronization of brain systems via correlations in neurophysiological measures of brain activity. When mea-

sured during resting state, it maps the intrinsic functional architecture of the brain (Varoquaux and Craddock (2013)). Accumulated evidence has indicated that, compared to a healthy brain, a connectivity network alters in the presence of numerous neurological disorders, including Alzheimer's disease, attention deficit hyperactivity disorder, autism spectrum disorder, and many others (Hedden et al. (2009); Tomasi and Volkow (2012); Rudie et al. (2013)). Such alternations in brain connectivity are associated with cognitive and behavioral functions, and hold crucial insights of pathologies of neurological disorders (Fox and Greicius (2010)). In this article, we tackle the problem of comparing brain functional connectivity patterns across multiple subject groups, e.g., the diseased versus the healthy control.

One of the mainstream imaging modalities to study brain functional connectivity is resting-state functional magnetic resonance imaging (fMRI). We focus on fMRI here, while the method we develop is applicable to other similar imaging modalities as well. For each study subject at rest during the scan, fMRI measures changes in blood flow and oxygenation at individual voxels of brain over time, yielding a 4-way data array (Lindquist (2008)). To overcome spurious correlations due to close spatial proximity, a common practice is to parcellate the brain and map brain voxels to a list of pre-specified brain regions, then average the time courses of voxels within the same region. This results in a region by time matrix for each fMRI scan. Based upon this spatial temporal matrix, an undirected graph is constructed to depict brain connectivity, where nodes represent neurological elements such as brain regions, and links measure pairwise interaction and dependence between nodes (Bullmore and Sporns (2009)). There have been numerous dependence metrics proposed in the brain connectivity literature, and among those, partial correlation is a well accepted and commonly used measure for functional connectivity (Ryali et al. (2012); Wang et al. (2016)).

Central to functional connectivity analysis are *estimation* and *inference* of connectivity patterns across multiple subject groups. The former can be formulated as a sparse precision

matrix estimation problem, and there have been a large number of graph estimation solutions proposed. Examples include precision matrix estimation under a vector normal distribution (Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Friedman et al. (2008); Cai et al. (2011)), or a matrix normal distribution (Yin and Li (2012); Leng and Tang (2012); Zhou (2014); Qiu et al. (2016)). The latter can be formulated as a graph-based hypothesis testing problem. Most existing studies transform this problem into the classical two-sample testing framework by summarizing a network as a set of network metrics (Kim et al. (2014)). Although this strategy has proven useful, the extent to which each network metric provides a meaningful representation of brain function requires substantial care (Fornito et al. (2013)), and there is no unanimous agreement on what network metrics best characterize brain functions. There have been relatively much fewer solutions that directly test precision matrices under graphical models, and those are emerging only recently (Liu (2013); Xia et al. (2015); Narayan et al. (2015); Chen and Liu (2015); Xia and Li (2017)).

In this article, we adopt the matrix Gaussian graphical model framework, and develop statistical inferential procedures for testing equality of individual entries of the partial correlation matrices across multiple groups. We mostly focus on the two-population scenario, and only briefly discuss the extension to more than two populations. Specifically, let $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)} \in \mathbb{R}^{p\times q}$ denote the spatial-temporal matrices of the two groups, e.g., the diseased and the healthy control. We assume $\boldsymbol{X}^{(g)}$ follows a matrix normal distribution, with the Kronecker product covariance structure, $\boldsymbol{\Sigma}\{\text{vec}(\boldsymbol{X}^{(g)})\} = \boldsymbol{\Sigma}_{S_g} \otimes \boldsymbol{\Sigma}_{T_g}, g = 1, 2$, and correspondingly, $\boldsymbol{\Sigma}^{-1}\{\text{vec}(\boldsymbol{X}^{(g)})\} = \boldsymbol{\Sigma}_{S_g}^{-1} \otimes \boldsymbol{\Sigma}_{T_g}^{-1} = \boldsymbol{\Omega}_{S_g} \otimes \boldsymbol{\Omega}_{T_g}, g = 1, 2$, where $\boldsymbol{\Sigma}_{S_g}, \boldsymbol{\Omega}_{S_g} \in \mathbb{R}^{p\times p}$ denote the spatial covariance and precision matrix, respectively, and $\boldsymbol{\Sigma}_{T_g}, \boldsymbol{\Omega}_{T_g} \in \mathbb{R}^{q\times q}$ denote the temporal ones. This matrix normal assumption has been frequently adopted in numerous finance, genetics, and biological applications (Yin and Li (2012); Leng and Tang (2012)), and is also scientifically plausible in the neuroimaging context. For instance, the standard neuroimaging

processing software, such as SPM (Friston et al. (2007)) and FSL (Smith et al. (2004)), commonly adopt a framework that assumes the data are normally distributed per location with a noise factor and an autoregressive structure, which shares a similar spirit as the matrix normal formulation. Moreover, Aston et al. (2016) has developed a test to check if the data conforms with the Kronecker product structure.

Under the matrix normal framework, let

$$\boldsymbol{\Omega}_{S_g} = \left(\omega_{S_g,i,j}\right)_{i,j=1}^{p}, \qquad \boldsymbol{R}_{S_g} = \boldsymbol{D}_{S_g}^{-1/2}\boldsymbol{\Omega}_{S_g}\boldsymbol{D}_{S_g}^{-1/2} = \left(\rho_{S_g,i,j}\right)_{i,j=1}^{p},$$

where $\boldsymbol{R}_{S_g}$ is the partial correlation matrix of the spatial locations, and $\boldsymbol{D}_{S_g}$ is the diagonal matrix of $\boldsymbol{\Omega}_{S_g}$. Our goal is to test, *simultaneously*, for $1 \leq i < j \leq p$,

$$H_{0,i,j} : \rho_{S_1,i,j} = \rho_{S_2,i,j} \quad \text{versus} \quad H_{1,i,j} : \rho_{S_1,i,j} \neq \rho_{S_2,i,j}. \tag{1}$$

Our solution is based upon a key observation that, in the context of brain connectivity analysis, the spatial precision matrix $\boldsymbol{\Omega}_{S_g}$, or more precisely, the spatial partial correlation matrix $\boldsymbol{R}_{S_g}$, is of the primary scientific interest, but the temporal precision matrix $\boldsymbol{\Omega}_{T_g}$ is not. We thus treat $\boldsymbol{\Omega}_{T_g}$, or equivalently $\boldsymbol{\Sigma}_{T_g}$, as a nuisance. Accordingly, we build our test statistic based on the linear transformation of the samples, $\boldsymbol{X}_g\boldsymbol{\Sigma}_{T_g}^{-1/2}$, and consider two scenarios: one assumes the temporal covariance $\boldsymbol{\Sigma}_{T_g}$ is known, and we term the method as an oracle procedure; the other uses a data-driven approach to estimate and plug in $\boldsymbol{\Sigma}_{T_g}$, and we term it a data-driven procedure. We show that, asymptotically, our proposed multiple testing procedures can control the false discovery at the pre-specified level. We compare in detail, both analytically in Section 4 and numerically in Section 5, with some alternative graph model hypothesis testing solutions, and demonstrate clear advantages of our approach. Our proposal provides a timely and useful inferential tool for brain connectivity alternation analysis.

The rest of the article is organized as follows. Section 2 develops the multiple testing procedure and Section 3 studies its asymptotic properties. Section 4 analytically compare our

method with some related solutions. Section 5 presents the numerical simulations, and Section 6 analyzes a real fMRI dataset. Section 7 extends the discussion to multiple populations. All technical proofs are relegated to an online supplement.

2 Testing Procedure

2.1 Data transformation

Let $\{\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)}\}$ and $\{\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)}\}$, each a matrix with dimension $p \times q$, denote two sets of i.i.d. random samples from two independent matrix normal distributions. The mean, without loss of generality, is assumed to be zero, and the covariance is of the form $\boldsymbol{\Sigma}_{S_g} \otimes \boldsymbol{\Sigma}_{T_g}$ for $g = 1, 2$. Assume $n_1 \asymp n_2$ and let $n = \max(n_1, n_2)$. Our goal is to detect spatial locations where the connectivity, in terms of the spatial partial correlation, differ across the two groups. Toward that goal, we separate the spatial and temporal dependence structures, and develop our testing procedure targeting the spatial partial correlation matrix $\boldsymbol{R}_{S_g}$, while treating $\boldsymbol{\Sigma}_{T_g}$ as a nuisance. Working with $\boldsymbol{R}_{S_g}$, rather than $\boldsymbol{\Omega}_{S_g}$, also avoids the identifiability issue between $\boldsymbol{\Sigma}_{S_g}$ and $\boldsymbol{\Sigma}_{T_g}$. We build the test based upon the linear transformation of the original samples. We first consider the scenario where $\boldsymbol{\Sigma}_{T_g}$ is known, and consider the transformed samples, $\boldsymbol{Y}_k^{(g)} = \boldsymbol{X}_k^{(g)} \boldsymbol{\Sigma}_{T_g}^{-1/2}, k = 1, \ldots, n_g, g = 1, 2$. We term the resulting test an oracle procedure. As $\boldsymbol{\Sigma}_{T_g}$ is rarely known in practice, we next consider the scenario where an estimator $\hat{\boldsymbol{\Sigma}}_{T_g}$ of $\boldsymbol{\Sigma}_{T_g}$ is plugged in, and consider the transformed samples, $\hat{\boldsymbol{Y}}_k^{(g)} = \boldsymbol{X}_k^{(g)} \hat{\boldsymbol{\Sigma}}_{T_g}^{-1/2}, k = 1, \ldots, n_g, g = 1, 2$. Accordingly, we term it as a data-driven procedure. There are multiple ways to estimate $\boldsymbol{\Sigma}_{T_g}$, or equivalently $\boldsymbol{\Omega}_{T_g}$. Examples include the usual sample covariance estimator, the banded estimator (Bickel and Levina (2008)), the adaptive thresholding estimator (Cai and Liu (2011)) for $\boldsymbol{\Sigma}_{T_g}$, or the Clime estimator (Cai et al. (2011)) for $\boldsymbol{\Omega}_{T_g}$. In Section 3.2, we give the necessary conditions for the estimators of $\boldsymbol{\Sigma}_{T_g}$ to guarantee the desired asymptotic properties.

In Section 5, we choose two estimators, the usual sample covariance estimator and the banded estimator, and numerically compare them in our context of matrix graph multiple testing.

2.2 Test statistics

We first develop the test statistics for the oracle case when $\mathbf{\Sigma}_{T_g}$ is known. The key is to describe the partial correlation matrix $\boldsymbol{R}_{S_g}$ in terms of a series of regression models (Anderson (2003, Sec 2.5)). Specifically, for the transformed samples $\boldsymbol{Y}_k^{(g)} = \boldsymbol{X}_k^{(g)} \mathbf{\Sigma}_{T_g}^{-1/2}$, we have,

$$Y_{k,i,l}^{(g)} = \boldsymbol{Y}_{k,-i,l}^{(g)\mathrm{T}} \boldsymbol{\beta}_{i,g} + \epsilon_{k,i,l}^{(g)}, \quad 1 \le i \le p, 1 \le l \le q, \tag{2}$$

where $\epsilon_{k,i,l}^{(g)} \sim N(0, \sigma_{S_g,i,i} - \mathbf{\Sigma}_{S_g,i,-i} \mathbf{\Sigma}_{S_g,-i,-i}^{-1} \mathbf{\Sigma}_{S_g,-i,i})$ and is independent of $\boldsymbol{Y}_{k,-i,l}^{(g)}$. The slope coefficient vector $\boldsymbol{\beta}_{i,g}$ and the error term $\epsilon_{k,i}^{(g)}$ satisfy that $\boldsymbol{\beta}_{i,g} = -\omega_{S_g,i,i}^{-1} \mathbf{\Omega}_{S_g,-i,i}$ and $r_{i,j,g} = \mathrm{cov}(\epsilon_{k,i,l}^{(g)}, \epsilon_{k,j,l}^{(g)}) = \frac{\omega_{S_g,i,j}}{\omega_{S_g,i,i}\omega_{S_g,j,j}}$. Therefore, the elements $\omega_{S_g,i,j}$ of the precision matrix $\mathbf{\Omega}_{S_g}$, and in turn the elements $\rho_{S_g,i,j}$ of the partial correlation matrix $\boldsymbol{R}_{S_g}$, can be represented in terms of $r_{i,j,g}$ from the regression model (2). A natural estimator of $r_{i,j,g}$ is the sample covariance between the residuals, $\tilde{r}_{i,j,g} = \frac{1}{n_g q} \sum_{k=1}^{n_g} \sum_{l=1}^{q} \hat{\epsilon}_{k,i,l}^{(g)} \hat{\epsilon}_{k,j,l}^{(g)}$, where $\hat{\epsilon}_{k,i,l}^{(g)} = Y_{k,i,l}^{(g)} - \bar{Y}_{i,l}^{(g)} - (\boldsymbol{Y}_{k,-i,l}^{(g)} - \bar{\boldsymbol{Y}}_{\cdot,-i,l}^{(g)})^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{i,g}$, and $\hat{\boldsymbol{\beta}}_{i,g}$ is an estimator of $\boldsymbol{\beta}_{i,g}$. We discuss the estimation of $\boldsymbol{\beta}_{i,g}$ in Section 2.3. When $i = j$, $\tilde{r}_{i,i,g}$ is a nearly unbiased estimator of $r_{i,i,g}$. However, when $i \neq j$, $\tilde{r}_{i,j,g}$ tends to be biased due to the correlation induced by the estimated parameters. We thus consider a bias-corrected estimator of $r_{i,j,g}$ of the form

$$\hat{r}_{i,j,g} = \begin{cases} -(\tilde{r}_{i,j,g} + \tilde{r}_{i,i,g}\hat{\beta}_{i,j,g} + \tilde{r}_{j,j,g}\hat{\beta}_{j-1,i,g}), & \text{when } \ i < j \\ \tilde{r}_{i,i,g}, & \text{when } \ i = j. \end{cases}$$

Based on this estimator $\hat{r}_{i,j,g}$ of $r_{i,j,g}$, we obtain an estimator of $\rho_{S_g,i,j}$,

$$\hat{\rho}_{S_g,i,j} = \frac{\hat{r}_{i,j,g}}{(\hat{r}_{i,i,g} \cdot \hat{r}_{j,j,g})^{1/2}}, \ \ 1 \le i < j \le p. \tag{3}$$

To estimate the variance of $\hat{\rho}_{S_g,i,j}$, we note that $\theta_{i,j,g} = \mathrm{var}\{\epsilon_{k,i}^{(g)} \epsilon_{k,j}^{(g)} / (r_{i,i,g} r_{j,j,g})^{1/2}\} / (n_g q) = (1 + \beta_{i,j,g}^2 r_{i,i,g} / r_{j,j,g}) / (n_g q)$. As such, the variance of $\hat{\rho}_{S_g,i,j}$ can be estimated by

$$\hat{\theta}_{i,j,g} = (1 + \hat{\beta}^2_{i,j,g}\hat{r}_{i,i,g}/\hat{r}_{j,j,g})/(n_g q). \tag{4}$$

Finally, for the hypothesis testing problem (1), we arrive at our test statistic,

$$W_{i,j} = (\hat{\rho}_{S_1,i,j} - \hat{\rho}_{S_2,i,j})/(\hat{\theta}_{i,j,1} + \hat{\theta}_{i,j,2})^{1/2}, \;\; 1 \le i < j \le p.$$

We next consider the data-driven case when $\mathbf{\Sigma}_{T_g}$ is unknown and estimated. With an estimator $\hat{\mathbf{\Sigma}}_{T_g}$ of $\mathbf{\Sigma}_{T_g}$ plugged in, the transformed samples are $\hat{\boldsymbol{Y}}_k^{(g)} = \boldsymbol{X}_k^{(g)}\hat{\mathbf{\Sigma}}_{T_g}^{-1/2}$. In addition, in this scenario, the regression coefficients vary at different time points. We thus replace (2) with

$$\hat{Y}^{(g)}_{k,i,l} = \hat{\boldsymbol{Y}}^{(g)\mathrm{T}}_{k,-i,l}\boldsymbol{\beta}_{i,l,g} + \epsilon^{(g)}_{k,i,l}, \quad 1 \le i \le p, 1 \le l \le q, \tag{5}$$

where $\boldsymbol{\beta}_{i,l,g}$ denotes the slope coefficient vector, and we estimate $\{\boldsymbol{\beta}_{i,l,g}, 1 \le l \le q\}$ by $\hat{\boldsymbol{\beta}}_{i,g}$, as will be discussed in Section 2.3. The rest of the setup is the same as in the oracle case, and we follow a similar process of building the test statistic $W_{i,j}$. We remark that, if one estimates $\mathbf{\Sigma}_{T_g}$ using the usual sample covariance estimator, it is biased up to a factor of $\mathrm{tr}(\mathbf{\Sigma}_{S_g})/p$. However, this bias does not affect the test statistic $W_{i,j}$, due to the two-step standardization employed in (3) and (4).

2.3 Estimation of slope

There are multiple ways to estimate the slope coefficient vector $\boldsymbol{\beta}_{i,g}$ in (2) and $\boldsymbol{\beta}_{i,l,g}$ in (5). To ensure the desired asymptotic properties, we require that the corresponding estimator satisfies the regularity condition (C4) in Section 3.1 for the oracle case, or the regularity condition (C5) in Section 3.2 for the data-driven case.

In our implementation, we use the Lasso to estimate the slope vector. Specifically, for the oracle case, we estimate $\boldsymbol{\beta}_{i,g}$ by

$$\hat{\boldsymbol{\beta}}_{i,g} = \boldsymbol{D}_{i,g}^{-\frac{1}{2}} \arg\min_{\boldsymbol{u}} \left\{ \frac{1}{2nq} \left| \left(\boldsymbol{Y}^{(g)}_{\cdot,-i} - \bar{\boldsymbol{Y}}^{(g)}_{(\cdot,-i)} \right) \boldsymbol{D}_{i,g}^{-1/2}\boldsymbol{u} - \left(\boldsymbol{Y}^{(g)}_{(i)} - \bar{\boldsymbol{Y}}^{(g)}_{(i)} \right) \right|_2^2 + \lambda_{n,i,g}|\boldsymbol{u}|_1 \right\}, \tag{6}$$

where $\boldsymbol{Y}^{(g)}$, $g = 1, 2$, is the $n_g q \times p$ data matrix by stacking the transformed samples $\boldsymbol{Y}_k^{(g)} = \boldsymbol{X}_k^{(g)} \boldsymbol{\Sigma}_{T_g}^{-1/2}$, $k = 1, \ldots, n_g$, $\boldsymbol{Y}_{(i)}^{(g)} = (Y_{1,i}^{(g)}, \ldots, Y_{n_g q,i}^{(g)})^{\mathrm{T}} \in \mathbb{R}^{n_g q \times 1}$, $\bar{\boldsymbol{Y}}_{(i)}^{(g)} = (\bar{Y}_i^{(g)}, \ldots, \bar{Y}_i^{(g)})^{\mathrm{T}} \in \mathbb{R}^{n_g q \times 1}$, with $\bar{Y}_i^{(g)} = \frac{1}{n_g q} \sum_{k=1}^{n_g q} Y_{k,i}^{(g)}$, $\bar{\boldsymbol{Y}}_{(\cdot,-i)}^{(g)} = (\bar{\boldsymbol{Y}}_{\cdot,-i}^{(g)\mathrm{T}}, \ldots, \bar{\boldsymbol{Y}}_{\cdot,-i}^{(g)\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n_g q \times (p-1)}$, with $\bar{\boldsymbol{Y}}_{\cdot,-i}^{(g)} = \frac{1}{n_g q} \sum_{k=1}^{n_g q} \boldsymbol{Y}_{k,-i}^{(g)}$, $\boldsymbol{D}_{i,g} = \mathrm{diag}(\hat{\boldsymbol{\Sigma}}_{S_g,-i,-i})$, $\hat{\boldsymbol{\Sigma}}_{S_g}$ is the sample covariance matrix with $n_g q$ transformed samples, $|\cdot|_2$ denotes the vector L_2 norm, and $|\cdot|_1$ denotes the vector L_1 norm. For the data-driven case, we replace the transformed samples $\boldsymbol{Y}_k^{(g)} = \boldsymbol{X}_k^{(g)} \boldsymbol{\Sigma}_{T_g}^{-1/2}$ with $\hat{\boldsymbol{Y}}_k^{(g)} = \boldsymbol{X}_k^{(g)} \hat{\boldsymbol{\Sigma}}_{T_g}^{-1/2}$, and estimate $\boldsymbol{\beta}_{i,l,g}$ accordingly. The tuning parameters $\lambda_{n,i,g}$ in (6) is selected adaptively given the data, following a similar procedure as in Xia and Li (2017).

2.4 Multiple testing

Next we develop a multiple testing procedure for $H_{0,i,j} : \rho_{S_1,i,j} = \rho_{S_2,i,j}$, so to identify spatial locations that have their conditional dependence changed between the two groups. We first describe the procedure in Algorithm 1, then the reasoning behind it. Once the test statistics are obtained, the testing algorithms are the same for the oracle and data-driven cases. As such, we use the same set of notations for both scenarios. We only differentiate them when we study their respective asymptotic properties in Section 3.

The key for our testing is to control the false discovery, since there are $(p^2 - p)/2$ simultaneous hypothesis tests. Let t be the threshold level such that $H_{0,i,j}$ is rejected if $|W_{i,j}| \geq t$. Then the false discovery proportion (FDP) and the false discovery rate (FDR) are defined as

$$\mathrm{FDP}(t) = \frac{\sum_{(i,j) \in \mathcal{H}_0} I(|W_{i,j}| \geq t)}{\sum_{1 \leq i < j \leq p} I(|W_{i,j}| \geq t) \vee 1}, \quad \mathrm{FDR}(t) = E\{\mathrm{FDP}(t)\}.$$

If the true nulls are known, we shall reject as many true positives as possible while controlling the false discovery proportion at the pre-specified level α. Here we choose the threshold level $t_0 = \inf\big\{0 \leq t \leq 2(\log p)^{1/2} : \ \mathrm{FDP}(t) \leq \alpha\big\}$. In practice, since the true nulls are unknown, we estimate $|\mathcal{H}_0|$ by $(p^2 - p)/2$, as it is at maximum $(p^2 - p)/2$ and is close to $(p^2 - p)/2$

Algorithm 1 Matrix graph hypothesis testing with FDR control.

1: Calculate the two-sample standardized test statistics $\{W_{i,j}, 1 \le i < j \le p\}$.

2: Estimate the false discovery proportion by

$$\widehat{\text{FDP}}(t) = \frac{2\{1-\Phi(t)\}(p^2-p)/2}{\sum_{1\le i<j\le p} I(|W_{i,j}| \ge t) \vee 1}.$$

3: For given $0 \le \alpha \le 1$, calculate

$$\hat{t} = \inf\left\{0 \le t \le 2(\log p)^{1/2} : \widehat{\text{FDP}}(t) \le \alpha\right\}. \tag{7}$$

If $\hat{t}$ does not exist, set $\hat{t} = 2(\log p)^{1/2}$.

4: For $1 \le i < j \le p$, reject $H_{0,i,j}$ if and only if $|W_{i,j}| \ge \hat{t}$.

when $\boldsymbol{R}_{S_1} - \boldsymbol{R}_{S_2}$ is sparse. Henceforth, we estimate the number of false rejections by $2\{1-\Phi(t)\}(p^2-p)/2$, where $\Phi(t)$ is the standard normal cumulative distribution function. This leads to the multiple testing Algorithm 1.

3 Theory

3.1 The oracle procedure

We first investigate the theoretical properties of the oracle multiple testing procedure. We begin with a set of regularity conditions.

(C1) There are constants $c_0, c_1 > 0$ such that, $c_0^{-1} \le \lambda_{\min}(\boldsymbol{\Omega}_{S_g}) \le \lambda_{\max}(\boldsymbol{\Omega}_{S_g}) \le c_0$, and $c_1^{-1} \le \lambda_{\min}(\boldsymbol{\Omega}_{T_g}) \le \lambda_{\max}(\boldsymbol{\Omega}_{T_g}) \le c_1$, for $g = 1, 2$. Besides, $\log p = o\{(nq)^{1/5}\}$.

(C2) Let $A_\tau = \big\{(i,j) : |\omega_{S_g,i,j}| \ge (\log p)^{-2-\tau}, 1 \le i < j \le p, \text{ for } g = 1 \text{ or } 2\big\}$. There exists some $\tau > 0$ such that $|A_\tau \cap \mathcal{H}_0| = o(p^\nu)$ for any $\nu > 0$.

(C3) Let

$$\mathcal{S}_\zeta = \left\{ (i,j) : 1 \le i < j \le p, \frac{|\rho_{S_1,i,j} - \rho_{S_2,i,j}|}{(\theta_{i,j,1} + \theta_{i,j,2})^{1/2}} \ge (\log p)^{1/2+\zeta} \right\}.$$

For some $\zeta, \delta > 0$, $|\mathcal{S}_\zeta| \ge [1/\{(8\pi)^{1/2}\alpha\} + \delta](\log\log p)^{1/2}$.

(C4) Let $\hat{\boldsymbol{\beta}}^o_{i,g}$ denote an estimator of $\boldsymbol{\beta}_{i,g}$ in (2) under the oracle scenario with the transformed samples $\boldsymbol{Y}^{(g)}_k = \boldsymbol{X}^{(g)}_k \boldsymbol{\Sigma}^{-1/2}_{T_g}, k = 1, \ldots, n_g, g = 1, 2$. Then $\max_{1\le i\le p} |\hat{\boldsymbol{\beta}}^o_{i,g} - \boldsymbol{\beta}_{i,g}|_1 = o_{\mathrm{p}}[\{\log\max(p,q,n)\}^{-1}]$, and $\max_{1\le i\le p} |\hat{\boldsymbol{\beta}}^o_{i,g} - \boldsymbol{\beta}_{i,g}|_2 = o_{\mathrm{p}}\left\{(nq\log p)^{-1/4}\right\}$.

A few remarks are in order regarding the regularity conditions. Condition (C1) is a technical condition that is commonly imposed in the high-dimensional hypothesis testing setting (Cai et al. (2013); Liu (2013); Xia et al. (2015)). Condition (C2) ensures that most of the regression residuals are not highly correlated with each other under the null hypothesis $H_{0,i,j} : \rho_{S_1,i,j} = \rho_{S_2,i,j}$. Condition (C3) is rather mild, because we have $(p^2-p)/2$ hypotheses in total, while this condition only requires a few entries of $\boldsymbol{R}_{S_1} - \boldsymbol{R}_{S_2}$ to have a standardized magnitude exceeding $\{(\log p)^{1/2+\rho}/(nq)\}^{1/2}$, for some constant $\rho > 0$. Condition (C4) places some requirements on the estimator $\hat{\boldsymbol{\beta}}^o_{i,g}$; it is easily satisfied by many estimation methods such as the Lasso and Dantzig selector. For instance, if one uses the Lasso, then (C4) is satisfied under (C1) and the sparsity condition $\max_{1\le i\le p} |\boldsymbol{\beta}_i|_0 = o[(nq)^{1/2}/\{\log\max(p,q,n)\}^{3/2}]$. These conditions are generally mild and reasonable.

The next theorem shows that our proposed oracle testing procedure controls the false discovery proportion and false discovery rate at the pre-specified level α, asymptotically.

Theorem 1. *Assume (C1) to (C4). Let $l_0 = |\mathcal{H}_0|$ and $l = (p^2-p)/2$. If $l_0 \ge c_0 p^2$ for some $c_0 > 0$, $p \le c_1(nq)^r$ for some $c_1, r > 0$, and $\hat{t}^o$ denotes the threshold value in* (7) *under the oracle case, then*

$$\lim_{(nq,p)\to\infty} \frac{\mathrm{FDR}(\hat{t}^o)}{\alpha l_0/l} = 1, \quad \textit{and} \quad \frac{\mathrm{FDP}(\hat{t}^o)}{\alpha l_0/l} \to 1 \ \textit{in probability, as } (nq,p) \to \infty.$$

By Theorem 1, when the number of brain network alternations between the two groups is small, i.e., $|\mathcal{H}_1| = o(p^2)$ where $\mathcal{H}_1$ is the set of alternatives where $\rho_{S_1,i,j} \neq \rho_{S_2,i,j}$, then we have $\lim_{(nq,p)\to\infty} \text{FDR}(\hat{t}^o) = \alpha$ and $\text{FDP}(\hat{t}^o) \to \alpha$ in probability, as $(nq, p) \to \infty$.

3.2 The data-driven procedure

We next derive the theoretical properties of the data-driven testing procedure. We continue to employ the regularity conditions (C1) to (C3), slightly modify the condition (C4) to (C5), and add a new condition, (C6), that places some constraint on the estimated temporal covariance matrix $\hat{\boldsymbol{\Sigma}}_{T_g}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, let the matrix element-wise infinity norm be $\|\boldsymbol{A}\|_\infty = \max_{1\le i,j\le p} |a_{i,j}|$, and the matrix 1-norm be $\|\boldsymbol{A}\|_{L_1} = \max_{1\le j\le p} \sum_{i=1}^p |a_{i,j}|$. Let $(\omega_{l,i,j}^{(g)})_{p\times p}$ denote the elements of $\boldsymbol{\Omega}_l^{(g)} = \text{cov}^{-1}\{(\boldsymbol{X}_k^{(g)}\hat{\boldsymbol{\Sigma}}_{T_g}^{-1/2})_{\cdot,l}\}$. Take $r_{1,n,p,q} = \{s_{p,q} q \log^{3/2} \max(p,q,n)\|\boldsymbol{\Omega}_{S_g}\|_{L_1}^2\}^{-1}$, $r_{2,n,p,q} = \{nqs_{p,q}^2 \log p \log^2 \max(p,q,n)\}^{-1/4}(q\|\boldsymbol{\Omega}_{S_g}\|_{L_1}^2)^{-1}$, and $s_{p,q} = \max_{g=1,2} \max_{1\le l\le q} \max_{1\le i\le p} \sum_{j=1}^p \max\{I(\omega_{S_g,i,j} \neq 0), I(\omega_{l,i,j}^{(g)} \neq 0)\}$.

(C5) Let $\hat{\boldsymbol{\beta}}_{i,g}^d$ denote an estimator of $\boldsymbol{\beta}_{i,l,g}$ under the data-driven scenario with $\hat{\boldsymbol{Y}}_k^{(g)} = \boldsymbol{X}_k^{(g)}\hat{\boldsymbol{\Sigma}}_{T_g}^{-1/2}$, $k = 1,\ldots,n_g, g = 1,2$. Assume $\max_{1\le i\le p, 1\le l\le q} |\hat{\boldsymbol{\beta}}_{i,g}^d - \boldsymbol{\beta}_{i,l,g}|_1 = o_{\text{p}}[\{\log\max(p,q,n)\}^{-1}]$, and $\max_{1\le i\le p, 1\le l\le q} |\hat{\boldsymbol{\beta}}_{i,g}^d - \boldsymbol{\beta}_{i,l,g}|_2 = o_{\text{p}}\left\{(nq\log p)^{-1/4}\right\}$.

(C6) Let $\hat{\boldsymbol{\Sigma}}_{T_g}$ be an estimator of $\boldsymbol{\Sigma}_{T_g}$ satisfying $\|\hat{\boldsymbol{\Sigma}}_{T_g}^{-1/2} - c\boldsymbol{\Sigma}_{T_g}^{-1/2}\|_\infty = O_{\text{p}}(r_{n,p,q})$ for an arbitrary constant $c > 0$. Assume $r_{n,p,q} = o\{\min(r_{1,n,p,q}, r_{2,n,p,q})\}$.

The estimator of $\boldsymbol{\beta}_{i,l,g}$ in condition (C5) can be obtained by the Lasso and many other approaches. The estimator of $\boldsymbol{\Sigma}_{T_g}$ in (C6) can be obtained in many ways too. To guarantee the desired theoretical properties, the estimator $\hat{\boldsymbol{\Sigma}}_{T_g}$ needs to satisfy the condition, $\|\hat{\boldsymbol{\Sigma}}_{T_g} - c\boldsymbol{\Sigma}_{T_g}\|_\infty = O_{\text{p}}[\{\log q/(np)\}^{1/2}]$, for an arbitrary constant $c > 0$. The banded estimator of Bickel and Levina (2008) and the adaptive thresholding estimator of Cai and Liu (2011) both satisfy this condition, and thus can be used in conjunction with our testing procedure. Alternatively, one can

directly estimate the precision matrix $\mathbf{\Omega}_{T_g}$ and base the testing procedures on $\{\boldsymbol{X}_k\hat{\mathbf{\Omega}}_{T_g}^{1/2}\}_{k=1}^n$, as long as the estimator $\hat{\mathbf{\Omega}}_{T_g}$ satisfies $\|\hat{\mathbf{\Omega}}_{T_g} - c\mathbf{\Omega}_{T_g}\|_\infty = O_{\mathrm{p}}\Big[\|\mathbf{\Omega}_{T_g}\|_{L_1}^2\{\log q/(np)\}^{1/2}\Big]$, for some constant $c > 0$. For instance, if the temporal precision matrix is sparse, in the sense that $\max_{g=1,2}\max_{1\le i\le q}\sum_{j=1}^q I(\omega_{T_g,i,j} \neq 0) \le c'$ for some constant $c' > 0$, then the Clime estimator of Cai et al. (2011) can be employed.

The next theorem summarizes the asymptotic properties of the data-driven procedure.

Theorem 2. *Assume (C1) to (C3), (C5) and (C6). If $\hat{t}^d$ denotes the threshold value in* (7) *under the data-driven case, and assuming the same conditions as in Theorem 1, then*

$$\lim_{(nq,p)\to\infty}\frac{\mathrm{FDR}(\hat{t}^d)}{\alpha l_0/l} = 1, \quad \textit{and} \quad \frac{\mathrm{FDP}(\hat{t}^d)}{\alpha l_0/l} \to 1 \ \textit{in probability, as } (nq,p)\to\infty.$$

This essentially shows that the data-driven multiple testing procedure performs asymptotically as well as the oracle procedure. That is, as long as the estimation of the temporal covariance structure is reasonably good, the data-driven procedure controls both the FDR and FDP at the pre-specified level α, asymptotically, under the same conditions as the oracle case.

4 Comparison

4.1 Estimation versus inference

Both sparse graph estimation and graph inference procedures can produce, in effect, a sparse representation of the network structure. However, the two classes of solutions differ in several ways. The key of graph estimation is to seek a bias-variance tradeoff, and many sparse graph estimators such as those of Yuan and Lin (2007) and Friedman et al. (2008) are biased. Our graph testing method, instead, requires and is built upon a nearly unbiased estimator. Moreover, graph estimation methods do not produce a direct quantification of statistical significance for individual network edges. Practically they may enjoy a high true positive discovery rate

(power), but there is no explicit control of false positive rate (significance level). Our solution both produces significance quantification and controls the false discovery explicitly. As such, it is distinct from the majority of existing sparse graph estimation solutions.

4.2 One-sample test versus two-sample test

Liu (2013) proposed a graph hypothesis testing procedure under the vector normal distribution, Chen and Liu (2015) and Xia and Li (2017) under the matrix normal, while all those tackled one-sample testing. Our method aims at two-sample testing. They are two different types of hypothesis testing problems. We next outline the major differences in terms of their research goals, test statistics, and theoretical tools.

Difference in goals: The two-sample testing method proposed in this article aims to detect the *alternation* of magnitude in spatial partial correlations across different groups. The one-sample solution produces evidence about the existence of spatial conditional dependence, i.e., whether $\omega_{S_g,i,j} = 0$, but no knowledge about the magnitude of such dependence. Thus the aims of the two approaches are completely different, and the conclusion of the two-sample test cannot be obtained from the one-sample test.

Difference in test statistics: In the one-sample test of Xia and Li (2017), the testing of the precision matrix and the partial correlation matrix are equivalent. As such, the test statistics were constructed based on the estimates of $\omega_{S_g,i,j}$. In the two-sample case, the precision matrices are unidentifiable, and the test aims at the equality of the two partial correlation matrices. Accordingly, the test statistics are based on the estimates of $\rho_{S_1,i,j} - \rho_{S_2,i,j}$. This requires two standardization processes: the standardization of the estimates of $\omega_{S_g,i,j}$, and the standardization of the difference of the estimates of $\rho_{S_g,i,j}$. This double standardization implies an increased level of complexity, and requires a different set of technical tools for the theoretical analysis of the test statistics, as shown next.

Difference in theoretical proofs: The asymptotic analysis of the two-sample procedure is technically much more involved than that of the one-sample procedure. First, in the one-sample case, $\omega_{S,i,j} = 0$ under the null hypothesis. As a result, it is relatively easy to establish the asymptotic normality of $\hat{\rho}_{i,j}$, then $W_{i,j}$. In the two-sample case, however, $\rho_{S_1,i,j}$ and $\rho_{S_2,i,j}$ are not necessarily equal to 0 under the null, and the standardized statistics $W_{i,j}$ is not asymptotically normal. To overcome this difficulty, we divide the set of indices into two subsets, one with a negligible correction and the other requires a major correction. Accordingly, different technical tools are employed for these two subsets to eventually establish the asymptotic normality of the corrected version of $W_{i,j}$. Second, in the one-sample setting, the residuals are weakly dependent with each other because $\boldsymbol{R}_S = \boldsymbol{I}$ under the null, whereas in the two-sample setting, the test statistics can be highly dependent since $\boldsymbol{R}_{S_g}$ is not necessarily an identity matrix. To show error rate control, reorganization of the set of test statistics according to the level of dependency is essential. Some special tools have been employed to show the negligibility of the highly dependent pairs. More details are given in the proofs in the online supplement.

4.3 Vector normal test versus matrix normal test

For the two-sample testing problem, Xia et al. (2015) developed a test under the vector normal distribution and established its asymptotic properties. Narayan et al. (2015) tackled the problem under the matrix normal distribution, but transformed the problem back to the vector normal case by employing a whitening preprocessing step to help induce independent columns of the matrix data. Our transformation of the data and the separation of the spatial and temporal covariances can be viewed, at the conceptual level, as a version of whitening for the matrix-valued quantity. However, our procedure differs from the classical whitening, resulting in different properties, both computationally and theoretically.

Difference in computation: Classical whitening seeks an unbiased estimator of the $q \times q$

temporal covariance matrix at every spatial location based on the n samples. As a result, it is computationally expensive, and requires $q < n$ if the usual sample covariance estimator is employed. This can be restrictive in brain connectivity analysis, since the temporal dimension q can easily exceed the sample size n. Our method in effect pools the np correlated samples to estimate $\mathbf{\Sigma}_{T_g}$, and as such it does *not* require $q < n$, and is much faster to compute. Due to the correlations among the np samples, the pooled estimator of $\mathbf{\Sigma}_{T_g}$ is unbiased only up to a constant. However, our test statistics, by construction, are not affected by this constant.

Difference in theory: To ensure the data-driven procedure performs asymptotically as well as the oracle one as if $\mathbf{\Sigma}_{T_g}$ were known, we require the estimator of $\mathbf{\Sigma}_{T_g}$ to satisfy (C6), in that some norm of $\widehat{\mathbf{\Sigma}}_{T_g}^{-1/2} - c\mathbf{\Sigma}_{T_g}^{-1/2}$, with $c > 0$ a constant, satisfies a given convergence rate. By pooling np samples, our estimator meets this requirement. However, the estimator of the conventional whitening procedure does not satisfy (C6), and thus cannot guarantee the asymptotic performance of the data-driven test.

5 Simulations

We have carried out intensive simulations to study the finite-sample performance of the proposed testing procedures. We have numerically compared the oracle and data-driven tests, compared the data-driven tests with the usual sample covariance estimator as the plug-in and the banded covariance estimator as the plug-in, and compared our proposed tests with the solution that first whitens and de-correlates the columns of the matrix data, then applies the two-sample test of Xia et al. (2015).

Specifically, we generated n i.i.d. samples from a matrix normal distribution with the precision matrix $\mathbf{\Omega}_{S_g} \otimes \mathbf{\Omega}_{T_g}$, $\mathbf{\Omega}_{S_g} \in \mathbb{R}^{p\times p}, \mathbf{\Omega}_{T_g} \in \mathbb{R}^{q\times q}, g = 1, 2$. We examined a range of spatial and temporal dimensions and the sample sizes: $p = \{50, 200, 800\}$, $q = \{50, 200\}$,

and $n = n_1 = n_2 = \{15, 50\}$. These values are consistent with the usual setup in functional connectivity analysis. We considered two temporal covariance structures: an autoregressive model, $\mathbf{\Sigma}_{T_g} = (\sigma_{T_g,i,j})$, with elements $\sigma_{T_1,i,j} = 0.4^{|i-j|}$, and $\sigma_{T_2,i,j} = 0.5^{|i-j|}$, $1 \leq i, j \leq p$; and a moving average model, $\mathbf{\Sigma}_{T_g} = (\sigma_{T_g,i,j})$, with nonzero elements $\sigma_{T_1,i,j} = 1/(|i-j|+1)$, for $|i-j| \leq 3$, and $\sigma_{T_2,i,j} = 1/(|i-j|+1)$, for $|i-j| \leq 4$. We also considered three spatial covariance structures: a banded graph, with bandwidth equal to 3 (Zhao et al. (2012)); a hub graph, with row and columns evenly partitioned into 20 disjoint groups; and a small-world graph, with 5 starting neighbors and 5% probability of rewiring (van Wieringen and Peeters (2014)). We first generated $\mathbf{\Omega}_{S_1}$ according to one of these spatial graph models, then constructed $\mathbf{\Omega}_{S_2}$ by randomly eliminating m percent of edges of $\mathbf{\Omega}_{S_1}$, with $m = 10\%$ and 50%. Tables 1 to 4 summarize the empirical FDR and the empirical power, in percentages, of various testing procedures based on 100 data replications. The significance level was set at $\alpha = 1\%$. The power was calculated as $100^{-1} \sum_{l=1}^{100} \{\sum_{(i,j)\in\mathcal{H}_1} I(|W_{i,j,l}| \geq \hat{t})\}/|\mathcal{H}_1|$, where $W_{i,j,l}$ denotes the test statistic for the l-th replication and $\mathcal{H}_1$ denotes the nonzero locations.

For the empirical FDR, we see from the tables that, the data-driven procedure based on the banded temporal covariance estimator achieves an FDR well under control of the specified significance level across all settings. Its performance further improves as (p, q) increases, and is close to that of the oracle procedure. The data-driven procedure based on the sample temporal covariance estimator is outperformed by the banded estimator based procedure when $n = 15$. This is because the sample covariance estimator is incapable of estimating the true covariance matrix well with a large q and a relatively small value of np. However, as n and p grow, its performance improves, and gets closer to the oracle procedure. On the other hand, the alternative whitening based testing solution suffers from some obvious FDR distortion, especially for the moving average temporal model when the sample size is relatively small ($n = 15$). Moreover, as (n, p, q) increases, for all testing procedures, the standard error of the

empirical FDR decreases, whereas our oracle and data-driven procedures perform similarly and achieve smaller standard errors than the whitening based solution.

For the empirical power, the proposed testing procedures are more powerful than the whitening based procedure. The data-driven procedure based on the banded estimator again performs similarly as the oracle, and outperforms the one based on the sample covariance estimator when (n, p) is small. Moreover, the empirical power decreases when p increases, since there are more edges to estimate, and the power increases when q increases, as there are more samples to estimate the graph structure. The empirical powers are low for the hub graph and the small-world graph when $n = 15$, $q = 50$, and $p = 800$. This is due to the fact that the magnitudes of the partial correlations generated are very small for these two graphs, and, as such, are difficult to detect when (n, q) is small but the spatial dimension p is large. Theoretically, one needs the difference of partial correlations to have the magnitude exceeding $c(\log p/(nq))^{1/2}$, for some constant $c > 0$. We see from Tables 1 and 2, when q grows to 200, the powers become much larger for those two graphs. Similarly, when the sample size becomes larger, i.e., when $n = 50$, we see from Tables 3 and 4 that the empirical powers are much higher than those when $n = 15$.

Table 5 reports the computation time, in seconds, for the four methods discussed, for a single data replication. We fix the spatial structure as banded, and the temporal structure as autoregressive, while we vary (n, p, q). The computation time for other spatial and temporal structures shows a similar qualitative pattern and is omitted here. From this table, we see that the data-driven procedure is slightly slower than the oracle procedure, as the former involves an additional step of temporal covariance estimation. However the overall computation times are comparable. The classical whitening based procedure is computationally much more expensive. Especially when the temporal dimension is large, e.g., $q = 200$, the computation time of the whitening-based procedure can be as large as 10 times of that of our method.

Table 1: Multiple testing empirical FDR (with standard error in parenthesis) and empirical power, both in percentage. Methods under comparison are the oracle procedure ("oracle"), the data-driven procedure based on the sample temporal covariance estimator ("data-driven-S"), the data-driven procedure based on the banded temporal covariance estimator ("data-driven-B"), and the two-sample test under vector normal after classical whitening ("whitening"). This table is for the autoregressive temporal structure, $\alpha = 1\%$, and $n = 15$.

m		$m = 10\%$						$m = 50\%$					
p		$q = 50$			$q = 200$			$q = 50$			$q = 200$		
	spatial structure	banded	hub	small	banded	hub	small	banded	hub	small	banded	hub	small
		Empirical FDR (SE) (in %), autoregressive temporal structure											
50	oracle	0.3 (1.5)	1.0 (5.3)	0.5 (1.7)	0.8 (2.2)	0.5 (3.5)	0.7 (1.6)	0.6 (0.8)	0.3 (1.4)	1.1 (1.4)	0.6 (1.0)	0.9 (2.4)	0.7 (0.7)
	data-driven-S	0.5 (1.8)	1.1 (5.3)	0.8 (2.4)	0.0 (0.0)	0.0 (0.0)	0.3 (1.1)	0.6 (0.9)	0.3 (1.3)	0.8 (1.4)	0.2 (0.6)	0.1 (0.8)	0.3 (0.5)
	data-driven-B	0.4 (1.6)	1.0 (4.9)	0.5 (1.6)	0.8 (2.3)	0.3 (2.5)	0.8 (1.9)	0.7 (0.8)	0.4 (1.4)	1.2 (1.8)	0.7 (1.0)	1.0 (2.4)	0.7 (0.7)
	whitening	1.2 (2.6)	1.8 (6.4)	0.8 (2.4)	1.3 (2.8)	3.6 (9.4)	0.8 (1.9)	1.2 (1.4)	0.9 (2.5)	1.5 (2.0)	1.1 (1.4)	1.6 (3.2)	1.1 (1.0)
200	oracle	0.4 (0.9)	1.0 (3.1)	0.4 (1.0)	0.7 (1.0)	0.7 (1.4)	0.7 (0.8)	0.6 (0.5)	0.7 (1.3)	1.2 (1.1)	0.9 (0.5)	1.0 (0.9)	0.8 (0.3)
	data-driven-S	0.4 (0.9)	1.1 (3.2)	0.5 (1.1)	0.5 (1.0)	0.3 (0.8)	0.6 (0.8)	0.6 (0.5)	0.6 (1.2)	1.1 (1.0)	0.8 (0.5)	0.5 (0.6)	0.7 (0.3)
	data-driven-B	0.4 (0.9)	1.0 (3.1)	0.5 (1.0)	0.7 (1.0)	0.8 (1.5)	0.7 (0.8)	0.7 (0.5)	0.7 (1.3)	1.2 (1.1)	0.9 (0.5)	1.0 (0.9)	0.8 (0.3)
	whitening	1.0 (1.5)	1.3 (3.7)	1.2 (1.8)	1.1 (1.3)	1.6 (2.8)	1.0 (1.0)	1.0 (0.6)	1.4 (1.8)	1.6 (1.3)	1.3 (0.6)	1.6 (1.1)	1.2 (0.4)
800	oracle	0.4 (0.4)	0.0 (0.0)	0.5 (0.7)	0.6 (0.4)	0.8 (1.0)	0.6 (0.3)	0.7 (0.2)	1.2 (4.2)	0.9 (0.7)	0.8 (0.2)	0.9 (0.6)	0.7 (0.2)
	data-driven-S	0.4 (0.4)	0.5 (1.5)	0.5 (0.7)	0.6 (0.4)	0.8 (1.1)	0.6 (0.3)	0.7 (0.2)	1.2 (4.2)	0.9 (0.7)	0.7 (0.2)	0.8 (0.6)	0.7 (0.2)
	data-driven-B	0.4 (0.4)	0.0 (0.0)	0.5 (0.7)	0.6 (0.4)	0.8 (1.0)	0.6 (0.3)	0.7 (0.2)	1.2 (4.2)	0.9 (0.7)	0.8 (0.2)	0.9 (0.6)	0.7 (0.2)
	whitening	0.8 (0.6)	1.6 (0.6)	0.9 (1.2)	1.1 (0.6)	2.0 (2.0)	1.0 (0.5)	1.1 (0.3)	1.7 (5.3)	1.4 (1.1)	1.2 (0.3)	1.7 (0.9)	1.2 (0.3)
		Empirical power (SE) (in %), autoregressive temporal structure											
50	oracle	92.1	99.6	62.4	100.0	100.0	100.0	93.1	99.7	37.9	100.0	100.0	100.0
	data-driven-S	87.4	98.7	53.0	100.0	100.0	100.0	88.9	99.5	28.7	100.0	100.0	99.8
	data-driven-B	91.4	99.3	60.6	100.0	100.0	100.0	92.4	99.8	36.6	100.0	100.0	100.0
	whitening	88.2	99.2	54.2	100.0	100.0	100.0	86.3	99.6	29.7	100.0	100.0	99.9
200	oracle	89.3	51.5	39.6	100.0	100.0	100.0	90.9	53.9	24.7	100.0	100.0	99.9
	data-driven-S	88.1	50.3	37.9	100.0	100.0	99.9	89.7	51.3	23.1	100.0	100.0	99.8
	data-driven-B	88.9	51.3	39.4	100.0	100 .0	100.0	90.8	53.1	24.5	100.0	100.0	99.9
	whitening	83.9	43.1	33.8	100.0	100.0	99.9	82.5	45.7	19.3	100.0	100.0	99.5
800	oracle	87.8	1.6	18.8	100.0	75.8	99.9	88.5	0.4	8.8	100.0	61.6	99.6
	data-driven-S	87.2	1.6	18.3	100.0	72.1	99.9	88.2	0.4	8.5	100.0	59.0	99.5
	data-driven-B	87.5	1.6	18.7	100.0	75.5	99.9	88.5	0.4	8.8	100.0	61.4	99.6
	whitening	81.7	1.4	15.2	100.0	62.2	99.7	78.1	0.3	6.3	100.0	50.8	98.2

Table 2: Multiple testing empirical FDR (with standard error in parenthesis) and empirical power, both in percentage. This table is for the moving average temporal structure, $\alpha = 1\%$, and $n = 15$.

m		$m = 10\%$						$m = 50\%$					
p		$q = 50$			$q = 200$			$q = 50$			$q = 200$		
	spatial structure	banded	hub	small	banded	hub	small	banded	hub	small	banded	hub	small
		Empirical FDR (SE) (in %), autoregressive temporal structure											
50	oracle	0.8 (2.1)	0.3 (2.5)	0.5 (1.7)	0.6 (1.9)	0.7 (3.7)	0.7 (1.4)	0.7 (1.0)	0.9 (2.5)	1.2 (1.5)	0.9 (1.1)	0.8 (2.1)	0.7 (0.8)
	data-driven-S	0.5 (1.8)	0.0 (0.0)	1.1 (2.5)	0.3 (1.6)	0.3 (2.5)	0.3 (0.9)	0.5 (1.0)	0.4 (1.6)	1.0 (1.5)	0.2 (0.6)	0.1 (0.8)	1.9 (0.4)
	data-driven-B	0.7 (2.1)	0.3 (2.5)	0.5 (1.7)	0.9 (2.0)	0.5 (3.5)	0.7 (1.4)	0.5 (1.0)	0.9 (2.5)	1.2 (1.6)	0.9 (1.1)	0.5 (1.7)	0.7 (0.8)
	whitening	1.1 (3.1)	1.9 (7.1)	1.5 (3.6)	1.3 (2.6)	1.5 (6.0)	1.4 (2.4)	1.3 (1.4)	1.7 (3.2)	2.2 (2.1)	1.6 (1.5)	1.7 (3.8)	1.6 (1.1)
200	oracle	0.6 (1.0)	0.7 (2.8)	0.7 (1.6)	0.8 (1.1)	0.7 (1.5)	0.7 (0.8)	0.7 (0.4)	1.0 (1.4)	1.1 (0.9)	1.0 (0.5)	1.0 (0.9)	0.9 (0.3)
	data-driven-S	0.7 (1.1)	0.9 (3.3)	0.7 (1.4)	0.6 (1.0)	0.3 (0.9)	0.6 (0.7)	0.7 (0.5)	0.8 (1.1)	1.1 (1.0)	0.8 (0.5)	0.5 (0.6)	0.7 (0.3)
	data-driven-B	0.6 (1.0)	0.8 (2.8)	0.8 (1.6)	0.8 (1.1)	0.8 (1.5)	0.7 (0.8)	0.7 (0.4)	1.0 (1.3)	1.1 (0.9)	1.0 (0.5)	1.0 (0.9)	0.9 (0.3)
	whitening	0.9 (1.4)	1.8 (4.6)	1.2 (2.3)	1.7 (1.7)	2.3 (3.2)	1.5 (1.3)	1.3 (0.8)	2.0 (2.3)	2.1 (1.5)	1.6 (0.7)	2.1 (1.6)	1.5 (0.6)
800	oracle	0.4 (0.4)	0.0 (0.0)	0.5 (0.6)	0.6 (0.4)	0.8 (1.1)	0.6 (0.4)	0.7 (0.2)	1.2 (4.2)	1.0 (0.7)	0.8 (0.2)	0.9 (0.6)	0.7(0.1)
	data-driven-S	0.4 (0.4)	0.5 (1.6)	0.5 (0.7)	0.6 (0.4)	0.8 (1.1)	0.6 (0.4)	0.7 (0.2)	1.2 (4.2)	0.9 (0.7)	0.7 (0.2)	0.8 (0.5)	0.7(0.1)
	data-driven-B	0.4 (0.4)	0.0 (0.0)	0.5 (0.6)	0.6 (0.4)	0.8 (1.1)	0.6 (0.4)	0.7 (0.2)	1.2 (4.2)	1.0 (0.7)	0.8 (0.2)	0.9 (0.6)	0.7(0.1)
	whitening	1.0 (0.8)	1.6 (3.2)	1.2 (1.4)	1.3 (0.7)	2.4 (2.1)	1.3 (0.5)	1.3 (0.4)	1.5 (4.7)	1.8 (1.8)	1.5 (0.3)	2.3 (1.1)	1.5 (0.3)
		Empirical power (SE) (in %), moving average temporal structure											
50	oracle	92.6	99.3	62.7	100.0	100.0	100.0	93.3	99.8	44.9	100.0	100.0	99.9
	data-driven-S	88.1	98.3	55.9	100.0	100.0	99.2	89.2	99.5	34.9	100.0	100.0	95.7
	data-driven-B	92.6	99.3	62.7	100.0	100.0	100.0	93.3	99.8	44.9	100.0	100.0	99.9
	whitening	83.3	97.9	49.9	100.0	100.0	99.9	82.9	98.9	31.5	100.0	100.0	98.7
200	oracle	90.4	51.4	31.9	100.0	100.0	100.0	91.5	54.5	23.1	100.0	100.0	99.9
	data-driven-S	88.9	49.1	30.1	100.0	100.0	99.9	90.3	52.5	21.1	100.0	100.0	99.8
	data-driven-B	90.1	51.4	31.8	100.0	100.0	100.0	91.4	54.2	22.8	100.0	100.0	99.9
	whitening	79.2	39.1	21.2	100.0	100.0	99.6	77.2	41.9	15.4	100.0	100.0	98.7
800	oracle	87.8	1.6	18.8	100.0	75.7	99.9	88.5	0.4	8.8	100.0	61.6	99.6
	data-driven-S	87.3	1.5	18.4	100.0	72.1	99.9	88.2	0.4	8.5	100.0	59.1	99.5
	data-driven-B	87.6	1.5	18.7	100.0	76.0	99.9	88.5	0.4	8.8	100.0	61.1	99.6
	whitening	75.3	1.3	11.7	100.0	55.3	99.0	71.6	0.3	5.1	100.0	45.5	96.1

Table 3: Multiple testing empirical FDR (with standard error in parenthesis) and empirical power, both in percentage. This table is for the autoregressive temporal structure, $\alpha = 1\%$, and $n = 50$.

m		$m = 10\%$						$m = 50\%$					
p		$q = 50$			$q = 200$			$q = 50$			$q = 200$		
	spatial structure	banded	hub	small	banded	hub	small	banded	hub	small	banded	hub	small
		Empirical FDR (SE) (in %), autoregressive temporal structure											
50	oracle	0.6 (1.9)	1.0 (4.9)	0.8 (1.5)	1.1 (2.7)	1.2 (5.4)	1.0 (2.0)	0.9 (1.2)	0.6 (1.8)	0.8 (1.0)	0.7 (1.0)	1.1 (2.4)	0.8 (0.8)
	data-driven-S	0.7 (2.0)	1.3 (5.5)	0.8 (1.6)	0.7 (2.0)	1.2 (5.4)	1.0 (2.0)	0.9 (1.2)	0.6 (1.8)	0.8 (1.0)	0.5 (0.8)	0.6 (1.8)	0.6 (0.6)
	data-driven-B	0.5 (1.7)	1.0 (4.9)	0.8 (1.5)	1.1 (2.7)	1.2 (5.4)	0.9 (2.0)	0.9 (1.1)	0.5 (1.7)	0.8 (1.0)	0.7 (1.2)	1.1 (2.4)	0.8 (0.8)
	whitening	0.8 (2.2)	1.5 (6.0)	1.0 (1.9)	1.1 (2.7)	2.2 (7.5)	1.1 (2.1)	1.3 (1.3)	0.7 (2.0)	1.2 (1.1)	0.9 (1.2)	1.1 (2.4)	1.0 (0.9)
200	oracle	0.4 (0.7)	1.0 (2.2)	0.7 (0.8)	0.8 (1.0)	1.1 (2.2)	0.7 (0.8)	0.8 (0.5)	0.9 (0.9)	0.7 (0.3)	0.8 (0.5)	1.0 (1.0)	0.8 (0.3)
	data-driven-S	0.3 (0.6)	1.0 (2.1)	0.7 (0.8)	0.9 (1.1)	0.8 (2.0)	0.8 (0.8)	0.8 (0.5)	0.8 (0.9)	0.7 (0.3)	0.8 (0.5)	0.8 (0.8)	0.8 (0.3)
	data-driven-B	0.3 (0.7)	0.9 (2.1)	0.7 (0.8)	0.9 (1.0)	1.1 (2.2)	0.7 (0.8)	0.8 (0.5)	0.9 (0.9)	0.7 (0.3)	0.8 (0.5)	1.0 (1.0)	0.8 (0.3)
	whitening	0.6 (0.9)	1.3 (2.5)	0.9 (1.1)	1.1 (1.4)	1.4 (2.3)	1.0 (0.8)	1.0 (0.6)	1.3 (1.0)	1.0 (0.4)	1.2 (0.6)	1.2 (1.0)	1.1 (0.4)
800	oracle	0.5 (0.4)	1.0 (1.4)	0.6 (0.3)	0.8 (0.5)	0.8 (0.9)	0.7 (0.4)	0.8 (0.2)	1.0 (0.7)	0.7 (0.2)	0.8 (0.2)	0.9 (0.4)	0.8 (0.1)
	data-driven-S	0.5 (0.4)	1.0 (1.4)	0.6 (0.3)	0.8 (0.5)	0.7 (0.8)	0.7 (0.4)	0.8 (0.2)	1.0 (0.7)	0.7 (0.2)	0.8 (0.2)	0.9 (0.4)	0.8 (0.1))
	data-driven-B	0.5 (0.4)	1.0 (1.4)	0.6 (0.3)	0.8 (0.5)	0.8 (0.9)	0.7 (0.4)	0.8 (0.2)	1.0 (0.7)	0.7 (0.2)	0.8 (0.2)	0.9 (0.4)	0.8 (0.1)
	whitening	0.9 (0.5)	1.5 (1.8)	0.9 (0.5)	1.1 (0.6)	1.3 (1.3)	1.0 (0.5)	1.1 (0.3)	1.5 (1.1)	1.1 (0.2)	1.1 (0.3)	1.3 (0.5)	1.0 (0.2)
		Empirical power (SE) (in %), autoregressive temporal structure											
50	oracle	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	data-driven-S	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	data-driven-B	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	whitening	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
200	oracle	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	99.6	100.0	100.0	100.0
	data-driven-S	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	99.6	100.0	100.0	100.0
	data-driven-B	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	99.6	100.0	100.0	100.0
	whitening	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	99.3	100.0	100.0	100.0
800	oracle	100.0	58.6	99.4	100.0	100.0	100.0	100.0	44.1	98.0	100.0	100.0	100.0
	data-driven-S	100.0	57.4	99.4	100.0	100.0	100.0	100.0	43.7	98.0	100.0	100.0	100.0
	data-driven-B	100.0	58.4	99.4	100.0	100.0	100.0	100.0	43.5	98.0	100.0	100.0	100.0
	whitening	100.0	53.2	99.1	100.0	100.0	100.0	100.0	40.2	96.9	100.0	100.0	100.0

Table 4: Multiple testing empirical FDR (with standard error in parenthesis) and empirical power, both in percentage. This table is for the moving average temporal structure, $\alpha = 1\%$, and $n = 50$.

m		$m = 10\%$						$m = 50\%$					
p		$q = 50$			$q = 200$			$q = 50$			$q = 200$		
	spatial structure	banded	hub	small	banded	hub	small	banded	hub	small	banded	hub	small
		Empirical FDR (SE) (in %), autoregressive temporal structure											
50	oracle	0.9 (2.2)	0.8 (4.3)	0.5 (1.3)	0.8 (2.3)	1.0 (4.9)	0.9 (1.7)	0.8 (1.1)	0.9 (2.2)	0.7 (0.6)	0.8 (1.0)	1.0 (2.6)	0.7 (0.6)
	data-driven-S	0.7 (2.2)	0.5 (3.5)	0.6 (1.6)	0.7 (2.1)	1.2 (5.4)	0.8 (1.5)	0.8 (1.1)	0.9 (2.2)	0.6 (0.6)	0.7 (0.9)	0.3 (1.3)	0.6 (0.6)
	data-driven-B	0.8 (2.2)	0.8 (4.3)	0.6 (1.6)	0.9 (2.4)	0.8 (4.2)	1.0 (1.8)	0.9 (1.1)	0.9 (2.2)	0.7 (0.6)	0.8 (1.0)	1.0 (2.6)	0.6 (0.6)
	whitening	0.8 (2.3)	0.8 (4.3)	0.8 (2.0)	1.2 (3.0)	1.0 (4.9)	1.1 (2.0)	0.9 (1.2)	1.1 (2.3)	1.0 (0.9)	1.2 (1.2)	1.4 (3.2)	0.9 (0.8)
200	oracle	0.6 (1.1)	0.6 (2.0)	0.6 (0.7)	0.8 (1.0)	1.0 (2.2)	0.7 (0.8)	0.7 (0.5)	0.8 (0.9)	0.8 (0.4)	0.8 (0.4)	1.0 (1.0)	0.8 (0.3)
	data-driven-S	0.7 (1.1)	0.8 (2.2)	0.6 (0.7)	0.9 (1.1)	0.8 (2.0)	0.8 (0.8)	0.7 (0.5)	0.8 (0.9)	0.8 (0.4)	0.8 (0.4)	0.8 (0.8)	0.8 (0.3)
	data-driven-B	0.6 (1.1)	0.5 (1.7)	0.6 (0.7)	0.8 (1.0)	1.0 (2.2)	0.8 (0.8)	0.7 (0.5)	0.8 (0.9)	0.8 (0.4)	0.8 (0.4)	1.0 (1.0)	0.8 (0.3)
	whitening	0.8 (1.2)	0.9 (2.2)	0.8 (0.9)	1.1 (1.1)	1.1 (2.2)	0.9 (0.8)	1.0 (0.5)	0.9 (1.1)	1.0 (0.5)	1.1 (0.5)	1.3 (1.0)	1.1 (0.4)
800	oracle	0.5 (0.4)	1.0 (1.4)	0.6 (0.3)	0.8 (0.5)	0.8 (0.9)	0.7 (0.4)	0.8 (0.2)	1.0 (0.7)	0.7 (0.2)	0.8 (0.2)	0.9 (0.4)	0.8 (0.2)
	data-driven-S	0.5 (0.4)	1.0 (1.4)	0.6 (0.3)	0.8 (0.5)	0.7 (0.8)	0.7 (0.4)	0.8 (0.2)	1.0 (0.7)	0.7 (0.2)	0.8 (0.2)	0.8 (0.4)	0.8 (0.2)
	data-driven-B	0.5 (0.4)	1.0 (1.4)	0.6 (0.3)	0.8 (0.5)	0.8 (0.9)	0.7 (0.4)	0.8 (0.2)	1.0 (0.7)	0.7 (0.2)	0.8 (0.2)	0.9 (0.4)	0.8 (0.2)
	whitening	0.8 (0.5)	1.4 (1.7)	0.7 (0.4)	0.9 (0.6)	1.1 (1.1)	0.9 (0.5)	1.0 (0.3)	1.3 (0.9)	0.9 (0.2)	1.0 (0.3)	1.2 (0.5)	1.0 (0.2)
		Empirical power (SE) (in %), moving average temporal structure											
50	oracle	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	data-driven-S	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	100.0
	data-driven-B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	whitening	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	99.7	100.0	100.0	100.0
200	oracle	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.4	100.0	100.0	100.0
	data-driven-S	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.4	100.0	100.0	100.0
	data-driven-B	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.4	100.0	100.0	100.0
	whitening	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	98.8	100.0	100.0	100.0
800	oracle	100.0	58.2	99.4	100.0	100.0	100.0	100.0	44.1	98.0	100.0	100.0	100.0
	data-driven-S	100.0	57.8	99.4	100.0	100.0	100.0	100.0	43.6	97.9	100.0	100.0	100.0
	data-driven-B	100.0	58.5	99.4	100.0	100.0	100.0	100.0	43.5	98.0	100.0	100.0	100.0
	whitening	100.0	51.8	98.9	100.0	100.0	100.0	100.0	39.3	96.5	100.0	100.0	100.0

		$n = 15$		$n = 50$	
p		$q = 50$	$q = 200$	$q = 50$	$q = 200$
50	oracle	7	29	27	89
	data-driven-S	7	45	29	102
	data-driven-B	7	44	30	119
	whitening	41	1319	68	4295
200	oracle	96	522	472	2033
	data-driven-S	99	600	532	2425
	data-driven-B	101	553	500	2370
	whitening	219	6310	691	18320
800	oracle	2051	7585	6300	25211
	data-driven-S	2063	7601	6658	26630
	data-driven-B	2196	7654	6756	25634
	whitening	2729	30413	7184	91440

Table 5: Computation time in seconds (rounded up to integers) for a single data replication. The spatial structure is fixed as a banded graph, and the temporal structure is fixed as an autoregressive covariance.

6 Data Analysis

Attention deficit hyperactivity disorder (ADHD) is one of the most commonly diagnosed child-onset neurodevelopmental disorders. It has an estimated childhood prevalence of $5 - 10\%$ worldwide, and an estimated annual cost in tens of billions of dollars (Pelham et al. (2007)). Symptoms of ADHD include difficulty in staying focused and paying attention, difficulty in controlling behavior, and over-activity. These symptoms can persist into adolescence and adulthood, resulting in a lifelong impairment (Biederman et al. (2000)). The understanding and diagnosis of ADHD are of great significance. We analyzed a dataset from the ADHD-200 Global Competition, which includes demographical information and resting-state fMRI of nearly one thousand children and adolescents, including combined types of ADHD and typically developing control (TDC). The data were collected from eight participating sites. To avoid potential site bias, we adopted the strategy of Ahn et al. (2015), and focused our analysis on the fMRI data from the New York University site only, which has the

largest number of subjects among all sites. A Siemens Allegra 3T scanner was used to acquire the 6-min resting-state fMRI scans. The scan parameters are: voxel size $= 3 \times 3 \times 4$mm, slice thickness $= 4$mm, number of slices $= 33$, repetition time $= 2$s, echo time $= 15$ms, flip angle $= 90^{\circ}$, and field of view $= 240$mm. During acquisition, each subject was asked to lie still, stay awake, and not to think about anything under a black screen. We excluded some subjects from further analysis, based on the criterion that, for each subject, one or two fMRI scans were acquired and, for each scan, a quality control assessment (pass or questionable) was given by the data curators. This information was provided in the phenotypic data. We only used the scan that passed the quality control; if both scans of a subject passed the quality control, we chose the first scan. If neither scan passes the quality control, we removed that subject from further analysis. We also removed the subjects with missing diagnostic status or missing scans. The resulting dataset consisted of 96 combined ADHD subjects and 91 TDC subjects. All fMRI scans were preprocessed using the Athena pipeline, including slice timing correction, motion correction, spatial smoothing, denoising by regressing out motion parameters and white matter and cerebrospinal fluid time courses. Each voxel time course was also band-pass filtered ($0.009 - 0.08$ Hz) to remove frequencies not related to resting-state brain activity. All the fMRI data were aligned in the MNI T1 template space, with the same spatial dimensions $49 \times 58 \times 47$. Then the brain was parcellated using the Anatomical Automatic Labeling (AAL) atlas (Tzourio-Mazoyer et al. (2002)). The resulting data is a spatial by temporal matrix for each subject, with the spatial dimension $p = 116$ and the temporal dimension $q = 172$. More information about this data competition can be found at http://fcon_1000.projects.nitrc.org/indi/adhd200/. The preprocessed version of the data can be found at http://neurobureau.projects.nitrc.org/ADHD200/Data.html.

We first applied the test of Aston et al. (2016) to check if the data conforms with the matrix normal distribution with a Kronecker product structure. The p-values of the test for the

ADHD group and the TDC group were 0.0059 and 0.0068, respectively. Considering that a very small significance level is typically used in analysis, these p-values suggest that, for this data, the deviation from the separable structure seems moderate. Nevertheless, we caution that any interpretation of our data analysis should be taken with a healthy skepticism even under this relatively mild violation of the model assumption. We then applied our proposed multiple testing procedure. Given that the banded estimator of $\mathbf{\Sigma}_{T_g}$ performed best in the simulations, we employed it for the data analysis as well, and the selected bandwidth for both groups was equal to 3. For ease of presentation, we report in Table 6 the top 35 links found to differentiate between the ADHD and TDC groups, and their associated p-values, which were all smaller than 1e-12. Figure 1 shows those top links and the associated brain regions visualized with the BrainNet Viewer (Xia et al. (2013)).

It is seen that the differentiating links between the two groups concentrate on the frontal gyrus, cingulate gyrus, cerebellum and cerebellar vermis, precentral gyrus, postcentral gyrus, and right insula areas. The prefrontal cortex is responsible for many higher-order mental functions, including those that regulate attention and behavior. It is commonly thought that ADHD is associated with alterations in the prefrontal cortex (Arnsten and Li (2005)). The cingulate gyrus is associated with cognitive process, and there are evidences of anterior cingulate dysfunctions in ADHD patients (Bush et al. (2005)). The cerebellum is responsible for motor control and cognitive functions such as attention and language, and dysfunction in the cerebellum and anomaly in the cerebellar vermis in ADHD patients have been reported (Toplak et al. (2006); Goetz et al. (2014)). The precentral gyrus is the site of the primary motor cortex, which is involved in the planning, control, and execution of voluntary movements. The postcentral gyrus is the location of the primary somatosensory cortex. Their possible involvement with ADHD has been noted previously (Fassbender et al. (2011)). The insula is involved in consciousness and plays a role in functions linked to emotion, perception, motor control and

self-awareness. Its dysfunction in ADHD has also been reported (Spinelli et al. (2011)). Our findings are in general consistent with the current clinical literature of ADHD.

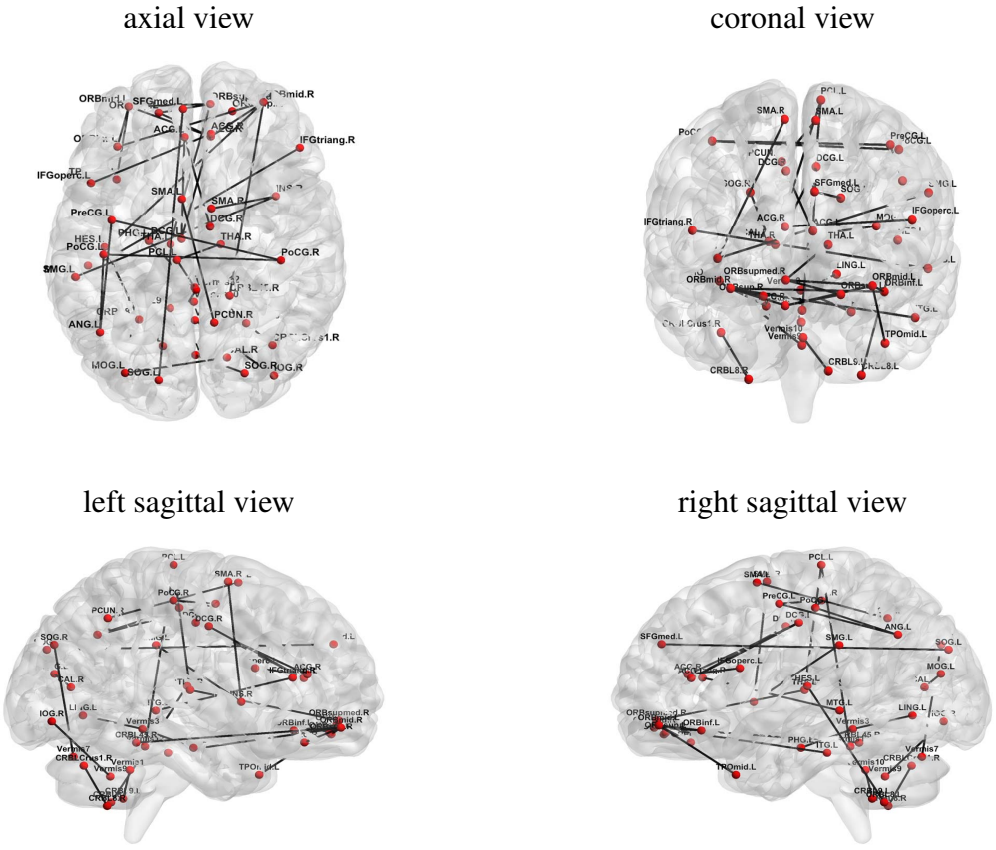
Table 6: Differentiating links and their associated p-values found by the proposed multiple testing procedure for the ADHD resting-state fMRI data. Reported are the links whose corresponding p-values smaller than 1e-12.

Differentiating links	p-value	Differentiating links	p-value
Frontal_Sup_Orb_L ↔ Frontal_Med_Orb_R	0	Frontal_Mid_L ↔ Temporal_Pole_Mid_L	2.54e-14
Frontal_Sup_Orb_R ↔ Frontal_Mid_Orb_R	0	Precentral_L ↔ Angular_L	2.75e-14
Frontal_Mid_Orb_L ↔ Frontal_Inf_Orb_L	0	Calcarine_R ↔ Occipital_Mid_L	3.10e-14
Frontal_Mid_Orb_L ↔ Frontal_Med_Orb_R	0	Frontal_Mid_Orb_R ↔ Frontal_Inf_Orb_L	5.00e-14
Frontal_Mid_Orb_L ↔ Rectus_R	0	Supp_Motor_Area_R ↔ Insula_R	1.22e-13
Frontal_Mid_Orb_R ↔ Cerebelum_4_5_R	0	Precentral_L ↔ Postcentral_R	1.40e-13
Frontal_Inf_Tri_R ↔ Temporal_Mid_L	0	Postcentral_L ↔ Angular_L	1.54e-13
Cingulum_Ant_L ↔ Cingulum_Mid_L	0	Occipital_Sup_R ↔ Vermis_7	1.88e-13
Cingulum_Ant_R ↔ Cingulum_Mid_L	0	Cerebelum_Crus2_R ↔ Cerebelum_8_R	2.11e-13
Postcentral_L ↔ Postcentral_R	0	Insula_R ↔ SupraMarginal_L	2.90e-13
Occipital_Inf_R ↔ Vermis_9	1.11e-16	Frontal_Inf_Oper_L ↔ Cingulum_Ant_R	3.09e-13
Cerebelum_9_L ↔ Vermis_10	2.22e-16	Paracentral_Lobule_L ↔ Thalamus_R	3.18e-13
Frontal_Mid_Orb_R ↔ Rectus_R	8.88e-16	Lingual_L ↔ Vermis_3	4.44e-13
Frontal_Mid_Orb_R ↔ Temporal_Inf_L	1.22e-15	Frontal_Med_Orb_R ↔ Thalamus_L	6.19e-13
Frontal_Sup_Orb_L ↔ Rectus_R	1.89e-15	Heschl_L ↔ Cerebelum_8_L	6.87e-13
Frontal_Sup_Medial_L ↔ Occipital_Sup_L	2.66e-15	Cingulum_Ant_L ↔ Cingulum_Mid_R	7.88e-13
Paracentral_Lobule_L ↔ Vermis_1_2	7.88e-15	Supp_Motor_Area_L ↔ Precuneus_R	9.65e-13
ParaHippocampal_L ↔ Vermis_3	2.10e-14		

7 Discussion and Extension

Motivated by applications in neuroscience research, we have proposed in this article a multiple hypothesis testing procedure for detecting the alternations of brain connectivities between two groups. Empirically it is demonstrated to enjoy a competitive performance, and it can handle both a small sample size ($n = 15$) as well as an adequately large network ($p = 800$). Theoretically it is shown to control the false discovery asymptotically. Moreover, since no

Figure 1: Differentiating links and the associated brain regions found by the proposed multiple testing procedure for the ADHD resting-state fMRI data. Reported are the top 35 links.



bootstrap or data permutation is required, the computation of our testing procedure is fast.

We have primarily focused on the two-sample testing scenario. In principle, our approach can be extended to multiple groups testing as well. Specifically, suppose we have $\{\boldsymbol{X}^{(g)}, g = 1, \ldots, K\}$ to denote the $p \times q$ spatial temporal matrices from K groups, $K \geq 2$. Each follows a matrix normal distribution, with the Kronecker product covariance structure, $\boldsymbol{\Sigma}\{\text{vec}(\boldsymbol{X}^{(g)})\} = \boldsymbol{\Sigma}_{S_g} \otimes \boldsymbol{\Sigma}_{T_g}, g = 1, \ldots, K$. Accordingly, $\boldsymbol{\Sigma}^{-1}\{\text{vec}(\boldsymbol{X}^{(g)})\} = \boldsymbol{\Sigma}_{S_g}^{-1} \otimes \boldsymbol{\Sigma}_{T_g}^{-1} = \boldsymbol{\Omega}_{S_g} \otimes \boldsymbol{\Omega}_{T_g}, g = 1, \ldots, K$. Let $\boldsymbol{R}_{S_g} = \boldsymbol{D}_{S_g}^{-1/2} \boldsymbol{\Omega}_{S_g} \boldsymbol{D}_{S_g}^{-1/2} = \left(\rho_{S_g,i,j}\right)_{i,j=1}^{p}$, we aim to simultaneously test,

$$H_{0,i,j} : \rho_{S_1,i,j} = \rho_{S_2,i,j} = \cdots = \rho_{S_K,i,j} \quad \text{versus} \quad H_{1,i,j} : \rho_{S_l,i,j} \neq \rho_{S_k,i,j}, \ 1 \leq l \neq k \leq K,$$

for $1 \leq i < j \leq p$. For each pair of groups, we define the standardized test statistic, $W_{i,j}^{(l,k)} =$

$(\hat{\rho}_{i,j,l} - \hat{\rho}_{i,j,k}) / \sqrt{\hat{\theta}_{i,j,l} + \hat{\theta}_{i,j,k}}, 1 \leq i < j \leq p$, where $\hat{\rho}_{i,j,l}$ and $\hat{\theta}_{i,j,l}$ for the l-th group can be obtained as in (3) and (4). Then we construct the sum-of-square type test statistic as $S_{i,j} = \sum_{1 \leq l < k \leq K} (W_{i,j}^{(l,k)})^2$. It can be shown that the limiting null distribution of $S_{i,j}$ is a mixture chi-square distribution. In order to develop a multiple testing procedure based on $S_{i,j}$, two dependence structures need to be taken into consideration. One is the dependence among different entries (i, j) for a given pair, namely, $\{W_{i,j}^{(l,k)}, 1 \leq i < j \leq p\}$, as studied in Section 3. The other is the dependence between different pairs (l, k) of the standardized statistics, $\{W_{i,j}^{(l,k)} 1 \leq l < k \leq K\}$. We leave this as our future research.

8 Supplementary Material

The proofs of the main theorems and some technical lemmas are available in the online supplementary material.

Acknowledgement

The research of Yin Xia was supported in part by NSFC Grants 11771094, 11690013, "The Recruitment Program of Global Experts" Youth Project from China, and the startup fund from Fudan University. The research of Lexin Li was supported in part by NSF Grants DMS-1310319 and DMS-1613137.

References

Ahn, M., Shen, H., Lin, W., and Zhu, H. (2015). A sparse reduced rank framework for group analysis of functional neuroimaging data. *Statistica Sinica*, 25:295–312.

Anderson, T. W. (2003). *An Introduction To Multivariate Statistical Analysis*. Wiley-Intersceince, 3rd ed, New York.

Arnsten, A. F. and Li, B.-M. (2005). Neurobiology of executive functions: Catecholamine influences on prefrontal cortical functions. *Biological Psychiatry*, 57:1377–1384.

Aston, J. A., Pigoli, D., and Tavakoli, S. (2016). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, accepted.

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

Biederman, J., Mick, E., and Faraone, S. V. (2000). Age-dependent decline of symptoms of attention deficit hyperactivity disorder: Impact of remission definition and symptom type. *American Journal of Psychiatry*, 157:816–818.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10:186–198.

Bush, G., Valera, E. M., and Seidman, L. J. (2005). Functional neuroimaging of attention-deficit/hyperactivity disorder: A review and suggested future directions. *Biological Psychiatry*, 57:1273–1284.

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 ℓ_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, S., Kang, J., Xing, Y., and Wang, G. (2015). A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. *Human Brain Mapping*, 36:5196–5206.

Chen, X. and Liu, W. (2015). Statistical inference for matrix-variate gaussian graphical models and false discovery rate control. *arXiv preprint arXiv:1509.05453*.

Fassbender, C., Schweitzer, J. B., Cortes, C. R., Tagamets, M. A., Windsor, T. A., Reeves, G. M., and Gullapalli, R. (2011). Working memory in attention deficit/hyperactivity disorder is characterized by a lack of specialization of brain function. *PLoS ONE*, 6:e27240.

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 Greicius, M. (2010). Clinical applications of resting state functional connectivity. *Frontiers in Systems Neuroscience*, 4.

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., editors (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.

Goetz, M., Vesela, M., and Ptacek, R. (2014). Notes on the role of the cerebellum in adhd. *Austin J Psychiatry Behav Sci*, 1:1013.

Han, F., Han, X., Liu, H., and Caffo, B. (2016). Sparse median graphs estimation in a high-dimensional semiparametric model. *The Annals of Applied Statistics*, 10:1397–1426.

Hedden, T., Van Dijk, K. R. A., Becker, J. A., Mehta, A., Sperling, R. A., Johnson, K. A., and Buckner, R. L. (2009). Disruption of functional connectivity in clinically normal older adults harboring amyloid burden. *The Journal of Neuroscience*, 29:12686–12694.

Kang, J., Bowman, F. D., Mayberg, H., and Liu, H. (2016). A depression network of functionally connected regions discovered via multi-attribute canonical correlation graphs. *NeuroImage*, 141:431–441.

Kim, J., Wozniak, J. R., Mueller, B. A., Shen, X., and Pan, W. (2014). Comparison of statistical tests for group differences in brain functional networks. *NeuroImage*, 101:681–694.

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

Lindquist, M. (2008). The statistical analysis of fmri data. *Statistical Science*, 23:439–464.

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*, pages 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*.

Pelham, W. E., Foster, E. M., and Robb, J. A. (2007). The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. *Ambulatory Pediatrics*, 7:121–131. Measuring Outcomes in Attention Deficit Hyperactivity Disorder.

Qiu, H., Han, F., Liu, H., and Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of Royal Statistical Society, Series B.*, 78:487–504.

Rudie, J., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P., Bookheimer, S., and Dapretto, M. (2013). Altered functional and structural brain network organization in autism. *NeuroImage*, 2:79–94.

Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012). Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59:3852–3861.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural {MR} image analysis and implementation as {FSL}. *NeuroImage*, 23, Supplement 1:S208–S219. Mathematics in Brain Imaging.

Spinelli, S., Vasa, R. A., Joel, S., Nelson, T. E., Pekar, J. J., and Mostofsky, S. H. (2011). Variability in post-error behavioral adjustment is associated with functional abnormalities in the temporal cortex in children with adhd. *Journal of Child Psychology and Psychiatry*, 52:808–816.

Tomasi, D. and Volkow, N. D. (2012). Abnormal functional connectivity in children with attention-deficit/hyperactivity disorder. *Biological Psychiatry*, 71:443–450.

Toplak, M. E., Dockstader, C., and Tannock, R. (2006). Temporal information processing in adhd: Findings to date and new methods. *Journal of Neuroscience Methods*, 151:15–29. Towards a Neuroscience of Attention-Deficit/Hyperactivity Disorder (ADHD).

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in {SPM} using a macroscopic anatomical parcellation of the {MNI} {MRI} single-subject brain. *NeuroImage*, 15:273–289.

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

Varoquaux, G. and Craddock, R. C. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405–415. Mapping the Connectome.

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. (2016). An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Frontiers in Neuroscience*, 10:1–17.

Xia, M., Wang, J., and He, Y. (2013). Brainnet viewer: A network visualization tool for human brain connectomics. *PLOS ONE*, 8:1–15.

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.

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

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.

Zhang, T., Wu, J., Li, F., Caffo, B., and Boatman-Reich, D. (2015). A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *Journal of the American Statistical Association*, 110:93–106.

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*, 13:1059–1062.

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

Department of Statistics, School of Management, Fudan University, Shanghai, 200433, P.R. China.

E-mail: (xiayin@fudan.edu.cn)

Division of Biostatistics, School of Public Health, University of California at Berkeley, Berkeley, CA 94720, U.S.A.

E-mail: (lexinli@berkeley.edu)