

**Statistica Sinica Preprint No: SS-2019-0361**

<b>Title</b>	Hypothesis Testing for Network Data with Power Enhancement
<b>Manuscript ID</b>	SS-2019-0361
<b>URL</b>	<a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a>
<b>DOI</b>	10.5705/ss.202019.0361
<b>Complete List of Authors</b>	Yin Xia and Lexin Li
<b>Corresponding Author</b>	Lexin Li
<b>E-mail</b>	lexinli@berkeley.edu
Notice: Accepted version subject to English editing.	

# Hypothesis Testing for Network Data with Power Enhancement

Yin Xia and Lexin Li

*Fudan University and University of California at Berkeley*

*Abstract:* Comparing two population means of network data is of paramount importance in a wide range of scientific applications. Numerous existing network inference solutions focus on global testing of entire networks, without comparing individual network links. The observed data often take the form of vectors or matrices, and the problem is formulated as comparing two covariance or precision matrices under a normal or matrix normal distribution. Moreover, many tests suffer from a limited power under a small sample size. In this article, we tackle the problem of network comparison, both global and simultaneous inferences, when the data come in a different format, i.e., in the form of a collection of symmetric matrices, each of which encodes the network structure of an individual subject. Such data format commonly arises in applications such as brain connectivity analysis and clinical genomics. We no longer require the underlying data to follow a normal distribution, but instead impose some moment conditions that are easily satisfied for numerous types of network data. Furthermore, we propose a power enhancement procedure, and show that it can control the false discovery, while it has the potential to substantially enhance the power of the test. We investigate the efficacy of our testing procedure through both an asymptotic analysis and a simulation study under a finite sample size. We further illustrate our method with examples of brain connectivity analysis.

*Key words and phrases:* Auxiliary information; False discovery rate; Multiple testing; Network data; Power enhancement.

## 1 Introduction

With recent prevalence of network data, the problem of comparing two populations of networks is gaining increasing attention. Our motivation is brain connectivity analysis, which studies functional and structural brain architectures through neurophysiological measures of brain activities and synchronizations (Fornito et al., 2013). Accumulated evidences have suggested that, compared to a healthy brain, the brain connectivity network alters in the presence of numerous neurological disorders, for example, Alzheimer’s disease, autism spectrum disorder, among many others. Such alternations are believed to hold crucial insights of disease pathologies (Fox and Greicius, 2010). A typical brain connectivity study collects imaging scans, such as functional magnetic resonance imaging, or diffusion tensor imaging, from groups of subjects with and without disorder. Based on the imaging scan, a network is constructed for each individual subject, with the nodes corresponding to a common set of brain regions, and the edges encoding the functional or structural associations between the regions. A fundamental scientific question of interest is to compare the brain networks and to identify local connectivity patterns that alter between the two populations. Network comparison is equally interesting in many other scientific areas as well, for instance, clinical genomics, where of crucial interest is to understand and compare gene regulatory networks of patients with and without cancer (Luscombe et al., 2004).

In the context of brain connectivity analysis, there has been a rich literature on network *estimation* methods (Ahn et al., 2015; Qiu et al., 2016; Wang et al., 2016; Zhu and Li, 2018, among many others). Recently, Zou et al. (2017) and Lan et al. (2018) studied estimation of the covariance matrix of a multivariate vector as a function of the similarity measure of the covariates, or a function of the adjacency matrix. There is, however, a relative paucity of *inference* methods,

## 1. INTRODUCTION

---

especially simultaneous inference for individual links. Even though both can produce, in effect, a concise representation of the network structure, network inference is a fundamentally different problem than network estimation. Among the few existing network inference solutions, Kim et al. (2014) studied a number of two-sample tests based on network summary metrics or generalized linear models. However, they only compared two networks globally, without any inference on the individual links of the networks. Besides, some of their tests resorted to bootstrap or permutation, which is computationally intensive and slow. Ginestet et al. (2017) characterized the geometry of the space of undirected networks with edge weights, and developed an analog of the classical two-sample test for network empirical means. However, they again focused on the global test of two entire networks. Chen et al. (2015) developed a method to detect differentially expressed connectivity subnetworks under different clinical conditions. They resorted to a permutation test, and controlled the family-wise error rate. Xia et al. (2015) first encoded the connectivity network by a partial correlation matrix computed from vector-valued data under a normal distribution. They then proposed a multiple testing procedure to compare the partial correlation matrices from the two populations, along with a proper false discovery control. Xia and Li (2019) further extended the test to matrix-valued data under a matrix normal distribution. In both cases, the test statistics were constructed based on the vector or matrix-valued data, which, as we explain next, may not be directly observable. Moreover, the underlying data distribution may not always be normal or matrix normal. Durante and Dunson (2018) developed a fully Bayesian solution for network comparison, which is very flexible and can handle the data format of our problem, but it requires specification of a series of prior distributions and can be computationally intensive.

Applications such as brain connectivity analysis actually raise new challenges for network

## 1. INTRODUCTION

---

inference. First, the observed data come in the form of  $p \times p$  matrices, where  $p$  is the number of network nodes. Each such matrix encodes the network structure for one individual subject, and a collection of network samples are observed. For instance, in brain structural connectivity, what one observes are the numbers of white matter fibers between pairs of brain anatomical regions. This matrix of counts forms a network observation for one subject, with brain regions constituting the nodes and the fiber counts the links, and we observe multiple such count-valued networks for multiple subjects. This is ultimately different from the data format studied in most existing network methods, where a network structure usually takes the form a covariance or precision matrix of some vector-valued or matrix-valued data. This fundamental difference in terms of the available data format would thus require a completely new problem formulation and inferential procedure. Second, in a multitude of applications including brain connectivity analysis, the sample size is usually very small, e.g., in tens. This calls for a testing procedure that is powerful enough to detect differentially expressed links under a limited sample size. In this article, we address the problem of comparing two populations of network data, more precisely, the two population means of networks. We aim to consider both global and simultaneous inferences, tackle the new data format, and explicitly enhance the power of the test.

Specifically, suppose we observe two groups of samples,  $\{\mathbf{S}_{1,1}, \dots, \mathbf{S}_{1,n_1}\}$  and  $\{\mathbf{S}_{2,1}, \dots, \mathbf{S}_{2,n_2}\}$ , where  $\mathbf{S}_{d,l}$  denotes the observed symmetric  $p \times p$  network data for the  $l$ th sample in the  $d$ th group,  $n_d$  is the total number of network samples in the  $d$ th group,  $l = 1, \dots, n_d$ , and  $d = 1, 2$ . Suppose  $\mathbf{S}_{d,l} = (S_{d,l,i,j})_{p \times p} \sim \mathcal{F}_d(\mathbf{s}_d)$ , where  $\mathcal{F}_d$  is some distribution with a symmetric mean matrix  $\mathbf{s}_d = (s_{d,i,j})_{p \times p}$ . Our goal is to test whether the two population means are the same:

$$H_0 : \mathbf{s}_1 = \mathbf{s}_2 \quad \text{versus} \quad H_1 : \mathbf{s}_1 \neq \mathbf{s}_2. \quad (1)$$

If the global null in (1) is rejected, we further aim to identify at which locations the two mean matrices are different. That is, we wish to simultaneously test:

$$H_{0,i,j} : s_{1,i,j} = s_{2,i,j} \quad \text{versus} \quad H_{1,i,j} : s_{1,i,j} \neq s_{2,i,j}, \quad \text{for } 1 \leq i < j \leq p. \quad (2)$$

In Xia et al. (2015), the observed data  $\mathbf{X}_{d,l} \in \mathbb{R}^p$  represents expressions of multiple genes for two groups of patients with long and short term survival. It is a vector, and is assumed to follow a normal distribution with the covariance matrix  $\Sigma_d$ . Let  $\mathbf{R}_d$  denote the corresponding partial correlation matrix, i.e., the standardized version of  $\Sigma_d^{-1}$ ,  $d = 1, 2$ . Then the network structure is encoded by  $\mathbf{R}_d$ , and the problem becomes testing if  $\mathbf{R}_1 = \mathbf{R}_2$ . Xia and Li (2019) followed a similar setup, except that the observed data  $\mathbf{X}_{d,l} \in \mathbb{R}^{p \times t}$  becomes a matrix, which represents brain temporal neural activity measures collected at multiple brain locations for two groups of patients with and without attention deficit hyperactivity disorder. It is assumed to follow a matrix normal distribution with the covariance  $\Sigma_d \otimes \Lambda_d$ , and the network is still encoded by the standardized version of  $\Sigma_d^{-1}$ . The key difference for our setting is that, we do *not* always observe  $\mathbf{X}_{d,l}$  directly, but instead  $\mathbf{S}_{d,l}$  *only*. This difference in data format completely distinguishes our method from nearly all existing solutions such as Xia et al. (2015) and Xia and Li (2019). Moreover, we do not impose that the underlying data follows a normal or matrix normal distribution. Instead, we consider a general class of distributions for  $\mathcal{F}_d$  satisfying some moment condition. Our method works for many different types of network links, for instance, binary links when  $\mathcal{F}_d$  follows a light tailed distribution, or count links when  $\mathcal{F}_d$  follows a heavy-tailed distribution.

For the global test (1), we develop a global test statistic taken as the maximum of a set of individual test statistics. We then derive its limiting null distribution, and show the resulting global test is power minimax optimal asymptotically. For the simultaneous test (2), we first develop a

## 1. INTRODUCTION

---

multiple testing procedure, and show that it can asymptotically control the false discovery at the pre-specified level. Next we propose a method to substantially enhance the power of the simultaneous inference procedure for (2). Specifically, we extend the grouping-adjusting-pooling idea of Xia et al. (2019a), and modify it for our inference of network data.

Our proposal differs from the existing solutions and makes several useful contributions. First, to the best of our knowledge, there has been no solution directly targeting simultaneous hypothesis testing of individual links for the network data in the format of  $S_{d,l}$ . Our method bridges this gap, and offers a timely solution to a range of scientific applications where this form of problem and data is commonly encountered. Second, our global test statistic is constructed as the maximum of the individual test statistics for all links. This type of maximum statistic enjoys various advantages and has been commonly employed in the hypothesis testing literature (e.g., Cai et al., 2013; Xia et al., 2019b). However, the derivation of its asymptotics, as well as the properties of the subsequent multiple testing procedure, are far from trivial in our new context of network comparison. Moreover, we remark that, in some network data applications, the individual test statistics may be correlated, and a global test statistic that utilizes such correlations may result in a more powerful test. However, this may not always be the case. For instance, in our brain connectivity application, the nodes are usually the brain anatomical regions, which can scatter at distant locations of the brain. As a result, there is no obvious correlation structure for the individual test statistics built on the pairs of brain regions. Therefore, we do not explicitly impose or employ any correlation structure when constructing the global test statistic. On the other hand, in our power enhancement procedure, we implicitly utilize the fact that some individual test statistics may be correlated and clustered. We then use a data driven approach to find such clusters and

## 2. MOMENT CONDITIONS AND EXAMPLES

---

incorporate this information in our test. Finally, the power enhancement approach we develop is particularly useful in numerous applications, e.g., brain connectivity analysis, where the sample size is limited. Although motivated by Xia et al. (2019a), our enhancement method differs from Xia et al. (2019a) considerably in several ways. We explicitly compare the two power enhancement procedures in Section 4.5. Overall, we feel our method provides a useful addition to the general toolbox of network inference.

We adopt the following notation throughout this article. For a symmetric matrix  $\mathbf{A}_d$ , let  $\lambda_{\max}(\mathbf{A}_d)$  and  $\lambda_{\min}(\mathbf{A}_d)$  denote the largest and smallest eigenvalues of  $\mathbf{A}_d$ , respectively. For a set  $\mathcal{H}$ , let  $|\mathcal{H}|$  denote its cardinality. For two sequences of real numbers  $\{a_n\}$  and  $\{b_n\}$ , write  $a_n = O(b_n)$  if there exists a constant  $C$  such that  $|a_n| \leq C|b_n|$  holds for all  $n$ , write  $a_n = o(b_n)$  if  $\lim_{n \rightarrow \infty} a_n/b_n = 0$ , and write  $a_n \asymp b_n$  if there are positive constants  $c$  and  $C$  such that  $c \leq a_n/b_n \leq C$  for all  $n$ . Write  $n = n_1 n_2 / (n_1 + n_2)$  and assume that  $n_1 \asymp n_2$ .

The rest of the article is organized as follows. Section 2 presents the moment conditions for the distribution of  $\mathcal{F}_d$  and show they are easily satisfied in numerous types of network data. Section 3 develops the global testing and the simultaneous testing for the two-sample network comparison, and Section 4 studies power enhancement, both of which are key to our proposal. Section 5 presents the simulations, and Section 6 presents two brain connectivity analysis examples as illustration. The Supplementary Material collects additional lemmas and the proofs.

## 2 Moment Conditions and Examples

We begin with some moment conditions imposed on  $\mathcal{F}_d$ . We then give a number of examples and show that those conditions are easily satisfied in numerous types of network data.



## 2.1 Moment conditions

We assume that the distribution  $\mathcal{F}_d$  of the network data  $\mathcal{S}_{d,l}$  satisfies one of the following two conditions: a sub-Gaussian-type tail, or a polynomial-type tail, as stated below.

(C1) (Sub-Gaussian-tail). Suppose that  $\log p = o(n^{1/5})$ , and that there exist some constants

$\eta > 0$  and  $K > 0$ , such that, for  $d = 1, 2$ ,

$$\mathbb{E} \left[ \exp \left\{ \eta (S_{d,l,i,j} - s_{d,i,j})^2 / \text{Var}(S_{d,l,i,j}) \right\} \right] \leq K, \quad \text{for } 1 \leq i < j \leq p, l = 1, \dots, n_d.$$

(C2) (Polynomial-tail). Suppose that  $p \leq cn^{\gamma_0}$  for some constants  $\gamma_0, c > 0$ , and that there exist some constants  $\epsilon > 0$  and  $K > 0$ , such that, for  $d = 1, 2$ ,

$$\mathbb{E} \left\{ \left| (S_{d,l,i,j} - s_{d,i,j}) / \text{Var}(S_{d,l,i,j})^{1/2} \right|^{4\gamma_0 + 2 + \epsilon} \right\} \leq K, \quad \text{for } 1 \leq i < j \leq p, l = 1, \dots, n_d.$$

We first comment that, both conditions are common, and similar conditions have been often assumed in the high-dimensional setting (Cai et al., 2014; Van de Geer et al., 2014). These moment conditions are much weaker than the Gaussian assumption as usually required in the testing literature (Schott, 2007). Next we discuss a number of network examples that satisfy the above moment conditions, including Bernoulli and mixture Bernoulli data, Poisson data, correlation and partial correlation data. Furthermore, we discuss some examples where the distributions are heavy-tailed, but after some data transformation, they still satisfy the moment conditions. Examples include transformed normal count data and transformed Wishart count data.

## 2.2 Network data examples

The first example is binary network, which is arguably the most commonly seen network data type, where each link is a binary indicator. The Bernoulli distribution is often assumed; i.e., for

## 2. MOMENT CONDITIONS AND EXAMPLES

$\mathbf{S}_{d,l} = (S_{d,l,i,j})_{p \times p}$ ,  $S_{d,l,i,j}$  follows a Bernoulli distribution with mean  $s_{d,i,j}$ ,  $u < s_{d,i,j} < 1 - u$  for a constant  $0 < u < 1$ ,  $l = 1, \dots, n_d$ ,  $d = 1, 2$ , and  $1 \leq i < j \leq p$ . In such case,  $\mathbf{S}_{d,l}$  satisfies the sub-Gaussian-tail condition in (C1), e.g., with  $\eta = 1$  and  $K = (1 - u) \exp\{u(1 - u)^{-1}\} + u \exp\{(1 - u)u^{-1}\}$ . The same holds true for the mixture Bernoulli distribution as discussed in Durrleman and Dunson (2018). That is, for some integer  $H > 0$  and randomly selected  $\{\phi_1, \dots, \phi_H\}$  subject to  $\sum_{h=1}^H \phi_h = 1$  and  $\phi_h > 0$ ,  $\mathbf{P}(S_{d,l,i,j} = x) = \sum_{h=1}^H \phi_h \left\{s_{d,i,j}^{(h)}\right\}^x \left\{1 - s_{d,i,j}^{(h)}\right\}^{1-x}$ , with  $u < s_{d,i,j}^{(h)} < 1 - u$  for some constant  $0 < u < 1$ ,  $x = 0, 1$ ,  $h = 1, \dots, H$ ,  $l = 1, \dots, n_d$ ,  $d = 1, 2$  and  $1 \leq i < j \leq p$ . For this example,  $\mathbf{S}_{d,l}$  again satisfies the sub-Gaussian-tail condition in (C1), with  $\eta = 1$  and  $K = (1 - u) \exp\{u(1 - u)^{-1}\} + u \exp\{(1 - u)u^{-1}\}$ .

The second example is correlation network, which is another equally common network data type. In brain functional connectivity analysis and many other applications, the network is often encoded by a Pearson correlation or a partial correlation matrix. Take the Pearson correlation network as an example. The functional imaging data is usually summarized as a spatial-temporal matrix. That is, for the  $l$ th subject in the  $d$ th group, the observed data is of the form  $\mathbf{X}_{d,l} \in \mathbb{R}^{p \times t_d}$ ,  $l = 1, \dots, n_d$ ,  $d = 1, 2$ , where  $p$  is the number of brain regions, and  $t_d$  is the number of repeated measures. Then the brain functional connectivity network is encoded by the sample correlation matrix  $\mathbf{S}_{d,l} = t_d^{-1} \sum_{j=1}^{t_d} \{\mathbf{X}_{d,l,(\cdot,j)} - \bar{\mathbf{X}}_{d,l}\} \{\mathbf{X}_{d,l,(\cdot,j)} - \bar{\mathbf{X}}_{d,l}\}^\top$ , where  $\mathbf{X}_{d,l,(\cdot,j)}$  denotes the  $j$ th column of the matrix  $\mathbf{X}_{d,l}$  and  $\bar{\mathbf{X}}_{d,l} = t_d^{-1} \sum_{j=1}^{t_d} \mathbf{X}_{d,l,(\cdot,j)}$  denotes the sample mean vector (Fornito et al., 2013). Next we show that, as long as  $\mathbf{X}_{d,l}$  satisfies one of the conditions in Lemma 1, then  $\mathbf{S}_{d,l}$  satisfies the sub-Gaussian-tail condition (C1).

**Lemma 1.** *Suppose  $\mathbf{X}_{d,l}$  satisfies one of the following conditions: (i)  $\log p = o(t_d^{1/5})$ , and there exist constants  $\eta' > 0$ ,  $K' > 0$  such that  $\mathbf{E}(\exp[\eta' \{X_{d,l,i,j} - \mathbf{E}(X_{d,l,i,j})\}^2 / \text{Var}(X_{d,l,i,j})]) \leq K'$ ,*

## 2. MOMENT CONDITIONS AND EXAMPLES

where  $t = \max\{t_1, t_2\}$  and  $t_1 \asymp t_2$ ; (ii)  $p \leq c't^{\gamma'_0}$ , for some  $\gamma'_0, c' > 0$ , and there exist constants  $\epsilon' > 0, K' > 0$  such that  $\mathbf{E} \left[ \left| \{X_{d,l,i,j} - \mathbf{E}(X_{d,l,i,j})\} / \text{Var}(X_{d,l,i,j})^{1/2} \right|^{4\gamma'_0+4+\epsilon'} \right] \leq K'$ , for  $i = 1, \dots, p, j = 1, \dots, t_d$ . Then  $\mathbf{S}_{d,l}$  satisfies the sub-Gaussian-tail condition in (C1), with  $\eta = 1/4$  and  $K = 2$ , as  $t \rightarrow \infty$ .

We remark that a similar result as Lemma 1 can be obtained for the partial correlation network, by using the inverse regression techniques as in Liu (2013). Xia and Li (2019) tackled network comparison assuming  $\mathbf{X}_{d,l}$  is directly observable and follows a matrix normal distribution. Lemma 1 suggests that, the test we develop later is still applicable when  $\mathbf{X}_{d,l}$  is available, even though it may not be as powerful as the test of Xia and Li (2019) in this case. On the other hand, the main focus of this article is to develop a test of comparing two networks even when  $\mathbf{X}_{d,l}$  is not observed, but only  $\mathbf{S}_{d,l}$  is. As such, our test is more general than that of Xia and Li (2019).

The third example is count network, another common network data type, where each link is a count. For instance, in brain structural connectivity analysis, the link is the number of white matter fibers between anatomical brain regions. The Poisson distribution is often imposed; i.e.,  $S_{d,l,i,j}$  follows a Poisson distribution with mean  $s_{d,i,j}$ ,  $0 < u_1 < s_{d,i,j} < u_2$ ,  $l = 1, \dots, n_d, d = 1, 2, 1 \leq i < j \leq p$ . For any constant  $\epsilon > 0$ , let  $M$  be the smallest integer that is no smaller than  $4\gamma_0 + 2 + \epsilon$ , where  $\gamma_0$  is as defined in (C2). Then  $\mathbf{S}_{d,l}$  satisfies the polynomial-tail condition (C2), with  $K$  upper bounded by  $u_1^{-(M-1)/2} \left[ \sum_{i=0}^M u_2^i \binom{M}{i} + u_2^M (u_2/2 + 1) \right]$ , and  $\binom{M}{i}$  is the number of ways to partition a set of  $M$  objects into  $i$  non-empty subsets.

We next consider some examples where the original network data  $\mathbf{G}_{d,l} = (G_{d,l,i,j})_{p \times p} \sim \tilde{\mathcal{F}}_d(\tilde{\mathbf{s}}_d)$ ,  $l = 1, \dots, n_d, d = 1, 2$ , and  $\tilde{\mathcal{F}}_d$  is some heavy-tailed distribution that only differs in the mean matrix  $\tilde{\mathbf{s}}_d = (\tilde{s}_{d,i,j}) \in \mathbb{R}^{p \times p}$  between the two groups. In such cases, testing the means

### 3. TWO-SAMPLE TEST ON NETWORK DATA

---

of the original samples are equivalent to testing the means of the transformed data,  $S_{d,l,i,j} = f(G_{d,l,i,j})$ , where  $f$  is some one-to-one transformation function. One example is the log-normal count network. After the logarithmic transformation of  $G_{d,l}$ , the transformed data  $S_{d,l}$  follows a normal distribution, and thus both (C1) and (C2) are satisfied. This can be further extended to the transformed normal mixture network. Another example is the transformed Wishart count network, where the transformed data  $S_{d,l}$  follows the Wishart distribution. For this case,  $S_{d,l}$  satisfies the sub-Gaussian-tail condition (C1). Moreover, in this case, the testing problems (1) and (2) are closely related to the covariance matrix testing problems studied in Li and Chen (2012) and Cai et al. (2013). The key difference between our method and the existing ones is that, we only observe  $S_{d,l}$ , but not the original vector samples. This example can be further extended to the case of the product of Gaussian mixtures network, or the Wishart mixtures network.

## 3 Two-sample Test on Network Data

We begin with the construction of a test statistic for the two testing problems (1) and (2). We then develop a global testing procedure for (1), and a simultaneous testing procedure for (2). For each test, we derive its corresponding asymptotic properties.

### 3.1 Test statistics

We first observe that the testing problem (1) is equivalent to the test,  $H'_0 : \max_{1 \leq i < j \leq p} |s_{1,i,j} - s_{2,i,j}| = 0$ . This motivates us to construct the test statistic based on

$$W_{i,j} = \bar{S}_{1,i,j} - \bar{S}_{2,i,j}.$$

### 3. TWO-SAMPLE TEST ON NETWORK DATA

where  $\bar{S}_{d,i,j} = n_d^{-1} \sum_{l=1}^{n_d} S_{d,l,i,j}$ . We standardize  $W_{i,j}$ , and estimate the variance of  $S_{d,l,i,j}$  by

$$V_{1,i,j} = n_1^{-1} \sum_{l=1}^{n_1} (S_{1,l,i,j} - \bar{S}_{1,i,j})^2, \quad \text{and} \quad V_{2,i,j} = n_2^{-1} \sum_{l=1}^{n_2} (S_{2,l,i,j} - \bar{S}_{2,i,j})^2, \quad (3)$$

respectively. This leads to our test statistic,

$$T_{i,j} = \frac{W_{i,j}}{(V_{1,i,j}/n_1 + V_{2,i,j}/n_2)^{1/2}}, \quad 1 \leq i < j \leq p. \quad (4)$$

## 3.2 Global test

In brain connectivity analysis and many other applications, it is generally postulated that the differences between two network structures concentrate on a small number of brain regions. This translates to a sparse alternative in our global test. Correspondingly, we construct the global test statistic as,

$$M_n = \max_{1 \leq i < j \leq p} T_{i,j}^2.$$

Let  $\Gamma_d \in \mathbb{R}^{q \times q}$  denote the covariance matrix of  $\text{vech}(\mathbf{S}_{d,l})$ , where  $q = p(p-1)/2$ , and  $\text{vech}(\cdot)$  is the operator that turns the upper triangular part of  $\mathbf{S}_{d,l}$  into a vector. Let  $\mathbf{R}_d = (r_{d,i,j}) \in \mathbb{R}^{q \times q}$  denote the corresponding correlation matrix. We introduce two conditions.

(A1)  $C_0^{-1} \leq \lambda_{\min}(\Gamma_d) \leq \lambda_{\max}(\Gamma_d) \leq C_0$  for some constant  $C_0 > 0$ ,  $d = 1, 2$ .

(A2)  $\max_{d=1,2} \max_{1 \leq i < j \leq q} |r_{d,i,j}| < r < 1$  for some constant  $0 < r < 1$ .

Both conditions are mild. Particularly, Condition (A1) implies that  $\max_j s_j(\alpha_0) \leq Kc_q^{-2}$  for some constant  $K > 0$ , where  $s_j(\alpha_0) = |\{i : \max_{d=1,2} |r_{d,i,j}| \geq c_q\}|$  and  $c_q$  is a correlation order that depends on  $q$ , with a common choice of  $(\log q)^{-1-\alpha_0}$  for some  $\alpha_0 > 0$ . In other words, it allows at most  $O\{qc_q^{-2}\}$  highly correlated pairs of network entries. For the high-dimensional

### 3. TWO-SAMPLE TEST ON NETWORK DATA

vector-valued data, such a condition on the eigenvalues of the covariance matrix is commonly imposed (Bickel et al., 2008; Rothman et al., 2008; Yuan, 2010; Cai et al., 2014). Condition (A2) is also mild, because if  $\max_{1 \leq i < j \leq q} |r_{d,i,j}| = 1$ , then  $\Gamma_d$  is singular. We next obtain the limiting distribution of our test statistic  $M_n$ .

**Theorem 1.** *Suppose that (A1)-(A2), and one of (C1) and (C2) hold. Then for any  $x \in \mathbb{R}$ ,*

$$P_{H_0}(M_n - 2 \log q + \log \log q \leq x) \rightarrow \exp\{-\pi^{-1/2} \exp(-x/2)\}, \text{ as } n_1, n_2, q \rightarrow \infty.$$

Based on this limiting null distribution, we define the asymptotic  $\alpha$ -level test as,

$$\Psi_\alpha = I(M_n \geq 2 \log q - \log \log q + q_\alpha),$$

where  $q_\alpha = -\log \pi - 2 \log \log(1 - \alpha)^{-1}$ .

We next study the power and the asymptotic optimality of the test  $\Psi_\alpha$ . Toward that end, define the sparsity of  $\mathbf{s}_1 - \mathbf{s}_2$  as  $k_q = |\{(i, j) : s_{1,i,j} - s_{2,i,j} \neq 0, 1 \leq i < j \leq p\}|$ . We also introduce a class of  $(\mathbf{s}_1, \mathbf{s}_2)$ ,

$$\mathcal{U}(c) = \left\{ (\mathbf{s}_1, \mathbf{s}_2) : \max_{1 \leq i < j \leq p} \frac{|s_{1,i,j} - s_{2,i,j}|}{\{\text{Var}(S_{1,l,i,j})/n_1 + \text{Var}(S_{2,l,i,j})/n_2\}^{1/2}} \geq c(\log q)^{1/2} \right\}.$$

**Theorem 2.** *Suppose that one of (C1) and (C2) holds. Then,*

$$\inf_{(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{U}(2\sqrt{2})} P(\Psi_\alpha = 1) \rightarrow 1, \text{ as } n_1, n_2, q \rightarrow \infty.$$

Furthermore, suppose that  $k_q = o(q^r)$  for some  $r < 1/2$ . Let  $\alpha, \beta > 0$  and  $\alpha + \beta = 1$ . Then there exists a constant  $c_0 > 0$  such that, for all sufficiently large  $n_d$  and  $q$ ,

$$\inf_{(\mathbf{s}_1, \mathbf{s}_2) \in \mathcal{U}(c_0)} \sup_{T_\alpha \in \mathcal{T}_\alpha} P(T_\alpha = 1) \leq 1 - \beta,$$

where  $\mathcal{T}_\alpha$  is the set of all  $\alpha$ -level tests, i.e.,  $P_{H_0}(T_\alpha = 1) \leq \alpha$  for all  $T_\alpha \in \mathcal{T}_\alpha$ .

### 3. TWO-SAMPLE TEST ON NETWORK DATA

This theorem shows that the null hypothesis in (1) can be rejected by  $\Psi_\alpha$  with a high probability if the pair of the network means belong to the class  $\mathcal{U}(2\sqrt{2})$ . In addition, with the mild sparsity condition  $k_q = o(q^r)$ , the lower bound rate of  $(\log q)^{1/2}$  cannot be further improved, because for a sufficiently small  $c_0$ , any  $\alpha$ -level test is unable to reject the null correctly uniformly over  $\mathcal{U}(c_0)$  with probability tending to 1. Henceforth, the global test  $\Psi_\alpha$  reaches the power minimax optimality asymptotically.

#### 3.3 Simultaneous test

We next develop a multiple testing procedure for (2) based on the test statistic  $T_{i,j}$  in (4). Let  $h$  be the threshold level such that  $H_{0,i,j}$  is rejected if  $|T_{i,j}| \geq h$ . Let  $\mathcal{H}_0 = \{(i, j) : s_{1,i,j} = s_{2,i,j}, 1 \leq i < j \leq p\}$  be the set of true nulls, and  $\mathcal{H}_1 = \mathcal{H} \setminus \mathcal{H}_0$  the set of true alternatives, where  $\mathcal{H} = \{(i, j) : 1 \leq i < j \leq p\}$ . Denote by  $R_0(h) = \sum_{(i,j) \in \mathcal{H}_0} I(|T_{i,j}| \geq h)$  and  $R(h) = \sum_{1 \leq i < j \leq p} I(|T_{i,j}| \geq h)$  the total number of false positives and rejections, respectively. Then we define the false discovery proportion and false discovery rate by

$$\text{FDP}(h) = \frac{R_0(h)}{R(h) \vee 1}, \quad \text{FDR}(h) = \mathbf{E}\{\text{FDP}(h)\}.$$

An ideal choice of  $h$  would reject as many true positives as possible while controlling the FDP at the pre-specified level  $\alpha$ . That is, we select  $h_0 = \inf \{h : 0 \leq h \leq (2 \log q)^{1/2}, \text{FDP}(h) \leq \alpha\}$ . Since  $R_0(h)$  is unknown, we estimate it conservatively by  $2q\{1 - \Phi(h)\}$ , where  $\Phi(h)$  is the standard normal cumulative distribution function. This leads to our multiple testing procedure as summarized in Algorithm 1.

We next show that this testing procedure controls the FDR and FDP asymptotically at the pre-specified level. For notation simplicity, we write  $\text{FDP} = \text{FDP}(\hat{h})$  and  $\text{FDR} = \text{FDR}(\hat{h})$ , where  $\hat{h}$

### 3. TWO-SAMPLE TEST ON NETWORK DATA

---

**Algorithm 1** Simultaneous inference with FDR control

---

Step 1: Estimate FDP by  $\widehat{\text{FDP}}(h) = 2q\{1 - \Phi(h)\}/\{R(h) \vee 1\}$ .

Step 2: For a given  $0 \leq \alpha \leq 1$ , calculate

$$\hat{h} = \inf \left\{ h : 0 \leq h \leq (2 \log q)^{1/2}, \widehat{\text{FDP}}(h) \leq \alpha \right\}.$$

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

Step 3: Reject  $H_{0,i,j}$  if and only if  $|T_{i,j}| \geq \hat{h}$ , for  $1 \leq i < j \leq p$ .

---

is obtained in Algorithm 1. Define  $\mathcal{A}_i(\xi) = \{j : \max(|r_{1,i,j}|, |r_{2,i,j}|) \geq (\log q)^{-2-\xi}\}$ , and  $\mathcal{S}_\rho = \{(i, j) : 1 \leq i < j \leq p, |s_{1,i,j} - s_{2,i,j}|/\{\text{Var}(S_{1,l,i,j})/n_1 + \text{Var}(S_{2,l,i,j})/n_2\}^{1/2} \geq (\log q)^{1/2+\rho}\}$ .

We further introduce some conditions.

(B1)  $|\mathcal{S}_\rho| \geq [1/\{\pi^{1/2}\alpha\} + \delta](\log q)^{1/2}$ , for some constant  $\delta > 0$  and any sufficiently small constant  $\rho > 0$ .

(B2)  $\max_{1 \leq i \leq q} |\mathcal{A}_i(\xi)| = o(q^\nu)$  for some constants  $\xi > 0$  and  $0 < \nu < (1 - r)/(1 + r)$ .

(B3)  $q_0 = |\mathcal{H}_0| \geq c_1 q$  for some constant  $c_1 > 0$ .

Condition (B1) on  $\mathcal{S}_\rho$  is mild, as it only requires a small number of  $s_1$  and  $s_2$  having standardized difference with the order of  $(\log q)^{1/2+\rho}$  for any sufficiently small constant  $\rho > 0$ . Condition (B2) is mild, as it requires that not too many  $S_{d,l,i,j}$  are highly correlated, but still allows the number of highly correlated pairs to grow in the order of  $o(q^{1+\nu})$ . Condition (B3) is also a natural and mild assumption, because if it does not hold, i.e.,  $q_0 = o(q)$ , then we can simply reject all the hypotheses. As a result, we would have  $|R_0| = q_0$ ,  $|R| = q$ , and the FDR would tend to zero. Under these conditions, we obtain the asymptotic properties of our multiple testing procedure in



terms of false discovery control.

**Theorem 3.** *Suppose that (A2), (B1)-(B3), and one of (C1) and (C2) hold with  $p \leq cn^{\gamma_0}$  for some constants  $\gamma_0, c > 0$ . Then,*

$$\lim_{(n_1, n_2, q) \rightarrow \infty} \frac{FDR}{\alpha q_0/q} = 1, \quad \text{and} \quad \frac{FDP}{\alpha q_0/q} \rightarrow 1 \text{ in probability, as } n_1, n_2, q \rightarrow \infty.$$

## 4 Power Enhancement

In brain connectivity analysis and many other applications, the sample size  $n_d$  is often small, whereas the number of nodes  $p$  can be moderate to large. This results in a limited power for the proposed test. We explore in this section an explicit power enhancement method that has potential to substantially improve the power of the simultaneous inference developed in Section 3.3. We borrow the idea of grouping, adjusting and pooling (GAP) that was first proposed in Xia et al. (2019a). However, our method differs from Xia et al. (2019a) in many ways, including a different, and actually less restrictive, assumption, a different set of primary and auxiliary statistics, and a different modification of the multiple testing procedure. We show that the modified procedure is asymptotically more powerful, meanwhile it can still control FDR and FDP asymptotically. We obtain these properties assuming the sub-Gaussian-tail condition (C1). Parallel results can be obtained under the polynomial-tail condition (C2) too, but are technically more involved. We begin by describing the intuition behind our power enhancement solution, then derive the proper auxiliary statistic for our inference problem. We then develop the modified simultaneous testing procedure, and study its asymptotic properties in terms of power improvement and false discovery control. We also compare in detail our method with the GAP method of Xia et al. (2019a).

## 4.1 Intuition

We recognize that there exists additional information in the data that is potentially useful to improve the simultaneous testing procedure of Algorithm 1. We first discuss our intuition, then use some simple example to illustrate where the auxiliary information is and how it can facilitate our multiple testing procedure.

In a multitude of applications including brain connectivity analysis, it is often believed that the difference between the two networks under different biological conditions is small. This means  $\mathbf{s}_1 - \mathbf{s}_2$  is sparse. Accordingly, one can find a baseline matrix  $\mathbf{s}_0$ , such that  $\mathbf{s}'_1 = \mathbf{s}_1 - \mathbf{s}_0$  and  $\mathbf{s}'_2 = \mathbf{s}_2 - \mathbf{s}_0$  are individually sparse. Let  $\mathcal{I}_d = \{(i, j) : s'_{d,i,j} \neq 0, 1 \leq i < j \leq p\}$  denote the support of  $\mathbf{s}'_d$ ,  $d = 1, 2$ , and  $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$  denote the union support. Note that the set of alternative hypotheses  $\mathcal{H}_1$  defined in Section 3.3 is the same as  $\mathcal{I}$ , if  $s_{1,i,j} \neq s_{2,i,j}$  for every  $(i, j) \in \mathcal{I}_1 \cap \mathcal{I}_2$ . In general,  $\mathcal{H}_1$  is a proper subset of  $\mathcal{I}$ . Since  $\mathbf{s}'_1$  and  $\mathbf{s}'_2$  are both sparse, we realize that the cardinality of  $\mathcal{I}$  is small. Moreover, the following relationship holds true:

$$(i, j) \notin \mathcal{I} \text{ implies that } s_{1,i,j} - s_{2,i,j} = 0, \quad 1 \leq i < j \leq p.$$

Therefore, the knowledge about  $\mathcal{I}$  is useful to help narrow down the search in multiple testing. In other words, if one can find a way to identify possible entries  $(i, j)$  in  $\mathcal{I}$ , it would provide useful information about the set of true alternatives  $\mathcal{H}_1$ , or equivalently, the set of true nulls  $\mathcal{H}_0$ . As a consequence, it can potentially increase the power of the testing procedure.

A key observation is then, while the test statistic is built on the difference between  $\bar{S}_{1,i,j}$  and  $\bar{S}_{2,i,j}$  as defined in Section 3.1, the sum of  $\bar{S}_{1,i,j}$  and  $\bar{S}_{2,i,j}$  can provide crucial information about  $\mathcal{I}$ . Consider a toy example where the network data is binary, and  $S_{d,l,i,j}$  follows a Bernoulli distribution with mean  $s_{d,l,i,j}$ ,  $l = 1, \dots, n_d$ ,  $d = 1, 2$ ,  $1 \leq i < j \leq p$ . Assume that  $s_{1,i,j} =$

## 4. POWER ENHANCEMENT

$s_{2,i,j} = s_{0,i,j} = 0.1$  for 80% of the  $(i, j)$  pairs,  $s_{1,i,j} = s_{2,i,j} = s_{0,i,j} = 0.9$  for 10% of the  $(i, j)$  pairs, and for the rest of the  $(i, j)$  pairs,  $s_{1,i,j}, s_{2,i,j} \sim \text{Uniform}(0.1, 0.9)$  and  $s_{0,i,j} = 0.1$ . In this example, for the pairs  $(i, j) \notin \mathcal{I}$ , the sum of  $s_{1,i,j}$  and  $s_{2,i,j}$  is either very small, which is 0.2, or very large, i.e., 1.8. Meanwhile, for the pairs  $(i, j) \in \mathcal{I}$ , the sum is in between. Henceforth, this sum contains useful information about  $\mathcal{I}$ , and can potentially enhance the power of the multiple testing procedure.

Based on the above discussion, we can see that, the more sparsity structure information the auxiliary statistics can capture, the more information they can provide about the union support  $\mathcal{I}$ , and the more substantial power gain the test can achieve. In general, the sparser the true difference  $s_1 - s_2$  is, the more information the auxiliary statistics can offer.

### 4.2 Auxiliary statistics

We next formally construct the auxiliary statistic that provides useful information about the union support  $\mathcal{I}$ . It is important to note that, the auxiliary statistic should be constructed so that they are asymptotically independent of the test statistic  $T_{i,j}$  in (4). This way the null distribution of  $T_{i,j}$  would not be distorted by the incorporation of the auxiliary statistic.

Recall  $V_{d,i,j}$  in (3) is the sample variance of  $S_{d,l,i,j}$ . We construct the auxiliary statistic as,

$$A_{i,j} = \frac{\bar{S}_{1,i,j} + \hat{\kappa}_{i,j} \bar{S}_{2,i,j}}{(V_{1,i,j}/n_1 + \hat{\kappa}_{i,j}^2 V_{2,i,j}/n_2)^{1/2}}, \quad 1 \leq i < j \leq p,$$

where  $\hat{\kappa}_{i,j} = (n_2 V_{1,i,j}) / (n_1 V_{2,i,j})$ . The next proposition shows that the test statistic  $T_{i,j}$  and the auxiliary statistic  $A_{i,j}$  are asymptotically independent under the null hypothesis. Define

$$a_{i,j} = \frac{s_{1,i,j} + \kappa_{i,j} s_{2,i,j}}{\{\text{Var}(S_{1,l,i,j}) + \kappa_{i,j}^2 \text{Var}(S_{2,l,i,j})\}^{1/2}}, \quad \text{where } \kappa_{i,j} = \frac{n_2 \text{Var}(S_{1,l,i,j})}{n_1 \text{Var}(S_{2,l,i,j})}.$$

**Proposition 1.** *Suppose (C1) holds with  $\log q = o(n^{1/c})$  for some  $c > 5$ . For any constants  $M > 0$  and  $C > 0$ , we have*

$$P_{H_{0,i,j}}(|T_{i,j}| \geq h, |A_{i,j}| \geq \lambda) = \{1 + o(1)\}G(h)\mathbf{P}(|N(0, 1) + a_{i,j}| \geq \lambda) + O(q^{-M}),$$

*uniformly for  $0 \leq h \leq C\sqrt{\log q}$ ,  $0 \leq \lambda \leq C\sqrt{\log q}$ , and  $1 \leq i < j \leq p$ , with  $G(h) = 2\{1 - \Phi(h)\}$ . Furthermore, for all  $0 \leq k \leq CN$  with an integer constant  $N$ ,*

$$P_{H_{0,i,j}}(|T_{i,j}| \geq h, |A_{i,j}| < \lambda_k) = \{1 + o(1)\}G(h)\mathbf{P}(|N(0, 1) + a_{i,j}| < \lambda_k) + O(q^{-M}),$$

*uniformly for  $0 \leq h \leq C\sqrt{\log q}$  and  $1 \leq i < j \leq p$ , where  $\lambda_k = (k/N)\sqrt{\log q}$ .*

### 4.3 Power enhanced simultaneous test

Based on  $(T_{i,j}, A_{i,j})$ , we now modify the simultaneous testing procedure of Algorithm 1. We first describe the main idea. We next summarize the modified testing procedure in Algorithm 2. Finally, we discuss some specific choices of the key parameters of the algorithm.

Since there are totally  $q = p(p-1)/2$  tests to carry out simultaneously, we rearrange the pairs of  $\{(T_{i,j}, A_{i,j}), 1 \leq i < j \leq p\}$  into  $\{(T_i, A_i), i = 1, \dots, q\}$ . After obtaining all the  $p$ -values,  $p_i = 2\{1 - \Phi(|T_i|)\}$ , from Algorithm 1, our basic idea is to adjust those  $p$ -values by  $p_i^w = \min\{p_i/w_i, 1\}$ , with  $w_i$  being the adjusting weights,  $i = 1, \dots, q$ . We utilize the auxiliary statistics  $A_i$  to help compute the adjusting weights  $w_i$ , by groups. Specifically, we consider a set of grid values,  $\mathcal{J} = \{(C_1N - 1)\sqrt{\log q}/N, C_1\sqrt{\log q}, \dots, (C_2N - 1)\sqrt{\log q}/N, C_2\sqrt{\log q}\}$ , where  $C_1, C_2$  and  $N$  are some pre-specified constants. We divide the index set  $\{1, \dots, q\}$  into  $K$  groups according to the auxiliary statistics  $(A_1, \dots, A_q)$ . As an example, we take  $K = 3$ . That is, we choose two grid points  $\mathcal{J}_K = \{\lambda_1, \lambda_2\}$  in  $\mathcal{J}$ , and obtain  $K = 3$  groups of indices,

#### 4. POWER ENHANCEMENT

$\mathcal{G}_1 = \{i : 1 \leq i \leq q, -\infty < A_i \leq \lambda_1\}$ ,  $\mathcal{G}_2 = \{i : 1 \leq i \leq q, \lambda_1 < A_i \leq \lambda_2\}$ , and  $\mathcal{G}_3 = \{i : 1 \leq i \leq q, \lambda_2 < A_i \leq \infty\}$ . For each group  $\mathcal{G}_k$ , we compute its cardinality,  $q_k = |\mathcal{G}_k|$ . We also estimate the proportion,  $\pi_k$ , of alternatives in  $\mathcal{G}_k$ ,  $k = 1, \dots, K$ . To do so, we employ the method of Schweder and Spjøtvoll (1982) and Storey (2002) to obtain an estimate  $\tilde{\pi}_k$  first, then stabilize it by  $\hat{\pi}_k = (\epsilon \vee \tilde{\pi}_k) \wedge (1 - \epsilon)$ , where  $\epsilon$  is a small positive number; we set  $\epsilon = 10^{-5}$ . Then for all the indices in  $\mathcal{G}_k$ , we compute the group-wise adjusting weight:

$$w_i = \left( \sum_{k=1}^K \frac{q_k \hat{\pi}_k}{1 - \hat{\pi}_k} \right)^{-1} \frac{q \hat{\pi}_k}{(1 - \hat{\pi}_k)}, \quad i \in \mathcal{G}_k, 1 \leq k \leq K. \quad (5)$$

This idea of adjusting the weights  $w_i$  by groups is motivated by our intuition in Section 4.1. After obtaining the weights, we adjust the  $p$ -values and apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995, BH) to the adjusted  $p$ -values  $p_i^w$ . Finally, we search all possible choices of  $\mathcal{J}_K$  among  $\mathcal{J}$ , and find the one that yields the largest number of rejections. We apply BH again to the adjusted  $p$ -values under this choice of  $\mathcal{J}_K$  to obtain the final adjusted rejection region. We summarize this modified simultaneous testing procedure in Algorithm 2.

We discuss some specific choices of the parameters in Algorithm 2. First, the number of groups  $K$  is usually set at  $K = 3$ . As shown in Xia et al. (2019a), when  $K \geq 4$ , there is little additional power gain, but a more expensive computation. Second, the constants  $C_1$  and  $C_2$  can be chosen so that  $C_1 \sqrt{\log q}$  is equal to the smallest value of the auxiliary statistics and  $C_2 \sqrt{\log q}$  is equal to the largest value of the auxiliary statistics. If the absolute values of the smallest and largest auxiliary statistics exceed  $16 \sqrt{\log q}$ , we truncate at  $C_1 = -16$  and  $C_2 = 16$  to stabilize and expedite the computation. We note here that, if the network data are non-negative, such as the binary and poisson network data, then both  $C_1$  and  $C_2$  are non-negative. By contrast, in Xia et al. (2019a),  $C_1$  and  $C_2$  were fixed at  $-4$  and  $4$ . Finally,  $N$  can be any integer for the theoretical

---

**Algorithm 2** Adjusted simultaneous inference with FDR control and power enhancement.

---

Step 1: Initialization:

Step 1.1: Compute the test statistics and auxiliary statistics  $\{(T_i, A_i), i = 1, \dots, q\}$ .

Step 1.2: Compute the  $p$ -values:  $p_i = 2\{1 - \Phi(|T_i|)\}, i = 1, \dots, q$ .

Step 1.3: Input the pre-specified constants  $K, C_1, C_2$  and  $N$ .

Step 1.4: Compute the grid set:

$$\mathcal{J} = \left\{ (C_1 N - 1)\sqrt{\log q}/N, C_1\sqrt{\log q}, \dots, (C_2 N - 1)\sqrt{\log q}/N, C_2\sqrt{\log q} \right\}.$$

Step 2: For each  $\mathcal{J}_K = \{\lambda_1, \dots, \lambda_{K-1}\}$  in  $\mathcal{J}$ , and  $\lambda_0 = -\infty, \lambda_K = \infty$ :

Step 2.1: Construct  $\mathcal{G}_k = \{i : 1 \leq i \leq q, \lambda_{k-1} < A_i \leq \lambda_k\}, 1 \leq k \leq K$ .

Step 2.2: For each  $\mathcal{G}_k$ , compute the cardinality,  $q_k = |\mathcal{G}_k|$ .

Step 2.3: For each  $\mathcal{G}_k$ , estimate the proportion,  $\hat{\pi}_k$ , of alternatives in  $\mathcal{G}_k$ .

Step 2.4: Compute the adjusting weights  $w_i, i = 1, \dots, q$ , according to (5).

Step 2.5: Adjust the  $p$ -values:  $p_i^w = \min\{p_i/w_i, 1\}, i = 1, \dots, q$ .

Step 2.6: Apply the BH procedure, and record the total number of rejections.

Step 3: Obtain the adjusted rejection region:

Step 3.1: Choose  $\mathcal{J}_K$  that yields the largest number of rejections.

Step 3.2: Compute the corresponding adjusted  $p$ -values:  $p_i^w, 1 \leq i \leq q$ .

Step 3.3: Reorder all the adjusted  $p$ -values:  $p_{(1)}^w \leq \dots \leq p_{(q)}^w$ .

Step 3.4: Output the rejection region  $\{i : i < \hat{\tau}\}$ , where  $\hat{\tau} = \max\{i : p_{(i)}^w \leq \alpha i/q\}$ .

---

validity. Numerically, a larger value of  $N$  implies a more precise grid search, but at the cost of a heavier computational burden. We choose  $N$  such that the gap between two adjacent grid points,  $(\log p)^{1/2}/N$ , equals 0.1 approximately.

#### 4.4 FDR control and power enhancement

We next show that the modified inference of Algorithm 2 is asymptotically more powerful than Algorithm 1, meanwhile it can still asymptotically control the false discovery.

Denote  $\{p_i^w : 1 \leq i \leq q\}$  the adjusted  $p$ -values from Algorithm 2, and  $\{p_{(i)}^w : 1 \leq i \leq q\}$  the ordered adjusted  $p$ -values. The corresponding adjusted FDP is:

$$\text{FDP}_{\text{adj}} = \frac{\sum_{i \in \mathcal{H}_0} I \left\{ p_i^w \leq p_{(\hat{\tau})}^w \right\}}{\sum_{i=1}^q I \left\{ p_i^w \leq p_{(\hat{\tau})}^w \right\} \vee 1},$$

where  $\hat{\tau}$  is the cutoff obtained from Step 3.4 of Algorithm 2, and  $I(\cdot)$  is the indicator function. Accordingly,  $\text{FDR}_{\text{adj}} = \mathbf{E}(\text{FDP}_{\text{adj}})$ . The next theorem shows that the modified procedure can still control FDR and FDP asymptotically.

**Theorem 4.** *Suppose (A2), (B1)-(B3), and (C1) hold with  $p \leq cn^{\gamma_0}$  for constants  $\gamma_0, c > 0$ . Then,*

$$\lim_{(n_1, n_2, q) \rightarrow \infty} \frac{\text{FDR}_{\text{adj}}}{\alpha q_0 / q} = 1, \quad \text{and} \quad \frac{\text{FDP}_{\text{adj}}}{\alpha q_0 / q} \rightarrow 1 \text{ in probability, as } n_1, n_2, q \rightarrow \infty.$$

Next, denote the power of the testing procedures of Algorithms 1 and 2 by  $\Psi$  and  $\Psi_{\text{adj}}$ , respectively. That is,

$$\Psi = \mathbf{E} \left\{ \frac{\sum_{(i,j) \in \mathcal{H}_1} I(|T_{i,j}| \geq \hat{h})}{|\mathcal{H}_1|} \right\}, \quad \Psi_{\text{adj}} = \mathbf{E} \left[ \frac{\sum_{i \in \mathcal{H}_1} I \left\{ p_i^w \leq p_{(\hat{\tau})}^w \right\}}{|\mathcal{H}_1|} \right].$$

Then the next theorem shows that, by incorporating the auxiliary statistics  $A_{i,j}$ , the modified simultaneous testing procedure of Algorithm 2 is asymptotically more powerful than Algorithm 1, which is solely based on the test statistics  $T_{i,j}$ .

**Theorem 5.** *Suppose the same conditions in Theorem 4 hold. Then,*

$$\Psi_{adj} \geq \Psi + o(1), \quad \text{as } q \rightarrow \infty.$$

## 4.5 Comparison to GAP

Although motivated by the GAP method of Xia et al. (2019a), our power enhancement procedure is also considerably different. While GAP tackled the problem of mean comparison of vector-valued samples, we target the problem of network mean comparison. This leads to a different set of test and auxiliary statistics, but a number of additional intrinsic differences as well.

First, the two methods impose different assumptions. A key requirement for GAP to enhance the power is that the parameters of interest from each group are individually sparse. In our setup, however, the parameters may all be non-negative. For instance, in a binary network or a count network, all the entries of both means  $\mathbf{s}_1$  and  $\mathbf{s}_2$  are usually non-negative. As such, the means may not be individually sparse. Our procedure instead only requires the difference of the two means  $\mathbf{s}_1 - \mathbf{s}_2$  is sparse, which reasonably holds and is often imposed in numerous applications including brain connectivity analysis (Zhu and Li, 2018).

Second, the two methods differ in terms of the range of the auxiliary statistics that contribute most to the power enhancement. Consider the case when  $K = 3$ . In Xia et al. (2019a), since both means are assumed to be individually sparse, the tests that are more likely to be adjusted and rejected are those with the corresponding auxiliary statistics either being negative and small, or positive and large. That is, the power enhancement hinges more on those tests in  $\mathcal{G}_1$  and  $\mathcal{G}_3$  with small or large auxiliary statistics. However, in our setup, the individual means  $\mathbf{s}_1$  and  $\mathbf{s}_2$  can both be dense and their entries are all positive. Instead we only assume that  $\mathbf{s}_1 - \mathbf{s}_2$  is sparse. Take a



#### 4. POWER ENHANCEMENT

---

binary brain connectivity network as an example. The observed networks are often sparse, in that most links are zero, since it is known that brain connections are energy consuming and biological units tend to minimize energy-consuming activities (Raichle and Gusnard, 2002; Bullmore and Sporns, 2009). This translates to small connection probabilities for most entries of  $s_1$  and  $s_2$ , while all these probabilities are positive. Moreover, the difference of the means between the two populations is often sparse, which translates to equal connection probabilities for most entries of  $s_1$  and  $s_2$ , or equivalently zero difference for most entries of  $s_1 - s_2$ . This is similar to the toy example we discuss in Section 4.1. For such cases, as a consequence of Algorithm 2, the tests whose corresponding auxiliary statistics are too small or too large would be adjusted so that they are less likely to be rejected. Instead, those tests whose auxiliary statistics are in between would be adjusted so that they are more likely to be rejected. In other words, the power enhancement in our setup may hinge more on  $\mathcal{G}_2$ , rather than  $\mathcal{G}_1$  and  $\mathcal{G}_3$ .

Third, due to the above difference, the grid construction in Step 1.4 of Algorithm 2 is noticeably different from that of GAP in Xia et al. (2019a). Specifically, in Xia et al. (2019a), to ensure the inclusion of important locations in  $\mathcal{G}_1$  and  $\mathcal{G}_3$ , the constants  $C_1$  and  $C_2$  can be simply fixed at  $-4$  and  $4$ , respectively, so that the upper bound of those small negative auxiliary statistics and lower bound of those large positive auxiliary statistics can be attained in the grid  $\mathcal{J}$ . By contrast, for our problem, the upper bound of the auxiliary statistics in the union support  $\mathcal{I}$  can go beyond the bound in Xia et al. (2019a), i.e.,  $4\sqrt{\log q}$ , and the lower bound of the auxiliary statistics in the union support can be non-negative. Since the tests in  $\mathcal{G}_2$  are more likely to be adjusted and rejected, we need to do a more thorough grid construction and choose the constants  $C_1$  and  $C_2$  based on the smallest and largest values of the auxiliary statistics as described in Section 4.3.

## 5 Simulations

We first present the simulation setup, where we consider different network structures, sparsity levels, network sizes and sample sizes. We then investigate the empirical performance of the global test, and compare the two simultaneous tests, Algorithms 1 and 2.

### 5.1 Setup

We consider  $p \times p$  networks, with two network sizes,  $p = 100$  and  $200$ . This results in  $q = 100(100 - 1)/2 = 4590$  and  $q = 200(200 - 1)/2 = 19900$  links, respectively. We consider five common network structures, the Bernoulli, Bernoulli mixture, and transformed Wishart distributions, and for the Bernoulli case, the binary links are generated from a power-law distribution, a stochastic block model, and an Erdős-Rényi model. For each network structure, we further consider three sparsity levels.

- + **Bernoulli:** Select the sets  $\mathcal{M}_{d,1}$  and  $\mathcal{M}_0$  from  $q$  hypotheses according to the following models generated by the R package `igraph`, with  $|\mathcal{M}_{d,1}| = |\mathcal{M}_0| = k_q/2$ ,  $d = 1, 2$ . Here  $k_q$  is a parameter that controls the sparsity level, and is specified later.
  - **Power-law distribution:** with  $p$  nodes,  $k_q/2$  edges, the power law exponent of the degree distribution is set to 2.1, and all other parameters are set to the default values.
  - **Stochastic block model:** with 2 blocks and the diagonal Bernoulli rates matrix, where the diagonal values are set to  $k_q/(2q)$ .
  - **Erdős-Rényi model:** with  $p$  nodes and  $k_q/2$  edges.

Let  $\mathcal{M}_d = \mathcal{M}_{d,1} \cup \mathcal{M}_0$ . For  $(i, j) \notin \mathcal{M}_d$ , generate  $S_{d,l,i,j} \sim \text{Bernoulli}(1, 0.3)$ . For  $(i, j) \in \mathcal{M}_d$ , generate  $S_{d,l,i,j} \sim \text{Bernoulli}(1, r_{d,i,j})$ , where  $r_{1,i,j}$  is set to 0.5 with probability 0.1, and 0.8 otherwise, whereas  $r_{2,i,j}$  is set to 0.8 with probability 0.1, and 0.5 otherwise.

+ **Bernoulli mixture:** Generate  $\mathcal{M}_d$  in the same way as before. Generate  $S_{d,l,i,j} \sim \text{Bernoulli}(1, r_{d,i,j})$ , where  $r_{d,i,j} = \pi_{i,j} * r_{d,1,i,j} + (1 - \pi_{i,j}) * r_{d,2,i,j}$ , with  $\pi_{i,j} \sim \text{Uniform}(0, 1)$ . For  $(i, j) \notin \mathcal{M}_d$ ,  $r_{d,1,i,j} = r_{d,2,i,j} = 0.3$ ,  $d = 1, 2$ . For  $(i, j) \in \mathcal{M}_d$ ,  $r_{1,1,i,j}$  is set to 0.5 with probability 0.1, and 0.7 otherwise, whereas  $r_{2,1,i,j}$  is set to 0.7 with probability 0.1, and 0.5 otherwise, and  $r_{d,2,i,j} = r_{d,1,i,j} + 0.2$ .

+ **Wishart with logarithm transformation:** Select the sets  $\mathcal{M}_{d,1}$  and  $\mathcal{M}_0$  from  $q$  hypotheses, uniformly and randomly, with  $|\mathcal{M}_{d,1}| = k_q/4$ , and  $|\mathcal{M}_0| = 3k_q/4$ ,  $d = 1, 2$ . Let  $\mathcal{M}_d = \mathcal{M}_{d,1} \cup \mathcal{M}_0$ . Generate  $\Sigma_d$  such that  $\Sigma'_{d,i,j} = \text{Uniform}(3, 5)$  if  $(i, j) \in \mathcal{M}_d$  and  $\Sigma'_{d,i,j} = 0$  otherwise. Let  $\Sigma'_{d,j,i} = \Sigma'_{d,i,j}$  and  $\Sigma_d = \Sigma'_d + \{|\lambda_{\min}(\Sigma'_d)| + 0.5\}I$ , where  $I$  is the identify matrix. Generate  $S'_{d,l} \sim \text{Wishart}(m^{-1}\Sigma_d, m)$ , with  $m = 300$ , and  $S_{d,l} = \log[\text{round}\{\exp(S'_{d,l})\}]$ , where  $\text{round}(\cdot)$  rounds a number to the nearest integer.

For each network structure, the parameter  $k_q$  controls the sparsity level, and we examine three levels,  $k_q = 0.2q, 0.15q$  and  $0.1q$ , where  $q$  is the total number of the network links.

## 5.2 Results

First, we investigate the empirical size of the proposed global test  $\Psi_\alpha$  for the global testing problem (1). For this problem, the population network means are equal to each other under the null hypothesis, and we set  $\mathcal{M}_1 = \mathcal{M}_2$ , and set the sample size  $n_1 = n_2 = 500$ . We also

## 5. SIMULATIONS

---

compare with the global testing method aSPU developed by Kim et al. (2014), implemented in the R package aSPU. Table 1 reports the empirical size of the two tests, in percentage, based on 1000 data replications under the significance level  $\alpha = 5\%$ . It is seen clearly from the table that our proposed global testing procedure controls the type I error reasonably well, while the aSPU method has a slight size inflation in some cases though not severe. In addition, we report the computation time of each method, in seconds, averaged over three sparsity levels and all replications. It is seen from the last two columns of the table that, the average computation time of aSPU is much longer than our method; e.g., for  $p = 100$  and  $p = 200$ , it is about 9 times and 15 times of our method, respectively. This is because Kim et al. (2014) did not derive the theoretical null distribution of their test statistics, but instead employed the permutations to obtain the critical value, which results in a more time consuming procedure. We also note that, in this setting, the sample size is much smaller than the total number of hypotheses  $q$ , but is larger than the sample size we use in the multiple testing simulations. This is due to the relatively slow convergence rate of the Bernoulli normal approximation and the maximum type statistics.

Next, we examine the empirical FDR and the empirical power of the simultaneous testing procedure for the multiple testing problem (2). We consider two sample sizes,  $n_1 = n_2 = 100$  and  $n_1 = n_2 = 25$ , and the latter mimics the real data setting where the sample size is very limited. We apply both Algorithms 1 and 2, one with the proposed power enhancement, and one without. Table 2 reports the empirical FDR and power, both in percentage, based on 100 replications under the significance level  $\alpha = 5\%$  for the Bernoulli network structure. Table 3 reports the results for the Bernoulli mixture and Wishart with logarithm transformation. It is seen that, in all cases, the empirical FDRs are generally controlled under the nominal level by both

## 5. SIMULATIONS

Table 1: The empirical size and computation time for our global test  $\Psi_\alpha$  and the aSPU test of Kim et al. (2014). The empirical size is in percentage based on 1000 data replications. The computation time is in seconds averaged over three sparsity levels and all replications. The significance level is  $\alpha = 5\%$ , and the sample size is  $n_1 = n_2 = 500$ .

	Method	$p = 100$			$p = 200$			Computation time	
		$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$	$p = 100$	$p = 200$
Power-law	$\Psi_\alpha$	4.1	5.1	5.6	5.0	3.6	4.9	0.52	1.99
	aSPU	5.8	6.7	6.0	5.8	5.6	5.7	4.86	32.6
Stochastic Block	$\Psi_\alpha$	4.3	5.4	5.2	5.9	4.4	5.1	0.56	2.00
	aSPU	6.2	6.1	5.5	6.2	7.0	5.4	4.66	32.8
Erdős-Rényi	$\Psi_\alpha$	5.3	5.4	4.1	4.8	5.4	5.4	0.52	2.01
	aSPU	5.9	7.5	5.2	6.1	5.3	6.4	4.74	32.3
Bernoulli mixture	$\Psi_\alpha$	6.2	5.1	4.9	4.8	3.4	5.5	0.51	2.06
	aSPU	6.8	6.3	5.4	7.0	5.4	6.0	4.77	32.9
Transformed Wishart	$\Psi_\alpha$	4.9	4.5	5.7	4.2	4.8	4.5	0.54	2.23
	aSPU	6.5	6.9	6.3	6.5	6.8	5.7	4.42	28.1

algorithms. Algorithm 2 is slightly more conservative than Algorithm 1, which is mainly due to the normalization step of the weight calculation as shown in (5). A similar phenomenon has also been observed in Xia et al. (2019a). For the empirical power, it is seen that Algorithm 2 achieves a clear power improvement over Algorithm 1, without sacrificing the size of the test. This is mainly due to the utilization of the auxiliary information in Algorithm 2. Furthermore,

## 5. SIMULATIONS

**Table 2:** The empirical FDR and the empirical power for the simultaneous testing procedures, Algorithms 1 and 2. The results are in percentages based on 100 data replications. The significance level is  $\alpha = 5\%$ . The network structure is Bernoulli.

Network structure	$p = 100$						$p = 200$					
	$n_1 = n_2 = 100$			$n_1 = n_2 = 25$			$n_1 = n_2 = 100$			$n_1 = n_2 = 25$		
	$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$
Bernoulli, power-law												
Empirical FDR												
Algorithm 1	4.1	4.5	4.7	6.2	6.3	7.2	4.2	4.4	4.9	5.8	6.5	7.5
Algorithm 2	2.6	2.6	2.9	3.5	4.6	5.3	2.2	2.3	2.5	3.5	4.9	5.1
Empirical power												
Algorithm 1	88.7	87.0	84.7	42.2	40.8	39.7	88.7	87.0	84.9	41.6	40.5	39.9
Algorithm 2	92.1	91.7	90.9	54.8	54.1	53.4	92.3	91.8	91.0	54.2	53.9	53.2
Bernoulli, stochastic block												
Empirical FDR												
Algorithm 1	4.3	4.4	4.8	6.1	6.4	7.7	4.2	4.5	4.9	5.8	6.3	7.6
Algorithm 2	2.8	2.7	3.0	3.5	4.5	5.5	2.2	2.3	2.5	3.5	5.0	5.0
Empirical power												
Algorithm 1	89.0	87.1	84.8	41.5	40.4	40.0	89.0	87.0	84.9	41.4	40.4	40.1
Algorithm 2	92.2	91.7	90.8	54.5	54.5	54.0	92.5	91.9	90.9	54.2	53.9	53.4
Bernoulli, Erdős-Rényi												
Empirical FDR												
Algorithm 1	4.1	4.4	4.8	6.0	6.0	7.3	4.0	4.4	4.8	5.9	5.2	7.3
Algorithm 2	2.3	2.6	2.9	3.9	4.1	5.7	2.1	2.2	2.4	3.7	4.5	5.1
Empirical power												
Algorithm 1	88.0	86.9	84.7	44.1	41.8	40.6	88.1	86.8	84.5	44.4	42.0	40.8
Algorithm 2	91.8	91.3	90.6	54.7	54.4	53.3	91.9	91.4	90.5	54.6	54.1	53.6

## 5. SIMULATIONS

Table 3: The empirical FDR and the empirical power for the simultaneous testing procedures, Algorithms 1 and 2. The results are in percentages based on 100 data replications. The significance level is  $\alpha = 5\%$ . The network structure is Bernoulli mixture and transformed Wishart.

	$p = 100$						$p = 200$					
	$n_1 = n_2 = 100$			$n_1 = n_2 = 25$			$n_1 = n_2 = 100$			$n_1 = n_2 = 25$		
	$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$	$0.2q$	$0.15q$	$0.1q$
Network structure												
Bernoulli mixture	Empirical FDR											
Algorithm 1	4.0	4.4	4.8	6.1	6.0	7.4	4.0	4.4	4.8	6.0	5.8	7.5
Algorithm 2	1.4	1.7	2.0	3.2	4.2	5.0	1.3	1.5	1.7	2.9	4.3	4.5
	Empirical power											
Algorithm 1	88.3	87.2	85.6	41.8	40.8	41.1	88.2	87.1	85.7	41.6	40.7	41.2
Algorithm 2	93.8	93.6	93.6	54.2	54.3	54.1	93.5	93.6	93.5	53.9	54.2	54.1
Transformed Wishart	Empirical FDR											
Algorithm 1	4.2	4.6	5.1	5.0	5.6	6.6	4.2	4.6	4.9	4.9	5.3	5.9
Algorithm 2	1.6	1.8	2.0	2.4	2.8	3.7	1.6	1.9	2.0	1.8	2.1	2.6
	Empirical power											
Algorithm 1	63.5	65.9	69.6	44.1	46.8	50.6	52.6	55.7	60.4	37.5	40.2	43.1
Algorithm 2	70.9	73.8	78.4	50.3	54.5	59.9	59.8	63.9	69.9	41.4	45.2	50.0

the performance under the varying sample size confirms the power enhancement of Algorithm 2 as theoretically revealed in Section 4.4. We also observe that, the power gain becomes more substantial when the true difference  $s_1 - s_2$  becomes more sparse, which agrees with our intuition explained in Section 4.1.

## 6 Brain Connectivity Analysis

We illustrate our method with two brain connectivity analysis examples.

### 6.1 Structural connectivity analysis

The first example is a brain structural connectivity analysis of diffusion tensor images (DTI). DTI is a magnetic resonance imaging technique that measures the diffusion of water molecules to map white matter tractography in the brain. The data we analyze is the KKI-42 dataset, available at <http://openconnecto.me/data/public/MR/archive/>, and its detailed description can be found in Landman et al. (2011). This data consists of 21 subjects with no history of neurological conditions, aging from 22 to 61 years old. Each subject received two resting-state DTI under a scan-rescan imaging session. For simplicity, we treat the data as if those images were from independent samples, which is common for the analysis of this dataset (Wang et al., 2017). It results in a total sample size of 42 for this study. Brain regions are constructed following the Desikan Atlas (Desikan et al., 2006), leading to  $p = 68$  regions equally divided in the left and right hemispheres. Each DTI image has been preprocessed and summarized in the form of a  $68 \times 68$  network, where the edges record the total number of white matter fibers between the pair of nodes. It is also equally common to focus on the form of a binary network, where the edges become the binary indicators of presence or absence of white matter fibers (Wang et al., 2017). We partition the subjects into two age groups, the ones whose are younger than 30 years, and the ones who are 30 or older. Age 30 is a transition period, usually known as the “age 30 transition”, when the first phase of early adulthood comes to a close, and the basis for the next life structure is formed. Moreover, this partition yields about the same number of subjects for each group, with



## 6. BRAIN CONNECTIVITY ANALYSIS

Table 4: Structural connectivity analysis of the KKI-42 dataset. Reported are the significantly different links found by Algorithms 1 and 2 for the binary and count network data, respectively.

Binary network		Count network	
Algorithm 1	Algorithm 2	Algorithm 1	Algorithm 2
r.posteriorcingulate ↔ l.superiorparietal	r.posteriorcingulate ↔ l.superiorparietal	r.corpuscallosum ↔ l.superiorparietal	r.corpuscallosum ↔ l.superiorparietal
r.posteriorcingulate ↔ l.supramarginal	r.precuneus ↔ l.postcentral	l.isthmuscingulate ↔ l.posteriorcingulate	l.isthmuscingulate ↔ l.posteriorcingulate
–	r.caudalanteriorcingulate ↔ r.lingual	r.caudalmiddlefrontal ↔ r.rostralmiddlefrontal	r.caudalmiddlefrontal ↔ r.rostralmiddlefrontal
–	r.posteriorcingulate ↔ l.caudalmiddlefrontal	l.lateralorbitofrontal ↔ l.superiorfrontal	l.lateralorbitofrontal ↔ l.superiorfrontal
–	l.lateraloccipital ↔ l.parsopercularis	–	r.posteriorcingulate ↔ l.precuneus
–	r.superiorparietal ↔ l.precentral	–	r.caudalmiddlefrontal ↔ r.parsotriangularis
–	r.paracentral ↔ l.superiorparietal	–	l.fusiform ↔ l.temporalpole
–	l.bankssts ↔ l.frontalpole	–	l.torhinal ↔ l.lateralorbitofrontal
		–	r.corpuscallosum ↔ l.precuneus
		–	l.caudalmiddlefrontal ↔ l.pericalcarine
		–	r.bankssts ↔ r.postcentral
		–	l.lateralorbitofrontal ↔ l.temporalpole
		–	l.parsopercularis ↔ l.rostralmiddlefrontal
		–	l.medialorbitofrontal ↔ l.temporalpole
		–	l.corpuscallosum ↔ l.superiorparietal

$n_1 = 22$  for the younger-than-30 age group, and  $n_2 = 20$  for the older-than-30 age group. We study the age-related difference in structural connectivity patterns, which is of universal interest, as aging is the main risk factor for progressive loss of both structures and functions of brain neurons (Morrison and Hof, 1997).

We apply both multiple testing procedures Algorithms 1 and 2 to this dataset, first the binary network, then the count network with a logarithm transformation. We set the significance level at 0.05. For the binary network, out of the total of 2278 links, Algorithm 1 identifies 2 significantly different links, whereas the power-enhanced Algorithm 2 identifies 8 links, including the first link found by Algorithm 1 plus 7 additional links. For the count network, Algorithm 1 identifies 4 significantly different links, whereas Algorithm 2 identifies 15 links, including all the links found

## 6. BRAIN CONNECTIVITY ANALYSIS

---

by Algorithm 1 plus 11 additional links. These results agree with both our theory and simulations, in that Algorithm 2 is usually able to recognize more significant links than Algorithm 1. Table 4 reports the identified links by the two algorithms for both types of network data. Some links found by our power-enhanced procedure agree with the neuroscience literature; for instance, the link between left fusiform and left temporal pole under the count network. The temporal pole, also known as Brodmanna area 38, is a paralimbic region involved in high-level semantic memories and socio-emotional processing. The fusiform gyrus is part of the temporal lobe and occipital lobe in Brodmann area 37, and is linked with various neural pathways related to recognition. Li et al. (2013) also found significant difference in structural connectivity patterns between the left fusiform and the left temporal pole, for the young subjects (18 to 23 years old) versus the middle-aged and old subjects (30 to 58, and 61 to 89 years old). Meanwhile, other links found by our procedure require further scientific validation; for instance the links between left temporal pole and orbitofrontal cortex. The latter is a prefrontal cortex region in the frontal lobe of the brain involved in the cognitive process of decision-making.

### 6.2 Functional connectivity analysis

The second example is a brain functional connectivity analysis of functional magnetic resonance images (fMRI). Functional MRI measures the blood oxygen level signals, and provides a tool to study brain functional connectivity network. The data we analyze is the ADHD-200 dataset, and is available at <http://neurobureau.projects.nitrc.org/ADHD200/Data.html>. More detailed description can be found in Ahn et al. (2015). Attention deficit hyperactivity disorder (ADHD) is one of the most commonly diagnosed child-onset neurodevelopmental disorders,

## 6. BRAIN CONNECTIVITY ANALYSIS

Table 5: Functional connectivity analysis of the ADHD-200 dataset. Reported are the significantly different links found by Algorithms 1 and 2 for the Pearson correlation and partial correlation network data, respectively.

Pearson correlation network		Partial correlation network	
Algorithm 1	Algorithm 2	Algorithm 1	Algorithm 2
–	r.frontal.sup ↔ r.frontal.med.orb	–	l.paracentral.lobule ↔ r.paracentral.lobule
–	r.cerebelum6 ↔ r.cerebelum.8	–	r.frontal.sup.orb ↔ r.frontal.mid.orb
–	l.cerebelum8 ↔ vermis7	–	r.frontal.inf.orb ↔ l.temporal.pole.sup
		–	r.fusiform ↔ r.cerebelum6
		–	l.frontal.sup ↔ l.frontal.mid

and has an estimated childhood prevalence of 5–10% worldwide (Pelham et al., 2007). This data consists of 96 subjects with ADHD, and 91 normal controls. Each subject received a resting-state fMRI scan, and each brain image is parcellated using the Anatomical Automatic Labeling (AAL) Atlas with  $p = 116$  regions (Tzourio-Mazoyer et al., 2002). The resulting data is a spatial by temporal matrix, which is then turned into a Pearson correlation matrix or a partial correlation matrix to represent the brain functional connectivity network. Actually, both correlation measures are frequently used in functional connectivity analysis (Bullmore and Sporns, 2009). We use both measures to study the difference in functional connectivity patterns between the two groups of subjects with and without ADHD.

We again apply both multiple testing procedures Algorithms 1 and 2 to this dataset, first the Pearson correlation network, then the partial correlation network. For the Pearson correlation network, Algorithm 1 identifies no significantly different links, whereas the power-enhanced

Algorithm 2 identifies 3 links. For the partial correlation network, Algorithm 1 again identifies no significantly different links, whereas the power-enhanced Algorithm 2 identifies 5 links. Table 5 reports the identified links by the two algorithms for both types of network data. One brain region that differentiating links concentrate is cerebellum. The cerebellum is responsible for motor control and cognitive functions such as attention and language, and dysfunction in the cerebellum in ADHD patients have been reported (Toplak et al., 2006). We also remark that, there are fewer links found here compared to those found in Xia and Li (2019). This is because the data in the format of spatial temporal matrix analyzed in Xia and Li (2019) carries more information than the data in the format of correlation matrix. Nevertheless, the focus of this article is to develop inferential tests for the scientific applications where only the data format of some symmetric network matrix is available.

## 7 Conclusion

In this article, we develop both global and simultaneous inference methods for network comparisons when the data are observed in the form of  $p \times p$  matrices, each of which encodes the network structure for an individual subject. This data format is different from those studied in the existing network literature, and leads to a different set of testing procedures and the associated theory. In addition, we propose a power enhancement approach to tackle the challenge of limited sample size in numerous applications.

We have primarily focused on the scenario of using a symmetric matrix to encode a network structure in this article. In principle, our methods can be extended to the asymmetric matrix scenario as well, with corresponding modifications of the total number of tests and the related

theoretical properties. In the interest of space, we leave it as our future research.

## 8 Supplementary Material

The additional lemmas and theorem proofs are available in the online supplementary material.

### Acknowledgement

Xia's research was partially supported by NSFC grants 11771094, 11690013 and The Recruitment Program of Global Experts Youth Project. Li's research was partially supported by NSF grant DMS-1613137 and NIH grants R01AG061303, R01AG062542 and R01AG034570.

### 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.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.*, 57:289–300.
- Bickel, P. J., Levina, E., et al. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36:199–227.
- Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10:186–198.

---

REFERENCES

- 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.
- Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependency. *Journal of the Royal Statistical Society, Series B.*, 76:349–372.
- 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.
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31:968–980.
- Durante, D. and Dunson, D. B. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13:29–58.
- 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.
- Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S., Kolaczyk, E. D., et al. (2017). Hypoth-

---

REFERENCES

- esis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750.
- 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.
- Lan, W., Fang, Z., Wang, H., and Tsai, C.-L. (2018). Covariance matrix estimation via network structure. *J. Bus. Econom. Statist.*, 36(2):359–369.
- Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A., Bogovic, J. A., Hua, J., Chen, M., Jarso, S., et al. (2011). Multi-parametric neuroimaging reproducibility: a 3-t resource study. *Neuroimage*, 54:2854–2866.
- Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40:908–940.
- Li, X., Pu, F., Fan, Y., Niu, H., Li, S., and Li, D. (2013). Age-related changes in brain structural covariance networks. *Frontiers in Human Neuroscience*, 7:98.
- Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41:2948–2978.
- Luscombe, N. M., Madan Babu, M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312.
- Morrison, J. H. and Hof, P. R. (1997). Life and death of neurons in the aging brain. *Science*, 278:412–419.

---

REFERENCES

- 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(1, Supplement):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.
- Raichle, M. E. and Gusnard, D. A. (2002). Appraising the brain’s energy budget. *Proceedings of the National Academy of Sciences*, 99:10237–10239.
- Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.
- Schott, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *Journal of Multivariate Analysis*, 98:1825–1839.
- Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502.
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B.*, 64:479–498.
- 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(1):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



---

REFERENCES

- a macroscopic anatomical parcellation of the {MNI} {MRI} single-subject brain. *NeuroImage*, 15(1):273 – 289.
- Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42:1166–1202.
- Wang, L., Zhang, Z., and Dunson, D. (2017). Common and individual structure of multiple networks. *arXiv preprint arXiv:1707.06360*.
- 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, 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., Cai, T. T., and Sun, W. (2019a). GAP: a general framework for information pooling in two-sample sparse inference. *Journal of the American Statistical Association*, to appear.
- Xia, Y. and Li, L. (2019). Matrix graph hypothesis testing and application in brain connectivity alternation detection. *Statistica Sinica*, 29:303–328.
- Xia, Y., Li, L., Lockhart, S. N., and Jagust, W. J. (2019b). Simultaneous covariance inference for multimodal integrative analysis. *Journal of the American Statistical Association*, accepted:1–30.

---

REFERENCES

- Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286.
- Zhu, Y. and Li, L. (2018). Multiple matrix gaussian graphs estimation. *Journal of the Royal Statistical Society, Series B.*, 80:927–950.
- Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. *J. Amer. Statist. Assoc.*, 112(517):266–281.