Covariance-based sample selection under heterogenous data for autism risk gene detection

Kevin Lin [*], Han Liu [†], Kathryn Roeder [‡]

March 4, 2018

Abstract

Given that autism spectral disorder (ASD) is a neurological disorder that affects roughly 1-2% of individuals in the United States, it is imperative for biologists to understand the genetic cause of ASD. Previous research in this direction analyze the BrainSpan dataset, which contains microarray gene expression samples from brain tissues from varying brain region and developmental periods. Since the covariance among gene expressions has been shown to vary with respect to the spatiotemporal properties of the brain tissue on average, previous research focused on only samples originating from a particular brain region and developmental period and discarded the remaining samples prior to analyzing the data. While this was done to avoid the

[*]Carnegie Mellon University, Statistics Department, Email: kevinl1@andrew.cmu.edu

[†]Northwestern University, Department of Electrical Engineering and Computer Science, Email: hanliu@northwestern.edu

[‡]Carnegie Mellon University, Statistics Department, Email: roeder@andrew.cmu.edu

issue of heterogeneity, it also leads to potential loss of statistical power when detecting risk genes. In this article, we develop a new method to find a subset of samples that share the same population covariance matrix in order to retain a larger and more homogenous set of samples for the downstream analysis. We apply an existing method on these selected samples to identify genes that are liable for developing ASD, and we see an improvement in the genes we identify.

1 Introduction

The genetic cause of autism spectrum disorder (ASD), a neurodevelopmental disorder that affects roughly 1-2% individuals in the United States, remains an open problem despite decades of research (Autism and Investigators, 2014). ASD is characterized primarily by impaired social functions and repetitive behavior (Kanner et al., 1943; Rutter, 1978). To better understand this disorder, scientists identify specific genes that are liable for increasing the chance of developing ASD when damaged or mutated (Sanders et al., 2015). These are genes are called risk genes. While breakthroughs in genomic technologies and the availability of large ASD cohorts have led to the discovery of dozens of risk genes, preliminary studies suggest there are hundreds of risk genes still unidentified (Buxbaum et al., 2012). In this work, we build upon the current statistical methodologies to further improve our ability to identify risk genes.

We focus on statistical methods that use gene co-expression networks to help identify risk genes. These networks are estimated from microarray expression data from brain tissue. Since these gene co-expression networks provide insight into genes that regulate normal biological mechanisms in fetal and early brain development, it was hypothesized that risk genes that alter these mechanisms should be clustered in these networks (Šestan et al., 2012). Early findings confirmed this hypothesis (Gilman et al., 2011; Parikshak et al., 2013;

Willsey et al., 2013). These results led to the development of the Detection Association With Networks (DAWN) algorithm, which identifies new risk genes based on their connectivity to previously identified risk genes (Liu et al., 2014, 2015). However, the previous DAWN analyses suffer from statistical limitations that we will investigate and resolve in this article.

We challenge DAWN's assumptions regarding the homogeneity of the covariance matrix in microarray expression data. Previous DAWN analyses assume that microarray expression samples from the same brain tissue-type must share the same covariance matrix. This assumption was influenced by the findings in Kang et al. (2011) and Willsey et al. (2013), which showed that gene co-expression patterns differ among different brain regions and developmental periods on average. Statistically, this means that the covariance matrix among the genes' microarray expressions may differ with respect to the spatio-temporal properties of the brain tissue. Despite the findings in Kang et al. (2011) and Willsey et al. (2013) however, no statistical analysis was performed in Liu et al. (2014) or Liu et al. (2015) to check how homogenous the specific samples used in previous DAWN analyses were. Furthermore, since previous DAWN analyses limited themselves to microarray samples of a specific brain tissue-type, many other microarray samples assumed to be heterogeneous are thrown out, leading to a potential loss of power when estimating the gene co-expression network and in identifying risk genes.

To overcome these limitations, we aim to select a subset of microarray expression dataset that is more homogenous and larger in sample size than ones used in previous analyses. We take advantage of the recent developments in high-dimensional covariance testing (Chang et al., 2015a; Cai et al., 2013) to determine whether two microarray expression datasets originating from different brain tissues share the same population covariance matrix. This is paired with a multiple-testing method called Stepdown that accounts for the dependencies among many hypothesis tests (Romano and Wolf, 2005; Chernozhukov et al., 2013). We

show that the tailoring the Stepdown method to perform many covariance tests leads to an improvement in identifying risk genes. This article addresses the numerous algorithmic challenges needed to implement this idea.

In Section 2, we describe the data and statistical model for heterogeneity in the covariance matrix. In Section 3, we provide a visual diagnostic to question the homogeneity assumptions of previous DAWN analyses. In Section 4, we describe the different stages of our procedure to find a subset of homogenous samples within a dataset. In Section 5, we illustrate the properties of our procedure on synthetic datasets. In Section 6, we apply our procedure on microarray expression data to show that, in the end, when combined with DAWN, we identify an improved set of risk genes. Section 7 provides an overall summary and discussion.

2 Data and model background

Datasets recording the gene expression patterns of brain tissue are hard to come by due to the difficulty to obtain and preserve brain tissue. One dataset part of the BrainSpan project (the "BrainSpan dataset" henceforth) contributes one of the largest transcriptome datasets in this direction, sampling tissues from 57 postmortem brains that showed no sign large-scale genomic abnormalities (Kang et al., 2011). Many studies have favored this dataset since its 1,340 microarray samples captures the the spatial and temporal changes in gene expression that occur in the brain during development (De Rubeis et al., 2014; Cotney et al., 2015; Dong et al., 2014).

The heterogeneity of gene expression due to the spatiotemporal differences in brain tissues presents statistical challenges. As documented in detail in Kang et al. (2011), the region and developmental period of the originating brain tissue contribute more to the heterogeneity than other variables such as sex and ethnicity. To understand this heterogeneity, we

partition the dataset using the following schema. Each microarray sample is categorized into one of 16 *spatio-temporal window*, or *window* for short, depending on which brain region and developmental period the brain tissue is derived from. Within each window, all microarray samples originating from the same brain are further categorized into one of 212 *Individual (ID) spatio-temporal partition*, or *partition* for short. Figure 1 summarizes how the 1,340 microarray samples are categorized into different windows and partitions. This figure highlights the importance of Window 1B. Willsey et al. (2013) found that the co-expression among known risk genes varies greatly from window to window, and are most co-expressed within the 107 samples from Window 1B, representing the prefrontal cortex and primary motor-somatosensory cortex from 10 to 19 postconceptual weeks. As a result, previous DAWN analyses focused on only these 107 samples, assuming that these samples from were all homogenous without further statistical investigation, and discarded the remaining 1235 samples, (Liu et al., 2014, 2015). We seek to improve upon this, first by formalizing a statistical model.

2.1 Modeling approach

We now describe a Gaussian mixture model that assumes that samples from the same partition are homogenous while samples from differing partitions could be heterogeneous. For the pth partition, let $\boldsymbol{X}_1^{(p)}, \ldots \boldsymbol{X}_{n_p}^{(p)} \in \mathbb{R}^d$ denote n_p i.i.d. microarray samples, and let $w(p)$ denote the window that partition p resides in. These n_p samples are drawn from either a Gaussian distribution with covariance $\boldsymbol{\Sigma}$, or a Gaussian distribution with a different covariance matrix $\boldsymbol{\Sigma}_p$. Our notation emphasizes that $\boldsymbol{\Sigma}$ is the covariance matrix shared among all partitions, while $\boldsymbol{\Sigma}_p$ may vary from partition to partition. A fixed but unknown parameter $\gamma_{w(p)} \in [0, 1]$ controls how frequently the partitions in window w are drawn from these two distributions, meaning it controls the amount of heterogeneity. For each partition p, this

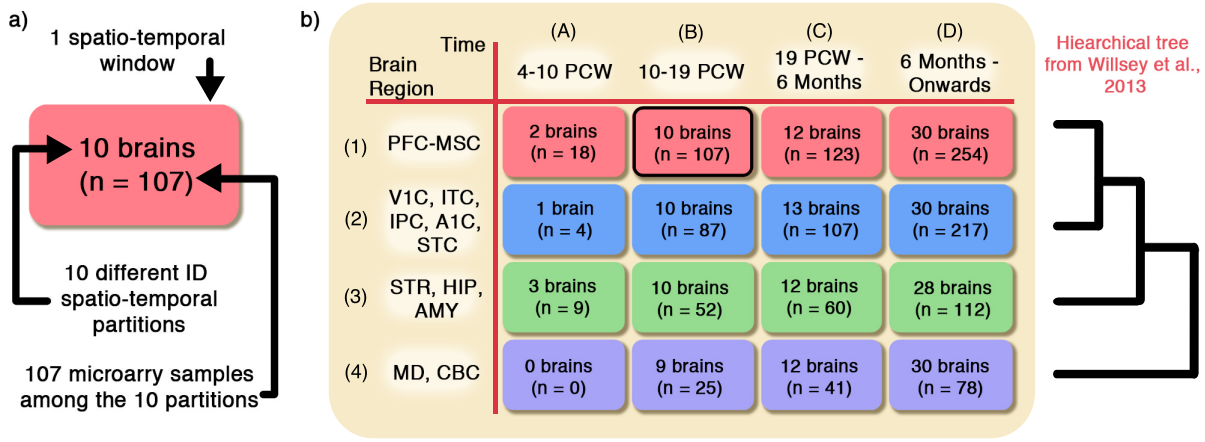

Figure 1: (Left) A schematic exemplifying the relation between the 107 microarray samples grouped by the originating 10 brains. This forms 10 different partitions. Since all these partitions originate from the same brain region and developmental period, they are further grouped into the same window. (Right) The number of partitions (r) and microarray samples (n) in each window (w) for the BrainSpan data. The 57 postmortem brains belong to 4 different developmental periods. Each brain is dissected and sampled at 4 different brain regions, contributing 6 to 12 microarray samples per region. In total, over the 212 partitions, there are 1,340 microarray samples, each measuring the expression of over 14,370 genes. Window 1B (outlined in black) is the window that previous work (Liu et al., 2015) focused on, and the hierarchical tree from Willsey et al. (2013) is shown to the right.

mixture model is succinctly describe as,

$$
\begin{aligned}
I^{(p)} &\sim \text{Bernoulli}(\gamma_{w(p)}), \\
\boldsymbol{X}_1^{(p)}, \ldots, \boldsymbol{X}_{n_p}^{(p)} &\overset{i.i.d.}{\sim} \begin{cases} N(\mathbf{0}, \boldsymbol{\Sigma}) & \text{if } I^{(p)} = 1 \\ N(\mathbf{0}, \boldsymbol{\Sigma}_p) & \text{otherwise,} \end{cases}
\end{aligned} \tag{2.1}
$$

where $I^{(p)}$ is the latent variable that determines whether or not the samples in partition p have covariance $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}_p$. With this model setup, our task becomes determining the set of partitions that originated from the covariance matrix $\boldsymbol{\Sigma}$, which we will call

$$
\mathcal{P} = \left\{ p \ : \ I^{(p)} = 1 \right\}. \tag{2.2}
$$

The findings of Kang et al. (2011) and Willsey et al. (2013) inform us on how much heterogeneity with a window to expect via $\gamma_{w(p)}$. While analyses such as Liu et al. (2015) assumed that all the samples in Window 1B were homogenous, it was noted in Kang et al. (2011) that sampling variability in brain dissection and in the proportion of white and gray matter in different brain tissues can cause variability in the gene expression patterns. This means that scientifically, we do not expect all the partitions in Window 1B to be homogenous (i.e., $\gamma_{w(p)} = 1$). Furthermore, Willsey et al. (2013) found a hierarchical clustering among the four brain regions. This is illustrated in Figure 1, where the gene expression patterns in the brain regions represented in first row are most similar to those in the second row and least similar to those in the fourth row. The authors also found a smooth continuum of gene expression patterns across different developmental periods, represented as the columns of the table in Figure 1. Hence, we expect $\gamma_{w(p)}$ to decrease smoothly as the window w becomes more dissimilar to Window 1B, in both the spatial and temporal direction.

2.2 Connections

Other works have used models similar to (2.1) on microarray expression data to tackle the different co-expression patterns between different tissues and subjects, but their methods differ from ours. One direction is to directly cluster the covariance matrices of each partition (Ieva et al., 2016). However, this approach does not factor in the variability in the empirical covariance matrix, unlike our hypothesis-testing based method. Another approach is to explicitly model the population covariance matrix for each partition as the summation of a shared component and a partition-specific heterogenous component. This is commonly used in batch-correction procedures where the analysis tries to remove the heterogeneous component from each partition (Leek and Storey, 2007). However, we feel such an additive model is too restrictive for analyzing the BrainSpan dataset, as we do not believe there is a shared covariance matrix across all regions of the brain. Instead, our approach will find specific set of partitions with statistically indistinguishable covariance matrices.

3 Elementary analysis

In this section, we develop a visual diagnostic to investigate if the 107 samples used in previous works (Liu et al., 2014, 2015) are as homogeneous as these previous analyses assumed. Using a hypothesis test for equal covariances, our diagnostic leverages the following idea: We divide the samples among all partitions into two groups and apply the hypothesis test to the samples between both groups. If all the partitions were truly drawn from distributions with equal covariances, then over many possible divisions, a QQ-plot of the resulting p-values should look roughly uniform. The less uniform the p-values look, the less we are inclined to interpret our partitions were all drawn from distributions with equal covariances.

Algorithm 1: Covariance homogeneity diagnostic

1. Loop over trials $t = 1, 2, \ldots, T$:

   (a) Randomly divide the selected partitions in the set $\widehat{\mathcal{P}}$ into two sets, $\widehat{\mathcal{P}}^{(1)}$ and $\widehat{\mathcal{P}}^{(2)}$, such that $\widehat{\mathcal{P}}^{(1)} \cup \widehat{\mathcal{P}}^{(2)} = \widehat{\mathcal{P}}$ and $\widehat{\mathcal{P}}^{(1)} \cap \widehat{\mathcal{P}}^{(2)} = \emptyset$.

   (b) For each partition $p \in \widehat{\mathcal{P}}^{(1)}$, center the samples $\boldsymbol{X}_1^{(p)}, \ldots, \boldsymbol{X}_{n_p}^{(p)}$. Then aggregate all samples in $\widehat{\mathcal{P}}^{(1)}$ to form the set of samples

$$\mathcal{X} = \left\{ \boldsymbol{X}_1^{(p)}, \ldots, \boldsymbol{X}_{n_p}^{(p)} \ : \ p \in \mathcal{P}^{(1)} \right\}.$$

   Similarly, form the set of samples $\mathcal{Y}$ from the set of partitions $\mathcal{P}^{(2)}$.

   (c) Compute the p-value for the hypothesis test that tests whether or not the samples in $\mathcal{X}$ and $\mathcal{Y}$ have the same covariance matrix.

2. Plot the QQ-plot of the resulting T p-values to see if empirical distribution of the p-values is close to a uniform distribution.

We remind the reader that the above procedure is a diagnostic, not necessarily a recipe for a goodness-of-fit test. This is because the T p-values are not independent, so it is difficult to analyze the theoretical properties of this diagnostic. However, as we will demonstrate in later sections of this article, this diagnostic is nonetheless able to display large-scale patterns in our dataset.

3.1 Specification of covariance hypothesis test

To complete the above diagnostic's description, we describe the procedure to test equality of covariance matrice. Let $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}\}$ and $\mathcal{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}\}$ be n_1 and n_2 i.i.d. samples from d-dimensional distributions with covariance $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ respectively, both with an empirical mean of $\mathbf{0}$. We define $\mathbb{X} \in \mathbb{R}^{n_1 \times d}$ and $\mathbb{Y} \in \mathbb{R}^{n_2 \times d}$ as the matrices

formed by concatenating these samples row-wise. Define empirical covariance matrices as $\widehat{\mathbf{\Sigma}}_X = \mathbb{X}^\top \mathbb{X}/n_1$, and $\widehat{\mathbf{\Sigma}}_Y = \mathbb{Y}^\top \mathbb{Y}/n_2$, where we denote individual elements of these matrices as $\widehat{\mathbf{\Sigma}}_X = [\widehat{\sigma}_{X,ij}]_{1\le i,j\le d}$ and likewise for $\widehat{\mathbf{\Sigma}}_Y$. We now discuss two possible hypothesis tests for equal covariance, $H_0 : \mathbf{\Sigma}_X = \mathbf{\Sigma}_Y$ that we consider in this article.

Method 1 (With normalization): The first method defines the test statistic according to Chang et al. (2015a) which extends Cai et al. (2013). In these works, the authors note that if $\mathbf{\Sigma}_X = \mathbf{\Sigma}_Y$, then the maximum element-wise difference between $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ is 0. Hence, Chang et al. (2015a) defines the test statistic as the maximum of element-wise differences squared between $\widehat{\mathbf{\Sigma}}_X$ and $\widehat{\mathbf{\Sigma}}_Y$, each normalized by its variance. Specifically, the test statistic is

$$\widehat{T} = \max_{ij} \left(t_{ij}\right) \quad \text{where } \widehat{t}_{ij} = \frac{\left(\widehat{\sigma}_{X,ij} - \widehat{\sigma}_{Y,ij}\right)^2}{\widehat{s}_{X,ij}/n_1 + \widehat{s}_{Y,ij}/n_2}, \quad i,j \in 1,\ldots,d, \tag{3.1}$$

where $\widehat{s}_{X,ij} = \sum_{m=1}^{n_1} (\mathbb{X}_{mi}\mathbb{X}_{mj} - \widehat{\sigma}_{X,ij})^2/n_1$ is the empirical variance of the variance-estimator $\widehat{\sigma}_{X,ij}$, and $\widehat{s}_{Y,ij}$ is defined similarly.

Then, Chang et al. (2015a) constructs an empirical null distribution of $\widehat{T}$ under H_0 : $\mathbf{\Sigma}_X = \mathbf{\Sigma}_Y$ using the multiplier bootstrap (Chernozhukov et al., 2013). On each of the $b \in \{1,\ldots,B\}$ trials, the multiplier bootstrap computes a bootstrapped test statistic $\widehat{T}^{(b)}$ by weighting each of the $n_1 + n_2$ observations by a standard Gaussian random variable drawn independently of all other variables, denoted collectively as $(g_1^{(b)},\ldots,g_{n_1}^{(b)},g_{n_1+1}^{(b)},\ldots,g_{n_1+n_2}^{(b)})$. Specifically, we construct the bootstrap statistic for the bth trial as

$$\widehat{T}^{(b)} = \max_{ij} \left(\widehat{t}_{ij}^{(b)}\right) \quad \text{where } \widehat{t}_{ij}^{(b)} = \frac{\left(\widehat{\sigma}_{X,ij}^{(b)} - \widehat{\sigma}_{Y,ij}^{(b)}\right)^2}{\widehat{s}_{X,ij}/n_1 + \widehat{s}_{Y,ij}/n_2}, \quad i,j \in 1,\ldots,d, \tag{3.2}$$

where $\widehat{\sigma}_{X,ij}^{(b)} = \sum_{m=1}^{n_1} g_m^{(b)}(\mathbb{X}_{mi}\mathbb{X}_{mj} - \widehat{\sigma}_{X,ij})/n_1$ and $\widehat{\sigma}_{Y,ij}^{(b)} = \sum_{m=1}^{n_2} g_{n+m}^{(b)}(\mathbb{Y}_{mi}\mathbb{Y}_{mj} - \widehat{\sigma}_{Y,ij})/n_2$. We compute the p-value by counting the proportion of bootstrap statistics are larger than the

test statistic, i.e.,

$$\text{p-value} = \frac{\text{Cardinality}\big(\{b : |\widehat{T}^{(b)}| \geq |\widehat{T}|\}\big)}{B}.$$

Chang et al. (2015b) proves that this test has asymptotically $1 - \alpha$ coverage under the null hypothesis for distributions with sub-Gaussian and sub-exponential tails, even in the high-dimensional regime where $d \gg \max(n_1, n_2)$.

Method 2 (Without normalization): The second method is similar to the first, except we replace the denominators shown in (3.1) and (3.2) with 1, meaning we do not normalize the element-wise squared difference between the two covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ by its variance. While Chang et al. (2015a) do not originally consider this formulation, as we will see in Subsection 4.2, this modification offers practical computational advantages for our partition selection procedure. Specifically,

$$\widehat{T} = \max_{ij} \big(t_{ij}\big) \quad \text{where } \widehat{t}_{ij} = \big(\widehat{\sigma}_{X,ij} - \widehat{\sigma}_{Y,ij}\big)^2, \quad i, j \in 1, \ldots, d, \tag{3.3}$$

and we make a similar modification for $\widehat{T}^{(b)}$. While this method will still yield a valid hypothesis test, we will see later on that the lack of normalizing the element-wise differences results in a less powerful test.

3.2 Application to BrainSpan

Equipped with a complete description of the diagnostic, we apply it to the BrainSpan dataset. Among the 10 partitions in the PFC-MSC 10-19 PCW window, we divide the partitions into two groups in 500 uniformly randomly chosen ways, and compute a p-value using Method 1 (with normalization) for each division. The QQ-plot of the resulting p-values are shown in Figure 2 (left), where we see that the empirical distribution of the p-values is right-skewed. Furthermore, we apply this diagnostic to all 125 partitions in the BrainSpan dataset, and we

see that the distribution of the p-values become even more right-skewed. This suggests that the 125 partitions in the BrainSpan dataset, specifically the 10 partitions in Window 1B, do not seem to all share the same covariance matrix, implying heterogeneity in the dataset. In the next section, we develop a method to resolve this issue by finding the largest subset of partitions possible among the 125 partitions in the BrainSpan dataset that share the same covariance matrix.

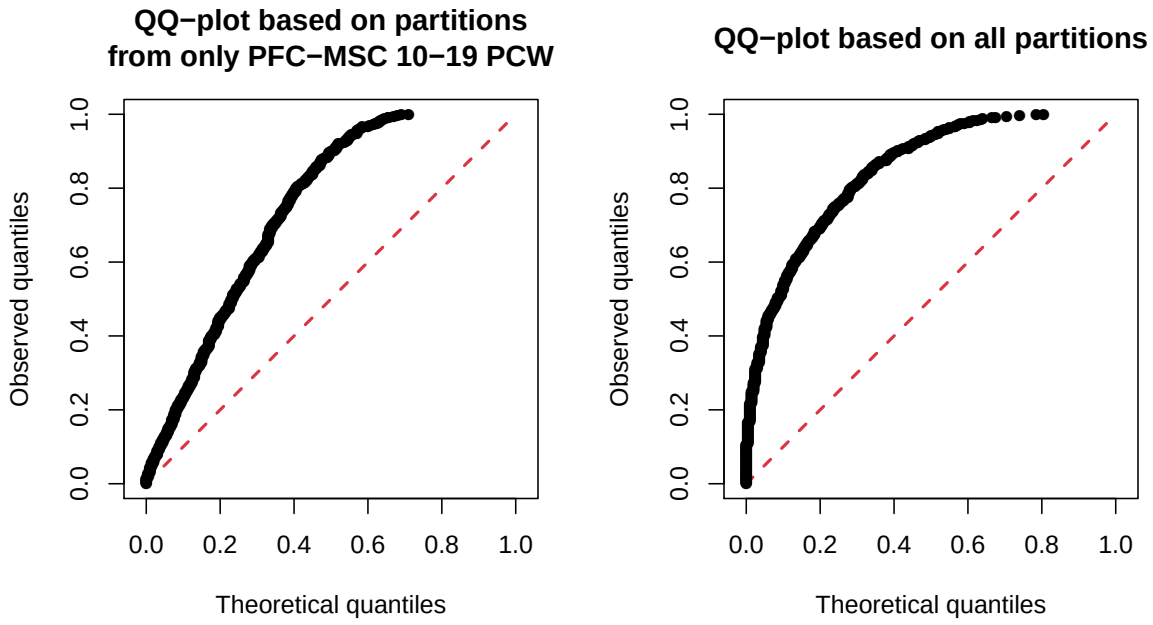

Figure 2: QQ-plots of the 500 p-values generated when applying our diagnostic to the BrainSpan dataset. Left: Using only the partitions in the PFC-MSC 10-19 PCW window. Right: Using all 125 partitions in the BrainSpan dataset.

4 Methods

While we have discussed methods to test for equivalent covariance matrices between any two partitions in Section 3, we cannot directly apply these methods to the BrainSpan dataset without suffering a loss of power. This because given there are $r = 125$ partitions, applying

the hypothesis test to each pair of partitions results in $\binom{r}{2} = 7,750$ dependent p-values. The p-values are dependent since each of the r partitions is involved in $r-1$ hypothesis tests. Hence, applying standard techniques such as a Bonferroni correction would ignore the dependencies among all the p-values, likely leading to a loss of power.

To preserve this dependency, we introduce our Stepdown procedure in Subsection 4.1 that simultaneously tests all $\binom{r}{2}$ hypothesis tests. This bootstrap-based procedure is computationally expensive. However, depending on the test statistic used in the $\binom{r}{2}$ hypothesis tests, we offer a computationally faster alternative in Subsection 4.2. Afterward determining which of the $\binom{r}{2}$ pairs of partitions do not have statistically significant differences in their covariance matrices, we develop a clique-based procedure in Subsection 4.3 to select a specific set of partitions $\widehat{\mathcal{P}}$.

4.1 Stepdown procedure: multiple testing with dependence

We use a Stepdown procedure developed in Chernozhukov et al. (2013) to control the family-wise error rate. We tailor the bootstrap-based procedure to our specific setting in the algorithm below. We denote $\widehat{T}_{(i,j)}$ as the test statistic formed using either of the two methods, (3.1) or (3.3), to test if the covariance of samples between those in partition i and partition j are equal, and $\widehat{T}^{(b)}_{(i,j)}$ is the corresponding bootstrap statistics on the bth bootstrap trial.

Algorithm 2: Stepdown procedure

1. Initialize the list enumerating all $\binom{r}{2}$ null hypotheses corresponding to the partition pairs
$$\mathcal{L}(1) = \Big\{(1,2), \ldots, (r-1, r)\Big\}.$$

2. Loop over steps $t = 1, 2, \ldots$

(a) Calculate $\widehat{T}_\ell$ for all $\ell \in \mathcal{L}(t)$, as stated in (3.1).

(b) For each bootstrap trial $b = 1, \ldots, B$:

i. Generate $N = \sum_p n_p$ i.i.d. Gaussian random variables, one for each sample in every partition, and compute $\widehat{T}_\ell^{(b)}$ for all $\ell \in \mathcal{L}(t)$, as stated in (3.2).

ii. Compute

$$\widehat{T}^{(b)} = \max\Big\{\widehat{T}_\ell^{(b)} : \ell \in \mathcal{L}(t)\Big\}. \tag{4.1}$$

(c) Remove any $\ell \in \mathcal{L}(t)$ if

$$\widehat{T}_\ell \geq \text{quantile}\Big(\{\widehat{T}^{(1)}, \ldots, \widehat{T}^{(b)}\}; 1 - \alpha\Big).$$

If not elements are removed from $\mathcal{L}(t)$, return the null hypotheses corresponding to $\mathcal{L}(t)$. Otherwise, continue to step $t + 1$.

Using techniques in Romano and Wolf (2005) and Chernozhukov et al. (2013), it can be shown that this algorithmic extension of the covariance test in Chang et al. (2015a) has the following familywise error guarantee,

$$\mathbb{P}\Big(\text{no true null hypothesis among } \mathcal{H} \text{ null hypotheses are rejected}\Big) \geq 1 - \alpha + o(1). \tag{4.2}$$

The reason the stepdown procedure is able to control the familywise error without using Bonferroni is because the $\binom{r}{2}$ bootstrapped statistics in each trial are derived the from same N Gaussian random variables, hence preserving the dependencies among the $\binom{r}{2}$ tests.

4.2 Computational extension for stepdown procedure

The largest drawback to the stepdown procedure as described above lies in its intensive computational cost. If the number of partitions r is large, then $\binom{r}{2}$ bootstrap statistics need

to be computed in each bootstrap trial. In this subsection, we reduce the computational cost by leveraging properties of the bootstrap statistics $\widehat{T}^{(b)}_{(i,j)}$'s.

Specifically, we consider only test statistics that satisfy the triangle inequality between datasets. In our context, a test statistic between datasets satisfies the triangle inequality if for any bootstrap trial b and for any partitions i, j and k,

$$\widehat{T}^{(b)}_{(i,k)} \leq \widehat{T}^{(b)}_{(i,j)} + \widehat{T}^{(b)}_{(j,k)}. \tag{4.3}$$

This property can potentially save expensive calculations when calculating (4.1) in Algorithm 2. Since we only care about the maximum bootstrap statistic $\widehat{T}^{(b)}$, the triangle inequality gives an upper bound on the bootstrap statistic $\widehat{T}^{(b)}_{(i,k)}$ between partitions i and k that leverages bootstrap statistics already calculated within a specific bootstrap trial. If this upper bound is smaller than the current maximum bootstrap statistic in a specific bootstrap trial, then we do not need to explicitly compute $\widehat{T}^{(b)}_{(i,k)}$.

One way to ensure that the bootstrap statistics $\widehat{T}^{(b)}_{(i,j)}$'s satisfy the triangle inequality is to ensure that the statistic is a distance metric between partitions, meaning in addition to (4.3), we require that $\widehat{T}^{(b)}_{(i,j)} \geq 0$ and $\widehat{T}^{(b)}_{(i,i)} = 0$ for partitions i and j. This is why we consider non-normalized test statistic (3.3) in Section 3. The normalized test statistic shown in (3.1), while resulting in more powerful tests, can not use this computational extension.

We describe the we developed algorithm below, which represents the bootstrap statistics as weighted edges in a graph. The algorithm uses Dijsktra's algorithm to find the shortest path between vertices. This implicitly computes the upper-bound in the bootstrap statistic between two partitions using the triangle inequality. This algorithm can provide substantial improvement in computational speed by leveraging the fact that determining the shortest path on a fully-dense graph has a computational complexity of $O(r^2)$, whereas computing

$T^{(b)}_{(i,j)}$ has a computational cost of $O(d^2 \cdot \max(n_1, n_2))$.

Algorithm 3: Distance metric-based procedure to compute $\widehat{T}^{(b)}$

1. Form graph $G = (V, E)$ with r nodes and all $\binom{r}{2}$ edges, and initialize each edge to have weight equal to (positive) infinity.

2. Arbitrarily construct a spanning tree $\mathcal{T}$ and compute all $\widehat{T}^{(b)}_{(i,j)}$ corresponding to edges $(i, j) \in \mathcal{T}$. Record $z = \max_{(i,j)\in\mathcal{T}} \widehat{T}^{(b)}_{(i,j)}$.

3. Construct a set of edges $\mathcal{S} = \mathcal{L}(t)\backslash\mathcal{T}$ which represents the bootstrap statistics between specific pairs of partitions that have yet to be computed.

4. While $\mathcal{S}$ is not empty:

   (a) Arbitrarily select an edge $(i, j) \in \mathcal{S}$ and remove it from $\mathcal{S}$. Compute the shortest-path distance from vertex i to j in G.

   (b) If the shortest-path distance is larger than z, update the edge (i, j) to have weight $\widehat{T}^{(b)}_{(i,j)}$, and update z to be $\max(z, \widehat{T}^{(b)}_{(i,j)})$.

5. Return z.

4.3 Largest partial clique: selecting partitions based on testing results

After applying the covariance testing with the stepdown procedure described in the previous two subsections, we have a subset of null hypotheses from $\mathcal{H}$ that we accept. In this subsection, we develop a clique-based method to estimate $\mathcal{P}$ defined in (2.2), the subset of partitions that share the same covariance matrix, from $\mathcal{H}$.

We conceptualize the task of selecting partitions as selecting vertices from a graph. Let $G = (V, E)$ be a graph with vertices V and edge set E such that

$$V = \{1, \ldots, r\}, \quad E = \{(i, j) : H_{0,(i,j)} \text{ is accepted by the stepdown procedure}\}. \tag{4.4}$$

Since each of the $\binom{|\mathcal{P}|}{2}$ pairwise tests among the partitions in $\mathcal{P}$ satisfy the null hypotheses, if none of the null hypotheses were incorrectly rejected, then the vertices corresponding to $\mathcal{P}$ will form a clique in graph G. However, due to the theoretical guarantee stated in (4.2), this event only occurs with limiting probability $1-\alpha$. This means the vertices corresponding to $\mathcal{P}$ have a non-zero probability to be missing edges to form a clique in graph G. Hence, loosely speaking, the goal becomes to select a subset of vertices in G that are highly connected. This task is exemplified in Figure 3.

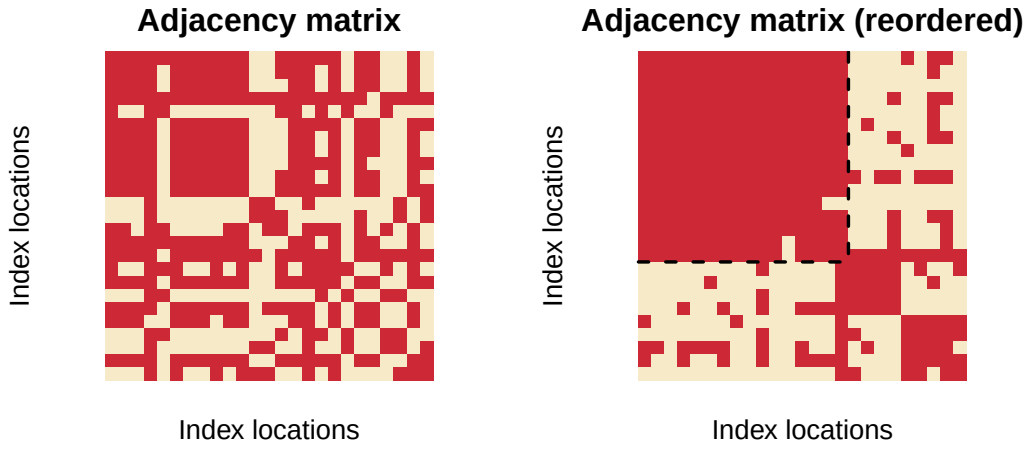

Figure 3: (Left) Visualization of an (example) adjacency matrix that can be formed using Equation (4.4), where the ith row from top and column from the left denotes the ith vertex. A red square in position (i, j) denotes an edge between vertex i and j. (Right) Illustration of the desired goal. The rows and columns were reordered from the left figure, and the dotted box denotes the vertices that were found to be highly connected.

There are many algorithms in statistics in computer science that can achieve this goal, but many such algorithms in practice suffer for a lack of monotonicity. Specifically, suppose

we have an algorithm $\mathcal{A}$ that takes in a graph G and outputs a set of partition, denoted by $\mathcal{A}(G)$, and for two graphs G' and G, let $G' \subseteq G$ denote that every edge in G' is in G. Since we are trying to select a subset of highly connected vertices in G, it would natural to have the following property:

$$G' \subseteq G \quad \Rightarrow \quad |\mathcal{A}(G')| \leq |\mathcal{A}(G)|. \tag{4.5}$$

This property is intuitive for whichever algorithm we use since in our context, intuitively, less partitions should be selected if our stepdown procedure deems more pairs of partitions to have statistically significant different covariance matrices. More importantly however, this property is critical in practice since the choice of α used by the stepdown procedure in Subsection 4.1 is decided by the user. Using the algorithmic description laid out in Subsection 4.1, it can be shown that

$$\alpha' \leq \alpha \quad \Rightarrow \quad G' \subseteq G.$$

Hence, given the above relationship, an algorithm that does not exhibit the property in (4.5) will be fragile as a larger α, meaning a stricter test, could counter-intuitively result in more partitions being selected. As we will demonstrate in Section 5 in simulation and continue in the appendix, many popular algorithms such a spectral clustering do not exhibit this property. Therefore, we develop a new algorithm that empirically exhibits the property (4.5).

Our algorithm finds the largest partial clique in the graph formed by (4.4). We say a set of k vertices form an γ-partial clique if there are at least $\gamma \cdot \binom{k}{2}$ edges among these k vertices. A largest γ-partial clique is the largest set of vertices that form a γ-partial clique. We justify the choice to search for the largest γ-partial clique since, by construction of our model in (2.1), the prevalent covariance structure among the r partitions is the desired covariance structure we wish to estimate.

We describe the algorithm we developed below. It starts by finding a list containing all maximal cliques in the graph based on (4.4). A maximal clique is a vertex set that form a clique but is not subset of a larger clique. The algorithm then proceeds by determining if the union of any two vertex sets form a γ-partial clique. If so, this union of vertices is added to the list of vertex sets. The algorithm returns the largest vertex set discovered when all pairs of vertex sets are tried and no new γ-partial clique is found. We demonstrate in Section 5 that this algorithm exhibits the monotonicity property (4.5).

Algorithm 4: Clique-based selection

1. Form graph G based on Equation (4.4).
2. Form $\mathcal{C}$, the set of all vertex sets that form a maximal clique in G.
3. While there are vertex sets $C_i, C_j \in \mathcal{C}$ the algorithm has not tried yet:
   (a) Determine if the union of vertices in C_i and C_j form a γ-partial clique in G. If so, add the union of vertices in C_i and C_j as a new vertex set in $\mathcal{C}$.
4. Return the largest vertex set in $\mathcal{C}$.

While a naive implementation of the above algorithm would require exponential time to complete in terms of r, by using a queue and three hash tables, our implementation dramatically reduces the computational cost. Mainly, one hash table is used to record all vertex sets tried on whether or not they form a γ-partial clique. This allows our implementation to exploit previous calculations using the following heuristic: we only check if a vertex set forms a γ-partial clique, if for some partitioning of the vertex set, at least one partition of vertices should form a a γ-partial clique. This idea is illustrated in Figure 4.

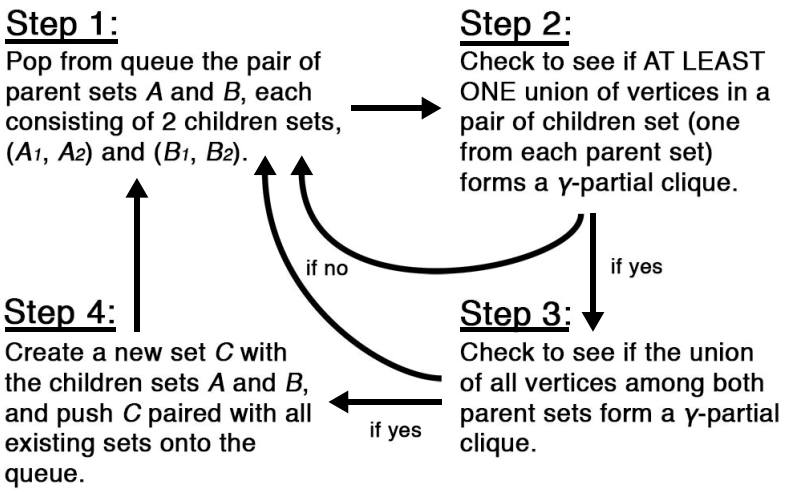

Figure 4: Schematic of Algorithm 4. Step 2 is able to leverage hash tables by checking previous calculations if the union of vertices in a pair of children sets forms a γ-partial clique, which takes near-constant calculations to access. This can save tremendous computation time since Step 3, which checks if the union of vertices in both parent sets form a γ-partial clique, takes quadratic time in the number of vertices to compute.

5 Simulation study

We perform empirical studies to show that our methods in Section 4 have more power and yield better estimation of the desired covariance matrix $\mathbf{\Sigma}$ over conventional methods as the samples among different partitions are drawn from increasingly different distributions.

Setup: We generate synthetic data in different partitions, where the data in each partition has $n = 25$ samples and $d = 50$ dimensions, using the following schema: we construct $\mathbf{\Sigma} = [\Sigma_{i,j}]_{i,j=1}^{d}$ by $\Sigma_{i,j} = i \cdot (d - j + 1)/(d + 1)$. We generate the first $r_1 = 15$ partitions, where each partition consists of n i.i.d. samples drawn from $N(0, \mathbf{\Sigma})$. For a fixed parameter $\beta \in (0, 1)$ which represents the flip percentage, we generate $\mathbf{\Sigma}'$ by uniformly randomly shuffling $(\beta \cdot 100)\%$ of the rows and their corresponding columns of $\mathbf{\Sigma}$. We then generate the next $r_2 = 5$ partitions, where each partition consists of n i.i.d. samples drawn from $N(0, \mathbf{\Sigma}')$. Lastly, we generate the last $r_3 = 5$ partitions in the same fashion, by generating

$\boldsymbol{\Sigma}''$ by shuffling $(\beta \cdot 100)\%$ of the rows and their corresponding columns of $\boldsymbol{\Sigma}$. Hence, we have a total of $r = r_1 + r_2 + r_3 = 25$ partitions where we desired goal is to find that the first r_1 partitions share the same covariance structure. In this simulation study, β parameterizes how "easy" this task is, as a larger β means the hypothesis test described in Section 3 has more power in distinguishing among samples drawn from the three covariance matrices $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}'$ and $\boldsymbol{\Sigma}''$. Figure 5 visualizes the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$.

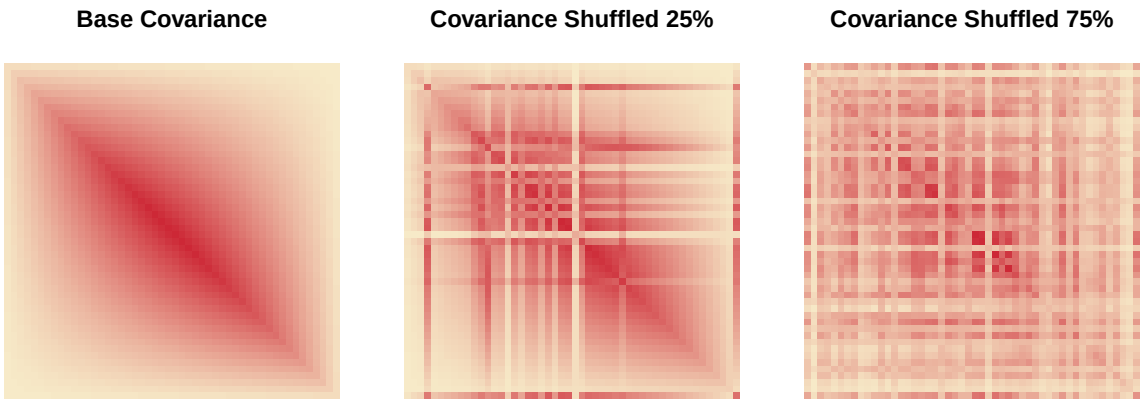

Figure 5: (Left) Visualization of the covariance matrix $\boldsymbol{\Sigma}$. (Middle) Visualization of one possible covariance matrix $\boldsymbol{\Sigma}'$, generated by swapping $\beta = 0.25$ fraction of the rows and respective columns of $\boldsymbol{\Sigma}$. (Right) Analogous to the middle plot, but swapping $\beta = 0.75$ fraction of the rows and respective columns of $\boldsymbol{\Sigma}$. As the swap percentage increases, the difference between covariance matrices becomes more apparent.

Multiple testing: We use the stepdown procedure described in Subsection 4.1 and Subsection 4.2 on our simulated data where $\beta = \{0, 0.1, 0.25, 0.75\}$ to see what how true positive rates and false positive rates vary with β. Let $\mathcal{L} = \{(i_1, j_1), (i_2, j_2), \ldots\}$ denote the set of partition pairs that correspond to the accepted null hypothesis. Since our goal is to find the first r_1 partitions, we define the true positive rate to be

$$\text{True positive rate for hypothesis} = \frac{\left|\left\{(i,j) \in \mathcal{L} : i \leq r_1 \text{ and } j \leq r_1\right\}\right|}{\binom{r_1}{2}}.$$

Similarly, we define the false positive rate to be

$$\text{False positive rate for hypothesis} = \frac{\left|\left\{(i,j) \in \mathcal{L} : i > r_1 \text{ or } j > r_1\right\}\right|}{\binom{r}{2} - \binom{r_1}{2}}.$$

We plot the RoC curves visualizing the true and false positive rates in Figure 6. Each curve traces out the median true and false positive rate over 20 simulations as α ranges from 0 (top-right of each plot) to 1 (bottom-left of each plot), where we use 1000 bootstrap trials per simulation. In all three plots, we see that as the flip percentage β increases, each method has more power. The left plot of Figure 6 represents the analysis that did not use the methods we develop in this article. There, we compute each $\binom{r}{2}$ p-values, one for each hypothesis test comparing two partitions, and accept hypotheses for varying levels of α after a Bonferroni correction. The right plot shows the curves for the stepdown procedure using the normalized statistic (3.1). As we mentioned in Subsection 4.1, there is a considerable loss of power from the stepdown procedure to the naive family-wise correction since the Bonferroni correction does not account for dependencies among hypothesis tests, there is a considerable loss of power.

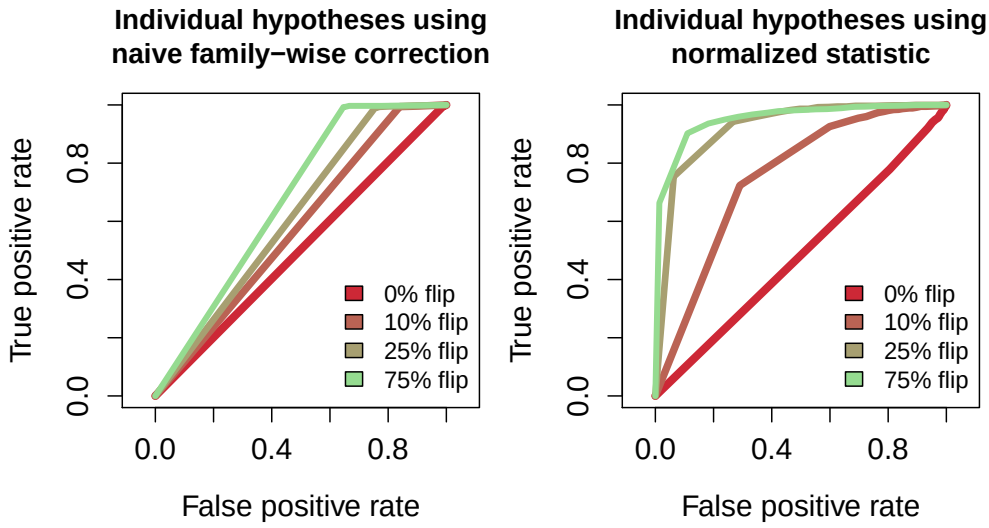

Figure 6: RoC curves for the accepted hypotheses, for settings where $\beta = (0, 0.1, 0.25, 0.75)$, where each curve traces out the results as α varies from 0 to 1. (Left) The curves resulting from using a Bonferroni correction to the $\binom{r}{2}$ individual hypothesis tests. (Right) The curves resulting from using the stepdown procedure with the normalized statistic (3.1).

Partition selection: Using the stepdown procedure using the normalized statistic (3.1), we proceed to select the partitions as in Subsection 4.3 to understand the monotonicity property and see how the true and false positive rates for partitions vary with the flip percentage β.

The left figure of Figure 7 shows how the certain methods to find highly connected vertices (4.4) fail the monotonicity property (4.5). Here, we compare our largest partial clique method, described in Subsection 4.3, against spectral clustering, a method used in network analyses designed to find highly connected vertices in degree-corrected stochastic block models (Lei and Zhu, 2017). Both methods are applied to the same simulated dataset and receive the same set of accepted hypotheses as the family-wise error rate α varies. Recall that since the stepdown procedure accepts more hypotheses as α decreases, the graph formed by (4.4) becomes denser as α decreases. However, as we see in left figure of Figure 7, the number of partitions selected by spectral clustering sometimes decreases as number of

accepted hypotheses increases, hence violating the desired monotonicity property. On the other hand, we see that our largest partial clique method works empirically satisfies the monotonicity property. Here, we set our algorithm to find the largest 0.95-partial clique, and it empirically satisfied the monotonicity property in the simulation suite.

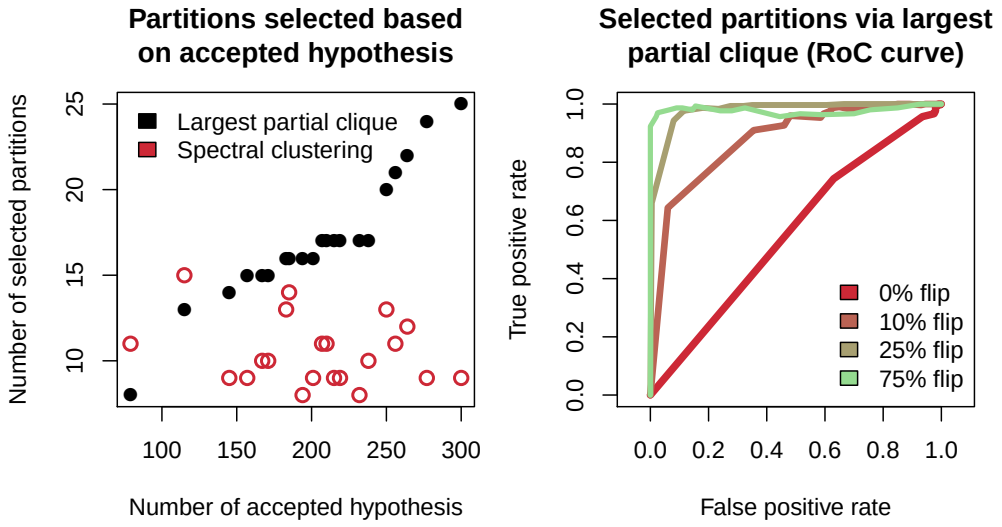

Figure 7: (Left) Number of selected partitions for a particular simulated dataset as the number of accepted hypotheses varies with the family-wise error rate α. (Right) Similar RoC curves to Figure 6, but for selected partitions after using the stepdown procedure with the normalized test statistic.

The right figure of Figure 7 shows the RoC curves for varying β as the family-wise error rate α varied during multiple testing. This figure is closely related to the middle plot of Figure 6. We use our largest partial clique method to find the largest 0.95-partial clique. Let $\widehat{\mathcal{P}}$ denote the selected set of partitions. Similar to before, we define the true and false positive rate in this setting as

$$\text{True positive rate for partitions} = \frac{\left|\left\{p \in \widehat{\mathcal{P}} : p \leq r_1\right\}\right|}{r_1},$$

$$\text{False positive rate for partitions} = \frac{\left|\left\{p \in \widehat{\mathcal{P}} : p > r_1\right\}\right|}{r_2 + r_3}.$$

We see that the power of the largest partial clique method increases as β increases, as expected.

Covariance estimation: Finally, we show that our method is able to improve the downstream covariance estimation compared to other approaches. To do this, we use four different methods to select partitions and compute the empirical covariance matrix among the samples in those partitions. The first three methods resemble analyses that could be performed on BrainSpan in practice. The first method is the method we develop with $\alpha = 0.7$. The second method always selects all the partitions, which resembles using all the partitions in the BrainSpan dataset. The third method always selects the same 5 partitions. These 5 partitions are fixed so the 3 partitions contains samples drawn from $N(0, \mathbf{\Sigma})$, while the other 2 partitions contain samples from each of the remaining two distributions. This resembles past work (Liu et al., 2015) that considered only partitions in Window 1B. For comparison, the last method resembles an oracle by selecting exactly the k_1 partitions contain samples drawn from $N(0, \mathbf{\Sigma})$.

Figure 8 shows that our partition selection method with $\alpha = 0.7$ performs almost as well as the oracle method over varying flip percentages β. This figure shows the average spectral error of the estimated covariance matrix for each method and flip percentage over 10 trials. Notice that for low β, our method (using a fixed α) and the method using all partitions yield a smaller spectral error than the method that knows exactly which samples are drawn from drawn from $N(0, \mathbf{\Sigma})$. This is because for low β, the covariance matrices $\mathbf{\Sigma}$, $\mathbf{\Sigma}'$, and $\mathbf{\Sigma}''$ are almost indistinguishable. However, as β increases, the differences among $\mathbf{\Sigma}$, $\mathbf{\Sigma}'$, and $\mathbf{\Sigma}''$ grows. This means methods that do not adaptively choose which partitions to select become increasingly worse. However, our method using $\alpha = 0.7$ remains competitive, performing almost as if it knew which partitions contain samples drawn from $N(0, \mathbf{\Sigma})$.

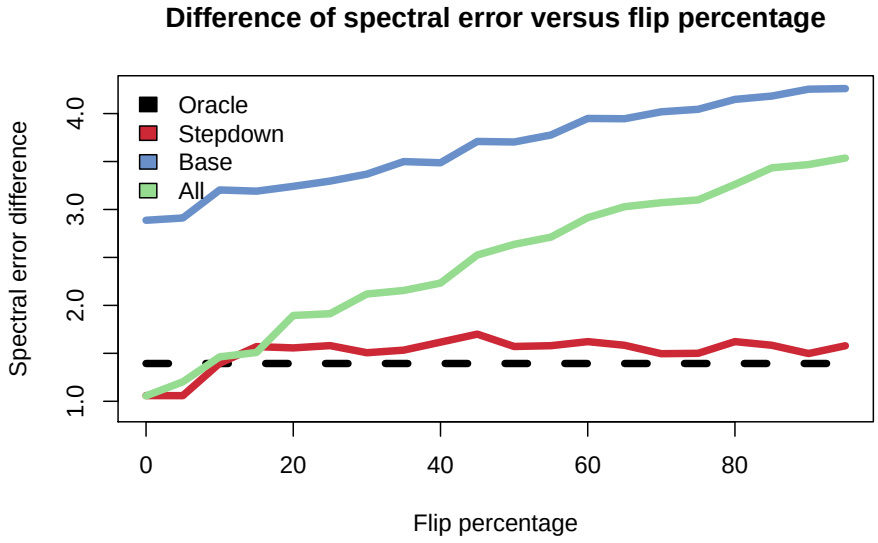

Figure 8: The average spectral error of each method's downstream estimated covariance matrix for varying flip percentage β. The four methods to select partitions shown are our method for $\alpha = 0.7$ (red), the method that selects all partitions (green), the method that selects a fixed set of 5 partitions (blue), and the method that selects exactly the partitions that contain samples drawn from $N(0, \mathbf{\Sigma})$ (black).

6 Application on BrainSpan study

6.1 Partition selection

We apply our entire selection procedure using the normalized statistic (3.1) on the BrainSpan dataset and find partitions that matches the scientific intuition described in Willsey et al. (2013) and obtain a better diagnostic compared to the ones performed in Section 3. To demonstrate this, we select partitions based on only the 200 genes with the largest risk score according to an external dataset (De Rubeis et al., 2014). Using 1000 bootstrap trials and familywise error level $\alpha = 0.1$, we use the stepdown procedure to find which null hypotheses are accepted among the $\binom{125}{2}$ hypotheses tested simultaneously. Based on these results, we

use our clique-based selection method to select the partitions that form the maximal 0.95-partial clique. To break ties between maximal partial cliques, we use the clique with the most partitions in Window 1B.

We visualize the results of the stepdown procedure in Figure 9 to illustrate that our method finds 43 partitions which do not have significantly different covariance matrices. Since each null hypothesis corresponds to a pair of partitions, we form the graph G connecting pairs of partitions corresponding to the accepted null hypotheses, as described in (4.4). The left figure in Figure 9 shows a subgraph of G as an adjacency matrix, while the right figure shows the graph with all 125 nodes. The nodes in this graph are laid out using a standard layout algorithm so highly connected sets of nodes are placed compactly (Fruchterman and Reingold, 1991). Hence, we can see that the 43 partitions we select correspond to 43 nodes in G that are highly connected.

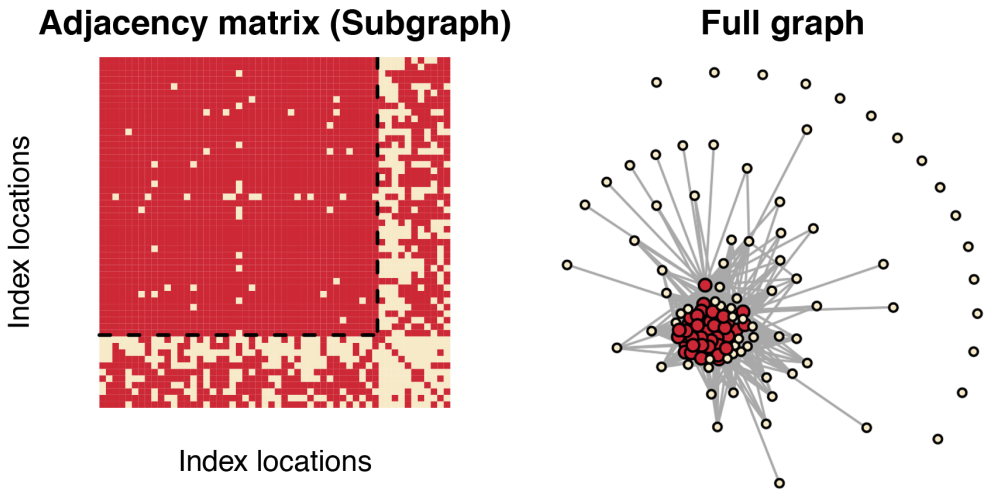

Figure 9: (Left) The adjacency matrix of a subgraph of G, where each row and corresponding column represents a different node, similar to Figure 3. A red pixel corresponds to an edge between two nodes, while a pale pixel represents no edge. The subgraph correspond to all the 43 selected partitions and 10 randomly chosen partitions not selected. (Right) The graph G containing all 125 nodes. Red nodes correspond to selected partitions, while pale nodes correspond to partitions not selected.

We visualize the proportion of selected partitions per window in the BrainSpan dataset in

Figure 10 to demonstrate that our findings are consistent with the findings in Willsey et al. (2013). As mentioned in Section 2, Willsey et al. (2013) found that partitions in Window 1B were mostly homogenous and were enriched for risk genes. The authors also found that the gene expression from different brain tissues are more correlated as the developmental periods are more similar. The authors also estimated a hierarchical clustering among the four brain regions. Indeed, our results match these finding as we select a large proportion of partitions in Window 1B. The proportion of selected partitions decreases as the window represents older developmental periods as well as brain regions more dissimilar to Window 1B.

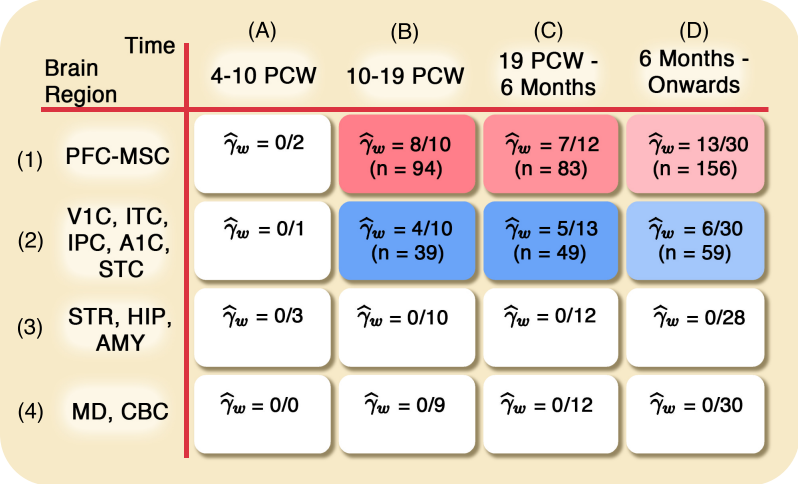

Figure 10: The number of partitions and samples (n) selected within each window. Partitions from 6 different windows are chosen, and the estimated γ_w is empirical fraction of selected partitions within each window. The more vibrant colors display a value of $\widehat{\gamma}_w$.

Lastly, we apply the same diagnostic as in Section 3 to show in Figure 11 that the samples within our 43 selected partitions are much more homogenous than the samples among all partitions in Window 1B. The p-values we obtain after 500 trials are much closer to uniform that those shown in Section 3. To re-emphasize, we interpret this result as a diagnostic, not as a formal goodness-of-fit test as our p-values are not independent, and the partitions were selected based on the BrainSpan data.

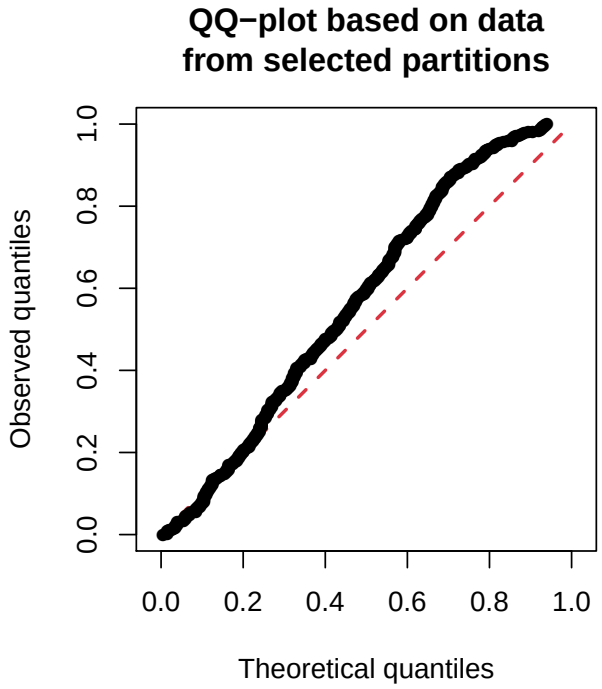

Figure 11: A QQ-plot of the 500 p-values generated when applying our diagnostic to the 43 selected partitions, similar in style to Figure 2.

6.2 Gene network and detected risk genes

DAWN uses two datasets of genetic information to identify risk genes. The first dataset has been the primary focus of this article so far. It contains the microarray samples that our method selected from the BrainSpan dataset. The second dataset contains risk scores for each gene that compare the amount of genetic variation found in individuals with ASD to individuals without ASD (He et al., 2013; De Rubeis et al., 2014). For example, if one type of genetic variation in a particular gene is found more commonly in individuals with ASD than individuals without ASD, this gene would have a higher risk score and be more likely to be a risk gene. As mentioned in Section 1, DAWN combines these two datasets by first estimating a gene co-expression network using the microarray samples, and then identifying

risk genes that either have a high risk score or are connected to many other genes with high risk scores.

Figure 12 illustrates the flowchart of how DAWN combines the gene co-expression network with the risk scores. The first step uses the method we developed in Section 4 to select 43 partitions from the BrainSpan dataset, as stated in Subsection 6.1. In the second step, DAWN estimates a Gaussian graphical model from the samples in these partitions to represent the gene co-expression network. We use neighborhood selection to estimate this Gaussian graphical model, where the tuning parameter was chosen via 5-fold cross validation (Meinshausen and Bühlmann, 2006). In the last step, DAWN identifies risk genes using a Hidden Markov random field model to combine the Gaussian graphical model with the risk scores. The details are in Liu et al. (2015), but in short, this assumes a mixture model of the risk scores between risk genes and non-risk genes, and the probability of being risk gene depends on the graph structure. To enable a fair comparison between our results and those in Liu et al. (2015), we apply the authors' screening method to analyze only 6670 genes, and identify 246 risk genes. The specific risk genes we identify are listed in our online supplement.

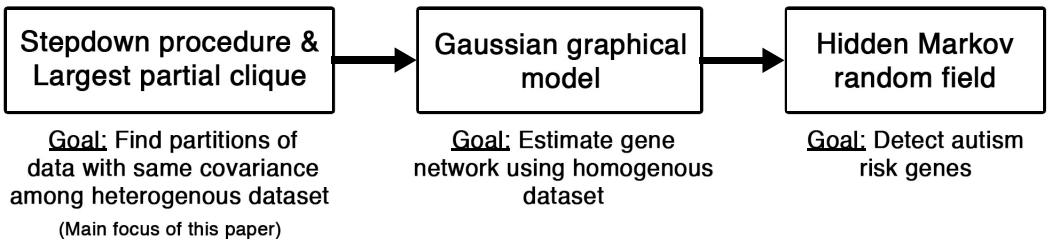

Figure 12: Flowchart of how our partition selection procedure (stepdown procedure and largest partial clique) is used downstream to find risk genes.

6.3 Investigation on detected risk genes

We demonstrate that the 246 risk genes we identify are more promising than those identified in Liu et al. (2015) since an independent study found new de novo loss-of-function (dnLoF) mutations in a higher percentage of our risk genes than before. Specifically, Iossifov et al. (2014) found 251 genes with dnLoF mutations not already factored into the risk scores used in our DAWN analysis. These 251 genes are natural candidates to compare our risk genes against since multiple separate studies found twice as many dnLoF mutations in individuals with ASD than individuals without (Neale et al., 2012; Iossifov et al., 2012; Sanders et al., 2012; O'Roak et al., 2012). This makes dnLoF mutations contain the most signal among all forms of genetic variation. Hence, we are hoping for as many of our 246 risk genes to overlap with these 251 genes with dnLoF mutations. However, since dnLoF mutations are rare events, we realistically do not expect a high overlap percentage. For example, De Rubeis et al. (2014) sequenced more than two thousand ASD trios but found less than two dozen genes with more than one dnLoF mutations.

We find that 19 of our 246 risk genes (7.8%) had additional dnLoF mutations in Iossifov et al. (2014), which is an improvement over the previous finding in Liu et al. (2015) where only 16 of 246 risk genes (6.5%) overlapped. These 19 genes are ADNP, ANK2, ARID1B, CHD8, DIP2A, DSCAM, DYRK1A, FOXP1, ILF2, KDM5B, KDM6B, MED13L, NCKAP1, PHF2, POGZ, RANBP17, RIMS1, SPAST, and WDFY3. If we modelled the number of risk genes that overlap as a Bernoulli random variable, our findings represents roughly a one standard deviation improvement. That is, an improvement of one standard deviation would require identifying $\sqrt{246 \cdot (16/246) \cdot (1 - 16/246)} \approx 3.86$ more overlapped risk genes, and we identify 3 more risk genes.

Furthermore, these 19 overlapped risk genes are robust to the familywise error control α used in our Stepdown procedure in Section 4. We apply our procedure to a range of α values

between 0.05 and 0.35 at intervals of 0.025. This results in 13 different sets of risk genes. In 8 or more of these sets of risk genes, we find the same 19 overlapped risk genes. In fact, the 246 risk genes themselves are also robust to α. Among the 13 sets of risk genes, 238 risk genes were identified 8 or more times. All together, our results show that our method is able to identify risk genes that are more promising than before, and this finding is not dependent on our method's tuning parameter α.

7 Conclusion and discussions

In this article, we develop a procedure to select partitions with statistically indistinguishable covariance matrices and apply it to help identify risk genes. Our procedure first applies a Stepdown method to simultaneously test all $\binom{r}{2}$ hypotheses, one for testing whether or not each pair of partitions shared the same population covariance matrix. The Stepdown method is critical since it can preserve the dependencies among all $\binom{r}{2}$ hypotheses via bootstrapping the joint null distribution. Then, our procedure uses a clique-based selection method to select the partitions based on the accepted null hypotheses. The novelty in this procedure is its ability to preserve monotonicity, a property stating that less partitions should be selected as the number of accepted null hypotheses gets smaller. We demonstrate empirically that our procedure achieves this property while common methods such as spectral clustering do not. When we apply our procedure to the BrainSpan dataset as part of the DAWN analysis, we find scientifically meaningful partitions based on the findings in Willsey et al. (2013). We also find a higher percentage of our risk genes overlap with genes identified in an independent study (Iossifov et al., 2014) compared to previous works (Liu et al., 2015). This result is not sensitive to the tuning parameter of our procedure.

The theoretical role of the familywise error level α is not well understood mathemati-

cally. Specifically, while (4.2) provides a theoretical guarantee on the set of null hypothesis accepted, what we would like to prove is a theoretical guarantee on the set of selected partitions $\widehat{\mathcal{P}}$. During the development of this article, it became clear that no guarantees on $\widehat{\mathcal{P}}$ can be provided unless we understand the power of the Stepdown procedure.

Our procedure is applied directly to help identify risk genes for ASD, but this line of work has broader implications in genetics. Due to the improvement of high throughput technologies, it has been increasingly accessible to gather large amounts of microarray expression data. However, as we have seen in this article, gene expression patterns can vary wildly among different tissues and subjects. This makes identifying samples that are relevant to the scientific task difficult. Beyond analyzing brain tissues, Greene et al. (2015) develop procedures to select relevant samples amongst a corpus of microarray expression data from many different tissue types for specific scientific tasks. While that work currently contains no statistical foundation, our work provides an possible statistical direction for this research field to move towards.

References

Autism and Investigators, D. D. M. N. S. Y. . P. (2014). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, united states, 2010. *Morbidity and Mortality Weekly Report: Surveillance Summaries*, 63(2):1–21.

Buxbaum, J. D., Daly, M. J., Devlin, B., Lehner, T., Roeder, K., State, M. W., and The Autism Sequencing Consortium (2012). The Autism Sequencing Consortium: Large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron*, 76(6):1052–1056.

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Chang, J., Zhou, W., and Zhou, W.-X. (2015a). Bootstrap tests on high dimensional covariance matrices with applications to understanding gene clustering. *arXiv preprint arXiv:1505.04493*.

Chang, J., Zhou, W., Zhou, W.-X., and Wang, L. (2015b). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *arXiv preprint arXiv:1505.04493*.

Chernozhukov, V., Chetverikov, D., Kato, K., et al. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Cotney, J., Muhle, R. A., Sanders, S. J., Liu, L., Willsey, A. J., Niu, W., Liu, W., Klei, L., Lei, J., and Yin, J. (2015). The autism-associated chromatin modifier chd8 regulates other autism risk genes during human neurodevelopment. *Nature communications*, 6.

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., Kou, Y., Liu, L., Fromer, M., and Walker, S. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–215.

Dong, S., Walker, M. F., Carriero, N. J., DiCola, M., Willsey, A. J., Adam, Y. Y., Waqar, Z., Gonzalez, L. E., Overton, J. D., Frahm, S., et al. (2014). De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell reports*, 9(1):16–23.

Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., and Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, 70(5):898–907.

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., and Sealfon, S. C. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*.

He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., and Buxbaum, J. D. (2013). Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes. *PLoS Genetics*, 9(8):e1003671.

Ieva, F., Paganoni, A. M., and Tarabelloni, N. (2016). Covariance-based clustering in multivariate and functional data analysis. *The Journal of Machine Learning Research*, 17(1):4985–5005.

Iossifov, I., ORoak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., and Patterson, K. E. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526):216–221.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-h., Narzisi, G., and Leotta, A. (2012). *De novo* gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–299.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., and Sedmak, G. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489.

Kanner, L. et al. (1943). Autistic disturbances of affective contact. *Nervous child*, 2(3):217–250.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161.

Lei, J. and Zhu, L. (2017). Generic sample splitting for refined community recovery in degree corrected stochastic block models. *Statistica Sinica*, 27:1639–1659.

Liu, L., Lei, J., and Roeder, K. (2015). Network assisted analysis to reveal the genetic basis of autism. *The Annals of Applied Statistics*, 9(3):1571–1600.

Liu, L., Lei, J., Sanders, S. J., Willsey, A. J., Kou, Y., Cicek, A. E., Klei, L., Lu, C., He, X., and Li, M. (2014). DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol Autism*, 5:22.

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

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., and Makarov, V. (2012). Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature*, 485(7397):242–245.

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., Levy, R., Ko, A., Lee, C., and Smith, J. D. (2012). Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature*, 485(7397):246–250.

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., Horvath, S., and Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, 155(5):1008–1021.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Rutter, M. (1978). Diagnosis and definition of childhood autism. *Journal of autism and childhood schizophrenia*, 8(2):139–161.

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., Murtha, M. T., Bal, V. H., Bishop, S. L., Dong, S., et al. (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron*, 87(6):1215–1233.

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., and Stein, J. L. (2012). *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241.

Šestan, N. et al. (2012). The emerging biology of autism spectrum disorders. *Science*, 337(6100):1301–1303.

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., Reilly, S. K., Lin, L., Fertuzinhos, S., and Miller, J. A. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, 155(5):997–1007.

A Dataset Details

There are four primary brain regions, each containing smaller, scientifically-interesting brain regions.

- **PFC-MSC**: The prefrontal cortex and primary motor-somatosensory cortex consists five smaller regions: primary motor cortex (M1C), primary somatosensory cortex (S1C), ventral prefrontal cortex (VFC), medial prefrontal cortex (MFC), dorsal prefrontal cortex (DFC) and orbital prefrontal cortex (OFC).
- **V1C, ITC, IPC, A1C, STC**: A region consisting of the primary visual cortex (V1C), inferior temporal cortex (ITC), primary auditory cortex (A1C), and superior temporal cortex (STC).
- **STR, HIP, AMY**: A region consisting of the stratum (STR), hippocampal anlage or hippocampus (HIP) and amygdala (AMY).
- **MD, CBC**: A region consisting of the mediodorsal nucleus of the thalamus (MD) and the cerebellar cortex (CD).

B Non-normalized analysis

to come, simulation results for non-normalized statistic. ROC curves as well as run-time improvement.

code?