CLUSTER ANALYSIS AND NETWORK COMMUNITY DETECTION WITH APPLICATION TO NEUROSCIENCE

by

Yun Zhang

B.S., Zhejiang University, 2009

M.S., the Ohio State University, 2011

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial

fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Yun Zhang

It was defended on

Sep. 30, 2016

and approved by

Allan R. Sampson, Ph.D., Department of Statistics, Professor

Kehui Chen, Ph.D., Department of Statistics, Assistant Professor

Sungkyu Jung, Ph.D., Department of Statistics, Assistant Professor

David Volk, M.D., Ph.D, Department of Psychiatry, Assistant Professor

Dissertation Advisors: Allan R. Sampson, Ph.D., Department of Statistics, Professor,

Kehui Chen, Ph.D., Department of Statistics, Assistant Professor

CLUSTER ANALYSIS AND NETWORK COMMUNITY DETECTION WITH APPLICATION TO NEUROSCIENCE

Yun Zhang, PhD

University of Pittsburgh, 2016

Sustained efforts have been devoted to understanding schizophrenia and related disorders. This dissertation is inspired from two conceptually important problems in schizophrenia research and we overcome statistical challenges inherent in solving these problems.

Basic neurobiological studies have unveiled distinct subtypes of schizophrenia. Moreover, genetic evidence shows certain core features are shared between schizophrenia and other disorders. It is of scientific interest to examine similarities in the profiles of subtypes in different disorders, which may help to develop novel therapeutic approaches. To address this challenge, we develop a statistical framework to assess whether or not clusters identified from independent populations exhibit commonalities. As an initial step, we formulate our hypotheses by borrowing the concept of bioequivalence under a finite normal mixture framework. We then propose testing procedures for univariate data based on the idea of two one-sided test (TOST). In an attempt to boost power, we propose to use a methodology based on bootstrap confidence intervals.

Neurocognitive research studies functional brain networks aiming to improve the understanding of the cognitive deficits in subjects with schizophrenia. One important problem in the inference for brain connectivity networks concerns brain segmentation problem which can be viewed as a community detection problem in network analysis. The stochastic block model (SBM) and its variants are popular models used in community detection for network data. In this research, we propose a feature adjusted stochastic block model (FASBM) to capture the impact of node features on the network links as well as to detect the residual

community structure beyond that explained by the node features. The proposed model can accommodate multiple node features and estimate the form of feature impacts from the data. Moreover, unlike many existing algorithms that are limited to binary-valued interactions, the proposed FASBM model and inference approaches are easily applied to relational data that generates from any exponential family distribution. We illustrate the methods on simulated networks and on three real world networks: a brain network, an US air-transportation network and a friendship network.

Keywords: Bioequivalence Testing, GABA Neuron-Related Biomarker Study, Stochastic Block Model, Community Detection, Node Features, Air-transportation Network, Brain Functional Connectivity Study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 MOTIVATION OVERVIEW

Considerable research has been devoted to the understanding of the neurobiology of schizophrenia and related disorders. The Conte Center for Translational Mental Health Research (CTMHR) at the University of Pittsburgh has been focusing on the mechanisms that link the pathology, pathophysiology and clinical features of schizophrenia and related disorders. In this dissertation, both parts of our research are inspired from conceptually important problems within predominant strands of CTMHR schizophrenia research and well generalizable to other contexts.

Basic neurobiological research concerning mental disorders focuses on studying neurobiological alterations in subjects with a mental disorder. Researchers measure neurobiological characteristics such as gene expression levels and protein levels from post-mortem brain tissue samples. For example, Volk et al. (Volk et al., 2012) identified a subset of schizophrenia subjects that consistently showed deficits in certain GABA neuron-related mRNAs through cluster analysis based on post-mortem brain tissue studies. On the other hand, recent advances in psychiatric genomics have given insight into the potential mechanisms underlying the overlap between schizophrenia and bipolar disorder. It has been found that schizophrenia and bipolar disorder shared deep genetic similarities (Craddock et al., 2005; Moskvina et al., 2009; Doherty & Owen, 2014), in support of the long-standing clinical observation of overlap in the symptoms. Making use of the data from multiple disorders in the Center studies may provide an opportunity to obtain new insights into the understanding of the potential mechanisms underlying the overlap between mental disorders. It is of scientific interest to examine similarities in the profiles of subtypes in different disorders, which may

help to develope novel therapeutic approaches (Doherty & Owen, 2014). In Chapter 2, we develop methodology to test whether or not clusters identified from independent populations exhibit commonalities.

In parallel, another strand of research is focused on studying the cognitive deficits in subjects with schizophrenia. Schizophrenia has often been conceived as a disorder of connectivity between components of brain networks (Lynall et al., 2010). In related studies for brain tissues of living subjects in schizophrenia, researchers are interested in understanding how schizophrenia affects brain networks. The advent of modern neuroimaging techniques such as fMRI makes it feasible to quantify different aspects of brains functional interactions. The study of functional brain networks may advance the understanding of how key functional networks are altered in schizophrenia, thus improving the understanding of cognitive difference in schizophrenia as revealed by fMRI. One important problem in the inference for brain connectivity networks concerns partitioning of functionally distinct brain regions, that is, brain segmentation. The brain segmentation problem is conceptually a community detection problem in network analysis. We propose a new framework for community detection in Chapter 3 that takes into account the topological structure of the network and the additional information on nodes. Although our work is motivated by inference for brain connectivity networks, the proposed models and algorithms pertain to a general setting and can be used in a variety of networks.

1.2 DISSERTATION OVERVIEW

The dissertation is organized as follows. In Chapter 2, we develop a new methodology to identify common clusters in independent populations. We start reviewing some existing literature on the topic of cluster validation in Section 2.1, particularly, the review focuses mainly on the work of Tibshirani et al. (2007) that addresses a related yet different problem. The hypotheses are formulated in Section 2.2. We propose two one-sided test (TOST) based approach for univariate data and multivariate data in Section 2.3.1. Because the testing procedures are overly conservative for multivariate data, we then propose a confidence interval

approach in Section 2.3.2 using various bootstrap methods: non-Studentized pivotal method, percentile method, bias-corrected percentile method and Normal method. A discussion of the asymptotic properties of the proposed testing procedures are presented in Section 2.4. We evaluate the performance of our proposed testing procedures in univariate data (Section 2.5.1) and bivariate data (Section 2.5.2) under a variety of scenarios using simulation. We then apply our methodology to a GABA neuron-related biomarker study (Section 2.6.1) We close the chapter with conclusions in Section 2.7.1 and discussions on some possible work for future research in Section 2.7.2.

In Chapter 3, we propose a new model to capture the impact of node features on the network links as well as to detect the residual community structure beyond that explained by the node features. Chapter 3 begins with a literature review of the community detection methods in the network analysis. This is followed by a more detailed introduction of the relevant background: single-index model (Section 3.2.1) and stochastic block model (Section 3.2.2) along with a brief review of commonly used algorithms for inference of block models (Section 3.2.3). We propose the feature adjusted stochastic block model (FASBM) in Section 3.3 and introduce the fitting algorithms for the proposed model in Section 3.4. The performance of the proposed method is demonstrated on a range of simulated networks in Section 3.5 and in Section 3.6 is applied to a functional brain network, an US air-transportation network and a friendship network. The chapter is concluded with a short discussion on future directions in Section 3.7.

We conclude the dissertation with some remarks in Chapter 4.

2.0 ARE THERE COMMON CLUSTERS IN INDEPENDENT POPULATIONS?

2.1 INTRODUCTION

Cluster analysis is a powerful technique that helps identify subtypes from heterogeneous data. Identification of subtypes may be important for research on mechanisms of disease with subjects in one cluster having similar characteristics. For example, using cluster analysis, Volk et al. (Volk et al., 2012) identified a subset of schizophrenia subjects that consistently showed deficits in certain GABA neuron-related mRNAs. Moreover, recent advances in psychiatric genomics have given more insight into the potential mechanisms underlying the overlap between schizophrenia and bipolar disorder. It has been found that schizophrenia and bipolar disorder share deep genetic similarities (Craddock et al., 2005; Moskvina et al., 2009), which supports the long-standing clinical observation of overlap in the symptoms. In addition, deficits in some GABA neuron-related mRNAs have been reported from subjects with schizophrenia disorder and bipolar disorder (Guidotti et al., 2000; Woo et al., 2008; Sibille et al., 2011). Taken together, these findings suggest that the subtype characterized by deficits in certain GABA neuron-related mRNA levels may identify a subset of subjects from each of these diagnostic groups. To test this hypothesis, researchers at the Conte Center for Translational Mental Health Research (CTMHR) measured in post-mortem tissue mRNA levels for four GABA neuron-related markers in the prefrontal cortex from subjects with a diagnosis of schizophrenia or bipolar disorder. Their goal is want to determine the extent to which a similar subtype may exist in subjects with these disorders. Motivated by this problem, our goal is to provide a statistical framework to examine similarities in the profiles of subtypes of different disorders, which may help to develop novel therapeutic approaches

(Doherty & Owen, 2014). This work has direct relevance to investigators pursuing new lines of research in light of the National Institute of Mental Health's (NIMH) Research Domain Criteria (RDoC) project.

If one is interested in identifying clusters in a single population, hierarchical algorithms using "bottom up" strategy inherently involves assessing similarities between clusters. The basic "bottom up" algorithm is very simple. Start with each point in a cluster of its own; then construct a hierarchy of clusters by examining a suitable notion of distance between two clusters using methods, such as Ward's method, Single-link Clustering or Complete-Link Clustering, and repeatedly merge the two most similar clusters together until there is only one remaining cluster.

However, there is relatively little literature concerning whether or not clusters identified from independent populations share commonalities. There is a seemingly related literature concerning cluster validation for multiple micro array studies (Chen et al., 2002; Datta & Datta, 2003; Kerr & Churchill, 2001; Yeung et al., 2001; Dudoit & Fridlyand, 2002; Dudoit et al., 2002; Tibshirani & Walther, 2005; Tibshirani et al., 2007).

Cluster validation aims to "assess the validity of classifications that have been obtained from the application of a clustering procedure" (Gordon, 1999). Clustering validation in one data set is concerned with evaluating the goodness of clustering results, aiding in determining which clustering analysis approach to use as well as the optimal cluster number (Liu et al., 2010). Cluster validation can be used in a slightly different way when in some studies, the goal of analyzing a new independent dataset (validating data set) is to identify the same clusters in the validation data that were defined in the previous data set (defining data set). If the cluster is present in the validating data set, then this cluster is validated because it is reproducible. With this goal in mind, in general, cluster validation procedures first define a cluster quality measure and then obtain p-values by computing how likely given values of that measure are to occur under an appropriate null model of no structure (Tibshirani et al., 2007).

It is suggested that, when the validating data set and the defining data set have the same variables, an appropriate approach for cluster validation analyses is to use a classifier made from the defining data (Dudoit & Fridlyand, 2002; Dudoit et al., 2002; Tibshirani &

Walther, 2005). For instance, Tibshirani et al. (2007) adopted the nearest centroid classifier, where the centroids represent the averages of all variables over subjects within each cluster in the defining data set. For every subject in the validating data set, the Pearson correlation coefficient between the subject and each centroid of the defining data is computed. If the correlations are all smaller than a cut-off value, then the subject is classified to a "below-cutof" group; otherwise, the subject is classified into the cluster whose correlation is the largest. In order to obtain a p-value for testing H_0: there is no cluster structure; vs H_a: the previously defined cluster is valid, it is important to compare a test statistic based on the cluster quality measure with the null distribution of the cluster quality measure. Tibshirani et al. (2007) proposed a new cluster quality measure called In-Group Proportion (IGP) defined as "the proportion of (new) observations classified to a cluster whose nearest neighbor is also classified to the same cluster", so that a high-quality cluster will have IGP close to 1 when the subject and its nearest neighbor are classified into the same cluster. Further, they proposed four difficult ways to generate a null distribution of IGPs and all of which are based on repeatedly generating new centroids that correspond to clusters that are placed randomly in the data, so that the p-value is defined as the proportion of null distribution IGPs that are greater than the actual IGP.

One shortcoming of Kapp and Tibshirani's approach is that, the number of clustering variables needs to be fairly large to generate a good null distribution. This is not an issue for microarrays studies, as the number of mRNA's being studied is usually quite large. However, in other setting, this may not hold. Additionally, despite the fact that their parameter-free hypothesis formulation is appealing, the precise definition of a null model is difficult to formalize, as one can argue that the opposite of reproducibility of a cluster in the data is not equivalent to no structure in the data, or vice versa.

To this end, we develop an approach with greater clarity for testing whether or not clusters identified from independent populations exhibit commonalities. The basic idea is to recast the formulation of the hypothesis tests in such a way that we can utilize some ideas from the analysis of pharmaceutical bioequivalence trials. Our method is tailored specifically for mixture model-based clustering and thus inherits the merits of this paradigm. Benefits of mixture model-based clustering in comparison to hierarchical clustering are discussed by

Raftery & Dean (2006). Due to the importance and broad applicability of finite normal mixture models, we consider the case where the mixture components have normal distributions.

2.2 HYPOTHESIS OF INTEREST

In this section, we formulate our approach for assessing whether or not there are common clusters in two independent populations. In the finite normal mixture framework, every cluster can be mathematically represented by a normal distribution. To simplify further the presentation, we suppose that there are only two populations under consideration and each population is distributed as a mixture of two p-variate normal distributions. The extension of our formulation to more than two populations and two mixtures is conceptually straightforward, but can be computationally intensive.

Suppose that we observe a random sample of p-dimensional variables $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_1}$ from population 1, which is a mixture of normals:

$$f(\boldsymbol{x}_i) = \pi_1 \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1 - \pi_1)\phi(\boldsymbol{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \tag{2.1}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $i = 1, \cdots, n_1$. Let $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{n_2}$ denote a random sample of size n_2 of p-dimensional vectors from population 2 distributed as

$$g(\boldsymbol{y}_j) = \pi_2 \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) + (1 - \pi_2)\phi(\boldsymbol{y}_j; \boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4) \tag{2.2}$$

for $j = 1, \cdots, n_2$, where both random samples are independent of each other.

In order to avoid the identifiability problem and unbounded likelihood as well as to assure that desirable asymptotic properties of maximum likelihood estimates hold (McLachlan & Peel, 2004), the following conditions are imposed:

$$\begin{aligned}
0 < \quad & \pi_1, \pi_2 \quad < 1 \\
\boldsymbol{\mu}_1^q \quad & > \quad \boldsymbol{\mu}_2^q, \quad q = \min\{j : \boldsymbol{\mu}_1^j \neq \boldsymbol{\mu}_2^j, j = 1, \cdots, p\} \\
\boldsymbol{\mu}_3^{q'} \quad & > \quad \boldsymbol{\mu}_4^{q'}, \quad q' = \min\{j : \boldsymbol{\mu}_3^j \neq \boldsymbol{\mu}_4^j, j = 1, \cdots, p\} \\
\boldsymbol{\Sigma}_1 \quad & = \quad \boldsymbol{\Sigma}_2 \quad = \quad \boldsymbol{\Sigma}_3 \quad = \quad \boldsymbol{\Sigma}_4 \quad \equiv \quad \boldsymbol{\Sigma}.
\end{aligned} \tag{2.3}$$

Under these conditions, the similarity of any two clusters from each population can be assessed in terms of a comparison of the means of two normal distributions.

In the context of our setting where we want to show a found cluster is common to both populations, we need to assess that two p-dimensional mean vectors are the same or different. The usual hypothesis testing paradigm is designed to protect Type I error, an error of incorrect rejection of a true null hypothesis. Consequently, lack of evidence to reject the null hypothesis does not imply sufficient evidence to support it. In light of that, the hypothesis one desires to prove should be stated as the alternative hypothesis. The formal hypothesis testing formulation in our clustering problem essentially becomes H_0: two mean vectors are unequal. *vs* H_A: two mean vectors are the same, or as we show in the following, a slight variation of the above formulation.

It turns out this type of hypothesis testing problem has been extensively studied in the context of demonstrating bioequivalence (BE) in the biopharmaceutical industry (Chow & Liu, 2000). Regulatory agencies require that a proposed generic drug be biosimilar to the approved and listed drug, i.e., the reference drug. Current US Food and Drug Administration (FDA) guidelines (FDA Guidance, 2001) declare the test and reference drug as average bioequivalent if a suitable measure of absorption differs by less than a (clinically) meaningful limit. In the same spirit, two clusters may be considered in common if characteristics of the two clusters differ by less than a scientifically meaningful and suitable cluster equivalence margin which we denote by Δ. In the other words, if a suitably measured univariate difference between any two clusters is less than Δ, the two clusters would be viewed equivalent or the same. The quantity Δ is the maximum allowable difference between any two clusters that from the scientific view can be ignored. The null hypotheses of no common cluster between two populations can be formulated as follows:

$$H_0 = H_{01} \cap H_{02} \cap H_{03} \cap H_{04} \tag{2.4}$$

where

$$\begin{aligned} H_{01} &: d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_3) \geq \Delta \quad H_{02} : d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_4) \geq \Delta \\ H_{03} &: d(\boldsymbol{\mu}_2, \boldsymbol{\mu}_3) \geq \Delta \quad H_{04} : d(\boldsymbol{\mu}_2, \boldsymbol{\mu}_4) \geq \Delta \end{aligned} \tag{2.5}$$

where $d(\boldsymbol{x}, \boldsymbol{y})$ is an appropriate measure of distance, $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) \geq 0$. Let R_r be the $\alpha_r = \alpha/4$ level rejection region for testing $H_{0r}, r = 1, \cdots, 4$; then an overall α level rejection region would be $R = \bigcup_{r=1}^{4} R_r$ by the Bonferroni multiplicity adjustment. The rejection of H_0 would lead to the conclusion that at least one common cluster is shared in the two independent populations. Precise descriptions of the testing procedures are given in the following section.

2.3 TESTING PROCEDURES

An appeal of fomulations (2.4) and (2.5) are that they are flexible enough to cover both the univariate case and the multivariate case by using different distance measures. The focus is to select an appropriate distance measure and find a way of constructing rejection regions for $H_{0r}, r = 1, \cdots, 4$. We first propose testing procedures for univariate data based on the idea of two one-sided tests. The two one-sided tests can be immediately extended to multivariate data by the intersection-union method. Although such an extension is straightforward, as we show later in Section 2.5.2, it suffers from low power even for bivariate data. We therefore further propose a more powerful testing approach for multivariate data using a methodology based on bootstrap confidence intervals.

2.3.1 Two one-sided test (TOST) based approach

2.3.1.1 Preliminaries

The two one-sided test (TOST) was first introduced by Schuirmann (1981, 1987) for assessing average bioequivalence and thereafter has been adopted by FDA (FDA Guidance, 1992) as its process for approving a new generic drug. In the context of the TOST setting, let θ denote a bioequivalence measure of interest, for example, the population mean difference of the AUC (areas under the blood serum curve) on the log scale between the reference drug and the test drug. The hypothesis of TOST is formulated as $H_0 : |\theta| \geq \Delta \quad \text{versus} \quad H_A : |\theta| < \Delta$, where Δ is a pre-defined clinically meaningful limit. Let D be an estimate of θ that is assumed

to follow a normal distribution with mean θ and variance σ_D^2. The α level TOST rejects H_0 if $|D| \leq \Delta - z_\alpha \sigma_D$, where z_α is the upper α quantile of the standard normal distribution. It also has been noted that TOST is operationally identical to the classic confidence interval approach: reject H_0 at level α if the $100(1-2\alpha)\%$ confidence interval for θ is contained in $(-\Delta, \Delta)$ (Berger et al., 1996). TOST has been shown to be a test of significance level α, and generally its actual size is smaller than the nominal significance level (Chow & Liu, 2000). A great deal of work has been devoted to improve the power of TOST. Brown et al. (1997), Munk et al. (2000) and Berger et al. (1996) proposed tests involving making adjustments to the TOST rejection region to improve power and keep the Type I error from exceeding the stated α level. Despite these tests are theoretically more powerful, the real advantage is too negligible for any practical use (Chow & Liu, 2000). Thus, from a practical point of view, TOST stands out for its simplicity and intuitive appeal.

The Expectation Maximization (EM) algorithm is widely used for maximum likelihood estimation in finite mixture models. Its main characteristics have been well studied (McLachlan & Krishnan, 2007). The major drawbacks of the EM algorithm are its slow convergence and the strong dependence of the estimates on the starting point. Many algorithms have been developed to speed up the convergence (Liu & Sun, 1997) and comparisons of different methods to choose sensible starting points for obtaining the highest likelihood are studied in (Karlis & Xekalaki, 2003; Biernacki et al., 2003). Our methodology does not use any of these updated aspects of the EM algorithm, but uses the most standard EM algorithm without any acceleration scheme. But clearly these other algorithms could be easily applied in our setting.

2.3.1.2 Univariate case

For motivation, we first consider $p = 1$, $d(x, y) = |x - y|$ and $n_1 = n_2 = n$. Our testing procedures are presented as follows:

Step 1: Apply the EM algorithm to the joint distribution of $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ to obtain maximum likelihood estimates $\widehat{\mu}_l, l = 1, \cdots, 4, \widehat{\pi}_1, \widehat{\pi}_2, \widehat{\sigma}$ subject to conditions (2.3). Parameters estimated at the g^{th} iteration are marked by a superscript g.

(a) Initializing Step: Apply k-means with pre-specified number of clusters for $x_1, \cdots, x_n$ and

$y_1, \cdots, y_n$ respectively to obtain initial values, set $g = 1$.

1. Initialize $\hat{\mu}_l^{(1)}$ by computing the average of the observations classified to cluster $l, l = 1, \cdots, 4$ and let $\hat{\mu}_1^{(1)} > \hat{\mu}_2^{(1)}$, $\hat{\mu}_3^{(1)} > \hat{\mu}_4^{(1)}$ for identifiability.
2. Initialize $\hat{\sigma}^{2(1)}$ by taking the average of sample variance $\hat{\sigma}_l^{2(1)}$ computed from the observations classified to cluster l.
3. Initialize $\hat{\pi}_1^{(1)}$, $\hat{\pi}_2^{(1)}$ by the proportion of data assigned to the corresponding clusters.

(b) E-step: Compute the posterior probabilities of cluster labels l for each point x_i and y_j, $i, j = 1, \cdots, n$,

$$p_{i,(l=1)} = \frac{\hat{\pi}_1^{(g)}\phi(x_i|\hat{\mu}_1^{(g)}, \hat{\sigma}^{2(g)})}{\hat{\pi}_1^{(g)}\phi(x_i|\hat{\mu}_1^{(g)}, \hat{\sigma}^{2(g)}) + (1-\hat{\pi}_1^{(g)})\phi(x_i|\hat{\mu}_2^{(g)}, \hat{\sigma}^{2(g)})} \qquad p_{i,(l=2)} = 1 - p_{i,(l=1)}$$

$$p_{j,(l=3)} = \frac{\hat{\pi}_2^{(g)}\phi(y_j|\hat{\mu}_3^{(g)}, \hat{\sigma}^{2(g)})}{\hat{\pi}_2^{(g)}\phi(y_j|\hat{\mu}_3^{(g)}, \hat{\sigma}^{2(g)}) + (1-\hat{\pi}_2^{(g)})\phi(y_j|\hat{\mu}_4^{(g)}, \hat{\sigma}^{2(g)})} \qquad p_{j,(l=4)} = 1 - p_{j,(l=3)}.$$

(c) M-step:

$$\hat{\pi}_1^{(g+1)} = \frac{\sum_{i=1}^n p_{i,(l=1)}}{n} \qquad \hat{\pi}_2^{(g+1)} = \frac{\sum_{j=1}^n p_{j,(l=3)}}{n}$$

$$\hat{\mu}_l^{(g+1)} = \frac{\sum_{i=1}^n p_{i,l}x_i}{\sum_{i=1}^n p_{i,l}}, l = 1, 2 \qquad \hat{\mu}_l^{(g+1)} = \frac{\sum_{j=1}^n p_{j,l}y_j}{\sum_{j=1}^n p_{j,l}}, l = 3, 4$$

$$\hat{\sigma}_l^{(g+1)} = \frac{\sum_{l=1}^2\sum_{i=1}^n p_{i,l}(x_i - \hat{\mu}_l^{(g+1)})^2 + \sum_{l=3}^4\sum_{j=1}^n p_{j,l}(y_j - \hat{\mu}_l^{(g+1)})^2}{2n}.$$

(d) Repeat E-step and M-step until there is suitable convergence.

Running k-means first to obtain initial values in mixture models is common used (McLachlan & Peel, 2004). Biernacki et al. (2003) also recommends a three step search-run-select strategy as follows: start with several different initial values in the initializing step, and follow step (b)-(d) with a relatively liberal convergence criteria in (d), then select the solution that leads to the largest value of loglikelihood and keep repeating (b)-(c) until a strict convergence criteria is met.

Step 2: Use either parametric or nonparametric bootstrap methods to obtain standard errors of the estimates. Draw bootstrap samples $x_1^{*(b)}, \cdots, x_n^{*(b)}$, $y_1^{*(b)}, \cdots, y_n^{*(b)}$ from the fitted mixture distributions if one uses the parametric bootstrap or from the observed data with replacement if one uses the nonparametric bootstrap, where superscript $*(b)$ denotes

the b^{th} bootstrap sample. Apply the EM algorithm to each bootstrap sample and obtain $\widehat{\mu}_l^{*(b)}, \widehat{\sigma}^{*(b)}$. Consequently, we approximate $\text{cov}(\widehat{\mu}_l, \widehat{\mu}_{l'})$ by $\widehat{\text{cov}}(\widehat{\boldsymbol{\mu}}_l^*, \widehat{\boldsymbol{\mu}}_{l'}^*)$, where $\widehat{\text{cov}}$ denotes the sample covariance matrix and $\widehat{\boldsymbol{\mu}}_l^* = [\widehat{\mu}_l^{*(1)}, \cdots, \widehat{\mu}_l^{*(B)}]^T$.

Step 3: Apply the Bonferroni procedure for control of the familywise error rate at level α. The rejection region for H_0 is $R = \cup_r^4 R_r$, where R_r is the rejection region for H_{0r} that can be construsted by TOST at a significance level $\alpha/4$ using the asymptotic normality properties of the parameter estimates (McLachlan & Peel, 2004). For example, R_1 is given as:

$$R_1 : |\widehat{\mu}_1 - \widehat{\mu}_3| \leq \Delta - z_{\alpha/4}\widehat{\sigma}_{\widehat{\mu}_1 - \widehat{\mu}_3}. \tag{2.6}$$

2.3.1.3 Multivariate case

The above testing procedures can be easily adapted to the multivariate case by considering two clusters as common if the population mean values for each of the p variables used in the cluster analysis differ by less than a meaningful limit denoted by $\Delta_j, j = 1, \cdots, p$. To simplify the notation, we assume that $\Delta_j = \Delta, j = 1, \cdots, p$. Thus, the null hypotheses of no common cluster between two populations in p-variate case can be formulated as

$$H_0 = H_{01} \cap H_{02} \cap H_{03} \cap H_{04} \tag{2.7}$$

where

$$\begin{aligned} H_{01} &: \bigcup_{j=1}^{p} |\mu_{1j} - \mu_{3j}| \geq \Delta \quad H_{02} : \bigcup_{j=1}^{p} |\mu_{1j} - \mu_{4j}| \geq \Delta \\ H_{03} &: \bigcup_{j=1}^{p} |\mu_{2j} - \mu_{3j}| \geq \Delta \quad H_{04} : \bigcup_{j=1}^{p} |\mu_{2j} - \mu_{4j}| \geq \Delta. \end{aligned} \tag{2.8}$$

The overall rejection region is still $R = \bigcup_r^4 R_r$, whereas R_r is constructed by $\bigcap_{j=1}^{p} R_{rj}$ using the intersection-union method with R_{rj} obtained by TOST. Despite its simplicity, we found that the intersection-union method can be very conservative. The drawback of this test is that it fails to account for correlations among the variables for the cluster analysis, so that the degree of conservativeness of the test depends on the correlations among the variables (Quan et al., 2001). However, for example, in our motivating data, it's likely that the

measured mRNA genes from post-mortem brain tissue are correlated to some extent. Thus, it seems that it would be best to consider a test that assesses all the variables simultaneously.

2.3.2 Confidence interval approach

As discussed in the preceding Section, the intersection-union method suffers from low power by examining commonality for each variable independently across clusters. In fact, as shown in the Table 2.6 , the power is very low even when only considering two variables in the cluster analysis. Considering that our goal is to demonstrate the overall similarity of two clusters, it seems unnecessarily strong to require similarity in every variable simultaneously. We instead propose to use the L_2 norm as the distance measure for multivariate data, i.e., $d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2 = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})}$, where $||\cdot||_2$ denotes the L_2 norm. L_2 norm is a scientifically meaningful measure of distance. The key step in our testing procedures is to construct confidence intervals for the L_2 norm of the difference of the normal means. However, identifying its exact confidence interval is complicated. In this situation, we use bootstrapping to obtain approximate confidence intervals. Note that bootstrapping implicitly takes into account of the correlation between variables.

Bootstrap methods for producing good approximate confidence intervals in complicated situations have been demonstrated and are well established, for example, Efron (1987), DiCiccio & Efron (1996), Hall (2013). There are a variety of ways to construct bootstrap confidence intervals. Some are computationally expensive, while some are computationally less demanding but with a lower degree of coverage accuracy. Here the less computationally intensive approaches are used in our work. However, confidence interval methods with higher degrees of coverage accuracy may be preferred when computation time is not at issue.

We provide an illustration of the use of four selected bootstrap methods in the context of our testing problem. They are (i) non-studentized pivotal method, (ii) percentile method, (iii) a variant of the percentile method and iv) normal method. For the sake of simplicity, we consider the bivariate case, and show, for example, how to construct the $100(1-\alpha)\%$ bootstrap confidence interval for $||\boldsymbol{d}_{13}||_2$, where $\boldsymbol{d}_{13} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$. Following the flow described in Section 2.3.1, we obtain the estimate $||\widehat{\boldsymbol{d}}_{13}||_2$ in Step 1, and have drawn B samples and

computed $||\widehat{\boldsymbol{d}}^*_{13}||_2^{(b)}$ in Step 2 for each bootstrap sample successively, $b = 1, \cdots, B$. If Δ is not contained in the $100(1-\alpha)\%$ bootstrap confidence interval of $||\boldsymbol{d}_{13}||_2$, then the null hypothesis H_{01} will be rejected in the formulation (2.5). For the simplicity of notation, we suppress the subscripts for the remainder of Section 2.3.2.

2.3.2.1 Non-Studentized pivotal method

The non-studentized pivotal method, sometimes referred to as the basic bootstrap interval method, is arguably a natural way of constructing bootstrap confidence intervals. It's based on the assumed pivotality of $||\widehat{\boldsymbol{d}}|| - ||\boldsymbol{d}||$. Define pivot $T = ||\widehat{\boldsymbol{d}}|| - ||\boldsymbol{d}||$. Let $H(t)$ denote the cumulative distribution function(CDF) of T, i.e., $H(t) = P(T \leq t)$, and define $c = ||\widehat{\boldsymbol{d}}|| - H^{-1}(r + \alpha - 1)$ with $r = H(||\widehat{\boldsymbol{d}}||)$. Thus, the $100(1-\alpha)\%$ confidence interval for $||\boldsymbol{d}||$ would be $(0, c)$ as

$$\begin{aligned} P(0 \leq ||\boldsymbol{d}|| \leq c) &= P(-c \leq -||\boldsymbol{d}|| \leq 0) \\ &= P(||\widehat{\boldsymbol{d}}|| - c \leq ||\widehat{\boldsymbol{d}}|| - ||\boldsymbol{d}|| \leq ||\widehat{\boldsymbol{d}}||) \\ &= H(||\widehat{\boldsymbol{d}}||) - H(||\widehat{\boldsymbol{d}}|| - c) \\ &= 1 - \alpha. \end{aligned}$$

The problem is that, in practice, the distribution H is not known. Nevertheless, the bootstrap principle allows us to learn about the relationship between the true parameter value $||\boldsymbol{d}||$ and the estimator $||\widehat{\boldsymbol{d}}||$ by looking at the relationship between $||\widehat{\boldsymbol{d}}||$ and $||\widehat{\boldsymbol{d}}^*||$, where $\widehat{\boldsymbol{d}}^*$ denotes the estimate based on the bootstrapping. Thus,the CDF of T can be approximated by the CDF of $T^* = ||\widehat{\boldsymbol{d}}^*|| - ||\widehat{\boldsymbol{d}}||$. Further, the CDF of T^*, $H^*(t) = P(T^* \leq t)$ can be estimated by $\frac{1}{B}\sum_{b=1}^{B} I(T_b \leq t)$, where $T_b = ||\widehat{\boldsymbol{d}}^*||^{(b)} - ||\widehat{\boldsymbol{d}}||$. Hence, the $100(1-\alpha)\%$ confidence interval for $||\boldsymbol{d}||$ would be $(0, \hat{c})$ with $\hat{c} = ||\widehat{\boldsymbol{d}}|| - \widehat{H}^{-1}(r + \alpha - 1) = 2||\widehat{\boldsymbol{d}}|| - ||\widehat{\boldsymbol{d}}^*||_{\alpha+r-1}$, where $||\widehat{\boldsymbol{d}}^*||_{\alpha+r-1}$ is the $\alpha + r - 1$ sample quantile of $(||\widehat{\boldsymbol{d}}^*||^{(1)}, \cdots, ||\widehat{\boldsymbol{d}}^*||^{(B)})$.

Note that there are two distinct sources of error in this procedure. The first error arises from the use of empirical CDF of T^* instead of its true CDF H^*. This error is usually negligible as long as B is sufficiently large. The second error, resulting from the assumption that the distribution of the statistic T is similar to the distribution of T^*, is much more critical. The coverage accuracy of the bootstrap confidence interval may be poor when the

two distributions differ substantially (Carpenter & Bithell, 2000). The coverage error of the non-studentized pivotal method has been shown to be of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ (Hall, 2013), where n is sample size.

2.3.2.2 Percentile method

Another commonly used bootstrap method is the percentile method introduced by Efron (1981). The rationale behind this approach is as follows. Suppose there exists a monotone transformation $g(\cdot)$ such that $\hat{\eta}^* - \hat{\eta} \sim \hat{\eta} - \eta \sim N(0, \sigma^2)$, where $\hat{\eta}^* = g(||\widehat{\boldsymbol{d}}^*||)$, $\hat{\eta} = g(||\widehat{\boldsymbol{d}}||)$ and $\eta = g(||\boldsymbol{d}||)$. Then, the one sided $1-\alpha$ interval for $||\boldsymbol{d}||$ is

$$(0, g^{-1}(\hat{\eta} - \sigma z_\alpha)) = (0, g^{-1}(F^{-1}_{\hat{\eta}^*}(1-\alpha)))$$

where $F_{\hat{\eta}^*}$ denotes the CDF of the random variable $\hat{\eta}^*$. Since a monotone transformation preserves quantiles, $g^{-1}(F^{-1}_{\hat{\eta}^*}(1-\alpha)) = g^{-1}(g(F^{-1}_{||\widehat{\boldsymbol{d}}^*||}(1-\alpha)))$. Therefore, the one-sided percentile interval is $(0, ||\widehat{\boldsymbol{d}}^*||_{1-\alpha}))$, where $||\widehat{\boldsymbol{d}}^*||_{1-\alpha}$ is the $1-\alpha$ quantile of $||\widehat{\boldsymbol{d}}^*||^{(1)}, \cdots, ||\widehat{\boldsymbol{d}}^*||^{(B)}$.

The validity of this method rests on the existence of $g(\cdot)$. However, for many problems an exact normalizing transformation will rarely exist. It has been shown that the coverage error can be substantial if the distribution of the estimate is not nearly symmetric (Efron & Tibshirani, 1994).

2.3.2.3 Bias corrected percentile method

The noted disadvantage of the percentile method led to the development of bias corrected (BC) method proposed by Efron (1987). Consider a monotone transformation $g(\cdot)$, such that $\hat{\eta}^* - \hat{\eta} \sim \hat{\eta} - \eta \sim N(-t\sigma, \sigma^2)$ for some constant t. The BC interval is $(0, F^{-1}_{||\widehat{\boldsymbol{d}}^*||}(\Phi(2t - z_\alpha)))$, where t is estimated by $\Phi^{-1}(\hat{P}(||\widehat{\boldsymbol{d}}^*|| \leq ||\widehat{\boldsymbol{d}}||))$ and Φ is the CDF of the standard normal distribution. The BC interval is still given by percentiles of the bootstrap samples, but the percentiles used are corrected for skewness and hence, provide improvement over the basic percentile approach in practice. In theory, the coverage errors of both the basic percentile interval and the BC interval are of order $\mathcal{O}(\frac{1}{\sqrt{n}})$ (Hall, 2013).

2.3.2.4 Normal method

The simplest way of constructing bootstrap confidence intervals is by assuming the underlying distribution of $||\widehat{\boldsymbol{d}}||$ is normal. This yields the normal interval $(0, ||\widehat{\boldsymbol{d}}|| + z_{1-\frac{\alpha}{2}}\hat{s})$, where $\hat{s}$ is the bootstrap estimate of the standard error. In view of the fact that there is no compelling evidence to assume normality for the L_2 norm or for some transformation of the L_2 norm, for example, the logarithm of the L_2 norm, the normal method may perform poorly in our context as later shown in Section 2.5.2 .

2.4 ASYMPTOTIC PROPERTIES

Obviously, the time efficiency, as well as the coverage accuracy of our proposed procedures, depends largely on the convergence of EM algorithm for normal mixtures and the construction of bootstrapping confidence intervals. For coverage accuracy, all the aforementioned bootstrap confidence interval methods we adopted are relatively computationally inexpensive and enjoy first-order accuracy (Hall, 1988).

It is well known that generally the EM algorithm can be trapped at local maxima, which are local maximum solutions of the likelihood function (Redner & Walker, 1984). The choice of initial values for the EM algorithm is of importance as to whether we will actually reach the global maxima. Practice has shown that it is preferable to start from several different initial values and then choose the solution that has the highest converged likelihood. Ma & Fu (2005) further studied the problem of correct convergence of the EM algorithm for normal mixtures, and found in both theory and practice that the EM algorithm can converge consistently to the true parameters when the starting point is suitably close to the true value. Moreover, the radius of this correct convergence starting point neighborhood becomes larger as the overlap of densities in a mixture becomes smaller. Therefore in practice, when overlap of the normal densities in the mixture is small enough and the sample size is large enough, the EM algorithm has a good chance to converge correctly. This, in turn implies that, our testing procedures should have satisfying performance in these cases. In the next section, we use simulations to demonstrate the empirical performance of our proposed procedures under

different degrees of overlap between the normal distributions in the mixture.

2.5 SIMULATIONS

In this section, we evaluate the performances of our testing procedure in both univariate and bivariate situations. The behaviors of our algorithms for cases ranging from easy-to-detect to hard-to-detect under different degrees of overlap between the normal distributions in the mixtures are illustrated.

2.5.1 Univariate simulation

Cluster analysis based on one variable is usually uninteresting in practice. The main purpose of doing simulations in the univariate case is to examine the testing characteristics of our procedures. We restrict our investigation to the case of two independent populations distributed as a mixture of two univariate normal distributions in Section 2.5.1. Hypotheses are expressible as hypotheses about shifts between four pairs of means as below with a pre-specified positive cluster equivalence margin Δ.

$$H_{01} : |\mu_1 - \mu_3| \geq \Delta \quad H_{02} : |\mu_1 - \mu_4| \geq \Delta$$
$$H_{03} : |\mu_2 - \mu_3| \geq \Delta \quad H_{04} : |\mu_2 - \mu_4| \geq \Delta$$

If a single test rejects the null hypothesis, then the true pairwise mean difference must be smaller than Δ. The chance of incorrectly declaring equivalence decreases as the true distances between the means increase. In the other words, the type I error rate for a single test decreases as the difference in means increases, so that the type I error is maximized when the true means are exactly Δ apart (Schuirmann, 1987) for a single comparison. Therefore, the scenario maximizing the Boneronni inequality bound on the family wise error rate (FWER) would be all the four pairs of means being exactly Δ apart.

However, it can be easily shown as follows that the scenario with four pairs of means being Δ apart is not attainable. Suppose there are four pairs of means being Δ apart in addition to the condition (2.3) which requires $\mu_1 > \mu_2$ and $\mu_3 > \mu_4$, then $|\mu_1 - \mu_4| = \Delta$ and

$|\mu_1 - \mu_3| = \Delta$ along with $|\mu_2 - \mu_4| = \Delta$ would lead to $\mu_3 = \mu_1 + \Delta, \mu_4 = \mu_1 - \Delta, \mu_2 = \mu_4 - \Delta$, resulting in $\mu_3 - \mu_2 = 3\Delta$ in contradiction of $|\mu_3 - \mu_2| = \Delta$. It can also be shown that the scenario with three pairs of means being Δ apart may not be reasonable. Suppose that there are three pairs of means being Δ apart, for example, H_{01}, H_{02} and H_{03} hold. By Triangle inequality, $2\Delta = |\mu_1 - \mu_3| + |\mu_1 - \mu_4| \geq |\mu_3 - \mu_4|$ and $2\Delta = |\mu_1 - \mu_3| + |\mu_2 - \mu_3| \geq |\mu_1 - \mu_2|$, i.e., the means of the two normal components within each population are at most 2Δ apart. However if the difference between the means of two clusters is less than Δ, the two clusters would be considered by our standard to be equivalent. Thus, we would want the distance between means of two clusters in the same population to be considerably larger than Δ for a reasonably separated mixture of normal distributions. With this in mind, an appropriate practical Δ should be chosen by the investigator in recognition of anticipated separation between the two normal components in the mixtures of normal distributions. In our motivating data, we are interested in examining similarities in the profiles of subtypes in populations with different mental disorders. By the formulation, two subtypes, regardless of which population they are identified from, would be considered equivalent if the difference between them is less than the cluster equivalence margin Δ. Therefore, we would want scientists to take account of the anticipated difference between subtypes identified in the same mental disorders in determining an appropriate cluster equivalence margin Δ.

Hence, in our simulations, we consider a "worst" case scenario as the configuration of two pairs of means being Δ apart. We study the FWER in the scenarios corresponding to $\mu_3 - \mu_1 = \delta$, $\mu_4 - \mu_2 = \delta$ setting δ equal to the cluster equivalence margin Δ. Suppose $\mu_1 = 1$, $\sigma = 1$, $\pi_1 = 0.4$, $\pi_2 = 0.3$ and $n_1 = n_2 = 500$. The separations of the normal components can be assessed by $h = |\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$ if the two components have means μ_1, μ_2 and variances σ_1^2, σ_2^2 respectively (McLachlan & Peel, 2004). Schilling et al. (2002) gives the values of h that separate unimodal and bimodal mixtures of normal distribution for specific values of the mixture proportion π_1 and σ_1/σ_2. In our setting, if $\mu_2 < -1.44$ then the mixture is bimodal. Three different true parameter values for μ_2 are chosen representing cases of well-separated (Well-sep), medium-separated(Med-sep), and poorly-separated (Poor-sep) mixture of normals: $\mu_2 = -3, -1.8, -1.44$ respectively. Under the null hypothesis, let $\mu_3 = \mu_1 + \delta$, $\mu_4 = \mu_2 + \delta$ and $\delta = \Delta$ with varying Δ taking values in $0.4, 0.5, 0.6$. It can be checked that

the distance between means of two clusters in the same population is at least 3 times greater than the cluster equivalence margin in our simulation settings. It has been suggested that, for $90\% - 95\%$ confidence intervals, the number of bootstrap samples B should be between 1000 and 2000 (Davison & Hinkley, 1997; Efron & Tibshirani, 1994; Carpenter & Bithell, 2000). In the simulation study, the bootstrap standard errors of the estimates are obtained based on 1000 bootstrap samples.

Simulation results displayed in Table 2.1 and Table 2.2 are based on 500 datasets generated for each parameter configuration. The simulation results presented in Table 2.1 show that the type I error rates are all controlled at the nominal level. The fact that type I error rates are far below than 0.05 can be explained by the conservativeness of TOST and the Bonferroni multiplicity adjustment. We also assess the performance of the maximum likelihood estimates obtained via the EM algorithm and the empirical coverage probability of the bootstrap confidence interval in Table 2.2. Bias and the mean square error of the estimates of $\mu_3 - \mu_1$ over 500 replications are computed given Δ under each setting. It is not surprising to find out that estimates from the EM algorithm have larger variance, therefore larger MSE, as normal components become less separated. In general, the empirical coverage rates are greater than the nominal level due to conservativeness of TOST.

Power under a set of alternative hypotheses based on 200 datasets are summarized in Table 2.3. Under the alternative hypothesis, we considered $\mu_3 = \mu_1 + \delta$, $\mu_4 = \mu_2 + \delta$ with $\delta = 0, 0.1, 0.3$ and $\mu_2 = -1.8$ in the scenario of medium-separated mixture of normals for power evaluation. It can be seen from Table 2.3 that given the cluster equivalence margin Δ, higher power is attained for smaller δ for the reason that as the true distance between two clusters defined in independent populations becomes smaller, it is more favorable to the alternative hypothesis, thereby resulting in greater power. On the other hand, when the true distance between clusters δ are fixed, larger Δ leads to higher power because larger values of the maximum allowable difference between any two clusters that would be viewed equivalent makes detection of common clusters easier. In Section 2.A.1, we demonstrate that power increases with $\Delta - \delta$.

Another important concern investigators and practitioners may raise is when common clusters are detected, how likely are we to conclude wrong pairs of common clusters. It can be

theoretically demonstrated that the probability is lower than 0.05 as Bonferroni multiplicity adjustment controls family error rate (FWER) in the strong sense (Dmitrienko et al., 2009), meaning that the chance of rejecting one or more correct null hypotheses is always less than α regardless of which and how many of the null hypotheses are correct. Here we use simulations to assess the empirical rejection rates of any subset of true null hypotheses conditional on correctly rejecting the overall null hypothesis. The simulation settings we consider are as follows: let $\mu_1 = 1$, $\sigma = 1$, $\mu_1 - \mu_2 = 2\Delta, \mu_4 - \mu_2 = \Delta$ and $\mu_3 - \mu_1 = \delta$ with δ taking values in $\frac{1}{4}\Delta, \frac{2}{4}\Delta, \frac{3}{4}\Delta$. Clearly $|\mu_1 - \mu_4|$ and $|\mu_2 - \mu_4|$ are on the boundary of the null hypotheses H_{02} and H_{04}, while $|\mu_1 - \mu_3|$ is in the alternative space of H_{01}. Note that in addition to being the cluster equivalence margin, Δ also determines the degree of separation between the normal distributions in the mixtures. It is direct to show that the mixture for the first population is bimodal if $\Delta > 1.22$ so that we let $\Delta = 1.2, 1.3$. Considering that the test has lower power when δ is approaching to the cluster equivalence margin Δ, we didn't consider δ value larger than $\frac{3}{4}\Delta$.

The results are given in Table 2.4. It can be seen that out of 200 simulations, there was no case of making wrong rejections. However, notice that when δ is approaching to $\frac{3}{4}\Delta$ and $\Delta = 1.2$, there are a handful of cases where the test statistic for testing $|\mu_1 - \mu_4|$, i.e., $|\widehat{\mu}_1 - \widehat{\mu}_4| + z_{\alpha/4}\widehat{\sigma}_{\widehat{\mu}_1 - \widehat{\mu}_4}$ are borderline yet not crossing Δ yet. In order to get a better sense of the likelihood of making wrong rejections, in Appendix 2.A.2, a exploratory analysis was conducted by increasing the significance level α beyond 0.05 and looking for points at which wrong rejections are made. Based on the results, we conclude that practitioners generally do not need to worry about claiming wrong pairs of common clusters using the proposed testing procedures.

Overall, the findings of the simulation studies suggest that our proposed test is able to control Type I error at the nominal level regardless of varying levels of overlap between normal components. The power of the test strongly depends on the size of Δ relative to other model parameters. In practice, however, the maximum allowable difference between any two clusters that from the scientific view can be ignored should be chosen in consultation with scientists familiar with the context of the data.

Table 2.1: Simulated type I error rate in univariate case. 500 datasets are simulated for each parameter configuration. True parameters used to generate the samples are set as: $f(x_i) = 0.4\phi(x_i; 1, 1) + 0.6\phi(x_i; \mu_2, 1), i = 1, \cdots, 500$ with $\mu_2 = -3, -1.8, -1.44$ and $f(y_j) = 0.3\phi(y_j; 1 + \delta, 1) + 0.7\phi(y_j; \mu_2 + \delta, 1), j = 1, \cdots, 500$ with $\delta = \Delta = 0.4, 0.5, 0.6$.

	$\Delta = 0.4$	$\Delta = 0.5$	$\Delta = 0.6$
Well-sep ($\mu_2 = -3$)	0.014	0.014	0.012
Med-sep ($\mu_2 = -1.8$)	0.006	0.006	0.008
Poor-sep ($\mu_2 = -1.44$)	0.006	0.0014	0.008

Table 2.2: Simulation results for the estimates of the $\mu_3 - \mu_1$. 500 datasets are simulated for each parameter configuration (see Table 2.1). $\widehat{|Bias|}$ denotes absolute value of the estimated bias. $\widehat{MSE}$ is the estimated mean square error of the estimates. $\widehat{Coverage}$ is the empirical coverage probability of the 97.5% bootstrap confidence interval.

		$\Delta = \delta = 0.4$	$\Delta = \delta = 0.5$	$\Delta = \delta = 0.6$
Well-sep	$\widehat{\lvert Bias \rvert}$	0.006	0.004	0.010
	$\widehat{MSE}$	0.0162	0.015	0.0162
	$\widehat{Coverage}$	0.984	0.982	0.990
Med-sep	$\widehat{\lvert Bias \rvert}$	0.002	0.006	0.003
	$\widehat{MSE}$	0.0249	0.0240	0.0285
	$\widehat{Coverage}$	0.984	0.984	0.972
Poor-sep	$\widehat{\lvert Bias \rvert}$	0.005	0.009	0.001
	$\widehat{MSE}$	0.0288	0.0378	0.0293
	$\widehat{Coverage}$	0.988	0.970	0.982

Table 2.3: Power evaluation. 200 datasets are simulated for each parameter configuration. True parameters used to generate the samples are set as: $f(x_i) = 0.4\phi(x_i; 1, 1) + 0.6\phi(x_i; -1.8, 1), i = 1, \cdots, 500$ and $f(y_j) = 0.3\phi(y_j; 1 + \delta, 1) + 0.7\phi(y_j; \mu_2 + \delta, 1), j = 1, \cdots, 500$ with $\delta = 0, 0.1, 0.3$.

	$\Delta = 0.4$	$\Delta = 0.5$	$\Delta = 0.6$
$\delta = 0$	0.32	0.76	0.95
$\delta = 0.1$	0.275	0.725	0.92
$\delta = 0.3$	0.055	0.185	0.44

Table 2.4: Rejection rates evaluation: P_1 =P(reject at least one H_{0i}), P_2 =P(reject at least one $H_{0i}, i \in T$|reject at least one H_{0i}) where T denotes the index set of the true null hypotheses. 200 datasets are simulated for each parameter configuration. True parameters used to generate the samples are set as: $f(x_i) = 0.4\phi(x_i; 1, 1) + 0.6\phi(x_i; \mu_2, 1), i = 1, \cdots, 500$ with $\mu_2 = 1 - 2\Delta$ and $f(y_j) = 0.3\phi(y_j; 1 + \delta, 1) + 0.7\phi(y_j; \mu_2 + \delta, 1), j = 1, \cdots, 500$ with $\delta = \frac{1}{4}\Delta, \frac{2}{4}\Delta, \frac{3}{4}\Delta$ and $\Delta = 1.2, 1.3$.

		P_1	P_2
$\Delta = 1.2$	$\delta = \frac{1}{4}\Delta$	0.795	0
	$\delta = \frac{1}{2}\Delta$	0.575	0
	$\delta = \frac{3}{4}\Delta$	0.200	0
$\Delta = 1.3$	$\delta = \frac{1}{4}\Delta$	0.885	0
	$\delta = \frac{1}{2}\Delta$	0.785	0
	$\delta = \frac{3}{4}\Delta$	0.275	0

2.5.2 Simulations in bivariate case

We evaluate and compare heuristic performances for bivariate data of the TOST-based method and the selected bootstrap confidence intervals approaches: non-Studentized pivotal method (pivotal), percentile method (percentile), bias-corrected(BC) percentile method, normal method assuming normality of L_2 norm(Normal1), normal method assuming normality of log(L_2 norm) (Normal2) in bivariate case. All the previously chosen bootstrap confidence interval methods are relatively computationally inexpensive and accessible to practitioners. Theoretically, they all have the same first-order accuracy (Hall, 1988). Our goal in this section is to evaluate their empirical performance using simulations and in order to try to recommend the methods that work well in our applications.

In the simulation study, we set $\pi_1 = 0.5$, $\pi_2 = 0.5$ and $n_1 = n_2 = 500$,

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 2.5 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_3 = \begin{pmatrix} 1+\delta \\ 2.5 \end{pmatrix}, \boldsymbol{\mu}_4 = \begin{pmatrix} -\delta \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We consider the Kullback-Leibler divergence between two distributions as the measure of separation between two bivariate normal distributions. The Kullback-Leibler divergence between two p-multivariate normal distributions $P(\boldsymbol{x})$ and $Q(\boldsymbol{x})$ with means $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ and covariance matrix Σ_1 and Σ_2 is defined as

$$KL(P,Q) = \frac{1}{2}\left\{\log\frac{|\Sigma_2|}{|\Sigma_1|} - p + Tr[\Sigma_2^{-1}\Sigma_1] + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right\}.$$

In our setting, $KL(P,Q) = \frac{1}{2}\left\{\frac{7.25 - 5\rho}{1-\rho^2} - 1\right\}$, which is monotonically decreasing in ρ if $\rho \leq 0$. Thus two different true parameter values for ρ are chosen to represent cases of different degree of overlap in the mixture of two normals: $\rho = -0.7, 0$, respectively.

In a similar manner to the univariate case, we study the FWER in the scenarios corresponding to $||\boldsymbol{\mu}_3 - \boldsymbol{\mu}_1|| = \delta = \Delta$, $||\boldsymbol{\mu}_4 - \boldsymbol{\mu}_2|| = \delta = \Delta$. We note that there are other points in the null space as extreme or more extreme than the point we chose. However, the chosen point is reasonably extreme and appealing to one's intuition. As shown in Table 2.5, most empirical type I errors are below the nominal level except for the pivotal method. Two normal methods and TOST-based method have empirical type I error rates equal to zero in most of the cases, indicating the conservativeness of these methods.

Powers are evaluated under the alternative hypothesis $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_4 = \boldsymbol{\mu}_2$ for cases ranging from easy-to-detect $\Delta = 0.6$ to hard-to-detect $\Delta = 0.2$ in Table 2.6. Similar to the univariate case, larger Δ leads to higher power. Generally, the pivotal method enjoys the highest power, and then followed by the Bias-Corrected percentile method, and the two normal methods. In almost all the cases, the TOST-based method and the percentile method have the lowest power. All the selected bootstrap confidence intervals approaches but the basic percentile method outperform the TOST-based method, indicating that the idea of evaluating the overall similarity of two clusters generally boost power substantially compared to examining commonality for each variable independently across clusters. Among the selected bootstrap confidence intervals methods, the pivotal method and Bias-Corrected percentile method perform well in terms of power. Taking into account the fact that the pivotal method may fail to protect type I error rate, the Bias-corrected method may be more robust and desirable and thereby recommended.

Table 2.5: Type I error rates evaluation in bivariate case. 500 datasets are simulated for each parameter configuration and the six methods are applied to the same data sets.

		Pivotal	Percentile	BC	Normal1	Normal2	TOST
$\rho = -0.7$	$\delta = \Delta = 0.2$	0.08	0	0.06	0	0.006	0
	$\delta = \Delta = 0.4$	0.006	0	0.02	0	0	0
	$\delta = \Delta = 0.6$	0.006	0	0.01	0	0	0
$\rho = 0$	$\delta = \Delta = 0.2$	0.12	0	0.07	0	0.01	0
	$\delta = \Delta = 0.4$	0	0	0.02	0	0	0
	$\delta = \Delta = 0.6$	0	0.006	0	0	0	0.004

Table 2.6: Power evaluation in bivariate case. 200 datasets are simulated for each parameter configuration.

		Pivotal	Percentile	BC	Normal1	Normal2	TOST
$\rho = -0.7$	$\Delta = 0.2$	0.77	0	0.56	0	0.18	0
	$\Delta = 0.4$	1	0	0.92	0.1	0.58	0.085
	$\Delta = 0.6$	1	0.66	1	0.98	0.88	1
$\rho = 0$	$\Delta = 0.2$	0.54	0	0.39	0	0.12	0
	$\Delta = 0.4$	0.94	0	0.88	0.13	0.39	0.015
	$\Delta = 0.6$	1	0.4	1	0.95	0.77	0.96

2.6 APPLICATION

In this section, we illustrate the cluster identification methods described in Section 2.3.2 by applying them to one data example: an interesting GABA neuron-related biomarker study that strongly motivated our research. We start with giving an overview of the studies and descriptions of data. We then discuss the application of out methodology to the data.

2.6.1 GABA neuron-related biomarker study

2.6.1.1 Overview of the published studies

Previously Volk et al. (2012) identified a subset of schizophrenia subjects that consistently showed deficits in certain GABA neuron-related mRNAs obtained from post-mortem brain tissues: GABA synthesizing enzyme glutamate decarboxylase (GAD67), calcium-binding protein parvalbumin (PV), neuropeptide somatostatin (SST) and Lhx6 which plays a critical role in the specification, migration, and maturation of neurons that express PV or SST. This subset of subjects was termed as the Low-GABA-Marker (LGM) molecular phenotype.

In a more recent study, Volk et al. (2016) quantified transcript levels for GAD67, PV, SST, and Lhx6 in the prefrontal cortex area of 184 subjects with a diagnosis of schizophrenia

($n = 39$), schizoaffective disorder ($n = 23$), bipolar disorder ($n = 35$), or control subjects ($n = 87$). Absence of any psychiatric diagnoses were confirmed in control subjects.

In this recent study, each subject with a psychiatric disorder was matched individually to one control subject by gender and as closely as possible for age. Ten control subjects were previously used as matching subjects in the published studies for bipolar disorder (Sibille et al., 2011; Kimoto et al., 2015) and also studies for schizophrenia or schizoaffective disorder (Volk et al., 2012, 2014) therefore the same pairing was retained. Samples from subjects in a pair were prepared and processed together in a blinded fashion throughout all stages of the study in order to control experimental variation. To account for significant effects of covariates, SST mRNA levels were adjusted by age and brain pH; PV mRNA levels were adjusted by brain pH; and GAD67 mRNA levels by brain pH. Further to account for varying scales among the four mRNAs, standardized mRNA levels of pH-adjusted GAD67, pH-adjusted PV, age and pH-adjusted SST were computed for all subjects by subtracting the overall mean and then dividing by the overall standard deviation. More details can be found in Volk et al. (2016).

In Volk et al. (2016), a cluster analysis was conducted using the standardized and appropriately adjusted GAD67, PV, SST, and Lhx6 expression levels based on all 184 subjects. The purpose of this analysis was to identify possible clusters among the subjects with schizophrenia, schizoaffective, bipolar and control subjects. The goal was to see if one cluster was connected to the subject with schizophrenia, schizoaffective disorder and bipolar disorders and not apparent in controls. The Ward method (Ward Jr, 1963) was used to do hierachical clustering on all 184 subjects. Two clusters were identified in their paper as displayed in Figure 2.1. One cluster was composed of 140 subjects (57 subjects with disorder and 83 control subjects who were generally intermixed), the other cluster of 44 subjects consisted mostly of subjects with a disorder ($n = 41$) and only 3 control subjects. It was found that the cluster with 44 subjects expresses low levels of GABA markers: mean adjusted transcript levels were lower for GAD67 ($-30\%; t_{182} = -14.5, p < .00001$), PV ($-28\%; t_{182} = -8.4, p < .00001$), SST ($-48\%; t_{182} = -12.7, p < .00001$), and Lhx6 ($-23\%; t_{182} = -10.2, p < .00001$) relative to the other cluster, which is consistent with the previous identification of LGM molecular phenotype identified by Volk et al. (2012) .

As further noted by Volk et al. (2016), excluding the control subjects and repeating the Ward cluster analysis on the 97 subjects with psychiatric disorder using the same adjusted mRNA values confirmed the presence of two unique clusters (Figure 2.2) of psychiatric disorder subjects with and without the LGM molecular phenotype: mean adjusted transcript levels were lower for GAD67 ($-32\%; t_{95} = -13.0, p < .0001$), PV($-24\%; t_{95} = -6.2, p < .0001$), SST ($-46\%; t_{95} = -8.1, p < .0001$), and Lhx6 ($-27\%; t_{95} = -9.6, p < .0001$) in the LGM phenotype relative to the non-LGM phenotype. There were 46.2% (18/39) of schizophrenia subjects, 47.8% (11/23) of schizoaffective disorder subjects, and 28.6% (10/35) of bipolar disorder subjects who were classified as having the LGM molecular phenotype, for a total of 39 subjects with a disorder.

These findings suggest that the subtype characterized by LGM molecular and the subtype characterized by non-LGM molecular could identify subsets of subjects from each of these diagnostic groups.

Figure 2.1: Two clusters identified in (Volk et al., 2016) based on 184 subjects.

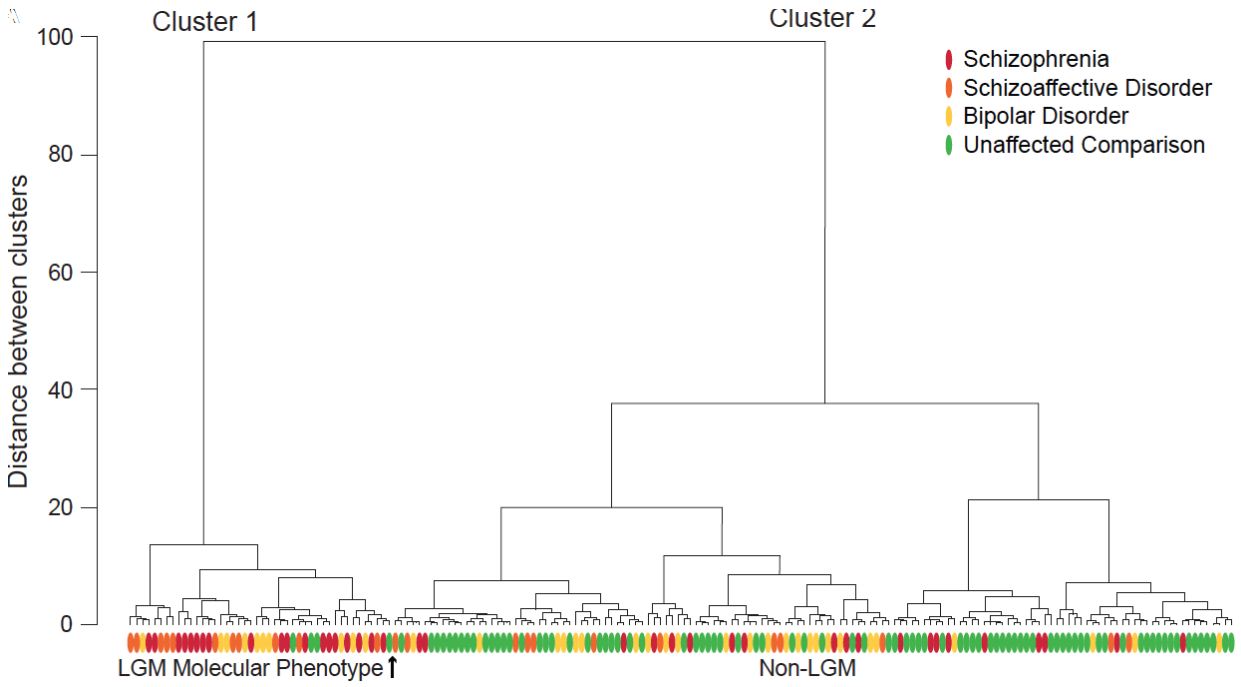
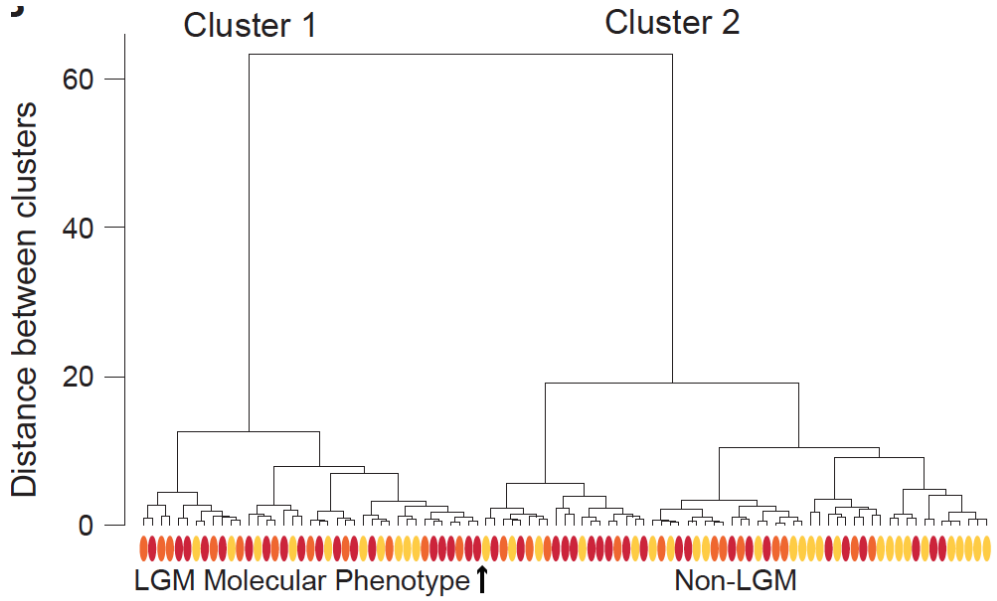


Figure 2.2: Two clusters identified in (Volk et al., 2016) based on 97 subjects with psychiatric disorder.



2.6.1.2 Using proposed testing procedures

As introduced in the proceeding section, the findings in Volk et al. (2016) suggest there may exist common clusters in subjects with different disorders. A major advantage of our approaches is that we provide a way to evaluate the strength of the statistical evidence in favor of the findings of existence of common clusters in independent populations. In this section, we apply our testing procedures using all four mRNAs of the 97 subjects with the psychiatric disorders. To simplify our hypothesis formulations, we pool schizophrenia and schizoaffective disorder subjects into one population. The presence of any subtype among the different schizophrenia-related disorders can be further evaluated in future work. The same adjusted mRNA values are used here to maintain the comparability of the results obtained from the proposed method with those published in Volk et al. (2016). In addition, standardization is again employed to account for varying scales among the four mRNAs.

In our testing frame, it is appropriate to assume that the observations of subjects with bipolar disorder and the observations of subjects with schizophrenia or schizoaffective dis-

order are sampled from two independent populations, population 1 and population 2. We further assume based on the earlier results (Volk et al., 2012) that each population is distributed as a mixture of two normals. We fit a two-component normal mixture model with common component-covariance matrices to each sample. The estimated mean of each component and the pairwise Euclidean distance between means are displayed in Table 2.7. Similar to what Volk et al. (2016) found, it is shown that the cluster with mean $\boldsymbol{\mu_2}$ and the cluster with mean $\boldsymbol{\mu_4}$ are characterized by lower expression levels for each gene in comparison with the other cluster identified from the same population of subjects. Based on the pairwise difference, we see that the cluster with mean $\boldsymbol{\mu_1}$ identified in the subjects with bipolar disorder seems to be close to the cluster with mean $\boldsymbol{\mu_3}$ identified in the subjects with a schizophrenia disorder, and the cluster with mean $\boldsymbol{\mu_2}$ and the cluster with mean $\boldsymbol{\mu_4}$ also look very similar.

Table 2.7: The estimated mean of each cluster and the pairwise differences.

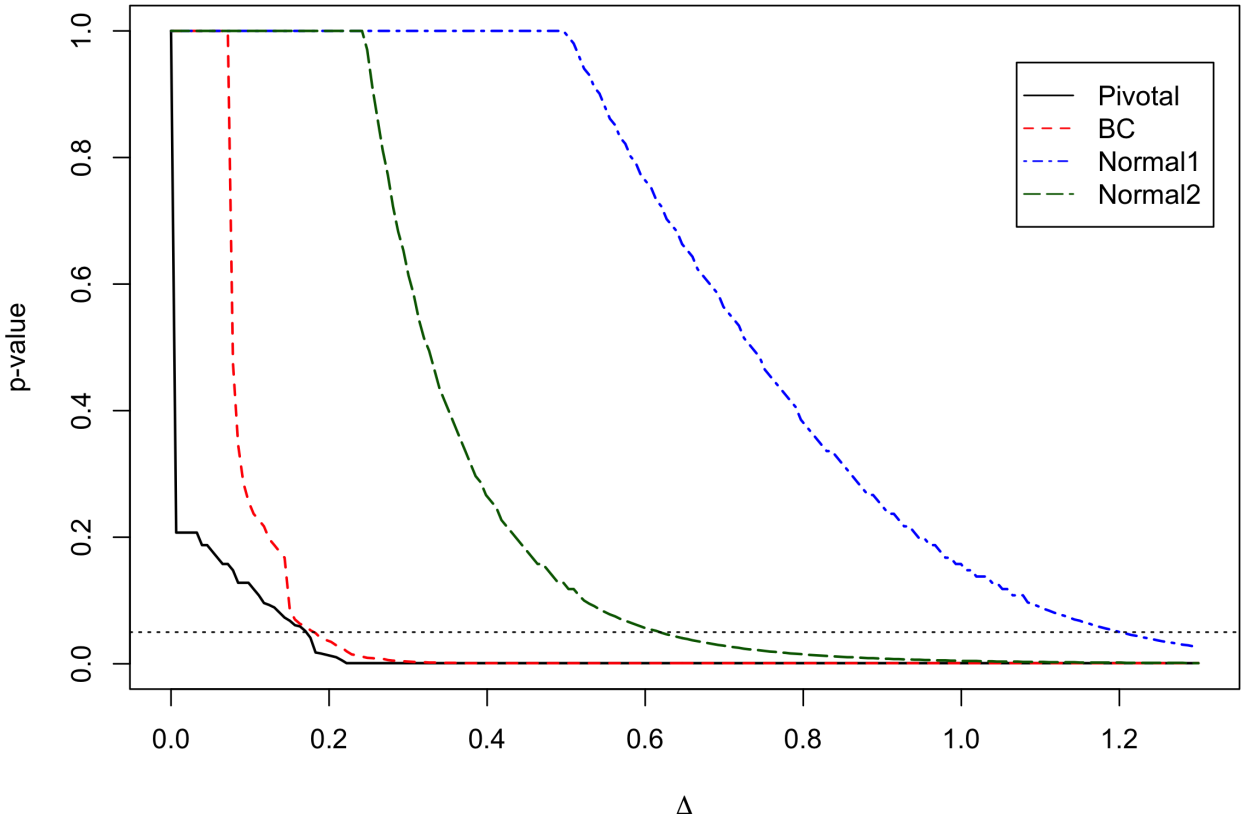
$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	$\boldsymbol{\mu_3}$	$\boldsymbol{\mu_4}$	Pairwise Euclidean distance of means
$\begin{bmatrix} 0.531 \\ 0.247 \\ 0.477 \\ 0.441 \end{bmatrix}$	$\begin{bmatrix} -1.086 \\ -0.505 \\ -0.976 \\ -0.902 \end{bmatrix}$	$\begin{bmatrix} 0.609 \\ 0.280 \\ 0.592 \\ 0.523 \end{bmatrix}$	$\begin{bmatrix} -0.914 \\ -0.421 \\ -0.890 \\ -0.785 \end{bmatrix}$	$\begin{bmatrix} 0 & 2.66 & 0.16 & 2.43 \\ & 0 & 2.83 & 0.24 \\ & & 0 & 2.59 \\ & & & 0 \end{bmatrix}$

The next step is to apply the proposed testing procedures to four mRNAs. We are faced with the different choices of bootstrap confidence interval approaches. As shown in Section 2.5.2, the confidence interval approaches using the pivotal method (Pivotal), the Bias-corrected method (BC), and the two normal methods(Normal1, Normal2) demonstrate relatively better power as compared to the TOST-based method and percentile method in the simulations. We illustrate their use in these data. In hypothesis testing, the p-value can be conceptualized as the smallest α level that would lead to a rejection of no common clusters. Therefore, for any given value of $\Delta \in (0, \infty)$, one can find the p-value by incrementally increasing α level from 0 until rejection is achieved, that is, until Δ is contained in the $100(1-\frac{\alpha}{4})\%$ bootstrap confidence interval of the L_2 norm of the difference between one pair

of means.

Figure 2.3 displays plots of p-values versus Δ for the selected bootstrap methods. The horizontal dotted line indicates where the p-values equal to 0.05. Then the smallest Δ leading to declaring common clusters at significance level 0.05 using each method is characterized by the corresponding Δ of the intersection point of each curve and the straight line $p = 0.05$. The respective values are 0.176, 0.18, 0.62 and 1.2, from left to right, obtained from Pivotal, Bias-corrected, normal method assuming normality of log(L2 norm) (Normal2), and normal method assuming normality of L2 norm (Normal1) respectively.

Figure 2.3: Plots of p-value versus Δ for PV, BC, Normal1 and Normal2 methods.



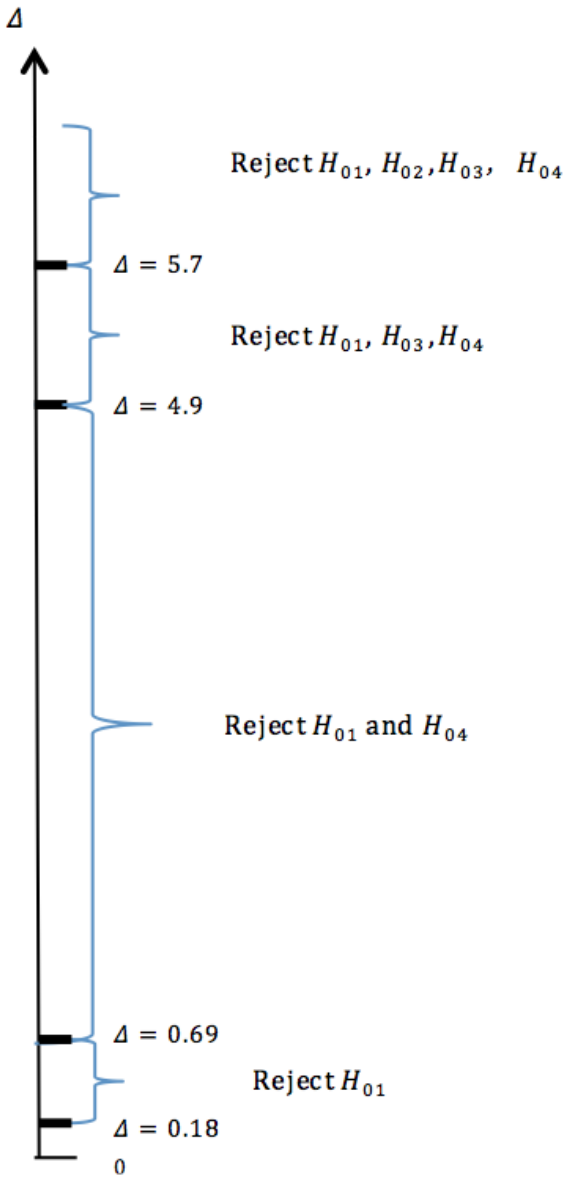
The smaller Δ is, the more stringent is the definition of the common cluster. Hence, rejecting the null hypothesis requires stronger evidence of the existence of common clusters. The findings that pivotal method and Bias-corrected method would reject the null hypothesis with relatively small Δ are consistent with our simulation results that show these two

methods have relatively higher power to detect common clusters. In the following discussion, we focus on the results from Bias-corrected method, as we recommended in Section 2.5.2.

For our data, at significant level 0.05, provided that two clusters can be considered as common clusters if the L_2 norm (i.e., Euclidean distance) between their means differs by less than 0.2, our approach concludes that we can identify at least one common cluster in subjects with bipolar disorder and subjects with schizophrenia. We note that the choice of Δ should be determined a priori by the subject-matter experts. Nevertheless, we argue that $\Delta = 0.2$ is a fairly stringent margin. With means of two clusters being 0.2 apart, the difference between each gene is at most 0.2, which is only one fifth of the standard deviation.

Among the four pairs of comparisons in our hypothesis testing, rejection of any comparison can lead to the rejection of the overall hypothesis. A natural question would be which pair(s) of clusters we deem common given that we conclude the existence of common clusters. In answering that, we further examine which null hypotheses of $H_{01}, H_{02}, H_{03}, H_{04}$ are rejected at significance level 0.05 given some Δ that is greater than 0.18. The results are provided in Figure 2.4. When Δ is between 0.18 and 0.69, the testing yields rejecting H_{01}, meaning that, the cluster with mean $\boldsymbol{\mu}_3$ and the cluster with mean $\boldsymbol{\mu}_1$ are recognized as common, which is the correct identification of the non-LGM phenotype. When Δ is between 0.69 and 4.9, the testing yields rejections of H_{01} and H_{04}, meaning that, in addition to the previous identified common cluster, the cluster with mean $\boldsymbol{\mu}_2$ and the cluster with mean $\boldsymbol{\mu}_4$ are also deemed common, which would support the LGM phenotypes being common. Further increasing Δ would lead to more rejections, but we argue that these would not be meaningful.

Figure 2.4: Schematic of Δ vs decision.



2.6.1.3 Summary of findings

In summary, we apply several bootstrap confidence interval approaches to the mRNAs measured in the GABA Neuron-Related biomarker study to assess if any subtype is shared in subjects with bipolar disorder and subjects with schizophrenia. Using the pivotal method

and Bias-corrected method one would reject the null hypothesis and declare existence of common clusters with a small cluster equivalence margin Δ. We further examine the results from the Bias-corrected method under different values of Δ. With a rather stringent choice of Δ, say 0.2, we would conclude a subset (cluster) featuring non-LGM molecular phenotype could identify a subset of subjects from each of the two diagnostic groups. With a slightly relaxed choice of Δ, for example, 0.7, we would additionally conclude a subset (cluster) featuring LGM molecular phenotype could identify a subset of subjects from each of the two diagnostic groups. In such case, we have statistically significant evidence at $\alpha = 0.05$, to show that the subtype characterized by LGM molecular and the subtype characterized by non-LGM molecular could identify subsets of subjects with bipolar disorder and subjects with schizophrenia.

2.7 CONCLUSIONS AND FUTURE WORK

2.7.1 Conclusions

In this chapter, we develop some methodologies to assess whether or not clusters identified from independent populations exhibit commonalities. There appears to be little literature that considers this problem. As an initial step in the research, we formulate our hypotheses by using concepts from bioequivalence issues in biopharmaceutical research combined with a finite normal mixture framework. Our layout of the formulations allows for univariate and multivariate data.

We first propose a testing procedures for univariate data based on the idea of the two one-sided test (TOST) that has been used in the analysis of pharmaceutical bioequivalence trials. The proposed test is directly extendable to multivariate data by the intersection-union method. The drawback of this multivariate approach is that it fails to account for correlations among the variables used in the cluster analysis, so that the intersection-union method can be very conservative. We show that it suffers from low power even for bivariate data. To address this issue, we propose to use the L_2 norm as the distance measure for multivariate data to establish the overall similarity of two clusters. We realize identifying an exact confidence interval for this measure is complex. We then propose to use a methodology based on bootstrap confidence intervals. We provide an illustration of the use of four selected bootstrap methods in the context of our testing problem. We show through multivariate simulations that all but one of the selected bootstrap confidence intervals outperform TOST-based method, indicating that the idea of evaluating the overall similarity between two clusters based on L_2 norm boosts power substantially.

Finally, we use our motivating data application to illustrate the use of the proposed tests in a biomarker study setting. In the GABA neuron-related biomarker study, we successfully confirm that the subtype characterized by LGM molecular and the subtype characterized by non-LGM molecular found by Volk et al. (2016) identify subsets of subjects with bipolar disorder and subjects with schizophrenia with reasonably chosen meaningful cluster equivalence margins.

2.7.2 Future work

Although many methods for multiple hypothesis testing have been developed, direct application to our context may not provide any tangible benefits or may require additional constraints. We note that there are some recent efforts (Röhmel, 2011; Lauzon & Caffo, 2009; Caffo et al., 2013) in studying multiplicity control in equivalence testing with three or more treatments. The scenario they considered is the clinical trial set up to show equivalence between all pairs of equivalent treatments. As our testing procedures involve testing cluster equivalence with clusters identified from different populations, we may be able to adapt their work to the case where the number of populations is greater than two.

Additionally, as we introduced in the very beginning, the proposed method applies to the-finite normal mixture framework. Although normal mixture models are widely used to model the distributions of a variety of random phenomena, in practice, data showing deviations from mixture of normals are inevitable. In the future, we would like to investigate the robustness of the proposed tests more thoroughly. For example, we could simulate observations from a mixture of Student t-distributions that are known to have fatter tails than the normal distributions, and assess how sensitive the proposed tests are to the deviations from the assumption of mixture of normals. We can also use simulations to assess the performance of the tests when the number of clusters is mis-specified.

Another direction we can explore is to extend mixture of normals to mixture of other exponential families distributions that may also have been extensively used in applications. As the very first step, one needs to concern the identification problems. Lack of identifiability happens when mixing two distributions from a parametric family just yields a third distribution from the same family. For example, it was demonstrated in (McLachlan & Peel, 2004) that the mixture of binomials is always just another binomial. Once identifiability is established, the EM algorithms can be used for maximum likelihood estimation in finite mixture models. A similar analysis flow as described in Section 2.3.1 can be utilized. It can be anticipated that the challenge lies in determining reasonable rejections regions, which may require a tremendous amount of research.

2.A APPENDIX

2.A.1 Examining relationship between power and $\Delta - \delta$

To simplify the presentation, below we derive the formula for computing power of testing $H_{01} : |\mu_1 - \mu_3| \geq \Delta$ given that $\mu_1 - \mu_3 = \delta$. It can be demonstrated that power increases with the size of $\Delta - \delta$ under some assumptions.

Assume $\hat{\mu_3} - \hat{\mu_1} \sim N(\mu_1 - \mu_3, \sigma^2_{\hat{\mu_3} - \hat{\mu_1}})$, the power of TOST evaluated at $|\mu_1 - \mu_3| = \delta$ is given by

$$P(|\hat{\mu_3} - \hat{\mu_1}| \leq \Delta - z_\alpha \sigma_{\hat{\mu_3} - \hat{\mu_1}} | \mu_1 - \mu_3 = \delta) = \Phi(\frac{\Delta - \delta}{\sigma_{\hat{\mu_3} - \hat{\mu_1}}} - z_\alpha) - \Phi(\frac{-\Delta - \delta}{\sigma_{\hat{\mu_3} - \hat{\mu_1}}} + z_\alpha).$$

When Δ, δ are relatively larger than $\sigma_{\hat{\mu_3} - \hat{\mu_1}}$, $\Phi(\frac{-\Delta - \delta}{\sigma_{\hat{\mu_3} - \hat{\mu_1}}} + z_\alpha) \approx 0$, then the power is approximately equal to $\Phi(\frac{\Delta - \delta}{\sigma_{\hat{\mu_3} - \hat{\mu_1}}} - z_\alpha)$. Obviously, power of the TOST, so as the overall testing procedure, increase with the size of $\Delta - \delta$.

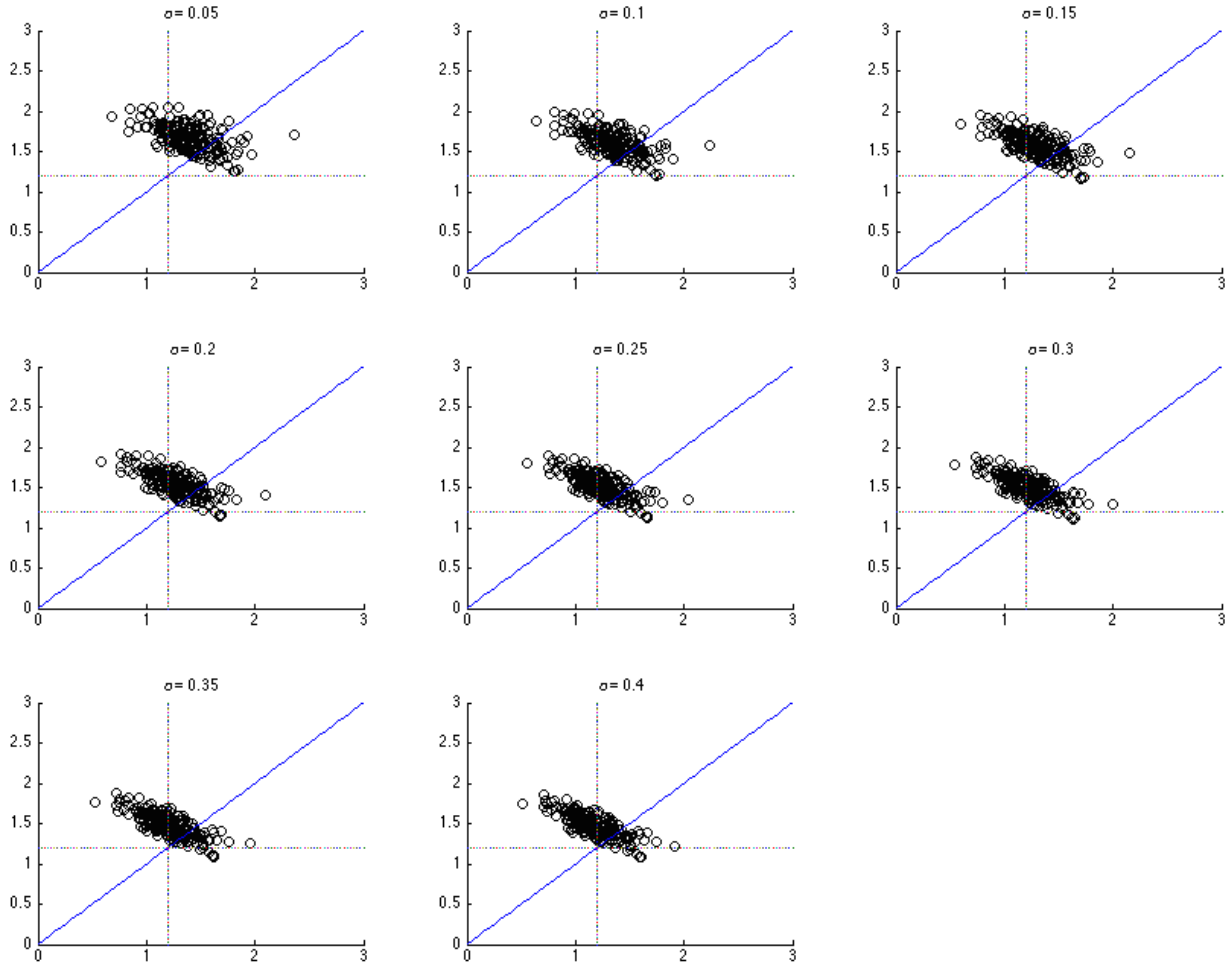
2.A.2 Likelihood of claiming wrong pairs of common clusters

As described in Section 2.5.1, we further explored the likelihood of claiming wrong pairs of common clusters for the simulation with $\delta = \frac{3}{4}\Delta$ and $\Delta = 1.2$ by increasing the significance level α beyond 0.05. For each α, 200 replicates of the test statistics for testing $|\mu_1 - \mu_4|$ are plotted against the test statistics for testing $|\mu_1 - \mu_3|$ in Figure 2.A.1 with diagonal line representing $y = x$ superimposed on them. The vertical line representing $x = \Delta$ and horizontal line indicating $y = \Delta$ divide the $x - y$ plane into four quadrants. Points lying in the upper-left quadrant correspond to the truth that H_{01} holds but H_{02} does not, whereas the points falling in the lower-right or lower-left quadrants lead to the incorrect rejection of H_{02}. Table 2.A.1 provides the incorrect rejection rates for varying α. It is notable that, even allowing the probability of making a type I error up to 0.4, the chance of declaring wrong common clusters is only about 5% given that one rejects the overall null hypothesis.

Table 2.A.1: Rejection rates evaluation under different significance levels: P_1 =P(reject at least one H_{0i}), P_2 =P(reject at least one $H_{0i}, i \notin T$|reject at least one H_{0i}). 200 datasets are simulated for each parameter configuration. True parameters used to generate the samples are set as: $f(x_i) = 0.4\phi(x_i; 1, 1) + 0.6\phi(x_i; \mu_2, 1), i = 1, \cdots, 500$ with $\mu_2 = 1 - 2\Delta$ and $f(y_j) = 0.3\phi(y_j; 1 + \delta, 1) + 0.7\phi(y_j; \mu_2 + \delta, 1), j = 1, \cdots, 500$ with $\delta = \frac{3}{4}\Delta$ and $\Delta = 1.2$.

α	P_1	P_2
0.05	0.20	0
0.10	0.28	0.018
0.15	0.36	0.042
0.20	0.40	0.038
0.25	0.45	0.056
0.30	0.47	0.054
0.35	0.54	0.047
0.40	0.58	0.052

Figure 2.A.1: Scatter plots of 200 replicates of $|\widehat{\mu}_1 - \widehat{\mu}_4| + z_{\alpha/4}\widehat{\sigma}_{\widehat{\mu}_1-\widehat{\mu}_4}$ vs $|\widehat{\mu}_1 - \widehat{\mu}_3| + z_{\alpha/4}\widehat{\sigma}_{\widehat{\mu}_1-\widehat{\mu}_3}$ for various α levels. The diagonal line is $y = x$. The vertical and horizontal dotted lines represent $x = \Delta$ and $y = \Delta$ respectively.



3.0 COMMUNITY DETECTION IN NETWORKS WITH NODE FEATURES

3.1 INTRODUCTION

In recent years, there has been increasing interest in statistical methodologies designed for network data. Network data takes the form of observed edges between nodes. Examples include brain networks (in which the nodes are segregated brain regions and edges are characterizations of white matter structural connectivity or brain's functional interactions) and social networks (in which the nodes are people and the edges may represent social interaction such as friendship or collaboration). The nodes and edges together define a network, often represented by an adjacency matrix, indicating the pairwise connection between nodes.

Community detection is a popular problem in network analysis. It has been a useful tool in identifying the important structures of many complex systems. Loosely speaking, network community refers to a subset of nodes that have similar profiles of connection to other nodes. Two classes of methods are commonly used for community detection.The first class of methods seeks community structure by optimizing a criterion that represents the quality of the partition of the network. These criteria come from a sense of what network communities should look like, lacking the interpretation of the data process that gives rise to the network.

The second class of methods involves fitting a probabilistic model that has well-defined communities, where community detection is achieved by optimizing some statistical criterion linked to the assumed model, for example, using the likelihood. One of the most popular models is the stochastic block model (SBM) (Holland et al., 1983; Snijders & Nowicki, 1997; Nowicki & Snijders, 2001). The important assumptions of the SBM model are that each

node belongs to one of the multiple blocks and the probability that an edge appears between any two nodes only depends on the memberships of the two nodes. Karrer & Newman (2011) proposed the degree corrected stochastic block model (DCBM) that allows degree inhomogeneity within blocks. Another popular model that shares the same goal of inferring node cluster labels is proposed in Handcock et al. (2007), where they extend the original latent space model proposed in Hoff et al. (2002) by combining a clustering model in the form of a mixture of Gaussians in the latent space so that inference on cluster labels is attainable along with the latent positions. For a survey of statistical models used in analysing network data, see Goldenberg et al. (2010) and Kolaczyk (2009).

Despite the extensive literature on community detection, most of the proposed methods only focus on the observed edges of the network without taking into account the additional information of node features (or node attributes). In many networks, the similarities and distinctions in the node features have considerable impact on the pattern of linking. The nodes in different communities are commonly assumed to have distinct connectivity patterns while the impact of node features is usually in a more continuous fashion. For example, in the global airline network, there are more connections between large airports, and, in the social network, individuals who are more similar to one another in age and education are more likely to have interconnections (McPherson et al., 2001). It is generally expected that integrating node features and network topology can help us understand the network structure better than using the adjacency matrix alone or node features information alone.

The primary focus is to take node features into account in network analysis in order to capture the impact of node features on the network links, as well as to detect the residual community structure beyond that explained by the node features. For instance, in the brain connectivity study, all the nodes are naturally embedded in a three-dimensional brain space. Connectivity between adjacent nodes is sometimes over-represented due to technical reasons (Stanley et al., 2013). One needs to account for the spurious connectivity in adjacent nodes by removing the effect of spatial location so as to recover functionally distinct brain regions ("communitie").

There are some recent attempts in integrating node features and network topology. Viennet et al. (2012) proposed to couple attribute information through inclusion of a pre-

determined similarity measure in the phase of constructing the k-nearest neighbour graph. Yang et al. (2013) developed an overlapping community detection method in networks with node features. Binkiewicz et al. (2014) introduced a covariate assisted spectral clustering that leverages both node covariates and the graph in spectral techniques. Zhang et al. (2015) proposed to include edge weights as a function of node features to an analogue of modularity so that nodes having more similarity are more likely to be grouped into the same community. Liu et al. (2014) extended Newman-Girvan modularity by defining a general null model. This model specifies some function to represent the effect of the node features on the network topology and, subsequently, take out the effect so as to reveal the hidden community structure. These efforts have provided great motivation for combining node features with community detection. However, the methods developed in these manuscripts or papers are mainly algorithmic approaches aiming at improving community detection, while we are trying to build a generative stochastic model that best captures the network structure. The non-generative models do not have a definite way to evaluate the performance of community detection since there is no unique "true model" even in the simulations. Moreover, these approaches are limited by their requirement for a pre-specified function form to describe the effects of node features in encoding network information.

In this chapter, we propose a feature adjusted stochastic block model (FASBM), which combines a block model component with community structures and a single-index function to incorporate node features. As a generative model, the FASBM model assumes that the connectivity probability between two nodes i and j is determined by their communities, and also a smooth function of the node features. The heterogeneity within a block is explained by the continuous effect of specific node features. The estimation of the FASBM model involves discovering the optimal block partition as part of the model estimation while capturing the impact of node features on the network links. The proposed model builds upon the stochastic block model (SBM) and, thus, inherits the merits of block models. With a semi-parametric single-index component, it is also adequately flexible to accommodate multiple node features with no prior information about the contribution of the features. Moreover, unlike many existing algorithms that are limited to binary-valued interactions, the proposed FASBM model and estimation approaches are applicable to relational data that are generated from

any exponential family distribution and are not restricted to being only Bernoulli.

3.2 BACKGROUND

3.2.1 Single-index model

Generalized linear model(GLM) is commonly used to explore the relationship between a response variable Y that follows from the exponential family distribution and covariates $\boldsymbol{z}$. A parametric GLM models a transformation $g(E(Y))$ as linear where g is known link function, i.e., $g(E(Y|\boldsymbol{z})) = \boldsymbol{\beta}^T\boldsymbol{z}$. In practice, however, the linear assumption may not hold. Hence, it's natural to consider the single index model: $g(E(Y|\boldsymbol{z})) = f(\boldsymbol{\beta}^T\boldsymbol{z})$, where f is unknown and $\boldsymbol{\beta}$ is the single-index coefficient. Single-index models have been proven to be an efficient way to avoid fitting multivariate nonparametric regression functions. Carroll et al. (1997) augmented the single-index model with additional covariates $\boldsymbol{x}$ taken into account, yielding a generalized partially linear single-index model: $g(E(Y|\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{\alpha}^T\boldsymbol{x} + f(\boldsymbol{\beta}^T\boldsymbol{z})$.

3.2.2 Stochastic Block Model (SBM)

Notation: A network G is defined in terms of nodes and edges $G = (V, E)$, where $V(G) = 1, \cdots, m$ is m number of nodes and $E(G)$ is set of edges. We consider undirected networks in our thesis. In an undirected network, all the edges are bi-directional. Most of the networks that have been studied have been binary in nature, that is, the edges between nodes indicate the presence or absence of an interaction. A network G can be represented by its random $m \times m$ adjacency matrix $Y = (Y_{ij})_{1\leq i,j\leq m}$. We assume self-loops are not allowed unless otherwise specified. Therefore, binary networks can be represented by a binary adjacency where $Y_{ij} = 1$ if there is an edge between node i and a different node j, and $Y_{ij} = 0$ otherwise.

The SBM has been developed in concordance with the notion of structural equivalence in a graph. A stochastic block model is a generative model for networks. Let K be the number of non-overlapping communities, m be the number of nodes and $\boldsymbol{r}$ be a vector of community

labels with $r_i = k$ if node $i, i = 1, \cdots, m$, belongs to the community k , $k = 1, \cdots, K$. Throughout Chapter 3, we assume that the number of communities K is pre-fixed. For the SBM, the adjacency matrix Y is generated by

$$Y_{ij} = \begin{cases} \text{independent Bernoulli with probability } \mu_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ Y_{ji} & \text{if } i > j \end{cases} \tag{3.1}$$

A stochastic block model is parameterized by a pair of $(\boldsymbol{r}, B)$, where B is a $K \times K$ symmetric matrix,

$$E(Y_{ij}) = \mu_{ij} = B_{r_i r_j}. \tag{3.2}$$

Under SBM, each node belongs to one of the multiple blocks, and the probablity that an edge appears between any two nodes only depends on the block memberships of the two nodes. Fitting a block model to any binary network involves partitioning nodes into blocks with each block representing a "community". Popular community detection algorithms for estimating the blocks in the SBM include likelihood-based algorithm, spectral clustering and algorithms based on the modularity score.

3.2.3 Commonly used algorithms

In this section, we provide a brief review of commonly used algorithms for inference of block models, including likelihood-based algorithms, spectral clustering and algorithms based on the modularity score. Spectral clustering and algorithms based on the modularity do not involve generative models. However, from a theoretical standpoint, it has been proved that simple spectral clustering (Von Luxburg, 2007) is consistent under SBM (Rohe et al., 2011; Lei et al., 2014). Bickel & Chen (2009) also proved that under some conditions, partitions obtained from the Newman-Girvan Modularity (Newman & Girvan, 2004) are consistent estimators of block partitions under the SBM.

3.2.3.1 Likelihood inference

The primary interest of community detection is concerned with estimating $\boldsymbol{r}$. It has been proved in Bickel & Chen (2009) that blockmodels and the corresponding likelihood-based algorithms are (asymptotically) unbiased and lead to the detection of the correct community structure. Let $L(Y; B, \boldsymbol{r})$ denote the log-likelihood function

$$L(Y; B, \boldsymbol{r}) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \{Y_{ij} \log(B_{r_i r_j}) + (1 - Y_{ij}) \log(1 - B_{r_i r_j})\}.$$

Finding the global maximum involves maximizing the likelihood function over all possible label assignments, which is computationally infeasible. Some types of greedy label-switching algorithms for maximizing the likelihood function have been proposed and work well in practice.

3.2.3.2 Spectral clustering

Spectral clustering includes all the techniques that partition the nodes into clusters by using the eigenvectors of adjacency matrix. There are different variants of spectral clustering in many applications (Jordan & Bach, 2004). Spectral clustering in our proposal refers to the basic spectral clustering (Von Luxburg, 2007) used for community recovery as described in the following. Define diagonal matrix D with $D_{ii} = \sum_{j=1}^{m} Y_{ij}, i = 1, \cdots, m$ and matrix $L = D^{-1/2} Y D^{-1/2}$. With predetermined K, the first step is to find the eigenvectors corresponding to the K largest absolute value of the eigenvalues of L, then choose the orthogonal eigenvectors and stack them in columns to form a matrix A, finally, treat each row of A as a point in $\mathbb{R}^k$ and partition them into K communities by k-means.

3.2.3.3 Newman-Girvan modularity

Modularity is a criterion for evaluating the quality of a partition of a network into communities, see Newman & Girvan (2004); Newman (2006). The basic idea of modularity is to compare the number of within-community edges in an observed network to the number of expected edges under some equivalent randomized network called null model and maximize

this difference over all possible community partitions. The general mathematical expression of a modularity criterion is

$$Q(\boldsymbol{r}) = \sum_{ij}[Y_{ij} - P_{ij}]I(r_i = r_j)$$

where P_{ij} is the probability of an edge between node i and j under the null model. The choice of the null model determines the exact form of modularity. A popular choice of the null model proposed by Newman & Girvan (2004) is $P_{ij} = D_{ii}D_{jj}/2m$. Searching over all possible partitions for modularity optimization is usually intractable, hence, practical algorithms are based on approximate optimization methods such as fast modularity optimization algorithm (Newman, 2004).

3.3 FEATURE ADJUSTED STOCHASTIC BLOCKMODEL (FASBM)

As introduced in the proceding section, the stochastic blockmodel is one of the most widely used models for community detection. In order to capture the impact of node features on the network structure as well as to detect the residual community structure beyond that explained by the node features, we aim to find a way to take feature information into account in the stochastic block model. Additionally, we aim to account for following three practical considerations:

1) There may be multiple node features influencing the connection probability.

2) In the general case, we may not have good knowledge of how node features impact connections.

3) Many networks have relational data indicating differing strengths of interactions. For example, in a brain network there may be stronger or weaker reactions between two regions of interest, or in a collaborative research network there may be more or fewer co-authored papers between two researchers. Dichotomizing the strength of interaction would clearly destroy potentially valuable information.

By embedding these ideas within the framework of the Stochastic BlockModel, we propose the Feature Adjusted Stochastic BlockModel (FASBM) that takes feature information explicitly into account on the basis of SBM. The adjacency matrix Y is generated by

$$Y_{ij} = \begin{cases} \text{independent exponential family with mean } \mu_{ij} & \text{if} \quad i < j \\ 0 & \text{if} \quad i = j \\ Y_{ji} & \text{if} \quad i > j \end{cases} \tag{3.3}$$

The distributions we consider here are mainly in one-parameter exponential family (uniquely determined by μ_{ij}). We allow for an unknown scaling parameter such as the variance in normal distribution. Our algorithm does not estimate the nuisance scaling parameter. Further specification of the mean μ_{ij} is as follows

$$E(Y_{ij}) = \mu_{ij} = g^{-1}(\boldsymbol{\theta}_{r_i r_j} + f(\boldsymbol{\beta}^T \boldsymbol{z_{ij}})), \quad \text{with } \|\boldsymbol{\beta}\| = 1. \tag{3.4}$$

where g is a known link function, $\boldsymbol{\theta}$ is a $K \times K$ symmetric matrix that captures the block-wise effect, f is an unknown smooth function that will be estimated nonparametrically, $\boldsymbol{z}_{ij}$ is a p-dimensional vector of covariates and $\boldsymbol{\beta}$ is the p-dimensional linear coefficient. Here $\boldsymbol{z_{ij}}$ is selected in a manner depending on the node features $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ and we assume $\boldsymbol{z_{ij}} = \boldsymbol{z_{ji}}$. Suppose that in a brain network, we are interested in assessing the impact of space on brain connectivity as well as recovering hidden communities, the physical distance between two brain regions represented by node i and node j may be a sensible choice. In such case, $z_{ij} = d(\boldsymbol{f}_i, \boldsymbol{f}_j)$ where d is a distance measure and the feature is node position. Our model encompasses many types of relational data generated from an exponential family distribution. If Y_{ij} is binary, common link functions g include logit $g(\mu) = \log(\frac{\mu}{1-\mu})$, probit and $g(\mu) = \phi^{-1}(\mu)$ where ϕ is the standard normal distribution function; for count Y_{ij} that follows a Poisson distribution, the common link function is $g(\mu) = log(\mu)$. For Gaussian data Y_{ij}, g is simply the identity link. Families that generate the well known class of generalized linear models are all extendable in the same way to the FASBM. The component $f(\boldsymbol{\beta}^T \boldsymbol{z_{ij}})$ can be refered to as a single-index component (Carroll et al., 1997). The restriction $\|\boldsymbol{\beta}\| = 1$ is required for identifiability and for easier interpretation, we set the first component of $\boldsymbol{\beta}$

to be positive. Single-index models have been proven to be an efficient way to avoid fitting multivariate nonparametric regression functions.

The proposed FASBM can be viewed as a generalized semi-parametric single index model (3.4), which consists of two parts: i) block model parameter $\boldsymbol{\theta}$ that enters the model as a parametric component, retaining the generality and tractability of the block model and ii) a single-index component $f(\boldsymbol{\beta}^T \boldsymbol{z_{ij}})$. The non-parametric function f is flexible to characterize nonlinear covariate effects, while $\boldsymbol{\beta}^T \boldsymbol{z}_{ij}$ reduces the dimension of the covariates. When no feature is concerned or covariates have no effect on node connections, FASBM becomes a generalization of the stochastic block model to accommodate relational data drawn from exponential families other than Bernoulli distribution. The classic SBM is obviously a special case of FASBM.

3.4 LIKELIHOOD INFERENCE FOR FASBM

In this section, we introduce the fitting algorithms for our proposed model. Consider $m(m-1)$ independent random variables Y_{ij} from exponential family distribution. The log-likelihood function in the canonical form with a canonical link is given as

$$\begin{aligned} L(Y; \boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{\beta}) &= \sum_{i=1}^{m} \sum_{j=i+1}^{m} \{(Y_{ij}\gamma_{ij} - b(\gamma_{ij}))/\phi + a(y_{ij}, \phi)\} \text{ and} \\ \gamma_{ij} &= \boldsymbol{\theta}_{r_i r_j} + f(\boldsymbol{\beta}^T \boldsymbol{z}_{ij}) \end{aligned} \tag{3.5}$$

for some known functions $b(\cdot), a(\cdot,\cdot)$, and a nuisance parameter ϕ. Our goal is to maximize the logarithm of the likelihood function with respect to the unknown model parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, f$, along with the node label assignment vector $\boldsymbol{r}$. Because an exact maximization of the (3.5) is computationally intractable, we propose an approach that alternates between two stages of maximization: first with respect to the parameters in the block model component, $\boldsymbol{r}$ and $\boldsymbol{\theta}$, and then with respect to the parameters in the single-index model component, f and $\boldsymbol{\beta}$. We adapt the likelihood-based algorithms for the SBM to stage 1 and the estimation procedures for fitting single-index models (Carroll et al., 1997) to stage 2. Note that we

used the canonical link function to explicitly write equation (3.5). In fact, the algorithm works for general link functions. Detailed descriptions of the algorithms are provided in the Subsection 3.4.2, and the code is publicly available on authors webpage.

In light of the fact that it has been proved in Bickel & Chen (2009) that partitions with likelihood-based algorithms for the SBM are consistent, we would expect a good chance of recovering membership consistently, as long as we can consistently estimate the single-index part f and $\boldsymbol{\beta}$. On the other hand, given $\boldsymbol{r}$, our model can be viewed as a generalized semi-parametric single-index model and consistency of the estimates $f, \boldsymbol{\beta}$ and $\boldsymbol{\theta}$ follows from Carroll et al. (1997). Empirically we show satisfactory performance of the algorithm as detailed in the Section 3.5.

3.4.1 Preliminaries

Local linear maximum likelihood estimation: We estimate f using local linear maximum likelihood estimation. Imagine for a moment that node membership $\boldsymbol{r}$, and $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ are fixed. We estimate the function f for each point x_0 by maximizing the local kernel-weighted log-likelihood

$$\begin{aligned} L(Y;\boldsymbol{\theta},\boldsymbol{r},\boldsymbol{\beta}) &= \sum_{i=1}^{m}\sum_{j=i+1}^{m}\{(Y_{ij}\gamma_{ij}-b(\gamma_{ij}))/\phi+a(y_{ij},\phi)\}K_h(\boldsymbol{\beta}^T\boldsymbol{z}_{ij}-x_0) \text{ and} \\ \gamma_{ij} &= \boldsymbol{\theta}_{r_i r_j}+B_0+B_1(\boldsymbol{\beta}^T\boldsymbol{z}_{ij}-x_0)). \end{aligned} \tag{3.6}$$

with respect to $\boldsymbol{B}=(B_0,B_1)^T$ and then $\hat{f}(x_0)=\hat{B}$ and $\hat{f}'(x_0)=\hat{B}_1$. Here $f(x)$ is locally approximated by a linear function near x_0:

$$f(x)\approx f(x_0)+f'(x_0)(x-x_0)\equiv B_0+B_1(x-x_0),$$

and $K_h(\cdot)=K(\cdot/h)/h$ is a rescaled kernel function $K(\cdot)$ with bandwidth h, which places more weight on those observations closer to x_0.

Fisher Scoring algorithm: Our estimation of $\boldsymbol{\theta}$ and $f(\cdot)$, $\boldsymbol{\beta}$ all use the Fisher Scoring algorithm for maximum likelihood estimation. Consider a random variable y with a distribution in the exponential family. The log-likelihood for one observation can be expressed as $l(y,\gamma,\phi)=[(y\gamma-b(\gamma))/\phi+a(y,\phi)]$ for known functions $b(\cdot),a(\cdot,\cdot)$, and it is easy to

show that $E(y) = \mu = b'(\gamma)$ and $\text{Var}(y) = b''(\gamma)\phi = V(\mu)$. When alternating between the estimation of $\boldsymbol{\theta}$, $f(\cdot)$ and $\boldsymbol{\beta}$, the proposed model $\mu = g^{-1}(\theta + f(\boldsymbol{\beta}^T\boldsymbol{z}))$ can be written as $g(\mu) = \eta(\boldsymbol{B})$, with its respective form of η and unknown parameters $\boldsymbol{B}$. By the chain rule and properties of exponential family, the score function $U(\boldsymbol{B})$ for N observations becomes

$$\begin{aligned} U(\boldsymbol{B}) = \sum_{s=1}^{N} \boldsymbol{u}_s = \sum_{s=1}^{N} \frac{\partial l_s(\boldsymbol{B})}{\partial \boldsymbol{B}} &= \sum_{s=1}^{N} \frac{\partial l_s}{\partial \gamma_s}\frac{\partial \gamma_s}{\partial \mu_s}\frac{\partial \mu_s}{\partial \eta_s}\frac{\partial \eta_s}{\partial \boldsymbol{B}} w_s \\ &= \sum_{s=1}^{N} \frac{y_s - \mu_s}{\phi}\frac{1}{V(\mu_s)} g^{-1\prime}(\eta_s)\frac{\partial \eta_s}{\partial \boldsymbol{B}} w_s \\ &= \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}} \boldsymbol{W_1} \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu}) \end{aligned}$$

with diagonal matrix $[\boldsymbol{W_1}]_{ss} = \dfrac{g^{-1\prime}(\eta_s)}{\phi V(\mu_s)}$ and diagonal weight matrix $[\boldsymbol{W}]_{ss} = w_s$. The weight matrix $\boldsymbol{W}$ is simply an identity matrix when maximizing the global log-likelihood for the estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. When considering the local kernel-weighted log-likelihood for estimating $f(x_0)$, the local kernel-weight w_s for each observation is specified in Section 3.4.1.

The Hessian matrix and Information become:

$$H(\boldsymbol{B}) = \frac{\partial U(\boldsymbol{B})}{\partial \boldsymbol{B}} = \sum_{s=1}^{N}(y_s - \mu_s)\frac{\partial\left([\boldsymbol{W_1}]_{ss} w_s \dfrac{\partial \eta_s}{\partial \boldsymbol{B}}\right)}{\partial \boldsymbol{B}} + [\boldsymbol{W_1}]_{ss} w_s \frac{\partial \eta_s}{\partial \boldsymbol{B}}\frac{\partial (y_s - \mu_s)}{\partial \boldsymbol{B}},$$

and

$$\begin{aligned} I(\boldsymbol{B}) = -E(H(\boldsymbol{B})) = \sum_{s=1}^{N}[\boldsymbol{W_1}]_{ss} w_s \frac{\partial \eta_s}{\partial \boldsymbol{B}}\frac{\partial \mu_s}{\partial \boldsymbol{B}} &= \sum_{s=1}^{N}[\boldsymbol{W_1}]_{ss} w_s \frac{\partial \eta_s}{\partial \boldsymbol{B}}\frac{\partial \mu_s}{\partial \eta_s}\frac{\partial \eta_s}{\partial \boldsymbol{B}} \\ &= \sum_{s=1}^{N}[\boldsymbol{W_1}]_{ss} w_s \frac{\partial \eta_s}{\partial \boldsymbol{B}} g^{-1\prime}(\eta_s)\frac{\partial \eta_s}{\partial \boldsymbol{B}} \\ &= \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}} \boldsymbol{W_2} \boldsymbol{W} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}}\right)^T \end{aligned}$$

with diagonal matrix $[\boldsymbol{W_2}]_{ss} = \dfrac{(g^{-1\prime}(\eta_s))^2}{\phi V(\mu_s)}$. Given $\boldsymbol{B}^{(l)}$ at the previous step, by the Fisher Scoring algorithm, the updated $\hat{\boldsymbol{B}}^{(l+1)} = \hat{\boldsymbol{B}}^{(l)} + I^{-1}(\hat{\boldsymbol{B}}^{(l)})U(\hat{\boldsymbol{B}}^{(l)})$,

$$\hat{\boldsymbol{B}}^{(l+1)} = \hat{\boldsymbol{B}}^{(l)} + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}} \boldsymbol{W_2} \boldsymbol{W} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}}\right)^T\right)^{-1} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}} \boldsymbol{W_1} \boldsymbol{W}(\boldsymbol{y} - \boldsymbol{\mu})\Big|_{(l)} \tag{3.7}$$

The approach for updating f, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ all fall into the above framework with its respective η and unknown parameters $\boldsymbol{B}$, which will be specified in Section 3.4.2. The weight matrices $\boldsymbol{W}$ are the kernel weights for local likelihood estimation, and only used in updating f. Given $\boldsymbol{\eta}$, the link function g, and the distribution of Y, matrices $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ can be computed according to the forumla given above.

3.4.2 The Algorithm

Before we demonstrate the detailed algorithms, we convert the upper triangle (excluding the diagonal) of Y into a vector $Y^*_{N\times 1} = \left(Y_{12}, \cdots, Y_{(m-1)m}\right)^T$ where $N = m(m-1)/2$, and accordingly let $\boldsymbol{Z}^*_{N\times p} = \left(\boldsymbol{z}_{12}, \cdots, \boldsymbol{z}_{(m-1)m}\right)^T$. We use $Y^*_{s(ij)}$ and $\boldsymbol{z}^*_{s(ij)}$ for the correspondance between s and the pair (i,j) when necessary.

(a) Initialization: Let $\hat{f}(\cdot) = 0$, each entry of $\hat{\boldsymbol{\beta}} = \sqrt{1/p}$, assign initial labels $\boldsymbol{r}$ by k-means on the rows of Y matrix.

(b) Updating $\boldsymbol{\theta}$ and $\boldsymbol{r}$: Given $\hat{f}^{(o)}$ and $\hat{\boldsymbol{\beta}}^{(o)}$, obtain $\hat{\boldsymbol{\theta}}^{(o+1)}$ and $\hat{\boldsymbol{r}}^{(o+1)}$ by repeating steps of updating $\boldsymbol{\theta}$ and $\boldsymbol{r}$ iteratively until $\boldsymbol{r}$ is unchanged.

Suppressing the superscript (o), given the current $\hat{f}$ and $\hat{\boldsymbol{\beta}}$, each iteration of updating $\boldsymbol{\theta}$ and $\boldsymbol{r}$ involves two steps:

(i) Given $\hat{\boldsymbol{r}}^{(q-1)}$, update $\hat{\boldsymbol{\theta}}^{(q)}$ through (3.7) by reparameterizing the upper triangle of $\boldsymbol{\theta}_{K\times K}$ into $\boldsymbol{B}_{P\times 1} = (\boldsymbol{\theta}_{11}, \cdots, \boldsymbol{\theta}_{1K}, \boldsymbol{\theta}_{22}, \cdots, \boldsymbol{\theta}_{KK})^T$ with $P = K(K+1)/2$. Here $\eta_{s(ij)} = \boldsymbol{x}^T_{s(ij)}\boldsymbol{B} + f(\boldsymbol{\beta}^T\boldsymbol{z}^*_{s(ij)})$ for $s = 1, \ldots, N$, where $\boldsymbol{x}_{s(ij)}$ has only one 1 indicating the memberships (r_i, r_j), otherwise zero.

(ii) Given $\hat{\boldsymbol{\theta}}^{(q)}$, the community label for ith node $r_i^{(q)}$ is updated by minimizing the negative log-likelihood through the greedy label-switching algorithm (Stephens, 2000) as follows:

$$\hat{r_i}^{(q)} = \arg\min_{k\in\{1,\cdots,K\}} \sum_{j=1}^{m} \{-Y_{ij}\log[g^{-1}(\hat{\boldsymbol{\theta}}^{(q)}_{k,r_j^{(q-1)}} + \hat{f}(\hat{\boldsymbol{\beta}}^T\boldsymbol{z_{ij}}))] - (1-Y_{ij})\log[1-g^{-1}(\hat{\boldsymbol{\theta}}^{(q)}_{k,r_j^{(q-1)}} + \hat{f}(\hat{\boldsymbol{\beta}}^T\boldsymbol{z_{ij}}))]\}.$$

(c) Updating $\boldsymbol{\beta}$ and f: Given $\hat{\boldsymbol{\theta}}^{(o+1)}$ and $\hat{\boldsymbol{r}}^{(o+1)}$, obtain $\hat{f}^{(o+1)}$ and $\hat{\boldsymbol{\beta}}^{(o+1)}$ by iterating between updating $\boldsymbol{\beta}$ and f until $\frac{\|\hat{f}^{(q)}(\boldsymbol{t}) - \hat{f}^{(q-1)}(\boldsymbol{t})\|}{\|\hat{f}^{(q-1)}(\boldsymbol{t})\|} \leq \epsilon$ for a suitably chosen small constant ϵ, where $\|\cdot\|$ denotes L_2 norm, $\boldsymbol{t}$ denotes a grid of points and q denotes the index of iteration consisting of updating $\boldsymbol{\beta}$ and $\boldsymbol{f}$.

Omitting the superscript (o), given the current $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{r}}$, each iteration of updating $\boldsymbol{f}$ and $\boldsymbol{\beta}$ involves two steps:

(i) Given $\hat{f}^{(q-1)}$, $\hat{\boldsymbol{\beta}}^{(q)}$ is obtained through (3.7) by viewing $\boldsymbol{B} = \boldsymbol{\beta}$. Here $\eta_{s(ij)} = \boldsymbol{\theta}_{r_i r_j} + f(\boldsymbol{B}^T \boldsymbol{z}^*_{s(ij)})$. Note that $\hat{\boldsymbol{\beta}}$ need to be normalized to meet $\|\boldsymbol{\beta}\| = 1$.

(ii) Given $\hat{\boldsymbol{\beta}}^{(q)}$, we fit $\hat{f}(\cdot)$ at a fixed but fine grid of points and subsequently using interpolation to get the other values. Take one of the grid points x_0 for example, $\hat{f}(x_0)$ is updated through (3.7) using the local likelihood approach. Here, $\boldsymbol{B} = (B_0, B_1)$, $\eta_{s(ij)} = \boldsymbol{\theta}_{r_i r_j} + B_0 + B_1(\boldsymbol{\beta}^T \boldsymbol{z}^*_{\boldsymbol{s(ij)}} - x_0))$, $[\boldsymbol{W}]_{ss} = K_h(\hat{\boldsymbol{\beta}} \boldsymbol{z}^*_{\boldsymbol{s}} - x_0)$ and $\hat{f}(x_0) = \hat{B}_0$.

(d) Iterate between steps (b) and (c) until $\frac{\|\hat{f}^{(o+1)}(\boldsymbol{t}) - \hat{f}^{(o)}(\boldsymbol{t})\|}{\|\hat{f}^{(o)}(\boldsymbol{t})\|} \leq \epsilon$ for a suitably chosen small constant ϵ.

3.5 SIMULATION STUDIES

In this section, we evaluate performance of our algorithm (FASBML) for fitting FASBM under different types and levels of node influence. We compared the community detection results with the likelihood-based inference of SBM (SBML) and the simple spectral clustering (SPEC). Although there are some papers that consider the incorporation of node features in community detection, as we discussed in the introduction, these methods are not model based and the influence of the node feature has to be in a known format. Therefore, we can not directly compare with these methods.

We consider two measures to quantify the performance in terms of the agreement between the true $\boldsymbol{r}$ and $\hat{\boldsymbol{r}}$. The first measure is the average misclassification rates, quantifying the overall proportion of mis-clustered nodes (Girvan & Newman, 2002). We also adopt

the normalized mutual information criterion (NMI) (Kvalseth, 1987) to measure clustering quality, where higher values indicate better matching.

In all the cases below, the network generation procedure takes the following steps: first, generate labels for m nodes independently with $P(r_i = 1) = \cdots = P(r_i = K) = 1/K$; second, randomly position nodes within the interval $(0, 1)$ and compute the distance d_{ij} between node i and node j; finally, the edges between node i and node j are generated independently as $Y_{ij} \sim \text{Bernoulli}\left(g^{-1}(\boldsymbol{\theta}_{r_i r_j} + f(d_{ij}))\right)$, where g is the logit function. The values of $\boldsymbol{\theta}$ for $K = 2$ and $K = 3$ are as follows,

$$\boldsymbol{\theta} = \text{logit}\begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\theta} = \text{logit}\begin{pmatrix} 0.5 & 0.2 & 0.2 \\ 0.2 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.1 \end{pmatrix}.$$

In simulation I, we let $f = a\sin(-8d_{ij})$, with a taking different values, $0, 1.4$ or 1.8, so as to investigate the robustness of different methods to varying influences of node features. Table 3.1 and Table 3.2 show the results of 100 simulation runs, for $K = 2$ and $K = 3$, where networks have varying community size, $m = 100, 200, 400$. In the situations where features have no impact on the network topology, i.e., $a = 0$, FASBML and SBML perform equally well followed by spectral clustering. All the methods perform better as m increases as, with more links, there is effectively more data to use for fitting the model. The performance of SBM deteriorates rapidly as the amount of node influence increases. On the other hand, the partition found by FASBML still has very good agreement with the actual partition in the presence of large feature influence, and the performance improves as m increases. The inferiority of SBM relative to FASBM in these scenarios is understandable as FASBM always uses both the network topology and the features, whereas SBM completely ignores feature influence. In addition, the fact that FASBML and SBML have equally good performance when $a = 0$ confirms the robustness of FASBML.

It is worth mentioning that all the three algorithms require the number of communities to be known in advance, and we used the true K in the simulation. Determining the number of communities is gaining increasing interest recently (Chen & Lei, 2014; Bickel & Sarkar, 2016). In Section 3.6, we used the network cross-validation (NCV) method proposed by

(Chen & Lei, 2014) to determine the number of communities. Also for spectral clustering method, we need to choose the dimension d of spectral embedding. In the simulation, we tried different d values for spectral method and reported the one with the best performance.

Table 3.1: Results of simulation I, $K = 2$. The average misclassification rates $\overline{Mis_p}$ and normalized mutual information (NMI) are shown together with their standard deviations enclosed in parentheses for varying a in $f = a\sin(-8d_{ij})$, and varying number of nodes m. Numbers in bold indicate the best performance.

		$m = 100$			$m = 200$			$m = 400$		
		FASBML	SBML	SPEC	FASBML	SBML	SPEC	FASBML	SBML	SPEC
$a = 0$	$\overline{Mis_p}$	**0.012**	**0.012**	0.041	**0.0004**	**0.0004**	0.006	**0**	**0**	0.0003
		(0.010)	(0.010)	(0.024)	(0.0015)	(0.0015)	(0.006)	(0)	(0)	(0.0009)
	NMI	**0.924**	**0.924**	0.783	**0.997**	**0.997**	0.955	**1**	**1**	0.998
		(0.06)	(0.06)	(0.100)	(0.013)	(0.013)	(0.040)	(0)	(0)	(0.008)
$a = 1.4$	$\overline{Mis_p}$	0.157	0.443	**0.128**	**0.012**	0.469	0.079	**0.0001**	0.481	0.045
		(0.200)	(0.087)	(0.041)	(0.067)	(0.024)	(0.031)	(0.0004)	(0.014)	(0.020)
	NMI	**0.592**	0.038	0.470	**0.962**	0.005	0.625	**0.999**	0.002	0.75
		(0.404)	(0.147)	(0.116)	(0.141)	(0.007)	(0.104)	(0.003)	(0.002)	(0.08)
$a = 1.8$	$\overline{Mis_p}$	**0.174**	0.461	0.182	**0.036**	0.469	0.132	**0.005**	0.482	0.105
		(0.192)	(0.028)	(0.057)	(0.119)	(0.023)	(0.046)	(0.049)	(0.015)	(0.030)
	NMI	**0.524**	0.007	0.349	**0.908**	0.004	0.464	**0.989**	0.002	0.549
		(0.375)	(0.009)	(0.115)	(0.251)	(0.007)	(0.118)	(0.100)	(0.002)	(0.090)

For simulation II, we use two more examples to illustrate the empirical performance of the non-parametric estimation for the function f. We set $K = 2$ in both examples. In the first example, f is an exponential function, $f(d_{ij}) = 2\exp(-8d_{ij}) - 2$; in the second example, f is a polynomial function, $f(d_{ij}) = 10d_{ij}^4 - 42d_{ij}^3 + 50d_{ij}^2 - 20d_{ij}$. A fitted curve randomly selected from 100 simulations is depicted in Figure 3.1 for each scenario. It is shown that, when the network is of moderate size, the fitted curve is remarkably close to the true curve, except some boundary effect near the endpoints. The proposed algorithm can also provide satisfying partition results as presented in Table 3.3.

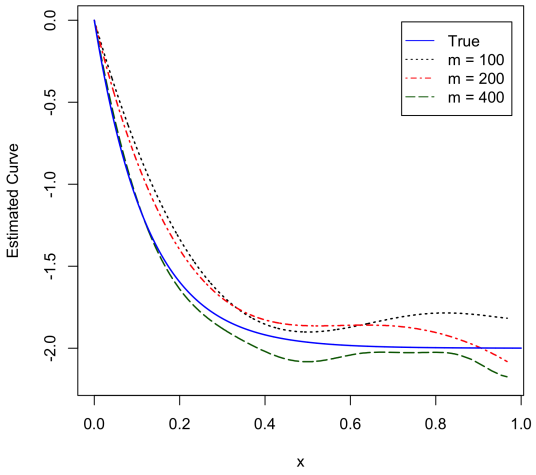
Table 3.2: Results of simulation I: $K = 3$. The average misclassification rates $\overline{Mis_p}$ and normalized mutual information (NMI) are shown together with their standard deviations enclosed in parentheses for varying a in $f = a\sin(-8d_{ij})$, and varying number of nodes m.

		$m = 100$			$m = 200$			$m = 400$		
		FASBML	SBML	SPEC	FASBML	SBML	SPEC	FASBML	SBML	SPEC
$a = 0$	$\overline{Mis_p}$	**0.262**	0.265	0.298	**0.073**	0.075	0.185	**0.011**	**0.011**	0.074
		(0.133)	(0.133)	(0.067)	(0.095)	(0.096)	(0.045)	(0.005)	(0.005)	(0.021)
	NMI	**0.546**	0.544	0.404	**0.825**	0.824	0.545	**0.954**	0.953	0.753
		(0.110)	(0.110)	(0.077)	(0.060)	(0.063)	(0.060)	(0.020)	(0.020)	(0.045)
$a = 1.4$	$\overline{Mis_p}$	**0.380**	0.535	0.407	**0.167**	0.524	0.352	**0.038**	0.534	0.335
		(0.098)	(0.052)	(0.065)	(0.149)	(0.041)	(0.057)	(0.089)	(0.037)	(0.055)
	NMI	**0.332**	0.099	0.272	**0.682**	0.117	0.331	**0.910**	0.117	0.351
		(0.142)	(0.071)	(0.072)	(0.176)	(0.057)	(0.056)	(0.076)	(0.056)	(0.050)
$a = 1.8$	$\overline{Mis_p}$	**0.421**	0.566	0.450	**0.197**	0.573	0.434	**0.020**	0.592	0.436
		(0.094)	(0.044)	(0.059)	(0.154)	(0.040)	(0.059)	(0.008)	(0.030)	(0.053)
	NMI	**0.260**	0.053	0.209	**0.625**	0.050	0.244	**0.919**	0.038	0.270
		(0.125)	(0.049)	(0.068)	(0.195)	(0.037	(0.056)	(0.025)	(0.033)	(0.038)

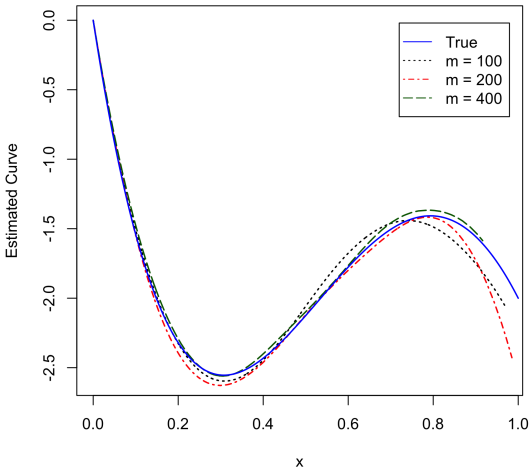
Table 3.3: Results of Simulation II. The average misclassification rates $\overline{Mis_p}$ and normalized mutual information (NMI) are shown for FASBML together with their standard deviations enclosed in parentheses for exponential f and polynomial f, with varying number of nodes m.

	$m = 100$		$m = 200$		$m = 400$	
	Exp f	Poly f	Exp f	Poly f	Exp f	Poly f
$\overline{Mis_p}$	0.098	0.172	0.021	0.046	0.002	0.006
	(0.056)	(0.090)	(0.012)	(0.040)	(0.003)	(0.004)
NMI	0.574	0.398	0.866	0.763	0.981	0.954
	(0.136)	(0.178)	(0.067)	(0.130)	(0.021)	(0.029)

Figure 3.1: Estimates of f for a randomly selected simulated network with varying f functions and varying number of nodes m. (a) $f(x) = 2\exp(-8x) - 2$. (b) $f(x) = 10x^4 - 42x^3 + 50x^2 - 20x$. (c) $f(x) = 1.4sin(-8x)$. (d) $f(x) = 1.8sin(-8x)$

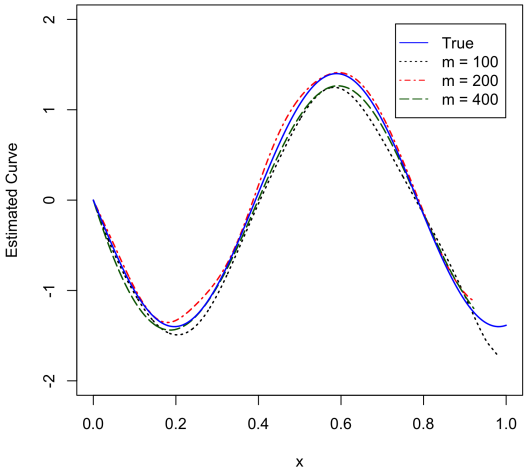


(a)

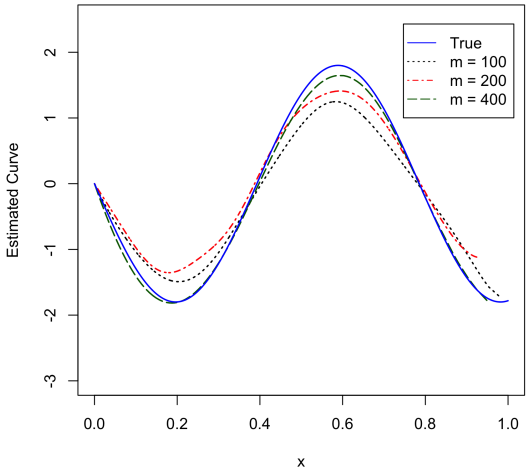


(b)

.



(c)



(d)

3.6 DATA APPLICATIONS

In this section, we show applications of our method to three actual world networks: a functional brain network, an US air-transportation network and Lazega lawyers friendship network, which are representative examples of biological, social and infrastructure systems. The proposed FASBM reveals interesting node feature effects, as well as interpretable communities.

3.6.1 Functional brain network

We first consider an application to brain functional connectivity study using resting-state functional magnetic resonance imaging (RS-fMRI) data. The data were collected by University of Pittsburgh Medical Center and detailed descriptions can be found in Hwang et al. (2012). Imaging data were preprocessed to reduce noise and artifacts using standard fMRI data processing methods.

RS-fMRI measures the intrinsic, high-amplitude, low-frequency blood-oxygen-level dependence signal (BOLD) fluctuations of the brain. The relationship between RS-fMRI signals from different regions is thought to reflect functional connectivity independent of any particular brain state (Van Dijk et al., 2010). Functional connectivity between a pair of voxels is usually estimated by calculating the Pearson correlation coefficient between their BOLD time series, treating the observations as coming from a single bivariate distribution.

The brain network in this analysis contains 448 nodes (voxels) in the basal ganglia mask. The data matrix Y_{ij} is the averaged Fisher's z-transformed correlation values based upon all subjects. The basal ganglia subserves a wide range of functions, including motor, cognitive, motivational, and emotional processes and has been implicated in numerous neurological and psychiatric disorders. There have been great interest in using RS-fMRI techniques to study the functional connectivity in basal ganglia (Di Martino et al., 2008; Robinson et al., 2009; Barnes et al., 2010).

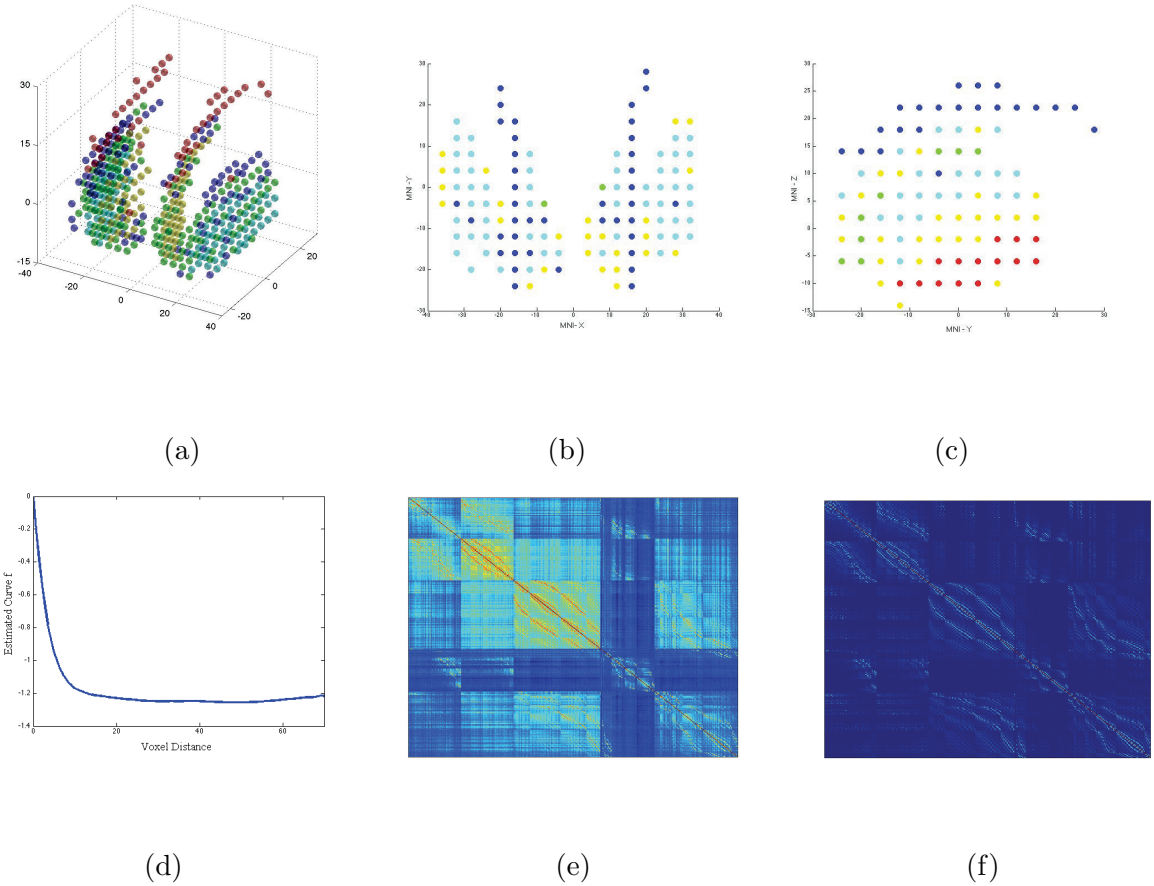
Given the fact that connectivity between adjacent nodes is sometimes over-represented due to inevitable technical reasons in fMRI data acquisition process and data processing

(Stanley et al., 2013), we consider the Euclidean distance between two voxels as the covariate z_{ij} in applying FASBM to discover the underlying block structure of the functional brain network. Here the spatial location of each node is defined as the coordinates of the center of the voxel in Montreal Neurological Institute (MNI) stereotactic space.

The estimated function f as shown in Figure 3.2d reflect the expected relationship between brain connectivity and spatial locations. Figure 3.2e reveals that the pairs of voxels within the same block are not exactly connected in the same way as evidenced by the noise patterns within blocks. Fitting of the simple stochastic block model to the brain network can not characterize the heterogeneity within blocks, whereas the proposed FASBM with spatial feature z_{ij} incorporated is a better approximation to the data by accounting for the spurious connection between adjacent nodes. As shown in Figure 3.2f: the nonparametric function f in our model captures the additive effect of the deviations from the block structure. It can be seen that the heterogeneity within the blocks are well explained by the effect of local correlations as modeled by the nonparametric function f.

As shown in the top panels of Figure 3.2, using FASBML yields functionally distinct but spatially coherent parcellations of the brain region. Previous studies have parcellated the basal ganglia based on its extrinsic functional connectivity with the cortex (Barnes et al., 2010; Choi et al., 2012). It is unknown that whether or not the basal ganglia can be successfully parcellated by only considering local, intrinsic functional information within the basal ganglia. Using the proposed method, we have successfully identified basal ganglia subdivisions by only considering functional connectivity pattern between basal ganglia voxels. This parcellation closely resembled those reported using structural anatomical information (Tziortzi et al., 2011). By visual examination: cluster 1(yellow) corresponds to the caudate body, cluster 3(green) corresponds closely to the putamen, cluster 5(cyan) closely to the pallidum , and cluster 2(red), 4(blue) partially correspond to the caudate head.

Figure 3.2: (a) The functional brain network: each voxel was represented by a single node at its spacial location with the color reflecting the inferred community membership by the proposed FASBML. (b) Projection of (a) in the x-y plane of the MNI stereotactic space. (c) Projection of (a) in the MNI y-z plane. (d) Estimated f function. (e) Connectivity matrix of the brain network data with voxels ordered by inferred community membership. (f) Fitted f evaluated on the distance matrix z_{ij} of the brain network data with voxels ordered by inferred community membership.



(a) (b) (c)

(d) (e) (f)

3.6.2 United States air-transportation network

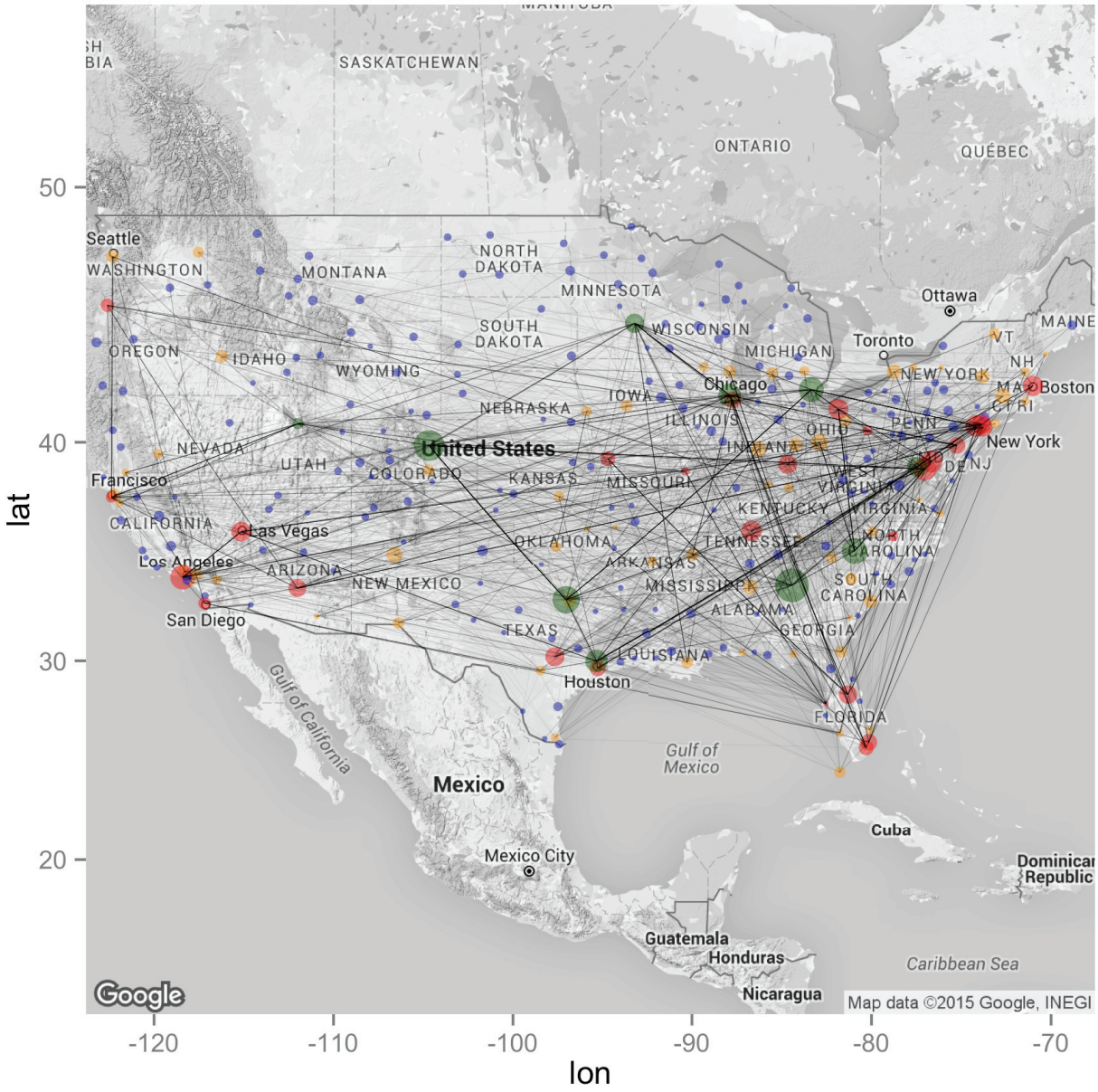
For the second example, we analyze a US airline network. We extracted information of the United States domestic airports and flights for the year 2012 from the OpenFlights/Airline Route Mapper Route Database. The resulting air-transportation network comprises 300 nodes denoting airports in the United States and about 6000 flight routes within the United States operated by the major airlines (United Airlines (UA), American Airlines(AA), Delta Air Lines and Southwest Airlines). The edges in the network indicate presence or absence of non-stop flights between two airports. The full data set can be downloaded from `http://openflights.org/data.html`.

The air-transportation network is a complex network with heterogeneous degrees: a handful of nodes in the air transportation network are busy airports having a significant number of connections to and from other airports. For example, *Chicago O'hare International (Int'l) airport, Hartsfield Jackson Atlanta Int'l, Charlotte Douglas Int'l*, and *Denver Int'l*, each has more than 70 connections. Therefore, it is expected that community-detection methods solely based on the adjacency matrix will tend to form communities characterized by different degrees. For instance, SBML split the network into four groups by degree: high, relatively high, medium and low, as shown in Figure 3.3.

In the following, we fit the proposed feature adjusted stochastic block model (FASBM) in the hope to discover community structures that are not merely due to the degree distribution. The node feature we consider is the number of airports each node has connections to, i.e., $f_i = \sum_{l=1}^{m} Y_{il}$, and let the covaraite $z_{ij} = f_i + f_j$. The use of FASBM requires a pre-specified number of communities as input, whereas it is unclear how many communities are in the airline network, 2-fold network cross-validation (NCV) was applied to determine the number of communities. NCV approach is recently proposed by Chen & Lei (2014) to select number of clusters through block-wise edge splitting. Using negative log-likelihood as loss functions, NCV method consistently selects $K = 4$ communities.

The resulting communities do not entirely correspond to groups of high and low degree, as shown in Figure 3.4. The community labeled in orange identifies almost all the "home base" airports of Southwest airline: *Las Vegas McCarran Int'l, Houston Hobby Int'l,*

Figure 3.3: The communities inferred by stochastic block model (SBM). Each vertex represents an airport, the size of which is proportional to the square root of its number of connections and the color of which reflects inferred community membership. SBML split the network into four groups by degree: high (green), relatively high (red), medium (orange) and low (blue).



Chicago Midway Int'l, Baltimore-Washington Int'l, Lambert-St. Louis Int'l, Nashville Int'l and Kansas City Int'l, Austin-Bergstrom Int'l and so on. The community labeled in red mainly consists of airports served as hubs for UA, AA or Delta airlines, including *Hartsfield Jackson Atlanta Int'l* and *Detroit Metropolitan Airport* as hubs for Delta, *Chicago O'hare Int'l*, *Newark Liberty Int'l* and *Washington Dulles Int'l* as hubs for UA, *Philadelphia Int'l*, *Charlotte Douglas Int'l* and *Ronald Reagan Washington National Airport* as hubs for AA.

The community labeled in green comprises airports characterized by varying node degrees, where the low degree airports have one of UA, AA or Delta airlines as the only carrier, and busy airports serve as hubs for one of the UA, AA and Delta airlines. The community labeled in blue corresponds to airports with low degree. Many members of this community are regional airports that serve air traffic within a relatively local or lightly populated regions. Additionally, we have shown in Figure 3.5 that, the shape of the estimated f function reflects the general relationship between connectivity probability and the sum of degrees for a pair of nodes - airports with high degrees tend to connect to other airports, and the opposite holds true for low degrees airports.

Our results are in agreement with the fact that Southwest, as the fourth largest airlines in the U.S., after the big three legacy carriers (UA, AA and Delta), was less assertive in big travel markets and chose to avoid competing with the "big three" in their hub airports, and instead focus on cities other than these big hubs. Southwest Airlines adopts a point-to-point (PP) configuration wherein airports are connected by direct routes. On the contrary, the "big thre" adopt the hub-and-spoke (HS) system (Aguirregabiria & Ho, 2010) , wherein most of the operations are concentrated in the hubs and all other cities in the network (i.e., the spokes) are connected to the hubs by non-stop flights. Although there is no "correct" way to partition air transportation network, compared to the partition by SBM, FASBM allows us to recover the hidden structural organization that is beyond the groups of degrees. The node feature information incorporated in the block model helps to provide more insights into the development and categorization of the air-transportation network.

Figure 3.4: The US Air-transportation Network. Airports were each replaced by a single vertex, the size of which is proportional to the square root of its number of connections and the color of which reflects inferred community membership by likelihood inference of Feature adjusted likelihood stochastic block model (FASBML): green labels the community comprising airports characterized by varying node degrees: low degree airports with one of UA, AA or Delta airlines as the only carrier or busy airports served as hubs for one of the UA, AA and Delta airlines; red labels the community of hubs for UA, AA or Delta; orange labels the community corresponding to almost all the home base airports of Southwest airline; blue labels the community of regional airports.

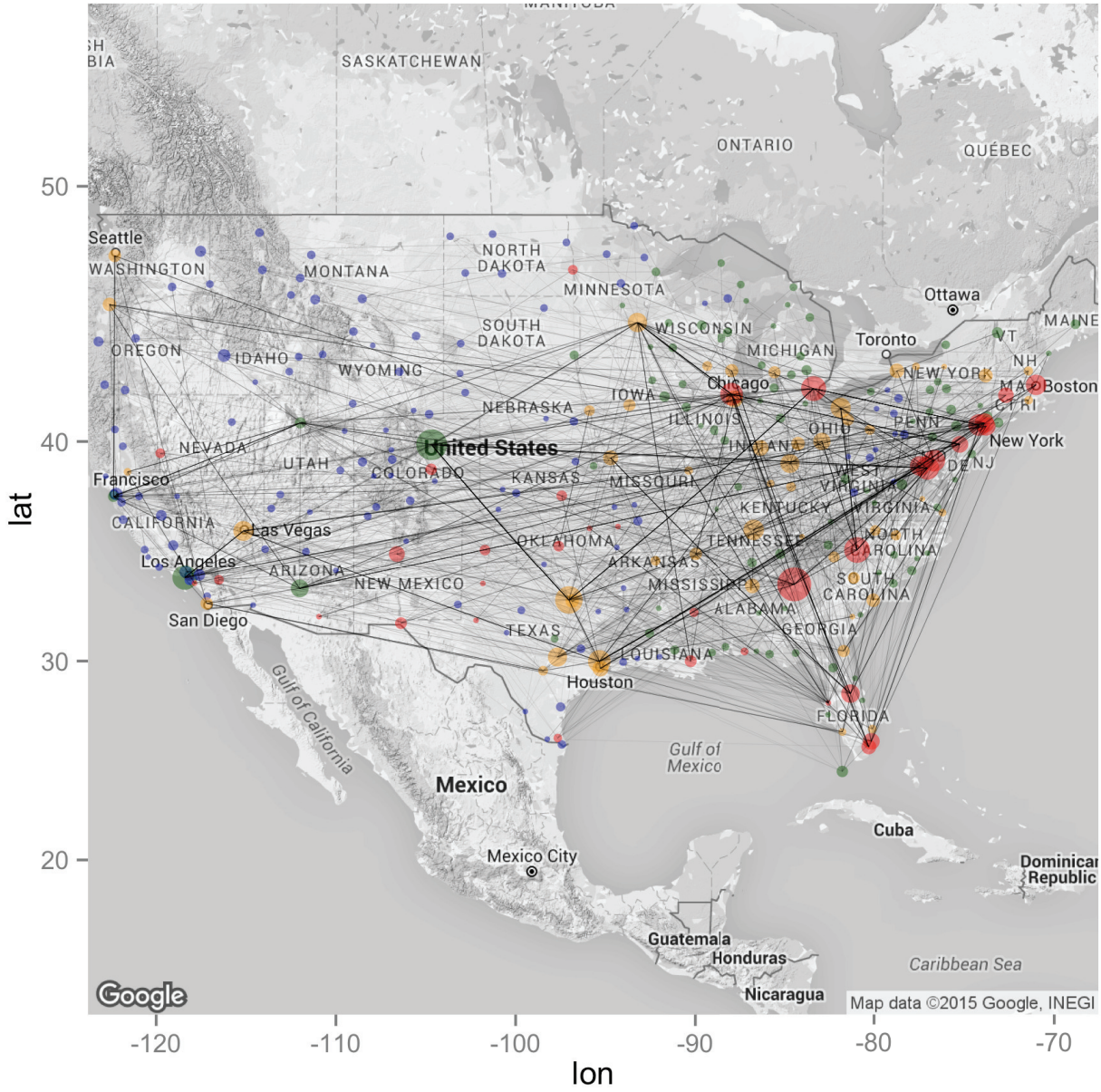
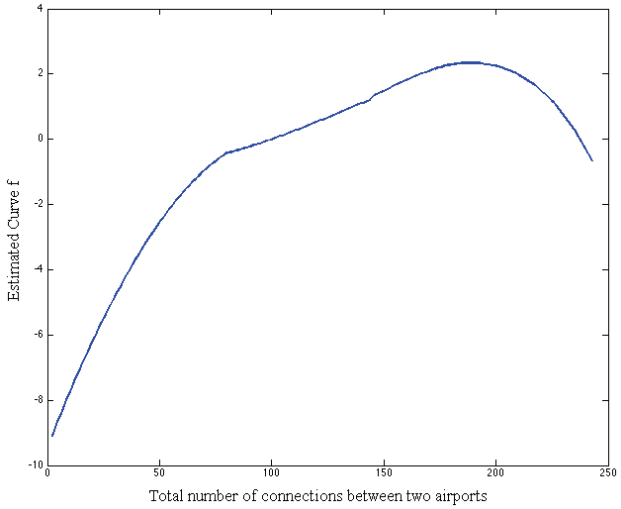


Figure 3.5: Estimates curve against the total number of the connections of the two airports for the US air-transportation network.



3.6.3 Lazega lawyers friendship network

The last example is the friendship network collected by Lazega (Lazega, 2001) among members of a New England law firm. The nodes of this network represent lawyers in the firm, and the edges indicate friendship ties between the lawyers. Additionally, we have information of age, gender, office location (Hartford, Providence, or Boston), the practice (litigation or corporate law), status (partner or associate) , law school attended (Harvard, Yale, University of Connecticut, or other), years with the firm, and seniority (i.e., the number of years the lawyers spent in the law firm) of each lawyer. The Providence office, which only includes two isolated nodes and two non-isolated nodes, is excluded in the analysis. The resulting friendship network contains 67 lawyers.

In this work, the attribute that is taken into account in the FASBM is the differences in seniority between two lawyers (nodes). This is in agreement with the findings in Snijders et al. (2006) that similarity effect of work locations and the effects associated with seniority are the two most important covariates on network topology. It is of interest in a study like this to assess the seniority similarity effect on friendship, and in the meanwhile, incorporate the information into community detection process to improve results of the selectivity as compared to the use of the network information alone.

Using the FASBML method, the lawyer are partitioned into two clusters: of the 31 lawyers partitioned into the first cluster, 30 of them work in the Boston office, whereas of the lawyers partitioned in the other cluster, half of them work in the Boston office and the other half are in the Hartford office as shown in Figure 3.6. The likelihood of friendship establishment between lawyers is affected by the differences between lawyers in how long they have served in the firm, as demonstrated in Figure 3.7, indicating that friendship is more likely to be established between members with similar length of service with the firm.

Figure 3.6: Partition by FASBM. Color reflects the inferred community membership.

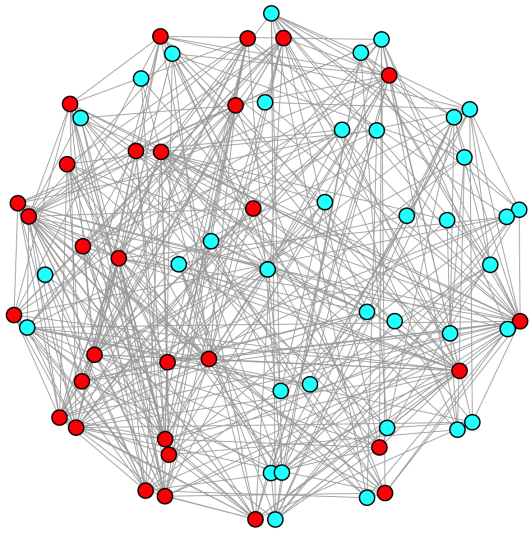
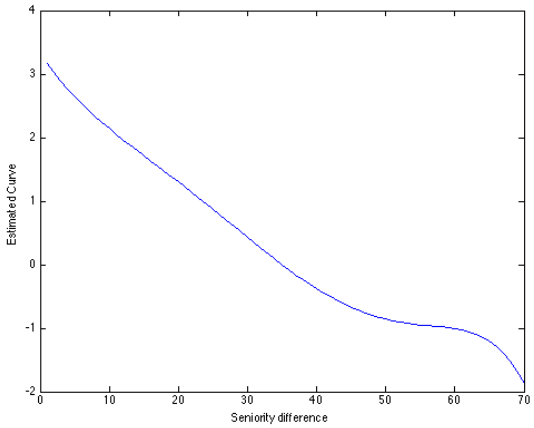


Figure 3.7: Estimated curve against the seniority difference for the Lazega lawyers network data.



3.7 CONCLUSIONS AND FUTURE WORK

In this chapter, we have demonstrated how one can incorporate node feature information on the basis of stochastic block models, focusing on the problem of community detection beyond that explained by the node features as well as learning the influence of features on the network topology. The empirical results show that the proposed method can estimate f non-parametrically, requiring no prior knowledge of how and the extent to which the network is affected by the features. The proposed feature adjusted stochastic block model (FASBM) can be used as generative models for estimation and prediction in networks, making probabilistic statements about the impact of features and so on. Useful extensions include models for directed networks and overlapped communities, and we leave these for future work. The proposed FASBM may be easily generalized to directed networks by relaxing the assumption that $\boldsymbol{\theta}$ in the model is symmetric, and then in the estimation steps we will use all the data Y_{ij} instead of the upper triangular.

In the following, we discuss several computational issues. First, the estimation of f using local likelihood smoothing can be computationally intensive. Fan & Chen (1999) and Cai et al. (2000) proposed to replace the iterative local MLE with the one-step Newton Ralphson estimator and proved in theory that the one-step local MLE does not deteriorate performance as long as the initial estimator is reasonably accurate. The choice of bandwidth in the estimation of f controls how smooth the fit is. Since we have $m \times (m-1)/2$ data points for the curve fitting, the design is very dense. Our practical experience suggests that use of one-tenth of the total range as bandwidth usually results in a relatively smooth f function. Other data-driven methods developed in kernel smoothing although time consuming can also be used. Given that the design can be extremely dense and the curve is usually fairly smooth, we implemented the option allowing one randomly sample a grid of points to fit the curve. Alternative methods such as binned and updated method (Fan & Marron, 1994) can also be considered. In addition to these accelerating methods, one can also adopt other non-parametric smoothing methods to estimate f. Second, like the classic SBM and its variants, the number of communities K in the FASBM has to be pre-specified. In the paper we adapted the network cross-validation (NCV) method for stochastic block model

proposed by Chen & Lei (2014) because the extension of NCV to FASBM is conceptually straightforward. Recently, there are other methods of choosing K developed for SBM based on likelihood approaches, which might also be useful for FASBM. Last but not least, we used greedy-algorithm to avoid a full search of the possible partitions in the model fitting. This algorithm works very well in practice but so far there is no theoretical guarantee of the convergence to the global maximum. We believe that the development of approximation theories for these greedy algorithms is of interest.

4.0 REMARKS

Schizophrenia is a chronic, severe, debilitating mental illness that affects about 1% of the population according to the introduction of the disease on the NIMH website. It is a complex and heterogeneous disorder spanning a broad range of clinical symptoms. There is an enormous amount of research to promote the understanding of this disease from broad aspects.

For example, basic neurobiological studies have focused on identifying possible abnormalities in the neurobiological characteristics linking the pathology, pathophysiology and clinical features of schizophrenia. Studies have unveiled distinct subtypes of schizophrenia. Moreover, there is increasing genetic evidence showing certain core features are shared between schizophrenia and other disorders. These findings point to the scientific hypothesis of shared disease mechanisms among different disorders and to the need for developing suitable approaches to testing statistical hypotheses that are generated from the scientific hypotheses of interest. In Chapter 2, we develop statistical methodologies to assess whether or not subtypes identified from independent populations exhibit commonalities and have successfully applied the proposed methods to a GABA neuron-related biomarker study consisting of subjects with bipolar disorder and subjects with schizophrenia. We provided statistical evidence that two subtypes characterized by differential neurobiological characteristics could identify subsets of subjects with bipolar disorder and subjects with schizophrenia.

In parallel, sustained efforts have been made on characterizing neurocognitive development in schizophrenia. The advent of modern neuroimaging techniques such as fMRI makes it feasible to quantify different aspects of brain functional interactions. The study of functional brain networks advances the understanding of the course of schizophrenia. The fundamental problem in the inference for brain connectivity networks concerns partitioning

of functionally distinct brain regions. We identify the brain segmentation problem conceptually as a community detection problem in network analysis. In Chapter 3, we applied the proposed community detection method in a brain functional connectivity study and have successfully identified basal ganglia subdivisions by only considering functional connectivity pattern between basal ganglia voxels. The resulting parcellation closely resembled that reported using structural anatomical information.

Both parts of research in this dissertation are inspired from conceptually important problems in schizophrenia research, but are equally applicable to other mental disorders. We are working on statistical challenges inherent in solving these problems in hoping to provide new insights into understanding of the disease and ultimately propel the development of early detection, new and more effective treatments and even prevention for schizophrenia.

Acknowledgment: This research was financially supported by Dr. Volk and Dr. Lewis's grant. The author would like to thank Dr. Volk and Dr. Lewis for their financial support and research help.

BIBLIOGRAPHY

Aguirregabiria, V., & Ho, C.-Y. (2010). A dynamic game of airline network competition: Hub-and-spoke networks and entry deterrence. *International Journal of Industrial Organization*, *28*(4), 377–382.

Barnes, K. A., Cohen, A. L., Power, J. D., Nelson, S. M., Dosenbach, Y. B., Miezin, F. M., Petersen, S. E., & Schlaggar, B. L. (2010). Identifying basal ganglia divisions in individuals using resting-state functional connectivity MRI. *Frontiers in systems neuroscience*, *4*.

Berger, R. L., Hsu, J. C., et al. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*(4), 283–319.

Bickel, P. J., & Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068–21073.

Bickel, P. J., & Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, *41*(3), 561–575.

Binkiewicz, N., Vogelstein, J. T., & Rohe, K. (2014). Covariate-assisted spectral clustering. *ArXiv e-prints*.

Brown, L. D., Hwang, J. G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, (pp. 2345–2367).

Caffo, B., Lauzon, C., & Röhmel, J. (2013). Correction to easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician*, *67*(2), 115–116.

Cai, Z., Fan, J., & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, *95*(451), 888–902.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine*, *19*(9), 1141–1164.

Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S. H., & Zhang, M. Q. (2002). Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statistica Sinica*, (pp. 241–262).

Chen, K., & Lei, J. (2014). Network Cross-Validation for Determining the Number of Communities in Network Data. *ArXiv e-prints*.

Choi, D. S., Wolfe, P. J., & Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, (p. asr053).

Chow, S.-C., & Liu, J.-P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies*. CRC Press.

Craddock, N., O'Donovan, M. C., & Owen, M. J. (2005). The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *Journal of Medical Genetics*, *42*(3), 193–204.

Datta, S., & Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, *19*(4), 459–466.

Davison, A., & Hinkley, D. (1997). *Bootstrap Methods and Their Application*, vol. 1. Cambridge University Press.

Di Martino, A., Scheres, A., Margulies, D., Kelly, A., Uddin, L., Shehzad, Z., Biswal, B., Walters, J., Castellanos, F., & Milham, M. (2008). Functional connectivity of human striatum: a resting state fmri study. *Cerebral cortex*, *18*(12), 2735–2747.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, (pp. 189–212).

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (2009). *Multiple testing problems in pharmaceutical statistics*. CRC Press.

Doherty, J. L., & Owen, M. J. (2014). Genomic insights into the overlap between psychiatric disorders: implications for research and clinical practice. *Genome Med*, *6*(4), 29.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology*, *3*(7), research0036.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*(457), 77–87.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *canadian Journal of Statistics*, *9*(2), 139–158.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.

Fan, J., & Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, (pp. 927–943).

Fan, J., & Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, *3*(1), 35–56.

FDA Guidance, F. (1992). Statistical procedures for bioequivalence studies using a standard two treatment crossover design. *US Food and Drug Administration Center for Drug Evaluation and Research*.

FDA Guidance, F. (2001). Statistical approaches to establishing bioequivalence. *Center for Drug Evaluation and Research. United States Food and Drug Administration*.

Girvan, M., & Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

Goldenberg, A., Zheng, A., Fienberg, S., & Airoldi, E. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, *2*(2), 129–233.

Gordon, A. (1999). *Classification*. Chapman and Hall/CRC.

Guidotti, A., Auta, J., Davis, J. M., et al. (2000). Decrease in reelin and glutamic acid decarboxylase67 (gad67) expression in schizophrenia and bipolar disorder: a postmortem brain study. *Archives of general psychiatry*, *57*(11), 1061–1069.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, (pp. 927–953).

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer.

Handcock, M. S., Raftery, A. E., & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301–354.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, *97*(460), 1090–1098.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, *5*(2), 109–137.

Hwang, K., Hallquist, M., & Luna, B. (2012). The development of hub architecture in the human functional brain network. *Cerebral Cortex*, *23*, 2380–2393.

Jordan, F., & Bach, F. (2004). Learning spectral clustering. *Adv. Neural Inf. Process. Syst*, *16*, 305–312.

Karlis, D., & Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, *41*(3), 577–590.

Karrer, B., & Newman, M. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, *83*(1), 016107.

Kerr, M. K., & Churchill, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, *98*(16), 8961–8965.

Kimoto, S., Zaki, M. M., Bazmi, H. H., & Lewis, D. A. (2015). Altered markers of cortical γ-aminobutyric acid neuronal activity in schizophrenia: role of the narp gene. *JAMA psychiatry*, *72*(8), 747–756.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Science & Business Media.

Kvalseth, T. O. (1987). Entropy and correlation: Some comments. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*(17), 517–519.

Lauzon, C., & Caffo, B. (2009). Easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician*, *63*(2), 147–154.

Lazega, E. (2001). The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership.

Lei, J., Rinaldo, A., et al. (2014). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, *43*(1), 215–237.

Liu, C., & Sun, D. X. (1997). Acceleration of EM algorithm for mixtures models using ECME. In *ASA proceedings of the stat. comp. session*, (pp. 109–114).

Liu, X., Murata, T., & Wakita, K. (2014). Detecting network communities beyond assortativity-related attributes. *Physical Review E*, *90*(1), 012806.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, (pp. 911–916). IEEE.

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., & Bullmore, E. (2010). Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, *30*(28), 9477–9487.

Ma, J., & Fu, S. (2005). On the correct convergence of the em algorithm for gaussian mixtures. *Pattern Recognition*, *38*(12), 2602–2611.

McLachlan, G., & Krishnan, T. (2007). *The EM Algorithm and Extensions*, vol. 382. John Wiley & Sons.

McLachlan, G., & Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, (pp. 415–444).

Moskvina, V., Craddock, N., Holmans, et al. (2009). Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Molecular Psychiatry*, *14*(3), 252–260.

Munk, A., Hwang, J. G., & Brown, L. D. (2000). Testing average equivalence finding a compromise between theory and practice. *Biometrical Journal*, *42*(5), 531–552.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, *69*(6), 066133.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Nowicki, K., & Snijders, T. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087.

Quan, H., Bolognese, J., & Yuan, W. (2001). Assessment of equivalence on multiple endpoints. *Statistics in Medicine*, *20*(21), 3159–3173.

Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, *101*(473), 168–178.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, *26*(2), 195–239.

Robinson, S., Basso, G., Soldati, N., Sailer, U., Jovicich, J., Bruzzone, L., Kryspin-Exner, I., Bauer, H., & Moser, E. (2009). A resting state network in the motor control circuit of the basal ganglia. *BMC neuroscience*, *10*(1), 137.

Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, *39*, 1878–1915.

Röhmel, J. (2011). On familywise type i error control for multiplicity in equivalence trials with three or more treatments. *Biometrical Journal*, *53*(6), 914–926.

Schilling, M. F., Watkins, A. E., & Watkins, W. (2002). Is human height bimodal? *The American Statistician*, *56*(3), 223–229.

Schuirmann, D. (1981). On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval. *Biometrics*, *37*(3), 617–617.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.

Sibille, E., Morris, H. M., Kota, R. S., & Lewis, D. A. (2011). Gaba-related transcripts in the dorsolateral prefrontal cortex in mood disorders. *International Journal of Neuropsychopharmacology*, *14*(6), 721–734.

Snijders, T., & Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, *14*(1), 75–100.

Snijders, T., Pattison, P., Robins, G., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological methodology*, *36*(1), 99–153.

Stanley, M. L., Moussa, M. N., Paolini, B. M., Lyday, R. G., Burdette, J. H., & Laurienti, P. J. (2013). Defining nodes in complex brain networks. *Frontiers in computational neuroscience*, *7*.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.

Tibshirani, R., et al. (2007). Are clusters found in one dataset present in another dataset? *Biostatistics*, *8*(1), 9–31.

Tziortzi, A. C., Searle, G. E., Tzimopoulou, S., Salinas, C., Beaver, J. D., Jenkinson, M., Laruelle, M., Rabiner, E. A., & Gunn, R. N. (2011). Imaging dopamine receptors in humans with [11 c]-(+)-phno: dissection of d3 signal and anatomy. *Neuroimage*, *54*(1), 264–277.

Van Dijk, K. R., Hedden, T., Venkataraman, A., Evans, K. C., Lazar, S. W., & Buckner, R. L. (2010). Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *Journal of neurophysiology*, *103*(1), 297–321.

Viennet, E., et al. (2012). Community detection based on structural and attribute similarities. In *ICDS 2012, The Sixth International Conference on Digital Society*, (pp. 7–12).

Volk, D., Sampson, A., Zhang, Y., Edelson, J., & Lewis, D. (2016). Cortical gaba markers identify a molecular subtype of psychotic and bipolar disorders. *Psychological Medicine*, (pp. 1–12).

Volk, D. W., Edelson, J. R., & Lewis, D. A. (2014). Cortical inhibitory neuron disturbances in schizophrenia: role of the ontogenetic transcription factor lhx6. *Schizophrenia bulletin*, *40*, 1053–1061.

Volk, D. W., Matsubara, T., Li, S., Sengupta, E. J., Georgiev, D., Minabe, Y., Sampson, A., Hashimoto, T., & Lewis, D. A. (2012). Deficits in transcriptional regulators of cortical parvalbumin neurons in schizophrenia. *American Journal of Psychiatry*, *169*(10), 1082–1091.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395–416.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Woo, T.-U. W., Kim, A. M., & Viscidi, E. (2008). Disease-specific alterations in glutamatergic neurotransmission on inhibitory interneurons in the prefrontal cortex in schizophrenia. *Brain research*, *1218*, 267–277.

Yang, J., McAuley, J., & Leskovec, J. (2013). Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining (ICDM)*, (pp. 1151–1156). IEEE.

Yeung, K. Y., Haynor, D. R., & Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics*, *17*(4), 309–318.

Zhang, Y., Levina, E., & Zhu, J. (2015). Community Detection in Networks with Node Features. *ArXiv e-prints*.