CLUSTERING METHODOLOGIES WITH APPLICATIONS TO INTEGRATIVE ANALYSES OF POST-MORTEM TISSUE STUDIES IN SCHIZOPHRENIA

by

Qiang Wu

B.S., University of Science and Technology of China, China, 2002

M.A., University of Pittsburgh, USA, 2007

Submitted to the Graduate Faculty of

the Department of Statistics in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF STATISTICS

This dissertation was presented

by

Qiang Wu

It was defended on

August 2, 2007

and approved by

Allan R. Sampson, Ph.D, Professor

David A. Lewis, M.D., Professor

Leon J. Gleser, Ph.D, Professor

Satish Iyengar, Ph.D, Professor

Dissertation Director: Allan R. Sampson, Ph.D, Professor

CLUSTERING METHODOLOGIES WITH APPLICATIONS TO INTEGRATIVE ANALYSES OF POST-MORTEM TISSUE STUDIES IN SCHIZOPHRENIA

Qiang Wu, PhD

University of Pittsburgh, 2007

There is an enormous amount of research devoted to the understanding of the neurobiology of schizophrenia. Basic neurobiological studies have focused on identifying possible abnormal neurobiological markers in subjects with schizophrenia. However, due to the many possible combinations of symptoms, schizophrenia is clinically thought not to be a homogeneous disease, so that this possible heterogeneity might be explained neurobiologically in various brain regions. Statistically, the interesting problem is to cluster the subjects with schizophrenia with these neurobiological markers. But, in attempting to combine the neurobiological measurements from multiple studies, several experimental specifics arise that lead to difficulties in developing statistical methodologies for the clustering analysis. The main difficulties are differing control subjects, effects of covariates and existence of missing data. We develop new parametric models to successively deal with these difficulties. First, assuming no missing data and no clusters we construct multivariate normal models with structured means and covariance matrices to deal with the differing control subjects and the effects of covariates. We obtain several parameter estimation algorithms for these models and the asymptotic properties of the resulting estimators. Using these newly obtained results, we then develop model based clustering algorithms to cluster the subjects with schizophrenia into two possible subpopulations while still assuming no missing data. We obtain a new more effective algorithm for clustering and show by simulations that our new algorithm provides the same results in a relatively faster manner as compared to direct applications of some existing

algorithms.

Finally, for some actual data obtained from three studies conducted in the Conte Center for the Neuroscience of Mental Disorders in the Department of Psychiatry at the University of Pittsburgh, to handle the missingness we conduct imputations to create multiply imputed data sets using certain regression methods. The new complete data clustering algorithm is then applied to the multiply imputed data sets. The resulting multiple clustering results are integrated to form one single clustering of the subjects with schizophrenia to represent the uncertainty due to the missingness. The results suggest the existence of two possible clusters of the subjects with schizophrenia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I am deeply indebted to my advisor, Dr. Allan R. Sampson, for his support, commitment, and patience. He encouraged me to think independently and develop research skills which I would benefit from in my whole career. He generously devoted his time to read and revise my draft, and to provide me stimulating advices on my research. He has a warm heart and treated me with kindness.

I would like to sincerely thank my committee members, Dr. David A. Lewis, Dr. Leon J. Gleser and Dr. Satish Iyengar, for spending their time reading my draft. Particularly, I thank Dr. Leon J. Gleser for his constructive comments on the first part of my research. And I thank Dr. David A. Lewis for his interest in my research and his generosity in allowing me to use the data from post-mortem tissue studies conducted in his lab.

I extend many thanks to my friends and collaborators, especially Dr. Takanori Hashimoto. He is not only my collaborator but also one of my best friends. He was of great help in clarifying certain basic neurobiological ideas and collecting the data. I thank Dr. Zhuoxin Sun, a former student of my advisor, for her generous help in the early stage of my research. As a successor of her graduate student researcher position, I learned a lot from her. I also would like to thank Ms. Ana-Maria Iosif for correcting my writing.

Finally, I would like to thank my family. Last year, my wife presented me a special and precious gift – our son Kevin. The happiness of having this lovely family was my backbone for pursuing my American dream and completing my PhD degree. I also owe a lot to my other family members, especially my mother and mother-in-law for taking care of baby Kevin, which provided me enough time for my study.

This research was financially supported both by my advisor, Dr. Allan R. Sampson, and by the Department of Statistics.

1.0 INTRODUCTION

Schizophrenia is a chronic, severe, and disabling brain disease, characterized mainly by the impairment of certain cognitive functions, such as working memory. Neuroscientists are using many approaches to understand the neurobiology of this disease with the ultimate goal to develop more effective clinical treatments. The Conte Center for the Neuroscience of Mental Disorders (CCNMD) in the Department of Psychiatry at the University of Pittsburgh is heavily involved in conducting basic neurobiological research concerning schizophrenia. A major research interest of the Center is to use post-mortem tissue samples to detect neurobiological alterations in subjects with schizophrenia (for example, see Konopaske et al. (2005)). These studies are conducted involving differing neurobiological measurements on various brain regions of subjects from the Brain Bank Core of the Center. While individual studies typically address the possible abnormality of a single neurobiological marker, the potential to combine the data from multiple studies would provide an opportunity to synthesize the data collected in the Center's studies and possibly produce new insights into the understanding of schizophrenia. We are aware of only one previous attempt at such a data synthesis in schizophrenia research. This study involved tissue studies from the Stanley Foundation based on a single cohort of subjects with psychiatric disorders and control subjects and focused on identifying various neurobiological markers which distinguished subjects from the different diagnostic groups (Knable et al., 2001, 2002). Their combined data set consisted of 60 subjects from four different diagnostic groups, including schizophrenia, bipolar disorder, non-psychotic depression and normal, and a total of 102 different neurobiological markers. The authors implemented a linear discriminant function (LDF) model and a classification and regression tree (CART) model, in addition to the regular analysis of variance (ANOVA) model, to identify subsets of neurobiological markers that discriminated

subjects with psychiatric disorders from normal controls. Use of the LDF and CART models instead of the ANOVA model helps to reduce the rate of false discovery. However, we are interested in another research direction. Due to the many combinations of symptoms, schizophrenia is clinically thought not to be a homogeneous disease, so that this heterogeneity might be explained neurobiologically in the various brain regions. As a result, another way of synthesizing the data is to develop new statistical methods to identify possible subpopulations of subjects with schizophrenia by examining these bio-markers. The ultimate clinical goal would then be to relate these subpopulations of subjects with schizophrenia to clinical information concerning the subjects. The statistical methodology we develop to address this synthesis is framed generally enough to be applicable in other settings with similar structures.

Model based clustering techniques have been widely studied and implemented in practice for decades, especially with the emergence of the Expectation and Maximization (EM) algorithm introduced by Dempster, Laird, and Rubin (1977). It enables us to model the heterogeneity of the data of various complicated structures where other clustering methodologies are less possible, e.g., when there exist both an outcome and some covariates. In addition, in the multivariate settings and when some data are missing, the distance metrics required by some procedures are very difficult to define, especially for those cases with nonidentical missing patterns. In this dissertation, we focus on the clustering problem for multivariate normal distributed data with structured means and covariance matrices. Fortunately, there has also been a fair amount of research since the 1960s conducted concerning estimation and testing for the multivariate normal distribution with structured means and covariance matrices (See, for example, Anderson, 1969, 1970, 1973; Szatrowski, 1979, 1980, 1983; Rubin and Szatrowski, 1982; Jennrich and Schluchter, 1986). The structured forms for means and covariance matrices arise in many settings, for example, educational or biological studies. In a biological setting, the patterned mean structures come from the existence of covariates, and the patterned covariance structures result from the biological symmetry within subjects and the consideration of random effects. The particular statistical distributional structures that are focused upon in this dissertation arise from our goal of synthesizing data across the Center's multiple post-mortem tissue studies concerning schizophrenia with the objec-

tive of identifying possible subpopulations of subjects with schizophrenia and associated bio-markers that show similar neurobiological characteristics.

The statistical work of clustering subjects with schizophrenia would be easier if the data on multiple bio-markers within or across various brain regions were simultaneously obtained on the same set of subjects. However, this usually cannot be achieved due to both the time constraints and the high costs of such kind of studies. In the attempt to synthesize the data from multiple studies in the Center, several specifics of the data arise and lead to distributional structures more general than those previously considered. For a number of studies that involve repeated measures, e.g., across different brain layers, there are pertinent ways to combine them into one single observation per subject. For instance, the sum over all layers is an appropriate choice for the total number of neurons, while the average is to be used for mRNA expression levels. In each study, every subject with schizophrenia has been matched with a control subject based on age at death, gender and post-mortem interval. As a result, we use appropriate within pair differences as the primary data in our analysis. The reason for doing this is to control for both experimental and demographical variations with details discussed in Chapter 2. However, in a number of cases, due to the availability of tissue samples and other experimental constraints, different controls might have been paired with the same subject with schizophrenia when that subject is used in different studies. This introduces covariance matrices with differing structures. Furthermore, various demographic measurements, such as duration of the disease, brain pH value and storage time of tissue sample, are also available for each subject in addition to age, gender and post-mortem interval. Some of them, e.g., age, are often informative about the neurobiological measurements, while others, e.g., post-mortem interval, brain pH and storage time, are only experimental adjustments to attempt to recover the tissue status at time of death. Hence, we consider a selected subset of the demographic characteristics as covariates in the clustering analysis. Finally, while some studies use the same sets of subjects with schizophrenia, others have overlapping sets, and yet others have disjoint sets. New subjects are frequently used in studies, while some older ones are much less frequently used. This partial usage of the subjects with schizophrenia creates much missingness in combining data from multiple studies. Specifically, for each subject with schizophrenia and its corresponding

controls, not all the observations over all studies are available. Moreover, if missing data occur, then the relationships between the missing and the observed control subjects matched with the same subject with schizophrenia are also unavailable. The details of our motivating data are provided in Chapter 2.

The outline of the remainder of this dissertation is as follows. In Chapter 3, we review some existing literature on the topic of structured means and covariance matrices, as well as some basic model-based clustering techniques including the classic mixture modeling. In Chapter 4, we develop and evaluate some new multivariate models to deal with the structured means and covariance matrices arising from our specific settings, assuming no clustering and no missing data. Following a more detailed literature review on some modern model based clustering techniques, in Chapter 5, model based clustering algorithms are then built upon these new generalizations of the structured models still with the assumption of no missing data. A new algorithm is shown to provide the same clustering result as the existing EM gradient algorithm in a relatively faster manner. Finally, in Chapter 6 we apply the new clustering algorithm to the combined data from multiple post-mortem tissue studies with help of some multiple imputation techniques to deal with the missingness.

The review chapter, Chapter 3, focuses mainly on the work of Anderson (1969, 1970, 1973), Szatrowski (1979, 1980, 1983) and Jennrich and Schluchter (1986). In addition, some general issues concerning classic mixture models and some computational issues including the EM algorithm are also reviewed, since they contain the basic idea of model based clustering. The review section in Chapter 5 reviews some recent work of DeSarbo and Corn (1988), Jones and McLachlan (1992), Arminger, Stein, and Wittenberg (1999) and Zhang (2003) regarding model based clustering.

As an initial step in clustering subjects with schizophrenia, we require using new multivariate models and developing their corresponding model fitting algorithms without the assumptions of clusters and missing data. We present these results in Chapter 4. While these models ultimately will be required to implement our clustering approaches, they are of interest in their own right. Several model fitting algorithms, including the Method of Scoring and the Newton-Raphson, are considered for parameter estimation and the relevant asymptotic distributions are obtained. In addition, a one-iteration estimator using the

Method of Scoring algorithm starting from a consistent starting point is show to be asymptotically equivalent to the MLE. Simulations are then provided to verify the key asymptotic results. In the analysis, the vector of the pairwise differences across different studies sharing the same subject with schizophrenia is treated as having a multivariate normal distribution with patterned mean and covariance structures. The particular structures, we develop, result from two specific factors concerning the Center's studies, that is, the differing control subjects and the existence of nonidentical covariates. The factor of differing control subjects creates patterned covariance structures, while the factor of nonidentical covariates results in patterned structures of the means.

Based on the new multivariate models with patterned mean and covariance structures, model based clustering techniques are built in Chapter 5 still with the assumption of no missing data. The data are now assumed to come from a mixture of two different multivariate normal distributions with patterned mean and covariance structures. Several existing algorithms, including the EM gradient algorithm (Lange, 1995) and Titterington's (1984) (Titterington, 1984) algorithm, are considered to cluster the subjects with schizophrenia into two possible subpopulations. A new algorithm is then developed and shown to provide the same clustering results in a relative faster manner. Simulations are given to compare this new algorithm to the existing ones.

The actual data obtained from multiple post-mortem tissue studies has a large scale of missingness. As a result, the clustering algorithms discussed in Chapter 5 cannot be directly applied. Directly working on the observed data is also intractable given the complicated structures of our data. Nevertheless, with the assumption of a missing completely at random (MCAR) missing mechanism, imputation techniques can be implemented to impute the missing data. Then, the clustering algorithms in Chapter 5 can be applied to the imputed data. In order to represent the uncertainty due to the missingness, multiple imputations are conducted and the clustering results from the multiple imputed data are combined to form a single clustering of the subjects with schizophrenia. Finally, some graphical summaries are obtained based on the observed data to understand the differences between the two clusters. The details of this application are discussed in Chapter 6.

2.0 MOTIVATING DATA

2.1 AN OVERVIEW OF POST-MORTEM TISSUE STUDIES

As of December 31, 2005, the post-mortem tissue data from subjects with psychiatric disorders in the Center consists of about 50 subjects with schizophrenia and 80 control subjects from the Brain Tissue Bank. Approximately 35 separate post-mortem tissue studies have been conducted in Dr. Lewis's lab. Limited historical information, such as diagnostic records, behavior pattern, usage of drugs and cause of death, as well as the demographic characteristics, have been obtained for these subjects. These subjects, especially the ones with schizophrenia, have been repeatedly used for studies conducted in Dr. Lewis's Lab. In each study one or more neurobiological characteristics in particular brain regions have been measured and analyzed mainly with the analysis of covariance (ANCOVA) model or its multivariate version (MANCOVA). The primary purpose of these studies is to detect possible neurobiological alterations in the subjects with schizophrenia as compared to the corresponding controls with the consideration of certain adjusting factors such as the demographic characteristics. In each study, every subject with schizophrenia has been matched with a control subject based upon certain demographic characteristics. In pairing, the matched subjects have their ages at death, gender and post-mortem intervals as close as possible. The tissue samples from the matched pairs are then blinded and processed together. However, due to the availability of tissue samples and other experimental constraints, different control subjects might have been paired with the same subject with schizophrenia across different studies. Also, different subsets, typically 10-30 subjects, of the subjects with schizophrenia have been used across different studies, which conceptually introduces a large amount of missingness when we want to combine the data.

With the opportunity of combining the post-mortem tissue data from multiple studies, two interesting questions can be raised. First, as we have mentioned, schizophrenia might not be a uniform disease. So the first question is whether we can identify some meaningful subclasses of the subjects with schizophrenia based on the post-mortem tissue data. Statistically, the problem of interest is to attempt to cluster the subjects with schizophrenia to examine the possible heterogeneity of the disease. Second, the Center's studies explore different bio-markers implicated with the disease, where there might be neurobiological relationships. As a result, it is more likely that different choices of studies would yield different clusterings of the subjects with schizophrenia. This suggests possibly needing to use some simultaneous clustering methods to find bio-markers showing similar neurobiological characteristics. Our far-reaching goal is then to find neurobiologically related bio-markers and relate the clustering of the subjects with schizophrenia with these bio-markers. A further goal would be to compare any clustering results with the limited amount of clinical data available for subjects in these post-mortem studies, by which we may be able to provide new insights to clinicians. In this research, we focus on clustering of the subjects with schizophrenia with a pre-selected subset of the bio-markers, and leave the bi-clustering for the future. We try to limit the pre-selection of the bio-markers to those showing significant alterations in previous studies and, in part, with the consultation of investigators in the Center.

In integrating data from multiple studies, several special features of the data require us to develop new statistical methodologies. The several difficulties include the existence of differing control subjects across different studies for the same subjects with schizophrenia, the existence of covariates for each subject and the large amount of missing data. In addition, in a number of studies there are repeated measurements over multiple brain regions. When this happens, there will be pertinent ways to combine the repeated measurements into one single observation per subject to reduce the complexity of computation without losing significant information. For instance, the sum over all layers will be an appropriate choice for neuron number, while the average is to be used for mRNA expression levels. In this dissertation we will focus on the parameter estimation and clustering of the subjects with schizophrenia with reasonable model assumptions by successively dealing with the problems of the differing controls, the existence of covariates and the missing data.

2.2 DIFFERING CONTROL SUBJECTS

In biological experiments, it is usually plausible to assume that observations from the same subjects are correlated, while observations from different subjects are not. However, since the tissue samples from paired subjects in one study are prepared and processed together, the corresponding observations might be affected by common experimental variations, such as the ambient temperature when processed, and the density of the staining solution used in the experiment. In order to control the experimental, as well as the demographical variations on which the pairing is defined, the pairwise differences of observations on a subject with schizophrenia and its corresponding controls are obtained and often in the neuroscience literature treated as the primary data. Then, the vector of the paired differences across studies for the same subject with schizophrenia is treated as a random vector having a multivariate normal distribution. The covariance between the pairwise differences involving the same subject with schizophrenia from two studies depends on whether or not the pairs share the same control subject. For instance, let $\{S_{i1}, S_{i2}, \cdots, S_{ip}\}$ be the measurements from p studies on the i^{th} subject with schizophrenia, $i = 1, \ldots, n$, and let $\{C_{i1}, C_{i2}, \cdots, C_{ip}\}$ be the corresponding measurements on the control subjects paired with the i^{th} subject with schizophrenia in these studies, where $\{C_{i1}, C_{i2}, \cdots, C_{ip}\}$ might not be from the same subject. Now consider the differences $\{S_{i1} - C_{i1}, S_{i2} - C_{i2}, \cdots, S_{ip} - C_{ip}\}$. It is clear that

$$Cov(S_{ij} - C_{ij}, S_{ij'} - C_{ij'}) = Cov(S_{ij}, S_{ij'}) + Cov(C_{ij}, C_{ij'}) \quad \text{for} \quad j \neq j', \tag{2.1}$$

since $Cov(S_{ij}, C_{ij}) = 0$, because S_{ij} and C_{ij} are always from different subjects. Then for those observations where $\{C_{ij}, C_{ij'}\}$ happen to be from the same control subject, we have $Cov(C_{ij}, C_{ij'}) \neq 0$; otherwise, we have $Cov(C_{ij}, C_{ij'}) = 0$. So the covariance matrices for the n vectors of the differences will not be identical. This feature of the covariance matrices causes difficulty in the analysis only when the assignments of control subjects cause the resulting covariance matrices for all the differences to have two or more different forms. Otherwise, we can treat the problem as an ordinary multivariate regression problem.

To clarify these ideas, we consider in Table 2.1 a prototypical example where there are $p = 3$ studies. As can be seen from the table, there are a total of 5 possible different covariance

Table 2.1: A prototype for dimension $p = 3$

Case	Controls in study 1	Controls in study 2	Controls in study 3	Covariance matrices
1	#1	#2	#3	$\begin{bmatrix} \sigma^s_{11}+\sigma^c_{11} & \sigma^s_{12} & \sigma^s_{13} \\ \sigma^s_{21} & \sigma^s_{22}+\sigma^c_{22} & \sigma^s_{23} \\ \sigma^s_{31} & \sigma^s_{32} & \sigma^s_{33}+\sigma^c_{33} \end{bmatrix}$
2	#1	#1	#2	$\begin{bmatrix} \sigma^s_{11}+\sigma^c_{11} & \sigma^s_{12}+\sigma^c_{12} & \sigma^s_{13} \\ \sigma^s_{21}+\sigma^c_{21} & \sigma^s_{22}+\sigma^c_{22} & \sigma^s_{23} \\ \sigma^s_{31} & \sigma^s_{32} & \sigma^s_{33}+\sigma^c_{33} \end{bmatrix}$
3	#1	#2	#1	$\begin{bmatrix} \sigma^s_{11}+\sigma^c_{11} & \sigma^s_{12} & \sigma^s_{13}+\sigma^c_{13} \\ \sigma^s_{21} & \sigma^s_{22}+\sigma^c_{22} & \sigma^s_{23} \\ \sigma^s_{31}+\sigma^c_{31} & \sigma^s_{32} & \sigma^s_{33}+\sigma^c_{33} \end{bmatrix}$
4	#1	#2	#2	$\begin{bmatrix} \sigma^s_{11}+\sigma^c_{11} & \sigma^s_{12} & \sigma^s_{13} \\ \sigma^s_{21} & \sigma^s_{22}+\sigma^c_{22} & \sigma^s_{23}+\sigma^c_{23} \\ \sigma^s_{31} & \sigma^s_{32}+\sigma^c_{32} & \sigma^s_{33}+\sigma^c_{33} \end{bmatrix}$
5	#1	#1	#1	$\begin{bmatrix} \sigma^s_{11}+\sigma^c_{11} & \sigma^s_{12}+\sigma^c_{12} & \sigma^s_{13}+\sigma^c_{13} \\ \sigma^s_{21}+\sigma^c_{21} & \sigma^s_{22}+\sigma^c_{22} & \sigma^s_{23}+\sigma^c_{23} \\ \sigma^s_{31}+\sigma^c_{31} & \sigma^s_{32}+\sigma^c_{32} & \sigma^s_{33}+\sigma^c_{33} \end{bmatrix}$

matrices for a single observation, where $\sigma^s_{jk} = Cov(S_j, S_k)$ and $\sigma^c_{jk} = Cov(C_j, C_k)$ when C_j and C_k are from the same control subject for $j, k = 1, 2, 3$. In general, there are a total of $2^p - p$ possible different covariance matrices determined by a total of p^2 free parameters. In Table 2.1, case 1 corresponds to the setting where there are three different controls for a subject with schizophrenia, so that the resulting covariance matrix is as shown; whereas in case 2, the same control subjects are used in studies 1 and 2, and a different control subject is used in study 3, so a term σ^c_{12} is added to the covariance term between study 1 and 2. The rest of the cases can be explained in the same way.

2.3 INCORPORATING COVARIATES

In our motivating data, each subject with schizophrenia has their own age, gender, post-mortem interval, tissue storage time and so forth. The typical primary ANCOVA or MAN-

COVA model used in analyzing individual studies has diagnostic group as the main effect, pair as a pairing effect and brain pH and storage time as covariates. In the typical secondary model employed in the analysis of an individual study, pair is replaced by the covariates age, gender and post-mortem interval and the interactions between the covariates and the main effect are also included. See Konopaske et al. (2005) for an example of these typical analytic approaches. This means that when we take the within-pair differences between the subjects with schizophrenia and the controls these covariates can still have impact. We build the mean structure of our model as a linear function of some of these covariates. The clustering then can be defined in terms of both the main effect, represented by the intercept, and the effects of some covariates, represented by the slopes. The details of choosing effective covariates and defining the clustering are discussed in Section 5.2.

2.4 MISSING DATA

To examine the degree of missingness, a graphic view of the post-mortem data from subjects with psychiatric disorders is constructed by only recording whether or not the data are available. By properly permuting the rows and columns we have Figure 2.1. The columns represent the studies and are labeled by their id numbers in the time orders of the studies. The row labels are the id numbers of the subjects with schizophrenia. An entry "1" in the matrix of Figure 2.1 means the data are observed with a corresponding control subject and "." means not. It can be seen that proper subsets of both the studies and the subjects with schizophrenia may be required in order to do the analysis due to the large scale of missing data.

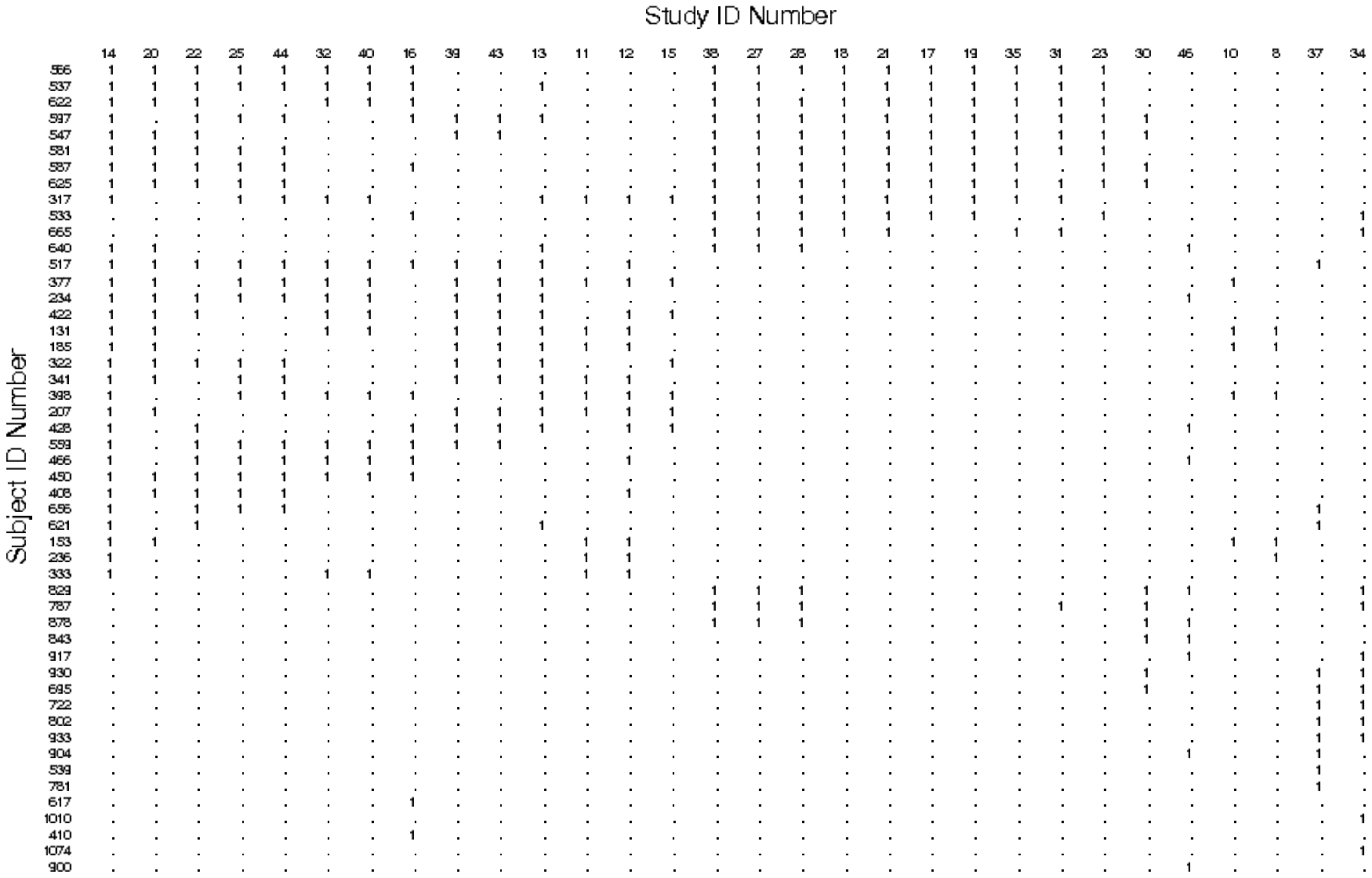


Figure 2.1: Missing Data Indices

When missing data occurs, both observations on the subjects with schizophrenia and on their corresponding control subjects are missing. As a result, the underlying relationship between the missing and the observed controls paired with the same subjects with schizophrenia is also unavailable, which is critical in constructing the covariance matrices. However, the relationship among the controls matched to the same subjects with schizophrenia belongs to the experimental design and should not affect the clustering result. So in our imputation, we assume that if a subject with schizophrenia is not used in a study, then hypothetically in the imputation for that subject in that study, the last corresponding control used in the previous study is assumed. This assumption is for simplicity. Here, it is reasonable to assume the missing mechanism is MCAR, because the subject selection in each study is conducted individually and not related to the neurobiological measurements. For example, one of the main concerns in the subject selection is the quality of the tissue samples. As a result, multiple imputation techniques can then be implemented to deal with the missing data. A detailed application is presented in Chapter 6.

3.0 LITERATURE REVIEW

A substantial amount of research has been focused on parameter estimation, obtaining the asymptotic distributions of the estimates, and deciding testing methodologies for the problem of patterned means and covariances structure. The maximum likelihood estimates of the parameters usually have no closed form, and iterative procedures have been given. The asymptotic properties of the maximum likelihood estimates have been considered in Anderson (1973) and Szatrowski (1983). The test considered is usually a likelihood ratio test. A discussion of models which generalize patterned means and covariances is given in Jennrich and Schluchter (1986). In addition to this literature, we also review the major results on the classic mixture models and some computational issues including the EM algorithm. The review of some more advanced results on clustering algorithms for regression models is given in Section 5.1.

3.1 PATTERNED MEANS AND COVARIANCES MODELS

Let Y_i, $i = 1, \ldots, n$, be independent observations, respectively, from p-dimensional normal distributions, $\mathcal{N}(\mu_i, \Sigma_i)$. Anderson (1969) discusses a model with a linear mean structure $\mu_i \equiv \mu = X\beta = \sum_{j=1}^{r} \beta_j x_j$ and a linear covariance structure $\Sigma_i \equiv \Sigma = \sum_{g=0}^{m} \sigma_g G_g$, $i = 1, \cdots, n$, where $\beta_1, \cdots, \beta_r$ and $\sigma_0, \cdots, \sigma_m$ are unknown coefficients, $x_1, \cdots, x_r$ are known, linearly independent, p-component vectors, and $G_0, \cdots, G_m$ are known, symmetric, linearly independent $p \times p$ matrices. It is assumed that the parameter space is not empty. The maximum likelihood estimates then have no closed form in general, but can be obtained by

solving the likelihood equations

$$\sum_{k=1}^{r} x_j' \Sigma^{-1} x_k \beta_k = x_j' \Sigma^{-1} \bar{Y}, \qquad j = 1, \ldots, r, \tag{3.1}$$

and

$$\sum_{f=0}^{m} tr \Sigma^{-1} G_g \Sigma^{-1} G_f \sigma_f = tr \Sigma^{-1} G_g \Sigma^{-1} C, \qquad g = 0, \ldots, m, \tag{3.2}$$

iteratively, where $C = (1/n) \sum_{i=1}^{n} (Y_i - \mu)(Y_i - \mu)'$. In Section 3.3.1, we discuss the corresponding computational details. Since the log-likelihood function approaches infinity when Σ approaches singularity or some of its elements tend to infinity, Anderson (1970) argued that there was at least one relative maximum in the set of $\sigma_0, \cdots, \sigma_m$ such that $\hat{\Sigma} = \sum_{g=0}^{m} \hat{\sigma}_g G_g$ was positive definite, and if multiple relative maximums existed, the absolute maximum to the likelihood function was attained on the set of solutions minimizing $|\hat{\Sigma}|$. However, in general the iterative solutions to (3.1) and (3.2) are not guaranteed to converge to the MLE. As a result, multiple starting points are required to find the global maximum. However, if the iterations converge, then the estimates are consistent, asymptotically efficient as $n \to \infty$ and have a limiting normal distribution with covariance matrix

$$[Cov(\hat{\beta}_i, \hat{\beta}_j)] = (1/n)[x_i' \Sigma^{-1} x_j]^{-1} \tag{3.3}$$

and

$$[Cov(\hat{\sigma}_g, \hat{\sigma}_h)] = (1/n)[\frac{1}{2} tr \Sigma^{-1} G_g \Sigma^{-1} G_h]^{-1} \tag{3.4}$$

with asymptotic independence between the two sets of estimators, i.e., for the $\hat{\beta}$'s and the $\hat{\sigma}$'s. As shown later by Szatrowski (1983), this iterative algorithm coincides with the Method of Scoring algorithm. Some special cases, e.g., $G_0, \cdots, G_m$ are simultaneously diagonalizable by the same orthogonal matrix, where the general problem is considerably simpler have also been considered by, for example, Srivastava (1966), Graybill and Hultquist (1961) and Herbach (1959). Some of these authors consider likelihood ratio tests which are usually used to test the goodness-of-fit of linear structures. The existence of explicit or one-iteration maximum likelihood estimates for certain cases was considered in Szatrowski (1980). Rubin and Szatrowski (1982) introduced cases where the data can be augmented with "artificial"

missing data so that the expanded problems have explicit solutions. In these cases, the EM algorithm for missing data can be easily implemented to find the MLEs.

In the case of multivariate data analysis, "missing data" or "incomplete data" is a common problem, because we may not observe every component of some observation vectors. For example, Szatrowski (1983) assumed that instead of observing Y_i, we observed $E_{\alpha(i)}Y_i$, $i = 1, \ldots, n$, where E_α, $\alpha = 1, \ldots, q$, were known $u_\alpha \times p$ matrices of full rank with $u_\alpha \leq p$. The function $\alpha(i)$ is given by $\alpha(i) = j$ for $i = m_{j-1}+1, \ldots, m_j; j = 1, \ldots, q, m_0 \equiv 0$. Furthermore, let $n_\alpha = m_\alpha - m_{\alpha-1}$ be the number of observations of the form $E_\alpha \mathbf{y}$ and $f_\alpha = n_\alpha/n$. The following condition was given for the estimability of parameters for this missing data pattern:

Condition 3.1.1 (Szatrowski (1983)). For each j, there exists an α such that $E_\alpha x_j \neq 0$, $j = 1, \ldots, r$, and for each g, there exists an α such that $E_\alpha G_g E'_\alpha \neq 0$, $g = 1, \ldots, m$.

The maximum likelihood estimates were then found using the Newton-Raphson, Method of Scoring, or the EM algorithms. Asymptotic distributions of the maximum likelihood estimates were given assuming another condition which was necessary due to the convergence requirements in the case of missing data:

Condition 3.1.2 (Szatrowski (1983)). $\lim_{n\to\infty}(n_s(t)/n) = \eta_{ts} \in (0, 1)$ for $s = 1$, $t = 1, \ldots, r$ and $s = 2$, $t = 1, \ldots, m$, with $n_1(j) = \sum_{\alpha=1}^{q} n_\alpha 1(E_\alpha x_j \neq 0)$, $j = 1, \ldots, r$, and $n_2(g) = \sum_{\alpha=1}^{q} n_\alpha 1(E_\alpha G_g E'_\alpha \neq 0)$, $g = 1, \ldots, m$, where $1(\cdot)$ is an indicator function.

Szatrowski (1983) extended the asymptotic results given in Anderson (1973) by allowing missing data. The limiting covariance matrices are $(\sum_1^q n_\alpha X'_\alpha \Sigma_\alpha^{-1} X_\alpha)^{-1}$ and $(1/2 \sum_1^q tr \Sigma_\alpha^{-1}$ $G_{g\alpha}\Sigma_\alpha^{-1}G_{h\alpha})^{-1}$ for both sets of the parameters, i.e. the $\hat{\beta}$'s and the $\hat{\sigma}$'s, respectively.

The assumptions in the above models are relatively restrictive. For example, the covariates $X = [x_1, \cdots, x_n]$ for the mean vector and the covariance matrix Σ are assumed to be the same across observations. More general models were discussed by Jennrich and Schluchter (1986), based on earlier work of Harville (1977), Laird and Ware (1982) and Ware (1985). They assumed, instead, that $\mu_i = X_i\beta$ and $\Sigma_i = \Sigma_i(\theta)$, where $\Sigma_i(\theta)$ depends on i only through the dimension of Σ_i. And furthermore, the dimensions of the Y_i's could be different, generally due to missing data. In general, the Newton-Raphson and Method of

Scoring algorithms can be implemented to maximize the relatively complicated log-likelihood function; however, these algorithms are very computationally intensive. In addition, the resulting estimates, $\hat{\Sigma}_i$, in each iteration are not guaranteed to be positive definite. If this happens, the algorithm will break down. In some cases a reparameterization, such as using the Cholesky decomposition, of the matrices is sufficient. However, this cannot be achieved in all circumstances. Step halving is then an alternative algorithmic method to ensure the positive definiteness of the covariance matrices and possibly the increase of the log-likelihood function. By cutting the step size in half, consecutively if necessary, one can always find the solution in the current iteration to ensure the positive definiteness of the covariance matrices or the increase of the log-likelihood function or both at the same time given some directional and monotonic conditions on the derivatives of the log-likelihood. For example, when the Newton-Raphson algorithm is implemented, to ensure that the log-likelihood function is increasing in each iteration when step halving is used the Hessian matrix of the log-likelihood function has to be negative definite for all parameter values in the parameter space. The positive definiteness of the new covariance matrices can always be achieved by using sufficiently small step sizes given the old ones are positive definite, since the new estimate is a linear interpolation of the old one and the update. However, the step halving can substantially increase the computational burden. And one often needs to differentiate between a solution on the boundary and a local maximum. Nevertheless, the idea of step halving is crucial in our application of the Method of Scoring algorithm in Chapter 4.

3.2 CLASSIC MIXTURE MODELS

Let $Y_1, \cdots, Y_n$ be n independent observations and $Z_1, \cdots, Z_n$ be n unobserved group indicators associated with the Y_i's. Marginally, $Z_i = (z_{i1}, \cdots, z_{ig})$ for $i = 1, \ldots, n$ are i.i.d. multinomial$(1; \pi_1, \cdots, \pi_g)$, with $0 \leq \pi_k \leq 1$, $k = 1, \ldots, g$, and $\sum_{k=1}^{g} \pi_k = 1$. The conditional density or mass function of Y_i given Z_i is given by

$$f(y_i | \, z_{ik} = 1) = f_k(y_i, \theta_k), \quad i = 1, \ldots, n, \quad k = 1, \ldots, g, \tag{3.5}$$

where the θ_k's are unknown parameters. Usually the distributions $\{f_k(\cdot, \theta_k)\}$ are from the same exponential family parameterized by a vector parameter θ and differ only in the value of the parameter. It follows then the marginal density or mass function of Y_i is

$$f(y_i) = \sum_{k=1}^{g} \pi_k f_k(y_i, \theta_k), \quad i = 1, \ldots, n. \tag{3.6}$$

The problem of assessing the order, g, of the mixtures without prior information is hard, particularly when some of the components are not widely separated (See McLachlan and Peel (2000), Section 6, and Titterington, Smith, and Makov (1985)). Some approaches for determining g that have been applied include assessing the number of modes of a distribution nonparametrically, using information criteria, such as AIC and BIC, and applying a likelihood ratio test.

However, sometimes the number g of groups is known a priori. The parameter estimation in mixture models for fixed g can then be achieved using maximum likelihood via the EM algorithm. In the EM framework, the Y_i's are viewed as incomplete data while $\{Y_i, Z_i\}$, $i = 1, \ldots, n$, are treated as the complete or augmented data. The E-step is then to compute the conditional expectation of the Z_i's given the observed data, i.e. the Y_i's, and the current estimated parameter values. The M-step involves finding the maximum likelihood estimates of the parameters with the Z_i's replaced by the conditional expectations in the E-step. In a more complicated situation where some components of the Y_i's are missing, the E-step then should also compute the conditional expectation of these missing components. The computational details are reviewed in Section 3.3.2.

In frequentist theory, the standard errors of the MLE can be estimated through either the Fisher information matrix or bootstrap. Let $\vartheta = \{\pi_k, \theta_k; k = 1, \ldots, g\}$ be the vector of unknown parameters. It is well known that the asymptotic covariance matrix of the MLE $\hat{\vartheta}$, that is, the inverse of the Fisher information matrix $I(\vartheta)$, can be estimated either by the observed information matrix $I(\hat{\vartheta}; Y)$, which is the Hessian of the negative log-likelihood function evaluated at $\hat{\vartheta}$, or by the plug-in estimator $I(\hat{\vartheta})$. In order to reduce the computational burden, Louis (1982) showed that the observed information matrix could be computed as

$$I(\hat{\vartheta}; Y) = E\{I_c(\vartheta; Y, Z) | Y\}|_{\vartheta=\hat{\vartheta}} - Var\{S_c(Y, Z; \vartheta) | Y\}|_{\vartheta=\hat{\vartheta}}, \tag{3.7}$$

where $I_c(\vartheta; Y, Z)$ and $S_c(Y, Z; \vartheta)$ are the information matrix and the score function based on the complete data, respectively. Moreover, it was shown by Efron and Hinkley (1978) that $I(\hat{\vartheta}; Y)$ was better than $I(\hat{\vartheta})$ in terms of estimating the standard errors of the MLE. According to Brasford, Greenway, McLachlan, and Peel (1997), the bootstrap method is preferred when the sample size is relatively small. By running the EM algorithm B times on the B bootstrapped samples and then combining the estimates of the parameters, the bootstrap is more time-consuming but yields estimates of the standard errors that are more stable than those of information-based (McLachlan and Peel, 2000, Section 2).

3.3 COMPUTATIONAL ISSUES

The main task of computation is to maximize the likelihood function, or equivalently the log-likelihood function, over the parameter space. Some desired properties of the algorithms include fast convergence and stability with respect to the choice of starting point.

3.3.1 Iterative Algorithms

The iterative procedure introduced in Anderson (1973) has been shown to be equivalent to the Method of Scoring algorithm Szatrowski (1983). Nevertheless, we review this algorithm because we use his idea of a one-iterate solution in deriving the asymptotic distributions of the MLE in Chapter 4 and it is useful in showing many nice properties of the Method of Scoring algorithm. Explicitly, the algorithm iterates between

$$\sum_{k=1}^{r} x_j' \hat{\Sigma}^{(t)-1} x_k \hat{\beta}_k^{(t+1)} = x_j' \hat{\Sigma}^{(t)-1} \bar{Y}, \quad j = 1, \ldots, r, \tag{3.8}$$

and

$$\sum_{f=0}^{m} tr \hat{\Sigma}^{(t)-1} G_g \hat{\Sigma}^{(t)-1} G_f \hat{\sigma}_f^{(t+1)} = tr \hat{\Sigma}^{(t)-1} G_g \hat{\Sigma}^{(t)-1} \hat{C}^{(t+1)}, \quad g = 0, 1, \ldots, m, \tag{3.9}$$

from an initial value $\{\sigma_0^{(0)}, \cdots, \sigma_m^{(0)}\}$ of $\{\sigma_0, \cdots, \sigma_m\}$. We iteratively solve (3.8) for the β's with $\{\sigma_0^{(0)}, \cdots, \sigma_m^{(0)}\}$ plugged in the Σ and then solve (3.9) for the σ's with $\{\sigma_0^{(0)}, \cdots, \sigma_m^{(0)}\}$ plugged in the Σ and the new estimates of the β's plugged in C. A starting point of the

β's is not necessary, since we always begin the iteration with (3.8). $\hat{\Sigma}^{(t)}$ is the estimate of the covariance matrix with $\hat{\sigma}_0^{(t)}, \hat{\sigma}_1^{(t)}, \cdots, \hat{\sigma}_m^{(t)}$ plugged in, and $\hat{C}^{(t+1)} = (1/n)\sum_{i=1}^{n}(Y_i - X\hat{\beta}^{(t+1)})(Y_i - X\hat{\beta}^{(t+1)})'$. It is shown that as long as $\hat{\Sigma}^{(t)}$ is nonsingular, the matrices of coefficients in (3.8) and (3.9) are positive definite, i.e., we would have a successive solution. Anderson (1973) showed that in order to obtain unbiased, consistent and asymptotically efficient estimates, only one iteration of (3.8) and (3.9) is necessary if the initial estimates are consistent. The asymptotic covariance matrices are given in Section 3.1.

When there are missing data, non-identical covariates or non-identical covariance matrices, the (observed) likelihood function becomes much more complicated. In these cases, direct numerical optimization of the log-likelihood function is desirable. The first and second order partial derivatives of the log-likelihood function with respect to the unknown parameters can usually be calculated analytically. Then we have the Newton-Raphson and Method of Scoring algorithms to maximize the log-likelihood function sharing a common form:

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + a^{(t)} H^{-1}(\hat{\theta}^{(t)}) S(\hat{\theta}^{(t)}) \tag{3.10}$$

with $a^{(t)}$ being a possible step size in the current iteration, where $S(\hat{\theta}^{(t)})$ is the score function and $H(\hat{\theta}^{(t)})$ is the negative of the Hessian matrix (for Newton-Raphson) or its expectation (for Method of Scoring) both evaluated at the current parameter values. There are variants of the Newton-Raphson algorithm that use numerical approximation to the Hessian matrix to avoid the calculation of the second derivatives of the log-likelihood function (see Berndt et al., 1974). When $\mathrm{E}[\partial^2 \log(L)/\partial\beta_j\partial\sigma_h] = 0$, the Method of Scoring algorithm has a simple form iterating through β and σ separately.

For problems involving missing data, the EM algorithm is usually preferred. For example, when the data are assumed to be normal, the conditional expectations of the missing values are just the linear regression predictions based on the observed data and the current parameter values. However, in certain cases, the M-step might still need an iterative algorithm to solve the likelihood equations, which can lead to computationally inefficiency of the EM algorithm.

None of these algorithms is guaranteed to converge for general starting points, patterns of mean and covariance structures and patterns of missing data. And none of them has

been shown to be superior to another. Szatrowski (1983) showed that the Newton-Raphson and EM algorithm are more vulnerable to the choice of starting points than the Method of Scoring algorithm in the case of patterned mean and covariance structures with missing data. When they do converge to a root of the likelihood equation, this root is not always the MLE.

3.3.2 The EM Algorithm for Classic Mixture Models

For incomplete-data problems, the EM algorithm, introduced by Dempster, Laird, and Rubin (1977), is an alternative iterative procedure to find the maximum of a log-likelihood function without computing or approximating the second derivatives. It maximizes the observed log-likelihood function with the help of the augmented log-likelihood function, which in many cases can be written in a simpler form. The EM algorithm has an E-step (conditional expectation) and an M-step (Maximization) in each iteration. The E-step involves evaluating the conditional expectation of the complete data log-likelihood function

$$Q\left(\vartheta|\,\vartheta^{(t)}\right) = E\left[l(\vartheta|\,Y)|\,Y_{obs},\vartheta^{(t)}\right], \tag{3.11}$$

where Y_{obs} is the observed part of the complete data Y and $\vartheta^{(t)}$ is the current estimate of the parameter ϑ. The M-step maximizes $Q(\vartheta|\vartheta^{(t)})$ with respect to ϑ to obtain $\vartheta^{(t+1)}$. The iteration can be stopped when either the parameter estimates or the observed log-likelihood function evaluated at the parameter estimates does not change more than a specified amount. Key results of Dempster, Laird, and Rubin (1977) state that the EM algorithm increases the observed log-likelihood function at each iteration, i.e., $l(\vartheta^{(t+1)}|\,Y_{obs}) \geq l(\vartheta^{(t)}|\,Y_{obs})$, and if $\vartheta_{(t)}$ converges, it converges to a stationary point. Multiple starting points might be needed in order to obtain the global maximum. However, the speed of convergence of the EM algorithm has been shown to be linear and comparatively slow, especially when the fraction of missing information is large. In some cases, there is no analytic solution in the M-step, and then the simplicity of the EM algorithm breaks down. But there are some extensions of the EM algorithm which can help to avoid these problems (McLachlan and Krishnan, 1977). Finally, since the EM algorithm does not calculate the information matrix in each iteration,

it does not share with the Newton-Raphson and Method of Scoring algorithm the property of yielding the asymptotic covariance matrix of the MLE at convergence.

In the classic mixture models, let $\mathbf{y} = \{y_1, \cdots, y_n\}$ and $\mathbf{z} = \{z_1, \cdots, z_n\}$ be a realization of $\{Y_1, \cdots, Y_n\}$ and $\{Z_1, \cdots, Z_n\}$, then we have the augmented likelihood function as

$$L(\vartheta|\,\mathbf{y}, \mathbf{z}) = \prod_{i=1}^{n}\prod_{k=1}^{g}\{\pi_k f_k(y_i, \theta_k)\}^{z_{ik}}. \tag{3.12}$$

As a result, (3.11) can be rewritten as

$$Q\left(\vartheta|\,\vartheta^{(t)}\right) = E\left[l(\vartheta|\,\mathbf{y}, \mathbf{z})|\,\mathbf{y}, \vartheta^{(t)}\right] = \sum_{i=1}^{n}\sum_{k=1}^{g}\tau_{ik}^{(t)}\{\log\pi_k + \log f_k(y_i, \theta_k)\} \tag{3.13}$$

where $\tau_{ik}^{(t)} = E[z_{ik}|\,\mathbf{y}, \vartheta^{(t)}] = \pi_k^{(t)} f_k(y_i, \theta_k^{(t)}) / \sum_{j=1}^{g}\pi_j^{(t)} f_j(y_i, \theta_j^{(t)})$. Then the new estimates of the π_k's in the following M-step can be obtained as

$$\pi_k^{(t+1)} = (1/n)\sum_{i=1}^{n}\tau_{ik}^{(t)}, \quad k = 1, \ldots, g, \tag{3.14}$$

while the new estimates of the θ_k's can be obtained by solving the equations

$$\sum_{i=1}^{n}\tau_{ik}^{(t)}\partial\log f_k(y_i, \theta_k)/\partial\theta_k = 0, \quad k = 1, \ldots, g. \tag{3.15}$$

We do not directly apply the EM algorithm to the mixture problem we have. We consider a variant called the EM gradient algorithm (Lange, 1995), since there is no explicit solutions in the M-step in our problem. The details are in Chapter 5. In the case of missing data, we propose to impute the incomplete data, apply the mixture models to the imputed data to identify the possible clusters of the subjects with schizophrenia, and then combine the clustering results from multiple imputations. This is one rather straightforward way of dealing with the missing data. And in the current stage of our research and with the amount of missing data we have, this is one feasible method.

4.0 STRUCTURED MODELING WITH ONE POPULATION

As an initial step in our goal for clustering subjects with schizophrenia based on post-mortem tissue data, we develop new multivariate normal models with patterned mean and covariance structures in this chapter. We provide several model fitting algorithms, including the Method of Scoring and the Newton-Raphson algorithms, to find the parameter estimates for these new structured models. These models generalize standard models considered by Anderson (1973), Szatrowski (1983) and Jennrich and Schluchter (1986). A one-iteration estimator using a Simplified Method of Scoring algorithm starting from a consistent starting point is used to derive the asymptotic distributions of the estimators. The model fitting algorithms, as well as the asymptotic distributions, are examined using simulated data, and are applied to data from post-mortem tissue studies in schizophrenia.

4.1 THE MODEL WITH STRUCTURED MEANS AND COVARIANCES

4.1.1 Model Specification

Let $Y_i = (Y_{i1}, \cdots, Y_{ip})'$, $i = 1, \ldots, n$, be n independent p-dimensional observations and let X_i, $i = 1, \ldots, n$, be matrices of covariates associated with each observation. Usually each X_i takes form

$$X_i = \begin{bmatrix} a'_{i1} & v'_{i1} & 0 & \cdots & 0 \\ a'_{i2} & 0 & v'_{i2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a'_{ip} & 0 & 0 & \cdots & v'_{ip} \end{bmatrix}, \tag{4.1}$$

where $\{a_{ij}\}_{j=1}^{p}$ are vectors of length $r \geq 0$ and $\{v_{ij}\}_{j=1}^{p}$ are vectors of length $s \geq 0$ such that $r + s > 0$. Here, $\{a_{ij}\}_{j=1}^{p}$ share the same effect over the p measurements, while $\{v_{ij}\}_{j=1}^{p}$ do not. In our neurobiological context, a representative a_{ij} is the constant 1, which represents the diagnostic effect; whereas the subject's age would be representative of the v_{ij}'s. To develop our models, we assume that $\{X_i\}_{i=1}^{n}$ are a random sample from a distribution with finite second moments; the actual form of this distribution is not of main interest.

Conditional on X_i, Y_i is assumed to have a multivariate normal distribution for $i = 1, \ldots, n$. However, in our notation we suppress the conditioning on X_i and only focus on this when necessary for the asymptotics. First, assume the mean vectors be

$$E[Y_i] = \mu_i = X_i\beta, \quad i = 1, \ldots, n, \tag{4.2}$$

where β is an unknown vector of dimension $r + sp$. A special case is $X_i \equiv X$ for $1 \leq i \leq n$. Then it is necessary that the columns of X to be linearly independent, so that all individual parameters in β are estimable (Anderson, 1973). In general, the columns of each X_i don't have to be linearly independent as long as $(X_1', \cdots, X_n')$ is of full rank.

Second, define

$$I_i = \begin{bmatrix} I_i^{11} & I_i^{12} & \cdots & I_i^{1p} \\ I_i^{21} & I_i^{22} & \cdots & I_i^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ I_i^{p1} & I_i^{p2} & \cdots & I_i^{pp} \end{bmatrix}, \quad i = 1, \ldots, n, \tag{4.3}$$

to be n known $p \times p$ symmetric matrices with $I_i^{kk} = 1$ for $1 \leq k \leq p$, and $I_i^{kl} = I_i^{lk} = 0$ and $I_i^{kl} = I_i^{lk} = 1$ representing two possible choices of the covariance between the kth and the lth measurements for $1 \leq k < l \leq p$ for the ith vector. Then the covariance matrices of the Y_i's are defined as

$$\begin{aligned} \Sigma_i = E[(Y_i - \mu_i)(Y_i - \mu_i)'] &= \begin{bmatrix} \sigma_{11}^s + \sigma_{11}^c & \sigma_{12}^s + \sigma_{12}^c I_i^{12} & \cdots & \sigma_{1p}^s + \sigma_{1p}^c I_i^{1p} \\ \sigma_{21}^s + \sigma_{21}^c I_i^{21} & \sigma_{22}^s + \sigma_{22}^c & \cdots & \sigma_{2p}^s + \sigma_{2p}^c I_i^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}^s + \sigma_{p1}^c I_i^{p1} & \sigma_{p2}^s + \sigma_{p2}^c I_i^{p2} & \cdots & \sigma_{pp}^s + \sigma_{pp}^c \end{bmatrix} \\ &= \Sigma_S + I_i \cdot \Sigma_C, \quad i = 1, \ldots, n, \end{aligned} \tag{4.4}$$

where the symbol "$\cdot$" represents a pointwise product of two matrices with compatible dimensions, and

$$\Sigma_S = \begin{bmatrix} \sigma^s_{11} & \sigma^s_{12} & \dots & \sigma^s_{1p} \\ \sigma^s_{21} & \sigma^s_{22} & \dots & \sigma^s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^s_{p1} & \sigma^s_{p2} & \dots & \sigma^s_{pp} \end{bmatrix} \quad \text{and} \quad \Sigma_C = \begin{bmatrix} \sigma^c_{11} & \sigma^c_{12} & \dots & \sigma^c_{1p} \\ \sigma^c_{21} & \sigma^c_{22} & \dots & \sigma^c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^c_{p1} & \sigma^c_{p2} & \dots & \sigma^c_{pp} \end{bmatrix},$$

are the two unknown covariance matrices, e.g., in our setting, one for the subjects with schizophrenia and one for the control subjects. We use this parameterization to represent the covariance structure arising from the differing controls as discussed in Section 2.2. Since $I_i^{kl} = I_i^{lk'} = 1$ implies $I_i^{kk'} = 1$ and $I_i^{kl} = 1 - I_i^{lk'} = 1$ implies $I_i^{kk'} = 0$, the total possible choices of I_i for $1 \le i \le n$ is $2^p - p$. In the Table 2.1 prototype, we have

$$I_1 = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right], I_2 = \left[\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right], I_3 = \left[\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{smallmatrix}\right], I_4 = \left[\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{smallmatrix}\right] \quad \text{and} \quad I_5 = \left[\begin{smallmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{smallmatrix}\right]$$

for the five cases, respectively. Clearly, these indicator matrices, $\{I_i\}_{i=1}^n$, are fixed by the experimental design. Let G_{kl} be $p \times p$ matrix with "0" entries except a "1" at both the (k, l) and (l, k) entries for $k = 1, \dots, p$, $l = 1, \dots, p$ and $k \le l$. Then we can rewrite Σ_i as

$$\Sigma_i = \sum_{k=1}^{p} (\sigma^s_{kk} + \sigma^c_{kk}) G_{kk} + \sum_{1 \le k < l \le p} (\sigma^s_{kl} + \sigma^c_{kl} I_i^{kl}) G_{kl}. \tag{4.5}$$

This representation is used in the estimation procedures introduced in Section 4.2.

In addition, we require that the parameters governing the marginal distribution of $\{X_i\}_{i=1}^n$ are functionally independent of β, Σ_C and Σ_S. As a result, we can focus on the conditional distribution of $\{Y_i\}_{i=1}^n$ given $\{X_i\}_{i=1}^n$ in estimating β, Σ_C and Σ_S using maximum likelihood.

4.1.2 Parameter Identifiability

For the parameterization in Section 4.1.1, the parameter space is

$$\Theta = \{\beta \in R^{r+sp},\ \Sigma_S > 0 \text{ and } \Sigma_C > 0\}, \tag{4.6}$$

where $\Sigma > 0$ indicates matrix positive definiteness. However, (4.6) is not identifiable. The parameterization of $\Sigma_i = \Sigma_S + I_i \cdot \Sigma_C$ for $i = 1, \ldots, n$ is intended to represent the covariance structures of the pairwise differences to reflect the differing controls for the subjects with schizophrenia; and Σ_S and Σ_C can be viewed as the underlying covariance matrices for the observations on the subjects with schizophrenia and the controls, respectively. Given the indicators $\{I_i\}_{i=1}^n$, $\{\Sigma_i\}_{i=1}^n$, as a function of Σ_S and Σ_C, is guaranteed to be positive definite, as long as the arguments are. However, this function is not invertible in the sense that knowing $\{\Sigma_i\}_{i=1}^n$ is not sufficient to reconstruct Σ_S and Σ_C. We can only estimate the sum $\sigma_{kk}^s + \sigma_{kk}^c$ for $1 \le k \le p$, but not the individual items. So there usually exist multiple Σ_S and Σ_C corresponding to each $\{\Sigma_i\}_{i=1}^n$. A trivial example is $\left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right] + \left[\begin{smallmatrix} 3 & 1 \\ 1 & 2 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 3 \end{smallmatrix}\right] + \left[\begin{smallmatrix} 4 & 1 \\ 1 & 1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 5 & 2 \\ 2 & 4 \end{smallmatrix}\right]$. Moreover, if we just require the Σ_i, $1 \le i \le n$, to be any positive definite matrices, then the inversion solution may not even exist. For example, let $\Sigma_1 = \left[\begin{smallmatrix} 2 & -2 \\ -2 & 4 \end{smallmatrix}\right]$ and $\Sigma_2 = \left[\begin{smallmatrix} 2 & 2 \\ 2 & 4 \end{smallmatrix}\right]$, then either $\sigma_{12}^c = -4$ or $\sigma_{12}^c = 4$ both of which are impossible for $\sigma_{11}^c < 2$ and $\sigma_{22}^c < 4$.

As another technical issue, the identifiability of σ_{kl}^s and σ_{kl}^c for $k \neq l$ is design dependent. Explicitly speaking, when $\sum_{i=1}^n I_i^{kl} = 0$ or $\sum_{i=1}^n I_i^{kl} = n$ for some $k \neq l$, some covariance parameters will not be identifiable. However, it can be solved analogously when the parameter space has been reduced accordingly. For instance, if $\sum_{i=1}^n I_i^{kl} = 0$, then σ_{kl}^c should be removed; and if $\sum_{i=1}^n I_i^{kl} = n$, we treat the sum $\sigma_{kl}^s + \sigma_{kl}^c$ as a free parameter and discard its individual items. In either case, the dimension of the parameter space is reduced by one and the resulting parameters are estimable. Here, for simplicity we assume $1 < \sum_{i=1}^n I_i^{kl} < n-1$ for all $1 \le k < l \le p$.

As a result, it is impractical to use the parameterization of (4.6). Given $1 < \sum_{i=1}^n I_i^{kl} < n-1$ for all $1 \le k < l \le p$, we redefine $\sigma_{kk} = \sigma_{kk}^s + \sigma_{kk}^c$, and also for notational ease let $\sigma_{kl} = \sigma_{kl}^s$. Then we define

$$\Theta' = \{\beta \in R^{r+sp} \text{ and } \sigma \in R^{p^2} \text{ s.t. } \Sigma_{[1]} > 0, \cdots, \Sigma_{[2^p-p]} > 0\} \tag{4.7}$$

to be the new parameter space, where $\{\Sigma_{[1]}, \cdots, \Sigma_{[2^p-p]}\}$ are the $2^p - p$ possible covariance matrices and $\sigma = (\sigma_{11}, \cdots, \sigma_{pp}, \sigma_{12}, \cdots, \sigma_{(p-1)p}, \sigma^c_{12}, \cdots, \sigma^c_{(p-1)p})'$. The parameters in (4.7) are now identifiable. Since we only require the Σ_i, $1 \leq i \leq n$, to be positive definite for parameters from (4.7), (4.7) is wider than (4.6) in the sense that for some $\{\Sigma_i\}_{i=1}^n$ whose parameters are in (4.7), there do not exist $\Sigma_S > 0$ and $\Sigma_C > 0$ such that (4.4) holds. Some *post hoc* methods might need to be implemented to ensure the MLE falling in (4.6) if one insists. However, maximizing the log-likelihood function over this expanded space won't cause any algorithmic problem.

In general, not all the p^2 parameters in σ are estimable, and not all the $2^p - p$ covariance matrices are necessary. In this case the parameter space can be represented as

$$\Theta'' = \{\beta \in R^{r+sp} \text{ and } \sigma \in R^k \text{ s.t. } \Sigma_{[1]} > 0, \cdots, \Sigma_{[q]} > 0\} \tag{4.8}$$

where k is the number of parameters in σ which are estimable, and $q \leq 2^p - p$ is the number of necessary covariance matrices.

4.1.3 An Illustrative Example

The mean and covariance structures introduced in Section 4.1.1 occur in the combined data from multiple post-mortem tissue studies in the Center. However, for the actual Center's data, the combined data are incomplete due to the availability of tissue samples and other experimental constraints. Techniques for handling the actual degree of missingness will ultimately be developed for our clustering approaches. As a result, we provide here a simple example to demonstrate the necessity of the mean and covariance structures in the integrative analysis and also provide a better sense of what these data look like.

The data shown in this example were collected and initially analyzed by Hashimoto et al. (2003, 2005). A total of 26 pairs of subjects with schizophrenia and controls were used. The original purpose was to determine the causality of a neurotrophic factor BDNF and its receptor TrkB on the altered expression of GABA-related genes in schizophrenia, since the down-regulation of GABA-related genes, such as GAD_{67} and PV, seems to be related to cognitive deficits in subjects with schizophrenia. In these two studies, tissue samples of

Table 4.1: Characteristics of Subjects and Data: Hashimoto et al. (2003, 2005)

	Gender		Age		PMI[1]		Brain pH		Storage[2]		Pairwise differences			
Pair	C	S	C	S	C	S	C	S	C	S	BDNF	TrkB	GAD_{67}	Case[3]
1	M	M	41	40	22.1	29.1	6.72	6.82	78	88	-0.22	-7.78	-19.43	3
2	F	F	46	37	15	14.5	6.72	6.68	82	87	-1.69	-20.04	-19.37	4
3	M	M	20	27	14	16.5	6.86	6.95	89	85	-10.78	-30.37	-72.51	4
4	M	M	65	63	21.2	18.3	6.95	6.8	72	82	-0.6	-26.35	-28.19	4
5	F	F	37	38	23.5	17.8	6.74	7.02	86	79	-10.81	-43.95	-48.11	2
6	M	M	47	48	6.6	8.3	6.99	6.07	83	134	-2.53	-5.66	-9.53	4
7	F	F	54	46	17.8	10.1	6.47	7.02	71	77	-10.77	-52.22	-91.2	2
8	M	M	61	49	16.4	23.5	6.63	7.32	85	73	-2.15	-4.9	-33.3	2
9	M	M	56	58	14.5	18.9	6.57	6.78	65	73	1.43	-1.62	-21.01	4
10	M	M	51	49	11.6	5.2	7.15	6.86	65	71	-2.39	-44.31	-55.41	3
11	M	M	57	59	24	28.1	6.94	6.92	44	68	-8.77	-1.78	-5.77	3
12	M	M	28	27	25.3	19.2	7.04	6.67	41	48	-5.97	-11.09	-11.7	3
13	M	M	19	25	7	5	7.15	6.8	48	39	-0.07	-17.94	-1.66	2
14	M	M	28	33	16.5	10.8	7.14	6.72	31	30	-6.7	-43.6	-49.28	2
15	F	F	55	48	11.3	3.7	6.81	6.69	91	100	-1.83	-7.77	-0.26	3
16	M	M	42	50	26.1	40.5	6.95	7.1	73	98	-15.24	-90.08	-26.69	4
17	M	M	82	83	22.5	16	6.24	7.33	20	84	-4.21	-39.29	-26.4	2
28	F	F	52	47	22.6	20.1	7.02	7.26	76	80	-4.42	-18.29	7.76	2
29	M	M	38	40	20.7	17.3	6.73	6.7	75	75	-0.59	1.18	11.08	3
20	M	M	52	45	16.2	9.1	7.04	6.71	82	70	3.47	57.61	14.56	3
21	M	M	54	52	8	8	6.77	6.69	45	60	-2.9	-74.28	-52.74	3
22	F	F	65	63	21.5	29	6.78	6.42	20	56	-12.88	-45.55	-9.47	4
23	M	M	39	33	24.2	29	7.15	6.19	40	37	-19.34	-79.34	-36.47	4
24	F	F	67	71	24	23.8	7.06	6.82	53	33	-3.76	-41.24	-13.18	4
25	M	M	48	47	16.6	15.7	6.74	6.22	44	30	-5.89	-53.11	-31.8	4
26	M	M	40	44	15.8	8.3	6.88	5.93	68	29	-19.09	-78.84	-55.4	3

1: post-mortem interval in hours; 2: Storage time (month) at -80°C; 3: refer to Table 1

the prefrontal cortex (PFC) from these 26 pairs of subjects were obtained, and the mRNA expression levels of BDNF, TrkB and GAD_{67} were simultaneously measured. In the actual studies, the subjects with schizophrenia were matched to the same control subjects for all 3 measurements. Because we want to illustrate a general setting where different controls exist, we randomly changed these matches so that now a subject with schizophrenia might be matched to 2 or 3 different controls for the 3 measurements. Table 4.1 shows some characteristics and the data (pairwise differences) from the subjects used in Hashimoto et al. (2003, 2005). As can be seen, gender was matched perfectly, while age and PMI were matched as close as possible. Brain PH and tissue storage time were not matched. The randomly changed matches are also shown in Table 4.1 using the five possible covariance structure cases listed in Table 2.1. In this example, we only used three out of the five possible cases since the sample size is small.

4.2 MAXIMUM LIKELIHOOD ESTIMATION AND DERIVATIVES

In this section, the first and second derivatives, as well as the expectation of the second derivatives, of the log-likelihood function are given. For completeness, the derivatives used in the restricted maximum likelihood estimation (REML) are also shown but not implemented in the algorithms because of their complexity in computation. Iterative algorithms are defined to find both the MLE and the one-iteration estimators. We are mainly interested in the Method of Scoring and the Newton-Raphson algorithms. The former is quite straight forward given the derivatives and is more computationally intensive than the latter. The latter, however, shares the same simple form as the iterative algorithm derived directly from the likelihood equations. Asymptotic distributions of the estimators are then derived. Only the Method of Scoring algorithm is implemented in the simulations.

4.2.1 Likelihood Function and Derivatives

With the assumptions in Section 4.1, the conditional likelihood function for the n realizations $\mathbf{y} = \{y_1, \cdots, y_n\}$ given $\{X_i\}_{i=1}^n$ is of the form

$$L(\beta, \sigma | \mathbf{y}) = (2\pi)^{-\frac{np}{2}} \prod_{i=1}^{n} |\Sigma_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2} tr \Sigma_i^{-1} C_i\} \tag{4.9}$$

with $C_i = (y_i - X_i\beta)(y_i - X_i\beta)'$ being the usual sample cross product matrix for y_i, $i = 1, \ldots, n$, where $\beta = (\beta_1, \cdots, \beta_{(r+sp)})'$ is the vector of unknown parameters in the mean structure, and $\sigma = (\sigma_{11}, \cdots, \sigma_{pp}, \sigma_{12}, \cdots, \sigma_{(p-1)p}, \sigma_{12}^c, \cdots, \sigma_{(p-1)p}^c)'$ is the vector of unknown variance-covariance parameters that are involved in each Σ_i. For convenience, sometimes we use $\theta' = (\beta', \sigma')$ in the following discussion. We need to maximize (4.9) or its logarithm with respect to β and σ.

Using standard well-known matrix derivative results, we find the first partial derivatives of $l(\beta, \sigma | \mathbf{y})$, the logarithm of (4.9), as

$$\partial l / \partial \beta = \sum_{i=1}^{n} X_i' \Sigma_i^{-1} (y_i - X_i \beta)\,, \tag{4.10a}$$

$$\partial l / \partial \sigma_{kl} = (1/2) \sum_{i=1}^{n} tr \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} (C_i - \Sigma_i),\ 1 \le k \le l \le p\,, \tag{4.10b}$$

$$\partial l / \partial \sigma_{kl}^c = (1/2) \sum_{i=1}^{n} tr \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} (C_i - \Sigma_i) I_i^{kl},\ 1 \le k < l \le p\,, \tag{4.10c}$$

In general, for the likelihood equations, $\partial l / \partial \theta = 0$, the solutions for the β and the σ depend on each other and have no closed form, so that iterative algorithms are required. Continue

to take partial derivatives of (4.10) to yield the second partial derivatives given by

$$-\partial^2 l/\partial\beta^2 = \sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i\,, \tag{4.11a}$$

$$-\partial^2 l/\partial\sigma_{kl}\partial\sigma_{st} = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}\Sigma_i^{-1}(2C_i-\Sigma_i), \quad 1\le k\le l\le p\,, 1\le s\le t\le p\,, \tag{4.11b}$$

$$-\partial^2 l/\partial\sigma_{kl}^c\partial\sigma_{st}^c = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}\Sigma_i^{-1}(2C_i-\Sigma_i)I_i^{kl}I_i^{st}, \quad 1\le k<l\le p\,, 1\le s<t\le p\,, \tag{4.11c}$$

$$-\partial^2 l/\partial\beta\partial\sigma_{kl} = \sum_{i=1}^{n} X'\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}(y_i-X_i\beta),\ 1\le k\le l\le p\,, \tag{4.11d}$$

$$-\partial^2 l/\partial\beta\partial\sigma_{kl}^c = \sum_{i=1}^{n} X'\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}(y_i-X_i\beta)I_i^{kl},\ 1\le k<l\le p\,, \tag{4.11e}$$

$$-\partial^2 l/\partial\sigma_{kl}\partial\sigma_{st}^c = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}\Sigma_i^{-1}(2C_i-\Sigma_i)I_i^{st}, \quad 1\le k\le l\le p\,, 1\le s<t\le p\,. \tag{4.11f}$$

Then, taking the expected values of the second partial derivatives after observing that $E[\mathbf{y}_i] = X_i\beta$ and $E[C_i] = \Sigma_i$, $i = 1,\ldots,n$, we have

$$-E[\partial^2 l/\partial\beta^2] = \sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i\,, \tag{4.12a}$$

$$-E[\partial^2 l/\partial\sigma_{kl}\partial\sigma_{st}] = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st},\ 1\le k\le l\le p\,, 1\le s\le t\le p\,, \tag{4.12b}$$

$$-E[\partial^2 l/\partial\sigma_{kl}^c\partial\sigma_{st}^c] = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}I_i^{kl}I_i^{st}, \quad 1\le k<l\le p\,, 1\le s<t\le p\,, \tag{4.12c}$$

$$-E[\partial^2 l/\partial\beta\partial\sigma_{kl}] = 0,\ 1\le k\le l\le p\,, \tag{4.12d}$$

$$-E[\partial^2 l/\partial\beta\partial\sigma_{kl}^c] = 0,\ 1\le k<l\le p\,, \tag{4.12e}$$

$$-E[\partial^2 l/\partial\sigma_{kl}\partial\sigma_{st}^c] = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}I_i^{st},\ 1\le k\le l\le p\,, 1\le s<t\le p\,. \tag{4.12f}$$

4.2.2 Restricted Maximum Likelihood (REML)

It is well know that the MLE underestimates variance parameters. An alternative to avoid the biases is to use the restricted maximum likelihood estimators which are obtained by maximizing the residual (log-)likelihood function. By linear model theory the residual log-likelihood function for our problem is, apart from an additive constant,

$$l_R(\beta, \sigma | \, \mathbf{y}) = -\frac{1}{2}[\log |\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i| + \sum_{i=1}^{n} \log |\Sigma_i| + \sum_{i=1}^{n} tr \Sigma_i^{-1} C_i]. \tag{4.13}$$

The first and second derivatives involving the β are the same as those in Section 4.2.1, while those with respect to the σ can be obtained by observing that

$$\partial \log |\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i| / \partial \sigma_{kl} = -\sum_{i=1}^{n} tr[H^{-1} X_i' \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i], \tag{4.14}$$

$$\partial \log |\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i| / \partial \sigma_{kl}^c = -\sum_{i=1}^{n} tr[H^{-1} X_i' \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i] I_i^{kl}, \tag{4.15}$$

and

$$\begin{aligned} \frac{\partial^2 \log |\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i|}{\partial \sigma_{kl} \partial \sigma_{st}} &= -tr[H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{st} \Sigma_i^{-1} X_i) H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i)] \\ &\quad + 2tr[H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{st} \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i)], \end{aligned} \tag{4.16}$$

$$\begin{aligned} \frac{\partial^2 \log |\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i|}{\partial \sigma_{kl}^c \partial \sigma_{st}^c} &= -tr[H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{st} \Sigma_i^{-1} X_i I_i^{st}) H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i I_i^{kl})] \\ &\quad + 2tr[H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{st} \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i I_i^{kl} I_i^{st})], \end{aligned} \tag{4.17}$$

$$\begin{aligned} \frac{\partial^2 \log |\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i|}{\partial \sigma_{kl} \partial \sigma_{st}^c} &= -tr[H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{st} \Sigma_i^{-1} X_i I_i^{st}) H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i)] \\ &\quad + 2tr[H^{-1} \sum_{i=1}^{n} (X_i' \Sigma_i^{-1} G_{st} \Sigma_i^{-1} G_{kl} \Sigma_i^{-1} X_i I_i^{st})], \end{aligned} \tag{4.18}$$

where $H = \sum_{j=1}^{n} X_j' \Sigma_j^{-1} X_j$ and k, l, s and t are in the same range as in Section 4.2.1. The preferable property of REML against MLE is that it automatically considers the loss of degrees of freedom due to the estimation of the β. It can be seen that the residual likelihood function does not depend on the β. However, REML is more computationally intensive than the usual maximum likelihood method. As a result, we only implement the MLE in the following sections.

4.2.3 Model Fitting Algorithms

Anderson (1973) proposed an iterative algorithm to solve the likelihood equations for his model, which was later shown to be equivalent to the Method of Scoring algorithm by Szatrowski (1983). A similar equivalence between directly solving the likelihood equations and the Method of Scoring algorithm can also be shown for our problem. As a result, we focus our discussion on the Newton-Raphson and the Method of Scoring algorithms.

The Newton-Raphson and the Method of Scoring algorithms share the same form, namely,

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + a^{(t)} H^{-1}(\hat{\theta}^{(t)}) S(\hat{\theta}^{(t)}) \tag{4.19}$$

with the only difference being the definition of the H matrix. The Newton-Raphson algorithm directly uses the negative Hessian matrix, while the Method of Scoring algorithm uses the corresponding expectation, where both are evaluated at the current parameter estimate $\hat{\theta}^{(t)}$. Here $a^{(t)}$ is a scalar used to adjust the step size in each iteration and $S(\hat{\theta}^{(t)})$ is the score function evaluated at the current estimate $\hat{\theta}^{(t)}$.

Lemma 4.2.1 (Newton-Raphson Approach). *The Newton-Raphson algorithm for finding the MLE of* (4.9) *is given by* (4.19) *with* $a^{(t)} \equiv 1$, $H = -\partial^2 l/\partial\theta^2$ *in* (4.11) *and* $S = \partial l/\partial\theta$ *in* (4.10).

Lemma 4.2.2 (Method of Scoring Approach). *The Method of Scoring algorithm for finding the MLE of* (4.9) *is given by* (4.19) *with* $a^{(t)} \equiv 1$, $H = -E[\partial^2 l/\partial\theta^2]$ *in* (4.12) *and* $S = \partial l/\partial\theta$ *in* (4.10).

Because there is no nice simplification of the equations in (4.11), in using the Newton-Raphson algorithm we must update the β and σ simultaneously. However, because the expected partial derivatives in (4.12d) and (4.12e) are zero, it is easy to show that the Method of Scoring algorithm in Lemma 4.2.2 updates β and σ separately. For this reason we focus much more extensively on the Method of Scoring algorithm. We now show that the Method of Scoring algorithm in Lemma 4.2.2 can be simplified as follows:

Corollary 4.2.3 (Simplified Method of Scoring Approach). *The Method of Scoring*

algorithm for finding the MLE of (4.9) *can be simplified as*

$$\hat{\beta}^{(t+1)} = \left[\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} y_i\right), \tag{4.20}$$

$$\hat{\sigma}^{(t+1)} = \left[\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} & \Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \\ J_i' \Phi(\hat{\Sigma}_i^{(t)})^{-1} & J_i' \Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \end{bmatrix}\right]^{-1} \left(\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} \\ J_i' \Phi(\hat{\Sigma}_i^{(t)})^{-1} \end{bmatrix} \langle \hat{C}_i^{(t+1)} \rangle\right), \tag{4.21}$$

with $\hat{C}_i^{(t+1)} = (y_i - X_i\hat{\beta}^{(t+1)})(y_i - X_i\hat{\beta}^{(t+1)})'$, $i = 1, \ldots, n$, *where the two matrix operations* $\langle\cdot\rangle$ *and* $\Phi(\cdot)$ *are defined in Definition A.0.1 and Definition A.0.2 in the Appendix. And* J_i, $1 \le i \le n$, *are* $p(p+1)/2 \times p(p-1)/2$ *matrices defined as*

$$J_i = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \\ I_i^{12} & 0 & \cdots & 0 \\ 0 & I_i^{13} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & I_i^{(p-1)p} \end{bmatrix},$$

where the top is a $p \times p(p-1)/2$ *zero matrix.*

Proof. To show the simplification resulting in (4.20), we have

$$\begin{aligned} \hat{\beta}^{(t+1)} &= \hat{\beta}^{(t)} + \left[\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} (y_i - X_i\hat{\beta}^{(t)})\right) \\ &= \hat{\beta}^{(t)} + \left[\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} y_i\right) - \hat{\beta}^{(t)} \\ &= \left[\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} X_i' \hat{\Sigma}_i^{(t)-1} y_i\right). \end{aligned}$$

To show the simplification resulting in (4.21), we need to notice that

$$\begin{aligned} tr\Sigma_i^{-1} G_{kl} \Sigma_i^{-1} G_{st} &= 2 \langle G_{kl} \rangle' \Phi^{-1}(\Sigma_i) \langle G_{st} \rangle, \\ tr\Sigma_i^{-1} G_{kl} \Sigma_i^{-1} C_i &= 2 \langle G_{kl} \rangle' \Phi^{-1}(\Sigma_i) \langle C_i \rangle, \quad \text{and} \\ tr\Sigma_i^{-1} G_{kl} \Sigma_i^{-1} \Sigma_i &= 2 \sum_{1 \le s \le t \le p} tr\Sigma_i^{-1} G_{kl} \Sigma_i^{-1} G_{st} \sigma_{st}, \end{aligned}$$

with Definition A.0.1, A.0.2 and Theorem A.0.3 in the Appendix. By remembering the specialness of the G_{kl}'s, and that stacking row vectors $\{\langle G_{kl}\rangle'\}$ according to the sequence of $(11, \cdots, pp, 12, 13, \cdots, (p-1)p)$ for kl, we find that it just results in an identity matrix of order $p(p+1)/2$. Then we have

$$\partial l/\partial\sigma = \sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} \langle C_i\rangle \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} \langle C_i\rangle \end{bmatrix} - \sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} & \Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} & J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \end{bmatrix} \sigma, \quad \text{and}$$

$$-E[\partial^2 l/\partial\sigma^2] = \sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} & \Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} & J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \end{bmatrix}.$$

As a result, we have

$$\begin{aligned} \hat{\sigma}^{(t+1)} &= \hat{\sigma}^{(t)} + \left[\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} & \Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} & J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \end{bmatrix}\right]^{-1} \left(\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} \langle C_i\rangle \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} \langle C_i\rangle \end{bmatrix}\right) - \hat{\sigma}^{(t)} \\ &= \left[\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} & \Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} & J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} J_i \end{bmatrix}\right]^{-1} \left(\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1} \langle C_i\rangle \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} \langle C_i\rangle \end{bmatrix}\right). \end{aligned}$$

□

Equation (4.20) uses the current estimate $\hat{\sigma}^{(t)}$ of σ to yield an updated estimate $\hat{\beta}^{(t+1)}$ of β. We then update the estimate of σ using (4.21) to get $\hat{\sigma}^{(t+1)}$. It can be shown that given $\hat{\Sigma}_i^{(t)} > 0$ for $1 \le i \le n$ and $(X_1', \cdots, X_n')$ is of full rank, the coefficient matrices (the first matrices on the right hand sides) in (4.20) and (4.21) are positive definite; hence their inverses exist and (4.20) and (4.21) provide the updates.

Result 4.2.4. *The coefficient matrices*

$$\sum_{i=1}^{n} X_i'\Sigma_i^{-1}X_i \quad \text{and} \quad \sum_{i=1}^{n} \begin{bmatrix} \Phi(\Sigma_i)^{-1} & \Phi(\Sigma_i)^{-1} J_i \\ J_i'\Phi(\Sigma_i)^{-1} & J_i'\Phi(\Sigma_i)^{-1} J_i \end{bmatrix}$$

in (4.20) *and* (4.21) *are positive definite when* Σ_i *is positive definite for* $i = 1, \ldots, n$.

Proof. For the special case where $X_i \equiv X$ for $1 \leq i \leq n$, we have

$$z' \left[\sum_{i=1}^{n} X' \Sigma_i^{-1} X \right] z = \sum_{i=1}^{n} (Xz)' \Sigma_i^{-1} (Xz) > 0 \quad \text{for} \quad z \neq 0,$$

since by assumption the columns of X are linearly independent. For the general case, we have

$$\sum_{i=1}^{n} X_i' \Sigma_i^{-1} X_i = \mathcal{X}' \Omega^{-1} \mathcal{X},$$

where

$$\mathcal{X}' = (X_1', \cdots, X_n') \quad \text{and} \quad \Omega = diag(\Sigma_1, \cdots, \Sigma_n).$$

Now, given $\Sigma_i > 0$, $i = 1, \ldots, n$, and X is of full rank, $\mathcal{X}' \Omega^{-1} \mathcal{X}$ is easily shown to be positive definite by using the similar arguments as above.

To show the second matrix is positive definite, let $x = (x_1', x_2')' \neq 0$, then we have

$$(x_1', x_2') \begin{bmatrix} \Phi(\Sigma_i)^{-1} & \Phi(\Sigma_i)^{-1} J_i \\ J_i' \Phi(\Sigma_i)^{-1} & J_i' \Phi(\Sigma_i)^{-1} J_i \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1' + x_2' J_i') \Phi(\Sigma_i)^{-1} (x_1 + J_i x_2).$$

And since $\Phi(\Sigma_i)$ is always positive definite if Σ_i is, the second matrix is positive definite and singular if and only if $(x_1 + J_i x_2) = 0$ for all $i = 1, \ldots, n$, which means $x = 0$ by remembering our assumption that $1 < \sum_{i=1}^{n} I_i^{kl} < n - 1$ for $i = 1, \ldots, n$. So it is a contradiction with the assumption of $x \neq 0$. □

Although the MLE can be found by iterating (4.20) and (4.21) until convergence, we are interested in a one-iteration estimator of the Simplified Method of Scoring algorithm starting from a consistent initial value. The idea of the one-iteration estimator has been introduced in Anderson (1973). It is simpler to obtain than the MLE, and it is consistent, asymptotically normal and efficient. So it is asymptotically equivalent to the MLE.

4.2.4 Computational Details

We can always start the Simplified Method of Scoring algorithm with $\sigma^{(0)}$, an initial value of σ, and use (4.20) to find the first update $\beta^{(1)}$ of β so that we can avoid a starting point for the β. And the starting point for the covariance matrices can be chosen arbitrarily. For example, we use the identity matrix as a starting point in our simulations in Chapter 4.4. The traditional Method of Scoring algorithm uses $\hat{C}_i^{(t)}$ in (4.21). Here we use $\hat{C}_i^{(t+1)}$ instead. There is no big advantage of doing this, but at least it won't be worse than just using $\hat{C}_i^{(t)}$ in terms of convergence speed. Although, we've shown that given $\hat{\Sigma}_i^{(t)} > 0$ for $1 \leq i \leq n$ and $(X_1', \cdots, X_n')$ is of full rank, the coefficient matrices in (4.20) and (4.21) are positive definite, the positive definiteness of the updated covariance matrices, $\{\hat{\Sigma}_i^{(t+1)}\}_{i=1}^n$, is not guaranteed. In the simplest case when the Σ_i, $1 \leq i \leq n$, are all identical, a reparameterization using the Cholesky decomposition is a direct and relatively easy way to ensure the positive definiteness of its consecutive updates. However, in our problem this approach does not work since our covariance matrices are not identical. The approach we use is to monitor each iteration and ensure the positive definiteness of the covariance matrices by using step-halving. It is applicable since in using step-halving, the new estimate of the parameter is a linear interpolation of the old one and the update given in Corollary 4.2.3. And if we let the interpolation be close enough to the old values, then we can guarantee the positive definiteness of the new covariance matrices because the old ones are. The step-halving technique is only implemented for (4.21). There is no need for step-halving for β, since it has no restrictions. When using the step-halving technique, the iteration (4.21) becomes

$$\hat{\sigma}^{(t+1)} = \hat{\sigma}^{(t)} + a^{(t)}\delta^{(t)}, \tag{4.22}$$

where

$$\delta^{(t)} = \left[\sum_{i=1}^{n} \begin{bmatrix} \Phi^{-1}(\hat{\Sigma}_i^{(t)}) & \Phi^{-1}(\hat{\Sigma}_i^{(t)})J_i \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1} & J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1}J_i \end{bmatrix}\right]^{-1} \left(\sum_{i=1}^{n} \begin{bmatrix} \Phi(\hat{\Sigma}_i^{(t)})^{-1}\langle\hat{C}_i^{(t+1)}\rangle \\ J_i'\Phi(\hat{\Sigma}_i^{(t)})^{-1}\langle\hat{C}_i^{(t+1)}\rangle \end{bmatrix}\right) - \hat{\sigma}^{(t)},$$

and $0 < a^{(t)} \leq 1$ is used to ensure the positive definiteness of the covariance matrices. Starting from $a^{(t)} = 1$, if the matrices are not positive definite, we cut $a^{(t)}$ in half, repeatedly if necessary. We can stop the iterations when the change in either the log-likelihood function

or the parameter value does not exceed a predefined limit. Here, it is problematic to use stopping criteria involving the first order derivatives of the the log-likelihood function, since the algorithm may stop on the boundary of the parameter space when we force each Σ_i to be positive definite.

In the Newton-Raphson algorithm, we have to update the β and σ simultaneously. Thus, we are not able to restrict step-halving only to σ. As a result, due to the inherent complications we do not recommend step-halving for the Newton-Raphson algorithm, and strongly prefer to use the Simplified Method of Scoring algorithm here. Other reasons to prefer the Method of Scoring algorithm are: (i) the expected information matrix is robust to possible outliers; (ii) the expected information matrix at the last iteration leads to a better estimate of the asymptotic covariance matrix than does the empirical information matrix (Demidenko and Spiegelman, 1997); and (iii) the Method of Scoring algorithm has a preferable simpler form.

For maximization problems with constrained parameter spaces, the step-halving technique is not the only approach and need not to be the best one for all circumstances. Some researchers have proposed to let the estimates in the middle of the iterations go outside the parameter space as long as the algorithm is computable. For example, in our Method of Scoring algorithm, as long as $\hat{\Sigma}_i^{(t)}$ and the resulting coefficient matrices in (4.20) and (4.21) are invertible, there should not be any problem to get the estimates for the next iteration. And when we continue the iteration the estimates may very well re-enter the parameter space to make the final estimates legitimate. The attractive features of this method include that it is computationally simpler and possibly faster than using step-halving. Nevertheless, we implement step-halving for both simulations and applications in this research.

4.3 ASYMPTOTIC DISTRIBUTIONS

In this section, the asymptotic distribution of the parameter estimator is derived as the total sample size n goes to infinity. Due to our specific settings, more constraints need to be placed on the sample size. Roughly speaking, we want the number of data points useful for

the estimation of each parameter to go to infinity at a comparable rate. Denote the total number of possible covariance matrices given by the indicator matrices $\{I_i\}_{i=1}^n$ as q, where $q \leq 2^p - p$ as we have shown in Section 4.1.1. In the following, we also use $\Sigma_{[\alpha]}$ and $J_{[\alpha]}$ to denote, respectively, the common covariance matrix and the J matrix for each structure $\alpha = 1, \ldots, q$. Let n_α, $\alpha = 1, \ldots, q$, be the number of observations having each of those covariance structures. Then the condition is explicitly stated as the following.

Condition 4.3.1. $\lim_{n\to\infty}(n_\alpha/n) = \eta_\alpha \in (0,1)$ for $\alpha = 1, \ldots, q$.

Given the above condition, an argument similar to those in Theorem 7.3.1 and Theorem 7.3.2 of Lehmann (1999) can be given to show that given certain regularity conditions any consistent sequence, $\hat{\theta}(n)$, of solutions to the likelihood equations is asymptotically normal and efficient, and with probability one tends to the MLE as $n \to \infty$. While the results in Lehmann (1999) require normality, we drop the normality assumption in the following derivation of the asymptotics. We derive a stronger asymptotic result by using a one-iteration estimator of the Simplified Method of Scoring algorithm without step-halving and show that as long as the starting point is consistent, the parameter estimator obtained by one iteration of the Simplified Method of Scoring algorithm without step-halving is consistent and asymptotically normal. If we keep the iterations, then by induction every updated parameter estimator in the whole sequence is consistent and asymptotically normal.

Let $\hat{\beta}(n) = [\sum_{i=1}^n X_i'\Sigma_i^{-1}X_i]^{-1}(\sum_{i=1}^n X_i'\Sigma_i^{-1}y_i)$ be the estimate of β when using the true values of σ in (4.20) and define $\hat{\beta}^*(n) = [\sum_{i=1}^n X_i'\hat{\Sigma}_i^{-1}X_i]^{-1}(\sum_{i=1}^n X_i'\hat{\Sigma}_i^{-1}y_i)$ to be the estimate of β when using a consistent estimate of σ in (4.20). As a reminder, we assume that the X_i's are i.i.d. with finite second moments. Then by the Strong Law of Large Numbers (SLLN), $(1/n)\sum_{i=1}^n X_i'AX_i \xrightarrow{p} E[X'AX]$ as $n \to \infty$ for any finite constant matrix A, where X has the same distribution as each X_i. We first consider the situation where the σ is known and let

$$D_n = \frac{1}{n}\sum_{i=1}^n X_i'\Sigma_i^{-1}X_i = \sum_{\alpha=1}^q \frac{n_\alpha}{n}\frac{1}{n_\alpha}\sum_{i\in I(\alpha)} X_i'\Sigma_\alpha^{-1}X_i, \tag{4.23}$$

where $I(\alpha)$ is a set of index i such that Y_i has covariance matrix Σ_α. And then consider

$$\sqrt{n}(\hat{\beta}(n) - \beta) = D_n^{-1}(\sum_{\alpha=1}^q \frac{\sqrt{n_\alpha}}{\sqrt{n}}\frac{\sqrt{n_\alpha}}{n_\alpha}\sum_{i\in I(\alpha)} X_i'\Sigma_\alpha^{-\frac{1}{2}}z_i), \tag{4.24}$$

where conditional on X_i, $z_i = \Sigma_i^{-\frac{1}{2}}(y_i - X_i\beta)$ are i.i.d. with mean 0 and an identity covariance matrix. Our first result is the following:

Theorem 4.3.2. *Let $\{X_i\}_{i=1}^n$ be i.i.d. random samples from a distribution with finite second moments, and conditional on the X_i's, let the y_i, $i = 1, \ldots, n$, be independently distributed with y_i having mean $X_i\beta$ and covariance matrix as defined in Section 4.1.1. Then, $\sqrt{n}(\hat{\beta}(n) - \beta)$ has a limiting normal distribution with mean vector 0 and covariance matrix $(\sum_{\alpha=1}^q \eta_\alpha E[X_1'\Sigma_\alpha^{-1}X_1])^{-1}$ as $n \to \infty$.*

Proof. By SLLN, $D_n \to \sum_{\alpha=1}^q \eta_\alpha E[X'\Sigma_\alpha^{-1}X]$ in probability. And since $E[X_i'\Sigma_\alpha^{-\frac{1}{2}}z_i|X_i] = 0$ and $Var(X_i'\Sigma_\alpha^{-\frac{1}{2}}z_i|X_i) = X_i'\Sigma_\alpha^{-1}X_i$, we have $E[X_i'\Sigma_\alpha^{-\frac{1}{2}}z_i] = 0$ and $Var(X_i'\Sigma_\alpha^{-\frac{1}{2}}z_i) = E[X_i'\Sigma_\alpha^{-1}X_i] < \infty$. Then by the Central Limit Theorem (CLT), $(\sqrt{n_\alpha}/n_\alpha)\sum_{i\in i(\alpha)} X_i'\Sigma_\alpha^{-\frac{1}{2}}z_i \to N(0, E[X'\Sigma_\alpha^{-1}X])$ in distribution. And then by Slutzky's theorem, we have the result. □

Before deriving the asymptotic property of $\hat{\beta}^*(n)$, the estimator from (4.20) with consistent covariance matrices plugged in, we introduce the following useful lemma.

Lemma 4.3.3. *Let X_i, as well as Y_i, $i = 1, \ldots, n$, be random matrices (vectors or scalars as special cases). Assume $\{X_i, Y_i\}_{i=1}^n$ to be i.i.d. and have finite second moments. Let $A(n)$ be a sequence of random matrices such that $A(n) \xrightarrow{p} A$ as $n \to \infty$, where A is a constant matrix. Then we have the following two results.*

1. *As long as the dimensions are compatible, $(1/n)\sum_{i=1}^n X_i'A(n)Y_i$ has the same limit in probability as $(1/n)\sum_{i=1}^n X_i'AY_i$ as $n \to \infty$, that is, the limit is $E[X'AY]$, where (X, Y) has the same distribution as all the (X_i, Y_i)'s;*
2. *Given $E[Y|X] = 0$ and $Cov(Y|X)$ is positive definite and finite, $(1/\sqrt{n})(\sum_{i=1}^n X_i'A(n)Y_i)$ has the same limiting distribution as $(1/\sqrt{n})(\sum_{i=1}^n X_i'AY_i)$, if the latter has one.*

Proof. First, for the case where both X_i and Y_i are random vectors (or scalars), assume A and $A(n)$ are both matrices with compatible dimension with X_i and Y_i. Then we have

$$\frac{1}{n}\sum_{i=1}^n X_i'A(n)Y_i = \frac{1}{n}\sum_{i=1}^n tr\left(X_i'A(n)Y_i\right) = \frac{1}{n}\sum_{i=1}^n tr\left(A(n)Y_iX_i'\right) = tr\left(A(n)\frac{\sum_{i=1}^n Y_iX_i'}{n}\right).$$

By the SLLN and Slutzky's theorem, we have $(1/n)\sum_{i=1}^n X_i'A(n)Y_i \xrightarrow{p} tr(AE[YX']) = E[tr(AYX')] = E[X'AY]$.

For the case where both X_i and Y_i are matrices, we use the above results and have

$$(1/n)\sum_{i=1}^{n} X_i'A(n)Y_i = [(1/n)\sum_{i=1}^{n} x_{ij}'A(n)y_{ik}]_{jk} \xrightarrow{p} [E[x_j'Ay_k]]_{jk} = E[X'AY],$$

where x_{ij} and y_{ik} are the jth and kth columns of matrices X_i and Y_i, respectively. The results for one of them being a vector and the other being a matrix can be proved in the same way.

To show the second argument, we focus on the case where both X_i and Y_i are vectors, because the case where both X_i and Y_i are scalars is simple and other cases involving matrices will follow easily if we have the result for the vector case. Now consider

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i'A(n)Y_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i'AY_i = tr\left((A(n)-A)\frac{\sum_{i=1}^{n} Y_iX_i'}{\sqrt{n}}\right). \tag{4.25}$$

Because $E[Y|X] = 0$ and $Cov(Y|X)$ is positive definite and finite, we have $E[YX'] = 0$ and $Var(YX') = E[Var(Y|X) \otimes XX'] < \infty$, where $\otimes$ is the Kronecker product. Then by the CLT, $\sum_{i=1}^{n} Y_iX_i'/\sqrt{n}$ converges in distribution to the $N(0, Var(YX'))$ distribution. As a result, (4.25) converges in probability to zero. Then the result follows by Theorem 2.3.5 in Lehmann (1999,Section 2.3). □

Now we apply Lemma 4.3.3 to

$$\sqrt{n}(\hat{\beta}^*(n) - \beta) = \hat{D}_n^{-1}(\frac{1}{n}\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}(y_i - X_i\beta)), \tag{4.26}$$

where $\hat{D}_n^{-1} = (1/n)\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}X_i$. The first part of Lemma 4.3.3 proves that $\hat{D}_n$ converges in probability the same as D_n does, and the second part gives that $(1/n)\sum_{i=1}^{n} X_i'\hat{\Sigma}_i^{-1}(y_i - X_i\beta)$ converges in distribution the same as $(1/n)\sum_{i=1}^{n} X_i'\Sigma_i^{-1}(y_i - X_i\beta)$ does. Then we have the following corollary.

Corollary 4.3.4. *For the same settings as in Theorem 4.3.2, let $\hat{\beta}^*(n)$ be the estimate of β from one iteration of (4.20) where $\hat{\Sigma}_i^{(0)}$, $i = 1, \ldots, n$, are consistent estimators, respectively, of Σ_i, $i = 1, \ldots, n$. Then $\sqrt{n}(\hat{\beta}^*(n) - \beta)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\beta}(n) - \beta)$.*

In practice, the matrices $E[X'\Sigma_\alpha X]$, $\alpha = 1, \ldots, q$, cannot be calculated explicitly, because we neither have a specific distributional assumption for X, nor do we know the true parameter values. However, as in standard practice in such settings, we can estimate them using the sample moments with the estimated parameter values, that is, we estimate $E[X'\Sigma_\alpha X]$ by $(1/n)\sum_{i=1}^n X_i'\hat{\Sigma}_\alpha X_i$ for $\alpha = 1, \ldots, q$.

Finally, when β is known, equations (4.21) and (4.20) share the same form. So we have the following result for the estimates of σ.

Theorem 4.3.5. *For the same settings as in Theorem 4.3.2, let $\hat{\Sigma}_i^{(0)}$, $i = 1, \ldots, n$, be consistent estimators, respectively, of Σ_i, $i = 1, \ldots, n$ and assume β is known. Let $\hat{\sigma}^{(1)}$ be the solution to (4.21). Then $\sqrt{n}(\hat{\sigma}^{(1)} - \sigma)$ has a limiting normal distribution with mean 0 and covariance matrix E^{-1}, where*

$$E = \sum_{\alpha=1}^{q} \eta_\alpha \begin{bmatrix} \Phi^{-1}(\Sigma_\alpha) & \Phi^{-1}(\Sigma_\alpha)J_\alpha \\ J_\alpha'\Phi^{-1}(\Sigma_\alpha) & J_\alpha'\Phi^{-1}(\Sigma_\alpha)J_\alpha \end{bmatrix}.$$

However, β is usually not known. We would like to substitute for it, its consistent estimate $\hat{\beta}^{(0)}$, and have the result in Theorem 4.3.5 still hold.

Corollary 4.3.6. *For the same settings as in Theorem 4.3.2, let $\hat{\Sigma}_i^{(0)}$, $i = 1, \ldots, n$, be consistent estimators, respectively, of Σ_i, $i = 1, \ldots, n$, and $\hat{\beta}^{(0)}$ be a consistent estimator of β. Let $\hat{\sigma}^{(1)}$ be the solution to (4.21). Then $\sqrt{n}(\hat{\sigma}^{(1)} - \sigma)$ has a limiting normal distribution with mean 0 and covariance matrix E^{-1}.*

Proof. For $\alpha = 1, \ldots, q$, consider

$$\begin{aligned}
&\frac{1}{\sqrt{n_\alpha}} \sum_{i \in i(\alpha)} (y_i - X_i\hat{\beta}^{(0)})(y_i - X_i\hat{\beta}^{(0)})' \\
&= \frac{1}{\sqrt{n_\alpha}} \sum_{i \in i(\alpha)} (y_i - X_i\beta)(y_i - X_i\beta)' + \frac{1}{\sqrt{n_\alpha}} \sum_{i \in i(\alpha)} X_i(\beta - \hat{\beta}^{(0)})(y_i - X_i\beta)' \\
&\quad + \frac{1}{\sqrt{n_\alpha}} \sum_{i \in i(\alpha)} (y_i - X_i\beta)(\beta - \hat{\beta}^{(0)})'X_i' + \frac{1}{\sqrt{n_\alpha}} \sum_{i \in i(\alpha)} X_i(\beta - \hat{\beta}^{(0)})(\beta - \hat{\beta}^{(0)})'X_i'.
\end{aligned}$$

It is not hard to see that by Lemma 4.3.3, all the last three terms in the above equation converge in probability to zero. Then it follows that $\sqrt{n}(\hat{\sigma}^{(1)} - \sigma)$ has the same limiting distribution as stated in Theorem 4.3.5. □

When both β and σ are unknown, technically we can start with a consistent estimate $\hat{\sigma}^{(0)}$ of σ, obtain a consistent estimate $\hat{\beta}^*$ of β using (4.20) and then find the new estimate $\hat{\sigma}^{(1)}$ of σ using (4.21) with $\hat{\beta}^*$ and $\hat{\sigma}^{(0)}$ plugged in. Now, if we keep the iterations until convergence, then by induction the parameter estimates in the whole sequence will be consistent and have asymptotic distributions as stated in Theorem 4.3.2 and 4.3.5.

Theorem 4.3.7. *For the same settings as in Corollary 4.3.6, except now we keep the iterations until convergence, then every parameter estimate in both sequences $\hat{\beta}^{(t)}$ and $\hat{\sigma}^{(t)}$ for $t = 1, 2, \cdots$ is consistent and $\sqrt{n}(\hat{\beta}^{(t)} - \beta)$ and $\sqrt{n}(\hat{\sigma}^{(t)} - \sigma)$ have limiting normal distributions as stated in Theorem 4.3.2 and 4.3.5.*

Given (4.12) and Condition 4.3.1, it is not hard to show that $\hat{\theta}^{(t)}(n)$, $t = 1, 2, \ldots$, are asymptotically efficient in the sense that their covariance matrices achieve the Cramer-Rao lower bound as $n \to \infty$. So the one-iteration estimator is asymptotically equivalent to the MLE. In fact, in practice we will always stop the iterations in finite steps. We show by simulations in Section 4.4 that there is no big advantage in running the iteration for more than one step if the sample size is large, whereas for small sample size situations the advantage of running the algorithm until convergence is significant.

Finding a consistent starting point for σ is not hard. For example, one iteration of (4.20) and (4.21) from identity covariance matrices yields the ordinary least square estimator for β and the method of moments estimator for σ, both of which are consistent. Furthermore, for the special case where $X_i \equiv X$ for $1 \leq i \leq n$, we treat X to be a constant matrix. As long as the columns of X are linearly independent, the conclusion in Theorem 4.3.7 still holds. For instance, in our neurobiological setting, if we decided not to use any covariates other than a "1" resulting from the diagnostic effect, then all the X_i's will be the same.

Finally, having the asymptotic distributions of the parameter estimates available, we are able to test the unknown parameters using Wald tests. For example, in our neurobiological context, we could test to see if age has a significant effect on the measurements.

4.4 SIMULATIONS STUDY

In this section, we provide simulation results which examine the properties of both the MLEs and the one-iteration estimates. The Simplified Method of Scoring algorithm is implemented using the identity covariance matrices as a starting value with a maximum of 200 iterations to obtain the MLEs. And to find the one-iteration estimates, we carry out the Simplified Method of Scoring algorithm for two iterations, thereby guaranteeing a starting value which is consistent.

We simulate multivariate normal data with dimension $p = 3$ and sample sizes of $n = 25, 50$ and 100. Since for $p = 3$ there are a total of 5 possible different covariance structures as illustrated in Table 2.1, we simulate the case where all five covariance structures occur equally often, that is, we simulate $n/5$ data points for each of them, so that $\eta_\alpha = 0.2$ for $\alpha = 1, \ldots, 5$, in Condition 4.3.1. For each of the three sample sizes, in order to obtain the sampling distributions of both the MLEs and the one-iteration estimates, we do 1000 simulations and obtain both estimates from every simulation. We consider one other setting where the five covariance structures do not occur equally often. In this simulation, the total sample size is 610, with $n_1 = n_5 = 0$, $n_2 = 10$, $n_3 = 100$ and $n_4 = 500$ corresponding to each of the five covariance matrices.

Using the parameterization of Section 4.1.1, we set the covariates $X_i = I \otimes x_i'$, $i = 1, \ldots, n$, where $\otimes$ is the Kronecker product and I is the identity matrix with dimension equal to three; and x_i is a three dimensional vector with its first element being 1, second element being an integer sampled from 20 to 80 with equal probabilities, and third element being sampled uniformly from $\{0, 1\}$. This setting attempts to imitate the pattern of the covariates in our motivating data. We set $\beta = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{31}, \beta_{32}, \beta_{33}) = (-8, 0.04, 0.1, -28, -0.6, 1, -60, 0.4, 15)$ and $\sigma = (\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}, \sigma^c_{12}, \sigma^c_{13}, \sigma^c_{23}) = (50, 900, 500, 120, 100, 400, 80, -100, -300)$.

In finding the MLEs, the Method of Scoring algorithm converges within 200 iterations in 84.6% of the simulations for $n = 25$, 98.5% for $n = 50$, and 100% for $n = 100$. Figure 4.1 shows the histograms for all the 1000 simulation results for both the MLE and the one-iteration estimate of $\beta_{22} = -0.6$ as compared to the normal asymptotic distribution based

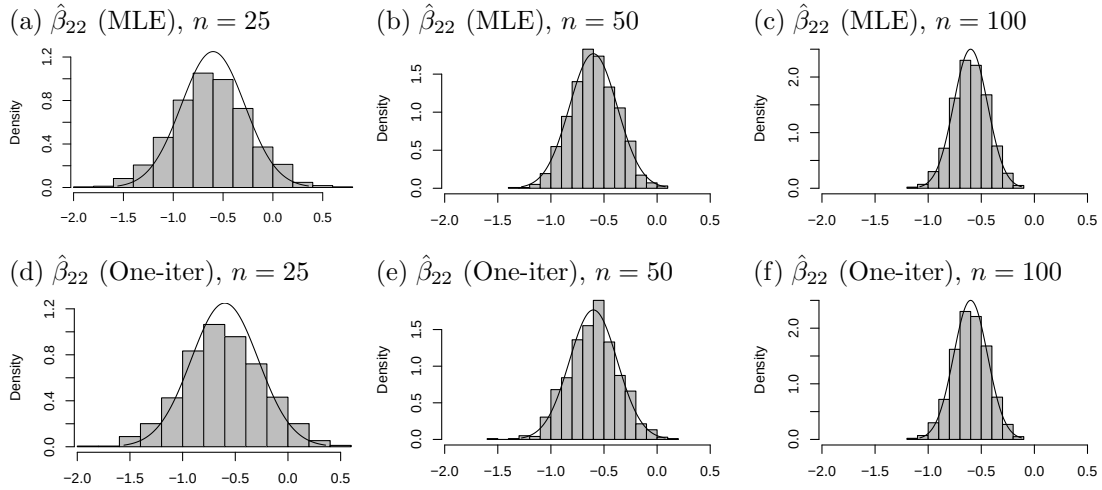


Figure 4.1: Simulation histograms and asymptotic distribution of $\hat{\beta}_{22}$: both the MLE and the one-iteration estimate have $\sqrt{n}(\hat{\beta}_{22} - \beta_{22}) \xrightarrow{D} N(0, 1.595^2)$; (a) the MLE for $n = 25$; (b) the MLE for $n = 50$; (c) the MLE for $n = 100$; (d) the one-iteration estimate for $n = 25$; (e) the one-iteration estimate for $n = 50$; (f) the one-iteration estimate for $n = 100$.

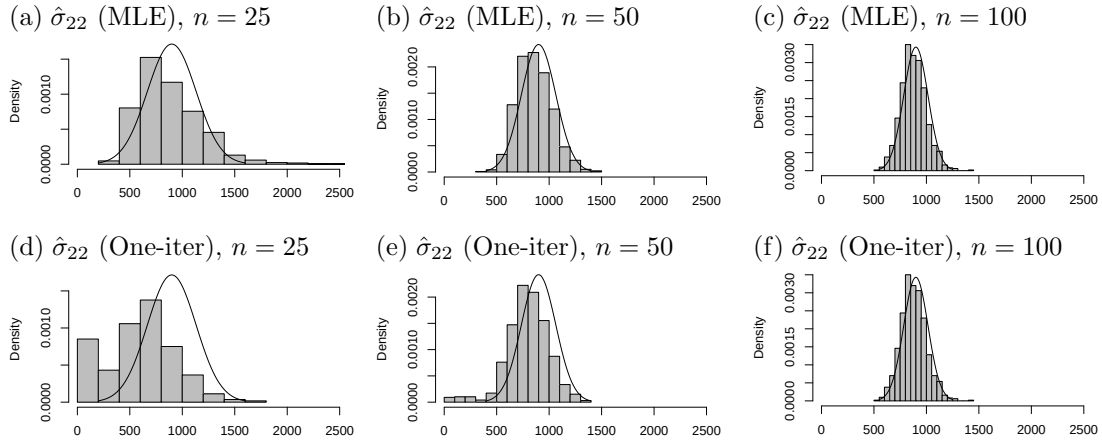


Figure 4.2: Simulation histograms and asymptotic distribution of $\hat{\sigma}_{22}$: both the MLE and the one-iteration estimate have $\sqrt{n}(\hat{\sigma}_{22} - \sigma_{22}) \xrightarrow{D} N(0, 1162^2)$; (a) the MLE for $n = 25$; (b) the MLE for $n = 50$; (c) the MLE for $n = 100$; (d) the one-iteration estimate for $n = 25$; (e) the one-iteration estimate for $n = 50$; (f) the one-iteration estimate for $n = 100$.

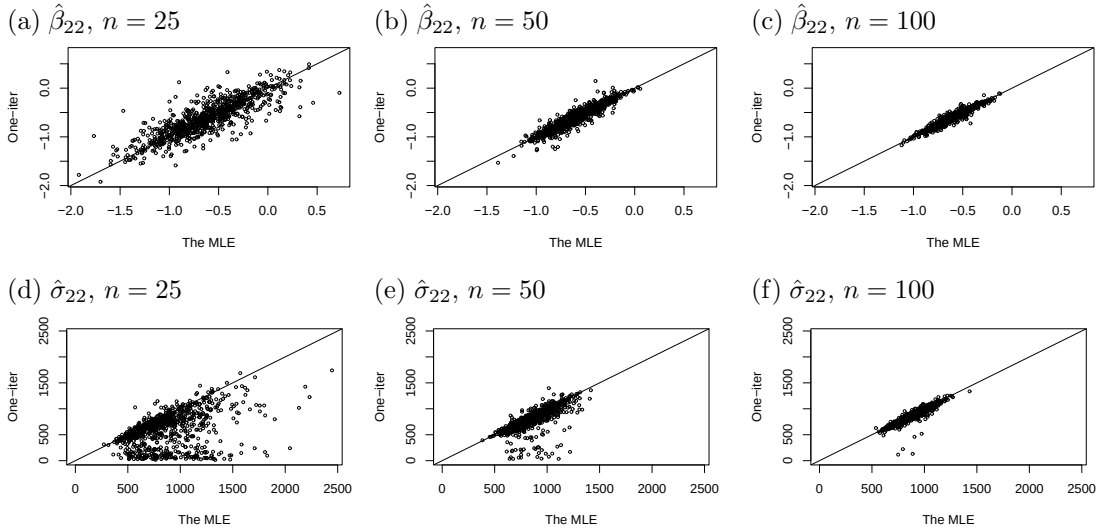


Figure 4.3: A pairwise comparison of the MLE and the one-iteration estimate: (a) $\hat{\beta}_{22}$ for $n = 25$; (b) $\hat{\beta}_{22}$ for $n = 50$; (c) $\hat{\beta}_{22}$ for $n = 100$; (d) $\hat{\sigma}_{22}$ for $n = 25$; (e) $\hat{\sigma}_{22}$ for $n = 50$; (f) $\hat{\sigma}_{22}$ for $n = 100$.

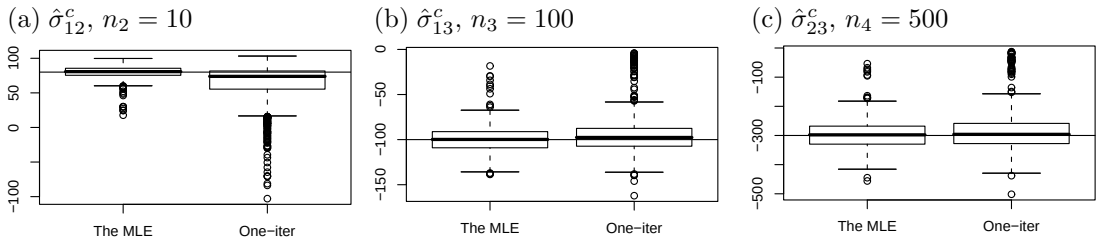


Figure 4.4: Boxplots of $\hat{\sigma}^c_{12}$, $\hat{\sigma}^c_{13}$ and $\hat{\sigma}^c_{23}$ in the unbalanced case: the MLEs have $\sqrt{n}(\hat{\sigma}^c_{12} - \sigma^c_{12}) \xrightarrow{D} N(0, 186^2)$, $\sqrt{n}(\hat{\sigma}^c_{13} - \sigma^c_{13}) \xrightarrow{D} N(0, 320^2)$ and $\sqrt{n}(\hat{\sigma}^c_{23} - \sigma^c_{23}) \xrightarrow{D} N(0, 1054^2)$, (a) $n_2 = 10$ is useful to estimate σ^c_{12}; (b) $n_2 = 100$ is useful to estimate $\hat{\sigma}^c_{13}$; (c) $n_2 = 500$ is useful to estimate $\hat{\sigma}^c_{23}$.

on the true parameter values (Note: the true value of β_{22} in this simulation is -0.6). The simulation histograms confirm the appropriateness of the asymptotic normal distribution even for small sample sizes.

We do the same thing for $\sigma_{22} = 900$ in Figure 4.2. However, this time there appears to be an underestimation, especially for small sample sizes, since the means of the simulation histograms are less than the true value. And the one-iteration estimate is worse in underestimating σ_{22} than the MLE. In order to directly compare the MLE and the one-iteration estimate, we plot the one-iteration estimate versus the MLE for each simulation for both β_{22} and σ_{22} in Figure 4.3. It shows that the one-iteration estimate and the MLE are comparable for estimating β_{22} for all sample sizes, whereas the one-iteration estimate is worse than the MLE in terms of underestimating σ_{22}. However, the difference diminishes as the sample size gets larger. Results for the unbalanced case are summarized in Figure 4.4. It shows that in a single data set, the accuracy of parameter estimate depends on the number of data points actually useful for the estimation of that parameter.

In addition, we conducted more simulations on some even larger sample sizes, n = 250, 500, 1000, 2500 and 10000, which we do not report on in detail here. We find that for both the β and the σ, the sample size n = 100 is large enough for the one-iteration estimator to approximate the MLE, and also for this sample size the histograms of the 1000 simulation results confirm the appropriateness of the asymptotic normal distribution based on the true parameter values. And as the sample size becomes larger, both estimators follow the asymptotic distributions even more closely.

In conclusion, for the parameters in the mean structure, the one-iteration estimates are close to the MLEs, and the simulation distributions of both estimators are close to the asymptotic normal distributions even for small sample sizes. On the contrary, for the parameters in the covariance structure, there are nonneglectable biases for both the one-iteration estimates and the MLEs for small sample sizes, with the one-iteration having greater bias. As a result, for estimation of the parameters in the mean structure, one iteration of the Simplified Method of Scoring algorithm is generally enough. To conduct hypothesis testing with respect to the parameters in the mean structure or to estimate the parameters in the covariance structure for small samples, there is an advantage to continue the iterations until

convergence thereby obtaining the MLEs.

4.5 APPLICATIONS TO THE ILLUSTRATIVE EXAMPLE

For the data given in Table 4.1, we find the estimates using the new model, where the dependent variables are the paired differences on BDNF, TrkB and GAD_{67} and the covariates include the constant "1", age and gender of the subjects with schizophrenia. The covariance structures are indexed by the last column of Table 4.1. As a comparison, we also show the parameter estimates from a multivariate regression model assuming all cases have the same covariance structure. These results are presented in Table 4.2.

Under the new model specification, we are able to do Wald testing for the parameters using the asymptotic distributions. For instance, in testing the hypothesis of whether or not age significantly affects GAD_{67}, that is, $H_0 : \beta_{32} = 0$ vs. $H_a : \beta_{32} \neq 0$, we have the estimate $\hat{\beta}^N_{32} = 0.163$ and the asymptotic standard error $s(\hat{\beta}^N_{32}) = 0.293$, so that $z = 0.163/0.293 = 0.556$ and the p-value $= 0.58$. We would conclude for these data that age has a nonsignificant positive effect on GAD_{67} at level $\alpha = 0.05$. Furthermore, one can test whether or not age has significant effects on all three dependent variables simultaneously by observing that $X'\Sigma^{-1}X \sim \chi^2_p$ for $X \sim N_p(0, \Sigma)$. For our example, the null hypothesis is $H_0 : \beta_{12} = \beta_{22} = \beta_{32} = 0$, and $(\hat{\beta}^N_{12}, \hat{\beta}^N_{22}, \hat{\beta}^N_{32})' \sim N_3(0, \Omega)$ under H_0, where Ω is the corresponding asymptotic covariance matrix. We calculate the test statitics $\chi^2_3 = 0.156$ and p-value $= 0.984$. So there is no statistical evidence that age has an effect on the dependent variables at the .05 level.

This example illustrates how the parameter estimates change when an inappropriate covariance matrix is used instead of an appropriate one. For example, whereas in the new model $\hat{\beta}^N_{12} = -0.035$, $\hat{\beta}^N_{13} = -0.549$ and $\hat{\beta}^N_{21} = -40.72$, in the multivariate regression model we have $\hat{\beta}^R_{12} = 0.036$, $\hat{\beta}^R_{13} = 0.096$ and $\hat{\beta}^R_{21} = -28.45$. When it comes to the clustering problem that we consider in Section 5, we anticipate that the changes in the parameter estimates might be even larger and could significantly alter the clustering results.

Table 4.2: Estimates for data given in the illustrative example

Param.	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}	β_{31}	β_{32}	β_{33}
New Model Est.	**-3.070**	**-0.035**	**-0.549**	**-40.72**	**0.083**	**5.143**	**-53.77**	**0.163**	**14.16**
Multi. Reg. Est.	-7.554	0.036	0.096	-28.45	-0.060	0.982	-62.67	0.364	15.06
Param.	σ_{11}	σ_{22}	σ_{33}	σ_{12}	σ_{13}	σ_{23}	σ^c_{12}	σ^c_{13}	σ^c_{23}
New Model Est.	**49.32**	**885.1**	**540.8**	**122.9**	**106.9**	**428.9**	**80.51**	**-103.8**	**-391.9**
Multi. Reg. Est.	35.33	986.6	575.1	138.7	68.805	493.7			

5.0 CLUSTERING OF SUBJECTS WITHOUT MISSING DATA

In this chapter, we develop methods for the clustering analysis of the subjects with schizophrenia by using the estimation procedures developed in the earlier chapters in conjunction with a generalization of the EM gradient algorithm (Lange, 1995) and the algorithm introduced in Titterington (1984). We assume the bio-markers used for the clustering analysis are preselected and the data are complete. A new clustering algorithm is developed and found to provide the same simulation results as a direct application of the EM gradient algorithm and Titterington's (1984) algorithm for our setting, but is more time efficient in comparison. A review of recent literature on the topic of regression clustering is also given.

5.1 CLUSTERING LITERATURE REVIEW

We have considered many types of clustering methods, including K-means and some hierarchical clustering algorithms. None of them, except a probabilistic clustering algorithm using finite mixture models, seems to be appropriate for our post-mortem tissue data. Thus, our efforts are focused on building a model-based clustering algorithm with a finite mixture of normal distributions appropriate to our specific settings. When there is both an outcome variable and covariates, different names, including *regression models for conditional normal mixtures* and *regression clustering*, have been given to the problem. The basic idea is to cluster the subjects according to the discrepancy in the regression parameters or in addition the covariance parameters. Choosing the number of mixture components is a hard and yet unsolved problem, but hopefully one can make use of some information external to the data to get a reasonable choice. The literature on this topic includes DeSarbo and Corn (1988),

Jones and McLachlan (1992), Arminger, Stein, and Wittenberg (1999) and Zhang (2003).

Let the normal density function with mean μ and variance σ^2 be denoted by $\phi(y;\mu,\sigma^2)$. DeSarbo and Corn (1988) defined a regression model for finite normal mixtures with a univariate outcome y_i given covariates x_i as

$$f(y_i|x_i) = \sum_{k=1}^{g} \pi_k \phi(y_i; x_i'\beta_k, \sigma_k^2) \tag{5.1}$$

where β_k is the column vector of regression parameters and σ_k^2 is the error variance for component k, $k = 1,\ldots,g$. The first covariate is possibly the constant 1. Given the number of components g, DeSarbo and Corn (1988) estimated the parameters using the EM algorithm. Their method was extended by Jones and McLachlan (1992) to a multivariate setting, i.e., y_i is a vector, as

$$f(y_i|x_i) = \sum_{k=1}^{g} \pi_k \phi(y_i; B_k x_i, \Sigma_k) \tag{5.2}$$

where B_k is the matrix of regression parameters and Σ_k is the covariance matrix for component k. Here, $\phi(y;\mu,\Sigma)$ is the density of a multivariate normal distribution with mean vector μ and covariance matrix Σ. The same EM algorithm is used to estimate the parameters. For the models with constrained or parameterized mean and covariance structures where

$$B_k = B_k(\theta) \quad \text{and} \quad \Sigma_k = \Sigma_k(\theta), \ k = 1,\ldots,g,$$

Arminger et al. (1999) introduced three likelihood based strategies for the estimation of the parameters. The first one is called a two stage procedure with the first stage carrying out an unconstrained estimation procedure using the direct EM algorithm introduced in DeSarbo and Corn (1988) and Jones and McLachlan (1992). Upon obtaining the estimates, $\hat{B}_k$ and $\hat{\Sigma}_k$, and their estimated asymptotic joint covariance matrix, $\hat{\Omega}$, the second stage estimates the parameter vector θ from $\hat{B}_k$ and $\hat{\Sigma}_k$ by minimizing

$$D(\theta) = [\hat{\kappa} - \kappa(\theta)]'\hat{\Omega}^{-1}[\hat{\kappa} - \kappa(\theta)] \tag{5.3}$$

over θ, where vector κ denotes collectively the parameters in $\{B_k\}$ and $\{\Sigma_k\}$. The resulting estimator $\hat{\theta}$ is shown to be asymptotically normal with mean θ and covariance matrix

$$V(\hat{\theta}) = \left[\left(\frac{\partial \kappa'(\theta)}{\partial \theta}\right)\Omega^{-1}\left(\frac{\partial \kappa'(\theta)}{\partial \theta}\right)'\right]^{-1}. \tag{5.4}$$

A consistent estimator $\hat{V}(\hat{\theta})$ of $V(\hat{\theta})$ can be found by replacing θ and Ω by $\hat{\theta}$ and $\hat{\Omega}$, respectively. The estimates $\hat{\pi}_k$ of π_k, $k = 1, \ldots, g$, are obtained in the first stage and remain unchanged in the second stage. The second procedure discussed in Arminger et al. (1999) is the direct EM algorithm as discussed in DeSarbo and Corn (1988) and Jones and McLachlan (1992). As noted by Arminger et al. (1999), sometimes one would need another iterative algorithm, e.g., Newton's algorithm, in each iteration of the M-step, which might significantly slow the computational speed. The last procedure introduced in Arminger, Stein, and Wittenberg (1999) is the EM gradient algorithm as proposed in Lange (1995). This algorithm proceeds in the same way as the direct EM algorithm, except that only one iteration of Newton's algorithm is carried out in its M-step. Its properties were discussed in Lange (1995). The asymptotic covariance matrix of the estimates can be found using the Fisher information matrix, the observed information matrix or Louis's method as described in Section 3.2. This last algorithm is of special interest in our proposal because it fits well into our settings. The details of implementing it are discussed in Section 5.3.

Additionally, Zhang (2003) introduced some related data mining strategies in doing regression clustering (RC), including RC-KM (K Means) and RC-KHM (K-Harmonic Means). These are hard boundary clustering algorithms. Let $Z = (X, Y) = \{(x_i, y_i); i = 1, \ldots, n\}$ be the data and $\{Z_k\}_{k=1}^g$ with $Z = \bigcup_{k=1}^g Z_k$ and $Z_k \bigcap Z_{k'} = \emptyset$ for $k \neq k'$ be any partition of the data. Zhang (2003) solves the problem by minimizing

$$f_{RC}(\{Z_k\}_{k=1}^g, \{f_k\}_{k=1}^g) = \sum_{i=1}^{n} \psi\{(f_k(x_i), y_i); 1 \leq k \leq g\} \tag{5.5}$$

over $\{Z_k\}_{k=1}^g$ and $\{f_k\}_{k=1}^g$, where $\{f_k\}_{k=1}^g$ are chosen from a set of functions Φ (which are typically linear regression on x). For RC-KM, $\psi\{(f_k(x_i), y_i); 1 \leq k \leq g\} = \min_{1 \leq k \leq g}\{e(f_k(x_i), y_i)\}$ and usually $e(f_k(x_i), y_i) = \|f_k(x_i) - y_i\|^p$ with $p = 1, 2$, while for RC-KHM, $\psi\{(f_k(x_i), y_i); 1 \leq k \leq g\}$ is the harmonic mean of $\{\|f_k(x_i) - y_i\|^p\}_{k=1}^g$ for $p \geq 2$. Algorithms are available in Zhang (2003) for finding a local optimum. Basically, these algorithms iteratively fit some multiple linear regression models within each cluster, move observations to the closest clusters for the next iteration and stop when the target function does not change much. Such algorithms are extremely valuable for a fast exploratory data analysis due to their straightforward nature and relatively simple forms.

5.2 SETTINGS FOR THE CURRENT PROBLEM

In order to model our problem in a similar fashion as those in DeSarbo and Corn (1988) and Jones and McLachlan (1992), both their mean and covariance structures need to be extended. The mean structure should be extended to encompass the covariate forms in Section 4.1.1, and the covariance structure should be defined as (4.4).

The specific mean structure we are considering results from the nature of post-mortem tissue studies and can obviously be applied in other similar settings. Most of the time, in these studies there are, associated with each subject, a number of demographic characteristics that may have different effects in possible subpopulations. Thus, covariates, when included in the model, sometimes act merely as adjustment factors to make the clustering based on the disease effect more appropriate, or sometimes more importantly, they, themselves, have effects defining the clusters. In our context, since we are using the pairwise differences as the outcome variables, to represent the diagnostic effect a constant 1 is the first covariate considered. Other than this, our analysis of previous individual studies has suggested that age might be associated with the disease effect. So the age effect is not merely an adjustment but may, in fact, be an effect defining the clustering. Another clustering covariate being considered is gender. Other characteristics, such as brain pH value and post-mortem time interval, can be covariates; however, we do not believe that they are related to the clusters. In addition, there may be study level effects related to unknown experimental factors that can affect all the measurements in a study in a similar way. In our approach, we try to eliminate these effects by computing the pairwise differences. Furthermore, we begin with the assumption that there are at most two clusters of the subjects with schizophrenia in our data. Thus, when we apply our methodologies to our data, we will only consider a mixture with two components.

The covariance structure proposed in Section 4.1.1 is specific to our setting. It results from the use of different control subjects across studies for some subjects with schizophrenia and treating pairwise differences as the dependent variables. Previous results from individual studies have confirmed the existence of significant correlations among different bio-markers within or across regions for both control and schizophrenic groups as also for the pairwise

differences (for example, see Hashimoto et al., 2003, 2005). And it is quite straightforward that the specification of the covariance structure can affect the parameter estimation and furthermore affect the clustering on the subjects. However, it seems to us that at this stage in our methodologies it makes little sense biologically to assume that covariance parameters differ across possible clusters. At the very least we do not believe we can determine this from the amount of data that we currently have. On the other hand, assuming the same covariance parameters across possible clusters does give us a number of statistical advantages. For example, we do not lose too much efficiency in parameter estimation in the case of small sample sizes. And it saves a lot of computational burden by reducing the number of free parameters.

5.3 CLUSTERING ALGORITHMS

While still assuming no missing data, we begin this section with introducing a mixture model for the heterogeneity of the subjects with schizophrenia followed by a discussion of the properties of some existing model fitting algorithms, including the EM algorithm, the EM gradient algorithm and Titterington's (1984) algorithm. We consider applying them to our specific mixture problem. A new algorithm is then developed and shown to have some nicer properties over the existing ones. Instead of assuming only 2 clusters, we derive the algorithms generally for $g \geq 2$ clusters. The results can be directly applied to the case of $g = 2$. However, for the applications to our actual data, it is not practical to assume $g > 2$.

5.3.1 Existing Algorithms

We consider $Z_i = (Z_{i1}, \cdots, Z_{ig})$, $i = 1, \ldots, n$, to be the unobserved group indicators for integer $g \geq 2$, i.e., we assume that, in general, the data are from a mixture of g subpopulations. Unconditionally, $\{Z_i\}_{i=1}^n$ are i.i.d. with a multinomial$(1, \pi_1, \cdots, \pi_g)$ distribution. And conditionally, for the observed data $\{y_1, \cdots, y_n\}$ we assume

$$f(y_i|z_{ik} = 1) = \phi(y_i; X_i\beta_k, \Sigma_i), \tag{5.6}$$

where $\phi(\cdot; X\beta, \Sigma)$ is the density function of a multivariate normal distribution with mean $X\beta$ and covariance matrix Σ. As we discussed in Section 5.2, the clusters are defined based on the parameters $\{\beta_k\}_{k=1}^g$ in the mean structure, and the parameters σ in the covariance structure are kept the same across clusters. The covariance matrices $\{\Sigma_i\}_{i=1}^n$ have the same forms as defined in Section 4.1.1 given the control indicators $\{I_i\}_{i=1}^n$. Then a straightforward method to obtain the estimates of the parameters and cluster the subjects with schizophrenia is the EM algorithm. Using the notations as in Section 3.3.2, the conditional expectation of the complete data log-likelihood function given the observed data and the parameter estimates from the previous iteration $\theta = \theta^{(t)}$, where θ is the collection of all the parameters in $\{\pi_k\}_{k=1}^g$, $\{\beta_k\}_{k=1}^g$ and σ, can be written as

$$Q\left(\theta|\theta^{(t)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{g} \tau_{ik}^{(t)}\{\log \pi_k + \log \phi(y_i; X_i\beta_k, \Sigma_i)\} \tag{5.7}$$

where

$$\tau_{ik}^{(t)} = \frac{\pi_k^{(t)}\phi(y_i; X_i\beta_k^{(t)}, \Sigma_i^{(t)})}{\sum_{j=1}^g \pi_j^{(t)}\phi(y_i; X_i\beta_j^{(t)}, \Sigma_i^{(t)})}. \tag{5.8}$$

The EM algorithm iterates between computing (the E-step) and maximizing (the M-step) the Q function over θ for $t = 0, 1, 2, \ldots$ until convergence under certain criterion. If the EM algorithm were applied to our problem, the updates of the subpopulation probabilities in the M-step would have explicit forms as

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)}}{n}, \quad k = 1, \ldots, g. \tag{5.9}$$

By restricting the variance-covariance components to be the same across different subpopulations, we find the first partial derivatives of the $Q(\theta|\theta^{(t)})$ function with respect to its first argument as

$$\partial Q/\partial\beta_j = \sum_{i=1}^{n} \tau_{ij}^{(t)} X_i'\Sigma_i^{-1}(y_i - X_i\beta_j),\ j = 1, \ldots, g, \tag{5.10a}$$

$$\partial Q/\partial\sigma_{kl} = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}(\sum_{j=1}^{g} \tau_{ij}^{(t)}C_{ij} - \Sigma_i),\ 1 \le k \le l \le p\,, \tag{5.10b}$$

$$\partial Q/\partial\sigma_{kl}^c = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}(\sum_{j=1}^{g} \tau_{ij}^{(t)}C_{ij} - \Sigma_i)I_i^{kl},\ 1 \le k < l \le p\,, \tag{5.10c}$$

where $C_{ij} = (y_i - X_i\beta_j)(y_i - X_i\beta_j)'$ for $1 \le i \le n$ and $1 \le j \le g$. By setting the quantities in (5.10) equal to zeros and solving, we can find the next updates of the parameters $\{\beta_k\}_{k=1}^g$ and σ. However, in our problem we do not have a closed form solution and require another iterative algorithm in the M-step. This fact sometimes renders this algorithm computationally ineffective in practice.

As a result, we consider a newer algorithm introduced in Lange (1995) – the EM gradient algorithm. In order to use this EM gradient algorithm, the first and second derivatives of the function $Q(\theta|\theta^{(t)})$ with respect to its first argument are required. Continuing to take partial derivatives of (5.10) yields the second partial derivatives of the Q function as

$$-\partial^2 Q/\partial\beta_j^2 = \sum_{i=1}^n \tau_{ij}^{(t)} X_i'\Sigma_i^{-1}X_i,\ 1 \le j \le g, \tag{5.11a}$$

$$-\partial^2 Q/\partial\beta_j\partial\beta_k = 0,\ 1 \le j \ne k \le g, \tag{5.11b}$$

$$-\partial^2 Q/\partial\sigma_{kl}\partial\sigma_{st} = (1/2)\sum_{i=1}^n tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}\Sigma_i^{-1}(2\sum_{j=1}^g \tau_{ij}^{(t)}C_{ij} - \Sigma_i), \quad 1 \le k \le l \le p\,, 1 \le s \le t \le p\,, \tag{5.11c}$$

$$-\partial^2 Q/\partial\sigma_{kl}^c\partial\sigma_{st}^c = (1/2)\sum_{i=1}^n tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}\Sigma_i^{-1}(2\sum_{j=1}^g \tau_{ij}^{(t)}C_{ij} - \Sigma_i)I_i^{kl}I_i^{st}, \quad 1 \le k < l \le p\,, 1 \le s < t \le p\,, \tag{5.11d}$$

$$-\partial^2 Q/\partial\beta_j\partial\sigma_{kl} = \sum_{i=1}^n \tau_{ij}^{(t)} X_i'\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}(y_i - X_i\beta_j),\ 1 \le j \le g, 1 \le k \le l \le p\,, \tag{5.11e}$$

$$-\partial^2 Q/\partial\beta_j\partial\sigma_{kl}^c = \sum_{i=1}^n \tau_{ij}^{(t)} X_i'\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}(y_i - X_i\beta_j)I_i^{kl},\ 1 \le j \le g, 1 \le k < l \le p\,, \tag{5.11f}$$

$$-\partial^2 Q/\partial\sigma_{kl}\partial\sigma_{st}^c = (1/2)\sum_{i=1}^n tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}\Sigma_i^{-1}(2\sum_{j=1}^g \tau_{ij}^{(t)}C_{ij} - \Sigma_i)I_i^{st}, \quad 1 \le k \le l \le p\,, 1 \le s < t \le p\,. \tag{5.11g}$$

Then in the M-step the EM gradient algorithm updates the parameter values with one iteration of the Newton's method by

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)}\left[\frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial\theta^2}\right]^{-1}\left(\frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta}\right)|_{\theta=\theta^{(t)}}, \tag{5.12}$$

with $\alpha^{(t)}$ being a possible step size. The E-step is carried through as usual. According to Lange et al. (2000), this EM gradient algorithm can also be derived from the view of "optimization transfer", for which we provide a brief introduction in the following. By Dempster et al. (1977), we have

$$l(\theta) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}), \tag{5.13}$$

$$\frac{\partial H(\theta|\theta^{(t)})}{\partial \theta}|_{\theta=\theta^{(t)}} = 0, \tag{5.14}$$

$$\frac{\partial^2 H(\theta|\theta^{(t)})}{\partial \theta^2}|_{\theta=\theta^{(t)}} < 0, \tag{5.15}$$

with $H(\theta|\theta^{(t)}) = E_{Z|Y,\theta^{(t)}}[\log(f_\theta(Y,Z)/f_\theta(Y))]$, where Y is the observed data, Z is the unobserved group indices and $f_\theta(\cdot)$ presents the density function with parameter θ. As a result, we have

$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial \theta}|_{\theta=\theta^{(t)}} = \frac{\partial l(\theta)}{\partial \theta}|_{\theta=\theta^{(t)}}, \tag{5.16}$$

$$\frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial \theta^2}|_{\theta=\theta^{(t)}} = \frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\theta^{(t)}} + \frac{\partial^2 H(\theta|\theta^{(t)})}{\partial \theta^2}|_{\theta=\theta^{(t)}} < \frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\theta^{(t)}}. \tag{5.17}$$

Thus, the EM gradient algorithm is merely an approximation to the Newton-Raphson algorithm in maximizing $l(\theta)$ by ignoring the $H(\theta|\theta^{(t)})$ part in the Hessian matrix. Given (5.17) and a number of regularity conditions, this algorithm has almost the same local convergence properties as the usual EM algorithm, i.e., in a neighborhood of a local optimum, it is ascending and converges linearly. As one of the regularity conditions, it is required that $\partial^2 Q(\theta|\theta^{(t)})/\partial\theta^2$ be negative definite, which secures the existence of its inverse and thus guarantees the convergence of the EM gradient algorithm. This condition of concavity is always satisfied near a local maximum due to (5.17), but is not guaranteed globally. For certain distributions, e.g., the exponential family with natural parameterization, the observed log-likelihood function is easily shown to be concave, and so is the Q function. Unfortunately, in general this is not true. So this EM gradient algorithm does not share the property of global monotonicity with the usual EM algorithm when it starts far away from the optimum. Thus, directly using this EM gradient algorithm in our setting is dangerous. Even if this algorithm did produce invertible matrices, it could be very time consuming because one large matrix needs to be evaluated and inverted in each iteration.

Lange (1995) proposed a variant, which he called the limited line search, to enforce the global monotonicity by adjusting $\alpha^{(t)}$ in each step. It maximizes $Q(\theta|\theta^{(t)})$ along the EM gradient direction $d(\theta^{(t)}) = -[\partial^2 Q(\theta|\theta^{(t)})/\partial\theta^2]^{-1}(\partial Q(\theta|\theta^{(t)})/\partial\theta)$ from the current point $\theta^{(t)}$ in the M-step. Lange (1995) also showed that there was a unique point $\theta^{(t)} + \alpha^{(t)} d(\theta^{(t)})$ maximizing $Q(\theta^{(t)} + \alpha d(\theta^{(t)})|\theta^{(t)})$ for $0 < \alpha < 1$. As another disadvantage of both the EM gradient algorithm and its limited line search version, (5.12) does not ensure that the estimate $\theta^{(t+1)}$ falls in the parameter space. Sometimes, a reparameterization can surmount this difficulty, but not always. And it seems to us that it is hard to maintain the global monotonicity and ensure the estimates falling in the parameter space simultaneously.

Since the actual data obtained by combining data from multiple post-mortem tissue studies has a large degree of missingness, we ultimately will need to consider implementing multiple imputation techniques to deal with the missing data. Given the degree of missingness, a large amount of imputation is necessary, so that time efficiency is an important characteristic of the algorithms that we must consider. We want an algorithm that is more time efficient and more stable in comparison to the EM gradient algorithm when applied to our problem. Thus, we consider applying Titterington's (1984) algorithm to our mixture problem. Titterington (1984) used the Fisher information matrix of the complete data instead of the matrix $-\partial^2 Q(\theta|\theta^{(t)})/\partial\theta^2$ in (5.12). That is

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \left[I_c(\theta^{(t)})\right]^{-1} \left(\frac{\partial Q(\theta|\theta^{(t)})}{\partial\theta}\right)|_{\theta=\theta^{(t)}}, \tag{5.18}$$

where $I_c(\theta)$ is the complete data information matrix. For a variety of models, for example, the mixtures with normal densities, $I_c(\theta)$ has a simpler form than $-\partial^2 Q(\theta|\theta^{(t)})/\partial\theta^2$, which is sometimes an intriguing feature. And $I_c(\theta)$ is guaranteed to be positive definite in the neighborhood of a local maximum. Furthermore, it is not hard to prove that

$$I_c(\theta^{(t)}) \equiv E_{Z,Y|\theta}\left[\frac{\partial^2 l_c(\theta)}{\partial\theta^2}\right]|_{\theta=\theta^{(t)}} = E_{Y|\theta}\left[\frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial\theta^2}\right]|_{\theta=\theta^{(t)}}, \tag{5.19}$$

where $l_c(\theta)$ is the complete data log-likelihood function. To see this, we have

$$\begin{aligned}
E_{Y|\theta}\left[\frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial\theta^2}\right]|_{\theta=\theta^{(t)}} &= E_{Y|\theta}\left[\frac{\partial^2}{\partial\theta^2}E_{Z|Y,\theta^{(t)}}\left[\log f_\theta(Y,Z)\right]\right]|_{\theta=\theta^{(t)}} \\
&= E_{Y|\theta}\left[E_{Z|Y,\theta^{(t)}}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(Y,Z)\right]\right]|_{\theta=\theta^{(t)}} \\
&= E_{Y|\theta^{(t)}}\left[E_{Z|Y,\theta^{(t)}}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(Y,Z)|_{\theta=\theta^{(t)}}\right]\right] \\
&= E_{Z,Y|\theta^{(t)}}\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(Y,Z)|_{\theta=\theta^{(t)}}\right] \\
&= E_{Z,Y|\theta}\left[\frac{\partial^2 l_c(\theta)}{\partial\theta^2}\right]|_{\theta=\theta^{(t)}}.
\end{aligned}$$

Due to (5.19), Titterington's (1984) algorithm works like a scoring version of the EM gradient algorithm or an approximation to the Method of Scoring algorithm in maximizing $l(\theta)$. In order to implement this algorithm, we find

$$-E[\partial^2 l_c/\partial\beta_j^2] = \sum_{i=1}^{n}\pi_j X_i'\Sigma_i^{-1}X_i,\ 1\le j\le g, \tag{5.20a}$$

$$-E[\partial^2 l_c/\partial\beta_j\partial\beta_k] = 0,\ 1\le j\ne k\le g, \tag{5.20b}$$

$$-E[\partial^2 l_c/\partial\sigma_{kl}\partial\sigma_{st}] = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st},\quad 1\le k\le l\le p\,, 1\le s\le t\le p\,, \tag{5.20c}$$

$$\begin{aligned}
-E[\partial^2 l_c/\partial\sigma_{kl}^c\partial\sigma_{st}^c] &= (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}I_i^{kl}I_i^{st}, \\
&\qquad 1\le k<l\le p\,, 1\le s<t\le p\,,
\end{aligned} \tag{5.20d}$$

$$-E[\partial^2 l_c/\partial\beta_j\partial\sigma_{kl}] = 0,\ 1\le j\le g, 1\le k\le l\le p\,, \tag{5.20e}$$

$$-E[\partial^2 l_c/\partial\beta_j\partial\sigma_{kl}^c] = 0,\ 1\le j\le g, 1\le k<l\le p\,, \tag{5.20f}$$

$$-E[\partial^2 l_c/\partial\sigma_{kl}\partial\sigma_{st}^c] = (1/2)\sum_{i=1}^{n} tr\Sigma_i^{-1}G_{kl}\Sigma_i^{-1}G_{st}I_i^{st},\quad 1\le k\le l\le p\,, 1\le s<t\le p\,. \tag{5.20g}$$

Due to (5.20e) and (5.20f), it is possible now to update β and σ separately. So this algorithm is simpler than the EM gradient algorithm. And since Titterington's (1984) algorithm uses the expected information matrix, one might expect that it is more robust to the choice of starting point than the EM gradient algorithm.

5.3.2 A New Clustering Algorithm

Now suppose the parameter θ can be partitioned as $\theta' = (\theta_1', \theta_2')$ such that in the M-step of the EM algorithm θ_1 has an explicit solution given the value of θ_2. For example, for mixture problems with normal densities the parameters in the mean structures are easier to be updated and usually have closed form solutions, i.e., the weighted least square estimates, given the variance-covariance parameters. In this case, it should be more efficient to update θ_1 with the closed form solution given $\theta_2^{(t)}$, i.e., an ECM step, and update θ_2 with a gradient method.

It is not hard to see that by setting the quantities in (5.10a) equal to zeros and solving, we obtain explicit solutions

$$\beta_j^{(t+1)} = \left[\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} y_i\right), \ j = 1, \ldots, g, \tag{5.21}$$

of the β given $\{\Sigma_i = \Sigma_i^{(t)}\}_{i=1}^n$, which would lead to an ECM update on the β part in the M-step. However, neither the EM gradient algorithm nor Titterington's (1984) algorithm when applied to our mixture problem provide an ECM update for β. This fact provides us an intuition that we probably can improve the convergence properties of both the EM gradient algorithm and Titterington's (1984) algorithm by using this ECM update on the β part in each iteration and updating the σ with a gradient method. In the following, we develop a new algorithm by modifying Titterington's (1984) algorithm and show that the new algorithm produces the ECM update on the β part and updates the σ with a gradient method in each iteration. By doing this, we provide a possible way to improve the convergence properties of iterative algorithms in similar settings.

In calculating $E[\partial^2 l_c(\theta)/\partial\theta^2]$ in Titterington's (1984) algorithm, we use the fact that $E[Z_{ik}] = \pi_k$ for $1 \le k \le g$ and $1 \le i \le n$. And in each iteration of Titterington's (1984) algorithm, π_k is estimated by $\pi_k^{(t)}$ for $1 \le k \le g$. By a careful inspection of (5.10a), (5.21) and (5.20a), we find that it is this fact does not permit the algorithm to yield an explicit solution for β. To see this, we rewrite the M-step of Titterington's (1984) algorithm related

to the β as

$$\begin{aligned}\beta_j^{(t+1)} &= \beta_j^{(t)} + \left[\sum_{i=1}^{n} \pi_j^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} (y_i - X_i \beta_j^{(t)})\right) \\ &= \left[\sum_{i=1}^{n} \pi_j^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} y_i\right) \\ &\quad + \beta_j^{(t)} - \left[\sum_{i=1}^{n} \pi_j^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right]^{-1} \left[\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right] \beta_j^{(t)}\end{aligned}$$

for $j = 1, \ldots, g$.

The change that we make to Titterington's (1984) algorithm is to replace $E_{\theta^{(t)}}[Z_{ik}]$ with its conditional expectation $E_{\theta^{(t)}}[Z_{ik}|Y] = \tau_{ik}^{(t)}$ for $1 \leq k \leq g$ and $1 \leq i \leq n$. Although we currently have little theoretical justification for this modification, encouraging simulation results provided in Section 5.4 suggests this being a modification that leads to faster convergence while providing the same results as the two existing algorithms. Consequently, the quantities in (5.20a) change to

$$\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{-1} X_i,\ 1 \leq j \leq g, \tag{5.22}$$

and everything else in (5.20) remain the same.

To explicitly write down the new algorithm, we substitute (5.22) and (5.20b) - (5.20g) into (5.12) and get

$$\beta_j^{(t+1)} = \beta_j^{(t)} - \left[\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right]^{-1} \left(\frac{\partial Q(\theta|\theta^{(t)})}{\partial \beta_j}\right) |_{\theta=\theta^{(t)}}, 1 \leq j \leq g, \tag{5.23}$$

$$\sigma^{(t+1)} = \sigma^{(t)} - \left[E\left[\frac{\partial^2 Q(\theta|\theta^{(t)})}{\partial \sigma^2}\right]\right]^{-1} \left(\frac{\partial Q(\theta|\theta^{(t)})}{\partial \sigma}\right) |_{\theta=\theta^{(t)}}. \tag{5.24}$$

It is not hard to show that (5.23) and (5.24) can be further simplified as

$$\beta_j^{(t+1)} = \left[\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} X_i\right]^{-1} \left(\sum_{i=1}^{n} \tau_{ij}^{(t)} X_i' \Sigma_i^{(t)-1} y_i\right), 1 \leq j \leq g, \tag{5.25}$$

$$\sigma^{(t+1)} = \left[\sum_{i=1}^{n} \begin{bmatrix} \Phi(\Sigma_i^{(t)})^{-1} & \Phi(\Sigma_i^{(t)})^{-1} J_i \\ J_i' \Phi(\Sigma_i^{(t)})^{-1} & J_i' \Phi(\Sigma_i^{(t)})^{-1} J_i \end{bmatrix}\right]^{-1} \left(\sum_{i=1}^{n} \begin{bmatrix} \Phi(\Sigma_i^{(t)})^{-1} \langle \sum_{j=1}^{g} \tau_{ij}^{(t)} C_{ij}^{(t+1)} \rangle \\ J_i' \Phi(\Sigma_i^{(t)})^{-1} \langle \sum_{j=1}^{g} \tau_{ij}^{(t)} C_{ij}^{(t+1)} \rangle \end{bmatrix}\right), \tag{5.26}$$

in the same way as we did in Section 4.2.1. As a conclusion, the new algorithm uses (5.9) to update the estimate of $\{\pi_j\}_{j=1}^g$ and uses (5.25) and (5.26) to update the estimates of the $\{\beta_j\}_{j=1}^k$ and the σ. Thus, we break the original big problem down to several smaller steps. The steps of (5.25) and (5.26) let us update the β's and the σ's separately so that we do not have to invert a larger matrix. The matrix inversion in (5.25) and (5.26) is guaranteed to exist by the results in Section 4.2.1. Furthermore, step halving can be easily applied to (5.26) to ensure that the new estimates falling in the parameter space. Some successful simulation results have been obtained using this new algorithm.

Upon convergence of the algorithms, the clusters can be formed by checking the estimated subpopulation probabilities for each subject, that is, we assign each of them to the cluster with highest estimated probability. The asymptotic covariance matrix of the parameter estimates can be obtained via the Fisher information matrix, observed information matrix, or Louis' method if we desire less computational burden.

Locally, the EM gradient, Titterington's (1984) and the new algorithm have comparable linear convergence speed, since they are all using one iteration of Newton type algorithms in the M-step. The EM gradient algorithm has a much longer mean time for each iteration because it calculates and inverts a larger matrix. It is important to note that our new algorithm leads to an explicit solution for the β in each iteration. It then increases the likelihood function more than the other two algorithms in the β part of each iteration. As a result, we anticipate that globally the new algorithm converges faster than both the other two. As we show in Section 5.4, this is actually confirmed by our simulations.

In general, for conditional densities of the form $f(y_i|z_{ik}=1) = \phi(y_i; X_i\beta_k, \Sigma_i(\sigma_k))$, this new algorithm will provide a closed form solution for $\{\beta_k\}_{k=1}^g$ and update $\{\sigma_k\}_{k=1}^g$ with a gradient method in each iteration, where $\Sigma_i(\sigma_k)$ represents a constrained covariance matrix for subject i with free parameter σ_k. Our problem, $\sigma_k \equiv \sigma$ for $1 \le k \le g$, is a special case.

Although it is still recommended to choose starting points carefully, it seems that the algorithm is much less sensitive to the starting point for the σ since the covariance parameters are the same across clusters. And completely random starting points for the β seem not to be a bad choice. As least in the following simulations presented in Section 5.4, starting with random clustering indices results in about 95% successful clustering results by which

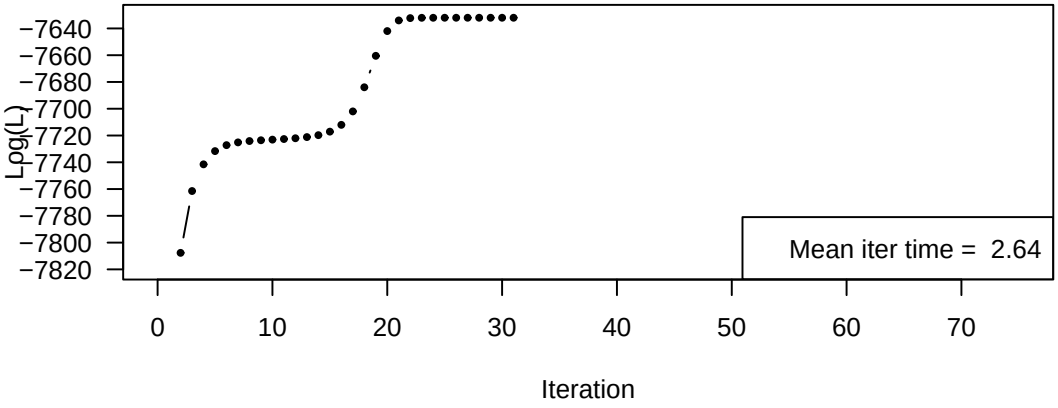
we mean we can cluster more than 95% data points correctly in a single data set. Currently, in the literature there are two existing primary methods for starting point selection for these types of clustering problems. The first one is using a simpler clustering method, such as K-Means or some hierarchical algorithms, to find reasonably good starting clustering indices. This method only works for relative simple problems, especially with no covariates. The second one is to implement multivariate regression models by ignoring the clusters and then simulate the starting points from the asymptotic distributions of the parameters. In this method, the required number of starting points to reach the global optimum should increase as the dimension of the covariates increase. For situations like we have in our problem, where only one or two covariates are associated with the clustering, a graphical visualization of our data, together with several iterations of the regression clustering algorithm introduced in Zhang (2003) might be helpful.

5.4 CLUSTERING SIMULATION RESULTS

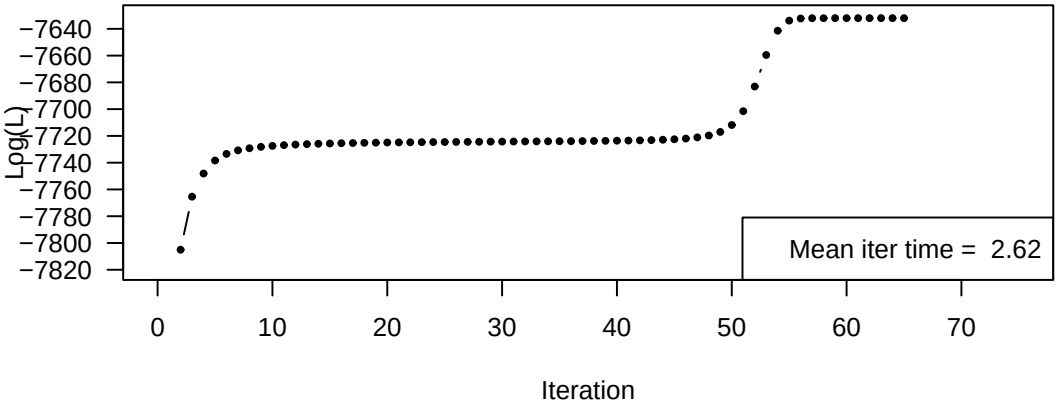
Using the same settings for the covariates as in Section 4.4, we simulate 500 data sets for the clustering analysis. Each of them contains 500 subjects, 250 for each cluster, within which there are 50 subjects for each of the five possible covariance structure as shown in Table 2.1. The two clusters differ only in the parameters for the mean structures, and let $\beta_1' = (\beta_{11}^1, \beta_{12}^1, \beta_{13}^1, \beta_{21}^1, \beta_{22}^1, \beta_{23}^1, \beta_{31}^1, \beta_{32}^1, \beta_{33}^1) = (-100, 2, 50; -50, 2, 50; -50, 1, 50)$ and $\beta_2' = (\beta_{11}^2, \beta_{12}^2, \beta_{13}^2, \beta_{21}^2, \beta_{22}^2, \beta_{23}^2, \beta_{31}^2, \beta_{32}^2, \beta_{33}^2) = (100, -2, 50; 50, 2, 50; 50, -1, 50)$ be those parameters for the two clusters, respectively. In addition, let $\sigma = (\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}, \sigma_{12}^c, \sigma_{13}^c, \sigma_{23}^c) = (1000, 1500, 1000, 400, 500, 600, 200, -100, -200)$. The five possible individual covariance matrices can be obtained the same way as in Table 2.1. Although negative correlations are less possible than positive ones in our motivating data, we use some negative ones here only for illustration. At least, algorithmically it does not matter if we use positive or negative correlations.

We first investigate the convergence speed of the three algorithms in which we are interested, the EM gradient algorithm, Titterington's (1984) algorithm and the new algorithm.

(a) The new algorithm



(b) Titterington's (1984) algorithm



(c) The EM gradient algorithm

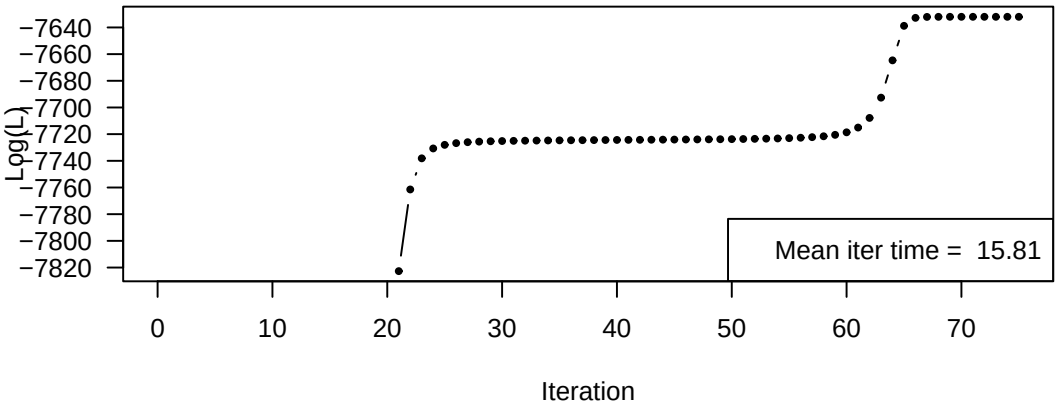


Figure 5.1: Speed of convergence of the clustering algorithms: (a) for the new algorithm; (b) for Titterington's (1984) algorithm; (c) for the EM gradient algorithm. This is an illustration based on one simulated data set.

We randomly pick 10 data sets from the total of 500 simulated data sets. Each of the three algorithms of interest is then implemented on these 10 data sets from the same starting values to find the parameter estimates. For the feasibility of comparison, the three algorithms are stopped according to the same criterion, that is, when the change in the parameter estimates does not exceed a pre-defined limit. We find that when we start the algorithms from near the true parameter values, the three algorithms converge in almost the same number of steps (≈ 11), except that the EM gradient algorithm requires more time (≈ 16 seconds) in each iteration as compared to the other two (< 3 seconds). However, when we start the algorithms far from the true parameter values, the three algorithms behave differently. A typical result from one of the 10 simulated data sets is shown in Figure 5.1. The x-axis represents the number of iterations, while the y-axis represents the value of the log-likelihood function evaluated at the parameter estimates in each iteration. For the feasibility of comparison, some beginning iteration history for all three algorithms with low values (large negative values) of the log-likelihood function is not shown in Figure 5.1. It can be seen that the EM gradient algorithm converges in about 80 steps and costs about 15.8 seconds for each iteration, while Titterington's (1984) algorithm converges in about 70 steps and costs about 2.6 seconds for each iteration. However, our new algorithm only requires about 30 steps to converge, which is a big advantage as compared to the others. And its mean iteration time is comparable to that of Titterington's (1984) algorithm, i.e., about 2.6 seconds, which is significantly lower than that of the EM gradient algorithm. From Figure 5.1, the main feature of the new algorithm is that it requires significantly fewer iterations in finding the region containing a maximum when starting randomly, while its number of steps for subsequent "local refinement" is actually comparable to the two existing algorithms. As we mentioned earlier, the reason for the fast global convergence of the new algorithm is that we used the weighted least square estimator for the β in each iteration.

Since the direct application of both the EM gradient algorithm and Titterington's (1984) algorithm to our simulated data is time consuming, for feasibility in making comparisons we ran all the three algorithms on the 10 simulated data sets and found that the new algorithm gives the same results as the other two. As a result, only the new algorithm is used for the parameter estimation for the rest of the simulated data sets.

As we mentioned early, a careful selection of starting points is still recommended. For our current simulations, we selected two different types of starting points for the purpose of demonstration. One is chosen to be close to the true parameter values, and the other one is by starting the algorithms from randomly generated clustering indices, i.e., a random starting point. In addition, for any single simulated data set, we define the final clustering result to be "successful" if the algorithm clusters more than 95% of its data points correctly. For the 500 simulated data sets, our computations show that by starting from near the true parameter values we get "successful" clustering results on 100% of the simulated data sets, while by starting randomly we get "successful" clustering results on about 95% of the simulated data sets. For the other 5% of the simulated data sets, the algorithm either does not converge (1.4%) or converges (3.6%) to a solution resulting in a random clustering in which the subjects are clustered complete randomly. For those data sets with "successful" clustering results when starting from random clustering indices, we summarize the results of the parameter estimation in Table 5.1 as compared to the true parameter values. It can been seen that the parameter estimation is reasonably accurate as long as the algorithm finds the correct clusters. In fact, these results are surprisingly good. After all, no one will rely on merely one random starting point if one has no information about where to start. For example, we can always start the algorithm with multiple random starting points and pick the solution maximizing the likelihood function as the result. The chance that we can find the correct clustering is high.

Table 5.1: A Summary of the parameter estimates in the clustering simulations

Param.	π	β^1_{11}	β^1_{12}	β^1_{13}	β^1_{21}	β^1_{22}	β^1_{23}	β^1_{31}	β^1_{32}	β^1_{33}
Truth[a]	0.5	-100	2	50	-50	2	50	-50	1	50
Mean[b]	0.50	-100.9	2.01	50.0	-50.5	2.02	49.1	-50.3	1.01	49.8
Std.[c]	0.01	6.47	0.11	4.10	7.67	0.14	5.25	6.33	0.11	4.16
Param.		β^2_{11}	β^2_{12}	β^2_{13}	β^2_{21}	β^2_{22}	β^2_{23}	β^2_{31}	β^2_{32}	β^2_{33}
Truth		100	-2	50	50	2	50	50	-1	50
Mean		99.7	-2.00	49.9	50.0	2.01	48.8	49.9	-1.00	50.1
Std.		6.74	0.12	4.21	8.00	0.15	4.97	6.80	0.12	4.15
Param.		σ_{11}	σ_{22}	σ_{33}	σ_{12}	σ_{13}	σ_{23}	σ^c_{12}	σ^c_{13}	σ^c_{23}
Truth		1000	1500	1000	400	500	600	200	-100	-200
Mean		1006.58	1502.33	971.43	458.32	483.26	544.26	160.24	-97.69	-155.43
Std.		69.85	127.95	64.09	69.85	56.52	69.72	71.45	64.06	78.00

[a]The true parameter values
[b]Means of the simulation estimates
[c]Standard deviations of the simulation estimates

6.0 STRUCTURED CLUSTERING WITH MISSING DATA AND APPLICATIONS TO POST-MORTEM TISSUE DATA

In this chapter, we demonstrate methods for clustering the subjects with schizophrenia into two possible subpopulations in the existence of missing data. Because the actual data are incomplete, the new clustering algorithm developed in Chapter 5 cannot be directly implemented. Directly working on the observed data likelihood function is also intractable due to the complexity of our model and the large degree of missingness. We consider using certain multiple imputation techniques to impute the missing data and then apply the complete data clustering algorithm to the imputed data. Finally, the multiple clustering results are integrated to form one single clustering of the subjects with schizophrenia. The integration incorporates the uncertainty due to the missingness.

6.1 INTRODUCTION

At this point in our research, we consider a limited set of the studies from the 35 possible studies for the application of our methods. We focus on several bio-markers showing significant alterations in subjects with schizophrenia in three individual studies. The first bio-marker is the expression level of a GABA-related gene, GAD_{67}, in the prefrontal cortex (PFC) which has been studied in Hashimoto et al. (2005). It is important because its down-regulation represents some dysfunction in the PFC which contributes to cognitive deficits in subjects with schizophrenia. And it has been shown to be significantly decreased in subjects with schizophrenia. The second selected bio-marker is the somal volume of pyramidal neurons (herein denoted by NISSL) in deep layers 3 of certain PFC region as studied in Pierri

et al. (2001). The somal volume of a neuron is associated with its functioning and pyramidal neurons in deep layers 3 of PFC play an important role in neuronal circuitry. A statistically significant decrease of NISSL in subjects with schizophrenia has also been observed in the original study. The somal size of a subpopulation of large pyramidal neurons (herein denoted by NNFP) also in deep layer 3 of PFC, as studied in Pierri et al. (2003), is selected to be the third important bio-marker, though a statistically nonsignificant decrease in subjects with schizophrenia is reported in the original paper. In the original studies, GAD_{67} has measurements on 27 pairs of subjects with schizophrenia and their corresponding controls, NISSL has measurements on 28 pairs, and NNFP has measurements on 13 pairs. When combined together, the total number of unique pairs is 41. Due to certain technical reasons, 4 pairs of subjects are excluded from our research. So the final number of usable pairs of subjects is 37. The data are shown in Table 6.1 with blanks represent missing data. The first column in Table 6.1 contains the internal artificial id numbers of the subjects with schizophrenia. And again, the last column in Table 6.1 represents the different covariance structures due to the differing controls as illustrated in Table 2.1.

The selection of NNFP is just for the purpose of providing a demonstration data set, and is not biologically attractive. This is due to the facts that in the study of Pierri et al. (2003) it was noted that NNFP measured the somal size of large pyramidal neurons which were a subset of the pyramidal neurons measured with NISSL, and it was shown in that study that the alteration in NNFP was not statistically significant. Furthermore, in that study the staining technique used in obtaining NNFP was shown to be confounded with the actual neuron size. In fact, in any application of our methodologies to a large post-mortem tissue data set, we must recognize the exploratory nature of our procedure and treat the final clustering result with great caution. A review of the clustering results by experienced neurobiologists and clinicians is necessary to determine its practical meaning. The purpose of this chapter is to provide a demonstration of the feasibility of our clustering approaches when the bio-markers are pre-selected.

Table 6.1: A combined Data of GAD_{67}, NISSL and NNFP

			Pairwise differences of			
Sch. ID	Age	Gender	GAD_{67}	NISSL	NNFP	Case
317	48	M	-91.203	0.19369	0.192	1
398	41	F		-0.09153	-0.42389	1
131	62	M		-0.29346	-0.53035	4
185	64	M		-0.47223	-0.25673	4
207	72	M		0.13115	-0.20188	4
234	51	M		-0.09278	-0.16416	4
236	69	M		0.16462		4
322	40	M		-0.33697	0.15357	4
333	66	F		0.01563		4
341	47	F		0.31749	0.12091	4
377	52	M		0.00584	0.39115	4
408	46	M		-0.15809		4
422	54	M		0.0305	0.13179	4
428	67	F		0.10574	0.14986	4
466	48	M		-0.06563		4
533	40	M	-19.433			4
539	50	M	-26.399			4
547	27	M	-72.513	-0.29217	-0.07861	4
559	61	F		-0.29772	0.04952	4
566	63	M	-28.187	-0.13604		4
581	46	M	-48.111	-0.189		4
587	38	F	-9.530	-0.15396		4
597	46	F	-33.295	-0.46205	-0.29125	4
621	83	M	7.760	0.13461		4
622	58	M	-55.412	-0.41927		4
656	47	F	11.078	-0.11994		4
665	59	M	-11.696			4
722	45	M	14.558			4
781	52	M	-52.737			4
787	27	M	-1.6649			4
802	63	F	-9.472			4
829	25	M	-49.284			4
878	33	M	-0.25724			4
904	33	M	-36.465			4
917	71	F	-13.178			4
930	47	M	-31.795			4
933	44	M	-55.398			4

6.2 MULTIPLE IMPUTATION APPROACHES

From Table 6.1, we see that more than 45% of the data are missing. As introduced in Chapter 2, we assume the data are missing completely at random. Though we are able to write down the observed likelihood function, it is intractable to directly maximize it due to the complexity of our model and the large degree of missingness. If the degree of missingness is relatively small and the clusters are well defined, the new complete data clustering algorithm we develop in Chapter 5 can be modified accordingly to account for the missing data and maximize the observed likelihood function. Similarly as in the missing data EM algorithm, this modification only requires calculation of the conditional expectations of the missing data given the observed ones and the current parameter estimates in the E-step. However, given the large degree of missingness and the high variability of the data, the observed likelihood function is highly irregular and has lots of modes so that it is hard to find its global maximum. As a result, directly modifying the new complete data clustering algorithm is also not preferred.

The way we choose to analyze this data is to multiply impute the missing data, analyze the imputed data with the new complete data clustering algorithm introduced in Chapter 5, and integrate the multiple clustering results based on each of the multiple imputations to form one single clustering of the subjects with schizophrenia. The last step of our integration approach incorporates the uncertainty due to the missingness. In multiple imputation of the missing data, Markov Chain Monte Carlo (MCMC) methods are usually the first choice, especially for complicated parametric models. However, to our knowledge, the Bayesian models concerning structured mixture models are not yet in the literature. As a result, we use a two-step regression method to impute the missing data in our research. The basics of imputing using regression methods are discussed in Little and Rubin (2002). In the first step, linear regression models

$$GAD_{67} = \beta_{10} + \beta_{11} \times Age + \beta_{12} \times Gender + \epsilon_1, \tag{6.1a}$$

$$NISSL = \beta_{20} + \beta_{21} \times Age + \beta_{22} \times Gender + \epsilon_2, \tag{6.1b}$$

$$NNFP = \beta_{30} + \beta_{31} \times Age + \beta_{32} \times Gender + \epsilon_3, \tag{6.1c}$$

where $\epsilon_i \sim N(0, \sigma^2 I_i)$ independently for $i = 1, 2, 3$, are fitted to the observed data, since all values of the covariates are available. As there might be correlations among GAD_{67}, NISSL and NNFP, in the second step regression models

$$\epsilon_1 = \xi_{10} + \xi_{11} \times \epsilon_2 + u_1, \tag{6.2a}$$

$$\epsilon_2 = \xi_{20} + \xi_{21} \times \epsilon_1 + u_2, \tag{6.2b}$$

$$\epsilon_3 = \xi_{30} + \xi_{31} \times \epsilon_1 + \xi_{32} \times \epsilon_2 + u_3, \tag{6.2c}$$

where $u_i \sim N(0, \tau_i^2 I_i)$ independently for $i = 1, 2, 3$, are fitted to the residuals obtained in the first step for the complete cases. Equations (6.2a) and (6.2b) use the complete cases of GAD_{67} and NISSL in estimating ξ_{10}, ξ_{11}, ξ_{20} and ξ_{21}, while equation (6.2c) uses the complete cases of GAD_{67}, NISSL and NNFP, with the previous imputed residuals of GAD_{67} and NNFP from (6.2a) and (6.2b) included, to estimate ξ_{30}, ξ_{31} and ξ_{32}. The residuals of NNFP are not included in (6.2a) and (6.2b) because of the missing pattern of our data. As can be seen from Table 6.1, there are only 3 complete cases on all three dependent variables and the subjects with NNFP are only a subset of those with NISSL.

Let $\beta_i = (\beta_{i0}, \beta_{i1}, \beta_{i2})'$ and X be the common design matrix in regression models (6.1). In order to reflect the estimation uncertainty in imputing the missing data, $\tilde{\beta}_i$ are sampled from $N(\hat{\beta}_i, \hat{\sigma}_i^2 (X'X)^{-1})$ for $i = 1, 2, 3$, where $\hat{\beta}_i$ and $\hat{\sigma}_i^2$ are the least square estimates. $\tilde{\xi}_{ik}$'s are sampled in the same fashion according to regression models (6.2). And $\tilde{u}_i$ are sampled from $N(0, \hat{\tau}_i I_i)$ for $i = 1, 2, 3$, where $\hat{\tau}_i$ are the least square estimates of τ_i. The missing data are then imputed using

$$GAD_{67} = \tilde{\beta}_{10} + \tilde{\beta}_{11} \times Age + \tilde{\beta}_{12} \times Gender + \tilde{\epsilon}_1, \tag{6.3a}$$

$$NISSL = \tilde{\beta}_{20} + \tilde{\beta}_{21} \times Age + \tilde{\beta}_{22} \times Gender + \tilde{\epsilon}_2, \tag{6.3b}$$

$$NNFP = \tilde{\beta}_{30} + \tilde{\beta}_{31} \times Age + \tilde{\beta}_{32} \times Gender + \tilde{\epsilon}_3, \tag{6.3c}$$

where

$$\tilde{\epsilon}_1 = \tilde{\xi}_{10} + \tilde{\xi}_{11} \times \epsilon_2 + \tilde{u}_1, \tag{6.4a}$$

$$\tilde{\epsilon}_2 = \tilde{\xi}_{20} + \tilde{\xi}_{21} \times \epsilon_1 + \tilde{u}_2, \tag{6.4b}$$

$$\tilde{\epsilon}_3 = \tilde{\xi}_{30} + \tilde{\xi}_{31} \times \epsilon_1 + \tilde{\xi}_{32} \times \epsilon_2 + \tilde{u}_3. \tag{6.4c}$$

However, we do understand that this two-step regression model is not the right model to analyze our mixture problem. First, we ignore the effect of differing controls in conducting the multiple imputations. Furthermore, since we treat the subjects as if they are from the same population, the imputations tends to make the subjects look more alike than they should be. We hope this effect of assimilation will not cover the possible interesting feature of the real data. Nevertheless, based on the above two-step regression model, 200 imputations are obtained for the purpose of the clustering analysis.

6.3 INTEGRATING MULTIPLE CLUSTERING RESULTS

After obtaining the multiple imputations, we apply the new complete data clustering algorithm introduced in Chapter 5 to every imputed data set. The algorithm converges in 400 iterations on 192 out of the 200 imputed data sets, and the corresponding results are then used. The subjects with schizophrenia are clustered according to the posterior probabilities that the subjects belong to mixture class k, that is,

$$P(z_{ik} = 1) = \frac{\hat{\pi}_k \phi(y_i; X_i\hat{\beta}_k, \hat{\Sigma}_i)}{\sum_{h=1}^{g} \hat{\pi}_h \phi(y_i; X_i\hat{\beta}_h, \hat{\Sigma}_i)} \tag{6.5}$$

for $i = 1, \ldots, n$ and $k = 1, 2$, where ϕ represents the normal density function. A subject is clustered into subpopulation 1 if $P(z_{i1} = 1) > P(z_{i2} = 1)$, and vice versa. If equality occurs, the subjects can be clustered into either subpopulation. However, it is problematic to directly compare the clustering results from the multiple imputations, since the order of the subpopulations may not be preserved for all of the multiple imputations, i.e., the subpopulation 1 in the clustering results of two imputations might be different. In fact, what is comparable is the pairwise relationships, that is, we can compare whether a pair of subjects with schizophrenia is clustered into the same subpopulation or not for two different imputations. See Larsen (2005) for a complete discussion.

For our data analysis, we focus on the total of 666 pairs resulting from the 37 subjects with schizophrenia used in the research, that is, $\binom{37}{2} = 666$. For each imputation, we record whether or not a particular pair is in the same subpopulation. A code "1" is given to a pair if

they are in the same group, and "0" otherwise. We then sum over the multiple imputations of the codes for each pair. We denote the resulting summations as S_{ij} for $1 \leq i < j \leq 37$. It is obvious that $0 \leq S_{ij} \leq 192$ for $1 \leq i < j \leq 37$. For a particular pair (i, j), a large S_{ij} gives an indicator that the pair of subjects are similar, and vice versa. And the randomness of each S_{ij} is from the multiple imputations, and the multiple imputations are conducted in a way so that they are independent from each other. So it is not hard to see that for each pair (i, j) for $1 \leq i < j \leq 37$, we have

$$S_{ij} \sim \text{Binomial}(192, p_{ij}) \tag{6.6}$$

where p_{ij} is the unknown probability that the pair (i, j) belongs to the same group. As a result, a hypothesis test can be conducted based on each S_{ij} to test

$$H_0 : p_{ij} = \frac{1}{2} \quad \text{vs.} \quad H_a : p_{ij} \neq \frac{1}{2}. \tag{6.7}$$

Accepting the null hypothesis provides no evidence to cluster the pair, while rejecting the null hypothesis suggests the existence of possible clusters. Figure 6.1 is the histogram of the $\{S_{ij}\}$ with a 95% acceptance interval based on the normal approximation, that is,

$$(0.5 - Z_{97.5\%}\sqrt{\frac{0.5(1-0.5)}{192}}, 0.5 + Z_{97.5\%}\sqrt{\frac{0.5(1-0.5)}{192}}) \times 192 = (82.4, 109.6).$$

As we can see, the histogram is not symmetric with a fair amount of observations on the right-hand side out of the right acceptance boundary, which means there are subjects with schizophrenia who tend to be clustered together. This provides evidence of existing possible clusters.

Given the hint from the histogram, we continue our investigation by creating hierarchical clusterings of the subjects with schizophrenia using $d_{ij} = 1 - S_{ij}/192$ for $1 \leq i < j \leq 37$ as a distance metric. Here, $0 \leq d_{ij} \leq 1$ for $1 \leq i < j \leq 37$ measure how likely that subject pairs (i, j) are in different clusters. A large value of d_{ij} means the subject pair (i, j) is apart, whereas a small value of d_{ij} shows that the subject pair (i, j) is similar. For example, if $S_{ij} = 192$ then $d_{ij} = 0$ which means subject pair (i, j) is always clustered together in the 192 imputations. As a result, we would like to conclude that the subject pair (i, j) should be in the same subpopulation. On the other hand, if $S_{ij} = 0$ such that $d_{ij} = 1$,

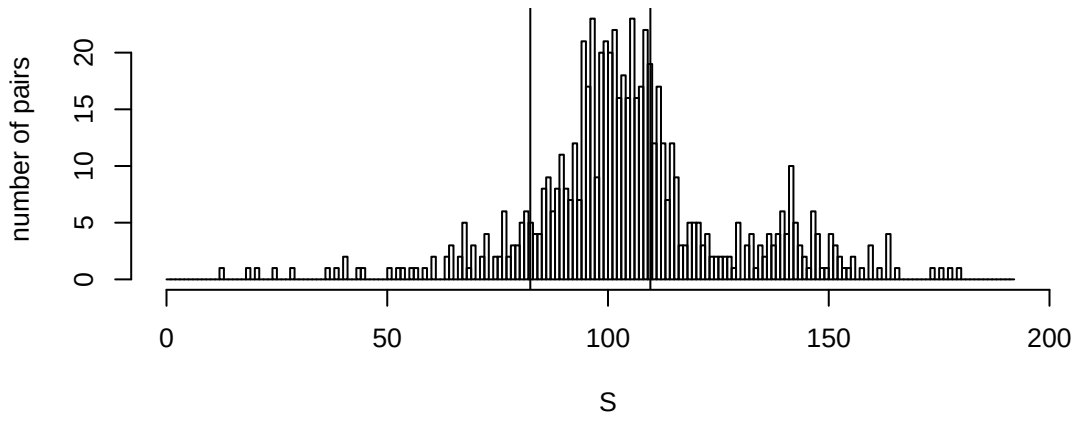
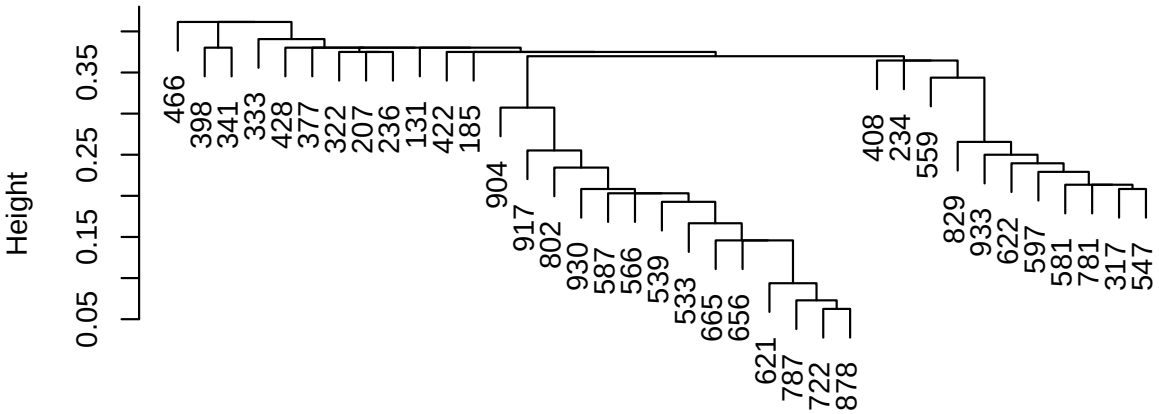


Figure 6.1: The histogram of the $\{S_{ij}\}$ with 95% acceptance interval

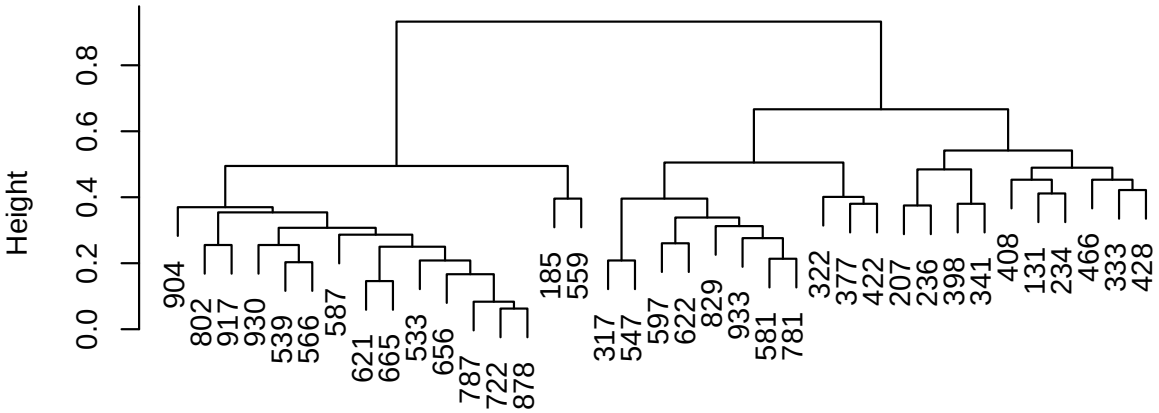
we would then conclude that subject pair (i, j) is from two different subpopulations. The clusterings are obtained using R package "hclust" with different agglomeration methods, i.e., single linkage, complete linkage and average linkage. The reason that we use multiple agglomeration methods is to check the consistency of the clustering of the subjects with schizophrenia with respect to the different distance methods. Single linkage defines the distance between two clusters to be the shortest distance between the subjects in the two clusters, complete linkage uses the largest distance between the subjects in the two clusters, and average linkage applies the average distance between the subjects in the two clusters. Figure 6.2 contains the dendrograms of the clusterings. All three dendrograms suggest a possible existence of two clusters with some subjects' memberships being not clear. For example, in using single linkage, two possible clusters are shown on the right-hand side of the dendrogram while those subjects on the left-hand side have no clear clustering grouping. However, we can identify two possible groups of subjects with schizophrenia as (904, 802, 917, 930, 539, 566, 587, 621, 665, 533, 656, 787, 722, 878, 185, 559) and (317, 547, 597, 622, 829, 933, 581, 781, 322, 377, 422, 207, 236, 398, 341, 408, 131, 234, 466, 333, 428) with the complete and average linkages. Possibly due to the large variability and the great degree of missingness of our data, the clustering uncertainty is high as shown by the large distance among the subjects within clusters in all three dendrograms.

It will be helpful to see what causes the difference between the clusters in interpreting

(a) Single linkage



(b) Complete linkage



(c) Average linkage

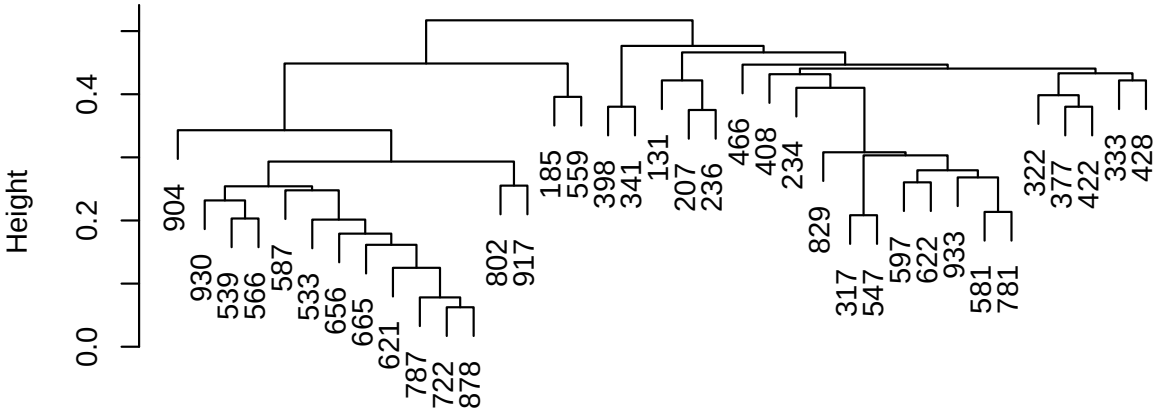


Figure 6.2: The dendrograms of clusterings with different agglomeration methods: (a) dendrogram using single linkage; (b) dendrogram using complete linkage; (c) dendrogram using average linkage

the results. To us, it seems unrealistic to directly compare the parameter estimates for the two clusters across the multiple imputations. The reasons include (i) the three dependent variables are on different scales; (ii) the clusters change from imputation to imputation; and (iii) the ordering of the clusters is not preserved in the multiple imputations. Instead, we try to figure out the differences between the two cluster based on only the observed data. However, our capability of identifying the difference is limited by the degree of missingness.

In Figure 6.3, we create box plots of the three dependent variables for both the two clusters. The box plots represent the overall difference between the two clusters. By comparing the box plots for the two clusters on each dependent variable, we find that the two clusters have a significant overall difference on GAD_{67} as shown in box plots 6.3 - (a). However, there seem to be no overall differences on NISSL and NNFP between the two clusters as shown in box plots 6.3 - (b) and (c).

In addition, in order to check whether age and gender have significant effects on defining the two clusters, we create scatter plots of GAD_{67}, NISSL and NNFP versus age and gender as shown in Figure 6.4 and Figure 6.5. In scatter plot 6.4 - (a), age seems to have different intercepts on GAD_{67} for the two clusters while there is no evidence to conclude the slopes are different. This result is consistent with that shown in box plots 6.3 - (a). However, there is no definite conclusion can be made on the scatter plots 6.4 - (b) and (c). Again, in Figure 6.5 - 4(a), similar differences on GAD_{67} between the two clusters for male and female are identified, while NISSL and NNFP exhibit no difference between the two clusters for both gender.

As a conclusion, in this example the two clusters differ mainly on the diagnostic effect of GAD_{67} while age and gender are not significant factors in defining the two clusters. Moreover, NISSL and NNFP seem not closely related to this clustering of the subjects with schizophrenia, because the two clusters exhibit no significant difference on them. However, this conclusion is highly limited by the degree of missingness of our data, so that it needs to be treated with great caution and subject to further examination perhaps in light of existing clinical information.

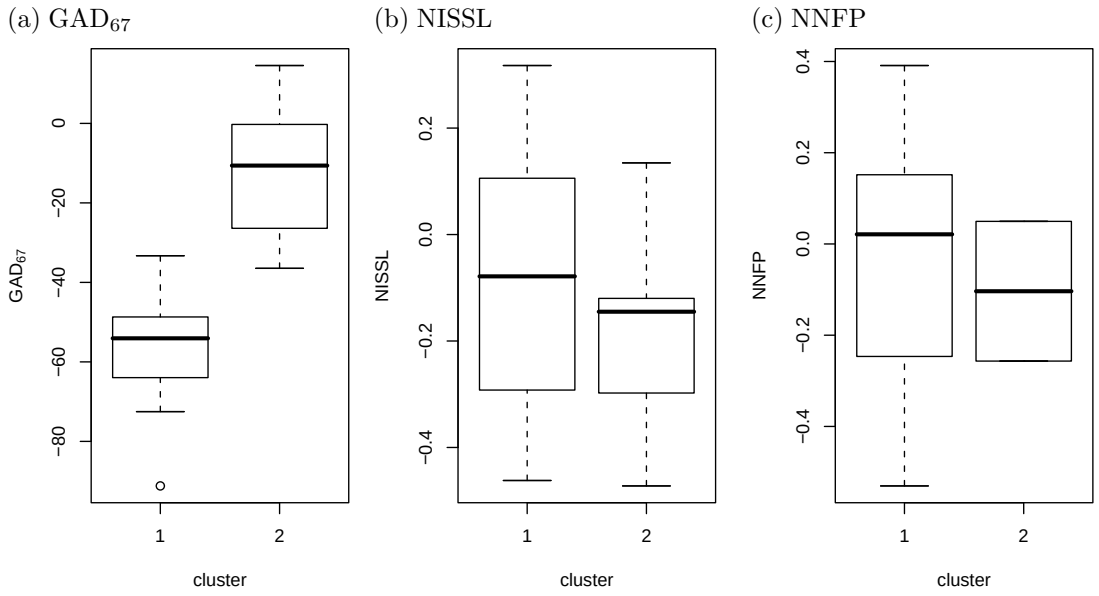


Figure 6.3: Boxplots of GAD_{67}, NISSL and NNFP for the two clusters for the available cases: (a) box plots of GAD_{67}; (b) box plots of NISSL; (c) box plots of NNFP.

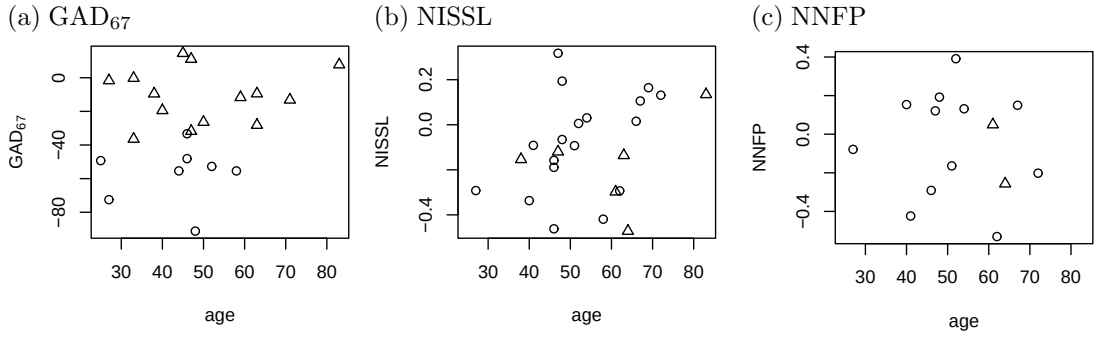


Figure 6.4: Scatter plots of GAD_{67}, NISSL and NNFP vs. age for the two clusters for the available cases: (a) scatter plots of GAD_{67} vs. age; (b) scatter plots of NISSL vs. age; (c) scatter plots of NNFP vs. age.

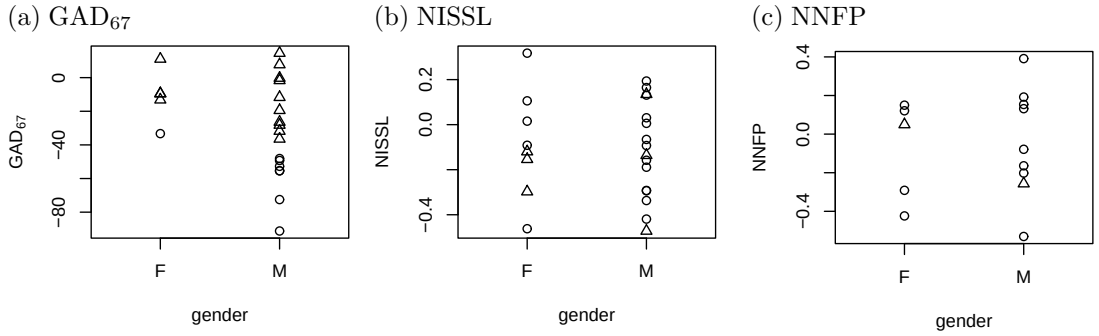


Figure 6.5: Scatter plots of GAD_{67}, NISSL and NNFP vs. gender for the two clusters for the available cases: (a) scatter plots of GAD_{67} vs. gender; (b) scatter plots of NISSL vs. gender; (c) scatter plots of NNFP vs. gender.

7.0 CONCLUSIONS

In this dissertation, we explore three major research steps with the goal of clustering subjects with schizophrenia into possible subpopulations by using the post-mortem tissue data obtained in the Conte Center for the Neuroscience of Mental Disorders in the Department of Psychiaty at the University of Pittsburgh. While these three steps are critical for the main goal of this research, each step is also of interest in its own right.

As an initial step in the research, we develop multivariate normal models with structured means and covariance matrices assuming no clusters and no missing data. The mean structures result from the inclusion of covariates, while the covariance structures represent the existence of differing control subjects. Several algorithms are considered to find the maximum of the likelihood function. A one-iteration estimator of the Simplified Method of Scoring algorithm starting from consistent initial values is used and shown to be asymptotically equivalent to the MLE. Simulations are conducted to verify the key asymptotic results. In general, it is shown that for large sample sizes there is no big advantage of continuing the Simplified Method of Scoring algorithm for more than one step if the starting point is constant, while for small sample sizes the advantage of finding the MLE is significant. In addition, Wald testing is suggested based on the asymptotic distributions of the parameter estimates.

In the second step, we treat the data as from a mixture of two multivariate normal distributions with patterned mean and covariance structures. We still assume no missing data. Several algorithms, including the EM gradient algorithm and Titterington's (1984) algorithm, are considered for model fitting. Though all these algorithms are applicable to our problem, we show that a new clustering algorithm we develop provides the same clustering results as others in a relatively faster manner. Simulations are conducted to both

compare the convergence speed of the algorithms and evaluate the clustering performance of the new algorithm.

For the actual data obtained from multiple post-mortem tissue studies in the Center, there is a large degree of missingness. As a result, the new clustering algorithm we develop in Chapter 5 cannot be directly applied. Directly maximizing the observed likelihood function is also intractable due to the complexity of our data and the degree of missingness. Instead, we choose to impute the missing data with certain regression method. This imputation model is not optimal for our problem, and we hope that the interesting feature of the real data will not be covered by the imputation method. After obtaining the multiple imputations, each imputed data set is analyzed with the new complete data clustering algorithm introduced in Chapter 5. The clustering results from the multiple imputations are then integrated to form a single clustering of the subjects with schizophrenia. The integration incorporates the uncertainty due to the missingess. The result suggests the existence of two possible clusters of the subjects with schizophrenia. Finally, some graphical summaries are obtained based on the observed data to understand the differences between the two clusters. In our research, the actual selection of the bio-markers might not be biologically attractive and is used only for the feasibility of demonstration. And our capability of identifying the differences between the two clusters is limited by the degree of missingness. Nevertheless, our results and applications together show that our methodologies are applicable in clustering the subjects with schizophrenia with data from post-mortem tissue studies in the Center and other similar settings.

There are a number of future research directions we plan to explore based upon the results we have obtained so far. As we mentioned, the multiple imputation technique we used in our research might be problematic. The two-step regression model ignores the covariance structures and the possible clusters of the subjects with schizophrenia, so that it might make the subjects more similar than they should be and cover the interesting features of the data. As a result, we would like to develop multiple imputation techniques more suitable for our settings in the future. To our knowledge, some other researchers are now working on developing Bayesian models for the model with structured means and covariances in a one population setting. The corresponding research for our clustering problem with structured

means and covariance matrices which requires more efforts would be based on any such new research.

Also as we mentioned earlier, different choices of bio-markers most probably will produce different clustering results. In the future, we would like to investigate some bi-clustering techniques such that we can cluster the subjects and the bio-markers simultaneously to show a bio-marker related clustering pattern of the subjects with schizophrenia. Since there are a tremendous amount of information available on all kinds of different bio-markers, we don't want to necessarily limit our searching for clusters of the subjects with schizophrenia on some pre-selected bio-markers. However, due to the special structure of our data, e.g., structured means and covariances and missing data, the bi-clustering is noticeably difficult and requires a tremendous amount of research.

In addition, in clustering the subjects with schizophrenia, we assumed that the unconditional probability for a subject to belong to a cluster was a constant. While this assumption is intuitively attractable and provides us a relatively simple model, it is rather restricted. Instead, the unconditional probabilities might dependent on some known characteristics of the subjects. For example, the mixture of experts models define the unconditional probabilities to be functions of the known covariates. For our problem, we already assumed that the clusters could be defined on the effect of covariate age. However, it is also possible that the subdivision of the subjects with schizophrenia shows different patterns in different age groups. In this case, it would then be necessary to assume the unconditional probabilities depending on the covariate age.

Finally, we are also interested in the regression clustering algorithms as introduced in Zhang (2003). The intriguing features include that (i) only multiple linear regression analyses are implemented in these algorithms; (ii) subjects were moved to the nearest regression subset based only on the regression results to form hard boundary clusters; and (iii) one simple target function is evaluated in each iteration. As a result, they are straightforward and possibly faster, so that they are suitable for exploratory studies such as our clustering problem. Again, the mean and covariance structures and the missing data in our problem would cause great difficulties in developing corresponding statistical models.

APPENDIX

USEFUL DEFINITIONS

Definition A.0.1 (Szatrowski (1980)). Let $\mathbf{A}$ be a symmetric $p \times p$ matrix. $\langle \mathbf{A} \rangle$ is defined to be a column vector consisting of the upper triangle of elements of $\mathbf{A}$, i.e.,

$$\langle \mathbf{A} \rangle = (a_{11}, a_{22}, \cdots, a_{pp}, a_{12}, a_{13}, \cdots, a_{1p}, a_{23}, \cdots, a_{p-1,p})'.$$

Definition A.0.2 (Anderson (1969)). Define $\mathbf{\Phi}$ as the $p(p+1)/2 \times p(p+1)/2$ symmetric matrix with elements $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{\Sigma}) = (\phi_{ij,\,kl}) = (\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$, $i \leq j, k \leq l$. The notation $\phi_{ij,\,kl}$ represents the element of $\mathbf{\Phi}$ with row in the same position as the element a_{ij} in $\langle \mathbf{A} \rangle$ where $\mathbf{A}$ is a $p \times p$ symmetric matrix and column in the same position as a_{kl} in $\langle \mathbf{A} \rangle'$

Theorem A.0.3 (Szatrowski (1980)). *If* $\mathbf{E}$ *and* $\mathbf{F}$ *are* $p \times p$ *symmetric matrix, then*

$$\langle \mathbf{E} \rangle' \mathbf{\Phi}^{-1}(\mathbf{\Sigma}) \langle \mathbf{F} \rangle = \frac{\mathbf{1}}{\mathbf{2}} \mathbf{tr} \mathbf{\Sigma}^{-1} \mathbf{E} \mathbf{\Sigma}^{-1} \mathbf{F}. \tag{.1}$$

BIBLIOGRAPHY

Anderson, T. W. (1969), "Statistical inference for covariance matrices with linear structure," in *Proceedings of the Second International Symposium on Multivariate Analysis*, ed. Krishnaiah, P. R., New York: Academic Press, pp. 55–66.

— (1970), "Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices," in *Probability and Statistics*, eds. Bose, R. C., Chakravarti, I. M., Mahalanobis, P. C., Rao, C. R., and Smith, J. K. C., Chapel Hill: University of North Carolina Press, pp. 1–24.

— (1973), "Asymptotic efficient estimation of covariance matrices with linear structure," *The Annals of Statistics*, 1, 135–141.

Arminger, G., Stein, P., and Wittenberg, J. (1999), "Mixtures of conditional mean- and covariance-structure models," *Psychometrika*, 64, 475–494.

Berndt, E. B., Hall, B., Hall, R., and Hausman, J. A. (1974), "Estimation and inference in nonlinear structural models," *Ann. Econ. Soc. Meas.*, 3, 653–665.

Brasford, K. E., Greenway, D. R., McLachlan, G. J., and Peel, D. (1997), "Standard erros of fitted component means of normal mixtures," *Computational Statistics*, 12, 1–17.

Demidenko, E. and Spiegelman, D. (1997), "A paradox: more Berkson measurement error can lead to more efficient estimates," *Communication in Statistics: Theory and Methods*, 26, 1649–1675.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39, 1–38.

DeSarbo, W. S. and Corn, L. W. (1988), "A maximum likelihood methodology for clusterwise linear regression," *Journal of Classification*, 5, 249–282.

Efron, B. and Hinkley, D. V. (1978), "Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion)," *Biometrika*, 65, 457–478.

Graybill, F. A. and Hultquist, R. A. (1961), "Theorems concerning Eisenhart's Model II," *Ann. Math. Satist.*, 32, 261–269.

Harville, D. A. (1977), "Maximum likelihood approaches to variance component estimation and to related problems (with discussion)," *Journal of American Statistical Association*, 72, 320–340.

Hashimoto, T., Bergen, S. E., Nguyen, Q. L., Xu, B., Monteggia, L. M., Pierri, J. N., Sun, Z., Sampson, A. R., and Lewis, D. A. (2005), "Relationship of brain-derived neurotrophic factor and its receptor TrkB to altered inhibitory prefrontal circuitry in schizophrenia," *The Journal of Neuroscience*, 25, 372–383.

Hashimoto, T., Volk, D. W., Eggan, S. M., Mirnics, K., Pierri, J. N., Sun, Z., Sampson, A. R., and Lewis, D. A. (2003), "Gene expression deficits in a subclass of GABA neurons in the prefontal cortex of subjects with schizophrenia," *The Journal of Neuroscience*, 23, 6315–6326.

Herbach, L. H. (1959), "Properties of model II-type analysis of variance tests, A: Optimum nature of the F-test for model II in the balanced case," *Ann. Math. Satist.*, 30, 939–959.

Jennrich, R. I. and Schluchter, M. D. (1986), "Unbalanced repeated-measures models with structured covariance matrices," *Biometrics*, 42, 805–820.

Jones, P. N. and McLachlan, G. J. (1992), "Fitting finite mixture models in a regression context," *Australian Journal of Statistics*, 32, 233–240.

Knable, M. B., Bacrci, B. M., Barko, J. J., Webster, M. J., and Torrey, E. F. (2002), "Molecular abnormalities in the major psychiatric illnesses: Classification and Regression Tree (CRT) analysis of post-mortem prefrontal markers," *Molecular Psychiatry*, 7, 392–404.

Knable, M. B., Torrey, E. F., Webster, M. J., and Bartko, J. J. (2001), "Multivariate analysis of prefrontal cortical data from the Stanley Foundation Neuropathology Consortium," *Brain Research Bulletin*, 55, 651–659.

Konopaske, G. T., Sweet, R. A., Wu, Q., Sampson, A. R., and Lewis, D. A. (2005), "Regional specificity of chandelier neuron axon terminal alterations in schizophrenia," Accepted for publish in Neuroscience.

Laird, N. M. and Ware, J. H. (1982), "Random-effect models for longitudinal data," *Biometrics*, 38, 963–974.

Lange, K. (1995), "A gradient algorithm locally equivalent to the EM algorithm," *Journal of the Royal Statistical Society*, 57, 425–437.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization transfer using surrogate objective functions," *Journal of Compuational and Graphical Statistics*, 9, 1–59.

Larsen, M. D. (2005), "Multiple imputation for cluster analysis," in *Proceedings of the INTERFACE*, Interface Foundation of North America.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, New York: Springer-Verlag.

Little, R. J. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, New Jersey: John Wiley & Sons, Inc., 2nd ed.

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society B*, 44, 226–233.

McLachlan, G. J. and Krishnan, T. (1977), *The EM Algorithm and Extensions*, New York: Wiley.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Pierri, J. N., Volk, C. L. E., Auh, S., Sampson, A., and Lewis, D. (2001), "Decreased somal size of deep layer 3 pyramidal neurons in the prefrontal cortex of subjects with schizophrenia," *Archives of General Psychiatry*, 58, 466–473.

— (2003), "Somal size of prefrontal cortical pyramidal neurons in schizophrenia: differential effects across neuronal subpopulations," *Biological Psychiatry*, 54, 111–120.

Rubin, D. B. and Szatrowski, T. H. (1982), "Finding maximum likelihood estimates of patterned covariance matrices by EM algorithm," *Biometrika*, 69, 657–660.

Srivastava, J. N. (1966), "On testing hypotheses regarding a class of covariance structures," *Psychometrika*, 31, 147–164.

Szatrowski, T. H. (1979), "Asymptotic nonnull distributions for likelihood ratio statistics in the multivariate normal patterned mean and covariance matrix testing problem," *The Annals of Statistics*, 7, 823–837.

— (1980), "Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances," *The Annals of Statistics*, 8, 802–810.

— (1983), "Missing data in the one-population multivariate normal patterned mean and covariance matrix testing and estimation problem," *The Annals of Statistics*, 11, 947–958.

Titterington, D. M. (1984), "Recursive parameter estimation using incomplete data," *Journal of the Royal Statistical Society*, 46, 257–267.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Ware, J. H. (1985), "Linear models for the analysis of longitudinal studies," *American Statistician*, 39, 95–101.

Zhang, B. (2003), "Regression Clustering," in *Proceedings of Third IEEE International Conference on Data Mining (ICDM'03)*, Los Alamitos, CA, USA: IEEE Computer Society, vol. 00, p. 451.