REPEATED MEASURES MIXTURE MODELING WITH APPLICATIONS TO NEUROSCIENCE

by

Zhuoxin Sun

M.S., Ocean University of Qingdao, 1996

B.S., Shandong Normal University, 1993

Submitted to the Graduate Faculty of

the Faculty of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Zhuoxin Sun

It was defended on

January 31th 2005

and approved by

Ori Rosen, Department of Statistics (co-advisor)

Allan R. Sampson, Department of Statistics (co-advisor)

Henry W. Block, Department of Statistics

Robert E. Kass, Department of Statistics, CMU

Robert A. Sweet, Department of Psychiatry

Dissertation Advisors: Ori Rosen, Department of Statistics (co-advisor),

Allan R. Sampson, Department of Statistics (co-advisor)

REPEATED MEASURES MIXTURE MODELING WITH APPLICATIONS TO NEUROSCIENCE

Zhuoxin Sun, PhD

University of Pittsburgh, 2005

In some neurological postmortem brain tissue studies, repeated measures are observed. These observations are taken on the same experimental subject and are therefore correlated within the subject. Furthermore, each observation can be viewed as coming from one of a pre-specified number of populations where each population corresponds to a possible type of neurons.

In this dissertation, we propose several mixture models with two components to model such repeated data. In the first model, we include subject-specific random effects in the component distributions to account for the within-subject correlation present in the data. The mixture components are generalized linear models with random effects, while the mixing proportions are governed by a logistic regression. In the second proposed model, the mixture components are generalized linear models, while the component-indicator variables are modeled by a multivariate Bernoulli distribution that depends on covariates. The within-subject observations are taken to be correlated through the latent component indicator random variables. As a special case of the second model, we focus on multivariate Bernoulli mixtures of normals, where the component-indicator variables are modeled by logistic regressions with random effects, and the mixture components are linear regressions. The third proposed model combines the first and second models, so that the within-subject correlation is built into the model not only through the component distributions, but also through the latent component indicator variables. The focus again is on a special case of the third model, where the mixture components are linear regressions with random effects while the mixing

proportions are logistic regressions with another group of random effects. For each model, model fitting procedures, based on MCMC methods for sampling from the posterior distribution of the parameters, are developed. The second and third model are used to compare schizophrenic and control subjects with regard to the somal volumes of deep layer 3 pyramidal cells in the auditory association cortex. As a preliminary analysis, we start by employing classic mixture models and mixtures-of-experts to analyze such data neglecting the within-subject correlation. We also provide a discussion of the statistical and computational issues concerning estimation of classic Poisson mixtures.

Keywords: Mixture models; Mixtures-of-experts; MCMC; Repeated measures; Multivariate Bernoulli distribution.

TABLE OF CONTENTS

PREFACE xii

1.0 INTRODUCTION 1

2.0 LITERATURE REVIEW 5

2.1 Classic Mixture Models 5

2.2 Mixtures-of-Experts 9

2.3 Mixture Models for Dependent Data 11

2.4 Computational Approaches to Fitting Mixture Models 15

2.4.1 The EM Algorithm 15

2.4.2 MCMC Methods 16

3.0 CLASSIC POISSON MIXTURE MODELING WITH APPLICATION TO NEUROSCIENCE 19

3.1 Introduction 19

3.2 Fitting Two-Component Mixtures of Poissons to the Grain Count Data . . 21

3.2.1 Fitting a Two-Component Mixture of Poissons to the Data in Each Slide Section 21

3.2.2 Repeated Measurement Analysis 23

3.3 Unknown Number of Components 25

3.3.1 Bootstrapping the LRTs to Test for the Number of Components . . . 25

3.3.2 Fitting Four-Component Mixtures of Poissons to the Grain Count Data 26

3.4 Some New Results on Mixtures of Poissons 30

3.5 Applying Mixtures-of-Experts to the Grain Count Data 34

4.0 MIXTURES OF GENERALIZED LINEAR MIXED MODELS (MIXTURES OF GLMMS) . . . 38
4.1 Introduction . . . 38
4.2 Mixtures of GLMMs . . . 40
4.2.1 The Model . . . 40
4.2.2 The Marginal Distribution of the Observed Data . . . 41
4.3 Normal Component Mixtures of GLMMs . . . 42
4.3.1 The Model . . . 42
4.3.2 The Marginal Distribution of the Observed Data . . . 43
4.4 Poisson Component Mixtures of GLMMs . . . 46
4.4.1 The Model . . . 46
4.4.2 The Expectation and Covariance Matrix of the Observations . . . 47
4.4.3 Applying MCMC Methods to the Poisson Component Mixtures of GLMMs . . . 49
4.4.3.1 The Likelihood and Conditional Distributions. . . . 49
4.4.3.2 Sampling from the Conditional Distribution of $\boldsymbol{\gamma}$. . . . 51
4.4.3.3 Sampling the Random Effects s_i. . . . 52
4.4.3.4 Sampling from the Conditional Distribution of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. . . . 52
4.5 Extensions . . . 54
4.5.1 Using Probit Regressions to Model the Mixing Proportions . . . 54
4.5.2 Different Subject-Specific Random Effects in the Mixture Components 55
4.5.3 Other Extensions . . . 57
4.6 Discussion . . . 57
5.0 MULTIVARIATE BERNOULLI MIXTURE MODELS WITH APPLICATION TO POSTMORTEM TISSUE STUDIES IN SCHIZOPHRENIA . . . 59
5.1 Introduction and Motivating Example . . . 59
5.1.1 Overview . . . 59
5.1.2 Motivating Example . . . 60
5.2 Multivariate Bernoulli Mixtures of Normals . . . 61

5.2.1 The Model . . . 61
5.2.2 The Joint Distribution of the Observed Data for Each Subject . . . 62
5.3 Inference . . . 66
5.3.1 Augmented Likelihood and Prior Distributions . . . 66
5.3.2 The Sampling Scheme . . . 66
5.4 Application . . . 68
5.5 Simulation Study . . . 73
5.6 Discussion and Summary . . . 79
6.0 MULTIVARIATE BERNOULLI MIXTURES OF GLMMS WITH APPLICATION TO POSTMORTEM TISSUE STUDIES IN SCHIZOPHRENIA . . . 81
6.1 Introduction and Motivating Example . . . 81
6.2 Multivariate Bernoulli Mixtures of Mixed Normals . . . 83
6.2.1 The Model . . . 83
6.2.2 The Joint Distribution of the Observed Data for Each Individual . . . 84
6.3 Inference . . . 87
6.3.1 Augmented Likelihood and Prior Distributions . . . 87
6.3.2 The Sampling Scheme . . . 88
6.4 Problem Encountered in a Simulation Study . . . 89
6.5 Discussion . . . 95
7.0 FUTURE RESEARCH . . . 97
7.1 Unknown Number of Components . . . 98
7.2 Other Approaches to Constructing Multivariate Bernoulli Distributions . . . 98
7.3 Some Extra New Models . . . 99
7.3.1 Extra New Model I . . . 99
7.3.2 Extra New Model II . . . 100
7.3.3 Extra New Model III . . . 101
APPENDIX A. NEW RESULTS FOR THE CLASSIC MIXTURES OF POISSONS: QQ-PLOTS . . . 104

APPENDIX B. APPLYING MIXTURES-OF-EXPERTS TO THE NEURON VOLUME DATA . . . 109
APPENDIX C. IDENTIFIABILITY OF MULTIVARIATE BERNOULLI MIXTURES OF NORMALS . . . 112
C.1 Order Restriction for Parameters . . . 113
C.2 Two Equivalent Conjectures . . . 113
C.3 Numerical Demonstration . . . 115
APPENDIX D. DETAILS OF THE SAMPLING SCHEME FOR THE MULTIVARIATE BERNOULLI MIXTURES OF NORMALS . . . 119
D.1 Updating $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ . . . 119
D.2 Updating $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ . . . 120
APPENDIX E. DETAILS OF THE SAMPLING SCHEME FOR MULTIVARIATE BERNOULLI MIXTURES OF MIXED NORMALS . . . 122
E.1 Updating $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ . . . 122
E.2 Updating $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{s}$ . . . 123
BIBLIOGRAPHY . . . 124

LIST OF TABLES

1 *Results from fitting two-component mixtures of Poissons to each slide* 24

2 *P-values using bootstrap LRT* 27

3 *Results from fitting four-component mixtures of Poissons to each slide* 28

4 *A simulation study for comparing three approaches to computing standard errors* 32

5 *Fitting two-component mixtures-of-experts to the grain count data* 37

6 *Results of model fitting to the neuron volume data. Estimates (posterior means) and 95% credible intervals. Results are based on 13,000 iterations after 2,000 burn-in iterations.* 69

7 *The average of estimates and percentages of coverage (cover) of 10 runs for each of well-separated, medium-separated, and poorly-separated cases in the simulation study. (Mean square errors are in parentheses.)* 74

8 *Results of model fitting to one of the simulated data sets. True values, estimates (posterior means) and the 95% credible intervals.* 90

9 *The estimates obtained by fitting the new model to one of the simulated data sets* 93

10 *Fitting two-component mixtures-of-experts to the neuron volume data* 111

LIST OF FIGURES

1 *For each subject, the overall mean of the log-transformed somal volume, the posterior mean of each of the smaller neuron population and larger neuron population, and the mixing proportion of smaller neurons vs. larger neurons.* . 72

2 *Well-separated mixtures of normals: Dot plots of the estimates from the 10 runs for diagnostic, age, gender, PMI, and storage time effects in smaller neuron population, larger neuron population and mixing proportions.* 76

3 *Medium-separated mixtures of normals: Dot plots of the estimates from the 10 runs for diagnostic, age, gender, PMI, and storage time effects in smaller neuron population, larger neuron population and mixing proportions.* 77

4 *Poorly-separated mixtures of normals: Dot plots of the estimates from the 10 runs for diagnostic, age, gender, PMI, and storage time effects in smaller neuron population, larger neuron population and mixing proportions.* 78

5 *True and estimated density plots for subject 566.* 94

6 *Mixture models for repeated measures* 103

7 *QQ-plots for very-well-separated classic mixtures of Poissons.* 105

8 *QQ-plots for well-separated classic mixtures of Poissons.* 106

9 *QQ-plots for poorly-separated classic mixtures of Poissons* 107

10 *QQ-plots for very-poorly-separated classic mixtures of Poissons* 108

PREFACE

I am really grateful for all the people who helped me during the last five and a half years. First I would like to thank my co-advisors: Prof. Ori Rosen and Prof. Allan R. Sampson. Without their support and encouragement, this dissertation would not have been be completed. Their broad knowledge and sharp insights impresses me and has given the breadth and depth of the research done in this dissertation. They are always very accessible, helpful, and supportive not only in my research, but also in other aspects of my life. I have been a graduate student researcher under the direction of Prof. Allan R. Sampson for almost five years and I have learned so much from him. He has guided and helped me to become a statistician.

I want to thank Prof. Henry W. Block, Prof. Robert E. Kass, and Prof. Robert A. Sweet for being my committee members. Their valuable comments improve this dissertation. I thank our collaborators, Prof. Takanori Hashimoto, Prof. Robert A. Sweet and Prof. David A. Lewis, from Department of Psychiatry, University of Pittsburgh, for helpful discussions and the permission to use their data.

All the other faculty members in the University of Pittsburgh, Department of Statistics, have been very supportive and I have gained a lot from their classes. I would like to mention Prof. Leon J. Gleser, Prof. Satish Iyengar, and Prof. Thomas H. Savits. During my study in Pittsburgh, I have made friends with a lot of great people, Ana-Maria Kupresanin, Hsiao-yun Huang, Sungyoung Auh, Gabriela Czanner, Qingxia Chen, Tulay K. Sengul, Lulu Ren and others. I cannot list all their names, but I appreciate all their kind help.

I would like to thank my parents, Yanzhao Sun and Guixiang Li, and my two sisters. I would not be able to pursue my dream in the U.S. without their endless love, support and encouragement. Finally, I want to thank my husband, Feng Wang, who is the one always on my side and giving me strength. He brings me love, joy, and hopes, and helps me to conquer

all kinds of difficulties. I am very lucky to have him in my life. This dissertation could not have been finished without him.

1.0 INTRODUCTION

As an extremely flexible way of modeling, finite mixture models have received a lot of attention for a century. In addition to being exploited as a convenient semiparametric framework to models with unknown distributional shapes, mixture models have obvious application to modeling heterogeneous data. In biological settings, heterogeneity can result from various sources, for example, species and geographical region. A mixture model is a natural choice to use when there is a group-structure in the data or when one wants to explore the data for such a structure.

In order to use mixture models in practice, much effort has gone into finding proper ways to estimate the parameters. It is only in the recent twenty years that substantial advances have been made in this area. This is due in large part to the seminal paper by Dempster, Laird, and Rubin (1977) on the EM algorithm. It is straightforward to obtain the maximum likelihood estimates (MLE) of mixture model parameters via the EM algorithm, which interprets the observed data as incomplete and introduces component indicator variables to simplify the problem. The EM algorithm has led to increased use of mixture models in various fields. For the analysis of complicated statistical models, more and more statisticians are turning to Bayesian methods. With the advent of high-speed computers, Markov Chain Monte Carlo (MCMC) methods have been developing rapidly and have become one of the most commonly used techniques for fitting complex finite mixture models.

As the simplest mixture models for univariate variables, classic mixture models, which do not include any covariates, have been already extensively studied, and the applications of the EM algorithm to fit these models are simple.

Mixtures-of-experts, proposed by Jacobs, Jordan, Nowlan and Hinton (1991), is a mixture model for univariate variables, where both the component distributions and the mixing

proportions are allowed to depend on covariates. In practice, the components are usually generalized linear models. The mixtures-of-experts model combines the properties of generalized linear models with those of standard mixture models. Peng, Jocobs, and Tanner (1996) illustrated how to use the EM algorithm and MCMC methodology to fit mixtures-of-experts and their extension, hierarchical mixtures-of-experts.

There is some literature on formulating extensions of mixture models for dependent data. The hidden Markov model is a convenient way as it employs a stationary Markovian model for the latent states. It allows successive observations to be dependent through the component states from which they are generated. In one dimension, the hidden states are distributed as a Markov chain, whereas in two or more dimensions, they are distributed as a Markov random field. As for estimating the parameters in hidden Markov models, the EM algorithm, referred to as the forward-backward algorithm in this context, is fairly commonly used in the one-dimension case even though it is time-consuming and numerically sensitive. However, the EM algorithm is extremely complicated for Markov random fields. MCMC methods turn out to be a powerful approach to parameter estimation in hidden Markov models. The applications of MCMC methods to hidden Markov models have been demonstrated in a number of papers, including Robert, Celeux, and Diebolt (1993) and Chib (1996).

In another approach for dependent data, Rubin and Wu (1997) suggested an "extra component mixture" model to fit to a data set concerning normal and schizophrenic eye-tracking behavior. In their proposed model, the repeated measurements in the schizophrenic subjects are modeled with a two-component mixture model where the components are linear regressions with random effects, and the mixing proportions are governed by a logistic regression. Rubin and Wu demonstrated that their model can be fitted by the ECM algorithm, an extension of the EM algorithm, as well as by MCMC methods. Rosen, Jiang and Tanner (2000) proposed mixtures of marginal models for dependent data, which combine the properties of mixtures-of-experts and those of generalized estimating equations. Parameter estimation in their models was performed by a generalization of the EM algorithm.

In this dissertation, we propose three mixture models for repeated measurements, all of which can be viewed as multivariate extensions of mixtures-of-experts.

Our models are motivated by a number of neurological postmortem brain tissue studies, where repeated measurements are often observed. These observations are taken on the same experimental subject and are therefore dependent within the subject. Furthermore, each observation can be viewed as coming from one of a pre-specified number of populations, and subject-level covariates impact both the mixing proportions and the locations of the mixture components. Two such motivating data sets involving grain count data and neuron volume data are used in this dissertation to illustrate our methodology. In addition to a wide variety of applications in quantitative neurobiology, our models can be applied to longitudinal studies where repeated measurements taken over time arise from more than one population.

The first model that we propose is a mixture of generalized linear mixed models (mixture of GLMMs), where we include subject-specific random effects in the component distributions to account for the within-subject correlation present in the data. The components are generalized linear models with random effects, while the mixing proportions are governed by logistic or probit regressions. In this model, the latent component indicator random variables are considered to be independent within a subject.

Our second proposed model is a multivariate Bernoulli mixture model, where the within-subject observations are taken to be correlated by introducing dependence among the unobservable component-indicator variables within each subject. We use multivariate Bernoulli random variables that depend on covariates to describe the component indicator variables, while the mixture components in this model are generalized linear models. We focus on multivariate Bernoulli mixture of normals, which is a multivariate Bernoulli mixture model where the mixture components are linear regressions and the mixing proportions are modeled by logistic regressions with random effects.

The third model that we propose combines mixtures of GLMMs and multivariate Bernoulli mixture models, so that the within-subject dependence is induced not only through the component distributions, but also through the hidden component indicator variables. We refer to it as multivariate Bernoulli mixtures of GLMMs. As a special case of this model, we focus on multivariate Bernoulli mixtures of mixed normals, where the mixture components are linear regressions with random effects while the mixing proportions are logistic regressions

with random effects.

In Chapter 2, we provide a literature review of some important results on classic mixture models, mixtures-of-experts, mixture models for dependent data, the EM algorithm, and MCMC methods. In Chapter 3, we employ classic Poisson mixture models with a pre-specified number of components to model the grain count data. The EM algorithm is implemented estimating the model parameters. We also provide a discussion of the statistical and computational issues concerning estimation of classic Poisson mixtures. In this chapter, we employ another existing mixture model, the Poisson component mixture-of-experts to model the grain count data as well. The first new model, mixture of GLMMs is presented in Chapter 4. We develop the estimation procedures for Poisson component mixtures of GLMMs. In Chapter 5, we describe the multivariate Bernoulli mixture of normals and its extensions. We illustrate how to use MCMC methods to fit this model to the neuron volume data. The third proposed model, the multivariate Bernoulli mixture of mixed normals and its application to neuron volume data are given in Chapter 6. In Chapter 7, we summarize the possible extensions of our models and some future topics.

2.0 LITERATURE REVIEW

2.1 CLASSIC MIXTURE MODELS

Let $Y_1, \ldots, Y_n$ denote a random sample of size n and suppose that the density or probability mass function of Y_i can be written in the form

$$f(y_i) = \sum_{k=1}^{g} p_k f_k(y_i, \theta_k), \tag{2.1}$$

where the $f_k(y_i, \theta_k)$, $k = 1, \ldots, g$, are densities or probability mass functions with parameters θ_k, and p_k are nonnegative quantities that sum to one, that is $0 < p_i < 1, k = 1, \ldots g$, and $\sum_{k=1}^{g} p_k = 1$. To ensure that the parameters are identifiable, we take $\theta_1 < \theta_2 < \ldots < \theta_g$.

In this part of the literature review, we concentrate on classic mixture models, where the components f_k and mixing proportion p_k are without covariates. However, most of the results here can be extended to any arbitrary component distributions.

A very important and difficult problem in mixture models is to assess the number of components g in a mixture. This has not been completely resolved. When we know the number of the groups in a population *a priori*, each component in a mixture corresponds to a distinct existing group. In this situation, where g is known, there is one difficulty noted by Donoho (1988) that a mixture with g components might be empirically indistinguishable from one with fewer than g components, because in some instances, two components are too close to be separated. Overfitting mixture models may cause nonidentifiability, as pointed out by Crawford (1994), which may lead to either one of the mixing proportions being equal to 0 or two component densities being the same. Hence, if either of these two situations occurs in fitting a mixture with g components, we know that some of the corresponding components are not widely apart enough to be separated.

On the other hand, in practice there are many examples involving the use of mixture models where we do not have information about the number of the groups in a population. McLachlan and Peel (2000, Section 6) gave a lucid account of approaches for assessing the number of components g in a mixture model in this situation. To avoid the nonidentifiability problem, it is reasonable in practice to assess the number of components in terms of estimating the mixture order, which is defined as the smallest value of g such that all components $f_k(y_i)$ are different and none of the mixing proportions p_k are zero. The order of a mixture model can be investigated nonparametrically in terms of assessing the number of modes of a distribution. Such inferential procedures were illustrated by Titterington, Smith, and Makov (1985). However, the drawback of this approach is that the components of the mixture have to be sufficiently wide apart in order to be detected as modes. Other than a few nonparametric methods, assessing the order of a mixture model has been mainly considered in two ways, both using the likelihood function, as described by McLachlan and Peel (2000). One approach is based on a penalized form of the log likelihood. As the log likelihood increases with an additional component, it is subtracted by a term, which penalizes the model for the number of components. The penalized log likelihood yields information criteria such as AIC and BIC for the choice of g. The shortcoming of these criteria is that they are unable to produce a number to quantify the confidence in the result, such as a p -value. The other main approach to estimating the order of a mixture model is to carry out a likelihood ratio test (LRT). In mixture models, the regularity conditions break down for the LRT to have its usual asymptotic χ^2 null distribution with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses. Hence some demanding resampling approaches such as the bootstrap have to be used in order to assess the p -value of the LRT. Karlis and Xekalaki (1999) proposed a method of bootstrapping LRTs to estimate the number of components in Poisson mixtures. For more details on their approach, see Section 3.3.1, where we carry out this approach to test the number of components in the grain count data.

As for estimating the parameters in mixture models for fixed g, maximum likelihood has been one of the most commonly used approaches since the advent of the EM algorithm, especially for the classic mixture model described previously. We review the EM algorithm

and illustrate how to fit classic mixture models via the EM algorithm in Section 2.4.1. Here we interpret the mixture model expressed in (2.1) in the EM framework. The observed data point y_i, is augmented with a g-dimensional vector $\boldsymbol{z}_i$ where for $j = 1, \ldots, g$, the jth element $z_{ij} = 1$, or 0, indicating whether or not the observation y_i came from the jth component. Since in (2.1), the random variables $Y_1, \ldots, Y_n$ are assumed to be independent, it follows that the random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{i.i.d.}{\sim}$ multinomial$(1, p_1, \ldots, p_g)$. In the EM framework, the y_i's can be viewed as the incomplete data, because the associated $\boldsymbol{z}_i$'s are unobservable. The pairs $\{y_i, \boldsymbol{z}_i\}, i = 1, \ldots, n$, can be treated as the complete or augmented data, where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are realizations from the multinomial$(1, p_1, \ldots, p_g)$. We denote the complete data by $\boldsymbol{Y_c}$.

After the number of the components have been determined and the parameter estimates have been obtained via the EM algorithm, the next natural problem considered is obtaining the standard errors of the parameter estimates. There are three main approaches in general. The first two methods are information-based. In this chapter, for ease of notation, we use $\boldsymbol{\Psi}$ to denote the unknown parameter vector and $\boldsymbol{y}$ to denote all the observed data. It is well known that the asymptotic covariance matrix of the MLE $\hat{\boldsymbol{\Psi}}$ is equal to the inverse of the Fisher information matrix, $\mathcal{I}(\boldsymbol{\Psi})$, and consequently, the standard error of $\hat{\boldsymbol{\Psi}}_r$, the rth entry of $\hat{\boldsymbol{\Psi}}$, is estimated by

$$\hat{SE}(\hat{\boldsymbol{\Psi}}_r) = (\mathcal{I}^{-1}(\boldsymbol{\Psi}))_{rr}^{1/2}. \tag{2.2}$$

The matrix $\mathcal{I}(\boldsymbol{\Psi})$ can be approximated by the observed information matrix $I(\hat{\boldsymbol{\Psi}}; \boldsymbol{y})$, which is the Hessian of the negative log likelihood function evaluated at $\hat{\boldsymbol{\Psi}}$. The standard error of $\hat{\boldsymbol{\Psi}}_r$, can be estimated by

$$\hat{SE}(\hat{\boldsymbol{\Psi}}_r) = (I^{-1}(\hat{\boldsymbol{\Psi}}; \boldsymbol{y}))_{rr}^{1/2}. \tag{2.3}$$

In order to evaluate $I(\hat{\boldsymbol{\Psi}}; \boldsymbol{y})$, one obvious approach is to analytically compute the second derivatives of the log likelihood. In practice, however, this may be difficult or tedious for most mixture models.

To simplify the problem of computing derivatives, a commonly used approach is Louis' method which computes the observed information matrix in the context of the EM algorithm.

Louis (1982) showed that the observed information matrix $I(\hat{\boldsymbol{\Psi}}; \boldsymbol{y})$, can be computed as

$$I(\hat{\boldsymbol{\Psi}}; \boldsymbol{y}) = E\left\{ I_c(\hat{\boldsymbol{\Psi}}, \boldsymbol{Y_c}) | \, \boldsymbol{y} \right\} - Var\left\{ S_c(\boldsymbol{Y_c}; \hat{\boldsymbol{\Psi}}) | \, \boldsymbol{y} \right\}, \tag{2.4}$$

where $I_c(\hat{\boldsymbol{\Psi}}, \boldsymbol{Y_c})$, $S_c(\boldsymbol{Y_c}; \hat{\boldsymbol{\Psi}})$ denote the observed information matrix and the score statistics, respectively, for the complete data introduced within the EM framework. More details about Louis' method are given by Tanner (1996, Section 4).

A second approach is to use the obvious approximation of $\mathcal{I}(\boldsymbol{\Psi})$, which is $\mathcal{I}(\hat{\boldsymbol{\Psi}})$, the plug-in estimator, that is, the expected information matrix evaluated at $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$. The standard error of $\hat{\boldsymbol{\Psi}}_r$, is estimated by

$$\hat{SE}(\hat{\boldsymbol{\Psi}}_r) = (\mathcal{I}^{-1}(\hat{\boldsymbol{\Psi}}))_{rr}^{1/2}. \tag{2.5}$$

The expected information matrix is usually more complicated to use than the observed information matrix, since it requires an expectation to be taken. Moreover, Efron and Hinkley (1978) have provided explanations that $I(\hat{\boldsymbol{\Psi}}, \boldsymbol{y})$ is better than $\mathcal{I}(\hat{\boldsymbol{\Psi}})$ in terms of estimating the standard error of the MLE.

The third method to obtain standard errors uses the bootstrap approach. The estimation of the standard errors of the elements of $\hat{\boldsymbol{\Psi}}$ can be implemented by the following steps using the bootstrap.

Step 1. A bootstrap sample $\boldsymbol{y_b^*}$, is generated from the original observed data $\boldsymbol{y}$.

Step 2. The EM algorithm is applied to the bootstrap sample $\boldsymbol{y_b^*}$ to obtain the MLE $\hat{\boldsymbol{\Psi}}_b^*$ for this new data set.

Step 3. Repeat Step 1-2 B times for $b = 1, \ldots, B$. Then the covariance matrix of $\hat{\boldsymbol{\Psi}}$ can be approximated by the sample covariance matrix of these B bootstrap realizations.

$$\hat{\text{cov}}(\hat{\boldsymbol{\Psi}}) = \sum_{b=1}^{B} (\hat{\boldsymbol{\Psi}}_b^* - \bar{\hat{\boldsymbol{\Psi}}}^*)(\hat{\boldsymbol{\Psi}}_b^* - \bar{\hat{\boldsymbol{\Psi}}}^*)^T / (B-1), \tag{2.6}$$

where,

$$\bar{\hat{\boldsymbol{\Psi}}}^* = \sum_{b=1}^{B} \hat{\boldsymbol{\Psi}}_b^* / B. \tag{2.7}$$

As in (2.2), the standard error of the rth element of $\hat{\boldsymbol{\Psi}}$ can be estimated by taking the square root of the rth diagonal element of (2.6). McLachlan and Peel (2000, Section 2) have provided

a detailed account of estimating standard errors using the bootstrap. Basford, Greenway, McLachlan, and Peel (1997) compared the bootstrap and information-based methods for some mixtures with normal components. They concluded that the standard errors given by information-based approaches are less stable than those obtained by the bootstrap unless the sample size is very large.

All the above approaches for obtaining the standard errors of the parameter estimates are based on frequentist theory. If a Bayesian approach is taken, we generate samples from the posterior distributions of the unknown parameters. The standard errors of the parameter estimates are therefore assessed by the sample standard deviations of the simulated samples.

2.2 MIXTURES-OF-EXPERTS

The mixtures-of-experts model was first introduced in the neural network literature by Jacobs et al. (1991); see also Jordan and Jacobs (1994). As a mixture model in which both the component densities and the mixing proportions are dependent on covariates, the mixtures-of-experts model has the properties of both generalized linear models and mixture models.

Suppose that we have n independent observations $Y_1, \ldots, Y_n$ with corresponding covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. In the mixtures-of-experts model, the density of Y_i can be modeled as

$$f(y_i|\,\boldsymbol{x}_i, \boldsymbol{\Psi}) = \sum_{k=1}^{g} p_k(\boldsymbol{x}_i; \boldsymbol{\gamma}) f_k(y_i|\,\boldsymbol{x}_i, \boldsymbol{\beta}_k, \phi_k), \tag{2.8}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_k^T, \phi_1, \ldots, \phi_n)^T$, and the number of components is fixed at g.

If the component densities $f_k(y_i|\,\boldsymbol{x}_i, \boldsymbol{\beta}_k, \phi_k)$ belong to the exponential family, (2.8) can be considered as a mixture of generalized linear models (GLM). In practice, the component densities of mixtures-of-experts usually belong to the exponential family.

The common model for expressing the mixing proportions $p_k(\boldsymbol{x}_i; \boldsymbol{\gamma}), k = 1, \ldots, g$ is a generalization of logistic regression:

$$p_k(\boldsymbol{x}_i; \boldsymbol{\gamma}) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma_k}}}{1 + \sum_{h=1}^{g-1} e^{\boldsymbol{x}_i^T \boldsymbol{\gamma_h}}}, \; k = 1, \ldots, g-1, \tag{2.9}$$

and $p_g(\boldsymbol{x}_i;\boldsymbol{\gamma}) = 1 - \sum_{h=1}^{g-1} p_h(\boldsymbol{x}_i;\boldsymbol{\gamma_l})$, where $\boldsymbol{\gamma} = (\boldsymbol{\gamma_1}^T, \ldots, \boldsymbol{\gamma_{g-1}}^T)^T$. Alternatively, we can model the mixing proportions by a generalization of probit regression as proposed in Albert and Chib (1993). We define the cumulative probabilities

$$q_k(\boldsymbol{x}_i;\boldsymbol{\gamma}) = \sum_{h=1}^{k} p_h(\boldsymbol{x}_i;\boldsymbol{\gamma}),\ k = 1, \ldots, g-1, \tag{2.10}$$

and q_k can be given by

$$q_k(\boldsymbol{x}_i;\boldsymbol{\gamma}) = \Phi(\gamma_k - \boldsymbol{x}_i^T\boldsymbol{\alpha}), \tag{2.11}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{g-1}, \boldsymbol{\alpha}^T)^T$ is the unknown parameter vector. To ensure that the parameters are identifiable, we take $\gamma_1 = 0$ without loss of generality.

A probabilistic motivation for mixtures-of-experts is as follows. Given covariates $\boldsymbol{x}_i$,

- A vector $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 at the kth position is selected from a multinomial distribution with probability $p_k(\boldsymbol{x}_i;\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the parameter vector underlying the multinomial distribution.
- Given e_k, a response y_i is generated from $f_k(y_i|\,\boldsymbol{x}_i, \boldsymbol{\beta}_k, \phi_k)$, where $\boldsymbol{\beta}_k$ is a parameter vector and ϕ_k is a dispersion parameter. Assume that $f_k(y_i|\,\boldsymbol{x}_i, \boldsymbol{\beta}_k, \phi_k)$ belongs to the exponential family and let μ_{ik} denote the expectation of $y_i|e_k$. If h is the canonical link, the natural parameter and dispersion parameter of the conditional distribution of $y_i|e_k$ are $\eta_{ik} = h(\mu_{ik}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_k$ and ϕ_k respectively.

Mixtures-of-experts can also be viewed as fitting piecewise regression functions to the input data $\{\boldsymbol{x}_i\}$. However, contrary to traditional piecewise regression where each function is based on disjoint input regions, mixtures-of-experts adopts functions that are defined on overlapping regions. In other words, an input point may lie in multiple regions simultaneously in mixtures-of-experts. These regions therefore have "soft" boundaries. Moreover, the boundaries between these regions are simple functions which are dependent on input points.

Estimating the parameters in mixtures-of-experts can be implemented via the EM algorithm. After the component-indicator variables are introduced, parameter estimates can be obtained by the Iteratively Re-weighted Least Squares (IRWLS) procedure of GLM. McLachlan and Peel (2000) gave a comprehensive review of this. We demonstrate how to fit mixtures-of-experts models via the EM algorithm with applications to the grain count

data and neuron volume data in Section 3.5 and in Appendix B. A full Bayesian approach is another way to do inference for mixtures-of-experts. Peng et al. (1996) discussed the fitting of mixtures-of-experts using MCMC methods. Our review of MCMC methods, including the Gibbs sampler and the Metropolis algorithm, is presented in Section 2.4.2.

2.3 MIXTURE MODELS FOR DEPENDENT DATA

Some work has been done in recent years to extend mixture models to dependent data. Most of this work has concentrated on hidden Markov models, which are mixture models whose component indicators are unobserved random variables distributed as finite state Markov chains. Hidden Markov models are most useful when the observations are serially correlated.

In classic mixture models, the y_i's are generated independently from the distribution in (2.1). Hidden Markov models relax the independence of the y_i's by imposing dependence for the component states from which the y_i's are generated. We use V_i to denote the component state random variables in the context of hidden Markov models. Let $V_i \in \{1, 2, \ldots, g\}$, be the unobservable state random variable associated with each y_i, where $i = 1, \ldots, n$. The V_i's are assumed to be distributed as a finite-state Markov chain, denoted by Markov$(A, \boldsymbol{\pi}_1)$, which means that,

$$V_i | \, V_{i-1} \sim \text{Markov}(A, \boldsymbol{\pi}_1) \tag{2.12}$$

where $A = (p_{kw})$ is the transition probability matrix and $\boldsymbol{\pi}_1 = (\pi_{11}, \ldots, \pi_{1g})$ is the initial probability distribution. We have $P(V_i = w | \, V_{i-1} = k) = p_{kw}$ and $P(V_1 = k) = \pi_{1k}$, for $k = 1, \ldots, g$. Given $V_i = k$, the observation y_i can be selected from a population with density $f_k(\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ is the parameter vector. The corresponding random variable Y_i is therefore conditionally distributed as follows:

$$f(y_i | \, V_{i-1} = v_{i-1}, \boldsymbol{\Psi}) = \begin{cases} \sum_{k=1}^{g} \pi_{1k} f_k(y_i, \boldsymbol{\theta_k}) & \text{for } i = 1 \\ \\ \sum_{k=1}^{g} p_{kv_{i-1}} f_k(y_i, \boldsymbol{\theta_k}) & \text{for } i > 1. \end{cases} \tag{2.13}$$

In (2.13), the unknown parameter vector is $\boldsymbol{\Psi} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g, A, \boldsymbol{\pi}_1\}$. Given all the component state random variables V_i's, the Y_i's are assumed to be conditionally independent; that is $f(y_1, \ldots, y_n | V_1, \ldots, V_n) = \prod_{i=1}^{n} f(y_i|V_i)$.

Note that the classic mixture model described in (2.1) is a special case of the hidden Markov model when the V_i's are distributed independently and identically over time. This happens if and only if the initial distribution $\boldsymbol{\pi}_1 = (p_1, \ldots, p_g)^T$, and the transition probability matrix satisfies $A = (\boldsymbol{\pi}_1 : \boldsymbol{\pi}_1 : \ldots : \boldsymbol{\pi}_1)^T$.

If the component state variables V_i's are assumed to arise from a Markov random field in two or more dimensions, the corresponding mixture model is referred to as hidden Markov random field. For more details, see Geman and Geman (1984).

As in other mixture models, parameter estimation in hidden Markov models usually can be done in either a frequentist or a Bayesian approach. In a series of papers, Baum and his colleagues (Baum and Petrie (1966), and Baum, Petrie, Soules, and Weiss (1970)) discussed a recursive algorithm to obtain the MLE's in hidden Markov models. Their work, which precedes the EM algorithm, is now referred to as the forward-backward algorithm. This recursive algorithm is actually an application of the EM algorithm to hidden Markov models. As noted by Leroux and Puterman (1992), the forward-backward algorithm and its subsequent modifications are unfortunately time-consuming and numerically unstable. Concerning the EM algorithm in hidden Markov random fields, Qian and Titterington (1991) explained that this problem increases with complex structures of the component states, and concluded that the EM algorithm may not be useful in hidden Markov random fields in most situations. A Bayesian approach can avoid the computation of the likelihood by treating the unobservable Markov states v_i as unknown parameters and simulating them along with the other unknown model parameters using the Gibbs sampler. Robert et al. (1993) showed how to perform Bayesian estimation for hidden Markov models through the Gibbs sampler. Chib (1996) improved previous methods and showed that the latent states v_i's can be simulated from their joint distribution simultaneously, instead of n individual conditional distributions. This greatly improved the convergence of the Gibbs sampler. Using the pseudolikelihood function in place of the likelihood, Rydén and Titterington (1998) circumvented the difficulties with hidden Markov random fields and applied the Gibbs sampler for this model. When

g is unknown, Robert, Rydén, and Titterington (2000) proposed reversible jump MCMC methods to estimate the parameters, as well as the number of components. For more recent work regarding Bayesian methods for hidden Markov models, see the review written by Scott (2002).

In some other work on mixture models that allow for dependent data, instead of imposing the dependence of the component-indicator variables through a Markov chain, random effects are incorporated into the component distributions to account for the correlation in the data. Aitkin (1996, 1999) considered repeated measurements selected from mixtures of GLMs, where the mixing proportions were not dependent on covariates. In his papers, the likelihood can be written as an integral over the random effect and is approximated numerically by Gaussian quadrature.

Rubin and Wu (1997) suggested the "Extra-Component Mixture Model" to fit a data set concerning normal and schizophrenic eye-tracking behavior. The data set includes repeated measurements of manual reaction times for each of 43 normal subjects and 43 schizophrenic subjects. On average, there are 34 observations for each subject. For susceptible schizophrenic subjects, the observed outcomes were considered as coming from one of two populations. One consists of responses which suffer from a deficit, and are similar to the normal observations but relatively slower; while the other consists of responses which suffer additionally from intermittent disruptions and are slower and more variable. In the Rubin and Wu model, the repeated measurements in the susceptible schizophrenic subjects were modeled with a two-component mixture model where the components were linear regressions with random effects and the mixing proportions were governed by logistic regressions. We term this model the Rubin-Wu model in this dissertation although it is less general than the "Extra-Component Mixture Model", where Rubin et al. also modeled the observations from the normal subjects and non susceptible schizophrenic subjects. Briefly speaking, let Y_{ij} be the jth measurement on subject i, the vector $\boldsymbol{x_{ij}}$ be the covariates and α_i be the random

effect. Then the Rubin-Wu model can be rewritten as:

$$
\begin{aligned}
(Y_{ij} |\, Z_{ij} = 0, \boldsymbol{x_{ij}}, \alpha_i) &\overset{indep}{\sim} N(\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1, \sigma_1^2), \\
(Y_{ij} |\, Z_{ij} = 1, \boldsymbol{x}_{ij}, \alpha_i) &\overset{indep}{\sim} N(\alpha_i + \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2, \sigma_2^2), \\
\alpha_i &\overset{indep}{\sim} N(0, \sigma_\alpha^2);
\end{aligned}
$$

and for the latent component indicator random variables,

$$
\begin{aligned}
Z_{ij} &\overset{indep}{\sim} \text{Bernoulli}(\lambda_{ij}), \\
\text{logit}(\lambda_{ij}) &= \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_\lambda,
\end{aligned}
$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_\lambda$, σ_1^2, σ_2^2, σ_α^2 are unknown parameters. The Z_{ij}'s and α_i's are mutually independent. The $\boldsymbol{x_{ij}}$'s are covariate vectors. The model was fitted by MCMC methods with the starting values chosen as the estimates from the ECM (expectation–conditional maximization) algorithm. ECM is an extension of the EM algorithm, which replaces the M-step in the EM algorithm by a sequence of conditional M steps, when a single M-step has no closed form.

As an extension of mixtures-of-experts, mixtures of marginal models were developed by Rosen et al. (2000) for repeated measurements. It combines the properties of mixtures-of-experts and those of generalized estimating equations, which was introduced by Liang and Zeger (1986) and Zeger and Liang (1986). In mixtures of marginal models, the marginal distributions for each observation are considered as mixtures-of-experts and written in the form of (2.8). To account for the correlation between observations on the same subject, Rosen et al. (2000) introduced the working correlation matrix for each component and employed generalized estimating equations in the M-step of the EM algorithm.

2.4 COMPUTATIONAL APPROACHES TO FITTING MIXTURE MODELS

2.4.1 The EM Algorithm

The EM algorithm was introduced by Dempster, Laird and Rubin (1977). It is an iterative method to find the mode of a likelihood function $L(\boldsymbol{\Psi}|\,\boldsymbol{y})$. By augmenting the data with the latent data $\boldsymbol{z}$, the augmented log likelihood function $\log L(\boldsymbol{\Psi}|\,\boldsymbol{y},\boldsymbol{z})$ can be written in a simpler form than the original log likelihood. In each iteration, there are two steps: Expectation (E-step) and Maximization (M-step). Let $\boldsymbol{\Psi}^{(t)}$ be the current guess of the mode at iteration t, and $p\,(\boldsymbol{z}|\,\boldsymbol{y},\boldsymbol{\Psi}^{(t)})$ denote the conditional distribution of the latent random variable $\boldsymbol{Z}$. The E-step consists of taking the expectation of the augmented log likelihood with respect to $p\,(\boldsymbol{z}|\,\boldsymbol{y},\boldsymbol{\Psi}^{(t)})$, that is,

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(t)}) = \int_{\boldsymbol{z}} \log L(\boldsymbol{\Psi}|\,\boldsymbol{y},\boldsymbol{z}) p\,(\boldsymbol{z}|\,\boldsymbol{y},\boldsymbol{\Psi}^{(t)}) d\boldsymbol{z}. \tag{2.14}$$

The M-step consists of maximizing the Q function (2.14) with respect to $\boldsymbol{\Psi}$ to obtain $\boldsymbol{\Psi}^{(t+1)}$. The E-step and M-step are repeated iteratively until $\|\,\boldsymbol{\Psi}^{(t+1)}-\boldsymbol{\Psi}^{(t)}\|$ or $\left|\,L(\boldsymbol{\Psi}^{(t+1)}|\,\boldsymbol{y}) - L(\boldsymbol{\Psi}^{(t)}|\,\boldsymbol{y})\right|$ is less than a pre-specified small number.

For the classic mixture models in (2.1), in the EM framework, the y_i's can be viewed as the incomplete-data, while the pairs $\{y_i,\boldsymbol{z}_i\}, i=1,\ldots,n$, can be treated as the complete or augmented data, where $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n$ are realizations from the multinomial$(1,p_1,\ldots,p_g)$. The augmented likelihood can be written as:

$$L(\boldsymbol{\Psi}|\,\boldsymbol{y},\boldsymbol{z}) = \prod_{i=1}^{n}\prod_{k=1}^{g}\{p_k f_k(y_i,\theta_k)\}^{z_{ik}}. \tag{2.15}$$

Given the current estimate $\boldsymbol{\Psi}^{(t)}$, in the E-step, the conditional expectation of the augmented log likelihood can be derived from (2.14) and expressed as:

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{g}\tau_{ik}^{(t)}\{\log\, p_k + \log\, f_k(y_i,\theta_k)\}. \tag{2.16}$$

where $\tau_{ik}^{(t)} = E(z_{ik}|\,\boldsymbol{y}, \boldsymbol{\Psi}^{(t)}) = p_k^{(t)} f_k(y_i, \theta_k^{(t)}) / \sum_{j=1}^{g} p_j^{(t)} f_j(y_i, \theta_j^{(t)})$. In the M-step, by maximizing $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(t)})$ in equation (2.16), the updated $p_k^{(t+1)}$ is given as

$$p_k^{(t+1)} = \sum_{i=1}^{n} \tau_{ik}/n \tag{2.17}$$

for $k = 1, \ldots, g$. Moreover, the updated $\theta_k^{(t+1)}$ $(k = 1, \ldots, g)$ can be obtained by solving the equation

$$\sum_{i=1}^{n} \tau_{ik}^{(t)} \partial \log\, f_k(y_i, \theta_k) / \partial \theta_k = 0,\ \ k = 1, \ldots, g. \tag{2.18}$$

We employ the EM algorithm with Poisson mixture components to the grain count data in Section 3.2 and Section 3.3. The parameters in other mixture models such as mixtures-of-experts can be estimated by the EM algorithm as well. We demonstrate this with applications to the grain count data and neuron volume data in Section 3.5 and Appendix B.

Dempster et al. (1977) noted that the EM algorithm increases the likelihood function $L(\boldsymbol{\Psi}|\,\boldsymbol{y})$ at each iteration, meaning $L(\boldsymbol{\Psi}^{(t+1)}|\,\boldsymbol{y}) \geq L(\boldsymbol{\Psi}^{(t)}|\,\boldsymbol{y})$. They also showed that if $\boldsymbol{\Psi}^{(t)}$ converges, it goes to some stationary points such as local maxima or saddle points. In order to find the global maximum, various starting values need to be tried. The convergence rate of the EM algorithm is linear and may converge very slowly in a neighborhood of the maximum point. For more results on the EM algorithm, see Tanner (1996, Section 4) and McLachlan and Peel (2000) .

2.4.2 MCMC Methods

When a Bayesian approach is taken, it is usually impossible to sample from the posterior distribution directly. To overcome this problem, MCMC methods generate samples from the posterior distribution by constructing a Markov chain with the posterior distribution as its stationary distribution.

The Gibbs sampler yields a Markov chain by simulating from the full conditional distribution for each parameter. Let $\boldsymbol{\Psi} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\}$ denote the parameter vector (which may include the parameters for the missing data), where $\boldsymbol{\theta}_i\,(i = 1, \ldots, d)$ is a subvector of the $\boldsymbol{\Psi}$. The full conditional distribution of $\boldsymbol{\theta}_i$ is denoted by $f(\boldsymbol{\theta}_i|\,\boldsymbol{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_d)$. After

a starting value of $\boldsymbol{\Psi}^{(0)}$ is chosen, the Gibbs sampler is implemented iteratively to obtain $\boldsymbol{\Psi}^{(t)}$ for t $(t = 1, 2, \ldots)$ as follows:

Step1. Simulate $\boldsymbol{\theta}_1^{(t)}$ from $f(\boldsymbol{\theta}_1 | \, \boldsymbol{y}, \boldsymbol{\theta}_2^{(t-1)}, \ldots, \boldsymbol{\theta}_d^{(t-1)})$

Step2. Simulate $\boldsymbol{\theta}_2^{(t)}$ from $f(\boldsymbol{\theta}_2 | \, \boldsymbol{y}, \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_3^{(t-1)}, \ldots, \boldsymbol{\theta}_d^{(t-1)})$

$\vdots$

Step d**.** Simulate $\boldsymbol{\theta}_d^{(t)}$ from $f(\boldsymbol{\theta}_d | \, \boldsymbol{y}, \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}, \ldots, \boldsymbol{\theta}_{d-1}^{(t)})$

The above loop is run for a burn-in period of N_1 iterations. The samples, $\{\boldsymbol{\Psi}^{(t)}, \, t > N_1\}$, can therefore be regarded as having been simulated from the posterior distribution of the unknown parameters.

The Metropolis algorithm is another important method based on Markov chain theory. It was first proposed by Metropolis, et al. (1953). Instead of simulating from the full conditional distributions of the parameters, it works on the posterior distribution $f(\boldsymbol{\Psi} | \, \boldsymbol{y})$ directly. In each iteration, the Metropolis algorithm chooses a candidate from a pre-specified distribution and accepts or rejects the candidate with probability defined in terms of $f(\boldsymbol{\Psi} | \, \boldsymbol{y})$. A generalization of the Metropolis algorithm, is the Metropolis-Hastings algorithm (Hastings, 1970), which is carried out as follows. Given the current value $\boldsymbol{\Psi}^{(t)}$, there are two steps in the process of choosing the next value of the Markov chain, say $\boldsymbol{\Psi}^{(t+1)}$, :

Step 1. Sample a candidate $\boldsymbol{\Psi}^*$ from a proposal density $q(\boldsymbol{\Psi}^{(t)}, \boldsymbol{\Psi}^*)$, which is an arbitrary transition probability function, for example, a multivariate normal or a multivariate t distribution.

Step 2. Accept $\boldsymbol{\Psi}^*$ and let $\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^*$ with probability $\alpha(\boldsymbol{\Psi}^{(t)}, \boldsymbol{\Psi}^*)$, otherwise reject $\boldsymbol{\Psi}^*$ and let $\boldsymbol{\Psi}^{(t+1)} = \boldsymbol{\Psi}^{(t)}$, where

$$\alpha(\boldsymbol{\Psi}^{(t)}, \boldsymbol{\Psi}^*) = \min \left\{ \frac{f(\boldsymbol{\Psi}^* | \, \boldsymbol{y}) q(\boldsymbol{\Psi}^*, \boldsymbol{\Psi}^{(t)})}{f(\boldsymbol{\Psi}^{(t)} | \, \boldsymbol{y}) q(\boldsymbol{\Psi}^{(t)}, \boldsymbol{\Psi}^*)} \, , \; 1 \right\} . \tag{2.19}$$

In the implementation of MCMC methods for most mixture models, component-indicator variables Z_i's defined in Section 2.4.1 are introduced. They are treated as missing data and simulated along with the unknown parameters from their full conditional distributions. If the full conditional distributions can be sampled directly, the Gibbs sampler can be easily

implemented. Otherwise, Müller (1993) suggested that Metropolis steps be used within the Gibbs sampler to handle this situation. For example, if the full conditional distribution of $\boldsymbol{\theta}_1$, i.e., $f(\boldsymbol{\theta}_1|\,\boldsymbol{y}, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d)$ is not a standard distribution, one can perform Metropolis (or Metropolis-Hastings) steps to obtain a realization from $\boldsymbol{\theta}_1$. The value of $\boldsymbol{\theta}_1^{(t)}$ can be chosen as the Kth value in the Metropolis subchain. Peng et al. (1996) used this strategy with $K = 40$ in the implementation of MCMC to mixtures-of-experts, even though Müller pointed out that $K = 1$, i.e., one pass through Step 1 and Step 2 described above is sufficient. In a non-mixture context, Chib, Greenberg, and Winkelmann (1998) and Chib and Jeliazkov (2001) employed Müller's method ($K = 1$) to fit Poisson regression with random effects. We apply Müller's method with $K = 1$ to fit our proposed models to the motivating data sets in Chapter 4, 5, and 6.

See Tierney (1994), Tanner (1996), and Gelman, Carlin, Stern, and Rubin (2003) for more complete reviews of MCMC methods.

3.0 CLASSIC POISSON MIXTURE MODELING WITH APPLICATION TO NEUROSCIENCE

3.1 INTRODUCTION

In this chapter, the main consideration is the application of classic Poisson mixtures for modeling some neuronal postmortem brain tissue data. In this application several additional novel issues are raised. Overall it appears that mixture models have not been previously applied to neuroscience data like these.

This research was motivated by the grain count data set, which was collected and described by Hashimoto, Volk, Eggan, Mirnics, Pierri, Sun, Sampson, and Lewis (2003). Impairments of certain cognitive functions, such as the working memory, are commonly observed in individuals with schizophrenia. Some neuroscientists have suggested that it might be caused by alterations in the circuitry of the prefrontal cortex (PFC). Previous studies have shown the altered γ- aminobutyric acid (GABA) neurotransmission in the PFC of individuals with schizophrenia. In order to understand the neural circuitry basis of impaired working memory in schizophrenia, it is important to identify the affected subset of GABA neurons. Most GABA neurons express one of three calcium-binding proteins: parvalbumin (PV), calretinin(CR), and calbindin D-28. These markers can be used to identify particular subsets of GABA neurons. One of the goals of this study is to see if the PV containing neurons have altered PV mRNA expression level in the PFC of individuals with schizophrenia.

In this study, brain tissues from fifteen pairs of schizophrenic and control subjects, matched for sex, age and postmortem interval (PMI), were examined. For each subject, the right PFC was blocked coronally, and serial sections were cut and then mounted onto glass slides. Three slide sections were selected by systematic random sampling and were

evenly spaced with the rostro-caudal locations of sections matched within each pair. These sections were hybridized with ^{35}S-labeled RNA probes in a hybridization buffer. These RNA probes are radioactive, that is, emit β particles and bind specially to PV mRNA. The section after the probe wash is coated with a photo-sensitive emulsion. The β particles emitted from the bound ^{35}S-labeled neurons react with the emulsion coating the sections and are visible as grains on the film. The magnitudes of the grain counts is a measure of the PV mRNA expression level. Thus, neurons with high visual grain counts are likely to be PV GABA neurons. However, there is a natural background of β particles striking the section and thus for all neurons in a section, there will typically be a non-negative grain count visible on the neuron cross-section. Thus, the grain count for a randomly chosen neuron can be viewed as coming from one of two populations: the grain counts for PV GABA neurons or the grain counts for non-PV GABA neurons. To randomly sample the neurons, sampling frames (approximately 40-80 per section) were placed by systematic random sampling within the area of interest for each section. In each sampling frame, the grains within each neuron were counted. There were approximately 1000 neurons counted for PV mRNA in each section, so that these neurons can be viewed as coming from a mixtures of grain counts from the PV-containing neurons and grain counts from non-PV-containing neurons. In postmortem tissue studies, age, gender, PMI, pair, storage time, and brain pH of each subject often affect the mRNA expression in the neurons. Therefore, we treat them as covariates, in addition to regarding the diagnostic effect (schizophrenia versus control) as the main effect.

Hashimoto, et al. (2003) only used the observations which are larger than a chosen cut-off point and treated these observations as coming from the PV-containing neuron population. In this chapter, we first employ a two-component mixture of Poissons to model the observed grain counts in each section, and then use a multivariate analysis of covariance model to detect the diagnostic effect. The results are given in Section 3.2.

While not discussed in Hashimoto, et al. (2003), it is known that neuroscientists suspect that there are at least two types of PV-containing neurons, including arbor neurons and chandelier neurons. Arbor neurons have larger neuronal size because of their larger dendrites while the chandelier neurons have smaller neuronal size as a result of their limited dendrites. It is suspected that each of the various types of PV-containing neurons might show different

numbers of grains due to their different neuronal sizes. In Section 3.3, we treat the number of neuron types as unknown and test the number of components in a Poisson mixture using bootstrap methods, and use an appropriate number of components in the mixture of Poisson to model the observed grain counts.

While exploring the methodology to fit the grain count data, we obtain some new results, given in Section 3.4 regarding estimating the standard errors of the parameter estimates in mixtures of Poissons.

In addition to the classic Poisson modeling for the grain count data, in Section 3.5, we employ another existing mixture model, mixtures-of-experts, to analyze the grain count data.

3.2 FITTING TWO-COMPONENT MIXTURES OF POISSONS TO THE GRAIN COUNT DATA

In this section, the analysis is carried out in two stages to determine the diagnostic effect on grain counts. In the first stage, a classic two-component mixture of Poissons is fitted to the grain counts for each slide section for each subject. Within slide section, the neurons were treated independently. In the second stage, repeated measurement analysis is used to evaluate the diagnostic effect upon the means of the components and mixing proportion, respectively. The estimates of each parameter, for example, the means of the first components on three sections of one subject, obtained in stage I, are treated as repeated measures taken on this subject and are assumed equally correlated with exchangeable covariance structure.

3.2.1 Fitting a Two-Component Mixture of Poissons to the Data in Each Slide Section

With the assumption that within a subject the grain counts within each section are independent identically distributed, the mixture of Poissons model can be applied to the grain count data. Recall that the data set consists of 15 pairs (a schizophrenic subject and a normal

subject in each pair), 3 sections in each subject and approximately 1,000 neurons within each section. For each subject and each of the three sections, let $Y_1, \ldots, Y_n$, denote the i.i.d. grain counts of the neurons, so that the density can be written in the form:

$$f(y_i) = \sum_{k=1}^{g} p_k \frac{e^{-\lambda_k} \lambda_k^{y_i}}{y_i!}, \tag{3.1}$$

where $\lambda_k > 0$, $k = 1, \ldots, g$, are unknown parameters, p_k, $k = 1, \ldots, g$, are unknown constant weights, and g is the number of mixture components. For identifiability, let $\lambda_1 < \lambda_2 < \ldots < \lambda_g$. We now treat the observed grain counts as coming from two population: the grain counts for non-PV GABA neurons or the grain counts for PV GABA neurons. Thus, we assume $g = 2$ in (3.1), and λ_1 and λ_2 are, respectively, the means of grain counts for non-PV neurons and PV neurons, while p_1 is the proportion of non-PV containing neurons on this specific slide section.

For each subject and each slide section, we use the EM algorithm to estimate $p_1, \lambda_1, \lambda_2$. Using the EM algorithm discussed in Section 2.4.1 for classic mixture models, in the E-step, the conditional expectation of the indicators is given by

$$\begin{aligned} \tau_i^{(t)} &= E(Z_{i1} | \, y_1, \ldots, y_n, p_1^{(t)}, \lambda_1^{(t)}, \lambda_2^{(t)}) \\ &= \frac{p_1^{(t)} e^{-\lambda_1^{(t)}} \lambda_1^{(t) y_i}}{p_1^{(t)} e^{-\lambda_1^{(t)}} \lambda_1^{(t) y_i} + (1 - p_1^{(t)}) e^{-\lambda_2^{(t)}} \lambda_2^{(t) y_i}}. \end{aligned}$$

In the M-step, the updated parameters are given by

$$p_1^{(t+1)} = \sum_{i=1}^{n} \tau_i^{(t)} / n, \tag{3.2}$$

$$\lambda_1^{(t+1)} = \sum_{i=1}^{n} \tau_i^{(t)} y_i / \sum_{i=1}^{n} \tau_i^{(t)}, \tag{3.3}$$

$$\lambda_2^{(t+1)} = \sum_{i=1}^{n} (1 - \tau_i^{(t)}) y_i / \sum_{i=1}^{n} (1 - \tau_i^{(t)}). \tag{3.4}$$

The E-step and M-step are repeated iteratively until $|p_1^{(t+1)} - p_1^{(t)}|$, $|\lambda_1^{(t+1)} - \lambda_1^{(t)}|$, and $|\lambda_2^{(t+1)} - \lambda_2^{(t)}|$ are all less than 0.0001.

By fitting the two-component Poisson mixture to the grain counts in each slide section, for section j in subject i, the estimates of $p_1, \lambda_1, \lambda_2$ are obtained and denoted by $\hat{p}_{1ij}, \hat{\lambda}_{1ij}, \hat{\lambda}_{2ij}$, where $i = 1, \ldots, 30$ and $j = 1, 2, 3$.

The averages (standard errors) of $\hat{\lambda}_{2ij}$ and $\hat{p}_{1ij}$ are 43.67 (1.77) and 0.95 (0.001) respectively. These values are very close to those obtained by using the cut-off point technique described in Hashimoto, et al. (2003), where the average (standard error) of the means of the PV-containing neurons across slide sections is 39.01 (1.48) and the average (standard error) of the proportions of the PV-containing neurons is 0.94 (0.002).

3.2.2 Repeated Measurement Analysis

To evaluate the diagnostic effect on the mean of grain count in non-PV containing neurons, we first perform the square root transformation on the $\hat{\lambda}_{1ij}$'s to stabilize the variance as they are the estimates of Poisson means. Then the $(\sqrt{\hat{\lambda}_{1i1}}, \sqrt{\hat{\lambda}_{1i2}}, \sqrt{\hat{\lambda}_{1i3}})^T$ can be considered as correlated and treated as repeated measures with a compound symmetric covariance structure; see Littell, Milliken, Struop, and Wolfinger (1996). In other words, $(\sqrt{\hat{\lambda}_{1i1}}, \sqrt{\hat{\lambda}_{1i2}}, \sqrt{\hat{\lambda}_{1i3}})^T$ is assumed to have a multivariate normal distribution with mean $(\mu_{i1}, \mu_{i2}, \mu_{i3})^T$ and compound symmetric covariance matrix in which the diagonal elements are σ^2 and the off-diagonal elements are $\rho\sigma^2$. The primary model employed to detect the diagnostic effect is a multivariate analysis of covariance (MANCOVA) with diagnostic group as the main effect, pair, slide section as categorical variables and storage time and brain pH as the other covariates. To further validate this model, we consider a secondary model with diagnostic group as the main effect, slide section as categorical variable and age, gender, PMI, storage time and Brain pH as the other covariates. The same variance stabilizing transformation and MANOVA models are employed for the $\hat{\lambda}_{2ij}$'s to examine the diagnostic effect on the mean of grain counts in PV containing neurons. Since the $\hat{p}_{1ij}$'s are binomial proportions, the arcsine square root transformation, $\arcsin\sqrt{\hat{p}_{1ij}}$, is performed to stabilize the variance before the MANCOVA models are used. The MANCOVA analysis are implemented in SAS Proc Mixed. The analysis is based on the transformed data. To provide the interpretability in neuroscience, we back-transform the least squares means (lsmeans). We recognize that the

Table 1: *Results from fitting two-component mixtures of Poissons to each slide*

Response variables	Cont. lsmeans[a]	Sz. lsmeans[b]	F test	% diff.[c]
mean of non-PV neurons	2.024	1.881	$F_{1,12} = 4.80, p = 0.049$	7.08%
mean of PV neurons	47.197	37.205	$F_{1,12} = 13.37, p = 0.003$	21.17%
prop. of non-PV neurons	0.946	0.950	$F_{1,12} = 1.03, p = 0.330$	–

[a]Back-transformed lsmeans for control group
[b]Back- transformed lsmeans for schizophrenic group
[c]percent difference of the back-transformed lsmeans relative to control

back-transformed estimates may alter some of the nice properties of the estimates obtained from the transformed data.

The results from the primary model, which are consistent with those from the secondary model for all response variables, are summarized in Table 1, where we report the back-transformed least squares means (lsmeans) for each diagnostic group (control and schizophrenic), the results of an F test for diagnostic effect based on type III sums of squares, and for the mean of grain count in each Poisson component, we give the percent difference of the back-transformed least squares means relative to the control group, which is $(\mathrm{C} - \mathrm{S})/\mathrm{C}$, where C denotes the back-transformed lsmeans for the control group and S denotes the back-transformed lsmeans for the schizophrenic group.

The MANCOVA results show that the grain counts are reduced in subjects with schizophrenia for both PV containing neurons, and in non-PV containing neurons. The diagnosis does not affect the proportions of PV containing neurons. The marginally significant difference between the two diagnostic groups of the non-PV containing neurons was initially somewhat surprising to the neuroscientists. However, there appears a scientific explanation for this. Recall that PV containing neurons enclose ^{35}S-labeled RNA probes which emit β particles. The β particles can travel from their binding site up to $100\mu m$ in emulsion and still remain visible as grains. However, the average diameter of the neurons is $22\mu m$. Thus, more PV in the PV containing neurons also increases the background grain counts for nearby neurons and seemingly there then is the impact of this on the grain counts in nearby non-PV containing neurons. Since in normal subjects there are higher grain counts in the PV containing

neurons, i.e., more ^{35}S-labeled RNA probes, there is the potential for more background grain counts for nearby non-PV containing neurons, and more grain counts for nearby PV containing neurons as well. This perhaps explains the slightly increased grain counts in normal subjects' non-PV containing neurons, and also suggests that the grain counts in normal subjects are probably slightly overestimated (this latter comment is not considered further in this dissertation).

3.3 UNKNOWN NUMBER OF COMPONENTS

Neuroscientists know that two types of PV-containing neurons exist, which are arbor neurons and chandelier neurons. However, it is not clear if there are more types of the PV-containing neurons. Instead of viewing the grain count data as observations from two populations, PV containing neurons and non-PV containing neurons, we now treat the grain count data as arising from a variety of neurons, that is, arising from a mixture of Poissons where we do not have information about the number of components.

3.3.1 Bootstrapping the LRTs to Test for the Number of Components

We again assume that the grain counts in each slide section are independent and identically distributed, and that for each subject and each of the three sections, the grain count of each neuron has the density written in the form (3.1). Note that approximately 1000 neurons were counted in each section for each subject. To determine the optimal number of components g in mixtures of Poisson, we use the methods proposed in Karlis and Xekalaki (1999).

Consider the null hypothesis H_0: the number of component is g and alternative hypothesis H_1: the number of component is $g+1$. The likelihood ratio test (LRT) for testing such a hypothesis has some difficulty since the standard asymptotic χ^2 distribution is not applicable. The reason for this is that $p_{g+1} = 0$ under the null hypothesis, is on the boundary of the parameter space, and as such the regularity conditions break down. Karlis et al. (1999) proposed using a bootstrap approach to construct the null distribution of the LRT statistic.

Testing for the optimal number of component g can be carried out in the following steps.

Consider fitting mixtures of Poissons with g components and $g+1$ components, respectively, to the grain count data, and obtain the MLE's of the parameters under each model by the EM algorithm, denoted by $\hat{\mathbf{\Psi}}_g$ and $\hat{\mathbf{\Psi}}_{g+1}$ respectively. Calculate the LRT, denoted by L_{obs}. Next simulate B bootstrap samples with size n, where n is the sample size of the observed grain counts, from a g-component mixture of Poissons with parameter $\hat{\mathbf{\Psi}}_g$. For each bootstrap sample, fit the g-component mixture of Poissons and $g+1$-component mixture of Poissons, and obtain the LRT statistic, denoted by $L_i, i = 1, \ldots, B$. Compute the assessed p-value for L_{obs} relative to the distribution of $L_i, i = 1, \ldots, B$. If the assessed p-value is smaller than the pre-specified level of significance, then continue the process comparing now $g+1$ and $g+2$ components. If the p-value is larger, then one concludes that there are exactly g components.

It is time-consuming to carry out this computationally intensive test for all 90 slide sections. To illustrate the procedure, we choose three subjects randomly from the 30 subjects, and for each of the three slide sections in each subject, we obtain the optimal number of components for fitting the mixtures of Poissons to the grain counts. In Table 2, we report for each slide section for the three chosen subjects the assessed p-values for each choice g, for testing g versus $g+1$ components. On the basis of these p-values at the 5% level of significance, for 5 slide sections, the optimal number of components would be chosen to be equal to 4; for 3 slide sections, it would be equal to 3; while for 1 slide section , it would be 5. In light of these results, we view as reasonable fitting a mixture of Poissons with four components for the grain counts in every slide section. We do so now for all 30 subjects.

3.3.2 Fitting Four-Component Mixtures of Poissons to the Grain Count Data

We employ the EM algorithm to estimate the unknown parameters λ_k and p_k, where $k = 1, \ldots, 4$ and $\sum_{k=1}^{4} p_k = 1$. In the E-step, conditional on the current estimates of the unknown

Table 2: *P-values using bootstrap LRT*

		P-value for g versus $g+1$				
Subject	Slide	1	2	3	4	5
1	1	0	0	0.035	0.383	–
1	2	0	0	0.005	0.254	–
1	3	0	0	0.239	–	–
2	1	0	0	0.005	0.259	–
2	2	0	0	0.005	0.100	–
2	3	0	0	0.005	0.313	–
3	1	0	0	0.418	–	–
3	2	0	0	0.189	–	–
3	3	0	0	0.025	0.045	0.234

parameters, we calculate the expectations of the component-indicator variables by

$$\begin{aligned} \tau_{ik}^{(t)} &= E(Z_{ik}|\, y_1, \ldots, y_n, p_1^{(t)}, \ldots, p_4^{(t)}, \lambda_1^{(t)}, \ldots, \lambda_4^{(t)}) \\ &= \frac{p_k^{(t)} e^{-\lambda_k^{(t)}} \lambda_k^{(t) y_i}}{\sum_{k=1}^4 p_k^{(t)} e^{-\lambda_k^{(t)}} \lambda_k^{(t)} y_i}, \quad k = 1, \ldots, 4. \end{aligned}$$

In the M-step, the updated parameters are given by

$$p_k^{(t+1)} = \sum_{i=1}^n \tau_{ik}^{(t)}/n, \tag{3.5}$$

$$\lambda_k^{(t+1)} = \sum_{i=1}^n \tau_{ik}^{(t)} y_i / \sum_{i=1}^n \tau_{ik}^{(t)}, \quad k = 1, \ldots, 4. \tag{3.6}$$

$$\tag{3.7}$$

As before, the E-step and M-step are repeated until the difference between the updated value and the current value is less than 0.0001 for all the λ_k's and p_k's.

After obtaining the parameter estimates for each slide section, denoted by $\hat{\lambda}_{kij}$ and $\hat{p}_{kij}$, where $i = 1, \ldots, 30$, $j = 1, 2, 3$, and $k = 1, \ldots, 4$, we again employ multivariate analysis of covariance on the $\hat{\lambda}_{1ij}$'s, $\hat{\lambda}_{2ij}$'s, $\hat{\lambda}_{3ij}$'s, $\hat{\lambda}_{4ij}$'s and $\hat{p}_{1ij}$'s, $\hat{p}_{2ij}$'s and $\hat{p}_{3ij}$'s respectively, to examine the diagnostic effect on the mean of grain counts in each component and the proportion of each component. As described in Section 3.2, for the λ_{kij}'s and p_{kij}'s, we use the square root transformations and arcsine square root transformations, respectively, to stabilize their

Table 3: *Results from fitting four-component mixtures of Poissons to each slide*

Response variables	Cont. lsmeans[a]	Sz. lsmeans[b]	F test	% diff.[c]
mean of first comp.[d]	1.392(0.025)	1.280(0.025)	$F_{1,12} = 1.79, p = 0.205$	8.02 %
mean of second comp.	7.043(0.119)	5.614(0.119)	$F_{1,12} = 2.65, p = 0.130$	20.29%
mean of third comp.	34.101(0.173)	25.699(0.173)	$F_{1,12} = 9.25, p = 0.010$	24.64%
mean of fourth comp.	93.281(0.288)	70.944(0.288)	$F_{1,12} = 8.56, p = 0.012$	23.95%
prop.[e] of first comp.	0.810	0.788	$F_{1,12} = 0.77, p = 0.399$	–
prop. of second comp.	0.137	0.160	$F_{1,12} = 0.82, p = 0.384$	–
prop. of third comp.	0.036	0.034	$F_{1,12} = 1.18, p = 0.299$	–
prop. of forth comp.	0.017	0.018	–	–

[a]Back-transformed lsmeans for control group (the standard error of the lsmeans in parrentheses)
[b]Back-transformed lsmeans for schizophrenic group (the standard error of the lsmeans in parrentheses)
[c]percent difference of the back-transformed lsmeans relative to control
[d]component
[e]proportion

variances, before carrying out repeated measures analysis. As in the two-component Poisson mixture case, for each of the 7 response variables, the primary model employed has diagnostic group as the main effect, pair, slide section as categorical factors and storage time and brain pH as the other covariates. The secondary model has diagnostic group as the main effect, slide section as a categorical factor and age, gender, PMI, storage time and Brain pH as the other covariates. In Table 3, we report the MANCOVA results for each response variable from the primary model. These results are consistent with those from the secondary model. We give the back-transformed least squares means (lsmeans) for each diagnostic group (control and schizophrenic), the results of an F test for the diagnostic effect, and the percent difference of the back-transformed least squares means relative to the control group.

We can see from the Table 3 that the back-transformed lsmeans in each of the four components decrease in subjects with schizophrenia. However, the diagnostic effects on the grain counts in the two larger components are strongly significant, while for the two smaller components, there are no significant diagnostic effects. However, the relative change of the second back-transformed component is 20.29%, which is very large, indicating that there are big differences in the second component between the two diagnostic groups, even though its F-test shows no difference due to the relative large standard errors in both groups compared

with their lsmeans. For the proportions of components, there are no differences between the two diagnostic groups.

Comparing the results in Table 3 with those in Table 1, we notice that the summation of the proportions of the first and second components in the four-component mixtures are 0.947 and 0.948, respectively, for control and schizophrenic subjects, which are close to their proportions of non-PV neurons, which are 0.946 and 0.950, given in Table 1. Furthermore, we calculated the weighted average of the mean of the first and second component, with their corresponding proportions as weights. The weighted averages are 2.208 and 1.010, respectively, for control and schizophrenic subjects, and these values are close to the lsmeans of the grain counts in non-PV neurons for both groups. Similarly, the weighted average of the third and fourth components are calculated, which are 52.678 and 41.526, respectively for control and schizophrenic groups, which are close to the lsmeans for grain counts in PV-containing neurons. These results seem to indicate that the first two components in the four-component mixtures are actually the components of the non-PV neurons. We suspect that the first component consists of the non-PV neurons which do not have PV-containing neurons nearby, and so do not contain the β particles traveling from the PV-containing neurons. In other words, the grain counts within these non-PV neurons are mostly from the natural background β particles. On the other hand, the second component consists of non-PV containing neurons which are close to the PV containing neurons, and thus these non-PV neurons contain not only the natural β particles, but also the traveling β particles from PV containing neurons. Moreover, the third and fourth components can be viewed as two types of PV containing neurons, which one could conjecture might be chandelier and arbor neurons.

The strong diagnostic effect on the grain counts in PV-containing neurons seen in Section 3.2 appears to be due to the reduction in both types of PV-containing neurons. The marginal diagnostic effect in non-PV neurons detected in Section 3.2 appears due to the reduction in the non-PV neurons nearby the PV containing neurons, which further confirms the suggested scientific explanation provided in Section 3.2.

3.4 SOME NEW RESULTS ON MIXTURES OF POISSONS

There are three main approaches to assessing the standard errors of the parameter estimates in mixture models obtained using the EM algorithm. These approaches are based on the Fisher information matrix, the observed information matrix via Louis' method, and on bootstrapping. For the normal component mixtures, Basford et. al. (1997) claimed that the standard errors obtained by bootstrap are more stable than those obtained by information-based approaches unless the sample size is very large. In this section, we compare these three methods for mixtures of Poisson components.

Let $Y_1, \ldots, Y_n$ denote a random sample of size n from a mixture of Poissons, with probability mass function given in (3.1), where we now assume that $g = 2$. The results can be extended to any finite mixture of Poissons.

The comparison of the three methods are based on a simulation study, except for assessing the standard errors of the parameter estimates by the Fisher information matrix, which can be calculated directly. Both the Louis' method and the bootstrap methods depend on simulated data. We only did the comparison when the components are mixed in equal proportions. Several combinations of true parameter values are formed by taking $p_1 = 0.5$, $\lambda_1 = 1$ and varying λ_2 from 1.2 to 7. In Table 4, we report the comparison results of the three methods when $\lambda_2 = 1.5$ and $\lambda_2 = 4$, as examples of well-separated Poisson mixture and poorly-separated Poisson mixture. A similar pattern can be seen in all combinations of parameter values we used.

For any given parameter values $p_1, \lambda_1, \lambda_2$, the Fisher information matrix $\mathcal{I}(p_1, \lambda_1, \lambda_2)$ can be calculated directly as follows. The second derivatives of the log likelihood based on (3.1) are first computed analytically, and then their expectations with respect to Y are approximated by finite sums. For instance, the first diagonal element of the Fisher information matrix is computed from

$$nE_Y \frac{\partial^2 \log L(p_1, \lambda_1, \lambda_2 | \, y)}{\partial \lambda_1^2} \approx n \sum_{y=0}^{N} \frac{\partial^2 \log L(p_1, \lambda_1, \lambda_2 | \, y)}{\partial \lambda_1^2} f(y | \, p_1, \lambda_1, \lambda_2). \tag{3.8}$$

In (3.8), $f(y | \, p_1, \lambda_1, \lambda_2) = p_1 \frac{e^{-\lambda_1} \lambda_1^y}{y!} + (1 - p_1) \frac{e^{-\lambda_2} \lambda_2^y}{y!}$; the sample size n is chosen as 1000, and N is a sufficiently large number such that $f(y | \, p_1, \lambda_1, \lambda_2) \approx 0$ for any $y \geq N$. The asymptotic

standard errors of the estimators of $p_1, \lambda_1, \lambda_2$ are obtained from (2.2). It can be seen in Table 4 that the standard errors of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ become large when λ_2 gets close to λ_1 for fixed p_1.

To avoid the tedious computation of the second derivatives of the log likelihood, Louis' method is employed to obtain the observed information matrix $I(\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2)$. For any given $p_1, \lambda_1, \lambda_2$, 250 random sample, each of size 1000, are simulated from a mixture of Poissons. For each sample, the parameter estimates $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ are obtained via the EM algorithm as shown in Section 3.2. The observed information matrix is obtained from (2.4) and the estimates of the standard errors of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ are then obtained from (2.3). It is shown that the standard error estimates from these 250 samples are very stable when the two underlying Poisson components are well separated, but are quite variable when the two components are poorly separated. For example, when $p_1 = 0.5$, $\lambda_1 = 1$, $\lambda_2 = 1.5$, the estimate of the standard error of $\hat{p}_1$ varies from 0.002 to 2535.73. The latter value corresponds to a simulated data with the estimates $\hat{p}_1 = .9964$, $\hat{\lambda}_1 = 1.235$, and $\hat{\lambda}_2 = 1.235002$. Eliminating this data set, the averages and sample standard deviations of the remaining 249 standard errors of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ are reported in Table 4. For other combinations of $p_1, \lambda_1, \lambda_2$ values, we did not encounter such a rare case, and the averages and sample standard deviations of the 250 standard errors are reported in Table 4.

To examine the behavior of the bootstrap method in estimating the standard errors of the parameter estimates, we simulate 250 random samples from a mixture of Poissons for given values of $p_1, \lambda_1, \lambda_2$. Each random sample is of size $n = 1000$ and the standard errors of the estimates of $p_1, \lambda_1, \lambda_2$ are obtained by carrying out the bootstrap procedure given in Section 2.1 with the number of bootstrap samples B chosen as 500. The means and sample standard deviation of the 250 standard errors of the estimates of $p_1, \lambda_1, \lambda_2$ are given in Table 4. It is noted that the estimates of the standard errors do not vary much even when the two underlying components are poorly separated.

To obtain the finite sample standard errors for the parameter estimates from random samples of size $n = 1000$, we also simulate 5000 random samples, each of size $n = 1000$ from mixtures of Poissons for given values of $p_1, \lambda_1, \lambda_2$. The parameter estimates $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ are obtained via the EM algorithm for each random sample. We give the sample standard deviations of the 5000 replicates of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ in Table 4, as simulated results.

Table 4: *A simulation study for comparing three approaches to computing standard errors*

Method	well-separated $p_1 = 0.5,\ \lambda_1 = 1,\ \lambda_2 = 4$				poorly-separated $p_1 = 0.5,\ \lambda_1 = 1,\ \lambda_2 = 1.5$			
	$\hat{se}(\hat{p}_1)$	$\hat{se}(\hat{\lambda}_1)$,	$\hat{se}(\hat{\lambda}_2)$	$\hat{se}(\hat{\mu})$	$\hat{se}(\hat{p}_1)$	$\hat{se}(\hat{\lambda}_1)$,	$\hat{se}(\hat{\lambda}_2)$	$se(\hat{\mu})$
Fisher Information[a]	0.038	0.096	0.160	0.069	1.914	0.95	0.98	0.036
Louis' Method[b]	0.038	0.096	0.160	0.069	0.500	0.426	0.994	0.037
	(0.004)	(0.009)	(0.013)	(0.002)	(0.669)	(0.395)	(1.028)	(0.001)
Bootstrap[c]	0.039	0.098	0.165	0.069	0.284	0.266	0.595	0.036
	(0.005)	(0.012)	(0.019)	(0.003)	(0.045)	(0.128)	(0.442)	(0.002)
Simulation[d]	0.039	0.100	0.160	0.070	0.296	0.306	0.768	0.036

[a]The entries for Fisher information are the true asymptotic standard errors of the parameter estimates.

[b]The entries for the Loius' method are the averages (and the corresponding sample standard deviations) of 250 asymptotic standard errors of parameter estimates, each of which are obtained from Louis' method for a simulated sample of size 1000. (For $p_1 = 0.5,\ \lambda_1 = 1,\ \lambda_2 = 1.5$, the presented results are the average and standard deviation of 249 simulated samples).

[c]The entries for the bootstrap are the averages (and the corresponding sample standard deviations) of 250 estimated standard errors of parameter estimates, each of which are obtained via bootstrapping for a simulated sample of size 1000.

[d]The entries for Simulation are the sample standard deviations of 5000 realizations of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$, each of which are obtained from a simulated sample of size 1000.

Comparing the standard errors obtained from the Fisher information matrix, Louis' method and the bootstrap with the simulated results, we conclude that bootstrapping provides the closest and most reasonable estimates of standard errors of parameter estimates for mixtures of Poissons when the components are poorly separated.

To better understanding the results of the comparison, for each combination of $p_1, \lambda_1, \lambda_2$, QQ-plots for the 5000 replicates of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ obtained from the simulated results are given in Appendix A. It is found that $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ are normally distributed when the two underlying Poisson components are well separated, and they are right (or left) skewed when the two components are poorly separated. This suggests that $n = 1000$ is not a sufficient sample size to make the sampling distributions of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ asymptotically normal when the two components are not far apart. Thus the information-based methods appear to require much larger sample size than 1000 to provide accurate estimates of the standard errors.

Another interesting result has to do with the expectation of Y_i, denoted by μ, where Y_i has the distribution (3.1). Since the μ is equal to $p_1\lambda_1 + (1-p_1)\lambda_2$, we estimate it by $\hat{\mu} = \hat{p}_1\hat{\lambda}_1 + (1-\hat{p}_1)\hat{\lambda}_2$. When using information-based methods to assess the standard errors of $\hat{\mu}$, we consider the asymptotic standard error of $\hat{\mu}$, obtained via the δ method as

$$\begin{aligned} se(\hat{\mu}) &= se(\hat{p}_1\hat{\lambda}_1 + (1-\hat{p}_1)\hat{\lambda}_2) \\ &= \left\{(\hat{\lambda}_1 - \hat{\lambda}_2, \hat{p}_1, 1-\hat{p}_1)\text{cov}(\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2)(\hat{\lambda}_1 - \hat{\lambda}_2, \hat{p}_1, 1-\hat{p}_1)^T\right\}^{1/2}. \end{aligned}$$

where $\text{cov}(\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2)$ denotes the asymptotic covariance matrix of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$. The $\text{cov}(\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2)$ can be calculated as the inverse of Fisher information matrix evaluated at the true parameter values, or approximated by the inverse of the observed information matrix evaluated at the true parameter values. Using the bootstrap, the standard error of $\hat{\mu}$ is given by the sample standard deviation of the B bootstrap realizations of $\hat{p}_1\hat{\lambda}_1 + (1-\hat{p}_1)\hat{\lambda}_2$. We notice in Table 4 that all methods provide almost identical estimates of the standard error of $\hat{\mu}$ regardless of the separation of the two Poisson components. In addition, it can be seen that the standard error of $\hat{\mu}$ is always small no matter what the standard errors of $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ are, which suggests that estimating μ is very stable in any situation.

As a summary of the simulation study, we find that a bootstrapping approach provides a better way to estimate the standard errors of the parameter estimates in comparison with

the information-based approaches for mixtures of Poissons where the components are poorly separated. We suspect this happens because the sample size used in our simulation study is not large enough. This conclusion is consistent with what Basford et al. (1997) reported in the normal mixture case. Furthermore, we find that the estimate of the expectation of the observations is very stable regardless of the separation of the two Poisson components.

3.5 APPLYING MIXTURES-OF-EXPERTS TO THE GRAIN COUNT DATA

With both the component densities and the mixing proportions depending on covariates, we use the EM algorithm to fit mixtures-of-experts models with two Poisson components to the grain count data.

To apply the mixtures-of-experts model to the grain count data, we only consider the grain data on slide section 1 for just the schizophrenic subjects. In fact, we can consider both schizophrenic and normal subjects by including the diagnostic effect in the mixtures-of-experts model as one of the covariates. However, in this chapter, with the main purpose of ensuring that the mixtures-of-experts model works for this grain count data set, we only use the data from the schizophrenic subjects to make the problem simpler. Later, when needed, we apply mixtures-of-experts to the whole grain count data set and the estimates are chosen as starting values for the unknown parameters in the model given in Chapter 4.

Let Y_{ij} denote the grain counts of the jth neuron from subject i, and let $\boldsymbol{x}_i$ denote the covariate vector associated with subject i, where $i = 1, \ldots, 15$; $j = 1, \ldots, l_i$. Assume that the Y_{ij}'s are independent and the density of Y_{ij} can be written as:

$$f(y_{ij}|\, \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = p(\boldsymbol{x}_i, \boldsymbol{\gamma}) f(y_{ij}|\, \boldsymbol{x}_i, \boldsymbol{\beta_1}) + (1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) f(y_{ij}|\, \boldsymbol{x}_i, \boldsymbol{\beta_2}) \tag{3.9}$$

where

$$f(y_{ij}|\, \boldsymbol{x}_i, \boldsymbol{\beta_k}) = \frac{e^{-\lambda(\boldsymbol{x}_i, \boldsymbol{\beta_k})} \lambda(\boldsymbol{x}_i, \boldsymbol{\beta_k})^{y_{ij}}}{y_{ij}!},\ k = 1, 2, \tag{3.10}$$

$$p(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}, \tag{3.11}$$

and

$$\lambda(\boldsymbol{x}_i, \boldsymbol{\beta_k}) = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k}, \ k = 1, 2. \tag{3.12}$$

In (3.11) and (3.12), $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ are the unknown parameter vectors, and $\boldsymbol{x}_i$ is the covariate vector for subject i. The log likelihood for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ is given by

$$\log L(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j=1}^{l_i} \log\{p(\boldsymbol{x}_i, \boldsymbol{\gamma}) f(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\beta_1}) + (1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) f(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\beta_2})\} \tag{3.13}$$

In order to implement the EM algorithm to estimate the parameters, we augment the data with component indicators Z_{ij}. Let $Z_{ij} = 1$ if y_{ij} comes from the first Poisson component, so that $P(Z_{ij} = 1) = p(\boldsymbol{x}_i, \boldsymbol{\gamma})$, and $Z_{ij} = 0$, otherwise with probability $1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})$. The augmented log likelihood can then be written as:

$$\begin{aligned} & \log L(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}, \boldsymbol{z}) \\ = & \sum_{i=1}^{n} \sum_{j=1}^{l_i} \{z_{ij} [\log p(\boldsymbol{x}_i, \boldsymbol{\gamma}) + \log f(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\beta_1})] \\ + & (1 - z_{ij}) [\log(1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) + \log f(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\beta_2})]\}. \end{aligned} \tag{3.14}$$

Therefore, for each iteration t, in the E–step, the conditional expectation of the Z_{ij}'s is given by:

$$\begin{aligned} \tau_{ij}^{(t)} &= E(Z_{ij} | y_{ij}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}) \\ &= \frac{p(\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}) e^{-\lambda(\boldsymbol{x}_i, \boldsymbol{\beta_1}^{(t)})} \lambda(\boldsymbol{x}_i, \boldsymbol{\beta_1}^{(t)})^{y_{ij}}}{p(\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)}) e^{-\lambda(\boldsymbol{x}_i, \boldsymbol{\beta_1}^{(t)})} \lambda(\boldsymbol{x}_i, \boldsymbol{\beta_1}^{(t)})^{y_{ij}} + (1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma}^{(t)})) e^{-\lambda(\boldsymbol{x}_i, \boldsymbol{\beta_2}^{(t)})} \lambda(\boldsymbol{x}_i, \boldsymbol{\beta_2}^{(t)})^{y_{ij}}}. \end{aligned}$$

The expectation of the augmented log likelihood can therefore be expressed as:

$$\begin{aligned} & Q(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}) \\ = & \sum_{i=1}^{n} \sum_{j=1}^{l_i} \{\tau_{ij}^{(t)} [\log(p(\boldsymbol{x}_i, \boldsymbol{\gamma})) + \log(f(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\beta_1}))] \\ + & (1 - \tau_{ij}^{(t)}) [\log(1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) + \log(f(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\beta_2}))]\}. \end{aligned} \tag{3.15}$$

In the M-step, $Q(\boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)})$ is maximized with respect to $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$. This is done by maximizing

$$\sum_{i=1}^{n} \sum_{j=1}^{l_i} \left\{ \tau_{ij}^{(t)} \log(p(\boldsymbol{x}_i, \boldsymbol{\gamma})) + (1 - \tau_{ij}^{(t)}) \log(1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) \right\} \tag{3.16}$$

with respect to $\boldsymbol{\gamma}$ and by maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{l_i}\tau_{ij}^{(t)}\log(f(y_{ij}|\,\boldsymbol{x}_i,\boldsymbol{\beta_1})) \tag{3.17}$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{l_i}(1-\tau_{ij}^{(t)})\log(f(y_{ij}|\,\boldsymbol{x}_i,\boldsymbol{\beta_2})) \tag{3.18}$$

with respect to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively. It can be seen that (3.16) has the same form as the log likelihood function of $\boldsymbol{\gamma}$ in logistic regression fitted to the response variables $\{\tau_{ij}^{(t)},\ i=1,\ldots,n,\ j=1,\ldots,l_i\}$, and so an updated $\boldsymbol{\gamma}^{(t+1)}$ can be calculated by a GLM fitting program such as the glm( ) function in S-PLUS with binomial error distribution and logit as the link function. Moreover, (3.17) is like the log likelihood function of $\boldsymbol{\beta}_1$ in a single Poisson regression fitted to the response $\{y_{ij},\ i=1,\ldots,n,\ j=1,\ldots,l_i\}$ with $\{\tau_{ij}^{(t)},\ i=1,\ldots,n,\ j=1,\ldots,l_i\}$ as known weights. The updated $\boldsymbol{\beta}_1^{(t+1)}$ can thus be computed again by the glm( ) function with Poisson error distribution and log as the link function. Similarly, $\boldsymbol{\beta}_2^{(t+1)}$ can be obtained by maximizing (3.18), which can be treated as the log likelihood function of a Poisson regression fitted to the response $\{y_{ij},\ i=1,\ldots,n,\ j=1,\ldots,l_i\}$ with $\{1-\tau_{ij}^{(t)},\ i=1,\ldots,n,\ j=1,\ldots,l_i\}$ as known weights.

The E-step and M-step are computed iteratively until $|\log L(\boldsymbol{\gamma}^{(t+1)},\boldsymbol{\beta}_1^{(t+1)},\boldsymbol{\beta}_2^{(t+1)}|\,\boldsymbol{y}) - \log L(\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_1^{(t)},\boldsymbol{\beta}_2^{(t)}|\,\boldsymbol{y})| < 10^{-5}$. The covariates included in the model are age, gender, PMI, storage time and brain pH. We run the EM algorithm with 50 different sets of starting values and compute the value of the log likelihood corresponding to the each set of starting values. The following solution is the one corresponding to the largest value of the log likelihood

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= (1.6042, -0.0003, -0.1077, -0.0024, 0.0052, 0.1762)^T \\
\hat{\boldsymbol{\beta}}_1 &= (0.2991, 0.0028, 0.0385, -0.0067, -0.0003, 0.0412)^T \\
\hat{\boldsymbol{\beta}}_2 &= (1.8670, 0.0043, -0.2041, 0.0049, -0.0032, 0.2769)^T.
\end{aligned}$$

The elements of each vector are the estimates of the intercept, the coefficients of age, gender, PMI, storage time and brain pH, respectively. The EM algorithm converges after 32 iterations.

Table 5: *Fitting two-component mixtures-of-experts to the grain count data*

	estimates		
Schizophrenia subjects	$\hat{p}_i$	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$
1	0.951	1.92	40.80
2	0.953	1.75	44.98
3	0.950	2.03	37.90
4	0.937	1.88	39.83
5	0.957	1.75	48.07
6	0.956	1.75	41.79
7	0.957	1.88	48.22
8	0.952	1.83	46.82
9	0.948	1.89	40.98
10	0.942	1.91	37.16
11	0.947	1.80	45.48
12	0.941	1.73	37.69
13	0.953	1.77	35.74
14	0.951	1.67	38.44
15	0.958	1.88	25.22

From (3.12), we calculate for each schizophrenic subject i, $\hat{p}_i$, the proportion of non-PV containing neurons, $\hat{\lambda}_{i1}$, the mean of the grain count in non-PV containing neurons, and $\hat{\lambda}_{i2}$, the mean of the grain count in PV containing neurons. These estimates are presented in Table 5. The averages (standard errors) of the $\hat{\lambda}_{i1}$, $\hat{\lambda}_{i2}$ and $\hat{p}_i$ are 1.83 (0.02), 40.61 (1.53), 0.95 (0.002) respectively, which are consistent with the results from applying the two-component mixtures of Poissons and the cut-off point technique. It can be seen that most neurons are non-PV containing neurons and that the two underlying Poisson components are well separated.

In Appendix B, we illustrate the procedures for fitting two normal component mixtures-of-experts using the EM algorithm to another neuronal postmortem brain tissue data, the neuron volume data.

4.0 MIXTURES OF GENERALIZED LINEAR MIXED MODELS (MIXTURES OF GLMMS)

4.1 INTRODUCTION

In this chapter, a new model, mixture of GLMMs is introduced for modeling repeated measurements. This model can be viewed as an extension of mixtures-of-experts for modeling repeated measurements which are observations taken on the same experimental subject and are correlated within the subject. It is a mixture model where the components are generalized linear models with random effects, and the mixing proportions are modeled by logistic or probit regression. The random effects are incorporated into the component distributions to account for the within-subject correlation present in the data.

In mixtures-of-experts (Jacobs,et al, (1991)), the observations are assumed to be independent and they follow a mixture distribution where the mixture components are typically generalized linear models, while the mixing proportions are linear logits.

The Rubin-Wu model proposed by Rubin and Wu (1997) is a two-component mixture model for repeated measures where the mixture components are linear regressions with random effects and the mixing proportions are logits. The random effects in the mixture components account for the correlation in the within-subject observations.

The model developed in this chapter extends the mixtures-of-experts by including subject specific random effects in the mixture components in order to account for the correlation in the data. The Rubin-Wu model is a special case of the mixture of GLMMs.

The motivating data is the grain count data described in detail in Section 3.1. To identify the affected subset of GABA neurons which might impair certain cognitive functions in subjects with schizophrenia, the brain tissue from 15 schizophrenic subjects and 15 normal

subjects were examined in this study. For each subject, the part of brain of interest was cut into serial sections which were hybridized with ^{35}S-labeled RNA probes. The RNA probes bind specifically with PV mRNA and emit β particles. The β particles react with the emulsion covering the slides and can be visible as grains, which are the measurements of the PV mRNA. On average, for each subject, the grain counts within 1000 neurons were counted. Since there was also background β particles, each neuron has a nonnegative grain count. Thus, we think of these grain counts as coming from two populations: PV-containing neurons and non-PV containing neurons. The diagnostic effect is the main interest in this study. The covariates age, gender, postmortem interval(PMI), storage time, and brain pH are associated with the subject. Within a subject, we treat the observations on the three slide sections as independent. Handling the correlation among the repeated sections within an individual is a more complex task, and requires an additional extension not considered in this dissertation.

Hashimoto, et al. (2003) used a cutoff point and assumed that all grain counts larger than the threshold are from the PV containing neurons. In Chapter 3, we applied classic mixtures of Poissons and mixtures-of-experts to the grain count data by ignoring the within-subject correlation. In this chapter, we propose a more appropriate model to fit such data.

After providing an overview of mixtures of GLMMs in Section 4.2, we study the joint distributions of the observed data to gain a better understanding of the structure of the data under this model. In Section 4.3, we present the normal component mixtures of GLMMs and its joint distribution. In Section 4.4, we give the Poisson component mixtures of GLMMs and outline the fitting procedures for applying the model to the grain count data. We have not implemented the sampling scheme given in Section 4.4 for the grain count data. This is left as future research.

4.2 MIXTURES OF GLMMS

4.2.1 The Model

Suppose n subjects are randomly selected, and l_i measurements are obtained on subject i. Assume the subject random effects, S_i's, are random samples from a normal distribution with mean 0 and variance σ_s^2.

Let Y_{ij} denote the jth measurement on subject i, $i = 1, \ldots, n$; $j = 1, \ldots, l_i$. Given the random effect S_i, we assume that the Y_{ij}'s are independently distributed with density

$$\begin{aligned} & f(y_{ij} | \, S_i = s_i, \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \\ = \; & p(\boldsymbol{x}_i, \boldsymbol{\gamma}) f_1(y_{ij}, \eta_1(\boldsymbol{x}_i, \boldsymbol{\beta}_1, s_i), \varphi_1) + (1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) f_2(y_{ij}, \eta_2(\boldsymbol{x}_i, \boldsymbol{\beta}_2, s_i), \varphi_2), \end{aligned} \tag{4.1}$$

where the distribution f_k is a member of the exponential family with natural parameter $\eta_k(\boldsymbol{x}_i, \boldsymbol{\beta}_k, s_i)$ and dispersion parameter φ_k. If the link function is chosen as canonical, the natural parameter is given by

$$\eta_k(\boldsymbol{x}_i, \boldsymbol{\beta}_k, s_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i. \tag{4.2}$$

In (4.1), the mixing proportion is modeled as

$$p(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}, \tag{4.3}$$

that is, $\text{logit}(p(\boldsymbol{x}_i, \boldsymbol{\gamma})) = \boldsymbol{x}_i^T \boldsymbol{\gamma}$. For this model, the parameter vectors $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2, \varphi_1, \varphi_2$ and the variance σ_s^2 are unknown. The vector of the covariates associated with subject i is $\boldsymbol{x}_i$.

4.2.2 The Marginal Distribution of the Observed Data

In mixtures of GLMMs, the distribution of $(Y_{i1}, \ldots, Y_{il_i})$, which are the observations on the ith subject, is assumed to be independent across subjects. In this section, we present the joint distribution of $(Y_{i1}, \ldots, Y_{il_i})$. To simplify the notation, we omit the subject index and denote the subject response vector $(Y_{i1}, Y_{i2}, \ldots, Y_{il_i})$ by $(Y_1, Y_2, \ldots, Y_m)$. This also leads to suppression of the covariate vector $\boldsymbol{x}_i$.

Suppose that given the subject-specific random effects, the conditional density of Y_j ($j = 1, \ldots, m$) is a two-component mixture with any kind of component distributions. Further suppose that the subject-specific random effects are assumed to have an arbitrary distribution. The following theorem gives the distribution of $(Y_1, \ldots, Y_m)$ in this general setting.

Theorem 1. *Assume that $Y_1|\, S$, $Y_2|\, S, \ldots$ $Y_m|\, S$ are i.i.d. and that the density of Y_j conditional on S can be written as*

$$f(y_j|\, S = s) = pf_1(y_j, \eta_1, \phi_1, s) + (1-p)f_2(y_j, \eta_2, \phi_2, s), \tag{4.4}$$

where f_1, f_2, f_S are any density or probability mass functions,

$$S \sim f_S(s, \sigma_s), \tag{4.5}$$

and η_1, ϕ_1, η_2, ϕ_2, σ_s are parameters. Then the joint distribution of Y_1, $Y_2, \ldots$ Y_m is given by:

$$\begin{aligned}
f_{y_1,\ldots,y_m}(y_1, \ldots, y_m) &= p^m G(y_1, \ldots, y_m; \eta_1, \eta_1, \eta_1, \ldots, \eta_1; \phi_1, \phi_1, \phi_1, \ldots, \phi_1; \sigma_s) \\
&+ (1-p)p^{m-1}[G(y_1, \ldots, y_m; \eta_2, \eta_1, \eta_1, \ldots, \eta_1; \phi_2, \phi_1, \phi_1, \ldots, \phi_1; \sigma_s) \\
&+ G(y_1, \ldots, y_m; \eta_1, \eta_2, \eta_1, \ldots, \eta_1; \phi_1, \phi_2, \phi_1, \ldots, \phi_1; \sigma_s) + \cdots \\
&+ G(y_1, \ldots, y_m; \eta_1, \eta_1, \eta_1, \ldots, \eta_2; \phi_1, \phi_1, \phi_1, \ldots, \phi_2; \sigma_s)] \\
&+ (1-p)^2 p^{m-2}[G(y_1, \ldots, y_m; \eta_2, \eta_2, \eta_1, \ldots, \eta_1; \phi_2, \phi_2, \phi_1, \ldots, \phi_1; \sigma_s) \\
&+ G(y_1, \ldots, y_m; \eta_2, \eta_1, \eta_2, \ldots, \eta_1; \phi_1, \phi_2, \phi_2, \ldots, \phi_1; \sigma_s) + \cdots \\
&+ G(y_1, \ldots, y_m; \eta_1, \eta_1, \ldots, \eta_2, \eta_2; \phi_1, \phi_1, \ldots, \phi_2, \phi_2; \sigma_s)] \\
&+ \cdots \\
&+ (1-p)^m G(y_1, \ldots, y_m; \eta_2, \eta_2, \eta_2, \ldots, \eta_2; \phi_2, \phi_2, \phi_2, \ldots, \phi_2; \sigma_s),
\end{aligned} \tag{4.6}$$

where

$$
\begin{aligned}
& G(y_1,\ldots,y_m;\eta_1,\eta_1,\eta_1,\ldots,\eta_1;\phi_1,\phi_1,\phi_1,\ldots,\phi_1;\sigma_s) \\
= & \int_s f_1(y_1,\eta_1,\phi_1,s)f_1(y_2,\eta_1,\phi_1,s)\ldots f_1(y_m,\eta_1,\phi_1,s)f_S(s,\sigma_s)d_S, \\
& G(y_1,\ldots,y_m;\eta_2,\eta_1,\eta_1,\ldots,\eta_1;\phi_2,\phi_1,\phi_1,\ldots,\phi_1;\sigma_s) \\
= & \int_s f_2(y_1,\eta_2,\phi_2,s)f_1(y_2,\eta_1,\phi_1,s)\ldots f_1(y_m,\eta_1,\phi_1,s)f_S(s,\sigma_s)d_S, \\
& G(y_1,\ldots,y_m;\eta_1,\eta_2,\eta_1,\ldots,\eta_1;\phi_1,\phi_2,\phi_1,\ldots,\phi_1;\sigma_s) \\
= & \int_s f_1(y_1,\eta_1,\phi_1,s)f_2(y_2,\eta_2,\phi_2,s)\ldots f_1(y_m,\eta_1,\phi_1,s)f_S(s,\sigma_s)d_S, \\
& \vdots \\
& G(y_1,\ldots,y_m;\eta_2,\eta_2,\eta_2,\ldots,\eta_2;\phi_2,\phi_2,\phi_2,\ldots,\phi_2;\sigma_s) \\
= & \int_s f_2(y_1,\eta_2,\phi_2,s)f_2(y_2,\eta_2,\phi_2,s)\ldots f_2(y_m,\eta_2,\phi_2,s)f_S(s,\sigma_s)d_S,
\end{aligned}
$$

There are 2^m *terms in (4.6).*

Proof: It follows directly from $f_{y_1,\ldots,y_m}(y_1,\ldots,y_m) = \int_s \left\{\prod_{j=1}^m f(y_j|\, S=s)\right\} f_S(s,\sigma_s)d_s$.

♢

4.3 NORMAL COMPONENT MIXTURES OF GLMMS

4.3.1 The Model

If the two components in (4.1) are normal distributions, the model is a mixture of GLMMs with normal components. The conditional density function of $y_{ij}|\, S_i$ is:

$$
\begin{aligned}
& f(y_{ij}|\, S_i = s_i, \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2) \\
= & \; p(\boldsymbol{x}_i,\boldsymbol{\gamma})\phi(y_{ij};\mu(\boldsymbol{x}_i,\boldsymbol{\beta}_1,s_i),\sigma_1^2) + (1-p(\boldsymbol{x}_i,\boldsymbol{\gamma}))\phi(y_{ij};\mu(\boldsymbol{x}_i,\boldsymbol{\beta}_2,s_i),\sigma_2^2), \qquad (4.7)
\end{aligned}
$$

where $\phi(y_{ij};\mu(\boldsymbol{x}_i,\boldsymbol{\beta}_k,s_i),\sigma_k^2)$ denotes the univariate normal density with mean $\mu(\boldsymbol{x}_i,\boldsymbol{\beta}_k,s_i)$ and variance σ_k^2, and $\mu(\boldsymbol{x}_i,\boldsymbol{\beta}_k,s_i)$ is given by

$$
\mu(\boldsymbol{x}_i,\boldsymbol{\beta}_k,s_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}_k + s_i, \;\; k=1,2. \qquad (4.8)
$$

The model described above is actually the Rubin-Wu model. Therefore, mixtures of GLMMs, where the mixture components can be any distribution belonging to the exponential family, can be viewed as an extension of the Rubin-Wu model, where the two mixture components are chosen as normal distributions. For more details on the Rubin-Wu model, see Section 2.3.

4.3.2 The Marginal Distribution of the Observed Data

We now give the joint distribution of the observations for each individual in the case of mixtures of GLMMs with normal components. As in Section 4.2.2, we again omit the subject index and denote the subject response vector $(Y_{i1}, Y_{i2}, \ldots, Y_{il_i})$ by $(Y_1, Y_2, \ldots, Y_m)$.

Theorem 2. *Assume that $Y_1|\, S$, $Y_2|\, S, \ldots$ $Y_m|\, S$ are i.i.d. and that the conditional density of $Y_j|\, S$ is written as*

$$f(y_j|\, S = s) = p\,\phi(y_j; \mu_1 + s, \sigma_1^2) + (1-p)\phi(y_j; \mu_2 + s, \sigma_2^2), \tag{4.9}$$

where

$$S \sim N(0, \sigma_s^2), \tag{4.10}$$

and μ_1, μ_2, σ_1^2, σ_2^2, σ_s^2 are parameters. Then the joint distribution of $(Y_1,\ Y_2, \ldots, Y_m)$ is

$$\begin{aligned}
f_{y_1,\ldots,y_m}(y_1, \ldots, y_m) &= p^m \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_1, \mu_1, \mu_1, \ldots, \mu_1)^T, \Sigma_1) \\
&+ (1-p)p^{m-1}[\phi_m(\,(y_1, \ldots, y_m)^T; (\mu_2, \mu_1, \mu_1, \ldots, \mu_1)^T, \Sigma_2) \\
&+ \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_1, \mu_2, \mu_1, \ldots, \mu_1)^T, \Sigma_3) + \cdots \\
&+ \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_1, \mu_1, \mu_1, \ldots, \mu_2)^T, \Sigma_{m+1})] \\
&+ (1-p)^2 p^{m-2}[\phi_m(\,(y_1, \ldots, y_m)^T; (\mu_2, \mu_2, \mu_1, \ldots, \mu_1)^T, \Sigma_{m+2}) \\
&+ \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_2, \mu_1, \mu_2, \ldots, \mu_1)^T, \Sigma_{m+3}) + \cdots \\
&+ \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_1, \mu_1, \ldots, \mu_2, \mu_2)^T, \Sigma_{m+m(m-1)/2})] \\
&+ \cdots \\
&+ (1-p)^m \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_2, \mu_2, \mu_2, \ldots, \mu_2)^T, \Sigma_{2^m}),
\end{aligned} \tag{4.11}$$

where $\phi_m(\boldsymbol{y}, \boldsymbol{\mu}, \Sigma)$ *is the multivariate normal density with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* Σ. *In (4.11),*

$$\Sigma_1 = \begin{bmatrix} \sigma_1^2+\sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_1^2+\sigma_s^2 & \cdots & \sigma_s^2 \\ \cdots & & & \\ \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_1^2+\sigma_s^2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} \sigma_2^2+\sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_1^2+\sigma_s^2 & \cdots & \sigma_s^2 \\ \cdots & & & \\ \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_1^2+\sigma_s^2 \end{bmatrix},$$

$$\Sigma_3 = \begin{bmatrix} \sigma_1^2+\sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2+\sigma_s^2 & \cdots & \sigma_s^2 \\ \cdots & & & \\ \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_1^2+\sigma_s^2 \end{bmatrix}, \quad \cdots,$$

$$\Sigma_{2^m} = \begin{bmatrix} \sigma_2^2+\sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2+\sigma_s^2 & \cdots & \sigma_s^2 \\ \cdots & & & \\ \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_2^2+\sigma_s^2 \end{bmatrix}. \tag{4.12}$$

It is noted that each mean vector in (4.11) is described by an element of $\times_{j=1}^{m}\{\mu_1, \mu_2\}$; while its corresponding covariance matrix has σ_s^2 as the off-diagonal elements and the diagonal elements are described by the corresponding element of $\times_{j=1}^{m}\{\sigma_1^2, \sigma_2^2\}$. If $\sigma_1^2 = \sigma_2^2$ in (4.7), then $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_{2^m}$ in Theorem 2.

Proof: By Theorem 1 and without loss of generality, we need to show that for example,

$$\begin{aligned} & \int_s \phi(y_1; \mu_1+s, \sigma_1^2)\phi(y_2; \mu_2+s, \sigma_2^2)\ldots\phi(y_m; \mu_1+s, \sigma_1^2) f_S(s, \sigma_s) d_s \\ = \; & \phi_m(\,(y_1, \ldots, y_m)^T; (\mu_1, \mu_2, \mu_1, \ldots, \mu_1)^T, \Sigma_3). \end{aligned} \tag{4.13}$$

We employ the moment generating function to show (4.13). Consider the random variables $W_1, \ldots, W_m$ and assume

$$\begin{aligned} W_j|\, S &\overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), \text{ for } j = 1, 3, \ldots, m \\ W_2|\, S &\overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2). \end{aligned}$$

where

$$S \sim N(0, \sigma_s^2).$$

We recognize that the left side of (4.13) is the joint density function of $(W_1, \ldots, W_m)$. Next we compute the moment generating function of $(W_1, \ldots, W_m)$. For any vector $\boldsymbol{t} = (t_1, t_2, \ldots, t_m)^T$,

$$\begin{aligned}
M(\boldsymbol{t}) &= E_S\left[\prod_{j=1}^{m} E(\exp\{t_j W_j\}|\, S)\right] \\
&= E_S\left[\exp\{t_1(\mu_1 + S) + \sigma_1^2 t_1^2/2\}\, \exp\{t_2(\mu_2 + S) + \sigma_2^2 t_2^2/2\} \ldots \exp\{t_m(\mu_1 + S) + \sigma_1^2 t_m^2/2\}\right] \\
&= \exp\left\{t_1\mu_1 + t_2\mu_2 + \ldots + t_m\mu_1 + \frac{\sigma_1^2}{2}(t_1^2 + t_3^2 + \ldots + t_m^2) + \frac{\sigma_2^2}{2}(t_2^2)\right\} E_S\left[\exp\left\{\sum_{j=1}^{m} t_j S\right\}\right] \\
&= \exp\left\{t_1\mu_1 + t_2\mu_2 + \ldots + t_m\mu_1 + \frac{\sigma_1^2}{2}(t_1^2 + t_3^2 + \ldots + t_m^2) + \frac{\sigma_2^2}{2}(t_2^2)\right\} \exp\left\{\frac{\sigma_s^2}{2}(\sum_{j=1}^{m} t_j)^2\right\} \\
&= \exp\left\{\boldsymbol{t}^T \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_1 \\ \cdots \\ \mu_1 \end{pmatrix} + \frac{1}{2}\boldsymbol{t}^T \begin{bmatrix} \sigma_1^2 + \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2 + \sigma_s^2 & \cdots & \sigma_s^2 \\ \cdots & & & \\ \sigma_s^2 & \sigma_s^2 & \cdots & \sigma_1^2 + \sigma_s^2 \end{bmatrix} \mathbf{t}\right\},
\end{aligned}$$

which is the moment generating function of the multivariate normal distribution with $(\mu_1, \mu_2, \mu_1, \ldots, \mu_1)^T$ as the mean vector and Σ_3 as the covariance matrix. It follows from the uniqueness of the moment generating function that (4.13) holds, and therefore Theorem 2 is true.

$\diamondsuit$

In both Theorem 1 and Theorem 2, the joint distributions of $(Y_1, Y_2, \ldots, Y_m)$ are 2^m component mixtures. These mixture models provide insight into the structure of the data under the mixtures of GLMMs, but do not aid in the estimation of the parameters. If we estimate the unknown parameters based on the joint distribution of $Y_1, \ldots, Y_m$ directly, the difficulty of the problem dramatically increases with the number of observations on each subject.

We now give an example of Theorem 2.

Example 1. *In Theorem 2, if the number of the observations $m = 2$, then the joint distribution of $(Y_1,\ Y_2)$ is*

$$\begin{aligned} f_{y_1,y_2}(y_1,y_2) &= p^2\phi_2(\,(y_1,y_2)^T;(\mu_1,\mu_1)^T,\begin{bmatrix}\sigma_1^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_1^2+\sigma_s^2\end{bmatrix}) \\ &+ (1-p)p\,\phi_2(\,(y_1,y_2)^T;(\mu_2,\mu_1)^T,\begin{bmatrix}\sigma_2^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_1^2+\sigma_s^2\end{bmatrix}) \\ &+ (1-p)p\,\phi_2(\,(y_1,y_2)^T;(\mu_1,\mu_2)^T,\begin{bmatrix}\sigma_1^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_2^2+\sigma_s^2\end{bmatrix}) \\ &+ (1-p)^2\phi_2(\,(y_1,y_2)^T;(\mu_2,\mu_2)^T,\begin{bmatrix}\sigma_2^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_2^2+\sigma_s^2\end{bmatrix}). \end{aligned}$$

4.4 POISSON COMPONENT MIXTURES OF GLMMS

4.4.1 The Model

If the two components in (4.1) are Poisson distributions, the model is a mixture of GLMMs with Poisson components, that is, the probability mass function of $Y_{ij}|\,S_i$ is

$$\begin{aligned} & f(y_{ij}|\,S_i = s_i,\boldsymbol{x}_i,\boldsymbol{\gamma},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \\ = \ & p(\boldsymbol{x}_i,\boldsymbol{\gamma})\frac{e^{-\lambda(\boldsymbol{x}_i,\boldsymbol{\beta_1},s_i)}\lambda(\boldsymbol{x}_i,\boldsymbol{\beta_1},s_i)^{y_{ij}}}{y_{ij}!} + (1-p(\boldsymbol{x}_i,\boldsymbol{\gamma}))\frac{e^{-\lambda(\boldsymbol{x}_i,\boldsymbol{\beta_2},s_i)}\lambda(\boldsymbol{x}_i,\boldsymbol{\beta_2},,s_i)^{y_{ij}}}{y_{ij}!} \end{aligned} \tag{4.14}$$

where

$$\lambda(\boldsymbol{x}_i,\boldsymbol{\beta_k},s_i) = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}_k+s_i},\ \ k=1,2. \tag{4.15}$$

4.4.2 The Expectation and Covariance Matrix of the Observations

In mixture of GLMMs with Poisson components, the marginal density of the multiple observations from each individual cannot be written as a mixture of standard densities because the density function G in (4.6) does not have a closed form in this case. To gain some insight into the joint distribution of the observed data, we give the expectation and covariance matrix of the observations on each individual under this model. We again denote the subject response vector $(Y_{i1}, Y_{i2}, \ldots, Y_{il_i})$ by $(Y_1, Y_2, \ldots, Y_m)$.

Theorem 3. *Assume $Y_1|\, S$, $Y_2|\, S, \ldots$ $Y_m|\, S$ are i.i.d. and the conditional density of $Y_j|\, S$ is*

$$f(y_j|\, S = s) = pf(y_j; \lambda_1 e^s) + (1-p)f(y_j; \lambda_2 e^s) \tag{4.16}$$

where $f(y_j; \theta_k) = e^{-\theta_k}\theta_k^{y_j}/y_j!$, $k = 1, 2$,

$$S \sim N(0, \sigma_s^2),$$

and λ_1, λ_2, σ_s^2 are parameters. Then,

$$E\{(Y_1,\, Y_2, \ldots\, Y_m)^T\} = [p\lambda_1 + (1-p)\lambda_2]e^{\sigma_s^2/2}1_{m\times 1} \tag{4.17}$$

$$Var\{(Y_1,\, Y_2, \ldots\, Y_m)^T\} = \begin{bmatrix} a+b & b & \cdots & b \\ b & a+b & \cdots & b \\ \cdots & & & \\ b & b & \cdots & a+b \end{bmatrix}_{m\times m}, \tag{4.18}$$

where

$$\begin{aligned} a &= [p\lambda_1 + (1-p)\lambda_2]e^{\sigma_s^2/2} + p(1-p)(\lambda_1 - \lambda_2)^2 e^{2\sigma_s^2} \\ b &= [p\lambda_1 + (1-p)\lambda_2]^2(e^{2\sigma_s^2} - e^{\sigma_s^2}). \end{aligned}$$

Proof: For $j = 1, \ldots, m$,

$$\begin{aligned} E(Y_j) &= E_S\{E(Y_j|\, S)\} \\ &= E_S\left\{\sum y_j[pf_1(y_j, \lambda_1 e^S) + (1-p)f_2(y_j, \lambda_2 e^S)]\right\} \\ &= E_S\{p\lambda_1 e^S + (1-p)\lambda_2 e^S\} \\ &= [p\lambda_1 + (1-p)\lambda_2]E_S(e^S) \\ &= [p\lambda_1 + (1-p)\lambda_2]e^{\sigma_s^2/2}, \end{aligned}$$

hence equation (4.17) holds.

To prove (4.18), we first show $Var(Y_j) = a+b$, for any $j = 1, \ldots, m$, using the well-known result,

$$Var(Y_j) = Var_S(E(Y_j|\, S)) + E_S(Var(Y_j|\, S)). \tag{4.19}$$

Since

$$\begin{aligned} Var_S(E(Y_j|\, S)) &= Var_S\{p\lambda_1 e^S + (1-p)\lambda_2 e^S\} \\ &= [p\lambda_1 + (1-p)\lambda_2]^2(e^{2\sigma_s^2} - e^{\sigma_s^2}) \\ &= b \end{aligned}$$

and

$$\begin{aligned} & E_S(Var(Y_j|\, S)) \\ = \; & E_S\left\{E(Y_j^2|\, S) - (E(Y_j|\, S))^2\right\} \\ = \; & E_S\left\{[p\lambda_1 + (1-p)\lambda_2]e^S + \left\{[p\lambda_1^2 + (1-p)\lambda_2^2] - [p\lambda_1 + (1-p)\lambda_2]^2\right\} e^{2S}\right\} \\ = \; & [p\lambda_1 + (1-p)\lambda_2]e^{\sigma_s^2/2} + p(1-p)(\lambda_1 - \lambda_2)^2 e^{2\sigma_s^2} \\ = \; & a \end{aligned}$$

It follows that $Var(Y_j) = a + b$. The result follows since for $j \neq j'$, it is easy to show that

$$\begin{aligned} Cov(Y_j, Y_{j'}) &= E_S\left\{E(Y_j|\, S)E(Y_{j'}|\, S)\right\} - \{E(Y_j)\}^2 \\ &= [p\lambda_1 + (1-p)\lambda_2]^2(e^{2\sigma_s^2} - e^{\sigma_s^2}) \\ &= b. \end{aligned}$$

$\diamondsuit$

In Theorem 3, the expectation and covariance of Y_j in (4.17) and (4.18) reduce to

$$E(Y_j) = p\lambda_1 + (1-p)\lambda_2, \tag{4.20}$$

$$Var(Y_j) = p\lambda_1 + (1-p)\lambda_2 + p(1-p)(\lambda_1 - \lambda_2)^2, \ j = 1, \ldots, m, \tag{4.21}$$

if and only if $\sigma_s^2 = 0$, i.e., if there is no subject random effect. Note that (4.20) and (4.20) are the expectation and variance of Y_j under classic mixtures of Poissons. Also, the covariance between Y_j and $Y_{j'}$ (for any $j' \neq j$), denoted by b in (4.18), is equal to 0. Hence, from this special case, it can be seen again that mixtures of GLMMs is an extension of classic mixture models and mixtures-of-experts by incorporating subject-specific random effects into the component distributions.

4.4.3 Applying MCMC Methods to the Poisson Component Mixtures of GLMMs

In this section, we employ MCMC methods to simulate from the posterior distribution of the parameters in the Poisson component mixtures of GLMMs given in Section 4.4.1.

4.4.3.1 The Likelihood and Conditional Distributions. Treating the component indicators and the subject-specific random effects as missing data, we obtain the augmented likelihood function, which has a simpler form than the original likelihood. Placing prior distributions on the unknown parameters, conditional distributions, for implementing the Gibbs sampler, are obtained in this section.

As before, we augment the data with indicators: $Z_{ij} = 1$ if Y_{ij} comes from the first component, and $Z_{ij} = 0$ if Y_{ij} comes from the second component. Treating the z_{ij}'s and the random effects s_i's as missing data, the augmented likelihood function is proportional to:

$$(\sigma_s^2)^{-\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{s_i^2}{2\sigma_s^2}} \prod_{j=1}^{l_i} \prod_{k=1}^{2} \left\{ p_k(\boldsymbol{x}_i, \boldsymbol{\gamma}) \exp\left\{ (\boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i) y_{ij} - e^{(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i)} \right\} \right\}^{z_{ijk}}, \tag{4.22}$$

where $p_1(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \frac{e^{\boldsymbol{x}_i^T \gamma}}{1+e^{\boldsymbol{x}_i^T \gamma}}$, $p_2(\boldsymbol{x}_i, \boldsymbol{\gamma}) = 1 - p_1(\boldsymbol{x}_i, \boldsymbol{\gamma})$, and $z_{ij1} = z_{ij}$, $z_{ij2} = 1 - z_{ij}$.

To simulate from the posterior distribution through the Gibbs sampler, we specify prior distributions on the unknown parameters and then derive conditional distributions of both the missing data and the unknown parameters. We place independent normal prior distributions on each of $\boldsymbol{\gamma}, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ with means 0 and variance matrices $\sigma_0^2 I$, where σ_0^2 is a large number. The prior on σ_s^2 is an inverse Gamma, IG(α_0, β_0). The sampling scheme for the Gibbs sampler is as follows.

1. Initialize the parameters $\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, \sigma_s^{2^{(0)}}$.

   For iterations $t = 1, 2, \cdots$:

2. Sample from $Z_{ij}|(\, y_{ij}, s_i, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2) \overset{indep}{\sim}$ Bernoulli(τ_{ij}), where

$$\begin{aligned} \tau_{ij} &= P(Z_{ij} = 1|\, y_{ij}, s_i, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2) \\ &= \frac{p_1(\boldsymbol{x}_i, \boldsymbol{\gamma}) e^{(\boldsymbol{\beta}_1^T \boldsymbol{x}_i + s_i) y_{ij} - e^{(\boldsymbol{\beta}_1^T \boldsymbol{x}_i + s_i)}}}{p_1(\boldsymbol{x}_i, \boldsymbol{\gamma}) e^{(\boldsymbol{\beta}_1^T \boldsymbol{x}_i + s_i) y_{ij} - e^{(\boldsymbol{\beta}_1^T \boldsymbol{x}_i + s_i)}} + (1 - p_1(\boldsymbol{x}_i, \boldsymbol{\gamma})) e^{(\boldsymbol{\beta}_2^T \boldsymbol{x}_i + s_i) y_{ij} - e^{(\boldsymbol{\beta}_2^T \boldsymbol{x}_i + s_i)}}}. \quad (4.23) \end{aligned}$$

3. $S_1|\,(\{Z_{ij}\}, \{Y_{ij}\}, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2),\ \ldots,\ S_n|\,(\{Z_{ij}\}, \{Y_{ij}\}, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2)$ are independent and for $i = 1, \ldots, n$,

$$f(S_i|\{z_{ij}\}, \{y_{ij}\}, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2) \propto e^{-\frac{s_i^2}{2\sigma_s^2}} \prod_{j=1}^{l_i} \prod_{k=1}^{2} \left\{ p_k(\boldsymbol{x}_i, \boldsymbol{\gamma}) \exp\{(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i) y_{ij} - e^{(\boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i)}\} \right\}^{z_{ijk}}. \quad (4.24)$$

4. Sample from

$$f(\boldsymbol{\gamma}|\{z_{ij}\}, \{y_{ij}\}, \{s_i\}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2) \propto e^{-\frac{\boldsymbol{\gamma}^T \boldsymbol{\gamma}}{2\sigma_0^2}} \prod_{i=1}^{n} \prod_{j=1}^{l_i} \prod_{k=1}^{2} p_k(\boldsymbol{x}_i, \boldsymbol{\gamma})^{z_{ijk}}. \quad (4.25)$$

5. Sample from

$$f(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2|\{z_{ij}\}, \{y_{ij}\}, \{s_i\}, \boldsymbol{\gamma}, \sigma_s^2) \propto \prod_{k=1}^{2} \exp\left\{ \sum_{i=1}^{n} \sum_{j=1}^{l_i} (y_{ij} z_{ij1} \boldsymbol{\beta}_k^T \boldsymbol{x}_i - z_{ij1} e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i + s_i}) - \frac{\boldsymbol{\beta}_k^T \boldsymbol{\beta}_k}{2\sigma_0^2} \right\}. \quad (4.26)$$

6. Sample from

$$f(\sigma_s^2|\,\{z_{ij}\}, \{y_{ij}\}, \{s_i\}, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \propto \frac{1}{(\sigma_s^2)^{\frac{n}{2}+1+\alpha_0}} e^{-\frac{1/2 \sum_{i=1}^{n} s_i^2 + \beta_0}{\sigma_s^2}}. \quad (4.27)$$

It follows that $f(\sigma_s^2|\,\{z_{ij}\},\{y_{ij}\},\{s_i\},\boldsymbol{\gamma},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ is the density of an IG$(\frac{n}{2}+\alpha_0, \frac{1}{2}\sum_{i=1}^n s_i^2+\beta_0)$ random variable.

It is noted that the conditional distributions of the Z_{ij}'s and σ_s^2 are standard densities and therefore can be sampled from directly. However, since the conditional distributions of the S_i's, $\boldsymbol{\gamma},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2$ cannot be written in closed form, samples from these distributions will be realized via a Metropolis-Hastings step with the proposal densities being multivariate t-distributions.

Next we illustrate in detail how to sample from the s_i's, $\boldsymbol{\gamma},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2$.

4.4.3.2 Sampling from the Conditional Distribution of $\boldsymbol{\gamma}$.

In mixture of GLMMs, we choose $p_i = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}$, so that the Z_{ij}'s are modeled by logistic regressions. We now follow Chib et al. (1998) and Chib and Jeliazkov (2001) to sample $\boldsymbol{\gamma}$ from (4.25). The basic idea here is to approximate $f(\boldsymbol{\gamma}|\{Z_{ij}\},\{y_{ij}\},\{S_i\},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma_s^2)$ by a multivariate t distribution with mean equal to the posterior mode and variance equal to the negative inverse of the second derivatives of the log posterior.

First, we take the logarithm of $f(\boldsymbol{\gamma}|\{Z_{ij}\},\{y_{ij}\},\{S_i\},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma_s^2)$ expressed in (4.25) and denote it by $\log f(\boldsymbol{\gamma})$. The mode of $\log f(\boldsymbol{\gamma})$ is obtained via the Newton-Raphson algorithm using the derivatives

$$\frac{\partial \log f(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = -\frac{1}{\sigma_0^2}\boldsymbol{\gamma} + \sum_{i=1}^n\sum_{j=1}^{l_i}(z_{ij} - \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}})\boldsymbol{x}_i \tag{4.28}$$

$$\frac{\partial^2 \log f(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} = -\frac{1}{\sigma_0^2}I - \sum_{i=1}^n\sum_{j=1}^{l_i}\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}})^2}\boldsymbol{x}_i\boldsymbol{x}_i^T. \tag{4.29}$$

Next we let $\boldsymbol{m_0}$ and V_0 denote the mode of $\log f(\boldsymbol{\gamma})$ and $(-\frac{\partial^2 \log f(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2}|_{\boldsymbol{m_0}})^{-1}$ respectively, and define the proposal density as

$$f_T(\boldsymbol{\gamma}|\,\boldsymbol{m_0},V_0,\upsilon) \propto |V_0|^{-1/2}\left\{1+\frac{1}{\upsilon}(\boldsymbol{\gamma}-\boldsymbol{m_0})^TV_0^{-1}(\boldsymbol{\gamma}-\boldsymbol{m_0})\right\}^{-(\upsilon+p)/2}, \tag{4.30}$$

which is the multivariate t-distribution with υ degrees of freedom (υ is a given number), location parameter vector $\boldsymbol{m_0}$ and scale matrix V_0. In (4.30), p denotes the dimension of

$\boldsymbol{\gamma}$, which is equal to 6 in our case. Using the Metropolis-Hastings algorithm described in Section 2.4.2, we propose $\boldsymbol{\gamma}^* \sim f_T(\boldsymbol{\gamma}|\, m_0, V_0, \upsilon)$ and accept it with probability

$$\alpha(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \min\left\{ \frac{f(\boldsymbol{\gamma}^*) f_T(\boldsymbol{\gamma}|\, \boldsymbol{m_0}, V_0, \upsilon)}{f(\boldsymbol{\gamma}) f_T(\boldsymbol{\gamma}^*|\, \boldsymbol{m_0}, V_0, \upsilon)},\ 1 \right\}. \tag{4.31}$$

4.4.3.3 Sampling the Random Effects s_i. We now simulate from the conditional distributions of S_i by the Metropolis-Hastings algorithm. Analogous to the case of $\boldsymbol{\gamma}$ in the logistic regression, we take the logarithm of $f(S_i|\{Z_{ij}\}, \{y_{ij}\}, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_s^2)$ expressed in (4.24) and denote it by $\log f(s_i)$. The mode of $\log f(s_i)$ is obtained via the Newton-Raphson algorithm using the derivatives

$$\frac{\partial \log f(s_i)}{\partial s_i} = -s_i/\sigma_s^2 + \sum_{j=1}^{l_i}\sum_{k=1}^{2}(y_{ij} - e^{\boldsymbol{x}_i^T\boldsymbol{\beta}_k + s_i}) z_{ijk} \tag{4.32}$$

$$\frac{\partial^2 \log f(s_i)}{\partial s_i^2} = -1/\sigma_s^2 + \sum_{j=1}^{l_i}\sum_{k=1}^{2}(-e^{\boldsymbol{x}_i^T\boldsymbol{\beta}_k + s_i}) z_{ijk}. \tag{4.33}$$

Next we let m_0 denote the mode of $\log f(s_i)$ and V_0 denote $(-\frac{\partial^2 \log f(s_i)}{\partial s_i^2}|_{m_0})^{-1}$, and define the proposal density as $f_T(s_i|\, m_0, V_0, \upsilon)$ given in (4.30), where p is the dimension of s_i, which is equal to 1 in our case. We propose $s_i^* \sim f_T(s_i|\, m_0, V_0, \upsilon)$ and accept it with probability

$$\alpha(s_i, s_i^*) = \min\left\{ \frac{f(s_i^*) f_T(s_i|\, m_0, V_0, \upsilon)}{f(s_i) f_T(s_i^*|\, m_0, V_0, \upsilon)},\ 1 \right\}. \tag{4.34}$$

4.4.3.4 Sampling from the Conditional Distribution of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$. To simplify the notation, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, and $f_k(\boldsymbol{\beta}_k) = f_k(\boldsymbol{\beta}_k|\{z_{ij}\}, \{y_{ij}\}, \{s_i\}, \boldsymbol{\gamma}, \sigma_s^2)$ for $k = 1, 2$, where

$$f_k(\boldsymbol{\beta}_k|\{z_{ij}\}, \{y_{ij}\}, \{s_i\}, \boldsymbol{\gamma}, \sigma_s^2) \propto \exp\left\{ \sum_{i=1}^{n}\sum_{j=1}^{l_i}(y_{ij} z_{ij1} \boldsymbol{\beta}_k^T \boldsymbol{x}_i - z_{ij1} e^{\boldsymbol{\beta}_k^T \boldsymbol{x}_i + s_i}) - \frac{\boldsymbol{\beta}_k^T \boldsymbol{\beta}_k}{2\sigma_0^2} \right\}. \tag{4.35}$$

Comparing with (4.26), we have

$$f(\boldsymbol{\beta}) = f_1(\boldsymbol{\beta}_1) f_2(\boldsymbol{\beta}_2). \tag{4.36}$$

Sampling $\boldsymbol{\beta}$ from its conditional distribution requires the use of a Metropolis-Hastings step again. We derive the gradient vector and Hessian matrix of the logarithm of the conditional distribution expressed as in (4.26) and we have

$$\frac{\partial \log f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left[\left(\frac{\partial \log f_1(\boldsymbol{\beta}_1)}{\partial \boldsymbol{\beta}_1} \right)^T, \left(\frac{\partial \log f_2(\boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2} \right)^T \right]^T, \tag{4.37}$$

where

$$\frac{\partial \log f_k(\boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k} = -\boldsymbol{\beta}_k/\sigma_0^2 + \sum_{i=1}^{n} \sum_{j=1}^{l_i} (y_{ij} - e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i}) z_{ijk} \boldsymbol{x}_i, \ k = 1, 2; \tag{4.38}$$

and

$$\frac{\partial^2 \log f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \begin{bmatrix} \frac{\partial^2 \log f_1(\boldsymbol{\beta}_1)}{\partial \boldsymbol{\beta}_1^2} & 0 \\ 0 & \frac{\partial^2 \log f_2(\boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2^2} \end{bmatrix}, \tag{4.39}$$

where

$$\frac{\partial^2 \log f_k(\boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k^2} = -I/\sigma_0^2 + \sum_{i=1}^{n} \sum_{j=1}^{l_i} (-e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k + s_i}) z_{ijk} \boldsymbol{x}_i \boldsymbol{x}_i^T, \ k = 1, 2. \tag{4.40}$$

We calculate the mode of $\log f(\boldsymbol{\beta})$, denoted by $\boldsymbol{m_0}$, using the Newton-Raphson algorithm; and compute $(-\frac{\partial^2 \log f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}|_{\boldsymbol{m_0}})^{-1}$, denoted by V_0. Defining the proposal density as $f_T(\boldsymbol{\beta}|\, \boldsymbol{m_0}, V_0, \upsilon)$ as in (4.30), where p is the dimension of $\boldsymbol{\beta}$, which is equal to 12 in our case, we propose $\boldsymbol{\beta}^* \sim f_T(\boldsymbol{\beta}|\, \boldsymbol{m_0}, V_0, \upsilon)$ and accept it with probability

$$\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \min \left\{ \frac{f(\boldsymbol{\beta}^*) f_T(\boldsymbol{\beta}|\, \boldsymbol{m_0}, V_0, \upsilon)}{f(\boldsymbol{\beta}) f_T(\boldsymbol{\beta}^*|\, \boldsymbol{m_0}, V_0, \upsilon)}, \ 1 \right\}. \tag{4.41}$$

In summary, to simulate from the posterior distribution of the parameters in the Poisson component mixtures of GLMMs, the Gibbs sampler will be run for N cycles beyond a burn-in of N_1 cycles after a given random starting value. Each cycle includes:

- Generate the z_{ij}'s from a Bernoulli distribution with probability τ_{ij} given in (4.23).
- Generate $\boldsymbol{\gamma}$ as described in Section 4.4.3.2.
- Generate the s_i's as described in Section 4.4.3.3.
- Generate $\boldsymbol{\beta}$ as described in Section 4.4.3.4.
- Generate σ_s^2 from the distribution IG$(\frac{n}{2} + \alpha_0, \frac{1}{2}\sum_{i=1}^{n} s_i^2 + \beta_0)$.

Rubin and Wu (1997) also employed the Gibbs sampler to estimate the model parameters. In their sampling scheme, the conditional distributions of the subject-specific random effects S_i's are standard densities and then S_i's were sampled directly. In our Poisson component mixtures of GLMMs, the conditional distributions of S_i's are more complicated and we have to implement Metropolis-Hastings steps to sample them.

4.5 EXTENSIONS

4.5.1 Using Probit Regressions to Model the Mixing Proportions

In Section 4.2, the mixing proportions are modeled as logits. An alternative to the mixing proportions is a probit model, so that (4.3) becomes $p(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \Phi(\boldsymbol{x}_i^T \boldsymbol{\gamma})$, where Φ denotes the standard normal cdf.

When we carry out MCMC methods to sample from $\boldsymbol{\gamma}$, probit modeling results in closed form distributions. However, it may not converge faster than the logit transformation. We give the sampling scheme for $\boldsymbol{\gamma}$ using the probit transformation as follows.

As proposed in Albert and Chib (1993), we incorporate unobserved normal random variables $W_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, l_i$, where $W_{ij} \sim N(\boldsymbol{x}_i^T \boldsymbol{\gamma}, 1)$. Let $Z_{ij} = 1$, if $W_{ij} > 0$; and $Z_{ij} = 0$, otherwise. It then follows that the Z_{ij}'s are independent Bernoulli random variables with probability $p_i = \Phi(\boldsymbol{x}_i^T \boldsymbol{\gamma})$.

The distribution of W_{ij}, conditional on $Z_{ij}, \boldsymbol{\gamma}$, is:

$$\begin{aligned} W_{ij} |\, Z_{ij}, \boldsymbol{\gamma} &\overset{i.i.d}{\sim} N(\boldsymbol{x}_i^T \boldsymbol{\gamma}, 1) \ \text{ truncated on the left by 0, if } Z_{ij} = 1; \\ W_{ij} |\, Z_{ij}, \boldsymbol{\gamma} &\overset{i.i.d}{\sim} N(\boldsymbol{x}_i^T \boldsymbol{\gamma}, 1) \ \text{ truncated on the right by 0, if } Z_{ij} = 0. \end{aligned} \tag{4.42}$$

Let $\boldsymbol{z}$ be the vector of $(z_{11}, \ldots, z_{1l_1}, z_{21}, \ldots, z_{2l_2}, \ldots, z_{n1}, \ldots, z_{nl_n})^T$ and $\boldsymbol{w}$ be the vector of $(w_{11}, \ldots, w_{1l_1}, w_{21}, \ldots, w_{2l_2}, \ldots, w_{n1}, \ldots, w_{nl_n})^T$. If the prior on $\boldsymbol{\gamma}$ is $N(0, \sigma_0^2 I)$, the conditional distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{Z}$ and $\boldsymbol{W}$, is given by

$$\boldsymbol{\gamma} |\, \boldsymbol{Z}, \boldsymbol{W} \sim N(\widetilde{\boldsymbol{\gamma}}, \widetilde{\Sigma}), \tag{4.43}$$

where

$$\widetilde{\boldsymbol{\gamma}} = (\sigma_0^{-2} I + \boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{w},$$
$$\widetilde{\Sigma} = (\sigma_0^{-2} I + \boldsymbol{X}^T \boldsymbol{X})^{-1},$$

and

$$\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_2^T, \ldots, \boldsymbol{x}_n^T, \ldots, \boldsymbol{x}_n^T)^T.$$

Note that if $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ is a vector of length r, then $\boldsymbol{X}$ is a vector of length $r \sum_{i=1}^n l_i$.

We now carry out the Gibbs sampler to obtain a deviate of $\boldsymbol{\gamma}$. The iterative scheme is described as follows.

Step 1: Choose as starting values $\boldsymbol{\gamma}^{(0)} = (X^T X)^{-1} X^T \boldsymbol{z}$.

Step 2: Generate $\boldsymbol{w}^{(t+1)}$ from (4.42), given $\boldsymbol{z}$ and $\boldsymbol{\gamma}^{(t)}$.

Step 3: Generate $\boldsymbol{\gamma}^{(t+1)}$ from (4.43), given $\boldsymbol{z}$ and $\boldsymbol{w}^{(t+1)}$.

Step 4: Iterate Step 2 and Step 3 .

4.5.2 Different Subject-Specific Random Effects in the Mixture Components

The mixture of GLMMs proposed in Section 4.2 has the same random effects S_i in both components. An extension of this model is to incorporate different but correlated random effects into the mixture components. Assume (S_{i1}, S_{i2})'s are random samples from a bivariate normal distribution with mean $(0, 0)$ and covariance matrix Σ, then (4.1) can be modified to

$$\begin{aligned} & f(y_{ij} | \, S_{i1} = s_{i1}, S_{i2} = s_{i2}, \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \\ = \; & p(\boldsymbol{x}_i, \boldsymbol{\gamma}) f_1(y_{ij}, \eta_1(\boldsymbol{x}_i, \boldsymbol{\beta}_1, s_{i1}), \varphi_1) + (1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) f_2(y_{ij}, \eta_2(\boldsymbol{x}_i, \boldsymbol{\beta}_2, s_{i2}), \varphi_2). \quad (4.44) \end{aligned}$$

As a special case of (4.44), we incorporate correlated random effects into the normal component mixtures of GLMMs, and give the following theorem where the number of observations on each subject is equal to 2. We denote the subject response vector (Y_{i1}, Y_{i2}) by (Y_1, Y_2).

Theorem 4. *Assume* $Y_1|\,(S_1, S_2)$, $Y_2|\,(S_1, S_2)$ *are i.i.d. and the density of* Y_j *conditional on* S_1, S_2 *is*

$$f(y_j|\,S_1 = s_1, S_2 = s_2) = p\,\phi(y_j; \mu_1 + s_1, \sigma_1^2) + (1-p)\phi(y_j; \mu_2 + s_2, \sigma_2^2), \tag{4.45}$$

where

$$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{1s}^2 & \sigma_{12s} \\ \sigma_{12s} & \sigma_{2s}^2 \end{bmatrix} \right), \tag{4.46}$$

and μ_1, μ_2, σ_1^2, σ_2^2, σ_{1s}^2, σ_{2s}^2, σ_{12s} *are parameters, then the joint distribution of* $(Y_1,\ Y_2)$ *is*

$$\begin{aligned} f_{y_1,y_2}(y_1, y_2) &= p^2\phi_2(\,(y_1, y_2)^T; (\mu_1, \mu_1)^T, \begin{bmatrix} \sigma_1^2 + \sigma_{1s}^2 & \sigma_{1s}^2 \\ \sigma_{1s}^2 & \sigma_1^2 + \sigma_{1s}^2 \end{bmatrix}) \\ &+ (1-p)p\,\phi_2(\,(y_1, y_2)^T; (\mu_2, \mu_1)^T, \begin{bmatrix} \sigma_2^2 + \sigma_{2s}^2 & \sigma_{12s} \\ \sigma_{12s} & \sigma_1^2 + \sigma_{1s}^2 \end{bmatrix}) \\ &+ (1-p)p\,\phi_2(\,(y_1, y_2)^T; (\mu_1, \mu_2)^T, \begin{bmatrix} \sigma_1^2 + \sigma_{1s}^2 & \sigma_{12s} \\ \sigma_{12s} & \sigma_2^2 + \sigma_{2s}^2 \end{bmatrix}) \\ &+ (1-p)^2\phi_2(\,(y_1, y_2)^T; (\mu_2, \mu_2)^T, \begin{bmatrix} \sigma_2^2 + \sigma_{2s}^2 & \sigma_{2s}^2 \\ \sigma_{2s}^2 & \sigma_2^2 + \sigma_{2s}^2 \end{bmatrix}). \end{aligned}$$

Proof: It is similar to the proof in Theorem 2. Here we need to use the moment generating function of $(S_1, S_2) = E(e^{t_1S_1+t_2S_2}) = \frac{1}{2}\,(t_1\ t_2) \begin{bmatrix} \sigma_{1s}^2 & \sigma_{12s} \\ \sigma_{12s} & \sigma_{2s}^2 \end{bmatrix} (t_1\ t_2)^T$.

♢

Note that Theorem 4 gives the joint distribution of (Y_1, Y_2), where each subject has only two replicates, as in Example 1 given in Section 4.3. Theorem 4 considers a more complicated structure for the subject-specific random effects. Each of the two component distributions in (4.45) has a different normally distributed random effect and the two random effects are correlated. It can be seen that Example 1 is a special case of Theorem 4 when $\sigma_{1s}^2 = \sigma_{2s}^2 = \sigma_{12s}$, i.e., $S_1 = S_2$.

Theorem 4 can be generalized to the case where the number of observations on each subject, $m \geq 3$, and again the density of $Y_1, \ldots, Y_m$ is a mixture of 2^m multivariate normal

distributions which have the same mean vectors as in (4.11) and more complex covariance matrices than those in (4.11). The mean vector of each multivariate normal component is again described by an element of $\times_{j=1}^{m}\{\mu_1, \mu_2\}$; while the diagonal elements of its corresponding covariance matrix are described by the corresponding element of $\times_{j=1}^{m}\{\sigma_1^2, \sigma_2^2\}$; and the off-diagonal elements are either σ_{1s}^2, σ_{2s}^2 or σ_{12s} depending on the structure of the mean vector.

4.5.3 Other Extensions

In mixtures of GLMMs proposed in Section 4.2, the covariates $\boldsymbol{x}_i$ for each subject can be replaced by $\boldsymbol{x}_{ij}$, which may vary with the responses within subject. Such a model can be applied to fit longitudinal data. Because in our motivating data set, all the covariates are at the subject level, we only use $\boldsymbol{x}_i$ as covariates for simplicity of notation.

The mixtures of GLMMs can be generalized to any number of finite components $g > 2$. The component distributions are modeled as before. To model the mixing proportions, the generalization of logistic regression in (2.9) or the generalization of probit regression in (2.10) and (2.11) can be employed.

The MCMC methods given in Section 4.4.3 can be applied to the Poisson component mixture of GLMMs where the covariates vary with the responses within subject by simply replacing $\{\boldsymbol{x}_i\}$ with the corresponding $\{\boldsymbol{x}_{ij}\}$. It can be modified to apply to mixtures of GLMMs with any finite component number $g > 2$ as well.

Note that in mixtures of GLMMs, there is no random effect in the mixing proportions. In Chapter 5 and Chapter 6 we introduce other new models which incorporate the subject-specific random effects into the mixing proportions.

4.6 DISCUSSION

In this chapter, we extend mixtures-of-experts by including subject-specific random effects into the mixture components to model repeated measures and propose mixtures of GLMMs.

Normal component mixture of GLMMs is actually the Rubin-Wu model. For Poisson component mixtures of GLMMs, we develop the estimation procedure using MCMC methods.

In the sampling scheme for Poisson component mixtures of GLMMs, we propose sampling the random effects s_i's and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ separately. We suspect that the Markov chain could move very slowly due to the high correlation between $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and the s_i's. If this happens in the application, we would propose sightly modifying the Metropolis-Hastings steps given in Section 4.4.3.3 and Section 4.4.3.4, and sample all s_i's and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ together.

As part of our future research, we plan to fit the Poisson component mixtures of GLMMs to the grain count data and implement the inference. The response variables Y_{ij}'s in the mixtures of GLMMs are the grain counts of the jth neuron from subject i, $i = 1, \ldots, n; j = 1, \ldots, l_i$. The covariate vector includes the indicator of diagnostic effect, age, indicator if gender, PMI, storage time, and brain pH associated with each subject. The reason for not implementing the fitting procedures to the grain count data is that the two components of the Poisson mixture in this particular data set are widely separated, as shown in Section 3.2 and in Section 3.5. The wide separation between the components poses a less challenging computational problem when fitting the mixture distribution. We plan to apply our methodology to a data set with discrete outcomes where the components of the mixture are medium or even poorly separated.

5.0 MULTIVARIATE BERNOULLI MIXTURE MODELS WITH APPLICATION TO POSTMORTEM TISSUE STUDIES IN SCHIZOPHRENIA

5.1 INTRODUCTION AND MOTIVATING EXAMPLE

5.1.1 Overview

In this chapter, we introduce a novel model for repeated measures where each repeated observation has a mixture distribution. This model is motivated by our work with neuronal postmortem brain tissue studies, where multiple neurons are sampled within a subject, and subject-level variables impact both the mixing proportions and the locations of the mixture components.

Our methodology is based on a multivariate extension of mixtures-of-experts, which is a mixture model for univariate variables proposed by Jacobs, Jordan, Nowlan and Hinton (1991). In mixtures-of-experts, the mixture components are commonly generalized linear models while the mixing proportions are modeled as multivariate linear logits. Both the mixture components and the mixing proportions are allowed to depend on covariates. Our multivariate model induces dependence by having the component indicator variables within a subject depend on both subject-specific random effects and experimental fixed effects.

In order to account for the dependence between repeated measures involving mixtures, several extensions have been recently proposed. To model multiple eye-tracking observations from susceptible schizophrenic subjects, Rubin and Wu (1997) proposed a two-component mixture model in which the components are linear regressions with random effects, and the mixing proportions are linear logits. The within-subject dependence is accounted for by

subject-specific random effects in the component distributions. We term this the Rubin-Wu model, although it is actually slightly less general than the "extra component mixture" model proposed in their paper for other purposes.

Other approaches for modeling dependent mixture response data include hidden Markov models (see McLachlan and Peel (2000), Chapter 13) and mixtures of marginal models (Rosen, Jiang and Tanner (2000)). The latter model combines the properties of mixtures-of-experts with those of generalized estimating equations (Liang and Zeger, 1986) and incorporates a working correlation matrix into each component to account for the dependence between observations on the same subject.

5.1.2 Motivating Example

One of the studies that strongly motivates our model is a neuronal postmortem tissue study comparing schizophrenic and control subjects with regard to the somal volumes of deep layer 3 pyramidal neurons in the auditory association cortex (Sweet, Pierri, Auh, Sampson, and Lewis (2003)). In subjects with schizophrenia, the precision of the auditory sensory memory is usually deficient. Earlier studies indicate that imprecision of the auditory sensory memory may be related to abnormalities in the auditory association cortex. To further explore this result, Sweet et al. (2003) examined the somal volumes of deep layer 3 pyramidal cells in the auditory association cortex (Brodmann Area 42, BA42), using postmortem brain tissues from eighteen schizophrenic subjects and eighteen normal subjects. For each subject, three slide sections containing the region BA42 were selected by systematic random sampling. To sample cells on a slide section, random systematic sampling boxes were placed on the region of interest in each section, and the sampled cell volumes were obtained using the nucleator method (Gundersen (1988)). Approximately 100 to 250 neurons were selected in this manner for each subject. In layer 3 of BA42, some of the neurons have a longer axon and project to distant cortical regions. Other neurons have a shorter axon and project to the adjoin cortical region, Brodmann Area 41 (BA41). There is evidence that neuron volume is correlated with the extent of its axonal projection. This suggests that for each subject there might be within region BA42 subgroups of neurons with different somal sizes. Sweet

et al. (2003) treated all the observed neuron volume as coming from one population and conducted a multivariate covariance analysis. They showed that the overall mean neuron volume decreases in schizophrenic subjects. However, they were not able to detect a subgroup of neurons that are affected in subjects with schizophrenia. This leads us to consider a mixture model for somal volumes from BA42, where somal volumes measured within a subject are dependent. In this area of neuroscience, it is often the case that a subject's age, gender, postmortem interval (PMI) and tissue storage time can affect neuron volumes and possibly the mixing proportions. Thus, in addition to the diagnosis main effect (schizophrenia or control), these additional covariates need to be taken into account for each subject.

In Section 5.2, we present our new model, multivariate Bernoulli mixtures of normals, and then compare it with normal component mixtures-of-experts and the Rubin-Wu model by examining the joint distribution of the observed data for each subject under each model. In Section 5.3, we develop a procedure for estimating the model parameters, using Markov chain Monte Carlo (MCMC) methods. In Section 5.4, we use our methodology to analyze the somal volume data. Simulation results are reported in Section 5.5, while possible extensions of our model and concluding remarks are given in Section 5.6.

5.2 MULTIVARIATE BERNOULLI MIXTURES OF NORMALS

5.2.1 The Model

Let Y_{ij} and $\boldsymbol{x}_{ij}$ denote, respectively, the jth observation on subject i, and the covariate vector associated with observation Y_{ij}, where $i = 1, \ldots, n$; $j = 1, \ldots, l_i$. A latent component indicator variable for each observation Y_{ij} is denoted by Z_{ij}, where Z_{ij} takes on values 0 and 1. To describe the joint distribution of $(Z_{i1}, Z_{i2}, \ldots, Z_{il_i})^T$, let $W_i, i = 1, \ldots, n$ be independent normally distributed random variables with mean 0 and variance σ_w^2. Conditional on $W_i = w_i$, we assume that the Z_{ij}'s, $j = 1, \ldots, l_i$, are independent Bernoulli random variables with mean

$$\pi(\boldsymbol{x}_{ij}, \boldsymbol{\gamma}, w_i) = \frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}}. \tag{5.1}$$

Thus, we assume that these Bernoulli means are logits which depend on the covariate vector $\boldsymbol{x}_{ij}$ and random effect w_i. Marginally, the Z_{ij}'s are correlated and $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{il_i})$ follows a multivariate Bernoulli distribution. The distribution of Y_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, l_i$, is given by

$$\begin{aligned} Y_{ij} | \, (Z_{ij} = 1) &\overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1, \sigma_1^2) \\ Y_{ij} | \, (Z_{ij} = 0) &\overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2, \sigma_2^2), \end{aligned} \tag{5.2}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2, \boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2, \sigma_w^2$ are unknown parameters. The identifiability of the model is discussed in Appendix C.

To better understand our model, we use notation for multivariate Bernoulli distributions, which have been extensively discussed in the literature. Let $F_{\boldsymbol{\theta}}(\boldsymbol{z})$, $\boldsymbol{\theta} \in \Theta$ be a family of distributions for an m-vector $\boldsymbol{Z}$, with components $Z_j = 0$ or 1, $j = 1, \ldots, m$. Such a family is called a multivariate Bernoulli distribution (with parameter $\boldsymbol{\theta}$); we denote this as MVB($\boldsymbol{\theta}$) or MVB($\boldsymbol{\theta}, \boldsymbol{x}$), when a covariate $\boldsymbol{x}$ is involved.

Cox (1972) discussed several approaches to constructing multivariate Bernoulli distributions. Our approach is based on latent structures where an underlying variable accounts for the interrelationship between the conditionally independent binary variables.

5.2.2 The Joint Distribution of the Observed Data for Each Subject

To compare our model with the normal-component mixtures-of-experts and the Rubin-Wu model, we now derive the joint distribution of the observed data for each subject. For simplicity of notation and without loss of generality, we assume two observations on each subject, i.e., $l_i = 2$.

For our model, given the subject-specific random effect W_i, the conditional density of Y_{ij} is given by $f(y_{ij} | \, W_i = w_i) = \pi_{ij} \phi(y_{ij}; \mu_{ij1}, \sigma_1^2) + (1 - \pi_{ij}) \phi(y_{ij}; \mu_{ij2}, \sigma_2^2)$, where $\phi(\cdot \, ; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2, and π_{ij} is as in (5.1), suppressing

$\boldsymbol{x}_{ij}, \boldsymbol{\gamma}$, and w_i. It follows that, the joint distribution of (Y_{i1}, Y_{i2}) is

$$
\begin{aligned}
& f_{y_{i1},y_{i2}}(y_{i1}, y_{i2}) \\
= & \int_{w_i} \left\{ \prod_{j=1}^{2} f(y_{ij} |\, W_i = w_i) \right\} \phi(w_i; 0, \sigma_w^2) dw_i \\
= & \left\{ \int \pi_{i1}\pi_{i2}\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\, (y_{i1}, y_{i2})^T; (\mu_{i11}, \mu_{i21})^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
+ & \left\{ \int (1-\pi_{i1})\pi_{i2}\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\, (y_{i1}, y_{i2})^T; (\mu_{i12}, \mu_{i21})^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
+ & \left\{ \int \pi_{i1}(1-\pi_{i2})\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\, (y_{i1}, y_{i2})^T; (\mu_{i11}, \mu_{i22})^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \\
+ & \left\{ \int (1-\pi_{i1})(1-\pi_{i2})\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\, (y_{i1}, y_{i2})^T; (\mu_{i12}, \mu_{i22})^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}),
\end{aligned}
\tag{5.3}
$$

where $\phi_2(\cdot\,; \boldsymbol{\mu}, \Sigma)$ is the bivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ and

$$
\mu_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k, \;\; j = 1, 2 \text{ and } k = 1, 2. \tag{5.4}
$$

For mixtures-of-experts (Jacobs et al., 1991), the latent component-indicator Z_{ij}'s are independent Bernoulli random variables with mean $p_{ij} = \frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma}}}$, so that, the joint distribution

of $(Y_{i1},\ Y_{i2})$ is

$$
\begin{aligned}
f_{y_{i1},y_{i2}}(y_{i1}, y_{i2}) &= \prod_{j=1}^{2} f(y_{ij}) \\
&= p_{i1}p_{i2}\phi_2(\,(y_{i1}, y_{i2})^T; (\mu_{i11}, \mu_{i21})^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
&+ (1-p_{i1})p_{i2}\phi_2(\,(y_{i1}, y_{i2})^T; (\mu_{i12}, \mu_{i21})^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
&+ p_{i1}(1-p_{i2})\phi_2(\,(y_{i1}, y_{i2})^T; (\mu_{i11}, \mu_{i22})^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \\
&+ (1-p_{i1})(1-p_{i2})\phi_2(\,(y_{i1}, y_{i2})^T; (\mu_{i12}, \mu_{i22})^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}), \qquad (5.5)
\end{aligned}
$$

where μ_{ijk} are the same as in (5.4).

In the Rubin-Wu model, the latent component-indicator random variable Z_{ij} is the same as in mixtures-of-experts. Given Z_{ij} and a normally distributed subject-specific random effect S_i with mean 0 and variance σ_S^2, where $S_1, \ldots, S_n$ are independent, the conditional distribution of Y_{ij} is:

$$
\begin{aligned}
(Y_{ij} |\, Z_{ij} = 0, S_i = s_i) &\overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + s_i, \sigma_1^2), \\
(Y_{ij} |\, Z_{ij} = 1, S_i = s_i) &\overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + s_i, \sigma_2^2),
\end{aligned}
$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}$, σ_1^2, σ_2^2, σ_S^2 are the unknown parameters. In this case, the joint distribution

of $(Y_{i1},\ Y_{i2})$ is

$$
\begin{aligned}
f_{y_{i1},y_{i2}}(y_{i1},y_{i2}) &= \int_{s_i}\left\{\prod_{j=1}^{2} f(y_{ij}|\ S_i=s_i)\right\}\phi(s_i;0,\sigma_s^2)ds_i \\
&= p_{i1}p_{i2}\phi_2(\ (y_{i1},y_{i2})^T;(\mu_{i11},\mu_{i21})^T,\begin{bmatrix}\sigma_1^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_1^2+\sigma_s^2\end{bmatrix}) \\
&+ (1-p_{i1})p_{i2}\phi_2(\ (y_{i1},y_{i2})^T;(\mu_{i12},\mu_{i21})^T,\begin{bmatrix}\sigma_2^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_1^2+\sigma_s^2\end{bmatrix}) \\
&+ p_{i1}(1-p_{i2})\phi_2(\ (y_{i1},y_{i2})^T;(\mu_{i11},\mu_{i22})^T,\begin{bmatrix}\sigma_1^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_2^2+\sigma_s^2\end{bmatrix}) \\
&+ (1-p_{i1})(1-p_{i2})\phi_2(\ (y_{i1},y_{i2})^T;(\mu_{i12},\mu_{i22})^T,\begin{bmatrix}\sigma_2^2+\sigma_s^2 & \sigma_s^2\\ \sigma_s^2 & \sigma_2^2+\sigma_s^2\end{bmatrix}),
\end{aligned}
\tag{5.6}
$$

where the μ_{ijk}'s are again as in (5.4).

Comparing the joint densities of the observed data under each model in (5.3), (5.5) and (5.6), we see that all three density functions are four-component mixtures of multivariate normals. In our model, the covariance matrices of the multivariate normal components are the same as those in the normal-component mixtures-of-experts; they have independent components with $\times_{j=1}^{2}\{\sigma_1^2,\sigma_2^2\}$ describing all possible variances. In the Rubin-Wu model, the covariance matrix of each multivariate normal component has a compound-symmetric structure with σ_s^2 as the off-diagonal elements with variances described by $\times_{j=1}^{2}\{\sigma_1^2+\sigma_s^2,\sigma_2^2+\sigma_s^2\}$. As shown in (5.5) and (5.6), the mixing proportions of the joint densities in the normal-component mixtures-of-experts and those in the Rubin-Wu model correspond to independent random variables. In (5.3), the mixing proportions correspond to dependent multivariate Bernoulli random variables.

These results can easily be extended to m observations on each subject, and all the density functions can be written as 2^m mixtures of multivariate normals.

5.3 INFERENCE

5.3.1 Augmented Likelihood and Prior Distributions

The hierarchical nature of our model lends itself naturally to Bayesian estimation via Markov chain Monte Carlo (MCMC) methods.

We augment the observed data with the component indicators Z_{ij}, $i = 1, \ldots, n, j = 1, \ldots, l_i$, and the subject-specific random effects w_i, $i = 1, \ldots, n$. Let $\boldsymbol{y} = (y_{11}, \ldots, y_{1l_1}, \ldots, y_{n1}, \ldots, y_{nl_n})^T$, $\boldsymbol{z} = (z_{11}, \ldots, z_{1l_1}, \ldots, z_{n1}, \ldots, z_{nl_n})^T$, and $\boldsymbol{w} = (w_1, \ldots, w_n)^T$, so that the augmented likelihood is proportional to

$$(\sigma_w^2)^{-\frac{n}{2}} \prod_{i=1}^{n} \exp\{-\frac{w_i^2}{2\sigma_w^2}\}$$
$$\prod_{j=1}^{l_i} \left[\pi_{ij} \frac{1}{\sqrt{\sigma_1^2}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1)^2}{2\sigma_1^2}\}\right]^{z_{ij}} \left[(1 - \pi_{ij}) \frac{1}{\sqrt{\sigma_2^2}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2)^2}{2\sigma_2^2}\}\right]^{1-z_{ij}}.$$

We place independent normal prior distributions on $\boldsymbol{\gamma}, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ with means 0 and variance matrices $\sigma_\gamma^2 I_{q\times q}$, $\sigma_{\beta_1}^2 I_{q\times q}$, and $\sigma_{\beta_2}^2 I_{q\times q}$ respectively, where q is the length of the covariate vector. The priors on σ_1^2, σ_2^2 and σ_w^2 are taken to be independent inverse Gamma distributions, denoted by IG(α_1, δ_1), IG(α_2, δ_2), IG(α_w, δ_w), respectively, and the priors on $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ are assumed independent of those on σ_1^2, σ_2^2, and σ_w^2. To obtain vague priors, the values of σ_γ^2, $\sigma_{\beta_1}^2$, and $\sigma_{\beta_2}^2$ are assumed large, while α_1, δ_1, α_2, δ_2, α_w, and δ_w are set to small values.

5.3.2 The Sampling Scheme

The Gibbs sampler is used for sampling from the posterior distribution of the parameters. A Metropolis-Hastings step (Hastings, 1970) is performed for nonstandard conditional distributions. To achieve good mixing, we treat $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ as a block, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as another block. Sampling individually from the full conditional distributions of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ and the random effects w_i's, results in slow convergence, because of high correlation between $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$, as well as between $\boldsymbol{\gamma}$ and $\boldsymbol{w}$. The sampling scheme we propose is as follows.

1. Initialize the parameters $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, \sigma_1^{2^{(0)}}, \sigma_2^{2^{(0)}}, \boldsymbol{\gamma}^{(0)}, \sigma_w^{2^{(0)}}$ and $\boldsymbol{w}^{(0)}$.

   For iterations $t = 1, 2, \cdots$:

2. Sample the component-indicators $z_{ij}^{(t+1)}$, $i = 1, \ldots, n, j = 1, \ldots, l_i$ from a Bernoulli random variable with mean $\tau_{ij}^{(t)}$, where

$$\tau_{ij}^{(t)} = \frac{\frac{1}{\sigma_1^{(t)}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1^{(t)})^2}{2\sigma_1^{2(t)}} + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma}^{(t)} + w_i\}}{\frac{1}{\sigma_1^{(t)}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1^{(t)})^2}{2\sigma_1^{2(t)}} + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma}^{(t)} + w_i^{(t)}\} + \frac{1}{\sigma_2^{(t)}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2^{(t)})^2}{2\sigma_2^{2(t)}}\}}, \tag{5.7}$$

3. Sample $\sigma_1^{2(t+1)}$ from $IG(\frac{1}{2}\sum_{i=1}^n l_i + \alpha_1, \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^{l_i} z_{ij}^{(t+1)}(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1^{(t)})^2 + \delta_1)$.
4. Sample $\sigma_2^{2(t+1)}$ from $IG(\frac{1}{2}\sum_{i=1}^n l_i + \alpha_2, \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^{l_i} (1 - z_{ij}^{(t+1)})(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2^{(t)})^2 + \delta_2)$.
5. Sample $\sigma_w^{2(t+1)}$ from $IG(\frac{n}{2} + \alpha_w, \frac{1}{2}\sum_{i=1}^n w_i^{2(t)} + \delta_w)$.
6. Sample $(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{w}^{(t+1)})$ as a block from their conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{w} | \boldsymbol{y}, \boldsymbol{z}^{(t+1)}, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}, \sigma_1^{2(t+1)}, \sigma_2^{2(t+1)}, , \sigma_w^{2(t+1)})$ via a Metropolis-Hastings step.
7. Sample $(\boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)})$ as a block from their conditional distribution $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \boldsymbol{y}, \boldsymbol{z}^{(t+1)}, \sigma_1^{2(t+1)}, \sigma_2^{2(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{w}^{(t+1)}, \sigma_w^{2(t+1)})$ via a Metropolis-Hastings step.

Details of the Metropolis-Hastings steps are given in Appendix D.

When implementing this sampling scheme, some of the updated τ_{ij} given in (5.7) might be close to 0 or 1, implying that no observations, corresponding to some values of the covariates, are allocated to a mixture component. If this happens, the Markov chain may move very slowly. To avoid this problem, we choose starting values $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, \sigma_1^{2^{(0)}}, \sigma_2^{2^{(0)}}, \boldsymbol{\gamma}^{(0)}, \sigma_w^{2^{(0)}}$ and $\boldsymbol{w}^{(0)}$ that guarantee that the initial τ_{ij}'s are away from 0 or 1, for all i and j. In the Metropolis-Hastings steps, the tuning constants of the proposal distributions, described in Appendix D, are selected such that the acceptance ratios for drawing the unknown parameters are larger than 0.20. Some starting values may also result in low acceptance ratios, leading to slow convergence. For this reason, we examine the acceptance ratios in short preliminary runs to obtain the tuning constants and the appropriate starting values.

5.4 APPLICATION

In this section, we apply our model to the neuron volume data described in Section 5.1.2, where a randomly chosen neuron can be viewed as coming from one of two populations: smaller neurons or larger neurons. Since somal volume distributions are typically right skewed, neuron volumes are first log-transformed. We then treat the transformed somal volumes the y_{ij}, $i = 1, \ldots, n, j = 1, \ldots, l_i$ as a mixture with two normal components and fit our model. Each subject's covariate vector $\boldsymbol{x_{ij}}$ consists of an intercept, an indicator of diagnostic group (normal=1, schizophrenic=2), age, gender (Female=1, male=2), postmortem interval (PMI), and the corresponding tissue storage time.

To obtain vague prior distributions, we set $\sigma^2_\gamma = \sigma^2_{\beta_1} = \sigma^2_{\beta_2} = 10$ in the normal priors on $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$, and take $\alpha_1 = \alpha_2 = \alpha_w = 0.01$ and $\delta_1 = \delta_2 = \delta_w = 0.02$ in the inverse Gamma priors on σ^2_1, σ^2_2, and σ^2_w. The values used for the tuning constants in the Metropolis-Hastings steps are reported in Appendix D. Random initial values are selected such that $0.01 < \tau^{(0)}_{ij} < 0.99$, for all i and j, where $\tau^{(0)}_{ij}$ is the initial probability of $Z_{ij} = 1$, given in (5.7). The MCMC algorithm is run for 13,000 cycles after a burn-in period of 2000 iterations, although the chain converges quickly and starts stabilizing after around 500 iterations. The algorithm is run three times starting from three different sets of random initial values. The results from all three runs agree very closely.

Table 6 presents the posterior means and 95% credible intervals for the different parameters. The posterior mean provides an estimate of each parameter, obtained as the average of the sample values excluding the burn-in iterations. The 95% credible interval for each parameter is obtained by ordering the sample values after discarding the burn-in samples, and finding the 0.025 and 0.975 sample quantiles. The "smaller neuron population" and "larger neuron population" correspond to the two normal components in the model. The "Mixing proportions" refer to the proportions of the smaller neurons.

For each of the diagnostic main effects, a p-value is also obtained by finding two times the posterior probability of the event. For the smaller and larger neuron populations, the p-values of the diagnostic effects are, respectively, < 0.001 and 0.18; for the mixing proportion, the p-value of the diagnostic effect is 0.58.

Table 6: *Results of model fitting to the neuron volume data. Estimates (posterior means) and 95% credible intervals. Results are based on 13,000 iterations after 2,000 burn-in iterations.*

	2.5%	mean	97.5%
Smaller neuron population			
intercept (β_{10})	6.998	7.302	7.622
diagnostic	-0.195	-0.133	-0.073
age	-0.0026	0.0021	0.0064
gender	0.019	0.010	0.178
PMI	-0.024	-0.017	-0.009
storage time	-0.00009	-0.00004	0.00001
σ_1^2	0.283	0.310	0.338
Larger neuron population			
intercept (β_{20})	6.662	7.282	7.875
diagnostic	-0.230	-0.095	0.040
age	0.0017	0.0116	0.0212
gender	-0.00033	0.170	0.346
PMI	-0.0143	0.0003	0.0141
storage time	-0.00009	0.00001	0.00011
σ_2^2	0.588	0.639	0.692
Mixing proportions			
intercept (γ_0)	-2.691	0.729	3.872
diagnostic	-0.581	0.200	0.96
age	-0.0599	-0.0079	0.0428
gender	0.275	1.183	2.136
PMI	-0.149	-0.062	0.023
storage time	-0.00122	-0.00058	0.00000
σ_w^2	0.416	0.823	1.519

The results from this analysis directly address the question in which Sweet et al. (2003) were interested. It is seen that for the smaller neuron population, the 95% credible interval for the diagnostic effect does not include zero ($p < 0.000$), indicating a significant diagnostic effect. The negative estimate indicates that subjects with schizophrenia have smaller volumes than controls for this population of neurons. For the larger neuron population there is no significant diagnostic effect ($p = 0.18$), and for the proportion of smaller neurons (versus larger), there is no significant diagnostic difference ($p = 0.58$). Our results suggest that the overall reduction for schizophrenic subjects in somal volume seen originally by Sweet et al. in the deep layer 3 pyramidal neurons (BA42) appears to be due to a reduction in somal volume of this region's smaller pyramidal neurons, a population presumably of locally projecting neurons. To further confirm this statement, other neurological studies need to be conducted.

Although of much less scientific interest, there were several other significant parameters. In the smaller neuron population, in addition to the strong diagnostic effect, there are significant gender and PMI effects and storage time effect is marginally significant. In the larger neuron population, the age effect is significant, and the gender effect is marginally significant. Moreover, for the mixing proportions, gender has a significant effect, whereas storage time has a marginal effect. The male gender is associated not only with increased neuron volumes in both smaller and larger neuron populations, but also with increased mixing proportions of smaller neurons. Increased PMI is associated with decreased neuron volumes in smaller neuron population. Longer storage time is connected to decreased neuron volumes in smaller neuron population and decreased mixing proportions of smaller neurons. Increased age is connected to increased neuron volumes in larger neuron population.

Notice that the covariate vectors $\boldsymbol{x}_{ij}$'s in our simulated data set only contain between-subject factors. Thus all neurons belonging to the same subject have a common mean for the first component, a common mean for the second component, and a common proportion for the first component, denoted by μ_{i1}, μ_{i2}, and p_i respectively. Hence, we can obtain the posterior means for μ_{i1}, μ_{i2}, and p_i for each subject. The dot plots of the estimates of μ_{i1}, μ_{i2}, and p_i by diagnostic group are given in Figure 1. In Figure 1, we also provide the dot plot of the overall mean of the log-transformed neuron volumes for each subject. It can be seen

from these dot plots that the somal volumes in both populations decrease for schizophrenic subjects. However, the overall reduction comes mainly from the smaller neurons.

In addition, we calculate the percent difference of the back-transformed means relative to the control group for each population, that is, $(\exp(\mathrm{C}) - \exp(\mathrm{S}))/\exp(\mathrm{C})$, where C denotes the average of the posterior means across subjects for the control group and S denotes the average of posterior means across subjects for the schizophrenic group. For the smaller and larger neuron populations, the percent differences are 15.30% and 8.86% respectively. A reanalysis of the observed neuron volumes was done treating all volumes as coming from one population. We employed a multivariate analysis of covariance (MANCOVA) with diagnostic group as the main effect, subject as random effect, age, gender, PMI and tissue storage time as covariates. This analysis yielded that the percent difference of the back-transformed least squares means relative to the control group is 13.4%. These results confirm our preceding findings.

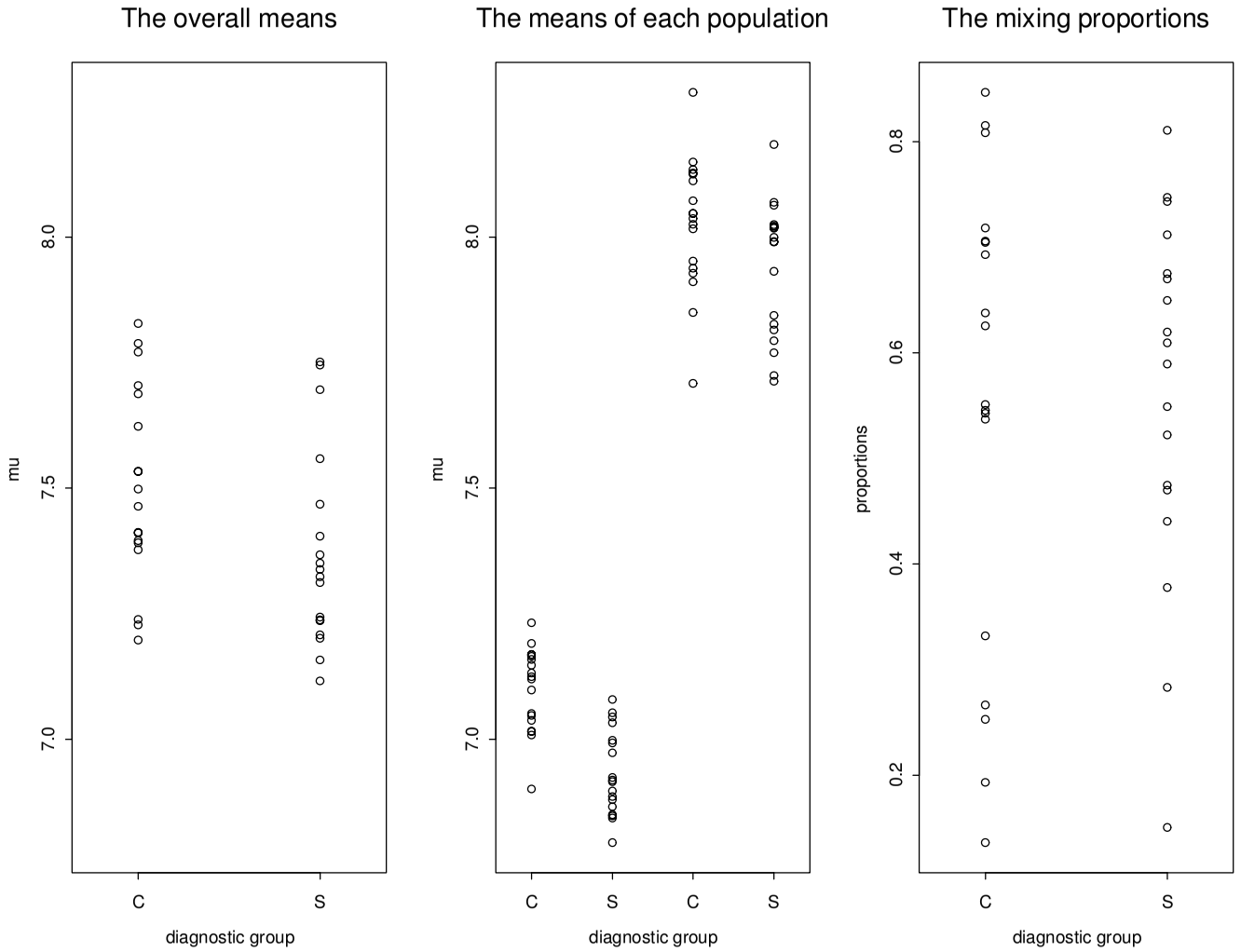

Figure 1: *For each subject, the overall mean of the log-transformed somal volume, the posterior mean of each of the smaller neuron population and larger neuron population, and the mixing proportion of smaller neurons vs. larger neurons.*

5.5 SIMULATION STUDY

In this section we report the results of a simulation study conducted to ensure that the results in Section 5.4, are valid. We simulate 30 data sets, each of which has the "same" data structure as the neuron volume data. That is, in each data set, there are 36 subjects, each having the same number of repeated measures as in the neuron volume data. The response variable y_{ij}, $i = 1, \ldots, 36, j = 1, \ldots, l_i$, in each simulated data set can be viewed as a simulated log-transformed neuron volume for the jth neuron in subject i. The corresponding covariate vector $\boldsymbol{x_{ij}}$ in the simulated data set is the same as the covariate vector in the neuron volume data.

Three sets of true parameter values for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, $\boldsymbol{\gamma}$ and σ_w^2 (See table 7) are first chosen representing cases of well-separated, medium-separated, and poorly-separated multivariate Bernoulli mixture of normals. Based on McLachlan et al. (2000, p 9) and Schilling, Watkins, and Watkins (2002), the separations of the normal components can be assessed by $\Delta = |\mu_1 - \mu_2|/(\sigma_1 + \sigma_2)$ if the two components have means μ_1, μ_2 and variances σ_1, σ_2 respectively. When $\sigma_2/\sigma_1 = 0.80$, Schilling et al. (2002) pointed out that the mixture density is bimodal if and only if $\Delta > 1.35, 1.26, 1.15, 1.01, 1.29$, respectively corresponding to the mixing proportions $p = 0.3, 0.4, 0.5, 0.6, 0.7$. When $\sigma_2/\sigma_1 = 0.90$, the mixture density is bimodal if and only if $\Delta > 1.36, 1.25, 1.11, 1.16, 1.34$, respectively corresponding to the mixing proportions $p = 0.3, 0.4, 0.5, 0.6, 0.7$. In our simulated data, each neuron can be viewed as coming from a normal mixture where the components have means μ_{ij1}, μ_{ij2} and variances σ_1, σ_2 respectively, where $\mu_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k, k = 1, 2$. The separation of the normal components for each neuron can be assessed by $\Delta_{ij} = |\mu_{ij1} - \mu_{ij2}|/(\sigma_1 + \sigma_2)$. Therefore, we assess the separation of the normal components in the simulated data by looking at the average of Δ_{ij} across all the neurons, denoted by $\bar{\Delta}$. The $\bar{\Delta}$'s corresponding to the three chosen sets of true parameter values are 4.1, 1.9, 1.1 respectively and the ratios of σ_2 to σ_1 are 0.79, 0.89 and 0.89 respectively. Moreover, for each chosen set of true parameters, the corresponding mixing proportion for each neuron has a wide range. Comparing their $\bar{\Delta}$'s with the minimum bimodal thresholds 1.01, 1.11, and 1.11, shows that the three sets of true values of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, $\boldsymbol{\gamma}$ and σ_w^2 correspond to well-separated, medium-separated, and

Table 7: *The average of estimates and percentages of coverage (cover) of 10 runs for each of well-separated, medium-separated, and poorly-separated cases in the simulation study. (Mean square errors are in parentheses.)*

Parameters	well-separated			medium-separated			poorly-separated		
	true	estimate	cover	true	estimate	cover	true	estimate	cover
Smaller neurons									
intercept (β_{10})	0.10	0.01(0.08)	1.0	2.70	2.77(0.16)	1.0	2.90	3.15(0.31)	0.9
diagnostic	0.30	0.30(0.01)	0.8	0.01	0.03(0.00)	1.0	0.08	0.10(0.01)	0.9
age	-0.50	-0.50(0.00)	1.0	0.08	0.08(0.00)	1.0	0.08	0.08(0.00)	1.0
gender	2.60	2.64(0.02)	0.8	-0.05	-0.08(0.01)	0.8	-0.01	-0.03(0.01)	0.9
PMI	0.80	0.80(0.00)	0.9	0.02	0.02(0.00)	0.8	0.04	0.04(0.00)	0.9
storage time	-0.00	-0.00(0.00)	1.0	-0.00	-0.00(0.00)	0.9	-0.00	-0.00(0.00)	1.0
σ_1^2	8.00	8.06(0.04)	0.9	2.50	2.51(0.01)	0.9	2.50	2.51(0.01)	1.0
Larger neurons									
intercept (β_{20})	0.01	0.17(0.08)	1.0	1.35	1.38(0.03)	1.0	1.35	1.34(0.03)	1.0
diagnostic	1.20	1.25(0.02)	0.9	0.12	0.14(0.00)	1.0	0.12	0.11(0.00)	1.0
age	-0.10	-0.10(0.00)	1.0	0.13	0.13(0.00)	1.0	0.10	0.10(0.00)	1.0
gender	2.00	1.98(0.02)	1.0	-0.10	-0.09(0.01)	0.8	-0.10	-0.09(0.00)	1.0
PMI	0.90	0.90(0.00)	1.0	0.13	0.13(0.00)	1.0	0.11	0.11(0.00)	1.0
storage time	-0.00	-0.00(0.00)	1.0	-0.00	-0.00(0.00)	0.9	-0.00	-0.00(0.00)	0.9
σ_2^2	5.00	5.01(0.01)	1.0	2.00	2.02(0.00)	1.0	2.00	2.02(0.01)	0.8
Mixing prop.									
intercept (γ_0)	-1.90	-1.53(2.71)	1.0	1.90	1.65(2.59)	1.0	1.90	1.10(4.03)	1.0
diagnostic	0.65	0.62(0.08)	1.0	-0.65	-1.04(0.61)	0.9	-0.65	-0.39(0.42)	1.0
age	0.13	0.12(0.00)	1.0	-0.13	-0.13(0.00)	1.0	-0.13	-0.12(0.00)	1.0
gender	-1.50	-1.37(0.74)	0.8	1.50	1.25(0.46)	1.0	1.50	1.31(0.67)	1.0
PMI	-0.21	-0.23(0.00)	1.0	0.21	0.25(0.02)	0.8	0.21	0.21(0.01)	0.9
storage time	0.00	0.00(0.00)	1.0	-0.00	-0.00(0.00)	0.9	-0.00	-0.00(0.00)	0.9
σ_w^2	2.00	2.09(0.24)	0.9	4.00	4.10(1.14)	1.0	4.00	4.77(3.11)	1.0

poorly-separated mixtures of normals respectively.

Corresponding to each set of given true parameter values, we simulate 10 data sets from our model. The Gibbs sampler for each data set takes more than 10 hours to run. We fit our model, to each simulated data set by using the sampling scheme described in Section 5.3.2. The posterior mean and 95% credible interval for each parameter are obtained based on 8000 iterations after a burn-in period. For most simulated data sets, the Markov chain stabilizes after 1000 iterations and so the burn-in period contains 1000 cycles. For some simulated data sets corresponding to poorly-separated mixtures, the algorithm requires more burn-in cycles, for example 3000 iterations, to achieve convergence.

For each given set of true values of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, σ_1^2, σ_2^2, $\boldsymbol{\gamma}$ and σ_w^2, the average values, and the mean square errors of each parameter across the 10 runs are obtained, and the coverage rates of the 95% credible intervals for each true parameter are calculated.

Table 7 presents the simulation results. For each setting of well-separated, medium-separated, and poorly-separated mixtures of normals, we report the average of the estimates, the mean square error, and the coverage rate over 10 runs for each parameter. It can be seen from Table 7 that the estimates have little bias, and the coverage rates are reasonable. The mean square errors are overall small, except for estimating the intercept in the mixing proportions and σ_w^2. For the well-separated case, Figure 2 gives dot plots of the estimates based on the 10 runs for the diagnostic, age, gender, PMI, and storage time effects for each of the smaller and larger neuron populations and the mixing proportions. Figure 3 and Figure 4 are respectively, dot plots for the medium-separated and poorly-separated mixtures of normals. From the dot plots, we can also see that the estimates are fluctuating roughly symmetrically around all the true values. The simulation results suggest that the estimates and the inference obtained in Section 5.4 for the neuron volume data are valid.

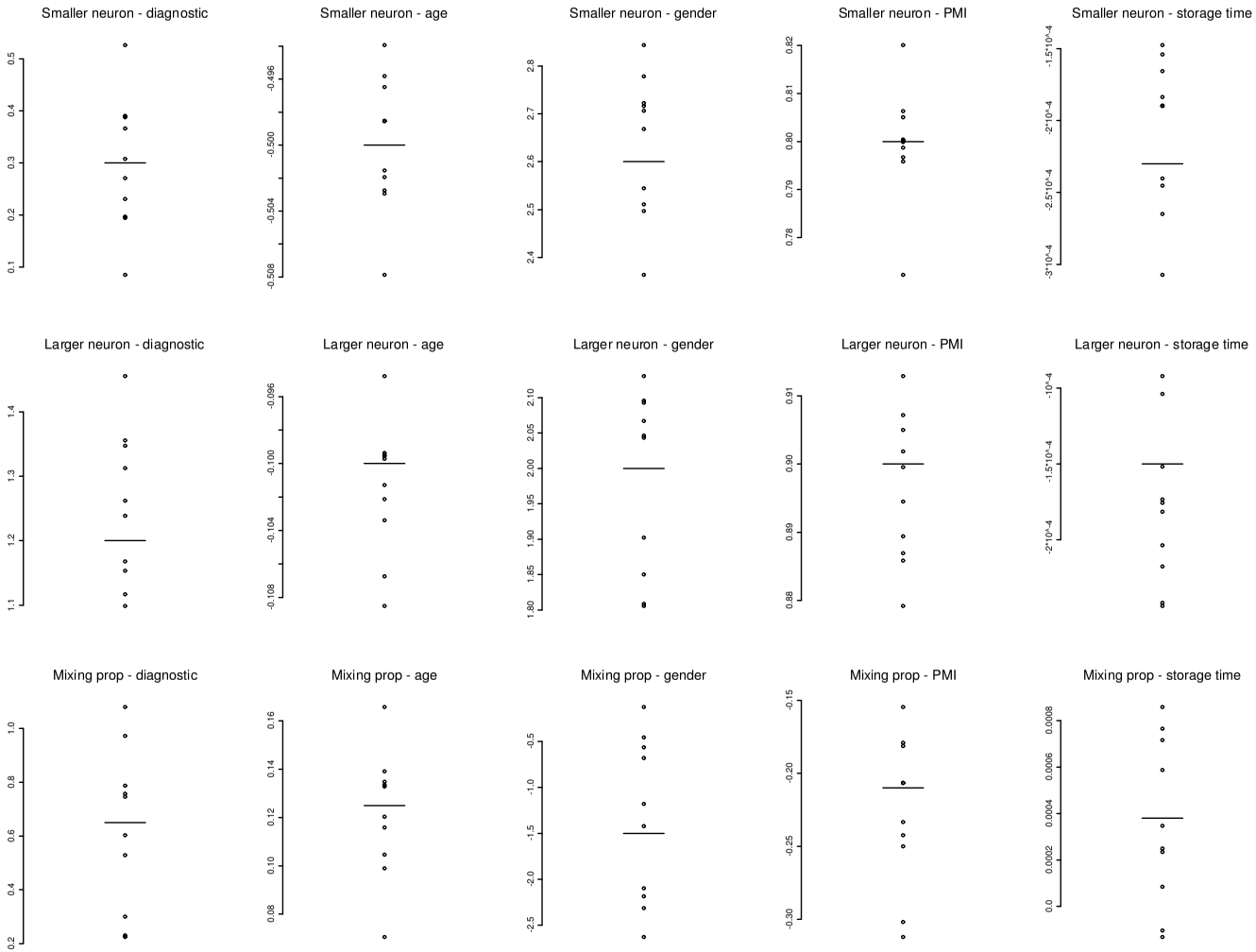

Figure 2: *Well-separated mixtures of normals: Dot plots of the estimates from the 10 runs for diagnostic, age, gender, PMI, and storage time effects in smaller neuron population, larger neuron population and mixing proportions.*

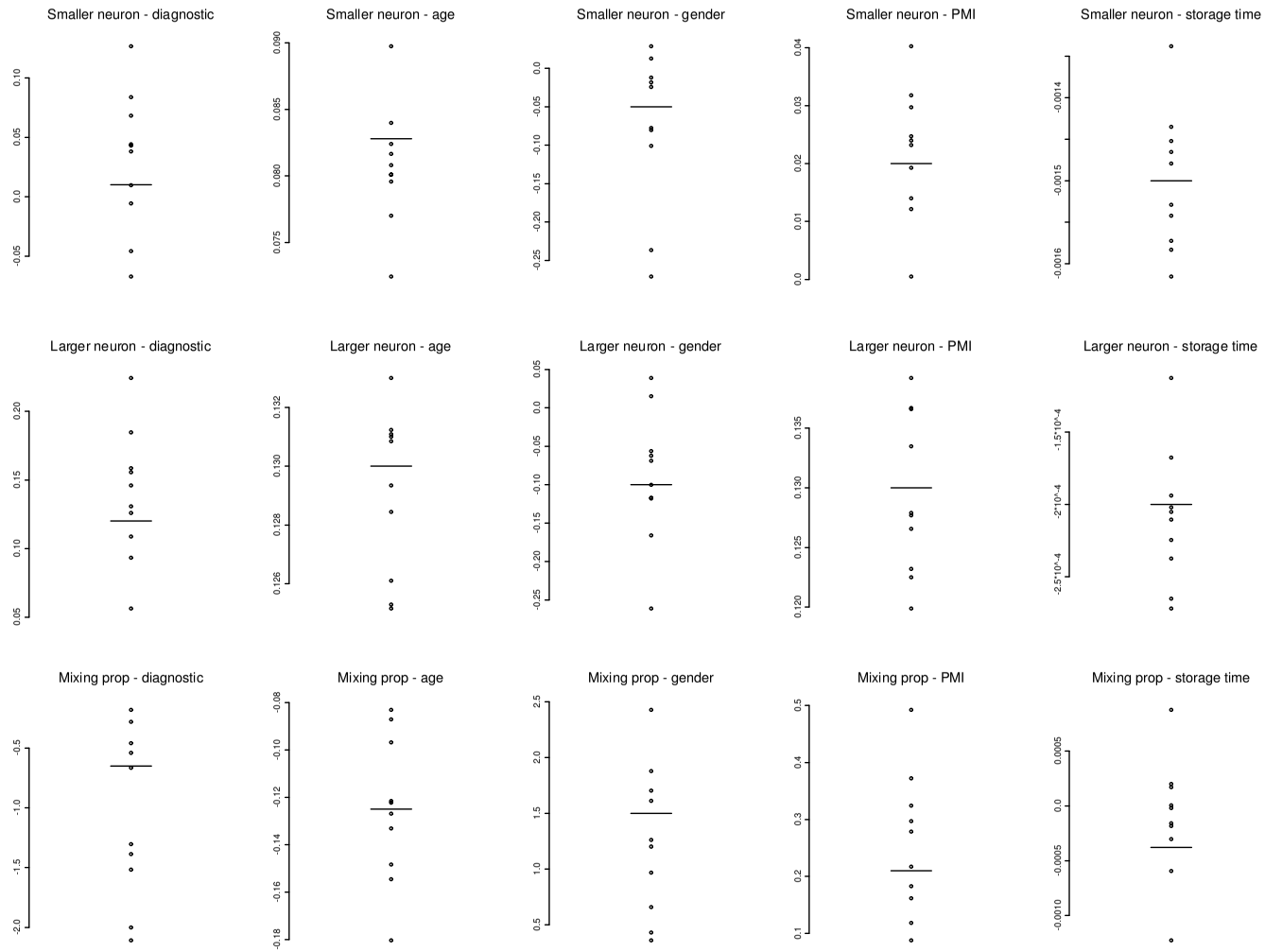

Figure 3: *Medium-separated mixtures of normals: Dot plots of the estimates from the 10 runs for diagnostic, age, gender, PMI, and storage time effects in smaller neuron population, larger neuron population and mixing proportions.*

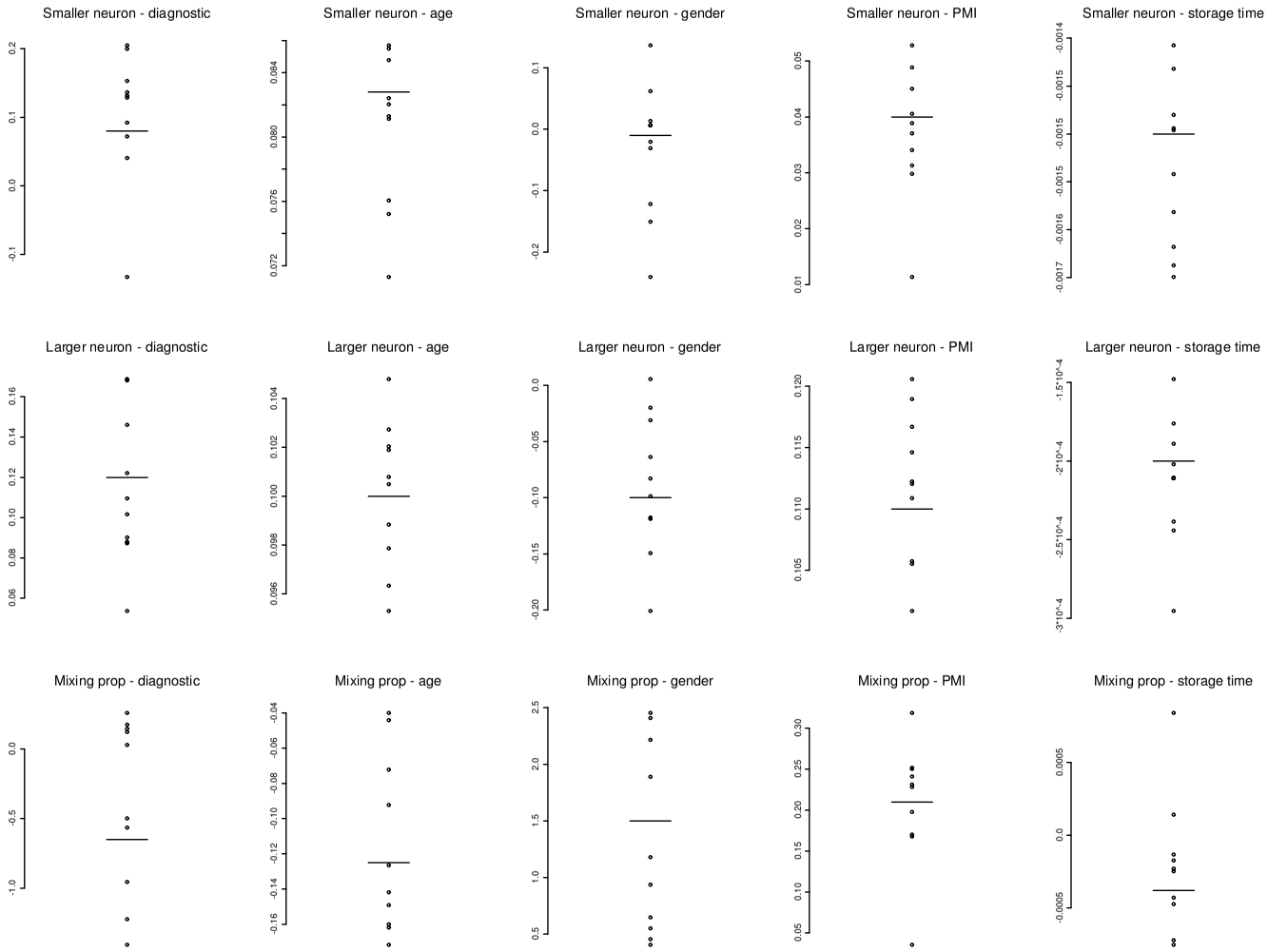

Figure 4: *Poorly-separated mixtures of normals: Dot plots of the estimates from the 10 runs for diagnostic, age, gender, PMI, and storage time effects in smaller neuron population, larger neuron population and mixing proportions.*

5.6 DISCUSSION AND SUMMARY

As mentioned in Section 5.2.1, modeling the component indicators by logits with random effects is only one way of constructing a multivariate Bernoulli distribution. Another possibility is to assume that conditional on the subject specific random effect W_i, the Z_{ij}'s are independent Bernoulli random variables with mean W_i, and model the W_i's to be independent beta random variables with parameters $e^{\boldsymbol{x}_i^T\boldsymbol{\alpha}}$ and $e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are unknown parameters. Note that the distribution of W_i depends only on $\boldsymbol{x}_i$, which consists of the between-subject factors.

In addition to using a latent variable to account for the dependence in $\boldsymbol{Z_i}$, an alternative to constructing a MVB($\boldsymbol{\theta}, \boldsymbol{x}$) is to assume that the $\boldsymbol{Z_i}$'s are distributed as a multivariate tolerance distribution, which we illustrate in the normal setting. Let $\boldsymbol{W_i} = (W_{i1}, \ldots, W_{il_i})$, $i = 1, \ldots, n$, have independent multivariate normal distributions with mean vector $(\boldsymbol{x}_{i1}^T\boldsymbol{\alpha}, \boldsymbol{x}_{i2}^T\boldsymbol{\alpha}, \ldots, \boldsymbol{x}_{il_i}^T\boldsymbol{\alpha})$ and a known covariance matrix Σ, and $\boldsymbol{\alpha}$ is an unknown parameter vector. Given $\boldsymbol{W_i}$, let $Z_{ij} = 1$ if $W_{ij} > 0$, and otherwise $Z_{ij} = 0$, for $j = 1, \ldots, l_i$.Then $\boldsymbol{Z_i}$=$(Z_{i1}, \ldots, Z_{il_i})$ has a multivariate Bernoulli distribution.

The multivariate Bernoulli mixtures of normals can be generalized to multivariate Bernoulli mixture models where the mixture components are any member of the exponential family. The model fitting procedures given in Section 5.3 can be modified to fit other examples of multivariate Bernoulli mixture models, such as multivariate Bernoulli mixtures of Poissons.

In this chapter, we only consider two-component multivariate Bernoulli mixtures of normals. Our results can also be extended to any finite number of components $g > 2$. In this case, in order to describe the component-indicator variables, we construct families of multivariate multinomial distributions that depend on covariates. One approach to constructing such a distribution is to incorporate subject-specific random effects into the mixing proportions which are modeled by multivariate linear logits. Let $\boldsymbol{Z_{ij}} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijg})$, where $Z_{ijk} = 1$ if y_{ij} comes from the kth component and $Z_{ijk} = 0$ otherwise. Given a normal distributed subject-specific random effect with mean 0 and variance σ_w^2, assume that $\boldsymbol{Z_{ij}}$ is distributed according to a multinomial distribution consisting of one draw from l_i categories

with probabilities $p_1(\boldsymbol{x}_{ij}), \ldots, p_g(\boldsymbol{x}_{ij})$, where

$$\begin{aligned} p_k(\boldsymbol{x}_{ij}) &= \frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma_k} + w_i}}{1 + \sum_{h=1}^{g-1} e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma_h} + w_i}}, \ k = 1, \ldots, g-1, \\ p_g(\boldsymbol{x}_{ij}) &= 1 - \sum_{h=1}^{g-1} p_h(\boldsymbol{x}_{ij}). \end{aligned}$$

Given $\boldsymbol{Z_{ij}}$, the conditional distribution of the observed y_{ij} is normal with mean $\boldsymbol{x}_{ij}^T \boldsymbol{\beta_k}$ and variance σ_k^2 if $z_{ijk} = 1$. In this model, $\boldsymbol{\gamma_k}$, $\boldsymbol{\beta_k}$, σ_k^2, $k = 1, \ldots, g-1$ and σ_w^2 are the unknown parameters.

In some cases where we do not have information about the number of components in the data, the number of components g might be treated as an unknown parameter and sampled from the posterior distribution using a reversible jump MCMC scheme, as proposed in Green (1995). This is a topic for further investigation.

In summary, in this chapter we present a novel mixture model for repeated measurements in which correlation among repeated observations on the same subject is induced by introducing correlation among the unobservable component-indicator variables within each subject. The mixture components in our model are linear regressions, and the mixing proportions are logits with random effects. Inference is facilitated by sampling from the posterior distribution of the parameters via MCMC methods. We fit this model to the neuron volume data to examine the diagnostic main effect. Although the neuron volume data contains between-subject factors only, our model can accommodate both between-subject and within-subject factors. Thus, our model can be applied to longitudinal studies, as well as to neurological studies.

6.0 MULTIVARIATE BERNOULLI MIXTURES OF GLMMS WITH APPLICATION TO POSTMORTEM TISSUE STUDIES IN SCHIZOPHRENIA

6.1 INTRODUCTION AND MOTIVATING EXAMPLE

We describe extensions of several recent repeated measures models which in turn can be viewed as generalizations of mixtures-of-experts (Jacobs, Jordan, Nowlan and Hinton (1991)). Our extended model was motivated by a neuronal postmortem human brain tissue study, in which the observations taken on the same subject are correlated and each observation arises from a mixture of two populations, each corresponding to a potential type of neuron.

Rubin and Wu (1997) proposed a two-component mixture model to model such data, which is a special case of their "extra component mixture" that we refer to as the Rubin-Wu model. To account for the within-subject dependence in the data, they introduce subject-specific random effects into the mixture components of their model, where the mixture components follow a linear regression model with subject-specific random effects, while the mixing proportions are logits linear in the covariates.

In Chapter 5 we proposed a different model in this setting. That multivariate Bernoulli mixtures of normals model, has linear regression models for the mixture components and includes multivariate Bernoulli distributions to model the mixing proportions. The approach for constructing a multivariate Bernoulli distribution was to model component-indicator variables by logistic regression with subject-specific random effects. The subject-specific random effects play the role of the latent variables which account for the dependence present in the data.

The model we propose in this chapter, which we refer to as multivariate Bernoulli mix-

tures of mixed normals, combines the Rubin-Wu model and the multivariate Bernoulli mixtures of normals. The mixture components in the proposed model are linear regressions with subject-specific random effects, as in the Rubin-Wu model, while the mixing proportions are governed by logistic regressions with another group of subject-specific random effects, as in the multivariate Bernoulli mixtures of normals of Chapter 5. The subject-specific random effects in the mixing proportions are independent of the random effects in the mixture components, and both account for the within-subject dependence in the data.

The motivating data, which was initially analyzed in Sweet, Pierri, Auh, Sampson, and Lewis (2003), is described in detail in Chapter 5. Brain tissue from eighteen normal subjects and eighteen schizophrenic subjects were selected. To examine the deficient auditory sensory memory in schizophrenic subjects, the somal volumes of deep layer 3 pyramidal cells in the auditory association cortex in all subjects were examined. For each subject, approximately 100 to 150 neurons were sampled and their neuronal volumes were measured. These are treated as repeated measures on each subject. Furthermore, two types of neurons exist in the auditory association cortex: one with shorter axons and consequently having smaller neuron volumes, and the other with longer axons and having larger somal volumes. Therefore, the somal volume of each neuron can be viewed as coming from one of two populations: a smaller neuron group and a larger neuron group. The known covariates associated with each subject which can affect the measured neuron volume are age, gender, postmortem interval (PMI) and brain storage time. The diagnostic effect is of primary interest in this study.

After describing the model in Section 6.2, we give the joint distribution of the observed data for each individual and compare this joint distribution with those of the Rubin-Wu model and the multivariate Bernoulli mixtures of normals. Section 6.3 outlines the estimation procedure, based on Markov chain Monte Carlo (MCMC) methods. In Section 6.4, we present simulation results and illustrate the difficulties we encounter in fitting the model to simulated data. Due to these problems, we are not currently able to implement our methodology to the neuronal auditory cortex data, and we indicate as future research what remains to be done. Extensions of the proposed model and other future research topics are discussed in Section 6.5.

6.2 MULTIVARIATE BERNOULLI MIXTURES OF MIXED NORMALS

In this section, we present the model and illustrate its relationship with the Rubin-Wu model and the multivariate Bernoulli mixtures of normals by considering the joint distribution of the observations for each individual under each of these three models.

6.2.1 The Model

Let Y_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, l_i)$ denote the jth observation on subject i, and let Z_{ij} and $\boldsymbol{x}_{ij}$ denote, respectively, the unobservable mixture component-indicator variable and the covariate vector associated with observation Y_{ij}. Given a subject-specific normal random effect W_i $(i = 1, \ldots, n)$ with mean 0 and variance σ_w^2, the missing component-indicator random variables Z_{ij} $(j = 1, \ldots, l_i)$ have independent Bernoulli distributions with mean

$$\pi_{ij} = \frac{e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}{1 + e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}. \tag{6.1}$$

The component-indicator variable Z_{ij} indicates whether the observation comes from the first component $(Z_{ij} = 1)$ or the second component $(Z_{ij} = 0)$.

Furthermore, given Z_{ij}, and another subject-specific normal random effect S_i $(i = 1, \ldots, n)$ with mean 0 and variance σ_s^2, which is assumed to be independent of W_i $(i = 1, \ldots, n)$, the conditional density of Y_{ij} has the following normal distributions.

$$\begin{aligned} Y_{ij}|\,(Z_{ij} = 1, S_i) &\overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + s_i, \sigma_1^2) \\ Y_{ij}|\,(Z_{ij} = 0, S_i) &\overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + s_i, \sigma_2^2), \end{aligned} \tag{6.2}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2, \boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \sigma_w^2$ are the unknown parameters.

The main feature of the proposed model is that the within-subject dependence is accounted for in both the mixing proportions and the mixture components. Thus, the model not only assumes that the normal components are correlated in one subject, but also assumes that the component a neuron is coming from is stochastically dependent on which components the other neurons come from in the same individual.

In this new model, the Z_{ij}'s are modeled by logistic regressions with random effects, which is one approach to constructing a multivariate Bernoulli distribution, MVB($\boldsymbol{\theta}, \boldsymbol{x}$). As suggested in Chapter 5, there are other mechanisms for building a MVB($\boldsymbol{\theta}, \boldsymbol{x}$), which are different variations of the model proposed in this chapter.

6.2.2 The Joint Distribution of the Observed Data for Each Individual

We now provide the joint distribution of the observed data for each individual for the multivariate Bernoulli mixtures of mixed normals model and compare it with those of the Rubin-Wu model and the multivariate Bernoulli mixtures of normals. For ease of notation and without loss of generality, assume that the number of observations on each subject are equal to 2, i.e., $l_i = 2$ $(i = 1, \ldots, n)$.

Given the subject-specific random effects W_i and S_i, the conditional density of the response variable Y_{ij} in the multivariate Bernoulli mixtures of mixed normals is written as

$$f(y_{ij}|\, S_i = s_i, W_i = w_i) = \pi_{ij}\phi(y_{ij}, \mu_{ij1} + s_i, \sigma_1^2) + (1 - \pi_{ij})\phi(y_{ij}, \mu_{ij2} + s_i, \sigma_2^2),$$

where $\phi(\cdot\,, \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2; π_{ij} is as in (6.1) and

$$\mu_{ijk} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k, \;\; j = 1, 2, \text{and } k = 1, 2. \tag{6.3}$$

Therefore, the joint distribution of $(Y_{i1},\, Y_{i2})$ is given by

$$\begin{aligned}
& f_{y_{i1}, y_{i2}}(y_{i1}, y_{i2}) \\
= \;& \left\{ \int \pi_{i1}\pi_{i2}\phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i11}, \mu_{i21})^T, \begin{bmatrix} \sigma_1^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_1^2 + \sigma_s^2 \end{bmatrix}) \\
+ \;& \left\{ \int (1 - \pi_{i1})\pi_{i2}\phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i12}, \mu_{i21})^T, \begin{bmatrix} \sigma_2^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_1^2 + \sigma_s^2 \end{bmatrix}) \\
+ \;& \left\{ \int \pi_{i1}(1 - \pi_{i2})\phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i11}, \mu_{i22})^T, \begin{bmatrix} \sigma_1^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2 + \sigma_s^2 \end{bmatrix}) \\
+ \;& \left\{ \int (1 - \pi_{i1})(1 - \pi_{i2})\phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i12}, \mu_{i22})^T, \begin{bmatrix} \sigma_2^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2 + \sigma_s^2 \end{bmatrix}),
\end{aligned} \tag{6.4}$$

where $\phi_2(\cdot\ , \boldsymbol{\mu}, \Sigma)$ denotes the bivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

In the Rubin-Wu model, the latent component-indicator random variable Z_{ij} has independent Bernoulli distributions with mean p_{ij}, where $p_{ij} = \frac{e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}}}{1+e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}}}$. Given Z_{ij} and the subject-specific random effect S_i which has a normal distribution with mean 0 and variance σ_s^2, the conditional distribution of Y_{ij} has the normal distribution given in (6.2). For the Rubin-Wu model, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,$\boldsymbol{\gamma}$, σ_1^2, σ_2^2 and σ_s^2 are the unknown parameters. As described in Chapter 5, the joint distribution of $(Y_{i1},\ Y_{i2})$ for the Rubin-Wu model is

$$
\begin{aligned}
f_{y_{i1},y_{i2}}(y_{i1}, y_{i2}) &= p_{i1}p_{i2}\phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i11}, \mu_{i21})^T, \begin{bmatrix} \sigma_1^2+\sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_1^2+\sigma_s^2 \end{bmatrix}) \\
&+ (1-p_{i1})p_{i2}\phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i12}, \mu_{i21})^T, \begin{bmatrix} \sigma_2^2+\sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_1^2+\sigma_s^2 \end{bmatrix}) \\
&+ p_{i1}(1-p_{i2})\phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i11}, \mu_{i22})^T, \begin{bmatrix} \sigma_1^2+\sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2+\sigma_s^2 \end{bmatrix}) \\
&+ (1-p_{i1})(1-p_{i2})\phi_2(\,(y_{i1}, y_{i2})^T, (\mu_{i12}, \mu_{i22})^T, \begin{bmatrix} \sigma_2^2+\sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2+\sigma_s^2 \end{bmatrix}),
\end{aligned}
\tag{6.5}
$$

where the μ_{ijk}'s are given in (6.3).

In the multivariate Bernoulli mixtures of normals, given the normally distributed subject-specific random effect W_i with mean 0 and variance σ_w^2, the conditional distribution of the component indicator Z_{ij} is Bernoulli with mean π_{ij}, as in (6.1). The conditional distribution of $Y_{ij}|Z_{ij}$ is a normal distribution with mean μ_{ij1} and variance σ_1^2 when $Z_{ij} = 1$, and otherwise, a normal distribution with mean μ_{ij2} and variance σ_2^2, where μ_{ij1} and μ_{ij2} are given in

(6.3). In Chapter 5, we give the joint distribution of (Y_{i1}, Y_{i2}) for this model , which is

$$
\begin{aligned}
& f_{y_{i1},y_{i2}}(y_{i1}, y_{i2}) \\
= & \left\{ \int \pi_{i1}\pi_{i2}\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T; (\mu_{i11}, \mu_{i21})^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
+ & \left\{ \int (1-\pi_{i1})\pi_{i2}\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T; (\mu_{i12}, \mu_{i21})^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
+ & \left\{ \int \pi_{i1}(1-\pi_{i2})\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T; (\mu_{i11}, \mu_{i22})^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \\
+ & \left\{ \int (1-\pi_{i1})(1-\pi_{i2})\phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T; (\mu_{i12}, \mu_{i22})^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}),
\end{aligned}
\tag{6.6}
$$

All the joint distributions of the observed data for each individual under the three models are four-component mixtures of bivariate normals. The joint distribution for each individual under the multivariate Bernoulli mixtures of mixed normals has the same mixture components as in the Rubin-Wu model, and the same mixing proportions as in the multivariate Bernoulli mixtures of normals. The mixture components in the multivariate Bernoulli mixtures of mixed normals are bivariate normals with compound-symmetric covariance matrices. The diagonal elements of each covariance matrix are described by pairs of elements from the cross-product set $\{\sigma_1^2 + \sigma_s^2, \sigma_2^2 + \sigma_s^2\} \times \{\sigma_1^2 + \sigma_s^2, \sigma_2^2 + \sigma_s^2\}$, while the off-diagonal elements are σ_s^2 in each covariance matrix. The non-zero off-diagonal elements are one source of the within-subject correlation. The mixing proportions in the multivariate Bernoulli mixtures of mixed models are the expectations of a function of $\boldsymbol{x}_{i1}^T\boldsymbol{\gamma}$ and w_i over the normally distributed random variable w_i, which provides another source of dependence for the within-subject observations. The above results can be easily extended to any number of observations for each subject.

6.3 INFERENCE

Inference for the proposed model is facilitated via MCMC methods. In this section, we outline the MCMC sampling scheme for the model and provide some further details in Appendix E.

6.3.1 Augmented Likelihood and Prior Distributions

We augment the observed data with the unobservable variables, i.e., the component indicators $Z_{ij}, i = 1, \ldots, n, j = 1, \ldots, l_i$, the subject-specific random effects in the mixing proportions $W_i, i = 1, \ldots, n$, and the subject-specific random effects in the mixture components $S_i, i = 1, \ldots, n$. We treat the missing values as unknown "parameters" and sample from them along with the unknown parameters. For ease of notation, we define $\boldsymbol{y} = (y_{11}, \ldots, y_{1l_1}, \ldots, y_{n1}, \ldots, y_{nl_n})^T$, $\boldsymbol{z} = (z_{11}, \ldots, z_{1l_1}, \ldots, z_{n1}, \ldots, z_{nl_n})^T$, $\boldsymbol{w} = (w_1, \ldots, w_n)^T$, and $\boldsymbol{s} = (s_1, \ldots, s_n)^T$. The augmented likelihood is then given by:

$$
\begin{aligned}
& L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \sigma_w^2 | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{s}) \\
\propto \; & (\sigma_w^2)^{-\frac{n}{2}} (\sigma_s^2)^{-\frac{n}{2}} \prod_{i=1}^{n} e^{-\frac{w_i^2}{2\sigma_w^2}} e^{-\frac{s_i^2}{2\sigma_s^2}} \\
& \prod_{j=1}^{l_i} \left[\frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \frac{1}{\sqrt{\sigma_1^2}} e^{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1 - s_i)^2}{2\sigma_1^2}} \right]^{z_{ij}} \left[\frac{1}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \frac{1}{\sqrt{\sigma_2^2}} e^{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2 - s_i)^2}{2\sigma_2^2}} \right]^{1 - z_{ij}} .
\end{aligned}
$$

We put independent priors on all the unknown parameter vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,$\boldsymbol{\gamma}$ and the unknown variances σ_1^2, σ_2^2, σ_w^2 and σ_s^2. The priors on the q-vectors $\boldsymbol{\gamma}, \boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ are taken to be q-variate normals with common mean vectors 0 and separate variance matrices $\sigma_\gamma^2 I_{q \times q}$, $\sigma_{\beta_1}^2 I_{q \times q}$, and $\sigma_{\beta_2}^2 I_{q \times q}$, where σ_γ^2, $\sigma_{\beta_1}^2$ and $\sigma_{\beta_2}^2$ are large numbers to ensure vague priors. For σ_1^2, σ_2^2, σ_w^2 and σ_s^2, we use inverse Gammas as priors, denoted by IG(α_1, δ_1), IG(α_2, δ_2), IG(α_w, δ_w) and IG(α_s, δ_s), respectively, where the shape parameters α_1, α_2, α_s, α_w and the rate parameters δ_1, δ_2, δ_s, δ_w are set to small numbers such as 0.01 to make the priors noninformative.

6.3.2 The Sampling Scheme

The Gibbs sampler is used to sample the unknown parameters from their posterior distribution. A Metropolis-Hasting step is performed for nonstandard conditional distributions. To achieve good mixing, we treat $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ as a block and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$, and $\boldsymbol{s}$ as another block.

1. Initialize the parameters $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, \boldsymbol{\gamma}^{(0)}, \sigma_1^{2^{(0)}}, \sigma_2^{2^{(0)}}, \sigma_s^{2^{(0)}}, \sigma_w^{2^{(0)}}$, and missing values $\boldsymbol{s}^{(0)}$, and $\boldsymbol{w}^{(0)}$.
2. Sample the component-indicators $z_{ij}^{(t+1)}, i = 1, \ldots, n, j = 1, \ldots, l_i$, from a Bernoulli random variable with mean $\tau_{ij}^{(t)}$, where

$$\tau_{ij}^{(t)} = \frac{\frac{1}{\sigma_1^{(t)}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1^{(t)} - s_i^{(t)})^2}{2\sigma_1^{2(t)}} + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma}^{(t)} + w_i^T\}}{\frac{1}{\sigma_1^{(t)}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1^{(t)} - s_i^{(t)})^2}{2\sigma_1^{2(t)}} + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma}^{(t)} + w_i^{(t)}\} + \frac{1}{\sigma_2^{(t)}} \exp\{-\frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2^{(t)} - s_i^{(t)})^2}{2\sigma_2^{2(t)}}\}}.$$

3. Sample $\sigma_1^{2(t+1)}$ from $IG(\frac{1}{2}\sum_{i=1}^n l_i + \alpha_1, \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^{l_i} z_{ij}^{(t+1)}(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1^{(t)} - s_i^{(t)})^2 + \delta_1)$.
4. Sample $\sigma_2^{2(t+1)}$ from $IG(\frac{1}{2}\sum_{i=1}^n l_i + \alpha_2, \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^{l_i} (1 - z_{ij}^{(t+1)})(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2^{(t)} - s_i^{(t)})^2 + \delta_2)$, such that $\sigma_2^{2(t+1)} < \sigma_1^{2(t+1)}$ for identifiability of the normal components.
5. Sample $\sigma_w^{2(t+1)}$ from $IG(\frac{n}{2} + \alpha_w, \frac{1}{2}\sum_{i=1}^n w_i^{2(t)} + \delta_w)$.
6. Sample $\sigma_s^{2(t+1)}$ from $IG(\frac{n}{2} + \alpha_s, \frac{1}{2}\sum_{i=1}^n s_i^{2(t)} + \delta_s)$.
7. Sample $(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{w}^{(t+1)})$ as a block from their conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{w} | \boldsymbol{y}, \boldsymbol{z}^{(t+1)}, \boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}, \boldsymbol{s}^{(t)}, \sigma_1^{2(t+1)}, \sigma_2^{2(t+1)}, \sigma_s^{2(t+1)}, \sigma_w^{2(t+1)})$ via a Metropolis-Hastings step.
8. Sample $(\boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)}, \boldsymbol{s}^{(t+1)})$ as a block from their conditional distribution $p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \boldsymbol{y}, \boldsymbol{z}^{(t+1)}, \sigma_1^{2(t+1)}, \sigma_2^{2(t+1)}, \sigma_s^{2(t+1)}, \boldsymbol{\gamma}^{(t+1)}, \boldsymbol{w}^{(t+1)}, \sigma_w^{2(t+1)})$ via a Metropolis-Hastings step.

For details of the Metropolis-Hastings steps, see Appendix E.

To obtain initial values for the Markov chain, we first fit mixtures-of-experts to the data via the EM algorithm neglecting the within-subject correlation. Starting values for σ_s^2, σ_w^2 are randomly selected from truncated normal distributions. Starting points for w_i and s_i are randomly sampled from normal distributions with means 0 and variances $\sigma_s^{2(0)}$ and $\sigma_w^{2(0)}$, respectively.

6.4 PROBLEM ENCOUNTERED IN A SIMULATION STUDY

Before attempting to apply our MCMC estimation method to the neuronal volume data, we investigated the properties of our estimators. To do so, we conducted a simulation study and encountered some problems in all three simulated data sets. For this reason, we did not conduct further simulations. In this section, we report our findings from these three simulated data sets and discuss a solution to the problem.

Given a set of true parameter values of ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}$, σ_1^2, σ_2^2, σ_s^2, σ_w^2), three data sets are simulated, where each has the "same" data structure as the neuronal volume data described as follows. The covariate vectors $\boldsymbol{x}_{ij}, i = 1, \ldots, 36, j = 1, \ldots, l_i$, used in the simulated data sets are taken from the motivating neuron volume data. Each vector $\boldsymbol{x}_{ij}$ consists of an intercept, the diagnostic effect indicator (normal=1, schizophrenic=2), age, gender effect indicator (female=1, male=2), postmortem interval (PMI), and storage time associated with subject i. In each simulated data, the response variables y_{ij}, $i = 1, \ldots, 36, j = 1, \ldots, l_i$, which can be viewed as the log-transformed neuron volume for the j th neuron in subject i, are simulated from the multivariate Bernoulli mixtures of mixed normals with the given parameter values ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}$, σ_1^2, σ_2^2, σ_s^2, σ_w^2).

We fit the multivariate Bernoulli mixtures of mixed normals, to each simulated data set by the sampling scheme given in Section 6.3. The Markov chain converges after 10,000 iterations. To see whether or not the parameters have reached a steady state, the chain is run for additional 30,000 iterations after the initial 10,000 iterations. To be conservative, we treat the first 11,000 iterations as burn-in cycles, and posterior means and 95% credible intervals for parameters of interest are calculated based on the remaining 29,000 iterations.

We study the behavior of the resulting estimates by comparing the estimates and the credible intervals with the true parameter values. For all three simulated data sets, almost all the true parameter values of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}$, σ_1^2, σ_2^2, σ_s^2, σ_w^2 are within their corresponding 95% credible intervals. In Table 8, we present results from one simulated data set as an example. The "smaller neuron population" and "larger neuron population" indicate the two normal components in the model and "Mixing proportions" indicate the proportions of the smaller neurons. It can be seen from the table that as far as the unknown parameters

Table 8: *Results of model fitting to one of the simulated data sets. True values, estimates (posterior means) and the 95% credible intervals.*

	true	2.5%	mean	97.5%
Smaller neurons population				
intercept ($\boldsymbol{\beta}_{10}$)	2.7000	-3.1426	0.5804	4.3998
diagnostic	0.0100	-1.6799	-0.3645	0.9192
age	0.0828	0.0212	0.0934	0.1682
gender	-0.0500	-1.0008	0.5092	1.9956
PMI	0.0200	-0.1166	0.0191	0.1571
storage time	-0.0015	-0.0019	-0.0010	-0.0002
σ_1^2	2.5000	2.3630	2.5150	2.6740
Larger neurons population				
intercept ($\boldsymbol{\beta}_{20}$)	1.3500	-3.8042	-0.0427	3.7691
diagnostic	0.1200	-1.5762	-0.2826	0.9881
age	0.1300	0.0579	0.1310	0.2048
gender	-0.1000	-1.0333	0.4559	1.9291
PMI	0.1300	-0.0030	0.1327	0.2680
storage time	-0.0002	-0.0008	0.0001	0.0010
σ_2^2	2.0000	1.9840	2.0890	2.1990
Mixing proportions				
intercept ($\boldsymbol{\gamma}_0$)	1.9000	-3.8925	1.0241	5.9954
diagnostic	-0.6500	-2.5304	-0.9260	0.4911
age	-0.1250	-0.2166	-0.1298	-0.0442
gender	1.5000	0.3463	2.0259	3.7908
PMI	0.2100	0.1412	0.3020	0.4899
storage time	-0.0004	-0.0020	-0.0009	0.0002
σ_w^2	4.0000	2.3160	4.5950	8.9510
σ_s^2	3.0000	2.3950	3.9820	6.5740

$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2, \sigma_s^2$, and σ_w^2 are concerned, the model fitting procedure provides reasonable estimates.

A problem emerges when we study parameter estimates of interest at the subject level. To further examine the performance of the MCMC algorithm, for each subject i, $i = 1, \ldots, 36$, we obtain the posterior means for the random effects s_i and w_i. Furthermore, by noticing that the covariate vectors $\boldsymbol{x}_{ij}$'s in our simulated data set only contain between-subject factors and so all neurons belonging to the same subject have a common mean for the first component, a common mean for the second component, and a common proportion for the first component, denoted by μ_{i1}, μ_{i2}, and p_i respectively, we obtain the posterior means for μ_{i1}, μ_{i2}, and p_i for each subject, by averaging respectively $\mu_{i1}^{(t)} = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1^{(t)} + s_i^{(t)}$, $\mu_{i2}^{(t)} = \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_2^{(t)} + s_i^{(t)}$, and $p_i^{(t)} = \frac{e^{\boldsymbol{x}_{i1}^T\boldsymbol{\gamma}^{(t)}+w_i^{(t)}}}{1+e^{\boldsymbol{x}_{i1}^T\boldsymbol{\gamma}^{(t)}+w_i^{(t)}}}$ after discarding the burn-in samples, where $\boldsymbol{\beta}_1^{(t)}$, $\boldsymbol{\beta}_2^{(t)}, \boldsymbol{\gamma}^{(t)}$, $s_i^{(t)}$, and $w_i^{(t)}$ are the sample values in iteration t. In Table 9, for the same simulated data set presented in Table 8, we give the true values and estimated values, denoted by "true" and "est" respectively, for s_i, μ_{i1}, μ_{i2}, w_i and p_i. We notice that, for some subjects, the estimated s_i, w_i and μ_{i1}, μ_{i2}, p_i are far away from their true values. For example, subject 566 has much smaller estimated first and second component means than their corresponding true values, and the estimated s_i and w_i are quite different from their true values. We present the true and estimated density plots of neuron volume for Subject 566 in Figure 6.4, which highlights the manner in which the estimated component means shift from the true values of the component means. For this subject, given the true values of the unknown parameters and the true values of w_i and s_i, the true density function can be expressed as the mixture of normals $0.96\phi(7.24, 2.5) + 0.04\phi(13.06, 2)$. Given the estimated parameters and the estimated w_i and s_i, the estimated density function is $0.01\phi(1.75, 2.52) + 0.99\phi(7.50, 2.09)$, where $\phi(\cdot\ , \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2. We can see that the first component in the true mixture is very close to the second component in the estimated mixture, and the estimated p_i is very close to the true value of $1 - p_i$. Because the mixing proportion of the second component in the true mixture, which is 0.04, and the mixing proportion of the first component in the estimated mixture, which is 0.01, are very small, the two density plots in Figure 6.4 are almost identical. This phenomenon happens in all three simulated data sets for different subjects. It is noted that, all of these problematic subjects have true mixing

proportions close to 0 or 1.

To better understand why the posterior means of μ_{i1}, μ_{i2}, p_i are away from their true values for some subjects, we calculate a group of pseudo posterior means of μ_{i1}, μ_{i2} and p_i, denoted by est* in Table 9, by using their true values of s_i, w_i, instead of the realizations of s_i, w_i in each iteration. These pseudo posterior means are quite close to the true μ_{i1}, μ_{i2} and p_i for subject 566, indicating that the estimated s_i, w_i cause the shifting, not the estimated $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}$. A similar examination of the other two simulated data sets confirms that stray s_i, w_i are responsible for the shifting.

We run another Gibbs sampler starting from the true parameter values. For all subjects, the estimated s_i, w_i, μ_{i1}, μ_{i2}, p_i are close to their true values.

We speculate that when the Markov chain starts from some other values than the true parameter values, it is easily trapped in local maxima, where for some subjects whose true mixing proportions close to 0 or 1, the estimated μ_{i1}, μ_{i2}, p_i stray away from their true values. To account for the incorrectly-estimated subject-specific μ_{i1}, μ_{i2}, p_i, the corresponding s_i and w_i are able to compensate each other, therefore the Markov chain can be trapped for an extremely long time. For most subjects, μ_{i1}, μ_{i2}, and p_i are estimated correctly which results in correct estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$.

In our motivating data set, there are only between-subject covariates, leading to subject-specific μ_{i1}, μ_{i2}, p_i. We conjecture that if the covariate vectors contain both between-subject and within-subject factors, the component shifting problem might disappear. For such data, if for some neurons of one subject, the Markov chain runs into a local maximum, there are still other neurons in the same subject which may be estimated well, leading to the correct estimates of w_i and s_i, allowing the chain to quickly move out of the local maximum.

Table 9: *The estimates obtained by fitting the new model to one of the simulated data sets*

	s_i		μ_{i1}			μ_{i2}			w_i		p_i		
Sub.	est	true	est	true	est*	est	true	est*	est	true	est	true	est*
178	-0.42	0.26	1.27	0.95	1.94	8.07	7.98	8.74	0.80	-0.60	0.08	0.10	0.03
234	0.72	0.91	2.77	2.47	2.95	9.93	9.85	10.11	-0.59	-2.00	0.03	0.03	0.01
250	-3.22	-3.27	-1.60	-1.82	-1.66	4.27	4.21	4.21	-1.72	-3.13	0.00	0.00	0.00
285	-1.19	-1.03	-1.32	-1.14	-1.12	5.17	5.28	5.38	-0.71	-1.67	0.43	0.45	0.27
317	-1.33	-1.18	0.64	0.44	0.77	6.86	6.85	6.99	-1.29	-2.49	0.01	0.01	0.00
322	1.36	1.69	2.59	2.66	2.91	8.52	8.70	8.85	3.17	1.61	0.67	0.67	0.33
341	2.30	1.89	3.81	3.65	3.40	10.71	10.69	10.31	-2.72	-4.34	0.00	0.00	0.00
377	-0.64	-0.72	2.01	1.67	1.91	8.26	8.10	8.16	-1.08	-2.04	0.01	0.02	0.01
395	1.62	2.00	3.88	3.80	4.27	9.77	9.72	10.16	1.52	0.82	0.77	0.78	0.62
396	-2.45	-1.99	-0.17	-0.14	0.32	6.25	6.28	6.74	-0.52	-0.88	0.71	0.69	0.62
398	2.37	1.80	3.67	3.61	3.10	9.38	9.37	8.82	0.51	-0.89	0.04	0.05	0.01
408	-1.42	-1.28	1.09	1.16	1.25	7.95	8.08	8.11	1.06	0.77	0.84	0.84	0.78
412	2.68	3.01	5.09	5.02	5.44	11.07	11.01	11.42	1.66	1.28	0.88	0.90	0.82
449	1.28	0.89	3.92	3.82	3.53	8.49	8.44	8.10	-1.30	-4.50	0.00	0.00	0.00
450	1.12	1.10	4.31	4.40	4.30	10.99	11.09	10.99	-1.48	-1.27	0.48	0.50	0.53
451	2.08	2.60	5.46	5.71	5.99	10.90	11.18	11.43	0.72	0.45	0.51	0.57	0.45
452	-0.25	-0.81	1.86	1.65	1.33	7.38	7.11	6.85	-0.10	-0.74	0.25	0.28	0.17
466	-2.14	-2.24	1.08	1.11	0.98	7.32	7.37	7.23	0.47	0.37	0.74	0.74	0.70
474	-0.02	-0.25	3.13	3.25	2.91	8.64	8.76	8.42	1.63	0.78	0.35	0.34	0.21
517	-3.35	-4.32	-0.93	-1.00	-1.90	3.04	2.96	2.07	-1.26	-2.08	0.00	0.00	0.00
537	1.18	0.00	3.42	3.34	2.26	8.06	7.83	6.91	1.24	0.60	0.60	0.61	0.45
559	-0.06	-0.95	4.63	4.65	3.73	10.25	10.33	9.36	2.00	1.42	0.25	0.23	0.20
566	-3.70	1.43	1.75	7.24	6.87	7.50	13.06	12.63	-3.96	4.24	0.01	0.96	0.95
567	-0.84	-1.41	2.78	2.90	2.23	7.55	7.56	7.00	-1.61	-2.25	0.09	0.06	0.06
568	1.28	0.68	6.13	6.09	5.52	10.76	10.76	10.15	2.94	2.02	0.22	0.21	0.14
575	0.84	0.18	5.05	4.91	4.39	9.93	9.80	9.28	1.06	0.06	0.11	0.09	0.06
581	-0.14	-0.71	3.80	3.72	3.25	10.17	9.97	9.64	2.11	2.99	1.00	1.00	1.00
587	-0.32	-1.52	2.15	2.08	0.98	7.11	6.89	5.95	3.30	2.69	0.96	0.96	0.92
592	-2.09	-2.15	1.60	1.70	1.58	7.05	6.98	7.03	-1.82	-1.10	0.89	0.88	0.93
597	2.28	1.29	5.38	5.44	4.40	9.73	9.74	8.75	-2.21	-2.60	0.01	0.01	0.00
620	0.37	0.31	6.52	6.54	6.46	11.83	11.88	11.77	1.58	2.09	0.81	0.81	0.84
625	3.49	2.91	7.85	7.83	7.29	13.54	13.42	12.99	0.99	2.55	0.97	0.99	0.99
634	1.22	1.33	6.12	6.40	6.25	10.97	11.18	11.10	0.78	1.23	0.85	0.86	0.88
643	0.46	0.45	5.35	5.55	5.38	10.98	11.05	11.01	-3.72	-2.54	0.48	0.49	0.72
656	-2.19	-3.13	1.52	1.78	0.61	6.67	6.81	5.77	-1.10	-0.99	0.24	0.26	0.29
681	-0.58	-1.11	4.38	4.14	3.87	8.40	8.06	7.88	-0.24	-0.20	0.43	0.41	0.45

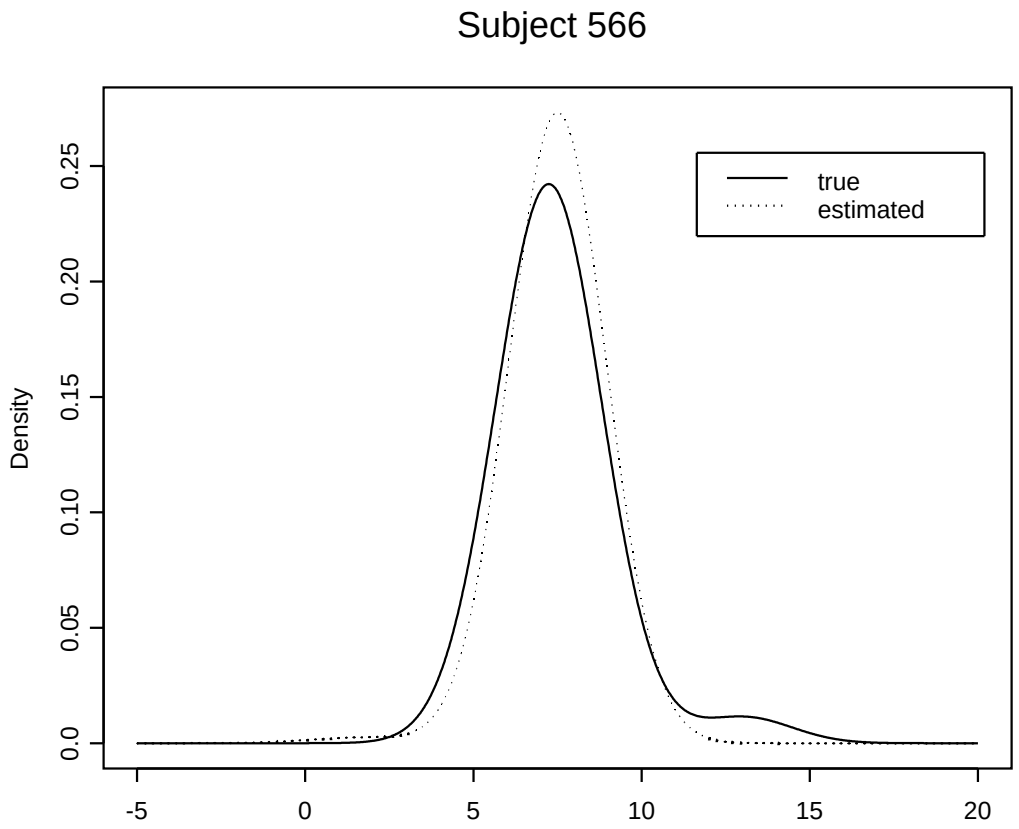

Figure 5: *True and estimated density plots for subject 566.*

True density function: $0.96\phi(7.24, 2.5) + 0.04\phi(13.06, 2)$*; estimated density function:* $0.01\phi(1.75, 2.52) + 0.99\phi(7.50, 2.09)$.

6.5 DISCUSSION

Due to the problems discussed in the Section 6.4, we have not been able to fully apply the proposed model to our motivating data sets. For the motivating data set, we are able to obtain reasonable inference for diagnostic effect in both populations of neurons and the mixing proportions; however, for a few subjects we appear to draw incorrect inferences about the mean of each neuron population and the proportion of smaller neurons. We feel that this model might be more useful for data with both within-subject and between-subject covariates. In our future research, we plan to do more simulations to confirm this speculation.

In Chapter 5, we suggested other possible approaches to building a MVB($\boldsymbol{\theta}, \boldsymbol{x}$) distribution, such as modeling the component indicator variables as Bernoullis with mean w_i, which is in turn distributed as a beta distribution. Similar extensions may be developed in the context of our proposed model.

The model proposed in this chapter can be extended to multivariate Bernoulli mixtures of generalized linear mixed models, where the mixture components are generalized linear mixed models, while the component-indicator variables are modeled by various multivariate Bernoulli distributions. Because the outcomes are continuous in our motivating example, we focus on normal mixture components in this chapter. For discrete outcomes, another instance of multivariate Bernoulli mixed mixtures models, multivariate Bernoulli mixtures of mixed Poissons can be used instead, where the mixture components are modeled by Poisson regressions with a subject-specific random effect.

We only consider two-component multivariate Bernoulli mixtures of mixed normals in this chapter. These can be extended to any finite number of components by using multivariate multinomial distributions to model the component-indicator variables as discussed in Chapter 5. When the number of components is unknown *a priori*, it can be treated as a missing variable and be sampled by a reversible jump MCMC algorithm, proposed by Green (1995). This is a topic for future research.

Note that the Rubin-Wu model is a special case of the multivariate Bernoulli mixtures of mixed normals proposed in this chapter when $\sigma_s^2 = 0$. On the other hand, the multivariate Bernoulli mixture of normals, proposed in Chapter 5, is a special case of the multivariate

Bernoulli mixture of mixed normals when $\sigma_w^2 = 0$. Testing which of the three models is a better fit to the data is a very challenging problem. Under the null hypothesis, $\sigma_w^2 = 0$ or $\sigma_s^2 = 0$, are on the boundary of the parameter space, so that standard asymptotic χ^2 distribution is not applicable to these likelihood ratio tests.

To summarize, in this chapter we propose a new mixture model to handle repeated measures. This model combines the Rubin and Wu model and the multivariate Bernoulli mixtures of normals proposed in Chapter 5. In the proposed model, the mixture components have a linear regression model with subject-specific random effects while the mixture proportions are modeled by a multivariate Bernoulli mixture distribution depending on covariates, denoted by MVB($\boldsymbol{\theta}, \boldsymbol{x}$). The specific MVB($\boldsymbol{\theta}, \boldsymbol{x}$) distribution we use is a model with linear logits having subject-specific random effects. The subject-specific random effects in both the mixing proportion and the mixture components model the dependence in the data. Our model can applied to neurological studies as well as to longitudinal studies.

7.0 FUTURE RESEARCH

In this dissertation, we propose three mixture models for repeated measurements. All of them can be viewed as multivariate extensions of mixtures-of-experts and have wide application in quantitative neurobiology.

As a special case of the first model, mixtures of GLMMs, we focus on normal-component mixtures of GLMMs and Poisson-component mixtures of GLMMs in Chapter 4. For Poisson-component mixtures of GLMMs, we outline the sampling scheme based on MCMC methods, but have not implemented the procedures for the motivating example, the grain count data, which has well-separated mixture components which makes it a less challenging computational problem for fitting the mixture model. We plan to employ our fitting procedure to a data set with medium or poorly separated Poisson mixture components in the future.

The second model, multivariate Bernoulli mixture model has been studied thoroughly with application to the neuronal volume data in Chapter 5.

The third model, multivariate Bernoulli mixtures of mixed normals, is proposed in Chapter 6. Due to problems encountered in a simulation study, we have not applied this model to our motivating data set, the neuronal volume data. We speculate that the model might be more appropriate for a data with both between-subject and within-subject covariates. We plan to do more simulations to confirm this speculation and then apply the model to fit an appropriate data set.

In addition to these preceding topics which we hope to revisit, there are some other extensions of the proposed models we would like to explore further.

7.1 UNKNOWN NUMBER OF COMPONENTS

In this dissertation, we assume *a priori* two distinct groups in a population and so only consider two mixture components in all three proposed models. If we do not have information on the number of components g, we can apply a reversible jump MCMC algorithm to fit the data by treating g as missing data and sampling it with the other unknown parameters.

If we detect more than two components, i.e., $g > 2$, the mixing proportions in the first proposed model are taken to be multivariate linear logits. To account for the dependency of the component-indicator variables in the second and third proposed model, the mixing proportions can be handled by multivariate multinomial distributions with random effects. One approach to constructing such a distribution is to model the latent component-indicator variables by multivariate linear logits with random effects.

7.2 OTHER APPROACHES TO CONSTRUCTING MULTIVARIATE BERNOULLI DISTRIBUTIONS

In the models proposed in Chapter 5 and Chapter 6, the mixing proportions are modeled by multivariate Bernoulli distributions to account for the within-subject correlation. The approach we adopt for constructing such a distribution is to model the mixing proportions as linear logits with random effects.

As pointed out in Section 5.6, there are other approaches to constructing multivariate Bernoulli distributions. For example, the component-indicator variables may be samples from a Bernoulli distribution with mean W_i, where W_i is a subject-specific random effect which follows a beta distribution dependent on covariates. Another approach to building a multivariate Bernoulli distribution is to model the component-indicator variables as multivariate tolerance distributions dependent on covariates.

7.3 SOME EXTRA NEW MODELS

In addition to the three mixture models proposed in the previous chapters, we plan to work on some other new models for repeated measures. We intend to further examine these possible extensions and apply them to appropriate data sets in the future.

7.3.1 Extra New Model I

The first new model extends mixtures-of-experts, while it introduces a less complicated correlation structure between the mixture components as compared to the mixtures of GLMMs. We now present this model with normal components as an example.

Let S_{1i} and S_{2i} be independent subject-specific random effects having normal distributions with mean 0 and variances σ^2_{1s} and σ^2_{2s} respectively. The conditional density of Y_{ij} is given as:

$$(Y_{ij}|\, Z_{ij}=0, S_{1i}=s_{1i}, S_{2i}=s_{2i}) \overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1 + s_{1i}, \sigma_1^2),$$
$$(Y_{ij}|\, Z_{ij}=1, S_{1i}=s_{1i}, S_{2i}=s_{2i}) \overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2 + s_{2i}, \sigma_2^2),$$

whereas the latent component indicator random variables, Z_{ij}'s are distributed as Bernoulli with mean p_{ij} where $\text{logit}(p_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\gamma}$. In this model, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,$\boldsymbol{\gamma}$, σ_1^2, σ_2^2, σ^2_{1s} and σ^2_{2s} are the unknown parameters.

Without loss of generality, we again assume that the number of observations on each

subject is equal to 2. The joint distribution of $(Y_{i1},\, Y_{i2})$ is

$$
\begin{aligned}
f_{y_{i1},y_{i2}}(y_{i1},y_{i2}) &= \int_{s_{1i}}\int_{s_{2i}} \left\{\prod_{j=1}^{2} f(y_{ij}|\, S_{1i}=s_{1i},\, S_{2i}=s_{2i})\right\} \phi(s_{1i},0,\sigma^2_{1s})\phi(s_{2i},0,\sigma^2_{2s})d_{s_{1i}}d_{s_{2i}} \\
&= p_{i1}p_{i2}\phi_2(\,(y_{i1},y_{i2})^T,(\mu_{i11},\mu_{i21})^T,\begin{bmatrix}\sigma^2_1+\sigma^2_{1s} & \sigma^2_{1s}\\ \sigma^2_{1s} & \sigma^2_1+\sigma^2_{1s}\end{bmatrix}) \\
&+ (1-p_{i1})p_{i2}\phi_2(\,(y_{i1},y_{i2})^T,(\mu_{i12},\mu_{i21})^T,\begin{bmatrix}\sigma^2_2+\sigma^2_{2s} & 0\\ 0 & \sigma^2_1+\sigma^2_{1s}\end{bmatrix}) \\
&+ p_{i1}(1-p_{i2})\phi_2(\,(y_{i1},y_{i2})^T,(\mu_{i11},\mu_{i22})^T,\begin{bmatrix}\sigma^2_1+\sigma^2_{1s} & 0\\ 0 & \sigma^2_2+\sigma^2_{2s}\end{bmatrix}) \\
&+ (1-p_{i1})(1-p_{i2})\phi_2(\,(y_{i1},y_{i2})^T,(\mu_{i12},\mu_{i22})^T,\begin{bmatrix}\sigma^2_2+\sigma^2_{2s} & \sigma^2_{2s}\\ \sigma^2_{2s} & \sigma^2_2+\sigma^2_{2s}\end{bmatrix}),
\end{aligned}
$$

where $\mu_{ijk}=\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k$, for $j=1,2$ and $k=1,2$.

7.3.2 Extra New Model II

The second model proposed in this section is an extension of the multivariate Bernoulli mixtures of normals with a less complicated correlation structure between the mixture components as compared to the multivariate Bernoulli mixtures of mixed normals.

In this model, the mixing proportions are modeled as in the multivariate Bernoulli mixtures of normals. Given a subject-specific normally distributed random effect W_i with mean 0 and variance σ^2_w, let Z_{ij} have a Bernoulli distribution with mean $\pi_{ij}=\frac{e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}{1+e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}$.

Let S_{1i} and S_{2i} be independent normally distributed subject-specific random effects with mean 0 and variances σ^2_{1s} and σ^2_{2s} respectively. The model is given by:

$$(Y_{ij}|\, Z_{ij}=0, S_{1i}=s_{1i}, S_{2i}=s_{2i}) \overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1+s_{1i},\sigma^2_1), \tag{7.1}$$

$$(Y_{ij}|\, Z_{ij}=1, S_{1i}=s_{1i}, S_{2i}=s_{2i}) \overset{indep}{\sim} N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2+s_{2i},\sigma^2_2), \tag{7.2}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,$\boldsymbol{\gamma}$, σ^2_1, σ^2_2, σ^2_w, σ^2_{1s} and σ^2_{2s} are the unknown parameters.

Without loss of generality, assume that the number of observations on each subject is equal to 2. The joint distribution of $(Y_{i1},\ Y_{i2})$ is

$$
\begin{aligned}
& f_{y_{i1},y_{i2}}(y_{i1},y_{i2}) \\
= & \int_{s_{1i}}\int_{s_{2i}}\int_{w_i}\left\{\prod_{j=1}^{2} f(y_{ij}|\ S_{1i}=s_{1i},\ S_{2i}=s_{2i})\right\}\phi(s_{1i},0,\sigma^2_{1s})\phi(s_{2i},0,\sigma^2_{2s})\phi(w_i,0,\sigma^2_w)ds_{1i}ds_{2i}dw_i \\
= & \left\{\int \pi_{i1}\pi_{i2}\phi(w_i;0,\sigma^2_w)dw_i\right\}\ \phi_2(\ (y_{i1},y_{i2})^T,(\mu_{i11},\mu_{i21})^T,\begin{bmatrix}\sigma^2_1+\sigma^2_{1s} & \sigma^2_{1s}\\ \sigma^2_{1s} & \sigma^2_1+\sigma^2_{1s}\end{bmatrix}) \\
+ & \left\{\int (1-\pi_{i1})\pi_{i2}\phi(w_i;0,\sigma^2_w)dw_i\right\}\ \phi_2(\ (y_{i1},y_{i2})^T,(\mu_{i12},\mu_{i21})^T,\begin{bmatrix}\sigma^2_2+\sigma^2_{2s} & 0\\ 0 & \sigma^2_1+\sigma^2_{1s}\end{bmatrix}) \\
+ & \left\{\int \pi_{i1}(1-\pi_{i2})\phi(w_i;0,\sigma^2_w)dw_i\right\}\ \phi_2(\ (y_{i1},y_{i2})^T,(\mu_{i11},\mu_{i22})^T,\begin{bmatrix}\sigma^2_1+\sigma^2_{1s} & 0\\ 0 & \sigma^2_2+\sigma^2_{2s}\end{bmatrix}) \\
+ & \left\{\int (1-\pi_{i1})(1-\pi_{i2})\phi(w_i;0,\sigma^2_w)dw_i\right\}\ \phi_2(\ (y_{i1},y_{i2})^T,(\mu_{i12},\mu_{i22})^T,\begin{bmatrix}\sigma^2_2+\sigma^2_{2s} & \sigma^2_{2s}\\ \sigma^2_{2s} & \sigma^2_2+\sigma^2_{2s}\end{bmatrix}),
\end{aligned}
$$

again $\mu_{ijk}=\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k$, for $j=1,2$ and $k=1,2$.

7.3.3 Extra New Model III

This model can be viewed as a generalization of all the mixture models proposed in this dissertation. It extends the multivariate Bernoulli mixtures of mixed normals by incorporating different subject-specific random effects into the mixture components. As in the multivariate Bernoulli mixtures of mixed normals, the mixing proportions in this mew model are linear logits with random effects.

Assume (S_{i1},S_{i2}) is a random variable following bivariate normal distribution with mean $(0,0)$ and covariance matrix $\Sigma=\begin{bmatrix}\sigma^2_{1s} & \sigma_{12s}\\ \sigma_{12s} & \sigma^2_{2s}\end{bmatrix}$. In this subsection's model, the distribution of Y_{ij} conditional on Z_{ij},S_{1i},S_{2i} is given as in (7.1) and (7.2) while the component-indicator variables Z_{ij}'s are modeled as in extra new model II given in Subsection 7.3.2. In this model, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,$\boldsymbol{\gamma}$, σ^2_1, σ^2_2, σ^2_w, and Σ are the unknown parameters.

Again, without loss of generality, assume that the number of observations on each subject is equal to 2. The joint distribution of $(Y_{i1},\ Y_{i2})$ under this new model is

$$
\begin{aligned}
& f_{y_{i1},y_{i2}}(y_{i1}, y_{i2}) \\
= & \int_{s_{1i}} \int_{s_{2i}} \int_{w_i} \left\{ \prod_{j=1}^{2} f(y_{ij} |\ S_{1i} = s_{1i},\ S_{2i} = s_{2i}) \right\} \phi(s_{1i}, 0, \sigma_{1s}^2) \phi(s_{2i}, 0, \sigma_{2s}^2) \phi(w_i, 0, \sigma_w^2) ds_{1i} ds_{2i} dw_i \\
= & \left\{ \int \pi_{i1}\pi_{i2} \phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\ (y_{i1}, y_{i2})^T, (\mu_{i11}, \mu_{i21})^T, \begin{bmatrix} \sigma_1^2 + \sigma_{1s}^2 & \sigma_{1s}^2 \\ \sigma_{1s}^2 & \sigma_1^2 + \sigma_{1s}^2 \end{bmatrix}) \\
+ & \left\{ \int (1 - \pi_{i1})\pi_{i2} \phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\ (y_{i1}, y_{i2})^T, (\mu_{i12}, \mu_{i21})^T, \begin{bmatrix} \sigma_2^2 + \sigma_{2s}^2 & \sigma_{12s} \\ \sigma_{12s} & \sigma_1^2 + \sigma_{1s}^2 \end{bmatrix}) \\
+ & \left\{ \int \pi_{i1}(1 - \pi_{i2}) \phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\ (y_{i1}, y_{i2})^T, (\mu_{i11}, \mu_{i22})^T, \begin{bmatrix} \sigma_1^2 + \sigma_{1s}^2 & \sigma_{12s} \\ \sigma_{12s} & \sigma_2^2 + \sigma_{2s}^2 \end{bmatrix}) \\
+ & \left\{ \int (1 - \pi_{i1})(1 - \pi_{i2}) \phi(w_i; 0, \sigma_w^2) dw_i \right\} \phi_2(\ (y_{i1}, y_{i2})^T, (\mu_{i12}, \mu_{i22})^T, \begin{bmatrix} \sigma_2^2 + \sigma_{2s}^2 & \sigma_{2s}^2 \\ \sigma_{2s}^2 & \sigma_2^2 + \sigma_{2s}^2 \end{bmatrix}).
\end{aligned}
$$

It is easy to see that extra model III reduces to multivariate Bernoulli mixtures of mixed normals when $S_{i1} = S_{i2}$, i.e., $\sigma_{1s}^2 = \sigma_{2s}^2 = \sigma_{12s}$. When $\sigma_{12s} = 0$, extra model III becomes extra model II, which in turn becomes extra model I when $\sigma_w^2 = 0$.

In Figure 6, we present all the mixture models given in this dissertation and illustrate their relationships. To compare the proposed models and test whether or not a model is a good fit to the data are very challenging problems, which we intend to explore further.

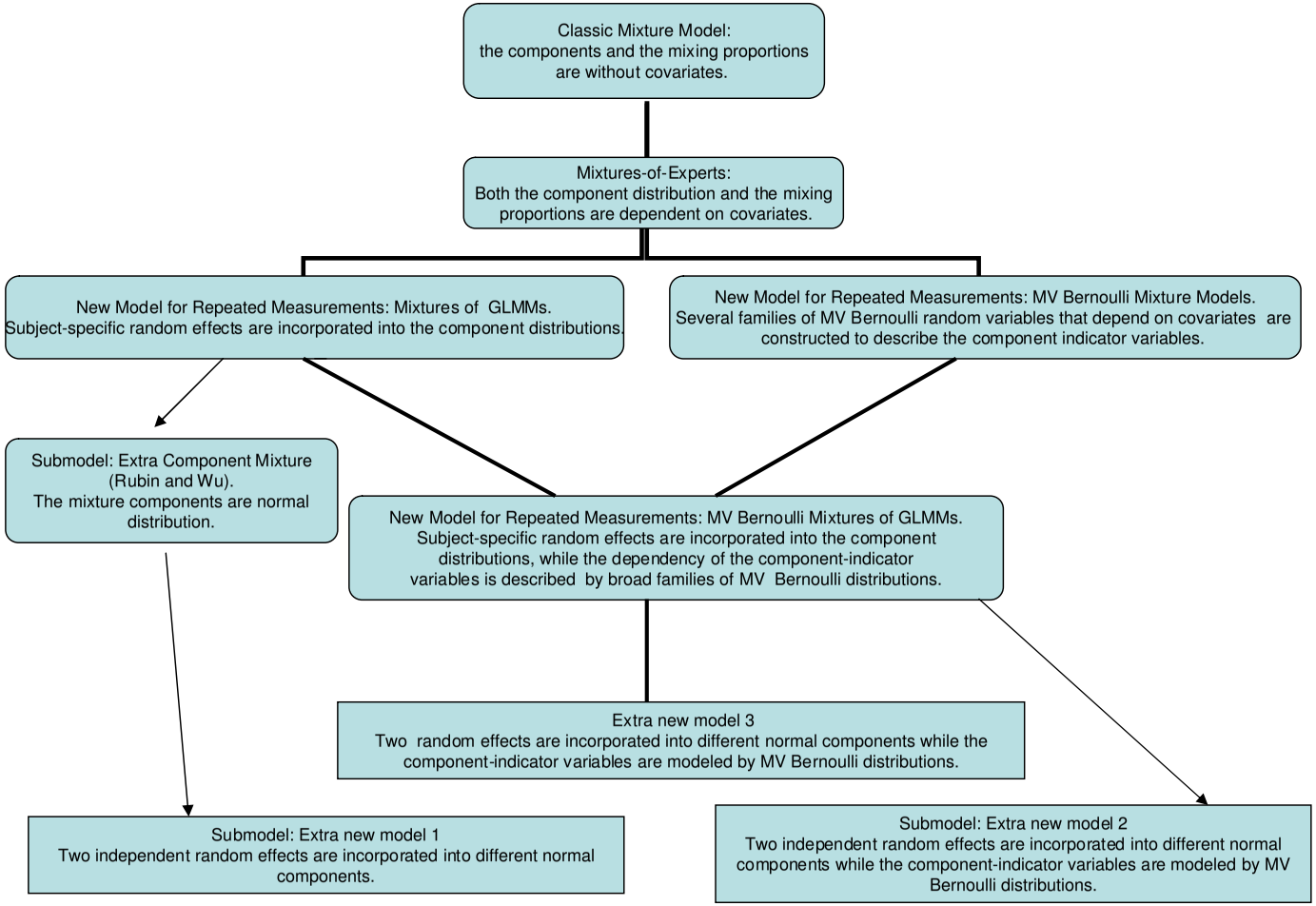

Figure 6: *Mixture models for repeated measures*

APPENDIX A

NEW RESULTS FOR THE CLASSIC MIXTURES OF POISSONS: QQ-PLOTS

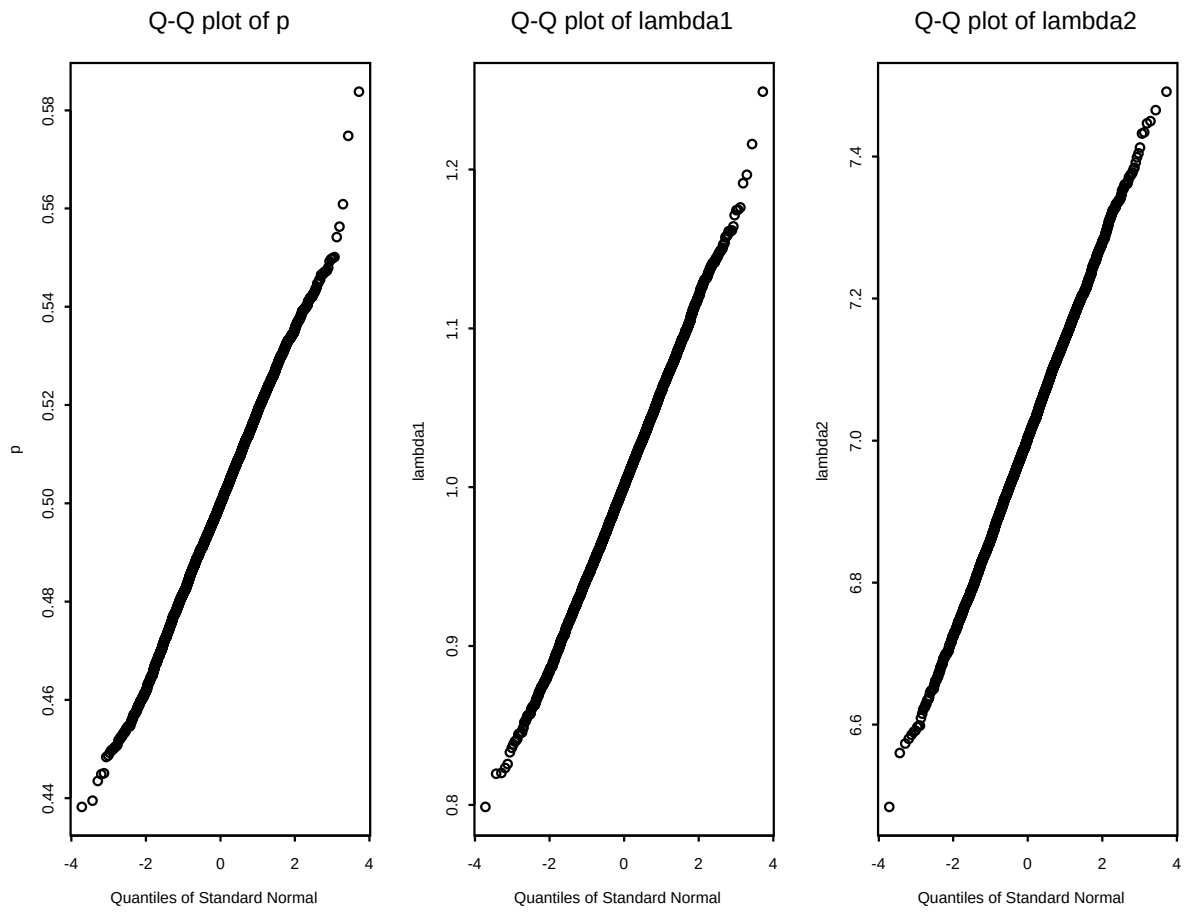

Figure 7: *QQ-plots for very-well-separated classic mixtures of Poissons.*

QQ plots for 5000 *replicates of* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *when* $p_1 = 0.5, \lambda_1 = 1, \lambda_2 = 7$. *We simulate* 5000 *random samples, each of size* $n = 1000$ *from classic two-component mixtures of Poissons. The parameter estimates* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *are obtained via the EM algorithm for each random sample.*

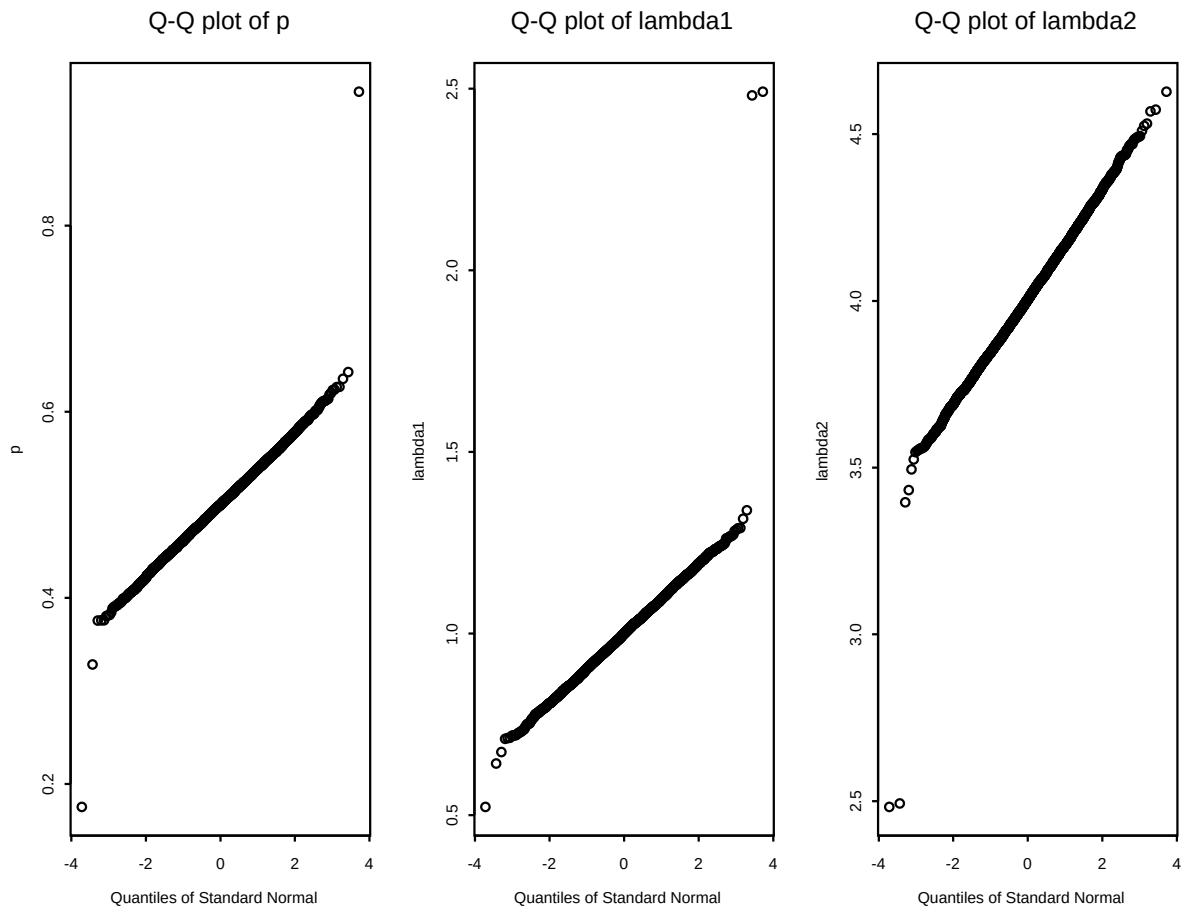

Figure 8: *QQ-plots for well-separated classic mixtures of Poissons.*

QQ plots for 5000 *replicates of* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *when* $p_1 = 0.5, \lambda_1 = 1, \lambda_2 = 4$. *We simulate* 5000 *random samples, each of size* $n = 1000$ *from classic two-component mixtures of Poissons. The parameter estimates* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *are obtained via the EM algorithm for each random sample.*

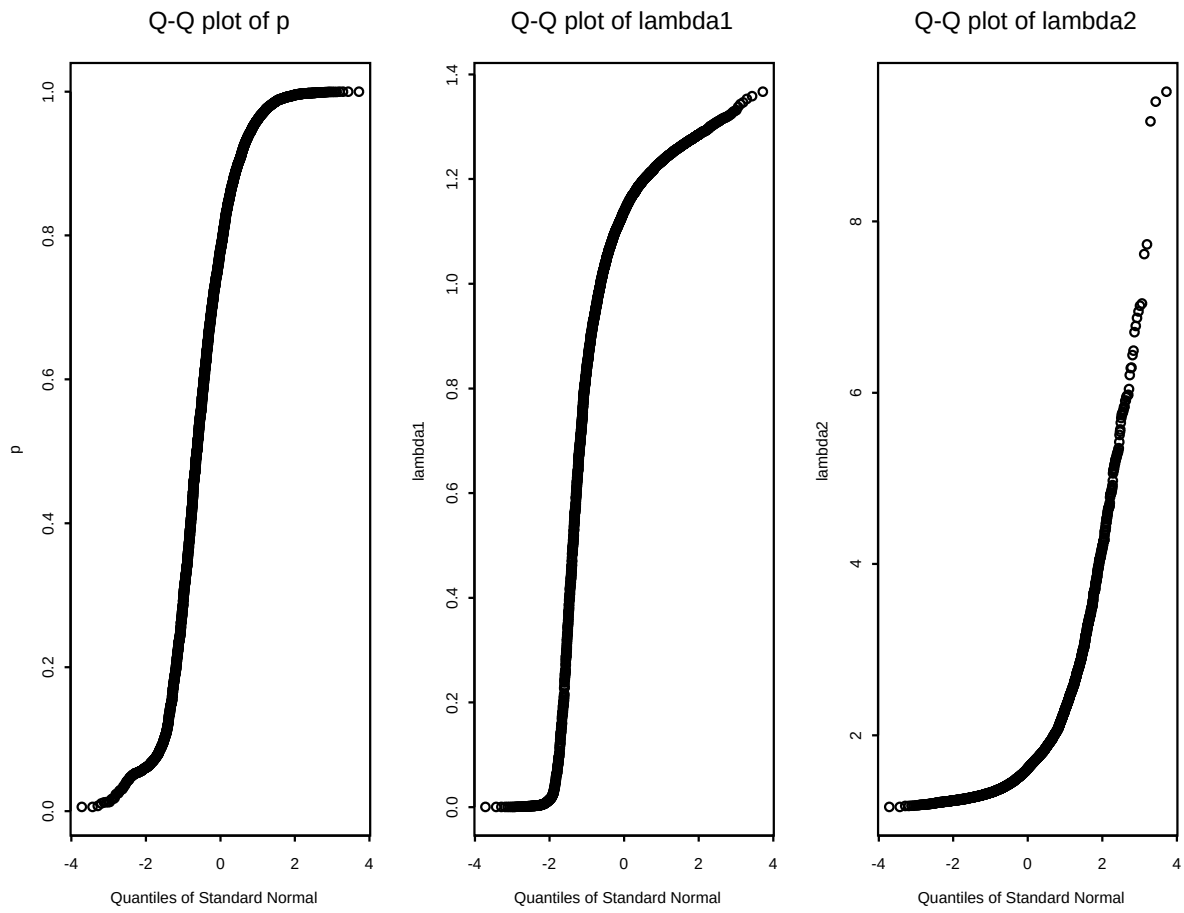

Figure 9: *QQ-plots for poorly-separated classic mixtures of Poissons*

QQ plots for 5000 *replicates of* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *when* $p_1 = 0.5, \lambda_1 = 1, \lambda_2 = 1.5$. *We simulate* 5000 *random samples, each of size* $n = 1000$ *from classic two-component mixtures of Poissons. The parameter estimates* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *are obtained via the EM algorithm for each random sample.*

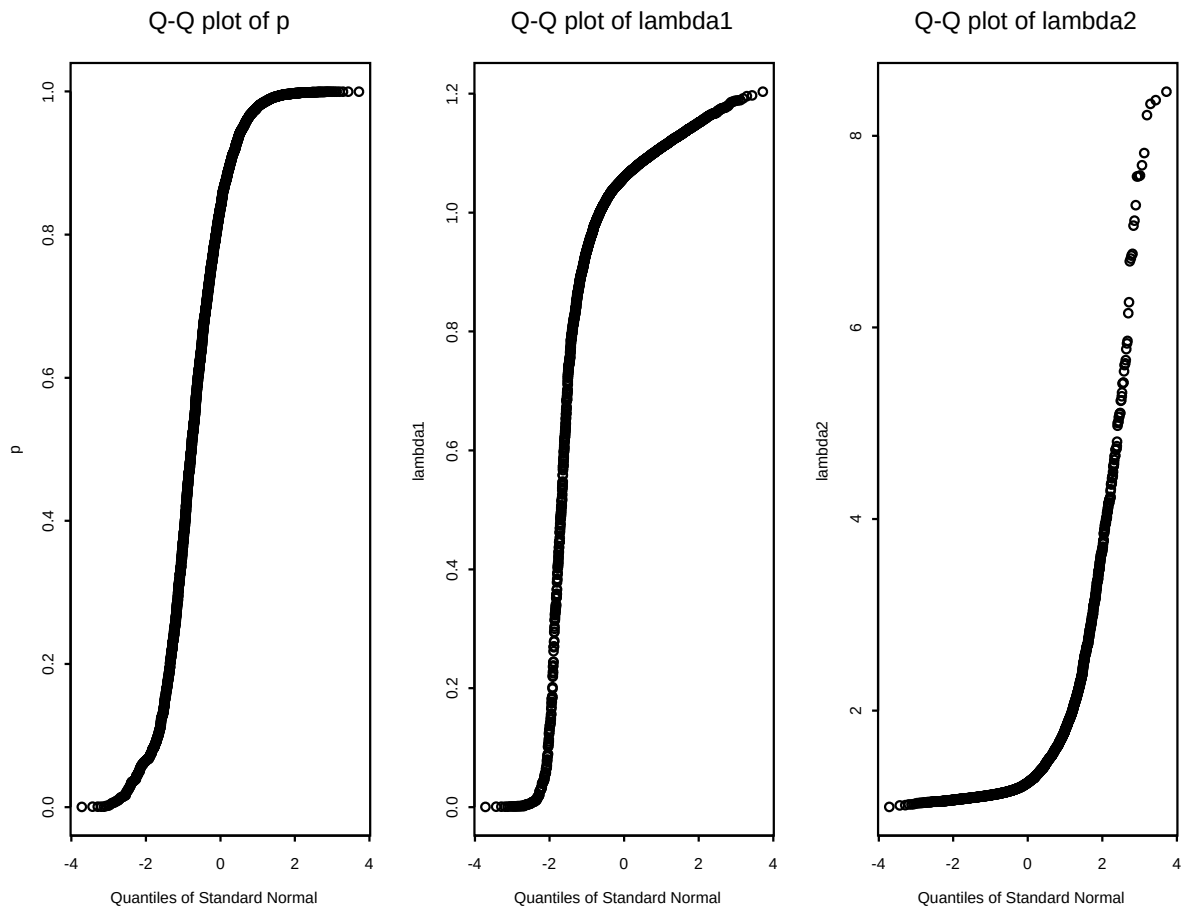

Figure 10: *QQ-plots for very-poorly-separated classic mixtures of Poissons*

QQ plots for 5000 *replicates of* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *when* $p_1 = 0.5, \lambda_1 = 1, \lambda_2 = 1.2$. *We simulate* 5000 *random samples, each of size* $n = 1000$ *from classic two-component mixtures of Poissons. The parameter estimates* $\hat{p}_1, \hat{\lambda}_1, \hat{\lambda}_2$ *are obtained via the EM algorithm for each random sample.*

APPENDIX B

APPLYING MIXTURES-OF-EXPERTS TO THE NEURON VOLUME DATA

As a preliminary model, mixtures-of-experts with two normal components is applied in this appendix to the neuron volume data set. We only consider the data on eighteen normal subjects. However, we can consider both schizophrenic and normal subjects by including the diagnostic effect in the mixtures-of-experts model as one of the covariates. In this appendix, where our main goal is to demonstrate the procedures for fitting normal component mixtures-of-experts to neuron volume data, we only use the data from the control subjects to make the problem simpler. The jth measurement on subject i, Y_{ij}, is taken to be the logarithm of the neuron volume. The covariates used here are age, gender, PMI and storage time. Assuming that all the neuron volumes within subjects and across subjects are independent, the probability density function of $\{Y_{ij}\}$ is given by

$$f(y_{ij}|\,\boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \sigma_1^2, \sigma_2^2) = p(\boldsymbol{x}_i, \boldsymbol{\gamma}) f(y_{ij}|\,\boldsymbol{x}_i, \boldsymbol{\beta_1}, \sigma_1^2) + (1 - p(\boldsymbol{x}_i, \boldsymbol{\gamma})) f(y_{ij}|\,\boldsymbol{x}_i, \boldsymbol{\beta_2}, \sigma_2^2),$$

where $f(y_{ij}|\,\boldsymbol{x}_i, \boldsymbol{\beta_k}, \sigma_k^2)$ is a normal density with mean $\mu(\boldsymbol{x}_i, \boldsymbol{\beta_k})$ and variance σ_k^2, and

$$\begin{aligned} p(\boldsymbol{x}_i, \boldsymbol{\gamma}) &= \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\gamma}}}, \\ \mu(\boldsymbol{x}_i, \boldsymbol{\beta_k}) &= \boldsymbol{x}_i^T \boldsymbol{\beta}_k, \ k = 1, 2. \end{aligned} \tag{B.1}$$

By introducing component-indicators, the EM algorithm can be implemented in the same way as in Poisson mixtures-of-experts to the grain count case, given in Section 3.5, except that the Poisson regression is replaced with linear regression to estimate the updated $\boldsymbol{\beta}_k^{(t+1)}$

and $\sigma_k^{2^{(t+1)}}$. This can be carried out by the lm( ) function of S-PLUS. We run the EM algorithm with various sets of starting values, and compute the value of the log likelihood corresponding to each set of starting values. The following solution is the one corresponding to the largest value of the log likelihood

$$\begin{aligned}
\hat{\boldsymbol{\gamma}} &= (13.7464, -0.2050, 3.1092, -0.3235, -0.0017)^T \\
\hat{\boldsymbol{\beta}}_1 &= (9.1215, -0.0197, -0.1078, -0.0321, -0.0001)^T \\
\hat{\boldsymbol{\beta}}_2 &= (11.3197, -0.0538, 1.2036, -0.1083, -0.0004)^T \\
\sigma_1^2 &= 0.3472 \\
\sigma_2^2 &= 0.6749.
\end{aligned}$$

The elements of each vector are the estimates of the intercept, the coefficients of age, gender, PMI and storage time respectively.

In Table 10, we report $\hat{p}_i$, the proportion of shorter axon neurons, $\hat{\mu}_{i1}$, the mean of the log volumes of the shorter axon neurons, and $\hat{\mu}_{i2}$, the mean of the log volumes of the longer axon neurons for each individual. It is seen from the table that these two neuron populations are not far apart. The procedures illustrated in this appendix are used in Chapter 6, where before implementing the MCMC algorithm to the model, we apply mixtures-of-experts to the neuron volume data and use the estimates as starting values for the unknown parameters in the multivariate Bernoulli mixtures of mixed normals.

Table 10: *Fitting two-component mixtures-of-experts to the neuron volume data*

	estimates		
Control subjects	$\hat{p}_i$	$\hat{\mu}_{i1}$	$\hat{\mu}_{i2}$
1	0.641	7.13	8.49
2	0.370	7.41	7.86
3	0.448	7.44	7.69
4	0.893	7.25	8.78
5	0.664	7.10	8.28
6	0.845	7.20	8.62
7	0.821	7.59	8.42
8	0.873	7.22	8.73
9	0.408	7.40	7.69
10	0.173	7.28	7.39
11	0.432	7.39	7.69
12	0.213	7.29	7.55
13	0.228	7.31	7.53
14	0.802	7.14	8.35
15	0.250	6.89	7.80
16	0.824	7.15	8.52
17	0.375	6.95	7.80
18	0.975	7.35	9.18

APPENDIX C

IDENTIFIABILITY OF MULTIVARIATE BERNOULLI MIXTURES OF NORMALS

In this appendix, the identifiability of the new model is justified numerically while it remains to be proved analytically. To simplify the problem, we assume two observations on each subject, i.e., $l_i = 2$, and only between-subject factors in the model, i.e., the covariate vector satisfies $\boldsymbol{x}_{ij} = \boldsymbol{x}_i$, for all j.

Under these assumptions, according to (5.3), the density function of (Y_{i1}, Y_{i2}) is

$$
\begin{aligned}
& f_{\boldsymbol{\theta}}(y_{i1}, y_{i2}) \\
= & \left\{ \int \frac{1}{(1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i})^2} \phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T, (\boldsymbol{x}_i^T\boldsymbol{\beta}_1, \boldsymbol{x}_i^T\boldsymbol{\beta}_1)^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
+ & \left\{ \int \frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i}} \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i}} \phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T, (\boldsymbol{x}_i^T\boldsymbol{\beta}_2, \boldsymbol{x}_i^T\boldsymbol{\beta}_1)^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}) \\
+ & \left\{ \int \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i}} \frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i}} \phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T, (\boldsymbol{x}_i^T\boldsymbol{\beta}_1, \boldsymbol{x}_i^T\boldsymbol{\beta}_2)^T, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \\
+ & \left\{ \int \frac{1}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i})^2} \phi(w_i, 0, \sigma_w^2) dw_i \right\} \phi_2((y_{i1}, y_{i2})^T, (\boldsymbol{x}_i^T\boldsymbol{\beta}_2, \boldsymbol{x}_i^T\boldsymbol{\beta}_2)^T, \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}), \qquad \text{(C.1)}
\end{aligned}
$$

where $\boldsymbol{\theta}$ denotes the unknown parameter vector $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\gamma}^T, \sigma_1^2, \sigma_2^2, \sigma_w^2)^T$. To prove identifiability, we essentially need to show that if $f_{\boldsymbol{\theta}}(y_{i1}, y_{i2}) = f_{\tilde{\boldsymbol{\theta}}}(y_{i1}, y_{i2})$, for all y_{i1}, y_{i2}, then $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Jiang and Tanner (1999) provided identifiability conditions for mixtures-of-experts. They showed that the two-normal-component mixtures-of-experts are identifiable if the components are ordered. However, their proof of identifiability cannot be applied to our model due to the complex structure of the component-indicator variables in our model.

C.1 ORDER RESTRICTION FOR PARAMETERS

As in any other mixture models, without parametric restrictions, we can show that the density f of our model is invariant under permutation of the components. Let $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\beta}_2^T, \boldsymbol{\beta}_1^T, -\boldsymbol{\gamma}^T, \sigma_2^2, \sigma_1^2, \sigma_w^2)^T$. It is easily shown that $f_{\boldsymbol{\theta}}(y_{i1}, y_{i2}) = f_{\tilde{\boldsymbol{\theta}}}(y_{i1}, y_{i2})$ from (C.1) and the following two facts

$$\begin{aligned}
\int \frac{1}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i})^2}\phi(w_i,0,\sigma_w^2)dw_i &= \int \frac{1}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i})^2}\phi(w_i,0,\sigma_w^2)dw_i, \\
\int \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i}}\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i}}\phi(w_i,0,\sigma_w^2)dw_i &= \int \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i}}\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i}}\phi(w_i,0,\sigma_w^2)dw_i.
\end{aligned}$$

The lack of the identifiability of $f_{\boldsymbol{\theta}}$ due to permutation of the components can be handled by imposing one of the following order restrictions:

1) $\sigma_1^2 < \sigma_2^2$; or

2) $\sigma_1^2 = \sigma_2^2$ and $(\boldsymbol{\beta}_1)_0 < (\boldsymbol{\beta}_2)_0$; or

3) $\sigma_1^2 = \sigma_2^2$, $(\boldsymbol{\beta}_1)_0 = (\boldsymbol{\beta}_2)_0$ and $(\boldsymbol{\beta}_1)_1 < (\boldsymbol{\beta}_2)_1$; or

$\cdots\cdots$;

q+1) $\sigma_1^2 = \sigma_2^2$, $(\boldsymbol{\beta}_1)_0 = (\boldsymbol{\beta}_2)_0, \ldots, (\boldsymbol{\beta}_1)_{q-2} = (\boldsymbol{\beta}_2)_{q-2}$, and $(\boldsymbol{\beta}_1)_{q-1} < (\boldsymbol{\beta}_2)_{q-1}$,

where $(\boldsymbol{\beta}_k)_m$ denotes the mth entry in vector $\boldsymbol{\beta}_k$ and q is the length of $\boldsymbol{\beta}_k$.

C.2 TWO EQUIVALENT CONJECTURES

We first give Conjecture 1, and then show that our model is identifiable if Conjecture 1 holds.

Conjecture 1. *For any* (a, b) *and* $(\tilde{a}, \tilde{b})$*, where* $b > 0$*,* $\tilde{b} > 0$*, if*

$$\int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{1+e^{-(a+bw)}} dw = \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{1+e^{-(\tilde{a}+\tilde{b}w)}} dw \tag{C.2}$$

$$\text{and} \quad \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{(1+e^{-(a+bw)})^2} dw = \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{(1+e^{-(\tilde{a}+\tilde{b})w})^2} dw, \tag{C.3}$$

then $a = \tilde{a}$ *and* $b = \tilde{b}$*.*

The following theorem shows that the new model is identifiable if Conjecture 1 is true.

Theorem 5. *Let* $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\gamma}^T, \sigma_1^2, \sigma_2^2, \sigma_w^2)^T$ *and* $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T, \tilde{\boldsymbol{\gamma}}^T, \tilde{\sigma_1^2}, \tilde{\sigma_2^2}, \tilde{\sigma_w^2})^T$*. Assume* $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$ *is of full rank. Given that Conjecture 1 holds, if* $f_{\boldsymbol{\theta}}(y_{i1}, y_{i2}) = f_{\tilde{\boldsymbol{\theta}}}(y_{i1}, y_{i2})$*, for all* y_{i1}, y_{i2}*, where* f *is given in (C.1), then* $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$*.*

Proof: Assume that $f_{\boldsymbol{\theta}}(y_{i1}, y_{i2}) = f_{\tilde{\boldsymbol{\theta}}}(y_{i1}, y_{i2})$, for all y_{i1}, y_{i2}, where f is a four-component mixture of bivariate normals. By the identifiability of mixtures of multivariate normals, shown in Yakowitz and Spragins (1968), we have

$$\boldsymbol{x}_i^T \boldsymbol{\beta}_1 = \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}_1; \quad \boldsymbol{x}_i^T \boldsymbol{\beta}_2 = \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}_2; \tag{C.4}$$

$$\sigma_1^2 = \tilde{\sigma_1^2}; \quad \sigma_2^2 = \tilde{\sigma_2^2}; \tag{C.5}$$

and

$$\int \frac{1}{(1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i})^2} \phi(w_i, 0, \sigma_w^2) dw_i = \int \frac{1}{(1+e^{-\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}-w_i})^2} \phi(w_i, 0, \tilde{\sigma_w^2}) dw_i; \tag{C.6}$$

$$\int \frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i}} \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i}} \phi(w_i, 0, \sigma_w^2) dw_i = \int \frac{1}{1+e^{\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}+w_i}} \frac{1}{1+e^{-\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}-w_i}} \phi(w_i, 0, \tilde{\sigma_w^2}) dw_i; \tag{C.7}$$

$$\int \frac{1}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\gamma}+w_i})^2} \phi(w_i, 0, \sigma_w^2) dw_i = \int \frac{1}{(1+e^{\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}+w_i})^2} \phi(w_i, 0, \tilde{\sigma_w^2}) dw_i. \tag{C.8}$$

It follows from (C.4) that $\boldsymbol{\beta}_1 = \tilde{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_2 = \tilde{\boldsymbol{\beta}}_2$ because $\boldsymbol{X}$ is of full rank. The equations in (C.6), (C.7), (C.8) are equivalent to

$$\int \frac{1}{1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i}} \phi(w_i, 0, \sigma_w^2) dw_i = \int \frac{1}{1+e^{-\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}-w_i}} \phi(w_i, 0, \tilde{\sigma_w^2}) dw_i; \tag{C.9}$$

$$\int \frac{1}{(1+e^{-\boldsymbol{x}_i^T\boldsymbol{\gamma}-w_i})^2} \phi(w_i, 0, \sigma_w^2) dw_i = \int \frac{1}{(1+e^{-\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}-w_i})^2} \phi(w_i, 0, \tilde{\sigma_w^2}) dw_i. \tag{C.10}$$

Let $a = -\boldsymbol{x}_i^T\boldsymbol{\gamma}$ and $\tilde{a} = -\boldsymbol{x}_i^T\tilde{\boldsymbol{\gamma}}$; $b = \sigma_w$ and $\tilde{b} = \tilde{\sigma_w}$. So that (C.9) and (C.10) reduces to (C.2) and (C.3), respectively. It then follows from Conjecture 1 that $a = \tilde{a}$ and $b = \tilde{b}$, which implies that $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$ as $\boldsymbol{X}$ is of full rank, and $\sigma_w^2 = \tilde{\sigma_w^2}$.

♢

Next we give Conjecture 2, which is equivalent to (C.9) and (C.10), so that in turn it is equivalent to Conjecture 1.

Conjecture 2. *For two normal random variables U and U' with means a and $\tilde{a}$ and variances σ^2 and $\tilde{\sigma^2}$ respectively, if*

$$E\left(\frac{1}{1+e^{-U}}\right) = E\left(\frac{1}{1+e^{-U'}}\right)$$
$$\text{and} \quad E\left(\frac{1}{1+e^{-U}}\right)^2 = E\left(\frac{1}{1+e^{-U'}}\right)^2,$$

then $a = \tilde{a}$ and $\sigma^2 = \tilde{\sigma^2}$.

C.3 NUMERICAL DEMONSTRATION

In this section we give a numerical "proof" of Conjecture 1. Let

$$\begin{aligned} g(a,b) &= \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{1+e^{-(a+bw)}} dw \\ \text{and} \qquad h(a,b) &= \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{(1+e^{-(a+bw)})^2} dw. \end{aligned}$$

The following lemmas provide some properties of $g(a,b)$, which are used later in the numerical demonstration of Conjecture 1.

Lemma 1. $\left| g(a,b) - \int_{-10}^{10} \frac{e^{-w^2/2}}{1+e^{-(a+bw)}} dw \right| < 10^{-22}$ *for all a and $b > 0$.*

Proof: Observe that $\int_{20}^{\infty} \frac{e^{-w^2/2}}{1+e^{-(a+bw)}} dw < \int_{20}^{\infty} e^{-w^2/2} dw < 10^{-23}$. Similarly $\int_{-\infty}^{-20} \frac{e^{-w^2/2}}{1+e^{-(a+bw)}} dw < 10^{-23}$.

♢

As in Lemma 1, we can approximate $h(a,b)$ by $\int_{-10}^{10} \frac{e^{-w^2/2}}{(1+e^{-(a+bw)})^2} dw$, and the difference is again less than 10^{-23}.

Lemma 2. *For fixed b, $g(a,b)$ increases with a.*

Proof: This follows since $\frac{\partial g(a,b)}{\partial a} = \int_{-\infty}^{\infty} \frac{e^{-w^2/2} e^{-(a+bw)}}{(1+e^{-(a+bw)})^2} dw > 0$.

♢

Lemma 3. *For fixed b, $g(0,b) = \sqrt{2\pi}/2$.*

Proof: Note that $\frac{\partial g(0,b)}{\partial b} = \int_{-\infty}^{\infty} \frac{we^{-w^2/2}e^{-bw}}{(1+e^{-bw})^2}dw = 0$, since the integrand is an odd function of w. Therefore, $g(0,b) = g(0,0) = \sqrt{2\pi}/2$.

♢

From Lemma 3, we can see that for any $b > 0$, $g(0,b)$ have the same values.

Lemma 4. *If $g(a,b) = g(\tilde{a},b)$, then $a = \tilde{a}$.*

Proof: Note that $g(a,b) - g(\tilde{a},b) = \int_{-\infty}^{\infty} e^{-w^2/2}e^{bw}\frac{e^{-\tilde{a}}-e^{-a}}{(e^{bw}+e^{-a})(e^{bw}+e^{-\tilde{a}})}dw$. If $g(a,b) = g(\tilde{a},b)$, then $e^{-\tilde{a}} = e^{-a}$ as $e^{-w^2/2}e^{bw}/\{(e^{bw}+e^{-a})(e^{bw}+e^{-\tilde{a}})\} > 0$ for all w, and therefore $a = \tilde{a}$.

♢

Lemma 5. $g(-a,b) = \sqrt{2\pi} - g(a,b)$.

Proof: Note that $g(-a,b) = \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{1+e^{(a-bw)}}dw = \int_{-\infty}^{\infty} \frac{e^{-w^2/2}}{1+e^{(a+bw)}}dw$, and so that $g(-a,b) + g(a,b) = \sqrt{2\pi}$.

♢

Lemma 6. *For any $a > 0$, $g(a,b)$ decreases with b and $\sqrt{2\pi}/2 < g(a,b) < \sqrt{2\pi}/(1+e^{-a}) < \sqrt{2\pi}$.*

For any $a < 0$, $g(a,b)$ increases with b and $0 < \sqrt{2\pi}/(1+e^{-a}) < g(a,b) < \sqrt{2\pi}/2$.

Proof: Note that

$$\frac{\partial g(a,b)}{\partial b} = \int_{-\infty}^{\infty} \frac{we^{-w^2/2}e^{-a-bw}}{(1+e^{-(a+bw)})^2}dw = \int_{0}^{\infty} we^{-w^2/2}\left\{\frac{e^{-(a+bw)}}{(1+e^{-(a+bw)})^2} - \frac{e^{-(a-bw)}}{(1+e^{-(a-bw)})^2}\right\}dw.$$

Next let $t = e^{-a}$ and $u = e^{-bw}$, so that

$$\Delta \equiv \frac{e^{-(a+bw)}}{(1+e^{-(a+bw)})^2} - \frac{e^{-(a-bw)}}{(1+e^{-(a-bw)})^2} = \frac{tu(1-t^2)(u^2-1)}{(1+tu)^2(u+t)^2}.$$

Note that $u^2 - 1 < 0$, since $0 < u < 1$ for all $w > 0$. It follows that $a > 0 \Rightarrow 1 - t^2 > 0 \Rightarrow \Delta < 0 \Rightarrow \frac{\partial g(a,b)}{\partial b} < 0$. Consequently, $\sqrt{2\pi}/2 = g(a,\infty) < g(a,b) < g(a,0) < \sqrt{2\pi}/(1+e^{-a}) < \sqrt{2\pi}$. Similarly, for $a < 0$, $g(a,b)$ increases with b and $0 < \sqrt{2\pi}/(1+e^{-a}) < g(a,b) < \sqrt{2\pi}/2$.

♢

Numerical Demonstration

We now give the numerical "proof" of Conjecture 1. According to Lemma 1, in this demonstration, for any a and $b > 0$, both $g(a, b)$ and $h(a, b)$ are approximated by integrals with integrating variable w varying from -10 to 10.

The essential idea of the numerical demonstration is now easy to convey. By Lemma 6, the range of function g is $(0, \sqrt{2\pi})$. For any value $c \in (0, \sqrt{2\pi})$, Let K_c={(a, b), g(a, b)=c}. If every two different points in K_c have different h values, then Conjecture 1 holds.

Based on Lemma 6, there are two cases of all possible K_c.

Case I For any $c \in (\sqrt{2\pi}/2, \sqrt{2\pi})$, all (a, b) in K_c are satisfying $a > 0$.

Case II For any $c \in (0, \sqrt{2\pi}/2)$, all (a, b) in K_c are satisfying $a < 0$. Furthermore, By Lemma 5, $g(-a, b) = \sqrt{2\pi} - c$.

It is apparent that we only need to focus on finding all possible K_c in Case I. After finding K_c in Case I, the K_c in Case II can be obtained directly from the sets in Case I by changing the sign of a.

For any chosen $c \in (\sqrt{2\pi}/2, \sqrt{2\pi})$, solving the corresponding K_c is a very difficult task. To get around this problem, we choose a grid of (a_0, b_0) and denote it by G, where $G = \{(a_0, b_0), a_0 = 0.5, b_0 = 0.1, 0.2, \ldots, 3$; or $a_0 = 1.5, b_0 = 0.1, 0.2, \ldots, 10$; or $a_0 = 5, b_0 = 0.1, 0.2, \ldots, 10\}$. The range of $g(a_0, b_0)$, for all $(a_0, b_0) \in G$, is (1.40, 2.49), which almost covers $(\sqrt{2\pi}/2, \sqrt{2\pi})$. We are unable to choose $g(a_0, b_0)$ closer to $\sqrt{2\pi}/2$, which is approximately 1.25, because of numerical instability.

For each $(a_0, b_0) \in G$, let $c = g(a_0, b_0)$ and then find the corresponding K_c. To do this, we use a grid of possible a value to examine, namely, $a = 0.1, 0.2, \ldots, 5$. For each a value a^*, we use Mathcad to find the unique b, $0 < b < 30$, so that $g(a^*, b) = g(a_0, b_0)$. For example, when $(a_0, b_0) = (1.5, 0.2)$ and $a^* = 1.6$, the b value is 0.613. In fact, for any $a > 0$, there is a unique b such that $(a, b) \in K_c$. By our algorithm, we only pick countable points in K_c. However, due to the continuity and monotonicity of h function, if there is no identical h value for any of these countable points, then it is true for any point in K_c.

We performed the above procedures in Mathcad Professional 2001. We did not find any two points in K_c having the same h values. Based on our numerical demonstration, we

conclude that Conjecture 1 is true, which "proves" that our new model is identifiable.

The preceding "proof" procedure is based on the assumption of two observations on each subject. All the results apply to any finite number of observations on each subject.

APPENDIX D

DETAILS OF THE SAMPLING SCHEME FOR THE MULTIVARIATE BERNOULLI MIXTURES OF NORMALS

In this appendix, we give details about the Metropolis-Hastings steps to sample from the conditional distribution of $(\boldsymbol{\gamma},\ \boldsymbol{w})$ block and from the conditional distribution of $(\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2)$ block. In both cases, our proposal distribution is multivariate t with mean and variance equal respectively to the mode of the appropriate conditional distribution and the inverse curvature of the log of this conditional distribution at the mode.

D.1 UPDATING $\boldsymbol{\gamma}$ AND $\boldsymbol{W}$

The logarithm of $p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)$ is given by

$$\sum_{i=1}^{n}\sum_{j=1}^{l_i}\{z_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i)-\log(1+e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i})\}-\frac{\boldsymbol{\gamma}^T\boldsymbol{\gamma}}{2\sigma_\gamma^2}-\frac{\sum_{i=1}^{n}w_i^2}{2\sigma_w^2}. \tag{D.1}$$

The mode of (D.1) is obtained via a Quasi-Newton algorithm maximization routine using the derivatives

$$\begin{aligned}
\frac{\partial \log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n}\sum_{j=1}^{l_i}\{z_{ij}-\frac{e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}{1+e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}\}\boldsymbol{x}_{ij}-\frac{1}{\sigma_\gamma^2}\boldsymbol{\gamma}\\
\frac{\partial \log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)}{\partial w_i} &= \sum_{j=1}^{l_i}\{z_{ij}-\frac{e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}{1+e^{\boldsymbol{x}_{ij}^T\boldsymbol{\gamma}+w_i}}\}-\frac{w_i}{\sigma_w^2}.
\end{aligned}$$

Let $\boldsymbol{m_0}$ denote the mode of $p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)$ and V_0 denote the inverse curvature of $\log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)$ at $\boldsymbol{m_0}$. The necessary second derivatives are

$$\begin{aligned}
\frac{\partial^2 \log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} &= -\sum_{i=1}^{n} \sum_{j=1}^{l_i} \left\{ \frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \frac{1}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \right\} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T - \frac{1}{\sigma_\gamma^2} I \\
\frac{\partial^2 \log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)}{\partial w_i^2} &= -\sum_{j=1}^{l_i} \frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \frac{1}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} - \frac{1}{\sigma_w^2} \\
\frac{\partial^2 \log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)}{\partial \boldsymbol{\gamma} \partial w_i} &= -\sum_{j=1}^{l_i} \frac{e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \frac{1}{1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i}} \boldsymbol{x}_{ij}.
\end{aligned}$$

Let the proposal density $f_T(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{m_0}, \tau V_0, \upsilon)$ be a multivariate t distribution with υ degrees of freedom, location parameter vector $\boldsymbol{m_0}$ and scale matrix τV_0, where υ and τ are tuning constants. In Section 5.4 and Section 5.5, we use $\upsilon = 4$ and $\tau = 1$.

We propose $(\boldsymbol{\gamma}^*, \boldsymbol{w}^*) \sim f_T(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{m_0}, \tau V_0, \upsilon)$ and accept it with probability

$$\min \left\{ \frac{p(\boldsymbol{\gamma}^*, \boldsymbol{w}^* | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2) f_T(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{m_0}, \tau V_0, \upsilon)}{p(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2) f_T(\boldsymbol{\gamma}^*, \boldsymbol{w}^* | \, \boldsymbol{m_0}, \tau V_0, \upsilon)}, \; 1 \right\}.$$

D.2 UPDATING $\boldsymbol{\beta}_1$ AND $\boldsymbol{\beta}_2$

The logarithm of $p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)$ is given by

$$\sum_{i=1}^{n} \sum_{j=1}^{l_i} \left\{ z_{ij} \frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1)^2}{2\sigma_1^2} - (1 - z_{ij}) \frac{(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2)^2}{2\sigma_2^2} \right\} - \frac{\boldsymbol{\beta_1}^T \boldsymbol{\beta_1}}{2\sigma_{\beta_1}^2} - \frac{\boldsymbol{\beta_2}^T \boldsymbol{\beta_2}}{2\sigma_{\beta_2}^2}. \tag{D.2}$$

The derivatives of (D.2) are as follows.

$$\begin{aligned}
\frac{\partial \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_1} &= \frac{1}{\sigma_1^2} \sum_{i=1}^{n} \sum_{j=1}^{l_i} z_{ij} (y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_1) \boldsymbol{x}_{ij} - \frac{1}{\sigma_{\beta_1}^2} \boldsymbol{\beta}_1 \\
\frac{\partial \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_2} &= \frac{1}{\sigma_2^2} \sum_{i=1}^{n} \sum_{j=1}^{l_i} (1 - z_{ij}) (y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_2) \boldsymbol{x}_{ij} - \frac{1}{\sigma_{\beta_2}^2} \boldsymbol{\beta}_2 \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} &= \frac{1}{\sigma_1^2} \sum_{i=1}^{n} \sum_{j=1}^{l_i} z_{ij} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T - \frac{1}{\sigma_{\beta_1}^2} I \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^T} &= \frac{1}{\sigma_2^2} \sum_{i=1}^{n} \sum_{j=1}^{l_i} (1 - z_{ij}) \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T - \frac{1}{\sigma_{\beta_2}^2} I.
\end{aligned}$$

The Metropolis-Hastings algorithm for this block is analogous to the one in Appendix D.1. The tuning constants in the multivariate t distribution for this block are chosen as $\upsilon = 2.5$ and $\tau = 10$ in Section 5.4 and Section 5.5.

APPENDIX E

DETAILS OF THE SAMPLING SCHEME FOR MULTIVARIATE BERNOULLI MIXTURES OF MIXED NORMALS

In this appendix, we give the details of the Metropolis-Hastings steps to sample from the condition distribution of ($\boldsymbol{\gamma}$, $\boldsymbol{w}$) block, and from the condition distribution of ($\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{s}$) block.

E.1 UPDATING $\boldsymbol{\gamma}$ AND $\boldsymbol{W}$

The logarithm of the conditional distribution $p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \sigma_w^2\right)$ is given by

$$\sum_{i=1}^{n} \sum_{j=1}^{l_i} \{z_{ij}(\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i) - \log(1 + e^{\boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + w_i})\} - \frac{\boldsymbol{\gamma}^T \boldsymbol{\gamma}}{2\sigma_\gamma^2} - \frac{\sum_{i=1}^{n} w_i^2}{2\sigma_w^2},$$

which is the same as $\log p\left(\boldsymbol{\gamma}, \boldsymbol{w} | \, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \sigma_w^2\right)$ given in Appendix D.1. Thus, the Metropolis-Hastings steps to sample $(\boldsymbol{\gamma}, \boldsymbol{w})$ in the proposed model are identical to the procedures to sample $(\boldsymbol{\gamma}, \boldsymbol{w})$ in the multivariate Bernoulli mixtures of normals given in Appendix D.1.

E.2 UPDATING $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, AND $\boldsymbol{S}$

The logarithm of $p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)$ can be written as

$$\sum_{i=1}^{n}\sum_{j=1}^{l_i}\left\{z_{ij}\frac{(y_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1-s_i)^2}{2\sigma_1^2}-(1-z_{ij})\frac{(y_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2-s_i)^2}{2\sigma_2^2}\right\}-\frac{\boldsymbol{\beta_1}^T\boldsymbol{\beta_1}}{2\sigma_{\beta_1}^2}-\frac{\boldsymbol{\beta_2}^T\boldsymbol{\beta_2}}{2\sigma_{\beta_2}^2}-\frac{\sum_{i=1}^n s_i^2}{2\sigma_s^2}$$

The gradient vector and Hessian matrix of $\log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)$ are obtained using

$$\begin{aligned}
\frac{\partial \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_1} &= \frac{1}{\sigma_1^2}\sum_{i=1}^{n}\sum_{j=1}^{l_i} z_{ij}(y_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1-s_i)\boldsymbol{x}_{ij}-\frac{1}{\sigma_{\beta_1}^2}\boldsymbol{\beta}_1 \\
\frac{\partial \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_2} &= \frac{1}{\sigma_2^2}\sum_{i=1}^{n}\sum_{j=1}^{l_i} (1-z_{ij})(y_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2-s_i)\boldsymbol{x}_{ij}-\frac{1}{\sigma_{\beta_2}^2}\boldsymbol{\beta}_2 \\
\frac{\partial \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{s}} &= \frac{1}{\sigma_1^2}\sum_{j=1}^{l_i} z_{ij}(y_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_1-s_i) \\
&\quad +\frac{1}{\sigma_2^2}\sum_{j=1}^{l_i}(1-z_{ij})(y_{ij}-\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_2-s_i)-\frac{s_i}{\sigma_s^2} \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T} &= \frac{1}{\sigma_1^2}\sum_{i=1}^{n}\sum_{j=1}^{l_i} z_{ij}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T-\frac{1}{\sigma_{\beta_1}^2}I \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^T} &= \frac{1}{\sigma_2^2}\sum_{i=1}^{n}\sum_{j=1}^{l_i} (1-z_{ij})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^T-\frac{1}{\sigma_{\beta_2}^2}I \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial s_i^2} &= -\frac{1}{\sigma_1^2}\sum_{j=1}^{l_i} z_{ij}-\frac{1}{\sigma_2^2}\sum_{j=1}^{l_i}(1-z_{ij})-\frac{1}{\sigma_s^2} \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_1 \partial s_i} &= \frac{1}{\sigma_1^2}\sum_{j=1}^{l_i} z_{ij}\boldsymbol{x}_{ij} \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_2 \partial s_i} &= \frac{1}{\sigma_2^2}\sum_{j=1}^{l_i}(1-z_{ij})\boldsymbol{x}_{ij} \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2^T} &= 0 \\
\frac{\partial^2 \log p\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{s} | \, \boldsymbol{y}, \boldsymbol{z}, \sigma_1^2, \sigma_2^2, \sigma_s^2, \boldsymbol{\gamma}, \boldsymbol{w}, \sigma_w^2\right)}{\partial s_i s_j} &= 0, \; i \neq j.
\end{aligned}$$

The Metropolis-Hastings algorithm for this block is analogous to the one in Appendix D.1.

BIBLIOGRAPHY

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251-262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117-128.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679.

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

Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical Statistics* **37**, 1554-1563.

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics* **41**, 164-171.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* **75**, 79-97.

Chib, S., Greenberg, E., and Winkelmann, R. (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics* **86**, 33-54.

Chib, S., and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96**, 270-281.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21**, 113-120.

Crawford, S. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association* **89**, 259-267.

Dempster, A. P., Laird, N. M., and Rubin, D. B.(1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1-22.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B* **56**, 363-375.

Donoho, D.L.(1988). One-sided inference about functionals of a density. *Annals of Statistics* **16**, 1390-1420.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1-26.

efro1 Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457-487.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. 2nd ed. Chapman & Hall, London.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Gundersen, H. J (1988). The nucleator. *Journal of Microscopy* **151**, 3-21.

Hashimoto, T., Volk, D.M., 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 prefrontal cortex of subjects with schizophrenia. *Journal of Neuroscience* **23**, 6315-6326.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Heinen, T. (1996) *Latent Class and Discrete Latent Trait Models*. Thousand Oaks, CA: SAGE.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Computation* **3**, 79-87.

Jiang, W. and Tanner, M. A. (1999). On the Identifiability of Mixtures-of-Experts. *Neural Networks* **12**, 1253-1258.

Jordan, M.I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**, 181-214.

Karlis, D. and Xekalaki, E. (1999). On testing for the number of components in a mixed Poisson model. *Annals of the Institute of Statistical Mathematics* (51), 149-162.

Lazarsfeld, P. F. (1950) The logical and mathematical foundation of latent structure analysis. In *Measurement and Prediction.*(S. A. Stouffer et al., eds), 362-412. Princeton, N.J.: Princeton University Press.

Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545-558.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Littell, R.C., Milliken, G.A., Struop, W. W., and Wolfinger, R.D.(1996. *SAS System for Mixed Models.* SAS Institute inc., Cary, NC.

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 Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley & Sons.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087-1091.

Müller, P. (1993). A generic approach to posterior integration and Gibbs sampling. Technical Report. Purdue University.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied Linear Statistical Models.* Burr Ridee, Illinois: IRWIN.

Peng, F., Jacobs, R.A., and Tanner, M.A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association* **91**, 953-960.

Qian, W. and Titterington, D. M. (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London A* **337**, 407-428.

Robert, C.P., Celeux, G., and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics & Probability Letters* **16**, 77-83.

Robert, C.P., Rydén, T., and Titterington, D.M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society B* **62**, 57-75.

Rosen, O., Jiang, W., and Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika* **87**, 391-404.

Rubin, D. B. and Wu, Y. (1997). Modeling schizophrenic behavior using general mixture components. *Biometrics* **53**, 243-261.

Rydén, T. and Titterington, D.M. (1998). Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics* **7**, 194-211.

Schilling M.F., Watkins A.E., Watkins W. (2002). Is Human Height Bimodal? *The American Statistician* **56**, 223-229.

Scott, S. L. (2002). Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association* **97**, 337-351.

Sweet, R.A., Pierri, J.N., Auh, S., Sampson, A.R., and Lewis, D.A. (2003). Reduced pyramidal cell somal volume in auditory association cortex of subjects with schizophrenia. *Neuropsychopharmacology* **28**, 599-609.

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions,* 3rd ed. New York: Springer.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics* **22**, 1701-1728.

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

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics* **39**, 209-214.

Zeger, S.L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.