FINITE MIXTURE REGRESSION MODELS AND APPLICATIONS: DETECTION LIMIT AND GOODNESS-OF-FIT TEST

BY JUNWU SHEN

A dissertation submitted to the

The School of Public Health

University of Medicine and Dentistry of New Jersey

and the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Biostatistics

Written under the direction of

Dr. Shou-En Lu

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2011

ABSTRACT OF THE DISSERTATION

Finite Mixture Regression Models and Applications: Detection Limit and Goodness-of-Fit Test

by Junwu Shen

Dissertation Director: Dr. Shou-En Lu

Finite mixture models have been used to analyze data in a heterogeneous population. In this dissertation, we aimed to propose two statistical methodologies: one is the mixture regression model analysis accommodating repeated measures and observations under detection limits; the other is the GOF test for evaluating the model fit of mixture regression models, with and without random effects. Both methodologies were applied to the analysis of the 8-isoprostane data in the HEART study.

A general framework for random effects finite mixture regression models was proposed to analyze repeatedly measured data with observations under a detection limit. The non-detectables were treated as left-censored observations and the regression parameters were estimated by the maximum likelihood method. Using normal mixture models as an example, we demonstrated that the parameter estimators are unbiased.

In addition, we proposed a prototype goodness-of-fit test method based on the principle of cumulative residuals. Cumulative pseudo-residuals were defined based on the score functions and then a GOF test was proposed accordingly for two-component normal mixture models with and without random effects. Extensive simulation studies showed that the proposed GOF tests maintained the type I error rate, and had a reasonable power to detect model deviations.

Acknowledgements

I would like to express my sincerest gratitude to my thesis advisor, Dr. Shou-En Lu, for her insightful guidance, constant encourage, diligence, expertise in this area which lead to the successful completion of the thesis. This thesis would not have come to fruition without the steady support, patience, and wise counsel of Dr. Lu.

I would also like to thank all my thesis committee members: Drs. Weichung (Joe) Shih, Yong Lin, Yujun Wu, Junfeng (Jim) Zhang. Thank you all for your support and suggestions throughout my dissertation work. In particular, I want to thank Dr. Zhang for his permission to use the HEART study data for my dissertation work, and his comments on how to interpret the data.

My deepest love and gratidue are reserved for my wife, Fengjuan Xuan and my lovely daughter, Hannah Shen. I would like to thank my wife for her support, encouragement and understanding during my Ph.D. study at UMDNJ. My dissertation would not have been completed without her consistent support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Usually data collected from a mixture of heterogeneous sub-populations cannot be sufficiently described by a single distribution. Depending on the degree of heterogeneity among all subpopulations, the data usually displays a unimodal, bimodal or even multimodal distribution. For example, in Li et al., 2006, the distribution of HIV RNA data showed a bimodal feature because the HIV-infected patients were mixed with two subpopulations with one receiving suboptimal background therapy and the other more potent therapy. When the "membership" of subpopulations was unobserved, these types of data are typically analyzed by a mixture model with a finite number of component distributions.

Detection limit is the lowest amount of a substance that can be detected in a sample. It is a common problem in bioassay analyses, radiochemistry and many other areas (Currie, 1968; Joos et al., 1985). For example, the quantification limit of the Ultrasensitive Roche Amplicor HIV-1 Monitor Assay was 50 copies/mL, so the viral load of an HIV infected patient less than 50 copies/mL would not be detected (Erali and Hillyard, 1999). Many methods have been used to handle the detection limit problem in statistical data analysis. The simplest method is to impute the nondetectables by the value of the detection limit or half of the detection limit. However, this method is known to give biased results, especially when the proportion of observations below detection limit is relatively high. Some improved, yet more statistically sophisticated approaches include treating observations below detection limit as left-censored data (Li et al., 2006), or viewing the entire sample as semi-continuous data to be analyzed by the two-part model approach (Berk and Lachenbruch, 2002; Moulton et al., 2002; Taylor et al., 2001). The former approach will be the focus of the first methodology studied

in this dissertation.

It is important to evaluate the goodness-of-fit (GOF) of a statistical model. Evaluating the GOF of a mixture model can be a complicated and challenging issue. Many statistical approaches have been proposed to evaluate the GOF based on the number of mixture components. The penalized model selection criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) were the easiest and most commonly used approach (Bozdogan, 1987; Leroux, 1992). McLachlan (1987) and Susko (2003) in turn proposed confirmatory statistical tests to determine the number of mixture component. Lindsay and Roeder (1992) proposed a non-parametric approach to evaluate the fit. However, few literature discussed how to evaluate the GOF of a mixture regression model to the data. In this dissertation, we proposed a second methodology, that is, to evaluate the GOF of a mixture regression model using the cumulative residual approach (Lin et al., 1993, Spiekerman and Lin, 1996, Su and Wei, 1991, Lin et al., 2002, Pan and Lin, 2005). Basic rationale of the cumulative residual approach is reviewed in Chapter 2.

In this dissertation, we aimed to propose two statistical methodologies: one is the mixture regression model analysis accommodating repeated measures and observations under detection limit; the other is the GOF test for evaluating the model fit of mixture regression models, with and without random effects.

1.1 A motivating example

The motivating example of our study came from the data in the Beijing Health Effects of Air Pollution Reduction Trial (HEART) Study, which aimed to evaluate physiological responses to drastic changes in air pollutant levels before, during and after the 2008 Olympic Games. It has been shown that short-term changes in air pollutant levels, particularly fine particulate matter $PM_{2.5}$ (particulate matter with aerodynamic diameter less than 2.5 μm), are associated with cardiopulmonary morbidity and mortality (Pope and Dockery, 2006). However, the precise mechanisms that underlie these associations are not well-understood. China is one of the few countries with highest

air pollution levels. To lower the air pollutant levels during the 2008 Beijing Olympic Games, a series of aggressive measures were implemented to reduce pollutant emissions in Beijing and its surrounding areas, especially from mid-July to mid-September, 2008. With significantly improved air quality in Beijing during 2008 Olympics Games, it provided a unique opportunity to study the biological mechanisms of cardiovascular and pulmonary responses in relation to ambient pollutant levels with wide variations.

HEART study is a panel design study aimed to examine changes in multiple biomarkers in response to drastic changes in air pollutant levels before, during and after the 2008 Olympic Games. A total of 128 healthy and non-smoking Chinese hospital medical residents (students) with age between 25 and 35 years old were recruited. The study was classified into three periods: "pre-Olympic period" (June 2 to July 7), "during-Olympic period" (August 1 to August 29), and "post-Olympic period" (September 30 to October 31). Subjects were measured for the biomarkers during their scheduled six clinical visits. Within each time period, each subject had a suite of health measurements made repeatedly on two days, separated by approximately two weeks. Each subject was scheduled to have measurements made on the same weekday within and across all three periods.

One biomarker, EBC 8-isoprostane (8-ISO), an index of oxidative stress, was measured with a detection limit of 1.56 pg/ml. The summary statistics stratified by clinical visit were shown in Table 1.1. Values below 1.56 pg/ml were replaced by 1.56/2 pg/ml, following a commonly used conventional practice. It can be observed that a substantial amount of data, nearly 26 to 56%, was below detection limit. Imputing observations below detection limit by the value or one half of the value of detection limit may create substantial bias especially when the percentage of nondetectables is high, such as in this example data.

Besides the detection limit issue, the data also displayed a mulit-modal distribution, as shown in Figure 1.1. Figure 1.2 shows the Q-Q plots stratified by clinical visit with the normal fit based on the data above the detection limit. It can be seen that even if the data below detection limit is ignored, there is a significant deviation from the straight line, and this suggests that a single normal distribution is not appropriate

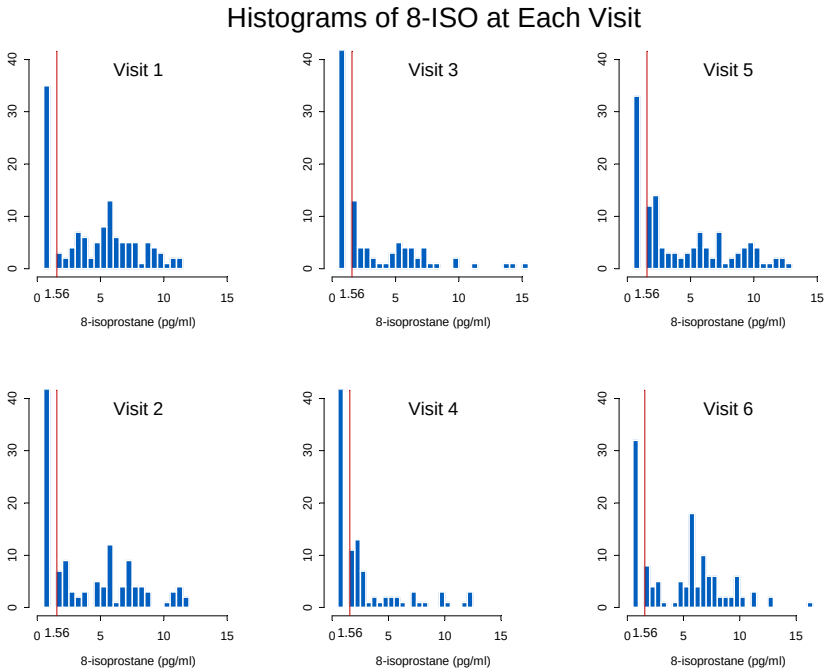

Figure 1.1: Frequency plots of 8-isoprostane data stratified by clinical visit (the vertical line at $x = 1.56$ denotes the detection limit of 1.56 pg/ml).

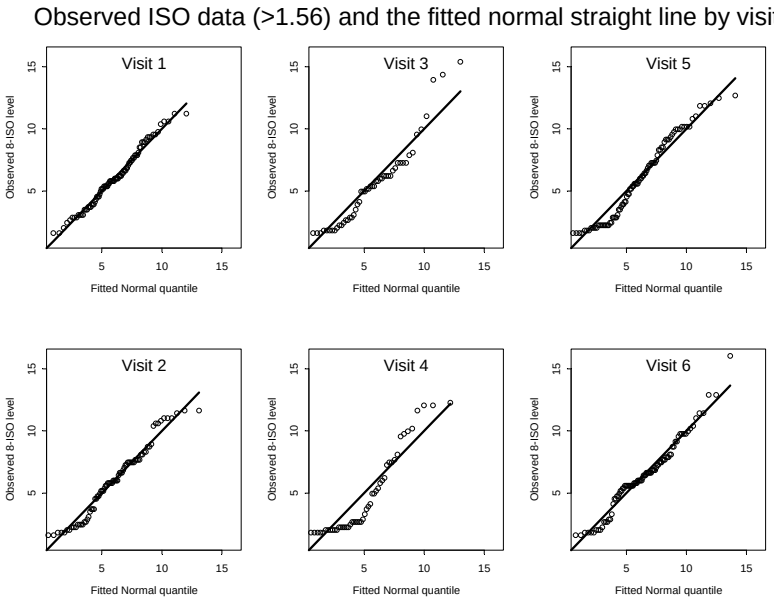

Figure 1.2: Observed 8-isoprostane data (>1.56 pg/ml) and the fitted normal straight line by clinical visit

Table 1.1: Summary statistics of EBC 8-isoprostane, stratified by clinical visit.

Visit	N	Mean	STD	MIN	Q1	Median	Q3	Max	%BDL*
1	128	4.42	3.12	0.78	0.78	4.54	6.63	11.22	30.47
2	124	4.07	3.38	0.78	0.78	2.77	6.74	11.64	36.29
3	125	2.67	3.08	0.78	0.78	0.78	3.91	15.40	56.00
4	124	2.37	2.83	0.78	0.78	0.78	2.35	12.27	56.45
5	124	4.31	3.58	0.78	0.78	2.87	7.16	12.69	28.23
6	124	4.88	3.48	0.78	0.78	5.58	7.16	16.03	26.61

*: BDL=below detection limit.

for this type of data. Ignoring the multi-modal distribution of the data may result in erroneous conclusions in both statistical and scientific inference.

1.2 Research questions and objectives

The data example in the previous section exhibited the issues of nondetectables, multi-modal distributions, and repeated measures. In this dissertation, we propose two methodologies. The first one deals with the issues listed above. Specifically, we proposed a random effect finite mixture regression model, accommodating measurements under detection limit. The second one deals with the GOF of mixture regression models. We proposed a prototype GOF test and outlined its extension to accommodate nondetectables as future work.

The rest of this dissertation is organized as follows. In Chapter 2, we review the literature for the analysis of finite mixture models, repeatedly measured data with a detection limit, and the model-checking techniques. Chapter 3 introduces the proposed finite mixture regression model for repeated measures with a detection limit, and its application in the analysis of a data example. A new goodness-of-fit test for finite mixture models with and without random effects is proposed in Chapter 4. In Chapter 5 we study the performance of the proposed goodness-of-fit tests via extensive simulation studies. Then, the proposed GOF tests are applied to the data example in Chapter 6. Chapter 7 summarizes the conclusions and future research work.

Chapter 2

Literature Review

The example data in the previous chapter have some important features: 1). Mixture of more than one distribution, 2).Observations below detection limit, 3). Repeated measures. The combination of these features complicates the data analysis. A brief literature review regarding statistical analysis of these types of data is given. In addition, goodness-of-fit tests for checking model adequacy are also reviewed.

2.1 Finite Mixture Models

Finite mixture models are commonly used to model data from a heterogeneous population. For example, Li, et al (2006) has shown that HIV RNA levels (in Log10 scale) in highly active antiretroviral therapy (HAART) treated population have a bimodal distribution, and a two component mixture model fitted the data significantly better than the model based on a single distribution.

To begin our discussion on finite mixture modeling, we first assumed that the study population of interest consists of i.i.d. (independently and identically distributed) subjects from a mixture of K subpopulations with different characteristics specified by parameter $\boldsymbol{\theta}_k$, for $k = 1, 2, ..., K$. Let Y be the random variable from that population, the density function of Y can be written as follows:

$$f(y; \mathbf{p}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f_k(y; \boldsymbol{\theta}_k) \tag{2.1}$$

where $\mathbf{p} = (p_1, ..., p_k)$ denotes the mixing proportions, or the probability of an individual belonging to the kth subpopulation with the constraint that $\sum_{i=1}^{K} p_k = 1$, $f_k(y; \boldsymbol{\theta}_k)$ is the conditional density given that the observation is from the kth subpopulation, and

$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$ and $\boldsymbol{\theta}_k$ is a parameter vector, for $k = 1, 2, ..., K$.

2.1.1 Parameter estimation

Under (2.1), $\Psi = (p_1, ..., p_{K-1}, \theta_1, ..., \theta_K)$ can be estimated with many different methods including the maximum likelihood approach, Bayesian estimation, inversion and error minimization, etc. (Everitt and Hand, 1981). For mixtures of normal distributions, Ψ can also be estimated using the method of moments (Everitt and Hand, 1981). In addition, a normal probability plot can be useful for evaluating the estimate of means, variances and mixing proportion of each component (Everitt and Hand, 1981). This graphical technique was useful before computers were widely available, and it is still used to provide good starting estimates when using some computer-based estimation procedures (Everitt and Hand, 1981).

2.1.2 Estimating the number of components

Estimating the number of components in a finite mixture model is one of the challenges in finite mixture model analysis. Two types of methods have been developed: exploratory graphical techniques and confirmatory hypothesis testing (Everitt and Hand, 1981). The simplest graphical method is sample histogram. A bimodal or multimodal histogram reveals a finite mixture distribution. However, a unimodal histogram does not necessarily mean that the sample data are from a single modal distribution. For example, when the means of two normal components are close to each other, the mixture normal distribution can be unimodal. A Fowlkes probability plot is a useful graphical tool compared to a sample histogram. It is more sensitive to detect the presence of a normal mixture than the regular quantile-quantile plot (Fowlkes, 1979). In a Fowlkes probability plot, standardized sample quantiles $x_{(i)}$ is plotted against the quantiles $y_i = \Phi(x_{(i)}) - b_i$, where Φ is the cdf of the standard normal distribution, and $b_i = (i - 1/2)/n$. An approximate horizontal line at $y = 0$ indicates a single normal distribution. An S-shape Fowlkes probability plot suggests that the data come from a mixture distribution (Fowlkes, 1979; Everitt and Hand, 1981).

In addition to graphical methods, some hypothesis tests, such as likelihood ratio

test, were proposed to determine the number of components in a finite mixture model. However, because Because the mixing proportions lie on the boundary of the parameter space and the regularity conditions do not hold, the regular likelihood ratio test statistic does not follow an asymptotic chi-square distribution and cannot be used to test the number of components in a mixture model (Wolfe, 1971; Binder, 1978; and Hartigan, 1985). Wolfe (1971) suggested a modified chi-square test to test the hypothesis $\mathrm{H}_0 : c = c_0$ against $\mathrm{H}_a : c = c_1$, where c represents the number of components in the mixture. The test statistic is $-\frac{2}{n}(n - 1 - d - \frac{c_1}{2})log(\lambda)$, where λ is the likelihood ratio L_{c_0}/L_{c_1}, n is the sample size and d is the number of parameters in the model. It was shown that this test statistic follows an approximate chi-square distribution with degrees of freedom $2d/(c_1 - c_0)$ under null hypothesis. Johnson (1973), Baker (1958) and Binder (1978) also proposed similar tests to determine the number of components in a mixture model. Penalized model selection criteria such as Akaike's information criterion (AIC) and Bayesian information criterion (BIC) were suggested to determine the number of components (Bozdogan, 1987; Leroux, 1992). McLachlan (1987) proposed a bootstrapping method to obtain the null distribution of the likelihood ratio statistic, but it is computationally intensive. Therefore, Susko (2003) proposed weighted tests of homogeneity for testing the number of components in a mixture.

2.1.3 Finite mixture regression models and applications

In many application areas, relevant risk covariates are introduced to finite mixture models, and this is called finite mixture regression model. Usually, these covariates are associated with the distribution parameters or even the mixing proportions in the mixture model (Dietz, 1968; Dietz and Bohning, 1996).

2.1.3.1 Univariate mixture regression models

A simple example of mixture regression models is defined as follows:

$$f(y_i, \mathbf{x}_i; \omega) = \sum_{k=1}^{K} p_k(\boldsymbol{\alpha}_k) f_k(y_i, \mathbf{x}_i; \boldsymbol{\theta}_k)$$

where ω denotes the vector for all regression parameters including $\boldsymbol{\alpha}_k$ and $\boldsymbol{\theta}_k$, y_i is the response for subject i, and x is the predictors (McLachlan and Peel, 2000). The mixing proportions are the probabilities in a multinomial distribution consisting of one draw of K categories, and they satisfy $\sum_{k=1}^{K} p_k(\boldsymbol{\alpha}_k) = 1$. The mixing proportions are usually modeled by a multinomial logit model.

Estimation of parameters in a mixture regression model is similar to that for a mixture model. They include maximum likelihood, method of moments, etc.. The most popular method for maximum likelihood estimation of the parameter vector is the EM algorithm (Dempster et al., 1977, Richardson and Green, 1997).

This prototype mixture models can be extended to model correlated responses. Different model components can be connected through correlated random effects.

2.1.3.2 Mixture regression models for correlated continuous data

Thompson et al. (1998) proposed a two-component finite mixture regression model to assess diagnostic criteria for diabetes. A generalized linear model is used for each of the two components and the mixing proportion is modeled by a logistic regression model. Based on the model by Thompson et al. (1998), Yau et al. (2003) introduced random effects into both the mixing proportion and the component distributions to handle correlated length of stay from the same hospital. Specifically, this two-component finite mixture regression model allows to simultaneously model the heterogeneity and dependency among observations. Specifically, the response Y_{ij}, the length of stay for the jth individual at the ith hospital, was modeled by a two-component mixture regression model:

$$f(y_{ij}; \mathbf{x}_{ij}) = p(\mathbf{x}_{ij}) f_1(y_{ij}; \mathbf{x}_{ij}) + [1 - p(\mathbf{x}_{ij})]\, f_2(y_{ij}; \mathbf{x}_{ij})$$

where $p(\mathbf{x}_{ij})$ is the probability of the patient belonging to the first component, $\mathbf{x}_{ij}$ is a vector of risk variables associated with y_{ij}, including constant 1 for intercept, and $f_k(y; x)$ is the pdf of the kth component distribution ($k = 1, 2$). The proportion p is

assumed to be a logistic function of $\mathbf{x}$:

$$p(\mathbf{x}_{ij}) = \frac{\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\gamma} + V_i)}{1 + \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\gamma} + V_i)}, i = 1, ..., n, j = 1, ..., J_i$$

where γ is a vector of unknown logistic coefficients, and V_i is an unobserved random effect due to the ith hospital affecting the proportion p, and is taken to be i.i.d. $N(0, \sigma_V^2)$, for $i = 1, 2, ..., n$. It is further assumed that the kth component density follows a normal distribution:

$$f_k(y_{ij}; \mathbf{x}_{ij}) = \exp\left\{-\frac{1}{2}\left[\log(2\pi\sigma_k^2) + (y_{ij} - \eta_{ij,k})^2/\sigma_k^2\right]\right\}$$

$$\eta_{ij,k} = \mathbf{x}_{ij}^{'}\boldsymbol{\beta}_k + U_{ki}$$

where σ_k^2 is the variance of the kth component normal distribution, $\boldsymbol{\beta}_k$ is a vector of regression parameters and U_{ki} is the unobservable random effect corresponding to the kth component distributed as iid $N(0, \sigma_{ki}^2)$ (Yau et al., 2003), for $i = 1, 2, ..., n$ and $k = 1, 2$. Moreover, V_i and U_{ki} are assumed to be independent for $i = 1, 2, ..., n$ and $k = 1, 2$. Yau et al. (2003) applied this model to analyzing hospital length of stay data, and in this particular case, the random effect is due to hospital and follows a normal distribution. Statistical inference was based on the maximum likelihood method. EMMIX software was used to fit the normal mixture model and the number of normal components was determined based on AIC, BIC and the p-value for bootstrapping likelihood ratio test (McLachlan, 1987). All selection criteria lead to a two-component normal mixture regression model and this model fits the data well.

2.1.3.3 Mixture regression models for semincontinuous data

Semicontinuous data is a combination of a point mass at one or more locations (typically at value of 0) and a continuous distribution for the remaining values. A mixture model with a degenerate component can be used to analyze this type of data. Olsen and Schafer (2001) applied a generalized two-part model to the analysis of longitudinal

semicontinuous data with clumping at zero. In this particular example, the probability of zero response and the nonzero response can be separately analyzed by logistic regression and linear regression, respectively. An indicator variable U_{ij} is defined for whether the response is non-zero: $U_{ij} = 1$ if $Y_{ij} \neq 0$, $U_{ij} = 0$ if otherwise. V_{ij} is defined as follows:

$$V_{ij} = \begin{cases} g(Y_{ij}), & \text{if } Y_{ij} \neq 0 \\ \text{irrelevant}, & \text{if } Y_{ij} = 0 \end{cases}$$

where g is a monotonically increasing function that makes V_{ij} approximately Gaussian. The responses Y_{ij} are modeled by two correlated random effects models: one for the logit probability $U_{ij} = 1$ and the other for the mean conditional response $\mathrm{E}(V_{ij}|U_{ij} = 1)$. Specifically, the logit probability was modeled via

$$\boldsymbol{\eta}_i = \mathbf{X}_i\beta + \mathbf{Z}_i c_i$$

where $\boldsymbol{\eta}_i$ is the vector with elements $\eta_{ij} = \text{logit}\left\{P(U_{ij} = 1)\right\}, j = 1, ..., J_i$ and $\mathbf{X}_i(J_i \times q_c)$ and $\mathbf{Z}_i(J_i \times p_c)$ are matrices of covariates for fixed and random effects, respectively. The conditional mean response $E(V_{ij}|U_{ij} = 1)$ is modeled via

$$\mathbf{V}_i = \mathbf{X}_i^*\gamma + \mathbf{Z}_i^* d_i + \epsilon_i$$

where $\mathbf{V}_i$ is the vector of length n_i^* containing all relevant values of non-zero $Y_{ij}(U_{ij} = 1)$ for subject i. The residuals ϵ_i are assumed to follow $N(0, \sigma^2 I)$ and $X_i^*(n_i^* \times q_d)$ and $Z_i^*(n_i^* \times p_d)$ are matrices of covariates. The random effects from the two parts c_i and d_i follow a bivariate normal distribution with non-zero correlation:

$$b_i = \begin{pmatrix} c_i \\ d_i \end{pmatrix} \sim N\left(\mathbf{0}, \varphi = \begin{pmatrix} \varphi_{cc} & \varphi_{cd} \\ \varphi_{dc} & \varphi_{dd} \end{pmatrix}\right)$$

It was studied that erroneously ignoring the correlation between these two random effects could cause bias in parameter estimation (Su et al., 2009). The likelihood for the

model was derived and approximated by high-order multivariate Laplace expansion developed by Raudenbush et al. (2000). The likelihood was maximized by an approximate Fisher scoring procedure (Olsen and Schafer, 2001).

The proposed two-part model in Olsen and Schafer (2001) was used to analyze the effect of parental monitoring and rebelliousness on the reported alcohol use in high school and middle school students. In addition to estimating the parameters for parental monitoring and rebelliousness using the two-parts model, a likelihood ratio test was conducted to test whether the logit part and the linear regression parts of the model are separable (i.e., whether the correlation between the two random effects is zero). It is found that a student's probability of alcohol use at one occasion is positively correlated with the level of use at other occasions. Simulation studies also show that the two-part model provided unbiased estimates for both the fixed effect parameters and variance components (Olsen and Schafer, 2001).

Tooze et al. (2002) proposed a similar correlation mixed-distribution model with correlated random effects to analyze repeated measures continuous data with clumping at zero. The positive observations in the semicontinuous data are modeled by a log-normal distribution instead of a normal distribution. It is also assumed that the random effects introduced in the two parts of the model are correlated to each other in order to account for the intra-subject correlation. The full likelihood is approximated by an adaptive Gaussian quadrature and maximized using quasi-Newton optimization in SAS PROC NLMIXED. A SAS macro was developed to analyze the data by three steps: 1). As an initial step, run the binomial model for the occurrence component and a lognormal model for non-zero responses separately using PROC GENMOD ignoring the intra-subject correlation. 2). Using the estimated parameters in step 1 as the starting values, estimate the occurrence and intensity with independent random effects using SAS PROC NLMIXED; 3) Using the estimated parameters in step 2, analyze the data using correlated random effects in SAS PROC NLMIXED. Simulation results based on the correlation mixed-distribution model showed that the estimated parameters are also unbiased (Tooze et al., 2002).

2.2 Analysis of data with observations below detection limit

As mentioned in the previous chapter, many statistical methods can be used to handle data measured with a detection limit. Observations below the detection limit are often imputed by the value of the detection limit or one half of the detection limit. However, results from these methods may be subject to bias. Other commonly used methods include a two-part model analysis (Berk and Lachenbruch, 2002; Moulton et al., 2002; Taylor et al., 2001) and treating data below detection limit as left-censored data (Hughes, 1999; Jacqmin-Gadda et al., 2000; Lyles et al., 2000; Thiebaut and Jacqmin-Gadda, 2004; Li et al., 2006).

2.2.1 Analysis by a two-part model

Data with a detection limit can be separately into two components: observations above the detection limit and observations below the detection limit. Therefore, these type of data can be analyzed by a two part model with two components. Observations below detection limit can be either the left-censored values from the distribution of the observations above the detection limit or a real zero observation. Therefore, a two part model is more flexible for modeling data with more or less than expected number of observations below the detection limit. Moulton and Halsey (1995) used a Bernoulli/log-normal mixture model to deal with cross-sectional quantitative assay data with observations below detection limit. A generalized gamma model was proposed for concentration distribution in order to account for the more or less skewness than a log-normal distribution (Moulton and Halsey, 1996). As an example, a Bernoulli/Gamma mixture model was used to analyze antibody concentration data (Moulton and Halsey, 1996). Model parameters are estimated by maximizing the likelihood based on the Bernoulli and Gamma mixture model, and it was implemented by a SAS macro program.

Moulton et al. (2002) extended a basic Bernoulli/log-gamma mixture model described in Moulton and Halsey (1996) to repeatedly measured HIV viral load data. A shared parameter is introduced to increase model efficiency and parsimony. There are two components in a basic Bernoulli/log-gamma mixture model: one is a point

distribution below the detection limit and the other is a log-gamma distribution for observations above the detection limit. Let p_i be the probability an observation issues from the log-gamma distribution and it is modeled by a logistic regression model:

$$\text{logit}(p_i) = \mathbf{x}_i'\boldsymbol{\beta}$$

where $\mathbf{x}_i$ is the covariates vector of the ith subject and $\boldsymbol{\beta}$ is the coefficients vector. The model for logarithm of the observations above the detection limit (y_i) is:

$$y_i = \mathbf{x}_i'\boldsymbol{\gamma} + \delta\epsilon_i$$

where ϵ_i is an i.i.d. error term with probability density, for $i = 1, 2, , n$:

$$f(\omega) = \begin{cases} |\delta|[\exp(\delta\omega)/\delta^2]^{1/\delta^2}\exp[-\exp(\delta\omega)/\delta^2]/\Gamma(1/\delta^2), & \delta \neq 0 \\ \exp(-\omega^2/2)/\sqrt{2\pi}, & \delta = 0 \end{cases}$$

A shared parameter θ is introduced between the covariates for the two components:$\gamma_i = \theta\beta_i, i = 1, 2$. This assumption is based on the expectation that covariates increasing the risk of ones HIV viral load to some degree might influence the attainted level of viral load to the same degree. A simulation study conducted based on the proposed shared parameter model shows that the introduction of the shared parameter improved the power. In order to account for the intra-subject correlation between different measurements, Moulton et al. (2002) proposed a population-averaged approach which consists of two steps: first the regression relationship was estimated assuming no correlation (independence working model), then adjusted for the intra-subject correlation in the variance-covariance matrix by the sandwich estimator (Moulton et al., 2002).

Taylor et al. (2001) developed a more complicated mixture model handling data with a detection limit. Specifically, the data under the detection limit contained true zeros. They used a similar idea to that in Moulton and Halsey (1995, 1996) and used the following models treating non-detectable values as a combination of true zeros and

left-censored outcomes. Assume the probability density function of the response Y is:

$$f_Y(y; w, \mu_y, \sigma_y^2) = wI(y;0) + (1-w)\left[1 - I(y;0)\right] g(y; \mu_y, \sigma_y^2)$$

where $w = \Pr(Y = 0)$, $I(y;0)$ is an indicator function for the event $y = 0$, and

$$g(y; \mu_y, \sigma_y^2) = \frac{1}{\sqrt{2\pi} y \sigma_y} \exp\left\{ -\frac{(\log y - \mu_y)^2}{2\sigma_y^2} \right\}, 0 < y < \infty$$

For this type of data with a limit of detection, a zero-inflated distribution is not sufficient as it fails to account for non-zero censored values. A simple censored model is also not valid as it fails to account for true zero responses.

2.2.2 Left-censored likelihood method

A more straightforward method for analyzing data with observations below detection limit is to treat the observations under a detection limit as left-censored data and use the likelihood based approach for statistical inference (Hughes, 1999; Jacqmin-Gadda et al., 2000; Lyles et al., 2000).

Li et al. (2006) used the principle of this likelihood approach and proposed a two-component mixture regression model to analyze bimodal HIV data with observations below a detection limit. In their study, the virologic responses of HIV patients treated by highly active antiretroviral therapy (HAART) were found to be heterogeneous: the higher mode corresponds to suboptimal virologic responses and the lower mode corresponds to optimal virologic responses. The study was under a longitudinal design with repeated measures, but the data was analyzed at each time point separately. A likelihood ratio test was applied to evaluate whether the HIV data was better described by a mixture or a single component normal distribution at each time point. Since the mixing proportion $p = 0$ falls on the boundary of the p, the standard regularity condition does not hold, and a bootstrapping resampling method was used to perform the likelihood ration test (Li et al., 2006).

2.3 Model-checking for Finite Mixture Models

Model-checking is important for statistical modelling. Usually the adequacy of a regression model is subjectively evaluated by residual plots. Many objective and quantitative goodness-of-fit techniques are also used for checking model assumptions. Most commonly used methods include likelihood ratio test, Akaike information criterion (AIC), and Bayesian information criterion (BIC), etc.. A common feature of these methods is to compare different models in fitting the data and choose the better model. However, these methods can only provide a comparison of goodness-of-fit between models, and they cannot be used to evaluate how well a specific model describes the data. Recently, Su and Wei (1991), Lin et al. (2002), and Pan and Lin (2005) introduced a modern approach and developed a Kolmogorov-type supremum goodness-of-fit test based on cumulative residuals to evaluate the overall fit of the model and the functional form of a covariate. Specifically, this method has been applied to generalized linear models (Su and Wei, 1991), Cox survival models (Lin et al., 1993, Spiekerman and Lin, 1996), marginal models (Lin et al., 2002) and generalized mixed linear models (Pan and Lin, 2005). In this dissertation, we adopted the principle of this approach and proposed a goodness of fit test for mixture regression models with and without random effects. Therefore, we gave a detailed review of this approach in section 2.3.3.

Checking the GOF for mixture models is more complicated than the "regular (non-mixture)" models because the component membership of each observation is unobserved, and thus it is difficult to apply the regular residual plots for model diagnostics. Lindsay and Roeder (1992) proposed diagnostic plots to detect the presence of mixing. Ngom (2005) evaluated the model fit for mixture models using divergence statistics. Wang et al. (2005) developed an exploratory graphical tool to evaluate the model adequacy of growth mixture models using a pseudoclass technique. Their proposed diagnostic plots can be used to detect model misspecification in the number of components, growth trajectory means and covariance structures. A detailed review of these methods are given in the following subsections.

2.3.1 Model checking using divergence statistics

Ngom (2005) studied the problem of goodness-of-fit in mixture models using power-divergence statistics. A power divergence is defined as follows:

$$\Delta_r\left[f(x,\theta_1), f(x,\theta_2)\right] = \frac{1}{r-1}\left[\sum_{l=1}^{M}\left\{\frac{f_l(x,\theta_1)^r}{f_l(x,\theta_2)^{r-1}} + \frac{f_l(x,\theta_2)^r}{f_l(x,\theta_1)^{r-1}}\right\} - 2\right],$$

where $f(x,\theta_1)$ and $f(x,\theta_2)$ are two probability density functions and $0 < r < +\infty$. This power-divergence was used to discriminate among alternative mixture regression models. The asymptotic distributions of the divergence statistics are studied by considering the parametric divergence measure:

$$\Delta_r\left[f, h(\theta)\right] = \frac{1}{r-1}\left[\sum_{l=1}^{M}\left\{\frac{f_l^r}{h_l^{r-1}(\theta)} + \frac{h_l^r(\theta)}{f_l^{r-1}}\right\} - 2\right].$$

A statistic divergence $\hat{\Delta}_r\left[f, h(\theta)\right] = \Delta_r\left[f, h(\hat{\theta}_n)\right]$ is defined by its estimator and the asymptotic distribution for this statistic is obtained. Under the null hypothesis, this test statistic is a chi-square type statistic, while under alternative hypothesis, it follows an asymptotically normal distribution. An asymptotic normal test was proposed based on the power-divergence that uses estimators from the EM algorithm. This test was used to determine whether the estimated completing mixture model is close to the true distribution against the alternative hypothesis.

2.3.2 Diagnostic plots for mixture models

Lindsay and Roeder (1992) proposed a method to assess the adequacy of a fitted mixture model by examining "residuals" based on the ratio of the observed to the expected fit. The main rationale of the diagnostic plots is that residuals based on the one-component model display a convexity property when the data actually follow a mixture model. For discrete mixture models, define the residual function by

$$r(y;Q) = \frac{\hat{p}(y)}{g(y;Q)} - 1,$$

where $\hat{p}(y) = n^{-1} \sum_{i=1}^{n} I(X_i = y)$ and Q is the unknown distribution of parameter θ. When Q is a point mass at θ, the residual function $r(y; \theta)$ measures departures from the homogeneity model in which Q is degenerate at θ, and it is called homogeneity residuals. For $r(y; \hat{Q})$, the residuals measure departures from the best-fitted mixture model, and is called mixture residuals. Then the presence of mixing can be detected by plotting the homogeneity residuals evaluated at the non-mixed maximum likelihood estimate. For discrete exponential family models, the residual plot can be used as a diagnostic tool to detect the presence of a mixture. A convex shape of the plot indicates the heterogeneity. Mixture diagnostic plots can be constructed similarly for continuous data by binning the data.

As a special case of finite mixture model, growth mixture modelling has become more widely used to study the heterogeneity of developmental or disease progression trajectories within a population. Wang et al. (2005) developed a graphical diagnostic tool for checking the number of mixture components, the mean response models and covariance structures. A pseudoclass technique, originally introduced by Bandeen-Roche et al. (1997), was used to generate pseudoclass adjusted residuals for checking model misspecifications. This technique included three steps: 1). Randomly assign each subject into a class based on the posterior distribution of class membership, and repeat the random draw for multiple times independently; 2). compute the pseudoclass adjusted residuals; 3) examine residuals using regular diagnostic tools. It has been justified that the empirical distributions of these pseudoclass adjusted residuals are asymptotically equivalent to the distributions of residuals from the actual latent classes (Wang et al., 2005).

Three different diagnostic plots were proposed to check the model adequacy using the distribution of pseudoclass adjusted residuals, and they were designed to assess different types of model misspecifications. A time trend plot showing the residual means as a function of time was created for each latent class to examine the mean growth structure in each component; quantile-quantile (Q-Q) norm plot for examining the adequacy of the number of mixture components; an empirical Q-Q plot for examining covariance structures. The advantage of these techniques is that they can compare models that

are not completed nested. However, these methods are only for exploration, and cannot be used as a confirmatory test.

2.3.3 Goodness-of-fit tests based on cumulative residuals

One of the problems with residual plots is that the variabilities of individual residuals are unknown. Therefore, Su and Wei (1991) developed a GOF test by taking the cumulative sums of residuals over certain covariates or fitted values. The main motivation for the use of cumulative residuals is that the cumulative residuals are centred at 0 if the assumed model is correct and the natural variation of cumulative residuals can be determined. Specifically, under the null hypothesis of the correct model specification, the distribution of the cumulative residuals can be approximated by a zero-mean Gaussian process whose realization can be generated by computer simulations. Each process can be compared, both graphically and numerically, with a number of realizations from the Gaussian process, then a p-value for testing model misspecification can be obtained from this comparison. Lin et al. (2002) extended this approach to both generalized linear models and marginal models with dependent observations. Pan and Lin (2005) further applied this method to generalized linear mixed models, and proposed supremum test statistics for checking the functional form of a covariate or the link function of the response.

Assume that the response Y follows an exponential family distribution with density function

$$f_Y(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\psi) + c(y,\psi)\},$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are some specific functions, θ is the parameter of interest, and ψ is a nuisance parameter. The mean function of Y, $\mu(\theta)$ is associated with covariate vector $\mathbf{X} = (1, X_1, ..., X_p)'$:

$$g\{\mu(\theta)\} = \boldsymbol{\beta}'\mathbf{X},$$

where g is a link function, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is a $(p+1) \times 1$ vector of unknown regression coefficients. Suppose that the data consist of n independent replicates of

$(Y, \mathbf{X})$. The likelihood score function $\mathbf{U}(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ is defined as:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} h(\boldsymbol{\beta}'\mathbf{X}_i)\mathbf{X}_i \left\{ Y_i - \nu(\boldsymbol{\beta}'\mathbf{X}_i) \right\},$$

where $\nu = g^{-1}(r)$, and $h(r) = \partial\{g \cdot \mu(r)\}^{-1}/\partial r$.

Define the residual for an exponential family distribution by $e_i = Y_i - \nu(\hat{\boldsymbol{\beta}}'\mathbf{X}_i), i = 1, ..., n$. The cumulative sum of the e_i over the jth component of $\mathbf{X}$, X_{ji} is

$$W_j(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I(X_{ji} \leq x) e_i,$$

and it is a stochastic process with possible jumps at distinct values of X_{ji}. Under the null hypothesis that the proposed model is correctly specified, the distribution of $W_j(x)$ has the same asymptotic distribution as $\hat{W}_j(x)$ (Su and Wei, 1991), which is defined as

$$\hat{W}_j(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \{ I(X_{ji} \leq x) + \eta'(x; \hat{\boldsymbol{\beta}}) I(\hat{\boldsymbol{\beta}})^{-1} \mathbf{X}_i h(\hat{\boldsymbol{\beta}}'\mathbf{X}_i) \} e_i G_i,$$

where

$$\eta(x; \boldsymbol{\beta}) = -\sum_{i=1}^{n} I(X_{ji} \leq x) \partial\nu(\boldsymbol{\beta}'\mathbf{X}_i)/\partial\boldsymbol{\beta},$$

and $(G_1, ..., G_n)$ are independent standard normal variables that are independent of $(Y_i, \mathbf{X}_i)$ (Lin et al., 2002). Because the null distribution of $W_j(x)$ can be approximated through simulating the corresponding zero-mean Gaussian process $\hat{W}_j(x)$, one may plot $W_j(\cdot)$ along with a few realizations from the $\hat{W}_j(\cdot)$ process to assess whether a trend seen in a residual plot reflects model misspecification or natural variation. Furthermore, one may objectively assess the goodness-of-fit for a model using a superum test. Define the Kolmogorov-type supremum statistic $S_j \equiv \sup_x |W_j(x)|$. An unusually large observed value of the test statistic would suggest incorrect function form of X_j. The P-value, $Pr(S_j \geq s_j)$ can be approximated by $Pr(\hat{S}_j \geq s_j)$ where $\hat{S}_j = \sup_x |\hat{W}_j(x)|$.

The previously proposed test statistic tends to be dominated by the residuals associated with small covariate values, therefore, another definition of the cumulative

residuals is proposed to address this issue (Lin et al., 2002):

$$W_j(x;b) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I(x-b < X_{ji} \leq x) e_i.$$

Similarly, it can be shown that the null distribution of $W_j(x;b)$ can be approximated by the conditional distribution of $\hat{W}_j(x;b)$ given data $(Y_i, \mathbf{X}_i)$ where

$$\hat{W}_j(x;b) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \{I(x-b < X_{ji} \leq x) + \eta'(x;b,\hat{\boldsymbol{\beta}}) I(\hat{\boldsymbol{\beta}})^{-1} \mathbf{X}_i h(\hat{\boldsymbol{\beta}}' \mathbf{X}_i)\} e_i G_i$$

where

$$\eta(x;b,\boldsymbol{\beta}) = -\sum_{i=1}^{n} I(x-b < X_{ji} \leq x) \partial \nu(\boldsymbol{\beta}' \mathbf{X}_i)/\partial \boldsymbol{\beta}.$$

The same supremum test can be used to assess the goodness-of-fit of the model. The advantage of this alternative test is that it is less dominated by the residuals associated with small covariate values. The test results may depend on the choice of b. Large values of b are preferred in detecting global misspecification, while small values of b are more sensitive to local deviations.

The same rationale was applied to a marginal model for dependent observations, and a similar supremum test statistic is proposed to assess whether a marginal model is appropriate for repeated measures data (Lin et al., 2002).

Pan and Lin (2005) extended this cumulative residual method to evaluate the goodness-of-fit test for generalized linear mixed models (GLMMs). Due to the existence of the random effects in a GLMM, more challenges need to be overcome to apply the goodness-of-fit test strategy.

A GLMM is defined as follows for repeated measures data $(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$:

$$g\{E(y_{ij}|\mathbf{b}_i)\} = \mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\mathbf{b}_i, i = 1, ..., n; j = 1, ..., J_i$$

where y_{ij} is the response for the ith subject on the jth occasion, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are the corresponding vectors of covariates associated with the fixed effects and random effects, respectively, $g(\cdot)$ is a known differentiable link function, $\boldsymbol{\beta}$ is $p \times 1$ vector of unknown

regression parameters, and $\mathbf{b}_i$ is a $q \times 1$ vector of unobservable random effects for the ith subject.

Three important assumptions were made for the defined GLMM: 1) the conditional distribution of y_{ij} given $\mathbf{b}_i$ follows an exponential family distribution with density function $f_{y|b}(y|\mathbf{b})$; 2) the repeated measures for the same subject are conditionally independent given $\mathbf{b}_i$; 3) the $\mathbf{b}_i$ are i.i.d. with density function $f_b(\mathbf{b})$.

Based on the defined GLMM, the marginal mean of the response y_{ij} is given by

$$E(y_{ij}) = m_{ij}(\boldsymbol{\theta}) = \int g^{-1}(\mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\mathbf{b}_i) f_b(\mathbf{b}_i) d\mathbf{b}_i.$$

Then the residual is defined as $e_{ij} = y_{ij} - \hat{m}_{ij} (i = 1, ..., n; j = 1, ..., J_i)$ where $\hat{m}_{ij} \equiv m_{ij}(\hat{\boldsymbol{\theta}})$. Two cumulative residuals are defined for assessing the model properties in terms of the functional form covariates and the link function of the response, respectively:

$$W(\mathbf{x}) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{J_i} I(\mathbf{X}_{ij} \leq \mathbf{x}) e_{ij}, \text{and}$$

$$W_g(r) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \sum_{j=1}^{J_i} I(\hat{m}_{ij} \leq r) e_{ij},$$

where $\mathbf{x} = (x_1, ..., x_p)' \in R^p, r \in R, I(\mathbf{X}_{ij} \leq \mathbf{x}) = I(X_{1ij} \leq x_1, ..., X_{pij} \leq x_p)$, and X_{kij} is the kth component of $\mathbf{X}_{ij}, (k = 1, ..., p)$. The null distribution of $W(\mathbf{x})$ and $W_g(r)$ can be approximated by $\hat{W}(\mathbf{x})$ and $\hat{W}_g(r)$, respectively, where.

$$\hat{W}(\mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J_i} I(\mathbf{X}_{ij} \leq \mathbf{x}) e_{ij} + \eta'(\mathbf{x}; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} \mathbf{U}_i(\hat{\boldsymbol{\theta}}) \right\} G_i,$$

$(G_1, ..., G_n)$ are independent standard normal variables independent of data, $\mathbf{U}_i(\boldsymbol{\theta}) = \frac{\partial log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, and

$$\eta(\mathbf{x}; \boldsymbol{\theta}) = -n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J_i} I(\mathbf{X}_{ij} \leq \mathbf{x}) \partial m_{ij}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta},$$

$$\hat{W}_g(r) = n^{-1/2} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J_i} I(\hat{m}_{ij} \leq r) e_{ij} + \eta'_g(r; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} \mathbf{U}_i(\hat{\boldsymbol{\theta}}) \right\} G_i,$$

where

$$\eta_g(r;\boldsymbol{\theta}) = -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i} I(\hat{m}_{ij} \leq r)\partial m_{ij}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}.$$

Simulation results show that W is the most sensitive to the misspecification of the functional form of covariates $\mathbf{X}$, and W_g to the misspecification of the link function of the response (Pan and Lin, 2005). However, they did not check the model assumption for the random effects and assumed that the random components, including the error distribution and the random effects, are correctly specified.

These GOF tests based on cumulative residuals have been widely accepted and incorporated into some standard statistical software including SAS and R (SAS, 2011). In SAS, this test method has been included in the GENMOD procedure for the analysis of generalized linear model, and in the PHREG procedure for survival analysis. An R package has been published to check the goodness-of-fit for generalized linear models (Holst, 2011). However, this cumulative residual-based model checking technique has not been applied to the analysis of generalized linear mixed models in SAS or R.

Chapter 3

Finite Mixture Regression Model for Repeated Measures with a Detection Limit

Recall that in the HEART study, the distribution of EBC 8-isoprostane data exhibited the issues of nondetectables, multi-modal distributions, and repeated measures. In this chapter, we propose a random effect finite mixture regression model, accommodating measurements under a detection limit. As shown in the previous chapter, Li et al. (2006) analyzed data with a detection limit using a mixture model of two left-censored normal components. However, the nature of the data is repeated measures, while the analysis was carried out at each specific time point separately. Yau et al. (2003) applied a mixture regression model to clustered data and added correlated random effects in the statistical model to account for the intra-cluster correlation. However, their data were not measures subject to left-censoring, e.g. detection limit. These applications can be considered special cases of our method. Moreover, it has been shown that, under the two-part model setting, bias would occur if the correlation between the random effects in different model components, e.g., logistic and linear regression models, is erroneously ignored (Su et al., 2009). Therefore, we aimed to look into how the correlation between random effects in a mixture regression model under non-degenerated setting would influence the data analysis through our proposed models. In addition, the mixture regression model in Yau et al. (2003) is implemented by the EMMIX software, which is not commonly used. In this chapter, we demonstrated the use of NLMIXED procedure in SAS, a commonly used statistical software package, to implement the proposed random effects finite mixture regression models accounting for left-censoring to handle data below detection limit. The computing code is attached in the appendix. Lastly, the EBC 8-isoprostane data from the HEART study is used to

illustrated the proposed methodology.

3.1 Proposed Method

In this chapter, we extended the model of Li et al. (2006) to a finite mixture regression model with K components and treated observations below a detection limit as left-censored data. In lieu of analyzing longitudinal data at each time point separately, we added subject-specific random effects to account for the intra-subject correlation. The proposed model will be used to analyze left-censored multimodal repeated measures data, like the 8-isoprostane data from the HEART study, and the parameters in the mixture model will be estimated using the maximum likelihood method. Our proposed model is an extension of the random effect two-component mixture regression model of Yau et al. (2003) and we allow the random effects in the mixing proportions and component distributions to be correlated. Because two-component mixture models involve only one mixing proportion, three-component mixture models require two mixing proportions and these two mixing proportions are correlated with each other, it can be more challenging to specify and fit a three-component mixture model. In addition to the normal distributions, we also discussed some other component distributions such as gamma and log-normal distributions and studied some related statistical properties of the models. We began this chapter by introducing a general framework for a K-component random effect mixture regression model assuming a general type of component distribution. Then we discussed the model properties using normal mixture regression model as an example. For normal mixture regression models, two special cases were examined: two-component and three-component random effects normal mixture regression models.

3.1.1 Finite mixture regression model

Let Y_{ij} denote the jth continuous response for subject i (e.g., the value of EBC 8-isoprostane at the j-th visit for subject i in the HEART study), and assume that Y_{ij}

follows a finite mixture distribution:

$$f(y_{ij};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k f_k(y_{ij};\boldsymbol{\theta}_k),$$

where p_k is the mixing proportion of the kth component with $\sum_{k=1}^{K} p_k = 1$, and $f_k(\cdot;\boldsymbol{\theta}_k)$ denotes the pdf of the kth component distribution with parameter (vector) $\boldsymbol{\theta}_k$, for $k = 1, 2, ..., K$, and $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_K, p_1, ..., p_K)$. Let $\mathbf{x}_{ij}$ be a vector of risk variables associated with y_{ij}, including constant 1 for intercept. These parameters $\boldsymbol{\theta}_k$ and the mixing proportion p_k are related to risk variables $\mathbf{x}_{ij}$ through regression models. To account for the intra-subject correlation between repeated measures, we added the random effects $v = (v_1, ..., v_K)'$ and $u = (u_1, ..., u_{K-1})'$ into the regression models. Specifically, $\mathbf{v}$ is added to the regression models related to $\boldsymbol{\theta}_k$ and $\mathbf{u}$ is added to the regression models related to p_k. Throughout the following discussion, we assumed that the random vectors $\mathbf{u}$ and $\mathbf{v}$ follow a multivariate normal distribution:

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{u}}, & \Sigma_{\mathbf{uv}} \\ \Sigma_{\mathbf{uv}}, & \Sigma_{\mathbf{v}} \end{pmatrix} \right)$$

where $\Sigma_{\mathbf{u}}$ and $\Sigma_{\mathbf{v}}$ are the variance-covariance matrix for random vectors $\mathbf{u}$ and $\mathbf{v}$, respectively, and $\Sigma_{\mathbf{uv}}$ is the covariance matrix for $\mathbf{u}$ and $\mathbf{v}$.

3.1.1.1 Exponential family mixture models

In a mixture of exponential family distributions, each component follows an exponential familty distribution with pdf

$$f_k(y_{ij};\vartheta^{(k)}) = h(y_{ij})\exp\left\{\vartheta^{(k)} g_k(y_{ij}) - A(\vartheta^{(k)})\right\}$$

where $\vartheta^{(k)}$ is the parameter vector for the kth component, $g_k(y_{ij})$ is a function of y_{ij}, and also a sufficient statistic for $\theta^{(k)}$, and $A(\vartheta^{(k)})$ is a known function of $\vartheta^{(k)}$.

Based on the statistical properties of an exponential family distribution, the mean

of $g_k(y_{ij})$ for the kth component is

$$E_{\vartheta^{(k)}}\left[g_k(y_{ij})\right] = \frac{\partial A(\vartheta^{(k)})}{\partial \vartheta^{(k)}}.$$

Then we allow the $E_{\vartheta^{(k)}}\left[g_k(y_{ij})\right]$ to depend on covariate information x_{ij} through a regression model:

$$E_{\vartheta^{(k)}}\left[g_k(y_{ij})\right] = \mathbf{x}'_{ij}\boldsymbol{\beta}^{(k)} + v_{ki}$$

where $\boldsymbol{\beta}^{(k)}$ is a vector of regression coefficients for the kth component, and v_{ki} is the random effect for the ith subject in kth component. The mixing proportions p_k is associated with risk covariate $\mathbf{x}_{ij}$ and random effect u_{ki} through a multi-logistic regression model:

$$p_k(\mathbf{x}_{ij}, u_{ki}) = \begin{cases} \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(k)}+u_{ki})}{1+\sum_{k^*=1}^{K}\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(k^*)}+u_{k^*i})}, \text{if } k = 1, ..., K-1 \\ \frac{1}{1+\sum_{k^*=1}^{K}\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(k^*)}+u_{k^*i})}, \text{if } k = K \end{cases}$$

where $\boldsymbol{\alpha}^{(k)}$ is a vector of coefficients corresponding to $\mathbf{x}_{ij}$, and u_{ki} is the random effect for the kth mixing proportion for subject i. In addition, the random vectors $\mathbf{u}_i = (u_{1i}, ..., u_{K-1,i})'$ and $\mathbf{v}_i = (v_{1i}, ..., v_{Ki})'$ are assumed to be i.i.d. copies from the multivariate distribution as described previously.

3.1.1.2 Normal component mixture models

In normal component mixture models, we assumed that each component follows a normal distribution with pdf:

$$f_k(y_{ij}; \mu_k, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_{ij}-\mu_k)^2}{2\sigma^2}\right\},$$

and note that σ^2 is assumed to be the same across all K components for identifiability. Then μ_k is further associated with covariates vector $\mathbf{x}_{ij}$ (including a constant 1 for intercept) and random effects v_{ki} in the following model:

$$\mu_k(\mathbf{x}_{ij}, v_{ki}) = \mathbf{x}'_{ij}\boldsymbol{\beta}_k + v_{ki},$$

where $\boldsymbol{\beta}_k$ is a vector of regression coefficients for the kth component, and v_{ki} is the random effect for the ith subject in kth component. The mixing proportions p_k is associated with risk covariate $\mathbf{x}_{ij}$ and random effect u_{ki} through a multi-logistic regression model:

$$p_k(\mathbf{x}_{ij}, u_{ki}) = \begin{cases} \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}_k + u_{ki})}{1+\sum_{k^*=1}^{K}\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}_{k^*} + u_{k^*i})}, \text{if } k = 1, ..., K-1 \\ \frac{1}{1+\sum_{k^*=1}^{K}\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}_{k^*} + u_{k^*i})}, \text{if } k = K \end{cases}$$

where $\boldsymbol{\alpha}_k$ is a vector of coefficients corresponding to $\mathbf{x}_{ij}$, and u_{ki} is the random effect for the kth mixing proportion for subject i. In addition, the random vectors $\mathbf{u}_i = (u_{1i}, ..., u_{K-1,i})'$ and $\mathbf{v}_i = (v_{1i}, ..., v_{Ki})'$ are assumed to be i.i.d. copies from the multivariate distribution. Note that the vector of risk covariates $\mathbf{x}_{ij}$ can be different from those in the mixing proportions and component specific means. For simplicity, we assume that both mixing proportions and component means are associated with the same set of covariates ($\mathbf{x}_{ij}$).

Commonly used mixture regression models are discussed as follows.

1). Two-component normal mixture regression model

Let $\mathbf{u}_i = u_{1i}$ and $\mathbf{v}_i = (v_{1i}, v_{2i})'$. Following the previous model development, a two-component normal mixture model ($K = 2$) for response Y_{ij}, given the random effects $\mathbf{u}_i$ and $\mathbf{v}_i$, can be expressed as:

$$f(y_{ij}; \mathbf{x}_{ij}, \mathbf{u}_i, \mathbf{v}_i) = p_1(\mathbf{x}_{ij}, u_{1i}) f_1(y_{ij}; \boldsymbol{\theta}_1(\mathbf{x}_{ij}, v_{1i})) + \{1 - p_1(\mathbf{x}_{ij}, u_{1i})\} f_2(y_{ij}; \boldsymbol{\theta}_2(\mathbf{x}_{ij}, v_{2i})),$$

where $f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_{ki}))$ is a normal density:

$$f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_{ki})) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[y_{ij} - \mu_k(\mathbf{x}_{ij}, v_{ki})]^2}{2\sigma^2} \right\},$$

and

$$\mu_k(\mathbf{x}_{ij}, v_{ki}) = \mathbf{x}'_{ij}\boldsymbol{\beta}^{(k)} + v_{ki}, \tag{3.1}$$

$\boldsymbol{\beta}^{(k)}$ is a vector of regression coefficients for the kth component ($k = 1, 2$). Under

constraint $\sum_{k=1}^{2} p_k = 1$, we only need to model the mixing proportion for the first component p_1 via a logistic regression model, given the random effect $u_{1i} = u_i$:

$$\text{logit}\left\{p_1(\mathbf{x}_{ij}, u_i)\right\} = \mathbf{x}'_{ij}\boldsymbol{\alpha} + u_i, \tag{3.2}$$

where $\boldsymbol{\alpha}$ is a vector of coefficients corresponding to $\mathbf{x}_{ij}$, u_i is the random effect corresponding to first component proportion for subject i. In addition, we assume that the random variables u_i and $\mathbf{v}_i$ follow a trivariate normal distribution with mean $\mathbf{0}$.

A special case of random effects two-component normal mixture regression models is considered by assuming $u_i = 0, v_{1i} = v_{2i} = v_i$, then

$$\mu_k(\mathbf{x}_{ij}, v_i) = \mathbf{x}'_{ij}\boldsymbol{\beta}^{(k)} + v_i, \tag{3.3}$$

$$\text{logit}\left\{p_1(\mathbf{x}_{ij})\right\} = \mathbf{x}'_{ij}\boldsymbol{\alpha}, \tag{3.4}$$

The random effect v is assumed to follow a normal distribution with mean zero and variance σ_v^2. Under this model specification, the mean, variance and correlation of Y_{ij} and Y_{ij*} can be expressed as:

$$\begin{aligned}
E(Y_{ij}) &= p_1\mathbf{x}'_{ij}\boldsymbol{\beta}^{(1)} + (1-p_1)\mathbf{x}'_{ij}\boldsymbol{\beta}^{(2)} \\
Var(Y_{ij}) &= \sigma_v^2 + \{p_1^2 + (1-p_1)^2\}\sigma^2 \\
CORR(Y_{ij}, Y_{ij^*}) &= \frac{\sigma_v^2}{\sqrt{\{\sigma_v^2 + \left[p_1^2 + (1-p_1)^2\right]\sigma^2\}\{\sigma_v^2 + \left[(p_1^*)^2 + (1-p_1^*)^2\right]\sigma^2\}}}
\end{aligned}$$

Details of the model properties are included in Appendix A.

2). Three-component normal mixture regression model

Assume the response Y_{ij} follows a three-component normal mixture model:

$$f(y_{ij}; \mathbf{x}_{ij}) = \sum_{k=1}^{3} p_k(\mathbf{x}_{ij}, u_{ki}) f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_{ki})),$$

$$\sum_{k=1}^{3} p_k(\mathbf{x}_{ij}, u_{ki}) = 1.$$

The proportions $p_k(\mathbf{x}_{ij}, u_{ki})$ can be modeled via a multilogistic regression model defined in the previous section. To ease the computational difficulties caused by too many random variables, some assumptions regarding random effects u_{1i} and u_{2i} can be used. For instance:

a). $u_{1i} = u_i, u_{2i} = 0$;

b). $u_{1i} = u_{2i} = u_i$;

c). $u_{1i} = d \times u_{2i} = u_i$ where d is a constant;

All of the above three scenarios are a special case of multivariate normal distributions with scenario c being the most flexible. Scenario a implies a positive correlation between $p_2(\mathbf{x}_{ij}, u_{2i})$ and $p_3(\mathbf{x}_{ij}, u_{3i})$, and scenario b implies a positive correlation between $p_1(\mathbf{x}_{ij}, u_{1i})$ and $p_2(\mathbf{x}_{ij}, u_{2i})$. Similarly, a common unobserved measurement error v_i is assumed for subject i in each component: $v_{1i} = v_{2i} = v_{3i} = v_i$.

3.1.1.3 Log-normal mixture model

Normal distribution has a range from minus infinity to infinity, thus it may not be appropriate for describing non-negative data, such as HIV RNA data, and the 8-isoprostane data in the motivating example. We explored other distributions with only positive support, such as log-normal and gamma distribution.

A log-normal mixture regression model is similar to a normal mixture model except that the component distribution is a log-normal distribution instead of a normal distribution. The probability density function of each component distribution is:

$$f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_{ki})) = \frac{1}{y_{ij}\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\log y_{ij} - \mu_k)^2}{2\sigma^2}\right\}$$

where μ_k is the parameter for the kth component, and σ^2 is the common scale parameter

for all components. The mean of the log-normal distribution is

$$\exp\left(\mu_k + \frac{\sigma^2}{2}\right)$$

According to the principle on p.26, we can model the association between μ_k and covariates $\mathbf{x}_{ij}$, random vector $\mathbf{v}_k$ via the following model:

$$\mu_k(\mathbf{x}_{ij}, v_{ki}) = \mathbf{x}'_{ij}\boldsymbol{\beta}^{(k)} + v_{ki}$$

The modeling of mixing proportions p_k can be specified by the same rationale described previously.

3.1.1.4 Gamma mixture model

The probability density function for each gamma component distribution is:

$$f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_{ki})) = \frac{y_{ij}^{a_k-1}\exp(-y_{ij}/b)}{\Gamma(a_k)b^{a_k}}$$

where b is the common scale parameter for all components, a_k is the shape parameter for the kth components, and Γ is a gamma function. The mean of the gamma distribution is $a_k b$, and it is assumed to be associated with covariates $\mathbf{x}_{ij}$ and random effect v_{ki} via a linear regression model. The modeling of mixing proportions p_k can be specified by the same rationale described previously.

3.1.2 Estimation of mixture model parameters accounting for detection limit

There are usually two different methods to account for the data below detection limit. One is to use a two-part model to model the data below and above the detection limit separately; and the other is to treat observations below the detection limit as left-censored. To ease the complexity of the mixture regression model structure, we use the left-censoring approach to handle data below detection limit. Let $\delta_{ij} = 1$ if $Y_{ij} \geq D_0$, $\delta_{ij} = 0$ if Y_{ij} is below the detection limit ($Y_{ij} < D_0$). Conditional on the random effects

u_i and v_i, the conditional density of Y_{ij} can be written as:

$$f(Y_{ij}|u_i, v_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k(u_i) f_k(y_{ij}|v_i, \boldsymbol{\theta}_k)$$

The conditional cumulative density function of Y_{ij} given u_i and v_i is:

$$F(y_{ij}|u_i, v_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k(u_i) F_k(y_{ij}|v_i, \boldsymbol{\theta}_k)$$

To integrate out the random effects u_i and v_i, the contribution to the likelihood for the ith subject is:

$$\begin{aligned}
& L_i(\boldsymbol{\theta}|y_{i1}, ..., y_{iJ_i}) \\
= & \int_{u_i}\int_{v_i} \prod_{j=1}^{J_i} \{f(y_{ij}; u_i, v_i)\}^{\delta_{ij}} \{F(D_0; u_i, v_i)\}^{1-\delta_{ij}} du_i dv_i \\
= & \int_{u_i}\int_{v_i} \prod_{j=1}^{J_i} \left\{\sum_{k=1}^{K} p_k(u_i) f_k(y_{ij}|v_i, \boldsymbol{\theta})\right\}^{\delta_{ij}} \left\{\sum_{k=1}^{K} p_k(u_i) F_k(D_0|v_i, \boldsymbol{\theta})\right\}^{1-\delta_{ij}} f_{u,v}(u_i, v_i) du_i dv_i
\end{aligned}$$

where J_i is the total number of observations for subject i. The marginal likelihood is then:

$$L(\boldsymbol{\theta}|Y_{ij}) = \prod_{i=1}^{n} \left\{ \int_{u_i}\int_{v_i} \prod_{j=1}^{J_i} \{f(y_{ij}; u_i, v_i)\}^{\delta_{ij}} \{F(D_0; u_i, v_i)\}^{1-\delta_{ij}} f_{u,v}(u_i, v_i) du_i dv_i \right\}$$

In the case of random effects normal mixture model, the marginal likelihood becomes:

$$\begin{aligned}
& L(\boldsymbol{\theta}|Y_{ij}) \\
= & \prod_{i=1}^{n} \int_{u_i}\int_{v_i} \prod_{j=1}^{J_i} \left\{\sum_{k=1}^{K} p_k(u_i) \phi\left(\frac{y_{ij} - \mu_k(v_i)}{\sigma}\right)\right\}^{\delta_{ij}} \left\{\sum_{k=1}^{K} p_k(u_i) \Phi\left(\frac{D_0 - \mu_k(v_i)}{\sigma}\right)\right\}^{1-\delta_{ij}} \\
& f_{u,v}(u_i, v_i) du_i dv_i
\end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{(1)}, ..., \boldsymbol{\alpha}^{(K-1)}, \boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(K)}, \Sigma_{\mathbf{u}}, \Sigma_{\mathbf{v}}, \Sigma_{\mathbf{uv}})$, ϕ and Φ are the pdf and cdf of the standard normal distribution, respectively. The marginal likelihood of Y_{ij} for the special cases of two-component and three-component normal mixture models can be

obtained by specifying $K = 2$, and 3, respectively.

In the case of random effects log-normal mixture regression model, the marginal likelihood is:

$$\begin{aligned} & L(\boldsymbol{\theta}|Y_{ij}) \\ = & \prod_{i=1}^{n} \int_{u_i} \int_{v_i} \prod_{j=1}^{J_i} \left\{ \sum_{k=1}^{K} p_k(u_i) \phi \left[\frac{\log(y_{ij}) - \mu_k(v_i)}{\sigma} \right] \right\}^{\delta_{ij}} \left\{ \sum_{k=1}^{K} p_k(u_i) \Phi \left[\frac{\log D_0 - \mu_k(v_i)}{\sigma} \right] \right\}^{1-\delta_{ij}} \\ & f_{u,v}(u_i, v_i) du_i dv_i \end{aligned}$$

In the case of random effects gamma mixture regression model, the marginal likelihood is:

$$\begin{aligned} & L(\boldsymbol{\theta}|Y_{ij}) \\ = & \prod_{i=1}^{n} \int_{u_i} \int_{v_i} \prod_{j=1}^{J_i} \left\{ \sum_{k=1}^{K} p_k(u_i) \frac{y_{ij}^{b-1} \exp\left[-y_{ij}/a_k(v_i)\right]}{\Gamma(b) a_k^b(v_i)} \right\}^{\delta_{ij}} \left\{ \sum_{k=1}^{K} p_k(u_i) \frac{\gamma(b, D_0/a_k(v_i))}{\Gamma(b)} \right\}^{1-\delta_{ij}} \\ & f_{u,v}(u_i, v_i) du_i dv_i \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{(1)}, ..., \boldsymbol{\alpha}^{(K-1)}, \boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(K)}, b, \Sigma_{\mathbf{u}}, \Sigma_{\mathbf{v}}, \Sigma_{\mathbf{uv}})$, Γ is the gamma function, and $\gamma(b, D_0/a_k(v_i))$ is the lower incomplete gamma function.

The maximum likelihood estimator of parameter $\boldsymbol{\theta}$ can be obtained by maximizing the (marginal) likelihood $L(\boldsymbol{\theta}|Y)$. The consistency and asymptotic normality can be studied using the maximum likelihood theory (McLachlan and Peel, 2000, Yau et al., 2003).

3.1.3 Computational issues

In practice, the MLE $\boldsymbol{\theta}$ can be obtained using the optimizer, PROC NLMIXED in SAS. In PROC NLMIXED, one only needs to specify the conditional likelihood given the random effects, and this conditional likelihood could be of any form by using a GENERAL option in the MODEL statement. Then the joint marginal distribution of the random effects (u, v) needs to be specified in a separate RANDOM statement. To handle left censoring, one may create a flag variable for censoring and assign the

likelihood based on the value of the censoring flag variable.

The Quasi-Newton optimization method was used as the default option in SAS PROC NLMIXED. This algorithm does not require computation of second-order derivatives and it provides a good balance between speed and computational stability (SAS manual). The default approximate method for integrals is Adaptive Gaussian Quadrature, however, PROC NLMIXED may be bogged down trying to determine the number of quadruture points, and makes the computation time extremely long. Therefore, we use nonadaptive Gaussian quadrature by specifying NOAD option, and specify the number of quaduature points using qpoints= option. Insufficient number of quadrature points could lead to inaccurate parameter estimates, however, the accuracy deficiency of NOAD option could be compensated by specifying a higher value of qpoints. Usually nonadaptive Gaussian approach requires more quadruture points to achieve the same accuracy compared to the adaptive Gaussian approach) (Dmitrienko, 2005). In this dissertation, we specify qpoint=20 for each random effect. Theoretically, higher values of quadrature points (qpoints) will give more accurate parameter estimates but also require more computation time (SAS manual).

Specifying appropriate initial values for all parameters to be estimated is very important in fitting the mixture regression models. A step-wise strategy was used in this proposal. To get initial estimates for the regression parameters, we started by fitting a simpler mixture regression model without any random effects because it is much easier to fit a model without random effect. To get the parameter estimates for the random effects, we started with the fit of an intermediate model with only one random effect (u or v) and then used this as a starting point to estimate new parameters as we expanded the structure of the correlated random effects.

3.2 Simulation

3.2.1 Data generation

Although the proposed mixture regression model can be used for any finite number of components, in this dissertation, we studied the performance of the proposed method

using a two-component normal mixture regression model. The objectives of the simulation work include:

1). Study the performance of the estimation procedure for analyzing multimodal repeated measures data with left-censored observations using a random effects mixture regression model;

2). Evaluate the bias of the estimated parameters using a two-component mixture model without modeling the between-component correlation ρ when the correlation ρ exists;

3). Study the impact of the between-component correlation coefficient ρ on the mean and variance of the response.

To mimic the data structure of 8-isoprostane in the HEART study, we assumed that the data followed a random effects two-component mixture model specified below For simplicity, we assumed $v_{1i} = v_{2i} = v_i$. Thus, the data were generated from the distributions with the following pdf, given random effects u_i and v_i:

$$f(y_{ij}; \mathbf{x}_{ij}, u_i, v_i) = p(\mathbf{x}_{ij}, u_i) f_1(y_{ij}; \boldsymbol{\theta}_1(\mathbf{x}_{ij}, v_i)) + [1 - p(\mathbf{x}_{ij}, u_i)]\, f_2(y_{ij}; \boldsymbol{\theta}_2(\mathbf{x}_{ij}, v_i))$$

where $f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_i))$ is a normal density with:

$$f_k(y_{ij}; \boldsymbol{\theta}_1(\mathbf{x}_{ij}, v_i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[y_{ij} - \mu_k(\mathbf{x}_{ij}, v_i)]^2}{2\sigma^2} \right\},$$

where $\mu_k(\mathbf{x}_{ij}, v_i) = \mathbf{x}'_{ij}\boldsymbol{\beta}^{(k)} + v_i, \mathbf{x}'_{ij} = (1, x_1, x_2), x_1, x_2$ are indicator variables corresponding to during-Olympic period (1 for yes and 0 for the otherwise) and post-Olympic period (1 for yes and 0 for the otherwise), respectively, and $\boldsymbol{\beta}^{(k)'} = (\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)})'$ is a vector of coefficients for the kth component ($k = 1, 2$). The mixing proportion for the first component $p(\mathbf{x}_{ij}, u_i)$ follows a logistic regression, given the random effect u_i:

$$\text{logit}\,\{p(\mathbf{x}_{ij}, u_i)\} = \mathbf{x}'_{ij}\boldsymbol{\alpha} + u_i$$

where $\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \alpha_2)'$ is a vector of coefficients corresponding to $\mathbf{x}_{ij}$, u_i is the random effect corresponding to the mixing proportion $p(\mathbf{x}_{ij}, u_i)$ for subject i. In addition, we

Table 3.1: The true parameters used for simulating left-censored longitudinal data.

	Scenario 1	Scenario 2	Scenario 3
α_0	-0.6	-0.6	-0.6
α_1	0.5	0.5	0.5
α_2	0.2	0.2	0.2
$\beta_0^{(1)}$	2	2	2
$\beta_1^{(1)}$	0.2	0.2	0.2
$\beta_2^{(1)}$	0.3	0.3	0.3
$\beta_0^{(2)}$	14	14	14
$\beta_1^{(2)}$	0.5	0.5	0.5
$\beta_2^{(2)}$	0.4	0.4	0.4
σ	4	4	4
σ_1	1	1	1
σ_2	0.8	0.8	0.8
ρ	0.5	0	-0.5

generated random effects u_i and v_i from a bivariate normal distribution with mean zero and variance-covariance

$$\begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix},$$

where $\sigma_u, \sigma_v > 0$, and $-1 \leq \rho \leq 1$. For each period, two observations from each subject were generated.

The true values used for the simulation are shown in Table 3.1. There are three scenarios with different choices of ρ while other parameters remain the same. The parameters were chosen so that each component of the mixture model has a reasonable proportion to enable the convergence of parameter estimation in SAS. For each run of the simulation, the total number of subjects was chosen as $n = 1000$. Each subject had 6 repeated measurements with two measurements per period. Any simulated values < 1.56 (the detection limit in the motivating example) were set to 1.56 and they are considered left censored in the estimation procedure. Based on these simulation settings, the simulated data have about 30% observations below detection limit in total.

3.2.2 Performance of the estimated parameters in various scenarios

Simulated data were analyzed using a **two-component normal mixture regression model**. A total of 1000 and 2000 times of simulation was conducted to evaluate the performance of the proposed estimating procedure for each scenario specified in Table 3.1. Tables 3.2-3.4 show the estimated parameters and the corresponding relative bias in different scenarios with different values of ρ. The numbers in the parenthesis are Monte Carlo variance estimate and the average of the variance estimates based on the likelihood. It has been shown that the estimated parameters were consistent to the true values, and there was little difference between 1000 simulations and 2000 simulations. The Monte Carlo variance estimates and the variance estimates based on the likelihood are approximately the same for all parameters.

Table 3.2: Estimated parameters using random effect two component normal mixture models for scenario 1 ($\rho = 0.5, m = 1000$ subjects).

		True	1000 simulation		2000 simulation	
		value	Estimate(variance)	Bias	Estimate(variance)	Bias
logit(p)	α_0	-0.6	$-0.602(0.006^a, 0.006^b)$	-0.4%	$-0.599(0.006^a, 0.006^b)$	-0.2%
	α_1	0.5	$0.500(0.009, 0.009)$	-0.1%	$0.500(0.010, 0.009)$	0%
	α_2	0.2	$0.204(0.009, 0.009)$	2.2%	$0.202(0.009, 0.009)$	1%
Lower	$\beta_0^{(1)}$	2	$2.002(0.070, 0.070)$	0.1%	$1.995(0.068, 0.070)$	-0.3%
comp	$\beta_1^{(1)}$	0.2	$0.189(0.105, 0.100)$	-5.6%	$0.204(0.097, 0.100)$	2%
mean	$\beta_2^{(1)}$	0.3	$0.297(0.110, 0.109)$	-1.0%	$0.300(0.108, 0.109)$	0%
Higher	$\beta_0^{(2)}$	14	$14.003(0.028, 0.028)$	0%	$14.006(0.028, 0.028)$	0%
comp	$\beta_1^{(2)}$	0.5	$0.493(0.050, 0.051)$	-1.4%	$0.498(0.052, 0.051)$	-0.4%
mean	$\beta_2^{(2)}$	0.4	$0.403(0.047, 0.048)$	0.8%	$0.401(0.0450.048)$	0.3%
Variances	σ	4	$3.998(0.006, 0.005)$	-0.1%	$3.996(0.006, 0.005)$	-0.1%
	σ_v	1	$0.979(0.052, 0.051)$	-2.1%	$0.974(0.048, 0.050)$	-2.6%
	σ_u	0.8	$0.797(0.005, 0.005)$	-0.3%	$0.801(0.005, 0.005)$	0.1%
	ρ	0.5	$0.505(0.028, 0.057)$	1.1%	$0.506(0.030, 0.044)$	1.2%

a: Average of the variance estimates based on the likelihood.
b: Monte Carlo variance estimate.

Table 3.3: Estimated parameters using random effect two component normal mixture models for scenario 1 ($\rho = 0, m = 1000$ subjects).

		True	1000 simulation		2000 simulation	
		value	Estimate(variance)	Bias	Estimate(variance)	Bias
logit(p)	α_0	-0.6	$-0.599(0.006^a, 0.005^b)$	0.2%	$-0.600(0.005^a, 0.005^b)$	0%
	α_1	0.5	$0.498(0.009, 0.008)$	-0.4%	$0.499(0.009, 0.008)$	-0.2%
	α_2	0.2	$0.200(0.009, 0.009)$	0%	$0.200(0.008, 0.009)$	0%
Lower	$\beta_0^{(1)}$	2	$2.007(0.071, 0.068)$	0.4%	$1.995(0.063, 0.068)$	-0.3%
comp	$\beta_1^{(1)}$	0.2	$0.181(0.103, 0.097)$	-9.5%	$0.198(0.097, 0.097)$	-1%
mean	$\beta_2^{(1)}$	0.3	$0.293(0.114, 0.106)$	-2.3%	$0.304(0.109, 0.106)$	1.3%
Higher	$\beta_0^{(2)}$	14	$14.007(0.027, 0.027)$	-0.1%	$14.002(0.027, 0.026)$	0%
comp	$\beta_1^{(2)}$	0.5	$0.500(0.056, 0.049)$	0%	$0.496(0.048, 0.049)$	-0.8%
mean	$\beta_2^{(2)}$	0.4	$0.394(0.0480.046)$	-1.5%	$0.400(0.0450.046)$	0%
Variances	σ	4	$4.000(0.005, 0.005)$	0%	$3.996(0.005, 0.005)$	-0.1%
	σ_v	1	$0.971(0.042, 0.042)$	-2.9%	$0.980(0.041, 0.042)$	-2%
	σ_u	0.8	$0.797(0.005, 0.005)$	-0.3%	$0.797(0.005, 0.005)$	-0.3%
	ρ	0	$-0.041(0.061, 0.072)$		$-0.036(0.058, 0.071)$	

a: Average of the variance estimates based on the likelihood.
b: Monte Carlo variance estimate.

3.2.3 Bias introduced when ignoring the correlation between components

In order to investigate the extent of bias in estimated parameters if the between-component correlation ρ is not modeled when it exists, simulated data were fitted by the correct **random effects two component normal mixture model** (described in section 5.1) and a misspecified model without the between-component correlation ($\rho = 0$). Table 3.5 shows the same analysis results for $n = 1000$ subjects and 2000 simulations. It can be seen that the incorrect model would generate bias in the estimated parameters for up to 20%. The bias is smaller when the sample size is bigger.

3.2.4 Effect of ρ on the mean, variance and covariance

We next examined the effect of ρ on the shape of the mixture distributions. Datasets with $n = 5000$ subjects were simulated based on the model described in Section 5.1 with different values of correlation between the random effects u and v (ρ): $\rho = 0.8$, 0, and -0.8. Histograms and fitted probability density curves based on the simulated

Table 3.4: Estimated parameters using random effect two component normal mixture models for scenario 1 ($\rho = -0.5, m = 1000$ subjects).

		True	1000 simulation		2000 simulation	
		value	Estimate(variance)	Bias	Estimate(variance)	Bias
logit(p)	α_0	-0.6	$-0.604(0.005^a, 0.005^b)$	-0.6%	$-0.601(0.005^a, 0.005^b)$	0.2%
	α_1	0.5	$0.506(0.008, 0.008)$	1.2%	$0.502(0.008, 0.008)$	0.5%
	α_2	0.2	$0.202(0.009, 0.008)$	0.9%	$0.204(0.008, 0.008)$	2.2%
Lower	$\beta_0^{(1)}$	2	$1.986(0.067, 0.065)$	-0.7%	$1.988(0.063, 0.064)$	-0.6%
comp	$\beta_1^{(1)}$	0.2	$0.214(0.099, 0.094)$	7.1%	$0.204(0.094, 0.094)$	1.9%
mean	$\beta_2^{(1)}$	0.3	$0.307(0.108, 0.103)$	2.3%	$0.316(0.105, 0.102)$	5.5%
Higher	$\beta_0^{(2)}$	14	$13.994(0.027, 0.025)$	0%	$14.005(0.026, 0.025)$	0%
comp	$\beta_1^{(2)}$	0.5	$0.506(0.048, 0.046)$	1.3%	$0.496(0.048, 0.046)$	-0.8%
mean	$\beta_2^{(2)}$	0.4	$0.408(0.045, 0.043)$	2.0%	$0.402(0.043, 0.042)$	0.5%
Variances	σ	4	$3.997(0.005, 0.005)$	-0.1%	$3.996(0.005, 0.005)$	-0.1%
	σ_v	1	$0.981(0.031, 0.032)$	-1.9%	$0.988(0.032, 0.032)$	0%
	σ_u	0.8	$0.797(0.005, 0.004)$	-0.4%	$0.800(0.005, 0.005)$	0%
	ρ	-0.5	$-0.547(0.063, 0.074)$	9.3%	$-0.537(0.062, 0.075)$	7.3%

a: Average of the variance estimates based on the likelihood.
b: Monte Carlo variance estimate.

data were shown on Figure 3.1.

We can see that a positive value of ρ tend to move the two peaks closer to each other, and a negative ρ tends to push the two peaks away from each other. This leads to the greater standard deviation as $\rho < 0$ and smaller standard deviation as $\rho > 0$. Table 3.6 displays the sample mean and standard deviation of the simulated data using different value of ρ: 0.8, 0 and -0.8. It can be seen that the sample means do not vary much, however, the sample standard deviation changes with ρ. The more positive value of ρ, the smaller the standard deviation; the more negative value of ρ, the higher the stand deviation. Therefore, it seemed that the value of the between-component correlation ρ controlled the distance between the two peaks of the mixture distribution, and thus the variance of the overall response.

3.3 Analysis of the Example Data

Some analysis results for the example data are presented in this section. Starting from some exploratory analysis of the data, we showed the analysis using the proposed

Table 3.5: Estimated parameters using the correct model and incorrect model with two independent random effects (m=1000 subjects, 2000 simulations).

		True	Two correlated random effects u and v		Two independent random effects u and v	
		value	Estimate(variance)	Bias	Estimate(variance)	Bias
logit(p)	α_0	-0.6	$-0.602(0.006^a, 0.006^b)$	-0.4%	$-0.584(0.006^a, 0.006^b)$	-2.6%
	α_1	0.5	$0.500(0.009, 0.009)$	-0.1%	$0.487(0.008, 0.009)$	-2.7%
	α_2	0.2	$0.204(0.009, 0.009)$	2.2%	$0.194(0.009, 0.009)$	-3.2%
Lower	$\beta_0^{(1)}$	2	$2.002(0.070, 0.070)$	0.1%	$2.204(0.064, 0.062)$	10.2%
comp	$\beta_1^{(1)}$	0.2	$0.189(0.105, 0.100)$	-5.6%	$0.173(0.098, 0.100)$	-13.5%
mean	$\beta_2^{(1)}$	0.3	$0.297(0.110, 0.109)$	-1.0%	$0.288(0.113, 0.109)$	-4.0%
Higher	$\beta_0^{(2)}$	14	$14.003(0.028, 0.028)$	0%	$13.882(0.026, 0.025)$	-0.8%
comp	$\beta_1^{(2)}$	0.5	$0.493(0.050, 0.051)$	0.8%	$0.464(0.053, 0.051)$	-7.3%
mean	$\beta_2^{(2)}$	0.4	$0.403(0.047, 0.048)$	0.8%	$0.385(0.045, 0.048)$	-3.8%
Variances	σ	4	$3.998(0.006, 0.005)$	-0.1%	$4.054(0.005, 0.005)$	1.3%
	σ_v	1	$0.979(0.052, 0.051)$	-2.1%	$0.588(0.071, 0.162)$	-41%
	σ_u	0.8	$0.797(0.005, 0.005)$	-0.3%	$0.696(0.003, 0.004)$	-13%
	ρ	0.5	$0.505(0.028, 0.057)$	1.1%	–	

a: Average of the variance estimates based on the likelihood.
b: Monte Carlo variance estimate.

Table 3.6: The sample mean and standard deviation of the simulated data using different values of ρ.

ρ	Mean	Standard deviation
0.8	8.83	5.99
0	8.75	6.37
-0.8	9.00	6.82

random effects two-component normal mixture regression model, followed by the three-component model. Then we determine the number of components in the mixture model by comparing these two mixture regression models using some criteria, such as AIC and BIC. The good-of-fit of the selected model was also evaluated.

3.3.1 Introduction of the data and descriptive statistics of the data

As described in the motivating example, this example data contains the EBC 8-isoprotane biomarker level for a total of 124 subjects at 3 different periods: "Pre-Olympic period",

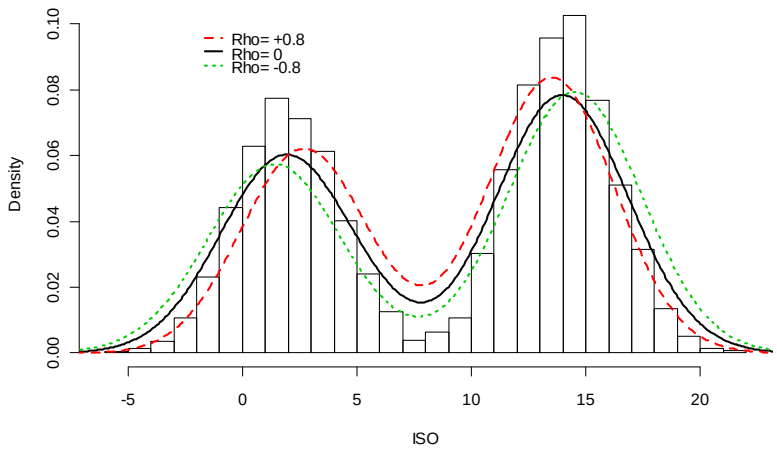

Figure 3.1: Empirical pdf for simulated data based on different values of ρ.

"During-Olympic period" and "Post-Olympic period". Each subject has two measurements of 8-isoprostane at each period.

As shown in Table 1.1, the levels of 8-isoprostane dropped substantially during the Olympic Games period and then rose up after the Games. Figure 3.2 shows the histogram of 8-isoprostane stratified by period. The vertical line at $x = 1.56$ pg/ml showed the detection limit of 8-isoprostane. It appeared that there was a substantial proportion of observations below detection limit and the remaining observations clustered around values of 5 and 10, which suggested that the data may not follow a standard normal distribution. A two-component or three-component mixture model may be more appropriate to describe the data.

3.3.2 Two-Component normal mixture regression model

We began the data analysis by fitting the random effect two-component normal mixture regression model described described on p.34, with the non-detectbles handled by the left-censoring approach. One of the challenges in fitting mixture regression models is to find the appropriate starting values. Therefore, we used a step-wise method to obtain starting values for parameters to be estimated: First, the data was fitted by a simple model without any random effects ($u = v = 0$), and then these estimated parameters were used as the starting values for fitting more complex models with more random

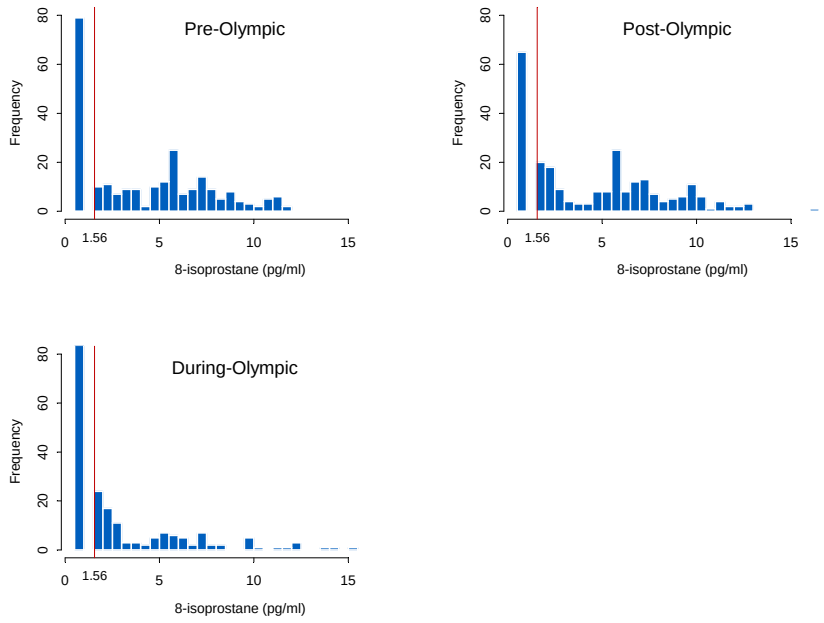

Figure 3.2: Histograms of 8-isoprostane biomarker at different periods.

effects. Table 3.7 shows the parameter estimates and their standard errors based on two-component random effect normal mixture models with different specifications of random effects u and v:

Model 1: No random effects, $u = v = 0$;

Model 2: Only one random effect, $u = 0, v \sim N(0, \sigma_v^2)$;

Model 3: Two independent random variables u and v, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, $\text{CORR}(u, v) = 0$,

Model 4: Two correlated random variables u and v, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, $\text{CORR}(u, v) = \rho \neq 0$.

The likelihood (-2logL), AIC and BIC for all models are also given in Table 3.7. Based on AIC, BIC criteria, model 2 with only one random effect $v(u = 0)$ is the best model as it has the smallest value of AIC and BIC. Note that model 3 or 4 may be preferred for a different data set although random variable v seems to be unnecessary for the example HEART study data.

Table 3.7: Data analysis results using two-component mixture regression models by a step-wise procedure: parameter estimates (SE).

	Parameters	Model 1[a]	Model 2[b]	Model 3[c]	Model 4[d]
logit(p)	α_0	-0.230 (0.230)	-0.006 (0.140)	-0.006 (0.146)	-0.005 (0.147)
	α_1	1.979 (0.362)	1.511 (0.224)	1.543 (0.230)	1.547 (0.231)
	α_2	0.320 (0.303)	0.041 (0.196)	0.042 (0.199)	0.038 (0.200)
Lower	$\beta_0^{(1)}$	0.242 (0.602)	0.688 (0.291)	0.686 (0.290)	0.670 (0.293)
comp	$\beta_1^{(1)}$	0.200 (0.654)	-0.319 (0.305)	-0.322 (0.304)	-0.320 (0.306)
mean	$\beta_2^{(1)}$	1.066 (0.725)	0.358 (0.333)	0.354 (0.331)	0.341 (0.333)
Higher	$\beta_0^{(2)}$	6.568 (0.363)	7.060 (0.226)	7.068 (0.224)	7.085 (0.227)
comp	$\beta_1^{(2)}$	1.818 (0.748)	0.668 (0.354)	0.671 (0.352)	0.697 (0.356)
mean	$\beta_2^{(2)}$	0.995 (0.468)	0.681 (0.258)	0.683 (0.257)	0.684 (0.258)
Variances	σ	2.484 (0.158)	1.646 (0.076)	1.643 (0.075)	1.644 (0.075)
	σ_v		1.464 (0.142)	1.451 (0.140)	1.466 (0.145)
	σ_u			0.353 (0.207)	0.371 (0.203)
	ρ				0.233 (0.300)
Test criteria	-2logL	3062.8	2986.2	2985.3	2984.7
	AIC	3082.8	3008.2	3009.3	3010.7
	BIC	3128.9	3039.2	3043.1	3047.3

a: Model 1, $u = v = 0$
b: Model 2, $u = 0, v \sim N(0, \sigma_v^2)$
c: Model 3, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, CORR$(u, v) = 0$
d: Model 4, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, CORR$(u, v) = \rho \neq 0$

Table 3.8 shows the estimated mixing proportion p and the estimated mean of each component normal distribution at different periods based on the fitted model with only one random effect v (model 2 in Table 3.7). Based on this model, all subjects are from either one of the two subpopulations with different mean values of 8-isoprostane biomarker level. For example, at the pre-Olympic period, 49.8% of the subjects are from the lower response subpopulation with a mean 8-isoprostane of 0.688, and the remaining subjects are from the higher response group with a higher mean 8-isoprostane level, 7.06. During the Olympic Games, this mixing proportion from the lower response group dramatically increases to 81.8%. The mean 8-isoprostane level of the lower response subpopulation reduces by about 50%, while the mean of the higher response group has only about 10% increase. At the post-Olympic period, the mixing proportion from the lower response group dropped back to around 50%, but the mean 8-isoprostane level of the lower response group keeps increasing to 1.045, and the mean of the higher response

group had a little change from during-Olympic period.

Table 3.8 also gives the estimated overall mean 8-isoprostane level at each period based on the fitted random effect two-component normal mixture regression model (model 2 in Table 3.7). The standard errors of the estimated overall means are computed from the standard error of the estimated model parameters using Δ-method. A likelihood ratio test is used to test the null hypothesis of constant overall mean 8-isoprostane level across different Olympic periods:

H_0: the overall mean of 8-isoprostane level does not change across different periods;

H_a: the overall mean of 8-isoprostane level is different across different periods.

Based on the fitted model (model 2 in Table 3.7), it is the same as testing:

$H_0 : \alpha_1 = \alpha_2 = \beta_1^{(1)} = \beta_2^{(1)} = \beta_1^{(2)} = \beta_2^{(2)} = 0$;

H_a: at least one of $\alpha_1, \alpha_2, \beta_1^{(1)}, \beta_2^{(1)}, \beta_1^{(2)}, \beta_2^{(2)}$ is not zero.

Under the above null hypothesis, the mixing proportion in the reduced two-component mixture model is

$$\text{logit}\left\{p(\mathbf{x}_{ij})\right\} = \alpha_0 + u_i$$

and the component means in the reduced two-component mixture model is $\mu_k(\mathbf{x}_{ij}, v_i) = \beta_0^{(k)} + v_i$, where $k = 1, 2$. As shown in Table 3.7, the -2LogL for the full model with one random effect v is 2986.2. For the reduced model under the null hypothesis, the -2logL is 3070.9, and the difference in the degree of freedom between these two models is 6. The difference in the -2logL between these two models is 84.7 and it follows a chi-square distribution with d.f.=6. The p-value < 0.001, and it is concluded that the overall mean 8-isoprostane level changes across different periods.

3.3.3 Three-component normal mixture regression model

Histograms in Figure 3.2 also suggested a three-component mixture distribution for the 8-isoprostane data in the HEART study. In this section, a random effects three-component normal mixture model assuming common random measurement error ($v_{1i} =$

Table 3.8: The mean proportion p and mean value of each mixture model component at different periods.

	Pre-Olympic	During Olympic	Post-Olympic	χ^2	d.f.	p-value
p (SE)	0.498(0.035)	0.818 (0.026)	0.509 (0.034)			
μ_1 (SE)	0.688(0.291)	0.369(0.231)	1.045(0.266)			
μ_2 (SE)	7.060(0.226)	7.728 (0.334)	7.741(0.226)			
Overall Mean (SE)	3.884(0.273)	1.707(0.269)	4.335(0.277)	84.7	6	< 0.0001

*SE: standard errors are calculated based on the standard errors of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}^{(k)}$ using Δ-method.

$v_{2i} = v_{3i} = v_i$) was used to analyze the 8-isoprostane data:

$$f(y_{ij}; \mathbf{x}_{ij}) = \sum_{k=1}^{3} p_k(\mathbf{x}_{ij}, u_i) f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_i))$$

where $\sum_{k=1}^{3} p_k(\mathbf{x}_{ij}, u_i) = 1$, and $f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_i))$ is a normal density with:

$$f_k(y_{ij}; \boldsymbol{\theta}_k(\mathbf{x}_{ij}, v_i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[y_{ij} - \mu_k(\mathbf{x}_{ij}, v_i)]^2}{2\sigma^2} \right\}$$

where $\mu_k(\mathbf{x}_{ij}, v_i) = \mathbf{x}'_{ij}\boldsymbol{\beta}^{(k)} + v_i$, $x'_{ij} = [1, x_1, x_2]$, x_1, x_2 are dummy variables corresponding to during-Olympic period and post-Olympic period, $\boldsymbol{\beta}^{(k)\prime} = (\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)})'$ is a vector of coefficients for the kth component ($k = 1, 2, 3$). The mixing proportions $p_k(\mathbf{x}_{ij}, u_i)$ follow a multi-logistic regression, given the random effect u_i:

$$p_k(\mathbf{x}_{ij}, u_i) = \begin{cases} \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(1)} + u_i)}{1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(1)} + u_i) + \exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(2)})}, & \text{if } k = 1, \\ \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(2)})}{1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(1)} + u_i) + \exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(2)})}, & \text{if } k = 2, \\ \frac{1}{1 + \exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(1)} + u_i) + \exp(\mathbf{x}'_{ij}\boldsymbol{\alpha}^{(2)})}, & \text{if } k = 3, \end{cases}$$

where $\boldsymbol{\alpha}^{(k)\prime} = (\alpha_0^{(k)}, \alpha_1^{(k)}, \alpha_2^{(k)})'$, $k = 1, 2$, is a vector of coefficients corresponding to $\mathbf{x}_{ij}$, u_i is the random effect corresponding to the mixing proportion $p_1(\mathbf{x}_{ij}, u_i)$ for subject i. A random variable u_i is only specified in $p_1(\mathbf{x}_{ij}, u_i)$ to simplify computation because the mixing proportion $p_2(\mathbf{x}_{ij}, u_i)$ and $p_3(\mathbf{x}_{ij}, u_i)$ seem to be positively correlated according

to the 8-isoprostane data. In addition, we assume that the random variables u_i and v_i follow a bivariate normal distribution with mean zero and variance-covariance

$$\begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix},$$

where $\sigma_u, \sigma_v > 0$, and $-1 \leq \rho \leq 1$.

A step-wise procedure was also used to find the appropriate starting values for the parameters to be estimated in the model. Table 3.9 shows the data analysis results using random effects three-component mixture regression models with different specifications of random effects u and v:

Model 1: No random effect, $u = v = 0$;

Model 2: Only one random effect, $u = 0, v \sim N(0, \sigma_v^2)$;

Model 3: Two independent random variables u and v, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, $\text{CORR}(u, v) = 0$;

Model 4: Two correlated random variables u and v, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, $\text{CORR}(u, v) = \rho \neq 0$; The parameter estimates from the simpler model are used as the initial values for the more complex model. Since model 2 has the smallest value of AIC and BIC, the mixture model with one random effect (model 2) is selected as the best model.

Based on the selected three-component mixture regression model with one random effect v (model 2 in Table 3.9), we can estimate the mixing proportions and normal component means at each period. Table 3.10 shows the estimated mixing proportions and the mean 8-isoprostane level for each one of the three subpopulations. It can be seen that there are three different subpopulations with the mean 8-isoprostane level around 1.2, 6.3, and 9.5 at pre-Olympic period, and these mean values changes slightly with different periods. However, the mixing proportions change significantly with period. For example, the probability of a subject from the lowest response level group increases from 49% at pre-Olympic period to 80% during Olympics Games. And this percentage reduces back to 49% at post-Olympic period. So it seems that some subjects jumped between the lowest response level group and higher response level groups as

Table 3.9: Data analysis results using random effects three-component mixture regression models: parameter estimates (SE)

	Parameters	Model 1[a]	Model 2[b]	Model 3[c]	Model 4[d]
logit(p_1)	$\alpha_0^{(1)}$	1.301 (0.284)	1.422 (0.299)	1.424 (0.302)	1.423 (0.302)
	$\alpha_1^{(1)}$	1.408 (0.416)	1.203 (0.414)	1.247 (0.417)	1.248 (0.417)
	$\alpha_2^{(1)}$	-0.139 (0.360)	-0.115 (0.361)	-0.118 (0.363)	-0.117 (0.363)
logit(p_2)	$\alpha_0^{(2)}$	1.156 (0.311)	1.186 (0.341)	1.188 (0.341)	1.187 (0.341)
	$\alpha_1^{(2)}$	-0.040 (0.464)	-0.312 (0.478)	-0.314 (0.478)	-0.314 (0.478)
	$\alpha_2^{(2)}$	-0.236 (0.395)	-0.160 (0.405)	-0.163 (0.405)	-0.161 (0.405)
Lower	$\beta_0^{(1)}$	0.974 (0.208)	1.176 (0.180)	1.184 (0.179)	1.182 (0.180)
comp	$\beta_1^{(1)}$	-0.228 (0.238)	-0.317 (0.163)	-0.321(0.163)	-0.321(0.162)
mean	$\beta_2^{(1)}$	0.257 (0.257)	0.078 (0.172)	0.075 (0.171)	0.074 (0.171)
Moderate	$\beta_0^{(2)}$	6.030 (0.229)	6.305 (0.198)	6.314 (0.199)	6.316 (0.200)
comp	$\beta_1^{(2)}$	0.114 (0.363)	-0.067 (0.268)	-0.061 (0.265)	-0.059 (0.266)
mean	$\beta_2^{(2)}$	0.310 (0.304)	0.435 (0.216)	0.439 (0.214)	0.440 (0.214)
Higher	$\beta_0^{(3)}$	9.743 (0.401)	9.500 (0.369)	9.509 (0.371)	9.508 (0.371)
comp	$\beta_1^{(3)}$	1.936 (0.593)	1.065 (0.502)	1.077 (0.499)	1.079 (0.499)
mean	$\beta_2^{(3)}$	0.714 (0.501)	0.967 (0.429)	0.963 (0.430)	0.965 (0.430)
Variances	σ	1.352 (0.067)	0.918 (0.042)	0.919 (0.042)	0.919 (0.042)
	σ_v		1.248 (0.127)	1.230 (0.134)	1.229 (0.135)
	σ_u			0.405 (0.176)	0.407 (0.176)
	ρ				0.046 (0.258)
Test criteria	-2logL	3005.4	2869.4	2867.6	2867.6
	AIC	3037.4	2903.4	2903.6	2905.6
	BIC	3111.2	2951.3	2954.4	2959.2

*: SAS options: noad qppoint=20, gconv=1E-13,
a: Model 1, $u = v = 0$
b: Model 2, $u = 0, v \sim N(0, \sigma_v^2)$
c: Model 3, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, CORR$(u, v) = 0$
d: Model 4, $u \sim N(0, \sigma_u^2), v \sim N(0, \sigma_v^2)$, CORR$(u, v) = \rho \neq 0$

the pollutant level changes.

The estimated overall means of 8-isoprostane level are also given in Table 3.10. A likelihood ratio test for testing constant overall means across different period is also performed:

$H_0 : \alpha_1^{(1)} = \alpha_2^{(1)} = \alpha_1^{(2)} = \alpha_2^{(2)} = \beta_1^{(1)} = \alpha_2^{(1)} = \alpha_1^{(2)} = \alpha_2^{(2)} = \alpha_1^{(3)} = \alpha_2^{(3)} = 0$;

H_a: at least one of $\alpha_1^{(1)}, \alpha_2^{(1)}, \alpha_1^{(2)}, \alpha_2^{(2)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_1^{(3)}, \beta_2^{(3)}$ is not zero. As shown in Table 3.9, the -2logL for the full model with one random effect v is 2869.4.

Table 3.10: The mean proportion and mean value of each mixture model component at different periods based on the final model (with one random effect v).

		Pre-Olympic	During Olympic	Post-Olympic	χ^2	d.f.	p-value
Lowest	p_1 (SE)	0.492 (0.033)	0.802 (0.025)	0.493 (0.032)			
response	μ_1 (SE)	1.176 (0.180)	0.859 (0.153)	1.254 (0.166)			
Moderate	p_2 (SE)	0.389 (0.039)	0.139 (0.022)	0.373 (0.033)			
response	μ_2 (SE)	6.035 (0.198)	6.238 (0.229)	6.740 (0.166)			
Highest	p_3 (SE)	0.119 (0.032)	0.058 (0.015)	0.134 (0.024)			
response	μ_3 (SE)	9.500 (0.369)	10.566 (0.352)	10.467 (0.250)			
Overall Mean (SE)		4.160 (0.234)	2.173 (0.226)	4.532 (0.252)	94.4	10	< 0.0001

*SE: standard errors are calculated based on the standard errors of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}^{(k)}$ using Δ-method.

For the reduced model under the null hypothesis, the -2logL is 2963.8, and the difference in the degree of freedom is 10. The difference in the -2logL between these two models is 94.4 and it follows a chi-square distribution with d.f.=10. The p-value < 0.001, and it is concluded that the overall mean 8-isoprostane level changes across different periods.

3.3.4 Determination of the number of components

Two-component and three-component normal mixture regression models have been applied to the analysis of the HEART study data in sections 6.3 and 6.4, respectively. Therefore, we would like to determine the optimum number of components in a mixture regression model for the example data. In this section, we compared mixture models with different number of components and select the best model.

Table 3.11 shows the goodness-of-fit criteria for random effect normal mixture regression models with different number of component distributions. Obviously one component mixture model is just a random effect model. All these models assume that there is a common random error v for each of the component distribution mean and no random effect ($u = 0$) for any mixing proportion. Based on the values of AIC, AIC3 and BIC, the three-component mixture regression model is the best one. AIC3 is a modified AIC with more penalty for the number of parameters and is equal to $-2\text{logL}+3n$. Since the regular likelihood ratio test is not valid for comparing two mixture models as the

regularity condition does not hold, a modified likelihood ratio test is performed (Wolfe, 1971). The modified LRT test statistic is computed by $-2/n(n-1-d-c_1/2)logL$, and it follows a chi-square with d.f. $= 2d(c_1 - c_0)$. The modified likelihood ratio test gives a p-value of < 0.0001, therefore, a three-component mixture regression model is needed to describe the data.

Table 3.11: Normal mixture distribution fits with different number of components. (with only one random effect v).

Parameters	One component	Two-Component	Three-component
-2logL	3076.4	2986.2	2869.4
AIC	3086.4	3008.2	2903.4
AIC3*	3091.4	3019.2	2920.4
BIC	3100.5	3039.2	2951.3
p-value for comparing to higher order mixture model	< 0.001	< 0.001	

*: AIC3 $= -2\text{logL} + 3n$.

3.3.5 Goodness-of-fit for the model

Figure 3.3 shows the histograms of the example data with the pdfs of the single-component, two-component (model 2 in Table 8), three-component normal mixture models (model 2 in Table 11) using the estimated parameters for each model. From these plots, we can see that the proposed three-component normal mixture model describes the data better than the two-component mixture model.

3.3.6 Computational issues

We used different optimization methods for fitting the mixture regression models using SAS PROC NLMIXED: The default Quasi-Newton approach, Newton-Raphson line search and Newton-Raphson Ridge optimization. These methods gave almost identical parameter estimates. However, the default Quasi-Newton approach was fastest to converge and the Newton-Raphson Ridge optimization was slowest. In addition, when we used the non-adaptive Gassian method, and specified a higher value of Qpoints= option, better parameter estimates were obtained. However, higher values of quadruture

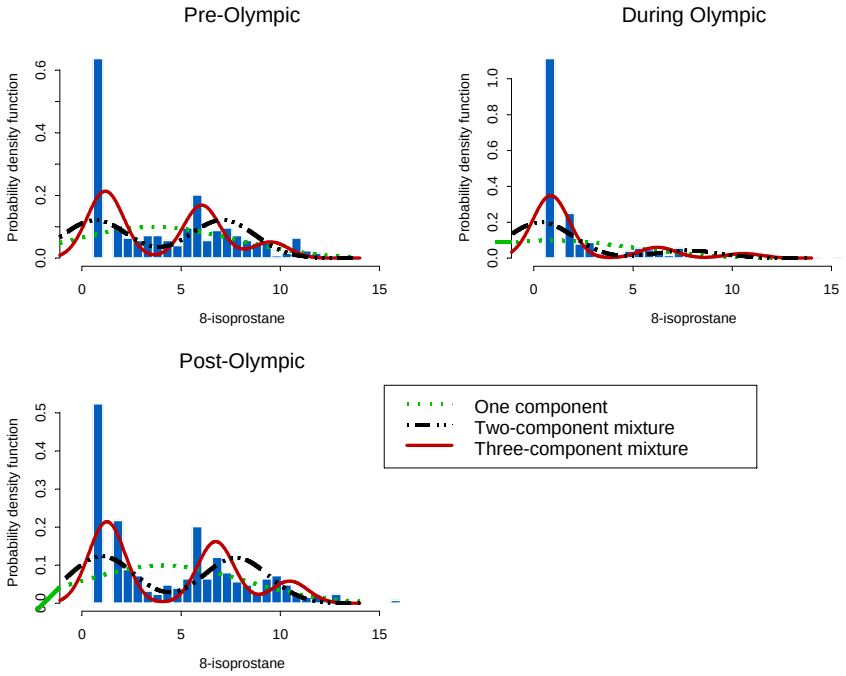

Figure 3.3: Histogram of the real data with the pdf of normal mixture model based on the estimated parameters.

points cost more computational time, especially when there are many random effects in the model. Therefore, it is important to balance better accuracy of parameter estimates and faster computation. For the example data, Qpoints=20 seemed to be sufficient.

Chapter 4

Goodness-of-fit Test for Mixture Regression Models

In this chapter, we proposed a goodness-of-fit (GOF) test for mixture models. We started with a two component normal mixture model without random effects, and extended the proposed test to a more complicated mixture model with random effects. We developed the prototype of the proposed methodology in this chapter and deferred the application to the detection limit problem in Discussion and Future Work.

4.1 Two component mixture models without random effects

4.1.1 Model settings

Let Y_i denote the response of subject i, for $i = 1, 2, ..., n$, and assume that the response Y_i follows a two-component normal mixture model with density:

$$f(y; \mu^{(1)}, \mu^{(2)}, \sigma) = p\phi(y; \mu^{(1)}, \sigma^2) + (1-p)\phi(y; \mu^{(2)}, \sigma^2),$$

where $\phi(\cdot; \mu^{(k)}, \sigma^2)$ denotes the normal density function with mean $\mu^{(k)}$ and variance σ^2, for $k = 1, 2$, and p denotes the mixing proportion. Moreover, for each subject i, we allow $\mu^{(1)}$, $\mu^{(2)}$, and p to depend on covariate information $\mathbf{X}_i$ and $\mathbf{T}_i$ via regression models:

$$\mu^{(k)}(\mathbf{X}_i; \boldsymbol{\beta}^{(k)}) = \mathbf{X}_i'\boldsymbol{\beta}^{(k)},$$

and

$$\text{logit}\{p(\mathbf{T}_i; \boldsymbol{\alpha})\} = \mathbf{T}_i'\boldsymbol{\alpha},$$

where $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\alpha}$ denote the regression parameter vectors ($k = 1, 2$), and $\mathbf{X}_i$ and $\mathbf{T}_i$ denote $(p+1)-$ and $(q+1)-$variate covariate vectors including 1's for intercepts in the

respective regression models. For simplicity of notation, we use $p(\boldsymbol{\alpha})$ to denote $p(\mathbf{T}_i;\boldsymbol{\alpha})$ in this dissertation. Note that $\mathbf{X}_i$ and $\mathbf{T}_i$ do not need to be identical, but may share some common covariates, and the link function for $p(\boldsymbol{\alpha})$ is not necessarily restricted to the logit function. However, in this dissertation, we use the logit link as an illustration.

4.1.2 Statistical properties of the two-component mixture model

Assume that each individual i is independent of each other. The likelihood based on the observed data $\{\mathbf{Y},\mathbf{X},\mathbf{T}\}=\{(Y_i,\mathbf{X}_i',\mathbf{T}_i')\}_{i=1}^n$ can be written as

$$L(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X},\mathbf{T})=\prod_{i=1}^{n}\left\{p(\boldsymbol{\alpha})\phi(y_i;\mathbf{X}_i'\boldsymbol{\beta}^{(1)},\sigma^2)+[1-p(\boldsymbol{\alpha})]\,\phi(y_i;\mathbf{X}_i'\boldsymbol{\beta}^{(2)},\sigma^2)\right\}$$

where $\boldsymbol{\theta}=(\boldsymbol{\alpha}',\boldsymbol{\beta}^{(1)'},\boldsymbol{\beta}^{(2)'},\sigma)'$ is a $(2p+q+4)\times 1$ regression parameter vector. The maximum likelihood estimator of the regression parameter $\hat{\boldsymbol{\theta}}$ can be obtained by solving the score equation:

$$U(\mathbf{Y};\boldsymbol{\theta})=\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X},\mathbf{T})=\mathbf{0}.$$

A commonly used alternative approach is to use the EM algorithm (Redner and Walker, 1984) that incorporates a latent variable ξ_i to indicate whether a subject i comes from the first component as indicated by $\xi_i=1$ (for subject i from the first component) with probability $p(\boldsymbol{\alpha})$, and $\xi_i=0$ (for subject i from the second component) with probability $1-p(\boldsymbol{\alpha})$. Let $\boldsymbol{\xi}=(\xi_1,...,\xi_n)'$. The likelihood based on the pseudo-complete data $\{\mathbf{Y},\boldsymbol{\xi},\mathbf{X},\mathbf{T}\}=\{(Y_i,\xi_i,\mathbf{X}_i',\mathbf{T}_i')\}_{i=1}^n$, as if ξ_i's are observable, can be constructed as:

$$L^c(\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\xi},\mathbf{X},\mathbf{T})=\prod_{i=1}^{n}\{p(\boldsymbol{\alpha})\phi(y_i;\mathbf{X}_i'\boldsymbol{\beta}^{(1)},\sigma^2)\}^{\xi_i}\{[1-p(\boldsymbol{\alpha})]\,\phi(y_i;\mathbf{X}_i'\boldsymbol{\beta}^{(2)},\sigma^2)\}^{1-\xi_i}.$$

It can be seen that the pseudo-complete likelihood L^c can be decomposed into three parts:

$$L^c(\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\xi},\mathbf{X},\mathbf{T})=L_1(\boldsymbol{\alpha})L_2(\boldsymbol{\beta}^{(1)},\sigma)L_3(\boldsymbol{\beta}^{(2)},\sigma),$$

where

$$L_1(\boldsymbol{\alpha}) = \prod_{i=1}^{n} p(\boldsymbol{\alpha})^{\xi_i} \left[1 - p(\boldsymbol{\alpha})\right]^{1-\xi_i},$$

$$L_2(\boldsymbol{\beta}_1, \sigma) = \prod_{i=1}^{n} \phi(y_i; \mathbf{X}_i'\boldsymbol{\beta}^{(1)}, \sigma^2)^{\xi_i},$$

$$L_3(\boldsymbol{\beta}^{(2)}, \sigma) = \prod_{i=1}^{n} \phi(y_i; \mathbf{X}_i'\boldsymbol{\beta}^{(2)}, \sigma^2)^{1-\xi_i}.$$

As shown in the above expressions, L_1 only involves $\boldsymbol{\alpha}$, L_2 only involves $\boldsymbol{\beta}^{(1)}$ and σ, and L_3 only involves $\boldsymbol{\beta}^{(2)}$ and σ. Then, the estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$ and σ can be easily derived by the following EM algorithm.

(a). Assign starting values to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$ and σ.

(b). E-step: Conditional on Y_i and the current values for $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}^{(1)}$, $\hat{\boldsymbol{\beta}}^{(2)}$ and $\hat{\sigma}$, calculate the expected value of ξ, $\hat{\xi}(\hat{\boldsymbol{\theta}})$, by

$$\begin{aligned}
\hat{\xi}_i(\hat{\boldsymbol{\theta}}) &= E(\xi_i | Y_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\sigma}) \\
&= \frac{\phi(y_i; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)}, \hat{\sigma}^2) p_i(\hat{\boldsymbol{\alpha}})}{\phi(y_i; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)}, \hat{\sigma}^2) p_i(\hat{\boldsymbol{\alpha}}) + \phi(y_i; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)}, \hat{\sigma}^2) \left[1 - p_i(\hat{\boldsymbol{\alpha}})\right]}
\end{aligned}$$

Note that $\hat{\xi}_i(\hat{\boldsymbol{\theta}})$ is a function of $\hat{\boldsymbol{\theta}}$. For simplicity of notation, we use $\hat{\xi}_i$ to denote $\hat{\xi}_i(\hat{\boldsymbol{\theta}})$.

(c). M-step: Update $\hat{\boldsymbol{\alpha}}$ by maximizing $L_1(\boldsymbol{\alpha})$ with ξ_i replaced by $\hat{\xi}_i$; Update $\hat{\boldsymbol{\beta}}^{(1)}$ by maximizing $L_2(\boldsymbol{\beta}^{(1)}, \sigma)$ with ξ_i replaced by $\hat{\xi}_i$; Update $\hat{\boldsymbol{\beta}}^{(2)}$ by maximizing $L_3(\boldsymbol{\beta}^{(2)}, \sigma)$ with ξ_i replaced by $\hat{\xi}_i$; Update $\hat{\sigma}$ by maximizing $L_2(\boldsymbol{\beta}^{(1)}, \sigma) L_3(\boldsymbol{\beta}^{(2)}, \sigma)$ with ξ_i replaced by $\hat{\xi}_i$. The score functions and the solutions to the score equations are included in the Appendix B.1.

(d). Repeat (b) and (c) until convergence.

Let $U^*(\mathbf{Y}, \boldsymbol{\xi}; \boldsymbol{\theta}) = \partial \log L^c(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. From the last M-step, it can be seen that the EM estimate $\hat{\boldsymbol{\theta}}$ is the final solution such that $U^*(\mathbf{Y}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\theta}}) = 0$. In addition, using the

method of Louis (1982), it can be shown that

$$U(\mathbf{Y};\boldsymbol{\theta}) = E\{U^*(\mathbf{Y},\boldsymbol{\xi};\boldsymbol{\theta})|\mathbf{Y},\boldsymbol{\theta}\},$$

where $E\{\cdot|\mathbf{Y},\boldsymbol{\theta}\}$ is the conditional expectation taken with respect to the joint distribution of all individuals given $\mathbf{Y}$ and $\boldsymbol{\theta}$. Thus, $U(\mathbf{Y};\hat{\boldsymbol{\theta}}) = E\{U^*(\mathbf{Y},\boldsymbol{\xi};\hat{\boldsymbol{\theta}})|\mathbf{Y},\hat{\boldsymbol{\theta}}\} = U^*(\mathbf{Y},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\theta}}) = 0$. Therefore, the EM solution is actually the solution to $U(\mathbf{Y};\hat{\boldsymbol{\theta}}) = 0$.

It can be shown that under the regularity conditions C1 and C2 in Redner and Walker (1984) and Atienza et al. (2007), $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean zero and covariance matrix $I(\boldsymbol{\theta})^{-1}$. The information of $\boldsymbol{\theta}$, $I(\boldsymbol{\theta})$, can be computed from the second derivative of the likelihood function:

$$I(\boldsymbol{\theta}) = -\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}^2}L(\boldsymbol{\theta}|\mathbf{Y},\mathbf{X},\mathbf{T}) = -\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}}U(\mathbf{Y};\boldsymbol{\theta}).$$

Redner and Walker (1984) further proved that regularity condition C1 is verified for finite mixture models from an exponential family, thus it is satisfied for the two-component normal mixture model defined in this chapter. Furthermore, for a mixture model of normal components with a common variance, the second regularity condition (C2) holds and thus the MLE exists as the global maximizer and is strongly consistent (McLachlan and Peel, 2000). From the asymptotic normality of the maximum likelihood estimator, we have the following theorem (McLachlan and Peel, 2000):

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{I(\boldsymbol{\theta})^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + R_n,$$

where $R_n \xrightarrow{P} 0$. Therefore, $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ is asymptotically equivalent to $\frac{1}{\sqrt{n}}I(\boldsymbol{\theta})^{-1}U(\mathbf{Y};\boldsymbol{\theta}) = \frac{1}{\sqrt{n}}I(\boldsymbol{\theta})^{-1}\sum_{i=1}^{n}U_i(Y_i;\boldsymbol{\theta})$, where $U_i(Y_i;\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\log f(y_i;\boldsymbol{\theta})$.

4.1.3 Goodness-of-fit test statistics for a two component mixture model without random effects

Because a mixture model involves both linear regression components and logistic regression for the component specific means and mixing proportion, respectively, we apply

the principle of cumulative residuals (Su and Wei, 1991, Lin et al., 2002, and Pan and Lin, 2005) and propose test statistics for testing the goodness-of-fit of the linear and logistic regression components separately and jointly . For each of these GOF tests, we can assess both the functional form of a covariate or the link function of the response in each component. In addition, an overall goodness-of-fit test based on the link functions can be used to evaluate the overall fit of the mixture model.

4.1.3.1 A GOF statistic for the linear components

For a two-component mixture regression model, the residual of each component cannot be directly computed. However, since the pseudo-complete likelihood $L^c(\boldsymbol{\theta}|\mathbf{Y}, \xi, \mathbf{X}, \mathbf{T})$ can be clearly separated into 3 parts $L_1(\boldsymbol{\alpha})$, $L_2(\boldsymbol{\beta}^{(1)}, \sigma)$, and $L_3(\boldsymbol{\beta}^{(2)}, \sigma)$, and each corresponding score has expectation zero, we define a pseudo-residual based on the score function from the decomposition of the pseudo-complete likelihood as follows, with the forms of the score functions given in Appendix B.1:

$$e_i^{(1)} = \hat{\xi}_i(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)})$$

$$e_i^{(2)} = (1 - \hat{\xi}_i)(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)})$$

where $e_i^{(1)}$ and $e_i^{(2)}$ are the pseudo-residuals corresponding to the first and second linear component, respectively. Clearly $\sum_{i=1}^n e_i^{(k)} = 0$, because $U(\mathbf{Y}; \hat{\boldsymbol{\theta}}) = U^*(\mathbf{Y}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\theta}}) = 0$.

Apply the principle of Su and Wei (1991), we define the cumulative pseudo-residuals with respect to the covariate values or predicted values for each linear component as:

$$W^{L_k}(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mathbf{X}_i < \mathbf{x}) e_i^{(k)}, k = 1, 2, \text{and}$$

$$W_g^{L_k}(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)} < r) e_i^{(k)}, k = 1, 2,$$

respectively, where $\mathbf{x} = (x_1, ..., x_p)'$.

Define the null hypothesis H_0 as the correct specification of all components in the two-component mixture model. Under H_0, the defined cumulative residuals are

expected to fluctuate around 0, and an unusually large value of $\sup_{\mathbf{x}}|W^{L_k}(\mathbf{x})|$ or $\sup_r|W_g^{L_k}(r)|$ would indicate model misspecification.

4.1.3.2 Null distributions of $W^{L_k}(\mathbf{x})$ and $W_g^{L_k}(r)$

To study the approximate null distribution of $W^{L_k}(\mathbf{x})$ and $W_g^{L_k}(r)$, we use $W^{L_1}(\mathbf{x})$ as an example, and the null distribution of $W^{L_2}(\mathbf{x})$ can be obtained using the same method. The null distribution of $W_g^{L_k}(r)$ is a little more complicated as the indicator function involves $\hat{\boldsymbol{\beta}}^{(k)}$, and will be discussed later. If the mixture model is correctly specified, by the Taylor series expansion, $W^{L_1}(\mathbf{x})$ is asymptotically equivalent to

$$V_n(\mathbf{x}) + \hat{\eta}_1'(\mathbf{x};\boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

where

$$V_n(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x})\left\{\hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\right\},$$

and

$$\hat{\eta}_1(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x})\frac{\partial}{\partial\boldsymbol{\theta}}\left\{\hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\right\}.$$

Because $\hat{\xi}_i(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$, the form of $\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\xi}_i(\boldsymbol{\theta})$ is complicated and the final expression of $\hat{\eta}_1(\mathbf{x};\boldsymbol{\theta})$ is included in Appendix B.2.

Because $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ can be approximated by $\frac{1}{\sqrt{n}}I(\boldsymbol{\theta})^{-1}\sum_{i=1}^{n}U_i(Y_i;\boldsymbol{\theta})$, the test statistic $W^{L_1}(\mathbf{x})$ can be approximated by:

$$W^{L_1}(\mathbf{x}) \approx V_n(\mathbf{x}) + \hat{\eta}_1'(\mathbf{x};\boldsymbol{\theta})\frac{1}{\sqrt{n}}I(\boldsymbol{\theta})^{-1}\sum_{i=1}^{n}U_i(Y_i;\boldsymbol{\theta}).$$

Because the expectations of $\hat{\xi}_i(\boldsymbol{\theta})(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(k)})$ and $U_i(Y_i;\boldsymbol{\theta})$ are zero (see Appendix B.1), the righthand side of the above equation is essentially a sum of n independent zero-mean random variables. By the multivariate central limit theorem, $W^{L_1}(\mathbf{x})$ converges in finite-dimension to a zero-mean Gaussian process. Define

$$\hat{W}^{L_1}(\mathbf{x};\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{I(\mathbf{X}_i \le \mathbf{x})e_i^{(1)} + \hat{\eta}_1'(\mathbf{x};\hat{\boldsymbol{\theta}})I(\hat{\boldsymbol{\theta}})^{-1}U_i(Y_i;\hat{\boldsymbol{\theta}})\right\}G_i,$$

where $(G_1, ..., G_n)$ are independent standard normal variables that are independent of the data. Conditional on the data $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$, it can be seen that the only random components in $\hat{W}^{L_1}(\mathbf{x})$ are the standard normal variables $G_1, ..., G_n$. Similar to Pan and Lin (2005), it can be shown that, using the conditional multiplier central limit theorem (van der Vaart and Wellner, 1996, Theorm 2.9.6), the conditional distribution of $\hat{W}^{L_1}(\mathbf{x}; \hat{\boldsymbol{\theta}})$ given the data $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$ is the same in the limit as the unconditional null distribution of $W^{L_1}(\mathbf{x})$. That is, the distribution of $W^{L_1}(\mathbf{x})$ can be approximated by that of $\hat{W}^{L_1}(\mathbf{x}; \hat{\boldsymbol{\theta}})$.

Therefore, to approximate the null distribution of $W^{L_1}(\mathbf{x})$, we can simulate a number of realizations of $\hat{W}^{L_1}(\mathbf{x}; \hat{\boldsymbol{\theta}})$ by repeatedly generating the normal random samples $(G_1, ..., G_n)$ while fixing the data at their observed values. Similarly, the asymptotic distribution of $W^{L_2}(\mathbf{x})$ can be approximated by that of $\hat{W}^{L_2}(\mathbf{x}; \hat{\boldsymbol{\theta}})$, where

$$\hat{W}^{L_2}(\mathbf{x}; \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(\mathbf{X}_i \leq \mathbf{x}) e_i^{(2)} + \hat{\eta}_2'(\mathbf{x}; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i$$

where

$$\hat{\eta}_2(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{X}_i \leq \mathbf{x}) \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \left[1 - \hat{\xi}_i(\boldsymbol{\theta}) \right] (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(2)}) \right\}.$$

Detailed expression of $\hat{\eta}_2(\mathbf{x}; \boldsymbol{\theta})$ is given in Appendix B.2.

To study the null distribution of $W_g^{L_k}(r)$, we define $\hat{W}_g^{L_k}(r)$ as:

$$\hat{W}_g^{L_k}(r; \hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(\mathbf{X}_i' \hat{\boldsymbol{\beta}}^{(k)} \leq r) e_i^{(k)} + \hat{\eta}_{g,k}'(r; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i.$$

Note that $\hat{W}_g^{L_k}(r; \hat{\boldsymbol{\theta}})$ is essentially $\hat{W}^{L_k}(\mathbf{x}; \hat{\boldsymbol{\theta}})$ with $I(\mathbf{X}_i \leq \mathbf{x})$ replaced by $I(\mathbf{X}_i' \hat{\boldsymbol{\beta}}^{(k)} \leq r)$. Now we establish the weak convergence of $W_g^{L_k}(r)$ and $\hat{W}_g^{L_k}(r)$. Assume that

$$\mathrm{E} \left\{ I(\mathbf{X}_i' \boldsymbol{\gamma}_1 < r_1) \left[\hat{\xi}_i(\boldsymbol{\theta}) (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(k)}) \right] - I(\mathbf{X}_i' \boldsymbol{\gamma}_2 < r_2) \left[\hat{\xi}_i(\boldsymbol{\theta}) (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(k)}) \right] \right\}^2 \longrightarrow 0$$

as $\boldsymbol{\gamma}_1 \to \boldsymbol{\gamma}_2$ and $r_1 \to r_2$ for all $\boldsymbol{\gamma}_2$ sufficiently close to $\boldsymbol{\beta}^{(k)}$. It follows

$$W_g^{L_k}(r) = \tilde{W}_g^{L_k}(r) + o_p(1),$$

where

$$\tilde{W}_g^{L_k}(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mathbf{X}_i' \boldsymbol{\beta}^{(k)} \leq r) e_{ij}^{(k)}.$$

Applying a similar argument that shows the weak convergence of $W^{L_k}(\mathbf{x})$, one can show that $\tilde{W}_g^{L_k}(r)$ converges in finite-dimension to a zero-mean Gaussian process. It then follows that the conditional distribution of $\hat{W}_g^{L_k}(r)$ given the data $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$ is the same in the limit as the unconditional null distribution of $W_g^{L_k}(r)$. Therefore, similar to $W^{L_k}(\mathbf{x})$, we can approximate the null distribution of $W_g^{L_k}(r)$ by simulating a number of realizations of $\hat{W}_g^{L_k}(r; \hat{\boldsymbol{\theta}})$ by repeatedly generating the normal random samples $(G_1, ..., G_n)$ while fixing the data at their observed values.

4.1.3.3 A GOF statistic for the mixing proportion

One of the challenges of checking a mixture model is the evaluation of the logistic regression of the mixing proportion. This is because that the component membership ξ_i is unobservable, and the residual cannot be defined as the difference between the observed and the predicted. Therefore, we used $\hat{\xi}_i$ to replace the observed component membership indicator and define the pseudo-residuals for the logistic regression component based on the score function $U(\mathbf{Y}; \hat{\boldsymbol{\theta}})$ (see Appendix B.1):

$$e_i^P = \hat{\xi}_i - \frac{\mathrm{e}^{\mathbf{T}_i' \hat{\boldsymbol{\alpha}}}}{1 + \mathrm{e}^{\mathbf{T}_i' \hat{\boldsymbol{\alpha}}}}.$$

From this definition, we can see that the pseudo-residual for the logistic regression is actually the difference between the predicted mixing proportion and the posterior probability of a subject from the first component. The posterior probability is the weighted average of the maximum likelihood estimator of the proportion (predicted mixing proportion) and the data. Therefore, the "pseudo-residual" indicates how well the logistic regression describes the mixing proportion, and will be used for the goodness-of-fit test for the logistic regression of mixing proportion. From the property of the EM estimators, we know that the summation of the pseudo-residuals equals zero, $\sum_{i=1}^{n} e_i^P = 0$ (see Appendix B.1). Then we can define the cumulative sum of the pseudo-residual for

the logistic regression $W^P(\mathbf{t})$ with respect to covariate values or predicted values as:

$$W^P(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mathbf{T}_i \leq \mathbf{t}) e_i^P,$$

$$W_g^P(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\mathbf{T}_i' \hat{\boldsymbol{\alpha}} \leq r) e_i^P,$$

where $\mathbf{T}_i$ is the subject-specific covariate vector for subject i and $\mathbf{t} = (t_1, ..., t_q)'$.

4.1.3.4 Null distributions of $W^P(\mathbf{t})$ and $W_g^P(r)$

Because $\sum_{i=1}^{n} e_i^P = 0$ and $E\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{\exp(\mathbf{T}_i'\boldsymbol{\alpha})}{1+\exp(\mathbf{T}_i'\boldsymbol{\alpha})}\right\} = 0$ (Appendix B.1), the null distribution of $W^P(\mathbf{t})$ and $W_g^P(r)$ can be obtained using the same method as for $W^{L_k}(\mathbf{x})$ and $W_g^{L_k}(r)$. Define

$$\hat{W}^P(\mathbf{t}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(\mathbf{T}_i \leq \mathbf{t}) e_i^P + \hat{\eta}_P'(\mathbf{t}; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i, \text{and}$$

$$\hat{W}_g^P(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(\mathbf{T}_i' \hat{\boldsymbol{\alpha}} \leq r) e_i^P + \hat{\eta}_{g,P}'(r; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i,$$

where $(G_1, ..., G_n)$ are independent standard normal variables that are independent of $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$, and

$$\hat{\eta}_P(\mathbf{t}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{T}_i \leq \mathbf{t}) \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \hat{\xi}_i(\boldsymbol{\theta}) - \frac{\exp(\mathbf{T}_i'\boldsymbol{\alpha})}{1 + \exp(\mathbf{T}_i'\boldsymbol{\alpha})} \right\},$$

$$\hat{\eta}_{g,P}(r, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} I(\mathbf{T}_i' \hat{\boldsymbol{\alpha}} \leq r) \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \hat{\xi}_i(\boldsymbol{\theta}) - \frac{\exp(\mathbf{T}_i'\boldsymbol{\alpha})}{1 + \exp(\mathbf{T}_i'\boldsymbol{\alpha})} \right\}.$$

Detailed expressions of $\hat{\eta}_P(\mathbf{t}, \boldsymbol{\theta})$ and $\hat{\eta}_{g,P}(r, \boldsymbol{\theta})$ are given in the Appendix B.3. Similar to the cumulative pseudo-residuals for linear components $W^{L_k}(\mathbf{x})$, it can be shown that the conditional distribution of $\hat{W}^P(\mathbf{t})$ and $\hat{W}_g^P(r)$ given the data $(Y_i, \mathbf{X}_i, \mathbf{T}_i)(i = 1, ..., n)$ is the same in the limit as the unconditional null distribution of $W^P(\mathbf{t})$ and $W_g^P(r)$,

respectively. To approximate the null distribution of $W^P(\mathbf{t})$ or $W_g^P(r)$, we can simulate a number of realizations from $\hat{W}^P(\mathbf{t})$ or $\hat{W}_g^P(r)$, respectively, by repeatedly generating the normal samples $(G_1, ..., G_n)$ while fixing the data at their observed values.

4.1.3.5 The link function GOF tests and the overall GOF test for a mixture regression model

As shown in the previous sections, the null distribution of $W_g^{L_k}(r)$ can be approximated by the conditional distribution of $\hat{W}_g^{L_k}(r)$ given the data $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$. Define a Kolmogorov supremum statistic $S_g^{L_k} \equiv \sup_r |W_g^{L_k}(r)|$. Because $W_g^{L_k}(\cdot)$ fluctuates randomly around mean zero under the null hypothesis, an unusually large value of $s_g^{L_k}$ (observed value of $S_g^{L_k}$) would indicate a poor link function for the kth linear component. The P-value can be approximated by $P(\hat{S}_g^{L_k} \geq s_g^{L_k})$, where $\hat{S}_g^{L_k} = \sup_r |\hat{W}_g^{L_k}(r)|$, and $s_g^{L_k}$ denotes the observed value of $S_g^{L_k}$.

To check the link function in the logistic regression of the mixing proportion, we consider the cumulative pseudo-residual $W_g^P(r)$. Using the same method in the previous section, it can be shown that the null distribution of $W_g^P(r)$ can be approximated by the conditional distribution of $\hat{W}_g^P(r)$ given the data. Therefore, the P-value for testing the goodness-of-fit for the link function of the mixing proportion can be approximated by $Pr(\hat{S}_g^P \geq s_g^P)$, where $\hat{S}_g^P = \sup_r |\hat{W}_g^P(r)|$ and s_g^P is an observed value of S_g^P.

Recall that the ultimate goal of the goodness-of-fit test for a mixture model is to test whether both the linear components and logistic regression for the mixing proportion adequately fit the data. Therefore, based on the afore-defined cumulative residuals using predicted values, we propose a Kolmogorov-type supremum statistic, $S_g^T = S_g^{L_1} + S_g^{L_2} + S_g^P$, for testing the overall goodness-of-fit (GOF) for a mixture model, where $S_g^{L_1} = \sup_r |W_g^{L_1}(r)|$, $S_g^{L_2} = \sup_r |W_g^{L_2}(r)|$, and $S_g^P = \sup_r |W_g^P(r)|$. Because $W_g^{L_k}(\cdot)$ and $W_g^P(\cdot)$ fluctuate randomly around zero under the null hypothesis, an unusually large value of S_g^T would indicate a poor fit of the mixture model. It has been shown that both $W_g^{L_k}(r)$ and $\hat{W}_g^{L_k}(r), k = 1, 2$ are asymptotically equivalent to a mean-zero Gaussian process, and both $W_g^P(r)$ and $\hat{W}_g^P(r)$ are also asymptotically equivalent to a

mean-zero Gaussian process:

$$\hat{W}_g^{L_k}(r) = W_g^{L_k}(r) + o_p(1), k = 1, 2$$

$$\hat{W}_g^{P}(r) = W_g^{P}(r) + o_p(1)$$

In Appendix B.4, we verified that $W_g^{L_1}(r) + W_g^{L_2}(r) + W_g^P(r) = \hat{W}_g^{L_1}(r) + \hat{W}_g^{L_2}(r) + \hat{W}_g^P(r) + o_p(1)$. Define

$$\hat{S}_g^T = \hat{S}_g^{L_1} + \hat{S}_g^{L_2} + \hat{S}_g^P = \sup_r|\hat{W}_g^{L_1}(r)| + \sup_r|\hat{W}_g^{L_2}(r)| + \sup_r|\hat{W}_g^P(r)|.$$

Therefore, under the null hypothesis that both the linear components and logistic regression in a mixture model are correctly specified, the conditional distribution of $\hat{S}_g^T$ given the data is the same in the limit as the unconditional distribution of S_g^T. The P-value $Pr(S_g^T \geq s_g^T)$ can be approximated by $Pr(\hat{S}_g^T \geq s_g^T)$ where s_g^T is an observed value of S_g^T.

Using the same principle, we propose a similar joint test statistic to explore the overall fit of the linear components as $S_g^L = S_g^{L_1} + S_g^{L_2}$. Using the same rationale, we can show that the unconditional distribution of S_g^L is the same in the limit as the conditional distribution of $\hat{S}_g^L = \hat{S}_g^{L_1} + \hat{S}_g^{L_2}$, respectively, given the data. The P-value $Pr(S_g^L \geq s_g^L)$ can be approximated by $Pr(\hat{S}_g^L \geq s_g^L)$ where s_g^L is an observed value of S_g^L.

4.1.3.6 Individual GOF tests for testing the functional form of a covariate

Very often, it is useful to evaluate the goodness-of-fit for the functional form of an individual covariate in the linear regression component or the logistic regression component, e.g., evaluate whether the covariate is linearly or quadratically associated with the kth component specific mean or the mixing proportion. To do this, define $W_j^{L_k}(x)$ as

$$W_j^{L_k}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(X_{ij} < x) e_i^{(k)}, k = 1, 2.$$

where X_{ij} is the jth covariate in the covariate vector $\mathbf{X}_i$. This statistic evaluates the fit of the functional form of covariate X_{ij} in the kth linear regression component, and note that $W_j^{L_k}(x)$ is actually a special case of $W^{L_k}(\mathbf{x})$ with $x_l = \infty$ for all $l \neq k$. Therefore, the null distribution of $W_j^{L_k}(x)$ can be approximated by the $\hat{W}_j^{L_k}(x)$ process which is a special case of $\hat{W}^{L_k}(\mathbf{x})$ with $x_l = \infty$ for all $l \neq k$:

$$\hat{W}_j^{L_k}(x) \quad = \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(X_{ij} \leq x) e_i^{(k)} + \hat{\eta}_k'(x; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i.$$

A Kolmogorov-type supremum statistic

$$S_j^{L_k} \equiv \sup_x |W_j^{L_k}(x)|$$

is used for a goodness-of-fit test, and an unusually large value of $s_j^{L_k}$ (observed value of $S_j^{L_k}$) would indicate a poor functional form of x_j. The P-value, $Pr(S_j^{L_k} \geq s_j^{L_k})$ can be approximated by $Pr(\hat{S}_j^{L_k} \geq s_j^{L_k})$ where $\hat{S}_j^{L_k} = \sup_x |\hat{W}_j^{L_k}(x)|$.

The same method can be used for checking the functional form of a covariate in the logistic regression model by defining a goodness-of-fit test statistic $S_j^P \equiv \sup_r |W_j^P(t)|$ where $W_j^P(t)$ is defined as

$$W_j^P(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(T_{ij} \leq t) e_i^P.$$

where T_{ij} is the jth covariate in the covariate vector $\mathbf{T}_i$. Define $\hat{W}_j^P(t)$ as

$$\hat{W}_j^P(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(T_{ij} \leq t) e_i^P + \hat{\eta}_P'(t; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i,$$

Then we can define a goodness-of-fit test statistic S_j^P for checking the functional form of a fixed covariate in the logistic regression model, $S_j^P \equiv \sup_r |W_j^P(t)|$. The P-value, $Pr(S_j^P \geq s_j^P)$ can be approximated by $Pr(\hat{S}_j^P \geq s_j^P)$ where $\hat{S}_j^P = \sup_r |\hat{W}_j^P(t)|$, and s_j^P is an observed value of S_j^P.

4.2 Two-component mixture models with random effects

In this section, we extend the goodness-of-fit testing strategy in the previous section to a repeated measures setting. Due to the multiple observations for each subject, we considered a two-component mixture model with random effects to account for the within-subject correlation.

4.2.1 Model settings

Let Y_{ij} denote the response of subject i at time point j $(j = 1, ..., J)$ and assume that the conditional density of Y_{ij} given a subject-specific random effect v_i can be written as :

$$f(y_{ij}; v_i, \mu_{ij}^{(1)}, \mu_{ij}^{(2)}, \sigma) = p\phi(y_{ij}; \mu_{ij}^{(1)}, \sigma^2) + (1-p)\phi(y_{ij}; \mu_{ij}^{(2)}, \sigma^2), i = 1, 2, ...n,$$

where $\mu_{ij}^{(k)} = \mathbf{X}_{ij}'\boldsymbol{\beta}^{(k)} + v_i$, $\mathbf{X}_{ij} = (1, X_{1ij}, ..., X_{pij})'$ is a $(p+1)$-variate covariate vector, $\boldsymbol{\beta}^{(k)} = \left(\beta_0^{(k)}, \beta_1^{(k)}, ..., \beta_p^{(k)}\right)'$ is a $(p+1)$-variate regression parameter vector, v_i follows a normal distribution with mean 0 and variance σ_1^2, $\phi(\cdot; \mu_{ij}^{(k)}, \sigma^2)$ denotes a normal density function with mean $\mu_{ij}^{(k)}$ and variance σ^2, $k = 1, 2$, and p is the mixing proportion. For simplicity of notation, we assume that all components share the same set of covariates $\mathbf{X}_{ij}$. Note that this model can be easily extended to cases where two linear components involve different sets of covariates. In this dissertation, we assume that the mixing proportion p only depends on subject-specific covariate information. As a result,

$$\text{logit}\left\{p_i(\mathbf{T}_i; \boldsymbol{\alpha})\right\} = \mathbf{T}_i'\boldsymbol{\alpha},$$

where $\mathbf{T}_i = (1, T_{1i}, ..., T_{qi})'$ is a subject-specific $(q+1) \times 1$ covariate vector, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_q)'$ is a $(q+1)$-variate regression parameter vector. In the remaining part of the dissertation, $p_i(\mathbf{T}_i; \boldsymbol{\alpha})$ is simplified as $p_i(\boldsymbol{\alpha})$.

It is reasonable to assume that the the mixing proportion only depends on subject-specific covariate information. Examples include the heterogeneous longitudinal trajectory of children's aggressive behavior development studied in Wang et al (2005).

Specifically, Wang et al (2005) used the growth mixture models, which can be considered as a special case of mixture models, to assess children's aggressive behavior before, duing, and after an intervention. Three types of growth patterns with different longitudinal trajectories of aggressive behavior have been identified, and each child was classified in one of these growth patterns. Another example is the heterogeneous longitudinal prostate-specific antigen (PSA) trajectories modelled by Lin et al (2002b) that involves the latent class models. Five different trajectory classes were identified and each prostate cancer patient was assumed to be classified into one of the trajectory class.

A few additional assumptions are made to the two-component mixture model with random effects: (i) each individual is independent of each other, (ii)conditional on v_i, the repeated measures $Y_{i1}, ..., Y_{iJ}$ are independent, and (iii) v_i are independently and identically distributed from normal distribution with mean 0 and variance σ_1^2. Based on these assumptions, it can be shown that the marginal density of $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iJ})$, after integrating out the random effect v_i, can be expressed as:

$$f(\mathbf{Y}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \Sigma) = p_i(\boldsymbol{\alpha})\phi^*(\mathbf{Y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(1)}, \Sigma) + \{1 - p_i(\boldsymbol{\alpha})\}\phi^*(\mathbf{Y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(2)}, \Sigma),$$

where $i = 1, 2, ...n$, $\mathbf{X}_i = (\mathbf{X}_{i1}, ..., \mathbf{X}_{iJ})$ is a $(p+1) \times J$ matrix, and $\phi^*(\mathbf{Y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(k)}, \Sigma)$ is a multivariate normal density with mean vector $\mathbf{X}_i'\boldsymbol{\beta}^{(k)}$ and variance-covariance matrix Σ. Σ is a compound symmetric matrix with $\Sigma = \sigma^2 I_n + \sigma_1^2 \mathbf{1}'\mathbf{1}$. This implies that $CORR(Y_{ij}, Y_{ij*}) = \rho = \sigma_1^2/(\sigma^2 + \sigma_1^2)$ for $j \neq j^*$ (Appendix C.1).

4.2.2 Statistical properties of the two-component mixture model with random effects

After integrating out the random effect v_i, the marginal likelihood for this two-component mixture model given the observed data $\{\mathbf{Y}, \mathbf{X}, \mathbf{T}\} = \{(\mathbf{Y}_i, \mathbf{X}_i', \mathbf{T}_i')\}_{i=1}^n$ can be expressed as:

$$L^m(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{T}) = \prod_{i=1}^{n} \left\{ p_i(\boldsymbol{\alpha})\phi^*(\mathbf{Y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(1)}, \Sigma) + [1 - p_i(\boldsymbol{\alpha})]\, \phi^*(\mathbf{Y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(2)}, \Sigma) \right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}^{(1)'}, \boldsymbol{\beta}^{(2)'}, \sigma, \sigma_1)'$ is a regression parameter vector. Then the maximum likelihood estimator of the regression parameter $\boldsymbol{\theta}$ can be obtained by solving the following score equation:

$$U^m(\mathbf{Y}; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L^m(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}, \mathbf{T}) = \mathbf{0}.$$

Similar to the previous section, an alternative approach involving a latent variable ξ_i^m can be used. That is $\xi_i^m = 1$ (for subjects from the first component) with probability $p_i(\boldsymbol{\alpha})$, and $\xi_i^m = 0$ (for subjects from the second component) with probability $1 - p_i(\boldsymbol{\alpha})$. Let $\boldsymbol{\xi}^m = (\xi_1^m, ..., \xi_n^m)'$. Define the pseudo-likelihood for the pseudo-complete data $\{\mathbf{Y}, \boldsymbol{\xi}^m, \mathbf{X}, \mathbf{T}\} = \{(Y_i, \xi_i^m, \mathbf{X}_i', \mathbf{T}_i')\}_{i=1}^n$ as if ξ_i^m's are observable, as:

$$L^{m,c}(\boldsymbol{\theta} | \boldsymbol{\xi}^m, \mathbf{Y}, \mathbf{X}, \mathbf{T}) = \prod_{i=1}^n L_i^{m,c}(\boldsymbol{\theta} | \xi_i^m, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{T}_i),$$

where

$$L_i^{m,c}(\boldsymbol{\theta} | \xi_i^m, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{T}_i) = \{p_i(\boldsymbol{\alpha}) \phi^*(\mathbf{Y}_i; \mathbf{X}_i' \boldsymbol{\beta}^{(1)}, \Sigma)\}^{\xi_i^m} \{(1 - p_i(\boldsymbol{\alpha})) \phi^*(\mathbf{Y}_i; \mathbf{X}_i' \boldsymbol{\beta}^{(2)}, \Sigma)\}^{1 - \xi_i^m}.$$

The pseudo-complete likelihood can be rewritten as

$$L^{m,c}(\boldsymbol{\theta} | \boldsymbol{\xi}^m, \mathbf{Y}, \mathbf{X}, \mathbf{T}) = L_1^m(\boldsymbol{\alpha}) L_2^m(\boldsymbol{\beta}^{(1)}, \sigma, \sigma_1) L_3^m(\boldsymbol{\beta}^{(2)}, \sigma, \sigma_1),$$

where:

$$L_1^m(\boldsymbol{\alpha}) = \prod_{i=1}^n \{p_i(\boldsymbol{\alpha})\}^{\xi_i^m} \{1 - p_i(\boldsymbol{\alpha})\}^{1 - \xi_i^m},$$

$$\begin{aligned} L_2^m(\boldsymbol{\beta}^{(1)}, \sigma, \sigma_1) &= \prod_{i=1}^n \frac{\sigma}{\sqrt{\sigma^2 + J\sigma_1^2}} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^J \\ &\quad \exp \left\{ \xi_i^m \left[\frac{\sigma_1^2 \left(\sum_{j=1}^J (Y_{ij} - \mathbf{X}_{ij}' \boldsymbol{\beta}^{(1)}) \right)^2}{2\sigma^2(\sigma^2 + J\sigma_1^2)} - \frac{\sum_{j=1}^J (Y_{ij} - \mathbf{X}_{ij}' \boldsymbol{\beta}^{(1)})^2}{2\sigma^2} \right] \right\}, \end{aligned}$$

$$
\begin{aligned}
& L_3^m(\boldsymbol{\beta}^{(2)}, \sigma, \sigma_1) \\
= & \prod_{i=1}^{n} \exp\left\{ (1-\xi_i^m) \left[\frac{\sigma_1^2 \left(\sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \right)^2}{2\sigma^2(\sigma^2 + J\sigma_1^2)} - \frac{\sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)})^2}{2\sigma^2} \right] \right\}.
\end{aligned}
$$

Detailed derivation of the pseudo-complete likelihood is given in Appendix C.1. The estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}$, σ, and σ_1 can be easily derived by the EM algorithm. In the E-step, conditional on $\mathbf{Y}_i$ and the current values for $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}^{(1)}$, $\hat{\boldsymbol{\beta}}^{(2)}$, $\hat{\sigma}$, and $\hat{\sigma}_1$, the expected value of ξ_i, $\hat{\xi}_i^m(\hat{\boldsymbol{\theta}})$ can be expressed as:

$$
\begin{aligned}
\hat{\xi}_i^m(\hat{\boldsymbol{\theta}}) &= E(\xi_i^m | \mathbf{Y}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\sigma}, \hat{\sigma}_1) \\
&= \frac{\phi^*(\mathbf{Y}_i; \mathbf{X}'_i\hat{\boldsymbol{\beta}}^{(1)}, \hat{\Sigma}) p_i(\hat{\boldsymbol{\alpha}})}{\phi^*(\mathbf{Y}_i; \mathbf{X}'_i\hat{\boldsymbol{\beta}}^{(1)}, \hat{\Sigma}) p_i(\hat{\boldsymbol{\alpha}}) + \phi^*(\mathbf{Y}_i; \mathbf{X}'_i\hat{\boldsymbol{\beta}}^{(2)}, \hat{\Sigma}) \left[1 - p_i(\hat{\boldsymbol{\alpha}})\right]},
\end{aligned}
$$

where $\hat{\Sigma} = \hat{\sigma}^2 I_n + \hat{\sigma}_1^2 \mathbf{1}'\mathbf{1}$. Note that $\hat{\xi}_i^m(\hat{\boldsymbol{\theta}})$ is a function of $\hat{\boldsymbol{\theta}}$. To ease the notation, we use $\hat{\xi}_i^m$ to denote $\hat{\xi}_i^m(\hat{\boldsymbol{\theta}})$. Then in the M-step, update $\hat{\boldsymbol{\alpha}}$ by maximizing $L_1^m(\boldsymbol{\alpha})$ with ξ_i^m replaced by $\hat{\xi}_i^m$; Update $\hat{\boldsymbol{\beta}}^{(1)}$ by maximizing $L_2^m(\boldsymbol{\beta}^{(1)}, \sigma, \sigma_1)$ with ξ_i^m replaced by $\hat{\xi}_i^m$; Update $\hat{\boldsymbol{\beta}}^{(2)}$ by maximizing $L_3^m(\boldsymbol{\beta}^{(2)}, \sigma, \sigma_1)$ with ξ_i^m replaced by $\hat{\xi}_i^m$; Update $\hat{\sigma}$ and $\hat{\sigma}_1$ by maximizing $L_2^m(\boldsymbol{\beta}^{(1)}, \sigma, \sigma_1) L_3^m(\boldsymbol{\beta}^{(2)}, \sigma, \sigma_1)$ with ξ_i^m replaced by $\hat{\xi}_i^m$. The score functions and the solutions to the score equations are included in Appendix C.2.

Using the same rationale in the previous section, it can be shown that the EM estimators $\hat{\boldsymbol{\theta}}$ are actually the solution to $U^m(\mathbf{Y}; \hat{\boldsymbol{\theta}}) = 0$ (Louis, 1982), and under the regularity conditions C1 and C2 (Redner and Walker, 1984), we have

$$
\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{I^{-1}(\boldsymbol{\theta})}{\sqrt{n}} \sum_{i=1}^{n} U_i^m(\mathbf{Y}_i; \boldsymbol{\theta}) + o_p(1),
$$

where $U_i^m(\mathbf{Y}_i; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{Y}_i; \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \Sigma)$. Therefore, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically equivalent to $\frac{1}{\sqrt{n}} I(\boldsymbol{\theta})^{-1} U^m(\mathbf{Y}; \boldsymbol{\theta}) = \frac{1}{\sqrt{n}} I(\boldsymbol{\theta})^{-1} \sum_{i=1}^{n} U_i^m(\mathbf{Y}_i; \boldsymbol{\theta})$.

4.2.3 Goodness-of-fit test statistics for a two-component mixture model with random effects

4.2.3.1 A GOF test statistic for linear components

Define the pseudo-residual based on the score function from the decomposition of $L^{m,c}$ as follows:

$$e_{ij}^{(m,1)} = \hat{\xi}_i^m (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}),$$

$$e_{ij}^{(m,2)} = (1 - \hat{\xi}_i^m)(Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(2)}),$$

where $e_{ij}^{(m,1)}$ and $e_{ij}^{(m,2)}$ are the pseudo-residuals for subject i at time point j corresponding to the first and second component, respectively. Clearly, $\sum_{i=1}^{n}\sum_{j=1}^{J} e_{ij}^{(m,k)} = 0$, for $k = 1, 2$ (see Appendix C.5). We then define the cumulative pseudo-residuals with respect to the covariate values or the predicted values by:

$$W^{m,L_k}(\mathbf{x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n}\sum_{j=1}^{J} I(\mathbf{X}_{ij} \leq \mathbf{x}) e_{ij}^{(m,k)}, k = 1, 2, \text{and}$$

$$W_g^{m,L_k}(r) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n}\sum_{j=1}^{J} I(\mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(k)} \leq r) e_{ij}^{(m,k)}, k = 1, 2,$$

respectively, where $\mathbf{x} = (x_1, ..., x_p)'$.

The null distribution of $W^{m,L_k}(\mathbf{x})$ and $W_g^{m,L_k}(r)$ can be obtained using the same method as in the previous section. Simply put, for large n, the distribution of $W^{m,L_k}(\mathbf{x})$ can be approximated by that of $\hat{W}^{m,L_k}(\mathbf{x};\hat{\boldsymbol{\theta}})$, where

$$\hat{W}^{m,L_k}(\mathbf{x};\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} I(\mathbf{X}_{ij} \leq \mathbf{x}) e_{ij}^{(m,k)} + \hat{\eta}_k^{m'}(\mathbf{x};\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i^m(\mathbf{Y}_i;\hat{\boldsymbol{\theta}}) \right\} G_i,$$

and $\{G_1, ..., G_n\}$ is a random sample from $N(0, 1)$ and is independent of $\{\mathbf{Y}, \mathbf{X}, \mathbf{T}\}$. In addition, the distribution of $W_g^{m,L_k}(r)$ can be approximated by that of $\hat{W}_g^{m,L_k}(r)$ where

$$\hat{W}_g^{m,L_k}(r;\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} I(\mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(k)} \leq r) e_{ij}^{(m,k)} + \hat{\eta}_{g,k}^{m'}(r;\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i^m(\mathbf{Y}_i;\hat{\boldsymbol{\theta}}) \right\} G_i.$$

Detailed expression of $\hat{\eta}_k^m(\mathbf{x};\boldsymbol{\theta})$ and $\hat{\eta}_{g,k}^m(r;\boldsymbol{\theta})$ is given in Appendix C.3.

4.2.3.2 A GOF test statistic for the mixing proportion

Because the logistic regression model of the mixing proportion only involve subject-level covariates, we can apply the same GOF test statistics as in the previous section and define the pseudo-residual for the logistic regression component as:

$$e_i^{m,P} = \hat{\xi}_i^m - \frac{\mathrm{e}^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}{1+\mathrm{e}^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}.$$

It can be shown that $\sum_{i=1}^n e_i^{m,P} = 0$ (see Appendix C.5). Then define the cumulative sum of the pseudo-residual for the logistic regression with respect to covariate values or predicted values as:

$$W^{m,P}(\mathbf{t}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n I(\mathbf{T}_i < \mathbf{t}) e_i^{m,P}, \text{and}$$

$$W_g^{m,P}(r) = \frac{1}{\sqrt{n}}\sum_{i=1}^n I(\mathbf{T}_i'\hat{\boldsymbol{\alpha}} < r) e_i^{m,P},$$

respectively, where $\mathbf{T}_i$ is the subject-specific covariate vector for subject i in the logistic regression, and $\mathbf{t} = (t_1, ..., t_q)'$.

Define

$$\hat{W}^{m,P}(\mathbf{t}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{ I(\mathbf{T}_i < \mathbf{t}) e_i^{m,P} + \hat{\eta}_P^{m\prime}(\mathbf{t};\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i^m(Y_i;\hat{\boldsymbol{\theta}}) \right\} G_i, \text{and}$$

$$\hat{W}_g^{m,P}(r) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{ I(\mathbf{T}_i'\hat{\boldsymbol{\alpha}} < r) e_i^{m,P} + \hat{\eta}_{g,P}^{m\prime}(r;\hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i^m(Y_i;\hat{\boldsymbol{\theta}}) \right\} G_i,$$

where $(G_1, ..., G_n)$ are independent standard normal variables that are independent of $(\mathbf{Y}, \mathbf{X}, \mathbf{T})$. Detailed expressions of $\hat{\eta}_P^m(\mathbf{t},\boldsymbol{\theta})$ and $\hat{\eta}_{g,P}^m(r,\boldsymbol{\theta})$ are given in Appendix C.4. It can be shown that the conditional distribution of $\hat{W}^{m,P}(\mathbf{t})$ or $\hat{W}_g^{m,P}(r)$ given the data

$\{\mathbf{Y}, \mathbf{X}, \mathbf{T}\}$ is the same in the limit as the unconditional null distribution of $W^{m,P}(\mathbf{t})$ or $W_g^{m,P}(r)$, respectively.

4.2.3.3 The link function GOF tests and the overall GOF test

As shown in the previous sections, we can use $\sup_r|W_g^{m,L_k}(r)|$ to evaluate the fit of the link function of the kth linear component. The P-value can be approximated by $Pr(\hat{S}_g^{m,L_k} \geq s_g^{m,L_k})$ where $\hat{S}_g^{m,L_k} = \sup_r|\hat{W}_g^{m,L_k}(r)|$, and s_g^{m,L_k} is an observed value of S_g^{m,L_k}. In addition, the P-value for testing the goodness-of-fit for the link function of the mixing proportion can be approximated by $Pr(\hat{S}_g^{m,P} \geq s_g^{m,P})$, where $\hat{S}_g^{m,P} = \sup_r|\hat{W}_g^{m,P}(r)|$ and $s_g^{m,P}$ is an observed value of $S_g^{m,P}$.

Then, the overall goodness-of-fit for a mixture model can be evaluated by a joint GOF test statistic based on the link function defined as $S_g^{m,T} = S_g^{m,L_1} + S_g^{m,L_2} + S_g^{m,P} = \sup_r|W_g^{m,L_1}(r)| + \sup_r|W_g^{m,L_2}(r)| + \sup_r|W_g^{m,P}(r)|$. Define

$$\hat{S}_g^{m,T} = \hat{S}_g^{m,L_1} + \hat{S}_g^{m,L_2} + \hat{S}_g^{m,P} = \sup_r|\hat{W}_g^{m,L_1}(r)| + \sup_r|\hat{W}_g^{m,L_2}(r)| + \sup_r|\hat{W}_g^{m,P}(r)|.$$

Using the same rationale as in section 4.1.3, it can be shown that the conditional distribution of $\hat{S}_g^{m,T}$ given the data is the same in the limit as the unconditional null distribution of $S_g^{m,T}$. A similar joint test statistic $S_g^{m,L}$ can be used to assess the overall fit of the linear components where $S_g^{m,L} = S_g^{m,L_1} + S_g^{m,L_2}$. Using the same rationale, we can show that the conditional distribution of $\hat{S}_g^{L} = \hat{S}_g^{L_1} + \hat{S}_g^{L_2}$ given the data is the same in the limit as the unconditional null distribution of $S_g^{m,L}$.

4.2.3.4 Individual GOF tests for testing the functional form of a covariate

A Kolmogorov-type supremum statistic $S_l^{m,L_k} \equiv \sup_x|W_l^{m,L_k}(x)|$ was used to evaluate the goodness-of-fit for the functional form of one of the covariates $x_l, l = 1, 2, .., p$, in the linear regression model, where $W_l^{m,L_k}(x)$ is defined as

$$W_l^{m,L_k}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J} I(X_{lij} < x)e_{ij}^{(m,k)}, k = 1, 2.$$

The null distribution of $W_l^{m,L_k}(x)$ can be approximated by the $\hat{W}_l^{m,L_k}(x)$ process:

$$\hat{W}_l^{m,L_k}(x) \quad = \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} I(X_{lij} \leq x) e_{ij}^{(m,k)} + \hat{\eta}_k^{m'}(x; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i^m(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}) \right\} G_i.$$

The P-value, $Pr(S_l^{m,L_k} \geq s_l^{m,L_k})$ can be approximated by $Pr(\hat{S}_l^{m,L_k} \geq s_l^{m,L_k})$ where $\hat{S}_l^{m,L_k} = \sup_x |\hat{W}_l^{m,L_k}(x)|$.

The same method can be used for checking the functional form of a covariate in the logistic regression model by defining a goodness-of-fit test statistic $S_j^{m,P} \equiv \sup_r |W_j^{m,P}(r)|$ where $W_j^{m,P}(t)$ is defined as

$$W_j^{m,P}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(T_{ij} < t) e_i^{m,P}.$$

Define $\hat{W}_j^{m,P}(t)$ as

$$\hat{W}_j^{m,P}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ I(T_{ij} < t) e_i^{m,P} + \hat{\eta}_P^{m'}(t; \hat{\boldsymbol{\theta}}) I(\hat{\boldsymbol{\theta}})^{-1} U_i^m(Y_i; \hat{\boldsymbol{\theta}}) \right\} G_i,$$

then the P-value, $Pr(S_j^{m,P} \geq s_j^{m,P})$ can be approximated by $Pr(\hat{S}_j^{m,P} \geq s_j^{m,P})$ where $\hat{S}_j^{m,P} = \sup_r |\hat{W}_j^{m,P}(r)|$, and $s_j^{m,P}$ is an observed value of $S_j^{m,P}$.

Chapter 5

Simulation Studies of Goodness-of-fit (GOF) Tests for Mixture Regression Models

In this chapter, we used simulation studies to evaluate the performance of the proposed GOF tests for mixture regression models described in the previous chapter. We started with the univariate two-component normal mixture models, followed by the multivariate two-component normal mixture models with random effects.

5.1 Univariate two-component normal mixture models

In this section, we evaluated the performance of the proposed goodness-of-fit (GOF) test for a two-component mixture model without random effects. Joint and separate GOF tests for the linear regression part and the logistic regression of the mixing proportion were conducted. For each of these tests, we tested either the link function of the response or the functional form of a specific covariate, where appropriate.

5.1.1 Data generation

Data were simulated from a two component normal mixture regression model as follows:

$$f(y;\mu_1,\mu_2,\sigma,p) = p\phi(y,\mu_1,\sigma^2) + (1-p)\phi(y,\mu_2,\sigma^2), \tag{5.1}$$

where $\phi(\cdot,\mu,\sigma^2)$ denotes the normal density function with mean μ and variance σ^2, and p denotes the mixing proportion. We evaluated the power of the proposed GOF tests under a number of chosen scenarios summarized in Table 5.1. To evaluate the empirical size of the type I error of the goodness-of-fit test, the data were simulated based on a null model. To evaluate the power of GOF test, the data were simulated

Table 5.1: The true models for data generation for the evaluation of type I error and empirical power (univariate normal mixture regression models)

Scenarios	Models	Difference from the null model
Null	$\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2}$, $\text{logit}\{p_i(\boldsymbol{\alpha})\} = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 X_{i2}$	None
Model 1	$\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2}$, $\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}$	Logistic component
Model 2a	$\mu_i^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_{i1} + \beta_2^{(1)} X_{i2} + \beta_3^{(1)} X_{i1}^2$ $\mu_i^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} X_{i1} + \beta_2^{(2)} X_{i2}$ $\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2T_{i2}$	First component mean model
Model 2b	$\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{i1}^2$ $\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2X_{i2}$	Component mean models
Model 3a	$\mu_i^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_{i1} + \beta_2^{(1)} X_{i2} + \beta_3^{(1)} X_{i1}^2$ $\mu_i^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} X_{i1} + \beta_2^{(2)} X_{i2}$ $\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}$	First component mean model and logistic component
Model 3b	$\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{i1}^2$ $\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}$	Component mean models and logistic component

k=1,2.

from different alternative models (models 1, 2a, 2b, 3a, and 3b) and analyzed using the assumed null model. Since a mixture model involves both component mean models and logistic regression for the mixing proportion, the null model and different alternative models used were described below:

Null model.

The null model assumes that for each response Y_i, for $i = 1, 2, \cdots, n$, $\mu^{(k)}$ is modeled by $\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2}$, for $k = 1, 2$, and $p(\boldsymbol{\alpha})$ by $\text{logit}\{p_i(\boldsymbol{\alpha})\} = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 T_{i2}$. To evaluate the empirical size of the test, we generated Y_i from the null model. Specifically, we generated Y_i, for the first component ($Y_i^{(1)}$) with probability $p_i(\boldsymbol{\alpha})$ from the normal distribution with mean $\mu_i^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_{i1} + \beta_2^{(1)} X_{i2}$ and variance σ^2, and for the second component ($Y_i^{(2)}$) with probability $1 - p_i(\boldsymbol{\alpha})$ from the normal distribution with mean $\mu_i^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} X_{i1} + \beta_2^{(2)} X_{i2}$ and variance σ^2; we set $X_{i2} = T_{i2}$, therefore the model of $p_i(\boldsymbol{\alpha})$ was essentially $\text{logit}\{p_i(\boldsymbol{\alpha})\} = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 X_{i2}$. Covariates X_{i1}, X_{i2} and T_{i1} were independent uniform random variables generated from $U(2, 5)$, $U(0, 1)$ and $U(2, 5)$, respectively. Because the relative position between the two component mean models could impact the performance of the GOF test (Figure 5.1-5.3),

data were simulated from three different cases where the two component mean models were severely overlapped, slightly overlapped, and non-overlapped, respectively. For case 1, the two component mean models were severely overlapped, and the regression parameters were chosen as: $\beta_0^{(1)}$=7, $\beta_1^{(1)}$=2, $\beta_2^{(1)}$=1, $\beta_0^{(2)}$=20, $\beta_1^{(2)}$=-3.2, $\beta_2^{(2)}$ = 1, α_0= -1.6, α_1= 0.5, and α_2= 0.2, σ = 0.8. For cases 2 and 3, where the two component mean models are slightly overlapped or completely separate, we chose $\beta_0^{(2)}$=15 and 11, respectively, and kept the other parameters the same as those in case 1. Different sample sizes (n) of 200, 300 and 400 were used for each case.

Alternative model 1: Both component mean models are correctly specified but the logistic model for the mixing proportion is misspecified.

The data of the response Y_i was generated from the same model as the null model with the exception that the mixing proportion was generated by

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}.$$

All the remaining parameters in the component mean models were chosen the same values as those in the null model scenario for each case. Then we fitted the assumed null model using the data generated from Model 1 and tested the overall fit of the model as well as the individual link function for the logistic regression and linear component mean models separately. In addition, we also tested the fit of the functional form of T_1 and X_1 in the logistic regression and linear component mean models, respectively.

Alternative models 2a and 2b: The component mean models are misspecified but the logistic model for the mixing proportion is correctly specified.

Under model 2a, Y_i was generated from the same model as the null model, except that the first component mean model $\mu^{(1)}$ was modeled by $\mu_i^{(1)} = \beta_0^{(1)} + \beta_1^{(1)}X_{i1} + \beta_2^{(1)}X_{i2} + \beta_3^{(1)}X_{i1}^2$. The mixing proportion was generated from the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2X_{i2}.$$

We chose $\beta_3^{(1)} = 0.3$, and the remaining parameters in the component mean model

$\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$ and σ were kept the same as those in the null model for each case. As shown in Table 5.1, only the first component mean model was misspecified when we fitted the data to the assumed null model. This scenario allowed us to evaluate the power performance of the proposed GOF tests when only one component mean model is misspecified and the other component mean model and the logistic regression for the mixing proportion are correctly specified.

Under model 2b, Y_i was generated based on the component mean models with the same basic form: $\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{i1}^2$, for $k = 1, 2$, and the mixing proportion was simulated according to the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2X_{i2}.$$

We chose $\beta_3^{(1)} = \beta_3^{(2)} = 0.3$ and kept the remaining parameters $\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$ and σ of the same values as those in the null model scenario for each case. This scenario allowed us to evaluate the power performance of the proposed GOF tests when both component mean models are misspecified and the logistic regression for the mixing proportion is correctly specified.

Alternative models 3a and 3b: Both the component mean models and the logistic model for the mixing proportion are misspecified.

Under model 3a, Y_i was generated from the same model as the null model scenario, except that the first component mean model μ_1 was modeled by $\mu_i^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_{i1} + \beta_2^{(1)} X_{i2} + \beta_3^{(1)} X_{i1}^2$. The mixing proportion was generated from the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}.$$

We chose $\beta_3^{(1)} = 0.3$, and the remaining parameters in the component mean model $\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$ and σ were kept the same as those in the null model for each case. As shown in Table 5.1, both the logistic regression model for the mixing proportion and the first component mean model were misspecified when we fitted the data to the assumed null model. This scenario allowed us to evaluate the power performance of the

proposed GOF tests when both the logistic regression for the mixing proportion and one component mean model were misspecified.

Under model 3b, Y_i was generated based on the component mean models with the same basic form: $\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{i1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{i1}^2$, for $k = 1, 2$, and the mixing proportion was simulated according to the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}.$$

We chose $\beta_3^{(1)} = \beta_3^{(2)} = 0.3$ and kept the remaining parameters $\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$ and σ of the same values as those in the null model for each case. This scenario allowed us to evaluate the performance of the proposed GOF tests when both the two component mean models and the logistic regression for the mixing proportion were misspecified.

A total of 5000 replicates of data was generated for the evaluation of the size of the tests. For each data generation, 1000 realizations (involving $G_i's$) from the null distributions were used in calculating the P-values. For the power assessment, 2000 replicates and 1000 realizations were used.

5.1.2 Simulation results

Figures 5.1-5.3 showed the degree of overlaps in the component specific means in the simulated data from the specified univariate two component mixture models under each case: severely overlapped, slightly overlapped, and non-overlapped.

We first evaluated the overall fit of the model based on the test statistic S_g^T under the null hypothesis that the two-component mixture model is correctly specified. We also evaluated the performance of these test statistics $S_g^{L_1}$, $S_g^{L_2}$ and S_g^P for their use to explore the individual fit of each model component, and S_g^L for its use to evaluate the overall fit in the component mean models. For the former, each test was evaluated under the sub-null hypothesis that the individual model component(s) was (were) correctly specified. We omit the complexity of multiple testing and set the nominal level of each test at 0.05. Results were summarized in Tables 5.2-5.5. Overall, the proposed tests showed reasonable power against model misspecification. When the entire model was

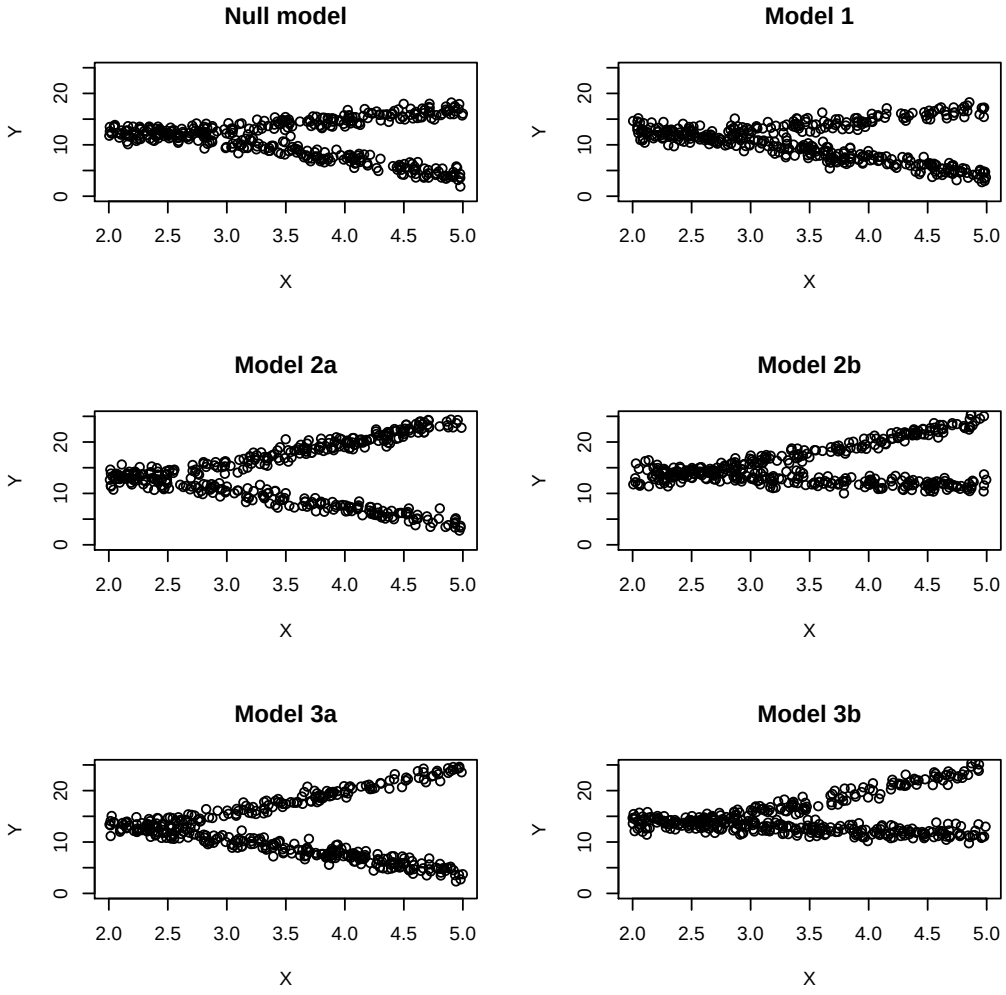

Figure 5.1: Simulated data for a univariate two component mixture model without random effects (case 1).

correctly specified (Table 5.2), the empirical size of the test was close to the nominal level 0.05 for the overall fit and the individual component fit, regardless of the sample size and the degrees of overlap in the component means. When some components in the models were misspecified (Tables 5.3-5.5), the power of the test increased as the sample size increased or the number of misspecified model components increased. As the degree of overlap in the mean models decreased, the power of the test slightly increased. It is interesting to note that when one or more components in the mixture model were misspecified, the size of the test for the correctly specified components was close to the

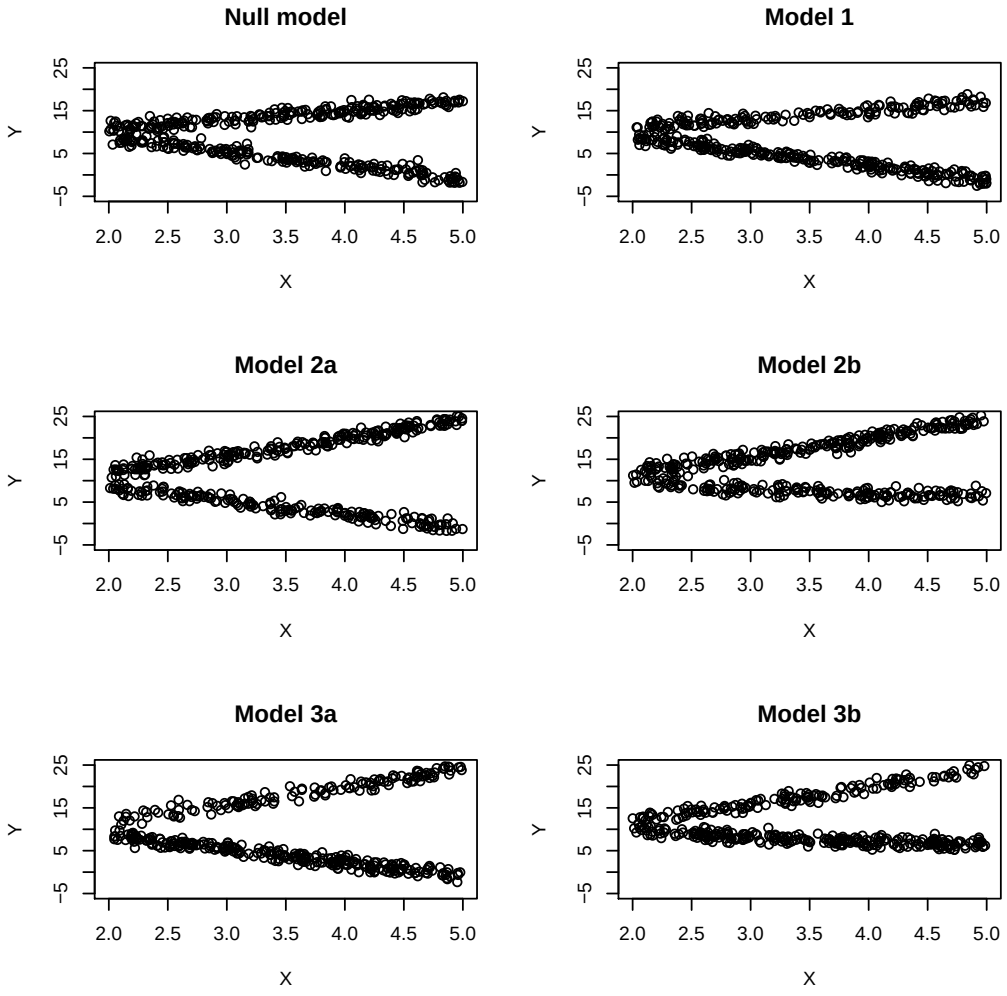

Figure 5.2: Simulated data for a univariate two component mixture model without random effects (case 2).

nominal level of 0.05. We conjecture that it is because the mixture models considered here implicitly assumed that the component responses $Y_i^{(1)}$ and $Y_i^{(2)}$ given the latent component membership were independent and the mixing proportion was independent of the component responses $Y_i^{(1)}$ and $Y_i^{(2)}$. The size of these tests was less affected by the model misspecification.

We next evaluated the performance of the test statistics S^{L_1}, S^{L_2} and S^P and used them to test the functional form of a covariate in the component mean models and logistic regression model for the mixing proportion. Specifically, we test the functional

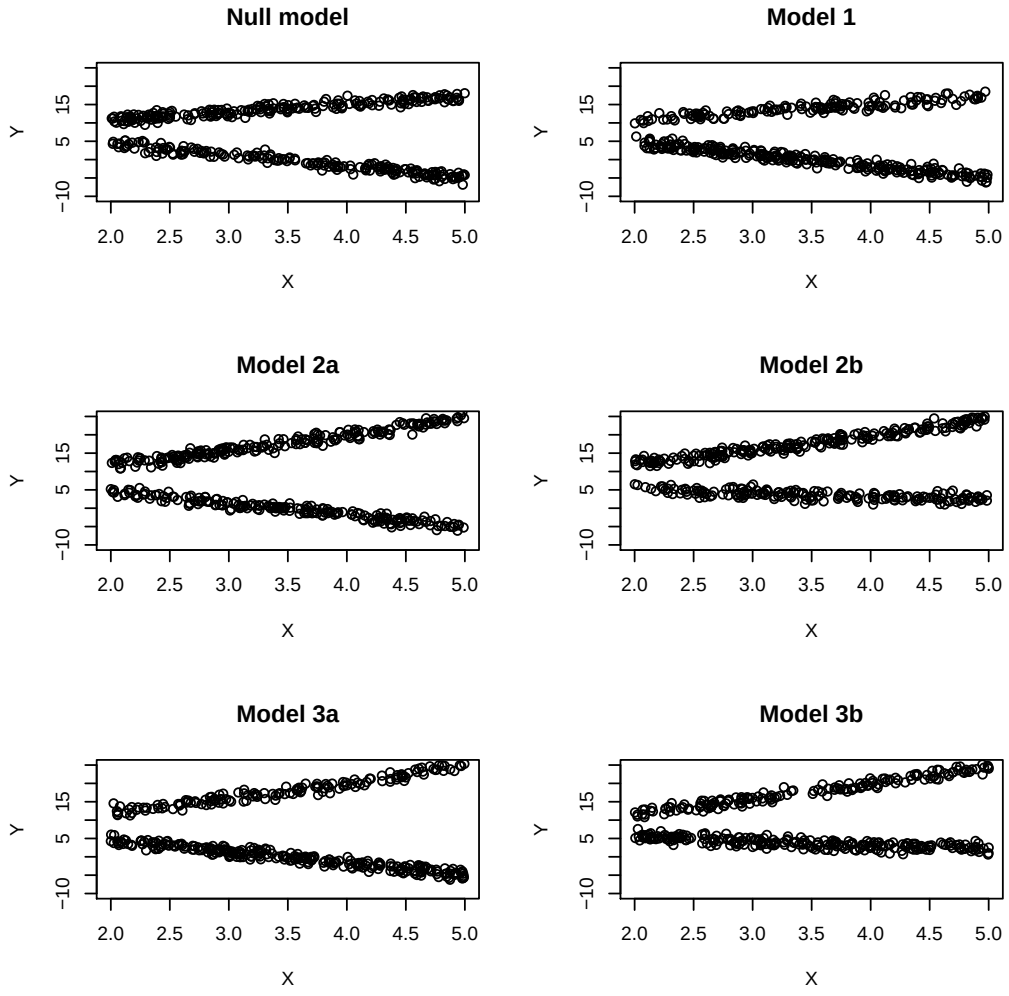

Figure 5.3: Simulated data for a univariate two component mixture model without random effects (case 3).

form of X_1 in the first and second linear component mean models, and T_1 in the logistic regression model for the mixing proportion. Results were summarized in Tables 5.6-5.9. Overall, the proposed tests showed good power and was sensitive to the misspecification of the functional form of a covariate. When the entire model was correctly specified (Table 5.6), the empirical size of the test was close to the nominal level 0.05 for each of the test, regardless of the sample size and the degrees of overlap in the component means. When some components in the models were misspecified, especially the functional forms for covariates were misspecified (Tables 5.7-5.9), the power of the test increased with the sample size and the deviation from the null model. As the degree of

Table 5.2: Empirical size of the link function GOF test for testing the overall fit and individual component fit for the univariate mixture regression model

Model parameter	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean component	Overall fit
Case 1	200	0.049	0.043	0.052	0.042	0.045
	300	0.046	0.049	0.051	0.052	0.051
	400	0.048	0.051	0.053	0.048	0.052
Case 2	200	0.049	0.046	0.051	0.049	0.048
	300	0.053	0.048	0.053	0.051	0.052
	400	0.046	0.047	0.052	0.046	0.053
Case 3	200	0.046	0.045	0.053	0.050	0.052
	300	0.051	0.052	0.048	0.056	0.055
	400	0.052	0.046	0.049	0.047	0.052

Table 5.3: Empirical power of the link function GOF test for testing the overall fit and individual component fit for the univariate mixture regression model (case 1)

Model	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean component	Overall fit
1	200	0.046	0.047	0.249	0.049	0.122
	300	0.047	0.053	0.390	0.049	0.196
	400	0.049	0.054	0.502	0.053	0.247
2a	200	0.433	0.049	0.052	0.329	0.288
	300	0.621	0.055	0.053	0.479	0.417
	400	0.778	0.054	0.054	0.651	0.575
2b	200	0.446	0.252	0.053	0.530	0.458
	300	0.638	0.360	0.048	0.737	0.653
	400	0.776	0.499	0.051	0.874	0.825
3a	200	0.292	0.053	0.265	0.191	0.293
	300	0.467	0.054	0.425	0.289	0.501
	400	0.585	0.052	0.569	0.374	0.651
3b	200	0.311	0.366	0.256	0.509	0.567
	300	0.452	0.500	0.398	0.699	0.791
	400	0.583	0.678	0.516	0.862	0.928

overlap in the mean models decreased, the power of the test increased.

5.2 Two-component mixture models with random effects

In this section, we evaluated the performance of the proposed goodness-of-fit test for a two-component mixture model with random effects.

Table 5.4: Empirical power of the link function GOF test for testing the overall fit and individual component fit for the univariate mixture regression model (case 2)

Model	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean component	Overall fit
1	200	0.049	0.043	0.331	0.041	0.159
	300	0.048	0.053	0.497	0.054	0.245
	400	0.051	0.051	0.612	0.050	0.328
2a	200	0.517	0.052	0.056	0.406	0.357
	300	0.723	0.051	0.058	0.595	0.508
	400	0.846	0.045	0.046	0.722	0.635
2b	200	0.471	0.299	0.050	0.582	0.514
	300	0.675	0.425	0.052	0.788	0.722
	400	0.799	0.522	0.048	0.893	0.839
3a	200	0.380	0.053	0.323	0.233	0.379
	300	0.578	0.044	0.494	0.353	0.604
	400	0.695	0.056	0.627	0.492	0.763
3b	200	0.306	0.379	0.322	0.525	0.636
	300	0.481	0.568	0.477	0.753	0.855
	400	0.616	0.701	0.626	0.882	0.953

5.2.1 Data generation

Data were simulated from a two component mixture regression model with random effects. The principle of the null model and different alternative models for the evaluation of type I error and power, is the same as that in the previous section, and is summarized in Table 5.10. It is assumed that all observations for a subject belong to the same component, and thus the mixing proportion only depends on subject-specific covariates. Therefore, the logistic regression component of the multivariate mixture model is the same as that for the univariate mixture model in the previous section. However, the component mean models are different due to subject-specific random effects. Details were described as follows:

Null model.

The null model assumes that for each response Y_{ij}, for $i = 1, 2, \cdots, n, j = 1, 2, 3$, the mixing proportion of each subject, p is modeled by $\text{logit}\{p_i(\boldsymbol{\alpha})\} = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 X_{i2}$, and given the random effect v_i, $\mu^{(k)}$ is modeled by $\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{ij1} + \beta_2^{(k)} X_{i2} + v_i$, for

Table 5.5: Empirical power of the link function GOF test for testing the overall fit and individual component fit for the univariate mixture regression model (case 3)

Model	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean component	Overall fit
1	200	0.048	0.054	0.326	0.056	0.170
	300	0.057	0.046	0.506	0.047	0.246
	400	0.053	0.052	0.634	0.054	0.332
2a	200	0.516	0.049	0.050	0.404	0.345
	300	0.744	0.048	0.047	0.608	0.516
	400	0.865	0.048	0.052	0.737	0.650
2b	200	0.529	0.324	0.054	0.635	0.560
	300	0.752	0.475	0.053	0.864	0.804
	400	0.864	0.622	0.052	0.952	0.911
3a	200	0.421	0.050	0.332	0.258	0.424
	300	0.573	0.049	0.476	0.368	0.604
	400	0.717	0.052	0.646	0.470	0.772
3b	200	0.382	0.437	0.342	0.620	0.701
	300	0.565	0.604	0.479	0.816	0.892
	400	0.694	0.774	0.624	0.927	0.978

Table 5.6: Empirical size of the functional form GOF tests under the null model for the univariate mixture regression model.

Model parameter	n	Mean component 1	Mean component 2	Mixing proportion
Case 1	200	0.048	0.047	0.054
	300	0.048	0.047	0.051
	400	0.048	0.051	0.054
Case 2	200	0.044	0.044	0.054
	300	0.053	0.054	0.051
	400	0.055	0.046	0.055
Case 3	200	0.044	0.048	0.051
	300	0.052	0.052	0.052
	400	0.052	0.049	0.052

$k = 1, 2$. In this null model scenario and the remaining model scenarios, we simulated the data for v_i from iid normal mean zero and variance σ_1^2. To evaluate the empirical size of the test, we generated Y_{ij} given v_i, for the first component ($Y_{ij}^{(1)}$) with probability $p_i(\boldsymbol{\alpha})$ from the normal distribution with mean $\mu_{ij}^{(1)}$ and variance σ^2, and for the second component ($Y_{ij}^{(2)}$) with probability $1-p_i(\boldsymbol{\alpha})$ from the normal distribution with mean $\mu_{ij}^{(2)}$ and variance σ^2. Covariates X_{i2} and T_{i1} were independent uniform random variables

Table 5.7: Empirical power of the functional form GOF test for the univariate mixture regression models (case 1)

Model	Sample size	Mean component 1	Mean component 2	Mixing proportion
1	200	0.046	0.050	0.371
	300	0.052	0.051	0.521
	400	0.048	0.052	0.613
2a	200	0.447	0.048	0.051
	300	0.638	0.056	0.050
	400	0.792	0.055	0.055
2b	200	0.453	0.334	0.051
	300	0.632	0.493	0.054
	400	0.791	0.651	0.055
3a	200	0.299	0.046	0.376
	300	0.469	0.054	0.548
	400	0.587	0.056	0.670
3b	200	0.315	0.477	0.352
	300	0.458	0.667	0.513
	400	0.608	0.812	0.665

generated from $U(0,1)$ and $U(2,5)$, respectively. Covariate X_{ij1} is a function of time point, is independent of X_{i2} and T_{i1}, and follows a uniform distribution $U(j+1, j+2)$, where j=1, 2, 3. Similarly, data were simulated from three different cases where the two component mean models were severely overlapped, slightly overlapped, and non-overlapped, respectively. The trend of response vs. covariate X_1 is similar as in the univariate mixture models. For case 1, the two component mean models were severely overlapped, and the regression parameters were chosen as: $\beta_0^{(1)}$=7, $\beta_1^{(1)}$=2, $\beta_2^{(1)}$=1, $\beta_0^{(2)}$=20, $\beta_1^{(2)}$=-3.2, $\beta_2^{(2)}$ = 1, α_0= -1.6, α_1= 0.5, and α_2= 0.2, σ = 0.8, σ_1 = 0.6. For cases 2 and 3, where the two component mean models are slightly overlapped or completely separate, we chose $\beta_0^{(2)}$=15 and 11, respectively. The values of the other parameters were kept the same as those in case 1. Different sample sizes (n) of 100, 150 and 200 subjects were used for each case. For each subject, a total of 3 measurements were collected.

Alternative model 1: The component mean models are correctly specified but the logistic model for the mixing proportion is misspecified.

Table 5.8: Empirical power of the functional form GOF test for the univariate mixture regression models (case 2)

Model	Sample size	Mean component 1	Mean component 2	Mixing proportion
1	200	0.041	0.048	0.437
	300	0.047	0.056	0.634
	400	0.045	0.052	0.746
2a	200	0.524	0.046	0.053
	300	0.733	0.051	0.058
	400	0.853	0.053	0.046
2b	200	0.476	0.407	0.056
	300	0.678	0.574	0.047
	400	0.802	0.702	0.055
3a	200	0.406	0.041	0.453
	300	0.569	0.051	0.604
	400	0.702	0.050	0.745
3b	200	0.326	0.520	0.458
	300	0.485	0.729	0.620
	400	0.614	0.843	0.742

Similarly, the data of the response Y_{ij} was generated from the same model as the null model with the exception that the mixing proportion was generated by

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}.$$

All the remaining parameters in the component mean models were chosen the same values as those in the null model for each case. Then we fitted the assumed null model using the data generated from model 1 and performed the same goodness-of-fit tests as in the univariate mixture models.

Alternative models 2a and 2b: The component mean models are misspecified but the logistic model for the mixing proportion is correctly specified.

Under model 2a, Y_{ij} was generated from the same model as the null model, except that the first component mean model $\mu^{(1)}$ was modeled by $\mu_{ij}^{(1)} = \beta_0^{(1)} + \beta_1^{(1)}X_{ij1} + \beta_2^{(1)}X_{i2} + \beta_3^{(1)}X_{ij1}^2 + v_i$. The mixing proportion was simulated from the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2X_{i2}.$$

Table 5.9: Empirical power of the functional form GOF test for the univariate mixture regression models (case 3)

Model	Sample size	Mean component 1	Mean component 2	Mixing proportion
1	200	0.048	0.051	0.462
	300	0.055	0.055	0.627
	400	0.048	0.055	0.754
2a	200	0.534	0.047	0.052
	300	0.749	0.052	0.054
	400	0.873	0.054	0.055
2b	200	0.523	0.428	0.053
	300	0.745	0.639	0.054
	400	0.871	0.770	0.055
3a	200	0.420	0.048	0.466
	300	0.582	0.048	0.610
	400	0.724	0.054	0.747
3b	200	0.394	0.569	0.467
	300	0.584	0.767	0.604
	400	0.702	0.890	0.745

We chose $\beta_3^{(1)} = 0.3$, and the remaining parameters in the component mean model $\beta_0^{(k)}, \beta_1^{(k)}, \beta_2^{(k)}$, σ_1 and σ were kept the same as those in the null model for each case.

Under model 2b, Y_{ij} was generated based on the component mean models with the same basic form: $\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{ij1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{ij1}^2 + v_i$, for $k = 1, 2$, and the mixing proportion was simulated according to the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2X_{i2}.$$

We chose $\beta_3^{(1)} = \beta_3^{(2)} = 0.3$ and kept the remaining parameters $\beta_0^{(k)}$, $\beta_1^{(k)}$, $\beta_2^{(k)}$, σ_1, and σ of the same values as those in the null model for each case.

Then we fitted the assumed null model using the data generated from model 2a or 2b and performed the same goodness-of-fit tests as in the univariate mixture models.

Alternative models 3a and 3b: Both the component mean models and the logistic model for the mixing proportion are misspecified.

Under model 3a or 3b, Y_{ij} was generated from the same model as the model 2a

Table 5.10: The true models for data generation for the evaluation of type I error and empirical power (random effects normal mixture regression models)

Scenarios	Models	Difference from the null model
Null	$\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{ij1} + \beta_2^{(k)} X_{i2} + v_i$, $\text{logit}\{p_i(\boldsymbol{\alpha})\} = \alpha_0 + \alpha_1 T_{i1} + \alpha_2 X_{i2}$	None
Model 1	$\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{ij1} + \beta_2^{(k)} X_{i2} + v_i$, $\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}$	Logistic component
Model 2a	$\mu_{ij}^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_{ij1} + \beta_2^{(1)} X_{i2} + \beta_3^{(1)} X_{ij1}^2 + v_i$ $\mu_{ij}^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} X_{ij1} + \beta_2^{(2)} X_{i2} + v_i$, $\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2T_{i2}$	First component mean model
Model 2b	$\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{ij1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{ij1}^2 + v_i$ $\text{logit}\{p_i(\boldsymbol{\alpha})\} = -1.6 + 0.5T_{i1} + 0.2X_{i2}$	Component mean models
Model 3a	$\mu_{ij}^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_{ij1} + \beta_2^{(1)} X_{i2} + \beta_3^{(1)} X_{ij1}^2 + v_i$ $\mu_{ij}^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} X_{ij1} + \beta_2^{(2)} X_{i2} + v_i$, $\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}$	First component mean model and logistic component
Model 3b	$\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_{ij1} + \beta_2^{(k)} X_{i2} + \beta_3^{(k)} X_{ij1}^2 + v_i$ $\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}$	Component mean models and logistic component

k=1,2.

or 2b, respectively, except that the mixing proportion was simulated from the logistic regression

$$\text{logit}\{p_i(\boldsymbol{\alpha})\} = 3.5 - 3T_{i1} + 0.5T_{i1}^2 + 0.2X_{i2}.$$

Then we fitted the assumed null model using the data generated from model 3a or 3b and performed the same goodness-of-fit tests as in the univariate mixture models.

A total of 5000 replicates of data was generated for the evaluation of the size of the tests. For each data generation, 1000 realizations (involving $G_i's$) from the null distributions were used in calculating the p-values. For the power assessment, 2000 replicates and 1000 realizations were used.

5.2.2 Simulation results

Similar to the previous section, we first evaluated the overall fit of the model based on the test statistic $S_g^{m,T}$ under the null hypothesis that the multivariate two-component mixture model is correctly specified. We also evaluated the performance of these test statistics S_g^{m,L_1}, S_g^{m,L_2} and $S_g^{m,P}$ separately and $S_g^{m,L}$ for their use to explore the individual fit of each model component and the overall fit in the component mean models.

Results were summarized in Tables 5.11-5.14. Similar to the univariate two-component mixture models in the previous section, the proposed tests showed reasonable power against model misspecification and the empirical size of the test was close to the nominal level 0.05 for the overall fit and the individual component fit, regardless of the sample size and the degrees of overlap in the component means.

Table 5.11: Empirical size of the link function GOF tests for testing the overall fit and individual component fit for the mixture regression model with random effects

Model parameter	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean component	Overall fit
Case 1	100	0.049	0.046	0.053	0.049	0.052
	150	0.047	0.049	0.055	0.053	0.052
	200	0.048	0.054	0.052	0.057	0.045
Case 2	100	0.045	0.046	0.053	0.050	0.050
	150	0.053	0.047	0.052	0.053	0.050
	200	0.050	0.051	0.050	0.056	0.051
Case 3	100	0.044	0.046	0.055	0.045	0.044
	150	0.049	0.043	0.053	0.049	0.047
	200	0.048	0.050	0.049	0.055	0.053

Table 5.12: Empirical power of the link function GOF tests for testing the overall fit and individual component fit for the mixture regression model with random effects (case 1)

Model	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean components	Overall fit
1	100	0.048	0.046	0.170	0.052	0.067
	150	0.043	0.055	0.249	0.056	0.091
	200	0.057	0.048	0.321	0.058	0.111
2a	100	0.584	0.044	0.054	0.458	0.426
	150	0.797	0.053	0.056	0.663	0.629
	200	0.904	0.051	0.048	0.796	0.769
2b	100	0.546	0.337	0.042	0.684	0.651
	150	0.772	0.538	0.056	0.890	0.880
	200	0.879	0.671	0.056	0.962	0.950
3a	100	0.427	0.050	0.166	0.279	0.307
	150	0.618	0.042	0.246	0.399	0.466
	200	0.785	0.051	0.308	0.550	0.627
3b	100	0.439	0.488	0.134	0.683	0.689
	150	0.617	0.701	0.242	0.891	0.907
	200	0.724	0.816	0.323	0.958	0.975

Table 5.13: Empirical power of the link function GOF tests for testing the overall fit and individual component fit for the mixture regression model with random effects (case 2)

Model	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean components	Overall fit
1	100	0.043	0.046	0.152	0.051	0.063
	150	0.044	0.043	0.236	0.050	0.076
	200	0.056	0.044	0.293	0.055	0.102
2a	100	0.577	0.041	0.045	0.439	0.420
	150	0.773	0.050	0.057	0.631	0.608
	200	0.877	0.044	0.045	0.770	0.731
2b	100	0.576	0.331	0.049	0.704	0.681
	150	0.779	0.539	0.050	0.898	0.882
	200	0.876	0.642	0.048	0.963	0.953
3a	100	0.435	0.052	0.175	0.278	0.316
	150	0.614	0.042	0.251	0.394	0.457
	200	0.785	0.043	0.324	0.571	0.658
3b	100	0.430	0.460	0.156	0.693	0.712
	150	0.612	0.673	0.223	0.876	0.888
	200	0.767	0.804	0.338	0.970	0.973

We next evaluated the performance of the test statistics S^{m,L_1}, S^{m,L_2}, and $S^{m,P}$ and used them to test the functional form of a covariate in the component mean models and logistic regression model for the mixing proportion. Specifically, we test the functional form of X_1 in the first and second linear component mean model, T_1 in the logistic regression model for the mixing proportion. Results were summarized in Tables 5.15-5.18. Similar to the univariate case, the proposed tests showed good power and was sensitive to the misspecification of the functional form of a covariate. When the entire model was correctly specified (Table 5.15), the empirical size of the test was close to the nominal level 0.05 for each of the test, regardless of the sample size and the degrees of overlap in the component means.

5.3 Summary of the simulation work

Based on the simulation results, the proposed GOF tests all showed good power and were sensitive to the model misspecification, no matter they were used to test the

Table 5.14: Empirical power of the link function GOF tests for testing the overall fit and individual component fit for the mixture regression model with random effects (case 3)

Model	n	Mean component 1	Mean component 2	Mixing proportion	Overall mean components	Overall fit
1	100	0.042	0.047	0.150	0.053	0.054
	150	0.045	0.051	0.233	0.055	0.080
	200	0.046	0.058	0.317	0.052	0.119
2a	100	0.562	0.046	0.051	0.441	0.416
	150	0.785	0.044	0.052	0.630	0.610
	200	0.887	0.041	0.054	0.764	0.734
2b	100	0.589	0.351	0.055	0.728	0.697
	150	0.776	0.534	0.057	0.898	0.887
	200	0.879	0.689	0.056	0.967	0.960
3a	100	0.450	0.055	0.154	0.308	0.331
	150	0.623	0.057	0.240	0.413	0.474
	200	0.757	0.054	0.345	0.500	0.595
3b	100	0.415	0.495	0.152	0.678	0.692
	150	0.639	0.681	0.250	0.883	0.906
	200	0.751	0.819	0.325	0.965	0.976

Table 5.15: Empirical size of the functional form GOF tests under the null model for the mixture regression model with random effects.

Model parameters	n	Mean component 1	Mean component 2	Mixing proportion
Case 1	100	0.044	0.041	0.059
	150	0.049	0.049	0.056
	200	0.048	0.055	0.054
Case 2	100	0.045	0.050	0.053
	150	0.050	0.050	0.052
	200	0.045	0.049	0.056
Case 3	100	0.048	0.047	0.051
	150	0.050	0.044	0.050
	200	0.048	0.049	0.049

overall fit, model component fit or the functional form of a covariate in a specific model component. When the entire model was correctly specified, the empirical size of these tests was close to the nominal level 0.05. When some components in the mixture models were misspecified, the power of the test increased with the sample size and the deviation from the assumed null model. When one or more components in the mixture model were misspecified, the size of the test for the correctly specified components was still

Table 5.16: Empirical power of the functional form GOF tests for mixture regression models with random effects (case 1)

		Mean component 1	Mean component 2	Mixing proportion
1	100	0.045	0.046	0.250
	150	0.047	0.056	0.336
	200	0.055	0.053	0.423
2a	100	0.613	0.046	0.058
	150	0.801	0.050	0.053
	200	0.910	0.050	0.054
2b	100	0.546	0.430	0.046
	150	0.786	0.673	0.052
	200	0.885	0.814	0.049
3a	100	0.428	0.041	0.229
	150	0.629	0.043	0.354
	200	0.792	0.053	0.427
3b	100	0.458	0.620	0.234
	150	0.624	0.815	0.345
	200	0.738	0.905	0.440

Table 5.17: Empirical power of the functional form GOF tests for multivariate mixture regression models with random effects (case 2)

		Mean component 1	Mean component 2	Mixing proportion
1	100	0.054	0.051	0.233
	150	0.043	0.044	0.337
	200	0.053	0.049	0.421
2a	100	0.572	0.049	0.055
	150	0.783	0.049	0.058
	200	0.887	0.048	0.055
2b	100	0.574	0.456	0.048
	150	0.797	0.678	0.043
	200	0.888	0.777	0.056
3a	100	0.438	0.046	0.226
	150	0.630	0.053	0.357
	200	0.783	0.047	0.440
3b	100	0.422	0.609	0.255
	150	0.634	0.796	0.324
	200	0.775	0.910	0.446

Table 5.18: Empirical power of the functional form GOF tests for multivariate mixture regression models with random effects (case 3)

		Mean component 1	Mean component 2	Mixing proportion
1	100	0.046	0.046	0.240
	150	0.047	0.044	0.342
	200	0.052	0.050	0.460
2a	100	0.587	0.042	0.059
	150	0.801	0.043	0.051
	200	0.904	0.048	0.053
2b	100	0.608	0.454	0.051
	150	0.787	0.671	0.053
	200	0.876	0.801	0.052
3a	100	0.449	0.048	0.240
	150	0.628	0.056	0.356
	200	0.778	0.046	0.467
3b	100	0.411	0.613	0.244
	150	0.632	0.812	0.353
	200	0.751	0.925	0.449

close to the nominal level of 0.05. In addition, the link function GOF test for the mixing proportion has a similar performance as the functional form test for the mixing proportion.

Chapter 6

Data Analysis

The proposed prototype goodness-of-fit method was used to analyze the HEART data.

6.1 Goodness-of-fit for univariate mixture regression model

In this section, we used the EBC 8-isoprostane data in the HEART study to demonstrate the use of the proposed GOF tests. Specifically, to demonstrate the use of the GOF tests for the univariate mixture regression models without random effects, we used the data from visit 5, which is the first visit right after the end of the 2008 Olympic Games. To deal with the non-detectables, we used the conventional approach that replaces the non-detectable values with one half detection limit (0.78 pg/ml) and acknowledge the limitation of this method. Extension of this GOF test to deal with non-detectables will be discussed in the next Chapter.

As shown in Figure 1.1 in Chapter 1, the 8-isoprostane data seemed to have a bimodal distribution, which implies that a two-component mixture model may better fit the data than the usual one component regression model involving the same set of covariates, and this is also supported by the evaluation of AIC and BIC values.

With the purpose of studying the relationship between PM2.5 exposure and 8-isoprostane, we fitted a two component normal mixture regression model with 8-isoprostane as the response variable and PM2.5 exposure on lag day 7 as the independent variable, adjusting for the average temperature and relative humidity 24 hours before the clinical visit. We specified the mean component models as:

$$\mu_i^{(k)} = \beta_0^{(k)} + \beta_1^{(k)}\text{PM2.5} + \beta_2^{(k)}\text{Temperature} + \beta_3^{(k)}\text{Humidity}, \text{for } k = 1, 2.$$

For the mixing proportion, we modelled it via the following logistic regression model:

$$\text{logit}\left[p_i(\boldsymbol{\alpha})\right] = \alpha_0 + \alpha_1 \text{BMI}.$$

The estimated regression paramters are summarized in Table 6.1. The mixing proportion slightly decreases with the increase of a subject's BMI with a p-value of 0.062. The mean 8-isoprostane level in the first component is slightly negatively affected by PM2.5, temperature and relative humidity, but none of these effects was statisitcally significant. In the second component, the mean 8-isoprostane level significantly decreases with the increase of PM2.5 with a p-value of 0.003.

Both functional form and link function GOF tests were performed for the proposed univaraite two component mixture regression model and the results were summarized in Table 6.1. The p-value for the overall GOF test is 0.327, which implies a good fit of the proposed mixture model to the data. Individual link function GOF tests also show the good fit of the linear components and logistic regression for the mixing proportion with p-values of 0.463, 0.188, and 0.509, respectively. We also generated the cumulative residual plots in Figure 6.2 for a visual examination of the fit for the link function for both linear components and logistic regression model for the mixing proportion. The red line in the plot is the observed cumulative residuals and black curves are 10 simulated realizations. The p-values pertain to the superum test with 10,000 realizations. Similarly, functional form GOF tests show that the functional forms of PM2.5 exposure in the mean component models and BMI in the regression model for the mixing proportion are good fit of the data with a p-value of 0.304, 0.754 and 0.508, respectively. The corresponding residual plots are shown in Figure 6.1.

6.2 Goodness-of-fit for multivariate mixture regression model with random effects

In this section, the 8-isoprostane data after 2008 Olympic Games were analzyed using a mixture regression models with random effects. Again, undetectable observations were imputed by one half the detection limit (0.78 pg/ml).

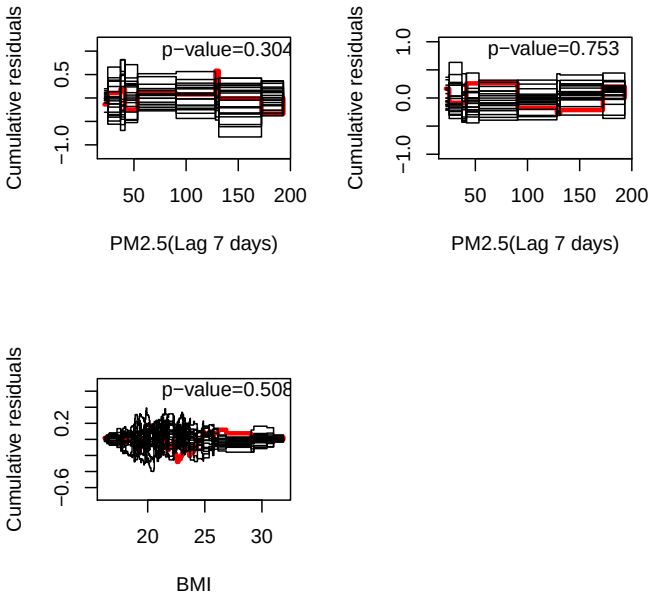

Figure 6.1: Plots of cumulative residual vs. covariate for the 8-isoprostane (visit 5) data example.

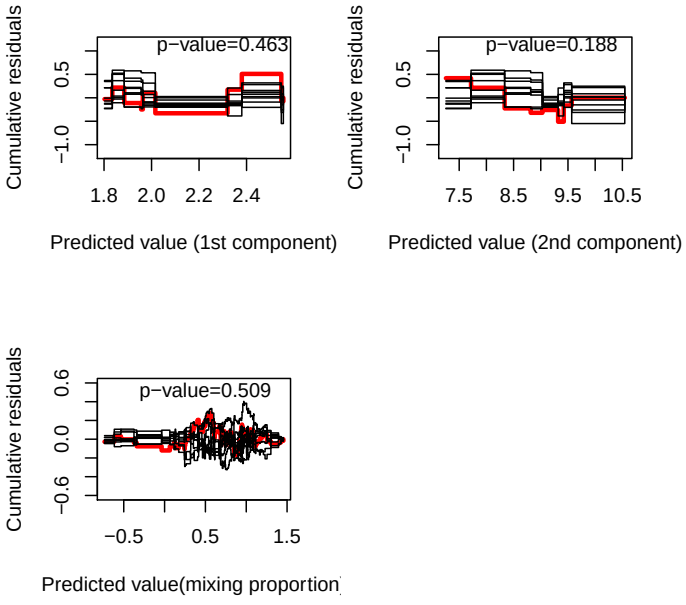

Figure 6.2: Plots of cumulative residual vs. predicted values for the 8-isoprostane (visit 5) data example.

Table 6.1: Data analysis results of the 8-isoprostane data using a univariate mixture regression model (visit 5 only).

Parameter estimation				
	Parameters	Estimate	SE	p value
Mixing proportion	α_0(Intercept)	3.72	1.65	0.026
	α_1(BMI)	-0.14	0.07	0.062
The first linear	$\beta_0^{(1)}$(Intercept)	6.99	4.72	0.141
Component	$\beta_1^{(1)}$(PM2.5)	-0.002	0.004	0.562
	$\beta_2^{(1)}$(Temperature)	-0.19	0.15	0.200
	$\beta_3^{(1)}$(Humidity)	-0.027	0.04	0.551
The second linear	$\beta_0^{(2)}$(Intercept)	18.8	5.48	0.0008
Component	$\beta_1^{(2)}$(PM2.5)	-0.019	0.006	0.003
	$\beta_2^{(2)}$(Temperature)	-0.35	0.197	0.081
	$\beta_3^{(2)}$(Humidity)	-0.042	0.04	0.300
	σ	1.66	0.12	< 0.0001
Link function GOF tests				
	Overall	First component	Second component	Mixing proportion
P value	0.327	0.463	0.188	0.509
Functional form GOF tests				
	Overall	First component (PM2.5 Lag 7)	Second component (PM2.5 Lag 7)	Mixing proportion (BMI)
P value		0.304	0.753	0.508

Based on the relative values of AIC and BIC, a two component mixture model provides a better fit than a one component regression model using the same set of covariates. To illustrate the proposed tests, we fitted a two component normal mixture regression model with random effects using the 8-isoprostane data measured in the post-Olympic Games period (visits 5 and 6) as the response variable and included PM2.5 exposure on lag day 4 as the independent variable, controlling for averaged temperature and relative humidity 24 hours prior to the clinical measurements. The mean component models were specified as:

$$\mu_{ij}^{(k)} = \beta_0^{(k)} + \beta_1^{(k)}\text{PM2.5} + \beta_2^{(k)}\text{Temperature} + \beta_3^{(k)}\text{Humidity}, \text{for } k = 1, 2,$$

and we modelled the mixing proportions via the logistic regression model:

$$\text{logit}\left\{p_i(\boldsymbol{\alpha})\right\} = \alpha_0 + \alpha_1\text{BMI}.$$

Table 6.2: Data analysis results of the 8-isoprostane data using mixture regression models with random effects (Post-Olympic visits only).

Parameter estimation				
	Parameters	Estimate	SE	p value
Mixing proportion p	α_0(Intercept)	19.5	9.59	0.044
	α_1(BMI)	-0.940	0.46	0.041
The first linear	$\beta_0^{(1)}$(Intercept)	2.82	2.22	0.207
Component	$\beta_1^{(1)}$(PM2.5)	0.018	0.006	0.003
	$\beta_2^{(1)}$(Temperature)	0.15	0.12	0.223
	$\beta_3^{(1)}$(Humidity)	-0.045	0.029	0.116
The second linear	$\beta_0^{(2)}$(Intercept)	9.23	1.90	< 0.0001
Component	$\beta_1^{(2)}$(PM2.5)	-0.0061	0.004	0.154
	$\beta_2^{(2)}$(Temperature)	-0.32	0.09	0.0008
	$\beta_3^{(2)}$(Humidity)	0.026	0.026	0.313
	σ	2.88	0.20	< 0.0001
	σ_1	1.66	0.32	< 0.0001
Link function GOF tests				
	Overall	First component	Second component	Mixing proportion
p value	0.977	0.967	0.897	0.912
Functional form GOF tests				
	Overall	First component (PM2.5 Lag 4)	Second component (PM2.5 Lag 4)	Mixing proportion (BMI)
p value		0.745	0.905	0.919

The estimated regression parameters were summarized in Table 6.2. As shown in the table, BMI is a significant predictor of the mixing proportion with p-value=0.041. A higher value of BMI seems to decrease the probability of a subject belonging to the first component with lower 8-isoprostane values. In the first component with lower 8-isoprostane values, the response increases with PM2.5 with a p-value=0.003, but it does not change significantly with temperature and humidity. In the second component with higher 8-isoprostane values, the response decreases with the increase of temperature with a p-value$<$ 0.001, but it does not seem to be associated with PM2.5 and relative humidity.

Both functional form and link function GOF tests were performed with results summarized in Table 6.2. The p-value for the overall GOF test is 0.977, implying a good fit of the proposed mixture model to the data. Individual link function GOF test also shows the good fit of the linear components and logistic regression for the mixing

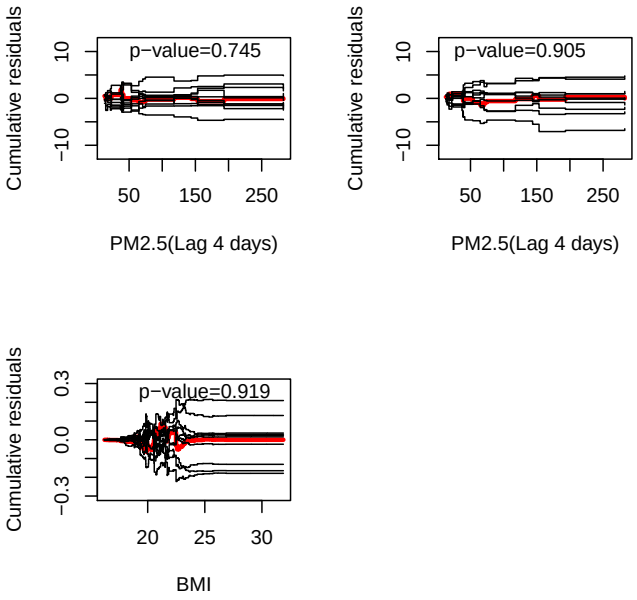

Figure 6.3: Plots of cumulative residual vs. covariate for the 8-isoprostane data example using a multivariate mixture regression model.

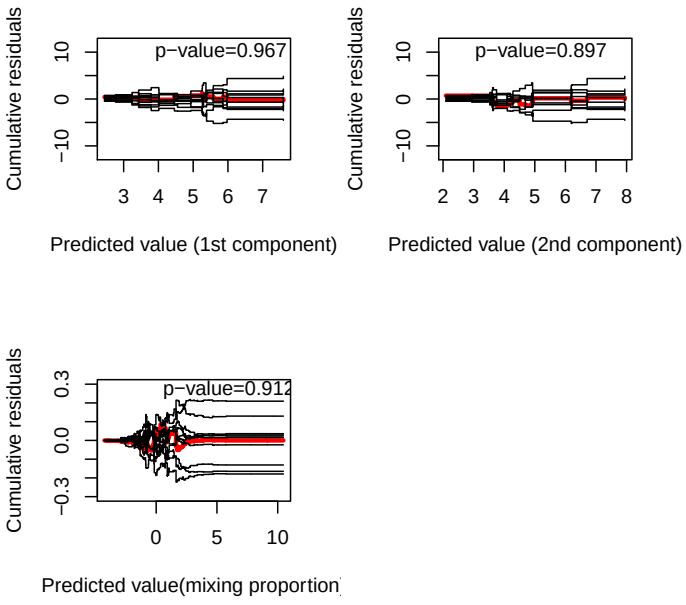

Figure 6.4: Plots of cumulative residual vs. predicted values for the 8-isoprostane data example using a multivariate mixture regression model.

proportion with a p-value of 0.967, 0.897,and 0.912, respectively. We also generated the cumulative residual plots in Figure 6.4 for a visual examination of the fit for the link function for both linear components and logistic regression model for the mixing proportion. Similarly, functional form GOF tests show that the functional forms of PM2.5 exposure in the mean component models and BMI in the regression model for the mixing proportion are also a good fit of the data with a p-value of 0.745, 0.905 and 0.919, respectively. The corresponding residual plots are shown in Figure 6.3.

Chapter 7

Conclusions and Future Research

7.1 Conclusions

In this dissertation, we proposed two statistical methodologies for mixture regression models: one is the mixture regression analysis applied to the analysis of repeatedly measured data with a detection limit, and the other is a prototype of GOF test to evaluate the fit of mixture regression models.

A general framework for finite mixture regression models was proposed to analyze repeatedly measured data with observations under a detection limit. Specifically, random effects were added to the standard mixture regression models to account for the intra-subject correlation for repeated measurements and the non-detectables were treated as left-censored observations in the likelihood-based statistical inference. The regression parameters were estimated by the maximum likelihood method. Using normal mixture models as an example, we demonstrated that the parameter estimators are unbiased. A two-component normal mixture regression model and a three-component normal mixture model were applied to analyze the 8-isoprotane data in the HEART study, and the three-component normal mixture model showed a better fit to the data than the two-component mixture model.

In order to assess the goodness-of-fit of the proposed mixture regression models to the data, we proposed a prototype goodness-of-fit test method based on the principle of cumulative residuals. Starting from a univariate two-component normal mixture model, we defined cumulative pseudo-residuals based on the score functions and then proposed a GOF test accordingly. Extensive simulation studies showed that the proposed GOF tests maintained the type I error rate, and had a reasonable power to detect model deviations. This GOF test method was also extended to a two-component normal

mixture model with random effects, and applied to the analysis of 8-isoprostane data in the HEART study as an illustration.

7.2 Discussions

Data below a detection limit are common in public health and biomedical research, which may cause complications for statistical analyses. For instance, the viral load of an HIV infected patient less than 50 copies/mL would not be detected due to the quantification limit of the measuring device, Roche Amplicor HIV-1 Monitor Assay (Erali and Hillyard, 1999). Another example is the measurement of microbiological assay of serum and urine samples (Joos et al., 1985). The simplest methods are to analyze the data by excluding observations below detection limit or impute the nondetectables by the value of the detection limit or half of the detection limit. However, these methods are known to give biased results, especially when the proportion of observations below detection limit is relatively high. Moreover, in some cases, the population under study is comprised of a mix of heterogeneous subpopulations with unknown membership for each subject. For instance, in (Li et al., 2006), the distribution of HIV RNA data showed a bimodal feature because the HIV-infected patients received diverse background therapies and patients responded to the therapies differently with different (unmeasured/unknown) susceptibility. For data that exhibit multiple features of multimodal distribution and detection limits, such as the EBC 8-isoprostane data in the HEART study, one could apply the proposed random effects mixture regression models, the proposed first methodology in this dissertation, to deal with these issues simultaneously. Random effects are introduced into the model to account for the intra-subject correlation in repeatedly measured data since the observations of the same subject are often correlated to each other. Based on the proposed methodology, one can obtain the mean response in each subpopulation, the overall mean response, and the proportions of each subpopulation (mixing proportions), as if the observations under the detection limit can be exactly measured. The proposed method also allows one to estimate and test the association of the mean response and the mixing proportion with a set of risk covariates.

In practice, to implement the mixture regression modeling technique, one can first use some graphical techniques, such as frequency histograms, to explore whether a mixture model is appropriate for the data to be analyzed and the possible number of mixture components needed. Then one can apply the mixture regression models with different numbers of components to the same dataset. For repeatedly measured data with multiple observation for a subject, random effects are introduced to account for the intra-subject correlation. Based on the values of model fitting criteria (such as AIC, and BIC), one can determine the number of components.

Next, the proposed GOF test method based on cumulative pseudo-residuals can be used to assess how well the fitted model describes the data. Although the best mixture model among a set of candidate mixture models can be identified by using many model checking techniques, the identified best model may not fit the data well enough. The proposed GOF test serves as a final model checking tool regarding whether the identified model fits the data well. This is critical for correctly interpreting public health data because an inappropriate statistical model could lead to incorrect interpretation of the data and misleading conclusions. If the GOF test indicates that the identified mixture model dose not have a good fit of the data, one should try to investigate other families of mixture distribution and/or new covariate information.

7.3 Future Research

7.3.1 Extension of GOF test methodology

In the proposed prototype GOF test, we considered the mixture models that assume all observations for a subject belonging to the same component. In practice, this may not always hold. For instance, teeth from the same patient may not respond equally to the medication that enhances alveolar bone density (Lu et al., 2004); some responded strongly and some did not. Such heterogeneous responses can be analyzed using mixture regression models. To account for intra-subject correlations, the random effects models or the marginal models with the generalized estimating equations approach (Zeger and Liang, 1986) can be used. For the latter, our approach can be straightforwardly applied

when the inference was based on the independence working model assumption (Lu et al., 2004). For the former, our approach may need substantial modifications.

In addition, the proposed prototype GOF test was introduced under the context of two-component normal mixture regression models. Following the same principle of this test, it can be easily extended to mixture regression models with more than two component distributions. The component distributions are also not limited to normal distributions, but can be extended to to other distributions, such as log-normal and gamma distributions. For this extension, one only needs to modify the likelihood function accordingly in its implementation. In addition, the derivation of the cumulative pseudo-residuals and the null distribution for random effects normal mixture regression models is based on the assumption that the intra-subject correlation structure is compound symmetric. When this assumption does not hold, one may need to reconstruct the likelihood, even use some numerical methods to approximate the new likelihood, and modify the form of cumulative pseudo-residuals and the corresponding null distributions accordingly.

7.3.2 Detection limit

In the first part of the dissertation, we proposed a mixture regression model to analyze repeatedly measured data with a detection limit. However the proposed GOF test in the second part of the dissertation has not accounted for the non-detectables yet. In this section, we address this issue using the univariate mixture regression model without random effects. Recall that for data with observations under a detection limit, we studied the statistical inference of mixture regression models in Chapter 3. Now continued from Chapter 3, we study the problem of non-detectables and lay out the fundamental theoretical framework. We defer the further study of the statistical inference and its performance as future work.

Let D_0 denotes the value of detection limit and δ_i denote the indicator for whether an observation is under detection limit or not; that is, $\delta_i = 1$ if $Y_i \geq D_0$, and $\delta_i = 0$ if Y_i is below the detection limit ($Y_i < D_0$). Then the pseudo-complete likelihood can be

constructed as follows:

$$L^{A,c}(\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\xi}^A,\mathbf{X},\mathbf{T}) = \prod_{i=1}^{n}\{p(\boldsymbol{\alpha})\psi(y_i;\mathbf{X}_i'\boldsymbol{\beta}^{(1)},\sigma^2)\}^{\xi_i^A}\{[1-p(\boldsymbol{\alpha})]\,\psi(y_i;\mathbf{X}_i'\boldsymbol{\beta}^{(2)},\sigma^2)\}^{1-\xi_i^A}.$$

where

$$\psi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)},\hat{\sigma}^2) = \phi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)},\hat{\sigma}^2)^{\delta_i}\Phi(D_0;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)},\hat{\sigma}^2)^{1-\delta_i}.$$

Because pseudo-complete likelihood can be clearly separated into 3 parts

$$L^{A,c}(\boldsymbol{\theta}|\boldsymbol{\xi}^A,\mathbf{Y},\mathbf{X},\mathbf{T}_i) = L_1^A(\boldsymbol{\alpha})L_2^A(\boldsymbol{\beta}^{(1)},\sigma)L_3^A(\boldsymbol{\beta}^{(2)},\sigma)$$

where

$$L_1^A(\boldsymbol{\alpha}) = \prod_{i=1}^{n}\{p_i(\boldsymbol{\alpha})\}^{\xi_i^A}\{[1-p_i(\boldsymbol{\alpha})]\}^{1-\xi_i^A},$$

$$L_2^A(\boldsymbol{\beta}^{(1)},\sigma) = \prod_{i=1}^{n}\{\phi(y_i;\boldsymbol{\beta}_1,\sigma^2)^{\delta_i}\Phi(D_0;\mathbf{X}_i'\boldsymbol{\beta}^{(1)},\sigma^2)^{1-\delta_i}\}^{\xi_i^A},$$

$$L_3^A(\boldsymbol{\beta}^{(2)},\sigma) = \prod_{i=1}^{n}\{\phi(y_i;\boldsymbol{\beta}_2,\sigma^2)^{\delta_i}\Phi(D_0;\mathbf{X}_i'\boldsymbol{\beta}^{(2)},\sigma^2)^{1-\delta_i}\}^{1-\xi_i^A},$$

and $U^A(\mathbf{Y},\boldsymbol{\xi}^A;\boldsymbol{\theta}) = \partial \log L^{A,c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be computed as follows:

$$\begin{aligned}\partial \log L^{A,c}(\boldsymbol{\theta})/\partial\boldsymbol{\alpha} &= \partial \log L_1^A(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\\ &= \sum_{i=1}^{n}\left\{\xi_i^A - \frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+e^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\}\mathbf{T}_i,\end{aligned}$$

$$\begin{aligned}\partial \log L^{A,c}(\boldsymbol{\theta})/\partial\boldsymbol{\beta}^{(1)} &= \partial \log L_2^A(\boldsymbol{\beta}^{(1)},\sigma)/\partial\boldsymbol{\beta}^{(1)}\\ &= \sum_{i=1}^{n}\frac{\xi_i^A\delta_i}{\sigma^2}(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\mathbf{X}_i\\ &\quad -\frac{\sqrt{2}\xi_i^A(1-\delta_i)}{\sqrt{\pi\sigma^2}}\exp\left\{-\frac{(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\}\mathbf{X}_i,\end{aligned}$$

$$\begin{aligned}\partial \log L^{A,c}(\boldsymbol{\theta})/\partial \boldsymbol{\beta}^{(2)} &= \partial \log L_3^A(\boldsymbol{\beta}^{(2)},\sigma)/\partial \boldsymbol{\beta}^{(2)} \\ &= \sum_{i=1}^{n} \frac{(1-\xi_i^A)\delta_i}{\sigma^2}(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\mathbf{X}_i \\ &\quad -\frac{\sqrt{2}(1-\xi_i^A)(1-\delta_i)}{\sqrt{\pi\sigma^2}}\exp\left\{-\frac{(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}\mathbf{X}_i,\end{aligned}$$

$$\begin{aligned}\partial \log L^{A,c}(\boldsymbol{\theta})/\partial \sigma &= \sum_{i=1}^{n} -\frac{1}{\sigma} + \sigma^{-3}\delta_i\{\xi_i^A(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2 + (1-\xi_i^A)(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\} \\ &\quad -(1-\delta_i)\sqrt{\frac{2}{\pi}}\frac{\xi_i^A}{\sigma^2}\exp\left\{-\frac{(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\}(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \\ &\quad -(1-\delta_i)\sqrt{\frac{2}{\pi}}\frac{1-\xi_i}{\sigma^2}\exp\left\{-\frac{(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(2)}).\end{aligned}$$

Based on the above score equations, we define pseudo-residuals for linear components as follows:

$$e_i^{A,1} = \hat{\xi}_i^A(\hat{\boldsymbol{\theta}})\delta_i(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)}) - \frac{\sqrt{2}\hat{\xi}_i^A(\hat{\boldsymbol{\theta}})(1-\delta_i)\sigma}{\sqrt{\pi}}\exp\left\{-\frac{(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\}$$

$$e_i^{A,2} = \left[1-\hat{\xi}_i^A(\hat{\boldsymbol{\theta}})\right]\delta_i(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)}) - \frac{\sqrt{2}\left[1-\hat{\xi}_i^A(\hat{\boldsymbol{\theta}})\right](1-\delta_i)\sigma}{\sqrt{\pi}}\exp\left\{-\frac{(D_0-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}$$

where $\hat{\xi}_i^A(\hat{\boldsymbol{\theta}})$ is the posterior probability of component membership, given $\hat{\boldsymbol{\theta}}$, and it can be computed from the EM algorithm:

$$\hat{\xi}_i^A(\hat{\boldsymbol{\theta}}) = \frac{\psi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)},\hat{\sigma}^2)p_i(\hat{\boldsymbol{\alpha}})}{\psi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)},\hat{\sigma}^2)p_i(\hat{\boldsymbol{\alpha}}) + \psi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)},\hat{\sigma}^2)\left[1-p_i(\hat{\boldsymbol{\alpha}})\right]}.$$

Note that when $Y_i \geq D_0$, $\psi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)},\hat{\sigma}^2) = \phi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)},\hat{\sigma}^2)$ for $k=1,2$, and

$$\hat{\xi}_i^A(\hat{\boldsymbol{\theta}}) = \hat{\xi}_i(\hat{\boldsymbol{\theta}}) = \frac{\phi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)},\hat{\sigma}^2)p_i(\hat{\boldsymbol{\alpha}})}{\phi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)},\hat{\sigma}^2)p_i(\hat{\boldsymbol{\alpha}}) + \phi(y_i;\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)},\hat{\sigma}^2)\left[1-p_i(\hat{\boldsymbol{\alpha}})\right]};$$

when $Y_i < D_0$, $\psi(y_i; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)}, \hat{\sigma}^2) = \Phi(D_0; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)}, \hat{\sigma}^2)$, and

$$\hat{\xi}_i^A(\hat{\boldsymbol{\theta}}) = \hat{\xi}_i^U(\hat{\boldsymbol{\theta}}) = \frac{\Phi(D_0; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)}, \hat{\sigma}^2) p_i(\hat{\boldsymbol{\alpha}})}{\Phi(D_0; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)}, \hat{\sigma}^2) p_i(\hat{\boldsymbol{\alpha}}) + \Phi(D_0; \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)}, \hat{\sigma}^2)\left[1 - p_i(\hat{\boldsymbol{\alpha}})\right]}.$$

The pseudo-residuals for the logistic regression of the mixing proportion is defined as:

$$e_i^{A,P} = \hat{\xi}_i^A(\hat{\boldsymbol{\theta}}) - \frac{\mathrm{e}^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}{1 + \mathrm{e}^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}.$$

Based on the score equations, it is easy to prove that $\sum_{i=1}^n e_i^{A,P} = 0$, and $\sum_{i=1}^n e_i^{A,k} = 0$, for $k = 1, 2$. Then the cumulative sum of the pseudo-residuals for each component can be defined using the similar rationale in Chapter 4 and the null distributions of the cumulative pseudo-residuals, i.e., the W and $\hat{W}$ statistics, can be derived accordingly. It can be seen that the expression of these W and $\hat{W}$ depends on this key element $\frac{\partial \hat{\xi}_i^A(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}$. We now show the steps to derive $\hat{\xi}^U(\hat{\theta})$ to conclude this dissertation.

According to the CDF of a normal distribution, $\hat{\xi}^U(\hat{\theta})$ can be expressed as

$$\hat{\xi}_i^U(\hat{\boldsymbol{\theta}}) = \frac{\left[1 + \mathrm{erf}\left(\frac{D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}}{\sqrt{2\sigma^2}}\right)\right] \exp(\mathbf{T}_i'\boldsymbol{\alpha})}{\left[1 + \mathrm{erf}\left(\frac{D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}}{\sqrt{2\sigma^2}}\right)\right] \exp(\mathbf{T}_i'\boldsymbol{\alpha}) + \left[1 + \mathrm{erf}\left(\frac{D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(2)}}{\sqrt{2\sigma^2}}\right)\right]},$$

where

$$\mathrm{erf}\left(\frac{D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(k)}}{\sqrt{2\sigma^2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(k)}}{\sqrt{2\sigma^2}}} e^{-t^2} dt.$$

For simplicity of notation, let

$$C_k = 1 + \mathrm{erf}\left(\frac{D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(k)}}{\sqrt{2\sigma^2}}\right).$$

Then

$$\hat{\xi}_i^U(\hat{\boldsymbol{\theta}}) = \frac{C_1 \exp(\mathbf{T}_i'\boldsymbol{\alpha})}{C_1 \exp(\mathbf{T}_i'\boldsymbol{\alpha}) + C_2}.$$

Then the derivatives of $\hat{\xi}_i^U(\hat{\boldsymbol{\theta}})$ with respective to $\boldsymbol{\theta}$ are as follows:

$$\frac{\partial \hat{\xi}_i^U(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \frac{C_1 C_2 \exp(\mathbf{T}_i'\boldsymbol{\alpha}) \mathbf{T}_i}{\left[\exp(\mathbf{T}_i'\boldsymbol{\alpha}) C_1 + C_2\right]^2},$$

$$\frac{\partial \hat{\xi}_i^U(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(1)}} = -\sqrt{\frac{2}{\pi\sigma^2}} \frac{C_2 \exp\left\{\mathbf{T}_i'\boldsymbol{\alpha} - \frac{\left[D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}\right]^2}{2\sigma^2}\right\} \mathbf{X}_i}{\left[\exp(\mathbf{T}_i'\boldsymbol{\alpha})C_1 + C_2\right]^2},$$

$$\frac{\partial \hat{\xi}_i^U(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(2)}} = \sqrt{\frac{2}{\pi\sigma^2}} \frac{C_1 \exp\left\{\mathbf{T}_i'\boldsymbol{\alpha} - \frac{\left[D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(2)}\right]^2}{2\sigma^2}\right\} \mathbf{X}_i}{\left[\exp(\mathbf{T}_i'\boldsymbol{\alpha})C_1 + C_2\right]^2},$$

$$\begin{aligned} & \frac{\partial \hat{\xi}_i^U(\boldsymbol{\theta})}{\partial \sigma} \\ = \quad & \sqrt{\frac{2}{\pi\sigma^2}} \frac{\exp\left(\mathbf{T}_i'\boldsymbol{\alpha}\right)}{\left[\exp(\mathbf{T}_i'\boldsymbol{\alpha})C_1 + C_2\right]^2} \\ & \left\{ C_1 \exp\left[-\frac{(D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right] - C_2 \exp\left[-\frac{(D_0 - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right] \right\}. \end{aligned}$$

With $\hat{\xi}_i^U(\hat{\boldsymbol{\theta}})$, we can define the $\hat{W}$ statistics as well as the Kolmogorov test statistics like those in Chapter 4. Then we can study the null distribution of these statistics to complete the statistical inference and we defer such task as future work.

Appendix A

Model Properties of Normal Mixture Regression Models

In this section we study the mean, variance, especially the intra-subject correlation under different specifications of random effects using normal mixture regression model as an example. To simplify the notation, we write $p_k(u_i)$ for $p_k(\mathbf{x}_{ij}, u_i)$ and $\mu_k(v_i)$ for $\mu_k(\mathbf{x}_{ij}, v_i)$, but keep in mind that these quantities are modeled as function of $\mathbf{x}_{ij}$.

A.1 A general random effects normal mixture regression model

1). The mean of response Y_{ij}

For a finite normal mixture model, the conditional expectation of Y_{ij} given random effects u_i and v_i is:

$$\mathrm{E}(Y_{ij}|u_i, v_i) = \sum_{k=1}^{K} p_k(u_i)\mu_k(v_i)$$

Denote the pdf of (u_i, v_i) by $f_{u,v}(u, v)$, and note that we assumed (u_i, v_i) follows a multivariate normal distribution. Then the expected mean of the response Y_{ij} is:

$$\begin{aligned}
\mathrm{E}(Y_{ij}) &= \mathrm{E}\left\{\mathrm{E}(Y_{ij}|u_i, v_i)\right\} \\
&= \mathrm{E}\left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i)\right\} \\
&= \iint\limits_{u,v} \left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i)\right\} f_{u,v}(u_i, v_i)du_i dv_i
\end{aligned}$$

If the random effects u and v are independent of each other, the expected mean of Y_{ij} can be simplified as follows:

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \sum_{k=1}^{K} \left\{ \int_u p_k(u_i) f_u(u_i) du_i \int_v \mu_k(v_i) f_v(v_i) dv_i \right\} \\ &= \sum_{k=1}^{K} \left\{ \mathbf{x}'_{ij} \boldsymbol{\beta}_k \int_u p_k(u_i) f_u(u_i) du_i \right\} \end{aligned}$$

Another special case to be considered is that only one random effect is needed: $u = 0$. In this case, the expected mean of Y_{ij} is:

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \sum_{k=1}^{K} \left\{ p_k \int_v \mu_k(v_i) f_v(v_i) dv_i \right\} \\ &= \sum_{k=1}^{K} p_k \mathbf{x}'_{ij} \boldsymbol{\beta}^{(k)} \end{aligned}$$

where p_k is an abbreviation for $p_k(\mathbf{x}_{ij})$, and is not a function of u_i.

It has been shown that the overall mean response $\mathrm{E}(Y_{ij})$ depends on the specifications of the random effects u_i and v_i. If the mixing proportions are not associated with any random variable ($u = 0$), then the overall mean response $\mathrm{E}(Y_{ij})$ is independent of the specifications of the random variable, and it can be computed analytically.

2). The variance of Y_{ij}

$$\begin{aligned} Var(Y_{ij}) &= Var\left\{E(Y_{ij}|u_i, v_i)\right\} + E\left\{Var(Y_{ij}|u_i, v_i)\right\} \\ &= Var\left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i)\right\} + E\left\{\sum_{k=1}^{K} p_k^2(u_i)\sigma^2\right\} \\ &= E\left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i) - E\left[\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i)\right]\right\}^2 + E\left\{\sum_{k=1}^{K} p_k^2(u_i)\sigma^2\right\} \\ &= \iint_{u,v} \left\{ \left[\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i) - E(Y_{ij})\right]^2 + \sum_{k=1}^{K} p_k^2(u_i)\sigma^2 \right\} f_{u,v}(u_i, v_i) du_i dv_i \end{aligned}$$

If there is only one random effect v in the mixture model and the mixing proportion p_k is not a function of any random variable, the variance of Y_{ij} can be simplified as

follows:

$$Var(Y_{ij}) = \int_v \left(\sum_{k=1}^{K} p_k \mu_k(v_i) - E(Y_{ij}) \right)^2 dv_i + \sum_{k=1}^{K} p_k^2(u_i)\sigma^2$$

where p_k is an abbreviation for $p_k(\mathbf{x}_{ij})$.

3). The covariance of Y_{ij} and Y_{ij^*} and within-subject correlation

For simplicity, we denote $p_k(\mathbf{x}_{ij}, u_i)$ by $p_k(u_i)$, $p_k(x_{ij^*}, u_i)$ by $p_k^*(u_i)$ and $\mu_k(\mathbf{x}_{ij}, v_i)$ by $\mu_k(v_i)$, $\mu_k^*(\mathbf{x}_{ij}, v_i)$ by $\mu_k^*(v_i)$. The response for subject i at jth measurement, Y_{ij}, and at the j^*th measurement, Y_{ij^*}, are conditionally independent given random effect u_i and v_i:

$$\begin{aligned}
E(Y_{ij}Y_{ij^*}) &= E\left\{E(Y_{ij}Y_{ij^*}|u_i, v_i)\right\} \\
&= E\{E(Y_{ij}|u_i, v_i)E(Y_{ij^*}|u_i, v_i)\} \\
&= E\left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i) \sum_{k=1}^{K} p_k^*(u_i)\mu_k^*(v_i)\right\} \\
&= \iint_{u,v} \left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i) \sum_{k=1}^{K} p_k^*(u_i)\mu_k^*(v_i)\right\} f_{u,v}(u_i, v_i) du_i dv_i
\end{aligned}$$

$$\begin{aligned}
& COV(Y_{ij}, Y_{ij^*}) \\
= & \iint_{u,v} \left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i) \sum_{k=1}^{K} p_k^*(u_i)\mu_k^*(v_i)\right\} f_{u,v}(u_i, v_i) du_i dv_i \\
- & \iint_{u,v} \left\{\sum_{k=1}^{K} p_k(u_i)\mu_k(v_i)\right\} f_{u,v}(u_i, v_i) du_i dv_i \iint_{u,v} \left\{\sum_{k=1}^{K} p_k^*(u_i)\mu_k^*(v_i)\right\} f_{u,v}(u_i, v_i) du_i dv_i
\end{aligned}$$

Then,

$$CORR(Y_{ij}, Y_{ij^*}) = \frac{COV(Y_{ij}, Y_{ij^*})}{\sqrt{Var(Y_{ij})Var(Y_{ij^*})}} = \frac{E(Y_{ij}Y_{ij^*}) - E(Y_{ij})E(Y_{ij^*})}{\sqrt{Var(Y_{ij})Var(Y_{ij^*})}}$$

A.2 Example: Two-component random effects normal mixture regression models

Consider a two-component random effects normal mixture regression model with model specifications (3.1) and (3.2), the mean and variance of the response Y_{ij} are as follows:

$$E(Y_{ij}) = \iint\limits_{u,v} \{p_1(u_i)\mu_1(v_{1i}) + [1 - p_1(u_i)]\,\mu_2(v_{2i})\}\, f_{u,v}(u_i, \mathbf{v}_i) du_i d\mathbf{v}_i$$

$$\begin{aligned} & Var(Y_{ij}) \\ = & \iint\limits_{u,v} \left\{ [p_1(u_i)\mu_1(v_{1i}) + [1 - p_1(u_i)]\,\mu_2(v_{2i}) - E(Y_{ij})]^2 + \left[p_1^2(u_i) + (1 - p_1(u_i))^2\right] \sigma^2 \right\} \\ & f_{u,v}(u_i, \mathbf{v}_i) du_i d\mathbf{v}_i \end{aligned}$$

The intra-subject correlation is as follows:

$$CORR(Y_{ij}, Y_{ij^*}) = \frac{COV(Y_{ij}, Y_{ij^*})}{\sqrt{Var(Y_{ij}) Var(Y_{ij^*})}}$$

$$\begin{aligned} & COV(Y_{ij}, Y_{ij^*}) \\ = & \iint\limits_{u,v} \{p_1(u_i)\mu_1(v_{1i}) + [1 - p_1(u_i)]\,\mu_2(v_{2i})\}\{p_1^*(u_i)\mu_1^*(v_{1i}) + [1 - p_1^*(u_i)]\,\mu_2^*(v_{2i})\} \\ & f_{u,v}(u_i, \mathbf{v}_i) du_i d\mathbf{v}_i - \iint\limits_{u,v} \{p_1(u_i)\mu_1(v_{1i}) + [1 - p_1(u_i)]\,\mu_2(v_{2i})\} f_{u,v}(u_i, \mathbf{v}_i) du_i d\mathbf{v}_i \\ & \cdot \iint\limits_{u,v} \{p_1^*(u_i)\mu_1^*(v_{1i}) + [1 - p_1^*(u_i)]\,\mu_2^*(v_{2i})\} f_{u,v}(u_i, \mathbf{v}_i) du_i d\mathbf{v}_i \end{aligned}$$

For a reduced two-component mixture model with only one random effect v with model specifications (3.3) and (3.4), the mean and variance of Y_{ij} can be rewritten as:

$$\begin{aligned} E(Y_{ij}) &= \iint\limits_{v} \{p_1\mu_1(v_i) + (1 - p_1)\mu_2(v_i)\} f_v(v_i) dv_i \\ &= p_1 \mathbf{x}_{ij}' \boldsymbol{\beta}^{(1)} + (1 - p_1)\mathbf{x}_{ij}' \boldsymbol{\beta}^{(2)} \end{aligned}$$

$$\begin{aligned} Var(Y_{ij}) &= Var\left\{E(Y_{ij}|v_i)\right\} + E\left\{Var(Y_{ij}|v_i)\right\} \\ &= Var\{p_1\mu_1(v_i) + (1-p_1)\mu_2(v_i)\} + E\{p_1^2\sigma^2 + (1-p_1)^2\sigma^2\} \\ &= \sigma_v^2 + \{p_1^2 + (1-p_1)^2\}\sigma^2 \end{aligned}$$

The intra-subject correlation is as follows for a reduced two-component mixture model with only one random effect v (model specifications (3.3) and (3.4)):

$$\begin{aligned} & E(Y_{ij}Y_{ij^*}) \\ = \; & E\{E(Y_{ij}Y_{ij^*}|v_i)\} \\ = \; & E\{E(Y_{ij}|v_i)E(Y_{ij^*}|v_i)\} \\ = \; & E\left\{\left[p_1(\mathbf{x}'_{ij}\boldsymbol{\beta}^{(1)} + v_i) + (1-p_1)(\mathbf{x}'_{ij}\boldsymbol{\beta}^{(2)} + v_i)\right]\left[p_1^*(\mathbf{x}'_{ij^*}\boldsymbol{\beta}^{(1)} + v_i) + (1-p_1^*)(\mathbf{x}'_{ij^*}\boldsymbol{\beta}^{(2)} + v_i)\right]\right\} \\ = \; & E\left\{\left[p_1\mathbf{x}'_{ij}\boldsymbol{\beta}^{(1)} + (1-p_1)(\mathbf{x}'_{ij}\boldsymbol{\beta}^{(2)} + v_i)\right]\left[p_1^*\mathbf{x}'_{ij^*}\boldsymbol{\beta}^{(1)} + (1-p_1^*)(\mathbf{x}'_{ij^*}\boldsymbol{\beta}^{(2)} + v_i)\right]\right\} \\ = \; & \{p_1\mathbf{x}'_{ij}\boldsymbol{\beta}^{(1)} + (1-p_1)\mathbf{x}'_{ij}\boldsymbol{\beta}^{(2)}\}\{p_1^*\mathbf{x}'_{ij^*}\boldsymbol{\beta}^{(1)} + (1-p_1^*)\mathbf{x}'_{ij^*}\boldsymbol{\beta}^{(2)}\} + \sigma_v^2 \end{aligned}$$

$$COV(Y_{ij}, Y_{ij^*}) = E(Y_{ij}Y_{ij^*}) - E(Y_{ij})E(Y_{ij^*}) = \sigma_v^2$$

Then,

$$CORR(Y_{ij}, Y_{ij^*}) = \frac{\sigma_v^2}{\sqrt{\{\sigma_v^2 + \left[p_1^2 + (1-p_1)^2\right]\sigma^2\}\{\sigma_v^2 + \left[(p_1^*)^2 + (1-p_1^*)^2\right]\sigma^2\}}}$$

It can be seen that the intra-subject correlation depends on the specifications of the random variables u_i and v_i, and in a reduced model with a reduced two-component mixture model with only one random effect v, the intra-subject correlation depends on the variances σ^2 and σ_v^2.

Appendix B

Computation Details for Normal Mixture Regression Models without Random Effects

B.1 Computation of score functions from the pseudo-complete likelihood

As shown in the section 4.1.2, we know that the EM estimators are the solutions to $U(\mathbf{Y};\hat{\boldsymbol{\theta}}) = U^*(\mathbf{Y},\hat{\boldsymbol{\xi}};\hat{\boldsymbol{\theta}}) = 0$ where $U^*(\mathbf{Y},\boldsymbol{\xi};\boldsymbol{\theta}) = \partial \log L^c(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. The log pseudo-complete likelihood is:

$$\begin{aligned}\log\{L^c(\boldsymbol{\theta}|\boldsymbol{\xi},\mathbf{Y},\mathbf{X},\mathbf{T}_i)\} &= \sum_{i=1}^{n}\xi_i\left\{\log\frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+e^{\mathbf{T}_i'\boldsymbol{\alpha}}} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\} \\ &\quad + (1-\xi_i)\left\{\log\frac{1}{1+e^{\mathbf{T}_i'\boldsymbol{\alpha}}} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}.\end{aligned}$$

Because pseudo-complete likelihood can be clearly separated into 3 parts $L^c(\boldsymbol{\theta}|\boldsymbol{\xi},\mathbf{Y},\mathbf{X},\mathbf{T}_i) = L_1(\boldsymbol{\alpha})L_2(\boldsymbol{\beta}^{(1)},\sigma)L_3(\boldsymbol{\beta}^{(2)},\sigma)$ where

$$L_1(\boldsymbol{\alpha}) = \prod_{i=1}^{n}\{p_i(\boldsymbol{\alpha})\}^{\xi_i}\{[1-p_i(\boldsymbol{\alpha})]\}^{1-\xi_i},$$

$$L_2(\boldsymbol{\beta}^{(1)},\sigma) = \prod_{i=1}^{n}\{\phi(y_i;\boldsymbol{\beta}_1,\sigma^2)\}^{\xi_i},$$

$$L_3(\boldsymbol{\beta}^{(2)},\sigma) = \prod_{i=1}^{n}\{\phi(y_i;\boldsymbol{\beta}_2,\sigma^2)\}^{1-\xi_i},$$

$U^*(\mathbf{Y}, \boldsymbol{\xi}; \boldsymbol{\theta}) = \partial \log L^c(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ can be computed as follows:

$$\begin{aligned} \partial \log L^c(\boldsymbol{\theta})/\partial \boldsymbol{\alpha} &= \partial \log L_1(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha} \\ &= \sum_{i=1}^{n} \left\{ \xi_i - \frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1 + e^{\mathbf{T}_i'\boldsymbol{\alpha}}} \right\} \mathbf{T}_i, \end{aligned}$$

$$\begin{aligned} \partial \log L^c(\boldsymbol{\theta})/\partial \boldsymbol{\beta}^{(1)} &= \partial \log L_2(\boldsymbol{\beta}^{(1)}, \sigma)/\partial \boldsymbol{\beta}^{(1)} \\ &= \sum_{i=1}^{n} \frac{\xi_i}{\sigma^2} (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(1)}) \mathbf{X}_i, \end{aligned}$$

$$\begin{aligned} \partial \log L^c(\boldsymbol{\theta})/\partial \boldsymbol{\beta}^{(2)} &= \partial \log L_3(\boldsymbol{\beta}^{(2)}, \sigma)/\partial \boldsymbol{\beta}^{(2)} \\ &= \sum_{i=1}^{n} \frac{(1 - \xi_i)}{\sigma^2} (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(2)}) \mathbf{X}_i, \end{aligned}$$

$$\partial \log L^c(\boldsymbol{\theta})/\partial \sigma = \sum_{i=1}^{n} \left\{ -\frac{1}{\sigma} + \sigma^{-3} \left[\xi_i (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(1)})^2 + (1 - \xi_i)(Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(2)})^2 \right] \right\}.$$

Then, $U(\mathbf{Y}; \hat{\boldsymbol{\theta}}) = U^*(\mathbf{Y}, \hat{\boldsymbol{\xi}}; \hat{\boldsymbol{\theta}})$ can be computed by replacing the $\xi_i, \boldsymbol{\alpha}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \sigma$ in $U^*(\mathbf{Y}, \boldsymbol{\xi}; \boldsymbol{\theta}) = \partial \log L^c(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ by $\hat{\xi}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \hat{\sigma}$, respectively:

$$U(\mathbf{Y}; \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \begin{pmatrix} \left\{ \hat{\xi}_i - \frac{e^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}{1 + e^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}} \right\} \mathbf{T}_i \\ \frac{\hat{\xi}_i}{\hat{\sigma}^2} (Y_i - \mathbf{X}_i' \hat{\boldsymbol{\beta}}^{(1)}) \mathbf{X}_i \\ \frac{(1 - \hat{\xi}_i)}{\hat{\sigma}^2} (Y_i - \mathbf{X}_i' \hat{\boldsymbol{\beta}}^{(2)}) \mathbf{X}_i \\ -\frac{1}{\hat{\sigma}} + \hat{\sigma}^{-3} \{ \hat{\xi}_i (Y_i - \mathbf{X}_i' \hat{\boldsymbol{\beta}}^{(1)})^2 + (1 - \hat{\xi}_i)(Y_i - \mathbf{X}_i' \hat{\boldsymbol{\beta}}^{(2)})^2 \} \end{pmatrix},$$

and

$$U(\mathbf{Y}; \boldsymbol{\theta}) = \sum_{i=1}^{n} U_i(\mathbf{Y}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \begin{pmatrix} \left\{ \xi_i - \frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1 + e^{\mathbf{T}_i'\boldsymbol{\alpha}}} \right\} \mathbf{T}_i \\ \frac{\xi_i}{\sigma^2} (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(1)}) \mathbf{X}_i \\ \frac{(1 - \xi_i)}{\sigma^2} (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(2)}) \mathbf{X}_i \\ -\frac{1}{\sigma} + \sigma^{-3} \{ \xi_i (Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(1)})^2 + (1 - \xi_i)(Y_i - \mathbf{X}_i' \boldsymbol{\beta}^{(2)})^2 \} \end{pmatrix}.$$

From the properties of score functions, we know that $U(\mathbf{Y}; \hat{\boldsymbol{\theta}}) = 0$ and $E\{U_i(\mathbf{Y}_i; \boldsymbol{\theta})\} = 0$.

B.2 The computation of $\hat{\eta}_k(\mathbf{x};\boldsymbol{\theta})$ and $\hat{\eta}_{g,k}(r;\boldsymbol{\theta})$

This section shows the computation of $\hat{\eta}_k(\mathbf{x};\boldsymbol{\theta})$ corresponding the GOF test statistic for the kth component, $W^{L_k}(\mathbf{x})$. We use $\hat{\eta}_1(\mathbf{x};\boldsymbol{\theta})$ as an example:

$$\hat{\eta}_1(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \leq \mathbf{x}) \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \right\}.$$

The derivative of the pseudo-residuals $\hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})$ with respect to each parameter is computed as follows:

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \left\{ \hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \right\} = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}^{(1)}} \left\{ \hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \right\} = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(1)}} - \hat{\xi}_i(\boldsymbol{\theta})\mathbf{X}_i$$

$$\frac{\partial}{\partial \boldsymbol{\beta}^{(2)}} \left\{ \hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \right\} = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(2)}}$$

$$\frac{\partial}{\partial \sigma} \left\{ \hat{\xi}_i(\boldsymbol{\theta})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \right\} = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)}) \frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \sigma}$$

Then, we need to compute the derivatives ($\frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}}$, $\frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(1)}}$, $\frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_2}$, $\frac{\partial \hat{\xi}_i(\boldsymbol{\theta})}{\partial \sigma}$) from the expression of $\hat{\xi}_i(\boldsymbol{\theta})$, which can be further simplified as:

$$\begin{aligned}
\hat{\xi}_i(\boldsymbol{\theta}) &= \frac{\exp(\mathbf{T}_i'\boldsymbol{\alpha})\exp\left\{-\frac{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\}}{\exp(\mathbf{T}_i'\boldsymbol{\alpha})\exp\left\{-\frac{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\} + \exp\left\{-\frac{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}} \\
&= 1 - \frac{\exp\left\{-\frac{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}}{\exp\left\{\mathbf{T}_i'\boldsymbol{\alpha} - \frac{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\} + \exp\left\{-\frac{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}}
\end{aligned}$$

For simplicity of notation, let $A_1 = \exp\left\{-\frac{(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2}{2\sigma^2}\right\}$, and $A_2 = \exp\left\{-\frac{(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2}{2\sigma^2}\right\}$. Then,

$$\hat{\xi}_i(\boldsymbol{\theta}) = 1 - \frac{A_2}{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2}$$

and the derivatives of $\hat{\xi}_i(\boldsymbol{\theta})$ are computed as follows:

$$\frac{\partial\hat{\xi}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}} = \frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})\mathbf{T}_i}{\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2}$$

$$\begin{aligned}\frac{\partial\hat{\xi}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^{(1)}} &= -A_2\frac{-A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha})\frac{(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(1)})}{\sigma^2}\mathbf{X}_i}{\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2} \\ &= \frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\mathbf{X}_i}{\sigma^2\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial\hat{\xi}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^{(2)}} &= A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha})\frac{-A_2\frac{(Y_i-\mathbf{X}_i'\boldsymbol{\beta}^{(2)})}{\sigma^2}\mathbf{X}_i}{\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2} \\ &= -\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\mathbf{X}_i}{\sigma^2\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial\hat{\xi}_i(\boldsymbol{\theta})}{\partial\sigma} &= \frac{A_2\left\{\exp(\mathbf{T}_i'\boldsymbol{\alpha})A_1(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2\sigma^{-3} + A_2(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\sigma^{-3}\right\}}{\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2} \\ &\quad - \frac{A_2(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\sigma^{-3}}{\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}} \\ &= \frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})\{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2 - (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\}}{\sigma^3\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2}\end{aligned}$$

Therefore the expression of $\hat{\eta}_1(\mathbf{x};\boldsymbol{\theta})$ is:

$$\hat{\eta}_1(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \le \mathbf{x})\,(\eta_\alpha, \eta_{\beta_1}, \eta_{\beta_2}, \eta_\sigma)$$

where

$$\eta_\alpha = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})\mathbf{T}_i}{\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\}^2}$$

$$\eta_{\beta_1} = -\hat{\xi}_i(\boldsymbol{\theta})\mathbf{X}_i + (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\mathbf{X}_i}{\sigma^2\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

$$\eta_{\beta_2} = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\mathbf{X}_i}{\sigma^2\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

$$\eta_{\sigma} = (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})\{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2 - (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\}}{\sigma^3\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

Then, $\hat{\eta}_1(\mathbf{x};\hat{\boldsymbol{\theta}})$ can be obtained by replacing $\boldsymbol{\theta}$, $\hat{\xi}_i(\boldsymbol{\theta})$, A_1 and A_2 with $\hat{\boldsymbol{\theta}}$, $\hat{\xi}_i(\hat{\boldsymbol{\theta}})$, $\hat{A}_1$ and $\hat{A}_2$, where

$$\hat{A}_1 = \exp\left\{-\frac{(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(1)})^2}{2\hat{\sigma}^2}\right\}$$

$$\hat{A}_2 = \exp\left\{-\frac{(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(2)})^2}{2\hat{\sigma}^2}\right\}$$

For the second component, $\hat{\eta}_2(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n I(\mathbf{X}_i \le \mathbf{x})\frac{\partial}{\partial\boldsymbol{\theta}}\left\{\left[1 - \hat{\xi}_i(\boldsymbol{\theta})\right](Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\right\}$. Then, the expression of $\hat{\eta}_2(\mathbf{x};\boldsymbol{\theta})$ is:

$$\hat{\eta}_2(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n I(\mathbf{X}_i \le \mathbf{x})\left(\eta_\alpha, \eta_{\beta_1}, \eta_{\beta_2}, \eta_\sigma\right)$$

where

$$\eta_{\alpha} = -(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})\mathbf{T}_i}{\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

$$\eta_{\beta_1} = -(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\mathbf{X}_i}{\sigma^2\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

$$\eta_{\beta_2} = -\left[1 - \hat{\xi}_i(\boldsymbol{\theta})\right]\mathbf{X}_i - (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\mathbf{X}_i}{\sigma^2\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

$$\eta_{\sigma} = -(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\frac{A_1A_2\exp(\mathbf{T}_i'\boldsymbol{\alpha})\{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2 - (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\}}{\sigma^3\left\{A_1\exp(\mathbf{T}_i'\boldsymbol{\alpha}) + A_2\right\}^2}$$

Similarly, $\hat{\eta}_2(\mathbf{x};\hat{\boldsymbol{\theta}})$ can be obtained by replacing $\boldsymbol{\theta}$, $\hat{\xi}_i(\boldsymbol{\theta})$, A_1 and A_2 with $\hat{\boldsymbol{\theta}}$, $\hat{\xi}_i(\hat{\boldsymbol{\theta}})$, $\hat{A}_1$ and $\hat{A}_2$.

In addition, $\hat{\eta}_{g,k}(r;\hat{\boldsymbol{\theta}})$ can be obtained from $\hat{\eta}_k(\mathbf{x};\hat{\boldsymbol{\theta}})$ by replacing $I(\mathbf{X}_i \le \mathbf{x})$ with $I(\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)} \le r)$.

B.3 The computation of $\hat{\eta}_P(\mathbf{t};\boldsymbol{\theta})$ and $\hat{\eta}_{g,P}(r;\boldsymbol{\theta})$

The computation of $\hat{\eta}_P(\mathbf{t};\boldsymbol{\theta})$ in the GOF test statistic for the logistic regression of a mixture model is given in this section.

$$\begin{aligned}\hat{\eta}_P(\mathbf{t},\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{T}_i \le \mathbf{t})\frac{\partial}{\partial\boldsymbol{\theta}}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(\mathbf{T}_i \le \mathbf{t})\left(\eta_\alpha^P, \eta_{\beta_1}^P, \eta_{\beta_2}^P, \eta_\sigma^P,\right)\end{aligned}$$

where

$$\begin{aligned}\eta_\alpha^P &= \frac{\partial}{\partial\boldsymbol{\alpha}}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\} \\ &= \frac{\partial\hat{\xi}_i(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}} - \frac{\partial}{\partial\boldsymbol{\alpha}}\frac{1}{1+\mathrm{e}^{-\mathbf{T}_i'\boldsymbol{\alpha}}} \\ &= \mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}\mathbf{T}_i\left\{\frac{A_1A_2}{(A_1\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}+A_2)^2} - \frac{1}{(1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}})^2}\right\}\end{aligned}$$

$$\begin{aligned}\eta_{\beta_1}^P &= \frac{\partial}{\partial\boldsymbol{\beta}^{(1)}}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\} \\ &= \frac{A_1A_2\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})\mathbf{X}_i}{\sigma^2(A_1\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}+A_2)^2}\end{aligned}$$

$$\begin{aligned}\eta_{\beta_2}^P &= \frac{\partial}{\partial\boldsymbol{\beta}^{(2)}}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\} \\ &= -\frac{A_1A_2\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})\mathbf{X}_i}{\sigma^2(A_1\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}+A_2)^2}\end{aligned}$$

$$\begin{aligned}\eta_\sigma^P &= \frac{\partial}{\partial\sigma}\left\{\hat{\xi}_i(\boldsymbol{\theta}) - \frac{\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\} \\ &= \frac{A_1A_2\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}\left\{(Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(1)})^2 - (Y_i - \mathbf{X}_i'\boldsymbol{\beta}^{(2)})^2\right\}}{\sigma^3(A_1\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}+A_2)^2}\end{aligned}$$

Similarly, $\hat{\eta}_P(\mathbf{t};\hat{\boldsymbol{\theta}})$ can be obtained by replacing $\boldsymbol{\theta}$, $\hat{\xi}_i(\boldsymbol{\theta})$, A_1 and A_2 with $\hat{\boldsymbol{\theta}}$, $\hat{\xi}_i(\hat{\boldsymbol{\theta}})$, $\hat{A}_1$ and $\hat{A}_2$.

In addition, $\hat{\eta}_{g,P}(r;\hat{\boldsymbol{\theta}})$ can be obtained from $\hat{\eta}_P(\mathbf{t};\hat{\boldsymbol{\theta}})$ by replacing $I(\mathbf{T}_i \leq \mathbf{t})$ with $I(\mathbf{T}_i'\hat{\boldsymbol{\alpha}} \leq r)$.

B.4 Proof of $W_g^{L_1}(r)+W_g^{L_2}(r)+W_g^P(r)=\hat{W}_g^{L_1}(r)+\hat{W}_g^{L_2}(r)+\hat{W}_g^P(r)+o_p(1)$

It has been shown earlier that $W_g^{L_k}(r)$ and $\hat{W}_g^{L_k}(r), k=1,2$ have the same asymptotic distribution, $W_g^P(r)$ and $\hat{W}_g^P(r)$ have the same asymptotic distribution. Let $W_g^{L_k}(r)=\hat{W}_g^{L_k}(r)+\hat{R}_{L_k,n}=\sum_{i=1}^n \hat{W}_{g,i}^{L_k}(r)+\hat{R}_{L_k,n}, k=1,2$, and $W_g^P(r)=\hat{W}_g^P(r)+\hat{R}_{P,n}=\sum_{i=1}^n \hat{W}_{g,i}^P(r)+\hat{R}_{P,n}$, where $\hat{W}_g^{L_k}(r)=\sum_{i=1}^n \hat{W}_{g,i}^{L_k}(r)$, $\hat{W}_g^P(r)=\sum_{i=1}^n \hat{W}_{g,i}^P(r)$, $\hat{R}_{L_k,n}=\sum_{i=1}^n \hat{R}_{L_k,i}(x)=o_p(1)$, and $\hat{R}_{P,n}=\sum_{i=1}^n \hat{R}_{P,i}(r)=o_p(1)$.

It can be shown that the expectation of $\hat{W}_{g,i}^{L_k}(r)$ and $\hat{W}_{g,i}^P(r)$ are both zero because the expectation of the pseudo-residual is zero and the expectation of the score function is zero.

$$\begin{aligned} E\left\{\hat{W}_{g,i}^{L_k}(r)\right\} &= \frac{1}{\sqrt{n}} I(\mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(k)} \le r) E(e_i^{(k)}) + \hat{\eta}_k'(r;\hat{\boldsymbol{\theta}}) I^{-1}(\hat{\boldsymbol{\theta}}) E\left\{U_i(Y_i;\hat{\boldsymbol{\theta}})\right\} \\ &= 0 \end{aligned}$$

$$\begin{aligned} E\left\{\hat{W}_{g,i}^{P}(r)\right\} &= \frac{1}{\sqrt{n}} I(\mathbf{T}_i'\hat{\boldsymbol{\alpha}} \le r) E(e_i^{P}) + \hat{\eta}_P'(r;\hat{\boldsymbol{\theta}}) I^{-1}(\hat{\boldsymbol{\theta}}) E\left\{U_i(Y_i;\hat{\boldsymbol{\theta}})\right\} \\ &= 0 \end{aligned}$$

In order to prove $W_g^{L_1}(r)+W_g^{L_2}(r)+W_g^P(r)=\hat{W}_g^{L_1}(r)+\hat{W}_g^{L_2}(r)+\hat{W}_g^P(r)+o_p(1)$, we need to show that $Cov\{W_g^{L_k}(r), W_g^P(r)\} = Cov\{\hat{W}^{L_k}(r), \hat{W}^P(r)\}, k=1,2$ and $Cov\{W_g^{L_1}(r), W_g^{L_2}(r)\} = Cov\{\hat{W}^{L_k}(r), \hat{W}^{L_2}(r)\}$. We will then show that

$$Cov\{W_g^{L_k}(r), W_g^P(r)\} = Cov\{\hat{W}^{L_k}(r), \hat{W}^P(r)\}$$

for $k=1,2$. Therefore, $Cov\{W_g^{L_1}(r), W_g^{L_2}(r)\} = Cov\{\hat{W}^{L_k}(r), \hat{W}^{L_2}(r)\}$ is also valid.

Given r,

$$\begin{aligned} Cov\{W_g^{L_k}(r), W_g^P(r)\} &= Cov\{\hat{W}^{L_k}(r)+\hat{R}_{L_k,n}(r), \hat{W}^P(r)+\hat{R}_{P,n}(r)\} \\ &= Cov\{\hat{W}^{L_k}(r), \hat{W}^P(r)\} + Cov\{\hat{W}^{L_k}(r), \hat{R}_{P,n}(r)\} \\ &\quad + Cov\{\hat{R}_{L_k,n}(r), \hat{W}^P(r)\} + Cov\{\hat{R}_{L_k,n}(x), \hat{R}_{P,n}(r)\} \\ &= (I)+(II)+(III)+(IV). \end{aligned}$$

What we need is to show is: all of the (II), (III), and (IV) are $o_p(1)$. Next we will show that $(II) = o_p(1)$:

Given $\varepsilon > 0$:

$$\begin{aligned}
& Cov\{\hat{W}^{L_k}(r), \hat{R}_{P,n}(r)\} \\
= \; & Cov\{\sum_{i=1}^{n} \hat{W}_{g,i}^{L_k}(r), \sum_{j=1}^{n} \hat{R}_{P,j}(r)\} \\
= \; & \sum_{i=1}^{n} Cov\{\hat{W}_{g,i}^{L_k}(r), \hat{R}_{P,i}(r)\} \text{ (because } \hat{W}_{g,i}^{L_k}(r) \text{ and } \hat{R}_{P,j}(r) \text{are independent for } i \neq j) \\
= \; & \sum_{i=1}^{n} E\{\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)\} - \sum_{i=1}^{n} E\{\hat{W}_{g,i}^{L_k}(r)\}E\{\hat{R}_{P,i}(r)\} \\
= \; & \sum_{i=1}^{n} E\{\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)\} \text{(because } E\{\hat{W}_{g,i}^{L_k}(r)\} = 0) \\
= \; & \sum_{i=1}^{n} \int \hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)dF,
\end{aligned}$$

where dF represents some probability measure of $\widehat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)$.

Then,

$$-\sum_{i=1}^{n} \int |\widehat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF \leq \sum_{i=1}^{n} \int \hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)dF \leq \sum_{i=1}^{n} \int |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF.$$

Note that

$$\begin{aligned}
& \sum_{i=1}^{n} \int |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF \\
= \; & \sum_{i=1}^{n} \int_{|\hat{R}_{P,i}(r)| \leq \varepsilon} |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF + \sum_{i=1}^{n} \int_{|\hat{R}_{P,i}(r)| > \varepsilon} |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF \\
\leq \; & \varepsilon \sum_{i=1}^{n} \int_{|\hat{R}_{P,i}(r)| \leq \varepsilon} |\hat{W}_{g,i}^{L_k}(r)|dF + \sum_{i=1}^{n} \int_{|\hat{R}_{P,i}(r)| > \varepsilon} |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF \\
= \; & (1) + (2).
\end{aligned}$$

Assume that $E\{\hat{W}^{L_k}(r)\}$ exists (i.e., there exists some $M > 0$ such that $E|\hat{W}^{L_k}(r)| < M$), $(1) \leq \varepsilon M$.

Because $\hat{R}_{P,n} = \sum_{i=1}^{n} \hat{R}_{P,i}(r) = o_p(1)$, ie, $P(|\sum_{i=1}^{n} \hat{R}_{P,i}(r)| > \varepsilon) \xrightarrow{P} 0$ as $n \xrightarrow{P} \infty$,

(2) can be made arbitrarily small as ε is made arbitrarily small.

Therefore, $\sum_{i=1}^{n} \int |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF \leq (1) + (2) \xrightarrow{P} 0$ as $n \xrightarrow{P} \infty$. Similarly, we can show that $-\sum_{i=1}^{n} \int |\hat{W}_{g,i}^{L_k}(r)\hat{R}_{P,i}(r)|dF \xrightarrow{P} 0$ as $n \xrightarrow{P} \infty$, thus

$$\sum_{i=1}^{n} \int \hat{W}_{g,i}^{L_k}(r)\hat{R}_{A,i}(r)dF \xrightarrow{P} 0$$

as $n \xrightarrow{P} \infty$.

Using the similar arguments, we can show that $(III) = o_p(1)$ and $(IV) = o_p(1)$. In addition, using the similar rationale, we can show that $W_g^{L_1}(r) + W_g^{L_2}(r) = \hat{W}_g^{L_1}(r) + \hat{W}_g^{L_2}(r) + o_p(1)$

Appendix C

Computation Details for Radon Effects Normal Mixture Regression Models

C.1 The pseudo-complete likelihood for a two component mixture model with random effects

The pseudo-likelihood for the pseudo-complete data $\{\mathbf{Y}, \boldsymbol{\xi}, \mathbf{X}, \mathbf{T}\} = \{(Y_i, \xi_i, \mathbf{X}_i', \mathbf{T}_i')\}_{i=1}^n$ as if ξ_i^m's are observable, can be written as:

$$L^{m,c}(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}^m, \mathbf{X}, \mathbf{T}) = \prod_{i=1}^{n} L_i^{m,c}(\boldsymbol{\theta}|\xi_i^m, \mathbf{y}_i, \mathbf{X}_i, \mathbf{T}_i),$$

where

$$L_i^{m,c}(\boldsymbol{\theta}|\xi_i^m, \mathbf{y}_i, \mathbf{X}_i, \mathbf{T}_i) = \{p_i(\boldsymbol{\alpha})\phi^*(\mathbf{y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(1)}, \Sigma)\}^{\xi_i^m} \{[1 - p_i(\boldsymbol{\alpha})]\, \phi^*(\mathbf{y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(2)}, \Sigma)\}^{1-\xi_i^m}.$$

Because the covariance matrix Σ is a compound symmetric matrix, $\Sigma = \sigma^2 I_n + \sigma_1^2 \mathbf{1}'\mathbf{1}$, the probability density function $\phi^*(\mathbf{y}_i; \mathbf{X}_i'\boldsymbol{\beta}^{(k)}, \Sigma), k = 1, 2$ can be further simplified:

$$\begin{aligned} \Sigma^{-1} &= (\sigma^2 I_n + \sigma_1^2 \mathbf{1}'\mathbf{1})^{-1} \\ &= \frac{I_n}{\sigma^2} - \frac{\sigma 1^2}{\sigma^2(\sigma^2 + J\sigma_1^2)} \mathbf{1}'\mathbf{1}, \end{aligned}$$

$$\begin{aligned} \det(\Sigma) &= \det(\sigma^2 I_n + \sigma_1^2 \mathbf{1}'\mathbf{1}) \\ &= (\sigma^2)^J \det(I_n + (\sigma_1^2/\sigma^2)\mathbf{1}'\mathbf{1}) \\ &= (\sigma^2)^{J-1}(\sigma^2 + J\sigma_1^2). \end{aligned}$$

Then we get:

$$\begin{aligned}\phi^*(\mathbf{Y}_i - \mathbf{X}_{ij}^{'}\boldsymbol{\beta}^{(k)}, \Sigma) &= \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_{ij}^{'}\boldsymbol{\beta}^{(k)})'\Sigma^{-1}(\mathbf{Y}_i - \mathbf{X}_{ij}^{'}\boldsymbol{\beta}^{(k)})\right\} \\ &= \frac{\sigma}{\sqrt{\sigma^2 + J\sigma_1^2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^J \exp\left\{\frac{\sigma_1^2\left(\sum_{j=1}^J(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(k)})\right)^2}{2\sigma^2(\sigma^2 + J\sigma_1^2)}\right\} \\ &\quad \times \exp\left\{-\frac{\sum_{j=1}^J(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(k)})^2}{2\sigma^2}\right\}.\end{aligned}$$

For simplicity of notations, let

$$B_{ki} = \exp\left\{\frac{\sigma_1^2\left[\sum_{j=1}^J(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(k)})\right]^2}{2\sigma^2(\sigma^2 + J\sigma_1^2)} - \frac{\sum_{j=1}^J(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(k)})^2}{2\sigma^2}\right\}, k = 1, 2,$$

then we have:

$$\phi^*(\mathbf{Y}_i - \mathbf{X}_{ij}^{'}\boldsymbol{\beta}_{(k)}, \Sigma) = \frac{\sigma}{\sqrt{\sigma^2 + J\sigma_1^2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^J B_{ki}.$$

Therefore, the pseudo-complete likelihood can be rewritten as:

$$L^{m,c}(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\xi}^m, \mathbf{X}, \mathbf{T}) = \prod_{i=1}^n \frac{\sigma}{\sqrt{\sigma^2 + J\sigma_1^2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^J \{p_i(\boldsymbol{\alpha})B_{1i}\}^{\xi_i^m}\{[1 - p_i(\boldsymbol{\alpha})]B_{2i}\}^{1-\xi_i^m}$$

where B_{1i} is a function of parameters $\sigma, \sigma_1, \boldsymbol{\beta}^{(1)}$, and B_{2i} is a function of parameters $\sigma, \sigma_1, \boldsymbol{\beta}^{(2)}$.

C.2 Computation of score functions from the pseudo-complete likelihood

As shown in the section 2.2, we know that the EM estimators are the solutions to $U^m(\mathbf{Y};\hat{\boldsymbol{\theta}}) = U^{m*}(\mathbf{Y},\hat{\boldsymbol{\xi}}^m;\hat{\boldsymbol{\theta}}) = 0$ where $U^{m*}(\mathbf{Y},\boldsymbol{\xi}^m;\boldsymbol{\theta}) = \partial \log L^{m,c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. The pseudo-complete likelihood is:

$$L^{m,c}(\boldsymbol{\theta}|\mathbf{Y},\boldsymbol{\xi}^m,\mathbf{X},\mathbf{T}) = \prod_{i=1}^{n} \frac{\sigma}{\sqrt{\sigma^2+J\sigma_1^2}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^J \{p_i(\boldsymbol{\alpha})B_{1i}\}^{\xi_i^m} \{[1-p_i(\boldsymbol{\alpha})]\, B_{2i}\}^{1-\xi_i^m}.$$

Then the pseudo-complete likelihood can be clearly separated into 3 parts

$$L^{m,c}(\boldsymbol{\theta}|\boldsymbol{\xi}^m,\mathbf{Y},\mathbf{X},\mathbf{T}) = L_1^m(\boldsymbol{\alpha}) L_2^m(\boldsymbol{\beta}^{(1)},\sigma,\sigma_1) L_3^m(\boldsymbol{\beta}^{(2)},\sigma,\sigma_1)$$

where:

$$L_1^m(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \{p_i(\boldsymbol{\alpha})\}^{\xi_i^m} \{1-p_i(\boldsymbol{\alpha})\}^{1-\xi_i^m},$$

$$L_2^m(\boldsymbol{\beta}^{(1)},\sigma,\sigma_1) = \prod_{i=1}^{n} \frac{\sigma}{\sqrt{\sigma^2+J\sigma_1^2}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^J (B_{1i})^{\xi_i^m},$$

$$L_3^m(\boldsymbol{\beta}^{(2)},\sigma,\sigma_1) = (B_{2i})^{1-\xi_i^m}.$$

$U^{m*}(\mathbf{Y},\boldsymbol{\xi};\boldsymbol{\theta}) = \partial \log L^{m,c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be computed as follows:

$$\begin{aligned} \partial \log L^{m,c}(\boldsymbol{\theta})/\partial\boldsymbol{\alpha} &= \partial \log L_1^m(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} \\ &= \sum_{i=1}^{n} \left\{ \xi_i^m - \frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+e^{\mathbf{T}_i'\boldsymbol{\alpha}}} \right\} \mathbf{T}_i, \end{aligned}$$

$$\begin{aligned}\partial \log L^{m,c}(\boldsymbol{\theta})/\partial \boldsymbol{\beta}^{(1)} &= \partial \log L_2^m(\boldsymbol{\beta}^{(1)}, \sigma, \sigma_1)/\partial \boldsymbol{\beta}^{(1)} \\ &= \sum_{i=1}^{n} \xi_i^m \lambda_{\boldsymbol{\beta}^{(1)}},\end{aligned}$$

where

$$\lambda_{\boldsymbol{\beta}^{(1)}} = \frac{\partial \log B_{1i}}{\partial \boldsymbol{\beta}^{(1)}} = \frac{1}{\sigma^2}\left\{-\frac{\sigma_1^2}{\sigma^2 + J\sigma_1^2}\sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(1)})\sum_{j=1}^{J}\mathbf{X}_{ij} + \sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(1)})\mathbf{X}_{ij}\right\},$$

$$\begin{aligned}\partial \log L^{m,c}(\boldsymbol{\theta})/\partial \boldsymbol{\beta}^{(2)} &= \partial \log L_3^m(\boldsymbol{\beta}^{(2)}, \sigma, \sigma_1)/\partial \boldsymbol{\beta}^{(2)} \\ &= \sum_{i=1}^{n} (1 - \xi_i^m) \lambda_{\boldsymbol{\beta}^{(2)}},\end{aligned}$$

where

$$\lambda_{\boldsymbol{\beta}^{(2)}} = \frac{\partial \log B_{2i}}{\partial \boldsymbol{\beta}^{(2)}} = \frac{1}{\sigma^2}\left\{-\frac{\sigma_1^2}{\sigma^2 + J\sigma_1^2}\sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(2)})\sum_{j=1}^{J}\mathbf{X}_{ij} + \sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(2)})\mathbf{X}_{ij}\right\},$$

$$\partial \log L^{m,c}(\boldsymbol{\theta})/\partial \sigma = \sum_{i=1}^{n}\left\{\frac{1-J}{\sigma} - \frac{\sigma}{\sigma^2 + J\sigma_1^2} + \xi_i^m \lambda_{1\sigma} + (1 - \xi_i^m)\lambda_{2\sigma}\right\},$$

where

$$\lambda_{1\sigma} = \frac{\partial \log B_{1i}}{\partial \sigma} = \frac{1}{\sigma^3}\left\{-\frac{\sigma_1^2(2\sigma^2 + J\sigma_1^2)}{(\sigma^2 + J\sigma_1^2)^2}\left[\sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(1)})\right]^2 + \sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(1)})^2\right\},$$

and

$$\lambda_{2\sigma} = \frac{\partial \log B_{2i}}{\partial \sigma} = \frac{1}{\sigma^3}\left\{-\frac{\sigma_1^2(2\sigma^2 + J\sigma_1^2)}{(\sigma^2 + J\sigma_1^2)^2}\left[\sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(2)})\right]^2 + \sum_{j=1}^{J}(Y_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}^{(2)})^2\right\},$$

$$\partial \log L^{m,c}(\boldsymbol{\theta})/\partial\sigma_1 \quad = \quad \sum_{i=1}^{n}\left\{-\frac{J\sigma_1}{\sigma^2+J\sigma_1^2}+\xi_i^m\lambda_{1\sigma_1}+(1-\xi_i^m)\lambda_{2\sigma_1}\right\},$$

where

$$\lambda_{1\sigma_1}=\frac{\partial \log B_{1i}}{\partial\sigma_1}=\frac{\sigma_1}{(\sigma^2+J\sigma_1^2)^2}\left\{\sum_{j=1}^{J}(Y_{ij}-\mathbf{X}_{ij}'\boldsymbol{\beta}^{(1)})\right\}^2,$$

and

$$\lambda_{2\sigma_1}=\frac{\partial \log B_{2i}}{\partial\sigma_1}=\frac{\sigma_1}{(\sigma^2+J\sigma_1^2)^2}\left\{\sum_{j=1}^{J}(Y_{ij}-\mathbf{X}_{ij}'\boldsymbol{\beta}^{(2)})\right\}^2.$$

Then, $U^m(\mathbf{Y};\hat{\boldsymbol{\theta}}) = U^{m*}(\mathbf{Y},\hat{\boldsymbol{\xi}}^m;\hat{\boldsymbol{\theta}})$ can be computed as follows by replacing the $\xi_i^m,\boldsymbol{\alpha},\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\sigma,\sigma_1$ in $U^{m*}(\mathbf{Y},\boldsymbol{\xi};\boldsymbol{\theta}) = \partial \log L^{m,c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ by $\hat{\xi}_i^m,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}^{(1)},\hat{\boldsymbol{\beta}}^{(2)},\hat{\sigma},\hat{\sigma_1}$, respectively. Therefore, the score function $U^m(\mathbf{Y};\boldsymbol{\theta})$ is:

$$U^m(\mathbf{Y};\boldsymbol{\theta})=\sum_{i=1}^{n}U_i^m(\mathbf{Y}_i;\boldsymbol{\theta})=\sum_{i=1}^{n}\begin{pmatrix}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})-\frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+e^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\}\mathbf{T}_i\\ \hat{\xi}_i^m(\boldsymbol{\theta})\lambda_{\boldsymbol{\beta}^{(1)}}\\ \left[1-\hat{\xi}_i^m(\boldsymbol{\theta})\right]\lambda_{\boldsymbol{\beta}^{(2)}}\\ \frac{1-J}{\sigma}-\frac{\sigma}{\sigma^2+J\sigma_1^2}+\hat{\xi}_i^m(\boldsymbol{\theta})\lambda_{1\sigma}+\left[1-\hat{\xi}_i^m(\boldsymbol{\theta})\right]\lambda_{2\sigma}\\ -\frac{J\sigma_1}{\sigma^2+J\sigma_1^2}+\hat{\xi}_i^m(\boldsymbol{\theta})\lambda_{1\sigma_1}+\left[1-\hat{\xi}_i^m(\boldsymbol{\theta})\right]\lambda_{2\sigma_1}\end{pmatrix},$$

and $U^m(\mathbf{Y};\hat{\boldsymbol{\theta}})$ is

$$U^m(\mathbf{Y};\hat{\boldsymbol{\theta}})=\sum_{i=1}^{n}\begin{pmatrix}\left\{\hat{\xi}_i^m-\frac{e^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}{1+e^{\mathbf{T}_i'\hat{\boldsymbol{\alpha}}}}\right\}\mathbf{T}_i\\ \hat{\xi}_i^m\hat{\lambda}_{\boldsymbol{\beta}^{(1)}}\\ (1-\hat{\xi}_i^m)\hat{\lambda}_{\boldsymbol{\beta}^{(2)}}\\ \frac{1-J}{\hat{\sigma}}-\frac{\hat{\sigma}}{\hat{\sigma}^2+J\hat{\sigma}_1^2}+\hat{\xi}_i^m\hat{\lambda}_{1\sigma}+(1-\hat{\xi}_i^m)\hat{\lambda}_{2\sigma}\\ -\frac{J\hat{\sigma}_1}{\hat{\sigma}^2+J\hat{\sigma}_1^2}+\hat{\xi}_i^m\hat{\lambda}_{1\sigma_1}+(1-\hat{\xi}_i^m)\hat{\lambda}_{2\sigma_1}\end{pmatrix},$$

where $\hat{\lambda}_{\boldsymbol{\beta}^{(1)}},\hat{\lambda}_{\boldsymbol{\beta}^{(2)}},\hat{\lambda}_{1\sigma},\hat{\lambda}_{2\sigma},\hat{\lambda}_{1\sigma_1},\hat{\lambda}_{2\sigma_1}$ can be obtained from $\lambda_{\boldsymbol{\beta}^{(1)}},\lambda_{\boldsymbol{\beta}^{(2)}},\lambda_{1\sigma},\lambda_{2\sigma},\lambda_{1\sigma_1},\lambda_{2\sigma_1}$ by replacing all $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$. From the properties of score functions, we know that $U^m(\mathbf{Y};\hat{\boldsymbol{\theta}}) = 0$ and $E\left\{U_i^m(\mathbf{Y}_i;\boldsymbol{\theta})\right\} = 0$.

C.3 The computation of $\hat{\eta}_k^m(\mathbf{x};\boldsymbol{\theta})$ and $\hat{\eta}_{g,k}^m(r;\boldsymbol{\theta})$

This section shows the computation of $\hat{\eta}_k^m(\mathbf{x};\boldsymbol{\theta})$ and $\hat{\eta}_{g,k}^m(r;\boldsymbol{\theta})$ corresponding the GOF test statistic for the kth component, $W^{m,L_k}(\mathbf{x})$. We use $\hat{\eta}_1^m(\mathbf{x};\boldsymbol{\theta})$ as an example:

$$\hat{\eta}_1^m(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} I(\mathbf{X}_{ij} \le \mathbf{x})\frac{\partial}{\partial \boldsymbol{\theta}}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(Y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\right\}.$$

The derivative of the pseudo-residuals $\hat{\xi}_i^m(\boldsymbol{\theta})(Y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})$ with respect to each parameter is computed as follows:

$$\frac{\partial}{\partial \boldsymbol{\alpha}}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\right\} = (\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}},$$

$$\frac{\partial}{\partial \boldsymbol{\beta}^{(1)}}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\right\} = -\hat{\xi}_i^m(\boldsymbol{\theta})\mathbf{X}_i + (\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(1)}},$$

$$\frac{\partial}{\partial \boldsymbol{\beta}^{(2)}}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\right\} = (\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(2)}},$$

$$\frac{\partial}{\partial \sigma}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\right\} = (\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \sigma},$$

$$\frac{\partial}{\partial \sigma_1}\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\right\} = (\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \sigma_1}.$$

Then, we need to compute the derivatives $(\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}}, \frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(1)}}, \frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(2)}}, \frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \sigma}, \frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \sigma_1})$ from the expression of $\hat{\xi}_i^m(\boldsymbol{\theta})$, which is a function of $\boldsymbol{\theta}$:

$$\hat{\xi}_i^m(\boldsymbol{\theta}) = 1 - \frac{B_{2i}}{\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i} + B_{2i}}.$$

Then the first derivatives of $\hat{\xi}_i^m$ with respect to $\boldsymbol{\theta}$ are as follows:

$$\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}\mathbf{T}_{ij}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2},$$

$$\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(1)}} = \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}\lambda_{\boldsymbol{\beta}^{(1)}},$$

$$\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^{(2)}} = \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}\lambda_{\boldsymbol{\beta}^{(2)}},$$

$$\begin{aligned}\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \sigma} &= \frac{B_{2i}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}\left\{\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}\lambda_{1\sigma}+B_{2i}\lambda_{1\sigma}\right\} - \frac{1}{\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i}}B_{2i}\lambda_{2\sigma} \\ &= \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}(\lambda_{1\sigma}-\lambda_{2\sigma}),\end{aligned}$$

$$\frac{\partial \hat{\xi}_i^m(\boldsymbol{\theta})}{\partial \sigma_1} = \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}(\lambda_{1\sigma_1}-\lambda_{1\sigma_2}).$$

The $\lambda_{\boldsymbol{\beta}^{(1)}}$, $\lambda_{\boldsymbol{\beta}^{(2)}}$, $\lambda_{1\sigma}$, $\lambda_{2\sigma}$, $\lambda_{1\sigma_1}$, $\lambda_{2\sigma_1}$ in the above equations are already defined in Appendix C.2. Therefore, $\hat{\eta}_1^m(\mathbf{x};\boldsymbol{\theta})$ can be computed as follows:

$$\hat{\eta}_1^m(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{=1}^{J} I(\mathbf{X}_{ij}\le \mathbf{x})\left(\eta_\alpha^m, \eta_{\beta_1}^m, \eta_{\beta_2}^m, \eta_\sigma^m, \eta_{\sigma_1}^m\right),$$

where

$$\eta_\alpha^m = (\mathbf{Y}_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{B_{1i}B_{2i}\exp(\mathbf{T}'_i\boldsymbol{\alpha})\mathbf{T}_i}{\{B_{1i}\exp(\mathbf{T}'_i\boldsymbol{\alpha})+B_{2i}\}^2},$$

$$\eta_{\beta_1}^m = -\hat{\xi}_i^m(\boldsymbol{\theta})\mathbf{X}_i + (\mathbf{Y}_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2\sigma^2}\lambda_{\boldsymbol{\beta}^{(1)}},$$

$$\eta_{\beta_2}^m = (\mathbf{Y}_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2\sigma^2}\lambda_{\boldsymbol{\beta}^{(2)}},$$

$$\eta_\sigma^m = (\mathbf{Y}_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}(\lambda_{1\sigma}-\lambda_{2\sigma}),$$

$$\eta_{\sigma_1}^m = (\mathbf{Y}_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i}+B_{2i})^2}(\lambda_{1\sigma_1}-\lambda_{2\sigma_1}).$$

Then, $\hat{\eta}_1^m(\mathbf{x};\hat{\boldsymbol{\theta}})$ can be obtained by replacing $\boldsymbol{\theta}$, $\hat{\xi}_i^m(\boldsymbol{\theta})$, B_{1i} and B_{2i} with $\hat{\boldsymbol{\theta}}$, $\hat{\xi}_i^m(\hat{\boldsymbol{\theta}})$, $\hat{B}_{1i}$ and $\hat{B}_{2i}$, where

$$\hat{B}_{1i} = \exp\left\{ \frac{\hat{\sigma}_1^2 \left(\sum_{j=1}^J (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)})\right)^2}{2\hat{\sigma}^2(\hat{\sigma}^2 + J\hat{\sigma}_1^2)} - \frac{\sum_{j=1}^J (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)})^2}{2\hat{\sigma}^2} \right\},$$

and

$$\hat{B}_{2i} = \exp\left\{ \frac{\hat{\sigma}_1^2 \left(\sum_{j=1}^J (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(2)})\right)^2}{2\hat{\sigma}^2(\hat{\sigma}^2 + J\hat{\sigma}_1^2)} - \frac{\sum_{j=1}^J (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(2)})^2}{2\hat{\sigma}^2} \right\}.$$

For the second component, $\hat{\eta}_2^m(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^J I(\mathbf{X}_{ij} \le \mathbf{x}) \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \hat{\xi}_i^m(\boldsymbol{\theta})(Y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \right\}$. Then the expression of $\hat{\eta}_2(\mathbf{x};\boldsymbol{\theta})$ is:

$$\hat{\eta}_2^m(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \sum_{=1}^J I(\mathbf{X}_{ij} \le \mathbf{x}) \left(\eta_\alpha^m, \eta_{\beta_1}^m, \eta_{\beta_2}^m, \eta_\sigma^m, \eta_{\sigma_1}^m\right),$$

where

$$\eta_\alpha^m = -(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \frac{B_{1i}B_{2i}\exp(\mathbf{T}'_i\boldsymbol{\alpha})\mathbf{T}_i}{\{B_{1i}\exp(\mathbf{T}'_i\boldsymbol{\alpha}) + B_{2i}\}^2},$$

$$\eta_{\beta_1}^m = -(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i} + B_{2i})^2\sigma^2}\lambda_{\boldsymbol{\beta}^{(1)}},$$

$$\eta_{\beta_2}^m = -\left[1 - \hat{\xi}_i^m(\boldsymbol{\theta})\right]\mathbf{X}_i - (\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i} + B_{2i})^2\sigma^2}\lambda_{\boldsymbol{\beta}^{(2)}},$$

$$\eta_\sigma^m = -(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i} + B_{2i})^2}(\lambda_{1\sigma} - \lambda_{2\sigma}),$$

$$\eta_{\sigma_1}^m = -(\mathbf{Y}_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}^{(2)}) \frac{B_{1i}B_{2i}\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}}{(\mathrm{e}^{\mathbf{T}'_{ij}\boldsymbol{\alpha}}B_{1i} + B_{2i})^2}(\lambda_{1\sigma_1} - \lambda_{2\sigma_1}).$$

Similarly, $\hat{\eta}_2^m(\mathbf{x};\hat{\boldsymbol{\theta}})$ can be obtained by replacing $\boldsymbol{\theta}$, $\hat{\xi}_i^m(\boldsymbol{\theta})$, B_{1i} and B_{2i} with $\hat{\boldsymbol{\theta}}$, $\hat{\xi}_i^m(\hat{\boldsymbol{\theta}})$, $\hat{B}_{1i}$ and $\hat{B}_{2i}$.

In addition, $\hat{\eta}_{g,k}^m(r;\hat{\boldsymbol{\theta}})$ can be obtained from $\hat{\eta}_k^m(\mathbf{x};\hat{\boldsymbol{\theta}})$ by replacing $I(\mathbf{X}_{ij} \le \mathbf{x})$ with $I(\mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(k)} \le r)$.

C.4 The computation of $\hat{\eta}_P^m(\mathbf{t};\boldsymbol{\theta})$ and $\hat{\eta}_{g,P}^m(r;\boldsymbol{\theta})$

The computation of $\hat{\eta}_P^m(\mathbf{t};\boldsymbol{\theta})$ and $\hat{\eta}_{g,P}^m(r;\boldsymbol{\theta})$ for the logistic regression of the mixing proportion is relatively easy.

$$\begin{aligned}\hat{\eta}_P^m(\mathbf{t},\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{T}_i \le \mathbf{t})\frac{\partial}{\partial\boldsymbol{\theta}}\left\{\hat{\xi}_i^m(\boldsymbol{\theta}) - \frac{\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}{1+\mathrm{e}^{\mathbf{T}_i'\boldsymbol{\alpha}}}\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{T}_i \le \mathbf{t})\left(\eta_\alpha^m, \eta_{\beta_1}^m, \eta_{\beta_2}^m, \eta_\sigma^m, \eta_{\sigma_1}^m\right)\end{aligned}$$

where

$$\begin{aligned}\eta_\alpha^m &= \frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}} - \frac{\partial}{\partial\boldsymbol{\alpha}}\frac{1}{1+e^{-\mathbf{T}_i'\boldsymbol{\alpha}}} \\ &= \frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}} - \frac{e^{\mathbf{T}_i'\boldsymbol{\alpha}}\mathbf{T}_i}{(1+e^{\mathbf{T}_i'\boldsymbol{\alpha}})^2}\end{aligned}$$

and $\eta_{\beta_1}^m = \frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^{(1)}}$, $\eta_{\beta_2}^m = \frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^{(2)}}$, $\eta_\sigma^m = \frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}}$, $\eta_{\sigma 1}^m = \frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}1}$. $\frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}}$, $\frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^{(1)}}$, $\frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}^{(2)}}$, $\frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}}$, $\frac{\partial\hat{\xi}_i^m(\boldsymbol{\theta})}{\partial\boldsymbol{\sigma}1}$ are all defined in Appendix C.3.

Similarly, $\hat{\eta}_P^m(\mathbf{t};\hat{\boldsymbol{\theta}})$ can be obtained by replacing $\boldsymbol{\theta}$, $\hat{\xi}_i^m(\boldsymbol{\theta})$, B_{1i} and B_{2i} with $\hat{\boldsymbol{\theta}}$, $\hat{\xi}_i^m(\hat{\boldsymbol{\theta}})$, $\hat{B}_{1i}$ and $\hat{B}_{2i}$.

In addition, $\hat{\eta}_{g,P}^m(\mathbf{t};\hat{\boldsymbol{\theta}})$ can be obtained from $\hat{\eta}_P^m(\mathbf{t};\hat{\boldsymbol{\theta}})$ by replacing $I(\mathbf{T}_i \le \mathbf{t})$ with $I(\mathbf{T}_i'\hat{\boldsymbol{\alpha}} \le r)$.

C.5 Proof for $\sum_{i=1}^{n} e_{ij}^{(m,k)} = 0, k = 1, 2$ and $\sum_{i=1}^{n} e_{i}^{m,P} = 0$

As shown earlier, $U^{m*}(\mathbf{Y}, \hat{\boldsymbol{\xi}}^{m}; \hat{\boldsymbol{\theta}}) = 0$ (Appendix C.2), then we have

$$\sum_{i=1}^{n} \frac{\hat{\xi}_i^m}{\hat{\sigma}^2} \left\{ -\frac{\hat{\sigma}_1^2}{\hat{\sigma}^2 + J\hat{\sigma}_1^2} \sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}) \sum_{j=1}^{J} \mathbf{X}_{ij} + \sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}) \mathbf{X}_{ij} \right\} = \mathbf{0}.$$

Because $\mathbf{X}_{ij}$ is a $(p+1) \times 1$ covariate vector including intercept, we can obtain the following by replacing $\mathbf{X}_{ij}$ with the intercept:

$$\begin{aligned} & \sum_{i=1}^{n} \frac{\hat{\xi}_i^m}{\hat{\sigma}^2} \left\{ -\frac{J\hat{\sigma}_1^2}{\hat{\sigma}^2 + J\hat{\sigma}_1^2} \sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}) + \sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}) \right\} \\ = & \sum_{i=1}^{n} \frac{\hat{\xi}_i^m}{\hat{\sigma}^2 + J\hat{\sigma}_1^2} \sum_{j=1}^{J} (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}) \\ = & 0. \end{aligned}$$

Then we have

$$\sum_{i=1}^{n} \sum_{j=1}^{J} e_{ij}^{(m,1)} = \sum_{i=1}^{n} \sum_{j=1}^{J} \hat{\xi}_i^m (Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(1)}) = 0.$$

Using the same argument, we can prove that

$$\sum_{i=1}^{n} \sum_{j=1}^{J} e_{ij}^{(m,2)} = \sum_{i=1}^{n} \sum_{j=1}^{J} (1 - \hat{\xi}_i^m)(Y_{ij} - \mathbf{X}'_{ij}\hat{\boldsymbol{\beta}}^{(2)}) = 0$$

.

Similarly, from the score function with respect to $\boldsymbol{\alpha}$, we know that

$$\sum_{i=1}^{n} \left\{ \hat{\xi}_i^m - \frac{e^{\mathbf{T}'_i\hat{\boldsymbol{\alpha}}}}{1 + e^{\mathbf{T}'_i\hat{\boldsymbol{\alpha}}}} \right\} \mathbf{T}_i = 0$$

Then, it is straightforward that $\sum_{i=1}^{n} e_i^{m,P} = \sum_{i=1}^{n} \left\{ \hat{\xi}_i^m - \frac{e^{\mathbf{T}'_i\hat{\boldsymbol{\alpha}}}}{1+e^{\mathbf{T}'_i\hat{\boldsymbol{\alpha}}}} \right\} = 0$

C.6 Proof for $E\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(Y_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}_{(k)})\right\}=0, k=1,2$ and $E\left\{\hat{\xi}_i^m(\boldsymbol{\theta})-\frac{e^{\mathbf{T}'_i\boldsymbol{\alpha}}}{1+e^{\mathbf{T}'_i\boldsymbol{\alpha}}}\right\}=0$

Because the expectation of score function $U_i^m(\mathbf{Y}_i;\boldsymbol{\theta})$ is zero, from the expression of the score function in Appendix C.2, we have

$$E\left\{\hat{\xi}_i^m(\boldsymbol{\theta})-\frac{e^{\mathbf{T}'_i\boldsymbol{\alpha}}}{1+e^{\mathbf{T}'_i\boldsymbol{\alpha}}}\right\}\mathbf{T}_i=\mathbf{0}$$

$$E\left\{\frac{\hat{\xi}_i^m(\boldsymbol{\theta})}{\sigma^2}\left[-\frac{\sigma_1^2}{\sigma^2+J\sigma_1^2}\sum_{j=1}^{J}(Y_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\sum_{j=1}^{J}\mathbf{X}_{ij}+\sum_{j=1}^{J}(Y_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}^{(1)})\mathbf{X}_{ij}\right]\right\}=\mathbf{0}$$

Then we know $E\left\{\hat{\xi}_i^m(\boldsymbol{\theta})-\frac{e^{\mathbf{T}'_i\boldsymbol{\alpha}}}{1+e^{\mathbf{T}'_i\boldsymbol{\alpha}}}\right\}=0$ and after some similar transformation shown in Appendix C.5, it is obvious that $E\left\{\hat{\xi}_i^m(\boldsymbol{\theta})(Y_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}_{(k)})\right\}=0, k=1,2$

Appendix D

Computing Codes

D.1 R code for simulation to evaluate the performance of the proposed GOF tests for a two-component mixture model without random effects

```
#########################################################################
#R code for computing p-values for combined data including class 1 and 2
#based on modified cumulative residual test for mixture models##########
#updated on Jan2011,to include GOF for joint testing of both linear comp
#components and logistic part, and separate GOF tests for linear
#components and logistic part; for each test, we use different indicators
#### either X_i (T_i) or X_i*beta (T_i*alpha)           ###
#### Add one more extra covariate in May 2011         #####
#########################################################################
seedno=rnorm(1)*1000000

library(gof)
library(MASS)
library(data.table)
library(zoo)

#########################################################################
###############read csv data exported from SAS##########################
test <- read.table("C:/Documents and Settings/JSHEN3/My Documents/Beijing
Olympic Thesis/data for R/testC.csv", header = TRUE, sep=",")
```

```
testcov <- read.table("C:/Documents and Settings/JSHEN3/My Documents/
Beijing Olympic Thesis/data for R/testcovC.csv", header = TRUE, sep=",")

#########################################################################
#************************Define function last.var*********************
#########################################################################
last.var=function(x){
  xx=x
  y=1:length(x)
  y[length(x)]=1
  for(i in 1:(length(x)-1)){
     if (xx[i]==xx[i+1]){y[i]=0}
     else {y[i]=1}
  }
  y
}

#########################################################################
#************************Define function cumresmixtureP****************
####################    for GOF test of the logistic part ###############
#########################################################################
cumresmixtureP=function(dat, ycol, x1col, sigcol, indi, datcov, R){
names(dat)[ycol]="y"
names(dat)[x1col]="x1"
d1=dat[order(dat[,indi]),] #sort by either xx or talpha
n=dim(dat)[1] #n: the number of obs in the data
inv=datcov[,2:11]

e=d1[13][[1]] - d1[6][[1]]
e1=d1[8][[1]]
```

```
e2=d1[9][[1]]
phat=d1[6][[1]]
ptild=d1[13][[1]]
T=d1[4][[1]]
T2=d1[14][[1]] #extra covariate
sigma=d1[1,7]

#combine intercept with all fixed covariates (age)
d1x=cbind(rep(1,n), d1$x1, d1$T2) #combine intercept with x1,remove y in d1
d1t=cbind(rep(1,n), d1$xx, d1$T2) # combine intercept with xx

h=(d1[1, sigcol])^2 #h=sigma^2 for each loop
eat= phat/(1-phat) #exp(alpha*T)
exp1= exp(-(e1^2)/(2*sigma^2)) #exp(-e1^2/2sigma^2)
exp2= exp(-(e2^2)/(2*sigma^2)) #exp(-e2^2/2sigma^2)
xialpha=(eat*exp1*exp2/(eat*exp1+exp2)^2)*d1t
xibeta1=(eat*exp1*exp2/(eat*exp1+exp2)^2)*e1*d1x/(sigma^2)
xibeta2= -(eat*exp1*exp2/(eat*exp1+exp2)^2)*e2*d1x/(sigma^2)
xisigma= (1/(sigma^3))* eat * exp1 * exp2 * (e1^2 - e2^2)
/((eat*exp1+exp2)^2)

eta.alpha= xialpha - eat*d1t/((1+eat)^2)
eta.beta1= xibeta1
eta.beta2= xibeta2
eta.sigma= xisigma
eta1=cbind(eta.alpha, eta.beta1, eta.beta2, eta.sigma)
eta=apply(eta1, 2, cumsum)

flag=last.var(d1[,indi])
```

```
w=cumsum(e)/sqrt(n)

mod.w=w*flag
w.sup=max(abs(mod.w))

What=matrix(0,n,R)
wh.sup=rep(0, R)

for (r in 1:R){
set.seed(seedno+r)
z=rnorm(n)      #generate n random variates N(0,1)
ez=e*z
what1=cumsum(ez)

Ualpha=(ptild - phat)*d1t
xe1z=d1x*e1*ptild/h
xe2z=d1x*e2*(1-ptild)/h
Usigma= -1/sigma + (ptild*(e1^2) + (1-ptild)*(e2^2))/(sigma^3)

tp=t(cbind(Ualpha, xe1z, xe2z, Usigma)) #temp: 4Xn matrix
tp2=t(z*t(tp))

what2=eta %*% as.matrix(inv) %*% apply(tp2, 1, sum)

what=(what1 + what2)
mod.what=(what*flag)/sqrt(n)

What[,r]=mod.what[,1]

wh.sup[r]=max(abs(mod.what))
```

```
}
p=sum(wh.sup>=w.sup)/R
list(W=mod.w, What=What, a=what1, b=what2, W.sup=w.sup,
What.sup=wh.sup, p=p)
}
#indi: the column number of the indicator in the data,
#eg 4 means dat[4]=xx is used in the indicator function
#eg 10 means data[10]=xx*alpha is used
g=cumresmixtureP(dat=test[test$loop==1,], ycol=5, x1col=3, sigcol=7,
 indi=4,datcov=testcov[testcov$loop==1, ], R=100)

#########################################################################
#*************************Define function cumresmixtureLin1**********
####################    for GOF test of the linear comp 1 ##########
#########################################################################
cumresmixtureLin1=function(dat, ycol, x1col, sigcol, indi, datcov, R){
names(dat)[ycol]="y"
names(dat)[x1col]="x1"
d1=dat[order(dat[,indi]),]
n=dim(dat)[1] #n: the number of obs in the data
inv=datcov[,2:11]

#e=d1[16][[1]]
e1=d1[8][[1]]
e2=d1[9][[1]]
phat=d1[6][[1]]
ptild=d1[13][[1]]
T=d1[4][[1]]
T2=d1[14][[1]] #extra covariate
sigma=d1[1,7]
```

```
e=ptild*e1 #\xi_i * e1

#combine intercept with all fixed covariates (age)
d1x=cbind(rep(1,n), d1$x1, d1$T2)
d1t=cbind(rep(1,n), d1$xx, d1$T2)  # combine intercept with xx

h=(d1[1, sigcol])^2 #h=sigma^2 for each loop
eat= phat/(1-phat) #exp(alpha*T)
exp1= exp(-(e1^2)/(2*sigma^2)) #exp(-e1^2/2sigma^2)
exp2= exp(-(e2^2)/(2*sigma^2)) #exp(-e2^2/2sigma^2)
xialpha=(eat*exp1*exp2/(eat*exp1+exp2)^2)*d1t
xibeta1=(eat*exp1*exp2/(eat*exp1+exp2)^2)*e1*d1x/(sigma^2)
xibeta2= -(eat*exp1*exp2/(eat*exp1+exp2)^2)*e2*d1x/(sigma^2)
xisigma= (1/(sigma^3))* eat * exp1 * exp2 * (e1^2 - e2^2)
 /((eat*exp1+exp2)^2)

#the followings are for comp1 only
eta.alpha= e1* xialpha
eta.beta1= -ptild*d1x + e1*xibeta1
eta.beta2=  e1*xibeta2
eta.sigma= e1*xisigma
eta1=cbind(eta.alpha, eta.beta1, eta.beta2, eta.sigma)
eta=apply(eta1, 2, cumsum)

flag=last.var(d1[,indi])

w=cumsum(e)/sqrt(n)

mod.w=w*flag
w.sup=max(abs(mod.w))
```

```
What=matrix(0,n,R)
wh.sup=rep(0, R)

for (r in 1:R){
set.seed(seedno+r)
z=rnorm(n)      #generate n random variates N(0,1)
ez=e*z
what1=cumsum(ez)

Ualpha=(ptild - phat)*d1t
xe1z=d1x*e1*ptild/h
xe2z=d1x*e2*(1-ptild)/h
Usigma= -1/sigma + (ptild*(e1^2) + (1-ptild)*(e2^2))/(sigma^3)

tp=t(cbind(Ualpha, xe1z, xe2z, Usigma)) #temp: 4Xn matrix
tp2=t(z*t(tp))

what2=eta %*% as.matrix(inv) %*% apply(tp2, 1, sum)

what=(what1 + what2)

mod.what=(what*flag)/sqrt(n)

What[,r]=mod.what[,1]

wh.sup[r]=max(abs(mod.what))
}
p=sum(wh.sup>=w.sup)/R
list(W=mod.w, What=What, a=what1, b=what2, W.sup=w.sup,
```

```
What.sup=wh.sup, p=p)
}
g2=cumresmixtureLin1(dat=test[test$loop==1,], ycol=5, x1col=3, sigcol=7,
 indi=3, datcov=testcov[testcov$loop==1, ], R=100)

#########################################################################
#****************Define function cumresmixture2**********************
#############    for GOF test of the linear comp 2 ###################
#########################################################################
cumresmixtureLin2=function(dat, ycol, x1col, sigcol, indi, datcov, R){
names(dat)[ycol]="y"
names(dat)[x1col]="x1"
d1=dat[order(dat[,indi]),]
n=dim(dat)[1] #n: the number of obs in the data
inv=datcov[,2:11]

#e=d1[16][[1]]
e1=d1[8][[1]]
e2=d1[9][[1]]
phat=d1[6][[1]]
ptild=d1[13][[1]]
T=d1[4][[1]]
T2=d1[14][[1]] #extra covariate
sigma=d1[1,7]
e=(1-ptild)*e2

#combine intercept with all fixed covariates (age)
d1x=cbind(rep(1,n), d1$x1, d1$T2)
d1t=cbind(rep(1,n), d1$xx, d1$T2)  # combine intercept with xx
```

```
h=(d1[1, sigcol])^2 #h=sigma^2 for each loop
eat= phat/(1-phat) #exp(alpha*T)
exp1= exp(-(e1^2)/(2*sigma^2)) #exp(-e1^2/2sigma^2)
exp2= exp(-(e2^2)/(2*sigma^2)) #exp(-e2^2/2sigma^2)
xialpha=(eat*exp1*exp2/(eat*exp1+exp2)^2)*d1t
xibeta1=(eat*exp1*exp2/(eat*exp1+exp2)^2)*e1*d1x/(sigma^2)
xibeta2= -(eat*exp1*exp2/(eat*exp1+exp2)^2)*e2*d1x/(sigma^2)
xisigma=(1/(sigma^3))* eat * exp1 * exp2 *(e1^2 - e2^2)
/((eat*exp1+exp2)^2)

eta.alpha= - e2* xialpha
eta.beta1= -e2*xibeta1
eta.beta2= -(1 - ptild)*d1x + -e2*xibeta2
eta.sigma= - e2*xisigma
eta1=cbind(eta.alpha, eta.beta1, eta.beta2, eta.sigma)
eta=apply(eta1, 2, cumsum)

flag=last.var(d1[,indi])

w=cumsum(e)/sqrt(n)

mod.w=w*flag
w.sup=max(abs(mod.w))

What=matrix(0,n,R)
wh.sup=rep(0, R)

for (r in 1:R){
set.seed(seedno+r)
z=rnorm(n)      #generate n random variates N(0,1)
```

```
ez=e*z
what1=cumsum(ez)

Ualpha=(ptild - phat)*d1t
xe1z=d1x*e1*ptild/h
xe2z=d1x*e2*(1-ptild)/h
Usigma= -1/sigma + (ptild*(e1^2) + (1-ptild)*(e2^2))/(sigma^3)

tp=t(cbind(Ualpha, xe1z, xe2z, Usigma)) #temp: 4Xn matrix
tp2=t(z*t(tp))

what2=eta %*% as.matrix(inv) %*% apply(tp2, 1, sum)

what=(what1 + what2)

mod.what=(what*flag)/sqrt(n)

What[,r]=mod.what[,1]

wh.sup[r]=max(abs(mod.what))
}
p=sum(wh.sup>=w.sup)/R
list(W=mod.w, What=What, a=what1, b=what2, W.sup=w.sup,
What.sup=wh.sup, p=p)
}
g2=cumresmixtureLin2(dat=test[test$loop==1,], ycol=5, x1col=3, sigcol=7,
indi=3, datcov=testcov[testcov$loop==1, ], R=100)

#########################################################################
#####Run Simulations N times for functional form tests  ###############
```

```
#########################################################################
N=max(test$loop) # total number of loops for simulation
n=max(test[test$loop==1,]$obs) #number of subjects in each loop

#pvalues for functional form tests
pL1=rep(0,N) # p-value for the GOF test of linear part 1
pL2=rep(0,N) # p-value for the GOF test of linear part 2
plog=rep(0,N) # p-value for the GOF test of logistic part
pL1L2=rep(0,N) # p-value for the joint GOF test of linear parts
pL1L2P=rep(0,N)
R=1000 #total number of realziations for each GOF test

for (i in 1:N){
W=matrix(0, n)
What=matrix(0, n, R) # W_hat for one random draws

loopdraw=test[test$loop==i,] #subset data for loop i and draw d
loopcov=testcov[testcov$loop==i,] #the covariance matrix for loop i

#functional form test for Linear comp 1 (gL2) and 2 (gL2) and logistic
part (gP)
gL1=cumresmixtureLin1(dat=loopdraw, ycol=5, x1col=3, sigcol=7, indi=3,
 datcov=loopcov, R=R)
gL2=cumresmixtureLin2(dat=loopdraw, ycol=5, x1col=3, sigcol=7, indi=3,
 datcov=loopcov, R=R)
glog= cumresmixtureP(dat=loopdraw, ycol=5, x1col=3, sigcol=7, indi=4,
datcov=loopcov, R=R)
#glinkOld=cumresmixtureLinComp(dat=loopdraw, ycol=5, x1col=3, sigcol=7,
indi=11, datcov=loopcov, R=R)
```

```
#test of linear comp1
pL1[i]=sum(gL1$What.sup>gL1$W.sup)/R

#test of linear comp2
pL2[i]=sum(gL2$What.sup>gL2$W.sup)/R

#test of logistic part
plog[i]=sum(glog$What.sup>glog$W.sup)/R

#Joint test of linear components (L1+L2)
Wnew=gL1$W.sup + gL2$W.sup
What=gL1$What.sup + gL2$What.sup
pL1L2[i]=sum(What>Wnew)/R

#Joint test of linear components and logistic regresssion
(L1+L2+P)
Wnew=gL1$W.sup + gL2$W.sup + glog$W.sup
What=gL1$What.sup + gL2$What.sup + glog$What.sup
pL1L2P[i]=sum(What>Wnew)/R
}
sum(pL1<0.05)/N
sum(pL2<0.05)/N
sum(plog<0.05)/N
#sum(pL1L2<0.05)/N
#sum(pL1L2P<0.05)/N

#########################################################################
############Run Simulations N times for Link function tests  #########
#########################################################################
N=max(test$loop) # total number of loops for simulation
```

```
n=max(test[test$loop==1,]$obs) #number of subjects in each loop
#pvalues for LINK FUNCTION tests
plk.L1=rep(0,N) # p-value for the GOF test of linear part 1
plk.L2=rep(0,N) # p-value for the GOF test of linear part 2
plk.log=rep(0,N) # p-value for the GOF test of logistic part
plk.L1L2=rep(0,N)
plk.L1L2P=rep(0,N)
R=1000 #total number of realziations for each GOF test

for (i in 1:N){
What=matrix(0, n, R) # W_hat for one random draws
loopdraw=test[test$loop==i,]
loopcov=testcov[testcov$loop==i,]

#LINK FUNCTION tests
glk.L1=cumresmixtureLin1(dat=loopdraw, ycol=5, x1col=3, sigcol=7, indi=11,
 datcov=loopcov, R=R)
glk.L2=cumresmixtureLin2(dat=loopdraw, ycol=5, x1col=3, sigcol=7, indi=12,
 datcov=loopcov, R=R)
glk.log= cumresmixtureP(dat=loopdraw, ycol=5, x1col=3, sigcol=7, indi=10,
 datcov=loopcov, R=R)

#test of linear comp1--link fucntion
plk.L1[i]=sum(glk.L1$What.sup>glk.L1$W.sup)/R

#test of linear comp2--link fucntion
plk.L2[i]=sum(glk.L2$What.sup>glk.L2$W.sup)/R

#test of logistic part--link fucntion
plk.log[i]=sum(glk.log$What.sup>glk.log$W.sup)/R
```

```
#Joint test of linear components (L1+L2)--link fucntion
Wnew=glk.L1$W.sup + glk.L2$W.sup
What=glk.L1$What.sup + glk.L2$What.sup
plk.L1L2[i]=sum(What>Wnew)/R

#Joint test of linear components and logistic regresssion (L1+L2+P)
--link fucntion
Wnew=glk.L1$W.sup + glk.L2$W.sup + glk.log$W.sup
What=glk.L1$What.sup + glk.L2$What.sup + glk.log$What.sup
plk.L1L2P[i]=sum(What>Wnew)/R
}
sum(plk.L1<0.05)/N
sum(plk.L2<0.05)/N
sum(plk.log<0.05)/N
sum(plk.L1L2<0.05)/N
sum(plk.L1L2P<0.05)/N
```

D.2 R code for simulation to evaluate the performance of the proposed GOF tests for a random effects two-component mixture model

```
#############################################################################
#R code for computing p-values for combined data including class 1 and 2
#based on modified cumulative residual test for mixture models##########
#Both functional form test and link function test ######################
#TwoCompMixed4a: covariates x_ij and xx2_i in the linear reg components#
#    covariates xx_i and xx2_i in the logistic reg for mix prop
#############################################################################
seedno=rnorm(1)*1000000
library(gof)
library(MASS)
library(data.table)
library(zoo)
library(mnormt)
R=500
#############################################################################
############read csv data exported from SAS###############################
test <- read.table("C:/Documents and Settings/JSHEN3/My Documents/Beijing
Olympic Thesis/data for R/testmix2.csv", header = TRUE, sep=",")
testcov <- read.table("C:/Documents and Settings/JSHEN3/My Documents/
Beijing Olympic Thesis/data for R/covmix2.csv", header = TRUE, sep=",")

#############################################################################
#####define a simple function cumresmixed1 for the first linear component
#############################################################################
cumresmixed1=function(dat, datcov, indi, R){
temp=dat
```

```
tempcov=datcov


#before sort data
n=max(temp$subject) #n: the number of subjects in the data
t=max(temp$t) #t: the number of repeated measures for each subject
sigma=temp[1,20]
sigma1=temp[1,21]
be1=temp[5][[1]] #be1: residual for the 1st component, length of (n*t)X1
be2=temp[6][[1]] #be2: residual for the 2nd component, length of (n*t)X1
subj=temp[2][[1]]
bphat=unique(temp[4][[1]])
d1t=cbind(rep(1,n), unique(temp$xx), unique(temp[,7]))
d1x=cbind(rep(1,n*t), temp[,9], temp[,7])
inv=tempcov[,2:12] #8X8 covariance matrix

##################compute the posterior prob ptild or xi#################
e1sq1=(aggregate(be1, by=list(subj), FUN="sum")[2][[1]])^2
e1sq2=aggregate(be1^2, by=list(subj), FUN="sum")[2][[1]]
B1=exp(sigma1^2 * e1sq1/(2*sigma^2*(sigma^2 + t*sigma1^2))
- e1sq2/(2*sigma^2) )
e2sq1=(aggregate(be2, by=list(subj), FUN="sum")[2][[1]])^2
e2sq2=aggregate(be2^2, by=list(subj), FUN="sum")[2][[1]]
B2=exp(sigma1^2 * e2sq1/(2*sigma^2*(sigma^2 + t*sigma1^2))
 - e2sq2/(2*sigma^2) )
ptild=bphat*B1/(bphat*B1 + (1-bphat)*B2) #posterior prob, nX1
ptild.long= rep(ptild, each=3)
sum(ptild.long*be1 +(1-ptild.long)*be2)

#########################compute the score function vectors############
e1overt =aggregate(be1, by=list(subj), FUN="sum")[2][[1]]
```

```
xovert =aggregate(d1x, by=list(subj), FUN="sum")[,2-3]
e1xovert =aggregate(be1*d1x, by=list(subj), FUN="sum")[,2-3]

e2overt =aggregate(be2, by=list(subj), FUN="sum")[2][[1]]
e2xovert =aggregate(be2*d1x, by=list(subj), FUN="sum")[,2-3]

Ualpha=(ptild - bphat)*d1t #same as without repeated measures
sigmafrac= sigma1^2/(sigma^2 + t*sigma1^2)
B1beta1=(-sigmafrac*e1overt*xovert + e1xovert)/(sigma^2)
Ubeta1= ptild*B1beta1
B2beta2=(-sigmafrac*e2overt*xovert + e2xovert)/(sigma^2)
Ubeta2= (1-ptild)*B2beta2
sigmafrac2= sigma1^2 * (2*sigma^2 + t*sigma1^2)/
(sigma^2 + t*sigma1^2)^2
B1sigma= (-sigmafrac2*e1sq1 + e1sq2)/(sigma^3)
B2sigma= (-sigmafrac2*e2sq1 + e2sq2)/(sigma^3)
Usigma= -(t-1)/sigma - sigma/(sigma^2 + t*sigma1^2) + ptild*B1sigma +
(1-ptild)*B2sigma
B1sigma1= sigma1*e1sq1/(sigma^2 + t*sigma1^2)^2
B2sigma1= sigma1*e2sq1/(sigma^2 + t*sigma1^2)^2
Usigma1= -t*sigma1/(sigma^2 + t*sigma1^2) + ptild*B1sigma1 +
(1-ptild)*B2sigma1
tp=t(cbind(Ualpha, Ubeta1, Ubeta2, Usigma, Usigma1))

########compute the partical derivatives of xi over theta, subject-based#
etalpha=bphat/(1-bphat) #etalpha: exp(T'alpha)=P_hat/(1-P_Hat)
xialpha= B1*B2*etalpha*d1t/(etalpha*B1 + B2)^2
xibeta1=  B1*B2*etalpha * B1beta1 /(etalpha*B1 + B2)^2
xibeta2=  B1*B2*etalpha * B2beta2 /(etalpha*B1 + B2)^2
xisigma= B1*B2*etalpha*(B1sigma -B2sigma)/(etalpha*B1 + B2)^2
```

```
xisigma1= B1*B2*etalpha*(B1sigma1 -B2sigma1)/(etalpha*B1 + B2)^2

xialpha.long= xialpha[rep(1:n, rep(t,n)), ]
xibeta1.long= as.matrix(xibeta1)[rep(1:n, rep(t,n)), ]
xibeta2.long= as.matrix(xibeta2)[rep(1:n, rep(t,n)), ]
xisigma.long= rep(xisigma, each=3)
xisigma1.long= rep(xisigma1, each=3)
xitheta.long=cbind(xialpha.long, xibeta1.long, xibeta2.long, xisigma.long,
 xisigma1.long)

###############expand the score function vectors to the size of###########
Ualpha.long= Ualpha[rep(1:n, rep(t,n)),]
Ubeta1.long= as.matrix(Ubeta1)[rep(1:n, rep(t,n)),]
Ubeta2.long= as.matrix(Ubeta2)[rep(1:n, rep(t,n)),]
Usigma.long= rep(Usigma, each=3)
Usigma1.long= rep(Usigma1, each=3)
Utheta.long=cbind(Ualpha.long, Ubeta1.long, Ubeta2.long, Usigma.long,
 Usigma1.long)

#########recombine the data with ptild and score function#################
newtemp=cbind(temp, ptild.long, xitheta.long, Utheta.long)

What=matrix(0,n*t,R)
wh.sup=rep(0, R)

d1=newtemp[order(newtemp[,indi]),] #sort by x(x=1) or xbeta1(indi=23)
e1=d1[5][[1]]
e2=d1[6][[1]]
phat=d1[4][[1]]
T=d1[3][[1]] #xx
```

```
e=d1[25]*e1
d1xsort=cbind(rep(1, n*t), d1$x, d1$xx2)

eta.alpha= e1*d1[,26:28] #xialpha.long
eta.beta1= -d1$ptild.long* d1xsort + e1* d1[,29:31]
eta.beta2= e1*d1[,32:34]
eta.sigma= e1*d1[,35]
eta.sigma1= e1*d1[,36]
eta1=cbind(eta.alpha, eta.beta1, eta.beta2, eta.sigma, eta.sigma1)
eta=apply(eta1, 2, cumsum)

flag=last.var(d1[,indi]) #sort by x(x=1) or xbeta1(indi=23)
w=cumsum(e)/sqrt(n)
mod.w=w*flag
w.sup=max(abs(mod.w))

for (r in 1:R){
set.seed(seedno+r)
z=rnorm(n)      #generate n random variates N(0,1)
zlong=rep(z, each=3)
newtemp2=cbind(newtemp, zlong)

d2=newtemp2[order(newtemp2[,indi]),]
ez=e*d2$zlong
what1=cumsum(ez)

tp2=t(z*t(tp))

what2=eta %*% as.matrix(inv) %*% apply(tp2, 1, sum)
what=(what1 + what2)
```

```
mod.what=(what*flag)/sqrt(n)
What[,r]=mod.what[,1]
wh.sup[r]=max(abs(mod.what))
}
p=sum(wh.sup>=w.sup)/R
list(W=mod.w, What=What, a=what1, b=what2, W.sup=w.sup,
What.sup=wh.sup, p=p)
}
#cumresmixed1(dat=test[test$loop==1,], datcov=testcov[testcov$loop==1, ],
indi=9, R=200)
g=cumresmixed1(dat=test[test$loop==1,], datcov=testcov[testcov$loop==1, ],
indi=1, R=500)

##########define a simple function###########
##########for the 2nd component################
cumresmixed2=function(dat, datcov, indi, R){
temp=dat
tempcov=datcov

#before sort data
n=max(temp$subject) #n: the number of subjects in the data
t=max(temp$t) #t: the number of repeated measures for each subject
sigma=temp[1,20]
sigma1=temp[1,21]
be1=temp[5][[1]]
be2=temp[6][[1]]
subj=temp[2][[1]]

e1sq1=(aggregate(be1, by=list(subj), FUN="sum")[2][[1]])^2
e1sq2=aggregate(be1^2, by=list(subj), FUN="sum")[2][[1]]
```

```
B1=exp(sigma1^2 * e1sq1/(2*sigma^2*(sigma^2 + t*sigma1^2))
- e1sq2/(2*sigma^2) )

e2sq1=(aggregate(be2, by=list(subj), FUN="sum")[2][[1]])^2
e2sq2=aggregate(be2^2, by=list(subj), FUN="sum")[2][[1]]
B2=exp(sigma1^2 * e2sq1/(2*sigma^2*(sigma^2 + t*sigma1^2))
- e2sq2/(2*sigma^2) )

bphat=unique(temp[4][[1]])
ptild=bphat*B1/(bphat*B1 + (1-bphat)*B2) #posterior prob, nX1
ptild.long= rep(ptild, each=3)
sum(ptild.long*be1 +(1-ptild.long)*be2)

d1t=cbind(rep(1,n), unique(temp$xx), unique(temp[,7]))
d1x=cbind(rep(1,n), temp[,9], temp[,7])

e1overt =aggregate(be1, by=list(subj), FUN="sum")[2][[1]]
xovert =aggregate(d1x, by=list(subj), FUN="sum")[,2-3]
e1xovert =aggregate(be1*d1x, by=list(subj), FUN="sum")[,2-3]

e2overt =aggregate(be2, by=list(subj), FUN="sum")[2][[1]]
e2xovert =aggregate(be2*d1x, by=list(subj), FUN="sum")[,2-3]

Ualpha=(ptild - bphat)*d1t #same as without repeated measures
sigmafrac= sigma1^2/(sigma^2 + t*sigma1^2)
B1beta1=(-sigmafrac*e1overt*xovert + e1xovert)/(sigma^2)
Ubeta1= ptild*B1beta1
B2beta2=(-sigmafrac*e2overt*xovert + e2xovert)/(sigma^2)
Ubeta2= (1-ptild)*B2beta2
sigmafrac2= sigma1^2 * (2*sigma^2 + t*sigma1^2)/
```

```
(sigma^2 + t*sigma1^2)^2
B1sigma= (-sigmafrac2*e1sq1 + e1sq2)/(sigma^3)
B2sigma= (-sigmafrac2*e2sq1 + e2sq2)/(sigma^3)
Usigma= -(t-1)/sigma - sigma/(sigma^2 + t*sigma1^2) + ptild*B1sigma
 + (1-ptild)*B2sigma
B1sigma1= sigma1*e1sq1/(sigma^2 + t*sigma1^2)^2
B2sigma1= sigma1*e2sq1/(sigma^2 + t*sigma1^2)^2
Usigma1= -t*sigma1/(sigma^2 + t*sigma1^2) + ptild*B1sigma1 +
(1-ptild)*B2sigma1
tp=t(cbind(Ualpha, Ubeta1, Ubeta2, Usigma, Usigma1))

#compute the partical derivatives of xi, subject-based
etalpha=bphat/(1-bphat) #etalpha: exp(T'alpha)=P_hat/(1-P_Hat)
xialpha= B1*B2*etalpha*d1t/(etalpha*B1 + B2)^2
xibeta1=  B1*B2*etalpha * B1beta1 /(etalpha*B1 + B2)^2
xibeta2=  B1*B2*etalpha * B2beta2 /(etalpha*B1 + B2)^2
xisigma= B1*B2*etalpha*(B1sigma -B2sigma)/(etalpha*B1 + B2)^2
xisigma1= B1*B2*etalpha*(B1sigma1 -B2sigma1)/(etalpha*B1 + B2)^2

xialpha.long= xialpha[rep(1:n, rep(t,n)), ]
xibeta1.long= as.matrix(xibeta1)[rep(1:n, rep(t,n)), ]
xibeta2.long= as.matrix(xibeta2)[rep(1:n, rep(t,n)), ]
xisigma.long= rep(xisigma, each=3)
xisigma1.long= rep(xisigma1, each=3)
xitheta.long=cbind(xialpha.long, xibeta1.long, xibeta2.long,
xisigma.long, xisigma1.long)

Ualpha.long= Ualpha[rep(1:n, rep(t,n)),]
Ubeta1.long= as.matrix(Ubeta1)[rep(1:n, rep(t,n)),]
Ubeta2.long= as.matrix(Ubeta2)[rep(1:n, rep(t,n)),]
```

```
Usigma.long= rep(Usigma, each=3)
Usigma1.long= rep(Usigma1, each=3)
Utheta.long=cbind(Ualpha.long, Ubeta1.long, Ubeta2.long, Usigma.long,
 Usigma1.long)

newtemp=cbind(temp, ptild.long, xitheta.long, Utheta.long)

inv=tempcov[,2:12] #8X8 covariance matrix

What=matrix(0,n*t,R)
wh.sup=rep(0, R)

d1=newtemp[order(newtemp[,indi]),]
e1=d1[5][[1]]
e2=d1[6][[1]]
phat=d1[4][[1]]
T=d1[3][[1]]
e=(1-d1[25])*e2
d1xsort=cbind(rep(1, n*t), d1$x, d1$xx2) #sorted x with intercet

eta.alpha= -e2*d1[,26:28] #xialpha.long
eta.beta1= -e2* d1[,29:31]
eta.beta2= -(1- d1$ptild.long)*d1xsort - e2*d1[,32:34]
eta.sigma= -e2*d1[,35]
eta.sigma1= -e2*d1[,36]

eta1=cbind(eta.alpha, eta.beta1, eta.beta2, eta.sigma, eta.sigma1)
eta=apply(eta1, 2, cumsum)

flag=last.var(d1[,indi])
```

```
w=cumsum(e)/sqrt(n)
mod.w=w*flag
w.sup=max(abs(mod.w))

for (r in 1:R){
set.seed(seedno+r)
z=rnorm(n)      #generate n random variates N(0,1)
zlong=rep(z, each=3)
newtemp2=cbind(newtemp, zlong)

d2=newtemp2[order(newtemp2[,indi]),]
ez=e*d2$zlong
what1=cumsum(ez)

tp2=t(z*t(tp))

what2=eta %*% as.matrix(inv) %*% apply(tp2, 1, sum)
what=(what1 + what2)
mod.what=(what*flag)/sqrt(n)
What[,r]=mod.what[,1]
wh.sup[r]=max(abs(mod.what))
}
p=sum(wh.sup>=w.sup)/R
list(W=mod.w, What=What, a=what1, b=what2, W.sup=w.sup,
What.sup=wh.sup, p=p)
}
g=cumresmixed2(dat=test[test$loop==2,], datcov=testcov[testcov$loop==2, ],
 indi=9, R=500)
```

```
#########################################################################
#define a simple function cumresmixedlog for the logistic regression of
##### mixting proportion########
#########################################################################
cumresmixedlog=function(dat, datcov, indi, R){
temp=dat
tempcov=datcov

#before sort data
n=max(temp$subject) #n: the number of subjects in the data
t=max(temp$t) #t: the number of repeated measures for each subject
sigma=temp[1,20]
sigma1=temp[1,21]
be1=temp[5][[1]]
be2=temp[6][[1]]
subj=temp[2][[1]]
bphat=unique(temp[4][[1]])
inv=tempcov[,2:12] #8X8 covariance matrix

d1t=cbind(rep(1,n), unique(temp$xx), unique(temp[,7]))
xx2=temp[,7]
d1x=cbind(rep(1,n), temp[,9], temp[,7])

################compute the posterior prob ptild or###################
e1sq1=(aggregate(be1, by=list(subj), FUN="sum")[2][[1]])^2
e1sq2=aggregate(be1^2, by=list(subj), FUN="sum")[2][[1]]
B1=exp(sigma1^2 * e1sq1/(2*sigma^2*(sigma^2 + t*sigma1^2))
- e1sq2/(2*sigma^2) )
e2sq1=(aggregate(be2, by=list(subj), FUN="sum")[2][[1]])^2
```

```
e2sq2=aggregate(be2^2, by=list(subj), FUN="sum")[2][[1]]
B2=exp(sigma1^2 * e2sq1/(2*sigma^2*(sigma^2 + t*sigma1^2))
- e2sq2/(2*sigma^2) )
ptild=bphat*B1/(bphat*B1 + (1-bphat)*B2) #posterior prob, nX1

################compute the score function vectors###############
e1overt =aggregate(be1, by=list(subj), FUN="sum")[2][[1]]
xovert =aggregate(d1x, by=list(subj), FUN="sum")[,2-3]
e1xovert =aggregate(be1*d1x, by=list(subj), FUN="sum")[,2-3]

e2overt =aggregate(be2, by=list(subj), FUN="sum")[2][[1]]
e2xovert =aggregate(be2*d1x, by=list(subj), FUN="sum")[,2-3]

Ualpha=(ptild - bphat)*d1t #same as without repeated measures
sigmafrac= sigma1^2/(sigma^2 + t*sigma1^2)
B1beta1=(-sigmafrac*e1overt*xovert + e1xovert)/(sigma^2)
Ubeta1= ptild*B1beta1
B2beta2=(-sigmafrac*e2overt*xovert + e2xovert)/(sigma^2)
Ubeta2= (1-ptild)*B2beta2
sigmafrac2= sigma1^2 * (2*sigma^2 + t*sigma1^2)/(sigma^2 + t*sigma1^2)^2
B1sigma= (-sigmafrac2*e1sq1 + e1sq2)/(sigma^3)
B2sigma= (-sigmafrac2*e2sq1 + e2sq2)/(sigma^3)
Usigma= -(t-1)/sigma - sigma/(sigma^2 + t*sigma1^2) + ptild*B1sigma
+ (1-ptild)*B2sigma
B1sigma1= sigma1*e1sq1/(sigma^2 + t*sigma1^2)^2
B2sigma1= sigma1*e2sq1/(sigma^2 + t*sigma1^2)^2
Usigma1= -t*sigma1/(sigma^2 + t*sigma1^2) + ptild*B1sigma1 +
(1-ptild)*B2sigma1
tp=cbind(Ualpha, Ubeta1, Ubeta2, Usigma, Usigma1) #temp: 4Xn matrix
```

```
etalpha=bphat/(1-bphat) #etalpha: exp(T'alpha)=P_hat/(1-P_Hat)
xialpha= B1*B2*etalpha*d1t/(etalpha*B1 + B2)^2
xibeta1=  B1*B2*etalpha * B1beta1 /(etalpha*B1 + B2)^2
xibeta2=  B1*B2*etalpha * B2beta2 /(etalpha*B1 + B2)^2
xisigma= B1*B2*etalpha*(B1sigma -B2sigma)/(etalpha*B1 + B2)^2
xisigma1= B1*B2*etalpha*(B1sigma1 -B2sigma1)/(etalpha*B1 + B2)^2
newtemp=cbind(unique(temp$xx), bphat, ptild, xialpha, xibeta1, xibeta2,
 xisigma, xisigma1, tp,unique(temp$xx2), unique(temp$talpha))
What=matrix(0,n,R)
wh.sup=rep(0, R)
d1=newtemp[order(newtemp[,indi]),]
phat=d1[2][[1]]
T=d1[1][[1]]
e= d1[,3]-d1[,2]
d1tsort=cbind(rep(1, n), T, d1[,26])
tpsort=d1[,15:25]

eat= phat/(1-phat)
eta.alpha= d1[,4:6] - eat*d1tsort/((1+eat)^2) #xialpha.long
eta.beta1= d1[,7:9]
eta.beta2= d1[,10:12]
eta.sigma= d1[,13]
eta.sigma1= d1[,14]
eta1=cbind(eta.alpha, eta.beta1, eta.beta2, eta.sigma, eta.sigma1)
eta=apply(eta1, 2, cumsum)

flag=last.var(d1[,indi])
w=cumsum(e)/sqrt(n)
mod.w=w*flag
w.sup=max(abs(mod.w))
```

```
for (r in 1:R){
set.seed(seedno+r)
z=rnorm(n)      #generate n random variates N(0,1)
ez=e*z
what1=cumsum(ez)

tp2=t(z*tpsort) #must be sorted tp by xx , then times z

what2=eta %*% as.matrix(inv) %*% apply(tp2, 1, sum)
what=(what1 + what2)
mod.what=(what*flag)/sqrt(n)
What[,r]=mod.what[,1]
wh.sup[r]=max(abs(mod.what))
}
p=sum(wh.sup>=w.sup)/R
list(W=mod.w, What=What, a=what1, b=what2, W.sup=w.sup,
 What.sup=wh.sup, p=p)
}
g=cumresmixedlog(dat=test[test$loop==2,], datcov=testcov[testcov$loop==2, ],
 indi=27, R=500)

############################################################
###Run Simulations N times using function cumresmixed#######
############################################################
N=max(test$loop) # total number of loops for simulation
#R=500 #total number of realziations for each GOF test

pL1=rep(0,N) # p-value for the GOF test of linear part 1
pL2=rep(0,N) # p-value for the GOF test of linear part 2
```

```
plog=rep(0,N) # p-value for the GOF test of logistic part
pL1L2=rep(0,N) # p-value for the joint GOF test of linear parts (1 & 2)
pL1L2P=rep(0,N)

for (i in 1:N){
loopdraw=test[test$loop==i,] #subset data for loop i and draw d
loopcov=testcov[testcov$loop==i,] #the covariance matrix for loop i

#########################################################################
#functional form test for Linear comp 1(gL2) and 2(gL2) and logistic
#part(gP) use X_i<x for linear components and T_i<r for logistic
#regression of P#####
#########################################################################
gL1=cumresmixed1(dat=loopdraw, datcov=loopcov, indi=9, R=R)
gL2=cumresmixed2(dat=loopdraw, datcov=loopcov, indi=9, R=R)
glog=cumresmixedlog(dat=loopdraw, datcov=loopcov, indi=1, R=R)

#########################################################################
#link function testfor Linear comp 1 (gL2) and 2 (gL2) and logistic
#part (gP) use X_ibeta_1<r for linear components and T_ialpha<r for
#logistic regression
#########################################################################
#gL1=cumresmixed1(dat=loopdraw, datcov=loopcov, indi=23, R=R)
#gL2=cumresmixed2(dat=loopdraw, datcov=loopcov, indi=24, R=R)
#glog=cumresmixedlog(dat=loopdraw, datcov=loopcov, indi=27, R=R)

#test of linear comp1
pL1[i]=sum(gL1$What.sup>gL1$W.sup)/R

#test of linear comp2
```

```
pL2[i]=sum(gL2$What.sup>gL2$W.sup)/R

#test of logistic part
plog[i]=sum(glog$What.sup>glog$W.sup)/R

#Joint test of linear components (L1+L2)
Wnew=gL1$W.sup + gL2$W.sup
What=gL1$What.sup + gL2$What.sup
pL1L2[i]=sum(What>Wnew)/R

Wnew=gL1$W.sup + gL2$W.sup + glog$W.sup
What=gL1$What.sup + gL2$What.sup + glog$What.sup
pL1L2P[i]=sum(What>Wnew)/R
}
sum(pL1<0.05)/N
sum(pL2<0.05)/N
sum(plog<0.05)/N
sum(pL1L2<0.05)/N
sum(pL1L2P<0.05)/N

sum(pL1<0.05, na.rm=T)/sum(pL1>=0, na.rm=T)
sum(pL2<0.05, na.rm=T)/sum(pL2>=0, na.rm=T)
sum(plog<0.05, na.rm=T)/sum(plog>=0, na.rm=T)
sum(pL1L2<0.05, na.rm=T)/sum(pL1L2>=0, na.rm=T)
sum(pL1L2P<0.05, na.rm=T)/sum(pL1L2P>=0, na.rm=T)


#####################################################################
Run Simulations N times using function cumresmixed###############
#####################################################################
```

```
N=max(test$loop) # total number of loops for simulation
#R=500 #total number of realziations for each GOF test

pL1=rep(0,N) # p-value for the GOF test of linear part 1
pL2=rep(0,N) # p-value for the GOF test of linear part 2
plog=rep(0,N) # p-value for the GOF test of logistic part
pL1L2=rep(0,N) # p-value for the joint GOF test of linear parts (1 & 2)
pL1L2P=rep(0,N)
# p-value for the joint GOF test of both logistic part and linear parts
(1 & 2 & P)

for (i in 1:N){
loopdraw=test[test$loop==i,] #subset data for loop i and draw d
loopcov=testcov[testcov$loop==i,] #the covariance matrix for loop i

#####################################################################
#functional form test for Linear comp 1(gL2) and 2(gL2) and
#logistic part(gP) use X_i<x for linear components and T_i<r for
#logistic regression of P#####
#####################################################################
#gL1=cumresmixed1(dat=loopdraw, datcov=loopcov, indi=9, R=R)
#gL2=cumresmixed2(dat=loopdraw, datcov=loopcov, indi=9, R=R)
#glog=cumresmixedlog(dat=loopdraw, datcov=loopcov, indi=1, R=R)

#####################################################################
    #link function tests
#####################################################################
gL1=cumresmixed1(dat=loopdraw, datcov=loopcov, indi=23, R=R)
gL2=cumresmixed2(dat=loopdraw, datcov=loopcov, indi=24, R=R)
```

```
glog=cumresmixedlog(dat=loopdraw, datcov=loopcov, indi=27, R=R)

#test of linear comp1
pL1[i]=sum(gL1$What.sup>gL1$W.sup)/R

#test of linear comp2
pL2[i]=sum(gL2$What.sup>gL2$W.sup)/R

#test of logistic part
plog[i]=sum(glog$What.sup>glog$W.sup)/R

#Joint test of linear components (L1+L2)
Wnew=gL1$W.sup + gL2$W.sup
What=gL1$What.sup + gL2$What.sup
pL1L2[i]=sum(What>Wnew)/R

#Joint test of linear components and logistic regresssion  (L1+L2+P)
Wnew=gL1$W.sup + gL2$W.sup + glog$W.sup
What=gL1$What.sup + gL2$What.sup + glog$What.sup
pL1L2P[i]=sum(What>Wnew)/R
}
sum(pL1<0.05)/N
sum(pL2<0.05)/N
sum(plog<0.05)/N
sum(pL1L2<0.05)/N
sum(pL1L2P<0.05)/N

sum(pL1<0.05, na.rm=T)/sum(pL1>=0, na.rm=T)
sum(pL2<0.05, na.rm=T)/sum(pL2>=0, na.rm=T)
sum(plog<0.05, na.rm=T)/sum(plog>=0, na.rm=T)
```

```
sum(pL1L2<0.05, na.rm=T)/sum(pL1L2>=0, na.rm=T)
sum(pL1L2P<0.05, na.rm=T)/sum(pL1L2P>=0, na.rm=T)
```

Bibliography

N. Atienza, J. Garcia-Heras, J.M. Munoz-Pichardo, and R. Villa. On the consistency of mle in finite mixture models of exponential families. *Journal of Statistical Planning and Inference*, 137:496–505, 2007.

G.A. Baker. Empiric investigation of a test of homogeneity for populations composed of normal distributions. *J. Amer. Statist. Assoc.*, 53:551–57, 1958.

K. Bandeen-Roche, D.L. Miglioretti, S. L. Zeger, and P. J. Rathouz. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92:1375–1386, 1997.

K.N. Berk and P.A. Lachenbruch. Repeated measures with zeros. *Statistical Methods in Medical Research*, 11:303–316, 2002.

D.A. Binder. Bayesian cluster analysis. *Biometrika*, 65:31–38, 1978.

H. Bozdogan. Model selection and akaikes information criterion (aic): the general theory and its analytical extensions. *Psychometrika*, 52:345–370, 1987.

L.A. Currie. Limits for qualitative detection and quantitative determination. application to radiochemistry. *Anal. Chem.*, 40(3), 1968.

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

E. Dietz. Limits for qualitative detection and quantitative determination. application to radiochemistry. *Anal. Chem.*, 40(3), 1968.

E. Dietz and D. Bohning. Statistical inference based on a general model of unobserved heterogeneity. In R Gilchrist G Tutz L Fahrmeir, F Francis, editor, *Advances in GLIM and Statistical Modeling. Lecture Notes in Statistics*, pages 75–82, Berlin, 1996. Springer.

A. Dmitrienko. *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Institute Inc., 2005.

M. Erali and D. R. Hillyard. Evaluation of the ultrasensitive roche amplicor hiv-1 monitor assay for quantitation of human immunodeficiency virus type 1 rna. *Journal of Clinical Microbiology*, pages 792–795, 1999.

B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Chapman and Hall, London and New York, 1981.

E.B. Fowlkes. Some methods for studying the mixture of two normal (lognormal) distributions. *J Am Stat Assoc*, pages 561–575, 1979.

J.A. Hartigan. A failure of likelihood asymptotics for normal mixtures, in:. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman*, 2:807–810, 1985.

K.K. Holst. Model-diagnostics based on cumulative residuals. http://cran.r-project.org/web/packages/gof/gof.pdf, 2011.

J. P. Hughes. Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55, 1999.

H. Jacqmin-Gadda, R. Thiebaut, G. Chene, and D. Commenges. Analysis of left-censored longitudinal data with application to viral load in hiv infection. *Biostatistics*, 1:355–368, 2000.

N. L. Johnson. Some simple tests of mixtures with symmetrical components. *Communications in Statistics*, 1973.

B. Joos, B. Ledergerber, M. Flepp, J.D. Bettex, R. Lthy, and W. Siegenthaler. Comparison of high-pressure liquid chromatography and bioassay for determination of ciprofloxacin in serum and urine. *Antimicrob Agents Chemother*, 27(3):353–356, 1985.

B. Leroux. Consistent estimation of a mixing distribution. *Ann. Statist.*, 20:1350–1360, 1992.

X. Li, H. Chu, J.E. Gallant, D.R. Hoover, W.J. Mack, J.S. Chmiel, and A. Muoz. Bimodal virological response to antiretroviral therapy for hiv infection: an application using a mixture model with left censoring. *Journal of Epidemiology Community Health*, 60:811–818, 2006.

D. Y. Lin, L. J. Wei, and Z. L. Ying. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80:557–572, 1993.

D.Y. Lin, L.J. Wei, and Z. Ying. Model-checking techniques based on cumulative residuals. *Biometrics*, 58(1):1–12, March 2002.

B.G. Lindsay and K. Roeder. Residual diagnostics for mixture models. *Journal of the American Statistical Association*, 87(419):785–794, Sept 1992.

R.H. Lyles, C. M. Lyles, and D. J. Taylor. Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. *J. R. Stat. Soc.*, 2000.

G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, New York, 2000.

G.J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, 36:318–324, 1987.

L. H. Moulton and N. A. Halsey. A mixture model with detection limit for regression analyses of antibody response to vaccine. *Biometrics*, 51:1570–78, 1995.

L. H. Moulton and N. A. Halsey. A mixed gamma model for regression analyses of quantitative assay data. *Vaccine*, 14:1154–58, 1996.

L. H. Moulton, F. C. Curriero, and P. F. Barroso. Mixture models for quantitative hiv rna data. *Statistical Methods in Medical Research*, 11:317–325, 2002.

P. Ngom. Goodness-of-fit in mixture models using power-divergence statistics. *Global Journal of Pure and Applied Mathematics*, 1(1):27–40, 2005.

M. K. Olsen and J. L. Schafer. A two-part random-effects model for semincontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745, 2001.

Z. Pan and D.Y. Lin. Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61:1000–1009, December 2005.

C. A. III Pope and D. W. Dockery. Health effects of fine particulate air pollution: Lines that connect. *J. Air and Waste Manage. Assoc.*, 56:709742, 2006.

S. W. Raudenbush, M. Yang, and M. Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics*, 9:141–157, 2000.

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

S. Richardson and P.J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society, Ser.B*, 59(4):731–792, 1997.

SAS. Sas online document, v9.1.3. `http://support.sas.com/documentation/onlinedoc/91pdf/`, 2011.

C. F. Spiekerman and D. Y. Lin. Checking the marginal cox model for correlated failure time data. *Biometrika*, 83:143–156, 1996.

J.Q. Su and L.J. Wei. A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, 86(414):420–426, June 1991.

L. Su, B.D.M. Tom, and V.T. Farewell. Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, 10(2):374–89, 2009.

E. Susko. Weighted tests of homogeneity for testing the number of components in a mixture. *Computational Statistics and Data Analysis*, 41:367–378, 2003.

D.J. Taylor, L.L. Kupper, S. M. Rappaport, and R. H. Lyles. A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics*, 57:681–688, 2001.

R. Thiebaut and H. Jacqmin-Gadda. Mixed models for longitudinal left-censored repeated measures. *Computer Methods and Programs in Biomedicine*, 74:255–260, 2004.

T. J. Thompson, P. J. Smith, and J. P. Boyle. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Appl. Statist.*, 47:393–404, 1998.

J. A. Tooze, G. K. Grunwald, and R.H. Jones. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, 11:341–355, 2002.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

C.P. Wang, C.H. Brown, and K. Bandeen-Roche. Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behaviour. *Journal of the American Statistical Association*, 100(471):1054–1076, September 2005.

J.H. Wolfe. A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Technical Report STB 72-2, Naval Personnel and Training Research Laboratory, Technical Bulletin, San Diego, California, USA, 1971.

K.K.W. Yau, A.H. Lee, and A.S.K. Ng. Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics and Data Analysis*, 41:359–366, 2003.

S.L. Zeger and K. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130, 1986.

Vita

Junwu Shen

Education

2011 Ph. D. in Biostatistics, University of Medicine and Dentistry of New Jersey, Piscataway, NJ

2006 M.S. in Statistics, The Pennsylvania State University, University Park, PA

2003 Ph.D. in Materials Science and Engineering, The Pennsylvania State University, University Park, PA

1996 M.S. in Mechanical Engineering, Tsinghua University, Beijing, China

1992 B.S. in Materials Science and Engineering, Beijing Institute of Technology, Beijing, China

Employment

2009-2011 Senior Research Statistician, Merck Research Laboratory, Kenilworth, NJ

2008-2009 Senior Research Statistician, Schering-Plough Research Institute, Kenilworth, NJ

2006-2008 Research Statistician, Schering-Plough Research Institute, Kenilworth, NJ

2006 Intern, Quintiles, Inc., Kansas City, MO

2005 Graduate Statistical Consultant, Department of Statistics, The Pennsylvania State University, University Park, PA

Publications

1. J Shen and SE Lu, "Goodness-of-Fit Tests of Finite Mixture Regression Models Based on Cumulative Pseudo-Residuals", *Joint Statistical Meetings (JSM)*, August 2011, Miami, FL.

2. JM Williams, KK Gandhi, SE Lu, S Kumar, J Shen and J Foulds, Higher Nicotine Levels in Schizophrenia Compared with Controls after Smoking a Single Cigarette", *Nicotine and Tobacco Research*, 2010, Vol. 12 (8), p855-859.