Bayesian Inference on Longitudinal Semi-continuous Substance Abuse/Dependence Symptoms Data

Dongyuan Xing
University of South Florida, dyxing2010@gmail.com

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the Biostatistics Commons

Scholar Commons Citation

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Bayesian Inference on Longitudinal Semi-continuous Substance Abuse/Dependence Symptoms Data

by

Dongyuan Xing

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
Department of Epidemiology and Biostatistics
College of Public Health
University of South Florida

Major Professor: Yangxin Huang, Ph.D.
            Henian Chen, M.D., Ph.D.
            Yiliang Zhu, Ph.D.
            Getachew A. Dagne, Ph.D.
            Julie Baldwin, Ph.D.

Date of Approval: June 9, 2015

Keywords: Two-part mixed-effects model, Substance abuse/dependence symptoms data, Bayesian analysis, Skewed distributions, Semi-continuous longitudinal data

Copyright © 2015, Dongyuan Xing
Dedication

This work is dedicated to my family, who have been supportive of my research all these years.
Acknowledgments

I would like to express my gratitude to everyone who supported me throughout the completion of this dissertation.

I would like to especially thank my major advisor, Dr. Yangxin Huang, for his guidance and support during the completion of this dissertation. Despite his busy working and research schedule, Dr. Huang has always been available to discuss my research and provide prompt responses to my questions. His invaluable guidance, constructive criticism and encouragement were crucial in the completion of this dissertation.

I am also very grateful to my dissertation committee members, Drs. Henian Chen, Yiliang Zhu, Getachew A. Dagne and Julie Baldwin for their insightful suggestions and comments. Special thanks are given to Dr. Chen for providing me the data used in my research.

I would also like to thank all of the faculty members and staff in the Department of Epidemiology and Biostatistics at College of Public Health for their help during my studies at the department.
Table of Contents

List of Tables .............................................................................................................................. iii

List of Figures............................................................................................................................. iv

Abstract...................................................................................................................................... v

Chapter 1: Introduction ............................................................................................................ 1
  1.1 Background of substance use................................................................................... 1
  1.2 Statistical inference in substance use................................................................. 3
  1.3 Skew-elliptical distributions..................................................................................... 5
    1.3.1 Skew-t distribution . . . . . . . . . . . . . . . . . . . . . . . . . 6
    1.3.2 Skew-normal distribution . . . . . . . . . . . . . . . . . . . . . . . . . 8
  1.4 Specific aims ............................................................................................................ 9

Chapter 2: Motivating substance use data................................................................................. 11

Chapter 3: Bayesian approach to linear mixed-effects models with application to
skewed-longitudinal alcohol data ............................................................................. 14
  3.1 Introduction............................................................................................................. 15
  3.2 Statistical models and methods............................................................................... 16
    3.2.1 Linear mixed-effects model with multivariate SN distribution . . . . . . . 16
    3.2.2 Three counterpart models . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
    3.2.3 Bayesian inference for parameter estimation . . . . . . . . . . . . . . . 19
  3.3 Alcohol data analysis............................................................................................... 20
    3.3.1 Specific models and implementation . . . . . . . . . . . . . . . . . . . . 20
    3.3.2 Comparison of model fitting results . . . . . . . . . . . . . . . . . . . . . 23
  3.4 Simulation study ..................................................................................................... 28
  3.5 Concluding discussion.............................................................................................. 30

Chapter 4: Bayesian inference on two-part mixed-effects models using skew distributions,
with application to longitudinal semi-continuous alcohol data................................ 32
  4.1 Introduction............................................................................................................. 32
  4.2 Two-part mixed-effects models with ST distribution .............................................. 36
    4.2.1 Model specification . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
    4.2.2 Simultaneous Bayesian inferential approach . . . . . . . . . . . . . . . . 38
  4.3 Alcohol data analysis............................................................................................... 40
    4.3.1 Specific models and implementation . . . . . . . . . . . . . . . . . . . . 40
    4.3.2 Comparison of model fitting results . . . . . . . . . . . . . . . . . . . . . 42
  4.4 Simulation study ..................................................................................................... 46
  4.5 Concluding discussion.............................................................................................. 48

Chapter 5: Bayesian inference on bivariate semi-continuous mixed-effects models with
application to longitudinal substance use data........................................................ 53
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>5.2</td>
<td>Joint models and Bayesian inference</td>
<td>56</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Joint models with ST distribution</td>
<td>56</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Simultaneous Bayesian inferential approach for joint modeling</td>
<td>58</td>
</tr>
<tr>
<td>5.3</td>
<td>Substance use data analysis</td>
<td>59</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Specific models and implementation</td>
<td>59</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Comparison of model fitting results</td>
<td>61</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Estimation results based on ST model</td>
<td>65</td>
</tr>
<tr>
<td>5.4</td>
<td>Simulation study</td>
<td>65</td>
</tr>
<tr>
<td>5.5</td>
<td>Concluding discussion</td>
<td>67</td>
</tr>
<tr>
<td>6</td>
<td>Chapter 6: Overall discussion and conclusions</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Appendices</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Appendix A: WinBUGS code for Chapter 4 - model ST1</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Appendix B: WinBUGS code for Chapter 4- simulation code</td>
<td>82</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 Bivariate frequency distribution of AADS and MADS . . . . . . . . . . 13
Table 3.1 Summary of parameter estimates based on Models N, SN, LN and G . . . . 27
Table 3.2 Summary of Monte Carlo simulation results for MC estimates of fixed-effects
\( \beta \) as well as MSE for Models N, SN, LN and G . . . . . . . . . . . . . . . . . 30
Table 4.1 Summary of deviance information criterion (DIC) based on Models N, SN
and ST under random-effects 2 \( \times \) 2 and 3 \( \times \) 3 structure. . . . . . . . . . . 44
Table 4.2 Summary of parameter estimates based on Models N, SN and ST under
random-effects 2 \( \times \) 2 and 3 \( \times \) 3 structures . . . . . . . . . . . . . . . . . . . . 45
Table 4.3 Summary of Monte Carlo simulation results for MC estimates of fixed-effects
as well as MSE for Models ST, SN and N . . . . . . . . . . . . . . . . . . . . . . 48
Table 4.4 Summary of parameter estimates based on Models N, SN and ST under
common random-effects structure . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
Table 4.5 Summary of parameter estimates based on Models SN1 and SN3 . . . . . . . 52
Table 5.1 Summary of parameter estimates based on Models N, SN and ST . . . . . . . 63
Table 5.2 Summary of Monte Carlo simulation results for MC estimates of fixed-effects
as well as MSE for Models ST, SN and N . . . . . . . . . . . . . . . . . . . . . . 66
Table 6.1 Summary of Monte Carlo simulation results for MC estimates of fixed-effects
\( \beta \) as well as MSE for one-part SN-LMM and two-part SN-LMM . . . . . . . 71
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Histograms of AADS (a,b) and MADS (c, d)</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>Convergence diagnostics with three Markov chains</td>
<td>23</td>
</tr>
<tr>
<td>3.2</td>
<td>Quantile-quantile (Q-Q) normal plots for residuals</td>
<td>25</td>
</tr>
<tr>
<td>3.3</td>
<td>Individual fitted curves of AADS scale for 4 representative patients</td>
<td>26</td>
</tr>
<tr>
<td>3.4</td>
<td>Prediction of change in AADS scale from Models N and SN</td>
<td>28</td>
</tr>
<tr>
<td>4.1</td>
<td>Convergence diagnostics with three Markov chains</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>Probability of occurrence of AADS and severity of AADS by gender given its occurrence</td>
<td>46</td>
</tr>
<tr>
<td>4.3</td>
<td>Probability of occurrence of AADS and severity of AADS by gene-5-HTTLPR given its occurrence</td>
<td>47</td>
</tr>
<tr>
<td>5.1</td>
<td>Convergence diagnostics with three Markov chains</td>
<td>61</td>
</tr>
<tr>
<td>5.2</td>
<td>Boxplots of asymmetry parameters for Models ST and SN</td>
<td>64</td>
</tr>
</tbody>
</table>
Abstract

Substance use data such as alcohol drinking often contain a high proportion of zeros. In studies examining the alcohol consumption in college students, for instance, many students may not drink in the studied period, resulting in a number of zeros. Zero-inflated continuous data, also called semi-continuous data, typically consist of a mixture of a degenerate distribution at the origin (zero) and a right-skewed, continuous distribution for the positive values. Ignoring the extreme non-normality in semi-continuous data may lead to substantially biased estimates and inference. Longitudinal or repeated measures of semi-continuous data present special challenges in statistical inference because of the correlation tangled in the repeated measures on the same subject.

Linear mixed-effects models (LMM) with normality assumption that is routinely used to analyze correlated continuous outcomes are inapplicable for analyzing semi-continuous outcome. Data transformation such as log transformation is typically used to correct the non-normality in data. However, log-transformed data, after the addition of a small constant to handle zeros, may not successfully approximate the normal distribution due to the spike caused by the zeros in the original observations. In addition, the reasons that data transformation should be avoided include: (i) transforming usually provides reduced information on an underlying data generation mechanism; (ii) data transformation causes difficulty in regard to interpretation of the transformed scale; and (iii) it may cause re-transformation bias. Two-part mixed-effects models with one component modeling the probability of being zero and one modeling the intensity of nonzero values have been developed over the last ten years to analyze the longitudinal semi-continuous data. However, log transformation is still needed for the right-skewed nonzero continuous values in the two-part modeling.

In this research, we developed Bayesian hierarchical models in which the extreme non-normality in the longitudinal semi-continuous data caused by the spike at zero and right skewness was accommodated using skew-elliptical (SE) distribution and all of the inferences were carried out through Bayesian approach via Markov chain Monte Carlo (MCMC). The substance abuse/dependence data,
including alcohol abuse/dependence symptoms (AADS) data and marijuana abuse/dependence symptoms (MADS) data from a longitudinal observational study, were used to illustrate the proposed models and methods. This dissertation explored three topics:

First, we presented one-part LMM with skew-normal (SN) distribution under Bayesian framework and applied it to AADS data. The association between AADS and gene serotonin transporter polymorphism (5-HTTLPR) and baseline covariates was analyzed. The results from the proposed model were compared with those from LMMs with normal, Gamma and LN distributional assumptions. Simulation studies were conducted to evaluate the performance of the proposed models. We concluded that the LMM with SN distribution not only provides the best model fit based on Deviance Information Criterion (DIC), but also offers more intuitive and convenient interpretation of results, because it models the original scale of response variable.

Second, we proposed a flexible two-part mixed-effects model with skew distributions including skew-t (ST) and SN distributions for the right-skewed nonzero values in Part II of model under a Bayesian framework. The proposed model is illustrated with the longitudinal AADS data and the results from models with ST, SN and normal distributions were compared under different random-effects structures. Simulation studies are conducted to evaluate the performance of the proposed models.

Third, multivariate (bivariate) correlated semi-continuous data are also commonly encountered in clinical research. For instance, the alcohol use and marijuana use may be observed in the same subject and there might be underlying common factors to cause the dependence of alcohol and marijuana uses. There is very limited literature on multivariate analysis of semi-continuous data. We proposed a Bayesian approach to analyze bivariate semi-continuous outcomes by jointly modeling a logistic mixed-effects model on zero-inflation in either response and a bivariate linear mixed-effects model (BLMM) on the positive values through a correlated random-effects structure. Multivariate skew distributions including ST and SN distributions were used to relax the normality assumption in BLMM. The proposed models were illustrated with an application to the longitudinal AADS and MADS data. A simulation study was conducted to evaluate the performance of the proposed models.
Chapter 1:

Introduction

1.1 Background of substance use

Substance use including alcohol use and drug use is common and costly in the United States and worldwide. Substance use is associated with a wide range of risk behaviors, results in substantial morbidity and mortality, and creates significant public health and socioeconomic problems.

Alcohol is the most commonly used addictive substance in the United States. The World Health Organization (WHO) in 1983 stated that “Problems related to alcohol consumption rank among the world’s major public health problems and constitute serious hazards for human health, welfare and life”. According to the 2013 National Survey on Drug Use and Health (NSDUH) in the United States, slightly more than half (52.2 percent) of Americans aged 12 or older reported being current drinkers of alcohol. This translates to an estimated 136.9 million persons. An estimate of 28.7 million persons (10.9 percent) drove under the influence of alcohol at least once in the past year. Alcohol use is associated with a variety of diseases and disorders and is the direct cause in many deaths. The long-term and excessive consumption of alcohol is the major cause of many cancers. Moreover, more common medical conditions such as hypertension and mental disorders are likely to be aggravated even by occasional and short-term alcohol consumption.

Illicit drug use is another important aspect in substance abuse. The 2013 NSDUH showed that an estimated 24.6 million (9.4 percent) Americans aged 12 or older reported illicit drug use in past month. Driving under the influence of illicit drugs occurred in 3.8% (9.9 million) of Americans aged 12 or older in the last year. The rate was highest in young adults aged 18 to 25. Marijuana was the most commonly used illicit drug. Surveys and clinical studies consistently indicate that drug use has strong associations with alcohol use disorder and other personality disorders (PDs).

Adolescence and young adulthood have been identified as particularly vulnerable periods for alcohol and drug use (Jackson et al., 2005; Sher et al., 1991). Information from the Monitoring
of Future Survey (MFS) shows that the percentage of teen use of tobacco, alcohol and illicit drug increases 150 to 200 percent when students move from 8th to 10th grade (Sloboda, 2005). Many studies have shown that consumption of alcohol steadily increases during the adolescent years until it reaches a peak in the early 20s, followed by a decline thereafter (Johnston et al., 2003; Muthen and Muthen, 2000). The finding that individuals who begin drinking before age 15 are four times more likely to develop alcohol dependence during their lifetimes than are those who begin drinking at age 21 (Grant and Dawson, 1998) indicates that the earlier an individual begins drinking, the greater his or her risk of developing alcohol-use problems in the future.

Numerous cross-sectional and longitudinal studies have been conducted to study the distributional patterns, risk factors and developmental trend of substance use, particularly on alcohol consumption (Bobo et al., 2010; Ledermann, 1956; Skog, 1980; Skog and Rossow, 1985; Kerr et al., 2002; Wilsnack et al., 2006; Wiesner et al., 2007). It has been recognized in recent years that there is considerable interindividual variability in developmental courses of drinking behavior (Muthen and Muthen, 2000; Schulenberg and Maggs, 2002). Clinical trials were conducted to research the prevention and treatment of alcohol abuse/dependence and drug use ( Anton et al., 2006; Chassin et al., 2004; Johnson et al., 2007). Several genomic studies were conducted to establish the association of genome and alcohol use disorder (Agrawal et al., 2009; Enoch et al., 2011; Liu et al., 2004; McHugh et al., 2010; Pagan et al., 2006; Prescott et al., 1999; Van der Zwaluw et al., 2010).

Statistical modeling helps to better understand the distribution and to identify predictive factors that lead to alcohol use and drug use. In order to draw the accurate statistical inference, it is important to understand the unique data features of alcohol and drug uses. Alcohol use and drug use could be assessed in either discrete or continuous measures. Examples of discrete (count) measures include the frequency of heavy drinking in the past month and the number of cigarettes containing marijuana the subject smoked in the last week. Examples of continuous measure include the total amount of alcohol consumption in liters and substance abuse/dependence symptoms scales based on Diagnosis and Statistical Manual of Mental Disorders (DSM) criteria. No matter how alcohol and drug uses are collected, there are typically an abundance of zeros in data. The nonzero positive values are usually right-skewed. The zero-inflated continuous data are also called semi-continuous data.
1.2 Statistical inference in substance use

The extreme non-normality in semi-continuous data leads to a lack of fit of commonly utilized parametric distributions. The traditional approach for modeling semi-continuous data is to apply a log-transformation on the response variable after adding a small constant to the data. It has been noted that log-transformation may not be able to satisfactorily approximate normal distribution due to the mode generated by the zeros in the original observations (Li et al., 2011). The selection of the small constant is arbitrary and could potentially provide rather different results. In addition, the reasons that data transformation should be avoided include: (i) reduced information; (ii) difficulty with regarding interpretation of the transformed scale; and (iii) potential re-transformation bias. One alternative is to categorize the responses into binomial or ordinal scales and then use the logistic or ordinal logit model to analyze the data, but this approach may lose power by discarding considerable outcome information. In recent years, models with Box-cox transformation, Gamma and Weibull distributions have been developed (Kehoe et al., 2012; Rehm et al., 2010; Skog, 1980; Skog and Rossov, 1993). Kehoe et al. (2012) compared the results from models with Gamma, Lognormal (LN) and Weibull distributional assumptions in a cross-sectional study, and recommended the model with Gamma distribution, so did Rehm et al. (2010).

Linear mixed-effects models (LMM) are frequently used to analyze repeatedly measured data due to the flexibility of modeling within-subject correlations and between-subject variations often arisen in longitudinal data. Normality of model errors is a routine assumption in LMM. However, this assumption may lack robustness against departures from normality and/or outliers and, thus, statistical inference and analysis with normal assumption may lead to misleading results (Arellano-Valle et al., 2007; Lin and Lee, 2008; Sahu et al., 2003). Due to non-normality of alcohol and drug uses data, LMM with log-transformation has been commonly used in longitudinal substance studies (Chassin et al., 2004; Colder et al., 2002; Jackson et al., 2005; Wiesner et al., 2007). Generalized LMMs (GLMM) with non-normal distribution such as Gamma distribution have been developed in recently years for modeling longitudinal skewed data. It is noted that GLMMs often involve complicated iterative procedures in estimation which may lead to intensive computation burden and non-convergence issue.

Two-part models, originating in econometrics (Cragg, 1971; Duan et al., 1983a), have been developed extensively in the last three decades to analyze zero-inflated data and have been ap-
plied to scientific fields other than economics such as clinical research and health services. There have been broad developments in zero-inflated models for count data. The models for count data usually mix a Poisson, negative binomial, binomial, multinomial ordinal distribution with a distribution degenerate at zero (Hall, 2000; Lee et al., 2006; Min and Agresti, 2005). Zero-inflated lognormal distribution is typically used for semi-continuous variable. In two-part models, we view the zero-inflated variable as the result of two processes: one binomial process determining whether the positive value occurs, and one continuous process determining the actual value if it is nonzero. Therefore, a two-part model consists of two components, with the first component (Part I) modeling the probability of a response being positive using probit or logistic regression, and the second component (Part II) modeling the conditional mean of the positive values (given positive values occurred) using linear regression. Various methods have been developed for analyzing cross-sectional and longitudinal semi-continuous data (Duan et al., 1983a; Hall and Zhang, 2004; Liu et al., 2008a; Moulton et al., 2002; Olsen and Schafer, 2001; Tooze et al., 2002). Olsen and Schafer (2001) first extended the two-part models developed by Duan et al. (1983a) and Manning et al. (1981) for cross-sectional data to the longitudinal setting by introducing correlated random-effects into the logit and log-normal components, respectively, and applied to longitudinal alcohol data. Manning et al. (2005) proposed a one-part generalized Gamma distribution that include the lognormal, Weibull and exponential distributions as special cases. Liu et al. (2010) extended the generalized Gamma model to a two-part mixed-effects model.

Parameter estimations in two-part modeling can be computationally difficult. For two-part models with uncorrelated random-effects, maximum likelihood estimates can be derived by fitting separate random-effects model for each component. For the correlated two-part mixed-effects model with lognormal distribution on the positive values, Olsen and Schafer (2001) suggested an approximate Fisher scoring procedure based on sixth-order Laplace approximations for obtaining maximum likelihood estimates; whereas Tooze et al. (2002) utilized adaptive Guassian quadrature together with quasi-Newton Raphson to obtain parameter estimates. Several authors have proposed Bayesian approaches to fit the two-part models (Cooper et al., 2007; Ghosh and Albert, 2009; Hatfield et al., 2011; Neelon et al., 2010, 2011; Zhang et al., 2006). For example, Cooper et al. (2007) used a Bayesian approach via Markov chain Monte Carlo (MCMC) to fit a probit-lognormal correlated two-part model on medical cost data.
In the literature, there have been very limited studies on the modeling of multivariate semi-continuous data. Duan et al. (1983a) discussed a four-part model to distinguish inpatient and outpatient costs in the analysis of cross-section medical cost data. The four-part model consists of four equations with two modeling the probability of occurrence of each outcome and the other two modeling the intensity of each outcome given its occurrence. Liu et al. (2008b) extended the four-part model to the repeated measures of medical cost data by introducing a correlated random-effects structure. The joint modeling of the survival time and bivariate zero-inflated PRO data where each PRO item assumed a zero-augmented beta (ZAB) distribution was evaluated by Hatfield et al. (2011, 2012). Zhu et al. (2009) proposed a joint modeling approach in which the alcohol and drug uses (in counts) shared one common latent structure and the model was estimated with an expectation and maximization (EM) algorithm. The unique data features of semi-continuous variable such as excessive zeros and right skewness in our AADS and MADS data increase the complexity in the multivariate analysis.

It would be most effective to use a flexible parametric asymmetric distribution to model the skewed data without the complication of data transformation. The skew distribution family, which accommodates asymmetry in a more flexible manner, can model both positively or negatively skewed data (depending on the sign of the skewness parameter) and includes normal and t distribution as special cases. There is very limited research on the application of the skew distribution in the modeling of semi-continuous data. As a result, we researched the skew distribution including skew-t (ST) or skew-normal (SN) distribution (Arellano-Valle and Genton, 2005; Arellano-Valle et al., 2007; Azzalini and Capitanio, 1999, 2003; Azzalini and Dalla-Valle, 1996; Jara et al., 2008; Sahu et al., 2003) in the application of one-part, two-part and bivariate mixed-effects models of semi-continuous substance use data.

1.3 Skew-elliptical distributions

Different versions of multivariate skew distributions have been proposed and used in literature (Arellano-Valle and Genton, 2005; Genton, 2004; Jara et al., 2008; Muthen and Asparouhov, 2014; Sahu et al., 2003). A new class of distributions by introducing skewness in multivariate elliptically distributions, referred to as skew-elliptical (SE) distributions, were developed in the literature (Genton, 2004; Sahu et al., 2003). The class, which is obtained by using transformation and
conditioning, contains many standard families including the multivariate skew-t (ST) and skew-

normal (SN) distribution as special case.

A \(k\)-dimensional random vector \(Y\) follows a \(k\)-variate SE distribution if its probability density function (pdf) is given by

\[
f(y|\mu, \Sigma; \Delta; m^{(k)}_{\nu}) = 2^k f(y|\mu, A; m^{(k)}_{\nu})P(V > 0),
\]

where \(A = \Sigma + \Delta^2\), \(\mu\) is a location parameter vector, \(\Sigma\) is a \(k \times k\) positive (diagonal) covariance matrix, \(\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_k)\) is a \(k \times k\) skewness matrix with the skewness parameter vector \(\delta = (\delta_1, \delta_2, \ldots, \delta_k)^T\); \(V\) follows the elliptical distribution \(El\left((\Delta A^{-1}(y - \mu), I_k - \Delta A^{-1}\Delta; m^{(k)}_{\nu})\right)\) and the density generator function \(m^{(k)}_{\nu}(u)\) provides the kernel of the original elliptical density and may depend on the parameter \(\nu\). This SE distribution is denoted by \(SE(\mu, \Sigma, \Delta; m^{(k)})\).

Two examples of \(m^{(k)}_{\nu}(u)\), leading to important special cases used throughout the dissertation, are \(m_{\nu}(u) = (1 + u/\nu)^{-(\nu+k)/2}\) and \(m_{\nu}(u) = \exp(-u/2)\), where \(\nu > 0\). These expressions lead to the multivariate ST and SN distributions, respectively. In the former case, \(\nu\) corresponds to the degrees of freedom parameter.

For completeness, this section briefly summarizes the multivariate ST and SN distributions introduced by Sahu et al. (2003) and Muthen and Asparouhov (2014) to be suitable for a Bayesian inference since it is built using the conditional method; see Sahu et al. (2003) and Muthen and Asparouhov (2014) for detailed discussions on properties of ST and SN distributions.

Assume a \(k\)-dimensional random vector \(Y\) follows a \(k\) variate ST or SN distribution with location vector \(\mu\), \(k \times k\) positive (diagonal) covariance matrix \(\Sigma\), \(k \times k\) skewness diagonal matrix \(\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_k)\), skewness parameter vector \(\delta = (\delta_1, \delta_2, \ldots, \delta_k)^T\) and degrees of freedom \(\nu\) (for ST).

### 1.3.1 Skew-t distribution

A \(k\)-dimensional random vector \(Y\) follows a \(k\) variate ST distribution, if its pdf is given by

\[
f(y|\mu, \Sigma, \Delta, \nu) = 2^k k_{k,\nu}(y; \mu, A)T_{1,\nu+k}(c^{-1}B; 0, 1),
\]
where $A = \Sigma + \Delta^2$, $B = \Delta^T A^{-1} (y - \mu) \sqrt{(\nu + k) / (1 - \Delta^T A^{-1} \delta)}$ and $c = \nu + (y - \mu)^T A^{-1} (y - \mu)$; $t_{k,\nu} (\cdot; \mu, \Sigma)$ is pdf of $k$-dimensional $t$-distribution with location vector $\mu$, scale matrix $\Sigma$ and degrees of freedom $\nu$; $T_{1,\nu} (\cdot; \mu, \Sigma)$ is the cumulative distribution function (cdf) of standard (univariate) $t$-distribution with degrees of freedom $\nu$. We denote the above distribution by $ST_{k,\nu} (\mu, \Sigma, \Delta)$. It can be shown that the mean and covariance matrix of $Y$ for the ST distribution $ST_{k,\nu} (\mu, \Sigma, \Delta)$ are given by

$$E(Y) = \mu + \frac{(\nu/\pi)^{1/2} \Gamma((\nu-1)/2)}{\Gamma(\nu/2)} \delta,$$

$$\text{cov}(Y) = \left[ \Sigma + \Delta^2 \right] \frac{\nu}{\nu - 2} - \frac{\nu}{\pi} \left[ \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)} \right]^2 \Delta^2. \quad (1.3)$$

It is noted that when $\Delta = 0$, the ST distribution reduces to the usual $t$-distribution. In order to have a zero mean vector, we should assume the location parameter $\mu = - (\nu/\pi)^{1/2} \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)} \Delta$.

According to the study by Muthen and Asparouhov (2014), if $Y$ follows $ST_{k,\nu} (\mu, \Sigma, \Delta)$, it can be expressed by a convenient stochastic representation as follows.

$$Y = \mu + \delta_{|X_0|} + \Sigma^{1/2} X_1, \quad (1.4)$$

where $X_1$ is $k$-dimensional vector with a multivariate $t$-distribution with zero mean, identity covariance matrix $I_k$ and degrees of freedom parameter $\nu$; $X_0$ is an univariate variable with a standard $t$-distribution with mean 0, variance 1 and degrees of freedom parameter $\nu$. The term $\delta_{|X_0|}$ can be thought of as a univariate skewness factor with factor loadings represented by the skew parameters of $\delta$, where the skewness is identified as that part of the $Y$ distribution not captured by the symmetric part $X_1$. Let $w = |X_0|$; then, $w$ follows a standard $t$-distribution truncated in the space $w > 0$ (i.e., a standard half $t$-distribution). Thus, a two-level hierarchical representation of (1.4) is given by

$$Y|w \sim t_{k,\nu} (\mu + \delta w, \Sigma), \quad w \sim t_{1,\nu} (0, 1) I(w > 0). \quad (1.5)$$

where $I(\cdot)$ denotes an indicator function. Note that when $\delta = 0$, the ST distribution $ST_{k,\nu} (\mu, \Sigma, \Delta)$ presented by the hierarchical expression (1.5) reduces to the multivariate $t$-distribution.
1.3.2 Skew-normal distribution

We briefly discuss a multivariate SN distribution introduced by Sahu, et al. (2003) here. An $k$-dimensional random vector $Y$ follows an $k$-variate SN distribution, if its pdf is given by

$$f(y|\mu, \Sigma, \Delta) = 2^k |A|^{-1/2} \phi_k \{ A^{-1/2}(y - \mu) \} P(V > 0),$$  

(1.6)

where $A = \Sigma + \Delta^2, V \sim N_m (\Delta A^{-1}(y - \mu), I_k - \Delta A^{-1} \Delta)$, and $\phi_k(\cdot)$ is the pdf of $N_k(0, I_k)$. We denote the above distribution by $SN_k(\mu, \Sigma, \Delta)$. An appealing feature of equation (1.6) is that it gives independent marginal when $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$. The pdf (1.6) thus simplifies to

$$f(y|\mu, \Sigma, \Delta) = \prod_{i=1}^k \left[ \frac{2}{\sigma_i^2 + \delta_i^2} \phi \left( \frac{y_i - \mu_i}{\sqrt{\sigma_i^2 + \delta_i^2}} \right) \Phi \left( \frac{\delta_i}{\sqrt{\sigma_i^2 + \delta_i^2}} \right) \right],$$  

(1.7)

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively.

The mean and covariance matrix are given by $E(Y) = \mu + \sqrt{2/\pi \delta}$, $\text{Cov}(Y) = \Sigma + (1 - 2/\pi) \Delta^2$. It is noted that when $\delta = 0$, the SN distribution reduces to usual normal distribution. In order to have a zero mean vector, we should assume the location parameter $\mu = -\sqrt{2/\pi \delta}$.

According to the study of Arellano-Valle et al. (2007), if $Y$ follows $SN_k(\mu, \Sigma, \Delta)$, it can be expressed by a convenient stochastic representation as follows.

$$Y = \mu + \delta |X_0| + \Sigma^{1/2} X_1,$$  

(1.8)

where $X_1$ is $k$-dimensional vector with a multivariate normal distribution $N_k(0, I_k)$; $X_0$ is a univariate variable with a standard normal distribution $N(0, 1)$. Let $w = |X_0|$; then, $w$ follows a standard normal distribution truncated in the space $w > 0$ (i.e., a standard half normal distribution). Thus, a two-level hierarchical representation of (1.8) is given by

$$Y|w \sim N_k(\mu + \delta w, \Sigma), \ w \sim N(0, 1) I(w > 0).$$  

(1.9)

Note that when $\delta = 0$, the hierarchical expression (1.9) presented for the SN distribution $SN_k(\mu, \Sigma, \Delta)$ reduces to a special case which is the multivariate normal distribution $N_k(\mu, \Sigma)$. 

8
1.4 Specific aims

Substance use data may exhibit skewness even after transformation. It will be valuable to explore whether a general skewed distribution such as ST or SN will bring a better model fitting. This study attempts to address three questions:

- First, it is important to explore if skew distribution family is suitable for analyzing semi-continuous substance use data that contain extra zeros and right skewness. The traditional approach of log-normal regression and Gamma regression involves data transformation and interpretation of coefficient estimates is for log-scaled response. If LMM with SN distribution is comparable to the traditional approach, the model with SN distribution would be preferable because SN distribution is modeling the original scale.

- Second, one-part LMM with skew distribution may not completely correct the extreme non-normality in semi-continuous data caused by the spike at zero, it is important to explore if the application of skew distribution in Part II of the two-part modeling will improve the model fit compared with the traditional two-part modeling approach.

- Third, longitudinal multivariate (bivariate) outcomes frequently arise in substance use studies, such as the alcohol abuse/dependence symptoms (AADS) and marijuana abuse/dependence symptoms (MADS) data that motivated our research. It is important to explore the bivariate analysis of semi-continuous data taking into account the correlation between the two response variables and the within-subject correlations in repeated measures.

This dissertation research is organized as follow:

- **Aim 1.** In Chapter 3, we present one-part LMM with SN distribution and compare to traditional models of LMM with LN and Gamma distribution in order to find which one has the best fit. The association of alcohol trajectory and gene and baseline covariates will be assessed. The effect of distributional assumptions on the parameter estimates will be evaluated. Simulation studies will be conducted to evaluate the proposed models.

- **Aim 2.** In Chapter 4, we develop two-part mixed-effects model with ST and SN distributions on the positive values taking into account the correlation between Part I and Part II, and the
within-subject correlation. A simulation studies will be conducted to evaluate the proposed models.

- **Aim 3.** In Chapter 5, we explore the statistical inference for bivariate semi-continuous variables by jointly modeling a logistic mixed-effects model on an indicator variable for zero-inflation and a bivariate LMM with ST and SN distributions. A simulation study will be conducted to evaluate the proposed models.
Chapter 2:

Motivating substance use data

The substance use data sets that motivated this research are from the Children in the Community (CIC) study (http://nyspi.org/child.com)(Kogan et al., 1977). The CIC study is a longitudinal study to evaluate the prevalence of mental health problems, the risk factors, and the outcomes in adulthood. The study goals include assessment of emotional and behavioral problems of childhood, and how children and families cope with these problems. It is one of the few studies that has conducted systematic, interview-based assessments of psychopathology over 30 years beginning in childhood in randomly-ascertained individuals. The CIC sample is a general population cohort based on households randomly sampled in 1975 when they lived in 100 rural, suburban, urban, and central city block groups cross-stratified by mean income and ethnic composition in two upstate New York counties. The sample’s racial distribution (91% Caucasian, 8% African-American) and socioeconomic status (21% with family income below the poverty line, 25% with upper middle class educational and income) represented the region that was selected in 1975 for its similarity in composition to the entire USA. The study began in 1975 (mean age of 5 for child participants) with more than 800 families enrolled and the follow-up interviews for alcohol and drug uses occurred in 1983 (mean age of 13), 1986 (mean age of 16), 1994 (mean age of 23) and 2003 (mean age 33). Deoxyribonucleic acid (DNA) samples and gene data including gene serotonin transporter polymorphism (5-HTTLPR) were collected in 2010 in 409 subjects who signed the consent form for DNA data collection.

The assessment of alcohol and drugs uses was consistent with Diagnosis and Statistical Manual of Mental Disorders (DSM) criteria. The Alcohol abuse/dependence symptoms (AADS) scale and the marijuana abuse/dependence symptoms (MADS) scale were derived for each participant at 4 follow-up interviews (year 1983, 1986, 1994 and 2003) . Cohen et al. (2007) studied the AADS/MADS data in more than 700 subjects using LMM with log-transformation and reported the association of AADS and MADS and gender, race, family SES and adolescent personality disorder assessed in
Previous studies suggest that gene 5-HTTLPR has an important role in alcohol consumption and dependence, but the findings are inconsistent. Van der Zwaluw et al. (2010) found that the adolescents with short allele in gene 5-HTTLPR had a larger increase in alcohol consumption than those without 5-HTTLPR short allele, consistent with the findings from two meta-analyses (Feinn et al., 2005; McHugh, 2010), but contradictory with the findings from some other studies where 5-HTTLPR short allele was reported to have protective effect or negative association with binge drinking in young adults (Olsson et al., 2005; Turker et al., 1998). Normal and binomial distributions are the typical assumptions in the statistical methods used in these studies.

In this dissertation, for the univariate one-part and two-part modeling, we are interested in assessing the relationship of alcohol dependence and gene 5-HTTLPR short allele with the longitudinal AADS data and the gene 5-HTTLPR data that is newly available. The baseline covariates such as age, gender, family SES, race and adolescent personality disorder will be assessed in the model. The dataset used in this dissertation included 1266 repeated measures of nonmissing ADDS and MADS scales from 318 subjects (41% male) with gene 5-HTTLPR data and nonmissing covariates. Note that missing mechanism was not considered in the sense that standard complete-data methods are adopted in the analysis. The AADS scales ranged from 0 to 17 with a mean of 2.3 (SD=2.6) and a median of 1.0 (inter-quartiles: 0 - 3.5). The data contain 26.2% of zeros and the positive values are right-skewed even after taking log-transformation (Figure 2.1a, 2.1b).

MADS data will be used in the bivariate semi-continuous modeling. MADS scales ranged from 0 to 9 with a mean of 0.7 (SD=0.9) and a median of 0.5 (inter-quartiles: 0.5 - 0.5). The data contains 21.2% of zeros (Figure 2.1c, 2.1d). In the joint distribution of AADS and MADS, both AADS and MADS are zero in 5.3% of pairs, AADS being zero and MADS being > zero in 20.9%, MADS being zero and AADS being > zero in 16.0%, and both responses being positive in 57.8% of data (Table 2.1).
Figure 2.1: Histograms of AADS (a,b) and MADS (c, d).

Table 2.1: Bivariate frequency distribution of AADS and MADS

<table>
<thead>
<tr>
<th></th>
<th>MADS</th>
<th></th>
<th>Total (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total (N)</td>
<td>0</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>AADS</td>
<td></td>
<td>332 (26.2%)</td>
<td>934 (73.8%)</td>
</tr>
<tr>
<td>0</td>
<td>67 (5.3%)</td>
<td>265 (20.9%)</td>
<td>332 (26.2%)</td>
</tr>
<tr>
<td>&gt; 0</td>
<td>202 (16.0%)</td>
<td>732 (57.8%)</td>
<td>934 (73.8%)</td>
</tr>
</tbody>
</table>
Chapter 3:

Bayesian approach to linear mixed-effects models with application to skewed-longitudinal alcohol data

Linear mixed-effects models (LMM) are frequently used to analyze repeated measurements due to the flexibility of modeling within-subject correlations and between-subject variations, often arisen in this type of data such as longitudinal alcohol data. However, the following two issues may stand out: (i) Normality of model errors is a routine assumption for LMM, but it may be unrealistic, obscuring important features of between- and within-subject variations; for example, alcohol data are often right skewed due to the large proportion of low or no alcohol use reported. (ii) Statistical modeling with lognormal (LN) or Gamma (G) distribution has been widely adopted in the analysis of alcohol data, though its validity is questionable and interpretation of results is inconvenient. This paper, under Bayesian framework, presents skew-normal (SN) LMM that relaxes normality assumption by using an SN distribution, which includes normal (N) distribution as a special case and provides flexibility in capturing a broad range of non-normal behavior. Alcohol data with features of skewness and excessive zeros are analyzed to demonstrate model performance by comparing SN-LMM with N-LMM, LN-LMM and G-LMM. Simulation studies are conducted to assess the performance of the four models for skewed data with varying proportions of zeros. Our findings suggest that, although all of SN-LMM, LN-LMM and G-LMM can be employed to handle skewed-longitudinal data, the proposed SN-LMM offers important advantages because the SN-LMM can not only capture skewed-longitudinal data well with more reasonable results, but can also offer a more intuitive and convenient interpretation of results.

Keywords: Bayesian analysis; longitudinal alcohol data; skew-normal distribution; lognormal distribution; Gamma distribution.
3.1 Introduction

Alcohol use is associated with a wide range of risk behaviors and leads to significant public health and social-economic problems. It is therefore important to understand the distribution patterns of alcohol use and to identify predictive factors that lead to alcohol use. Numerous studies have been conducted on various aspects of alcohol use, including longitudinal studies on the trajectory of alcohol use (Kerr et al., 2002; Bobo et al., 2010), clinical trials on the prevention and treatment of alcohol abuse/dependence (Johnson et al., 2007; Antón et al., 2006) and genomic studies on the association of genome and alcohol use (Van der Zwaluw et al., 2010; Enoch et al., 2011). In addition, adolescents and young adults are particularly vulnerable to alcohol use. Several studies have been conducted in this age group to research the developmental trend of alcohol use and its association with other substance uses (Wiesner et al., 2007; Cheadle and Whitback, 2011).

Alcohol use may be assessed in either discrete measures, such as the number of drinks last week, or continuous measures, such as alcohol abuse/dependence symptoms scale. Data are generally right-skewed because a large proportion of participants may report low or no alcohol use. The statistical modeling of count data often assumes Poisson distribution (Collins et al., 2012). Since Lederman (1956) hypothesized that alcohol consumption in a population follows a lognormal (LN) distribution, the continuous measures of alcohol use are generally assumed LN distribution in statistical modeling. Other parametric distributions that have been studied in alcohol data include Gamma and Weibull (Kehoe et al., 2012; Rehm et al., 2010; Skog, 1980; Skog and Rossow, 1993). In longitudinal studies, standard techniques for correlated data, such as linear mixed-effects models (LMM) are frequently used due to the exibility of modeling within-subject correlations and between-subject variations. Normality of model errors is a routine assumption in LMM. Due to the skewness of alcohol data, the analysis of longitudinal alcohol data often involves the log-transformed alcohol data as outcomes (Cohen et al., 2007). However, the log-transformed data, after the addition of a small constant to handle zeros, may not successfully approximate the normal distribution. In addition, the reasons that data transformation should be avoided include: (i) reduced information; (ii) difficulty with regarding interpretation of the transformed scale; and (iii) potential re-transformation bias. One alternative is to categorize the continuous measures and analyze the data with logistic/probit models; however, this approach may lose power by loss of information.
It would be most effective to use one parametric distribution to model the extreme non-normality in the alcohol data caused by zeros and right skewness without complication of data-transformation. The skew-normal (SN) distribution introduced by Azzalini (1985) has been studied intensively in the last three decades because it is flexible to accommodate extreme asymmetry in data and can model both positively and negatively skewed data (Arellano-Valle and Genton, 2005; Genton, 2004; Lin and Lee, 2008; Jara et al., 2008). Models and methods with SN distribution have been applied in many scientific fields, such as HIV studies (Ghosh et al., 2007; Huang and Dagne, 2010). However, the applications of LMM with SN distribution in the alcohol studies, to our best knowledge, are very limited. In this article, we will explore an LMM with SN distribution (SN-LMM) via Bayesian MCMC approach and apply this to the set of longitudinal alcohol data that motivated this research. We will consider a multivariate SN distribution introduced by Sahu et al. (2003) (see Section 1.3 in detail), which is suitable for a Bayesian inference. We will then conduct a systemic comparison of SN-LMM with various possible models generally used in alcohol studies, including (i) LMM with normal distribution (N-LMM), (ii) LMM with lognormal distribution (LN-LMM) and (iii) generalized LMM with Gamma distribution (G-LMM).

The remainder of this paper is organized as follows. In Section 3.2, we introduce SN-LMM model and the three counterpart models, and discuss the associated Bayesian inference method. In Section 3.3, longitudinal alcohol data described in Chapter 2 are analyzed to provide a systemic comparison of various scenarios-based potential models investigated. In Section 3.4, Simulation studies are conducted to assess the performance of the proposed models for skewed longitudinal data with different proportions of zeros. Finally, we conclude the paper with discussion in Section 3.5.

3.2 Statistical models and methods

3.2.1 Linear mixed-effects model with multivariate SN distribution

Denote the number of subjects by $n$ and the number of measurements on the $i$th subject by $n_i$. Let $y_i = (y_{i1}, \ldots, y_{in_i})^T$ be the vector of observed outcomes on the $i$th individual at times $t_{ij}$ ($i = 1, 2, \ldots, n; j = 1, 2, \ldots, n_i$). An LMM with SN distribution can be expressed as

$$y_i = x_i \beta + z_i b_i + e_i, \quad e_i \sim SN_{n_i}(-\sqrt{2/\pi} \delta_i, \sigma^2 I_{n_i}, \Delta_i), \quad b_i \sim N_q(0, \Sigma),$$

(3.1)
where \( x_i (n_i \times p) \) is the design matrix including time \( t_{ij} \) and possibly other time-independent and/or time-varying covariates corresponding to the fixed-effects, and \( z_i (n_i \times q) \) is the design matrix corresponding to random-effects \( b_i \). The vector of random errors \( e_i = (e_{i1}, \ldots, e_{in_i})^T \) follows a multivariate SN distribution with unknown variance parameter \( \sigma^2 \) and skewness diagonal matrix \( \Delta_i = \text{diag}(\delta_{i1}, \ldots, \delta_{in_i}) \) with skewness parameter vector \( \delta_i = (\delta_{i1}, \ldots, \delta_{in_i})^T \) (see Section 1.3 for details). In particular, if \( \delta_{i1} = \cdots = \delta_{in_i} = \delta \), \( \Delta_i = \delta I_{n_i} \) and \( \delta_i = \delta 1_{n_i} \) with \( 1_{n_i} = (1, \ldots, 1)^T \); this indicates that we are interested in skewness of overall data set and this is the case to be used in real data analysis in this paper. The vector of random-effects \( b_i = (b_{i0}, \ldots, b_{i,q-1})^T \) follows multivariate normal distribution with unknown variance-covariance matrix of \( \Sigma \). The conditional distribution of \( y_i \) given \( b_i \) is expressed as

\[
y_i|b_i \sim^{iid} SN_{n_i}(x_i\beta + z_i b_i, \sigma^2 I_{n_i}, \delta I_{n_i}),
\]

with

\[
E(y_i|b_i) = x_i\beta + z_i b_i - \delta \sqrt{2/\pi} 1_{n_i}, \quad \text{cov}(y_i|b_i) = [\sigma^2 + \delta^2 (1 - 2/\pi)]I_{n_i}.
\]

### 3.2.2 Three counterpart models

An LMM model with normal random error is a special case of SN-LMM defined above with skewness parameter \( \delta \) being zero. Under the assumption of multivariate normal distribution for random error and random-effects, model (3.1) reduces to the following formulation, denoted by N-LMM.

\[
y_i = x_i\beta + z_i b_i + e_i, \quad e_i \sim^{iid} N_{n_i}(0, \sigma^2 I_{n_i}), \quad b_i \sim^{iid} N_q(0, \Sigma),
\]

where the representations of \( x_i, z_i, \beta, b_i, e_i, \Sigma \) and variance parameter \( \sigma^2 \) remain the same as those in model (3.1).

A variable \( y \) is assumed to follow LN distribution if its log-transformation \( \log(y) \) follows normal distribution \( N(\mu, \sigma^2) \) where \( \mu \) and \( \sigma^2 \) are the mean and variance parameters on log-scale. The probability density function of \( y \) for LN distribution is expressed as

\[
f(y) = \frac{1}{y\sqrt{2\pi}\sigma} e^{-\frac{\log(y)-\mu}{2\sigma^2}}, \quad y > 0,
\]

where the expected value and variance of the \( y \) on original scale are given by \( E(y) = e^{(\mu+\sigma^2/2)} \) and
\[ \text{Var}(y) = e^{(2\mu + \sigma^2)}(e^\sigma - 1) = [E(y)]^2 \times (e^\sigma - 1) \] respectively (Zellner, 1971). We obtain lognormal linear mixed-effects model (LN-LMM) if we assume that the log-transformation of \( y_i \) follows the distribution assumption in model (3.3) as described below

\[ \log(y_i) = \mathbf{x}_i \beta + \mathbf{z}_i \mathbf{b}_i + e_i, \quad e_i \overset{iid}{\sim} N_n(0, \sigma^2 \mathbf{I}_n), \quad \mathbf{b}_i \overset{iid}{\sim} N_q(0, \mathbf{\Sigma}) \]

where the representations of \( \mathbf{x}_i, \mathbf{z}_i, \beta, \mathbf{b}_i, e_i, \mathbf{\Sigma} \) and variance parameter \( \sigma^2 \) remain the same as those in model (3.3). The conditional mean and variance for the untransformed response \( y_{ij} \) are given by

\[ u_{ij} = E(y_{ij}|\mathbf{b}_i) = \exp(\mathbf{x}_{ij} \beta + \mathbf{z}_{ij} \mathbf{b}_i + \sigma^2/2), \quad \text{Var}(y_{ij}|\mathbf{b}_i) = u_{ij}^2 (e^\sigma - 1), \]

where \( \mathbf{x}_{ij} \) is 1 \( \times \) \( p \) vector from the \( j \)th row elements in \( \mathbf{x}_i \) and \( \mathbf{z}_{ij} \) is defined similarly.

In this paper, we will assume homoscedasticity (i.e, \( \sigma^2 \) is constant) in data for simplicity. Under this assumption, a constant coefficient of variation (CV) on original scale is implied by

\[ CV^2 = \frac{\text{Var}(y)}{[E(y)]^2} = e^\sigma - 1. \]

Gamma probability distribution is commonly used to model right-skewed data. A variable \( y \) is assumed to follow Gamma distribution with a shape parameter \( \alpha \) and scale parameter \( \beta \) if it has probability density as follows.

\[ f(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta), \quad y > 0, \quad \alpha > 0, \quad \beta > 0, \]

where the mean and variance of Gamma distribution are given by \( E(y) = \alpha/\beta \) and \( \text{Var}(y) = \alpha/\beta^2 \), respectively. Assuming \( y_{ij} \) follows Gamma distribution, a generalized linear mixed-effects model with Gamma distribution (G-LMM) (McCullagh, 1989) can be defined as follows.

\[ g(\mu_{ij}) = g(E(y_{ij}|\mathbf{b}_i)) = \mathbf{x}_{ij} \beta + \mathbf{z}_{ij} \mathbf{b}_i, \quad g(\mu_{ij}) = \log(\mu_{ij}), \]

where \( \mathbf{b}_i \sim N_q(0, \mathbf{\Sigma}) \) as described in model (3.1) and \( g(.) \) is a link function to connect the random component \( y_{ij} \) and the covariate vectors \( \mathbf{x}_{ij} \) and \( \mathbf{z}_{ij} \). To ensure \( \mu_{ij} > 0 \), a log-link function \( g(\mu_{ij}) = \log(\mu_{ij}) \) is used. \( \mathbf{x}_{ij}(1 \times p) \) and \( \mathbf{z}_{ij}(1 \times q) \) are in the same as those in equation (3.5). The G-LMM can be reparameterized with mean \( \mu_{ij} \) and dispersion parameter \( 1/\upsilon \) as follows (McCullagh, 1989).

\[ f(y_{ij}|\mathbf{b}_i) = \frac{1}{\Gamma(\upsilon)} \left( \frac{\upsilon}{\mu_{ij} y_{ij}} \right)^\upsilon \exp\left( -\frac{\upsilon}{\mu_{ij}} y_{ij} \right) \frac{1}{y_{ij}}, \quad y_{ij} > 0, \quad \upsilon > 0, \quad \mu_{ij} > 0, \]
where \( \mu_{ij} = E(y_{ij} | b_i) = \exp(x_{ij} \beta + z_{ij} b_i) \) and \( \text{Var}(y_{ij} | b_i) = \mu_{ij}^2 / \nu \). Under Gamma distribution assumption, a constant coefficient of variation in data is assumed as \( CV^2 = 1 / \nu \).

### 3.2.3 Bayesian inference for parameter estimation

Following Sahu et al. (2003) and properties of SN distribution which is given in detail in Section 1.3, it can be shown that, by introducing a random vector \( w_i \), \( y_i \) and \( b_i \) in model (3.1) can be hierarchically formulated as

\[
\begin{align*}
  y_i | b_i, w_i &\sim N_n(x_i \beta + z_i b_i + \delta[w_i - \sqrt{2/\pi}1_{n_i}], \sigma^2 I_{n_i}), \\
  w_i &\sim N_n(0, I_{n_i})I(w_i > 0), \quad b_i \sim N_q(0, \Sigma), \\
  \end{align*}
\]

(3.8)

where \( 1_{n_i} = (1, \ldots, 1)^T \), \( I(\cdot > 0) \) is an indicator function and \( w_i \sim N_n(0, I_{n_i}) \) truncated in the space \( w_i > 0 \) (i.e., standard half-normal distribution).

Let \( \theta = \{\beta, \Sigma, \sigma^2, \delta\} \) be the collection of unknown population parameters in model (3.8). To complete the Bayesian formulation, we specify prior distributions for all the unknown parameters as follows.

\[
\begin{align*}
  \beta &\sim N_p(\beta_0, \Lambda), \quad \Sigma \sim IW(\Omega, \omega_1), \quad \sigma^2 \sim IG(\omega_2, \omega_3), \quad \delta \sim N(0, \omega_4), \\
\end{align*}
\]

(3.9)

where the mutually independent Normal \( (N) \), Inverse Gamma \( (IG) \) and Inverse Wishart \( (IW) \) prior distributions are chosen to facilitate computations. The full conditional posterior distribution of each parameter in \( \theta \) can be specified (Arellano-Valle et al., 2007), so that the MCMC approach via Gibbs sampler will be used.

Let the observed data \( D = \{(y_i, x_i, z_i); \ i = 1, \ldots, n; \ j = 1, \ldots, n_i\} \), \( f(\cdot | \cdot) \) be a conditional density function and \( \pi(\cdot) \) be a prior density function. We assume that \( \beta, \Sigma, \sigma^2, \) and \( \delta \) are independent of each other; in other words, \( \pi(\theta) = \pi(\beta)\pi(\Sigma)\pi(\sigma^2)\pi(\delta) \). After we specify the models for the observed data and prior distributions of the unknown model parameters, we can draw samples for the parameters based on their posterior distributions under the Bayesian framework. Thus, the joint posterior density of \( \theta \), conditional on \( D \), can be given by

\[
f(\theta | D) \propto \prod_{i=1}^{n} \int f(y_i | b_i, w_i) f(w_i | w_i > 0) f(b_i) db_i \pi(\theta).
\]

(3.10)

Because N-LMM is a special case of SN-LMM when \( \delta = 0 \), the Bayesian inference for N-LMM will be derived accordingly. The Bayesian inference for LN-LMM is defined similarly with log-
In the Bayesian inference for G-LMM, let $\theta = \{\beta, \upsilon, \Sigma\}$ be the collection of unknown parameters in model (3.6). The joint posterior distribution of the model parameters is

$$f(\theta | D) \propto \left\{ \prod_{i=1}^{n} \int f(y_i | b_i) f(b_i) db_i \right\} \pi(\theta),$$

where $\pi(\theta) = \pi(\beta)\pi(\upsilon)\pi(\Sigma)$. The prior distributions of $\beta$ and $\Sigma$ in (3.9) as well as $1/\upsilon \sim IG(\omega_1, \omega_2)$ will be assumed. There is no closed form for (3.11) and the more complex Metropolis-Hastings MCMC approach will be involved.

### 3.3 Alcohol data analysis

#### 3.3.1 Specific models and implementation

In this section, we analyze the alcohol data from CIC study described in Chapter 2. As depicted in Figure 2.1(a, b), AADS scales are right-skewed with excessive zeros. Toward this end, we will illustrate the performance of SN-LMM by comparing its counterpart LMMs with different distributional assumptions below.

- **Model SN**: A linear mixed-effects model with skew-normal (SN) distribution for model error;
- **Model N**: A linear mixed-effects model with normal distribution for model error;
- **Model LN**: A linear mixed-effects model with normal distribution for model error on log-scale response;
- **Model G**: A generalized mixed-effects model with a log-link function and Gamma distribution for response.

Multivariate normal distribution for random-effects is assumed for all models. Because LN and Gamma distribution can be only applied to model positive values, a small constant of 0.1 was added to outcome $y_{ij}$ in all models to make the results comparable.

To evaluate impact of Model SN on parameter estimation in comparison with the traditional approach-based models, we used Model N as illustration for the model selection to determine the set of covariates to be included in a final model. In addition to gene 5-HTTLPR which is a primary
predictor, gender, race, family SES and adolescent personality disorders that have been shown to be significantly associated with ADDS (Cohen et al., 2007) were evaluated by procedure described below and only the covariates which showed significant association with AADS were included in the final model. First, a basic unconditional Model N was assessed with age as the only covariate to estimate the AADS trajectory. The individual trajectory plots from randomly selected subjects depicted that linear changes over time was not sufficient to describe the trajectories (data not shown), the quadratic age term was then included in the basic model. Both the intercept and slope were considered with random-effects, which is given by \( b_i = (b_{i0}, b_{i1})^T \), allowing for parameter estimation at both the intra-individual and inter-individual levels. Second, covariates and their interactions with age were added to the basic model to determine any potential influence on the development of AADS. Gene, gender, race, SES, interaction between gene and age, and interaction between gender and age were kept in the final model in addition to age and quadratic age terms after model selection. To compare the performance of models with different distributional assumptions, the same set of covariates were included in all models. The centered age variable (i.e., age - mean age) was used in modeling such that the intercept coefficient estimate can be interpreted as AADS level at the mean age. Thus, the linear function with covariates and random-effects is defined as follows.

\[
x_{ij} \beta + z_{ij} b_i = \beta_0 + \beta_1 \times \text{age}_{ij} + \beta_2 \times \text{age}_{ij}^2 + \beta_3 \times \text{gene}_i + \beta_4 \times \text{gene}_i \times \text{age}_{ij} + \beta_5 \times \text{gender}_i + \beta_6 \times \text{gender}_i \times \text{age}_{ij} + \beta_7 \times \text{race}_i + \beta_8 \times \text{SES}_i + b_{i0} + b_{i1} \times \text{age}_{ij}.
\]

We conducted the following scenarios. First, because a normal distribution is a special case of an SN distribution when the skewness parameter is zero, we investigated how an asymmetric (SN) distribution for model error (Model SN) impacts parameter estimation in comparison with a symmetric (normal) distribution for model error (Model N). Second, we studies how the assumption of SN distribution for model error contributes to model results compared with the traditional assumption of LN distribution and Gamma distribution for model error. The Deviance Index Criterion (DIC) proposed by Spiegelhalter et al. (2002) was used as model selection criteria to evaluate and select the best model. The DIC is defined as \( DIC = \bar{D}(\theta) + p_D = 2\bar{D}(\theta) - D(\bar{\theta}) \), where \( \bar{D}(\theta) = E[D(\theta)|y] \) is the posterior mean of the deviance with a smaller value indicating better fit, and \( p_D = \bar{D}(\theta) - D(\bar{\theta}) = E[D(\theta)|y] - D[E(\theta|y)] \) is the effective number of parameters, defined
as the difference in the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters. The structure of DIC allows for automatic computation in WinBUGS software (Lunn et al., 2000). For Model LN, the DIC was calculated on original scale as well as the residual (i.e., residual \( r_{ij} = y_{ij} - u_{ij} \)) where \( u_{ij} \) is given by equation (3.5).

To carry out the Bayesian inference, we need to specify the values of the hyper-parameters in the prior distributions. We use the weakly informative prior distribution for the parameters in models. In particular, (i) each element in fixed-effects \( \beta \) was taken to be independent normal distribution \( N(0, 100) \); (ii) for the dispersion parameter \( \sigma^2 \) and \( 1/\nu \), we assume a weakly informative inverse Gamma prior distribution \( IG(0.01, 0.01) \) so that the distribution has a mean value of 1 and a variance of 100; (iii) the prior for the variance-covariance matrix of the random-effects \( \Sigma \) was taken to be inverse Wishart distributions with \( \Omega = \text{diag}(1, 1) \) and degrees of freedom (df) \( \omega_1 = 2 \); and (iv) the skewness parameter \( \delta \) was taken to be a normal distribution \( N(0, 100) \).

The MCMC sampler was implemented using WinBUGS software (Lunn et al., 2000) and the program code is available from the corresponding author upon request. When the MCMC procedure was applied to the alcohol data, convergence of the generated samples was assessed using standard tools within WinBUGS software such as trace plots and Gelman-Rubin (GR) diagnostics (Gelman and Rubin, 1992). Figure 3.1 shows the trace plots, dynamic version of GR diagnostic plots and autocorrelation plots based on Model SN for the representative parameters \( \beta_1, \beta_2, \beta_3 \) and \( \beta_4 \). We observe from trace plots (left panel) that the lines of three different chains mix or cross in trace, implying that convergence is ensured. For the plots of GR diagnostics (middle panel) where the three curves are given: the middle and bottom curves below the dashed horizontal line (indicated by the value one) represent the pooled posterior variance \( \hat{V} \) and average within-sample variance \( \hat{W} \), respectively, and the top curve represents their ratio \( \hat{R} \). It is seen that \( \hat{R} \) is generally expected to be higher than 1 at the initial stage of the algorithms, but \( \hat{R} \) tends to 1, and \( \hat{V} \) and \( \hat{W} \) stabilize as the number of iterations increase, indicating that the algorithm has approached convergence. We further monitor convergence using autocorrelation plots (right panel) that autocorrelations are very low with a lag being 20, implying that convergence is obtained. When these criteria suggested the convergence of chains, we proposed that the three chains were run with the following considerations. For each model, we ran three, initially dispersed chains for 40,000 iterations each, discarding the first 20,000 as a burn-in period, and retained every 20th sample for a total of 3,000 samples of
3.3.2 Comparison of model fitting results

Table 3.1 presents the posterior mean (PM) and the 95% credible intervals (CI) for fixed-effects parameters, the skewness and precision parameters as well as DIC from the four models with different distributional assumptions. In the comparison of Models N, SN, LN and G, Model SN provided the smallest DIC (712 for Model SN vs. 5465, 4054, and 4216 for Models N, LN and G, respectively). The quantile-quantile (Q-Q) normal plots from the four models were displayed in Figure 3.2. It can be seen that Model SN, where random error was SN distribution, provided the
best fit to the observed data, compared with the other three models. Meanwhile, the Q-Q plots for both Models LN and G showed the existence of both negative and positive outliers and the Q-Q plot for Model N showed outliers on the positive side. The estimates of individual trajectories displayed in Figure 3.3 showed that fitted values from Model SN traced the observed values almost perfectly compared with those from the other models. The estimated trajectories from Model SN lack smoothness are understandable because the predicted function involves a random vector \( w_i \) (see equation (3.8) in detail) to “chase the data” to this extent, according to the stochastic representation feature of SN distribution.

In the comparison of Model N with Model SN, we found that: (i) The posterior mean estimate of skewness parameter in SN distribution is 3.410 with 95% CI (3.245, 3.587) which is significantly positive (Table 3.1). This confirms the skewness with heavy right-tail of AADS scale shown in Figure 2.1(a, b), suggesting that accounting for skewness is important when the data exhibit skewness. (ii) The DIC for Model SN is significantly smaller than that for Model N (712 vs. 5465), indicating Model SN fits data better than Model N. (iii) The parameter estimates from the two models vary substantially, in particular, (a) the posterior mean estimate and 95% CI of \( \beta_5 \) for male gender are 1.150 (0.806, 1.475) and 0.508 (0.311, 0.729) from Models N and SN, respectively (Figure 3.4a and 3.4b); (b) the posterior mean estimate of \( \beta_4 \) from Model N for the interaction between gene 5-HTTLPR and age is -0.042 with 95% CI (-0.084, 0.001), which is marginally significant, indicating that gene 5-HTTLPR short allele is associated with an increased rate of change (slope) in AADS; this is contradictory with the insignificant finding from Model SN where the posterior mean estimate of \( \beta_4 \) is -0.023 with 95% CI (-0.055, 0.008) (Figure 3.4c and 3.4d).

In comparing Model SN with the traditional Models LN and G, the DIC showed that Model G provided better fit than Model LN (4054 vs. 4216); however, the proposed Model SN is superior to both Models G and LN. It is noted that the parameter estimates of covariate coefficients from both Models LN and G are on log-scale; while they may be compared to each other, they can not be compared directly with parameter estimates from Models N and SN. The parameter estimates from Models LN and G are very similar except for the estimate on intercept which is 0.269 for Model G and -0.278 for Model LN.

Based on model selection criteria of DIC and Q-Q plots, Model SN is the best model for the observed skewed alcohol data with excessive zeros. In Model SN, we found a positive association
Figure 3.2: Quantile-quantile (Q-Q) normal plots for residuals (in original scale) from Models N, SN, LN and G.
between AADS and linear age determined by $\beta_1$ (posterior mean=0.062 with 95% CI (0.036, 0.100)) and a negative association between AADS and quadratic age ($\beta_2$ = -0.006 with 95% CI (-0.008, -0.004)). This indicates that the mean level of AADS increased at early age and began to drop after subjects reaching certain age. We found that male gender is associated with higher mean AADS level ($\beta_5$=0.508 with 95% CI (0.311, 0.729)) and AADS increased more rapidly in male as indicated by the interaction between age and male gender ($\beta_6$=0.061 (0.034, 0.088)). The increased AADS is associated with a lower family SES ($\beta_8$ = -0.098 with 95% CI (-0.203, -0.007)). There is no significant association between AADS and gene 5-HTTLPR determined by $\beta_3$ (-0.081 with 95% CI (-0.277, 0.140)) and $\beta_4$ (-0.023 with 95% CI (-0.055, 0.008)).

![Figure 3.3: Individual fitted curves of AADS scale for 4 representative patients based on Models N, SN, LN, and G. The observed values were represented with diamond. Fitted curves for Model LN were presented on original scale.](image-url)
Table 3.1: Summary of parameter estimates based on Models N, SN, LN and G, including posterior mean (PM) for parameters of population (fixed-effects), variance, skewness, lower limit (L CI) and upper limit (U CI) of 95% equal-tail credible interval (CI), and deviance information criterion (DIC).

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>$\beta_6$</th>
<th>$\beta_7$</th>
<th>$\beta_8$</th>
<th>$\delta$</th>
<th>$\sigma^2$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>PM</td>
<td>2.010</td>
<td>0.141</td>
<td>-0.018</td>
<td>-0.249</td>
<td>-0.042</td>
<td>1.150</td>
<td>0.072</td>
<td>1.192</td>
<td>-0.192</td>
<td>.</td>
<td>3.335</td>
</tr>
<tr>
<td></td>
<td>L CI</td>
<td>1.252</td>
<td>0.102</td>
<td>-0.020</td>
<td>-0.637</td>
<td>-0.084</td>
<td>0.806</td>
<td>0.033</td>
<td>0.473</td>
<td>-0.372</td>
<td>.</td>
<td>3.030</td>
</tr>
<tr>
<td></td>
<td>U CI</td>
<td>2.729</td>
<td>0.180</td>
<td>-0.015</td>
<td>0.123</td>
<td>0.001</td>
<td>1.475</td>
<td>0.110</td>
<td>1.924</td>
<td>-0.019</td>
<td>.</td>
<td>3.680</td>
</tr>
<tr>
<td>SN</td>
<td>PM</td>
<td>2.200</td>
<td>0.062</td>
<td>-0.006</td>
<td>-0.081</td>
<td>-0.023</td>
<td>0.508</td>
<td>0.061</td>
<td>0.606</td>
<td>-0.098</td>
<td>3.410</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>L CI</td>
<td>1.710</td>
<td>0.036</td>
<td>-0.008</td>
<td>-0.277</td>
<td>-0.055</td>
<td>0.311</td>
<td>0.034</td>
<td>0.203</td>
<td>-0.203</td>
<td>3.245</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>U CI</td>
<td>2.624</td>
<td>0.100</td>
<td>-0.004</td>
<td>0.140</td>
<td>0.008</td>
<td>0.729</td>
<td>0.088</td>
<td>1.099</td>
<td>-0.007</td>
<td>3.587</td>
<td>0.024</td>
</tr>
<tr>
<td>LN</td>
<td>PM</td>
<td>-0.278</td>
<td>0.112</td>
<td>-0.010</td>
<td>-0.084</td>
<td>-0.017</td>
<td>0.499</td>
<td>0.023</td>
<td>0.821</td>
<td>-0.068</td>
<td>.</td>
<td>1.150</td>
</tr>
<tr>
<td></td>
<td>L CI</td>
<td>-0.706</td>
<td>0.087</td>
<td>-0.011</td>
<td>-0.297</td>
<td>-0.043</td>
<td>0.302</td>
<td>-0.002</td>
<td>0.397</td>
<td>-0.168</td>
<td>.</td>
<td>1.043</td>
</tr>
<tr>
<td></td>
<td>U CI</td>
<td>0.146</td>
<td>0.137</td>
<td>-0.009</td>
<td>0.119</td>
<td>0.011</td>
<td>0.697</td>
<td>0.048</td>
<td>1.246</td>
<td>0.028</td>
<td>.</td>
<td>1.270</td>
</tr>
<tr>
<td>G</td>
<td>PM</td>
<td>0.269</td>
<td>0.113</td>
<td>-0.011</td>
<td>-0.096</td>
<td>-0.019</td>
<td>0.470</td>
<td>0.021</td>
<td>0.785</td>
<td>-0.059</td>
<td>.</td>
<td>0.793</td>
</tr>
<tr>
<td></td>
<td>L CI</td>
<td>-0.113</td>
<td>0.088</td>
<td>-0.012</td>
<td>-0.283</td>
<td>-0.046</td>
<td>0.292</td>
<td>-0.003</td>
<td>0.409</td>
<td>-0.148</td>
<td>.</td>
<td>0.726</td>
</tr>
<tr>
<td></td>
<td>U CI</td>
<td>0.656</td>
<td>0.137</td>
<td>-0.009</td>
<td>0.085</td>
<td>0.008</td>
<td>0.645</td>
<td>0.046</td>
<td>1.158</td>
<td>0.026</td>
<td>.</td>
<td>0.865</td>
</tr>
</tbody>
</table>

Notes: $\sigma^2$ = variance on log-scale in Model LN and $\sigma^2=1/\nu = variance/(mean)^2$ in Model G.
3.4 Simulation study

To assess the performance of the proposed four models (Models N, SN, LN and G) and associated method under skewed-longitudinal data with varying proportions of zeros, as an illustration, we conducted the following two simulation studies. The first one in comparison of Model SN with Model N was conducted as follows.

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + b_{i0} + b_{i1} t_{ij} + e_{ij}, \]  

(3.13)

where the measurement time points are given as \( t_{ij} = 0, 0.1, 0.2, 0.3, 0.5, 0.75, 1, 2, 3, 4 \) and 5, a binary covariate \( x_i = 0 \) or 1, true parameter values are chosen as \( \beta = (\beta_0, \beta_1, \beta_2)^T = (2.2, 2, -0.5)^T, \)
and the random-effects \((b_{ij0}, b_{ij1})^T \sim N_2(\mathbf{0}, \Sigma)\) and \(\Sigma = \text{diag}(0.3^2, 0.3^2)\). We generated \(e_{ij} = \varepsilon_{ij} - 1\), in which \(\varepsilon_{ij}\) follows a Gamma distribution Gamma(1,1), yielding a skewed distribution with \(E(\varepsilon_{ij}) = 0\) and \(\text{Var}(\varepsilon_{ij}) = 1\). Under this specification, data generated from model (3.13) exhibit right-skewed feature with a range of \(0 \sim 20\). Trajectories of 500 individuals with the first 250 subjects assigned to \(x_i = 0\) and others to \(x_i = 1\) are generated. To evaluate the performance of Models SN and N under skewed data with different proportions of zeros, for each trajectory, we randomly set 0%, 10%, or 20% of data to zero. The MCMC strategies in real data analysis are adopted in this simulation study and the weakly informative prior distributions used in real data are adopted here. Since the priors are weakly informative, we expect the results to be somewhat robust with respect to prior distributions.

Models N and SN are fitted to simulated data set with 100 repetitions, respectively. Table 3.2 summarizes simulation results which include the MC mean of fixed-effects \(\beta\), as well as associated percent mean-square-error (MSE), defined by \(100 \times \sqrt{\text{MSE}^l/|\beta|^l} \) \((l = 0, 1, 2)\). We found that the proposed Model SN provided more accurate estimates on \(\beta_1\) for time measurements and \(\beta_2\) for the binary covariate, compared with Model N when the proportion of zeros is increased from 0% to 10% and 20% (Table 3.2). Our simulation results suggest that Model SN is powerful in handling the data exhibiting skewness with excessive zeros.

Because the data generation for Models LN and G involve log-transformation, a separate simulation study was conducted to compare Models LN and G following model below.

\[
y_{ij} = \exp(\beta_0 + \beta_1 t_{ij} + \beta_2 x_i + b_{i0} + b_{i1} t_{ij}) \times (1 + e_{ij}), \tag{3.14}
\]

where the measurement time points are given as \(t_{ij} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\), and 1, a binary covariate \(x_i = 0\) or \(1\), true parameter values are \(\beta = (\beta_0, \beta_1, \beta_2)^T = (0.5, 1, -0.5)^T\), and the random effects \((b_{i0}, b_{i1})^T \sim N_2(\mathbf{0}, \Sigma)\) and \(\Sigma = \text{diag}(0.3^2, 0.3^2)\). Because the error is multiplicative in model (3.14), we use Beta distribution to generate samples between 0 and 1 to keep the simulated \(y_{ij}\) in the range of 0 and 20 approximately. We generated \(e_{ij} = \varepsilon_{ij} - 0.5/(0.5 + 1)\), in which \(\varepsilon_{ij}\) follows a beta distribution Beta(0.5, 1), yielding a skewed distribution with \(E(\varepsilon_{ij}) = 0\) and \(\text{Var}(\varepsilon_{ij}) = 0.09\). The other specifications are identical to those in the first simulation. Under this specification, data generated from model (3.14) exhibit highly right-skewed feature with a range of \(0 \sim 20\). We found that Model G works better than Model LN on the estimate of \(\beta_1\) for
time measurements and the estimate of $\beta_2$ for the binary covariate, when the proportion of zeros is increased from 0% to 10% and 20% (Table 3.2). Our simulation results suggest that Model G may perform better than Model LN under the data exhibiting skewness with excessive zeros.

Table 3.2: Summary of Monte Carlo simulation results for MC estimates of fixed-effects $\beta$ as well as MSE for Models N, SN, LN and G based on 100 simulated datasets under different proportions of zeros.

<table>
<thead>
<tr>
<th>Model</th>
<th>true values</th>
<th>no zeros</th>
<th>10% zeros</th>
<th>20% zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>N-LMM</td>
<td>2.00</td>
<td>-0.50</td>
<td>1.80</td>
<td>-0.46</td>
</tr>
<tr>
<td></td>
<td>1.05</td>
<td>6.28</td>
<td>10.23</td>
<td>13.06</td>
</tr>
<tr>
<td>SN-LMM</td>
<td>1.99</td>
<td>-0.50</td>
<td>1.93</td>
<td>-0.48</td>
</tr>
<tr>
<td></td>
<td>1.05</td>
<td>6.28</td>
<td>3.50</td>
<td>12.06</td>
</tr>
<tr>
<td>LN-LMM</td>
<td>0.95</td>
<td>-0.47</td>
<td>0.85</td>
<td>-0.43</td>
</tr>
<tr>
<td></td>
<td>5.49</td>
<td>8.40</td>
<td>15.26</td>
<td>16.30</td>
</tr>
<tr>
<td>G-LMM</td>
<td>0.95</td>
<td>-0.47</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>5.18</td>
<td>8.34</td>
<td>5.04</td>
<td>5.04</td>
</tr>
</tbody>
</table>

Notes: MC mean=Monte Carlo mean
MSE=Percent $\sqrt{MSE} = 100 \times \sqrt{MSE}/||\hat{\beta}||$.
MC mean estimate of $\delta = 1.50, -2.54, and -3.44$ from Model SN for scenarios with 0%, 10% and 20% of zero.

Because the data generation algorithm is different, no direct comparison between Model SN and Model G was considered in simulation studies.

3.5 Concluding discussion

Model N has been known to give inaccurate estimates under non-normal data and was included here as comparison purpose. The posterior mean estimate of the skewness parameter $\delta$ (3.410) from Model SN confirmed that AADS scales are highly right-skewed as displayed in Figure 2.1. We were able to repeat the finding by Van der Zwaluw et al. (2010) on the association between gene 5-HTTLPR short allele and alcohol dependence using Model N in which normality is assumed for random-effects and model errors. However, in our Model SN (a linear mixed-effects model with SN assumption which clearly fits the data better), the association between gene 5-HTTLPR and
alcohol use is not significant. This controversy suggests that parameter estimates from models with normality assumption may be unreasonable and even be misleading when data exhibit extreme skewness, as demonstrated by our simulation study.

Although models LN and G provided better model fits to the real data than did Model N, the parameter estimates are for the log-scale of the response, which makes interpretation complicated. In addition, a small constant had to be added to the outcome variable in order to apply a log-transformation on the zeros. The selection of the constant could potentially provide rather different results.

In the comparisons with Models LN and G, the proposed Model SN not only provides the best fit to the real data in terms of DIC, but also offers more intuitive and convenient interpretation of results on the original scale.

We found a significant interaction of male gender and linear age in both Models N and SN. The interaction between quadratic age and gender was also evaluated. We found that, in both Models N and SN, adding the interaction of quadratic age and gender did not improve model fitting based on DIC, the parameter estimate for this interaction term was not significant.
Chapter 4:

Bayesian inference on two-part mixed-effects models using skew distributions, with application to longitudinal semi-continuous alcohol data

Semi-continuous data featured with an excessive proportion of zeros and right-skewed continuous positive values arise frequently in practice. One example would be the substance abuse/dependence symptoms data for which a substantial proportion of subjects investigated may report zero. Two-part mixed-effects models have been developed to analyze repeated measures of semi-continuous data from longitudinal studies. In this paper, we propose a flexible two-part mixed-effects model with skew distributions for correlated semi-continuous alcohol data under the framework of a Bayesian approach. The proposed model specification consists of two mixed-effects models linked by the correlated random-effects: i) a model on the occurrence of positive values using a generalized logistic mixed-effects model (Part I); and ii) a model on the intensity of positive values using a linear mixed-effects model where the model errors follow skew distributions including skew-$t$ and skew-normal distributions (Part II). The proposed method is illustrated with an alcohol abuse/dependence symptoms data from a longitudinal observational study and the analytic results are reported by comparing potential models under different random-effects structures. Simulation studies are conducted to assess the performance of the proposed models and method.

Keywords: Bayesian analysis; semi-continuous data; two-part mixed-effects model; skew distributions; alcohol abuse/dependence symptoms data.

4.1 Introduction

Semi-continuous data featured with an excessive proportion of zeros and right-skewed positive values arise frequently in clinical research. Examples include alcohol consumption, medical cost, and substance abuse symptom scales. Statistical models with normality assumption ignoring the skewness and the spike at zero are not suitable for this type of data and may lead to substantial bias and incorrect statistical inferences. Two-part models, originating in econometrics (Cragg,
1971; Duan et al., 1983a), have been developed extensively in the last three decades to analyze this type of data and have been applied to scientific fields other than economics such as clinical research and health services. In two-part models, we view a semi-continuous variable as the result of two processes: one binomial process determining whether the positive value occurs and one continuous process determining the actual value given it is nonzero. Therefore, a two-part model consists of two components, with the first component (Part I) modeling the probability of a response being positive using probit or logistic regression, and the second component (Part II) modeling the conditional mean of the positive values (given positive values occurred) using linear regression. Various methods have been developed for analyzing cross-sectional and longitudinal semi-continuous data (Duan et al., 1983a; Hall and Zhang, 2004; Moulton et al., 2002; Olsen and Schafer, 2001; Tooze et al., 2002).

Olsen and Schafer (2001) first extended the two-part models developed by Duan et al. (1983a) and Manning et al. (1981) for cross-sectional data to the longitudinal setting by introducing correlated random-effects into the logit and log-normal components, respectively, and applied to longitudinal alcohol data. In the two-part mixed-effects models, the binomial process is typically modeled with mixed-effects logistic or probit regression, and the continuous process is naturally modeled via linear mixed-effects models (LMM). The random-effects in the two components are generally assumed to be correlated through a multivariate normal distribution structure. Ignoring the between-component association mistakenly can yield biased estimates in the second part of the model (Su et al., 2009). The correlated random-effects can capture not only the between-component association, but also the within-subject correlation among repeated measurements collected from the same individual. A between-component correlation means that the process giving rise to the positive values is related to the magnitude of observed value given that a positive response occurred. For example, in a data collection of substance abuse/dependence symptom scales (DSM-IV) where zero represents no symptoms and the continuous positive values reflect the severity of symptom, a positive correlation suggests that an individual who tends to have a symptom is likely to have a severe symptom.

For the positive part of a semi-continuous variable, LMMs with a normality assumption were used by Husted et al. (2007) and Su et al. (2009) where in their studies the residual plots did not reveal skewness of data. However, the positive part of a semi-continuous variable is often right skewed. The logarithmic transformation was the most commonly used approach to correct the
skewness (Olsen and Schafer, 2001; Tooze et al., 2002) and other monotone increasing functions such as Box-cox transformation that would make the positive component approximately normal were also explored (Kipnis et al., 2009; Liu et al., 2010). The limitations with data transformation in Part II include reduced information, difficulty in interpreting the results and possible heteroscedasticity ((Liu et al., 2010; Duan et al., 1983b). An alternative approach is to use a generalized linear mixed-effects model (GLMM) with distributions in the exponential family that can model skewed data, such as Gamma (Liu et al., 2010), beta (Hatfield et al., 2011) and Weibull distributions. Manning et al.(2005) proposed a one-part generalized Gamma distribution that include the lognormal, Weibull and exponential distributions as special cases. Liu et al. (2010) extended it to the two-part mixed-effects model and concluded that model with generalized Gamma distribution in Part II provided the best fit to the alcohol consumption data in the comparisons with models with log-skew-normal distribution and Box-Cox transformation (Liu et al., 2012). It is noted that GLMMs often involve complicated iterative procedures in estimation which may lead to intensive computation burden and non-convergence issue.

It would be most effective to use a parametric distribution to model the right skewed positive values in two-part models. As a result, we propose a two-part mixed-effects model with an asymmetric distribution such as skew-t (ST) or skew-normal (SN) distribution (Arellano-Valle and Genton, 2005; Arellano-Valle et al., 2007; Azzalini and Capitanio, 1999, 2003; Azzalini and Genton, 2008; Ho and Lin, 2010; Jara et al., 2008; Sahu et al., 2003) for continuous positive values. The skew distributions including ST and SN distributions were first introduced by Azzalini (1985), and a detailed review on skew distributions is provided by Arellano-Valle and Genton (2005) and Genton (2004). In particular, Sahu et al. (2003) developed Bayesian inference for skewed data using SN and ST distributions followed by Jara et al. (2008). Lin and his collaborators (2007a, 2007b, 2010) developed maximum likelihood method for LMMs with ST and SN distributions. Arellano-Valle et al.(2007) developed Bayesian approach for LMMs with SN distribution. The skew distribution family, which accommodates asymmetry in a more flexible manner, can model both positively and negatively skewed data (depending on the sign of the skewness parameter), and includes normal and $t$ distribution as special cases. There is very limited research on the application of the skew distribution in two-part modeling. Chai and Bailey (2008) proposed a probit/log-skew-normal mixture model for cross-sectional semi-continuous data in which the skew-normally distributed error
was assumed for log-transformed positive values and estimated with likelihood approach. Mahmud et al. (2010) extended the research to the longitudinal setting. In both studies the skew-normal distribution was assumed on log-transformed positive values. The problem with log-skew-normal distribution is that there is a potential to overcorrect the skewness of data (Liu et al., 2012). Moreover, the re-transformation to the original scale is complicated with the log-transformed data (Duan et al., 1983b; Manning, 1998; Manning and Mullahy, 2001).

Parameter estimations in two-part modeling could be computationally difficult. For two-part models with uncorrelated random-effects, maximum likelihood estimates can be derived by fitting separate mixed-effects model for each component. For the correlated two-part mixed-effects model with lognormal distribution on the positive values, Olsen and Shafter (2001) and Tooze et al. (2002) developed different maximum likelihood approaches. Several authors have proposed Bayesian approaches to fit the two-part models (Cooper et al., 2007; Ghosh and Albert, 2009; Hatfield et al., 2011; Neelon et al., 2010, 2011; Zhang et al., 2006). For example, Cooper et al. (2007) used a Bayesian approach via Markov chain Monte Carlo (MCMC) to fit a probit-lognormal correlated two-part model on medical cost data.

In this paper, we propose a two-part mixed-effects model with logistic mixed-effects model on the occurrence of positive values and LMM with ST and SN distributions on the continuous positive values using Bayesian approach via MCMC procedure with application to a longitudinal alcohol data that motivated this research. We consider a multivariate ST distribution introduced by Sahu et al. (2003) and Muthen and Asparouhov (2014) (see Section 1.3 in detail), which is suitable for a Bayesian inference. It is noted that the ST distribution is approximate to the SN distribution when its degrees of freedom approach infinity, the ST distribution reduces to a \( t \) distribution if skewness parameter is zero and the SN distribution reduces to a normal distribution if skewness parameter is zero. Thus, we use an ST distribution to develop the two-part model and associated statistical methodologies, as it can be easily reverts to the other distributions such as the SN and N distributions. We model the between-component association by introducing correlated multivariate normal random-effects in both parts (Olsen and Schafer, 2001). Different random-effects structure will be evaluated including: i) two random intercept effects with one for each of the two components, and ii) two random intercept effects and a random slope effect for Part II of the model. It is easy to extend the model to include more random-effects. It is noted that including one additional
random-effect will increase the number of parameters by the number of subjects. In this paper, we assume the random-effects follow multivariate normal distribution and model errors follow skew distribution.

Previous studies suggest that gene 5-HTTLPR has an important role in alcohol consumption and dependence, but the findings are inconsistent. Van der Zwaluw et al. (2010) found that the adolescents with short allele in gene 5-HTTLPR had a larger increase in alcohol consumption than those without 5-HTTLPR short allele, consistent with the findings from two meta-analyses (Feinn et al., 2005, McHugh et al., 2010), but contradictory with the findings from some other studies where 5-HTTLPR short allele was reported to have protective effect or negative association with binge drinking in young adults (Olsson et al., 2005; Turker et al., 1998). Normal and binomial distributions are the typical assumptions in the statistical methods used in these studies. In this paper, we are interested in, using our proposed model with consideration of extreme non-normality in alcohol data, to re-evaluate the association of alcohol dependence and gene 5-HTTLPR with the AADS data and the newly available gene 5-HTTLPR data from CIC study.

The rest of this article is organized as follows. In Section 4.2, we present the two-part mixed-effects model with model errors to have ST distribution which includes SN and N distributions as special cases and discuss the associated Bayesian inference method. In Section 4.3, we apply the proposed methodologies to an alcohol data set described in Chapter 2 and report the analysis results. In Section 4.4, simulation studies are conducted to assess the performance of the proposed models. Finally, we conclude the article with a discussion in Section 4.5.

### 4.2 Two-part mixed-effects model with ST distribution

#### 4.2.1 Model specification

Let \( y_{ij} \) be a semi-continuous variable from the \( i \)th \((i = 1, \ldots, n) \) subject at time \( t_{ij} \)(\( j = 1, \ldots, n_i \)). The variable \( y_{ij} \) can be represented as two variables: the occurrence variable

\[
    r_{ij} = \begin{cases} 
        1 & \text{if } y_{ij} > 0 \\
        0 & \text{if } y_{ij} = 0
    \end{cases}
\]

and the intensity variable \( y_{ij} \) given \( y_{ij} > 0 \).

We model these two variables by a pair of correlated mixed-effects models with one for probability \( P(r_{ij} = 1) \) and one for the mean conditional response \( E(y_{ij}|r_{ij} = 1) \). The occurrence variable
$r_{ij}$ is assumed by a generalized logistic mixed model and is expressed as below

$$\logit(r_{ij} = 1) = \log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \mathbf{x}_{ij}\boldsymbol{\alpha} + \mathbf{z}_{ij}\boldsymbol{a}_i,$$

(4.1)

where $P_{ij} = P(r_{ij} = 1|\boldsymbol{\alpha}, \boldsymbol{a}_i)$ is the probability of $y_{ij}$ being positive conditional on random-effects $\boldsymbol{a}_i$, $\mathbf{x}_{ij}(1\times p)$ is the vector of covariates including time which may be associated with the probability of occurrence and $\mathbf{z}_{ij}(1\times q)$ is the design matrix corresponding to random-effects $\boldsymbol{a}_i$ which may include random intercept, slope and so on to vary by subject. The continuous positive part is modeled by an LMM where the model errors are assumed to follow skew-$t$ distribution as below

$$y_{ij}|y_{ij} > 0 = \mathbf{x}_{ij}^*\boldsymbol{\beta} + \mathbf{z}_{ij}^*\boldsymbol{b}_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} ST_\nu(-J(\nu)\delta, \sigma^2, \delta),$$

(4.2)

where $\mathbf{x}_{ij}^*(1\times p^*)$ and $\mathbf{z}_{ij}^*(1\times q^*)$ represent the design matrices for the fixed and random effects that may or may not be identical to those effects in Part I, $J(\nu) = (\nu/\pi)^{1/2}\{\Gamma[(\nu - 1)/2]/\Gamma(\nu/2)\}$ where $\Gamma(\cdot)$ is a Gamma function. The error term $\epsilon_{ij}$ is assumed to be distributed as ST distribution with degrees of freedom $\nu$, unknown variance parameter $\sigma^2$ and skewness parameter $\delta$. A common skewness parameter $\delta$ is assumed for overall data set (see Section 1.3 for details).

Thus, we have conditionally

$$y_{ij}|y_{ij} > 0 \sim ST_\nu(\mathbf{x}_{ij}^*\boldsymbol{\beta} + \mathbf{z}_{ij}^*\boldsymbol{b}_i - J(\nu)\delta, \sigma^2, \delta).$$

(4.3)

The random-effects $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ from the two parts are assumed jointly normal and possibly correlated:

$$\begin{pmatrix} \boldsymbol{a}_i \\ \boldsymbol{b}_i \end{pmatrix} \sim N\left(\mathbf{0}, \Psi = \begin{pmatrix} \Psi_a & \Psi_{ab} \\ \Psi_{ab} & \Psi_b \end{pmatrix}\right),$$

(4.4)

where $\Psi_a$ and $\Psi_b$ are the variance-covariance matrix for random-effects in the logistic mixed model and LMM, respectively, and $\Psi_{ab}$ denotes the covariance matrix between two models. In this paper, because we are restricted by the 4 time points per subject in real data we used, we will consider two-part models with variance-covariance matrix $\Psi$ of two $(2\times2$ structure) or three $(3\times3$ structure) random-effects defined as below (Olsen and Schafer, 2001)

$$\begin{pmatrix} a_{i0} \\ b_{i0} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{a_0} & \rho\sigma_{a_0b_0} \\ \rho\sigma_{a_0b_0} & \sigma^2_{b_0} \end{pmatrix}\right),$$

37
and
\[
\begin{pmatrix} a_{i0} \\ b_{i0} \\ b_{i1} \end{pmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{a0} & \rho \sigma_{a0b0} & \gamma \sigma_{a0b1} \\ \rho \sigma_{a0b0} & \sigma^2_{b0} & \lambda \sigma_{b0b1} \\ \gamma \sigma_{a0b1} & \lambda \sigma_{b0b1} & \sigma^2_{b1} \end{bmatrix} \right).
\]

The conditional probability density function (pdf) for \( y_{ij} \) is expressed as
\[
f(y_{ij}|a_i, b_i) = \{[1 - P_{ij}]^{1-r_{ij}} \times P_{ij} \times ST(y_{ij}|b_i)]^{r_{ij}} \}
\]
\[
= \{[1 - P_{ij}]^{1-r_{ij}} P_{ij}^{r_{ij}} \} \times [ST(y_{ij}|b_i)]^{r_{ij}}.
\]

(4.5)

Generally, the estimation of parameters \( \alpha, \beta, \sigma^2, \delta, \nu \) and \( \Psi \) is based on the likelihood function of data given as
\[
L = \prod_{i=1}^{n} \int_{a_i} \int_{b_i} \prod_{j=1}^{n_i} f(y_{ij}|a_i, b_i) f(a_i, b_i) da_i db_i
\]
\[
= \prod_{i=1}^{n} \int_{a_i} \int_{b_i} \prod_{j=1}^{n_i} \{[1 - P_{ij}]^{1-r_{ij}} P_{ij}^{r_{ij}} \} \times [ST(y_{ij}|b_i)]^{r_{ij}} f(a_i, b_i) da_i db_i,
\]

(4.6)

where the vector of random-effects \((a_i^T, b_i^T)^T\) follow multivariate normal distribution. As can be seen from (4.6), the likelihood function involves the integral with respect to the multivariate normal probability density function. Parameter estimation in the proposed models can be computationally difficult as the likelihood function depends on an analytically intractable integrals of a non-linear function with respect to the multivariate normal distribution of random-effects.

### 4.2.2 Simultaneous Bayesian inferential approach

We propose a Bayesian inference method via MCMC procedure to estimate the parameters in Parts I and II. Following the studies by Sahu et al. (2003) and Muthen and Asparouhov (2014), it can be shown, that by introducing a random variable \( w_{ij} \) based on the stochastic representation for the ST distribution (see Section 1.3 for details), \( y_{ij}|y_{ij} > 0 \) in Part II can be hierarchically formulated as

\[
y_{ij}|b_i, w_{ij} \sim t_{1, \nu}(\mathbf{x}_{ij}^* \beta + \mathbf{z}_{ij}^* \mathbf{b}_i + \delta[w_{ij} - J(\nu)], \sigma^2),
\]

\[
w_{ij} \sim t_{1, \nu}(0, 1) I(w_{ij} > 0),
\]

(4.7)

where \( w_{ij} \) follows a standard half-t distribution truncated in space \( I(w_{ij} > 0) \). Note that when \( \delta = 0 \), the ST distribution reduces to \( t \) distribution and it reduces to \( N \) distribution when \( \nu \) approaches infinity.
Let $\theta = (\alpha, \beta, \sigma^2, \delta, \nu, \Psi)$ be the collection of unknown population parameters in models (4.1), (4.4) and (4.7). To complete the Bayesian formulation, we specify prior distributions for all the unknown parameters as follows:

$$\alpha \sim N_p(\alpha_0, \Lambda_1), \beta \sim N_p(\beta_0, \Lambda_2), \sigma^2 \sim IG(\omega_1, \omega_2),$$

$$\nu \sim Exp(\omega_3)I(\nu > 2), \delta \sim N(0, \omega_4), \Psi \sim IW(\Omega, \omega_5),$$

where the mutually independent Normal ($N$), Inverse Gamma ($IG$), Inverse Wishart ($IW$) and exponential ($Exp$) distributions are chosen to facilitate computations. The prior distribution of skewness parameter $\delta$ is chosen to follow $N(0, .)$. Therefore, parameter estimate of $\delta$ can be either positive or negative determined by the data.

Let the observed data $D = \{(r_{ij}, y_{ij}, x_{ij}, z_{ij}, x_{ij}^{*}, z_{ij}^{*}); i = 1, ..., n; j = 1, ..., n_i\}$, $f(\cdot)$ be a density function, $f(\cdot|\cdot)$ be a conditional density function and $h(\cdot)$ be a prior density function. We assume that the parameters in $\theta$ are independent of each other; in other words, $h(\theta) = h(\alpha)h(\beta)h(\sigma^2)h(\nu)h(\delta)h(\Psi)$.

After we specify the models for the observed data and prior distributions of the unknown model parameters, we can draw samples for the parameters based on their posterior distributions under the Bayesian framework. Thus, the joint posterior density of $\theta$, conditional on $D$, can be given by

$$f(\theta|D) \propto \left\{ \prod_{i=1}^{n_i} \int_{a_i}^{b_i} \prod_{j=1}^{n_i} f(r_{ij}|a_i)[f(y_{ij}|b_i, w_{ij})f(w_{ij}|w_{ij} > 0)]^{r_{ij}} f(a_i, b_i)da_i db_i \right\} h(\theta),$$

where $f(r_{ij}|a_i) = [1 - P_{ij}]^{1-r_{ij}}P_{ij}^{r_{ij}}$.

In general, the integral in (4.9) is of high dimension and does not have a closed form. Analytic approximations to the integrals may not be sufficiently accurate. Therefore, it is prohibitive to directly calculate the posterior distribution of $\theta$ based on the observed data $D$. As an alternative, posterior computation of $\theta$ proceeds using an MCMC algorithm through Gibbs sampling along with Metropolis-Hastings (M-H) algorithm. In particular, Gibbs sampling is used to iteratively generate updates for parameters whose full conditional distributions are available (Sahu et al., 2003; Cooper et al., 2007; Huang and Wu, 2006; Davidian and Giltinan, 1995) and the M-H algorithm is used to update parameters where the full conditional distributions can not be expressed explicitly given the remaining parameters and data (Davidian and Giltinan, 1995; Ntzoufras, 2009).
4.3 Alcohol data analysis

4.3.1 Specific models and implementation

In this section, we analyze the AADS data from CIC study described in Chapter 2. As depicted in Figure 2.1 (a, b), AADS scales are right-skewed with excessive zeros. Toward this end, we will illustrate the performance of two-part ST and SN mixed-effects model by comparing the following statistical models with different distributional assumptions under various random-effects structures.

- Random intercept in Parts I and II denoted as $2 \times 2$ structure:
  - Model ST1: A two-part mixed-effects model with ST distribution for positive continuous values;
  - Model SN1: A two-part mixed-effects model with SN distribution for positive continuous values;
  - Model N1: A two-part mixed-effects model with normal distribution for positive continuous values;

- Random intercept in Parts I and II and random slope in Part II denoted as $3 \times 3$ structure:
  - Model ST2: A two-part mixed-effects model with ST distribution for positive continuous values;
  - Model SN2: A two-part mixed-effects model with SN distribution for positive continuous values;
  - Model N2: A two-part mixed-effects model with normal distribution for positive continuous values;

The primary predictive factor is gene 5-HTTLPR; the covariates include linear and quadratic age, gender, race and family SES as suggested by previous publications on the same study (Cohen et al., 2007). The interaction between gene and age and the interaction between gender and age are included. Although the two-part models allow different covariates in Parts I and II, we use same set of covariates for two parts so that we can assess whether there are differentiated effects of one covariate on the probability of occurrence and severity of AADS scales given the symptoms occurred. To compare the performance of models with different distributional assumptions, the
same set of covariates are included in all models. The centered age variable (i.e., age-mean age) is used in model such that the intercept coefficient estimate can be interpreted as AADS scale at the mean age. The corresponding parameters in Parts I and II are $\alpha_0$ and $\beta_0$ for intercept, $\alpha_1$ and $\beta_1$ for age, $\alpha_2$ and $\beta_2$ for quadratic age, $\alpha_3$ and $\beta_3$ for gene, $\alpha_4$ and $\beta_4$ for the interaction of gene and age, $\alpha_5$ and $\beta_5$ for gender, $\alpha_6$ and $\beta_6$ for the interaction of gender and gene, $\alpha_7$ and $\beta_7$ for race, and $\alpha_8$ and $\beta_8$ for family SES.

We conducted the following scenarios. **First**, we investigated how an asymmetric (ST and SN) distribution for model error (Models ST and SN) impacts parameter estimation in comparison with a symmetric (normal) distribution for model error (Model N). **Second**, we investigated the performance of proposed models under different random-effects structure. The Deviance Index Criterion (DIC) proposed by Spiegelhalter et al. (2002) was used as model selection criteria to evaluate and select the best model. As with other model selection criteria, we caution that DIC is not used for identifying the “correct” model, only as a tool to find the model that fits the data better. The DIC is defined as $DIC = \bar{D}(\theta) + p_D = 2\bar{D}(\theta) - D(\theta)$, where $\bar{D}(\theta) = E[D(\theta)|y]$ is the posterior mean of the deviance with a smaller value indicating better fit, and $p_D = \bar{D}(\theta) - D(\theta) = E[D(\theta)|y] - D[E(\theta|y)]$ is the effective number of parameters, defined as the difference in the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters. The structure of DIC allows for automatic computation in WinBUGS software.

To carry out the Bayesian inference, we need to specify the values of the hyper-parameters in the prior distributions. We take weakly informative prior distribution for the parameters in models. In particular, (i) each element in fixed-effects $\alpha$ and $\beta$ was taken to be independent normal distribution $N(0, 100)$; (ii) for the dispersion parameter $\sigma^2$, we assume a weakly informative inverse gamma prior distribution $IG(0.01, 0.01)$ so that the distribution has a mean value of 1 and a variance of 100; (iii) the prior for the variance-covariance matrix of the random-effects $\Psi$ was taken to be inverse Wishart distributions with $\Omega = \text{diag}(1, 1)$ and degrees of freedom (df) $\omega_5 = 2$ for $2 \times 2$, and $\Omega = \text{diag}(1, 1, 1)$ and $\omega_5 = 4$ for $3 \times 3$ random-effects structures; (iv) the skewness parameter $\delta$ was taken to be a normal distribution $N(0, 100)$; and (v) the degrees of freedom parameter $\nu$ followed a truncated exponential distribution, $\nu \sim \text{Exp}(0.1)I(\nu > 2)$.

The MCMC sampler was implemented using WinBUGS software (Lunn et al., 2000) and the program code is available from the corresponding author upon request. When the MCMC pro-
cedure was applied to the alcohol data, convergence of the generated samples was assessed using standard tools within WinBUGS software such as Gelman-Rubin (GR) diagnostics (Gelman and Rubin, 1992). Figure 4.1 shows the dynamic version of GR diagnostic plots, the trace plots and autocorrelation plots based on Model SN1 for the representative parameters $\alpha_1$, $\beta_1$ and $\delta$. The interpretation of these convergence checking tools was described in Section 3.3.1. When these criteria suggested the convergence of chains, we proposed that the three chains were run with the following considerations. For each model, we ran three, initially dispersed chains for 100,000 iterations each, discarding the first 50,000 as a burn-in period, and retained every 50th sample for a total of 3,000 samples of targeted posterior distributions of the unknown parameters to make inference.

### 4.3.2 Comparison of model fitting results

DIC is calculated separately for Parts I and II; overall DIC is also provided (Table 4.1). It is found that DIC for Part I is very similar across models. The overall DICs for Models N1, SN1, ST1, N2, SN2, and ST2 are 5119, 1999, 2810, 5147, 3112, and 4006, respectively. Therefore, based on DIC, Model SN1 is the best-fitting model, followed by Model ST1, supporting the contention of a departure from normality.

Model SN fits data better than model ST under both random-effects structures, implying that the data are not heavily tailed for the more complex Model ST to be necessary. The models with ST, SN and normal assumptions under $2 \times 2$ random-effects structure fit data better than their counterparts under $3 \times 3$ random-effects structure, indicating that adding random slope did not improve model fit. This may reflect the fact that removing zeros from LMMs reduces the individual variations in slope (i.e., rate of change).

The posterior mean (PM), the corresponding standard deviation (SD) and 95% credible interval (CI) for population parameters based on Models N1, SN1, ST1, N2, SN2, and ST2 are presented in Table 4.2. The parameter estimates from Part I are very similar across all models. In the comparison of Parts I and II in Model SN1, the covariates significantly associated with probability of the occurrence of AADS were linear age, quadratic age, gender and race. The same set of covariates were also associated with the severity of AADS given AADS occurred. Moreover, there was a significant interaction between the age of a subject and gender in Part II. The gene 5-HTTLPR and interaction between gene 5-HTTLPR and age were not significant in either part.
The positive estimate of $\alpha_1$ for linear age and negative estimate of $\alpha_2$ for quadratic age indicated that the odds of having AADS increase with subjects growing older and decrease after subjects reaching certain age. A similar phenomena was observed for severity of AADS given the symptom occurred. Male gender is associated with the higher odds of AADS occurrence and the increased severity of AADS given the occurrence of the symptoms. The significant interaction between age and gender in Part II indicates that male gender is associated with more rapid increase in severity given its occurrence (see Figure 4.2b).

Figure 4.1: Convergence diagnostics with three Markov chains as obtained from the WinBUGS software for representative parameters based on Model SN1: Gelman-Rubin(GR) diagnostic plots (top panel), trace plots (middle panel), and autocorrelation plots (bottom panel).

The estimates of the skewness parameter of Models SN1 and SN2 are, respectively, 3.59 with 95% CI (3.37, 3.78) and 3.32 with 95% CI (3.07, 3.56) which are significantly different from zero. This suggests the importance of using the skew distribution to adjust for the skewness of data, using
Table 4.1: Summary of deviance information criterion (DIC) based on Models N, SN and ST under random-effects $2 \times 2$ and $3 \times 3$ structure.

<table>
<thead>
<tr>
<th>Random-effects</th>
<th>2 × 2</th>
<th>3 × 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>N1</td>
<td>SN1</td>
</tr>
<tr>
<td>Part I - DIC</td>
<td>1047</td>
<td>1069</td>
</tr>
<tr>
<td>Part II-DIC</td>
<td>4072</td>
<td>930</td>
</tr>
<tr>
<td>Total - DIC</td>
<td>5119</td>
<td>1999</td>
</tr>
</tbody>
</table>

normality assumption would not be adequate. The estimates of 2.66 and 2.54 for the skewness parameter based on Models ST1 and ST2, respectively, are less than those from Models SN1 and SN2. This is probably because that the parameter $\nu$ for heaviness in the tails estimated to be 4.87 and 6.23 for Models ST1 and ST2 respectively causes trading-off the effect of skewness.

The parameter estimates on fixed-effects from Models SN1, SN2, ST1 and ST2 where the skewness was taken into account are basically consistent (Table 4.2). The parameter estimates from Models N1 and N2 where normality assumed are similar. In the comparison of Models SN1 and N1, a significant gender effect is observed in both models though the magnitude of gender effect is much smaller in Model SN1 (Figure 4.2). The finding that gene 5-HTTLPR is not significantly associated with AADS in Model SN1 is contradicted with that from Model N1. In Model N1, the interaction of age and gene 5-HTTLPR is significant in Part II indicating that, under normality assumption, subjects with the short allele in gene 5-HTTLPR have a larger increase in alcohol use over time than those with the long allele given the symptom occurred (Figure 4.3).

Moreover, variances of the random intercept effects $a_{i0}$ and $b_{i0}$ in the $2 \times 2$ structure of Models N1, SN1 and ST1 were highly significant indicating there are substantial variations in individuals and the random-effects were needed in the model to account for the correlation among measurements on the same subject.
Table 4.2: Summary of parameter estimates based on Models N, SN and ST under random-effects 2 × 2 and 3 × 3 structures, including estimated posterior mean (PM) for parameters of population (fixed-effects), variance, correlation structure of random-effects, skewness, lower limit (L_CI) and upper limit (U_CI) of 95% equal-tail credible interval (CI), and deviance information criterion (DIC).

<table>
<thead>
<tr>
<th>Random-effects Structure</th>
<th>Model</th>
<th>2 × 2</th>
<th>2 × 2</th>
<th>2 × 2</th>
<th>3 × 3</th>
<th>3 × 3</th>
<th>3 × 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PM</td>
<td>L_CI</td>
<td>U_CI</td>
<td>PM</td>
<td>L_CI</td>
<td>U_CI</td>
</tr>
<tr>
<td>Part I</td>
<td></td>
<td>PM</td>
<td></td>
<td></td>
<td>PM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td></td>
<td>0.14</td>
<td>0.03</td>
<td>0.02</td>
<td>0.99</td>
<td>0.03</td>
<td>0.96</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td></td>
<td>0.22</td>
<td>0.22</td>
<td>0.27</td>
<td>0.22</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td></td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td></td>
<td>-0.10</td>
<td>-0.62</td>
<td>0.42</td>
<td>-0.08</td>
<td>-0.61</td>
<td>0.45</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td></td>
<td>-0.02</td>
<td>-0.08</td>
<td>0.03</td>
<td>-0.02</td>
<td>-0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td></td>
<td>0.77</td>
<td>0.28</td>
<td>1.28</td>
<td>0.73</td>
<td>0.26</td>
<td>1.25</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td></td>
<td>0.02</td>
<td>-0.31</td>
<td>0.15</td>
<td>0.02</td>
<td>-0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td></td>
<td>1.43</td>
<td>-0.02</td>
<td>0.08</td>
<td>1.46</td>
<td>0.56</td>
<td>2.38</td>
</tr>
<tr>
<td>$\alpha_8$</td>
<td></td>
<td>-0.07</td>
<td>0.49</td>
<td>2.49</td>
<td>-0.07</td>
<td>-0.30</td>
<td>0.17</td>
</tr>
<tr>
<td>Part II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td></td>
<td>1.99</td>
<td>1.04</td>
<td>2.95</td>
<td>2.65</td>
<td>2.06</td>
<td>3.16</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td></td>
<td>0.14</td>
<td>0.10</td>
<td>0.18</td>
<td>0.04</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td></td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td></td>
<td>-0.22</td>
<td>-0.66</td>
<td>0.22</td>
<td>0.04</td>
<td>-0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td></td>
<td>-0.05</td>
<td>-0.09</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td></td>
<td>1.19</td>
<td>0.80</td>
<td>1.58</td>
<td>0.36</td>
<td>0.16</td>
<td>0.62</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td></td>
<td>0.06</td>
<td>0.03</td>
<td>0.10</td>
<td>0.02</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td></td>
<td>1.41</td>
<td>0.48</td>
<td>2.34</td>
<td>0.58</td>
<td>0.08</td>
<td>1.16</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td></td>
<td>-0.22</td>
<td>-0.41</td>
<td>-0.04</td>
<td>-0.03</td>
<td>-0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>$\delta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.59</td>
<td>3.37</td>
<td>3.78</td>
</tr>
<tr>
<td>$\nu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.87</td>
<td>3.36</td>
<td>7.30</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td></td>
<td>3.93</td>
<td>3.53</td>
<td>4.38</td>
<td>0.02</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>$\Psi$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td></td>
<td>0.82</td>
<td>0.61</td>
<td>0.93</td>
<td>0.52</td>
<td>0.11</td>
<td>0.80</td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon}$</td>
<td></td>
<td>1.87</td>
<td>1.12</td>
<td>2.85</td>
<td>1.73</td>
<td>0.98</td>
<td>2.69</td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon}$</td>
<td></td>
<td>1.50</td>
<td>1.06</td>
<td>2.04</td>
<td>0.13</td>
<td>0.05</td>
<td>0.33</td>
</tr>
<tr>
<td>$\gamma$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The significant positive correlation coefficient $\rho$ ranging from 0.52 to 0.82 has an intuitive interpretation: the high probability of incurring a symptom is associated with a high mean intensity, and vice versa. The variance of random slope $b_{i1}$ in Part II is very small compared with that of two random intercepts, and the correlation between random intercepts and random slope denoted by $\gamma$ and $\lambda$ were insignificant. This suggests that adding random slope to model does not increase model fitting meaningfully, which is consistent with the results by DIC. For interpretation of random-effects, a large and positive $a_{i0}$ implies that the $ith$ subject tends to have AADS more often than average; a large value of $b_{i0}$ implies that the $ith$ subject has more severe AADS at baseline and a large value of $b_{i1}$ implies that the $ith$ subject tends to increase AADS scale more quickly than average.

4.4 Simulation study

In this section we conduct a simulation study under longitudinal setting to evaluate the performance of Models ST, SN and N and associated method. The simulated data were generated following
Figure 4.3: Probability of occurrence of AADS and severity of AADS by gene-5-HTTLPR given its occurrence: a) from Models N1; and b) from Model SN1.

algorithm below.

$$\logit[ Pr(y_{ij} > 0) ] = \alpha_0 + \alpha_1 t_{ij} + \alpha_2 x_i + a_{i0},$$

$$y_{ij}|y_{ij} > 0 = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i + b_{i0} + e_{ij},$$

(4.10)

where the measurement time points were given as $t_{ij} = 0, 0.1, 0.2, 0.3, 0.5, 0.75, 1, 2, 3, 4$ and 5; a binary covariate $x_i = 0$ or 1; the true values of parameters in Part I were chosen as $(\alpha_0, \alpha_1, \alpha_2)^T=(-2.4, 1.0, 2.0)$ such that the percentage of zeros was approximately 50%; the true values of $\beta$ vector were $\beta=(\beta_0, \beta_1, \beta_2)^T=(1.8, 2.0, 2.0)^T$; the random-effects were set to a multivariate normal distribution $\left( \begin{array}{c} a_{i0} \\ b_{i0} \end{array} \right) \sim N\left( \begin{array}{c} 0 \\ 0 \end{array} \right, \begin{bmatrix} \sigma_a^2 & \rho \sigma_a \sigma_b \\ \rho \sigma_a \sigma_b & \sigma_b^2 \end{bmatrix} )$ with $\sigma_a=0.3, \sigma_b=0.2, \sigma_a \sigma_b = 1$ and $\rho=0.5$. We generated $e_{ij} = \varepsilon_{ij} - 1$, in which $\varepsilon_{ij}$ follows a Gamma distribution Gamma(1,1), yielding a skewed distribution with $E(e_{ij}) = 0$ and $\text{var}(e_{ij}) = 1$. Under this specification, data generated from (4.10) exhibit right skewness and a range of 0 ~ 20. We generated trajectories of 500 individuals with the first 250 subjects assigned to $x_i = 0$ and others to $x_i = 1$. The weakly informative prior distributions used in real data were adopted here.

Models ST, SN and N were fit to the simulated data set with 100 repetitions each. In each repetition, we ran a chain of 20,000 iterations with 10,000 discarded as burnin. The every 10th MCMC
sample was retained to form a total of 1000 samples for inference. Table 4.3 summarizes simulation results which include the MC mean of fixed-effects, as well as percent mean-square-errors (MSE), defined by $100 \times \sqrt{\text{MSE}_l/|\beta_l|}$ ($l = 0$, 1, 2). The results show that the proposed Models ST and SN provide similar accuracy and provide more accurate estimate of $\beta_1$ for time measurements and the estimate of $\beta_2$ for the binary covariate, compared with Model N. Our simulation results suggest that models with ST/SN distribution is more powerful in handling the zero-inflated continuous data compared with models with normal distribution.

Table 4.3: Summary of Monte Carlo simulation results for MC estimates of fixed-effects as well as MSE for Models ST, SN and N based on 100 simulated datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True value</td>
<td>-2.400</td>
<td>1.000</td>
<td>2.000</td>
<td>1.800</td>
<td>2.000</td>
<td>2.000</td>
</tr>
<tr>
<td>ST</td>
<td>MC mean</td>
<td>-2.378</td>
<td>0.987</td>
<td>1.964</td>
<td>1.820</td>
<td>2.000</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>2.381</td>
<td>1.946</td>
<td>2.809</td>
<td>3.093</td>
<td>0.249</td>
</tr>
<tr>
<td>SN</td>
<td>MC mean</td>
<td>-2.383</td>
<td>0.988</td>
<td>1.967</td>
<td>1.900</td>
<td>2.000</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>2.191</td>
<td>1.799</td>
<td>2.758</td>
<td>6.331</td>
<td>0.238</td>
</tr>
<tr>
<td>N</td>
<td>MC mean</td>
<td>2.375</td>
<td>0.984</td>
<td>1.960</td>
<td>1.805</td>
<td>2.004</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>2.283</td>
<td>2.117</td>
<td>2.910</td>
<td>3.984</td>
<td>0.620</td>
</tr>
</tbody>
</table>

Notes: MC mean=Monte Carlo mean
MSE=Percent $\sqrt{\text{MSE}} = 100 \times \sqrt{\text{MSE}_l/|\beta_l|}$.

4.5 Concluding discussion

In this paper, we evaluated two-part mixed-effects models with different distributional assumptions for positive continuous values, assuming that the positive values follow (i) skew-t distribution; (ii) skew-normal distribution and (iii) normal distribution. Based on the histogram of data, we knew that the model with normal distribution is likely to provide inaccurate estimates. The Models N1 and N2 are included for comparison purpose. The estimate of skewness parameter ranges from 2.54 to 3.59 in Models SN1, SN2, ST1 and ST2, indicating that it is necessary to take into account the skewness of positive values in modeling. Model SN fits data better than Model ST under both random-effects structures indicating that the real data are not heavily tailed. We were able to repeat the results of significant association between gene 5-HTTLPR short allele and alcohol dependence.
as reported in other publications (Van der Zwaluw et al., 2010) using the model with normality assumption. In contrast, no significant association was found using our proposed model where the normality assumption was relaxed with skew distributions. This implied that parameter estimates from model with (incorrect) normality assumption might be biased or misleading and emphasized the importance of taking into account the skewness of data in statistical modeling. Ignoring the skewness of data may lead to biased parameter estimates and falsely significant results. It is noted that the interpretation of the fixed-effects in the mixed-effects models is conditional on the level of random-effects. Marginal two-part models using the generalized estimating equation (GEE) approach were discussed by Hall and Zhang (2004) and Moulton et al. (2002) and this will be part of our future work.

We carried out sensitivity analyses to examine the dependence of parameter estimates on the choices of initial values and hyper-parameters in the prior distributions and found that the parameter estimates from proposed models were reasonable and robust under different initial values and hyper-parameter values.

We compared one-part LMMs with lognormal (LN), Ggamma, SN and normal distributions for model errors (results not presented here) and concluded that model with SN assumption is superior because it can not only capture skewed-longitudinal data well with more reasonable results, but also offers more intuitive and convenient interpretation of results because there is no data transformation involved in LMM with SN distribution. We run the two-part models under Gamma distribution (Model G) and LN (Model LN) distribution under the $2 \times 2$ random-effects structure on the real data and found that DIC values are 4808 for Model G and 4535 for Model LN, which are smaller than that from Model N but much larger than those from Models SN and ST (Table 4.2). The results from Models G and LN were not summarized in this paper because the parameter estimates are on log scale of $y_{ij}$ and can not be compared directly to those from models we proposed. We also fit Models SN and ST on the log-transformed response variable and found that the estimates of skewness parameter $\delta$ are negative which implies a left skewess. This indicated that log-transformation may over correct the skewness of data as noted by Liu et al. (2012).

An alternative way to capture the between-component correlation is to employ common random-effects in Parts I and II (e.g., random intercept $a_{i0}$ in Part I and $K \times a_{i0}$ in Part II where K is a constant); equivalently, the model specifies that the correlation of $a_{i0}$ and $b_{i0}$ is equal to one.
An advantage of the common random-effects model is that the number of parameter estimates is dramatically reduced. However, this type of model is criticized because the correlation between random-effects is forced to be one where the meaning and numerical effect of random-effects in Parts I and II are in fact different (Liu et al., 2010; Zhang and Strawderman, 2006). As a sensitivity analysis, we re-run the three models (N, SN and ST) under a common random-effects structure where $a_{i0}$ and $b_{i0}$ are random intercepts in Parts I and II, respectively, with a relation of $b_{i0} = \theta \times a_{i0}$; and in addition we include a separate and correlated random slope effect $b_{i1}$ in Part II (i.e., $(a_{i0}, b_{i1})^T \sim N(0, \Sigma)$). The results are summarize in Table 4.4. Under this random-effects structure, we found that Model SN is the best fit followed by Model ST and Model N based on DIC, which is consistent with our findings under 2 x 2 and 3 x 3 random-effect structures for the proposed models. We also found that the parameter estimates from models with common random-effects structure (Table 4.4) are very similar to those estimates described in Table 4.2 under 3 x 3 structure.

It is noted that the random-effects could also assume to follow multivariate skew distributions. Huang and Dagne (2011) have studied models where both model error and random-effects were assumed to have SN distribution and found that in comparison of random-effects with normal and random-effects with SN distribution, that the modeling results were very similar and not significantly different. In this paper, we separately developed Model SN3 in which both random-effects and model errors assumed SN distribution and found that the parameter estimates from Model SN3 are very similar to those from Model SN1 we proposed (Table 4.5) but DIC is much larger for Model SN3. Along with this finding, we assumed random-effects to have a multivariate normal distribution and model errors to have skew distribution in the two-part ST/SN models considered in this paper.

We assumed that zeros in AADS data correspond to “true zeros” indicating no presence of any symptoms. This assumption seems reasonable for self-reported symptom data. In clinical research where detection limits apply, artificial zeros may be generated where positive outcome observed are below detection limitation. Our proposed models can be easily extended to data with artificial zeros caused by under-detection limit by joint modeling with a logistic regression for “true zeros” (Gurmu and Dagne, 2012).
Table 4.4: Summary of parameter estimates based on Models N, SN and ST under common random-effects structure.

<table>
<thead>
<tr>
<th>DIC</th>
<th>N</th>
<th>SN</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PM</td>
<td>L_{c1}</td>
<td>U_{c1}</td>
</tr>
<tr>
<td>θ</td>
<td>1.029</td>
<td>0.759</td>
<td>1.388</td>
</tr>
<tr>
<td>Part I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α₀</td>
<td>0.950</td>
<td>0.054</td>
<td>1.804</td>
</tr>
<tr>
<td>α₁</td>
<td>0.210</td>
<td>0.054</td>
<td>0.263</td>
</tr>
<tr>
<td>α₂</td>
<td>-0.014</td>
<td>-0.018</td>
<td>-0.011</td>
</tr>
<tr>
<td>α₃</td>
<td>-0.116</td>
<td>-0.622</td>
<td>0.406</td>
</tr>
<tr>
<td>α₄</td>
<td>-0.024</td>
<td>-0.079</td>
<td>0.028</td>
</tr>
<tr>
<td>α₅</td>
<td>0.747</td>
<td>0.282</td>
<td>1.256</td>
</tr>
<tr>
<td>α₆</td>
<td>0.022</td>
<td>-0.027</td>
<td>0.073</td>
</tr>
<tr>
<td>α₇</td>
<td>1.453</td>
<td>0.635</td>
<td>2.345</td>
</tr>
<tr>
<td>α₈</td>
<td>-0.071</td>
<td>-0.283</td>
<td>0.148</td>
</tr>
<tr>
<td>Part II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β₀</td>
<td>1.849</td>
<td>0.908</td>
<td>2.823</td>
</tr>
<tr>
<td>β₁</td>
<td>0.150</td>
<td>0.098</td>
<td>0.201</td>
</tr>
<tr>
<td>β₂</td>
<td>-0.020</td>
<td>-0.022</td>
<td>-0.017</td>
</tr>
<tr>
<td>β₃</td>
<td>-0.219</td>
<td>-0.629</td>
<td>0.193</td>
</tr>
<tr>
<td>β₄</td>
<td>-0.047</td>
<td>-0.099</td>
<td>0.004</td>
</tr>
<tr>
<td>β₅</td>
<td>1.236</td>
<td>0.847</td>
<td>1.635</td>
</tr>
<tr>
<td>β₆</td>
<td>0.057</td>
<td>0.010</td>
<td>0.103</td>
</tr>
<tr>
<td>β₇</td>
<td>1.565</td>
<td>0.633</td>
<td>2.483</td>
</tr>
<tr>
<td>β₈</td>
<td>-0.229</td>
<td>-0.416</td>
<td>-0.037</td>
</tr>
<tr>
<td>δ</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>ν</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Table 4.5: Summary of parameter estimates based on Models SN1 and SN3.

<table>
<thead>
<tr>
<th></th>
<th>Model SN1</th>
<th></th>
<th>Model SN3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIC 1999</td>
<td></td>
<td>DIC 2946</td>
<td></td>
</tr>
<tr>
<td>Part I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>0.99</td>
<td>0.03</td>
<td>1.96</td>
<td>1.13</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>0.22</td>
<td>0.17</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>-0.01</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>-0.08</td>
<td>-0.61</td>
<td>0.45</td>
<td>-0.10</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>-0.02</td>
<td>-0.08</td>
<td>0.03</td>
<td>-0.02</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>0.73</td>
<td>0.26</td>
<td>1.25</td>
<td>0.74</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>0.02</td>
<td>-0.03</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td>1.46</td>
<td>0.56</td>
<td>2.38</td>
<td>1.33</td>
</tr>
<tr>
<td>$\alpha_8$</td>
<td>-0.07</td>
<td>-0.30</td>
<td>0.17</td>
<td>-0.06</td>
</tr>
<tr>
<td>Part II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>2.65</td>
<td>2.06</td>
<td>3.16</td>
<td>2.69</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.04</td>
<td>0.02</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.04</td>
<td>-0.16</td>
<td>0.23</td>
<td>0.03</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.01</td>
<td>-0.03</td>
<td>0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.36</td>
<td>0.16</td>
<td>0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.02</td>
<td>0.00</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>0.58</td>
<td>0.08</td>
<td>1.16</td>
<td>0.54</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-0.03</td>
<td>-0.13</td>
<td>0.06</td>
<td>-0.03</td>
</tr>
<tr>
<td>$\delta_e$</td>
<td>3.59</td>
<td>3.37</td>
<td>3.78</td>
<td>3.57</td>
</tr>
<tr>
<td>$\delta_{\alpha_i}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.53</td>
</tr>
<tr>
<td>$\delta_{\beta_i}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Chapter 5:

Bayesian inference on bivariate semi-continuous mixed-effects models with application to longitudinal substance use data

Multivariate (bivariate) correlated data encountered frequently in longitudinal studies are often analyzed using a multivariate linear mixed-effects model with normality assumption. Semi-continuous variables in the form of a mixture of high proportion of zeros and right-skewed positive values bring special challenges to the field of multivariate modeling. In this paper, we propose a Bayesian approach to analyze bivariate semi-continuous outcomes by jointly modeling a generalized logistic mixed-effects model on zero-inflation in either response and a bivariate linear mixed-effects model (BLMM) on the positive values through a correlated random-effects structure. Multivariate skew distributions including skew-\(t\) and skew-normal distributions are used to relax the normality assumption in BLMM. The proposed models are illustrated with an application to the correlated alcohol and drug uses data from a longitudinal observational study. A simulation study is conducted to evaluate the performance of the proposed models.

Keywords: Bayesian analysis; semi-continuous data; joint modeling; bivariate mixed-effects model; skew distributions.

5.1 Introduction

Semi-continuous data in the form of a mixture of zeros and skewed nonzero values arise frequently in economic, biological, and social science studies. Multivariate (bivariate) semi-continuous variables that are usually correlated are the primary outcomes in many cross-sectional and longitudinal studies. In a longitudinal study on substance abuse, for instance, the alcohol and illicit drug uses are often collected on the same subject simultaneously at each follow-up. The traditional approach for analyzing bivariate correlated continuous outcomes is via multivariate linear mixed-effect model (LMM) with normality assumption for the random-effects and within-subject errors. However,
those analyses may not provide robust inference when data exhibit skewness particularly in the situation of semi-continuous data where the extra zeros and skewed nonzero values may generate an extreme non-normality in data. One common approach for analyzing data with skewness is to apply data transformation such as log transformation to make the transformed response approximately Gaussian. Data transformation may not be appropriate in multivariate analysis because component-wise transformation does not lead to joint normality (Jara et al., 2008). Besides, the transformations may not be universal, i.e. transformation used for one variable may not be adapted for the others. Moreover, results based on transformed data are typically hard to interpret and retransformation may not be straightforward (Duan et al., 1983b). This motivates the researchers to explore the alternative parametric distributions that have flexibility in distributional assumptions of random-effects and/or model errors to produce robust parameter estimates. Ghosh et al. (2007) developed a bivariate linear mixed-effects model (BLMM) and relaxed the normality assumption by using multivariate skew-normal distributions and applied the method to HIV data, followed by more research from Bandyopadhyay et al. (2010) and Huang et al. (2014).

Two-part models originated in econometrics (Cragg, 1971; Duan et al., 1983a) have been developed extensively in the last three decades to analyze zero-inflated count and semi-continuous data. In two-part modeling, we view the zero-inflated outcome variable as the result of two processes: one binomial process determining whether the positive value occurs and one continuous process determining the actual value if it is nonzero. Zero-inflated Poisson (ZIP) model and zero-inflated Negative Binomial (ZINB) model are generally used for zero-inflated count data. Duan et al. (1983a) developed a two-part model for semi-continuous outcome with a pair of equations where a probit model was used for the probability of cost being positive and a linear model was used for the level of the positive values (log-transformed) and applied the model to a cross-sectional medical cost data. Olsen and Schafer (2001) and Tooze et al. (2002) extended the model to longitudinal setting by introducing the correlated random-effects into each component of two-part models. Log-normal distribution has been the most commonly used distributional assumption for Part II of two-part semi-continuous model.

In the analysis of multivariate zero-inflated data, bivariate ZIP and bivariate ZINB models have been developed on the bivariate zero-inflated count data (Lee et al., 2006; Li et al., 1999). Gurmu and Dagne (2012) developed a zero-inflated bivariate ordered probit regression model for the zero-
inflated ordinal data in a cross-sectional setting. However, there is limited research on bivariate semi-continuous models, particularly in longitudinal setting. Duan et al. (1983a) discussed a four-part model to distinguish inpatient and outpatient costs in the analysis of cross-section medical cost data. Liu et al. (2008b) extended the four-part model to the repeated measures of medical cost data by introducing a correlated random-effects structure and used likelihood approach for parameter estimation. The bivariate outcomes were assumed to be independent, conditional on the random-effects in their research.

Our research is motivated by a substance use data from a longitudinal observational study, in which alcohol abuse/dependence symptom (AADS) scales and marijuana abuse/dependence symptom (MADS) scales were measured simultaneously at each of four follow-up visits. Both AADS and MADS are zero-inflated followed by positive continuous values. We are interested in the correlation between the trajectories of the alcohol and drug uses and their associations with the shared underlying demographic and social-economical characteristics. Theoretically, the joint distribution of two correlated semi-continuous variables could have the following categories: i) both responses are zero (i.e., neither symptom), named as zero-zero state; ii) only AADS is zero (i.e., no alcohol abuse/dependence symptom but has marijuana abuse/dependence symptom) named as zero-one state; iii) only MADS is zero (i.e., no marijuana abuse/dependence symptom but has alcohol abuse/dependence symptom), named as one-zero state; and iv) both AADS and MADS are positive continuous (i.e., both symptoms occurred), named as one-one state. In the joint distribution of AADS and MADS in our real data, both AADS and MADS are zero in 5.3% of data, AADS being zero and MADS being > zero in 20.9%, MADS being zero and AADS being > zero in 16.0%, and both responses being positive in 57.8% of data (Table 2.1). Therefore, zero-inflation occurred at the three states where either or both responses are zero in a total of 42.2% data. Consequently, we create an indicator variable for concurrence of both symptoms (i.e., both responses are positive) versus zero in either response. It is possible that there are common underlying factors that drive the concurrence of AADS and MADS and the magnitude of the symptoms given they occurred. This leads to the correlation between the two data processes that can be jointly modeled with a random-effects structure. In addition, the correlation among repeated measurements on the same subject and the correlation among different response variables should be considered in longitudinal multivariate analysis.
In this paper, we propose a flexible joint modeling approach to the following two types of models: 
i) a generalized logistic mixed-effects model on zero-inflation; and ii) a BLMM given both responses are positive; while the two models are linked through a correlated random-effects structure. We relax the normality assumption in BLMM for the model errors with skew distributions including skew-t (ST) and skew-normal (SN) distributions in the family of skew-elliptical (SE) distribution introduced by Azzalini (1985). A detailed review on skew distributions is provided by Arellano-Valle and Genton (2005) and Genton (2004). In particular, Sahu et al. (2003) developed Bayesian inference for skewed data using SN and ST distributions followed by Jara et al. (2008). We consider a multivariate ST distribution introduced by Sahu et al. (2003) and Muthen and Asparouhov (2014) (see Section 1.3 in detail), which is suitable for a Bayesian inference. It is noted that the ST distribution is approximate to the SN distribution when its degrees of freedom approach infinity, the ST distribution reduces to a t distribution if skewness parameter is zero and the SN distribution reduces to a normal (N) distribution if skewness parameter is zero. Thus, we use an ST distribution to develop the two-part model and associated statistical methodologies, as it can be easily reverts to the other distributions such as the SN and N distributions. Three (N, SN and ST) distributions will be considered in this paper.

The rest of the paper proceeds as follows. In Section 5.2, we introduce the BLMM with a multivariate ST distribution and present the associated Bayesian inferential method. In Section 5.3, we apply the proposed method to the real data set described in Chapter 2 and report the analysis results. Section 5.4 conducts a simulation study aimed to evaluate the performance of the proposed models and method. We conclude the paper with discussion in Section 5.5.

5.2 Joint models and Bayesian inference

5.2.1 Joint models with ST distribution

Let $y_{ij}^{(a)}$ be the measurement of AADS from the $i$th ($i = 1, \cdots, n$) subject at $j$th time point ($j = 1, \cdots, n_i$) and $y_{ij}^{(m)}$ be the measurement of MADS defined similarly. An indicator variable denoting both responses being positive versus at least one being zero is defined by

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij}^{(a)} > 0 \text{ and } y_{ij}^{(m)} > 0 \\ 0 & \text{otherwise} \end{cases}$$
We model probability with \( r_{ij} = 1 \) using the following generalized (logistic) mixed-effects model.

\[
\text{logit}(r_{ij} = 1) = \log \left( \frac{P_{ij}}{1 - P_{ij}} \right) = \mathbf{x}_{ij} \alpha + \mathbf{z}_{ij} \mathbf{a}_{i},
\]

(5.1)

where \( P_{ij} = \Pr(r_{ij} = 1 | \alpha, \mathbf{a}_i) \) is the probability of both responses being positive conditional on random-effects \( \mathbf{a}_i \), \( \mathbf{x}_{ij}(1 \times p) \) is the vector of covariates including time which may be associated with the probability of occurrence of \( r_{ij} = 1 \) and \( \mathbf{z}_{ij}(1 \times q) \) is the design matrix corresponding to random-effects \( \mathbf{a}_i \), which may include random intercept, slope and so on to vary by subject.

Let \( \mathbf{x}_{ij}^{(a)} \) and \( \mathbf{x}_{ij}^{(m)} \) be the \( 1 \times p^a \) vectors of covariates associated with the fixed-effects \( \beta^{(a)} \) and \( \beta^{(m)} \) of the two responses, respectively, and \( \mathbf{z}_{ij}^{(a)} \) and \( \mathbf{z}_{ij}^{(m)} \) be the \( 1 \times q^a \) vectors of covariates associated with the random-effects \( \mathbf{b}_i^{(a)} \) and \( \mathbf{b}_i^{(m)} \) of the two responses, respectively; and the fixed and random effects may or may not be identical to those effects in model (5.1). To make notations more compact, given \( r_{ij} = 1 \), let \( \mathbf{y}_{ij} = (\mathbf{y}_{ij}^{(a)}, \mathbf{y}_{ij}^{(m)})^T \), \( \mathbf{x}_{ij}^* = \text{diag}(\mathbf{x}_{ij}^{(a)}, \mathbf{x}_{ij}^{(m)}) \), \( \mathbf{z}_{ij}^* = \text{diag}(\mathbf{z}_{ij}^{(a)}, \mathbf{z}_{ij}^{(m)}) \), \( \beta = (\beta^{(a)}T, \beta^{(m)}T)^T \), \( \mathbf{b}_i = (\mathbf{b}_i^{(a)}T, \mathbf{b}_i^{(m)}T)^T \) and \( \mathbf{e}_{ij} = (e_{ij}^{(a)}, e_{ij}^{(m)})^T \) where \( e_{ij}^{(a)} \) and \( e_{ij}^{(m)} \) are the within-subject residuals for two responses, respectively and \( \text{diag}(\mathbf{A}, \mathbf{B}) \) denotes a block diagonal matrix.

The bivariate continuous responses given \( r_{ij} = 1 \) can be modeled by a BLMM with model error following multivariate ST distribution defined as below.

\[
\mathbf{y}_{ij} | r_{ij} = 1 = \mathbf{x}_{ij}^* \beta + \mathbf{z}_{ij}^* \mathbf{b}_i + \mathbf{e}_{ij}, \quad \mathbf{e}_{ij} \sim \text{ST}_{2, \nu}(-J(\nu)\delta, \Sigma, \delta),
\]

(5.2)

where \( J(\nu) = (\nu/\pi)^{1/2} \left\{ \Gamma[(\nu - 1)/2]/\Gamma(\nu/2) \right\} \) where \( \Gamma(\cdot) \) is a gamma function. The model errors \( \mathbf{e}_{ij} \) follow a multivariate ST distribution with degrees of freedom \( \nu \) and a variance-covariance structure \( \Sigma = (\sigma_{ll'}^2)_{2 \times 2}, (l, l' = 1, 2); \) and \( \delta = (\delta^{(a)}, \delta^{(m)})^T \) with elements denoting the skewness parameters of AADS and MADS, respectively.

The random-effects in equations (5.1) and (5.2) follow a multivariate normal distribution with unknown variance-covariance matrix of \( \Psi : \)

\[
\mathbf{c}_i = (\mathbf{a}_i^T, \mathbf{b}_i^T)^T \sim N(\mathbf{0}, \Psi).
\]

In this paper, we include random intercept in each component of equations (5.1) and (5.2) such as \( a_{i0} \) in equation (5.1) and \( b_{i0}^{(a)} \) and \( b_{i0}^{(m)} \) for AADS and MADS, respectively, in equation (5.2). The random-effects follow a 3-dimensional multivariate normal distribution with unknown
variance-covariance matrix $\Psi$ as follows.

$$c_i = \begin{pmatrix} a_i \ 0 \\ b_i^{(a)} \ b_i^{(m)} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Psi \right),$$

It is easy to extend the model to include more random-effects, for example, random slope in each component of BLMM.

The conditional probability density function (pdf) for bivariate responses $y_{ij}$ is expressed as

$$f(y_{ij}|a_i, b_i) = [1 - P_{ij}]^{1-r_{ij}} \times [P_{ij} \times ST_{2, \nu}(y_{ij}|b_i)]^{r_{ij}}$$

$$= \left\{[1 - P_{ij}]^{1-r_{ij}} P_{ij}^{r_{ij}} \right\} \times \left[ ST_{2, \nu}(y_{ij}|b_i) \right]^{r_{ij}}. \quad (5.4)$$

### 5.2.2 Simultaneous Bayesian inferential approach for joint modeling

We propose a Bayesian inferential method via Markov chain Monte Carlo (MCMC) procedure to estimate the parameters in equations (5.1) and (5.2). Following the studies by Sahu et al. (2003) and Muthen and Asparouhov (2014), it can be shown that, by introducing a random vector $w_{ij} = (w_{ij}^{(a)}, w_{ij}^{(m)})^T$ with $w_{ij}^{(a)}$ and $w_{ij}^{(m)}$ being two random variables based on the stochastic representation for the ST distribution (see Section 1.3 for details), equation (5.2) can be hierarchically formulated as

$$y_{ij}|b_i, w_{ij} \sim T_{2, \nu}(x_{ij}^* \beta + z_{ij}^* b_i + \delta | w_{ij} - J(\nu)1_2), \Sigma),$$

$$w_{ij} \sim T_{2, \nu}(0, I_2) I(w_{ij} > 0), \quad (5.5)$$

where $1_2 = (1, 1)^T$, $I(w_{ij} > 0)$ is an indicator function and $w_{ij}$ is truncated in the space $w_{ij} > 0$ (i.e., standard standard half-t distribution).

Let $\theta = (\alpha, \beta, \Sigma, \Psi, \nu, \delta)$ be the collection of unknown population parameters in equations (5.1), (5.3) and (5.5). To complete the Bayesian formulation, we specify prior distributions for all the unknown population parameters as follows.

$$\alpha \sim N_p(\alpha_0, \Lambda_1), \beta \sim N_{2p^*}(\beta_0, \Lambda_2), \Sigma \sim IW(\Omega_1, \omega_1),$$

$$\Psi \sim IW(\Omega_2, \omega_2), \delta \sim N_2(\delta_0, \Lambda_3), \nu \sim Exp(\omega_3) I(\nu > 2), \quad (5.6)$$

where the mutually independent Normal ($N$), Inverse Gamma ($IG$), Inverse Wishart ($IW$) and exponential ($Exp$) distributions are chosen to facilitate computations.

We assume the same design matrices in equations (5.1) and (5.2) (i.e., $x_{ij} = x_{ij}^*$, $z_{ij} = z_{ij}^*$)
and let the observed data $\mathcal{D} = \{(r_{ij}, y_{ij}, x_{ij}, z_{ij}); i = 1, ..., n; j = 1, ..., n_i\}$, $f(\cdot|\cdot)$ be a conditional density function and $h(\cdot)$ be a prior density function. We assume that $\alpha, \beta, \Sigma, \nu, \Psi$ and $\delta$ are independent of each other; in other words, $h(\theta) = h(\alpha)h(\beta)h(\Sigma)h(\nu)h(\delta)h(\Psi)$. After we specify the models for the observed data and prior distributions of the unknown model parameters, we can draw samples for the parameters based on their posterior distributions under the Bayesian framework. Thus, the joint posterior density of $\theta$, conditional on $\mathcal{D}$, can be given by

$$f(\theta|\mathcal{D}) \propto \prod_{i=1}^{n} \int_{a_i} \int_{b_i} \prod_{j=1}^{n_i} f(r_{ij}|a_i)[f(y_{ij}|b_i, w_{ij})f(w_{ij}|w_{ij} > 0)]^{r_{ij}}f(a_i, b_i)\,da_idb_i \cdot h(\theta),$$

where $f(r_{ij}|a_i) = [1 - P_{ij}]^{1-r_{ij}}P_{ij}^{r_{ij}}$.

In general, the integral in equation (5.7) is of high dimension and does not have a closed form. Analytic approximations to the integrals may not be sufficiently accurate. Therefore, it is prohibitive to directly calculate the posterior distribution of $\theta$ based on the observed data $\mathcal{D}$. As an alternative, the MCMC procedure can be used to sample population parameters, $\theta$, and random-effects $(a_i^T, b_i^T)^T$, from conditional posterior distributions, based on equation (5.7) using the Gibbs sampler along with Metropolis-Hastings (M-H) algorithm (Cooper et al., 2007; Huang and Wu, 2006; Davidian and Giltinan, 1995). The advantage of the above representation based on hierarchical models is that they can be very easily implemented using the freely available WinBUGS software (Lunn et al., 2000).

### 5.3 Substance use data analysis

#### 5.3.1 Specific models and implementation

We illustrate proposed models and method by applying them to the AADS and MADS data described in Chapter 2. To evaluate the impact of skewness on model estimation, the models with ST and SN distributions are compared with the model with normal distribution.

- **ST Model**: Model with ST distribution for model errors.
- **SN Model**: Model with SN distribution for model errors.
- **N Model**: Model with normal distribution for model errors.
As we described in Chapter 2, zero-inflation in our real data occurred in 42.2% of data at three states (zero-zero: 5.3%, zero-one: 20.9% and one-zero: 16.0%) (Table 2.1). We are interested in evaluating the association of AADS/MADS and the baseline variables including linear and quadratic age, gender, and the interaction between gender and age, family SES and race based on a previous publication on the same data (Cohen et al., 2007). The same set of baseline variables were used in different parts of joint modeling so that we can assess whether there are differentiated effects of covariates on the probability of concurrence of AADS and MADS and severity of the responses given the symptoms occurred. To compare the performance of models with different distributional assumptions, the same set of covariates are included in all models. The centered age variable (i.e., age-mean age) is used in model such that the intercept coefficient estimate can be interpreted as AADS/MADS scale at the mean age. The corresponding parameters are $\alpha_0$, $\beta_0^{(a)}$ and $\beta_0^{(m)}$ for intercept, $\alpha_1$, $\beta_1^{(a)}$ and $\beta_1^{(m)}$ for age, $\alpha_2$, $\beta_2^{(a)}$ and $\beta_2^{(m)}$ for quadratic age, $\alpha_3$, $\beta_3^{(a)}$ and $\beta_3^{(m)}$ for gender, $\alpha_4$, $\beta_4^{(a)}$ and $\beta_4^{(m)}$ for the interaction between age and gender, $\alpha_5$, and $\beta_5^{(a)}$ and $\beta_5^{(m)}$ for race, $\alpha_6$, $\beta_6^{(a)}$ and $\beta_6^{(m)}$ for family SES.

In the absence of historical data/experiment, we specify practical weakly informative priors for all model parameters. In particular, (i) The fixed-effects are taken to be independent normal distribution $N(0,100)$ for each component of the population parameter vectors $\alpha$ and $\beta$. (ii) The prior for the variance-covariance matrices of the model errors $\Sigma$ is taken to be inverse Wishart distribution with covariance matrix $\Omega_1 = \text{Diag}(1,1)$ and degrees of freedom $\omega_1 = 2$. (iii) The prior for the variance-covariance matrix of the random-effects $\Psi$ is taken to be inverse Wishart distribution with covariance matrix $\Omega_1 = \text{Diag}(1,1,1)$ and degrees of freedom $\omega_2 = 3$. (iv) For skewness parameters $\delta^{(a)}$ and $\delta^{(m)}$, independent normal distribution $N(0,100)$ is used to accommodate either positive or negative skewness and allows the data to determine it. (v) The degrees of freedom parameter $\nu$ follows a truncated exponential distribution, $\nu \sim \text{Exp}(0.1)I(\nu > 2)$.

The MCMC sampler was implemented using WinBUGS software (Lunn et al., 2000) and the program code is available from the corresponding author upon request. When the MCMC procedure was applied to the substance use data, convergence of the generated samples was assessed using standard tools within WinBUGS software such as Gelman-Rubin (GR) diagnostics (Gelman and Rubin, 1992). Figure 5.1 shows the dynamic version of GR diagnostic plots, the trace plots and autocorrelation plots based on Model ST for the representative parameters $\alpha_2$, $\beta_2^{(a)}$ and $\delta^{(a)}$. The
interpretation of these convergence diagnosis tools is described in Section 3.3.1. When these criteria suggested the convergence of chains, we proposed that the three chains were run with the following considerations. For each model, we ran three, initially dispersed chains for 100,000 iterations each, discarding the first 50,000 as a burn-in period, and retained every 50th sample for a total of 3,000 samples of targeted posterior distributions of the unknown parameters to make inference.

Figure 5.1: Convergence diagnostics with three Markov chains as obtained from the WinBUGS software for representative parameters based on Model ST: Gelman-Rubin(GR) diagnostic plots (top panel), trace plots (middle panel), and autocorrelation plots (bottom panel).

5.3.2 Comparison of model fitting results

We conducted the following scenarios. First, we investigated how asymmetric (ST and SN) distributions for model errors (Models ST and SN) impact parameter estimation in comparison with a symmetric (normal) distribution for model error (Model N). Second, for the best model obtained above, we will report the results in detail. For selecting the best model that fits the data adequately, the Deviance Index Criterion (DIC) proposed by Spiegelhalter et al. (2002) was used. The DIC is
defined as $\text{DIC} = \bar{D}(\theta) + p_D = 2\bar{D}(\theta) - D(\bar{\theta})$, where $\bar{D}(\theta) = E[D(\theta)|y]$ is the posterior mean of the deviance with a smaller value indicating better fit, and $p_D = \bar{D}(\theta) - D(\bar{\theta}) = E[D(\theta)|y] - D(E(\theta|y)]$ is the effective number of parameters, defined as the difference in the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters. The structure of DIC allows for automatic computation in WinBUGS software. As with other model selection criteria, we caution that DIC is not intended for identification of the "correct" model, but rather merely as a method of comparing a collection of alternative formulations.

Bayesian modeling approach was used to fit the data. DIC, the posterior mean (PM), the corresponding standard deviation (SD) and 95% credible interval (CI) for fixed-effects parameters, skewness parameters, df and variance-covariance matrices based on Models N, SN and ST are presented in Table 5.1. The following findings are observed based on the modeling results.

The estimates of the vector $\alpha$ are similar across three models. The estimates of the vector $\beta$ are similar between Models SN and ST and are smaller than that from Model N. The estimates of the skewness parameters $\delta^{(a)}$ and $\delta^{(m)}$ are significantly positive from Models SN and ST providing evidence of high to moderate right-skewness of our data. The estimates (95% CI) of $\delta^{(a)}$ are 2.126 (1.812, 2.419) and 3.565 (3.292, 3.834) from Models ST and SN, respectively, and the estimates (95% CI) of $\delta^{(m)}$ are 0.476 (0.426, 0.532) and 1.313 (1.243, 1.387) from Models ST and SN, respectively. Figure 5.2 shows the boxplots for the skewness parameters, $\delta^{(a)}$ and $\delta^{(m)}$ based on 3,000 posterior samples of Models ST and SN. Note that the 95% CI does not include zero for both models, confirming that positive asymmetry. Thus, incorporating a skewness parameter in the modeling of the data is recommended. We also found that the skewness for AADS is significantly higher than MADS. The posterior mean estimates of $\nu$ for ST model is 3.036 with 95% CI (3.001, 3.136), which indicates a certain level of heavy tail in data.

The estimates of within-subject variances $\sigma^2_{11}$, $\sigma^2_{22}$ and covariance $\sigma^2_{12}$ for the two asymmetric Models (SN and ST) are smaller than those of Model N. This is expected because high variability, heaviness of the tails and skewness are interrelated to a certain extent.

The DIC values are 6540, 4764 and 4233 for Models N, SN and ST, respectively (Table 5.1). Based on DIC, Model ST is the best-fitting model, followed by Model SN, supporting the contention of a departure from normality.
Table 5.1: Summary of parameter estimates based on Models N, SN and ST, including deviance information criterion (DIC), the estimated posterior mean (PM) of parameters of population (fixed-effects), skewness, degrees of freedom, variance-covariance matrix lower limit (L CI) and upper limit (U CI) of 95% equal-tail credible interval (CI) based on Model N, SN and ST.

<table>
<thead>
<tr>
<th></th>
<th>Model N</th>
<th></th>
<th>Model SN</th>
<th></th>
<th>Model ST</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DIC</td>
<td>PM</td>
<td>L CI</td>
<td>U CI</td>
<td>PM</td>
<td>L CI</td>
</tr>
<tr>
<td>Pr(both symptoms)</td>
<td></td>
<td>6540</td>
<td>4764</td>
<td>4233</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( a_0 )</td>
<td></td>
<td>1.136</td>
<td>0.227</td>
<td>2.027</td>
<td>1.167</td>
<td>0.401</td>
</tr>
<tr>
<td>( a_1 )</td>
<td></td>
<td>0.015</td>
<td>-0.012</td>
<td>0.044</td>
<td>0.014</td>
<td>-0.014</td>
</tr>
<tr>
<td>( a_2 )</td>
<td></td>
<td>-0.023</td>
<td>-0.027</td>
<td>-0.020</td>
<td>-0.023</td>
<td>-0.027</td>
</tr>
<tr>
<td>( a_3 )</td>
<td></td>
<td>0.734</td>
<td>0.314</td>
<td>1.170</td>
<td>0.752</td>
<td>0.311</td>
</tr>
<tr>
<td>( a_4 )</td>
<td></td>
<td>0.034</td>
<td>-0.005</td>
<td>0.074</td>
<td>0.035</td>
<td>-0.002</td>
</tr>
<tr>
<td>( a_5 )</td>
<td></td>
<td>0.511</td>
<td>-0.351</td>
<td>1.399</td>
<td>0.476</td>
<td>-0.423</td>
</tr>
<tr>
<td>( a_6 )</td>
<td></td>
<td>0.003</td>
<td>-0.204</td>
<td>0.206</td>
<td>-0.028</td>
<td>-0.241</td>
</tr>
<tr>
<td>AADS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0^{(a)} )</td>
<td></td>
<td>2.878</td>
<td>2.029</td>
<td>3.757</td>
<td>3.167</td>
<td>2.535</td>
</tr>
<tr>
<td>( \beta_1^{(a)} )</td>
<td></td>
<td>0.142</td>
<td>0.104</td>
<td>0.181</td>
<td>0.061</td>
<td>0.024</td>
</tr>
<tr>
<td>( \beta_2^{(a)} )</td>
<td></td>
<td>-0.020</td>
<td>-0.023</td>
<td>-0.016</td>
<td>-0.010</td>
<td>-0.014</td>
</tr>
<tr>
<td>( \beta_3^{(a)} )</td>
<td></td>
<td>1.160</td>
<td>0.738</td>
<td>1.566</td>
<td>0.636</td>
<td>0.336</td>
</tr>
<tr>
<td>( \beta_4^{(a)} )</td>
<td></td>
<td>0.055</td>
<td>0.003</td>
<td>0.107</td>
<td>0.040</td>
<td>-0.001</td>
</tr>
<tr>
<td>( \beta_5^{(a)} )</td>
<td></td>
<td>0.631</td>
<td>-0.204</td>
<td>1.474</td>
<td>0.308</td>
<td>-0.241</td>
</tr>
<tr>
<td>( \beta_6^{(a)} )</td>
<td></td>
<td>0.003</td>
<td>-0.187</td>
<td>0.193</td>
<td>-0.052</td>
<td>-0.185</td>
</tr>
<tr>
<td>( \delta_1^{(a)} )</td>
<td></td>
<td>. . .</td>
<td>. . .</td>
<td>3.565</td>
<td>3.292</td>
<td>3.834</td>
</tr>
<tr>
<td>MADS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0^{(m)} )</td>
<td></td>
<td>0.665</td>
<td>0.286</td>
<td>1.043</td>
<td>1.138</td>
<td>0.871</td>
</tr>
<tr>
<td>( \beta_1^{(m)} )</td>
<td></td>
<td>0.007</td>
<td>-0.010</td>
<td>0.026</td>
<td>0.000</td>
<td>-0.009</td>
</tr>
<tr>
<td>( \beta_2^{(m)} )</td>
<td></td>
<td>-0.003</td>
<td>-0.004</td>
<td>-0.001</td>
<td>0.000</td>
<td>-0.001</td>
</tr>
<tr>
<td>( \beta_3^{(m)} )</td>
<td></td>
<td>0.131</td>
<td>-0.057</td>
<td>0.318</td>
<td>0.038</td>
<td>-0.041</td>
</tr>
<tr>
<td>( \beta_4^{(m)} )</td>
<td></td>
<td>0.015</td>
<td>-0.010</td>
<td>0.039</td>
<td>0.005</td>
<td>-0.008</td>
</tr>
<tr>
<td>( \beta_5^{(m)} )</td>
<td></td>
<td>0.290</td>
<td>-0.074</td>
<td>0.657</td>
<td>0.091</td>
<td>-0.084</td>
</tr>
<tr>
<td>( \beta_6^{(m)} )</td>
<td></td>
<td>0.026</td>
<td>-0.062</td>
<td>0.116</td>
<td>0.002</td>
<td>-0.037</td>
</tr>
<tr>
<td>( \delta_1^{(m)} )</td>
<td></td>
<td>. . .</td>
<td>. . .</td>
<td>1.313</td>
<td>1.243</td>
<td>1.387</td>
</tr>
<tr>
<td>( \nu ) (df)</td>
<td></td>
<td>. . .</td>
<td>. . .</td>
<td>. . .</td>
<td>3.036</td>
<td>3.001</td>
</tr>
<tr>
<td>( \Sigma ) (model errors)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sigma_{11} )</td>
<td></td>
<td>4.165</td>
<td>3.625</td>
<td>4.782</td>
<td>0.207</td>
<td>0.086</td>
</tr>
<tr>
<td>( \sigma_{22} )</td>
<td></td>
<td>0.895</td>
<td>0.793</td>
<td>1.009</td>
<td>0.020</td>
<td>0.015</td>
</tr>
<tr>
<td>( \sigma_{12} )</td>
<td></td>
<td>0.667</td>
<td>0.482</td>
<td>0.860</td>
<td>0.005</td>
<td>-0.012</td>
</tr>
<tr>
<td>Variable</td>
<td>SN</td>
<td>ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta_a$</td>
<td>1.6, 1.8, 2.0, 2.2, 2.4, 2.6</td>
<td>1.20, 1.25, 1.30, 1.35, 1.40, 1.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta_m$</td>
<td>0.40, 0.45, 0.50, 0.55</td>
<td>3.2, 3.4, 3.6, 3.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.2: Boxplots of asymmetry parameters for Models ST and SN. The upper and lower panels are boxplots for $\delta^{(a)}$ and $\delta^{(m)}$, respectively.
5.3.3 Estimation results based on ST model

Based on DIC, the best fit model is the ST model, in which model errors are assumed to have ST distribution. In Model ST, the estimated skewness parameters for AADS and MADS are 2.126 and 0.476, respectively (Table 5.1). Because the 95% CIs for both skewness parameters do not include zero, this confirms the positive skewness observed from the data. As indicated in Table 5.1, there is a positive association between linear age and the odds of concurrence of AADS and MADS, although not significant ($\alpha_1 = 0.014$ with 95% CI (-0.013, 0.041)). There is a significant negative association between quadratic age and the odds, indicating that the odds of concurrence of both symptoms decrease after subject reaching certain age ($\alpha_2 = -0.023$ with 95% CI (-0.027, -0.020)). Male gender is associated with higher odds of concurrence of AADS and MADS determined by $\alpha_3 = 0.762$ with 95% CI (0.333, 1.196), which is equivalent to an odds ratio of 2.147 for male versus female with 95% CI (1.40, 3.307). The odds of concurrence of AADS and MADS are not significantly associate with family SES or race.

The severity of AADS given both symptoms occurred increases with growing age ($\beta_1 = 0.090$ with 95% CI (0.059, 0.123)) and decreases after subjects reach certain age determined by the negative parameter estimate for quadratic age ($\beta_2 = -0.013$ with 95%CI (-0.017, -0.010)). There is a significant association between male gender and the severity of AADS ($\beta_3 = 0.691$ with 95% CI (0.388, 1.006)). Male gender is associated with a larger rate of change in severity of AADS determined by $\beta_4 = 0.035$ with 95% CI (-0.004, 0.075) although the p-value may be slightly above 0.05.

There is a trend for the association between male gender and increased MADS, although it is not significant (0.026 with 95% CI (-0.026, 0.078)). The severity of MADS is not significantly associated with any other covariates in Model ST.

5.4 Simulation study

In this section we conduct a simulation study under longitudinal setting to evaluate the performance of Models ST, SN and N and associated method. The simulated data were generated following
algorithm below.

\[
\begin{align*}
\logit[Pr(r_{ij} = 0)] &= \alpha_0 + \alpha_1 t_{ij} + \alpha_2 x_i + a_{i0}, \\
y_{ij}^{(a)} | r_{ij} = 1 &= \beta_0^{(a)} + \beta_1^{(a)} t_{ij} + \beta_2^{(a)} x_i + b_{i0}^{(a)} + e_{ij}^{(a)}, \\
y_{ij}^{(m)} | r_{ij} = 1 &= \beta_0^{(m)} + \beta_1^{(m)} t_{ij} + \beta_2^{(m)} x_i + b_{i0}^{(m)} + e_{ij}^{(m)}, \\
(a_{i0}, b_{i0}^{(a)}, b_{i0}^{(m)})^T &\sim N(0, \Psi_{3x3}),
\end{align*}
\]

where the measurement time points are given as \(t_{ij} = 0, 0.1, 0.2, 0.3, 0.5, 0.75, 1, 2, 3, 4 \text{ and } 5\); a binary covariate \(x_i = 0 \text{ or } 1\); the true values of parameter vector \(\alpha\) chosen as \((\alpha_0, \alpha_1, \alpha_2)^T = (-2.4, 1.0, 2.0)\) such that the percentage of zeros is approximately 50%; the true values of parameter vector \((\beta_0^{(a)}, \beta_1^{(a)}, \beta_2^{(a)})^T = (1.8, 2.0, 2.0)^T\) and \((\beta_0^{(m)}, \beta_1^{(m)}, \beta_2^{(m)})^T = (2.0, 1.0, 1.0)^T\), the true values are 0.09 on the diagonal of \(\Psi\) and 0.045 for off-diagonal which means that the variance of random-effects is 0.09 and correlation between random-effects is 0.5. We generated \(e_{ij}^{(a)} = \epsilon_{ij} - 1\), in which \(\epsilon_{ij}\) follows a gamma distribution Gamma(1,1), yielding a skewed distribution with \(E(\epsilon_{ij}) = 0\) and \(\text{var}(\epsilon_{ij}) = 1\). The generated data \(y_{ij}^{(a)}\) ranged approximately from 0 to 20. Similarly, we generated \(e_{ij}^{(m)} = \epsilon_{ij} - 0.5\), in which \(\epsilon_{ij}\) follows a gamma distribution Gamma(0.5,1), yielding a skewed distribution with \(E(\epsilon_{ij}) = 0\) and \(\text{var}(\epsilon_{ij}) = 0.5\). The generated \(y_{ij}^{(m)}\) ranged approximately from 0 to 10. We generated trajectories of 500 individuals with the first 250 subjects assigned to \(x_i = 0\) and others to \(x_i = 1\). The weakly informative prior distributions used in real data are adopted here.

Table 5.2: Summary of Monte Carlo simulation results for MC estimates of fixed-effects as well as MSE for Models ST, SN and N based on 100 simulated datasets.

<table>
<thead>
<tr>
<th>Model</th>
<th>(\alpha_0)</th>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
<th>(\beta_0^{(a)})</th>
<th>(\beta_1^{(a)})</th>
<th>(\beta_2^{(a)})</th>
<th>(\beta_0^{(m)})</th>
<th>(\beta_1^{(m)})</th>
<th>(\beta_2^{(m)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>True value</td>
<td>-2.40</td>
<td>1.00</td>
<td>2.00</td>
<td>1.80</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>ST</td>
<td>MC mean</td>
<td>-2.4304</td>
<td>1.0098</td>
<td>2.0309</td>
<td>1.8211</td>
<td>1.9993</td>
<td>2.0020</td>
<td>1.9626</td>
<td>0.9994</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>3.9243</td>
<td>3.4941</td>
<td>4.7861</td>
<td>2.2695</td>
<td>0.2320</td>
<td>1.6435</td>
<td>2.3032</td>
<td>0.2106</td>
</tr>
<tr>
<td>SN</td>
<td>MC mean</td>
<td>-2.4320</td>
<td>1.0109</td>
<td>2.0273</td>
<td>1.8762</td>
<td>1.9983</td>
<td>2.0159</td>
<td>2.0770</td>
<td>0.9994</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>3.5492</td>
<td>3.4773</td>
<td>4.3018</td>
<td>4.6494</td>
<td>0.2520</td>
<td>1.6971</td>
<td>4.1581</td>
<td>0.2814</td>
</tr>
<tr>
<td>N</td>
<td>MC mean</td>
<td>-2.4215</td>
<td>1.0082</td>
<td>2.0245</td>
<td>1.8126</td>
<td>1.9960</td>
<td>1.9932</td>
<td>2.0082</td>
<td>0.9988</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>3.7588</td>
<td>3.4107</td>
<td>4.6070</td>
<td>2.8589</td>
<td>0.5629</td>
<td>2.4112</td>
<td>1.8727</td>
<td>0.8096</td>
</tr>
</tbody>
</table>

Notes: MC mean=Monte Carlo mean 
\(\text{MSE} = \text{Percent } \sqrt{\text{MSE}} = 100 \times \sqrt{\text{MSE}/|\beta|}\).
We are interested in assessing the impact of distributional assumptions on the estimate of the vector $\beta$. Models ST, SN and N are fit to simulated data set with 100 repetitions each. In each repetition, we ran a chain of 20,000 iterations with 10,000 discarded as burn-in. The every 10th MCMC sample is retained for inference. Table ?? summarizes simulation results which include the MC mean of fixed-effects, as well as associated percent mean-square-errors (MSE), defined by $100 \times \sqrt{\text{MSE}_i/|\beta_l|}$ ($l = 0, 1, 2$). The parameter estimates on $\alpha_0, \alpha_1$ and $\alpha_2$ are similar across models. The proposed model ST provides the most accurate estimate of $\beta_1^{(a)}$ (time measurements) with a percent MSE of 0.232% followed by Model SN (0.252%) and Model N (0.5629%), similar pattern was observed for $\beta_1^{(m)}$. The three models provide similar accuracy on the estimates of $\beta_2^{(a)}$ and $\beta_2^{(m)}$ for the binary covariate. Our simulation results suggest that models with ST and SN distributions are more powerful in handling the zero-inflated continuous data compared with model with normal distribution.

5.5 Concluding discussion

In this paper, we propose a joint modeling approach to analyze the correlated bivariate semi-continuous outcomes in a longitudinal setting under a Bayesian framework. In bivariate semi-continuous data, the zero-inflation may occur at any or all of three states: zero-zero, zero-one and one-zero. Theoretically, we can develop the specific model for each type of zero-inflation. We can create one categorical (or binomial) variable to describe the zero-inflation status and model it with an appropriate generalized linear model. For example, if zero-inflation occurs at the zero-zero state only, a binomial variable $r_{ij}$ denoting either or both symptoms versus no symptoms can be generated and analyzed with logistic mixed model. In the situation of zero-inflation occurs at all three states, we can create an ordinal variable $r_{ij}$ representing no symptoms, one symptom versus both symptoms and analyze $r_{ij}$ with a proportional odds mixed model. Similarly, a categorical variable $r_{ij}$ can be created to denote no symptoms, alcohol symptom only, marijuana symptom only, and both symptoms and analyzed with multinomial baseline-logit model. The bivariate continuous responses conditional on $r_{ij} = 1$ can be modeled with a BLMM in which we propose to use a multivariate skew distribution to account for the skewness in the joint distribution. The generalized mixed model for $r_{ij}$ and the BLMM are then jointly modeled through a correlated random-effects structure. Therefore, the correlation between the correlated variables is accounted
for by both the random-effects structure and the variance-covariance structure of model errors. The joint modeling of proportional odds mixed model or multinomial mixed model with BLMM is more complicated and involves extra equations and parameter estimations. This topic will be explored in our future work.

To our best knowledge, there is very limited research on bivariate semi-continuous models. The four-part modeling developed by Duan et al. (1983a) and Liu et al. (2008b) used random-effects for the correlation between the two correlated bivariate semi-continuous responses and assumed independence between model errors. The independence in model errors of two response variables indicates that there is no correlation between the bivariate responses conditional on the random-effects. This is a very strong assumption. Fieuws and Verbeke (2004) found that this random-effects approach may yield misleading results in evaluating the relationship of the correlated bivariates. Our proposed models utilize both the random-effects structure and variance-covariance of model errors to account for the bivariate correlation. Relaxing the normality assumption in BLMM with multivariate ST/SN distribution reduces bias and improves efficiency in parameter estimation.
Chapter 6:

Overall discussion and conclusions

In this study, we researched the application of skew distributions including ST and SN distributions in the analyses of longitudinal semi-continuous substance use data. This chapter summarizes the new development arising from this study, the contributions of this study in terms of methodology and application, the study’s limitations, and further research directions.

Longitudinal (repeated) measures of continuous variables are routinely analyzed with LMM in which random-effects and model errors are commonly assumed a normal distribution. Although this assumption will bring convenience in the computation, it may be unrealistic and may lead to substantial bias in parameter estimations. Semi-continuous variable in the form of a spike at zero followed by right-skewed positive values may cause extreme non-normality to the data and the traditional approach of LMM is not proper for this type of data. Data transformation, including log-transformation and Box-cox transformation, is typically used to make the data approximately normal. There are several limitations to using transformation, including reduced information, difficulty in interpreting of the results, and heterogeneity in the back-transformation to the original scale. Two-part mixed-effects model based on a mixture distribution of zeros and a skewed distribution of positive values has been developed in the last ten years to analyze longitudinal semi-continuous data. However, non-normality is still an issue in Part II of the model and data transformation is commonly used. In recent years, the class of the skew distributions has attracted significant research interest for its flexibility to relax normality assumption in statistical modeling. There are different versions of skew distribution. In this paper, we use those introduced by Sahu et al. (2003) because it can be easily applied via Bayesian approach in WinBUGS. The following summarizes the main contributions of this dissertation.

In Chapter 3, we studied the application of SN distribution in the one-part LMM on alcohol use data under a Bayesian framework. We evaluated the association of gene 5-HTTLPR and alcohol use. Based on DICs, SN-LMM fits data better than N-LMM. With LMM with normality assumption
we found a marginally significant association between gene and AADS, which is consistent with findings in previous studies. However, the results can not be repeated with our proposed SN-LMM in which the normality assumption was relaxed with SN distribution. This indicates that the significant association between gene 5-HTTLPR and AADS from N-LMM might be artificial. The traditional log-transformation and Gamma regression models were also compared in the same paper. We concluded that our proposed Model SN-LMM is superior to these models in terms of DIC and data interpretation since there is no data transformation involved in SN-LMM. Our findings suggest that, although all of SN-LMM, LN-LMM and G-LMM can be employed to handle skewed-longitudinal data, the proposed SN-LMM offers important advantages in the sense that the SN-LMM can not only capture skewed longitudinal data well with more reasonable results, but can also offer more intuitive and convenient interpretation of results.

Two-part mixed-effects models are typically used for data set containing excessive zeros. However, it is understandable that if the amount of zero in data is small, two-part mixed-effects model may not be advantageous because its difficulties in computation and the potential to lead to contradictory interpretation of the covariate coefficients in the two components of two-part modeling. Su et al. (2009) suggested that when the number of zeros in longitudinal semi-continuous data, standard regression methods for the marginal distribution of outcomes, either truncated or bounded, should be considered. We conducted a simulation study to evaluate to what extent the amount of zeros increases the two-part modeling is required to handle the semi-continuous data. We used a data generation algorithm similar to the one in Section 4.4 with $\alpha_0$ varying to let the percentage of zeros to be approximately 5%, 10% and 20%. One-part SN-LMM and two-part SN-LMM are fitted to simulated data following the same Bayesian approach as described in Chapter 4. Table 6.1 summarizes simulation results for fixed effects $\beta$s. As expected, two-part SN-LMM model provided much more accurate estimates for fixed effects. For one-part SN-LMM, we found that, when the amount of zeros increases, the percent MSE for the estimate of intercept $\beta_0$ increases dramatically; the percent MSE for $\beta_1$ (time measurement) and $\beta_2$ (binary covariate) is within a reasonable range (5%) when the amount of zeros was increased to 20%. So, cautions must be used in applying one-part LMM to semi-continuous data.

Another simulation study was conducted to evaluate the performance of two-part modeling under datasets with increasing percentage of zeros. Using the same data generation algorithm in
Table 6.1: Summary of Monte Carlo simulation results for MC estimates of fixed-effects $\beta$ as well as MSE for one-part SN-LMM and two-part SN-LMM under different proportions of zeros.

<table>
<thead>
<tr>
<th>Model</th>
<th>5% zeros</th>
<th>10% zeros</th>
<th>20% zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>true values</td>
<td>1.80</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>1-part SN</td>
<td>MC mean</td>
<td>1.55</td>
<td>2.06</td>
</tr>
<tr>
<td>MSE</td>
<td>14.13</td>
<td>2.91</td>
<td>2.77</td>
</tr>
<tr>
<td>2-part SN</td>
<td>MC mean</td>
<td>1.90</td>
<td>2.00</td>
</tr>
<tr>
<td>MSE</td>
<td>5.85</td>
<td>0.12</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Notes: MC mean=Monte Carlo mean
MSE=Percent $\sqrt{MSE} = 100 \times \sqrt{MSE_l/|\beta_l|}$.

Section 4.4, we found that the parameter estimates of $\beta$s are relatively accurate even when the percentage of zeros was increased to 75% (percent MSE=6.0%,0.40% and 1.84% for $\beta_0$, $\beta_1$ and $\beta_2$, correspondingly), compared to the results in Table 6.1.)

In Chapter 4, we applied SN distribution and the more flexible ST distribution to the two-part mixed-effects model. We compared models with ST, SN and normal distributions and found that model with SN distribution best fit the real data and this is confirmed by results from simulation study. The fact that model with ST distribution is not superior to model with SN distribution is probably because that real data is not heavily tailed.

In Chapter 5, we studied the bivariate semi-continuous modeling. Multiple correlated outcomes such as AADS and MADS should be estimated in a model where their dependence on the independent variables can be considered. The application of data transformation in multivariate analysis is restricted due to reasons such as reduced information, no guarantee of joint normality, no universal transformation, or difficulty in interpreting of the results. In the joint distribution of two semi-continuous responses, the zero-inflation could occur at different states. Therefore, we proposed a joint modeling approach in which two types of models were joined with one logistic mixed-effects model for zero-inflation and one bivariate LMM on the correlated responses given both of them are nonzero. Multivariate ST and SN distributions were used in the bivariate mixed-effects model. The estimation was carried out under Bayesian framework.

A main concern in Bayesian analysis is the uncertainty regarding the prior distributions and the initial values selection. Sensitivity analysis should be conducted to check if the results are
robust against the different prior distributions and various initial values. The sensitivity analyses in this study show that the estimated parameters were not sensitive to either prior parameters or the initial values.

The AADS and MADS data from CIC study were previously studied by Cohen et al. (2007). The primary interest in Chapters 3 and 4 of this dissertation was to evaluate the association of AADS and gene 5-HTTLPR for which the data were newly available in 2010 in a subset of cohort. The findings on the association of gene 5-HTTLPR and AADS were inconsistent based on the same CIC data using different methods. We repeated the finding that gene 5-HTTLPR short allele is related to alcohol consumption using existing method, however, no significant association was found from our proposed models where ST/SN distribution was used to relax the normality assumption using the same dataset. Based on DICs from real data analyses, we know that models with ST/SN distribution provided better model fit compared with the existing models with normal, log-normal or gamma distributions. In addition, simulation studies showed that results from LMM with ST/SN distribution are more accurate than those from other existing methods. Therefore, we conclude that the significant association between gene 5-HTTLPR and AADS from the existing methods are inaccurate.

The unique data features of semi-continuous data (the spike at zero and right-skewed positive values) motivated us to combine a new parametric asymmetric distribution family and Bayesian inference. From our literature review and to our best knowledge, this study contributes to the methodology and application in the analysis of substance use in the following way: (i) extend the application of skew distributions to substance use analysis; (ii) extend the application of skew distribution family into two-part modeling; and (iii) explore bivariate analysis on alcohol and drug uses with consideration of zero-inflation and data skewness with a parametric skew distribution and hence eliminate the need for ad hoc data transformations. The proposed methods provide researchers more flexibility and more accurate inference in analyzing longitudinal semi-continuous data. The proposed methods may have a significant impact on substance use research. Our proposed method can be readily applied in other fields as long as data are zero-inflated continuous or in even more broader concept of any density function in the format of a mixture of a degenerate distribution at one point and a continuous distribution. In addition, the proposed hierarchical modeling approach can be easily implemented using the WinBUGS package that is publicly available.
These factors make our approach quite powerful and appealing to statisticians.

There are several limitations in this dissertation. Limited by 4 repeated measures per subject in the real data, we cannot test models with more than three random-effects in Chapter 4. We did not consider any missing mechanism, in longitudinal observational study on substance use, the dropout is likely to be informative, in future work, we consider to add model for nonignorable missing to the joint modeling. Marginal effects of covariates are not considered in this dissertation. This should be part of our future works. In future work, we also consider to explore models based on a mixture distribution of a degenerated density at one or more points (may not be at zero) and multivariate ST/SN for continuous distribution.
References


Ntzoufras, I. *Bayesian Modeling Using WinBUGS*, 2009; Wiley, New Jersey


Appendices

Appendix A: WinBUGS code for Chapter 4 - model ST1

```
Variables in dataset alcohol:
• y[,1] = patid
• y[,2] = AADS
• y[,3] = AADS positive flag
• y[,5] = MADS
• y[,6] = MADS positive flag
• y[,8] = age
• y[,9] = age-centered
• y[,10] = sex
• y[,11] = race
• y[,12] = family SES
• y[,15] = http_new2 (gene)

Begin of model
model {
    for (i in 1 :n) {  
        c[i,1]<-0
        # random effects for linear mixed effects
        c[i,2]<-0
        # correlated through Omega
        b[i,1:2] ~ dnorm( c[i,1:2], Omega[,])
    }
    for (j in 1:M) {  
        y[j,3] ~ dbin(p.bound[j],1)
        p.bound[j] <- max(0.001,min(p[j],0.9999))
        logit.p[j] <- logit(p[j])
        + alpha[9]*y[j,12]
    }
    for (j in 1:K) {  
        # Main components of LME model
        + delta*(w.e[j] -mue)
```

81
w.e[j] ~ dt(0,1,nu)I(0, )
ypos[j]<-ynot0[j,2]
ypos[j]~ dt(mu[j], tau,nu)
}

# Prior distributions of the hyperparameters
## (1) Coefficients
for (l in 1:9) beta[l] ~ dnorm(0,1.0E-2)
for (l in 1:9) alpha[l] ~ dnorm(0,1.0E-2)

## (2) Precision parameter
tau ~ dgamma(0.01,0.01)
sigma <- 1/tau

## (3) Variance-covariance matrice
Omega[1:2,1:2] ~ dwish(R[,],2)
v[1:2,1:2] < - inverse(Omega[,])
sigma.a <-v[1,1]
sigma.b <-v[2,2]
 rho <-v[1,2]/sqrt(sigma.a*sigma.b)

## (4) Skewness parameters
delta ~ dnorm(0, 0.01)

## (5) Degrees of freedom
nu ~ dexp(0.1) I(2,)
nue <- exp(loggam(0.5*(nu -1.))-loggam(0.5*nu ))*sqrt(nu /3.14159)

} #End of model

Appendix B: WinBUGS code for Chapter 4 - simulation code

beta.mean <-NULL
alpha.mean <-NULL
sim.summary <-data.frame(sim.nv=numerical(0),mean=numerical(0),sd=numerical(0),X2.5.=numerical(0), X25.=numerical(0), X50.=numerical(0), X75.=numerical(0),X97.5.=numerical(0))
n<500 # the number of subjects
sim.N<100 #the number of simulated datasets
shape<1
scale<1
beta<-c(1.8, 2, 2)
alpha<-c(-2.4 ,1, 2) #50% zeros
time<-c(0, 0.1, 0.2, 0.3, 0.5, 0.75, 1, 2 ,3.4,5 )
SIG1<- matrix(c(0.09, 0.03, 0.03, 0.04),2,byrow=T) #rho=0.5

# generate alcohol data

82
for(sim.n in c(1:sim.N)){
g.seed<-NULL

g.seed<-sim.n*10+1234
set.seed(g.seed)

generate random-effects
b.i<-mvtnorm(n, c(0,0),SIG1)
ID<-rep(c(1:n),length(time))
ID<--ID[order(ID)]

generate alcohol data without errors
y.simu<-NULL
for (i in c(1:n)){
x<- ifelse(i<=n/2,0,1)  # gender
xv<-rep(x,length(time))
beta.ij<-matrix(
c( rep(beta[1]+b.i[i,2], length(time)),
rep(beta[2], length(time)),
rep(beta[3]*x ,length(time))), 3,length(time),byrow=T)
y.simu <-c(y.simu, beta.ij[1,]+beta.ij[2,]*time+beta.ij[3,])
}
print(range(y.simu ))
hist(y.simu )

generate model error
ran.err <-rgamma(n*length(time), shape=shape ,scale=scale )-shape*scale
print(range(ran.err))
hist(ran.err)

generate proportion of zeros
ypos.simu<-NULL
for (i in c(1:n)){
x<-ifelse( i<=n/2, 0, 1)  # gender
for (j in c(1:length(time)) )
{
    p.ij<-exp(xB)/(1+exp(xB))  # probability of positive
    ypos.ij<-rbinom(1,1, p.ij)
    ypos.simu<-c(ypos.simu, ypos.ij)
}

print(table(ypos.simu)/length(ypos.simu))
# generate gender vector
x.simu<-NULL
for (i in c(1:n)) {
    x<- ifelse( i<=n/2,0,1)  # gender
    x.simu<-c(x.simu, rep(x, length(time)) )
}
# generate final data
DATA.simu<-cbind(ID, rep(time, n), x.simu, data.frame(y.simu+ ran.err, ypos.simu )
DATA.simu<-data.frame(DATA.simu)
names(DATA.simu)<-c('ID', 'time', 'x', 'y.orig', 'ypos')
range(DATA.simu$y.orig)
hist(DATA.simu$y.orig)
print(table(DATA.simu$y.orig<0)/length(DATA.simu$y.orig))
DATA.simu$y <- ifelse(DATA.simu$ypos==0,0,DATA.simu$y.orig )
hist(DATA.simu$y)
# # # #
R<-diag(rep(1,2))
Omega.init <- diag(rep(1,2))
M<-nrow(DATA.simu)
# order data from zero to one
y<-DATA.simu[order(DATA.simu$ypos, DATA.simu$ID, DATA.simu$time),]
y<-as.matrix(y)
K <- nrow(DATA.simu[DATA.simu$ypos == 0,]) + 1  # = positive values starting from K+1

data <- list ( n=n, M=M, K=K, y=y, R=R )

inits <- function() {
  list (delta=1, beta=c(2.2, 0.8, 1), alpha=c(1.2, 1, 0.5), Omega=Omega.init, tau=1)}

sim.model <- bugs(data, inits,
  model.file = "sim2pSN.txt", parameters = c('alpha', "beta" ),
  n.chains = 1, n.iter = 20000, n.burnin=10000, n.thin=10, debug=FALSE, codaPkg=FALSE,
  bugs.seed=654321, program=c("WinBUGS", "OpenBUGS", "winbugs", "openbugs") )

print(sim.n)
print(sim.model)

MCMC.mean <- sim.model$mean

beta.mean <- c(beta.mean, MCMC.mean$beta)

alpha.mean <- c(alpha.mean, MCMC.mean$alpha)

sim.nv <- rep(sim.n, 7)

MCMC.summary <- sim.model$summary

MCMC2.summary <- data.frame(sim.nv, sim.model$summary, row.names = NULL)

sim.summary <- rbind( sim.summary, MCMC2.summary )
}

# end loop of simulation

# estimates of beta and alpha based on sim.N simulations

alpha.mat <- matrix(alpha.mean, sim.N, 3, byrow=T)

beta.mat <- matrix(beta.mean, sim.N, 3, byrow=T)

# mean estimates of simulation

beta.final <- c(rep(0, 3))

beta.sd <- c(rep(0, 3))

beta.var <- c(rep(0, 3))

for (i in c(1:3))
  { beta.final[i] <- mean(beta.mat[, i])
    beta.sd[i] <- var(beta.mat[, i])
    beta.var[i] <- var(beta.mat[, i])
  }
beta.sd[i]<-sd(beta.mat[,i])
beta.var[i]<-var(beta.mat[,i])
}
beta.bias<-c(rep(0,3))
beta.mse<-c(rep(0,3))
for (i in c(1:3))
{beta.bias[i]<-100*(beta.final[i]-beta[i])/abs(beta[i])
  beta.mse[i]<-100*sqrt(var(beta.mat[,i])+(beta.final[i]-beta[i])**2)/abs(beta[i])
}
beta.par<-matrix(c(beta,beta.final,beta.sd,beta.bias,beta.mse),5,3,byrow=T)

# mean estimates of simulation
alpha.final<-c(rep(0,3))
alpha.sd<-c(rep(0,3))
alpha.var<-c(rep(0,3))
for (i in c(1:3))
{alpha.final[i]<-mean(alpha.mat[,i])
alpha.sd[i]<-sd(alpha.mat[,i])
alpha.var[i]<-var(alpha.mat[,i])
}
alpha.bias<-c(rep(0,3))
alpha.mse<-c(rep(0,3))
for (i in c(1:3))
{alpha.bias[i]<-100*(alpha.final[i]-alpha[i])/abs(alpha[i])
  alpha.mse[i]<-100*sqrt(var(alpha.mat[,i])+(alpha.final[i]-alpha[i])**2)/abs(alpha[i])
}
alpha.par<-matrix(c(alpha,alpha.final,alpha.sd,alpha.bias,alpha.mse),5,3,byrow=T)