Title
Estimation of Complex Generalized Linear Mixed Models for Measurement and Growth

Author
Jeon, Minjeong

Publication Date
2012

Peer reviewed|Thesis/dissertation

Estimation of Complex Generalized Linear Mixed Models for Measurement and Growth

by

Minjeong Jeon

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sophia Rabe-Hesketh , Chair
Professor Mark Wilson
Professor Cari Kaufman

Fall 2012

The dissertation of Minjeong Jeon, titled Estimation of Complex Generalized Linear Mixed Models
for Measurement and Growth, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Estimation of Complex Generalized Linear Mixed Models for Measurement and Growth

Abstract

Estimation of Complex Generalized Linear Mixed Models
for Measurement and Growth

by

Minjeong Jeon

Doctor of Philosophy in Education

University of California, Berkeley

Professor Sophia Rabe-Hesketh , Chair

Maximum likelihood (ML) estimation of generalized linear mixed models (GLMMs) is technically challenging because of the intractable likelihoods that involve high dimensional integrations over random effects. The problem is magnified when the random effects have a crossed design and thus the data cannot be reduced to small independent clusters. A variety of methods have been developed for approximating the intractable likelihood functions, but there seems no method yet that is both computationally efficient and accurate in a wide range of situations.

In this dissertation, I consider new estimation methods and applications of complex GLMMs for measurement and growth. The dissertation consists of three papers, 1) Variational maximization-maximization (MM) algorithm, 2) Monte Carlo local likelihood (MCLL) estimation, and 3) Autoregressive item response theory (IRT) growth model for longitudinal item analysis. In the first and second papers, I develop two ML methods for estimating GLMMs with crossed random effects. The variational MM algorithm is a modified expectation-maximization (EM) algorithm where a variational density is introduced in the expectation (E) step to approximate the true posterior density of the random effects given the data. The E-step is replaced by another maximization step that minimizes the Kullback-Leibler (KL) divergence between the posterior and the variational density, or equivalently, maximizes the lower bound of the log-likelihood with respect to the variational distribution. The MCLL algorithm uses the posterior samples of model parameters obtained from Markov chain Monte Carlo (MCMC) for likelihood inference. The posterior density is estimated by local likelihood density estimation and the likelihood function is approximated up to a constant by the local likelihood density estimate of the posterior divided by the prior. The performance of these new algorithms is evaluated using simulation and empirical studies and compared with other ML and Bayesian estimators. In the third paper, a new autoregressive IRT growth model is proposed to take into account serial correlations among responses to the same items over time. The consequences of ignoring serial dependence and

the initial conditions problem are investigated using simulations. The new model is applied to longitudinal data of Korean students' self-esteem.

Key words: Maximum likelihood estimation; Generalized liner mixed model; Crossed random effects; Variational approximation; MM algorithm; Local likelihood density estimation; MCLL; Autoregressive models; Local dependence; Initial conditions problem

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Sophia Rabe-Hesketh, for supporting me intellectually and emotionally over the past five years. Without her kind guidance and encouragement, none of this would have been possible. I also would like to thank Mark Wilson for his insightful comments and suggestions that have enriched my dissertation with a more scientific perspective. I would like to thank Cari Kaufman for adding a more statistical aspect to my dissertation. Her keen inspiration has helped me to discover the joy of learning statistics. I am also deeply grateful to my other mentors and formal advisors, Wim van der Linden, Juliet Shaffer, Sang-Jin Kang, and Guemin Lee for their support throughout my academic journey. I must also acknowledge Frank Rijmen's invaluable contributions to this project. Finally, my thanks go out to all my family and friends for all of their support and faith in me.

Chapter 1

General Introduction

Generalized linear mixed models (GLMMs), also known as multilevel or hierarchical generalized linear models (Raudenbush & Bryk, 2002; Goldstein, 2003; Rabe-Hesketh & Skrondal, 2012), are popular models for multilevel data with units nested in clusters. The canonical examples of multilevel data are students nested within schools and repeated measurements nested within subjects. Item response theory (IRT) models can be conceptualized as generalized linear mixed models (Rijmen et al., 2003).

Crossed random effects can be incorporated in GLMMS to handle data with two or more non-nested classifications such as students nested within schools cross-classified with neighborhoods (e.g., Goldstein, 1987; Raudenbush, 1993; McCaffrey et al., 2004). In psychometrics, crossed random effects models are also used e.g., for IRT measurement models with random item parameters (Van den Noortgate & De Boeck, 2003; De Boeck, 2008). Unlike typical IRT models that consider person as random and items as fixed, random item IRT models treat persons and items as random and the resulting model becomes a crossed random effects model. Random item IRT models are found to be useful in various settings, for instance, to account for random sampling of items from an item bank, to model item families, and to represent differential item functioning (for more examples, see e.g., De Boeck, 2008).

Maximum likelihood estimation of GLMMs is technically challenging because likelihoods often involve high dimensional intractable integrations over random effects (or latent variables). The problem is magnified when the random effects have a crossed design and thus the data cannot be reduced to small independent clusters (Vaida & Meng, 2005).

Various methods have been proposed for approximating the intractable likelihood function. For instance, the Laplace approximation (Tierney & Kadane, 1986; Lindstrom & Bates, 1988; Wolfinger, 1993) and adaptive quadrature (Naylor & Smith, 1982; Rabe-Hesketh et al., 2005; Schilling & Bock, 2005) have been widely used. The Laplace approximation and similarly, penalized quasi-likelihood (PQL; Breslow & Clayton, 1993) are known to perform poorly for small cluster sizes and for large variance components (Breslow & Lin, 1995; Joe, 2008). Adaptive quadrature is more accurate but computationally more demanding than Gaussian quadrature (Pinheiro & Bates, 1995; Rabe-Hesketh et al., 2005).

Monte Carlo (MC) methods have also been utilized in various ways for ML estimation. Most methods are based on sampling the random effects given fixed parameter estimates. Several MC expectation maximization (MCEM) algorithms have been proposed using various sampling methods: e.g., a Metropolis-Hastings (McCulloch, 1997), an independent sampler based on importance sampling or rejection sampling (Booth & Hobert, 1999), and a slice sampler (Vaida & Meng, 2005). The basic idea is to use MC samples to approximate the intractable conditional expectation for the E-step of the EM algorithm. MCEM requires samples at each iteration of the algorithm. In addition, the algorithm needs a method for calculating standard errors of the parameter estimates because it does not evaluate the likelihood function or its derivatives. A method for monitoring convergence may also be required (e.g., Booth & Hobert, 1999).

In addition, Bayesian methods have been suggested using diffuse priors to approximate ML estimates (Tanner, 1993; Diggle et al., 1994; McCulloch, 1997). However, this is often inappropriate for models with random effects because the posterior may not exist for diffuse priors (Natarajan & McCulloch, 1995; Hobert & Casella, 1996).

In this dissertation, I consider new estimation methods and applications of complex GLMMs for measurement and growth. The dissertation consists of three papers:

1. Variational maximization-maximization (MM) algorithm
2. Monte Carlo local likelihood (MCLL) method
3. Autoregressive IRT growth model for longitudinal item analysis

In the first and second papers, I develop two methods for estimating GLMMs with crossed random effects. In the third paper, I propose a new autoregressive IRT growth model that takes into account serial correlations among responses to the same items over time and apply it to longitudinal data of Korean students' self esteem. The three papers correspond to Chapters 2, 3, and 4, respectively. An abstract of each paper is provided below.

Chapter 2: Variational maximization-maximization algorithm

A variational maximization-maximization (MM) algorithm is developed for approximate maximum likelihood estimation of generalized linear mixed models with crossed random effects. The variational MM algorithm is a modified EM algorithm where the true posterior is approximated by a variational density in the E-step. The variational density function is found by minimizing the KL divergence between the posterior and the variational distribution or equivalently, maximizing the lower bound of the log-likelihood with respect to the variational distribution. The variational MM algorithm does not require a pre-specified form for the variational distribution. Models with crossed random effects can be estimated by the mean-field approximation that assumes the latent variables are conditionally independent given

the data. Adaptive quadrature is incorporated to improve the accuracy of the algorithm. Methods for estimating standard errors, evaluating the marginal likelihood, and predicting the random effects are provided. Performance of the algorithm is evaluated and compared with approximate maximum likelihood estimation based on the Laplace approximation using empirical and simulation examples.

Chapter 3:
Monte Carlo local likelihood method

A Monte Carlo local likelihood (MCLL) method is developed for estimating generalized linear mixed models (GLMMs) with crossed random effects. MCLL initially treats model parameters as random variables and samples them from the posterior for a particular prior. The likelihood function is approximated up to a constant by fitting a density to the posterior samples and dividing it by the prior. In the MCLL algorithm, the posterior density is approximated using local likelihood density estimation (Loader, 1996), where the log-likelihood is locally approximated by a polynomial function. In his Monte Carlo kernel likelihood (MCKL) method, De Valpine (2004) proposed such an approach but using kernel density estimation instead of local likelihood density estimation. A novel method to compute standard errors is developed for the MCLL method. Using empirical and simulation examples, we evaluate the MCLL algorithm and compare it to other maximum likelihood and Bayesian estimators.

Chapter 4:
Autoregressive IRT growth model for longitudinal item analysis

A first-order autoregressive or dynamic IRT growth model is proposed for longitudinal binary item analysis where responses to the same items are conditionally dependent across time given the latent trait. We show that the proposed model is equivalent to a local dependence IRT model that includes interaction parameters for responses at adjacent time points. The initial conditions problem is addressed using the method suggested by Heckman (1981) and Aitkin & Alfo (2003). The implication of this treatment is discussed with respect to measurement invariance. The proposed model is applied to longitudinal data on Korean students' self esteem. We investigate the consequences of ignoring local dependence and the initial conditions problem when the data are generated from a first-order autoregressive IRT growth model.

Notes

Some methods and applications overlap and the notation is not necessarily consistent across the three chapters.

Chapter 2

Variational Maximization-Maximization Algorithm

2.1 Introduction

Maximum likelihood estimation of generalized linear mixed models (GLMMs) is technically challenging because the likelihoods often involve high dimensional intractable integrals over random effects (or latent variables). The problem is magnified when the random effects have a crossed design and thus the data cannot be reduced to small independent clusters.

Various methods have been proposed for approximating the intractable likelihood functions. For instance, the Laplace approximation makes use of a second-order Taylor expansion of the integrand around the mode of the random effects (Tierney & Kadane, 1986; Lindstrom & Bates, 1988; Wolfinger, 1993). Penalized quasi-likelihood (PQL) uses the Laplace approximation but includes a penalty term in the approximate likelihood function (Breslow & Clayton, 1993). These approximate methods are known to perform poorly for small cluster sizes and for large variance components (Breslow & Lin, 1995; Joe, 2008).

Gaussian quadrature (Bock & Lieberman, 1970; Butler & Moffitt, 1982) has been used, which approximates integrals by a weighted average of the integrand evaluated at predetermined abscissas. The Gaussian quadrature rule can be viewed as a deterministic version of Monte Carlo integration in which random samples of the random effects are generated from a normal prior distribution (Pinheiro & Bates, 1995). Adaptive quadrature (Naylor & Smith, 1982; Pinheiro & Bates, 1995; Rabe-Hesketh et al., 2005; Schilling & Bock, 2005) is equivalent to using importance sampling in the context of Gaussian quadrature where the grid of abscissas is centered around the conditional modes or means of the random effects rather than zero. Adaptive quadrature with one quadrature point is equivalent to the Laplace approximation. For satisfactory results, Gaussian quadrature methods would require many abscissas. Adaptive quadrature is more accurate but computationally more demanding than Gaussian quadrature (Pinheiro & Bates, 1995; Rabe-Hesketh et al., 2005).

An expectation-maximization (EM) algorithm has been utilized for GLMMs where the random effects are treated as missing data (Dempster et al., 1977). To approximate the conditional expectation in the E-step, Monte Carlo (MC) methods have been used with various sampling methods: e.g., a Metropolis-Hastings (McCulloch, 1997), an independent sampler based on importance sampling or rejection sampling (Booth & Hobert, 1999), and a slice sampler (Vaida & Meng, 2005). However, MCEM is computationally demanding because it requires samples at each iteration of the algorithm and a method for monitoring convergence. Schafer (1987) used a scaled normal density function to approximate the posterior in the E-step and Steele (1996) suggested a second-order Laplace approximation for the integrals.

Variational approximation methods have been used in machine learning (Jordan et al., 1999; Jordan, 2004; Bishop, 2006). Humphreys & Titterington (2003) and Ormerod (2010) applied these ideas to statistical inference. Recently, Gaussian variational approximation methods have been proposed (Opper, 2009; Ormerod & Wand, 2012) for estimating GLMMs with nested random effects. The idea of the Gaussian variational approximation is to use a Gaussian density as a variational distribution to approximate the exact conditional distribution of the random effects given the observed data. However, the Gaussian variational approximation can be poor if the posterior is not close to Gaussian. Importantly, this method is restricted to models with nested random effects.

In this paper, we present a different version of the variational approximation method. Unlike the Gaussian variational approximation, no pre-specified form for the variational distribution is required in our algorithm. In addition, by using the mean-field approximation which treats the latent variables as conditionally independent given the data, we can estimate models with crossed random effects.

The outline of this chapter is as follows. In Section 2.2, we define the type of models that we consider. In Sections 2.3 and 2.4, the variational MM algorithm is described in detail. In Section 2.5, related issues are discussed such as estimating standard errors, evaluating the marginal likelihood, and predicting the random effects. Empirical and simulation studies are provided in Sections 2.6 and 2.7 to evaluate the proposed variational MM algorithm. The paper ends with some concluding remarks.

2.2 Model

To illustrate the proposed method, we consider a Rasch model with random item effects (e.g., De Boeck, 2008). The model is a generalized linear mixed model with crossed random effects for binary data and can be written as

$$\text{logit}\left(p\left(y_{is}=1|\theta_s,\delta_i\right)\right)=\text{logit}(\pi_{is})=\beta+\theta_s+\delta_i, \tag{2.1}$$

where y_{is} denotes the binary response for item i and person s with $i=1,...,I$ and $s=1,...,N$. β is a fixed intercept, θ_s is the person ability with density $p(\theta_s;\boldsymbol{\gamma})$, and $-\delta_i$ is the item

difficulty with density $p(\delta_i;\boldsymbol{\xi})$ where $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ are the parameter vectors that characterize the distributions of θ_s and δ_i, respectively.

The likelihood function for model (2.1) is obtained by integrating over the vectors of latent variables $\boldsymbol{\theta} = (\theta_1, ..., \theta_N)'$ and $\boldsymbol{\delta} = (\delta_i, ..., \delta_I)'$

$$L(\mathbf{y};\boldsymbol{\Psi}) = \int_{\theta_1}\cdots\int_{\theta_N}\int_{\delta_1}\cdots\int_{\delta_I} p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\delta})\left(\prod_s p(\theta_s;\boldsymbol{\gamma})\right)\left(\prod_i p(\delta_i;\boldsymbol{\xi})\right) d\delta_I\cdots d\delta_1\, d\theta_N\cdots d\theta_1,$$

where $\mathbf{y}$ is the vector of responses for all persons and items, $\boldsymbol{\Psi}$ the vector of all parameters, $\boldsymbol{\Psi} = (\beta, \boldsymbol{\xi}', \boldsymbol{\gamma}')'$ and $p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\delta})$ is the joint probability of all observed responses given the latent variables

$$p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\delta}) = \prod_i\prod_s p(y_{is}=1|\theta_s,\delta_i).$$

Later we will specify discrete or normal prior distributions for $p(\delta_i;\boldsymbol{\xi})$ and $p(\theta_s;\boldsymbol{\gamma})$.

2.3 Variational MM algorithm

The EM algorithm is a powerful tool for maximum likelihood estimation of models with missing data or latent variables (Dempster et al., 1977). The algorithm alternates between an E-step and an M-step: In the E-step, the expectation of the complete data log-likelihood, $\log f(\mathbf{y},\mathbf{z};\boldsymbol{\Psi})$ is computed over the posterior distribution of the latent variables $\mathbf{z} = (\boldsymbol{\theta},\boldsymbol{\delta})$ or missing data given the observed data $\mathbf{y}$ and given current parameter estimates. In the M-step, the posterior expectation computed in the E-step (often called Q function) is maximized with respect to the model parameters to produce updated estimates. The steps are repeated until convergence.

In the variational MM algorithm, the traditional E-step is modified by using a variational approximation. To describe the algorithm, we define the Q function at the mth iteration as

$$\begin{aligned} Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(m)}) &= E\left\{\log f(\mathbf{y},\mathbf{z};\boldsymbol{\Psi})|\mathbf{y};\boldsymbol{\Psi}^{(m)}\right\} \\ &= \int_{\mathbf{z}} p\left(\mathbf{z}|\mathbf{y};\boldsymbol{\Psi}^{(m)}\right)\log f\left(\mathbf{y},\mathbf{z};\boldsymbol{\Psi}\right)d\mathbf{z}, \end{aligned}$$

where $\boldsymbol{\Psi}^{(m)}$ are the current parameter estimates and $p\left(\mathbf{z}|\mathbf{y};\boldsymbol{\Psi}^{(m)}\right)$ is the probability density of the latent variables given the data for the current parameter estimates. The Q function cannot be evaluated analytically due to the integral over the posterior distribution $p(\mathbf{z}|\mathbf{y};\boldsymbol{\Psi}^{(m)})$. The variational MM algorithm replaces the posterior distribution $p(\mathbf{z}|\mathbf{y};\boldsymbol{\Psi}^{(m)})$ by a tractable alternative probability density function $g(\mathbf{z})$. The variational density function $g(\mathbf{z})$ is found by minimizing the Kullback-Leibler (KL) divergence (Shorack & Wellner, 1986,

p.159) between $p(\mathbf{z}|\mathbf{y};\mathbf{\Psi}^{(m)})$ and $g(\mathbf{z})$

$$\mathrm{KL}\left(g(\mathbf{z}),p(\mathbf{z}|\mathbf{y};\mathbf{\Psi}^{(m)})\right)=\int_{\mathbf{Z}}g(\mathbf{z})\log\frac{g(\mathbf{z})}{p(\mathbf{z}|\mathbf{y};\mathbf{\Psi}^{(m)})}\,d\mathbf{z}. \tag{2.2}$$

$\mathrm{KL}\left(g(\mathbf{z}),p(\mathbf{z}|\mathbf{y};\mathbf{\Psi}^{(m)})\right)$ is strictly positive and zero if and only if $g(\mathbf{z})=p(\mathbf{z}|\mathbf{y};\mathbf{\Psi}^{(m)})$ almost everywhere (Kullback & Leibler, 1951).

Equivalently, it can be shown that minimizing the KL in (2.2) is the same as maximizing a lower bound of the log-likelihood. The lower bound can be derived using Jensen's inequality

$$\begin{aligned}
l(\mathbf{y};\mathbf{\Psi}) &\equiv \log\int_{\mathbf{Z}} f\left(\mathbf{y},\mathbf{z};\mathbf{\Psi}\right)d\mathbf{z}\\
&=\log\int_{\mathbf{Z}} g(\mathbf{z})\frac{f\left(\mathbf{y},\mathbf{z};\mathbf{\Psi}\right)}{g(\mathbf{z})}d\mathbf{z}\\
&=\log E_g\left\{\frac{f\left(\mathbf{y},\mathbf{z};\mathbf{\Psi}\right)}{g(\mathbf{z})}\right\}\\
&\geq E_g\left\{\log\frac{f\left(\mathbf{y},\mathbf{z};\mathbf{\Psi}\right)}{g(\mathbf{z})}\right\}\\
&=E_g\{\log f(\mathbf{y},\mathbf{z};\mathbf{\Psi})\}-E_g\{\log g(\mathbf{z})\}\\
&\equiv \underline{l}\left(\mathbf{y};\mathbf{\Psi}\right),
\end{aligned} \tag{2.3}$$

where $l(\mathbf{y};\mathbf{\Psi})$ is the log-likelihood and E_g denotes the expectation over the latent variables $\mathbf{z}$ with density $g(\mathbf{z})$. The first term in the fifth line of (2.3) is an approximation to the Q function.

In order to show the relationship between the KL divergence and the lower bound, rewrite the KL divergence in (2.2)

$$\begin{aligned}
\mathrm{KL}\left(g(\mathbf{z}),p(\mathbf{z}|\mathbf{y})\right) &= \int_{\mathbf{Z}} g(\mathbf{z})\log\frac{g(\mathbf{z})}{p(\mathbf{z}|\mathbf{y};\mathbf{\Psi})}\,d\mathbf{z}\\
&=E_g\left\{\log g(\mathbf{z})\right\}-E_g\left\{\log p(\mathbf{z}|\mathbf{y};\mathbf{\Psi})\right\}\\
&=E_g\left\{\log g(\mathbf{z})\right\}-E_g\left\{\log\left(\frac{f(\mathbf{y},\mathbf{z};\mathbf{\Psi})}{p(\mathbf{y};\mathbf{\Psi})}\right)\right\}\\
&=E_g\left\{\log g(\mathbf{z})\right\}-E_g\left\{\log f(\mathbf{y},\mathbf{z};\mathbf{\Psi})\right\}+\log p(\mathbf{y};\mathbf{\Psi}),
\end{aligned}$$

where the third line is based on Bayes theorem. In the last line, the first two terms are $E_g\left\{\log g(\mathbf{z})\right\}-E_g\left\{\log f(\mathbf{y},\mathbf{z};\mathbf{\Psi})\right\}=-\underline{l}\left(\mathbf{y};\mathbf{\Psi}\right)$ and the third term $\log p(\mathbf{y};\mathbf{\Psi})$ is the marginal log-likelihood $l(\mathbf{y};\mathbf{\Psi})$. Therefore, the following decomposition holds for the marginal log-

likelihood

$$l(\mathbf{y};\boldsymbol{\Psi}) = \underline{l}\,(\mathbf{y};\boldsymbol{\Psi}) + \mathrm{KL}\,(g(\mathbf{z}), p(\mathbf{z}|\mathbf{y}))\,.$$

That is, the KL divergence $\mathrm{KL}\,(g(\mathbf{z}), p(\mathbf{z}|\mathbf{y}))$ describes the difference between the marginal log-likelihood and the lower bound. Thus, minimizing KL is equivalent to maximizing the lower bound $\underline{l}\,(\mathbf{y};\boldsymbol{\Psi})$ (Bishop, 2006, p.451).

The maximization-maximization (MM) algorithm (MM-algorithm; Neal & Hinton, 1998) consists of two maximization steps. The first M-step involves maximizing the lower bound $\underline{l}\,(\mathbf{y};\boldsymbol{\Psi}^{(m)})$ with respect to $g(\mathbf{z})$ given the current parameter estimates $\boldsymbol{\Psi}^{(m)}$ and the second M-step involves maximizing $\underline{l}\,(\mathbf{y};\boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$ given the current variational approximation $g(\mathbf{z})$.

It is clear that the quality of the variational MM-algorithm depends on the choice of $g(\mathbf{z})$. Ideally, $g(\mathbf{z})$ should resemble the true model-based posterior distribution $p\,(\mathbf{z}|\mathbf{y};\boldsymbol{\Psi})$ and make the integrals computationally tractable. The mean-field approximation assumes complete factorizability (or independence) of the latent variables $\mathbf{z}$ under the posterior (Hall et al., 2002; Bishop, 2006). The lower bound $\underline{l}\,(\mathbf{y};\boldsymbol{\Psi}^{(m)})$ then takes a relatively simple form $g(\mathbf{z}) = \prod_i g_i(z_i)$, where z_i is the ith element of $\mathbf{z}$ and $g_i(z_i)$ is the corresponding marginal density.

For model (2.1), the mean-field approximation is

$$\begin{aligned} g(\mathbf{z}) &= g\,(\boldsymbol{\delta}, \theta) \\ &\approx \left(\prod_i g_i\,(\delta_i)\right)\left(\prod_s g_s\,(\theta_s)\right). \end{aligned}$$

As a refinement of the mean-field approximation, one may use a different type of approximation, e.g., based on a mixture distribution where each of the component distributions is an independent distribution (Bishop et al., 1998; Humphreys & Titterington, 2003). Although these alternative refinements may give sharper lower bounds, they introduce extra complications to the algorithm, for example, requiring extra variational parameters. In addition, they may work only for particular problems (Humphreys & Titterington, 2003).

Hence, the mean-field approximation is a practical choice. It is easy to implement and works well for models with complex random effect structures. For instance, Rijmen & Jeon (in press) adopted a discrete mean-field approximation for estimating a complex generalized linear mixed model with crossed random effects and reported good precision of the method.

In a later section, the appropriateness of the mean-field approximation is investigated by examining posterior correlations of the random effects as a function of sample sizes and prior variances.

2.4 Implementation

We derive the first M-step of the algorithm without specifying functional forms for the latent variable distribution or for the variational approximation. The lower bound to the log-likelihood for model (2.1) can be written as

$$\begin{aligned}
\underline{l} &= \int_{\boldsymbol{\theta},\boldsymbol{\delta}} \left[\sum_s \log p(\theta_s) + \sum_i \log p(\delta_i) + \sum_i \sum_s \log p(y_{is}|\theta_s, \delta_i) - \sum_s \log g_s(\theta_s) - \sum_i \log g_i(\delta_i) \right] \\
&\times g(\boldsymbol{\theta}, \boldsymbol{\delta}) d(\boldsymbol{\theta}) d(\boldsymbol{\delta}) \\
&= \sum_s \int_{\theta_s} g_s(\theta_s) \log p(\theta_s) d\theta_s + \sum_i \int_{\delta_i} g_i(\delta_i) \log p(\delta_i) d\delta_i \\
&+ \sum_i \sum_s \int_{\theta_s} \int_{\delta_i} g_i(\delta_i) g_s(\theta_s) \log p(y_{is}|\theta_s, \delta_i) d\delta_i d\theta_s \\
&- \sum_s \int_{\theta_s} g_s(\theta_s) \log g_s(\theta_s) d\theta_s - \sum_i \int_{\delta_i} g_i(\delta_i) \log g_i(\delta_i) d\delta_i. \qquad (2.4)
\end{aligned}$$

Here we have used the mean-field approximation by assuming a fully factorized form for $g(\boldsymbol{\theta}, \boldsymbol{\delta})$. We maximize (2.4) with respect to $g_s(\theta_s)$ and $g_i(\delta_i)$, by means of the calculus of variations (or functional derivatives), subject to the constraints that these densities integrate to 1. Rewriting (2.4) and adding Lagrange multipliers for the constraints, we obtain

$$\begin{aligned}
F &= \sum_s \int_{\theta_s} g_s(\theta_s) \log p(\theta_s) d\theta_s + \sum_i \int_{\delta_i} g_i(\delta_i) \log p(\delta_i) d\delta_i \\
&+ \sum_i \sum_s \int_{\theta_s} \int_{\delta_i} g_i(\delta_i) g_s(\theta_s) \log p(y_{is}|\theta_s, \delta_i) d\delta_i d\theta_s \\
&- \sum_s \int_{\theta_s} g_s(\theta_s) \log g_s(\theta_s) d\theta_s - \sum_i \int_{\delta_i} g_i(\delta_i) \log g_i(\delta_i) d\delta_i \\
&+ \sum_s \lambda_s \left[\int_{\theta_s} g_s(\theta_s) d\theta_s - 1 \right] + \sum_i \lambda_i \left[\int_{\delta_i} g_i(\delta_i) d\delta_i - 1 \right]. \qquad (2.5)
\end{aligned}$$

Here λ_s and λ_i are the Lagrange multipliers for the normalization constraints on $g_s(\theta_s)$ and $g_i(\delta_i)$.

We optimize this functional F with respect to $g_s(\theta_s)$ and $g_i(\delta_i)$. In Appendix A, the idea of a functional derivative is illustrated with a simple example. For more information on the calculus of variations, see Sagan (1969) and Bishop (2006, Appendix D). The solutions for

$\log g_s(\theta_s)$ and $\log g_i(\delta_i)$ can be obtained as

$$\log g_s(\theta_s) = \log p(\theta_s) + \sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i - 1 + \lambda_s, \tag{2.6}$$
$$\log g_i(\delta_i) = \log p(\delta_i) + \sum_s \int_{\theta_s} g_s(\theta_s) \log p(y_{is}|\theta_s, \delta_i) d\delta_i - 1 + \lambda_i.$$

By exponentiating the first equation and integrating over θ_s, we obtain

$$1 = \exp(-1 + \lambda_s) \int_{\theta_s} p(\theta_s) \exp\left(\sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i \right) d\theta_s,$$
$$-1 + \lambda_s = -\log \int_{\theta_s} p(\theta_s) \exp\left(\sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i \right) d\theta_s. \tag{2.7}$$

Substituting (2.7) for $-1 + \lambda_s$ in (2.6), we obtain a solution for $\log g_s(\theta_s)$

$$\log g_s(\theta_s) = \log p(\theta_s) + \sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i$$
$$- \log \int_{\theta_s} p(\theta_s) \exp\left(\sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i \right) d\theta_s.$$

Thus, a solution for $g_s(\theta_s)$ can be obtained as

$$g_s(\theta_s) = \frac{p(\theta_s) \exp\left(\sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i \right)}{\int_{\theta_s} p(\theta_s) \exp\left(\sum_i \int_{\delta_i} g_i(\delta_i) \log p(y_{is}|\theta_s, \delta_i) d\delta_i \right) d\theta_s}. \tag{2.8}$$

Similarly, a solution for $g_i(\delta_i)$ can be obtained as

$$g_i(\delta_i) = \frac{p(\delta_i) \exp\left(\sum_s \int_{\theta_s} g_s(\theta_s) \log p(y_{is}|\theta_s, \delta_i) d\theta_s \right)}{\int_{\delta_i} p(\delta_i) \exp\left(\sum_s \int_{\theta_s} g_s(\theta_s) \log p(y_{is}|\theta_s, \delta_i) d\theta_s \right) d\delta_i}. \tag{2.9}$$

Note that Equations (2.8) and (2.9) represent a set of consistency conditions for the maximum of the lower bound subject to the factorization constraint (Bishop, 2006, p.466). These are not explicit solutions yet because $g_s(\theta_s)$ and $g_i(\delta_i)$ depend on expectations computed with respect to $g_i(\delta_i)$ and $g_s(\theta_s)$, respectively. Therefore, consistent solutions can be obtained by first initializing and then iteratively updating the variational approximations. Convergence is guaranteed because the lower bound is convex with respect to the factors of

$g(\mathbf{z})$ (Boyd & Vandenberghe, 2004).

The general expressions for the solutions in (2.8) and (2.9) involve integrals over the prior and the approximate posterior distribution of the latent variables. The explicit solutions for these and thus the rest of the algorithm (the second M-step) differ according to the choice of the prior distributions for θ_s and δ_i. In the next subsections, we describe two general choices for the prior, continuous (normal) and discrete prior distributions.

2.4.1 Normal Priors

Here we specify normal priors $p(\theta_s) = \phi(\theta_s; 0, \tau_\theta)$ and $p(\delta_i) = \phi(\delta_i; 0, \tau_\delta)$, where $\phi(\cdot; \mu, \sigma)$ denotes a normal density with mean μ and standard deviation σ. Then, we rewrite model (2.1) as

$$\text{logit}\left(p\left(y_{is} = 1 | u_{\theta s}, u_{\delta i}\right)\right) = \beta + \tau_\theta u_{\theta s} + \tau_\delta u_{\delta i},$$

where $u_{\theta s}$ and $u_{\delta i}$ are standard normal variables. The solutions for the random effects $u_{\theta s}$ and $u_{\delta i}$ are given in (2.8) and (2.9), where θ_s is replaced by $u_{\theta s}$, and δ_i is replaced by $u_{\delta i}$.

The integrals in both expressions can be approximated by Gaussian quadrature. For example, the integral in the numerator of (2.8) becomes

$$\begin{aligned}
&\int_{u_{\delta i}} g_i(u_{\delta i}) \text{log} p(y_{is} | u_{\theta s}, u_{\delta i}) du_{\delta i} \\
&= \int_{u_{\delta i}} \frac{g_i(u_{\delta i}) \text{log} p(y_{is} | u_{\theta s}, u_{\delta i})}{\phi(u_{\delta i})} \phi(u_{\delta i}) du_{\delta i} \\
&\approx \sum_d \frac{g_i(l_d) \text{log} p(y_{is} | u_{\theta s}, u_{\delta i} = l_d)}{\phi(l_d)} w_d,
\end{aligned}$$

where the prior density is used as a weight function in the second line. In the third line, $\phi(\cdot)$ is a standard normal density, and the Gauss-Hermite quadrature rule is applied where l_d and w_d are the quadrature locations and corresponding weights for integrating over $u_{\delta i}$.

Similarly, the integral in the denominator of (2.8) becomes

$$\begin{aligned}
&\int_{u_{\theta s}} \phi(u_{\theta s}) \exp\left(\sum_i \int_{u_{\delta i}} g_i(u_{\delta i}) \text{log} p(y_{is} | u_{\theta s}, u_{\delta i}) du_{\delta i}\right) du_{\theta s} \\
&\approx \int_{u_{\theta s}} \phi(u_{\theta s}) \exp\left(\sum_i \sum_d \frac{g_i(l_d) \text{log} p(y_{is} | u_{\theta s}, u_{\delta i} = l_d)}{\phi(l_d)} w_d\right) du_{\theta s} \\
&\approx \sum_t w_t \exp\left(\sum_i \sum_d \frac{g_i(l_d) \text{log} p(y_{is} | u_{\theta s} = l_t, u_{\delta i} = l_d)}{\phi(l_d)} w_d\right),
\end{aligned}$$

where l_t and w_t are the quadrature locations and corresponding weights for integrating over

$u_{\theta s}$. Similarly, we approximate the integrals in (2.9) using Gaussian quadrature.

Note that the variational parameters are the posterior probabilities $g_i(l_d)$ and $g_s(l_t)$ at the locations defined by the quadrature points.

In the second M-step, the lower bound is optimized with respect to the model parameters, $\Psi = (\beta, \tau_\theta, \tau_\delta)'$. For example, with respect to β, the solution for $\hat{\beta}$ is found by solving

$$\begin{aligned}
\frac{d\underline{l}}{d\beta} &= \frac{d}{d\beta}\int_{\boldsymbol{\theta},\boldsymbol{\delta}} \left[\sum_i \sum_s \log p(y_{is}|u_{\theta s}, u_{\delta i})\right] g(\boldsymbol{\theta}, \boldsymbol{\delta}) d\boldsymbol{\theta} d\boldsymbol{\delta} \\
&= \sum_i \sum_s \int_{u_{\delta i}} \int_{u_{\theta s}} g_i(u_{\delta i}) g_s(u_{\theta s}) \frac{d}{d\beta} \log p(y_{is}|u_{\theta s}, u_{\delta i}) du_{\theta s} du_{\delta i} \\
&\approx \sum_i \sum_s \sum_d \sum_t \frac{g_i(l_d) g_s(l_t)}{\phi(l_d)\phi(l_t)} \frac{d}{d\beta} \log p(y_{is}|u_{\theta s} = l_t, u_{\delta i} = l_d) w_d w_t = 0. \qquad (2.10)
\end{aligned}$$

The solutions for the variance parameters τ_θ and τ_δ can be obtained in a similar way. Notice that Equation (2.10) corresponds to the score function of a generalized linear model with frequencies $g_i(l_d)g_s(l_t)/\phi(l_d)\phi(l_t)$.

Adaptive Quadrature

A more efficient numerical integration method is adaptive quadrature which takes into account the location and spread (mean and standard deviation or mode and curvature) of the integrand. The quadrature locations are scaled and translated to be placed under the peak of the integrand (Rabe-Hesketh et al., 2005). Specifically, adaptive quadrature can be applied to the numerator of (2.8)

$$\begin{aligned}
&\int_{u_{\delta i}} g_i(u_{\delta i}) \log p(y_{is}|u_{\theta s}, u_{\delta i}) du_{\delta i} \\
&= \int_{u_{\delta i}} \frac{g_i(u_{\delta i}) \log p(y_{is}|u_{\theta s}, u_{\delta i})}{\phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}})} \phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}}) du_{\delta i} \\
&\approx \sum_d \frac{g_i(l_{id}) \log p(y_{is}|u_{\theta s}, u_{\delta i} = l_{id})}{\phi(l_d)} \sigma_{u_{\delta i}} w_{id} \\
&= \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s}, u_{\delta i} = l_{id}) w_{id}, \qquad (2.11)
\end{aligned}$$

where

$$\begin{aligned}
l_{id} &= \mu_{u_{\delta i}} + \sigma_{u_{\delta i}} l_d, \\
w_{id} &= \frac{\sigma_{u_{\delta i}} w_d}{\phi(l_d)},
\end{aligned}$$

are the item-specific quadrature locations and weights for integrating over $u_{\delta i}$. In the second line in (2.11), the variational approximation to the posterior $g_i(u_{\delta i})$ is approximated by $\phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}})$ where $\mu_{u_{\delta i}}$ and $\sigma_{u_{\delta i}}$ are the posterior mean and standard deviation for $u_{\delta i}$. In the third line, the variable of integration was changed to a standard normal variable.

The adaptive quadrature method works well if the ratio in the third line of (2.11) is well approximated by a low-order polynomial (Liu & Pierce, 1994). In (2.11), $\log p(y_{is}|u_{\theta s}, u_{\delta i})$ is (mirrored) S-shaped as a function of $u_{\delta i}$ and the denominator is a normal approximation of $g_i(u_{\delta i})$. Thus, the integrand in the first line of (2.11) is likely to be a unimodal and smooth function.

The variational parameters are now the posterior probabilities $g_i(l_{id})$ of the item-specific adaptive quadrature locations l_{id}. Applying the same logic to the integral in the numerator of (2.9), the variational parameters are the posterior probabilities of the person-specific adaptive quadrature locations.

Similarly, adaptive quadrature can be applied to the denominator of (2.8)

$$
\begin{aligned}
&\int_{u_{\theta s}} \left[\phi(u_{\theta s}) \exp\left(\sum_i \int_{u_{\delta i}} g_i(u_{\delta i}) \log p(y_{is}|u_{\theta s}, u_{\delta i}) du_{\delta i} \right) \right] du_{\theta s} \\
&= \int_{u_{\theta s}} \frac{\left[\phi(u_{\theta s}) \exp\left(\sum_i \int_{u_{\delta i}} g_i(u_{\delta i}) \log p(y_{is}|u_{\theta s}, u_{\delta i}) du_{\delta i} \right) \right] \phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})}{\phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})} du_{\theta s} \\
&\approx \int_{u_{\theta s}} \frac{\left[\phi(u_{\theta s}) \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s}, u_{\delta i} = l_{id}) w_{id} \right) \right] \phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})}{\phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})} du_{\theta s} \\
&\approx \sum_t \phi(l_{st}) \frac{w_t \sigma_{u_{\theta s}}}{\phi(l_r)} \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s} = l_{st}, u_{\delta i} = l_{id}) w_{id} \right) \\
&= \sum_t \phi(l_{st}) w_{st} \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s} = l_{st}, u_{\delta i} = l_{id}) w_{id} \right),
\end{aligned}
\tag{2.12}
$$

where

$$
\begin{aligned}
l_{st} &= \mu_{u_{\theta s}} + \sigma_{u_{\theta s}} l_t, \\
w_{st} &= \frac{\sigma_{u_{\theta s}} w_t}{\phi(l_r)},
\end{aligned}
$$

are the person-specific quadrature locations and the corresponding weights for integrating over $u_{\theta s}$, and $\mu_{u_{\theta s}}$ and $\sigma_{u_{\theta s}}$ are the posterior means and standard deviations for $u_{\theta s}$. Details on deriving (2.11) and (2.12) are provided in Appendix B.

Here the integrals over the person random effects $u_{\theta s}$ are evaluated using the same locations and weights as for evaluating the integral in the numerator of (2.9). The integrand in (2.12) is proportional to the variational distribution $g_s(u_{\theta s})$, which is the approximate

normal density $\phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})$. Therefore, the adaptive quadrature method is expected to work well.

The second M-step also changes by applying the adaptive quadrature method. That is, (2.10) becomes

$$\begin{aligned}
\frac{d\underline{l}}{d\beta} &= \frac{d}{d\beta}\int_{\boldsymbol{\theta},\boldsymbol{\delta}} \left[\sum_i \sum_s \log p(y_{is}|u_{\theta s}, u_{\delta i})\right] g(\boldsymbol{\theta},\boldsymbol{\delta}) d(\boldsymbol{\theta}) d(\boldsymbol{\delta}) \\
&= \sum_i \sum_s \int_{u_{\delta i}} \int_{u_{\theta s}} g_i(u_{\delta i}) g_s(u_{\theta s}) \frac{d}{d\beta} \log p(y_{is}|u_{\theta s}, u_{\delta i}) du_{\theta s} du_{\delta i} \\
&= \sum_i \sum_s \sum_d \sum_t g_i(l_{id}) g_s(l_{st}) \\
&\times \frac{d}{d\beta} \log p(y_{is}|u_{\theta s} = l_{st}, u_{\delta i} = l_{id}) w_{id} w_{st}.
\end{aligned}$$

The score functions for the the variance parameters τ_θ and τ_δ can be derived in a similar way.

The cluster-specific means and variances of $g_i(u_{\delta i})$ and $g_s(u_{\theta s})$ can be obtained by an iterative procedure. For example, for $g_s(u_{\theta s})$, first initialize $l_{st}^{(0)}$ and $w_{st}^{(0)}$ using starting values $\mu_{u_{\theta s}}^{(0)}$ and $\sigma_{u_{\theta s}}^{(0)}$. Then $\mu_{u_{\theta s}}$ and $\sigma_{u_{\theta s}}^2$ are at the kth iteration

$$\begin{aligned}
\mu_{u_{\theta s}}^{(k)} &= \int_{u_{\theta s}} u_{\theta s} g_s(u_{\theta s}) du_{\theta s} \\
&\approx \sum_t \frac{l_{st}^{(k-1)} g_s\left(l_{st}^{(k-1)}\right) w_{st}^{(k-1)}}{\phi\left(l_{st}^{(k-1)}\right)}, \\
{\sigma_{u_{\theta s}}^{(k)}}^2 &= \int_{u_{\theta s}} {u_{\theta s}}^2 g_s(u_{\theta s}) du_{\theta s} - {\mu_{u_{\theta s}}}^2 \\
&\approx \sum_t \left\{ \frac{l_{st}^{(k-1)} g_s\left(l_{st}^{(k-1)}\right) w_{st}^{(k-1)}}{\phi\left(l_{st}^{(k-1)}\right)} l_{st}^{(k-1)} \right\}^2 - {\mu_{u_{\theta s}}}^2,
\end{aligned} \tag{2.13}$$

where $l_{st}^{(k-1)}$, $l_{id}^{(k-1)}$ and $w_{st}^{(k-1)}$, $w_{id}^{(k-1)}$ are the cluster-specific quadrature locations and corresponding weights at the $(k-1)$th iteration. This sequence is repeated until convergence. The mean and variance for $g_i(u_{\delta i})$ can be derived similarly. Note that this is the method by Naylor & Smith (1982) and Rabe-Hesketh et al. (2005).

Alternatively, the mode and curvature at the mode can be used as in Pinheiro & Bates (1995) and Schilling & Bock (2005). In this case, an integration is not required.

Both methods of using the cluster-specific means and variances and using modes and

curvatures were implemented in the variational MM algorithm.

2.4.2 Discrete Priors

Assuming a normal density may not be optimal e.g., for non-normal or skewed normal latent variables. Without assuming a specific parametric form for the distribution, the non-parametric maximum likelihood estimator (NPMLE) of the distribution for the latent variables becomes a discrete distribution (de Leeuw & Verhelst, 1986; Lindsay et al., 1991; Heinen, 1996; Aitkin, 1999). To interpret the discrete distribution as the NPMLE, the number of masses must maximize the likelihood (Simar, 1976; Laird, 1978; Lindsay, 1983).

If discrete priors are used, the posteriors also have discrete distributions with the same support points as the priors and the variational approximation is discrete with masses as variational parameters.

The discrete distribution of the random effects is characterized by a finite set of locations and probabilities at these locations and the integrals in (2.8) and (2.9) become sums. For example, (2.8) becomes

$$v^u_{\theta_s} = \frac{w^u_{\theta_s} \exp\left(\sum_i \sum_d \log p(y_{is}|\theta_s, \delta_i = l^d_{\delta_i}) v^d_{\delta_i}\right)}{\sum_t w^t_{\theta_s} \exp\left(\sum_i \sum_d \log p(y_{is}|\theta_s = l^t_{\theta_s}, \delta_i = l^d_{\delta_i}) v^d_{\delta_i}\right)},$$

where $l^d_{\delta_i}$ $(d = 1, ..., D)$ and $l^t_{\theta_s}$ $(t = 1, ..., T)$ indicate locations for the discrete latent variable δ_i and θ_s, respectively. $v^d_{\delta_i} = g\left(\delta_i = l^d_{\delta_i}\right)$ and $v^t_{\theta} = g\left(\theta_s = l^t_{\theta_s}\right)$ are the masses of the variational approximation and $w^t_{\theta_s} = p\left(\theta_s = l^t_{\theta_s}\right)$ are the prior probabilities at the locations.

Note that with discrete priors, the prior locations and masses are model parameters and the posterior probabilities are variational parameters. The estimates of the variational parameters can be used to compute posterior moments of the random effects.

2.5 Related Issues

In this section, we discuss 1) the lower bound and marginal likelihood, 2) estimation of standard errors, 3) dependence structure of the random effects, and 4) prediction of the random effects.

2.5.1 Lower Bound and Marginal Likelihood

In the variational MM algorithm, the lower bound to the log-likelihood is maximized rather than the likelihood function itself. For valid inferences based on the lower bound, it should have the same shape as the log-likelihood, i.e., the same mode and curvature at the mode (Hall et al., 2002).

To evaluate the lower bound, we need to compute the marginal log-likelihood which is not feasible for GLMMs in general. Using a sampling method, however, we can approximate the marginal likelihood $L(\mathbf{y}; \mathbf{\Psi})$ as follows: First obtain posterior samples of the random effects using Markov chain Monte Carlo (MCMC) with the model parameters treated as fixed constants and set equal to the variational MM estimates. Then obtain the sample mean and covariance matrix of the posterior samples and use the corresponding multivariate normal distribution as importance density. Sample the random effects $\mathbf{z}$ from the importance density. Then the marginal likelihood can be approximated as

$$\begin{aligned} L(\mathbf{y}; \mathbf{\Psi}^*) &= \int \frac{p(\mathbf{y}|\mathbf{z}, \mathbf{\Psi}^*)p(\mathbf{z}; \mathbf{\Psi}^*)}{\tilde{g}(\mathbf{z}|\mathbf{y}, \mathbf{\Psi}^*)} \tilde{g}(\mathbf{z}|\mathbf{y}, \mathbf{\Psi}^*) d\mathbf{z} \\ &\approx \frac{1}{m} \sum_{j=1}^{m} \frac{p(\mathbf{y}|\mathbf{z}^{(j)}, \mathbf{\Psi}^*)p(\mathbf{z}^{(j)}; \mathbf{\Psi}^*)}{\tilde{g}(z^{(j)}|\mathbf{y}, \mathbf{\Psi}^*)}, \end{aligned}$$

where $p(\mathbf{y}|\mathbf{z}, \mathbf{\Psi}^*)$ is the joint probability of the responses given the latent variables $\mathbf{z}$, $p(\mathbf{z}; \mathbf{\Psi}^*)$ is the prior distribution given the parameter estimates $\mathbf{\Psi}^*$ from the variational MM method and $\tilde{g}(\mathbf{z}|\mathbf{y}, \mathbf{\Psi}^*)$ is a normal approximation to the posterior distribution (as the important density) that has the same support as the prior $p(\mathbf{z}; \mathbf{\Psi}^*)$. $\mathbf{z}^{(j)}$ $(j = 1, ..., m)$ is identically and independently drawn from $\tilde{g}(\mathbf{z}|\mathbf{y}, \mathbf{\Psi}^*)$.

By the strong law of large numbers, the importance approximation of the likelihood $\hat{L}(\mathbf{y}; \mathbf{\Psi}^*)$ is unbiased and consistent as $m \to \infty$, as long as the support of $\tilde{g}(\cdot)$ contains the support of $L(\cdot)$ (Geweke, 1989). A similar idea of using importance sampling has been adopted to evaluate a likelihood surface on which maximum likelihood estimation is carried out (Durbin & Koopman, 1997; Shephard & Pitt, 1997).

2.5.2 Standard Errors

As in the traditional EM algorithm, standard errors are not a by-product of the variational MM algorithm. In this section, we discuss two ways of approximating standard error estimates.

Hessian Matrix

A straightforward way of obtaining standard errors is to use the Hessian matrix. It can be directly obtained by solving the second derivatives of the lower bound, evaluated at the final estimates of the variational parameters with respect to the model parameters. Alternatively, the score functions in the second M-step (e.g., (2.10)) can be numerically differentiated with respect to the corresponding parameters.

We include only the model parameters in the Hessian matrix while treating the variational parameters as fixed. In the Gaussian variational approximation by Ormerod & Wand (2012),

the variational parameters (mean and variance parameters in their case) were all included in the Hessian matrix.

Bootstrap Standard Error

A bootstrap method can be used to estimate approximate standard errors (e.g., Efron, 1979). Data are simulated from the model given the parameter estimates for the real data. We denote the parameter estimates for the bth simulated dataset $\hat{\Psi}^*(b)$. The bootstrap standard error ($\widehat{\text{se}}_B$) can be computed as

$$\widehat{\text{se}}_B = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left[\hat{\Psi}^*(b) - \hat{\Psi}^*(\cdot)\right]^2},$$

where $\widehat{\Psi}^*(\cdot) = \sum_{b=1}^{B}\frac{1}{B}\hat{\Psi}^*(b)$ and B is the number of the bootstrap replicates. The Monte Carlo error (MCE) involved in the bootstrap standard error can be computed as described by Koehler et al. (2009). First define the bootstrap squared error as

$$d(b) = \left[\hat{\Psi}^*(b) - \hat{\Psi}^*(\cdot)\right]^2.$$

The squared bootstrap standard error can then be expressed as

$$\widehat{\text{se}}_B^2 = \frac{1}{B}\sum_{b=1}^{B} d(b).$$

An estimate of MCE for the squared bootstrap standard error ($\widehat{\text{se}}_B^2$) can be obtained as

$$\widehat{\text{MCE}}(\widehat{\text{se}}_B^2) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}[d(b) - d(\cdot)]^2},$$

where $d(\cdot) = \frac{1}{B}\sum_{b=1}^{B} d(b)$.

Finally, an MCE estimate for the bootstrap standard error can be obtained using the Delta method

$$\widehat{\text{MCE}}(\widehat{\text{se}}_B) = \left|\frac{1}{2\sqrt{\widehat{\text{se}}_B^2}}\right| \widehat{\text{MCE}}(\widehat{\text{se}}_B^2).$$

2.5.3 Dependence Structure of the Random Effects

The variational MM algorithm is based on the mean-field approximation that assumes posterior independence of the random effects. The performance of the algorithm may be affected by the degree to which the independence assumption is violated. In this section, we investigate the dependence among the random effects under the posterior as a function of the sample sizes and prior variances.

To derive analytical solutions, we assume a linear mixed model

$$\mathbf{y} = X\boldsymbol{\beta} + W\mathbf{z} + \boldsymbol{\epsilon},$$

where $\mathbf{z} \sim N(\mathbf{0}, \Psi)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Theta)$ with $\Theta = I\sigma^2$ and the identity matrix I.

The posterior covariance matrix can be computed for the linear mixed model as (Laird & Ware, 1982)

$$\text{Cov}(\mathbf{z}|\mathbf{y}, X) = \Psi - \Psi' W' \Sigma^{-1} W \Psi, \tag{2.14}$$

where $\Sigma = W\Psi W' + \Theta$ for model (2.1).

For the model with crossed random effects in (2.1), denote $\mathbf{z} = (\theta_1, ..., \theta_s, ..., \theta_N, \delta_1, ..., \delta_i, ..., \delta_I)'$ where θ_s and δ_i are the two crossed random effects with $s = 1, ..., N$ and $i = 1, ..., I$, respectively. The posterior covariance in (2.14) was computed as a function of the number of persons $N = (50, 100)'$ and the number of items $I = (10, 20, 40)'$, and the prior standard deviations $\tau_\theta = (0.5, 1.0, 1.5)'$ and $\tau_\delta = (0.2, 0.5, 1.0)'$. Figures 2.1 and 2.2 summarize the results.

The results suggest that for given prior variances ($\tau_\theta = 0.5, \tau_\delta = 0.2$), the correlations between θ_s increase as I increases. Similarly, the correlations between δ_i increase as N increases. For given sample sizes ($I = 10, N = 50$), the correlations between θ_s and between δ_i increase as the variance of θ_s or δ_i increases. This shows that either when the sample size or prior variance for θ_s or δ_i increases, the dependence among the random effects under the posterior increases.

2.5.4 Prediction of the Random Effects

Assigning values to (or prediction of) the random effects for individual clusters is useful for inference about particular clusters (Skrondal & Rabe-Hesketh, 2009), e.g., to assess the effectiveness of schools or hospitals (Raudenbush & Willms, 1995; Goldstein & Rasbash, 1996), in small area estimation or disease mapping (Rao, 2003), or for finding outlying clusters (Langford & Lewis, 1998). Prediction of abilities is also the main purpose of item response theory (IRT). For more information, see Skrondal & Rabe-Hesketh (2009).

Prediction of the random effects is a difficult problem for GLMMs because of the integral in the denominator of the posterior distribution. Here, we suggest using the variational approximation to the posterior, to derive posterior means (μ_{θ_s}, μ_{δ_i}) (expected a posteriori; EAP) or modes (maximum a posteriori; MAP). For instance, assuming normal priors with

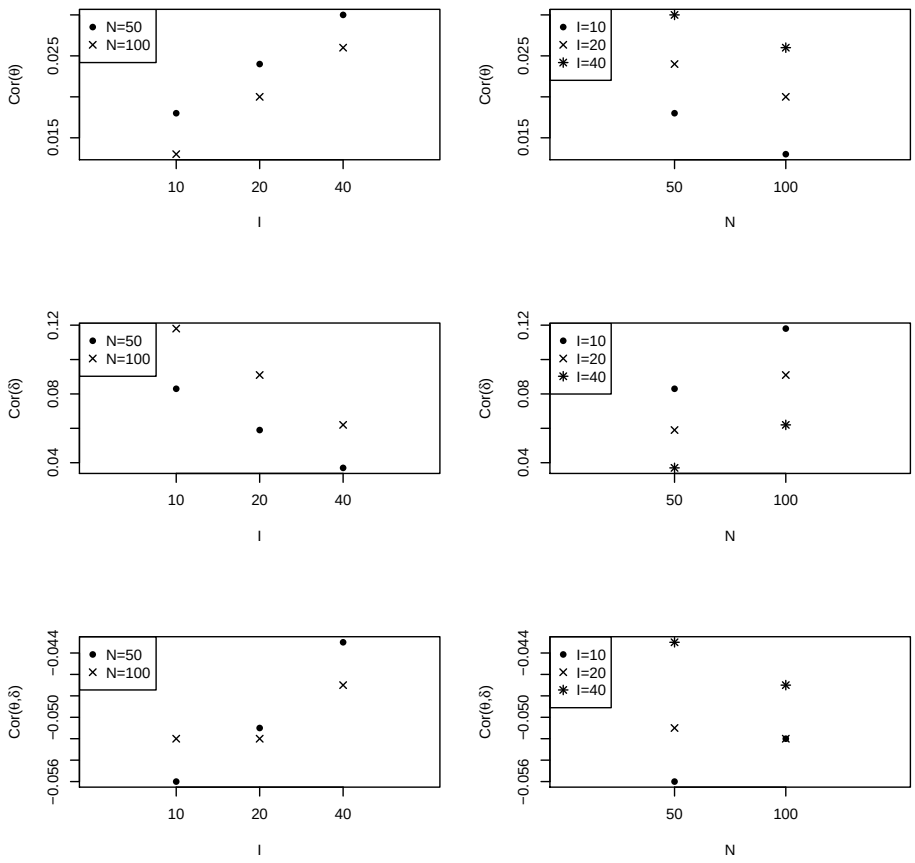

Figure 2.1: Posterior correlations among the random effects by the sample sizes (N=50,100, I=10,20,40) for given prior variances $\tau_\theta = 0.5$ and $\tau_\delta = 0.2$

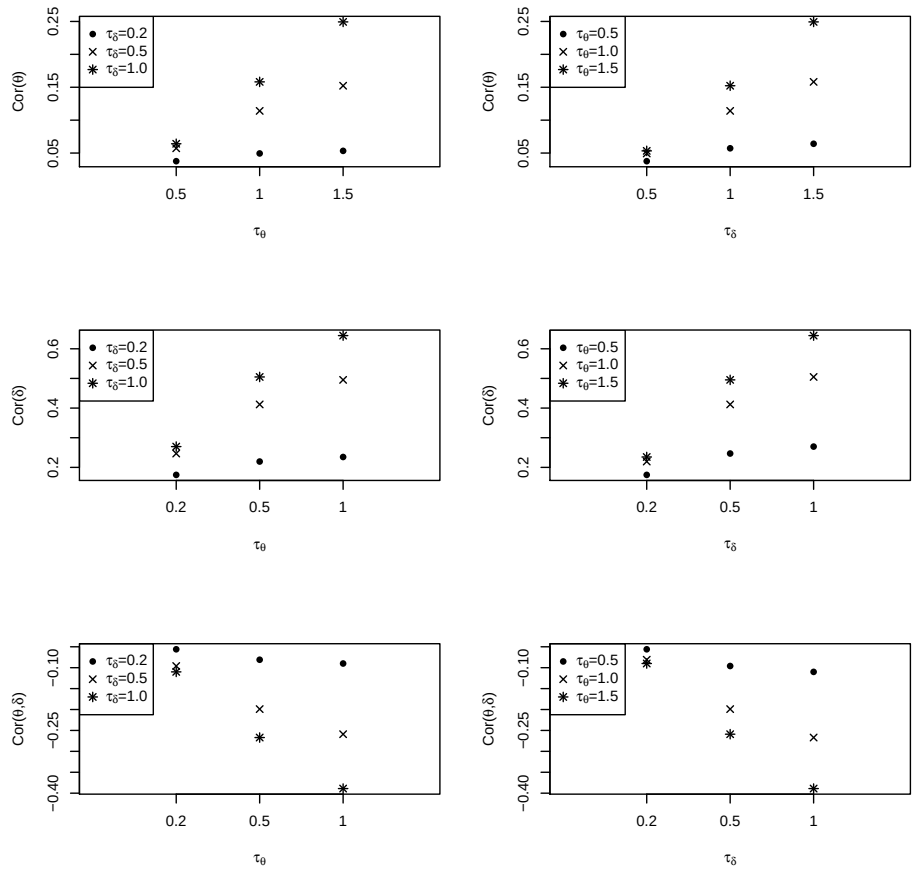

Figure 2.2: Posterior correlations among the random effects by the prior variances for given sample sizes $I = 10$ and $N = 50$

adaptive quadrature, the mean and standard deviation using (2.13) can be seen as the EAP and its standard error for θ_s. If the mode and curvature are used instead, the MAP and its standard error can be obtained.

2.6 Empirical Study

To illustrate the proposed algorithm, we use the salamander mating data (McCullagh & Nelder, 1989). This dataset is a benchmark that has been used to compare many different estimation methods for GLMMs with crossed random effects (e.g., Karim & Zeger, 1992; Breslow & Clayton, 1993; Booth & Hobert, 1999; Lee & Nelder, 2006; Cho & Rabe-Hesketh, 2011).

The dataset consists of three separate experiments, each involving matings among salamanders of two different populations, called Rough Butt (RB) and White Side (WS). Sixty females and sixty males of the two populations of salamander were paired by a crossed, blocked, and incomplete design in an experiment studying whether the two populations have developed generic mechanisms which would prevent inter-breeding. The response is binary, indicating whether the mating was successful between female i and male j. We adopted model A used in Karim & Zeger (1992)

$$\text{logit}(p(y_{ij} = 1 | z_i^f, z_j^m)) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2j} + \beta_3 x_{1i} x_{2j} + z_i^f + z_j^m, \tag{2.15}$$

where the covariates are dummy variables for White Side female (x_{1i}), White Side male (x_{2j}), and the interaction ($x_{1i}x_{2j}$). The two crossed random effects are random intercepts $z_i^f \sim N(0, \sigma_f^2)$ for females and $z_i^m \sim N(0, \sigma_m^2)$ for males. Each salamander participates in six matings, resulting in 360 matings in total.

Model (2.15) was fit to the dataset using the variational MM algorithm with adaptive quadrature (10 quadrature points). In order to check the independence assumption of the mean-field approximation for the data, we examined the posterior correlations among the female and male random effects. The posterior samples of the random effects were obtained by MCMC using WinBUGS 1.4 (Spiegelhalter et al., 2003) with model parameter fixed to the estimates of the variational MM algorithm. The posterior correlations among the random effects (among females, among males, and between females and males) appeared negligible, all being close to zero.

We compared the parameter and standard error estimates from the MM algorithm with those from the Laplace approximation implemented using `lmer` in the R package `lme4` (Bates & Maechler, 2009). For standard errors, the Hessian matrix was obtained by numerically differentiating the score functions. We also computed the bootstrap standard errors (based on 100 replicates) as well as the Monte Carlo errors computed as described in Section 2.5.2. In addition, we report the estimates from PQL (Breslow & Clayton, 1993) and MCEM (Booth & Hobert, 1999) from the literature. Table 2.1 lists the results.

Table 2.1: Comparison of several estimators for the salamander mating data. Standard errors are given in parentheses if reported. Laplace: `lmer`; PQL: Breslow & Clayton (1993); MCEM: Booth & Hobert (1999). For the variational MM algorithm, bootstrap standard errors (Boot.SE) and Monte Carlo errors (MCE) are reported.

Method	β_0	β_1	β_2	β_3	σ_m	σ_f
Variational MM	0.97	-2.84	-0.67	3.49	1.07	1.00
	(0.39)	(0.55)	(0.45)	(0.62)	-	-
(Boot.SE)	(0.35)	(0.46)	(0.37)	(0.59)	-	-
(MCE)	(0.02)	(0.03)	(0.04)	(0.05)	-	-
Laplace	1.00	-2.90	-0.70	3.59	1.08	1.02
	(0.39)	(0.56)	(0.46)	(0.64)	-	-
PQL	0.79	-2.29	-0.54	2.82	0.79	0.72
	(0.32)	(0.43)	(0.39)	(0.50)	-	-
MCEM	1.02	-2.96	-0.69	3.63	1.18	1.12

We do not report standard errors for the variance parameters because the use of standard errors for Wald-type tests and confidence intervals may be inappropriate for these parameters (e.g., Berkhof & Snijders, 2001). The parameter estimates for the variational MM method are close to those from the Laplace approximation and MCEM. Our standard error estimates are close to the standard errors from the Laplace approximation. The bootstrap standard errors are slightly smaller than the standard errors from the variational MM and the Laplace approximation. The difference is less than 2 MCEs.

To assess the lower bound of the log-likelihood, we compared the lower bound with the approximate marginal log-likelihood obtained using 1) importance sampling described in Section 2.5.1, and 2) adaptive quadrature (with three quadrature points) with `gllamm` (Rabe-Hesketh et al., 2005). For simplicity, the log-likelihood was plotted for each parameter with the other parameters fixed to the estimates from the variational MM algorithm. Figure 2.3 presents the results.

In the figure, circles represent the marginal log-likelihood obtained using importance sampling, triangles the lower-bound, and "x" the log-likelihood from adaptive quadrature. The dashed vertical lines indicate the parameter estimates obtained from the variational MM algorithm. For all parameters, the lower-bounds show shapes similar to the marginal log-likelihoods. The approximate marginal log-likelihood using importance sampling is very close to that from adaptive quadrature.

Finally, predictions of the random effects obtained from the variational MM algorithm were compared with 1) the MAP from the Laplace approximation and 2) the EAP from MCMC. The EAP from MCMC was obtained as the mean of the posterior samples of the random effects with the parameters fixed to the estimates. Figure 2.4 shows the results.

The sub-panels compare the EAP and MAP estimates from the variational MM algorithm with the Laplace approximation (MAP) and the MCMC method (EAP) for females (first row) and males (second row). The 45 degree line indicates that the two methods produce equivalent results. The results show that the variational MM algorithm provides the predictions close to those from the Laplace approximation and MCMC methods.

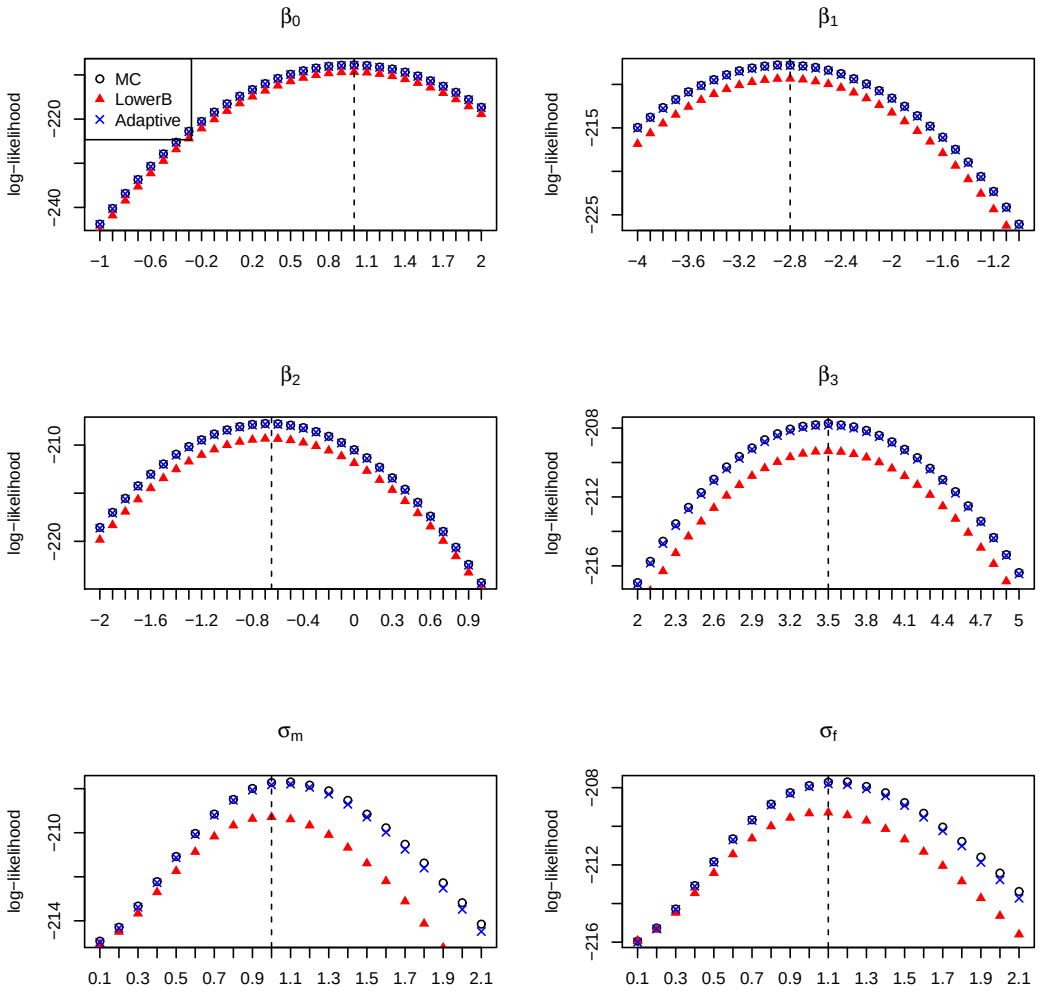

Figure 2.3: Log-likelihoods and lower bounds as a function of each parameter for the salamander mating model with other parameters set equal to estimates from the variational MM algorithm. MC is the marginal log-likelihood from the importance sampling method, LowerB is the lower bound from the variational MM algorithm, and Adaptive is the marginal log-likelihood from adaptive quadrature (3 quadrature points).

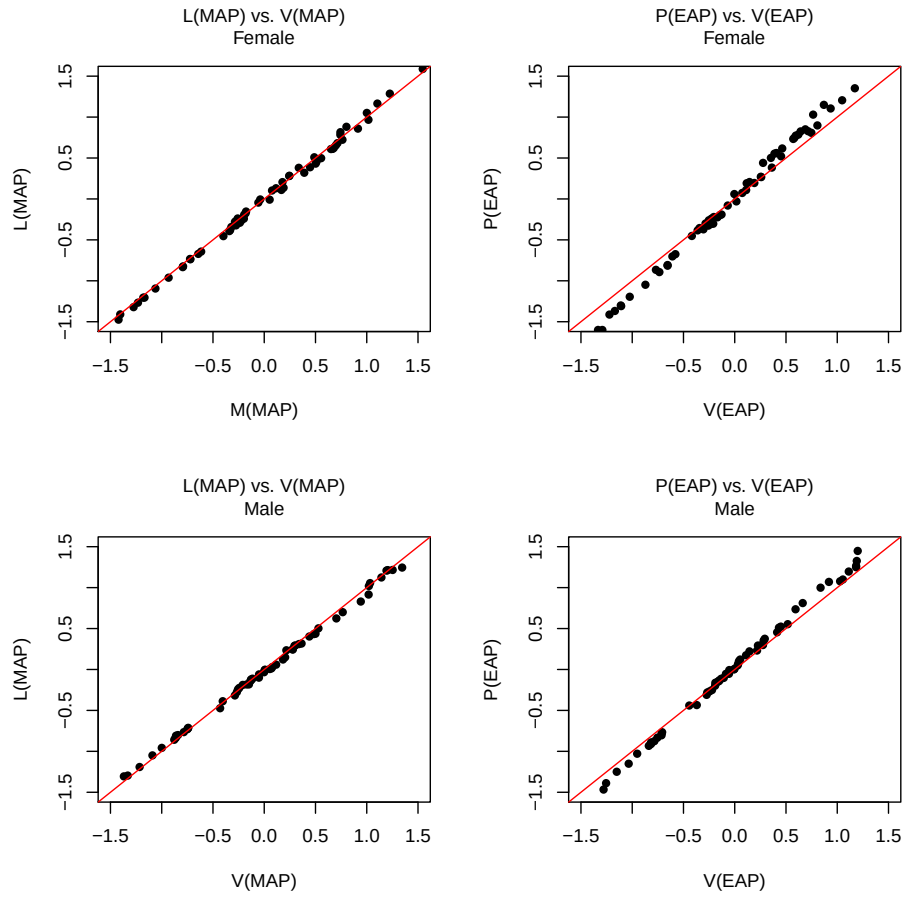

Figure 2.4: Comparison of predictions (EAP and MAP) from the variational MM algorithm with the Laplace approximation (MAP) and MCMC (the posterior mean; EAP) methods for female and male random effects. L(MAP): Laplace MAP, P(EAP): Bayesian EAP, V(MAP): variational MAP, V(EAP): variational EAP.

2.7 Simulation Studies

Simulation studies were carried out to evaluate the performance of the variational MM algorithm (with adaptive quadrature, 10 quadrature points) and to compare it with the Laplace approximation. Two examples were considered using 1) the crossed random effects model for the salamander mating data and 2) the random item Rasch model.

2.7.1 Crossed Random Effects Model for Salamander Mating Data

The first simulation study is closely related to the model for the salamander mating data used in the empirical study. We simulated 50 datasets based on model (2.15) using the true values that have been used by other researchers (e.g., Lin & Breslow, 1996), $\boldsymbol{\beta} = (1.06, -0.72, -3.05, 3.77)'$ and $(\sigma_f^2, \sigma_m^2)' = (.50, .50)'$. We also generated datasets that are ten times as large in terms of the total sample size as the original dataset (called large datasets from now on).

Figure 2.5 shows the estimated bias and root mean squared error (RMSE) for the parameter estimates from the variational MM algorithm and the Laplace approximation.

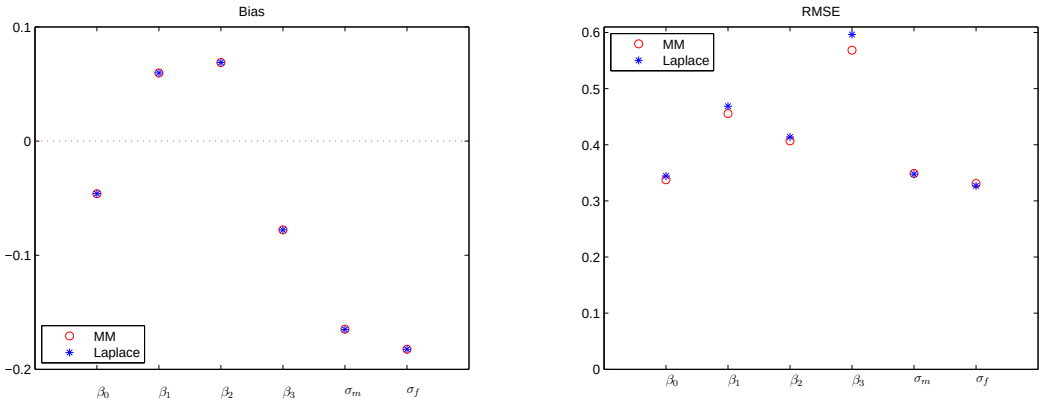

Figure 2.5: Bias and RMSE for the salamander simulation. MM is the variational MM algorithm and Laplace is the Laplace approximation.

In terms of bias, there are negligible differences between the two methods. In terms of RMSE, the variational MM algorithm tends to show somewhat smaller RMSE for the fixed effects parameters than the Laplace approximation. A similar pattern is observed for the large datasets in Figure 2.6.

For the large datasets, there is little difference in the estimated bias between the two methods. The RMSE is still smaller for the variational MM algorithm for the fixed effects parameters than the Laplace approximation, but the differences are somewhat smaller than

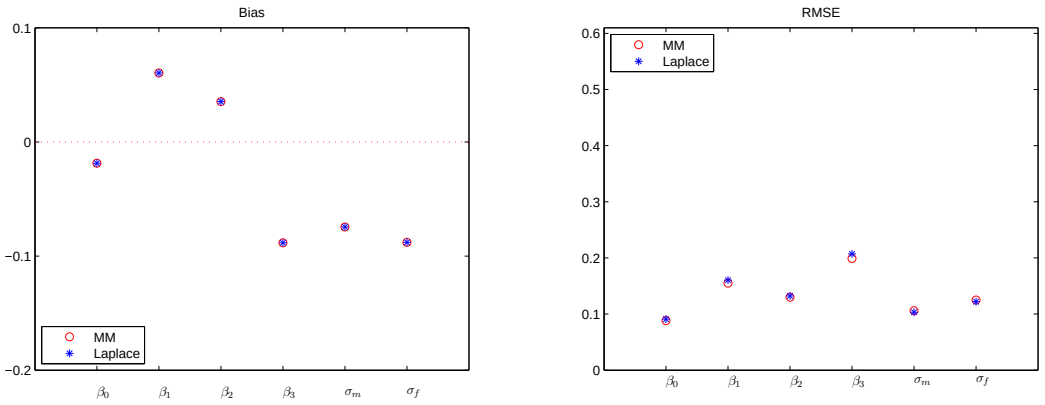

Figure 2.6: Bias and RMSE for the salamander simulation for large datasets. MM is the variational MM algorithm and Laplace is the Laplace approximation.

those in the smaller datasets. This result makes sense given that the Laplace approximation produces less bias for data with large cluster sizes (Joe, 2008).

2.7.2 Random Item Rasch Model

The second simulation study uses the random item Rasch model described in Section 2.2. Small and large datasets were generated based on model (2.1) under various conditions. For small datasets, we generated data with $N = (50, 100)'$ persons and $I = (20, 50)'$ items, and with standard deviations $\tau_\theta = 0.5$ for person abilities and $\tau_\delta = (0.2, 0.6, 1.2, 1.5)'$ for item difficulties. The intercept β was set to 0. For large datasets, we considered $N = (200, 300)'$ and $I = (20, 50)'$ for the sample sizes, and $\tau_\theta = (0.2, 0.5)'$, and $\tau_\delta = (0.2, 0.6)'$ for the standard deviations. 50 replicates were simulated for each condition.

Figures 2.7 to 2.10 present the estimated bias and RMSE for the model parameters for the intercept $\hat{\beta}$ and the person and item standard deviations $\hat{\tau}_\theta$ and $\hat{\tau}_\delta$.

Each figure corresponds to four item difficulty standard deviations $\tau_\delta = (0.2, 0.6, 1.2, 1.5)'$. In each figure, the first row presents the estimated bias and the second row the estimated RMSE. In each sub-panel, the solid line is for the variational MM algorithm and the dotted line for the Laplace approximation. The x-axis represents the sample sizes N and I (four combinations by $N1 = 50$, $N2 = 100$ and $I1 = 20$, $I2 = 50$).

In condition 1 ($\tau_\theta = 0.5$, $\tau_\delta = 0.2$) in Figure 2.7, the estimated bias and RMSE tend to decrease as the sample size increases for either person N or item I. Between the methods, the Laplace approximation tends to show larger bias for $\hat{\tau}_\theta$ and $\hat{\tau}_\delta$ than the variational MM algorithm, in particular with $N = 50$. In terms of RMSE, the Laplace approximation is larger than the variational MM algorithm across all sample sizes.

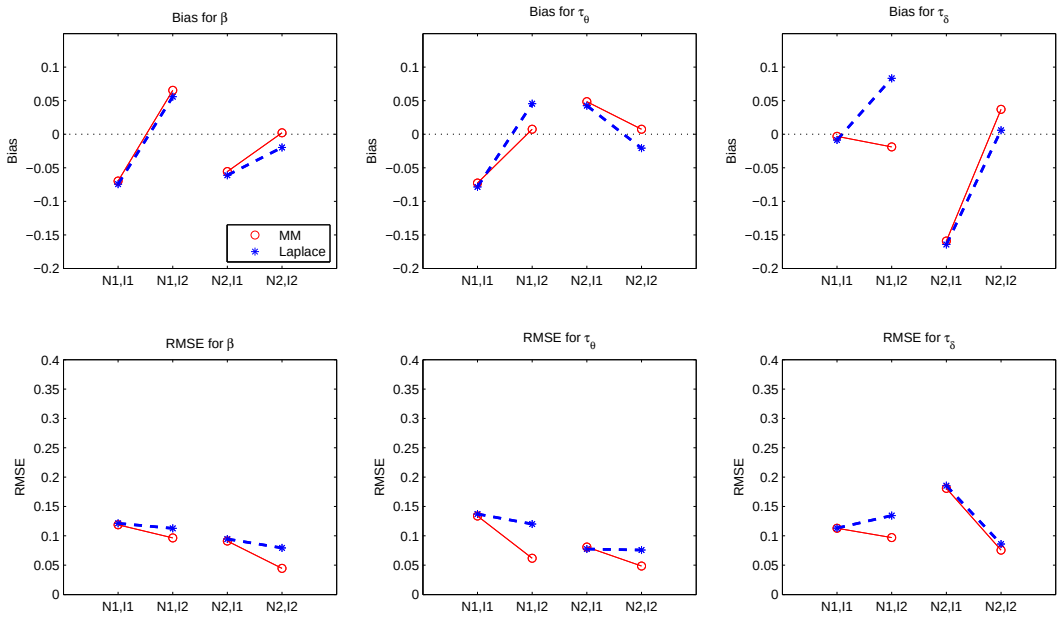

Figure 2.7: Bias and RMSE for the random item Rasch model simulation for small datasets in condition 1 ($\tau_\theta = 0.5$, $\tau_\delta = 0.2$). $N = (50, 100)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

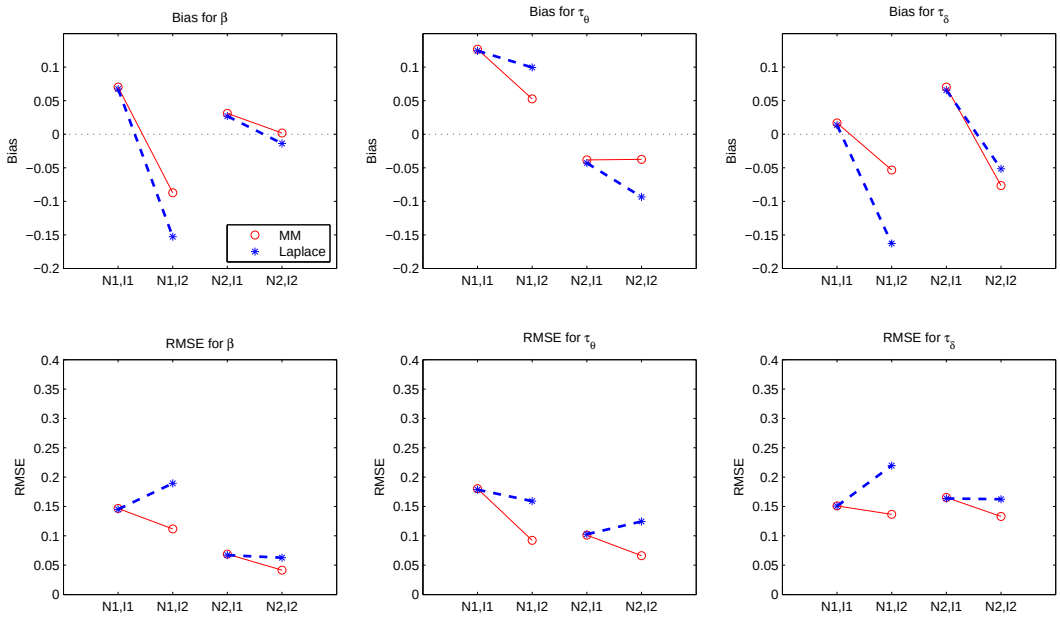

Figure 2.8: Bias and RMSE for the random item Rasch model simulation for small datasets in condition 2 ($\tau_\theta = 0.5$, $\tau_\delta = 0.6$). $N = (50, 100)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

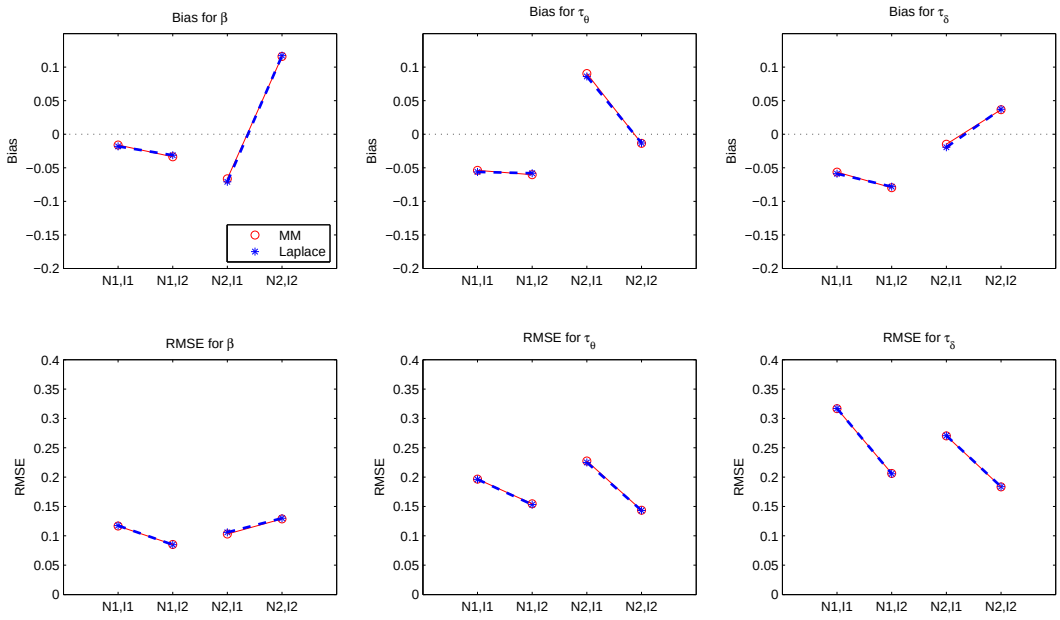

Figure 2.9: Bias and RMSE for the random item Rasch model simulation for small datasets in condition 3 ($\tau_\theta = 0.5$, $\tau_\delta = 1.2$). $N = (50, 100)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

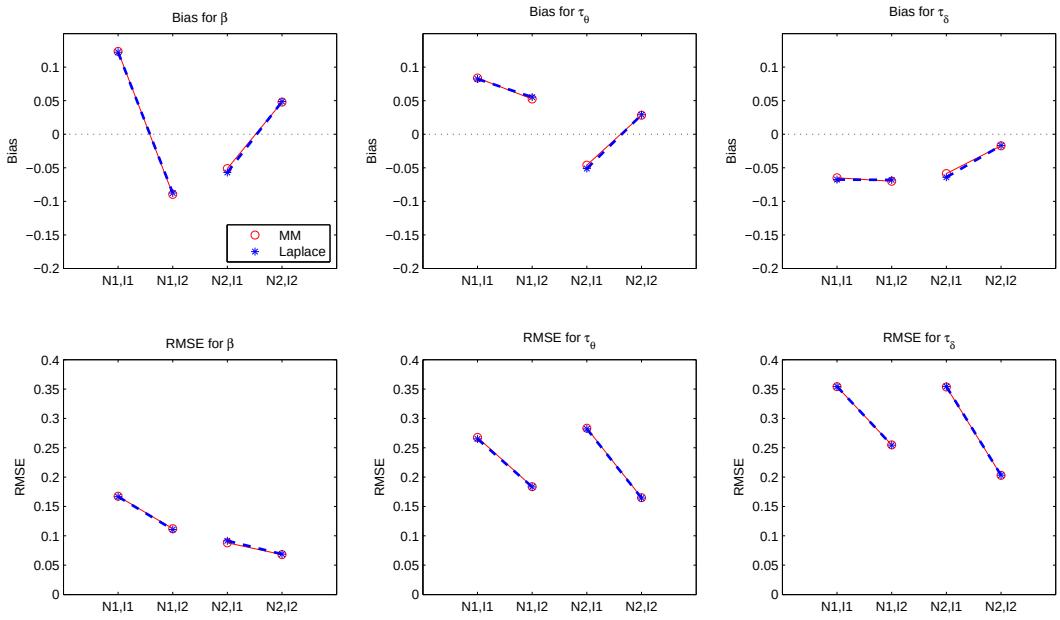

Figure 2.10: Bias and RMSE for the random item Rasch model simulation for small datasets in condition 4 ($\tau_\theta = 0.5$, $\tau_\delta = 1.5$). $N = (50, 100)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

In condition 2 ($\tau_\theta = 0.5$, $\tau_\delta = 0.6$) in Figure 2.8, a similar pattern is observed except that with $I = 50$, the Laplace approximation tends to show larger bias and RMSE than the variational MM algorithm. For $\hat{\beta}$ and $\hat{\tau}_\delta$, the estimated bias and RMSE increase for the Laplace approximation as the item sample size increases, from the conditions (N1,I1) to (N1,I2).

In condition 3 ($\tau_\theta = 0.5$, $\tau_\delta = 1.2$) in Figure 2.9 and condition 4 ($\tau_\theta = 0.5$, $\tau_\delta = 1.5$) in Figure 2.10, similar patterns are observed in general. As the sample sizes increase, the estimated bias and RMSE tend to decrease. However, the differences between the methods are smaller with large standard deviations τ_θ and τ_δ than in conditions 1 and 2.

Figures 2.11 to 2.14 present results for the large datasets with $N = (200, 300)'$ and $I = (20, 50)'$.

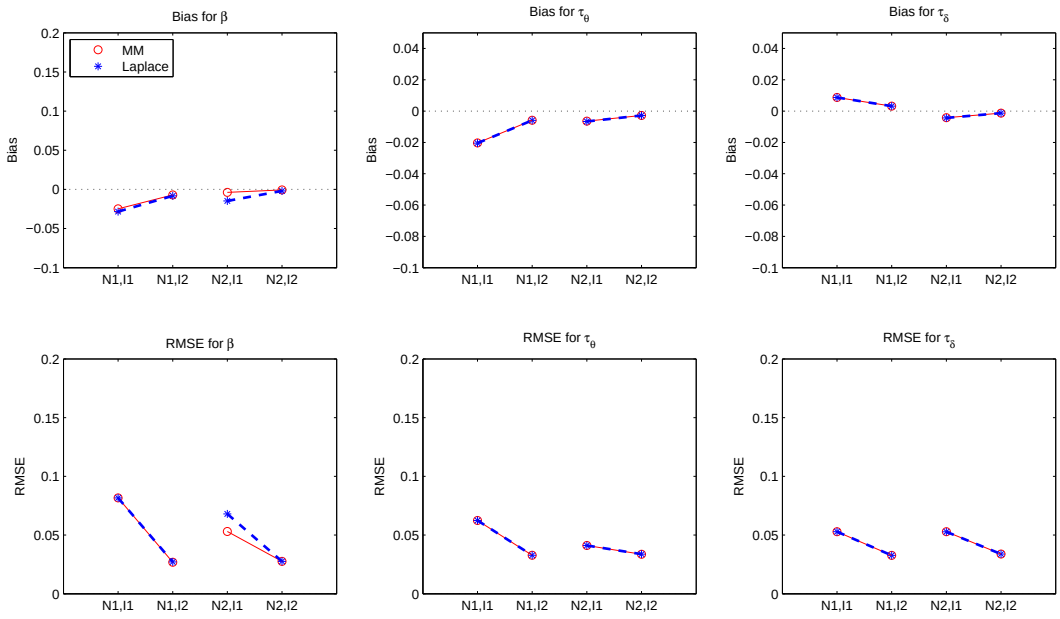

Figure 2.11: Bias and RMSE for the random item Rasch model simulation for large datasets in condition 1 ($\tau_\theta = 0.2$, $\tau_\delta = 0.2$). $N = (200, 300)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

Each figure corresponds to four conditions according to $\tau_\theta = (0.2, 0.5)'$, and $\tau_\delta = (0.2, 0.6)'$. The x-axis represents the four combinations by the sample sizes N and I. In condition 1 ($\tau_\theta = 0.2$, $\tau_\delta = 0.2$) in Figure 2.11, the overall pattern is the same as in the small datasets. The estimated bias and RMSE tend to decrease as the sample size increases for either N or I. The estimated bias and RMSE are quite similar between the methods, except for $\hat{\beta}$, the

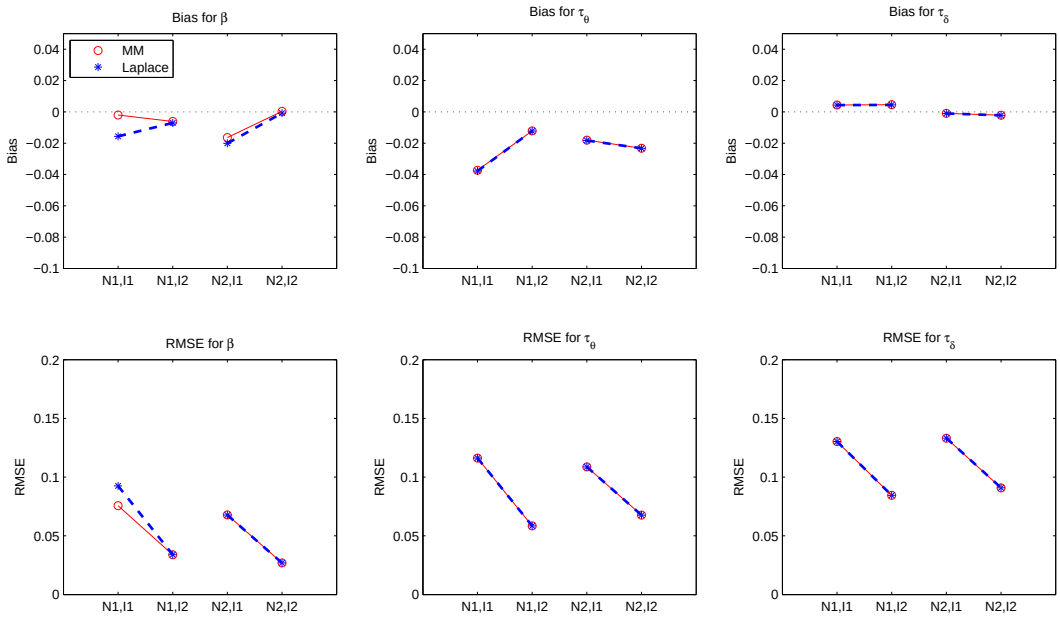

Figure 2.12: Bias and RMSE for the random item Rasch model simulation for large datasets in condition 2 ($\tau_\theta = 0.2$, $\tau_\delta = 0.6$). $N = (200, 300)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

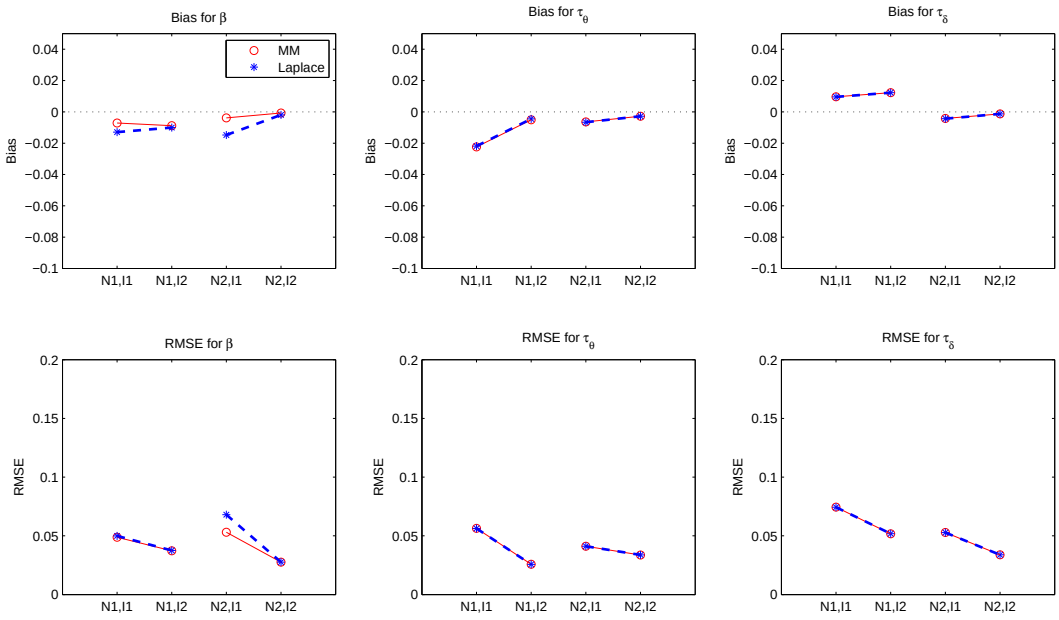

Figure 2.13: Bias and RMSE for the random item Rasch model simulation for large datasets in condition 3 ($\tau_\theta = 0.5$, $\tau_\delta = 0.2$). $N = (200, 300)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

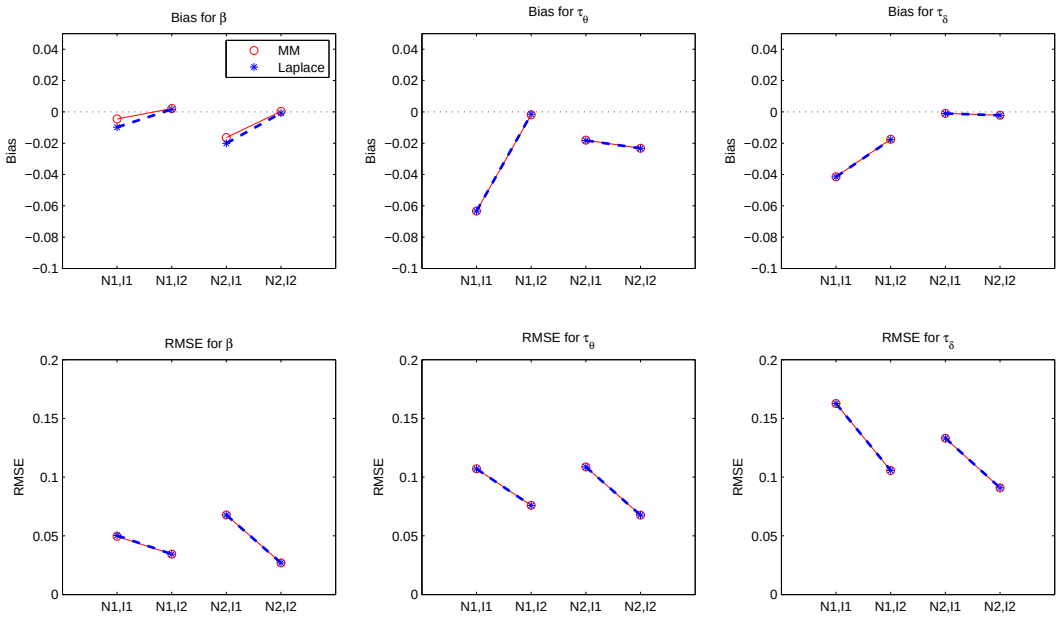

Figure 2.14: Bias and RMSE for the random item Rasch model simulation for large datasets in condition 4 ($\tau_\theta = 0.5$, $\tau_\delta = 0.6$). $N = (200, 300)'$ and $I = (20, 50)'$. The solid line is for the variational MM algorithm and the dotted line for the Laplace approximation.

Laplace approximation shows greater estimated bias and RMSE, particularly with N=300 and I=20. In condition 2 ($\tau_\theta = 0.2$, $\tau_\delta = 0.6$) in Figure 2.12, condition 3 ($\tau_\theta = 0.5$, $\tau_\delta = 0.2$) in Figure 2.13, and condition 4 ($\tau_\theta = 0.5$, $\tau_\delta = 0.6$) in Figure 2.14, similar results are observed. For $\hat{\beta}$, the Laplace approximation shows somewhat larger estimated bias and RMSE with N=200 and I=20, and with N=300 and I=20.

2.8 Concluding Remarks

Variational approximations have been mostly used for Bayesian inference in machine learning. Recently, Gaussian variational approximations (Opper, 2009; Ormerod & Wand, 2012) have been proposed for ML estimation of GLMMs. Hall et al. (2011) investigated theoretical properties of the Gaussian variational approximation, deriving asymptotic normality of the estimators and establishing root-m consistency of the estimates under relatively mild assumptions. However, this work was restricted to models with nested random effects.

In this paper, we proposed a variational MM algorithm for ML inference of GLMMs with crossed random effects. The variational approximation comes into play in approximating the posterior distribution of the random effects to make the integrals tractable. Accordingly, the algorithm involves finding a variational density function. The E-step is replaced by another M-step, minimizing the KL distance between the variational distribution and the true posterior distribution. This new M-step is equivalent to maximizing the lower bound to the log-likelihood with respect to the variational density function.

Our variational MM algorithm is more general and flexible than the Gaussian variational approximation because our algorithm does not require a pre-specified functional form for the variational distribution. The general form for the variational density function is derived so that different types of priors for the random effects can be handled. Importantly, we can estimate models with crossed random effects based on the mean-field approximation that assumes conditionally independent latent variables given the data. We found that with reasonable sample sizes and prior variances, the posterior correlations between the random effects are negligible. In addition, the lower bound was quite close to the marginal log-likelihood in the examples that we considered in this paper.

Several simulation examples were provided to evaluate the performance of the variational MM algorithm and compare it with the Laplace approximation for GLMMs with crossed random effects. The results show that overall, the variational MM algorithm performs as well as the Laplace approximation. With small cluster sizes, however, our algorithm performs better than the Laplace approximation especially for the variance parameters. Therefore, the variational MM method could be an effective alternative to the Laplace approximation.

Appendix A

Here the functional derivative is illustrated for the first M-step of the variational MM algorithm in Section 2.4. A fixed item (or regular) Rasch model is used for illustration. The model can be formulated as

$$\text{logit}\left(p\left(y_{is}=1|\theta_s\right)\right)=\beta_i+\theta_s,$$

where y_{is} denotes the binary response for person s to item i with $i=1,...,I$ and $s=1,...,N$. β_i is the item easiness parameter for item i and θ_s for person s is the person ability with a normal distribution $\theta_s \sim N(0,\tau_\theta^2)$.

The marginal probability for the response vector $\mathbf{y}_s$ for person s can be written as

$$p(\mathbf{y}_s)=\int_{\theta_s} p(\mathbf{y}_s|\theta_s)\phi(\theta_s;0,\tau_\theta)d\theta_s,$$

where $\phi(\cdot;\mu,\sigma)$ denotes the normal density with mean μ and standard deviation σ. The marginal log-likelihood function for all persons can be written as

$$L(\mathbf{y})=\sum_s \log \int_{\theta_s} p(\mathbf{y}_s|\theta_s)\phi(\theta_s;0,\tau_\theta)d\theta_s.$$

The lower bound to the log-likelihood now can be derived as

$$\begin{aligned}
\underline{l} &= \int_{\theta_s} \log\left[p(\mathbf{y}_s|\theta_s)\phi(\theta_s;0,\tau_\theta)\right] g_s(\theta_s)d\theta_s - \int_{\theta_s} \log\left[g_s(\theta_s)\right] g_s(\theta_s)d\theta_s \\
&= \int_{\boldsymbol{\theta}} \left[\sum_s \log p(\theta_s) + \sum_s \log p(\mathbf{y}_s|\theta_s) - \sum_s \log g_s(\theta_s)\right] g_s(\boldsymbol{\theta})d(\boldsymbol{\theta}) \\
&= \sum_s \int_{\theta_s} g_s(\theta_s)\log p(\theta_s)d\theta_s + \sum_s \int_{\theta_s} g_s(\theta_s)\log p(\mathbf{y}_s|\theta_s)d\theta_s - \sum_s \int_{\theta_s} g_s(\theta_s)\log g_s(\theta_s)d\theta_s,
\end{aligned}$$

where $g_s(\theta_s)$ is the variational distribution for θ_s. Note that the functional form for $g_s(\theta_s)$ is not required here.

To apply the functional derivative, we need to define a functional. The functional is obtained here by rewriting the lower bound and adding the constraint that variational dis-

tribution integrates to 1

$$F = \sum_s \int_{\theta_s} g_s(\theta_s) \log p(\theta_s) d\theta_s + \sum_s \int_{\theta_s} g_s(\theta_s) \log p(\mathbf{y}_s|\theta_s) d\theta_s - \sum_s \int_{\theta_s} g_s(\theta_s) \log g_s(\theta_s) d\theta_s$$
$$+ \sum_s \lambda_s \left[\int_{\theta_s} g_s(\theta_s) d\theta_s - 1 \right],$$

where λ is the Lagrange multiplier for the constraint, $\int_{\theta_s} g_s(\theta_s) d\theta_s = 1$.

Now perform the functional derivative of the functional F with respect to the variational density function $g_s(\theta_s)$. Note that as in this case when the functional is defined by integrals whose integrands take the form of $F(g_s(\theta_s))$ and does not depend on the derivatives of $g_s(\theta_s)$, stationarity simply requires $\frac{\partial F}{\partial g_s(\theta_s)} = 0$ for all values of θ_s (Bishop, 2006, p.705).

This implies

$$\log \left[p(\mathbf{y}_s|\theta_s) \phi(\theta_s; 0, \tau_\theta) \right] - \log \left[g_s(\theta_s) \right] - 1 + \lambda = 0.$$

Then we obtain

$$g_s(\theta_s) = p(\mathbf{y}_s|\theta_s) \phi(\theta_s; 0, \tau_\theta) \exp(-1 + \lambda). \tag{2.16}$$

By integrating (2.16) over θ_s, we obtain

$$1 = \int_{\theta_s} \exp(-1 + \lambda) p(\mathbf{y}_s|\theta_s) \phi(\theta_s; 0, \tau_\theta) d\theta_s,$$
$$\lambda = 1 - \log \int_{\theta_s} p(\mathbf{y}_s|\theta_s) \phi(\theta_s; 0, \tau_\theta) d\theta_s$$
$$= 1 - \log p(\mathbf{y}_s).$$

By substituting λ back to (2.16), we obtain the general solution for $g_s(\theta_s)$

$$g_s(\theta_s) = p(\mathbf{y}_s|\theta_s) \phi(\theta_s; 0, \tau_\theta) \exp(-1 + 1 - \log p(\mathbf{y}_s))$$
$$= \frac{p(\mathbf{y}_s|\theta_s) \phi(\theta_s; 0, \tau_\theta)}{p(\mathbf{y}_s)}$$
$$= p(\theta_s|\mathbf{y}_s).$$

It shows that for the ordinary Rasch model, the optimal solution for the variational density function $g_s(\theta_s)$ is the same as the true posterior density $p(\theta_s|\mathbf{y}_s)$.

Appendix B

Here we provide details on how to approximate the integrals in (2.8) and (2.9) using adaptive quadrature in Section 2.4.1. To approximate the numerator of (2.8) using adaptive quadrature, consider the second line in (2.11)

$$\int_{u_{\delta i}} \frac{g_i(u_{\delta i})\text{log}p(y_{is}|u_{\theta s}, u_{\delta i})}{\phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}})} \phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}}) du_{\delta i}.$$

To change the variable of integration to $a_i \sim N(0,1)$, we need the following changes:

$$\begin{aligned}
u_{\delta i} &= \mu_{u_{\delta i}} + \sigma_{u_{\delta i}} a_i, \\
a_i &= \frac{u_{\delta i} - \mu_{u_{\delta i}}}{\sigma_{u_{\delta i}}}, \\
du_{\delta i} &= \sigma_{u_{\delta i}} da_i, \\
\phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}}) &= \frac{1}{\sigma_{u_{\delta i}}\sqrt{2\pi}} e^{-\frac{(u_{\delta i} - \mu_{u_{\delta i}})^2}{2\sigma^2_{u_{\delta i}}}} \\
&= \frac{1}{\sigma_{u_{\delta i}}\sqrt{2\pi}} e^{-\frac{a_i^2}{2}} \\
&= \frac{1}{\sigma_{u_{\delta i}}} \phi(a_i).
\end{aligned}$$

Plug them all in (2.11) and obtain

$$\begin{aligned}
&\int_{u_{\delta i}} \frac{g_i(u_{\delta i})\text{log}p(y_{is}|u_{\theta s}, u_{\delta i})}{\phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}})} \phi(u_{\delta i}; \mu_{u_{\delta i}}, \sigma_{u_{\delta i}}) du_{\delta i} \\
&= \int_{a_i} \frac{g_i(\mu_{u_{\delta i}} + \sigma_{u_{\delta i}} a_i)\text{log}p(y_{is}|u_{\theta s}, u_{\delta i} = \mu_{u_{\delta i}} + \sigma_{u_{\delta i}} a_i)}{\frac{1}{\sigma_{u_{\delta i}}}\phi(a_i)} \frac{1}{\sigma_{u_{\delta i}}} \phi(a_i) \sigma_{u_{\delta i}} da_i \\
&= \int_{a_i} \frac{g_i(\mu_{u_{\delta i}} + \sigma_{u_{\delta i}} a_i)\text{log}p(y_{is}|u_{\theta s}, u_{\delta i} = \mu_{u_{\delta i}} + \sigma_{u_{\delta i}} a_i)}{\phi(a_i)} \phi(a_i) \sigma_{u_{\delta i}} da_i \\
&\approx \sum_d \frac{g_i(l_{id})\text{log}p(y_{is}|u_{\theta s}, u_{\delta i} = l_{id})}{\phi(l_d)} \sigma_{u_{\delta i}} w_{id} \\
&= \sum_d g_i(l_{id})\text{log}p(y_{is}|u_{\theta s}, u_{\delta i} = l_{id}) w_{id},
\end{aligned}$$

where

$$l_{id} = \mu_{u_{\delta i}} + \sigma_{u_{\delta i}} l_d,$$
$$w_{id} = \frac{\sigma_{u_{\delta i}} w_d}{\phi(l_d)},$$

are the item-specific quadrature locations and weights for integrating over $u_{\delta i}$, and $\mu_{u_{\delta i}}$ and $\sigma_{u_{\delta i}}$ are the posterior means and standard deviations for $u_{\delta i}$.

Similarly, to approximate the denominator of (2.8) using adaptive quadrature, consider the third line in (2.12)

$$\int_{u_{\theta s}} \frac{\left[\phi(u_{\theta s}) \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is} | u_{\theta s}, u_{\delta i} = l_{id}) w_{id}\right)\right] \phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})}{\phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})} du_{\theta s}.$$

To change the variable of integration to $a_s \sim N(0,1)$, we need the following changes:

$$\begin{aligned}
u_{\theta s} &= \mu_{u_{\theta s}} + \sigma_{u_{\theta s}} a_s, \\
a_s &= \frac{u_{\theta s} - \mu_{u_{\theta s}}}{\sigma_{u_{\theta s}}}, \\
du_{\theta s} &= \sigma_{u_{\theta s}} da_s, \\
\phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}}) &= \frac{1}{\sigma_{u_{\theta s}} \sqrt{2\pi}} e^{-\frac{(u_{\theta s} - \mu_{u_{\theta s}})^2}{2\sigma^2_{u_{\theta s}}}} \\
&= \frac{1}{\sigma_{u_{\theta s}} \sqrt{2\pi}} e^{-\frac{a_s^2}{2}} \\
&= \frac{1}{\sigma_{u_{\theta s}}} \phi(a_s).
\end{aligned}$$

Plug them all in (2.12) and obtain

$$
\begin{aligned}
&\int_{u_{\theta s}} \frac{\left[\phi(u_{\theta s}) \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s}, u_{\delta i} = l_{id}) w_{id}\right)\right] \phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})}{\phi(u_{\theta s}; \mu_{u_{\theta s}}, \sigma_{u_{\theta s}})} du_{\theta s} \\
&= \int_{a_s} \frac{\left[\phi(\mu_{u_{\theta s}} + \sigma_{u_{\theta s}} a_s) \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s} = \mu_{u_{\theta s}} + \sigma_{u_{\theta s}} a_s, u_{\delta i} = l_{id}) w_{id}\right)\right] \frac{1}{\sigma_{u_{\theta s}}} \phi(a_s)}{\frac{1}{\sigma_{u_{\theta s}}} \phi(a_s)} \sigma_{u_{\theta s}} da_s \\
&= \int_{a_s} \frac{\left[\phi(\mu_{u_{\theta s}} + \sigma_{u_{\theta s}} a_s) \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s} = \mu_{u_{\theta s}} + \sigma_{u_{\theta s}} a_s, u_{\delta i} = l_{id}) w_{id}\right)\right] \phi(a_s)}{\phi(a_s)} \sigma_{u_{\theta s}} da_s \\
&\approx \sum_t \phi(l_{st}) \frac{w_t \sigma_{u_{\theta s}}}{\phi(l_r)} \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s} = l_{st}, u_{\delta i} = l_{id}) w_{id}\right) \\
&= \sum_t \phi(l_{st}) w_{st} \exp\left(\sum_i \sum_d g_i(l_{id}) \log p(y_{is}|u_{\theta s} = l_{st}, u_{\delta i} = l_{id}) w_{id}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
l_{st} &= \mu_{u_{\theta s}} + \sigma_{u_{\theta s}} l_t, \\
w_{st} &= \frac{\sigma_{u_{\theta s}} w_t}{\phi(l_r)},
\end{aligned}
$$

are the person-specific quadrature locations and the corresponding weights for integrating over $u_{\theta s}$, and $\mu_{u_{\theta s}}$ and $\sigma_{u_{\theta s}}$ are the posterior means and standard deviations for $u_{\theta s}$.

Chapter 3

Monte Carlo Local Likelihood Method

3.1 Introduction

Maximum likelihood estimation for generalized linear mixed models (GLMMs) is hindered by high dimensional intractable integrals involved in the likelihood function. The problem is magnified when the random effects have a crossed design and thus the data cannot be reduced to small independent clusters (Vaida & Meng, 2005). For instance, a logistic mixed model for a binary outcome y_{ij} can be written as

$$\text{logit}(p(y_{ij} = 1|z_i, u_j)) = \mu + z_i + u_j,$$

where $z_i \sim N(0, \sigma^2)$ with $i = 1, ..., m$ and $u_j \sim N(0, \tau^2)$ with $j = 1, ..., n$ are independent random effects that are crossed with each other. If all combinations of i and j exist in the data, this likelihood function involves $m + n$ dimensional integrals and its integrand involves a product of $m \times n$ terms.

Various methods have been proposed for approximating the intractable likelihood function. For instance, the Laplace approximation (Tierney & Kadane, 1986; Lindstrom & Bates, 1988; Wolfinger, 1993) and adaptive quadrature (Naylor & Smith, 1982; Rabe-Hesketh et al., 2005; Schilling & Bock, 2005) have been widely used. The Laplace approximation and similar penalized quasi-likelihood (PQL; Breslow & Clayton, 1993) are known to perform poorly for small cluster sizes and for large variance components (Breslow & Lin, 1995; Joe, 2008). Adaptive quadrature is more accurate but computationally more demanding than Gaussian quadrature (Pinheiro & Bates, 1995). For more reviews, see e.g. Pinheiro & Bates (1995).

Monte Carlo (MC) methods have also been utilized in various ways in ML estimation. Most methods are based on sampling the random effects given fixed parameter estimates. These methods can be distinguished by whether a 'single sample' or 'many samples' are used per evaluation of the objective function (for this distinction, see Geyer, 1996). The 'single sample' method is computationally more efficient than the 'many samples' method because it uses the same samples for all evaluation of the objective function. For instance, Geyer

& Thompson (1992), Geyer (1994), and Sung & Geyer (2007) used MC simulations of the random effects for an importance sampling approximation of the likelihood (or the likelihood ratio). The efficiency of the 'single sample' method highly depends on the importance sampling distribution. If the initial guess of parameters is far from the true parameter values, this method can perform poorly (Geyer, 1994; McCulloch, 1997). Geyer (1996) suggested iterating the procedure so that the objective function is maximized around a true parameter region. However, it requires many MC samples per each iteration of the algorithm.

MC expectation maximization (MCEM) is an example of a 'many sample' method. Several MCEM algorithms have been proposed using various sampling methods: e.g., a Metropolis-Hastings (McCulloch, 1997), an independent sampler based on importance sampling or rejection sampling (Booth & Hobert, 1999), and a slice sampler (Vaida & Meng, 2005). The basic idea is to use MC samples to approximate the intractable conditional expectation. MCEM requires samples at each iteration of the algorithm. In addition, the algorithm needs a method for calculating standard errors because it does not evaluate the likelihood function or its derivatives. A method for monitoring convergence may also be required (e.g., Booth & Hobert, 1999).

Compared to the MC methods described above, an MC kernel likelihood (MCKL) algorithm (De Valpine, 2004) takes a unique position in that it jointly samples the parameters and random effects to approximate the likelihood function. MCKL is a 'single sample' method because once the posterior samples of the parameters (along with samples of the random effects) are obtained, they are used during all iterations of the algorithm. MCKL is different from the typical 'single sample' method that samples the random effects given particular parameter values. Specifically, the MCKL algorithm initially treats parameters as having probability densities and samples them from a posterior density as in Bayesian methods. The likelihood is estimated up to a constant as a weighted kernel density estimate where the weights are obtained by considering the posterior as an importance sampling density. The likelihood can also be estimated up to a constant as an unweighted kernel density estimated divided by the prior. De Valpine demonstrated the efficiency of the MCKL method in estimating the parameters of population dynamic models. However, a method for standard errors has not been provided yet for MCKL.

In this paper, we propose a MC local likelihood (MCLL) method for estimating GLMMs. MCLL is similar to MCKL in spirit: The algorithm begins with treating the parameters as random variables and sampling them jointly with random effects from a posterior distribution for a particular prior distribution (we discuss how to choose the prior later in this paper). The likelihood function is then approximated up to a constant by fitting a density to the posterior samples of the parameters and dividing it by the prior. In contrast to MCKL, we approximate the posterior density using local likelihood density estimation (Hjort & Jones, 1996; Loader, 1996), where the log-likelihood is locally approximated by a polynomial. An unweighted version of MCKL can be seen as a special case of MCLL with a polynomial of degree zero. One motivation for MCLL is that the kernel density estimate usually shows a substantial bias in near peaks (Loader, 1999, Ch.2). Furthermore, MCLL can exploit the

form of the local likelihood density estimate to provide estimates of standard errors that are accurate and easy to calculate.

The outline of this chapter is as follows. In Section 3.2, we introduce the general idea of local likelihood density estimation. The MCLL algorithm is then described in detail as well as some implementation issues. In Section 3.3, we discuss computation of standard errors and marginal likelihoods. Empirical and simulation studies are provided in Sections 3.4 and 3.5 to evaluate the proposed MCLL algorithm. The paper ends with some concluding remarks.

3.2 Monte Carlo Local Likelihood Method

The key idea of MCLL is to use local likelihood density estimation in order to approximate a likelihood function. In this section, we begin by outlining the general idea of local likelihood density estimation. The procedure of the MCLL algorithm is then described in detail and some implementation issues are discussed.

3.2.1 Local Likelihood Density Estimation

Suppose X is a random variable having unknown density $f(x)$ and $x_1, ..., x_n$ are n independent observations of X. Given a parametric family $f(x;\psi)$, we approximate $f(x)$ by $\hat{f}(x) = f(x;\hat{\psi}(x))$ as proposed by Hjort & Jones (1996). $\hat{\psi}$ is obtained by maximizing a local likelihood for $f(x)$, which is defined as

$$l(x,\psi) = \sum_{j=1}^{n} w(x_j - t)\log f(x_j;\psi) - n\int_R w(u-t)f(u;\psi)du,$$

where the nonnegative weight function is $w(u) = \frac{K(u/h)}{h}$, where K is a symmetric unimodal density function (or kernel function) and h is a bandwidth. When h goes to infinity, maximizing $l(x,\psi)$ is equivalent to maximizing the usual likelihood. With moderate h, maximizing $l(x,\psi)$ covers a semi-parametric version of the likelihood.

The local polynomial approximation supposes that $\log f(x)$ can be well approximated by a low-degree polynomial in a neighborhood of the fitting point t (Loader, 1996). That is,

$$\log f(x) \approx P_{\boldsymbol{a}}(x-t).$$

In the one dimensional case, we can write

$$P_{\boldsymbol{a}}(x-t) = a_0 + a_1(x-t) + \cdots + a_p(x-t)^p,$$

where $\boldsymbol{a} = (a_0, a_1, \cdots, a_p)'$ is the parameter vector for the polynomial function with degree

p. The localized log-likelihood $\hat{l}(x, \psi)$ can then be approximated as

$$\sum_{j=1}^{n} K\left(\frac{x_j - t}{h}\right) P_{\boldsymbol{a}}(x_j - t) - n \int K\left(\frac{u-t}{h}\right) \exp\left(P_{\boldsymbol{a}}(u-t)\right) du, \tag{3.1}$$

where the parameter space for $\boldsymbol{a}$ is assumed to be an open set which holds if K is continuous with bounded support.

If a maximizer of (3.1) exists, it satisfies the system of local likelihood equations

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n} A\left(\frac{x_j - t}{h}\right) K\left(\frac{x_j - t}{h}\right) \\ &= \int A\left(\frac{u-t}{h}\right) K\left(\frac{u-t}{h}\right) \exp\left(P_{\boldsymbol{a}}(u-t)\right) du, \end{aligned}$$

where $A(v) = (1, v, \cdots, v^p)'$. This equation shows the moment matching property between sample and population moments of the local likelihood density estimator (Loader, 1996).

Theoretical properties of local likelihood density estimation have been examined by e.g., Eguchi & Copas (1998), Hall et al. (2002), and Park et al. (2002). Recently, Delicado (2006) has proposed a local likelihood density estimation based on smooth truncation using a uniform kernel. Kauermann & Tutz (2000) and Wu & Zhang (2002) used local likelihood estimation for linear mixed and generalized linear mixed models, respectively, but in the context of approximating non-parametric functions.

3.2.2 MCLL Procedure

MCLL begins with obtaining Markov chain Monte Carlo (MCMC) samples of model parameters from the posterior for a particular set of priors. Then the algorithm involves two nested maximization steps: Maximizing an approximate likelihood $\hat{L}(\mathbf{y}|\boldsymbol{\theta})$ over $\boldsymbol{\theta}$, with each evaluated value of $\boldsymbol{\theta}$ requiring a maximization over parameters in the local polynomial function involved in calculating $\hat{L}(\mathbf{y}|\boldsymbol{\theta})$. These two maximization steps iterate until convergence.

Specifically, assuming a d-dimensional parameter space $\boldsymbol{\theta}$ with observed data vector $\mathbf{y}$, the MCLL algorithm proceeds as follows:

Step 1. Choose a prior $p(\boldsymbol{\theta})$ and use an MCMC method to obtain samples from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{C_s},$$

where the normalizing constant is $C_s = \int L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$.

Step 2. Maximize an approximate likelihood, defined up to the unknown constant C_s by

$$\hat{L}(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{p(\boldsymbol{\theta})}\mathrm{Ps}_p(\boldsymbol{\theta}|\mathbf{y}), \tag{3.2}$$

where $\mathrm{Ps}_p(\boldsymbol{\theta}|\mathbf{y})$ is the local likelihood estimate of the posterior density. Specifically, for a given value of $\boldsymbol{\theta}$, this is obtained by assuming that the log-posterior density can be locally approximated by a polynomial function $P_{\boldsymbol{a}}(\mathbf{u}-\boldsymbol{\theta})$ with parameters $\boldsymbol{a}$. For example, in the three dimensional case (d=3), in the vicinity of $\boldsymbol{\theta}$ the log-posterior can be approximated by a quadratic function

$$\begin{aligned} P_{\boldsymbol{a}}(\mathbf{u}-\boldsymbol{\theta}) = {} & a_0 + a_1(u_1-\theta_1) + a_2(u_2-\theta_2) + a_3(u_3-\theta_3) \\ & + \frac{1}{2}a_4(u_1-\theta_1)^2 + \frac{1}{2}a_5(u_2-\theta_2)^2 + \frac{1}{2}a_6(u_3-\theta_3)^2 \\ & + a_7(u_1-\theta_1)(u_2-\theta_2) + a_8(u_1-\theta_1)(u_3-\theta_3) \\ & + a_9(u_2-\theta_2)(u_3-\theta_3), \end{aligned} \tag{3.3}$$

where $\boldsymbol{a} = (a_0, a_1, ..., a_9)'$.

The $\boldsymbol{a}$ parameters are estimated for a particular $\boldsymbol{\theta}$ by maximizing a localized version of the log-likelihood as in (3.1), which in this case is

$$l(\boldsymbol{\theta}, \boldsymbol{a}) = \sum_{j=1}^{m} K\left(\frac{\boldsymbol{\theta}^{(j)}-\boldsymbol{\theta}}{\mathbf{h}}\right) P_{\boldsymbol{a}}(\boldsymbol{\theta}^{(j)}-\boldsymbol{\theta}) - m\int K\left(\frac{\mathbf{u}-\boldsymbol{\theta}}{\mathbf{h}}\right)\exp\left(P_{\boldsymbol{a}}(\mathbf{u}-\boldsymbol{\theta})\right)d\mathbf{u}, \tag{3.4}$$

where $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{m}$ are the posterior sample points.

The approximate likelihood function (3.2) in Step 2 can be seen as an unweighted estimate of the posterior density (De Valpine, 2004). A weighted version can be formulated as

$$\begin{aligned} \hat{L}(\mathbf{y}|\boldsymbol{\theta}) &= \frac{1}{m}\sum_{j=1}^{m}\mathrm{Ps}_p(\boldsymbol{\theta}|\mathbf{y})w^{(j)}, \\ w^{(j)} &= \frac{1}{p(\boldsymbol{\theta}^{(j)})}, \end{aligned}$$

where $w^{(j)}$ is the weight for $\mathrm{Ps}_p(\boldsymbol{\theta}|\mathbf{y})$ and $p(\boldsymbol{\theta}^{(j)})$ is the prior density evaluated at $\boldsymbol{\theta}^{(j)}$.

In the MCKL case, the weighted version may be preferable because it can be seen as an unnormalized, importance-sampled kernel estimate of the true likelihood $L(\mathbf{y}|\boldsymbol{\theta})$. However, when local density estimation is used as in MCLL, it is no longer clear how the weighted version can be seen as an importance-sampled estimate of the true likelihood. The unweighted

version may have an issue with narrow priors since a maximum may not exist in such a case; but with wide priors, there is little difference in performance between the weighted and unweighted versions (De Valpine, 2004). In addition, the unweighted version is easier to implement in practice than the weighted version. Therefore, we adopt the unweighted version as the main device for MCLL.

3.2.3 Implementation Issues

There are several issues to be discussed in implementing the MCLL algorithm. First, the bandwidth is chosen in Step 2 by considering the bias-variance trade-off. We choose a bandwidth at each data point so that the local neighborhood contains a specified number of data points. For a smoothing parameter α between 0 and 1, the nearest neighbor bandwidth is chosen as the kth smallest distance d, where $k = \lfloor n\alpha \rfloor$ and $d(x, x_i) = |x - x_i|$.

The degree of the local polynomial function can also affect the bias-variance trade-off. Fitting a high degree will usually lead to less bias but large variability of an estimate. We choose a quadratic function as a default because it is often sufficient to choose a low degree polynomial and focus on choosing the bandwidth to obtain a satisfactory fit (Loader, 1999, Ch.2).

The weight function affects the visual quality of the fitted shape rather than the bias-variance trade-off. A spherically symmetric weight function is usually used. We choose a tricube weight function, $K(u) = \frac{\exp(-|u|)}{1+|u|^3}$ as a default. Hjort & Jones (1996) suggested the Gaussian function for which closed-form evaluation of the integrals is available. But for local quadratic fitting, the parameters are constrained, which limits the ability of the estimate to reproduce troughs in the data (Loader, 1996).

Second, we consider orthogonal transformation of the posterior samples $\boldsymbol{\theta}^{(j)}$. Assuming multivariate normality, the posterior samples can be transformed as

$$\tilde{\boldsymbol{\theta}}^{(j)} = L^{-1}(\boldsymbol{\theta}^{(j)} - \mathbf{b}),$$

where $\mathbf{b}$ is the mean of the posterior samples $\boldsymbol{\theta}^{(j)}$ and L is the Cholesky decomposition of the empirical covariance matrix $\widehat{\mathrm{Cov}}(\boldsymbol{\theta})$ of the posterior samples $\boldsymbol{\theta}^{(j)}$ so that $\widehat{\mathrm{Cov}}(\boldsymbol{\theta}) = LL^T$. The transformed $\tilde{\boldsymbol{\theta}}^{(j)}$ have an identity covariance matrix and a zero mean vector.

This orthogonal transformation is also called data presphering (Wand & Jones, 1993; Duong & Hazelton, 2003). Preshering posterior samples is useful in implementing MCLL because it simplifies the integral term in (3.4). Specifically, for multidimensional parameter $\boldsymbol{\theta}$, if the components are approximately independent in the posterior, then interactions terms in $P_{\boldsymbol{a}}(\mathbf{u} - \boldsymbol{\theta})$ can be dropped. In addition, a product kernel can be used, with

$$K\left(\frac{\mathbf{u} - \boldsymbol{\theta}}{\mathbf{h}}\right) = \prod_{i=1}^{d} K_0\left(\frac{u_i - \theta_i}{h_i}\right), \tag{3.5}$$

where K_0 is a one-dimensional kernel. With these two simplifications, the multidimensional integral can be factorized as a product of one-dimensional integrals due to the orthogonality of the parameter space. In addition, the orthogonal transformation standardizes a bandwidth choice by transforming the parameter space to be on the same scale.

If the parameter space is not orthogonal, a product kernel (3.5) may be inappropriate to use. In such cases, the interaction terms can be included in the polynomial function $P_{\boldsymbol{a}}(\mathbf{u}-\boldsymbol{\theta})$, for which multidimensional-integration is needed. Ordinary quadrature rules are then no longer practical because for instance, 30^6 evaluations are required with 30 quadrature points when $d = 6$. Instead, Halton sequences can be used to reduce the number of evaluations (Sándor & Train, 2004). Draws derived from Halton sequences have the advantage of both improving coverage of the domain of integration and inducing a negative correlation between the draws from different observations. Other quasi-random integration rules could also be used (Fang & Wang, 1993). Halton draws are more effective than quasi-random draws because the same accuracy can be achieved with Halton draws with a smaller number of draws, thereby saving computer time (Train, 2003).

Third, we use a log-transformation of variance parameters. This has several advantages: First, it avoids need for a modified kernel for variance parameters in Step 2. Second, the posterior distributions are closer to normal so that the data presphering operation works better for a symmetric distribution.

Fourth, non-informative priors can be chosen for the fixed and log standard deviation parameters, in which case the posterior mean estimates (automatically obtained in Step 1) are also close to ML estimates. Note that even if informative priors are used, however, the MCLL algorithm provides results close to the ML estimates, unlike the posterior mean estimates. Informative priors are useful for improving mixing in MCMC in Step 1 but some care is required. We illustrate the choice of priors for given problems in the empirical study section.

We wrote an R package `mcll` that implements the MCLL algorithm (Step 2 maximizations).

3.3 Inference

Standard error estimates and the values of maximized log-likelihoods are standard tools for likelihood-based inference. Since they are not by-products of the MCLL algorithm, we develop methods for obtaining standard errors and marginal likelihoods. We also show how to compute the Bayes factor in a relatively simple way with MCLL.

3.3.1 Standard Errors

Asymptotic theory for the ML estimation (MLE) suggests obtaining standard error estimates using the Hessian matrix of the log-likelihood function evaluated at the ML estimates. Analo-

gously, one could calculate the Hessian matrix of $\log\hat{L}(\mathbf{y}|\boldsymbol{\theta})$ through numerical differentiation of the log-likelihood function. However, curvatures obtained by numerical differentiation will be sensitive to the bandwidth choice.

Therefore, we derive an alternative way of computing the Hessian matrix for MCLL. First write down the log-likelihood function $\log L(\mathbf{y}|\boldsymbol{\theta})$

$$\log L(\mathbf{y}|\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|\mathbf{y}) - \log p(\boldsymbol{\theta}) + C_s, \tag{3.6}$$

where $\log p(\boldsymbol{\theta}|\mathbf{y})$ is the log-posterior, $\log p(\boldsymbol{\theta})$ is the log-prior, and C_s is a constant.

Take the second derivatives with respect to $\boldsymbol{\theta}$ on both sides in (3.6) as

$$\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log L(\mathbf{y}|\boldsymbol{\theta})\bigg|_{\hat{\boldsymbol{\theta}}} = \frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log p(\boldsymbol{\theta}|\mathbf{y})\bigg|_{\hat{\boldsymbol{\theta}}} - \frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log p(\boldsymbol{\theta})\bigg|_{\hat{\boldsymbol{\theta}}},$$

evaluated at $\hat{\boldsymbol{\theta}}$. Simply rewrite this as

$$H_L = H_{\mathrm{Ps}} - H_{\mathrm{Pr}},$$

where H_L, H_{Ps}, and H_{Pr} are the Hessian matrices of the approximate log-likelihood $\log\hat{L}(\mathbf{y}|\boldsymbol{\theta})$, the log-posterior $\log p(\boldsymbol{\theta}|\mathbf{y})$, and the log-prior $\log p(\boldsymbol{\theta})$, respectively.

Typically H_{Pr} can be solved analytically. To obtain H_{Ps}, we use the quadratic approximation of the log-posterior obtained using local likelihood density estimation, assuming the log-posterior can be well approximated by a quadratic polynomial in the neighborhood of the mode. For example, in the case $d = 3$ and a quadratic function as given in (3.3), the Hessian matrix of the approximation is

$$H_{\mathrm{Ps}} \approx \begin{bmatrix} \hat{a}_4 & \hat{a}_7 & \hat{a}_8 \\ \hat{a}_7 & \hat{a}_5 & \hat{a}_9 \\ \hat{a}_8 & \hat{a}_9 & \hat{a}_6 \end{bmatrix}. \tag{3.7}$$

The coefficients for the interaction terms in (3.3) correspond to the off-diagonal terms ($\hat{a}_7$ to $\hat{a}_9$) in (3.7) and are zero if the elements of $\boldsymbol{\theta}$ are uncorrelated in the posterior. This will be approximately true if the orthogonal transformation has been used. Thus in practice, these off-diagonal terms are set to zero and not estimated.

3.3.2 Likelihood Inference

Suppose there are n observed responses $\mathbf{y}_i$ for n subjects i with random effects (or missing data) $\mathbf{z}_i$ in the context of GLMMs. Assuming a d-dimensional parameter vector $\boldsymbol{\theta}$, the

marginal (normalized) likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ can be written as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \int p(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i) d\mathbf{z}_i, \tag{3.8}$$

where $p(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta})$ is the joint distribution of $\mathbf{y}_i$ given the random effects $\mathbf{z}_i$ and the parameter values $\boldsymbol{\theta}$. $p(\mathbf{z}_i)$ is the prior distribution for $\mathbf{z}_i$.

In this section, we show how to approximate the marginal likelihood in (3.8) and how the likelihood-ratio (LR) statistic and the Bayes factor can be readily computed with MCLL.

Marginal Likelihoods

In general, computation of the marginal likelihood is not feasible for GLMMs. Using a sampling method, however, we can approximate the marginal likelihood as follows. First obtain posterior samples of the random effects using MCMC with model parameters treated as known constants set equal to the MCLL estimates. Then obtain the sample mean and covariance matrix of the posterior samples and use the corresponding multivariate normal distribution as importance density. Sample the random effects $\mathbf{z}$ from the importance density. Then the marginal likelihood can be approximated as

$$\begin{aligned} f(\mathbf{y}|\hat{\boldsymbol{\theta}}) &= \int \frac{p(\mathbf{y}|\mathbf{z}, \hat{\boldsymbol{\theta}}) p(\mathbf{z}; \hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{z}|\mathbf{y}, \hat{\boldsymbol{\theta}})} \tilde{p}(\mathbf{z}|\mathbf{y}, \hat{\boldsymbol{\theta}}) d\mathbf{z} \\ &\approx \frac{1}{m} \sum_{j=1}^{m} \frac{p(\mathbf{y}|\mathbf{z}^{(j)}, \hat{\boldsymbol{\theta}}) p(\mathbf{z}^{(j)}; \hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{z}^{(j)}|\mathbf{y}, \hat{\boldsymbol{\theta}})} = \hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}), \end{aligned}$$

where the importance density $\tilde{p}(\mathbf{z}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ is the normal approximation to the posterior samples of the random effects, which has the same support as the prior $p(\mathbf{z}; \hat{\boldsymbol{\theta}})$. $\mathbf{z}^{(j)}$ $(j = 1, ..., m)$ is identically and independently drawn from $\tilde{p}(\mathbf{z}|\mathbf{y}, \hat{\boldsymbol{\theta}})$. We use the multivariate normal assumption with the mean and covariance matrix given by the empirical mean and covariance matrix of the MCMC samples.

By the strong law of large numbers, the importance approximation of the likelihood $\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}})$ is unbiased and consistent as $m \to \infty$, as long as the support of $\tilde{p}(\cdot)$ contains the support of $f(\cdot)$ (Geweke, 1989). A similar idea of using importance sampling was adopted to evaluate a likelihood surface by Durbin & Koopman (1997) and Shephard & Pitt (1997).

The approximate marginal likelihood $\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}})$ almost surely converges to the true likelihood $f(\mathbf{y}|\hat{\boldsymbol{\theta}})$ no matter which importance density is chosen, but the rate of convergence depends on the accuracy of the importance density used. To measure the accuracy of the importance density, the effective sample size (ESS) can be computed following Liu (2001)

$$\text{ESS} = \frac{m}{1 + \text{var}(w_j)},$$

where m is the MCMC sample size and

$$w_j = \frac{p(\mathbf{z}^{(j)};\hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{z}^{(j)}|\mathbf{y},\hat{\boldsymbol{\theta}})} \Big/ \sum_{j=1}^{m} \frac{p(\mathbf{z}^{(j)};\hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{z}^{(j)}|\mathbf{y},\hat{\boldsymbol{\theta}})}.$$

Here $\text{var}(w_j)$ is the variance of the m importance weights over the distribution defined by $\tilde{p}(\cdot)$. A large variance leads to low efficiency relative to the sample size m and results in low ESS. In practice, m can be chosen to produce a sufficiently large value (close to ESS) to ensure small $\text{var}(w_j)$.

Following Shephard & Pitt (1997), the approximate log-likelihood is asymptotically unbiased as $m \to \infty$. With finite m, the bias can be expressed as

$$\log\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}) = \log f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \log\frac{1}{m}\sum_{j=1}^{m} \frac{p(\mathbf{z}^{(j)};\hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{z}^{(j)}|\mathbf{y},\hat{\boldsymbol{\theta}})}.$$

Here the term $\log\frac{1}{m}\sum_{j=1}^{m}\frac{p(\mathbf{Z}^{(j)}|\hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{Z}^{(j)}|\mathbf{y},\hat{\boldsymbol{\theta}})}$ is biased to $O(m^{-1})$ and thus disappears as m increases to infinity. The bias-corrected log-likelihood can be derived as

$$\log\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \frac{1}{2m}\frac{1}{m-1}\sum_{j=1}^{m}\{f(\mathbf{y}|\hat{\boldsymbol{\theta}})^{(j)} - \hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}})\}^2, \tag{3.9}$$

where $f(\mathbf{y}|\hat{\boldsymbol{\theta}})^{(j)} = \frac{p(\mathbf{y}|\mathbf{Z}^{(j)},\hat{\boldsymbol{\theta}})p(\mathbf{Z}^{(j)};\hat{\boldsymbol{\theta}})}{\tilde{p}(\mathbf{Z}^{(j)}|\mathbf{y},\hat{\boldsymbol{\theta}})}$.

Likelihood Ratio Statistics

The approximate marginal log-likelihood can be used to compute likelihood ratio test statistics. For example, denote $\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)$ and $\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)$ the approximate likelihoods for the two models M_1 and M_2, where M_1 is nested in M_2. The likelihood ratio statistic $\hat{\lambda}(\mathbf{y})$ can be computed as

$$\hat{\lambda}(\mathbf{y}) = -2[\log\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1) - \log\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)].$$

Since $\hat{\lambda}(\mathbf{y})$ converges in probability to $\lambda(\mathbf{y})$ as $m \to \infty$, under the null hypothesis

$$\begin{aligned}\lim_{n\to\infty}\lim_{m\to\infty} p(\hat{\lambda}(\mathbf{y}) > \lambda_\alpha) \\ = \lim_{n\to\infty} p(\lambda(\mathbf{y}) > \lambda_\alpha) \\ = p(\chi^2_{p-q} > \lambda_\alpha) = \alpha,\end{aligned}$$

where n is the sample size, α is the critical point, and p and q are the number of parameters in M_1 and M_2.

With finite m, $\hat{\lambda}(\mathbf{y})$ is still biased because of the bias in $\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)$ and $\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)$. An unbiased estimator of $\tilde{\lambda}(\mathbf{y})$ can be obtained as

$$\begin{aligned}\tilde{\lambda}(\mathbf{y}) &= -2[\log\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1) + \widehat{\text{bias}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1) - \log\hat{f}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2) - \widehat{\text{bias}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)] \\ &= \hat{\lambda}(\mathbf{y}) - 2[\widehat{\text{bias}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1) - \widehat{\text{bias}}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)],\end{aligned}$$

where $\widehat{\text{bias}}(\mathbf{y}|\boldsymbol{\theta}_k^*, M_k)$ for M_k was defined in (3.9).

3.3.3 Bayes Factors

Bayes factors are an important tool for Bayesian inference and can also be useful in a frequentist context. For example, the null hypothesis is rejected when the Bayes factor is small where the magnitude depends on the distribution of the Bayes factor under the null hypothesis and the significance level desired for the test (Chacón et al., 2007). Moreover, Bayes factors allow comparisons of nonnested models, irregular models, and more than two models (Kass & Raftery, 1995).

A Bayes factor can be defined as the ratio of the marginal likelihoods for model M_1 and M_2

$$\text{BF}_{12} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)},$$

where the marginal likelihood for M_k is defined as

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k.$$

Here $p(\mathbf{y}|\boldsymbol{\theta}_k, M_k)$ is the joint density of the responses given model M_k with parameter values $\boldsymbol{\theta}_k$ and $p(\boldsymbol{\theta}_k|M_k)$ is the prior density for the model parameters $\boldsymbol{\theta}_k$ in model M_k.

Estimation of the Bayes factor is a difficult problem because the marginal likelihoods are not easily computed from the output of the MCMC algorithm. The MCLL method provides a relatively simple way to compute the Bayes factor. To show that, first write down the posterior densities of the model parameters for models M_1 and M_2

$$\begin{aligned}p(\hat{\boldsymbol{\theta}}_1|\mathbf{y}, M_1) &= p(\hat{\boldsymbol{\theta}}_1|M_1)\frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)}{p(\mathbf{y}|M_1)}, \\ p(\hat{\boldsymbol{\theta}}_1|\mathbf{y}, M_2) &= p(\hat{\boldsymbol{\theta}}_2|M_2)\frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)}{p(\mathbf{y}|M_2)},\end{aligned}$$

where $p(\hat{\boldsymbol{\theta}}_k|M_k)$ is the prior and $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, M_k)$ is the likelihood given the MCLL estimates $\hat{\boldsymbol{\theta}}_k$

for model M_k. Dividing both sides by their priors, we obtain

$$\frac{p(\hat{\boldsymbol{\theta}}_1|\mathbf{y}, M_1)}{p(\hat{\boldsymbol{\theta}}_1|M_1)} = \frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)}{p(\mathbf{y}|M_1)}, \tag{3.10}$$

$$\frac{p(\hat{\boldsymbol{\theta}}_2|\mathbf{y}, M_2)}{p(\hat{\boldsymbol{\theta}}_2|M_2)} = \frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)}{p(\mathbf{y}|M_2)}. \tag{3.11}$$

Notice that the left hand sides in (3.10) and (3.11) are the unnormalized MCLL likelihoods $\hat{L}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)$ (the posterior density divided by the prior) for M_1 and M_2 as computed in (3.2). Dividing (3.11) by (3.10), obtain

$$\frac{\hat{L}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)}{\hat{L}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \times \frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)}.$$

Notice that $\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \widehat{\mathrm{BF}}_{12}$. That is, the Bayes factor $\widehat{\mathrm{BF}}_{12}$ is obtained as

$$\widehat{\mathrm{BF}}_{12} = \frac{\hat{L}(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)}{\hat{L}(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)} \times \frac{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, M_1)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, M_2)}, \tag{3.12}$$

where $\hat{L}(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, M_k)$ $(k = 1, 2)$ are by-products of the MCLL algorithm and the likelihood $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, M_k)$ can be obtained as described in Section 3.3.2. Note that this method works for any method that provides an unnormalized likelihood such as MCKL.

3.4 Empirical Studies

To illustrate the proposed algorithm, we consider three empirical examples: 1) the salamander mating data (McCullagh & Nelder, 1989), 2) the birth weight data (Rabe-Hesketh et al., 2008), and 3) the longitudinal data of Korean students (Jeon & Rabe-Hesketh, 2012).

3.4.1 Salamander Mating Data

The salamander mating dataset is a benchmark that has been used to compare many different estimation methods for GLMMs with crossed random effects (e.g., Karim & Zeger, 1992; Breslow & Clayton, 1993; Booth & Hobert, 1999; Lee & Nelder, 2006; Cho & Rabe-Hesketh, 2011). This dataset consists of three separate experiments, each involving matings among salamanders of two different populations, called Rough Butt (RB) and White Side (WS). Sixty females and sixty males of two populations of salamander were paired by a crossed, blocked, and incomplete design in an experiment studying whether the two populations have

developed generic mechanisms which would prevent inter-breeding. The response is a binary variable indicating whether mating was successful between female i and male j. We adopted the model A used in Karim & Zeger (1992)

$$\text{logit}(p(y_{ij} = 1|z_i^f, z_j^m)) = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2j} + \beta_4 x_{1i} x_{2j} + z_i^f + z_j^m, \tag{3.13}$$

where the covariates are dummy variables for White Side female (x_i), White Side male (x_j), and the interaction ($x_{1i}x_{2j}$). The two crossed random effects are random intercepts $z_i^f \sim N(0, \sigma_f^2)$ for females and $z_i^m \sim N(0, \sigma_m^2)$ for males. Each salamander participates in six matings, resulting in 360 matings in total. The two variance components in model (3.13) were reparameterized as $\tau_f = \log\sigma_f$ and $\tau_m = \log\sigma_m$.

The MCLL parameter estimates were compared with the Laplace approximation and Bayesian estimates (posterior means). They were also compared with the estimates from the literature, such as PQL (Breslow & Clayton, 1993), MCEM (Booth & Hobert, 1999), and MCMLE (Sung & Geyer, 2007). In addition, the MCKL method (De Valpine, 2004) was implemented for another comparison with MCLL.

The methods for likelihood inference described in Section 3.3 were implemented. First, for each parameter the marginal likelihood was computed in the neighborhood of the MCLL estimate with other parameters set equal to the MCLL estimates. Second, a reduced model was fit without the interaction parameter (β_4 coefficient of $x_{1i}x_{2j}$). The likelihood-ratio statistic and the Bayes factor were calculated to compare the reduced model with the full model.

Finally, the standard errors were computed both using diagonal and full Hessian matrices. For an efficient multivariate integration with the full Hessian matrix, we used Halton sequences with 20,000 draws. The computation time was compared between the diagonal and full Hessian methods.

3.4.2 Implementation

To obtain the MCMC samples from the posterior distribution in Step 1, diffuse normal priors were specified for the fixed effect (regression coefficient) parameters (with mean 0, standard deviation 100) and for the log standard deviation parameters τ_f and τ_m (with mean -0.98 and standard deviation 0.76). These specific values were chosen by noting that the mean and standard deviation can be analytically solved for the untransformed parameters σ_f and σ_m using the moments of the corresponding log-transformed variables. The mean and variance for the log-transformed variable can be obtained using

$$E(\sigma) = \log E(\tau) - \frac{1}{2}\log\left(1 + \frac{\text{Var}(\tau)}{E(\tau)^2}\right),$$
$$\text{Var}(\sigma) = \log\left(1 + \frac{\text{Var}(\tau)}{E(\tau)^2}\right),$$

where $E(\tau)$ and $\mathrm{Var}(\tau)$ are the mean and variance for the log-transformed variable. For example, to obtain the mean $E(\tau) = 0.5$ and variance $\mathrm{Var}(\tau) = 0.44^2$ for τ, we use $E(\sigma) = -0.98$ and $\mathrm{Var}(\sigma) = 0.76^2$ for σ.

The Bayesian software `WinBUGS` (Spiegelhalter et al., 2003) was used to obtain the posterior samples in Step 1, which was run by the R package `R2WinBUGS` (Sturtz et al., 2005). Three chains were used with relatively diffuse starting values. Each chain was run for 5,000 iterations after a 2,000 iteration burn-in period. For convergence assessment, the Gelman-Rubin statistic (Gelman & Rubin, 1992) was used in addition to graphical checks such as trace plots and autocorrelation plots. For Step 2, we use the R package `mcll` that we developed. For the bandwidth selection in Step 2, we used the default smoothing parameter $\alpha = 0.7$. A different choice of the smoothing parameter ($0 \leq \alpha \leq 1$) did not make much of a difference in the results.

To implement the Laplace approximation, we used the R function `lmer` in the `lme4` package (Bates & Maechler, 2009). For adaptive quadrature, `xtmelogit` in `Stata` (StataCorp, 2009) was used. To implement the MCKL method, we followed the procedure taken by Jeon (2011) using the same posterior samples as in the MCLL method. Specifically, for the bandwidth choice for MCKL, we took diagonal elements of the covariance matrix of the kernel density to be proportional to the marginal posterior variances in each dimension of the posterior space. For a proportionality constant, we used $q = 0.5$ although 10 different values (0.1 to 1.0) were all tried out for q. We also adjusted for smoothing bias in the MCKL estimates using posterior cumulants as suggested by De Valpine (2004).

Results

Table 3.1 lists the parameter estimates for model (3.13) from a variety of estimators in the literature. Standard errors for the regression coefficient parameters were included when they were reported in the original papers. Standard errors for the standard deviations σ_f and σ_m were not considered because the Wald-type tests and confidence intervals are inappropriate for these parameters (e.g., Berkhof & Snijders, 2001).

Overall, the results from MCLL were comparable to the other estimators. The regression coefficient estimates were a bit smaller than other estimates except PQL and MCMLE. The standard deviation estimates were close to the estimates from adaptive quadrature with three quadrature points. The MCKL parameter estimates were a bit smaller than the MCLL estimates. With a different bandwidth choice, the MCKL estimates also varied somewhat.

Our standard error estimates were quite close to those from the other estimators. With the full Hessian matrix, we obtained $(0.41, 0.56, 0.30, 0.48)'$ for the standard errors for the regression coefficient parameters in order. These were a bit smaller than those from the diagonal Hessian matrix. As for computation time, it took 54,956 seconds with the full Hessian matrix compared with 360 seconds with the diagonal Hessian matrix on a Intel Pentium Dual-Core 2.5-GHz processor computer with 3.2 GB of memory.

The approximate log-likelihood was computed using importance sampling with m=3,000

Table 3.1: Comparison of several estimators for the salamander mating data. Standard errors are given in parentheses if reported. MCEM: Booth & Hobert (1999); PQL: Breslow & Clayton (1993); Laplace: `lmer`; Adaptive quad(3): `xtmelogit` with 3 quadrature points; MCMLE: Sung & Geyer (2007); MCLL: MCLL method; Post.m: Posterior means (the posterior samples that were used for MCLL); MCKL: MCKL method after cumulant bias correction ($q = 0.5$).

Method	β_1	β_2	β_3	β_4	σ_m	σ_f
MCEM	1.02	-2.96	-0.69	3.63	1.18	1.12
PQL	0.79(0.32)	-2.29(0.43)	-0.54(0.39)	2.82(0.50)	0.79	0.72
Laplace	1.00(0.39)	-2.90(0.56)	-0.70(0.46)	3.59(0.64)	1.08	1.02
Adaptive quad(3)	1.01(0.41)	-2.95(0.58)	-0.70(0.48)	3.62(0.64)	1.16	1.10
MCMLE	1.00 (0.35)	-2.78(0.36)	-0.47(0.33)	3.52(0.53)	1.17	1.10
MCLL	0.93 (0.44)	-2.87(0.59)	-0.65(0.53)	3.59(0.72)	1.16	1.08
Post.m	1.01 (0.41)	-2.92 (0.58)	-0.69(0.49)	3.58(0.66)	1.09	1.02
MCKL	0.84	-2.77	-0.54	3.47	1.13	1.08

for a range of values for each parameter with the other parameters set equal to the MCLL estimates. The bias was close to zero for all parameters. The approximate log-likelihood was compared with that from adaptive quadrature. The results are shown in Figure 3.1.

In all sub-panels, our log-likelihood surfaces were close to those from adaptive quadrature, in terms of the overall shape, mode, and curvature at the mode. This shows that our method using importance sampling works quite well in approximating the log-likelihood.

To compare the full and reduced models, we computed the marginal log-likelihood, the LR statistic, and the Bayes factor. The marginal log-likelihood was -207.61 for the full model and -228.43 for the reduced model. For both models, our estimate of bias was close to zero and ESS was numerically the same as the MC sample size. Adaptive quadrature (with five quadrature points) provided similar marginal log-likelihoods, -207.62 and -228.44 for the full and reduced models, respectively. As for computation time, it took about 21,000 seconds with adaptive quadrature but only a few seconds with the importance sampling method.

The LR statistic between the full and reduced model was $\hat{\lambda} = -2(-228.41 + 207.62) = 41.58$ ($p < 0.001, df = 1$) and the Bayes factor was computed as 1.20 using (3.12). These results are strong evidence for the inclusion of the interaction term.

3.4.3 Birth Weight Data

In order to assess performance of MCLL for data where the true ML estimates are easy to obtain, we consider a linear mixed model. Specifically, we use the linear mixed model that was proposed by Rabe-Hesketh et al. (2008) to analyze nuclear family birth weight data

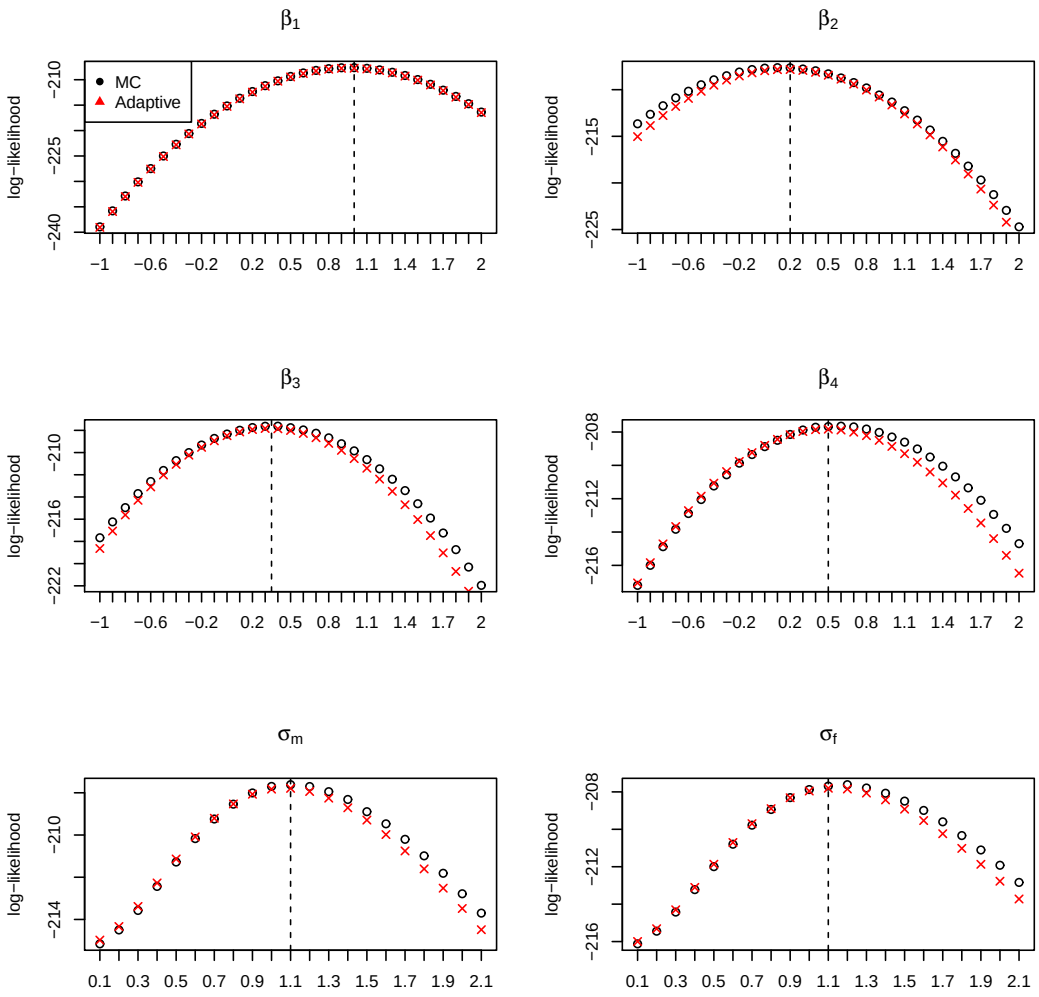

Figure 3.1: Log-likelihood surfaces obtained using importance sampling (MC) and adaptive quadrature (Adaptive). The vertical dashed lines indicate the MCLL estimates for the corresponding parameters.

from the Medical Birth Registry of Norway described in Magnus et al. (2001). In the original dataset, there were 1,000 nuclear families each comprising mother, father, and a single child (not necessarily the only child in the family). A two-level liner mixed model was formulated for family members i nested in families j with three uncorrelated random coefficients

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \alpha^{(2)}_{1j}[M_i + K_i/2] + \alpha^{(2)}_{2j}[F_i + K_i/2] + \alpha^{(2)}_{3j}[K_i/\sqrt{2}] + \epsilon_{ij}, \tag{3.14}$$

where $\mathbf{x}_{ij}$ is a vector of covariates with regression coefficients $\boldsymbol{\beta}$. M_i, K_i, and F_i are dummy variables for mother, child, and father, respectively. The covariates were male, a dummy variable for being male (x_{1ij}), midage, a dummy variable for mother aged 20-35 at time of birth (x_{2ij}), and highage, a dummy variable for mother older than 35 at time of birth (x_{3ij}). The three random effects at level 2, $\alpha^{(2)}_{1j}$, $\alpha^{(2)}_{2j}$, and $\alpha^{(2)}_{3j}$ are i.i.d as $\alpha^{(2)}_{kj} \sim N(0, \sigma^2_A)$ with $k = 1, 2, 3$. The level-1 residuals have a normal distribution, $\epsilon^{(2)}_{ij} \sim N(0, \sigma^2_E)$ and are independent of the random effects. Here σ_A can be interpreted as the additive genetic standard deviation and σ_E as the unique environment standard deviation.

To implement the MCLL method, the same settings were used as in the first example. Diffuse normal priors were specified for the regression coefficient parameters (mean 0, standard deviation 1,000) and for the log standard deviations logσ_A and logσ_E (mean 6.17, standard deviation 0.27). The MCLL estimates and standard errors (using the diagonal Hessian matrix) were compared with the Bayesian estimates and standard errors (posterior means and standard deviations) and with the true ML estimates which were obtained from the R function `lme` in the package `nlme` (Pinheiro et al., 2012).

Results

Table 3.2 lists the results for model (3.14) to the birth weight dataset.

Table 3.2: Parameter estimates (Est) and standard errors (SE) for the birth weight data. MLE is the true maximum likelihood estimates and Post.m is the posterior mean estimates

	MLE		Post.m		MCLL	
	Est	SE	Est	SE	Est	SE
β_1	3368.10	31.14	3366.00	31.50	3369.93	31.73
β_2	155.35	17.53	155.33	17.85	154.97	18.45
β_3	126.95	30.98	129.26	31.28	125.62	31.74
β_4	213.44	52.64	214.39	52.66	216.52	53.68
σ_E	374.67	-	375.94	-	375.58	-
σ_A	311.21	-	309.58	-	311.09	-

For regression coefficients, $\hat{\beta}_1$ indicates the estimated mean birth weight for female babies of mothers aged younger than 20 at the time of birth, and $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ represent the

estimated differences in the mean birth weight between male and female babies, between mothers, and between old and young mothers, controlling for the other variables. In the random part, the estimated genetic standard deviation ($\hat{\sigma}_A$) was a bit smaller than the estimated unique environment standard deviation ($\hat{\sigma}_E$).

The MCLL estimates were close to the true ML estimates both for the regression coefficient and standard deviation parameters. The MCLL standard errors were slightly larger than the ML standard errors although the differences were negligible. Compared with the posterior mean estimates, the MCLL estimates were closer to the ML estimates for the two regression coefficient parameters β_1, β_3, and both standard deviations σ_A, and σ_E. The standard errors were slightly larger than the posterior standard deviations.

3.4.4 Longitudinal Data on Self-esteem

The third example is the linear crossed random effects model proposed by Jeon & Rabe-Hesketh (2012) to investigate Korean students' growth in self-esteem. The data were taken from the Korea Youth Panel Survey (KYPS; Lee et al., 2010) from 2003 to 2006 where students were in middle school in the first two waves and in high school in the last two waves. There were 3,281 students in 104 middle schools at waves 1 and 2 and 2,924 students followed up after dispersing into 860 high schools at waves 3 and 4.

About 2.7% students switched their school membership during the middle school or high school years and these students were excluded from the data for simplicity. The response variable was self-esteem which is a mean-composite variable computed from six 5-point Likert-scale items. The mean (standard deviation) of self-esteem was 3.16 (0.62), 3.26 (0.62), 3.31 (0.60), 3.33 (0.61) at waves 1, 2, 3, and 4, respectively. The internal consistency of the measures (Cronbach's α) was on average 0.734.

We formulated a three-level linear mixed model with crossed random effects for the self-esteem, Y_{tsmh} at time t for student s who attended middle school m and high school h

$$Y_{tsmh} = \beta_1 + \beta_2 \text{time2} + \beta_3 \text{time3} + \beta_4 \text{time4} + \delta_s + \delta_m \mu_t + \delta_h \eta_t + e_{tsmh}, \tag{3.15}$$

where β_1 is an intercept and β_2, β_3, and β_4 are coefficients for time 2, 3, and 4 dummy variables. The random part of the model consists of a student-level random effect $\delta_s \sim N(0, \sigma_s^2)$, a middle school random effect $\delta_m \sim N(0, \sigma_m^2)$, a high school random effect $\delta_h \sim N(0, \sigma_h^2)$, and a time- and student-specific residual $e_{tsmh} \sim N(0, \sigma_e^2)$. The model contains occasion-specific weights, $\boldsymbol{\mu} = (1, 1, 1, 1)'$ and $\boldsymbol{\eta} = (0, 0, 1, 1)'$ that represent the contribution of school effects on student outcomes at each time point. η_1 and η_2 were set to zero because the future high school is assumed not to affect students while they are still in middle school. Jeon & Rabe-Hesketh (2012) considered $\boldsymbol{\mu} = (1, \mu_2, \mu_3, \mu_4)'$ and $\boldsymbol{\eta} = (0, 0, 1, \eta_4)'$ as model parameters, but here we simplified the model by treating them as fixed for illustration purposes.

To implement the MCLL method, the same settings were used as in the first two exam-

ples. Diffuse normal priors were specified for the regression coefficient parameters (mean 0, standard deviation 100) and for the log standard deviations, logσ_s, logσ_m, logσ_h, and logσ_e (mean -4.37, standard deviation 2.35). The MCLL estimates and standard errors (using the diagonal Hessian matrix) were compared with the posterior means and standard deviations and with the true ML estimates which were obtained from the `xtmixed` function in `Stata`.

Results

Table 3.3 lists the results for model (3.15) to the KYPS dataset.

Table 3.3: Parameter estimates (Est) and standard errors (SE) for the Korea Youth Panel Survey (KYPS) data. MLE is the true maximum likelihood estimates and Post.m is the posterior mean estimates.

	MLE		Post.m		MCLL	
	Est	SE	Est	SE	Est	SE
β_1	3.16	0.02	3.16	0.01	3.16	0.02
β_2	0.11	0.01	0.11	0.01	0.11	0.01
β_3	0.16	0.01	0.16	0.01	0.16	0.01
β_4	0.18	0.02	0.18	0.01	0.17	0.02
σ_s	0.46	-	0.46	-	0.46	-
σ_m	0.38	-	0.38	-	0.38	-
σ_h	0.09	-	0.09	-	0.09	-
σ_e	0.12	-	0.12	-	0.12	-

For regression coefficients, $\hat{\beta}_1$ indicates the estimated mean self-esteem of students at wave 1 (second grade in middle school). The coefficients for the time dummy variables, $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ represent the estimated differences in the mean self-esteem between each wave and wave 1. The mean growth from waves 1 and 2 was estimated as 0.11, from waves 2 and 3 was 0.05, and from waves 3 and 4 as 0.02. Students' self esteem tended to increase, but the rate of the growth decreased over time. In the random part, the estimated within-student and the estimated between-student standard deviations ($\hat{\sigma}_e$, $\hat{\sigma}_s$) were larger than the between-school standard deviations ($\hat{\sigma}_m$, $\hat{\sigma}_h$). All the regression coefficient parameter estimates and standard errors, and the standard deviation estimates from the MCLL method were close to the ML estimates. There was little difference between the MCLL estimates and posterior means and standard deviations in this example.

3.5 Simulation Studies

Two simulation studies were conducted to evaluate the performance of the MCLL method using 1) simulated salamander mating data for a generalized linear mixed model with crossed random effects and 2) simulated birth weight data for a linear mixed model. A linear mixed model is considered to evaluate the MCLL method when the true ML estimates are available.

3.5.1 Simulation Design

The first example was closely related to the salamander mating data in Section 3.4.1. 100 datasets were generated based on model (3.13) using the same true parameter values considered by other researchers (e.g., Lin & Breslow, 1996), which are $\boldsymbol{\beta} = (1.06, -3.05, -0.72, 3.77)'$ and $(\sigma_f^2, \sigma_m^2)' = (.50, .50)'$. The second example was related to the birth weight data in Section 3.4.3. 100 datasets were generated based on model (3.14) using the ML estimates of the original data as true values, $\boldsymbol{\beta} = (3368.09, 155.34, 126.94, 213.43)'$ and $(\sigma_A, \sigma_E)' = (311.21, 374.66)'$. To implement the MCLL method, the same settings were used as in the corresponding empirical studies.

In addition, Monte Carlo error (MCE) involved in all simulation estimates were estimated. Based on Koehler et al. (2009), MCE for the mean of the estimates ($\hat{\beta}$) can be defined as

$$\widehat{\text{MCE}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}(b) - \hat{\beta}(\cdot))^2},$$

where $\hat{\beta}(b)$ is the estimate at the bth simulated data and $\hat{\beta}(\cdot)$ is the mean of the estimates of the B replicates $\hat{\beta}(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}(b)$.

`Stata` command `simsum` (White, 2010) was used to compute the MCE for the means of the parameter and standard error estimates in the simulation studies.

3.5.2 Results

Table 3.4 lists the estimated bias and mean squared error (MSE) for the first simulation study mimicking the salamander mating dataset.

The MCLL method performed well compared with the Bayesian and the Laplace approximation methods. For the regression coefficient estimates, the bias and MSE were quite similar between the methods. For the standard deviations, however, the MCLL method showed smaller bias and MSE than the other two methods.

Table 3.5 lists the average standard error estimates compared with the standard deviations of the parameter estimates (or the empirical standard errors) for the regression coefficient parameters.

Table 3.4: Bias and mean squared error (MSE) of the MCLL, Laplace approximation, and posterior mean (Post.m) estimates for 100 simulated salamander datasets.

		Bias			MSE		
	True	Laplace	Post.m	MCLL	Laplace	Post.m	MCLL
β_1	1.06	-0.03	-0.03	-0.03	0.12	0.12	0.12
β_2	-3.05	0.04	0.01	0.01	0.22	0.21	0.21
β_3	-0.72	0.06	0.06	0.05	0.17	0.17	0.16
β_4	3.77	-0.04	-0.01	-0.04	0.35	0.33	0.35
σ_m	0.71	-0.17	-0.21	-0.13	0.10	0.09	0.07
σ_f	0.71	-0.15	-0.19	-0.12	0.12	0.09	0.08

Table 3.5: Average standard error estimates for 100 simulated salamander datasets. SD is the empirical standard error (standard deviation of the parameter estimates), SE is the average of the standard error estimates, and SE/SD is the ratio of SE to SD.

	MCLL			Laplace		
	SE	SD	SE/SD	SE	SD	SE/SD
β_1	0.31	0.34	0.91	0.29	0.34	0.85
β_2	0.50	0.46	1.09	0.44	0.46	0.96
β_3	0.40	0.40	1.00	0.37	0.41	0.90
β_4	0.61	0.59	1.03	0.53	0.59	0.90

The results show that the means of the standard error estimates over replicates were quite close to the empirical standard errors for all methods. Our standard error estimates tend to be more conservative than those from the Laplace approximation. Monte Carlo errors (MCEs) were about 10% of the means of the parameter estimates and less than 0.1 for the means of the standard errors in all methods.

For the linear mixed model example with the simulated birth weight datasets, we compared the MCLL and Bayesian estimates (posterior means) with the ML estimates. Figure 3.2 compares the distances from the ML estimates between the two methods for each parameter.

Figure 3.2 shows that the MCLL estimates are closer to the true ML estimates than the posterior mean estimates. In particular, the posterior mean estimates display a marked bias (defined relative to the ML estimates), which is evident in the point clouds being shifted away from zero on the x-axis. The second step of the MCLL algorithm adjusts the estimates and we observe that they are no longer biased relative to the ML estimates. This is a strong evidence that MCLL estimates are closer to the ML estimates than the posterior mean estimates. The MCEs were about 10% of the means of the parameter estimates for all methods.

Table 3.6 compares the average standard error estimates with the empirical standard errors.

Table 3.6: Average standard error estimates for 100 simulated birth weight datasets. SD is the empirical standard error (standard deviation of the parameter estimates), SE is the average of the standard error estimates, and SE/SD is the ratio of SE to SD.

	MCLL			ML		
	SE	SD	SE/SD	SE	SD	SE/SD
β_1	32.22	31.70	1.02	31.11	31.49	0.99
β_2	17.77	15.35	1.16	17.52	15.31	1.14
β_3	31.89	32.00	1.00	30.94	32.00	0.97
β_4	53.92	58.05	0.93	52.58	57.99	0.91

Table 3.6 shows that both the ML and MCLL standard errors are good approximations to the empirical standard deviations. As in the first simulation example, the MCLL standard error estimates tend to be a bit more conservative than the ML standard errors. The MCEs for the means of the standard error estimates were less than 0.05 for MCLL, less than 0.10 for the ML method, and less than 0.08 for the Bayesian method.

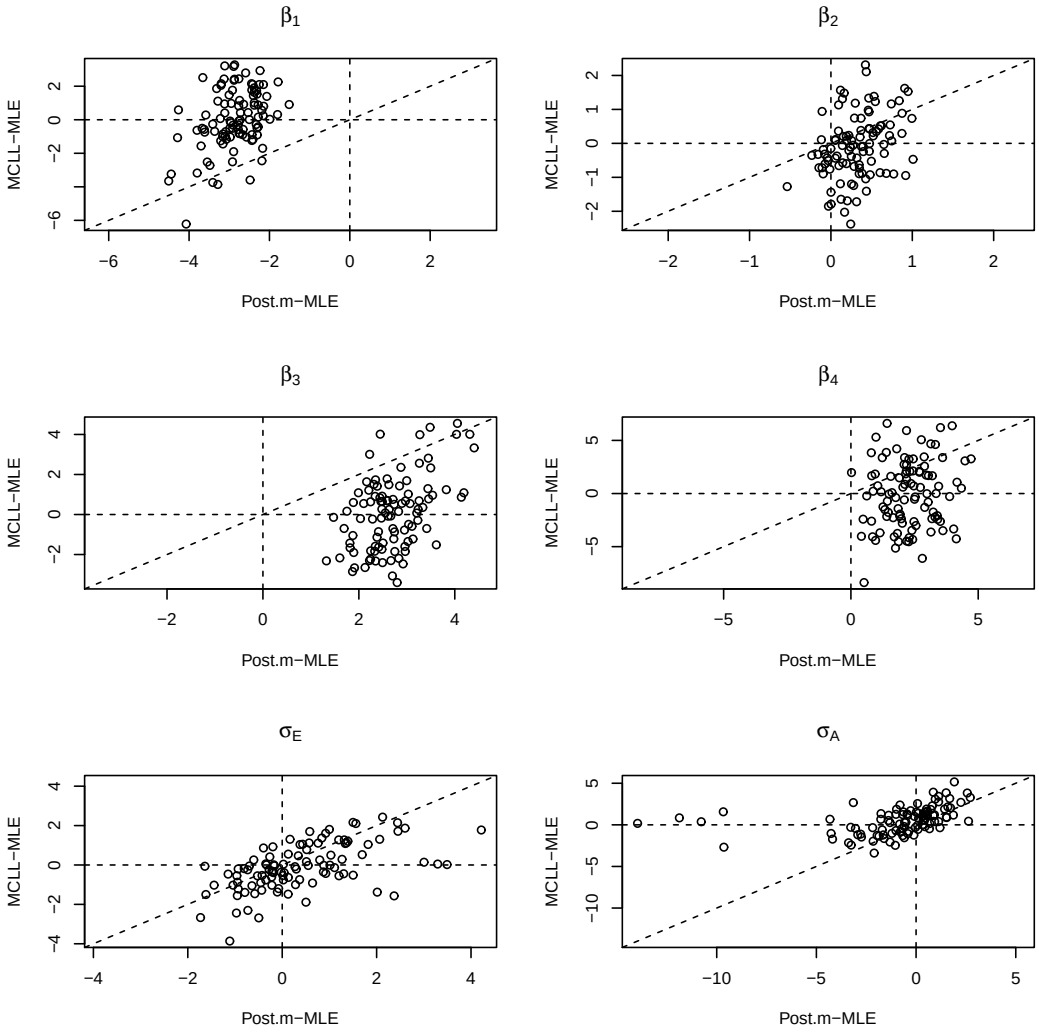

Figure 3.2: Distances from the ML estimates for MCLL estimates (MCLL-MLE) and for posterior mean estimates (Post.m-MLE) for 100 simulated birth weight datasets

3.6 Concluding Remarks

In this paper, the Monte Carlo local likelihood (MCLL) method was proposed for maximum likelihood estimation of GLMMs with crossed random effects. The MCLL method initially treats the model parameters as random variables and samples them jointly with random effects, from the posterior distribution for a particular prior. The likelihood function is then approximated up to a constant as a local likelihood density estimate of the posterior divided by the prior.

The MCLL method is similar to the MC kernel likelihood method (MCKL; De Valpine, 2004), which uses kernel density estimation to approximate the posterior. The key advantage of MCLL is that it provides methods for obtaining standard errors whereas MCKL does not. MCLL is also less sensitive to bandwidth selection than MCKL.

De Valpine (2004) showed convergence of the MCKL estimator based on proofs of convergence of kernel mode estimates. The proofs for MCKL may not be directly applied to the MCLL method with a polynomial higher than degree zero. Unlike kernel density estimation, there has been no proof yet on the convergence of mode estimates in local density estimation, which would be an intermediate step in proving convergence of the MCLL estimator.

Finally, it is important to note that MCLL allows likelihood inference for any complex models for which ML estimation may be infeasible but MCMC methods are possible. For example, in addition to GLMMs with crossed random effects considered here, the MCLL algorithm could be used to fit state-space models with higher dimensional latent variables. Potential applications for MCLL are therefore far beyond the models discussed in this paper. We have shown that the MCLL method provides results close to the ML estimates. Even if informative priors are specified, MCLL provides estimates close to the ML estimates, whereas the posterior mean estimates could be quite different. When ML inference is desired for highly complex models, the MCLL method seems to be an effective and practical choice.

Chapter 4

Autoregressive IRT Growth Model

4.1 Introduction

This paper considers longitudinal data where a latent construct is measured by multiple items at multiple time points. In measuring psychological traits such as engagement or self-esteem, typically the same scales with the same set of items are used over time. In ability testing, a set of common items are often included in different tests for the purpose of vertical equating. Responses to the same items over time may not be conditionally independent given the latent trait.

When the measures of the latent construct are continuous, curve-of-factors models or second-order latent growth models are often used in structural equation modeling (SEM) (e.g., Hancock & Kuo, 2001; Sayer & Cumsille, 2001). A common strategy in such models is to deal with violations of conditional independence by allowing residuals for the same items to be correlated over time (Loehlin, 1998; Sayer & Cumsille, 2001). In econometrics, correlated errors have also been used in probit models (e.g., Hyslop, 1999; Varin & Czado, 2010).

In item response theory (IRT), many methods have been developed to deal with local dependence within tests. Testlet-type models were suggested which use additional dimensions to capture dependence within item bundles or testlets (e.g., Gibbons & Hedeker, 1992; Wilson & Adams, 1995; Bradlow et al., 1999; Wang & Wilson, 2005; Jeon et al., in press). Such approaches are computationally demanding in general because the number of latent variables required increases as the number of item clusters increases. Hoskens & De Boeck (1997) present a fixed effects approach using interaction parameters for within-test local dependence. Alternatively, marginal models have been proposed e.g., by utilizing copula functions to capture local dependence among items (e.g., Braeken et al., 2007; Braeken, 2011). However, these marginal methods appear to be difficult to implement in practice.

For longitudinal data, multidimensional models have typically been used in IRT without much consideration for serial dependence. See for example, Andersen (1985), Embretson

(1991), and McGuire (2010). Recently, Cai (2010) suggested a two-tier IRT model that uses additional latent variables (or dimensions) to take into account local dependence among item responses, and also discussed an application to a longitudinal setting. A combination of the multilevel model and the IRT model has also been used to analyze longitudinal data. For example, a one-parameter logistic (1PL) IRT measurement model was applied in three-level growth models for binary and categorical data (e.g., Fox, 2005; Pastor & Beretvas, 2006). Segawa (2005) presented a multilevel IRT model including a two-parameter logistic (2PL) IRT measurement model for ordinal responses. For categorical responses in SEM, Serrano (2010) presented a second-order model for binary item responses using extra latent variables to allow for autocorrelations between the responses over time. Eid & Hoffmann (1998) proposed a multistate-multitrait model that includes latent factors for serial correlations among ordinal responses.

In this paper, we present an autoregressive IRT growth model that takes into account serial dependence. Autoregressive or dynamic models for binary panel data have been actively investigated in econometrics (e.g., Heckman, 1981; Hsiao, 2003; Bartolucci & Nigro, 2010) but rarely in psychometrics or educational measurement. A dynamic Rasch model has been proposed by Verhelst & Glas (1993) but in a different context, to model learning effects throughout tests.

The autoregressive IRT growth model that we present here allows the current response to an item to depend on the previous response in addition to the latent trait. In the measurement model, the coefficients for the lagged responses allow to study state dependence, i.e., how the past response can influence the response to the same item in the future. We will show that this autoregressive model is equivalent to a model that includes interaction parameters for item responses at adjacent time points. The initial conditions problem needs to be addressed because initial outcomes have no lagged variables (see e.g., Heckman, 1981; Wooldridge, 2005). We adopt the treatment suggested by Heckman (1981) and Aitkin & Alfo (2003) to deal with the initial conditions problem.

A linear growth curve model is specified for the latent trait in the structural model. The full model can be estimated using standard maximum likelihood (ML) software. ML estimation of the proposed model involves only three-dimensional integrals and the dimensionality of the integrals stays the same regardless of both the number of time-points and the number of items.

The outline of this chapter is as follows: We first review how local dependence has been treated in IRT. In Sections 4.2 and 4.3, we present IRT models with interaction parameters to capture local dependence for cross-sectional data. An autoregressive IRT growth model is then introduced for longitudinal data. Equivalence of this model to an IRT model with interaction parameters is shown. In Section 4.4, we discuss the treatment of the initial conditions problem and its implications for measurement invariance. In Section 4.5, we investigate the consequences of ignoring serial dependence and the initial conditions problem using simulations. An empirical study is provided in Section 4.6 to illustrate the proposed model. We end with some concluding remarks.

4.2 Treatment of Local Dependence in IRT

IRT models are not robust to violations of local stochastic independence, called local item dependence or residual dependence (Tuerlinckx & De Boeck, 2001; Braeken et al., 2007). Local item dependence can seriously affect estimation of model parameters on both item and person sides, the test information function, and the diagnostics that assume conditional independence (see e.g., Yen, 1984; Sireci et al., 1991; Yen, 1993; Chen & Thissen, 1997; Tuerlinckx & De Boeck, 2001; Braeken et al., 2007).

Several methods have been suggested to deal with local dependence in IRT. Typically local dependence is violated for items nested in subtests (Andrich, 1985), testlets (Wainer & Kiely, 1987), or item bundles (Wilson & Adams, 1995; Rosenbaum, 1999). To deal with the local dependence, the sum scores of testlets can be used as polytomous items. Alternatively, additional latent variables (or dimension) can be introduced to capture the dependence within testlets (Gibbons & Hedeker, 1992; Bradlow et al., 1999; Wang & Wilson, 2005; Cai, 2010; Jeon et al., in press).

Hoskens & De Boeck (1997) present a fixed effects approach that directly models local dependence using interaction parameters. Drawbacks of this approach are first, the marginal item characteristic curves are not reproducible (Fitzmaurice et al., 1993), i.e., the curves are no longer logistic functions. Second, the item parameters lose their usual interpretations (Ip, 2002; Wang & Wilson, 2005; Braeken et al., 2007; Braeken, 2011).

To avoid these problems, marginal models have been proposed such as the Bahadur-Ip model (Ip, 2000, 2001), the hybrid kernel model (Ip, 2002), and copula models (Braeken et al., 2007; Braeken, 2011). The main idea is to keep marginal probabilities intact by accounting for local dependence by separate tools. However, these methods are difficult to implement in practice and require modeling choices to be made. For example, for copula models, users have to choose an appropriate copula function.

4.3 Local Dependence IRT Models with Interaction Parameters

A natural way of modeling dependence among correlated item responses is to include interaction parameters in the IRT model. This approach was suggested by Hoskens & De Boeck (1997) and Adams et al. (1997) to capture local dependence in cross-sectional data, but it can also be extended to a longitudinal setting. In this section, starting from the local dependence model within tests, a serial dependence model is introduced for longitudinal data.

4.3.1 Local Dependence IRT Model within Tests

A 2PL IRT model is considered as the basic model. Assuming local independence, the 2PL model specifies the conditional probability of binary response y_{is} for item i and person s

given ability θ_s as

$$\Pr(Y_{is} = y_{is}|\theta_s) = \frac{\exp\left[y_{is}(\alpha_i\theta_s - \beta_i)\right]}{1 + \exp\left[(\alpha_i\theta_s - \beta_i)\right]}, \tag{4.1}$$

where $\theta_s \sim N(0, \sigma^2)$, and β_i and α_i are the item intercept and discrimination parameters, respectively. The item difficulty is β_i/α_i. Under the local independence assumption, the joint probability for a particular realization of responses (y_{1s}, y_{2s}) to items 1 and 2 for person s can be written as

$$\Pr(Y_{1s} = y_{1s}, Y_{2s} = y_{2s}|\theta_s) = \frac{\exp\left[y_{1s}(\alpha_1\theta_s - \beta_1) + y_{2s}(\alpha_2\theta_s - \beta_2)\right]}{\sum_{\{d_1,d_2\}} \exp\left[d_1(\alpha_1\theta_s - \beta_1) + d_2(\alpha_2\theta_s - \beta_2)\right]}, \tag{4.2}$$

where the sum in the denominator is over all possible response patterns with (d_1, d_2) equal to (0,0), (1,0), (0,1), and (1,1).

To model local dependence, interaction parameters can be incorporated for locally dependent items. For example, model (4.2) can be extended as

$$\Pr(Y_{1s} = y_{1s}, Y_{2s} = y_{2s}|\theta_s) = \frac{\exp\left[y_{1s}(\alpha_1\theta_s - \beta_1) + y_{2s}(\alpha_2\theta_s - \beta_2) + y_{2s}y_{1s}(-\lambda_{21})\right]}{\sum_{\{d_1,d_2\}} \exp\left[d_1(\alpha_1\theta_s - \beta_1) + d_2(\alpha_2\theta_s - \beta_2) + d_2d_1(-\lambda_{21})\right]}, \tag{4.3}$$

where λ_{21} is the parameter that quantifies the interaction between items 1 and 2. Note that in model (4.3), the marginal probability for y_{is} given θ_s is not reproducible (not the inverse logit function as in (4.1)), and α_i and β_i lose their usual interpretations as item discrimination and intercept parameters if $\lambda_{21} \neq 0$ (Braeken et al., 2007; Braeken, 2011).

4.3.2 Serial Dependence IRT Model for Longitudinal Data

The local dependence model in Section 4.3.1 can be extended to longitudinal settings to capture serial dependence. With longitudinal data, local dependence arises among the responses to the same items used repeatedly over time. Let the response pattern for item i and person s across T occasions be denoted $\mathbf{y}_{is} \equiv (y_{1is}, y_{2is}, ..., y_{Tis})'$, where y_{tis} is the response to item i at occasion t, $t = 1, \ldots, T$. Then the probability for $\mathbf{y}_{is}$ can be modeled as

$$\Pr(\mathbf{y}_{is}|\boldsymbol{\theta}_s) = \frac{\exp\left[-\beta_i \sum_{t=1}^{T} y_{tis} + \alpha_i \sum_{t=1}^{T} \theta_{ts} y_{tis} - \sum_{t=1}^{T} \lambda_i y_{tis} y_{(t-1)is}\right]}{\sum_{\{d\}} \exp\left[-\beta_i \sum_{t=1}^{T} d_t + \alpha_i \sum_{t=1}^{T} \theta_{ts} d_t - \sum_{t=1}^{T} \lambda_i d_t d_{t-1}\right]}, \tag{4.4}$$

where $\boldsymbol{\theta}_s = (\theta_{1s}, ..., \theta_{Ts})'$ is the vector of latent traits across time for person s and $\{d\}$ indicates the set of all possible response patterns. λ_i is an interaction parameter for the responses to item i between at adjacent occasions t and $t-1$. We assume that λ_i, β_i and α_i are constant across time for item i. Here y_{0is} does not exist and is set to 0 (See Section 4.5.1 on the initial conditions problem).

4.4 Autoregressive IRT Growth Model

We now introduce a first-order autoregressive IRT growth model for longitudinal analysis. Equivalence of this model to the serial dependence IRT model is shown.

4.4.1 Measurement Model

The measurement model corresponds to a first-order autoregressive or dynamic 2PL model. The conditional probability for binary response y_{tis} at time t for item i and person s can be written as

$$\text{logit}(\Pr(y_{tis} = 1 | y_{(t-1)is}; \theta_{ts})) = \alpha_i \theta_{ts} - \beta_i + \lambda_i y_{(t-1)is}, \tag{4.5}$$

where λ_i is the lag parameter for state dependence and the lagged variable $y_{(t-1)is}$ for item i.

It is useful to note that

$$\log \frac{\Pr(y_{tis} = 1 | \theta_{ts}, y_{(t-1)is} = 1) / \Pr(y_{tis} = 0 | \theta_{ts}, y_{(t-1)is} = 1)}{\Pr(y_{tis} = 1 | \theta_{ts}, y_{(t-1)is} = 0) / \Pr(y_{tis} = 0 | \theta_{ts}, y_{(t-1)is} = 0)} = \lambda_i,$$

when time $t > 1$. That is, the lag parameter λ_i is the log-odds ratio for current responses due to the previous responses changing from 0 to 1 (see e.g., Bartolucci & Nigro, 2010).

The dynamic Rasch model (Verhelst & Glas, 1993) was presented in IRT in the context of modeling learning effects. Instead of lagged responses, the cumulative number of correct responses preceding the item in question was used where the effect of the cumulative sum was considered as a learning effect induced by previous successes (De Boeck et al., 2011). This model was later extended by Verguts & De Boeck (2000), allowing for a different learning rate for each person.

In order to show equivalence of model (4.5) to the serial dependence model in (4.4), rewrite model (4.5) using a log-linear formulation. For example, at time 2

$$\Pr(y_{2is} | y_{1is}; \theta_{2s}) = \frac{\exp\left[y_{2is}(\alpha_i \theta_{2s} - \beta_i) - \lambda_i y_{2is} y_{1is}\right]}{\sum_{\{d\}} \exp\left[d_2(\alpha_i \theta_{2s} - \beta_i) - \lambda_i d_2 d_1\right]}.$$

Similarly, write model (4.5) at time 3

$$\Pr(y_{3is}|y_{2is};\theta_{3s}) = \frac{\exp\left[y_{3is}(\alpha_i\theta_{3s}-\beta_i)-\lambda_i y_{3is}y_{2is}\right]}{\sum_{\{d\}}\exp\left[d_3(\alpha_i\theta_{3s}-\beta_i)-\lambda_i d_3 d_2\right]}.$$

The joint conditional probability of the pair of the responses (y_{2is}, y_{3is}) at time 2 and 3 can then be written as

$$\Pr(y_{2is},y_{3is}|y_{1is};\theta_{ts}) = \frac{\exp\left[y_{2is}(\alpha_i\theta_{2s}-\beta_i)+y_{3is}(\alpha_i\theta_{3s}-\beta_i)-\lambda_i y_{2is}y_{1is}-\lambda_i y_{3is}y_{2is}\right]}{\sum_{\{d\}}\exp\left[d_2(\alpha_i\theta_{2s}-\beta_i)+d_3(\alpha_i\theta_{3s}-\beta_i)-\lambda_i d_2 d_1-\lambda_i d_3 d_2\right]}.$$

Notice that if we set $y_{0is}=0$ so that

$$\Pr(y_{1is}|\theta_{1s}) = \frac{\exp\left[y_{1is}(\alpha_i\theta_{1s}-\beta_i)\right]}{\sum_{\{d\}}\exp\left[d_1(\alpha_i\theta_{1s}-\beta_i)\right]},$$

we obtain the joint probability for $\mathbf{y}_{is}\equiv(y_{1is},y_{2is},y_{3is})$

$$\Pr(\mathbf{y}_{is}|\boldsymbol{\theta}_s) = \frac{\exp\left[-\beta_i\sum_{t=1}^{3}y_{tis}+\alpha_i\sum_{t=1}^{3}\theta_{ts}y_{tis}-\sum_{t=1}^{3}\lambda_i y_{tis}y_{(t-1)is}\right]}{\sum_{\{d\}}\exp\left[-\beta_i\sum_{t=1}^{T}d_t+\alpha_i\sum_{t=1}^{T}\theta_{ts}d_t-\sum_{t=1}^{T}\lambda_i d_t d_{t-1}\right]},$$

which is the serial dependence model in (4.4) with $T=3$.

4.4.2 Structural Model

In the measurement model (4.5), θ_{ts} is the latent trait for person s at occasion t. Andersen (1985) specifies a longitudinal IRT model with

$$\boldsymbol{\theta}_s = (\theta_{1s},\theta_{2s},...,\theta_{Ts})',$$

where $\boldsymbol{\theta}_s \sim N(\mathbf{0},\Sigma)$. Because Andersen's model does not contain change parameters, Embretson (1991) suggests specifying

$$\theta_{ts} = \sum_{r=1}^{t}\theta'_{rs},$$

where changes θ'_{rs} from the previous status to time r are modeled for person s, and $\boldsymbol{\theta}'_s = (\theta'_{1s}, \theta'_{2s}, ..., \theta'_{Ts})'$ with $\boldsymbol{\theta}'_s \sim N(\mathbf{0}, \Sigma)$.

However, Andersen's and Embretson's models can be computationally demanding because they require an increasing number of latent variables as the number of time points increases. McGuire (2010) proposed a simplified version of Embretson's model, which requires only two latent variables, one for the baseline and the other for the growth factor, by specifying

$$\theta_{ts} = \delta_{s1} + (b_1 + \delta_{s2})\text{time}_t, \tag{4.6}$$

where b_1 is the mean slope or the mean growth rate and time_t is the time associated with occasion t. δ_{s1}, the random intercept (or initial status) and δ_{s2}, the random slope (or growth rate) for person s are assumed to have a bivariate normal distribution

$$\begin{pmatrix} \delta_{s1} \\ \delta_{s2} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{s1} & \sigma_{s12} \\ \sigma_{s21} & \sigma^2_{s2} \end{bmatrix} \right).$$

This formulation relies on a strong assumption that the latent trait for each person follows a perfect straight line trajectory (or a higher-order polynomial if powers of time are added to (4.6)). In addition, it does not allow for a random influence on the latent trait (or a deviation from the line trajectory) at each time point. We extend this model by allowing for an individual and time-specific error term ϵ_{ts}. This extension with a 1PL measurement model was presented in Pastor & Beretvas (2006) among others. The structural model can be specified as

$$\theta_{ts} = \delta_{s1} + (b_1 + \delta_{s2})\text{time}_t + \epsilon_{ts}, \tag{4.7}$$

where $\epsilon_{ts} \sim N(0, \sigma^2_\epsilon)$. Specifying different time-specific residual variances $\sigma^2_{\epsilon_t}$ corresponds to weak factorial invariance. Specifying constant time-specific residual variances σ^2_ϵ corresponds to strict factorial invariance (for more information, see, Meredith, 1993).

Note that in this formulation, only three-dimensional integrals are required regardless of the number of time points and items. This gives us computational advantages over previous approaches such as: 1) Anderson's and Embretson's IRT growth models where the number of latent variables grows as the number of time points increases, and 2) the random effects approaches for handling serial dependence (e.g., Serrano, 2010) that require an increasing number of latent variables as the number of items increases. Cai (2010) showed how to reduce the number of latent variables for his two-tier model, but it still requires more latent variables as the number of time points increases.

The full model is obtained by combining the measurement model in (4.5) with the structural model in (4.7). Figure 4.1 illustrates the model for person s assuming I items at each of four time points.

In the figure, the frame represents person s, ovals latent variables, rectangles observed variables, and arrows connecting latent and/or observed variables represent regression relations. The double-headed curved arrows between the observed variables represent pairwise

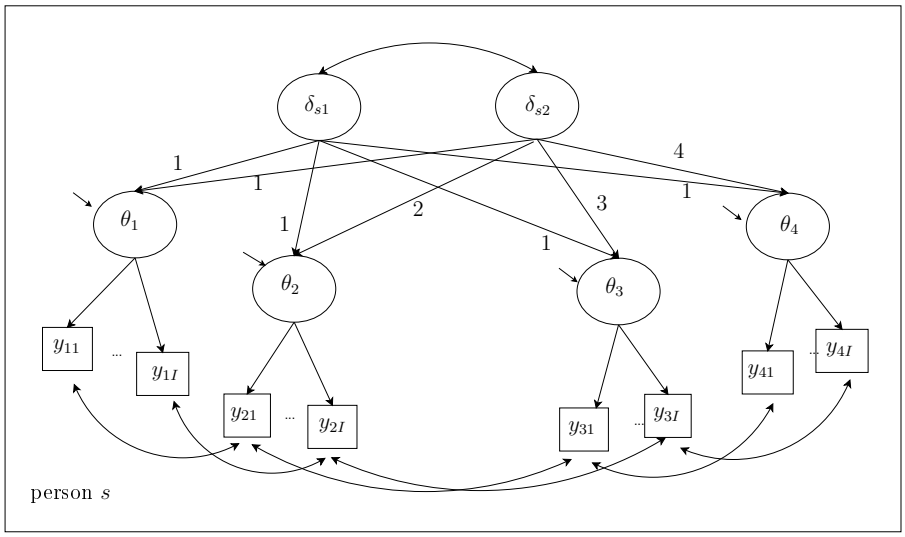

Figure 4.1: A serial dependence linear growth model with the random intercept, random slope, and time-specific random effects

interactions between adjacent time points. θ_1 to θ_4 represent the latent trait at each of the four time points, measured by the same I items. The short arrows pointing at the latent traits at each time point indicate time-specific random effects ϵ_{ts}, and the ovals δ_{s1} and δ_{s2} represent the random intercept (or initial status) and the random slope (or growth rate), respectively. The double-headed curved arrow between δ_{s1} and δ_{s2} represents the covariance σ_{s12}. The values associated with the arrows pointing at θ_1 to θ_4 indicate factor loadings $(1, 1, 1, 1)'$ for δ_{s1} and $(1, 2, 3, 4)'$ (for a linear growth) for δ_{s2} in this example.

4.5 Treatment of the Initial Conditions Problem

The initial conditions problem is an important theoretical and practical problem in dynamic models (Wooldridge, 2005). The consequences of ignoring the initial conditions problem have been studied in detail in econometrics (e.g., Anderson & Hsiao, 1981; Heckman, 1981; Wooldridge, 2005). For example, simply dropping the first outcome from the analyzed data produces inconsistent estimates (Hsiao, 1986; Fotouhi & Davies, 1997; Aitkin & Alfo, 1998).

In this section, the initial conditions problem is illustrated and the treatment of the initial conditions problem is discussed for the proposed autoregressive IRT growth model.

4.5.1 Initial Conditions Problem

To begin with, first write down the autoregressive model (4.5) at time 1

$$\text{logit}(\Pr(y_{1is} = 1|y_{0is}; \theta_{1s})) = \alpha_i\theta_{1s} - \beta_i + \lambda_i y_{0is}.$$

The initial conditions problem is that the lagged response y_{0is} does not exist for the initial outcome y_{1is}. There are two simple options to deal with this problem: First, to treat y_{0is} as missing so that y_{1is} is not modeled as a response variable. This leads to an endogeneity problem because the association between y_{1is} and θ_{ts} is not modeled. All the association between y_{2is} and y_{1is} will be attributed to λ_i, whereas some of the association is due to the correlation between θ_{2s} and θ_{1s}. Consequently λ_i will be over-estimated particularly when there are few time points (T is small) because the first time point will have a larger impact. The second method is to set $y_{0is} = 0$. The problem with this approach is that the model for y_{tis} is conditional on the previous response when time $t > 1$, but is marginal (with respect to the hypothetical previous response) at time $t = 1$. It does not make sense to assume that the parameters β_i and α_i are the same in the conditional and marginal models.

Aitkin & Alfo (2003) suggested specifying an approximate model for the marginal (not conditional on the previous response) probability of the initial outcome given the latent trait. Heckman (1981) also proposed a similar method that approximates the distribution of the first outcome. Following Aitkin & Alfo (2003), model (4.5) at time 1 can be formulated as

$$\text{logit}(\Pr(y_{1is} = 1|\theta_{1s})) = \alpha_i'\theta_{1s} - \beta_i'. \tag{4.8}$$

For time $t > 1$, we retain the model

$$\text{logit}(\Pr(y_{tis}|y_{(t-1)is}; \theta_{ts})) = \alpha_i\theta_{ts} - \beta_i + \lambda_i y_{(t-1)is}. \tag{4.9}$$

It is important to allow $\beta_i' \neq \beta_i$ and $\alpha_i' \neq \alpha_i$ in models (4.8) and (4.9). Note therefore that α_i' is still needed for time 1 even if the model is otherwise a 1PL model (with $\alpha_i = 1$).

The joint probability of the item responses given the latent trait can then be written as

$$\Pr(\mathbf{y}_s|\theta_{ts}) = \prod_{i=1}^{I} \Pr(y_{1is}|\theta_{1s}) \prod_{t=2}^{T} \Pr(y_{tis}|y_{(t-1)is}; \theta_{ts}).$$

As an alternative solution to the initial conditions problem, Wooldridge (2005) considered the distribution of the latent variable, conditional on the initial response

$$\theta_{ts} = \gamma y_{1is} + \theta_{ts}', \tag{4.10}$$

where θ_{ts}' is uncorrelated with the initial response y_{1is} in (4.10). The full model conditional

on y_{1is} then becomes

$$\text{logit}(\Pr(y_{tis} = 1|y_{(t-1)is}; \theta'_{ts}, y_{1is})) = \alpha_i\theta'_{ts} - \beta_i + \alpha_i\gamma y_{1is} + \lambda_i y_i,$$

when time $t > 1$. Note that θ'_{ts} is different from the original latent variable θ_{ts}.

Figure 4.2 visualizes the differences between Aitkin and Alfo's and Wooldridge's approaches using a simple example with one item at four time points in a unidimensional model with one latent variable.

In the figures, y_{11} to y_{41} indicate the responses to item 1 and x_{11} to x_{41} represent the times associated with the measurement occasions 1 to 4. In Aitkin and Alfo's model in the upper panel, the item parameters β'_1 and α'_1 at time 1 are different from β_1 and α_1 at later time points. Wooldridge's model in the lower panel contains an arrow from y_{11} to θ with coefficient γ.

We adopt Aitkin and Alfo's approach to deal with the initial conditions problem in the proposed model. With this treatment, the latent trait vector $\boldsymbol{\theta}_s$ can be left intact in both marginal and conditional models (with respect to the lagged responses) both at time 1 and time $t > 1$. Thus, when the main interest is on modeling growth of the latent trait, Aitkin and Alfo's approach is preferable to Wooldridge's method that specifies the distribution of the latent trait conditional on the initial response.

4.5.2 Identification and Measurement Invariance

In the proposed model, the item parameters at time 1 are allowed to be different from those at later time points ($\beta'_i \neq \beta_i$ and $\alpha'_i \neq \alpha_i$). For model identification, δ_{s1}, δ_{s2}, and ϵ_{ts} are set to have mean zero, and α_i for the first item is fixed to 1. For measurement invariance, the item parameters α_i and β_i are set equal when time $t > 1$. If the lag parameters of one or more items are set to zero, these items serve as anchor items, allowing analysis of change in $\boldsymbol{\theta}_s$ from time 1 to time 2. An iterative procedure may be used to find anchor items, similarly to item purification procedures for finding anchor items in detecting differential item functioning (Rogers & Swaminathan, 1993; Zumbo, 1999). Even if $\lambda_i \neq 0$ for all items, however, the model is still identified because linearity is assumed for the mean of $\boldsymbol{\theta}_s$ over time.

At a glance, allowing for $\beta'_i \neq \beta_i$ and $\alpha'_i \neq \alpha_i$ at time 1 may look like violating the measurement invariance assumption for longitudinal item analysis. Recall that at time 1, however, the model is marginal, i.e., it does not include the lagged responses, whereas at later time points the model is conditional on the lagged responses. Therefore, imposing the same item parameters at time 1 and at later time points actually forces the item characteristic curves (ICCs) to be different across time, which is a violation of measurement invariance (Mellenbergh, 1989; Meredith & Millsap, 1992; Millsap, 2010).

The marginal probabilities $\Pr(y_{tis}|\theta_{ts})$ (or ICCs) are no longer logistic curves at time $t > 1$ if $\lambda_i \neq 0$. We can still compute the marginal probability for person s to binary item i

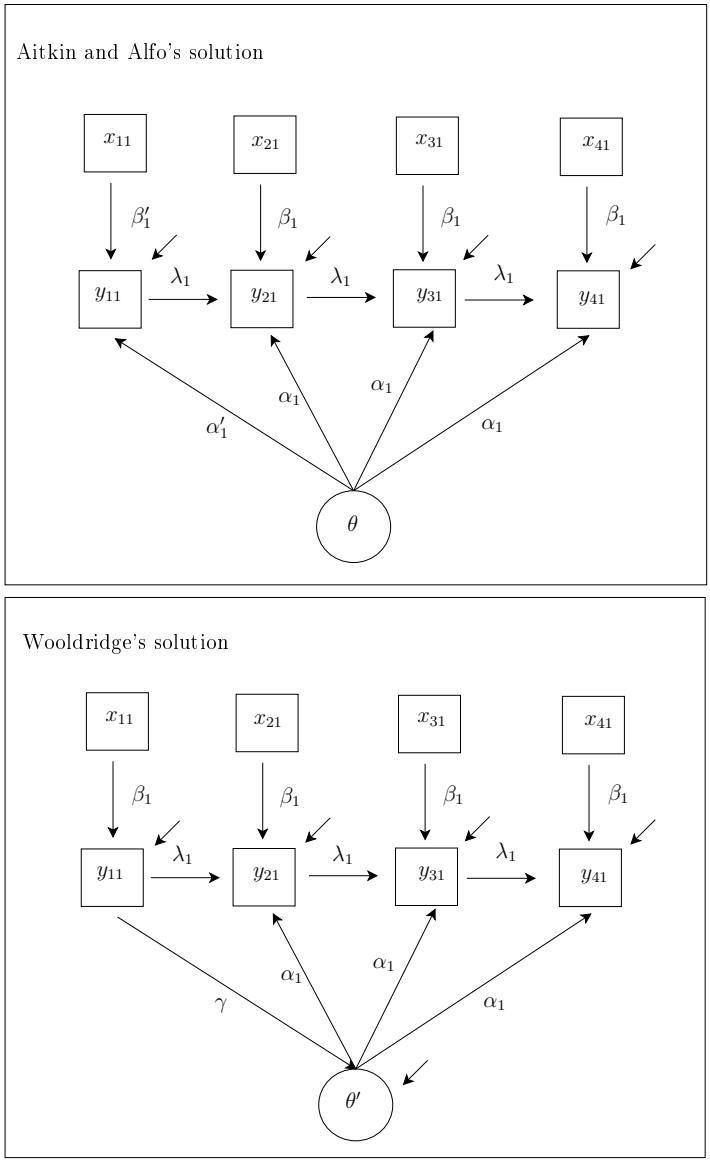

Figure 4.2: Solutions to the initial conditions problem by Aitkin & Alfo (2003) and Wooldridge (2005)

at time t recursively for $t = 2, 3, \ldots, T$ as

$$\Pr(y_{tis}|\theta_{ts}) = \sum_{\{d=0,1\}} \Pr(y_{tis}|\theta_{ts}, y_{(t-1)is} = d)\Pr(y_{(t-1)is} = d|\theta_{(t-1)s}).$$

That is, it is the sum of the probabilities of all possible sequences of responses to item i prior to time t when $t > 1$. Note that having free parameters at time 1 allows the logistic curve to be close to the not-quite-logistic curves at $t > 1$. Assuming constant λ_i at time $t > 1$ ensures that the curves are equivalent when $t > 1$.

4.6 Simulation Study

Instead of simulating datasets for a finite number of persons, we generate "population" data by computing the response probabilities for all possible response patterns and using them to weight the log-likelihood contributions of the response patterns for maximum likelihood estimation. A similar approach was used by Rotnitzky & Wypij (1994) and Heagerty & Kurland (2001).

We investigate the asymptotic bias of the maximum likelihood estimators using the population data when the model is incorrectly specified by 1) ignoring serial dependence, and 2) ignoring the initial conditions problem.

4.6.1 Generating Population Data

Suppose there are three binary items at three time points and hence $2^9 = 512$ response patterns in total. For each response pattern vector $\mathbf{y}_k$ $(k = 1, 2, ..., 2^9)$, we first obtain the response probability $\pi(\mathbf{y}_k) = g(\mathbf{y}_k; \boldsymbol{\psi}_0)$ under the true model with parameters $\boldsymbol{\psi}_0$. Given $\pi(\mathbf{y}_k)$ for all k, we treat the probabilities as frequency weights (possibly after multiplying by a large number and rounding to integers, but the software can handle non-integer frequency weights). Using the weights, we fit the model to the pseudo response vectors using the weighted log-likelihood. The maximum likelihood estimates of the specified model minimizes the Kullback-Leibler divergence between the true model $g(\mathbf{y}; \boldsymbol{\psi}_0)$ and the mis-specified, fitted model $f(\mathbf{y}; \boldsymbol{\psi})$

$$\mathrm{KL}\left(g(\mathbf{y}; \boldsymbol{\psi}_0), f(\mathbf{y}; \boldsymbol{\psi})\right) = E_g\left\{\log\frac{g(\mathbf{y}; \boldsymbol{\psi}_0)}{f(\mathbf{y}; \boldsymbol{\psi})}\right\}, \tag{4.11}$$

Let $\boldsymbol{\psi}^*$ be the ML estimates of the model parameters $\boldsymbol{\psi}$ for the population data. White (1982) shows that $\sqrt{\mathrm{N}}(\hat{\boldsymbol{\psi}}_{\mathrm{N}} - \boldsymbol{\psi}^*) \rightarrow N\left(0, A(\boldsymbol{\psi}^*)^{-1}B(\boldsymbol{\psi}^*)A(\boldsymbol{\psi}^*)^{-1}\right)$ where $A(\boldsymbol{\psi}^*)^{-1}B(\boldsymbol{\psi}^*)A(\boldsymbol{\psi}^*)^{-1}$

is the sandwich estimator applied to the population data and

$$A(\boldsymbol{\psi}^*) = \lim_N \frac{1}{N} \frac{\partial^2}{\partial \boldsymbol{\psi}^2} \log f(\mathbf{y}; \boldsymbol{\psi}) \Big|_{\boldsymbol{\psi}^*},$$
$$B(\boldsymbol{\psi}^*) = \lim_N \frac{1}{N} \sum_{s=1}^{N} \mathrm{Var}_{\boldsymbol{\psi}} \left\{ \frac{\partial}{\partial \boldsymbol{\psi}} \log f(\mathbf{y}_s; \boldsymbol{\psi}) \Big|_{\boldsymbol{\psi}^*} \right\},$$

where $\frac{\partial^2}{\partial \boldsymbol{\psi}^2} \log f(\mathbf{y}; \boldsymbol{\psi})$ is the Hessian of the marginal log-likelihood, and $\mathrm{Var}_{\boldsymbol{\psi}} \left\{ \frac{\partial}{\partial \boldsymbol{\psi}} \log f(\mathbf{y}_s; \boldsymbol{\psi}) \Big|_{\boldsymbol{\psi}^*} \right\}$ is the covariance matrix of the subject-specific contributions to the score vector. Hence, the sandwich estimator applied to the population data gives us the asymptotic sampling variance of the ML estimators for the mis-specified models. We do not need replicates, as in a conventional simulation study, to obtain sampling variances.

4.6.2 Simulation Design

To generate population data, we consider a simple example with three items at three time points. At time t, the measurement part for the generating model can be written as

$$\mathrm{logit}(\Pr(y_{tis} = 1 | y_{(t-1)is}; \theta_{ts})) = \alpha_i \theta_{ts} - \beta_i + \lambda_i y_{(t-1)is},$$

where $t = 1, 2, 3$ and $i = 1, 2, 3$. We assume that the lagged effect λ_1 for item 1 is the same across time, and that $\lambda_2 = 0$ and $\lambda_3 = 0$. We generate the lagged response y_{0is} by initially generating item responses at four time points (including at $t = 0$) with the same values for α_i, β_i, and λ_i.

The structural model is written as

$$\theta_{ts} = \delta_{s1} + (b_1 + \delta_{s2}) \mathrm{time}_t + \epsilon_{ts}, \tag{4.12}$$

where $\epsilon_{ts} \sim N(0, \sigma^2_\epsilon)$, and

$$\begin{pmatrix} \delta_{s1} \\ \delta_{s2} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{s1} & \sigma_{s12} \\ \sigma_{s21} & \sigma^2_{s2} \end{bmatrix} \right).$$

The true values for the model parameters are as follows:

- Item parameters $\boldsymbol{\alpha} = (1.0, 1.2, 0.8)'$ (α_1 is fixed)
- Item parameters $\boldsymbol{\beta} = (-1.0, 1.5, 0)'$
- Five different values for lag parameter $\lambda_1 = 0.2$, 0.4, 0.6, 0.8, and 1.0
- Mean slope $b_1 = 0.2$

- Variance parameters $\sigma_\epsilon = 0.2$, $\sigma_{s1} = 1.0$, $\sigma_{s2} = 0.5$, and $\sigma_{s12} = 0.0$

Using the generated population data, we estimate and compare the following models:

Model 1: Proposed model. The item parameters are allowed to be different at time 1 from those at $t > 1$ ($\beta_i' \neq \beta_i$ and $\alpha_i' \neq \alpha_i$).

Model 2: Independence model that ignores serial dependence. λ_i is not estimated.

Model 3: Constrained model that ignores the initial conditions problem. The item parameters are constrained to be the same across all time points ($\beta_i' = \beta_i$ and $\alpha_i' = \alpha_i$).

Figure 4.3 illustrates these three estimated models in addition to the data generating model for item 1.
In the figure, y_{01} to y_{31} indicate the responses to item 1 and x_{01} to x_{31} represent the times associated with occasions 0 to 3. β_1 and α_1 are the item parameters and λ_1 is the coefficient for the lagged response for item 1.

Note that in the proposed model, the measurement model at time 1 for item 1 is parameterized as

$$\text{logit}(\Pr(y_{1is} = 1; \theta_{1s})) = (\alpha_1 + \alpha_1^*)\theta_{1s} - (\beta_1 + \beta_1^*),$$

where α_1^* and β_1^* are the free parameters that represent the differences in α_1 and β_1 between at time 1 and $t > 1$. That is, the item parameters at time 1 are $\alpha_1' = \alpha_1 + \alpha_1^*$ and $\beta_1' = \beta_1 + \beta_1^*$. At time points 2 and 3, the item parameters are α_1 and β_1.

For simulation conditions, we consider different values for the autoregression coefficient $\lambda_1 = 0.2, 0.4, 0.6, 0.8$, and 1.0. Let ψ denote one of the model parameters, ψ^* the maximum likelihood estimates for the population data, and $\text{se}^*(\psi^*)$ and $\text{se}_R^*(\psi^*)$ denote the model-based and robust (sandwich estimator) standard errors for the population data when the weights add to 1 (if the weights add to N_{pop}, we multiply the standard errors by $\sqrt{N_{\text{pop}}}$). In each condition, we compute the asymptotic bias as $\psi^* - \psi_0$ and the asymptotic root mean squared error (RMSE) as $\sqrt{(\psi^* - \psi_0)^2 + (\text{se}_R^*(\psi^*)/\sqrt{N})^2}$ for sample size N. To see how well these asymptotic results hold in finite samples, we also simulate 200 datasets for N=200 and estimate model parameters and compute the finite-sample bias and RMSE in the usual way.

The software `gllamm` (Rabe-Hesketh et al., 2005) in `Stata` was used for the simulation study.

4.6.3 Power Calculation

We assess the power of the proposed model to detect the lagged effect λ_i. In principle, power can be estimated by carrying out a Monte Carlo study that records the proportion of replications in which the null hypothesis is rejected. The procedure can be tremendously simplified, however, by following a technique often used in SEM. One method, introduced

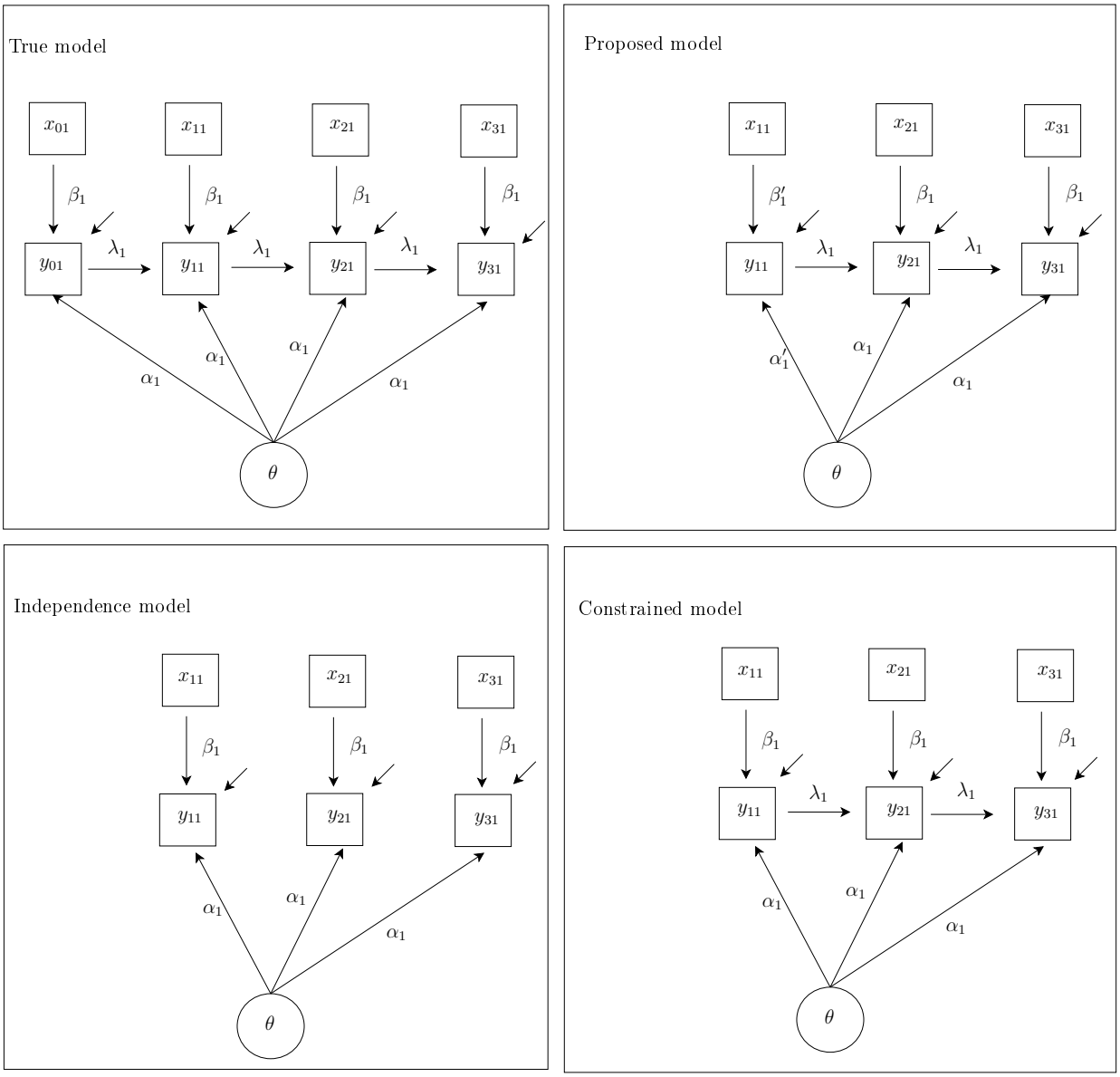

Figure 4.3: Data generating model and three estimated models

by Satorra & Saris (1985), is based on the fact that the likelihood ratio (LR) statistic has an asymptotic noncentral chi-square distribution $\chi^2(\omega, df)$ with degrees of freedom df and noncentrality parameter ω when the alternative model H_a is correct and the null model H_0 is tested, where H_0 corresponds to df constraints. The covariance matrix implied by the assumed model under H_a is used as population data and the models corresponding to H_a and H_0 are fit to this matrix. The corresponding LR statistic multiplied by N/N_{pop} is then used as noncentrality parameter ω, where N is the desired sample size and N_{pop} is the sample size specified for ML estimation. Instead of the LR statistic, Satorra & Saris (1985) also suggested using the Wald statistic for the model under H_a.

We adopt the Satorra-Saris method: To compute the power of the test of $H_0 : \lambda_{i_a} = \lambda_{i_0}$, we estimate the noncentrality parameter ω based on the Wald statistic with $df = 1$ (Bollen, 1989, p.338-349). Specifically, for sample size N

$$\omega = N \left(\frac{\lambda_i^*}{\text{se}^*(\lambda_i^*)} \right)^2,$$

where λ_i^* is the estimate obtained by fitting the H_a model to the population data generated under H_0, and $\text{se}^*(\lambda_i^*)$ is the asymptotic standard error when the weights for the population data add to 1. The asymptotic power of the test with significance level α is calculated as

$$\Pr\{\chi^2(\omega, df) > c_\alpha\},$$

where c_α is a critical point.

Asymptotic power is calculated for the different values of λ_i and a range of sample sizes (with 10 quadrature points). In order to assess how well the asymptotic power agrees with the finite-sample power, we also simulate 200 datasets for N=200 subjects and estimate the power based on the proportion of replicates where the null hypothesis is rejected in the likelihood ratio test (with df=1).

4.6.4 Results

Tables 4.1 to 4.5 list the parameter estimates, standard errors, robust standard errors, and log-likelihoods for the three models in each of the five simulation conditions. The standard errors are the asymptotic standard errors for sample size N=100, i.e., $\text{se}^*(\psi^*)/\sqrt{100}$.

The asymptotic bias of the proposed model is mostly zero or less than 0.01 across all conditions. The independence model that ignores serial dependence produces some degree of bias in most parameters. The size of bias appears relatively large for the mean slope b_1 and the standard deviations σ_ϵ, σ_{s1}, and σ_{s2} (the item parameters cannot be compared because their interpretation differs in the proposed model). The constrained model that ignores the initial conditions problem also produces some bias in most parameters. The bias appears large for the $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, λ_1, b_1, and σ_ϵ parameters, in particular. The asymptotic standard errors

Table 4.1: Population parameter estimates $\hat{\boldsymbol{\psi}}^*$ for condition 1 (λ_1=0.2). In the data generating model, $\beta_1^* = 0$ and $\alpha_1^* = 0$. ρ_{s12} is the correlation, $\frac{\sigma_{s12}}{\sigma_{s1}\sigma_{s2}}$.

Parameters	True	Proposed			Independence			Constrained		
		Est	SE	SE_R	Est	SE	SE_R	Est	SE	SE_R
β_1	-1.0	-1.00	0.31	0.31	-0.96	0.23	0.23	-0.97	0.22	0.22
β_1^*	-	0.06	0.41	0.41	-	-	-	-	-	-
β_2	1.5	1.50	0.28	0.28	1.49	0.26	0.26	1.51	0.28	0.28
β_3	0.0	0.00	0.18	0.18	0.00	0.17	0.17	0.01	0.17	0.17
α_1^*	-	0.02	0.64	0.64	-	-	-	-	-	-
α_2	1.2	1.20	0.55	0.55	1.09	0.36	0.37	1.18	0.48	0.48
α_3	0.8	0.80	0.35	0.35	0.73	0.24	0.24	0.79	0.31	0.31
λ_1	0.2	0.20	0.50	0.50	-	-	-	0.17	0.46	0.46
b_1	0.2	0.20	0.15	0.15	0.21	0.13	0.13	0.19	0.13	0.13
σ_ϵ	0.2	0.20	-	-	0.02	-	-	0.18	-	-
σ_{s1}	1.0	1.00	-	-	1.10	-	-	1.03	-	-
σ_{s2}	0.5	0.50	-	-	0.56	-	-	0.51	-	-
ρ_{s12}	0.0	0.00	-	-	0.00	-	-	0.00	-	-
Log-likelihood		-527.09			-527.16			-527.10		

Table 4.2: Population parameter estimates $\hat{\boldsymbol{\psi}}^*$ for condition 2 (λ_1=0.4). In the data generating model, $\beta_1^* = 0$ and $\alpha_1^* = 0$. ρ_{s12} is the correlation, $\frac{\sigma_{s12}}{\sigma_{s1}\sigma_{s2}}$.

Parameters	True	Proposed			Independence			Constrained		
		Est	SE	SE_R	Est	SE	SE_R	Est	SE	SE_R
β_1	-1.0	-1.00	0.32	0.32	-0.93	0.24	0.24	-0.93	0.22	0.22
β_1^*	-	0.12	0.41	0.41	-	-	-	-	-	-
β_2	1.5	1.50	0.29	0.28	1.48	0.25	0.25	1.52	0.28	0.28
β_3	0.0	0.00	0.18	0.18	0.01	0.17	0.17	0.02	0.17	0.17
α_1^*	-	0.04	0.65	0.65	-	-	-	-	-	-
α_2	1.2	1.20	0.55	0.55	0.98	0.33	0.35	1.16	0.47	0.48
α_3	0.8	0.80	0.35	0.35	0.66	0.22	0.23	0.78	0.31	0.31
λ_1	0.4	0.40	0.49	0.49	-	-	-	0.34	0.46	0.46
b_1	0.2	0.20	0.15	0.15	0.23	0.13	0.14	0.18	0.13	0.13
σ_ϵ	0.2	0.20	-	-	0.00	-	-	0.16	-	-
σ_{s1}	1.0	1.00	-	-	1.21	-	-	1.06	-	-
σ_{s2}	0.5	0.50	-	-	0.61	-	-	0.52	-	-
ρ_{s12}	0.0	0.00	-	-	0.00	-	-	0.00	-	-
Log-likelihood		-527.20			-527.51			-527.26		

Table 4.3: Population parameter estimates $\hat{\boldsymbol{\psi}}^*$ for condition 3 (λ_1=0.6). In the data generating model, $\beta_1^* = 0$ and $\alpha_1^* = 0$. ρ_{s12} is the correlation, $\frac{\sigma_{s12}}{\sigma_{s1}\sigma_{s2}}$.

Parameters	True	Proposed			Independence			Constrained		
		Est	SE	SE_R	Est	SE	SE_R	Est	SE	SE_R
β_1	-1.0	-1.00	0.32	0.32	-0.89	0.25	0.26	-0.89	0.22	0.22
β_1^*	-	0.17	0.40	0.40	-	-	-	-	-	-
β_2	1.5	1.50	0.29	0.29	1.47	0.25	0.25	1.53	0.28	0.28
β_3	0.0	0.00	0.18	0.18	0.01	0.17	0.17	0.03	0.17	0.17
α_1^*	-	0.02	0.66	0.65	-	-	-	-	-	-
α_2	1.2	1.20	0.56	0.56	0.88	0.36	0.37	1.14	0.47	0.48
α_3	0.8	0.80	0.36	0.36	0.59	0.24	0.24	0.77	0.31	0.31
λ_1	0.6	0.60	0.48	0.48	-	-	-	0.50	0.45	0.46
b_1	0.2	0.20	0.15	0.15	0.25	0.13	0.13	0.17	0.14	0.13
σ_ϵ	0.2	0.20	-	-	0.00	-	-	0.14	-	-
σ_{s1}	1.0	1.01	-	-	1.33	-	-	1.09	-	-
σ_{s2}	0.5	0.51	-	-	0.67	-	-	0.53	-	-
ρ_{s12}	0.0	0.00	-	-	0.00	-	-	0.00	-	-
Log-likelihood		-526.89			-527.55			-527.00		

Table 4.4: Population parameter estimates $\hat{\boldsymbol{\psi}}^*$ for condition 4 (λ_1=0.8). In the data generating model, $\beta_1^* = 0$ and $\alpha_1^* = 0$. ρ_{s12} is the correlation, $\frac{\sigma_{s12}}{\sigma_{s1}\sigma_{s2}}$.

Parameters	True	Proposed			Independence			Constrained		
		Est	SE	SE_R	Est	SE	SE_R	Est	SE	SE_R
β_1	-1.0	-1.00	0.32	0.32	-0.86	0.27	0.27	-0.86	0.22	0.22
β_1^*	-	0.23	0.40	0.40	-	-	-	-	-	-
β_2	1.5	1.50	0.29	0.29	1.46	0.24	0.24	1.54	0.28	0.28
β_3	0.0	0.00	0.18	0.18	0.01	0.16	0.16	0.03	0.17	0.17
α_1^*	-	0.07	0.66	0.66	-	-	-	-	-	-
α_2	1.2	1.20	0.56	0.56	0.78	0.27	0.30	1.12	0.46	0.48
α_3	0.8	0.80	0.36	0.36	0.52	0.18	0.20	0.76	0.30	0.31
λ_1	0.8	0.80	0.48	0.48	-	-	-	0.67	0.45	0.46
b_1	0.2	0.20	0.15	0.15	0.28	0.16	0.16	0.16	0.14	0.13
σ_ϵ	0.2	0.20	-	-	0.02	-	-	0.11	-	-
σ_{s1}	1.0	1.01	-	-	1.10	-	-	1.12	-	-
σ_{s2}	0.5	0.51	-	-	0.56	-	-	0.54	-	-
ρ_{s12}	0.0	0.00	-	-	0.00	-	-	0.01	-	-
Log-likelihood		-526.14			-527.27			-526.33		

Table 4.5: Population parameter estimates $\hat{\boldsymbol{\psi}}^*$ for condition 5 (λ_1=1.0). In the data generating model, $\beta_1^* = 0$ and $\alpha_1^* = 0$. ρ_{s12} is the correlation, $\frac{\sigma_{s12}}{\sigma_{s1}\sigma_{s2}}$.

Par	True	Proposed			Independence			Constrained		
		Est	SE	SE_R	Est	SE	SE_R	Est	SE	SE_R
β_1	-1.0	-1.00	0.32	0.32	-0.83	0.29	0.30	-0.82	0.22	0.23
β_1^*	-	0.29	0.40	0.40	-	-	-	-	-	-
β_2	1.5	1.51	0.29	0.29	1.45	0.23	0.23	1.55	0.28	0.28
β_3	0.0	0.00	0.18	0.18	0.01	0.16	0.16	0.04	0.17	0.17
α_1^*	-	0.08	0.66	0.66	-	-	-	-	-	-
α_2	1.2	1.20	0.56	0.56	0.68	0.24	0.28	1.10	0.46	0.48
α_3	0.8	0.80	0.36	0.36	0.46	0.16	0.18	0.74	0.30	0.31
λ_1	1.0	1.00	0.47	0.47	-	-	-	0.83	0.45	0.46
b_1	0.2	0.20	0.15	0.15	0.31	0.17	0.17	0.16	0.14	0.13
σ_ϵ	0.2	0.20	-	-	0.02	-	-	0.06	-	-
σ_{s1}	1.0	1.01	-	-	1.10	-	-	1.15	-	-
σ_{s2}	0.5	0.51	-	-	0.56	-	-	0.55	-	-
ρ_{s12}	0.0	0.00	-	-	0.00	-	-	0.01	-	-
Log-likelihood		-524.95			-526.64			-525.25		

are close to the robust standard errors in all models across all conditions (with the differences less than 0.01). The independence and constrained models tend to somewhat underestimate the standard errors for all model parameters across all conditions.

For the lag parameter, we compared the asymptotic standard errors for N=200 with the standard deviations of the parameter estimates and the means of the estimated standard errors, based on 200 simulated datasets. With 5 quadrature points, the standard deviations of the parameter estimates are a bit larger than the means of the standard error estimates, but both are smaller than the asymptotic standard errors. Specifically, the standard deviations of the estimates are about 18%, 24%, 31%, 32%, 29%, and 56% smaller than the asymptotic standard errors, and the means of the standard error estimates are about 31%, 37%, 43%, 46%, 48%, and 51% smaller than the asymptotic standard errors for $\lambda_1 = 0$, 0.2, 0.4, 0.6, 0.8, and 1.0, respectively. We also tried 10 and 15 quadrature points with $\lambda_1 = 0.2$, but the results hardly changed.

Figures 4.4 to 4.8 compare the asymptotic bias for each parameter between the models across conditions (except σ_{s12} that shows little bias (close to 0) in all models). For the parameters λ_1 and b_1, the estimated 95% confidence intervals for the finite-sample bias are computed based on 200 simulated datasets for N=200 and for different values of λ_1.

Overall, the asymptotic bias tends to increase as the true value for λ_1 increases from 0 to 1. The asymptotic bias for λ_1 lies in the estimated 95% confidence interval for the finite-sample bias in all conditions. Ignoring serial dependence produces particulary large bias for all $\boldsymbol{\alpha}$, b_1, σ_{s1}, and σ_{s2} parameters, and ignoring the initial conditions problem produces larger bias for β_3 and σ_ϵ than the other parameters. The asymptotic bias for b_1 lies in the estimated 95% confidence interval in all conditions except when $\lambda_1 = 1.0$.

Figure 4.9 presents the asymptotic RMSE for the mean slope parameter b_1 between the three models when N=200, 1,000, and 3,000.

With the sample size N=200, the asymptotic RMSE is larger in the proposed model than in the independence model when $\lambda_1 < 0.4$ and in the constrained model $\lambda_1 < 1.0$. This is because the asymptotic standard errors are underestimated in the independence and constrained models when the sample size is small. With the sample size N=1,000, the asymptotic RMSE is larger in the independence model when $\lambda_1 > 0.2$ and in the constrained model when $\lambda_1 > 0.4$ than in the proposed model. With the sample size N=3,000, the asymptotic RMSE is larger in both independence and constrained models when $\lambda_1 > 0.2$ than in the proposed model.

Now we illustrate the marginal item characteristic curves (ICCs) for the three models. Figure 4.10 shows ICCs in condition 5 ($\lambda_1 = 1.0$).

In the figure, the dashed curves represent the true ICCs from the data generating model, the solid curves represent the ICCs for the estimated models at time points 2 and 3, and the dashed-dotted curves represent the estimated ICC at time 1.

For the proposed model, there is nearly no gap between the estimated curves across time. When serial dependence is ignored, the estimated ICCs are the same across all time points, but they are all off from (lower than) the true ICCs. When the initial conditions problem is

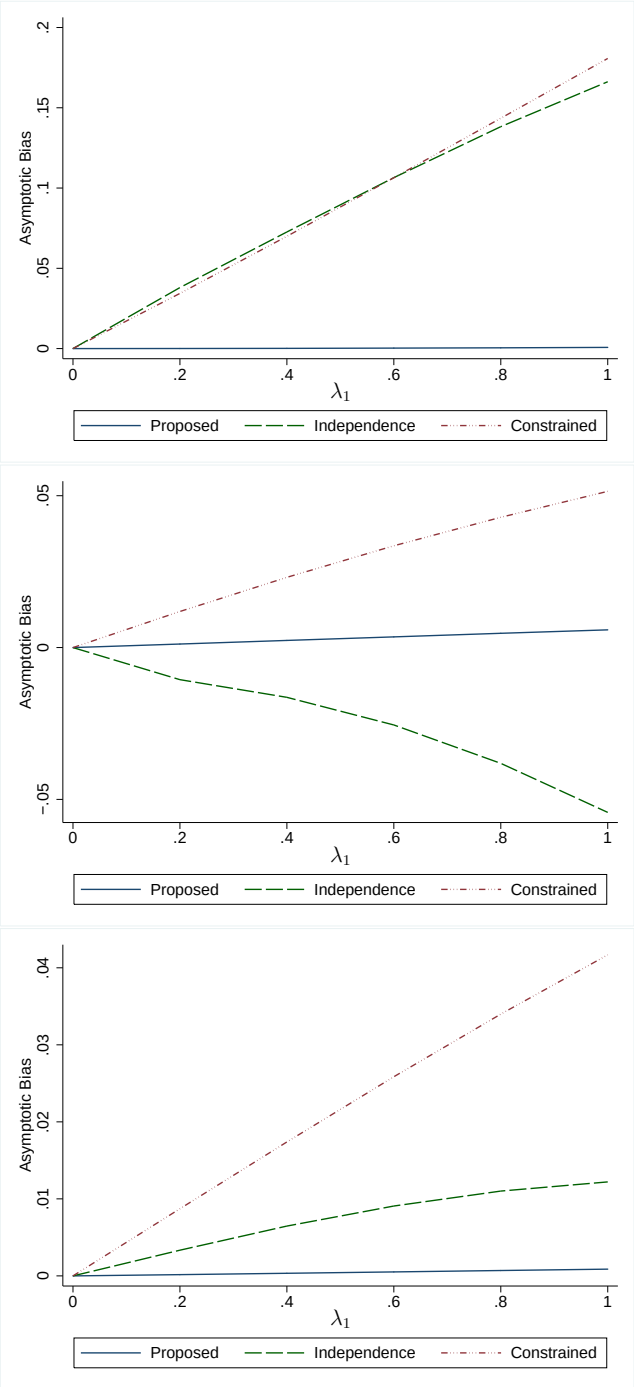

Figure 4.4: Asymptotic bias for the item parameters β_1 (top), β_2 (middle), and β_3 (bottom). Note that $\boldsymbol{\beta}$ have a different meaning in the proposed model and the constrained model that include the lag parameter.

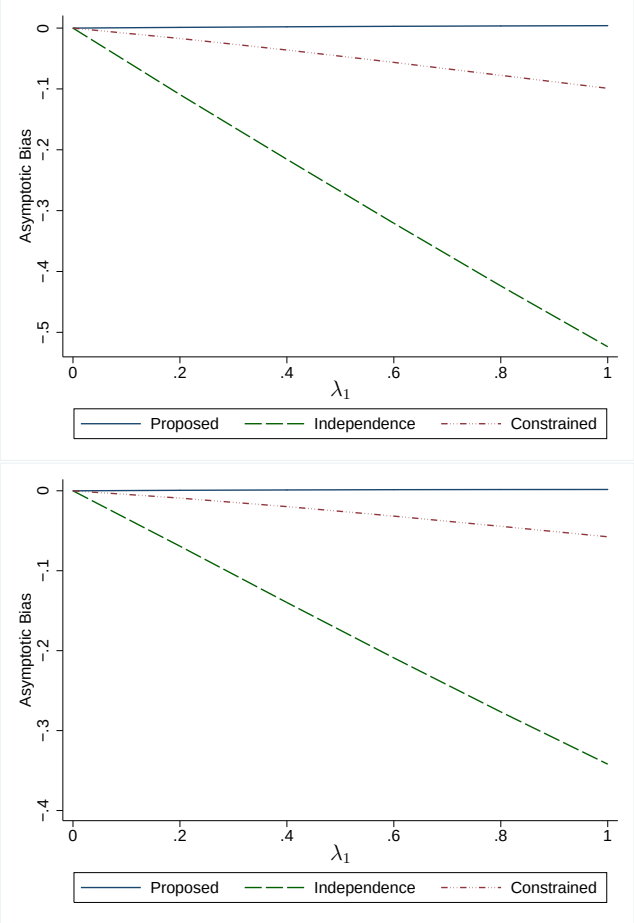

Figure 4.5: Asymptotic bias for the item parameters α_2 (top) and α_3 (bottom). Note that $\boldsymbol{\alpha}$ have a different meaning in the proposed model and the constrained model that include the lag parameter.

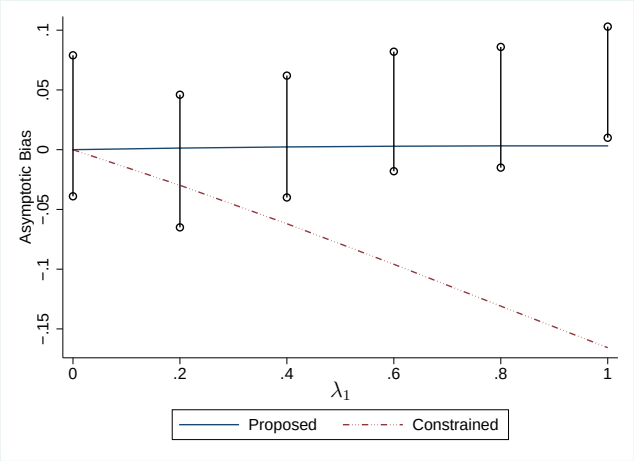

Figure 4.6: Asymptotic bias for the lag parameter λ_1. The estimated 95% confidence intervals for the finite-sample bias based on 200 replicates (N=200) are presented for the proposed model.

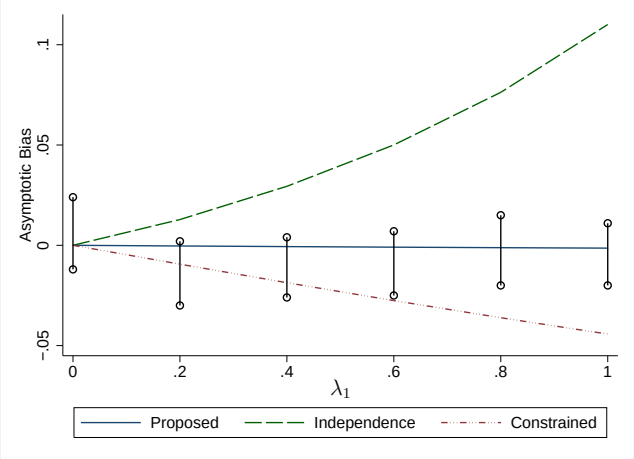

Figure 4.7: Asymptotic bias for the mean slope b_1. The estimated 95% confidence intervals for the finite-sample bias based on 200 replicates (N=200) are presented for the proposed model.

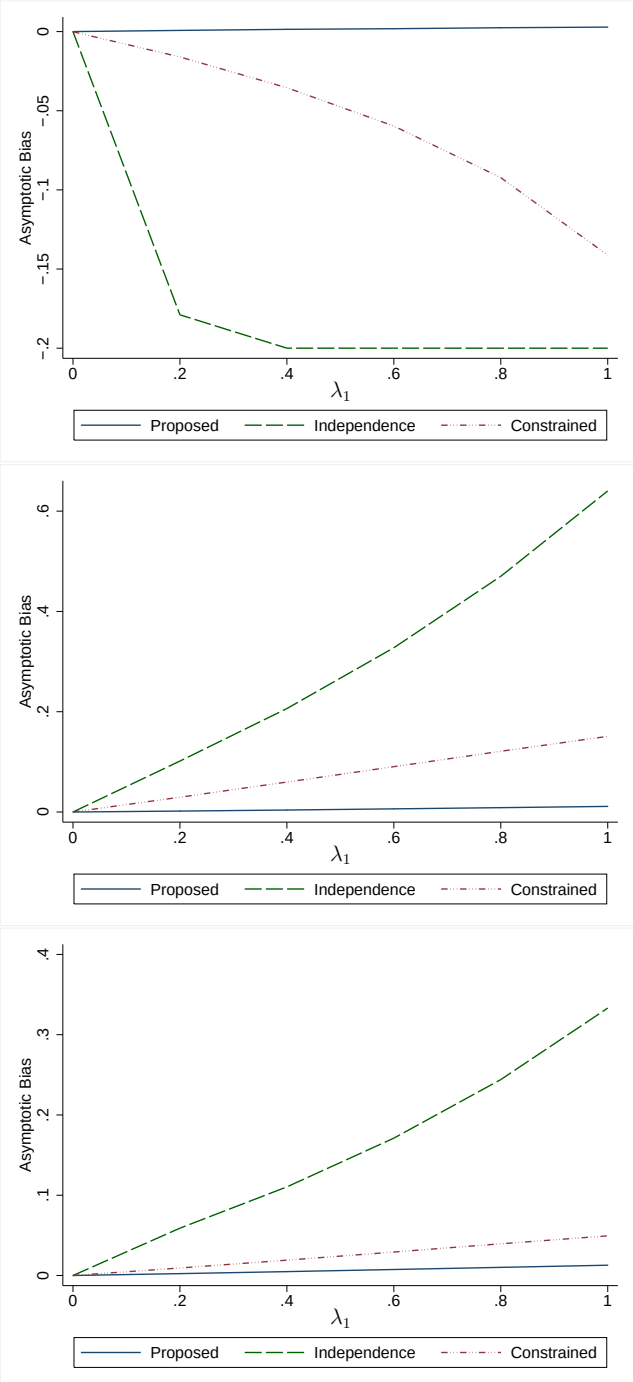

Figure 4.8: Asymptotic bias for the standard deviations of time effects σ_ϵ (top), initial status σ_{s1} (middle), and growth rate σ_{s2} (bottom).

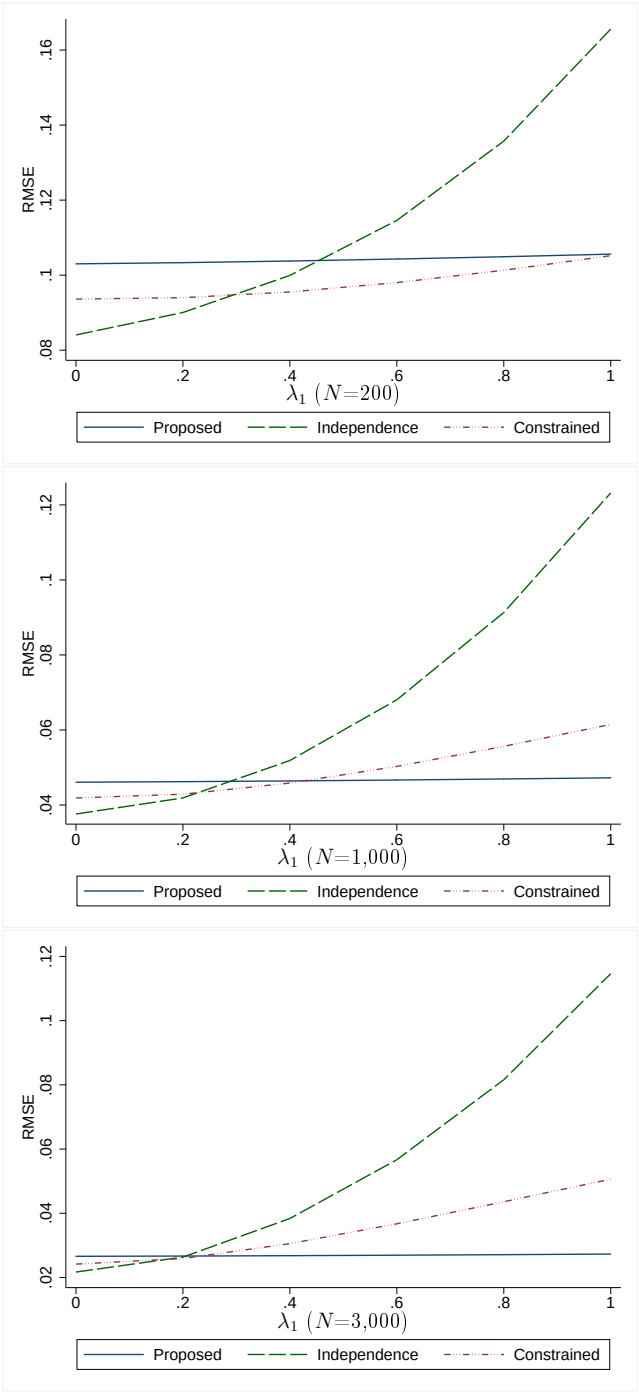

Figure 4.9: Asymptotic root mean squared error (RMSE) for the mean slope b_1 when N=200, 1000, and 3000.

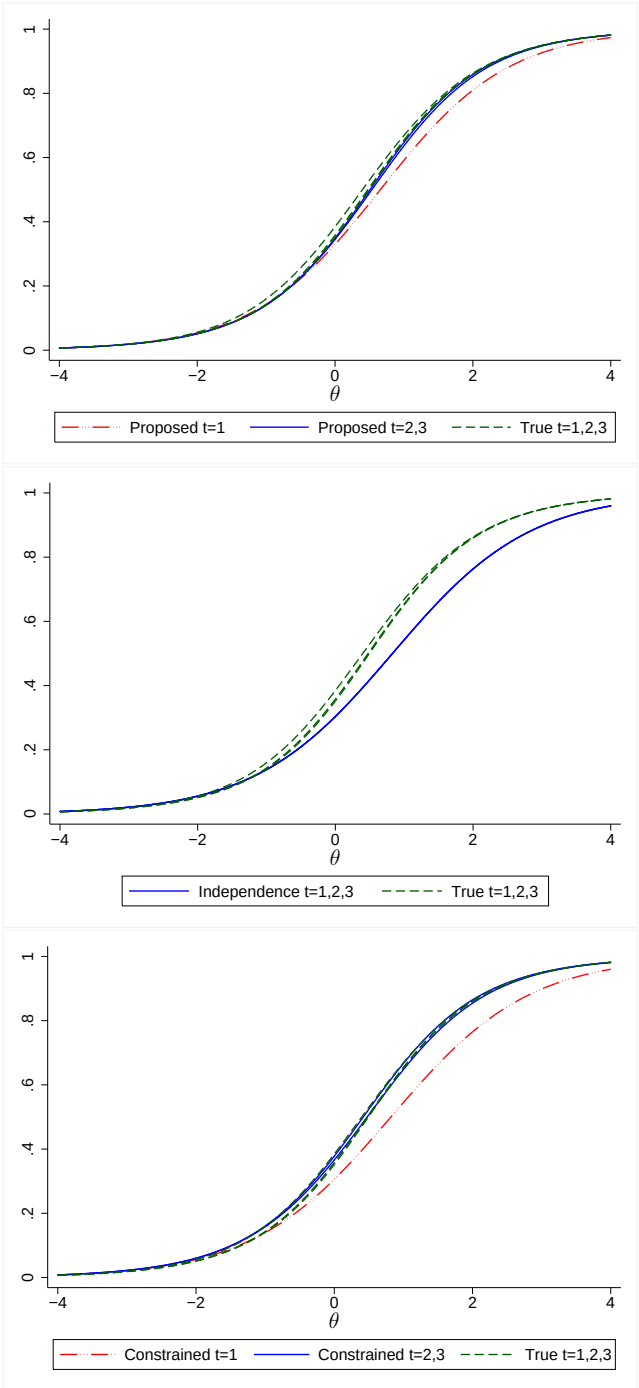

Figure 4.10: Item characteristic curves for the proposed model (top), independence model (middle), and constrained model (bottom)

ignored, the estimated ICC at time 1 is different from (lower than) the estimated ICCs at time points 2 and 3 and the true ICCs.

This result shows that imposing the invariance assumption on the item parameters across time actually forces the ICC to be different at time 1 from ICCs at later time points. Freeing the item parameters at time 1 helps the ICCs to resemble each other across all time points. Ignoring serial dependence results in bias in the ICCs at all time points.

Asymptotic power curves to detect the lagged effect for the proposed model is shown in Figure 4.11 as a function of the sample size. The estimated 95% confidence intervals for the finite-sample power are computed for the proposed model based on the LR test (df=1) using 200 simulated datasets (with 5 quadrature points). The same estimated confidence intervals are obtained with more quadrature points (10 and 15) at N=200.

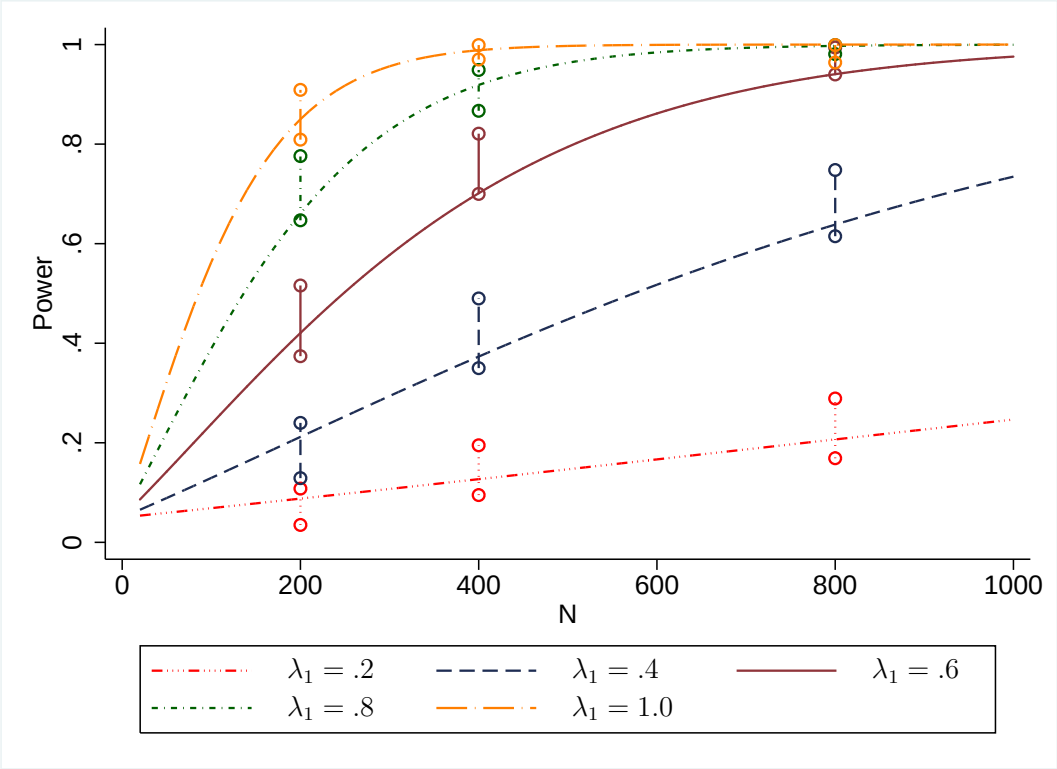

Figure 4.11: Asymptotic power to detect the lagged effect as a function of the sample size in varying values for the lagged effect. The estimated 95% confidence intervals for the finite samples using LR tests (based on 200 replicates) are shown for the proposed model for λ_1 = 0.2, 0.4, 0.6, 0.8 and 1.0 at N=200, 400, and 800.

The sample size required to achieve a power of 0.80 is about 200, 300, 600 for λ_1 = 1.0, 0.8, and 0.6, respectively. For $\lambda_1 \leq 0.4$, $N \geq 1,000$ is required. The estimated 95%

confidence intervals include the asymptotic curves at N=200, 400, and 800, based on the LR test. When the Wald test was used (df=1) for the finite samples, the lower bounds of the estimated 95% confidence intervals tend to be placed somewhat higher than the asymptotic power curves. It makes sense given that the standard errors were a bit underestimated for the finite samples. It is also a known fact that the likelihood ratio test is more conservative and reliable than the Wald test for finite samples (e.g., Engle, 1980; Buse, 1982)

4.7 Empirical Study

The Korea Youth Panel Survey (KYPS; Lee et al., 2010) tracked a nationally representative sample of second year middle school students every year from 2003 to 2008. Six waves of the data were collected where students progressed from middle school to high school at wave 3 and were out of high school at wave 6. There are 3,449 students in 103 middle schools at wave 1 and 3,125 students in 911 high schools at wave 3. At wave 6, there are 2,833 students. For simplicity, students who switched their school membership during the middle school or high school years were excluded from the data (less than 2% each year). The self-esteem scale was used which consisted of 12 items on a 5-point Likert scale (from strongly disagree to strongly agree). We chose seven items that appear more closely related to each other (e.g., all negatively worded), which was confirmed by Cronbach's alpha (about 0.65, each year). These items are: 1) I sometimes think I am a useless person, 2) I sometimes think I am a bad person, 3) I sometimes feel like I am a failure, 4) I think I am a trouble maker, 5) I think I am a juvenile delinquent, 6) Other people think I am a trouble maker, and 7) Other people think I am a juvenile delinquent. To measure a positive self-image (or self-esteem), the response categories were reversed and dichotomized.

We fit the full model (Ma) including the lag parameters for all items using `gllamm` (Rabe-Hesketh et al., 2005) with 5 quadrature points. Seven separate models (M1 to M7) were also fit, each including a lag parameter for one item (items 1 to 7), respectively. In addition, a reduced model (M0) without the lag parameter (λ_i) and without the two free parameters (α_i^*, β_i^* at time 1) was fit for comparison. Tables 4.6 and 4.7 summarize the results.

In all tables, the model-based standard errors are presented since there is not much difference (less than 0.01) between the model-based and robust standard errors.

In Table 4.6, the estimates of the lag parameter and free item parameters are listed. For the lag parameter, the Wald and likelihood ratio (LR) test statistics are also given. To compute the LR statistic, the log-likelihood for the reduced model (M0) is compared with the log-likelihoods for models M1 to M7, each with 3 degrees of freedom. For the full model (Ma), only the Wald statistics are presented. The lagged effects (λ_i) are significant and the estimates are quite large for all items, ranging from 0.65 to 1.03 (odds ratio 1.91 to 2.80) in the separate models (M1 to M7). The p-values are smaller than 0.0001 based on both LR and Wald tests, and the LR statistics appear similar to or slightly larger than the Wald statistics. In the full model (Ma), the lagged effects are also all significant and somewhat

Table 4.6: Parameter estimates and standard errors (in the parentheses) for the lag and free item parameters for the Korea Youth Panel Survey (KYPS) data. The estimates from the full model (Ma) and separate models (M1 to M7) are presented for each item. $\beta_i'(=\beta_i+\beta_i^*)$ and $\alpha_i'(=\alpha_i+\alpha_i^*)$ are also presented.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
λ_i							
Ma	0.85	0.92	0.97	1.21	2.05	2.29	2.51
	(0.04)	(0.04)	(0.04)	(0.07)	(0.13)	(0.14)	(0.15)
(Wald)	(389.66)	(365.57)	(473.49)	(267.64)	(231.04)	(257.92)	(258.24)
M1-M7	0.65	0.67	0.75	0.67	0.75	1.03	1.01
	(0.04)	(0.04)	(0.04)	(0.06)	(0.11)	(0.11)	(0.12)
(Wald)	(239.01)	(208.80)	(309.41)	(100.80)	(42.25)	(82.08)	(64.32)
(LR)	(243.36)	(208.00)	(330.68)	(195.24)	(51.84)	(116.96)	(106.50)
β_i^*							
Ma	0.33	0.28	0.63	0.13	1.61	0.17	1.81
	(0.06)	(0.07)	(0.06)	(0.12)	(0.42)	(0.30)	(0.47)
M1-M7	0.26	0.19	0.58	-0.28	0.37	-0.54	0.77
	(0.05)	(0.06)	(0.05)	(0.09)	(0.21)	(0.24)	(0.37)
β_i'							
Ma	-0.69	-1.16	-0.04	1.5	5.18	3.55	5.57
M1-M7	-0.79	-1.26	-0.16	1.24	4.42	3.20	5.16
α_i^*							
Ma	-0.25	-0.23	-0.24	-0.5	-0.64	-1.53	-0.95
	(0.04)	(0.05)	(0.04)	(0.09)	(0.24)	(0.20)	(0.26)
M1-M7	-0.08	-0.04	-0.02	-0.26	-0.33	-1.13	-0.47
	(0.04)	(0.05)	(0.05)	(0.01)	(0.22)	(0.20)	(0.24)
α_i'							
Ma	0.75	0.94	0.86	1.44	2.55	1.93	2.54
M1-M7	0.92	1.07	1.03	1.70	2.82	2.41	2.92

Table 4.7: Parameter estimates and standard errors (in the parentheses) for the structural and measurement parts of the model for the Korea Youth Panel Survey (KYPS) data. Reduced model (M0) and full model (Ma) are presented in addition to the separate models (M1 to M7). ρ_{s12} is the correlation, $\frac{\sigma_{s12}}{\sigma_{s1}\sigma_{s2}}$.

	M0	Ma	M1	M2	M3	M4	M5	M6	M7
Structural									
b_1	0.27	0.24	0.26	0.27	0.28	0.26	0.28	0.26	0.27
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
σ_ϵ	1.41	1.46	1.35	1.40	1.40	1.42	1.39	1.42	1.39
	(0.08)	(0.09)	(0.08)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)
σ_{s1}	0.83	0.96	0.79	0.82	0.80	0.76	0.82	0.75	0.84
	(0.03)	(0.06)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)
σ_{s2}	0.28	0.26	0.24	0.26	0.26	0.24	0.28	0.24	0.26
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
ρ_{s12}	0.05	-0.26	0.03	0.04	0.05	0.43	0.04	0.47	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Measurement									
β_1	-0.85	-1.02	-1.05	-0.80	-0.83	-0.80	-0.81	-0.83	-0.79
	(0.04)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03	(0.03)	(0.03)
β_2	-1.31	-1.44	-1.27	-1.45	-1.31	-1.25	-1.27	-1.29	-1.24
	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04	(0.04)	(0.04)
β_3	-0.35	-0.67	-0.31	-0.31	-0.74	-0.30	-0.31	-0.33	-0.29
	(0.03)	(0.05)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03	(0.03)	(0.04)
β_4	1.76	1.37	1.82	1.83	1.76	1.52	1.81	1.80	1.87
	(0.06)	(0.10)	(0.07)	(0.07)	(0.06)	(0.08)	(0.06)	(0.06)	(0.07)
β_5	4.46	3.57	4.80	4.80	4.69	4.56	4.05	4.58	4.87
	(0.13)	(0.22)	(0.17)	(0.17)	(0.16)	(0.14)	(0.30)	(0.14)	(0.18)
β_6	4.31	3.38	4.38	4.38	4.29	4.47	4.38	3.74	4.32
	(0.16)	(0.25)	(0.15)	(0.15)	(0.15)	(0.16)	(0.16)	(0.20)	(0.16)
β_7	5.39	3.76	5.38	5.38	5.29	5.55	5.30	5.28	4.39
	(0.19)	(0.25)	(0.18)	(0.18)	(0.17)	(0.20)	(0.18)	(0.18)	(0.23)
α_2	1.17	1.17	1.22	1.11	1.18	1.16	1.17	1.16	1.17
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
α_3	1.09	1.10	1.13	1.09	1.05	1.09	1.09	1.09	1.09
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
α_4	2.01	1.94	2.05	1.97	1.98	1.96	1.96	2.01	1.97
	(0.06)	(0.06)	(0.06)	(0.05)	(0.05)	(0.06)	(0.05)	(0.06)	(0.05)
α_5	3.16	3.19	3.37	3.24	3.25	3.08	3.15	3.18	3.21
	(0.11)	(0.14)	(0.13)	(0.12)	(0.12)	(0.10)	(0.12)	(0.11)	(0.12)
α_6	3.49	3.46	3.45	3.31	3.35	3.45	3.33	3.54	3.17
	(0.16)	(0.17)	(0.14)	(0.13)	(0.14)	(0.16)	(0.14)	(0.17)	(0.12)
α_7	3.76	3.49	3.66	3.52	3.56	3.70	3.47	3.58	3.39
	(0.18)	(0.17)	(0.15)	(0.14)	(0.14)	(0.17)	(0.14)	(0.16)	(0.15)
Log-likelihood	-55211	-54278	-55089	-55107	-55045	-55113	-55185	-55152	-55157

larger than those in M1 to M7. In particular, the lag parameter estimates for items 5 to 7 (odds ratio 7.38 to 12.18) are relatively larger than for other items (odds ratio 2.34 to 3.35) in the full model. This suggests that items 5 to 7 are more strongly influenced by the previous responses to the same items. The responses to these items are also more stable across time. These items are somewhat more negative than the other items and related to juvenile delinquency and other people's judgement of the student's behaviors.

The free item parameter estimates range from -0.54 to 0.77 for β_i^* and -0.02 to -1.13 for α_i^* in the separate models (M1 to M7). The estimates in the full model (Ma) tend to be somewhat larger for β_i^* and smaller for α_i^* than those in M1 to M7. We also present the item parameters β_i' ($=\beta_i + \beta_i^*$) and α_i' ($=\alpha_i + \alpha_i^*$) in Ma and M1 to M7. The differences between these estimates in the full and separate models are smaller in β_i' and α_i' than in β_i^* and α_i^*.

Table 4.7 lists parameter estimates and standard errors in the structural and measurement parts in all models (Ma, M0, and M1 to M7). Overall, there is not much difference between the models in the structural model parameters except ρ_{s12} is larger in absolute values in Ma than in other models. Specifically, the estimated mean slope (b_1) is about 0.7, and the estimated standard deviations of the time specific effects (σ_ϵ), the initial status (σ_{s1}), and the growth rate (σ_{s2}) are quite large, about 1.40, 0.80, and 0.26, respectively.

Based on the LR test, the full model (Ma) fits significantly better than all separate models ($p < 0.001$, df=20) and the reduced model (M0) ($p < 0.001$, df=23). The Wald tests for the lag parameters in the full model suggest that the full model fits better than each of the seven models with one lag parameter set to zero (and six lag parameters freely estimated) as well as the seven separate models with a lag parameter for one item at a time. These model comparisons correspond to the first steps of forward selection and backward elimination for model selection. Based on these results, the full model is chosen over the 15 competing models.

Figure 4.12 illustrates growth trajectories for 11 hypothetical students over six time points based on the full model (Ma). The latent trait values were generated using (4.12) where random effects were drawn from the corresponding multivariate and univariate normal distributions.

Overall, the Korean students' self-esteem tends to increase over time from grade 2 in middle school through one year after high school. The initial status and growth rate vary between students, but the variation in the initial status appears somewhat larger than the variation in the growth rate.

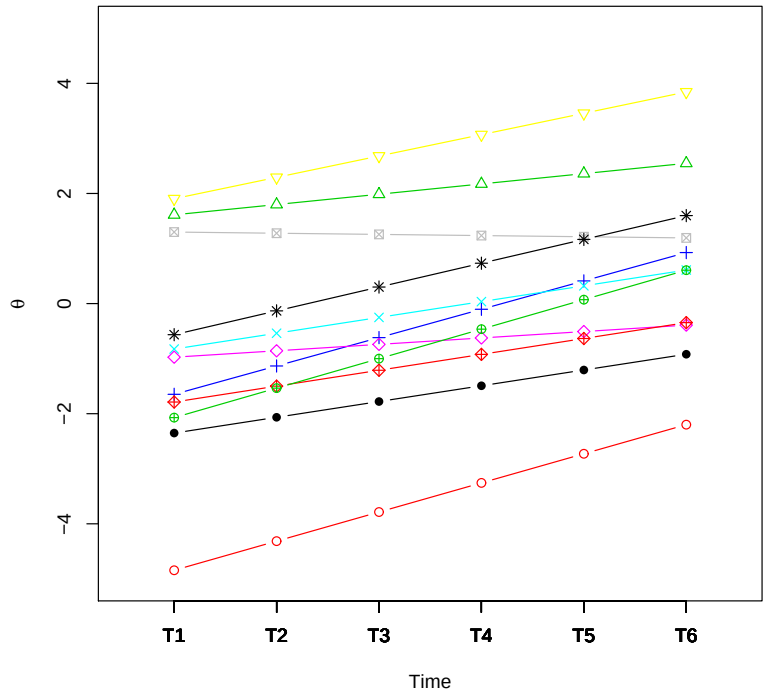

Figure 4.12: Growth trajectories for 11 hypothetical students (based on the full model) with randomly drawn random effects in the Korea Youth Panel Survey (KYPS) data.

4.8 Concluding Remarks

In this paper, we presented a first-order autoregressive IRT growth model for longitudinal binary item analysis. The proposed model for studying growth of a latent trait over time accommodates serial dependence between responses to the same items across time. Our model was illustrated with a linear growth trajectory, but an extension to a polynomial growth trajectory is straightforward. For polytomous responses, we can apply techniques that have been developed for categorical time series data (e.g. Fahrmeir & Kaufmann, 1987). However, such extensions increase the number of parameters to estimate and thus are computationally demanding.

We showed that the first-order autoregressive measurement model is equivalent to an IRT model with interaction parameters for responses at adjacent time points. Higher order interactions can also be considered. For example, an AR(2) autoregressive model is equivalent to allowing for interactions among the item responses two time-points apart.

Our model deals with serial correlations in longitudinal item analysis which has often been neglected in IRT. Standard ML software can be used for estimating the proposed model. Estimation requires only three-dimensional integrals and the dimensionality of the integrals stays the same regardless of the number of time points and items.

The importance of addressing the initial conditions problem in autoregressive IRT models were discussed and illustrated using simulations. We showed that constraining the item parameters to be equal across time can actually force the ICCs to differ across time, resulting in a violation of measurement invariance. A proper way of achieving approximate measurement invariance is to free the item parameters at time 1 so that the ICCs can resemble each other across time.

The proposed model can be estimated using existing ML software such as `gllamm` (Rabe-Hesketh et al., 2005) and `M-Plus` (Muthén & Muthén, 2008). However, when the data have a more complex data structure, such as a cross-classification of students by middle school and high school (Jeon & Rabe-Hesketh, 2012), such software may no longer be available to fit the model.

Chapter 5

Conclusion

In this dissertation, I considered new estimation methods and applications of complex generalized linear mixed models (GLMMs) for measurement and growth. The dissertation consists of three papers that correspond to Chapters 2, 3, and 4. Below I provide a brief summary for each chapter.

In Chapter 2, the variational maximization-maximization (MM) algorithm was presented for estimating GLMMS with crossed random effects. The variational MM algorithm is a modified version of the traditional EM algorithm where the E-step is replaced by another M-step that minimizes the KL distance between the variational distribution and the true posterior distribution. This new M-step is equivalent to maximizing the lower bound to the log-likelihood with respect to the variational distribution.

The variational MM algorithm is more general and flexible than the Gaussian variational approximation because our algorithm does not require a pre-specified functional form for the variational distribution. The general form for the variational density function is derived so that different types of priors for the random effects can be handled. Importantly, we can estimate models with crossed random effects based on the mean-field approximation that assumes conditionally independent latent variables given the data. We found that with reasonable sample sizes and prior variances, the posterior correlations between the random effects are negligible. In addition, the lower bound was quite close to the marginal log-likelihood in the examples that we considered in this paper.

Several simulation examples were provided to evaluate the performance of the variational MM algorithm and to compare it with the Laplace approximation for GLMMs with crossed random effects. The results show that overall, the variational MM algorithm performs as well as the Laplace approximation. With small cluster sizes, however, our algorithm performs better than the Laplace approximation especially for the variance parameters. Therefore, the variational MM method could be an effective alternative to the Laplace approximation.

In Chapter 3, the Monte Carlo local likelihood (MCLL) method was presented for maximum likelihood estimation of GLMMs with crossed random effects. The MCLL method

initially treats the model parameters as random variables and samples them jointly with random effects, from the posterior distribution for a particular prior. The likelihood function is then approximated up to a constant as a local likelihood density estimate of the posterior divided by the prior.

The MCLL method is similar to the MC kernel likelihood method (MCKL; De Valpine, 2004), which uses a kernel density estimation to approximate the posterior. The key advantage of MCLL is that it provides methods for obtaining standard errors whereas MCKL does not. MCLL is also less sensitive to bandwidth selection than MCKL.

It is important to note that MCLL allows likelihood inference for any complex models for which ML estimation may be infeasible but MCMC methods are possible. For example, in addition to GLMMs with crossed random effects considered here, the MCLL algorithm could be used to fit state-space models with higher dimensional latent variables. Potential applications for MCLL are therefore far beyond the models discussed in this paper. We have shown that the MCLL method provides results close to the ML estimates. Even if informative priors are specified, MCLL provides estimates close to the ML estimates, whereas the Bayesian estimates could be quite different. When ML inference is desired for highly complex models, the MCLL method seems to be an effective and practical choice.

In Chapter 4, a new autoregressive IRT growth model was proposed for longitudinal binary item analysis. The proposed model for studying growth of a latent trait accommodates serial dependence between responses to the same items across time. We showed that the first-order autoregressive measurement model is equivalent to an IRT model with interaction parameters for responses at adjacent time points. Higher order interactions can also be considered. For example, an AR(2) autoregressive model is equivalent to allowing for interactions among the item responses two time-points apart.

The proposed model deals with serial correlations in longitudinal item analysis which has often been neglected in IRT. Standard ML software can be used for estimating the proposed model. Estimation requires only three-dimensional integrals and the dimensionality of the integrals stays the same regardless of the number of time points and items.

The importance of addressing the initial conditions problem in autoregressive IRT models were discussed and illustrated using simulations. We showed that constraining the item parameters to be equal across time can actually force the ICCs to differ across time, resulting in a violation of measurement invariance. A proper way of achieving approximate measurement invariance is to free the item parameters at time 1 so that the ICCs can resemble each other across time.

Bibliography

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.

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

Aitkin, M., & Alfo, M. (1998). Regression models for longitudinal binary responses. *Statistics and Computing*, *8*, 289–307.

Aitkin, M., & Alfo, M. (2003). Longitudinal analysis of repeated binary data using autoregressive and random effect modelling. *Statistical Modelling*, *3*, 291–303.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Anderson, T. W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, *76*, 598–606.

Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.) *Test Design: Developments in Psychology and Psychometrics*, (pp. 245–275). New York: Academic Press.

Bartolucci, F., & Nigro, V. (2010). A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica*, *78*, 719–733.

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes R package version 0.999375-31*. Downloadable from http://CRAN.Rproject.org/package=lme4.

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*, 132–152.

Bishop, C., Lawrence, N., Jaakkola, T., & Jordan, M. (1998). Approximating posterior distributions in belief networks using mixtures. In M. Jordan, M. Kearns, & S. Solla (Eds.) *Advances in Neural Information Processing Systems*, (pp. 416–422). Cambridge, MA: MIT Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Booth, J., & Hobert, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society Series B*, *61*, 265–285.

Boyd, S., & Vandenberghe, L. (2004). *Convex Inference in Statistical Analysis*. Cambridge: Cambridge University Press.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Braeken, J. (2011). A boundary mixture approach violations of conditional independence. *Psychometrika*, *76*, 57–76.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, *72*, 393–411.

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Breslow, N., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, *82*, 81–91.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier tests: An expository note. *The American Statistician*, *36*, 153–157.

Butler, J. S., & Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, *50*, 761–764.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581–612.

Chacón, J. E., Montanero, J., Nogales, A. G., & Pérez, P. (2007). On the use of Bayes factor in frequentist testing of a precise hypothesis. *Communications in Statistics*, *36*, 2251–2261.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics and Data Analysis*, *55*, 12–25.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, *11*, 183–196.

De Valpine, P. (2004). Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. *Journal of the American Statistical Association*, *99*, 523–535.

Delicado, P. (2006). Local likelihood density estimation based on smooth truncation. *Biometrika*, *93*, 472–480.

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

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Duong, T., & Hazelton, M. L. (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, *15*, 17–30.

Durbin, J., & Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state-space models. *Biometrika*, *84*, 669–684.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. *The Annals of Statistics*, *7*, 1–26.

Eguchi, S., & Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society Series B*, *60*, 709–724.

Eid, M., & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics*, *23*, 193–215.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Engle, R. F. (1980). Wald, likelihood ratio and Lagrange Multiplier test in econometrics. In Z. Griliches, & M. Intriligator (Eds.) *Handbook of Econometrics*, (pp. 775–826). Amsterdam: North-Holland Science Publishers.

Fahrmeir, L., & Kaufmann, H. (1987). Regression models for non-stationary categorical time series. *Journal of Time Series Analysis*, *8*, 147–160.

Fang, K.-T., & Wang, Y. (1993). *Number-Theoretic Methods in Statistics*. London: Chapman & Hall.

Fitzmaurice, G. M., Laird, N. M., & Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, *8*, 284–309.

Fotouhi, A. R., & Davies, R. B. (1997). Modelling repeated durations: The initial conditions problem. In A. Forcina, G. Marchetti, & G. Hatzinger, R. Galmacci (Eds.) *Statistical Modelling, Proceedings of the 11th International Workshop on Statistical Modelling*, (pp. 159–166). Graphos: Citta di Castello.

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous items. *British Journal of Mathematical and Statistical Psychology*, *58*, 145–172.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317–1339.

Geyer, C. J. (1994). On the convergene of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society Series B*, *56*, 261–274.

Geyer, C. J. (1996). Estimation and optimization of functions. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.) *Markov Chain Monte Carlo in Practice*, (pp. 241–258). New York: Chapman & Hall.

Geyer, C. J., & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society Series B*, *54*, 657–699.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*, 430–431.

Goldstein, H. (2003). *Multilevel Statistical Models (3rd. ed.)*. London: Arnold.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A*, *159*, 505–513.

Hall, P., Humphreys, K., & Titterington, D. M. (2002). On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *Journal of the Royal Statistical Society Series B*, *2002*, 549–564.

Hall, P., Ormerod, J. T., & Wand, M. P. (2011). Theory of Gaussian variational approximation for a Poisson linear mixed model. *Statistica Sinica*, *21*, 369–389.

Hancock, G. R., & Kuo, W. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling*, *8*, 470–489.

Heagerty, P., & Kurland, B. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*, 973–985.

Heckman, J. J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In C. F. Manski, & D. MacFadden (Eds.) *Structural Analysis of Discrete Data with Econometric Applications*, (pp. 179–195). Cambridge: MIT Press.

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

Hjort, N., & Jones, M. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, *24*, 1619–1647.

Hobert, J., & Casella, G. (1996). The effect of improper priors on gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*, 1461–1473.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local item dependence among test items. *Psychological Methods*, *2*, 261–277.

Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Hsiao, C. (2003). *Analysis of Panel Data (2nd ed.)*. New York: Cambridge University Press.

Humphreys, K., & Titterington, D. (2003). Variational approximations for categorical causal modeling with latent variables. *Psychometrika*, *68*, 391–412.

Hyslop, D. R. (1999). State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. *Econometrica*, *67*, 1255–1294.

Ip, E. (2000). Adjusting for information inflation due to local dependence in moderately large item clusters. *Psychometrika*, *65*, 73–91.

Ip, E. (2001). Testing for local dependence in dichotomous and polutomous item response models. *Psychometrika*, *66*, 109–132.

Ip, E. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*, 367–386.

Jeon, M. (2011). *Monte Carlo kernel likelihood method for generalized linear mixed models with crossed random effects*. Master's thesis, University of California, Berkeley.

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behavioral Statistics*, *37*, 518–542.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (in press). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*.

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis*, *52*, 5066–5074.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, *19*, 140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.) *Learning in Graphical Models*, (pp. 105–161). Cambridge: MIT Press.

Karim, M., & Zeger, S. (1992). Generalized linear models with random effects: Salamander mating revisited. *Biometrics*, *48*, 631–644.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kauermann, G., & Tutz, G. (2000). Local likelihood estiamtion in varying-coefficient models including additive bias correction. *Nonparametric Statistics*, *12*, 343–371.

Koehler, E., Brown, E., & Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, *63*, 155–162.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *The Annals of Statistics*, *73*, 805–811.

Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data (with discussion). *Journal of the Royal Statistical Society Series A*, *34*, 1–41.

Lee, K.-S., Lim, H.-J., & Ahn, S.-Y. (2010). *Korea Youth Panel Study*. Seoul: National Youth Policy Institute. Available: http://archive.nypi.re.kr.

Lee, Y., & Nelder, J. A. (2006). Double-hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series C*, *55*, 1–29.

Lin, X., & Breslow, N. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, *91*, 1007–1016.

Lindsay, B. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, *11*, 86–94.

Lindsay, B. G., Clogg, C. C., & Grego, J. M. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*, 96–107.

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*, 1014–1022.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, *81*, 624–629.

Loader, C. (1996). Local likelihood density estimation. *The Annals of Statistics*, *24*, 1602–1618.

Loader, C. (1999). *Local Regression and Likelihood*. New York: Springer.

Loehlin, J. C. (1998). *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Magnus, P., Gjessing, H. K., Skrondal, A., & Skjaerven, R. (2001). Paternal contribution to birth weight. *Journal of Epidemiology and Community Health*, *55*, 873–877.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, *92*, 162–170.

McGuire, L. W. (2010). *Practical formulation of the latent growth item response model*. Ph.D. thesis, University of California, Berkeley.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research,*, *13*, 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*, 289–311.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*, 5–9.

Muthén, L., & Muthén, B. (2008). *Mplus User's Guide*. Angeles, CA: Muthen & Muthen.

Natarajan, R., & McCulloch, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, *82*, 639–643.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, *31*, 214–225.

Neal, R. M., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, (pp. 355–368). Dordrecht: Kluwer Academic Publishers.

Opper, A. C., M. (2009). Variational Gaussian approximation revisited. *Neural Computation*, *21*, 786–792.

Ormerod, J. T. (2010). Explaining variational approximations. *The American Statistician*, *64*, 140–153.

Ormerod, J. T., & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *21*, 2–17.

Park, B. U., Kim, W. C., & Jones, M. C. (2002). On local likelihood density estimation. *The Annals of Statistics*, *30*, 1480–1495.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy. *Applied Psychological Measurement*, *30*, 100–120.

Pinheiro, J., & Bates, D. (1995). Approximation to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphics and Statistics*, *4*, 12–35.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team (2012). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-104.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata (3rd ed.)*. College Station, TX: Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Gjessing, H. (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics*, *64*, 280–288.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Rao, C. R. (2003). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics*, *31*, 545–554.

Raudenbush, S. (1993). A cross random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, *18*, 321–349.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Willms, J. D. (1995). Estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307–335.

Rijmen, F., & Jeon, M. (in press). Fitting an item response theory model with random item effects across groups by a variational approximation method. *The Annals of Operations Research*.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116.

Rosenbaum, P. R. (1999). Item bundles. *Psychometrika*, *53*, 349–359.

Rotnitzky, A., & Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, *50*, 1163–1170.

Sagan, H. (1969). *Introduction to the Calculus of Variations*. New York: Dover.

Sándor, Z., & Train, K. (2004). Quasi-random simulation of discrete choice models. *Transportation Research Part B*, *38*, 313–327.

Satorra, A., & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *51*, 83–90.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth model. In L. M. Collins, & A. G. Sayer (Eds.) *New Methods For the Analysis of Change*, (pp. 179–199). Washington DC: American Psychological Association.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, *78*, 719–727.

Schilling, S. G., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Segawa, E. (2005). A growth model for multilevel ordinal data. *Journal of Educational and Behavioral Statistics*, *30*, 369–396.

Serrano, D. (2010). *A second-order growth model for longitudinal item response data*. Ph.D. thesis, University of North Carolina, Chapel Hill.

Shephard, N., & Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, *84*, 653–667.

Shorack, G. R., & Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *The Annals of Statistics*, *4*, 1200–1209.

Sireci, S., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.

Skrondal, A., & Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society Series A*, *172*, 659–687.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS version 1.4 [Computer program]*. UK: MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

StataCorp (2009). *Stata Statistical Software: Release 11*. TX: StataCorp LP.

Steele, B. M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics*, *52*, 1295–1310.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*, 1–16.

Sung, Y., & Geyer, C. J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, *35*, 990–1011.

Tanner, M. A. (1993). *Tools for Statistical Inference: Observed Data and Data Augmentation (2nd ed.)*. Berlin: Springer-Verlag.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and densities. *Journal of the American Statistical Association*, *81*, 82–86.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*, 181–195.

Vaida, F., & Meng, X.-L. (2005). Two slice-EM algorithms for fitting generalized linear mixed models with binary response. *Statistical Modelling*, *5*, 229–242.

Van den Noortgate, W., & De Boeck, P. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.

Varin, C., & Czado, C. (2010). A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics*, *11*, 127–138.

Verguts, T., & De Boeck, P. (2000). A Rasch model for learning while solving an intelligence test. *Applied Psychological Measurement*, *24*, 151–162.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, *58*, 395–415.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wand, M. P., & Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, *88*, 520–528.

Wang, W., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–26.

White, I. R. (2010). `simsum`: Analyses of simulation studies including Monte Carlo error. *The Stata Journal*, *10*, 369–385.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60*, 181–198.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika*, *80*, 791–795.

Wooldridge, J. F. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, *20*, 39–54.

Wu, H., & Zhang, J.-T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, *97*, 883–897.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods for Differential Item Functioning: Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.