Bayesian Approach Dealing with Mixture Model Problems

Huaiye ZHANG

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Statistics

Inyoung Kim, Committee Chair

Feng Guo

Scotland Leman

Eric P. Smith

George Terrell

April 23^{rd}, 2012

Blacksburg, Virginia

Keywords: Adaptive Rejection Metropolis Sampling, Simulated Annealing, Dirichlet Process, Hierarchical Model, Nonlinear Mixed Effects Model, Infinite Mixture Model

Bayesian Approach Dealing with Mixture Model Problems

Huaiye ZHANG

ABSTRACT

In this dissertation, we focus on two research topics related to mixture models. The first topic is "Adaptive Rejection Metropolis Simulated Annealing for Detecting Global Maximum Regions", and the second topic is "Bayesian Model Selection for Nonlinear Mixed Effects Model".

In the first topic, we consider a finite mixture model, which is used to fit the data from heterogeneous populations for many applications. An Expectation Maximization (EM) algorithm and Markov Chain Monte Carlo (MCMC) are two popular methods to estimate parameters in a finite mixture model. However, both of the methods may converge to local maximum regions rather than the global maximum when multiple local maxima exist. In this dissertation, we propose a new approach, Adaptive Rejection Metropolis Simulated Annealing (ARMS annealing), to improve the EM algorithm and MCMC methods. Combining simulated annealing (SA) and adaptive rejection metropolis sampling (ARMS), ARMS annealing generate a set of proper starting points which help to reach all possible modes. ARMS uses a piecewise linear envelope function for a proposal distribution. Under the SA framework, we start with a set of proposal distributions, which are constructed by ARMS.

This method finds a set of proper starting points, which help to detect separate modes. We refer to this approach as ARMS annealing. By combining together ARMS annealing with the EM algorithm and with the Bayesian approach, respectively, we have proposed two approaches: an EM ARMS annealing algorithm and a Bayesian ARMS annealing approach. EM ARMS annealing implement the EM algorithm by using a set of starting points proposed by ARMS annealing. ARMS annealing also helps MCMC approaches determine starting points. Both approaches capture the global maximum region and estimate the parameters accurately. An illustrative example uses a survey data on the number of charitable donations.

The second topic is related to the nonlinear mixed effects model (NLME). Typically a parametric NLME model requires strong assumptions which make the model less flexible and often are not satisfied in real applications. To allow the NLME model to have more flexible assumptions, we present three semiparametric Bayesian NLME models, constructed with Dirichlet process (DP) priors. Dirichlet process models often refer to an infinite mixture model. We propose a unified approach, the penalized posterior Bayes factor, for the purpose of model comparison. Using simulation studies, we compare the performance of two of the three semiparametric hierarchical Bayesian approaches with that of the parametric Bayesian approach. Simulation results suggest that our penalized posterior Bayes factor is a robust method for comparing hierarchical parametric and semiparametric models. An application to gastric emptying studies is used to demonstrate the advantage of our estimation and evaluation approaches.

Acknowledgments

My sincere thanks go to my advisor, Inyoung Kim, whose expertise in Bayesian statistics gave me a solid foundation for this research. I would also like to thank Inyoung Kim for her countless hours of editing and suggestions to help me complete this dissertation. I would like to express my thanks to each member of my committee, Feng Guo, Scotland Leman, Eric P. Smith, and George Terrell for taking an interest in my research and giving the direction for my further research.

Many people on the faculty, staff and graduate students of the Department of Statistics assisted and encouraged me in various ways. I would also like to thank them.

I thank my wife Kitty for all of the sacrifices she made on my behalf. This dissertation would never have been possible without her love, support and understanding throughout my graduate studies. I would like to thank my mother, father and brother for inspiring me through all of life's challenges.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation consists of two independent topics, which are briefly introduced in Chapter 1.1 and 1.2, respectively. The outline is described in Chapter 1.3.

1.1 Adaptive Rejection Metropolis Simulated Annealing

The first topic is "Adaptive Rejection Metropolis Simulated Annealing for Detecting Global Maximum Regions", or simply written as "ARMS Annealing". An Expectation Maximization (EM) algorithm is the most popular method to estimate parameters in finite mixture models. However, the EM algorithm often converges to the local maximum regions, and the resulting mode is sensitive to the starting points. In the Bayesian approach, the

Markov Chain Monte Carlo (MCMC) sometimes converges to the local mode and is difficult to move to another mode. Thus, we combine simulated annealing (SA) and adaptive rejection metropolis sampling (ARMS) to generate a set of proper starting points which help to reach all possible modes. The limitation of SA is the difficulty in choosing sequences of proper proposal distributions for the target distribution. Since ARMS uses a piecewise linear envelope function for a proposal distribution, we incorporate ARMS into an SA approach so that we can start a set of more proper proposal distributions. As a result, we can detect the maximum region and estimate parameters for this global region. We refer to this approach as "ARMS annealing". By putting together ARMS annealing with the EM algorithm and with the Bayesian approach, respectively, we have proposed two approaches: an EM ARMS annealing algorithm and a Bayesian ARMS annealing approach. We compare two approaches using simulation, showing that the two approaches are comparable to each other. Both approaches detect the global maximum region and estimate the parameters correctly in this region. We demonstrate the advantage of our approaches using an example of the mixture model of two Poisson regressions. This mixture model is used to analyze survey data on the number of charitable donations.

1.2 Bayesian Model Selection for Nonlinear Mixed Effects Model

The second topic is "Bayesian Model Selection for Nonlinear Mixed Effects Model." Nonlinear mixed effects model (NLME) is a popular model in agricultural, environmental, and biomedical applications to analyze repeated measurement data. Continuous responses evolve over time within individuals from a population of interest. NLME accommodates both the variation among measurements within individuals and the individual-to-individual variation. However, it requires strong assumptions, which make the model less flexible, that are often not satisfied in real applications. To allow the NLME model be more flexible on assumptions, we present three semiparametric Bayesian hierarchical models on NLME, constructed with Dirichlet process (DP) priors. We propose a "penalized posterior Bayes factor" for comparing three semiparametric models with the parametric model. Using simulation studies, we compare the performance of our semiparametric hierarchical Bayesian approaches with that of the parametric Bayesian hierarchical approach. Simulation result suggests that our semiparametric approaches are more robust and flexible. Longitudinal data, with application to gastric emptying studies, is used to demonstrate the advantage of our approaches.

1.3 Outline of This Dissertation

The structure of the dissertation is as following: we introduced the "ARMS Annealing" method in Chapter 2. This topic is organized as follows: we give a brief review of adaptive rejection metropolis sampling (ARMS) and simulated annealing (SA) in Chapter 2.1. In Chapter 2.2, we propose a new approach which is named as ARMS annealing. We describe how ARMS annealing can be incorporated into an EM algorithm and a Bayesian approach, respectively. In Chapter 2.3 and 2.4, we report the simulation results comparing our approach with an EM algorithm alone and a Bayesian approach alone, respectively. We apply our approach to an example of a mixture of two Poisson regression models. This mixture model is used to find what covariates affect the number of charitable donations. The survey was conducted in year 2001 by Volunteer 21, a nonprofit organization in South Korea.

In Chapter 3, "Bayesian Model Selection for Nonlinear Mixed Effects Model" is presented. In Chapter 3.2, we describe four Bayesian hierarchical NLME models: the first with parametric priors, the second with DP measurement errors, and the third and fourth with DP random effects. In Chapter 3.3, we describe how to estimate parameters in each model. In Chapter 3.4, we discuss possible model selection procedures and proposed a penalized posterior Bayes factor. In Chapter 3.5, we report the results of simulations comparing three Bayesian hierarchical models. In Chapter 3.6, we apply our approaches to the data from an equine gastric emptying study.

Chapter 4 contains concluding remarks and discussions of both Chapter 2 and 3. We

further discuss our finding, the limitations, and direction of our future research.

Chapter 2

Adaptive Rejection Metropolis Sampling Annealing

2.1 Introduction

2.1.1 Background

A mixture of finite regression models has been used to model the data from heterogeneous populations in many applications (McLachlan and Peel, 2000). An Expectation Maximization (EM) algorithm is the most popular method to estimate parameters in a finite mixture model. The Bayesian approach is another method to fit a finite mixture model. However, an EM algorithm is often convergent to a local maximum region and is sensitive to the choice of starting points. In the Bayesian approach, the Markov Chain Monte Carlo

(MCMC) sometimes converges to a local mode so that it is difficult to move from one mode to another. Hence, in this paper we propose a new method to improve the limitation of the EM algorithm so that EM can estimate parameters in the global maximum region, and to help MCMC chain to converge more quickly in a finite mixture model.

Our approach is developed by incorporating adaptive rejection metropolis sampling (ARMS) into simulated annealing (SA). ARMS is a combination of adaptive rejection sampling (Von Neumann, 1951) and a Metropolis-Hastings sampling (Hastings, 1970). Adaptive rejection sampling (ARS) was proposed for sampling from univariate log-concave distributions (Gilks and Wild, 1992). Gilks further proposed ARMS, which is an extension of ARS on sampling from non-log-concave full conditional distributions (Gilks et al., 1995). It generalizes adaptive rejection sampling by including a Metropolis-Hasting algorithm step for non log-concave distributions. A piecewise linear envelope function is used for a proposal distribution in ARMS. Another method which we have used to develop our approach is simulated annealing (SA). It was introduced by Kirkpatrick et al. (1983) as a way of handling multiple modes in an optimization context. Although the SA is a well-known approach for detecting isolated modes, the limitation of SA is that it poses a difficulty in choosing sequences of proper proposal distributions for a target distribution. By combining ARMS into an SA approach, we use the piecewise linear envelope function for a proper proposal distribution in SA. As a result, we can detect a global maximum region. By yielding a more proper proposal distribution, this new approach has a high acceptance rate in Metropolis-Hastings sampling.

The goal of our study is to propose a new method for improving the limitation of the

EM algorithm so that EM can estimate parameters in the global maximum region and also to develop a more effective Bayesian approach so that the MCMC chains converge faster in the mixture model. Our new method is based on simulated annealing (SA) and adaptive rejection metropolis sampling (ARMS). We refer to this approach as ARMS annealing. By putting together ARMS annealing with the EM algorithm and with the Bayesian approach, respectively, we propose two approaches: One approach is an EM ARMS annealing algorithm and the other is a Bayesian ARMS annealing approach. Using these approaches, we can detect the global maximum region and estimate parameters in this region.

2.1.2 Adaptive Rejection Metropolis Sampling (ARMS)

Let $f(\beta_p|\beta_{(-p)})$ be a target distribution which assumes a univariate function, where $p = 1, \ldots, s$ and $\beta_{(-p)}$ is the parameters except β_p. For ease of notation we write $f(\beta)$ instead of $f(\beta_p|\beta_{(-p)})$.

Let $S_m = \{\beta_i;\, i = 0, \ldots, m+1\}$ denote a current set of abscissae in ascending order, where β_0 and β_{m+1} are the possibly infinite lower and upper limits of domain D of $f(\beta)$. For $1 \leq i \leq m$, let $L_{ij}(\beta_i, \beta_j;\, S_m)$ denote the straight line through points $[\beta_i, lnf(\beta_i)]$ and $[\beta_j, lnf(\beta_j)]$. Define a piecewise linear function $h_{i,i+1}(\beta) = \max[L_{i,i+1}, \min\{L_{i-1,i}, L_{i+1,i+2}\}]$ and $h_m(\beta) = \{h_{i,i+1};\, i = 1, m-1\}$. Let $h_{0,1}(\beta) = L_{0,1}$ and $h_{m,m+1}(\beta) = L_{m,m+1}$ define the first and the last piecewise linear functions, respectively. The sampling distribution is then $g_m(\beta) = \exp\{h_m(\beta)\}/M_m$, where $M_m = \int \exp\{h_m(\beta)\}d\beta$. We further define β_{cur} and β_M

as the current value and new sample from $f(\beta)$, respectively. The algorithm for ARMS is described below:

- Step 0: Initialize m and S_m;
- Step 1: Sample β from $g_m(\beta)$;

  $g_m(\beta) = \exp\{h_m(\beta)\}/M_m$ can be sampled by inverse CDF. The detail procedure is:

  - Sample U from uniform(0,1);
  - Find the corresponding value β from cumulative function of $g_m(\beta)$;
- Step 2: Sample U from Uniform$(0, 1)$;
- Step 3: If $U > \frac{f(\beta)}{\exp\{h_m(\beta)\}}$, then { rejection step;

  set $S_{m+1} = S_m \cup \{\beta\}$;

  relabel points in S_{m+1} in ascending order;

  increment m and go back to step 1;}

  else{

  acceptance step;

  set $\beta_A = \beta$;}
- Step 4: Sample U from Uniform$(0, 1)$;
- Step 5: If $U > \min\left\{1, \frac{f(\beta_A)\min[f(\beta_{cur}),\exp(h_m(\beta_{cur}))]}{f(\beta_{cur}),\min[f(\beta_A)\exp(h_m(\beta_A))]}\right\}$,

  then {Metropolis-Hastings rejection step;

  set $\beta_M = \beta_{cur}$;}

  else{Metropolis-Hastings rejection step;

  set $\beta_M = \beta_A$ }

- Step 6: Return β_M.

The procedure ARMS algorithm is summarized in Figure 2.1: ARMS starts from $h_{0,1}(\beta) = L_{0,1}$ and construct $h_{i,i+1}(\beta)$; we include a new point and relabel points; we then reconstruct $h_{i,i+1}$.

2.1.3 Simulating Annealing (SA)

SA employs a sequence of distributions, with probabilities or probability densities given by $f_T(\beta), f_{T-1}(\beta), \ldots f_1(\beta), f_0(\beta)$, in which each $f_j(\beta)$ differs slightly from $f_{j+1}(\beta)$. The main interest of SA is to obtain distribution $f_0(\beta)$. A scheme of SA has $f_{j+1}(\beta) \propto f_0(\beta)^{t_j+1}$ and $f_j(\beta) \propto f_0(\beta)^{t_j}$, for $\{t_j; 0 < t_T < \ldots < t_0 = 1\}$. The distribution $f_{t_j}(\beta)$ is obtained using the MCMC simulation. Let denote $f_0(\beta)$ as $f(\beta)$ for ease notation.

The annealing run is started at an initial state, from which we first simulate a Markov chain designed to converge to $f_T(\beta)$, for fixed T, e.g., in our example $T = 8$. We next simulate a certain number of iterations of a Markov chain designed to converge to $f_{T-1}(\beta)$, starting from the final state of the previous simulation. In a similar way, we simulate a number of iterations of a Markov chain designed to converge to $f_j(\beta)$, using the final state of the previous simulation for $f_{j-1}(\beta)$. Finally, we simulate the chain designed to converge to $f(\beta)$. The distribution of the final state produced by this process is close to $f(\beta)$.

When our target distribution, $f(\beta)$, has several multimodes, the simulated Markov chain starting from some arbitrary initial proposal distribution might converge to the local mode,

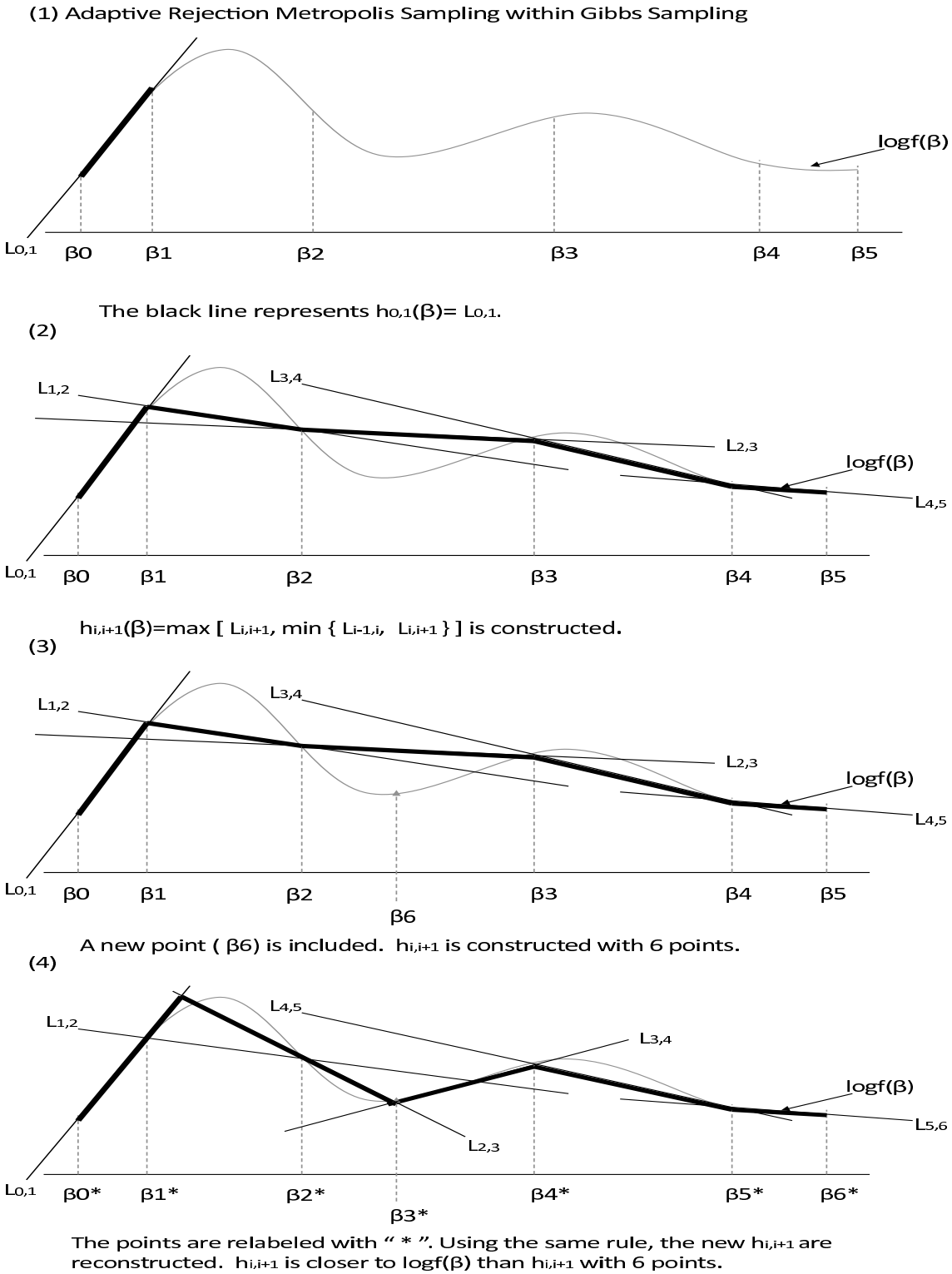


Figure 2.1: The Procedure of Adaptive Rejection Metropolis Sampling Within Gibbs sampling for a non-log-concave function $\log f(\beta) = \log f(\beta_p|\boldsymbol{\beta}_{(-p)})$, $k = 1, ..., s$: ARMS procedure starts from $h_{0,1}(\beta) = L_{0,1}$ and reconstruct adaptive rejection function $h_{i,i+1}(\beta) = \max[L_{i,i+1}, \min\{L_{i-1,i}, L_{i,i+1}\}]$ (black line)

which is close to an initial distribution. The annealing process can help to avoid this problem by considering the freer movement possibility. However, the limitation of SA is that it is difficult to choose sequences of proper proposal distributions for a target distribution.

2.2 Method of ARMS Annealing

2.2.1 ARMS Annealing

By incorporating ARMS into SA, we can overcome this limitation of SA. We refer to this approach as ARMS annealing approach. A traditional scheme of annealing is to set $f_j(\beta) = C_j f(\beta)^{t_j}$, where $C_j = \int f(\beta)^{t_j} d\beta$ and $t_j \in (0, 1]$. We note that $f_j(\beta)$ converges to a uniform distribution when t_j is close to zero. In our ARMS annealing approach, we select 10 points for t_j that are equally spaced points in (0,1]. The important step is that we substitute $f_j(\beta)$ and $f(\beta)$ to a piecewise linear function $h_{mj}(\beta)$ and $h_{0,1,j}$, where $h_{mj}(\beta) = \{h_{i,i+1,j}; i = 1, \ldots, m-1\}$ and m is the number of linear lines which construct the 'envelop'. In our study, we set m equals to 4. We define $h_{0,1,j} = L_{0,1,j}$ and $h_{m,m+1,j} = L_{m,m+1,j}$ as the first and last piecewise linear functions in the jth annealing. We can have $\exp\{h_{mj}(\beta)\} = C_j \exp\{h_{m0}(\beta)\}^{t_j}$ and $g_{mj}(\beta) = \exp\{h_{mj}(\beta)\}/M_{mj}$, where $M_{mj} = \int \exp\{h_{mj}(\beta)\} d\beta$.

Outline of the algorithm for ARMS annealing is described as follows. After then, the sub-algorithm of ARMS annealing is separately explained in detail:

- Step 0: Initialize β;

- Loop $p = 1$ to s {
  - Step 1: Set $f(\beta) = f(\beta_p|\beta_{(-p)})$;
  - Step 2: Initialize m and S_m;
  - Step 3: Set t_j, $t_j = \{t_T \text{ to } t_0 = 1\}$;
  - Step 4: Go to sub-algorithm of ARMS annealing, return β_M, $j = j - 1$;

    if $j \geq 0$ { go back to step 3;}

    else {keep $\beta_p = \beta_M$;

    go back to step 1; }

}

We now describe sub-algorithm of ARMS annealing in Step 4 as follows:

- Step 4.1: Sample β from $g_{mj}(\beta)$, where $g_{mj}(\beta) = \frac{1}{M_{mj}} \exp\{h_{mj}(\beta)\}$ and $M_{mj} = \int \exp(h_{mj}(\beta))$; $g_{mj}(\beta)$ can be sampled by inverse CDF. The detailed procedure is:
  - Sample U from uniform (0,1);
  - Find the corresponding value β from cumulative function of $g_{mj}(\beta)$;
- Step 4.2: Sample U from Uniform$(0, 1)$;
- Step 4.3: If $U > \frac{f_j(\beta)}{\exp\{h_{m,j}(\beta)\}}$, where $\exp\{h_{m,j}(\beta)\} = C_j \exp\{h_{m,0}(\beta)^{t_j}\}$,

then {rejection step;

set $S_{m+1} = S_m \cup \{\beta\}$

relabel points in S_{m+1} in ascending order;

increment m, and go back to Step 1;}
else{acceptance step;

set $\beta_A = \beta$;}

- Step 4.4: Sample U from Uniform$(0, 1)$;
- Step 4.5: If $U > \min\left\{1, \frac{f_j(\beta_A)\min[f_j(\beta_{cur}),\exp\{h_{m,j}(\beta_{cur})\}]}{f_j(\beta_{cur})\min[f_j(\beta_A),\exp\{h_{m,j}(\beta_A)\}]}\right\}$,
  then { Metropolis-Hastings rejection step;

  set $\beta_M = \beta_{cur}$.}
  else {Metropolis-Hastings rejection step;

  set $\beta_M = \beta_A$}
- Step 4.6: Return β_M.

In Step 4.5, we have $f_j(\beta) = C_j f(\beta)^{t_j}$ and $\exp\{h_{mj}(\beta)\} = C_j \exp\{h_{m0}(\beta)\}^{t_j}$, which mean that they are obtained using $f(\beta)$ and $h_{m0}(\beta)$. In addition, we note that C_j is cancelled out. The difference between ARMS and ARMS annealing algorithm is displayed Figure 2.2. ARMS annealing has been shrunk compared to ARMS.

2.2.2 EM ARMS Annealing Algorithm

In this section, we describe how ARMS annealing can be incorporated into the EM algorithm to estimate the parameters in the mixture model. Let us assume that our model

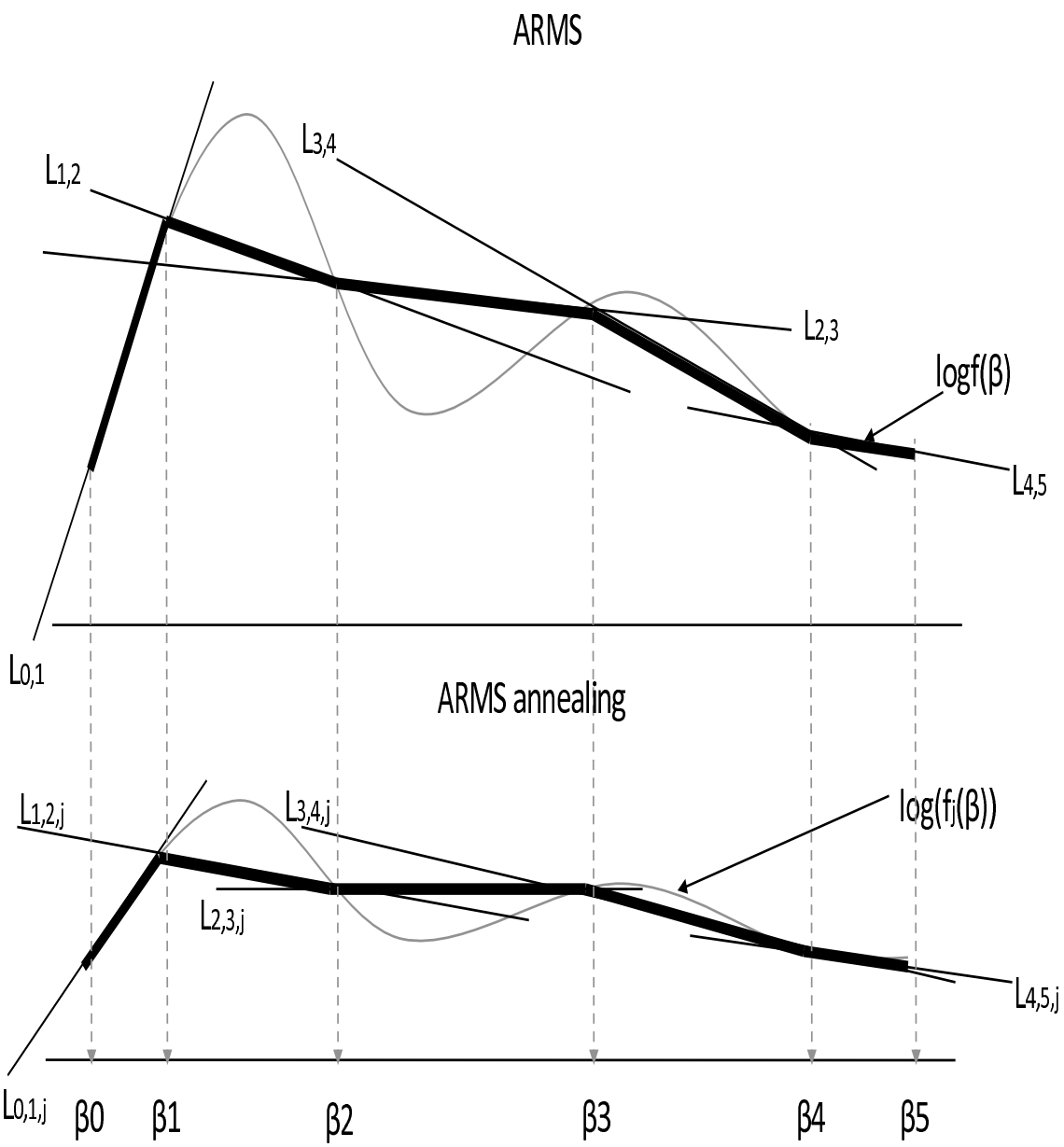


Figure 2.2: ARMS and ARMS annealing procedures: the figures (a) and (b) are ARMS and ARMS annealing procedures, respectively. The difference between the two procedures is that ARMS annealing has been shrunk compared to ARMS.

is a mixture of two Poisson regression models. Let y_i follow the mixture of two Poisson regression models and $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$ be the p vectors of the explanatory variables associated with the univariate outcome y_i, where $i = 1, \ldots, n$. Define $z_{1i} = 1$ if y_i is from Poisson distribution $f_1(\lambda_{1i})$ and $z_{1i} = 0$ if y_i is from the Poisson distribution $f_2(\lambda_{2i})$, where $\log(\lambda_{1i}) = \beta_{10} + \beta_{11}x_{1i} + \ldots + \beta_{1p}x_{pi}$ and $\log(\lambda_{2i}) = \beta_{20} + \beta_{21}x_{2i} + \ldots + \beta_{2p}x_{pi}$. The distribution of z_{ki} is $\text{Ber}(\pi_i)$, $k = 1, 2$, where $\pi_1 = \Pr(z_{1i} = 1)$ represents the proportion of observations following Poisson distribution $f_1(\lambda_1)$, and $\pi_2 = 1 - \pi_1$ stands for the proportion of observations following the other Poisson distribution, $f_2(\lambda_2)$. Define $\beta_1 = (\beta_{10}, \ldots, \beta_{1p})^T$, $\beta_2 = (\beta_{20}, \ldots, \beta_{2p})^T$, and $\theta = (\beta_1, \beta_2)$.

Let $Y_{obs} = (Y_i)$, $Y_{miss} = (Z_{ki})$, $Y_{complete} = (Y_{obs}, Y_{miss})$. We calculate the likelihood of complete data Y_c,

$$L_c(\theta, \pi) = \prod_{i=1}^{n} \{\pi_1 f_1(y_i|\lambda_{1i})\}^{z_{1i}} \{\pi_2 f_2(y_i|\lambda_{2i})\}^{1-z_{1i}}$$

We then incorporate ARMS annealing into the EM algorithm to estimate (θ, π). The procedures are summarized as follows:

- Step 1: Initialize π^0 and θ^0 using ARMS annealing;
- Step 2: We perform EM algorithm to estimate (θ, π);

  At the $(e+1)$th EM iteration,

– E-step is to obtain

$$E(z_{1i} = 1|Y_i = y_i) = Pr(z_{1i} = 1|Y_i = y_i) = \frac{\pi_1^e f_1(\lambda_{1j}^e)}{\pi_1^e f_1(\lambda_{1i}^e) + \pi_2 f_2(\lambda_{2i}^e)} = \tau_1(y_i; \theta^e)$$

where $\log(\lambda_{1j}^e) = \beta_{10.a}^e + \beta_{11.a}^e x_{1j} + \ldots + \beta_{1p}^e x_{pj}$ and $\log(\lambda_{2i}^e) = \beta_{20}^e + \beta_{21}^e x_{2j} + \ldots + \beta_{2p}^e x_{pj}$. Hence, we have $\pi_k^e = \frac{\sum_{i=1}^n \tau_k(y_i;\theta^e)}{n}$

The E-step can then be computed as

$Q(\theta^e, \pi_k^e) = \sum_{k=1}^2 \sum_{i=1}^n \tau_k(y_i; \theta^e)\{\log \pi_k^e + \log f_i(y_i; \theta^e, \pi_k^e)\}$

– M-step is to calculate the maximum likelihood estimates of θ using $Q(\cdot)$. These are solutions of the following equations, $\sum_{k=1}^2 \tau_k(y_i; \theta^e)\frac{\partial \log f_k(y;\theta)}{\partial \theta} = \mathbf{0}$.

– Repeat EM steps until (θ, π) converge;

- Step 3: Update the likelihood;
- Step 4: Repeat Steps 1-3, r times, e.g., $r = 50$ in our study;
- Step 5: Estimate global maximum likelihood estimators that have the largest likelihood.

2.2.3 Bayesian ARMS Annealing Approach

In this section, we describe how to incorporate ARMS annealing into the Bayesian approach to develop more effective MCMC simulation. We assume that our model is a mixture of two Poisson regression models. To complete the model specification, independent priors

are assumed, $z_{ki} \sim \text{Ber}(0.5)$ and $\theta \sim N(0, \sigma^2 I)$, where σ^2 is large fixed values, e.g., 100. The full conditional likelihoods are

$$[\beta_1|Y_{obs}, Z_{ji}, \beta_2] \propto \prod_{i=1}^{n} \{\pi_1 f_1(y_i|\lambda_{1i})\}^{z_{1i}} \{\pi_2 f_2(y_j|\lambda_{2i})\}^{1-z_{1i}} \pi(z_{1i}, \beta_1, \beta_2),$$

$$[\beta_2|Y_{obs}, Z_{ji}, \beta_1] \propto \prod_{i=1}^{n} \{\pi_1 f_1(y_i|\lambda_{1i})\}^{z_{1i}} \{\pi_2 f_2(y_j|\lambda_{2i})\}^{1-z_{1i}} \pi(z_{1i}, \beta_1, \beta_2),$$

$$[z_{ji}|Y_{obs}, \beta_1, \beta_2] \propto \prod_{i=1}^{n} \{\pi_1 f_1(y_i|\lambda_{1i})\}^{z_{1i}} \{\pi_2 f_2(y_j|\lambda_{2i})\}^{1-z_{1i}} \pi(z_{1i}, \beta_1, \beta_2).$$

We implement ARMS annealing in the Bayesian approach in the following way:

- Step 1: Initialize π^0 and θ^0 using ARMS annealing;
- Step 2: Using the initial values obtained from ARMS annealing we perform a Metropolis-Hastings algorithm to sample from the full conditional distributions;
  - At the rth ARMS annealing iteration,
    - Draw $\beta_1^{(r)}$ from $[\beta_1^{(r)}|Y_{obs}, Z_{ki}^{(r-1)}, \beta_2^{(r-1)}]$ using Metropolis Hastings;
    - Draw $\beta_2^{(r)}$ from $[\beta_2^{(r)}|Y_{obs}, Z_{ki}^{(r-1)}, \beta_1^{(r)}]$ using Metropolis Hastings;
    - Draw $z_{ki}^{(r)}$ from $[z_{ki}^{(r)}|Y_{obs}, \beta_1^{(r)}, \beta_2^{(r)}]$ using Metropolis Hastings;
  - Update $\pi^{(r)}$;
  - Increase r until the required the number of iterations;
- Step 3: Obtain the mode of each full conditional likelihood;

- Step 4: Update likelihood;
- Step 5: Repeat Steps 2-4;
- Step 6: Estimate parameters that give the largest likelihood.

2.3 Simulation

We have performed a simulation study to understand whether our two approaches capture the true underlying model well. For each approach, we compare two estimates obtained at two different modes: one is a global maximum mode, and the other is the second maximum local mode.

We use two covariates, x_1 and x_2 from our example data, where x_1 represents a income and x_2 stands for the status of volunteer, and standardize them by subtracting the mean and dividing the variance. We then generate y from the mixture of two Poisson regression models, $y_i \sim \pi_1 f_1(\lambda_{1i}) + (1-\pi_1) f_2(\lambda_{2i})$, $i = 1, \ldots, 1440$, where $\lambda_{1i} = \beta_{11} x_{1i} + \beta_{21} x_{2i}$ and $\lambda_{2i} = \beta_{12} x_{1i} + \beta_{22} x_{2i}$. Since we use the standardized covariates, we do not include the constant parameters in the model. We set the true value of the parameter as $(\pi_1, \beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}) = (0.7, 4, -1, 19, 0.8)$. We simulate 100 data sets. Each set has 1440 observations. For each simulated data set, the parameters were estimated using the EM ARMS annealing algorithm and the Bayesian ARMS annealing approach. For Bayesian approach, the priors are selected as follows: $z_{ki} \sim \text{Ber}(0.5)$ and $\theta \sim N(0, 100I)$, where $\theta = (\beta_1, \beta_2)$.

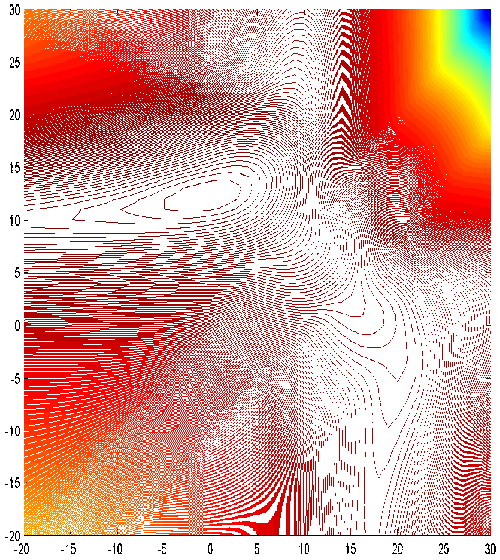


Figure 2.3: The contour plot of simulation data: The contour plot of β_{11} and β_{21} for the loglikelihood when $\hat{\pi}$, $\hat{\beta}_{12}$, and $\hat{\beta}_{22}$ are given.

The EM converges two different modes based on the initial values, including one global maximum mode. Given $\hat{\pi}$, $\hat{\beta}_{12}$, and $\hat{\beta}_{22}$, the contour plot of β_{11} and β_{21} for the loglikelihood is displayed in Figure 2.3.

This plot shows there are two modes. We additionally had several contour plots based on different given values. They showed two modes. Hence we do not display other contour plots in this article. We calculate the mean square errors of the parameters at two different modes. The global maximum estimates are $(\hat{\pi}_1, \hat{\beta}_{11}, \hat{\beta}_{21}, \hat{\beta}_{12}, \hat{\beta}_{22}) = (0.699, 3.936, -0.990, 19.012, 0.735)$, which are close to the true values; but the local maximum estimates are (0.563, 18.242, -2.611, 9.172, 1.969), which are far away from the true value. The average MSEs values are given in Table 2.1.

Method	Mode	true	0.7	4	-1	19	0.8
EM ARMS annealing	mode 1	est	0.6993	3.9363	-0.9901	19.0116	0.7353
		var	0.0002	0.1547	0.0053	0.0299	0.0007
		bias	-0.0007	-0.0637	0.0099	0.0116	-0.0047
		mse	0.0002	0.1587	0.054	0.0300	0.0007
		95%CI	[0.67,0.72]	[3.29,4.49]	[-1.10,-0.88]	[18.72,19.27]	[0.75,0.84]
	mode 2	est	0.5631	18.2421	-2.6107	9.1717	1.9694
		var	0.0002	0.0459	0.0043	0.0792	0.0011
		bias	-0.137	14.2420	-1.61070	-9.8280	1.1690
		mse	0.019	202.8840	2.5990	96.6740	1.3690
		95%CI	[0.54,0.58]	[17.83,18.53]	[-2.71,-2.51]	[8.73,9.65]	[1.92,2.02]
Bayesian ARMS annealing	mode 1	est	0.6779	3.8912	-0.9869	19.0031	0.7970
		var	0.0002	0.2683	0.0090	0.0375	0.0021
		bias	-0.0021	-0.1088	0.0131	0.0031	-0.0031
		mse	0.0002	0.2801	0.0092	0.0375	0.0021
		95%CI	[0.67,0.72]	[3.13,4.52]	[-1.12,-0.86]	[18.67,19.30]	[0.73,0.87]
	mode 2	est	0.5639	18.1972	-2.5953	9.1319	1.9720
		var	0.0002	0.1270	0.0334	0.1528	0.0034
		bias	0.5639	18.1972	-2.5953	9.1319	1.9720
		mse	0.3182	331.2664	6.7688	83.5445	3.8920
		95%CI	[0.54,0.59]	[17.78,18.57]	[-2.74,-2.48]	[8.66,9.61]	[1.89,2.05]

Table 2.1: Comparison of Average MSE using EM and Bayesian ARMS annealing: Average mean square errors of estimated parameters at two different modes using EM ARMS annealing and Bayesian ARMS annealing: mode 1 is the global maximum mode and mode 2 is the second maximum local mode.

The average MSEs of the estimated parameters at the global maximum mode are (0.0002, 0.158, 0.054, 0.0300, 0.001), which are quite small. However, these values at the local maximum mode are (0.019, 202.884, 2.599, 96.674, 1.369). These results suggest that EM alone is sensitive on initial values and can converge to the local maximum points. However, using the EM ARMS annealing, we can detect the global maximum mode, which gives the global maximum estimates. We also fit the model using the Bayesian ARMS annealing approach. The average MSEs values of the Bayesian ARMS annealing are summarized in Table 2.1. The average MSEs values at two different modes are (0.0002, 0.280, 0.009, 0.038, 0.002) and (0.318, 331.266, 6.769, 83.545, 3.892), respectively. The 95% Bayesian credible intervals are also included in Table 2.1.

In order to compute the 95% confidence interval or Bayesian credible intervals, 100 data

sets of y are generated from Poisson distribution for given x_1, x_2, and β's; then ARMS annealing is implemented to get proper start points for each data set; then the point estimation for β's is computed by both EM algorithm and Bayesian approach. For Bayesian approach we use median value of posterior samples as point estimator; By repeating this procedure for each data set we can construct the percentile based intervals.

These results also explain to us that the Bayesian approach alone can converge to the local mode, but the Bayesian ARMS annealing can also detect the maximum global mode.

Overall, EM ARMS annealing and Bayesian ARMS annealing are comparable to one another in terms of mean squares error although EM ARMS annealing has a slightly smaller MSE than Bayesian ARMS annealing.

2.4 Application

The data used in this paper are from a survey of household giving conducted from July 3 through July 17, 2002, in Korea. The survey was conducted on a nation wide sample of 1,456 individuals over 20 years old by means of individual interviews. The sample was made based on the proportions of gender and regions across the country, except for Jeju island in Korea. The data on charitable giving refers to the 2001 monetary giving of the respondent's household for the calendar year. The percentages of respondents in terms of seven covariates and people participating in charitable giving are summarized in Tables 2.2-2.3.

In Korea, a growing number of people involved in charitable activities have become

Variable	Category	The percentage of respondent	The percentage of people participating in charitable giving
Sex	Male	50.4%	52.7%
	Female	49.6%	49.7%
Age	20s	22.6%	44.7%
	30s	31.9%	49.8%
	40s	24.4%	57.5%
	50s	21.2%	50.3%
Income (million)	Less than 1.5	27.4%	27.4%
	1.5-2.5	40.1%	49.5%
	2.5-4.0	24.8%	60.1%
	More than 4	5.4%	53.9%
	No response	2.3%	
Education	Less than junior high school	18.3%	
	Less than senior high school	45.9%	
	College or more	35.9%	
Religion	Buddhism	26.0%	
	Protestant	20.0%	
	Catholic	8.2%	
	None	45.8%	
Region	Large cities	49.0%	
	Small&medium cities	37.4%	
	Rural areas	13.6%	
Occupation	Agriculture/Forestry/Fishery	6.6%	
	Self-employed	18.8%	
	Bule collar	17.5%	
	White collar	17.0%	
	Housewives	25.8%	
	Students	8.9%	
	Other	5.3%	

Table 2.2: The percentage of people in participating charitable giving in survey data

interested in issues, such as how to promote charitable giving among Korean people. That is, how can one help nonprofit organizations conduct successful fund raising, and how can those organizations be supported in fund raising by more effective government policies? This is related to a recent trend in Korean society in which nonprofit activities are no longer regarded to be sole responsibility of the government, and the civil sector should be more active in carrying out nonprofit activities to meet social demand more efficiently in a partnership with the government. Therefore, research needs to be done concerning the charitable giving behavior of the Korean people to provide some empirical findings that will give solid answers to those questions related to charitable giving. Several papers have

Types of organizations	The percentage of people participating in charitable giving in each organization
Religious organizations(for the purpose of help for the poor)	43.0%
Social welfare organizations	23.2%
Educational institutions	8.4%
Environmental organizations	4.5%
Private public-interest organizations	4.1%
International organizations	3.7%
Health and medical organizations	2.7%
Corporate and private foundations	2.8%
Arts, culture, and sports organizations	2.2%
Youth organizations	1.9%
Recreation organizations	1.1%
Others	58.8%

Table 2.3: The percentage of people participating in charitable giving to organizations in survey data

explained what covariates affect the amount of giving (Smith et al., 1999; Duncan, 1999). However, no one has discovered what covariates affect the number of charitable giving. The aim of this study is to find what these covariates affect the number of charitable giving. There are six covariates: income (with four categories), volunteering experience (1: yes, 0: no), attitude based on religious belief (1: yes, 0: no), education (1: college or more, 0: otherwise), age (1: 50s, 0: otherwise), and sex (1: male, 0: female).

The histogram of the number of charitable donations shows that there are two different modes (Figure 2.4). We consider the mixture of the two Poisson distributions to model the number of charitable donations based on the histogram of the data. The mixture of the two Poisson distributions implies that the whole population is subdivided into two groups, one with a lesser number of charitable donations and the other with a larger number of charitable donations. We fit the mixture of Poisson regression models into the number of charitable donations to identify significant covariates. The EM ARMS annealing algorithm and the Bayesian ARMS annealing approach are employed to estimate the parameters. The

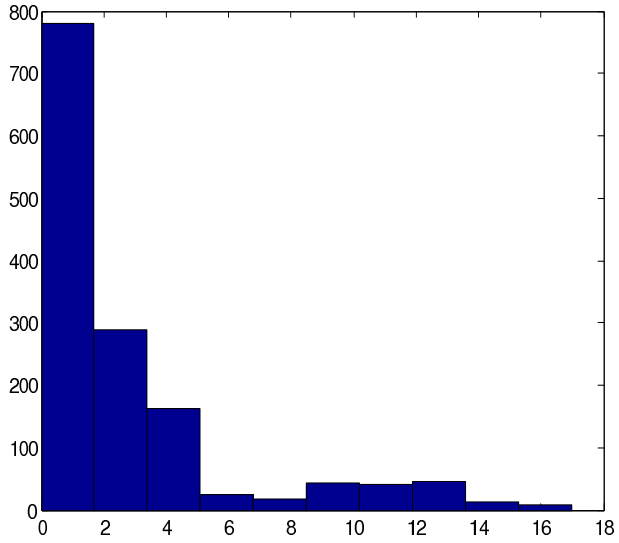


Figure 2.4: Histogram of the number of charitable donations in survey data

estimated values obtained by the EM ARMS are given in Table 2.4.

The 95% Bayesian credible interval obtained by the Bayesian ARMS annealing is shown in Table 2.5. Although EM alone converges to seven different modes based on initial values, our EM ARMS annealing can detect the global maximum mode and can estimate parameters at the maximum mode. We found that the Bayesian approach alone tends to stay around the local mode, but our Bayesian ARMS annealing approach converges to the global maximum mode more quickly because our approach can more easily move one mode to the other. We note that the two results obtained from our two approaches are similar to each other. Figure 2.6 has shown the scatter plot between the posterior samples of the mixing proportion π_1 and the posterior samples of each parameter. The points "x" represent the posterior samples of each parameter obtained from the Bayesian ARMS annealing approach, and the rectangular points represent the estimator obtained from the EM ARMS annealing approach. We notice

		Global max. mode	local max. mode					
Variable	parameters	1	2	3	4	5	6	7
	π_1	0.6293	0.5562	0.5389	0.5124	0.5192	0.5225	0.5841
Intercept	β_{11}	-2.9088	-2.8693	1.3027	-2.6039	1.2451	-2.4366	1.3200
Income	β_{21}	9.9875	2.3137	1.4103	3.6046	1.9400	2.5158	1.2743
Volunteer	β_{31}	1.4823	4.0462	0.1279	0.4741	0.1924	0.2535	0.1814
Religion	β_{41}	-0.2814	0.5715	0.2634	0.3997	-2.5980	0.2915	0.2033
Education	β_{51}	-0.0971	-0.1053	-0.0799	-0.0604	-0.0961	3.4661	-2.1776
Age	β_{61}	0.4508	0.1468	0.0232	3.6119	0.0070	-0.0727	-2.0575
Sex	β_{71}	0.4612	0.2508	-2.8044	-0.0583	0.0123	0.0745	-0.0293
Intercept	β_{12}	1.3291	1.4401	-3.2417	1.3487	-2.8483	1.4943	-3.5462
Income	β_{22}	2.3648	1.0974	1.1376	1.2924	1.0038	0.5847	2.7997
Volunteer	β_{32}	0.3937	-1.8581	0.2861	0.1998	0.2862	0.1915	0.1873
Religion	β_{42}	0.1352	0.1929	0.2390	0.2454	4.4964	0.1925	0.2928
Education	β_{52}	-0.0764	-0.0781	-0.0153	-0.0422	-0.0106	-3.2835	4.7407
Age	β_{62}	0.0709	-0.0491	-0.0341	-2.7979	0.0420	-0.0123	4.8975
Sex	β_{72}	0.0414	-0.0395	4.6384	0.0012	-0.0196	-0.0555	0.0266
	log-likelihood	-2955.9	-2971.9	-3007.0	-3009.5	-3011.4	-3012.3	-3186.5

Table 2.4: Parameter estimation using EM ARMS annealing algorithm for fitting the mixture of two Poisson regressions in survey data: income with four categories, volunteering experience (1: yes, 0: no), attitude based on the religious belief (1: yes, 0: no), Education (1: College or more, 0: otherwise) ,age (1 : 50s, 0:otherwise), and sex (1:male, 0:female).

that the areas in which the posterior samples are located include the estimators of the EM ARMS annealing approaches, which implies that the two approaches give similar results.

Figure 2.5 shows estimated parameters obtained using both EM ARMS annealing and Bayesian ARMS annealing, and 95% Bayesian credible interval using Bayesian ARSM annealing approach. We found the two methods give us similar estimation results. 95% Bayesian credible interval using Bayesian ARSM annealing approach covered the estimation from EM ARMS annealing approach.

The 95% Bayesian credible intervals obtained using Bayesian ARMS annealing are shown in Table 2.5. As a result, we found that the income variable with four categories and the volunteering variable (1: experience of volunteering, 0: otherwise) turned out to be significant with positive regression coefficients in both the lesser and larger donation groups.

		Bayesian	Credible	Interval	
	Parameter	Bay-low bound 5%	Bay-global mode	Bay-upper bound 95%	EM
	π	0.609	0.626	0.653	0.6293
Intercept	β_{11}	-3.16	-2.68	-2.04	-2.9088
Income	β_{21}	1.58	6.14	13.19	9.9875
Volunteer	β_{31}	1.19	1.43	1.61	1.4823
Religion	β_{41}	-0.33	-0.13	0.12	-0.2814
Education	β_{51}	-0.2	-0.03	0.13	-0.0971
Age	β_{61}	0.07	0.27	0.46	0.4508
Sex	β_{71}	0.24	0.39	0.59	0.4612
Intercept	β_{12}	1.20	1.38	1.57	1.3291
Income	β_{22}	0.27	1.53	3.17	2.3648
Volunteer	β_{32}	0.31	0.39	0.46	0.3937
Religion	β_{42}	0.09	0.15	0.27	0.1352
Education	β_{52}	-0.14	-0.06	0.02	-0.0764
Age	β_{62}	-0.06	0.05	0.15	0.0709
Sex	β_{72}	-0.04	0.03	0.1	0.0414

Table 2.5: 95% Bayesian credible interval obtained from Bayesian ARMS annealing approach in survey data

The credible intervals of income in the lesser and larger donation groups are [1.58, 13.19] and [0.27, 3.17], respectively. These intervals of the volunteering variable are [1.19,1.61] and [0.31,0.46], respectively. We also found that age (1: 50s, 0: otherwise) and sex (1: male, 0: female) are significant, with credible intervals [0.07,0.46], [0.24,0.59], respectively, and with positive regression coefficients in the lesser donation group, but not in the larger donation group. On the other hand, in the larger donation group, attitude based on religious belief (1: yes, 0: no) is identified as a significant variable with credible interval [0.09,0.27] and with the positive regression coefficients, but not in the lesser group.

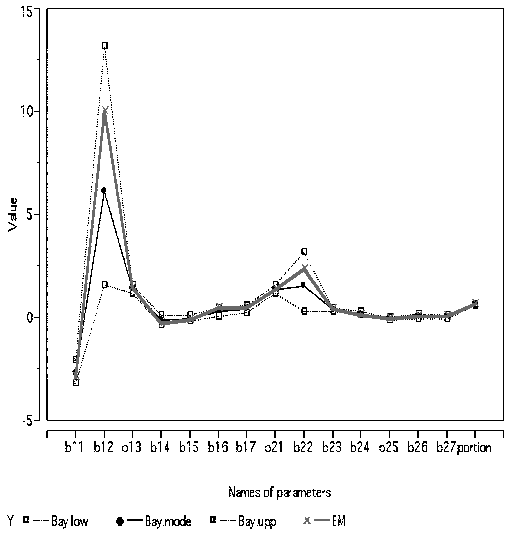


Figure 2.5: Estimated parameters obtained using both EM ARMS annealing and Bayesian ARMS annealing, and 95% Bayesian credible interval using Bayesian ARSM annealing approach in survey data

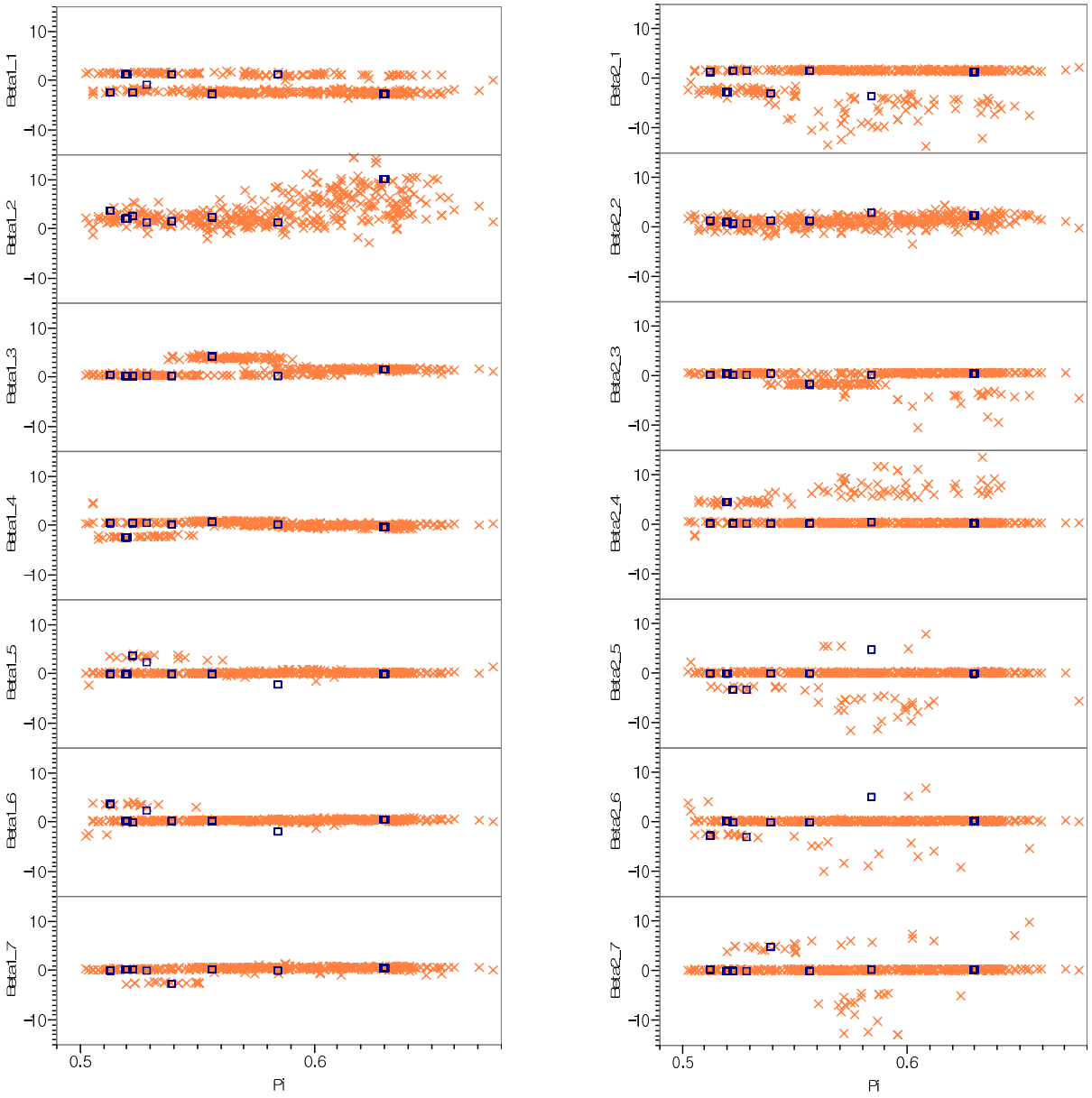


Figure 2.6: The scatter plot between posterior samples of the mixing proportion π and posterior samples of each parameter: the point "x" represents posterior samples of each parameter obtained from Bayesian ARMS annealing approach and the rectangular points represent estimator obtained from EM ARMS annealing approach. The area located posterior samples includes the estimators of EM ARMS annealing approaches which implies that the two approaches give similar results.

Chapter 3

Bayesian Model Selection for the NLME Model

3.1 Introduction

The nonlinear mixed effects models (NLME) are commonly used in agricultural, environmental, and biomedical applications to analyze repeated measurement data (Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997). Continuous responses evolve over time within individuals from a population of interest. The NLME model accommodates both the variation among measurements within individuals and the individual-to-individual variation. The NLME model is a mixed effects model in which some or all of the fixed and random effects occur nonlinearly in the model function.

Different methods have been proposed to estimate the parameters in the NLME model. Because the model function is nonlinear in the random effects, the integral for the marginal likelihood generally does not have a closed-form expression. To make the numerical optimization of the likelihood function a tractable problem, different approximations have been proposed. Some of these methods consist of a first-order Taylor expansion of the model function around the expected value of the random effects (Sheiner and Beal, 1980; Vonesh and Carter, 1992) or around the conditional modes of the random effects (Lindstrom and Bates, 1990). Gaussian quadratic rules are also used (Davidian and Gallant, 1992). The NLME model can be fitted using a global two-stage method (Steimer et al., 1984), an Expectation Maximization algorithm (Dempster et al., 1977), and a Bayesian approach (Gelman et al, 1998; Ibrahim et al, 2001). However, most of these methods require strong assumptions on both measurement errors and individual-specific parameters which limit the ability to measure heterogeneity errors and the variation among subjects. Furthermore, these assumptions on measurement errors or individual-specific parameters are often not satisfied in real applications.

Kleinmanm et al (1998) proposed a semiparametric Bayesian approach to the linear random effects model, which assume a Dirichlet process prior (Ferguson, 1973) on the parameters for individual subjects. We further developed this method for nonlinear mixed effects models and refer to it as a semiparametric nonlinear model with one-layer DP random effects. Further research works on semiparametric linear models with DP measurement errors (Escobar and West, 1995; Chib et al, 2010), and we call this method a linear DP

measurement error model. NLME models with DP random effects or measurement errors are rarely discussed in the previous research because of several reasons. Since one-layer DP random effects models cluster parameters on the individual parameter level, it is not suitable for comparison with a parametric random effects model. There is another difficulty when implementing one-layer DP random effects. The posterior distributions of individual parameters involve the marginal likelihood computation, which does not have a closed form for NLME. An approximation of the marginal likelihood is time consuming and sometimes is not stable, which make less the attraction.

Different from one-layer DP random effects model, we propose a two-layer DP random effects model, where subjects are from several subgroups, and random effects exist within each subgroup. In the other words, the population parameters come from a mixing distribution, instead of assuming a mixing distribution on individual parameters. Hence, it is reasonable to assume a Dirichlet process prior on the population parameter. This two-layer DP random effects model is suitable for the comparison with parametric NLME models and NLME DP measurement error models. The advantage of this model obtained is the closed form of the marginal likelihood if we choose the proper priors.

The motivation to develop semiparametric Bayesian hierarchical models is from our gastric emptying studies which are important in human and veterinary medical research. They evaluate medications or diets for promoting gastrointestinal motility and to examine unintended side-effects of new or existing medications, diets, and other procedures or interventions. The way gastric emptying data is summarized is important for establishing the

validity of gastric emptying studies, for allowing easier comparison between treatments or between groups of subjects, and for comparing results among studies. For the analysis of the data from gastric emptying studies, several nonlinear models were proposed based on an exponential, a double exponential, a power exponential, and a modified power exponential functions (Elashoff, 1982; 1983). The power exponential model is one of the most popular models for analyzing the gastric emptying data. The power exponential model is

$$y_{ij} = A_{0i} 2^{-(\frac{t_{ij}}{t_{50_i}})^{\beta_i}} + \epsilon_{ij}, \tag{3.1}$$

where y_{ij} represents the meal, drug, or other types remaining in the stomach at j^{th} time point of the i^{th} subject where $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$. The t_{ij} is the corresponding sampling time and A_{0i} stands for the amount of the remaining in the stomach at time 0. The t_{50_i} represents the parameter of time at which one-half of the meal or other types present at time 0 remains in the stomach of the i^{th} subject. The β_i is the shape parameter of the decreasing curve of the i^{th} subject. For $\beta_i = 1$, the power exponential is the same as the exponential model. A value of $\beta_i > 1$ describes a curve with an initial lag in emptying the gastric content. This type of curve is often seen for gastric emptying of a solid meal emptying (solid-phase emptying), where the initial lag phase may represent the time required to grind the food or treatment into smaller particles. A value of $\beta_i < 1$ describes a curve with a very rapid initial emptying, followed by a second slower emptying phase. Such a pattern is often seen for liquid-phase emptying (i.e., the emptying of a liquid meal).

While the power exponential model may provide an adequate summary of the gastric emptying for an individual patient, one loses information and statistical power by not using all of the observations made for an individual. This is particularly important when studies involve a relatively small number of subjects to evaluate the effects of drugs or diets. Therefore, we considered the application of a random coefficient regression model (Davidian and Giltinan, 1995) which accommodates both the variation among measurements within individuals and the individual-to-individual variation. In addition, because individuals studied may not be derived from a homogeneous subpopulation, a single regression model likely cannot adequately fit the data from heterogeneous subpopulations. In our study from equine medicine, we actually observed heterogeneity of measurement errors and also observed the variation among subjects.

To handle these problems, we propose semiparametric Bayesian NLME models: the first has a DP prior on measurement errors which may vary from subject to subject, and we refer to this model as the "NLME model with DP measurement errors". The second model has a DP prior on individual random effects parameters and we refer to this model as the "NLME model with one-layer DP random effects". The third model has a DP prior on population random effects parameters, and we call this model the "NLME model with two-layer DP random effects". A parametric NLME is also proposed in terms of the purpose of model comparisons.

Thus, our goal for this study is to propose three semiparametric Bayesian hierarchical models for NLME, to propose Gibbs sampling to estimate parameters in the NLME models,

and to develop a unified approach for model selection. Different model selection methods are discussed, such as Bayes factor, cross validation, posterior Bayes factor, and our proposed "penalized posterior Bayes factor" for our NLME models. Our semiparametric Bayesian hierarchical models are based on DP priors on measurement errors and on random effects. We compare them with a parametric Bayesian hierarchical model.

The topic is organized as follows. In Chapter 3.2, we describe four Bayesian hierarchical models with NLME: the first has parametric priors, the second has DP measurement errors, the third has one-layer DP random effects, and the fourth has two-layer DP random effects. In Chapter 3.3, we describe how to estimate parameters in each model. In Chapter 3.4, we discuss different model selection methods. In Chapter 3.5, we report the results of simulations comparing three Bayesian hierarchical models (not including one-layer DP random effects model). In Chapter 3.6, we apply our approaches to an equine gastric emptying study.

3.2 Nonlinear Mixed Effects Models

Based on (3.1), let $f(t_{ij}, \phi_i)$ represent a nonlinear function characterizing the relationship between y_{ij} and t_{ij}, with ϕ_i being the subject-specific regression parameters, where $\phi_i = (\beta_i, t_{50_i})^{\mathrm{T}}$ and $f_{ij} \equiv f(t_{ij}|\phi_i) = A_{0i}2^{-(\frac{t_{ij}}{t_{50_i}})^{\beta_i}}$. The measurement error ϵ_{ij} is assumed to be independent and identically distributed, $[\epsilon_{ij}|\lambda_e] \sim \mathrm{N}(0, \lambda_e^{-1})$ with $\lambda_e = \sigma_e^{-2}$. In this Chapter, we will use $[X]$ as the sampling distribution of X, and $p(X)$ as the probability of X. Individual parameters $(\beta_i, t_{50_i})^{\mathrm{T}}$ follow a lognormal distribution with population parameters

$(\beta, t_{50})^{\mathrm{T}}$ and Λ,

$$\begin{pmatrix} \beta_i \\ t_{50_i} \end{pmatrix} \sim \mathrm{LN}\left\{ \begin{pmatrix} \log(\beta) \\ \log(\mathrm{t}_{50}) \end{pmatrix}, \Lambda^{-1} \right\},$$

where $\beta_i \geq 0$, $t_{50_i} \geq 0$, and a 2×2 covariance matrix $\Sigma = \Lambda^{-1}$. $(\beta_i, t_{50_i})^{\mathrm{T}}$ follows the lognormal distribution which is represented by "LN". Then $\{\log(\beta_{\mathrm{i}}), \log(\mathrm{t}_{50_{\mathrm{i}}})\}^{\mathrm{T}}$ is from a normal distribution with the mean, $\{\log(\beta), \log(\mathrm{t}_{50})\}^{\mathrm{T}}$, and the variance, Λ^{-1}.

Let us define $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $Y_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})^{\mathrm{T}}$, $f_i = (f_{i1}, \ldots, f_{ij}, \ldots, f_{in_i})^{\mathrm{T}}$, $\Phi = (\phi_1, \ldots, \phi_i, \ldots, \phi_k)^{\mathrm{T}}$, and $\theta = (\beta, t_{50})^{\mathrm{T}}$.

We consider the following four Bayesian hierarchical models:

- Model 1: Parametric Bayesian hierarchical model

  We assume that the measurement error is normally distributed with homoscedastic errors having $\lambda_e = \sigma_e^{-2}$, that is, $[\epsilon_{ij}|\lambda_e] \sim \mathrm{N}(\epsilon_{ij}|0, \lambda_e^{-1})$. Prior distributions of the mean and precision of random effects and λ_e are following,

$$\begin{aligned} \begin{pmatrix} \beta \\ t_{50} \end{pmatrix} &\sim \mathrm{LN}\left\{ \begin{pmatrix} \beta_0 \\ t_{50_0} \end{pmatrix}, (\Lambda_{00})^{-1} \right\}, \\ [\Lambda] &\sim \mathrm{Wishart}(\Lambda|\frac{\Lambda_0}{\tau}, \tau), \\ [\lambda_e] &\sim \mathrm{Gamma}(\lambda_e|\frac{v_0}{2}, \frac{v_0 s_0^2}{2}). \end{aligned}$$

The parameters for the lognormal distribution are given by $(\beta_0, t_{50_0})^{\mathrm{T}}$ and Λ_{00}, and both of them are constants which we gave. The prior for the mean of the population precision, Λ, is given by Λ_0. The degree of freedom for the prior distribution is τ. τ need to be larger than 1 for our case. λ_e is the precision of measurement error. The prior mean of λ_e is ${s_0}^{-2}$, and $(v_0 s_0^2)/2$ is the rate parameter for the gamma distribution.

- Model 2: Semiparametric Bayesian NLME model with DP measurement errors

  We consider a situation, where the measurement error may be a heteroscadesitic error, follow a mixture distribution, or may be clustered into several groups with a certain mean shift and different variance. That is, $\epsilon_{ij} \sim \mathrm{N}(\mu_i, \lambda_{ei}^{-1})$ with $\lambda_{ei} = \sigma_i^{-2}$, where $\psi_i = (\mu_i, \lambda_{ei})$ is a hyerparameter. The hyperparameter, ψ_i, is following unknown distribution, which is an unknown probability measure over $(-\infty, +\infty) \times (0, +\infty)$. We consider G is the unknown distribution and the prior of G is following the Dirichlet process (DP) with the concentration parameter, α, and the base distribution, G_0. This leads to an error distribution that is an arbitrary location-variance mixture of normal distributions and can be clustered into several groups. This Bayesian hierarchical model can be written as follows:

$$\begin{aligned} [\epsilon_{ij}|\psi_i] &\sim \mathrm{N}(\epsilon_{ij}|\mu_i, \lambda_{ei}^{-1}), \\ [\psi_i|G] &\sim G, \\ [G] &\sim \mathrm{DP}(\alpha, \mathrm{G}_0), \end{aligned}$$

where $G_0 = \mathrm{N}\{\mu_i|0,(g\lambda_{ei})^{-1}\}\mathrm{Gamma}\{\lambda_{ei}|v_0/2,(v_0 s_0^2)/2\}$, and g, v_0, and s_0^2 are given. Note that if we rewrite $\psi_i = \mu_i + \mathrm{N}(0,\lambda_{\mathrm{ei}}^{-1})$, we consider μ_i plays a role as a random effect.

- Model 3: Semiparametric Bayesian NLME model with one-layer DP random effects

  We assume that random effect $\phi_i = (t_{50i},\beta_i)^{\mathrm{T}}$ follows an unknown distribution, G, which comes from a Dirichlet process, $\mathrm{DP}(\alpha, G_0)$. The base distribution, G_0, of DP is defined as $G_0 = \mathrm{LN}\{\theta_0,(\Lambda_0)^{-1}\}$, where $\theta_0 = (\beta_0, t_{50_0})^{\mathrm{T}}$. β_0, t_{50_0} and Λ_0 are all specified in advance. We summarize our model as follows,

$$\begin{aligned} \phi_i &\sim G, \\ G &\sim \mathrm{DP}(\alpha, G_0), \\ G_0 &= \mathrm{LN}\{\theta_0,(\Lambda_0)^{-1}\}. \end{aligned}$$

- Model 4: Semiparametric Bayesian NLME model with two-layer DP random effects

  Random effects are often variant from subject to subject. Hence we assume that the random effect $\phi_i = (t_{50_i},\beta_i)^{\mathrm{T}}$ follows $\mathrm{LN}\{\phi_{\mathrm{i}}|\omega_{\mathrm{i}} = (\theta_{\mathrm{i}},\Lambda_{\mathrm{i}})^{\mathrm{T}}\}$, and the prior distribution of ω_i is a Dirichlet process prior, $\mathrm{DP}(\alpha, G_0)$, where $\phi_i \sim \mathrm{LN}\{\phi_{\mathrm{i}}|\omega_{\mathrm{i}} = (\theta_{\mathrm{i}},\Lambda_{\mathrm{i}})^{\mathrm{T}}\}$, $\omega_i \sim G$, and $G \sim \mathrm{DP}(\alpha, G_0)$.

$$\begin{aligned} G_0 &= [\theta_i][\Lambda_i] \\ &= \mathrm{LN}\{\theta_{\mathrm{i}}|\theta_0,(\Lambda_{00})^{-1}\}\mathrm{Wishart}(\Lambda_{\mathrm{i}}|\Lambda_0/\tau,\tau). \end{aligned}$$

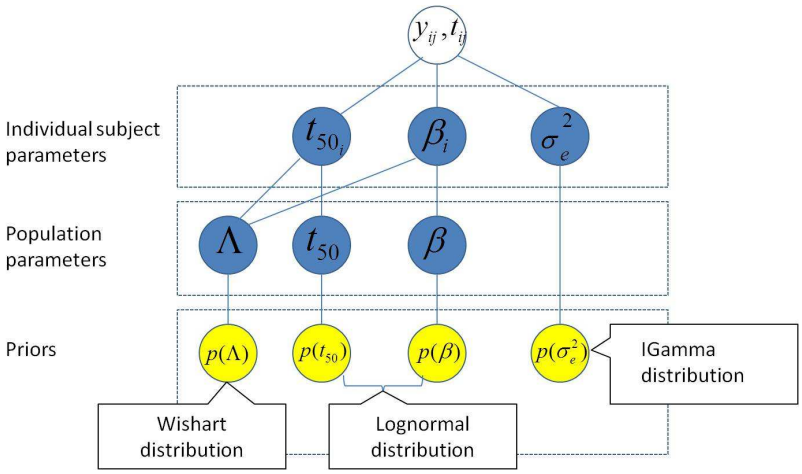


Figure 3.1: The relationship among the data, unknown parameters and their priors for Model 1 are graphically represented.

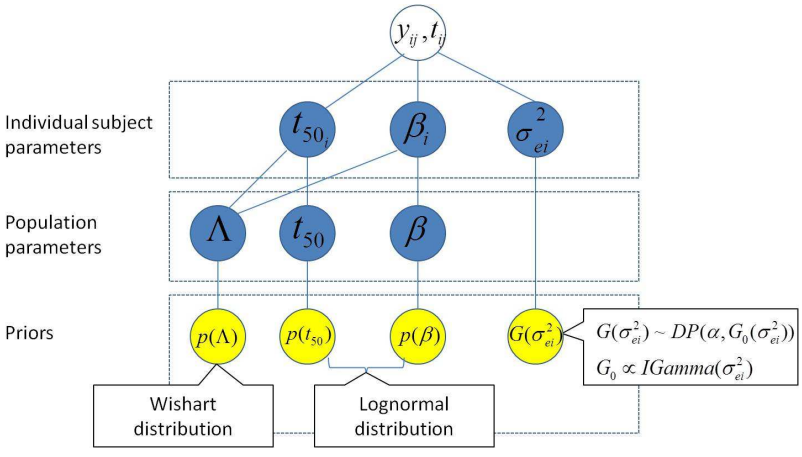


Figure 3.2: The relationship among the data, unknown parameters and their priors for Model 2 are graphically represented.

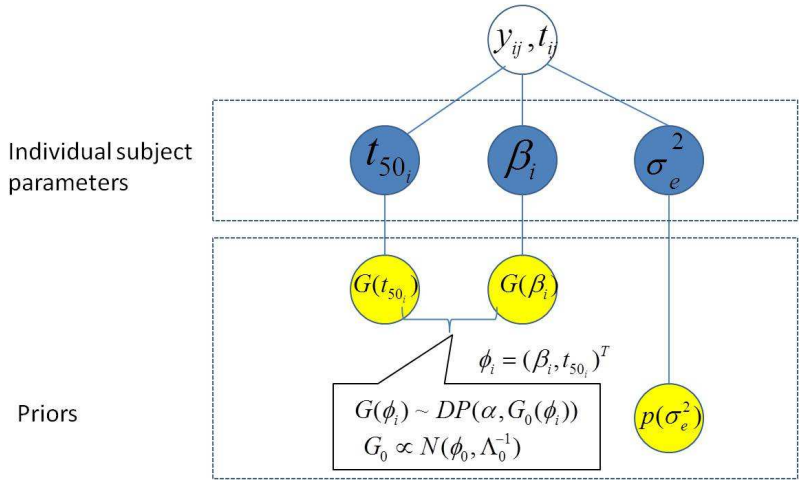


Figure 3.3: The relationship among the data, unknown parameters and their priors for Model 3 are graphically represented.

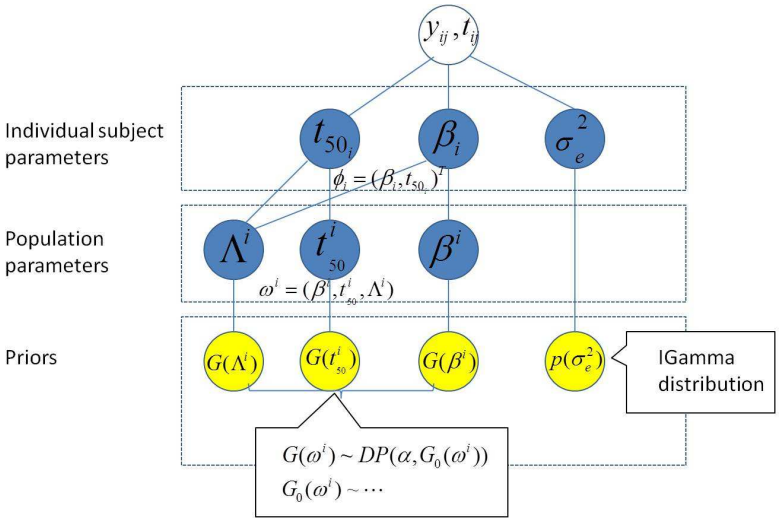


Figure 3.4: The relationship among the data, unknown parameters and their priors for Model 4 are graphically represented.

3.3 Estimation of NLME Model

3.3.1 Model 1: Parametric Bayesian Model

We first derive the full conditional distribution of parameters and explain how to sample from the full conditional distributions. The joint distributions of Y, Φ, θ, Λ, and λ_e can then be written as:

$$\begin{aligned}
[Y,\Phi,\theta,\Lambda,\lambda_e] &= [Y|\Phi,\lambda_e]\times[\Phi|\theta,\Lambda]\times[\theta|\Lambda]\times[\Lambda]\times[\lambda_e] \qquad (3.2)\\
&\sim (\lambda_e)^{\frac{\sum n_i}{2}}\exp\left\{-\lambda_e\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-f_{ij})^2}{2}\right\}\times\\
&\quad |\Lambda|^{\frac{k}{2}}(\prod^{k}\phi_i)^{-1}\exp[-\frac{1}{2}\sum_{i=1}^{k}\{\log(\phi_i)-\log(\theta)\}^{\mathrm{T}}\Lambda\{\log(\phi_i)-\log(\theta)\}]\times\\
&\quad |\Lambda_{00}|^{1/2}(\theta)^{-1}\exp[-\frac{1}{2}\{\log(\theta)-\theta_0\}^{\mathrm{T}}\Lambda_{00}\{\log(\theta)-\theta_0\}]\times\\
&\quad |\Lambda|^{\frac{\tau-d-1}{2}}\exp\{-\frac{1}{2}\mathrm{Tr}(\tau\Lambda_0^{-1}\Lambda)\}\times\\
&\quad (\lambda_e)^{\frac{v_0}{2}-1}\exp(-\frac{v_0 s_0^2}{2}\lambda_e).
\end{aligned}$$

Marginal distributions, $[\Phi|Y]$, $[\theta|Y]$, $[\Lambda|Y]$, and $[\lambda_e|Y]$, are obtained using both Metropolis-Hastings sampling and Gibbs sampling. The full conditional distributions of Φ, λ_e, θ, and Λ are obtained as follows:

The full conditional distribution of λ_e has a closed form,

$$[\lambda_e|\theta,\Phi,\Lambda,Y]=[\lambda_e|\Phi,Y]=(\lambda_e)^{\frac{\sum_{i=1}^{k}n_i+v_0}{2}-1}\exp\{-\lambda_e[\frac{v_0 s_0^2}{2}+\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\{y_{ij}-f(t_{ij}|\phi_i)\}^2}{2}]\},$$

which follows Gamma(a^d, b^d) and the posterior mean for λ_e is a^d/b^d. $a^d = (\sum_{i=1}^{k} n_i + v_0)/2$ and $b^d = (v_0 s_0^2)/2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \{y_{ij} - f(t_{ij}|\phi_i)\}^2/2$.

The full conditional distribution of Λ also has a closed form,

$$[\Lambda|\theta, \Phi, \lambda_e, Y] = [\Lambda|\Phi, \theta] = |\Lambda|^{\frac{k+\tau-d}{2}} \exp[-\frac{1}{2}\mathrm{Tr}\{(\mathrm{S} + \tau {\Lambda_0}^{-1})\Lambda\}],$$

which follows a wishart distribution with $\Lambda^d = (S + \tau {\Lambda_0}^{-1})^{-1}$ and $k + \tau$ degree of freedom, where $S = \sum_{i=1}^{k} \{\log(\phi_i) - \log(\theta)\}\{\log(\phi_i) - \log(\theta)\}^{\mathrm{T}}$. The posterior mean of Λ is $\Lambda^d/(k + \tau)$.

The full conditional distribution of θ has a closed form as well,

$$[\theta|\Phi, \lambda_e, \Lambda, Y] = [\theta|\Phi, \Lambda] \sim \mathrm{LN}(\theta_{post}, \Lambda_{post}^{-1}),$$

where $\theta_{post} = (\Lambda_{post})^{-1}\{\Lambda \sum_{i=1}^{k} \log(\phi_i) + \tau\Lambda_{00}\theta_0\}$ and $\Lambda_{post} = k\Lambda + \Lambda_{00}$.

Sampling Φ from The Full Conditional Distribution

Auxiliary variable sampling methods have been widely discussed as a substitution of the Metropolis-Hastings algorithm. In order to obtain a sample from our target variable, we need to sample from the joint distribution of our target variables and auxiliary variables, then collect samples from target variables to obtain the marginal distribution of interest.

In our example, target variables are $\Phi = (\phi_1, \ldots, \phi_k)^{\mathrm{T}}$ and $\phi_i = (\beta_i, t_{50_i})^{\mathrm{T}}$. First let us

consider target variable β_i, and define $p_{\beta_i} \equiv p(\beta_i|t_{50_i}, \lambda_e, Y)$ by dropping unrelated variables $\phi_{(-i)}$, where $\phi_{(-i)}$ stands for $(\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_k)^{\mathrm{T}}$. This p_{β_i} can be written as

$$\begin{aligned} p_{\beta_i} &\equiv (\lambda_e)^{\frac{n_i}{2}} \exp\left\{-\lambda_e \frac{\sum^{n_i}(y_{ij} - f_{ij})^2}{2}\right\} \times \\ &\quad |\Lambda|^{\frac{1}{2}} (\beta_i \times t_{50_i})^{-1} \exp[-\frac{1}{2}\{\log(\phi_i) - \log(\theta)\}^{\mathrm{T}} \Lambda \{\log(\phi_i) - \log(\theta)\}], \end{aligned}$$

where p_{β_i} depends on the observations for subject i with repeated measurements. The target distribution of Φ would be sampled by the order $\beta_1 \rightarrow \mathrm{t}_{50_1} \rightarrow \ldots \rightarrow \beta_k \rightarrow \mathrm{t}_{50_k}$. By defining an auxiliary variable z, the joint distribution over $(\beta_i, z)^{\mathrm{T}}$ is as an uniform distribution within the region $U = \{(\beta_i, z) : 0 < z < p_{\beta_i}(\beta_i)\}$. The sampling procedure for this joint distribution consists of two steps: $[z|\beta_i] \sim \mathrm{Unif}\{0, p_{\beta_i}(\beta_i)\}$ and $[\beta_i|z] \sim \mathrm{Unif}(U_z)$, where $U_z = \{\beta_i : z < p_{\beta_i}(\beta_i)\}$. After dropping all auxiliary variables z's, we obtain the marginal distribution for β_i. When the inverse function of p_{β_i} is not available, the alternative is multiple auxiliary variable methods (Higdon, 1998; Swendsen and Jian-Sheng, 1987; Edwards and Sokal, 1988; Besag and Green, 1993). Damien et al. (1999) proposed n_i auxiliary variables $(z_1, \ldots, z_{n_i})^{\mathrm{T}}$ and $U_z \equiv [\phi_i : \bigcap_{j=1}^{n_i}\{z_i < p^j_{\beta_i}(\beta_i)\}]$, where

$$\begin{aligned} p^j_{\beta_i} &= (\lambda_e)^{\frac{1}{2}} \exp\left\{-\lambda_e \frac{(y_{ij} - f_{ij})^2}{2}\right\} \times \\ &\quad |\Lambda|^{\frac{1}{2}} (\beta_i \times t_{50i})^{-1} \exp[-\frac{1}{2}\{\log(\phi_i) - \log(\theta)\}^{T} \Lambda \{\log(\phi_i) - \log(\theta)\}], \end{aligned}$$

and $p^j_{\beta_i}$ is based on the j^{th} measurement of subject i. Damien et al. (1999) extended this idea by choosing more effective sampling distributions rather than the uniform distribution.

One possible difficulty of n_i auxiliary variable methods is that U_z may be an empty set when n_i is relatively large.

To overcome this problem, Neal (2003) proposed an auxiliary variable method called Slice sampling. This method avoids computing U_z by adding a rejection step after proposing β_i. Slice sampling needs only one auxiliary variable, z, and the procedure is summarized in Figure 3.5. We briefly describe Slice sampling as the following: when the distribution of interest is univariate, only one variable is being updated in each run. More often, the single-variable slice sampling will be used to sample from a multivariate distribution for $\Phi = (\beta_1, t_{50_1}, \ldots, \beta_k, t_{50_k})^{\mathrm{T}}$ by sampling repeatedly for each variable in turn. To update β_i, we must have $p_{\beta_i}(\beta_i)$, which we have defined before. The single-variable Slice sampling method discussed here replaces the current value, β_i^0, with a new value, β_i^1, which is found by following three-step procedure:

- Step 0: Initialize β_i^0.
- Step 1: Draw an auxiliary variable, z, uniformly distributed from $[0, p_{\beta_i}(\beta_i^0)]$, thereby defining a horizontal "slice", $S_h = \{\beta_i : z \leq p_{\beta_i}(\beta_i)\}$. From Figure 1, we can see S_h contains β_i^0.
- Step 2: Find an interval, $I = (L, R)$, around β_i^0 that contains β_i^0 as well.
- Step 3: Draw a new point, β_i^1, from $S_h \bigcap I$.

Figure 3.5 illustrates the three step procedure which we explain above. After a value for the

auxiliary variables has been drawn, slice S_h is defined in Step 1. Step 2 is to find an interval $I = (L, R)$, containing the current point, β_i^0, from which the new point, β_i^1, will be drawn. The procedure called "stepping out" is one common way to apply Step 2 as below:

- Step 2.1: Input, p_{β_i} (sampling density), β_i^0 (current value), z (the vertical level defining slice), and w (size of one step) .
- Step 2.2: Initialize $U \sim \text{Unif}(0, 1)$, $L = \beta_i^0 - w \times U$, and $R = L + w$.
- Step 2.3: While $z \leq p_{\beta_i}(L)$, do $L = L - w$.
- Step 2.4: While $z \leq p_{\beta_i}(R)$, do $R = R + w$.
- Step 2.5: Output, $I = (L, R)$ interval we have found.

We can randomly pick an initial interval of size, w, containing β_i^0, and then extend the interval to both sides until the evaluation conditions fails, that is either $z \geq p_{\beta_i}(L)$ or $z \geq p_{\beta_i}(R)$. After the interval has been decided, we perform Step 3 to draw new point $p_{\beta_i}(\beta_i^1)$ which is called a "shrinkage" procedure:

- Step 3.1: Input p_{β_i} (sampling density), β_i^0 (current value), z (the vertical level defining slice), w (size of one step), and $I = (L, R)$ interval we obtained from Step 2.
- Step 3.2: Initialize $U \sim \text{Unif}(0, 1)$ and $\beta_i^1 = L + (R - L) \times U$.
- Step 3.3: If $z \leq p_{\beta_i}(\beta_i^1)$, then
  - If $\beta_i^1 \geq \beta_i^0$, then $R = \beta_i^1$; else $L = \beta_i^1$.
  - Go back to Step 3.2, "Initialize" step.
- Step 3.4: Output β_i^1 if $z > p_{\beta_i}(\beta_i^1)$.

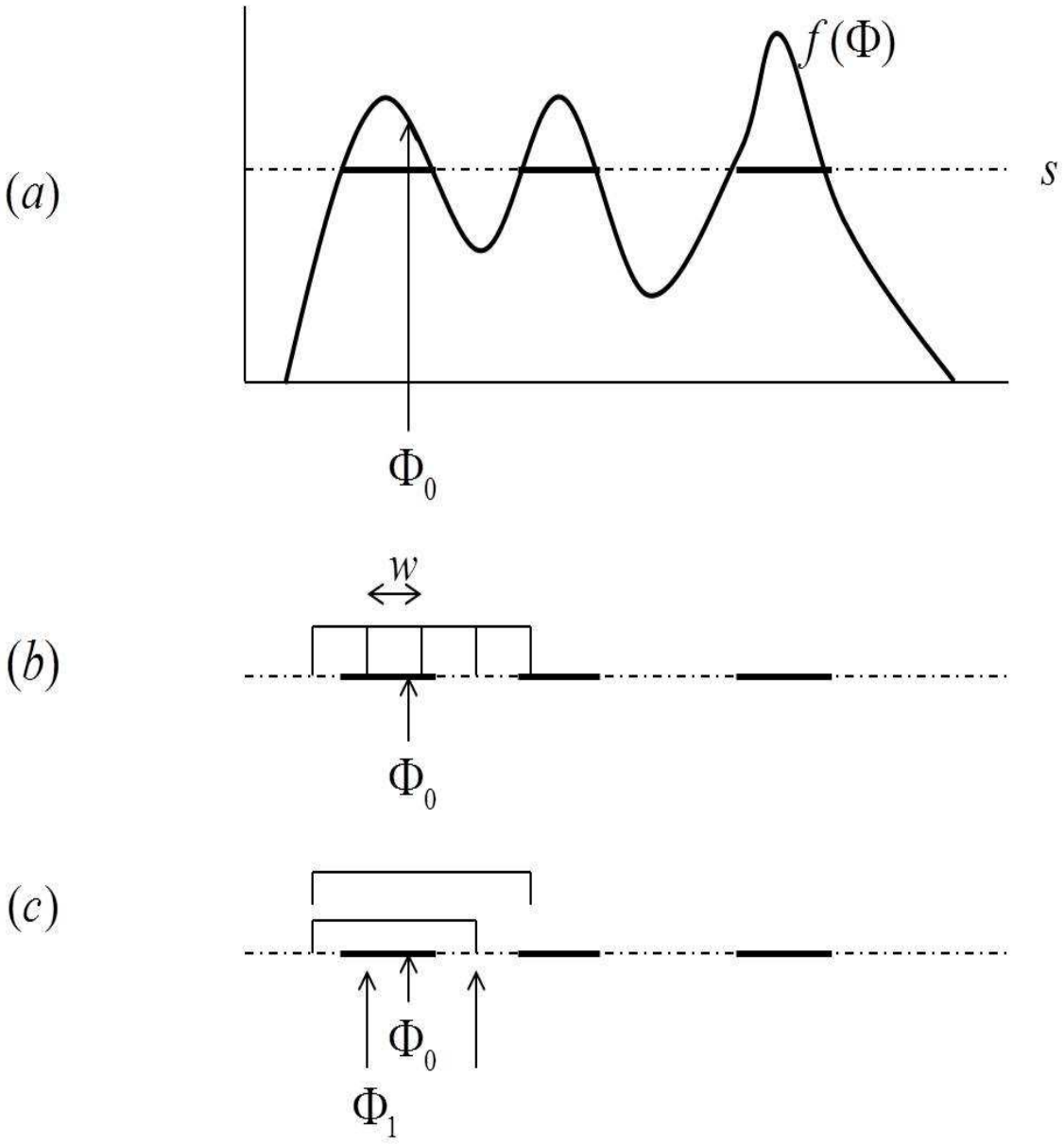


Figure 3.5: The procedure for Slice sampling: (a) Draw an auxiliary variable, defining a horizontal "slice", S_h; (b) Find an interval, I; (c) Draw a new point from $S_h \cap I$.

3.3.2 Model 2: Semiparametric Model with DP Measurement Errors

The joint distribution of Y, Φ, θ, Λ, and Ψ under DP error framework can be written as

$$\begin{aligned}
[Y,\Phi,\theta,\Lambda,\Psi] &= [Y|\Phi,\Psi]\times[\Phi|\theta,\Lambda]\times[\theta|\Lambda]\times[\Lambda]\times[\Psi] \qquad (3.3)\\
&\sim (\prod_{i=1}^{k}{\lambda_{ei}}^{n_i/2})\exp\{-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\lambda_{ei}\frac{(y_{ij}-\mu_i-f_{ij})^2}{2}\}\times\\
&\quad |\Lambda|^{\frac{k}{2}}(\prod_{i=1}^{k}\phi_i)^{-1}\exp[-\frac{1}{2}\sum_{i=1}^{k}\{\log(\phi_i)-\log(\theta)\}^{\mathrm{T}}\Lambda\{\log(\phi_i)-\log(\theta)\}]\times\\
&\quad |\Lambda_{00}|^{1/2}(\theta)^{-1}\exp[-\frac{1}{2}\{\log(\theta)-\theta_0\}^{\mathrm{T}}\Lambda_{00}\{\log(\theta)-\theta_0\}]\times\\
&\quad |\Lambda|^{\frac{\tau-d-1}{2}}\exp\{-\frac{1}{2}\mathrm{Tr}(\tau\Lambda_0^{-1}\Lambda)\}\times[\Psi|G].
\end{aligned}$$

- The full conditional distribution, $[\Phi|\theta,\Lambda,\Psi,Y]$, are sampled using Metropolis-Hastings sampling. The function to be evaluated is

$$\begin{aligned}
&[\phi_i=\{\beta_i,t_{50i}\}|\theta,\Lambda,\Psi,Y]\\
&\sim (\lambda_{ei})^{\frac{n_i}{2}}\exp\{-\sum_{j=1}^{n_i}\lambda_{ei}\frac{(y_{ij}-f_{ij})^2}{2}\}\times\\
&\quad (\beta_i\times t_{50i})^{-1}\exp[-\frac{1}{2}\{\log(\phi_i)-\log(\theta)\}^{\mathrm{T}}\Lambda\{\log(\phi_i)-\log(\theta)\}].
\end{aligned}$$

- The full conditional distribution, $[\Lambda|\theta,\Phi,\Psi,Y]$, has a closed form

$$[\Lambda|\theta,\Phi,\Psi,Y] = [\Lambda|\Phi,\theta]$$

$$\sim \quad |\Lambda|^{\frac{k+\tau-d}{2}} \exp[-\frac{1}{2}\mathrm{Tr}\{(S+\tau\Lambda_0^{-1})\Lambda\}],$$

where $S = \sum_{i=1}^{k} \{\log(\phi_i) - \log(\theta)\}\{\log(\phi_i) - \log(\theta)\}^{\mathrm{T}}$.

- The full conditional distribution, $[\theta|\Phi, \lambda_e, \Lambda, Y]$, has a closed form

$$[\theta|\Phi, \lambda_e, \Lambda, Y] = [\theta|\Phi, \Lambda] \sim \mathrm{LN}(\theta_{post}, \Lambda_{post}),$$

where $\theta_{post} = (\Lambda_{post})^{-1}(\Lambda \sum \phi_i + \Lambda_{00}\theta_0)$ and $\Lambda_{post} = k\Lambda + \Lambda_{00}$.

Sampling from Full Conditional Distribution of Ψ

The posterior distribution of Ψ is sampled by following the three steps:

- Step1: Sampling initial values of Ψ from the Dirichlet process prior to Ψ.

  Initial values for $\psi_i = (\mu_i, \lambda_{ei})$, which is following G, are sampled by the "stick-breaking" construction of the Dirichlet process (Sethuraman, 1994), which are

$$\begin{aligned} G(\psi) &= \sum_{l=1}^{\infty} p_l \delta_{\psi_l}; \\ \psi_l &\sim G_0; p_1 = V_1; p_l = V_l \prod_{j=1}^{l-1}(1-V_j); \\ V_l &\sim \mathrm{Beta}(1, \alpha), \mathrm{l} = 1, 2, \ldots, \end{aligned}$$

  where δ_a is the unit point mass at a and the summation is truncated at a large integer N as in Ishwaran and James (2001).

- Step2: Sampling the posterior distribution of ψ_i using a mixture distribution.

  Under Gibbs sampling, we can take a sample from $[\psi_i|Y, \Phi, \psi_{-i}, \alpha, G_0]$ using the following mixture distribution,

$$[\psi_i|Y, \phi, \psi_{-i}, \alpha, G_0] \propto q_{i,0}\pi_i(\psi_i|\underline{Y_i}, \underline{f_i}, \phi_i, G_0) + \sum_{r=1}^{s_{-i}} q_{-i,r}\delta_{\psi^*_{-i,r}},$$

  where $\pi_i(\psi_i|\underline{Y_i}, \underline{f_i}, \phi_i, G_0) = \mathrm{N}(\mu_i|\mu_{post}, (g_{post}\lambda_{ei})^{-1})\mathrm{Gamma}(\lambda_{ei}|a_{post}/2, b_{post}/2)$, $q_{i,0} = \frac{\alpha}{\alpha+k-1}t_{v_0}(\underline{Y_i}|\underline{f_i}, \Sigma)$, $\Sigma = s_0^2/C$, and $q_{-i,r} = \frac{k_{-i,r}}{(\alpha+k-1)}L(\underline{Y_i}|\underline{f_i}, \phi_i, \psi_i)$. The more details are derived in Appendix A.

- Step 3: Sampling α from a mixture of gamma distributions.

  In order to sample a proper α value, Escobar and West (1995) proposed a full structure of sampling α by using Gibbs sampling, which assume $p(\alpha) \propto \mathrm{Gamma}(a_1, b_1)$, and the prior mean for α is a_1/b_1. Then, the posterior of α is $[\alpha|k, s, \text{others}] = p(\alpha|k, s) \propto p(\alpha)p(s|k, \alpha)$. The following formula,

$$p(s|k, \alpha) = C_k(s)k!\alpha^s\frac{\Gamma(\alpha)}{\Gamma(\alpha+k)},$$

  assumes that the value of α only depends on the number of components, s, and $C_k(s) = p(s|k, \alpha = 1)$, where k is the total number of Φ and s is the number of unique Φ. Finally we can sample from the following:

$$[\eta|\alpha, s] \quad \propto \quad \mathrm{Beta}(\alpha+1, k),$$

$$[\alpha|k,s,\eta] \propto \pi_\eta \text{Gamma}\{a_1+s, b_1-\log(\eta)\} + (1-\pi_\eta)\text{Gamma}\{a_1+s-1, b_1-\log(\eta)\},$$

where $\pi_\eta/(1-\pi_\eta) = (a_1+s-1)/[k\{b_1-\log(\eta)\}]$.

Based on this derivation, α can be sampled as follows:

- Sampling η value from a beta distribution, conditional on α and k fixed at their most recent values.
- Sampling a new α value from the mixture of gamma distributions based on the same s and the η value just generated.

3.3.3 Model 3: Semiparametric Model with One-layer DP Random Effects

Defining $\Phi = (\phi_1, \ldots, \phi_k)^{\mathrm{T}}$, the joint distribution of Y, Φ, λ_e, and α under DP random effects can be written as

$$\begin{aligned}
[Y,\Phi,\lambda_e,\alpha] &= [Y|\Phi,\lambda_e]\times[\Phi|G]\times[\alpha]\times[\lambda_e] \\
&\sim (\lambda_e)^{\frac{\sum n_i}{2}} \exp\left\{-\lambda_e \frac{\sum\sum (y_{ij}-f_{ij})^2}{2}\right\} \times \\
&\quad (\lambda_e)^{\frac{v_0}{2}-1} \exp(-\frac{v_0 s_0^2}{2}\lambda_e) \times \\
&\quad [\Phi|G]\times[\alpha].
\end{aligned}$$

The full conditional distributions of λ_e, Φ, and α can be derived in the following way:

- The full conditional distribution of λ_e is:

$$[\lambda_e|\Phi, Y] = (\lambda_e)^{\frac{\sum_{i=1}^{k} n_i + v_0}{2} - 1} \exp\left(-\lambda_e [\frac{v_0 s_0^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \{y_{ij} - f(t_{ij}|\phi_i)\}^2}{2}]\right),$$

which is a gamma distribution,Gamma(a^d, b^d), with $a^d = \sum_{i=1}^{k} n_i + v_0/2$ and $b^d = \{v_0 s_0^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - f_{ij})^2\}/2$.

- The posterior distribution of Φ and α are sampled by following three steps:

  - Step 1: Sampling initial values of Φ from the Dirichlet process prior $[\Phi|G]$

    Initial values for $\phi_i = (\beta_i, t_{50_i})^{\mathrm{T}}$, which is following G, are sampled by the "stick-breaking" construction of the Dirichlet process Sethuraman(1994), which are

$$\begin{aligned} G(\phi_i) &= \sum_{l=1}^{\infty} p^l \delta_{\phi_i^l}; \\ \phi_i^l &\sim G_0; p^1 = V^1; \ldots; p^l = V^l \prod_{j=1}^{l-1} (1 - V^j); \\ V^l &\sim \text{Beta}(1, \alpha), \text{l} = 1, 2, \ldots, \end{aligned}$$

    and δ^a is the unit point mass at a and the summation is truncated at a large integer N as in Ishwaran et al.(2001).

  - Step 2: Sampling the posterior of ϕ_i using a mixture distribution

    Under Gibbs sampling, we can sampling from the following mixture distributions,

$[\phi_i|Y, \phi_{-i}, \alpha, G_0]$.

$$\begin{aligned}
[\phi_i|Y, \phi_{-i}, \alpha] &\propto [\phi_i|\phi_{-i}, \alpha]\mathrm{L}(\underline{Y_i}|\underline{f_i}, \lambda_e) \\
&= (\frac{\alpha}{\alpha+k-1}G_0 + \frac{1}{\alpha+k-1}\sum_{r=1}^{s_{-i}} k_{-i,r}\delta_{\phi^*_{-i,r}})\mathrm{L}(\underline{Y_i}|\underline{f_i}, \lambda_e), \\
&= \frac{\alpha}{\alpha+k-1}p(\underline{Y_i}|\lambda_e)G(\phi_i|\underline{f_i}, \underline{Y_i}, \lambda_e) \\
&\qquad + \sum_{r=1}^{s_{-i}} \frac{k_{-i,r}}{\alpha+k-1}\mathrm{L}(\underline{Y_i}|\underline{f_i}, \lambda_e)\delta_{\phi^*_{-i,r}},
\end{aligned}$$

where $\underline{Y_i} = (y_{i1}, \ldots, y_{in_i})$ and $\underline{f_i} = (f_{i1}, \ldots, f_{in_i})$.

* The exact result for $P(\underline{Y_i}|\lambda_e)$ is hard to obtain, since the integration is involved in this computation,

$$p(\underline{Y_i}|\lambda_e) \propto \int \mathrm{L}\{\underline{Y_i}|\underline{f_i}, \lambda_e\}G_0(\phi_i)d\phi_i.$$

  There are several ways to estimate this integration. One way uses $\hat{p}(\underline{Y_i}|\lambda_e) = [\sum_{g=1}^{G} \mathrm{L}\{\underline{Y_i}|\underline{f_i}(\phi_i^g), \lambda_e\}]/G$, giving ϕ_i^g from the prior distribution. The other way uses $\hat{p}(\underline{Y_i}|\lambda_e) = ([\sum_{g=1}^{G} \mathrm{L}\{\underline{Y_i}|\underline{f_i}(\phi_i^g), \lambda_e\}^{-1}]/G)^{-1}$, giving ϕ_i^g from the posterior distribution,

* ϕ_i from $G(\phi_i|\underline{f_i}, \underline{Y_i}, \lambda_e)$ can be sampled from a Metropolis-Hastings algorithm or Slice sampling scheme,

$$G(\phi_i|\underline{f_i}, \underline{Y_i}, \lambda_e) \propto \mathrm{L}\{\underline{Y_i}|\underline{f_i}, \lambda_e\}G_0(\phi_i)$$

$$\propto \quad (\lambda_e)^{\frac{\sum n_i}{2}} \exp\left\{-\frac{\lambda_e}{2}\sum\sum(y_{ij}-f_{ij})^2\right\}\times$$
$$|\Lambda_0|^{\frac{1}{2}}(\beta_i \times t_{50i})^{-1}\exp[-\frac{1}{2}\{\log(\phi_i)-\theta_0\}^{\mathrm{T}}\Lambda_0\{\log(\phi_i)-\theta_0\}].$$

– Step 3: Sampling α from a mixture of gamma distributions.

In order to sample a proper α value, Escobar and West (1995) proposed a full structure for sampling α by using Gibbs sampling, which assume $p(\alpha) \propto$ Gamma(a_3, b_3). Then, the posterior of α is $[\alpha|k, s, \text{others}] = p(\alpha|k, s) \propto p(\alpha)p(s|k, \alpha)$. The following formula,

$$p(s|k,\alpha) = C_k(s)k!\alpha^s\frac{\Gamma(\alpha)}{\Gamma(\alpha+k)},$$

assumes that the value of α only depends on the number of components, s, and $C_k(s) = p(s|k, \alpha = 1)$, where k is the total number of Φ and s is the number of unique Φ. Finally we can sample from the following:

$$[\eta|\alpha, s] \quad \propto \quad \text{Beta}(\alpha+1, k),$$
$$[\alpha|k, s, \eta] \quad \propto \quad \pi_\eta \text{Gamma}\{a_1+s, b_1-\log(\eta)\} + (1-\pi_\eta)\text{Gamma}\{a_1+s-1, b_1-\log(\eta)\},$$

where $\pi_\eta/(1-\pi_\eta) = (a_1+s-1)/[k\{b_1-\log(\eta)\}]$.

Based on the above derivation, α can be sampled as follows:

∗ Sampling η from a beta distribution, conditional on α and k fixed at their

most recent values.

* Sampling a new α value from the mixture of gamma distributions based on the same s and the η value just generated.

3.3.4 Model 4: Semiparametric Model with Two-layer DP Random Effects

The joint distributions of Y, Φ, θ, Λ, and λ_e can then be written as:

$$\begin{aligned}
[Y, \Phi, \theta, \Lambda, \lambda_e] &= [Y|\Phi, \lambda_e] \times [\Phi|\theta, \Lambda] \times [\theta|\Lambda] \times [\Lambda] \times [\lambda_e] \qquad (3.4)\\
&\sim (\lambda_e)^{\frac{\sum n_i}{2}} \exp\left\{-\lambda_e \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-f_{ij})^2}{2}\right\} \times \\
&\quad |\Lambda_i|^{\frac{k}{2}} (\prod^{k} \phi_i)^{-1} \exp[-\frac{1}{2}\sum_{i=1}^{k}\{\log(\phi_i)-\log(\theta_i)\}^{\mathrm{T}}\Lambda_i\{\log(\phi_i)-\log(\theta_i)\}] \times \\
&\quad (\lambda_e)^{\frac{v_0}{2}-1} \exp(-\frac{v_0 s_0^2}{2}\lambda_e) \times \\
&\quad [\omega_i = (\theta_i, \Lambda_i)|G],
\end{aligned}$$

where $G \sim \mathrm{DP}(\alpha, G_0)$ and $G_0 = [\theta_i][\Lambda_i] = \mathrm{LN}\{\theta_i|\theta_0, (\Lambda_{00})^{-1}\}\mathrm{Wishart}(\Lambda_i|\Lambda_0/\tau, \tau)$. Under this framework, the sampling procedure are as follows:

- $[\phi_i|\lambda_e, \theta_i, \Lambda_i, Y]$ can be sampled by Metropolis-Hastings sampling. In each step, one parameter is sampled out. That is, we first sample β_1 and then t_{50_1}, and second sample β_2 and then t_{50_2} and so forth.

The function to be evaluated by Metropolis-Hastings sampling is

$$\begin{aligned}
&[\beta_i|\lambda_e,\theta_i,\Lambda_i,Y,t_{50_i}] \\
&\sim \ (\lambda_e)^{\frac{n_i}{2}}\exp[-\lambda_e\frac{\sum_j^{n_i}\{y_{ij}-f(t_{ij}|\beta_i,t_{50_i})\}^2}{2}]\times \\
&\quad |\Lambda_i|^{\frac{1}{2}}(\beta_i\times t_{50_i})^{-1}\exp\left[-\frac{1}{2}\{\log(\phi_i)-\log(\theta_i)\}^{\mathrm{T}}\Lambda_i\{\log(\phi_i)-\log(\theta_i)\}\right].
\end{aligned}$$

- The full conditional distribution, $[\lambda_e|\theta,\Phi,\Lambda,Y]$, is

$$\begin{aligned}
[\lambda_e|\theta,\Phi,\Lambda,Y] &= [\lambda_e|\Phi,Y] \\
&\sim (\lambda_e)^{\frac{\sum n_i}{2}+a_1-1}\exp\left(-\lambda_e[b_1+\frac{\sum_{i=1}^k\sum_{j=1}^{n_i}\{y_{ij}-f(t_{ij}|\phi_i)\}^2}{2}]\right),
\end{aligned}$$

  which is Gamma(a_e,b_e) with $a_e=(\sum_{i=1}^k n_i)/2+a_1$ and $b_e=b_1+\sum_{i=1}^k\sum_{j=1}^{n_i}\{y_{ij}-f(t_{ij}|\phi_i)\}^2/2$.

- $\omega_i=(\theta_i,\Lambda_i)$ is sampled from a Dirichlet process mixture,

$$\omega_i \ \sim \ \frac{\alpha}{\alpha+k-1}p(\phi_i)p(\omega_i|\phi_i)+\sum_{r=1}^{s_{-i}}\frac{k_{-i,r}}{\alpha+k-1}p(\phi_i|\omega_r)\delta_{\omega^*_{-i,r}}.$$

  - The samples of the continuous distribution, $p(\omega_i|\phi_i)$, are obtained by Gibbs sampling procedure as described as follow;

    * The full conditional distribution of Λ_i has a closed form,

$$[\Lambda_i|\theta_i,\Phi,\lambda_e,Y]=[\Lambda_i|\Phi,\theta_i]=|\Lambda_i|^{\frac{1+\tau-d}{2}}\exp[-\frac{1}{2}\mathrm{Tr}\{(\mathrm{S_i}+\tau\Lambda_0^{-1})\Lambda_\mathrm{i}\}],$$

which follows a Wishart distribution with $\Lambda_i^d = S_i + \tau\Lambda_0^{-1}$ and $\tau + 2$ degree of freedom, where $S_i = \{\log(\phi_i) - \log(\theta_i)\}\{\log(\phi_i) - \log(\theta_i)\}^{\mathrm{T}}$. The details would be shown in (3.6).

* The full conditional distribution of θ_i has a closed form as well,

$$[\theta_i|\Phi, \lambda_e, \Lambda_i, Y] = [\theta_i|\Phi, \Lambda_i] \sim \mathrm{LN}(\theta_{i_{post}}, \Lambda_{i_{post}}^{-1}),$$

where $\theta_{i_{post}} = (\Lambda_{i_{post}})^{-1}\{\Lambda_i \log(\phi_i) + \Lambda_{00}\theta_0\}$ and $\Lambda_{i_{post}} = \Lambda_{00} + \Lambda_i$. It would be demonstrated in (3.7).

– $p(\phi_i|\omega_r) = \mathrm{LN}\{\phi_{\mathrm{i}}|\log(\theta_{\mathrm{r}}), (\Lambda_{\mathrm{r}})^{-1}\}$,

– The sample ϕ_i is from the distribution,

$$\begin{aligned} p(\phi_i) &= \int_{\Lambda_i}\int_{\theta_i} G_0(\omega_i)L(\phi_i|\omega_i)d\theta_i d\Lambda_i \\ &\sim \int_{\Lambda_i}\int_{\theta_i} \mathrm{LN}(\phi_i|\omega_i)\mathrm{LN}\{\theta_i|\theta_0, \Lambda_{00}^{-1}\}\mathrm{Wishart}(\Lambda_i|\Lambda_0/\tau, \tau)d\theta_i d\Lambda_i. \quad (3.5) \end{aligned}$$

We know that (3.5) has a closed form of $p(\phi_i)$, and resulting value of $p(\phi_i)$ is in Appendix B. When (3.5) is not a closed form, we use a method proposed by Chib (1995, 2001), which is described as follows:

* $\hat{p}(\phi_i)$ is estimated as follows,

$$\hat{p}(\phi_i) = \frac{p(\phi_i|\omega_i^*)p(\omega_i^*)}{p(\omega_i^*|\phi_i)}$$

$$= \frac{\mathrm{LN}(\phi_i|\omega_i^*)\mathrm{LN}\{\theta_i^*|\theta_0,\Lambda_{00}^{-1}\}\mathrm{Wishart}(\Lambda_{\mathrm{i}}^*|\Lambda_0/\tau,\tau)}{\mathrm{p}(\Lambda_{\mathrm{i}}^*|\phi_{\mathrm{i}},\theta_{\mathrm{i}}^*)\hat{\mathrm{p}}(\theta_{\mathrm{i}}^*|\phi_{\mathrm{i}})},$$

where ω_i^* is any point of parameter space ω_i. $\mathrm{LN}(\phi_i|\omega_i^*) = \mathrm{LN}[\phi_{\mathrm{i}}|\omega_{\mathrm{i}}^* = \{\log(\theta_{\mathrm{i}}^*),(\Lambda_{\mathrm{i}}^*)^{-1}\}]$.

* $p(\Lambda_i^*|\phi_i,\theta_i^*)$ can be obtained from multivariate t distribution which is derived as follows,

$$\begin{aligned}
[\Lambda_i|\theta_i,\phi_i,\lambda_e,Y] &= [\Lambda_i|\phi_i,\theta_i,Y] \\
&\sim |\Lambda_i|^{\frac{1}{2}}(\beta_i t_{50i})^{-1}\exp\left[-\frac{1}{2}\{\log(\phi_i)-\log(\theta_i)\}^{\mathrm{T}}\Lambda_i\{\log(\phi_i)-\log(\theta_i)\}\right]\times \\
&\quad |\tau\Lambda_i|^{1/2}(\beta_i t_{50i})^{-1}\exp\left[-\frac{\tau}{2}\{\log(\theta_i)-\theta_0\}^{\mathrm{T}}\Lambda_i\{\log(\theta_i)-\theta_0\}\right]\times \\
&\quad |\Lambda_i|^{\frac{\tau-d-1}{2}}\exp\{-\frac{1}{2}\mathrm{Tr}(\Lambda_0\Lambda_{\mathrm{i}})\} \\
&\sim |\Lambda_i|^{\frac{1+\tau-d}{2}}\exp\left[-\frac{1}{2}\mathrm{Tr}\{(\mathrm{S_i}+\tau\Lambda_0^{-1})\Lambda_{\mathrm{i}}\}\right], \qquad (3.6)
\end{aligned}$$

where $S_i = \{\log(\phi_i)-\log(\theta_i)\}\{\log(\phi_i)-\log(\theta_i)\}^{\mathrm{T}}$.

* We calculate $\hat{p}(\theta_i^*|\phi_i) = \sum_{g=1}^{G} p(\theta_i^*|\phi_i,\Lambda_i^g)$, where Λ_i^g is sampled from $p(\Lambda_i|\phi_i)$. The samples from the marginal distribution, $p(\Lambda_i|\phi_i)$, are generated from two full conditional distributions $[\Lambda_i|\phi_i,\theta_i,]$ and $[\theta_i|\phi_i,\Lambda_i]$. $p(\theta_i^*|\phi_i,\Lambda_i^g)$ is the following lognormal distribution,

$$\begin{aligned}
[\theta_i|\phi_i,\lambda_e,\Lambda_i,Y] &= [\theta_i|\phi_i,\Lambda_i] \\
&\sim |\Lambda_i|^{\frac{1}{2}}(\phi_i)^{-1}\exp\left[-\frac{1}{2}(\log(\phi_i)-\log(\theta_i))^{\mathrm{T}}\Lambda_i(\log(\phi_i)-\log(\theta_i))\right]\times
\end{aligned}$$

$$
\begin{aligned}
& |\Lambda_{00}|^{1/2}(\theta_i)^{-1}\exp[-\frac{\tau}{2}\{\log(\theta_i)-\theta_0\}^{\mathrm{T}}\Lambda_{00}\{\log(\theta_i)-\theta_0\}] \\
\sim\ & \mathrm{LN}(\theta_{i_{post}}, \Lambda_{i_{post}}^{-1}),
\end{aligned} \tag{3.7}
$$

where we have $\theta_{i_{post}} = (\Lambda_{i_{post}})^{-1}(\Lambda_i \log\phi_i + \Lambda_{00}\theta_0)$ and $\Lambda_{i_{post}} = \Lambda_{00} + \Lambda_i$.

- Sampling α from the mixture of gamma distributions which exactly follows the method in Chapter 3.3.2.

3.4 Model Selection

In Chapter 3.3, we propose four Bayesian hierarchical models. In Chapter 3.4, in terms of model comparisons, we propose a unified approach to calculate penalized posterior Bayes factors. We are interested in selecting a model from a set of candidate models M_1, M_2, and M_4. More specifically, M_1 indicates the parametric NLME model, M_2 indicates the NLME with DP measurement errors, and M_4 indicates the NLME with two-layer DP random effects.

The idea of marginal likelihood based methods is to compare the expected values of the likelihood functions for a pair of interest models. The marginal likelihood is used to calculate the Bayes factor. In general, the Bayes factor can be very sensitive to variations in the prior information of target parameters (Aitkin, 1991). If the prior accurately represents one's subjective belief about the target distribution, then such sensitivity is not a problem. However, the semiparametric Bayesian structure itself (Dirichlet process prior) may bring the variation in the priors, comparing a parametric prior model. Some researchers (Petrone

and Raftery, 1997) argue that the Bayes factor measures the degree of the improvement from the prior. The variation of the priors may come from the discrete nature of the Dirichlet process and infinite parameter spaces. Hence, Bayes factor may not be a proper option to compare a parametric model and a semiparametric model. The other difficulty to calculate Bayes factors comes from our nonlinear likelihood function form and DP prior structure, which increase the complexity of the approximations of the marginal likelihood.

To overcome the difficulties and drawbacks of the Bayes factor, Aitkin (1991) proposed a posterior Bayes factor, which is the ratio of the pseudo marginal likelihood. The pseudo marginal likelihood weights the likelihood functions by the posterior probabilities of target parameters, rather than the prior probabilities of them. The use of the posterior mean has several advantages, including the reduced sensitivity to variations in the prior and the avoidance of the Lindley paradox (Lindley, 1957) in testing point null hypotheses. In our study, we incorporate posterior Bayes factor by adding a penalty term into likelihood function, and propose penalized posterior Bayes factor to stably estimate variance components in our nonlinear mixed model.

In the rest of this chapter, we first describe the general procedure of calibrating the Bayes factor, and indicate the difficulties of the computation. We then discuss the Bayes factor of hierarchical models with DP random effects or DP measurement errors. At the end of this chapter, we show how to compute the penalized posterior Bayes factors and further discuss the cross predictive densities.

3.4.1 Bayes Factor

We wish to compare two models, M_r and M_r', using Bayes factor,

$$B_{rr'} = \frac{p(Y|M_r)}{p(Y|M_{r'})}, \tag{3.8}$$

where $r = 1, 2, 4$, $r' = 1, 2, 4$, and $r \neq r'$. Recall that M_1 stands for a NLME model with parametric priors, M_2 stands for a NLME model with DP measurement errors, and M_4 stands for a NLME model with two-layer DP random effects. $p(Y|M_1)$ is the marginal likelihood given M_1. The marginal likelihood of the NLME model with parametric priors can be expressed as:

$$p(Y|M_1) = \int_{\xi}\int_{\Phi} p(Y|\Phi, \xi, M_1)p(\Phi, \xi|M_1)d\Phi d\xi,$$

where $\xi = (\theta, \Lambda, \lambda_e)$. $p(\Phi, \xi|M_1)$ is the prior density of Φ and ξ under model M_1. Since we know the posterior of Φ is not given as a closed form, it is hard to obtain the exact value for $p(Y|M_1)$. There are two types of approximate analytic calculations available: Type I, involving the computation of the probability of $p(\Phi|\xi, Y, M_1)$; Type II, not involving the probability of $p(\Phi|\xi, Y, M_1)$.

Recall that $Y = (Y_1, \ldots, Y_i, \ldots, Y_k)$ and $\Phi = (\phi_1, \ldots, \phi_i, \ldots, \phi_k)$. Then $p(Y) = \prod_{i=1}^{k} p(Y_i)$.

We note $q(\phi_i|\xi, Y_i, M_0)$ represents an unnormalized function of the subject i^{th},

$$\begin{aligned} q(\phi_i|\xi, Y_i, M_0) &= (\lambda_e)^{\frac{n_i}{2}} \exp\left\{-\lambda_e \frac{\sum_{j=1}^{n_i}(y_{ij} - f_{ij})^2}{2}\right\} \times \\ & |\Lambda|^{\frac{1}{2}} (\beta_i \times t_{50i})^{-1} \exp[-\frac{1}{2}\{\log(\phi_i) - \log(\theta)\}^T \Lambda \{\log(\phi_i) - \log(\theta)\}] \quad (3.9) \end{aligned}$$

and $p(\phi_i|\xi, Y_i, M_0) = q(\phi_i|\xi, Y_i, M_0)/c$, where $c = \int q(\phi_i|\xi, Y_i, M_0) d\Phi$. Although MCMC allows us to sample from $p(\phi_i|\xi, Y_i, M_0)$ without knowing the normalized constant c, however, calculating the probability of $p(\phi_i|, Y_i, M_0)$ takes more effort.

For ease notation, we use $p(Y_i)$ for $p(Y_i|M_1)$. Then the model becomes

$$p(Y_i) = \int\int p(Y_i|\phi_i, \xi) p(\phi_i, \xi) d\phi_i d\xi.$$

One of type I methods (Newton and Raftery, 1994) is proposed based on an importance sampling scheme. The marginal likelihood can be estimated as

$$\hat{p}_{NR_1}(Y_i) = \left\{\frac{1}{s}\sum_{s=1}^{m} p(Y_i|\Phi^s, \xi^s)^{-1}\right\}^{-1},$$

where Φ^s and ξ^s are sampled from the posterior density $[\phi_i, \xi|Y_i]$. In other words the marginal likelihood can be estimated by the harmonic mean of the likelihood of a sample from the posterior distribution. It was shown that $\hat{p}_{NR1}(Y_i)$ converges to the correct value, $p(Y_i)$, as $s \to \infty$. However, $\hat{p}_{NR1}(Y_i)$ does not, in general, satisfy a Gaussian central limit theorem. $\hat{p}_{NR_1}(Y_i)$ manifests itself by the occasional occurrence of a value of (ϕ_i, ξ) with small likelihood

and hence has large effect on the final result. Two other estimators (Newton and Raftery, 1994) were proposed to avoid the instability of $\hat{p}_{NR_1}(Y_i)$.

One estimator is $\hat{p}_{NR_2}$. The sampling density for $\hat{p}_{NR_2}$ is a mixture of the prior and posterior densities, $[\phi_i, \xi|Y_i, \delta] = \delta p(\phi_i, \xi) + (1-\delta)p(\phi_i, \xi|Y_i)$, where δ is small, $\hat{p}_{NR_2}(Y_i)$ is defined by

$$\hat{p}_{NR_2}(Y_i) = \frac{\sum_{s=1}^{m} \frac{p(Y_i|\phi_i,\xi)}{\delta \times \hat{p}_{NR_2}(Y_i)+(1-\delta)\times p(Y_i|\phi_i,\xi)}}{\sum_{s=1}^{m}\{\delta \times \hat{p}_{NR_2}(Y_i) + (1-\delta) \times p(Y_i|\phi_i,\xi)\}^{-1}}.$$

The estimator $\hat{p}_{NR_2}(\phi_i, \xi)$ retains the efficiency of $\hat{p}_{NR_1}$ due to being based mostly on high likelihood values of ϕ_i and ξ, but avoids its instability. However it is required to simulate samples from both the prior and posterior densities.

The other is $\hat{p}_{NR_3}$ which simulates all s values from the posterior distribution, and $\hat{p}_{NR_i}$ is defined by

$$\hat{p}_{NR_3}(Y_i) = \frac{\frac{\delta \times m}{1-\delta} + \sum_{s=1}^{m} \frac{p(Y_i|\phi_i,\xi)}{\delta \times \hat{p}_{NR_3}(Y_i)+(1-\delta)\times p(Y_i|\phi_i,\xi)}}{\frac{\delta \times m}{(1-\delta)\times \hat{p}_{NR_3}(Y_i)} + \sum_{s=1}^{m}\{\delta \times \hat{p}_{NR_3}(Y_i) + (1-\delta) \times p(Y_i|\phi_i,\xi)\}^{-1}}.$$

Here δ is a small value, which is suggested to be as small as 0.01. Both of the estimator $\hat{p}_{NR_2}$ and $\hat{p}_{NR_3}$ are evaluated using an iterative scheme and both satisfy the Gaussian central theorem. The calculating of posterior Φ are not involved in these three estimators.

The type II method (Chib, 1995, 2001) estimates the marginal likelihood function based

on Bayes formula,

$$p(Y_i) = \frac{p(Y_i|\phi_i, \xi)p(\phi_i, \xi)}{p(\phi_i, \xi|Y_i)}. \tag{3.10}$$

The proposed estimator of the marginal density on logarithm scale is

$$\log\{\hat{p}(Y_i)\} = \underbrace{\log\{p(Y_i|\phi_i^*, \xi^*)\}}_{Element\ 1} + \underbrace{\log\{p(\phi_i^*, \xi^*)\}}_{Element\ 2} - \underbrace{\log\{p(\phi_i^*, \xi^*|Y_i)\}}_{Element\ 3}, \tag{3.11}$$

and this equation always holds for any ϕ_i^* and ξ^*. An arbitrary point in the parameter space is marked by "*". This approach requires us to evaluate the likelihood function, the prior, and an estimate of the posterior.

The difficulty for the computation of (3.11) comes from Element 3. The ϕ_i does not have a closed from, and the unnormalized posterior function is written as (3.9). In terms of the parametric prior setting for all the target parameters, Chib (1995, 2001) proposed a method to estimate the posterior probability of ϕ_i based on the unnormalized posterior function, (3.9). Our model 2 and model 4 assume that either λ_e or $\omega_i = (\theta_i, \Lambda_i)$ is from a Dirichlet process prior. This prior structure increase the computation complexity of Element 3, associating the unnormalized posterior function form. Kong et al. (1994) gave a sequential importance sampling scheme to help the estimation of DP mixture posterior, which was improved by Basu and Chib (2003), called as the collapsed sequential importance sampling. When a Dirichlet process prior and the unnormalized posterior function are involved in the model, Element 3 needs to be estimated by the collapsed sequential importance sampling by

combining with the method of Chib (1995, 2001). The complexity makes a slow process for estimating parameters.

3.4.2 Penalized Posterior Bayes Factor

As we mentioned before, Aitkin (1991) proposed a posterior Bayes factor using the ratio of the pseudo marginal likelihood. The pseudo marginal likelihood weights the likelihood functions by the posterior probabilities of target parameters, rather than the prior probabilities of them. The use of the posterior mean has several advantages, including less sensitivity to variations in the prior and the avoidance of the Lindley paradox (Lindley, 1957) in testing point null hypotheses. In our study, we propose a penalized posterior Bayes factor by adding a penalty term into likelihood function because our model is hierarchical structure, while a posterior Bayes factor can not used in such complicated structure..

For our case, the penalized posterior Bayes(PPB) factor with penalty term can expressed as:

$$PPB_{rr'} = \frac{p(Y|M_r)}{p(Y|M_{r'})}, \tag{3.12}$$

where

$$p(Y|M_r) = \int_{M_r} \underbrace{p(Y|\Phi, \lambda_e)}_{Likelihood} \underbrace{p(\Phi|\theta, \Lambda)}_{Penalty} \underbrace{p(\Phi, \theta, \Lambda, \lambda_e|Y)}_{Posterior} d\Phi d\lambda_e d\theta d\Lambda. \tag{3.13}$$

It can be estimated by evaluating $\sum_{s=1}^{m} p(Y|\Phi^s, \lambda_e^s)p(\Phi^s|\theta^s, \Lambda^s)/m$, where $(\Phi^s, \lambda_e^s, \theta^s, \Lambda^s)$ are generated from the posterior distribution $p(\Phi, \theta, \Lambda, \lambda_e|Y)$.

For the prediction problem, it is natural to assess the predictive ability of the model by using the cross validation. The cross validation predictive densities (Gelfand et al., 1992) are proposed as

$$CVPD = \int_{\Phi}\int_{\xi}\prod_{i=1}^{k}\prod_{j=1}^{n_i} p(y_{ij}|\Phi, \xi)p\{\Phi, \xi|Y_{(-ij)}\}d\Phi d\xi, \tag{3.14}$$

where $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, and $n = \sum_1^k n_i$. In (3.14), y_{ij} stands for a single observation from a subject, and $Y_{(-ij)}$ stands for all observations except y_{ij}.

3.5 Simulation

In Chapter 3.5, simulation studies are conducted to evaluate the performances of M1, M2, and M4.

Nine scenarios, defined by the sample size and true model setting, are considered to examine the properties of the model estimations and penalized posterior Bayes factors. In this simulation study, n individuals are simulated with m observation points per individual, expressed as t_{ij}, at [1,450] minutes with 50-minute increments. The individuals, $\phi_i = (\beta_i, t_{50_i})^{\mathrm{T}}$, are generated from multivariate lognormal distributions with the mean $(\beta, t_{50})^{\mathrm{T}}$ and precision matrix Λ. This profile is generated for the individual i by adding independent normal

errors with precision λ_e.

Three true model (TM) settings are considered:

- In TM setting 1, the measurement errors are from a non-mixture distribution and the random effect parameters are from a non-mixture population distribution as well. The population means are $\beta = 1.5$ and $t_{50} = 1.1$, and the population precision is $\Lambda = \begin{pmatrix} 84 & -1 \\ -1 & 88 \end{pmatrix}$. The parameters $(\beta_i, t_{50_i})^{\mathrm{T}}$ for individuals are generated from the lognormal distribution with the population means and precision. Within subjects, the observations share the same individual parameters β_i, t_{50_i}, and the same error precision, $\lambda_e = 15$.

- In TM setting 2, we would assume the measurement errors are from a mixture of two distributions and the random effect is still from a non-mixture population distribution. The population means are $\beta = 1.5$ and $t_{50} = 1.1$, and the population precision is $\Lambda = \begin{pmatrix} 84 & -1 \\ -1 & 88 \end{pmatrix}$. The parameters $(\beta_i, t_{50_i})^{\mathrm{T}}$ for individuals are generated from the lognormal distribution with the population means and precision. Within subjects, the observations share the same individual parameters β_i, t_{50_i}, and the mixture of error precisions, $\lambda_e = 5$ and $\lambda_e = 20$ with the probability, 0.5, for each.

- In TM setting 3, we assign the measurement errors are from a non-mixture distribution and the random effect is from a mixture of two distributions.The population means are a mixture of $(\beta = 1.8, t_{50} = 1.6)^{\mathrm{T}}$ and $(\beta = 0.6, t_{50} = 0.6)^{\mathrm{T}}$ with an equal probability,

0.5, and the population precision is $\Lambda = \begin{pmatrix} 84 & -1 \\ -1 & 88 \end{pmatrix}$. The parameters $(\beta_i, t_{50_i})^{\mathrm{T}}$ for individuals are generated from the lognormal distribution with the population means and precision. Within subjects, the observations share the same individual parameters β_i, t_{50_i}, and the same error precision, $\lambda_e = 15$.

We assign three different sample sizes for each of the three true models: (10 subjects & 10 observations for each subject), (10 subjects & 20 observations for each subject), and (20 subjects & 20 observations for each subject).

Hence, the simulation study examines all nine combinations of the sample sizes and true model settings. The nine combinations of scenarios are summarized below.

- Scenario 1: TM setting 1 and (10 subjects & 10 observations for each subject).
- Scenario 2: TM setting 2 and (10 subjects & 10 observations for each subject).
- Scenario 3: TM setting 3 and (10 subjects & 10 observations for each subject).
- Scenario 4: TM setting 1 and (10 subjects & 20 observations for each subject).
- Scenario 5: TM setting 2 and (10 subjects & 20 observations for each subject).
- Scenario 6: TM setting 3 and (10 subjects & 20 observations for each subject).
- Scenario 7: TM setting 1 and (20 subjects & 20 observations for each subject).
- Scenario 8: TM setting 2 and (20 subjects & 20 observations for each subject).

- Scenario 9: TM setting 3 and (20 subjects & 20 observations for each subject).

For each scenario we would simulate 100 datasets. We use these simulated data sets to compare three Bayesian hierarchical models, M1, M2, and M4.

The model evaluation results based on penalized posterior Bayes factors (3.12) are summarized in Table 3.1. The penalized posterior marginal likelihood (3.13) of each candidate

	Scenario 1	Scenario 2	Scenario 3
Model 1	0	0	0
Model 2	5	99	0
Model 4	0	0	89
No difference	95	1	11
Total groups	100	100	100
	Scenario 4	Scenario 5	Scenario 6
Model 1	0	0	0
Model 2	3	100	0
Model 4	0	0	91
No difference	97	0	9
Total groups	100	100	100
	Scenario 7	Scenario 8	Scenario 9
Model 1	0	0	0
Model 2	8	100	0
Model 4	0	0	100
No difference	92	0	0
Total groups	100	100	100

Table 3.1: 9 scenarios are evaluated by the penalized posterior marginal likelihood, model 1 is the parametric model, model 2 is the model with DP error, and model 4 is the model with two-layer DP random effects. 100 datasets for each scenario are sampled and parameters are estimated by those three models. The count of a certain model, which is significantly better than other two, is collected. The criterion is whether the log-penalized likelihood of the model is larger than the other two by 4 respectively.

model is calculated for a simulated dataset. 100 datasets are generated from each scenario and three models are implemented for each of those datasets. The logarithmic value of the penalized posterior marginal likelihood of each model for a given dataset is calculated. As Kass and Raftery (1995) mentioned, the first model beats the second model with a strong evidence where Bayes factor value is larger than 10, and with a decisive evidence where the

value is larger than 100. In this dissertation, we choose 50 as the cutoff of the significance. In other words, we consider a certain model is better than other two as its natural logarithmic value is larger than other two by 4. If none of them is larger than other two by 4, we consider that there is no difference among the three models. The comparison results are summarized in Table 3.1. We found that the three models have similar performances when the dataset is generated from non-mixture setting. Scenario 1, 4, and 7 are generated from non-mixture setting with different number of subjects and observations within each subjects. The count of no differences are 95, 97, and 92, respectively. At the same time, the counts, which model 2 is better than the other two, are 5, 3, and 8. Based on those results, we summarize that the three models have similar performances under Scenario 1, 4, and 7. Taking into account the complexities of the three models, model 1 is the proper model because of less computations. When the datasets are simulated from the scenarios with a mixture of measurement errors and non-mixture of random effects, the scenario 2, 5, and 8 are considered. We found the model 2 are always the best model, where the counts are 99, 100, and 100. The model 2, the semiparametric model with DP measurement error, can fit the data with a mixture of measurement error best. In scenario 3, 6, and 9, the datasets are with the non-mixture measurement errors and a mixture of random effects. The counts, where the model 4 is the best model, are 89, 91, and 100. We notice that the model 4, the semiparametric model with two-layer DP random effects, is the best model when the random effects are a mixture distribution. We found that the results are improved when the number of observations of each subject is getting larger (from 10 observations to 20 observations). Figure 3.6 shows

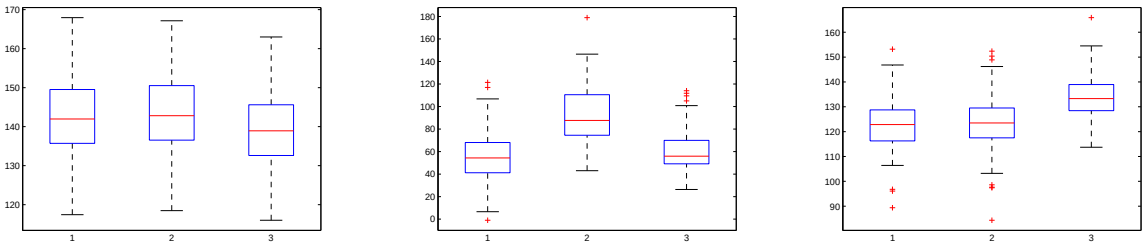
the confidence intervals of penalized posterior marginal likelihood in the simulation study.

We calculate cross validation predictive densities as a different model evaluation method. The results are summarized in Table (3.2).

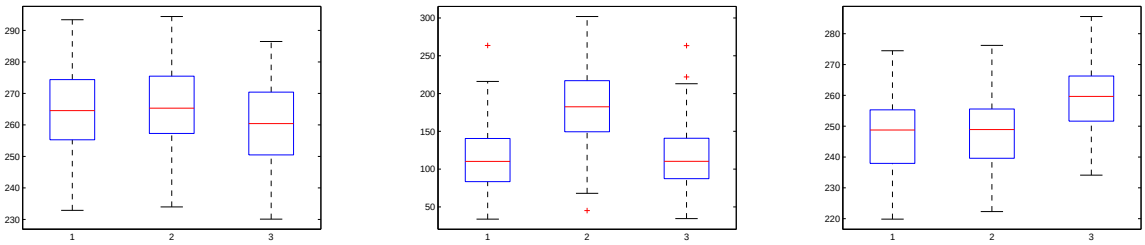
	Scenario 1	Scenario 2	Scenario 3
Model 1	0	0	14
Model 2	0	67	1
Model 4	0	0	3
No difference	100	33	82
Total groups	100	100	100

Table 3.2: 3 scenarios are evaluated by the 10-fold cross validation predictive density: model 1 is the parametric model, model 2 is the model with DP error, and model 4 is the model with DP random effects. 100 datasets are sampled from each scenario. The count of a certain model, which is significantly better than other two, are recorded. The criterion is where the logarithmic value of the cross validation predictive density likelihood is larger than the other two by 4.

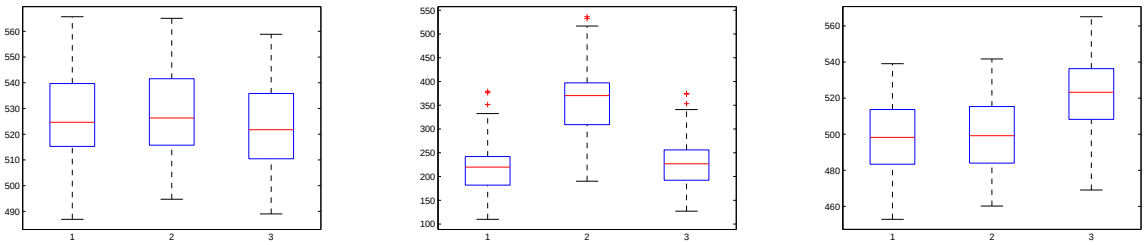
Scenario 1, 2, and 3 are evaluated by the 10-fold cross validation predictive density (CVPD). 100 datasets are sampled for each scenario. The count is the number of a certain model which is significantly better than other two. The criterion is defined as whether the logarithmic value of the cross validation predictive density is larger than the other two by 4. We found the counts for no difference are 100, 33, and 82 for scenario 1, 2, and 3, respectively. CVPD could not help to identify the right model in terms of the high percentage of no difference, while our penalized posterior Bayes factor can. This is because the penalized posterior Bayes factor (3.12) can evaluate the model structure including the penalty term, $p(\Phi|\theta, \Lambda)$, which is important for our model, while CVPD don't contain this penalty term. Therefore, the penalized posterior Bayes factor is more proper approach for the evaluation of the three models. Furthermore this posterior Bayes factor method is useful as we compare



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3



(d) Scenario 4 (e) Scenario 5 (f) Scenario 6



(g) Scenario 7 (h) Scenario 8 (i) Scenario 9

Figure 3.6: The confidence intervals of penalized posterior marginal likelihood: 100 datasets are generated from each of 9 scenarios. The penalized posterior marginal likelihood of three models are calculated for each dataset. The boxplots are to compare the confidence intervals of penalized posterior marginal likelihoods of three models under each scenario.

other parametric and semiparametric hierarchical models as well.

3.6 Application

The example for our study, which was described in Chapter 3.1, is from equine medicine. We take the time of measurement as t_{ij}, and scale of t_{ij} as $t^s_{ij} = t_{ij}/100$. Then the amount of contents remaining in the stomach is expressed as y_{ij}, and scale y_{ij} as $y^s_{ij} = y_{ij}/10000$. A_{0i} stands for the amount of contents remaining in the stomach at time 0 for each horse.

3.6.1 Estimation Results for the Parametric NLME Model

We start with (3.4). In this case, $[\Lambda] = \text{Wishart}(\Lambda|\Lambda_0, \tau)$ is the prior for the precision of the population mean. We set τ as 4, since $\tau \geq d+1$, where $d = 2$ in (3.4). We would like to make τ small so that the variance can be large. Λ_0 is a 2×2 identity matrix. $[(\beta, t_{50})^T|\Lambda] \sim \text{LN}\{(\beta_0, \text{t}_{50_0})^{\text{T}}, (\tau\Lambda)^{-1}\}$ is the prior for the population mean, where $t_{50_0} = -0.7$ and $\beta_0 = 0$. The prior for the measurement error is $\lambda_e = \text{Gamma}(\lambda_e|a_1, b_1)$, where $a_1 = 1$ and $b_1 = 1$.

We fit the NLME with parametric priors, and the results are summarized in Tables 3.3, 3.4, 3.5, 3.8, and 3.11. These results are based on 1000 runs after burn-in time. The medians of β_i and t_{50_i} in (3.4) are given in the left half of Tables 3.3 and 3.4. We denote that "BCI25" and "BCI975" stand for 95% Bayesian lower and upper credible interval. From Table 3.4, we found the smallest value for t_{50_i} is 0.35 ($=t_{50_9}$), which is from horse 1916; the largest one

	Parametric Prior Model (Model 1)			DP Error Model (Model 2)			DP Random Effect Model (Model 4)		
	BCI25	Median	BCI975	BCI25	Median	BCI975	BCI25	Median	BCI975
β_1	0.23	0.73	1.90	0.62	0.89	1.48	0.63	1.03	1.60
β_2	0.97	1.15	1.33	1.03	1.15	1.29	0.98	1.13	1.30
β_3	0.88	1.09	1.25	0.98	1.08	1.17	0.92	1.08	1.27
β_4	1.11	2.15	4.48	1.22	1.87	2.28	0.76	2.08	3.28
β_5	3.43	4.39	5.58	3.62	4.46	5.34	3.29	4.03	5.21
β_6	0.95	1.22	1.50	1.01	1.21	1.43	0.98	1.20	1.48
β_7	1.12	1.28	1.52	1.13	1.29	1.50	1.10	1.26	1.46
β_8	0.84	0.99	1.16	0.90	1.02	1.13	0.90	1.04	1.23
β_9	0.77	1.25	1.99	1.20	1.36	1.55	0.98	1.37	1.97
β_{10}	1.09	1.64	2.74	1.65	1.87	2.12	1.10	1.47	2.00
β_{11}	0.59	0.74	0.91	0.57	0.72	0.87	0.60	0.74	0.88
β_{12}	0.83	1.28	1.03	0.97	1.06	1.15	0.87	1.07	1.28
β_{13}	0.75	0.92	1.11	0.76	0.92	1.19	0.75	0.94	1.12
β_{14}	0.65	1.26	0.91	0.84	0.96	1.05	0.80	1.07	1.36
β_{15}	3.38	3.80	4.30	3.09	3.67	4.45	3.42	3.86	4.36
β_{16}	2.81	3.13	3.44	2.80	3.10	3.44	2.82	3.12	3.46
β_{17}	0.61	0.78	0.97	0.61	0.79	1.00	0.62	0.78	0.97
β_{18}	2.73	3.09	3.46	2.78	3.06	3.43	2.73	3.07	3.42
β_{19}	2.64	3.03	3.46	2.69	3.04	3.38	2.63	3.03	3.55
β_{20}	1.69	1.93	2.27	1.68	1.91	2.27	1.70	1.99	2.27
β_{21}	2.08	2.40	2.72	2.14	2.40	2.69	2.14	2.45	2.79

Table 3.3: The estimation of the individual shape parameters: β_i, $i = 1, \ldots, 21$, represents the shape parameter for the i^{th} individual. The median of β_i from the parametric model, DP error model, and DP random effects model are shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for β_i.

	Parametric Prior Model (Model 1)			DP Error Model (Model 2)			DP Random Effect Model (Model 4)		
	BCI25	Median	BCI975	BCI25	Median	BCI975	BCI25	Median	BCI975
t_{50_1}	0.18	0.52	0.93	0.36	0.45	0.57	0.36	0.59	1.07
t_{50_2}	1.06	1.19	1.30	1.11	1.20	1.29	1.06	1.19	1.30
t_{50_3}	0.63	0.73	0.81	0.68	0.73	0.78	0.66	0.73	0.81
t_{50_4}	1.80	2.24	2.82	2.02	2.24	2.42	1.61	2.21	2.76
t_{50_5}	2.49	2.61	2.72	2.53	2.61	2.70	2.48	2.61	2.75
t_{50_6}	0.38	0.45	0.51	0.40	0.45	0.49	0.38	0.44	0.52
t_{50_7}	1.50	1.59	1.69	1.50	1.59	1.68	1.50	1.58	1.68
t_{50_8}	0.55	0.63	0.72	0.59	0.64	0.68	0.56	0.65	0.73
t_{50_9}	0.26	0.35	0.44	0.32	0.35	0.37	0.28	0.37	0.45
$t_{50_{10}}$	0.37	0.45	0.53	0.42	0.45	0.47	0.37	0.43	0.52
$t_{50_{11}}$	1.31	1.50	1.67	1.32	1.50	1.67	1.29	1.50	1.68
$t_{50_{12}}$	0.57	0.67	0.78	0.64	0.67	0.71	0.59	0.68	0.78
$t_{50_{13}}$	0.70	0.81	0.91	0.70	0.81	0.92	0.70	0.82	0.90
$t_{50_{14}}$	0.32	0.42	0.53	0.39	0.42	0.46	0.36	0.46	0.54
$t_{50_{15}}$	2.86	2.93	3.01	2.79	2.92	3.10	2.86	2.93	3.02
$t_{50_{16}}$	3.28	3.36	3.44	3.26	3.35	3.44	3.28	3.36	3.44
$t_{50_{17}}$	1.20	1.34	1.49	1.21	1.35	1.52	1.23	1.38	1.52
$t_{50_{18}}$	2.56	2.62	2.69	2.54	2.62	2.69	2.55	2.62	2.69
$t_{50_{19}}$	2.05	2.13	2.20	2.07	2.12	2.18	2.06	2.12	2.21
$t_{50_{20}}$	1.73	1.83	1.94	1.71	1.84	1.93	1.73	1.85	1.97
$t_{50_{21}}$	2.80	2.91	3.01	2.81	2.92	3.02	2.80	2.91	3.02

Table 3.4: The estimation of the individual half-meal time parameters: t_{50i}, $i = 1, \ldots, 21$, represents the half-meal time parameter for the i^{th} individual. The median of t_{50i} from the parametric model, DP error model, and DP random effects model are shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for t_{50i}.

	Parametric Prior Model (Model 1)		
	BCI25	median	BCI975
β	1.27	1.48	1.69

Table 3.5: The estimation of the population shape parameters (Model 1): β represents the population shape parameter. The median of β from the parametric model is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for β.

	DP Error Model (Model 2)		
	BCI25	median	BCI975
β	1.30	1.49	1.73

Table 3.6: The estimation of the population shape parameters (Model 2): β represents the population shape parameter. The median of β from the DP error model is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for β.

	DP Random Effects Model (Model 4)		
	BCI25	median	BCI975
β^1	0.77	1.05	1.37
β^2	0.88	1.07	1.74
β^3	0.89	1.05	1.38
β^4	0.94	2.37	3.31
β^5	1.85	3.20	3.31
β^6	0.87	1.05	1.42
β^7	0.84	1.21	2.68
β^8	0.90	1.05	1.37
β^9	0.87	1.05	1.44
β^{10}	0.87	1.07	1.44
β^{11}	0.67	1.05	1.46
β^{12}	0.87	1.05	1.41
β^{13}	0.87	1.05	1.38
β^{14}	0.87	1.05	1.42
β^{15}	1.89	3.20	3.31
β^{16}	1.85	3.20	3.31
β^{17}	0.69	1.05	1.36
β^{18}	1.85	3.20	3.31
β^{19}	1.60	3.20	3.31
β^{20}	1.13	2.36	3.31
β^{21}	1.64	3.01	3.31

Table 3.7: The estimation of the population shape parameters (Model 4): β^i, $i = 1, \ldots, 21$, represents the population shape parameter. The median of β^i from the DP random effects model is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for β^i.

	Parametric Prior Model (Model 1)		
quantile	BCI25	median	BCI975
t_{50}	0.98	1.12	1.28

Table 3.8: The estimation of the population half-meal time parameters (Model 1): t_{50} represents the population shape parameter. The median of t_{50} from the parametric prior is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for t_{50}.

	DP Error Model (Model 2)		
	BCI25	median	BCI975
t_{50}	0.98	1.11	1.27

Table 3.9: The estimation of the population half-meal time parameters (Model 2): t_{50} represents the population half-meal time parameter. The median of t_{50} from the DP error model is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for t_{50}.

	DP Random Effects Model (Model 4)		
	BCI25	median	BCI975
t_{50}^1	0.41	0.74	1.21
t_{50}^2	0.66	0.78	1.81
t_{50}^3	0.54	0.74	1.04
t_{50}^4	0.74	2.27	2.86
t_{50}^5	1.80	2.64	2.98
t_{50}^6	0.41	0.69	0.83
t_{50}^7	0.68	1.36	2.34
t_{50}^8	0.53	0.74	1.04
t_{50}^9	0.41	0.70	0.99
t_{50}^{10}	0.41	0.69	0.90
t_{50}^{11}	0.66	0.85	1.89
t_{50}^{12}	0.53	0.74	1.04
t_{50}^{13}	0.54	0.76	1.22
t_{50}^{14}	0.41	0.69	0.88
t_{50}^{15}	1.80	2.64	3.02
t_{50}^{16}	1.80	2.64	3.06
t_{50}^{17}	0.66	0.78	1.77
t_{50}^{18}	1.80	2.64	2.98
t_{50}^{19}	1.68	2.64	2.95
t_{50}^{20}	1.17	2.18	2.84
t_{50}^{21}	1.80	2.64	3.02

Table 3.10: The estimation of the population half-meal time parameters (Model 4): t_{50}^i, $i = 1, \ldots, 21$, represents the half-meal time parameter for the i^{th} individual. The median of t_{50}^i from the DP random effects model are shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for t_{50}^i.

	Parametric Prior Model (Model 1)		
	BCI25	median	BCI975
λ_e	0.31	0.35	0.39

Table 3.11: The estimation of the population measurement precision parameters (Model 1): λ_e represents the population measurement precision parameter. The median of λ_e from the parametric model is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for λ_e.

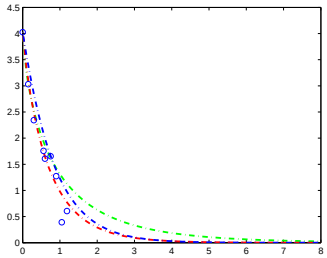
is 3.36 ($=t_{50_{16}}$) which is from horse 1937. The population parameters are provided in Tables 3.5, 3.8, and 3.11. The median for t_{50}, β, and Λ_e are 1.12, 1.48, and 0.35, respectively.

We also display the fitted curves of the NLME model with parametric priors in Figure 3.7; the gray dash-dot lines are fitted curves of the NLME model with parametric priors, and the circles are the observed data. We found that the curves fit the data well. Horses 1890, 1936, 1937, 1947, 1953 and 1965 have a non-decreasing curve pattern which cannot be captured by the power exponential models.

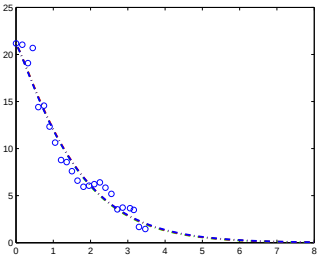
3.6.2 Estimation Results for the NLME Model with DP Errors

We consider (3.3). In this case, $G_0 = \mathrm{N}(\mu_i|0,(g\lambda_{ei})^{-1})\mathrm{Gamma}(\lambda_{ei}|a_2/2,b_2/2)$, where $a_2 = 1$, $b_2 = 1$, and $g = 0.1$. The prior for α is a Gamma(a_3, b_3), where $a_3 = 1$ and $b_3 = 1$. Like the parametric Bayesian model, $[\Lambda] = \mathrm{Wishart}(\Lambda|\Lambda_0,\tau)$ is the prior for the precision of the population mean. We set τ as 4 and Λ_0 as a 2×2 identity matrix. We also assume that $[(\beta,t_{50})^T|\Lambda] = \mathrm{LN}((\beta_0,\mathrm{t}_{50_0})^{\mathrm{T}},(\tau\Lambda)^{-1})$ is the prior for the population mean, where $t_{50_0} = -0.7$ and $\beta_0 = 0$.

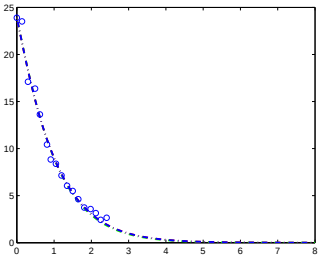
We fit the NLME model with DP errors, and the results are summarized in Tables 3.3, 3.4, 3.6, 3.9, and 3.12. All results are again based on 1000 runs after burn-in time. The medians and the Bayesian credible intervals of β_i and t_{50_i} in (3.3) are given in Table 3.3 and 3.4. From Table 3.4, we notice the smallest value for t_{50_i} is 0.35 ($=t_{50_9}$) which is from horse 1916, while the largest one is 3.35 ($=t_{50_{16}}$) which is from horse 1937. The population



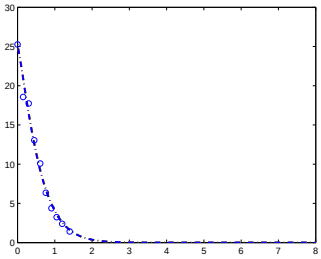
(a) Horse 1852



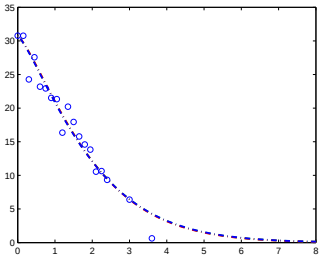
(b) Horse 1881



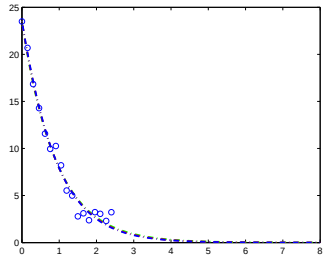
(c) Horse 1884



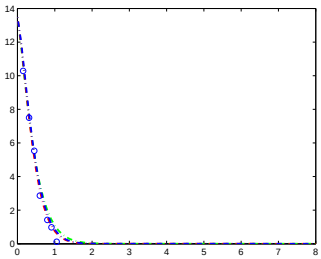
(d) Horse 1891



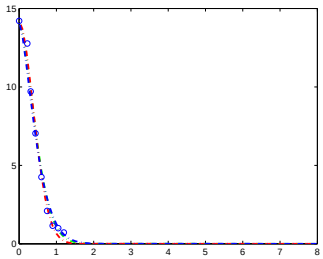
(e) Horse 1894



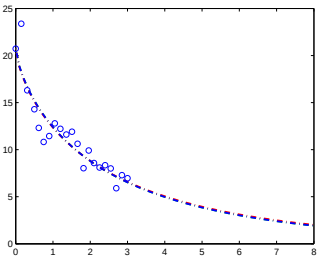
(f) Horse 1914



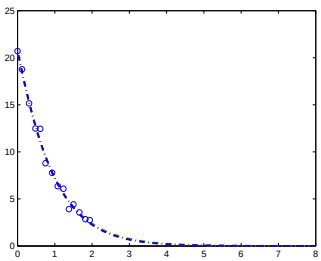
(g) Horse 1916



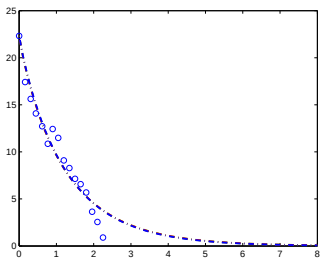
(h) Horse 1920



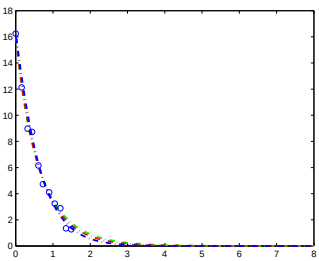
(i) Horse 1925



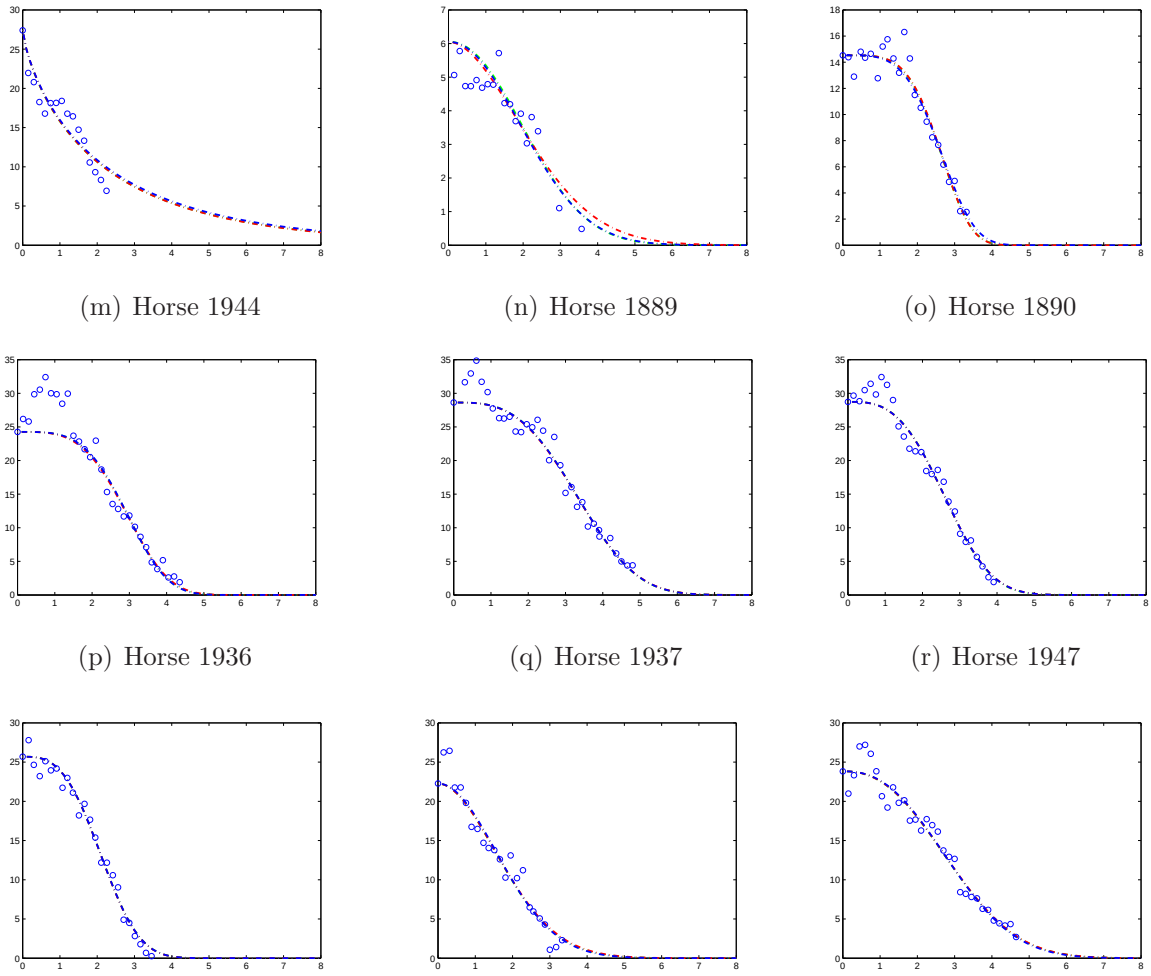
(j) Horse 1927



(k) Horse 1928



(l) Horse 1929

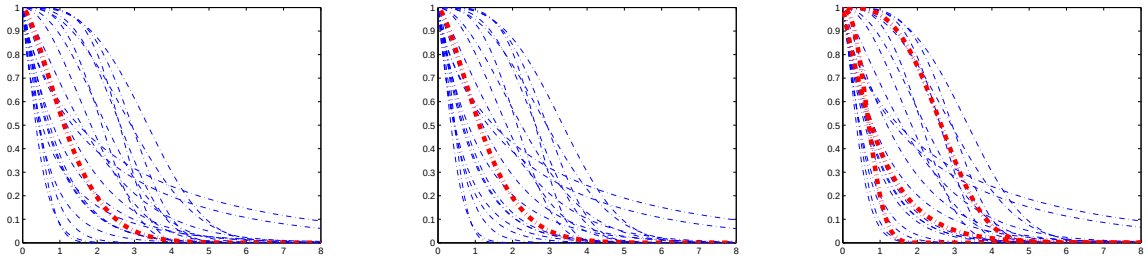


(m) Horse 1944 (n) Horse 1889 (o) Horse 1890

(p) Horse 1936 (q) Horse 1937 (r) Horse 1947

(s) Horse 1949 (t) Horse 1953 (u) Horse 1965

Figure 3.7: The estimated individual curves obtained from fitting NLME: The estimated individual curves obtained from fitting the nonlinear mixed effects model using 21 horses: The gray dash-dot line curves are from the NLME model with parametric priors, the black dash-dot line curves are from the NLME model with DP errors, and the dashed line curves are from the NLME model with DP errors. The circles are the observed points from individual horses.



(a) Model 1 (b) Model 2 (c) Model 4

Figure 3.8: The estimated individual and population curves: The estimated individual curves, which are obtained from fitting the nonlinear mixed effects model using 21 horses, are shown by thin dash lines. The population curves from the three models is shown by thick dash lines.

	DP Error Model (Model 2)		
quantile	BCI25	median	BCI975
λ_e^1	2.81	8.14	29.72
λ_e^2	0.31	0.63	1.02
λ_e^3	0.63	1.19	2.69
λ_e^4	0.96	2.48	4.14
λ_e^5	0.38	0.80	1.28
λ_e^6	0.29	0.69	1.46
λ_e^7	0.22	0.33	0.45
λ_e^8	0.45	0.97	2.10
λ_e^9	1.65	4.37	14.09
λ_e^{10}	1.28	3.40	8.05
λ_e^{11}	0.26	0.33	0.56
λ_e^{12}	1.26	3.05	5.80
λ_e^{13}	0.27	0.37	0.88
λ_e^{14}	1.50	3.60	8.13
λ_e^{15}	0.07	0.12	0.19
λ_e^{16}	0.17	0.29	0.38
λ_e^{17}	0.17	0.31	0.40
λ_e^{18}	0.23	0.33	0.46
λ_e^{19}	0.32	0.69	1.08
λ_e^{20}	0.19	0.32	0.40
λ_e^{21}	0.27	0.34	0.54

Table 3.12: The estimation of the population measurement precision parameters (Model 2): λ_e^i, $i = 1, \ldots, 21$, represents the shape parameter for the i^{th} individual. The median of λ_e^i from the parametric prior and DP error models are shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for λ_e^i.

parameters are also presented in Tables 3.6, 3.9, and 3.12, and the median for t_{50} and β are 1.49 and 3.05 respectively.

We also display individual fitted curves in Figure 3.7. The black dash-dot dash lines are the fitted curves of the NLME model with DP errors, and the circles are the observed data. We found that the curves fit the data well. We also note that Horses 1890, 1936, 1937, 1947, 1953 and 1965 have a non-decreasing curve pattern which cannot be captured by the power exponential models.

3.6.3 Estimation Results for the NLME Model with DP Random Effects

We then fit the NLME model with DP random effects, and the results are summarized in Tables 3.3, 3.4, 3.7, 3.10, and 3.13.

These results are based on 1000 runs after burn-in time. The medians of β_i and t_{50_i} are given in the right part of Table 3.3 and 3.4. The Bayesian credible intervals are also provided as "BCI25" and "BCI975". From Table 3.4, we found the smallest value for t_{50_i} is 0.37 ($=t_{50_9}$), which is from horse 1916; the largest one is 3.36 ($=t_{50_{16}}$) which is from horse 1937. The population parameters are summarized in Tables 3.7, 3.10, and 3.13, and the median for λ_e is 0.35. We also display individual fitted curves in Figure 3.7. The dashed lines are the fitted curves of the NLME model with DP random effects, and the circles are the observed data. We found the curves fit the data well. Horses 1890, 1936, 1937, 1947,

	DP Random Effects Model (Model 4)		
	BCI25	median	BCI975
λ_e	0.31	0.35	0.40

Table 3.13: The estimation of the population measurement precision parameters (Model 4): λ_e represents the measurement precision parameter for the population model. The median of λ_e from the DP random effects model is shown. BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals for λ_e.

1953 and 1965 have a non-decreasing curve pattern. Since the power exponential model has only decreasing pattern, it can not capture a non-decreasing curve pattern.

From Figure 3.9, The estimated modes for the population half-meal time parameters, t_{50}^i, are indicated by Lower triangles, and upper triangle represents the estimated modes for the population shape parameters. In this application, Horse 1-14 had the control meal, octanoic acid with egg, before the test. Horse 15-21 had the test meal along with atropine. According to Model 4, we found that Horses with test meals belong to a cluster with high β^i and t_{50}^i, except Horse 17. Horses with control meals belong to a cluster with low β^i and t_{50}^i, except Horse 4 and 5. Horse 7 has a special pattern with moderate values for β^i and t_{50}^i. Clustering results from Model 4 gave us a clear picture for the relationship between the gastric emptying pattern and the type of meal.

3.6.4 Comparison among Parametric, DP Errors, and DP Random Effects Models

Based on the results shown in Tables 3.3 and 3.4, we found that the estimated medians from the three models are very close to each other, while the credible intervals for the NLME

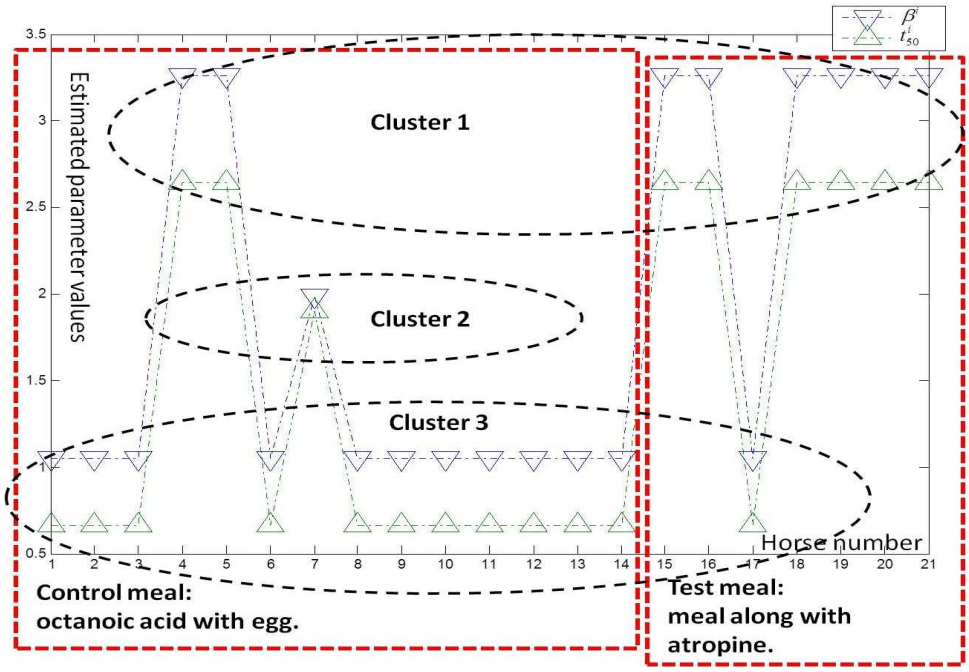


Figure 3.9: Estimated modes for the half-meal time and shape parameters (Model 4): Lower triangles represent the estimated modes for the population half-meal time parameters, t_{50}^i, $i = 1, \ldots, 21$. Upper triangles represent the estimated modes for the shape parameters, β^i for the i^{th} individual. Horse 1-14 had the control meal before the test. Horse 15-21 had the test meal along with atropine before test.

	Log-penalized Posterior Likelihood		
	BCI25	Median	BCI975
Parametric Model (Model 1)	-836.55	-821.28	-811.29
DP Measurement Error Model (Model 2)	-744.51	-728.42	-716.16
DP Random Effects Model (Model 4)	-799.34	-786.07	-773.93

Table 3.14: Logarithmic values for the penalized posterior likelihood: BCI25 and BIC975 represent the lower and upper bounds of 95% Bayesian credible intervals.

with DP errors and random effects are narrower than those of the parametric NLME. We further compare the three models using the penalized posterior Bayes factor.

From the Figure 3.7, we also observe that the individual fitted curves from the three models are overlapped. Hence the estimated variance could be crucial factor in identifying the right model.

Using the population parameters, we found the NLME with random effects (Model 4) clustered the individual subjects into roughly two groups in Tables 3.7 and 3.10. We also notice that model 1 and model 4 gave similar estimated values of β and t_{50}. Model 2 estimated the measurement error based on individual subjects, which is in Table 3.12.

For the penalized posterior Bayes factor, we calculated the logarithmic values of the penalized posterior likelihood, which are shown in Table 3.14. The credible intervals are shown in Figure 3.10.

We found the log-likelihood value of the parametric model (M1), measurement error model (M2), and random effects model (M4) are -821.28, -728.42, and -786.07. The logarithmic penalized posterior Bayes factor of model 2 and model 1 is 92.86 and the value of model 2 and model 4 is 57.65. Since the cutoff value is 4, we conclude model 2 is better than both of model 1 and 4. At the same time, model 4 is better than model 1.

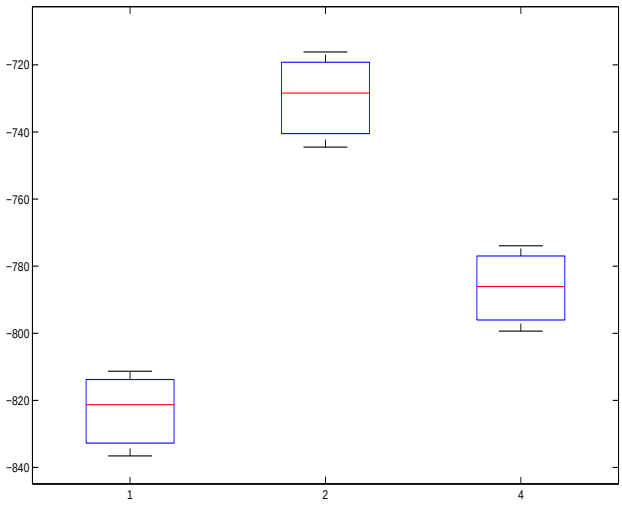


Figure 3.10: The log-penalized posterior marginal likelihood in the application: The box-plots for the log-penalized posterior marginal likelihoods of Model 1, 2, and 4. Model 2 is the best model, and model 4 is also better than model 1.

Chapter 4

Conclusion/Discussion

We have described two topics related to finite and infinite mixture models using Bayesian methods. The key contributions in this dissertation lie in the development of a computational algorithm for hierarchical mixture models. In Chapter 4.1 and 4.2, we describe our main contributions to those two topics and illustrate further discussion on the current and future extensions of this dissertation.

4.1 Summary and Discussion on ARMS Annealing

In Chapter 2, we have described the ARMS annealing method and demonstrated how it works. Though the simulated annealing technique has been known for many years, the real utility is realized only with a set of proper proposal distributions, which allow the samples to travel all over the parameter space. In this dissertation, we incorporate an ARMS sampling

method to help automatically select a set of proper proposal distributions. We have proposed an ARMS annealing approach to reduce the limitations of the EM algorithm so that EM can estimate parameters at the global maximum region. It also develops a more effective Bayesian approach so that the MCMC chain can more easily move from one mode to another. Two approaches were developed to estimate parameters by incorporating ARMS annealing into the EM algorithm and the Bayesian approach, respectively. Using simulation, we compared the two approaches, showing that they are comparable to each other. We found that EM alone often converged to the local mode near to the initial values, but our EM ARMS annealing can detect the global maximum mode and estimate parameters at the maximum mode. The Bayesian approach alone tends to stay around one local mode, but our Bayesian ARMS annealing approach converges to the global maximum mode more quickly because our approach can easily move from one mode to the other.

Our approach is especially useful when the data are from several heterogeneous populations. The mixture of finite models has been used to model heterogeneous populations. For computation time, our approaches are computationally more expensive than EM alone or the Bayesian approach alone, but it is a trade-off for accurate parameter estimation.

In the future, ARMS Annealing can be extended into a ARMS tempering model, which can directly sample from the target distribution with multiple modes. In the dissertation, we already explained how to propose a set of proper proposal distributions. The benefit of the simulated annealing method is to capture the global maximum within the relatively limited time of our Monte-Carlo run. However, the problem is that we cannot guarantee that the

final samples are from the true distribution of our target distribution.

One of the effective ways to overcome this drawback is to perform the simulated tempering method. In simulated tempering, we can perform a random walk in the parameter space, like in the standard Monte-Carlo (MC) method, as well as in the temperature space (t_j in Chapter 2.2.1). After many MC steps in the parameter space we can update the current temperature to perform a random walk in the temperature space. The current temperature could be increased or decreased to the value of the nearest neighbor in the temperature space. In this case, the system visits all points of the temperature space and has a long computation run at each value. Eventually, the final samples will be from the true target distribution, rather than only to detect the global maximum of the parameter space. We are currently working to develop a new method by combining ARMS and tempering method.

4.2 Summary and Discussion on Model Selection of NLME Models

In Chapter 3, we have proposed the semiparametric Bayesian methods for the nonlinear mixed effects model for longitudinal data. Semiparametric Bayesian methods have been widely used for more than 10 years. However, the model selection methods are rarely discussed. A simple and robust model selection method is needed for the model comparison among parametric and semiparametric Bayesian methods.

In this dissertation, we have proposed three semiparametric Bayesian models and a parametric Bayesian model as a baseline model. Based on our best knowledge, the semiparametric Bayesian nonlinear model with two-layer DP random effects, which is proposed in this dissertation, has never been studied. This model is useful because it assumes that the population parameters of random effects have a Dirichlet process prior. When we assume the random effects have a log-normal distribution and the base distribution of Dirichlet process prior is also from a log-normal distribution, the Dirichlet process posterior has a closed form. The probability of the marginal distributions from the Dirichlet process posterior is fairly easy to solve. Moreover, we also proposed a semiparametric Bayesian nonlinear model with one-layer DP random effects and a semiparametric Bayesian nonlinear model with DP measurement errors, which are also rarely discussed under nonlinear modeling settings.

The semiparametric Bayesian nonlinear model with two-layer DP random effects has two attractive properties. When we implement a nonlinear model with one-layer DP random effects, the marginal likelihood from the DP posterior is very difficult to compute in terms of the nonlinear form of the likelihood function. However, the nonlinear model with two-layer DP random effects constructs the marginal likelihood which does not depend on the likelihood function. The marginal likelihood in a DP posterior has a closed form as long as we assign the proper prior for the individual subjects and the base distribution for the population DP prior. This property can dramatically reduce the computation complexity of the nonlinear random effects model. The second attractive property of this model is that the model structure allows us to perform model selection on the semiparametric random effects

model, DP measurement error model, and the parametric random effects model. All of the three models have two layer structure on random effects. We point out that a semiparametric model with one-layer random effects is not comparable with the DP measurement error model and the parametric random effects model because one-layer random effects models don't have population parameters.

In this study, we also proposed a new model selection method, the penalized posterior Bayes factor. The constraint on random effects is added into the penalized posterior Bayes factor as a penalty term. Penalized posterior Bayes factor marginalizes the likelihood function by the posterior distribution of the two-layer parameter. The prior density is one component for computing Bayes factor and it is very flexible under the DP prior setting, where it makes the Bayes factor result sensitive to the value of prior density. In order to compare two-layer DP random effects model, DP measurement error model, and parametric random effects model, we calculate the marginal penalized likelihood over the posterior distribution of the two-layer parameter. Based on the simulation study, the penalized posterior Bayes factor is a robust and stable method to use to compare the parametric and semiparametric hierarchical nonlinear models.

In the application, we found both of the models with DP measurement errors (Model 2) and with two-layer DP random effects (Model 4) are better than parametric model (Model 1). It means the other possible model may need to have both mixing random effects and mixing measurement errors. However, the current model only allows us to choose one of them, which is mixing measurement errors. Future research suggested by this study can

develop a more flexible model using two DP priors for both random effects and measurement errors. The model should consider both mixing random effects and mixing measurement errors. We may consider the model that includes two DP priors, one on measurement error term and the other on random effects term. How to arrange the DP priors and how to construct a better computational structure will be the challenges, because this model might have slow convergence.

REFERENCES

Aitkin, M. (1995) Posterior Bayes Factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53, 111-142.

Basu, S. and Chib, S. (2003) Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models. *Journal of the American Statistical Association*, 98, 224-235.

Besag, J. and Green, P. J. (1993) Spatial Statistics and Bayesian Computation. *Journal of the Royal Statistical Society*, B 55, 25-37.

Chib, S. (1995) Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90, 1313-1321.

Chib, S. and Greenberg, E. (2010) Additive Cubic Spline Regression with Dirichlet Process Mixture Errors. *Journal of Econometrics*, 156, 322-336.

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

Damien, P. and Wakefield, J. and Walker, S. (1999) Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables. *Journal of the Royal Statistical Society*, B 61, 331-344.

Davidian, M. and Gallant, A. R. (1992) Smooth Nonparametric Maximum Likelihood Estimation for Population Pharmacokinetics, with Application to Quinidine. *Journal of*

Pharmacokinetics and Biopharmaceutics, 20, 529-556.

Davidian, M. and Giltinan, D. M. (1995) *Nonlinear Models for Repeated Measurement Data* , Chapman and Hall/CRC.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society Series B.*, 39, 1-38.

Duncan, B. (1999) Modeling Charitable Contributions of Time and Money. *Journal of Public Economics*, 77, 213-242.

Edwards, R. G. and Sokal, A. D. (1988) Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm.*Physical Review D (Particles and Fields)* , 38, 2009-2012.

Elashoff, J. D. and Reedy T. J. and Meyer J. H. (1982) Analysis of Gastric Emptying Data. *Gastroenterology*, 83, 1306-1312.

Elashoff, J. D. and Reedy T. J. and Meyer J. H. (1983) Methods for Gastric Emptying Data (Letter to the Editor). *Gastrointestinal Liver Physiology*, 7, G701-G702.

Escobar, M. D. and West, M. (1995) Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90, 577-588.

Ferguson, T. S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The Annals*

of Statistics , 1, 209-230.

Gilks, W. R. and Best, N. G. and Tan, K. K. C (1995) Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Applied Statistics*, 44, 455-472.

Gilks, W. R. and Wild, P. (1992) Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, 41, 337-348.

Hastings, W. K. (1970) Monte Carlo Sampling Methods Using Markove Chains and Their Applications. *Biometrika*, 57, 97-109.

Higdon, D. M. (1998) Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications. *Journal of the American Statistical Association*, 93, 585-595.

Ishwaran, H. and James, L. F. (2001) Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96. Theory and Methods, 161-173.

Von Neumann, J. (1951) Various Techniques Used in Connection with Random Digits. Monte Carlo Methods, *National Bureau of Standards*, 12, 36-38.

Kass, R. E. and Raftery, A. E. (1995) Bayes Factors. *Journal of American Statistical Association*. 90, 773-595.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983) Optimization by Simulated Annealing. *Science, New Series*, 220, 671-680.

Kleinmanm, K. P. and Ibrahim, J. G. (1998) A Semiparametric Bayesian Approach to the Random Effects Model. *Biometrics*, 54, 921-938.

Kong, A. and Liu, J. S. and Wong, W. (1994) Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association*, 89, Theory and Methods, 278-288.

Lindley, D. V. (1957) A Statistical Paradox. *Biometrika*, 44, 187-192.

Lindstrom, M. J. and Bates, D. M. (1990) Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, 46, 673-687.

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

Neal, R. M. (2003) Slice Sampling. *Ann. Statist*, 31, 705-767.

Newton, M. A. and Raftery, A. E. (1994) Approximate Bayesian Inference by the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Association*, 56, 3-48.

Sethuraman, J. (1994) A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4, 639-650.

Steimer, J. L., Mallet, A., Golmard, J. L., and Boisvieux, J. F. (1984) Alternative Approaches to Estimation of Population Pharmacokinetic Parameters: Comparison with the Nonlinear Mixed-Effect Model. *Drug Metabolism Reviews*, 15, 265-292.

Sheiner, L. B. and Beal, S. L. (1980) Evaluation of Methods for Estimating Population Pharmacokinetic Parameters. I. Michaelis-menten Model: Routine Clinical Pharmacokinetic Data. *Journal of Pharmacokinetics and Pharmacodynamics*, 8, 553-571.

Smith, V. H., Kchoe, M. R., and Creamer, M. E. (1999) The Private Provision of Public Goods: Altruism and Voluntary Giving. *Journal of Public Economics*, 77, 213-242.

Swendsen, R. H. and Wang J. S. (1987) Nonuniversal Critical Dynamics in Monte Carlo Simulations. *Physical Review Letters* , 58, 86-88.

Petrone, S. and Raftery, A. E. (1997) A Note on the Dirichlet Process Prior in Bayesian Nonparametric Inference with Partial Exchangeability. *Statistics & Probability Letters* 36, 69-83.

Vonesh, E. F. and Carter, R. L. (1992) Mixed-Effects Nonlinear Regression for Unbalanced Repeated Measures. *Biometrics*, 48, 1-17.

Vonesh, E. F. and Chinchilli, V. M. (1997) *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, CRC Press.

Appendix A

Bayesian Analysis of the Multivariate Normal Distribution

This appendix gives detailed derivations in Chapter 3.3.2. Suppose both the mean μ_i and precision $\lambda_{ei} = \sigma_i^{-2}$ for the measurement error of the i^{th} subject unknown. We define $y_{ij}^* = y_{ij} - f_{ij}$, $\underline{Y_i^*} = \underline{Y_i} - \underline{f_i}$, where $\underline{Y_i} = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})^{\mathrm{T}}$ and $\underline{f_i} = (f_{i1}, \ldots, f_{ij}, \ldots, f_{in_i})^{\mathrm{T}}$. The likelihood of $\underline{Y_i^*}$ can be written as

$$\begin{aligned} p(\underline{Y_i^*}|\mu_i, \lambda_{ei}) &= \frac{1}{(2\pi)^{\frac{n_i}{2}}} {\lambda_{ei}}^{\frac{n_i}{2}} \exp\{-\frac{\lambda_{ei}}{2} \sum_{j=1}^{n_i} (y_{ij}^* - \mu_i)^2\} \\ &= \frac{1}{(2\pi)^{\frac{n_i}{2}}} {\lambda_{ei}}^{\frac{n_i}{2}} \exp[-\frac{\lambda_{ei}}{2} \{n_i(\mu_i - \overline{\underline{Y_i^*}})^2 + \sum_{j=1}^{n_i} (y_{ij}^* - \overline{\underline{Y_i^*}})^2\}]. \end{aligned}$$

A.1 The Marginal Prior for the Measurement Error Mean, μ_i

The conjugate prior for μ_i and λ_{ei} is the Normal-Gamma prior:

$$\begin{aligned}
\mathrm{NG}(\mu_i, \lambda_{ei}|\mu_0 = 0, g, v_0, s_0) &= \mathrm{N}\{\mu_i|(g\lambda_{ei})^{-1}\}\mathrm{Gamma}\{\lambda_{ei}|\frac{v_0}{2}, rate = \frac{v_0 s_0^2}{2}\} \\
&= \frac{1}{Z_{NG}}{\lambda_{ei}}^{1/2}\exp(-\frac{g\lambda_{ei}}{2}\mu_i^2){\lambda_{ei}}^{v_0/2-1}e^{-\lambda_{ei}v_0 s_0^2/2} \\
&= \frac{1}{Z_{NG}}{\lambda_{ei}}^{\frac{v_0+1}{2}-1}\exp\{-\frac{\lambda_{ei}}{2}(g\mu_i^2 + v_0 s_0^2)\},
\end{aligned}$$

where

$$Z_{NG}(\mu_0, g, a_2, b_2) = \frac{\Gamma(\frac{v_0}{2})}{(\frac{v_0 s_0^2}{2})^{\frac{v_0}{2}}}(\frac{2\pi}{g})^{1/2}.$$

We can compute the marginal prior on μ_i, by integrating out λ_{ei}, as follows:

$$\begin{aligned}
p(\mu_i) &\propto \int p(\mu_i, \lambda_{ei})d\lambda_{ei} \\
&= \int {\lambda_{ei}}^{\frac{v_0+1}{2}-1}\exp\{-\lambda_{ei}(\frac{v_0 s_0^2 + g\mu_i^2}{2})\}d\lambda_{ei}.
\end{aligned}$$

Since this formula is an unnormalized Gamma$\{a^* = (v_0+1)/2, b^* = (v_0 s_0^2 + g\mu_i^2)/2\}$ distribution, we can write down

$$\begin{aligned} p(\mu_i) \propto \frac{\Gamma(a^*)}{(b^*)^{a^*}} &\propto (b^*)^{-a^*} \\ &= (\frac{v_0 s_0^2 + g\mu_i^2}{2})^{-(v_0+1)/2} \\ &= \{1 + \frac{1}{v_0}\frac{g\mu_i^2}{s_0^2}\}^{-(v_0+1)/2}, \end{aligned}$$

where is a $\mathrm{T}_{v_0}\{\mu_i|\mu_0 = 0, s_0^2/g\}$ distribution.

A.2 The Posterior Distributions for the Measurement Error Parameters, μ_i and λ_{ei}

The posterior can be derived as follows:

$$p(\mu_i, \lambda_{ei}|\underline{Y_i^*}) \propto \mathrm{NG}(\mu_i, \lambda_{ei}|\mu_0, g, v_0, s_0) p(\underline{Y_i^*}|\mu_i, \lambda_{ei}) \tag{A.1}$$

$$\propto {\lambda_{ei}}^{1/2} e^{-\frac{g\lambda_{ei}\mu_i^2}{2}} {\lambda_{ei}}^{\frac{v_0}{2}-1} e^{-\frac{v_0 s_0^2}{2}\lambda_{ei}} {\lambda_{ei}}^{n_i/2} e^{-\frac{\lambda_{ei}}{2}\sum^{n_i}(y_{ij}^*-\mu_i)^2} \tag{A.2}$$

$$\propto {\lambda_{ei}}^{1/2} {\lambda_{ei}}^{(v_0+n_i)/2-1} e^{-(v_0 s_0^2/2)\lambda_{ei}} e^{-\frac{\lambda_{ei}}{2}\{g(\mu_i)^2+\sum(y_{ij}^*-\mu_i)^2\}}. \tag{A.3}$$

From the previous equation A.1, we have

$$\sum_{j=1}^{n_i}(y_{ij}^* - \mu_i)^2 = n_i(\mu_i - \overline{\underline{Y_i^*}})^2 + \sum_{j=1}^{n_i}(y_{ij}^* - \overline{\underline{Y_i^*}})^2.$$

Then we can show that

$$\begin{aligned} g\mu_i^2 + \sum (y_{ij}^* - \mu_i)^2 &= g\mu_i^2 + n_i(\mu_i - \overline{Y_i^*})^2 + \sum (y_{ij}^* - \overline{Y_i^*})^2 \\ &= (g+n_i)(\mu_i - \mu_{post})^2 + \frac{g n_i (\overline{Y_i^*})^2}{g+n_i} + \sum (y_{ij}^* - \overline{Y_i^*})^2. \end{aligned}$$

Finally we obtain:

$$\begin{aligned} & p(\mu_i, \lambda_{ei} | \underline{Y_i^*}) \\ \propto\ & {\lambda_{ei}}^{1/2} e^{-(\lambda_{ei}/2)(g+n_i)(\mu_i - \mu_{post})^2} \times {\lambda_{ei}}^{(v_0+n_i)/2-1} \\ & e^{-(\lambda_{ei}/2)\sum (y_{ij}^* - \overline{Y_i^*})^2} e^{-(\lambda_{ei}/2)\frac{g n_i (\overline{Y_i^*} - \mu_0)^2}{g+n_i}} e^{-\lambda_{ei} v_0 s_0^2/2} \\ \propto\ & \mathrm{N}\{\mu_i | \mu_{post}, (g_{post}\lambda_{ei})^{-1}\} \times \mathrm{Gamma}(\lambda_{ei} | a_{post}/2, b_{post}/2), \end{aligned}$$

where $\mu_{post} = (g\mu_0 + n_i\overline{Y_i^*})/(g+n_i)$, $g_{post} = g + n_i$, $a_{post} = v_0 + n_i$, and $\{b_{post} = v_0 s_0^2 + \sum^{n_i} (y_{ij}^* - \overline{Y_i^*})^2 + g n_i (\overline{Y_i^*})^2\}/(g+n_i)$.

A.3 The Marginal Distribution for $\underline{Y_i}$

Assume $Y_i^* = (y_{i1}, y_{i2}, \ldots, y_{in_i})^{\mathrm{T}}$, and $\mu_0 = 0$. Then

$$\begin{aligned} p(\underline{Y_i^*}) &\propto \int\int NG(\mu_i, \lambda_{ei} | \mu_0 = 0, g, v_0, s_0) p(\underline{Y_i^*} | \mu_i, \lambda_{ei}) d\mu_i d\lambda_{ei} \\ &\propto (\beta_n/2)^{-\alpha_n/2} \\ &\propto [1/2\{v_0 s_0^2 + \sum^{n_i} (y_{ij}^* - \overline{Y_i^*})^2 + \frac{g n_i (\overline{Y_i^*})^2}{(g+n_i)}\}]^{-(v_0+n_i)/2}. \end{aligned}$$

It can be shown:

$$\begin{aligned} p(\underline{Y_i^*}) &\propto \{b_2 + \sum^{n_i}(y_{ij} - \overline{\underline{Y_i^*}})^2 + \frac{gn_i(\overline{\underline{Y_i^*}})^2}{g+n_i}\}^{-(v_0+n_i)/2} \\ &\propto (1 + \frac{1}{v_0}\frac{(\underline{Y_i^*})'C(\underline{Y_i^*})}{s_0^2})^{-(v_0+n_i)/2}, \end{aligned}$$

where

$$C = \begin{pmatrix} 1 - \frac{1}{g+n_i} & \dots & \frac{-1}{g+n_i} \\ \dots & \dots & \dots \\ \frac{-1}{g+n_i} & \dots & 1 - \frac{1}{g+n_i} \end{pmatrix}.$$

Therefore, we can know that $\underline{Y_i^*}$ is from a multivariate t distribution $\mathrm{T}_{v_0}\{0, \Sigma = s_0^2/C\}$. By assuming $\underline{Y_i^*} = \underline{Y_i} - \underline{f_i}$, it is easy to obtain marginal distribution for $\underline{Y_i}$ is from a multivariate t distribution $\mathrm{T}_{v_0}\{\underline{f_i}, \Sigma = s_0^2/C\}$.

Appendix B

Conjugate Bayesian Analysis for the Random Effects

We assume that $\phi_i = (\beta_i, t_{50_i})$ is from a multivariate lognormal distribution, where $\ln(\phi_i)$ is from the normal distribution with the mean, $\ln(\theta_i)$ and the precision, Λ_i. We also assign a prior for $\ln(\theta_i)$, which is a multivariate normal with the mean, θ_0 and the precision, Λ_{00}. The prior for Λ_i is a Wishart distribution with a positive definite matrix, Λ_0/τ and the degree of freedom, τ.

B.1 Likelihood of the Individual Parameter, ϕ_i

We assume ϕ_i is a 2×1 vector, and the likelihood of ϕ_i is

$$p(\phi_i|\theta_i, \Lambda_i) = (2\pi)^{-h/2}\frac{1}{\beta_i t_{50_i}}|\Lambda_i|^{1/2}\exp[-(\ln\phi_i - \ln\theta_i)^{\mathrm{T}}\Lambda_i(\ln\phi_i - \ln\theta_i)]. \tag{B.1}$$

B.2 Prior of Population Parameters, θ_i and Λ_i

The prior of θ_i and Λ_i can be written as,

$$\begin{aligned} p(\ln\theta_i, \Lambda_i) &= p(\ln\theta_i)p(\Lambda_i) \\ &= \mathrm{N}\{\ln\theta_i|\theta_0, (\Lambda_{00})^{-1}\}\mathrm{Wishart}(\Lambda_i|\Lambda_0/\tau, \tau) \\ &= \frac{1}{Z}|\Lambda_{00}|^{1/2}\exp\{-\frac{1}{2}(\ln\theta_i - \theta_0)^{\mathrm{T}}\Lambda_{00}(\ln\theta_i - \theta_0)\} \\ &\times \ |\Lambda_i|^{(\tau-h-1)/2}\exp[-\frac{1}{2}\mathrm{Tr}\{(\Lambda_0/\tau)^{-1}\Lambda_i\}], \end{aligned} \tag{B.2}$$

where $Z = (\frac{1}{2\pi})^{h/2}|\Lambda_0/\tau|^{\tau/2}2^{\tau h/2}\Gamma_h(\tau/2)$.

B.3 Posterior of Population Parameters, θ_i and Λ_i

The posterior distributions of θ_i and Λ_i can be written as,

$$p(\ln\theta_i, \Lambda_i|\phi_i) = \mathrm{N}\{\ln\theta_i|\theta_{ipost}, \Lambda_{ipost}^{-1}\}\mathrm{Wishart}\{\Lambda_i|\Lambda_i^d/(\tau^d), \tau^d\}, \tag{B.3}$$

and

$$\begin{aligned}
\theta_{ipost} &= (\Lambda_{ipost})^{-1}\{\Lambda_i \ln\phi_i + \Lambda_{00}\theta_0\}, \\
\Lambda_{ipost} &= \Lambda_{00} + \Lambda_i, \\
\Lambda_i^d &= S_i + \tau\Lambda_0^{-1}, \\
\tau^d &= \tau + 1.
\end{aligned}$$

B.4 Marginal Likelihood of the Individual Parameter, ϕ_i.

The marginal likelihood of conditional variance parameters, ϕ_i, can be written as a ratio of constants,

$$\begin{aligned}
p(\phi_i) &= \frac{p(\phi_i|\theta_i,\Lambda_i)p(\theta_i)p(\Lambda_i)}{p(\theta_i,\Lambda_i|\phi_i)} \\
&= \frac{Z^*}{Z}\frac{1}{(2\pi)^{h/2}\beta_i t_{50_i}} \\
&= \frac{1}{\pi^{h/2}}\frac{1}{\beta_i t_{50_i}}\frac{\Gamma_h(\tau^d/2)}{\Gamma_h(\tau/2)}\frac{|\Lambda_{ipost}/\tau^d|^{\tau^d/2}}{|\Lambda_{00}/\tau|^{\tau/2}}.
\end{aligned}$$