UNIVERSITY OF CALIFORNIA,
IRVINE

Bayesian Cluster Analysis with Longitudinal Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Yan He

Dissertation Committee:
Professor Wesley O. Johnson, Chair
Professor Babak Shahbaba
Professor Dan Gillen

2014

DEDICATION

To my parents

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor Wes Johnson for letting me work on intersting problems. His guidance and patience enabled me to finish the thesis.

CURRICULUM VITAE

Yan He

EDUCATION

Ph.D. in Statistics **June 2014**
University of California, Irvine *Irvine, California*

M.S. in Physics **March 2007**
University of California, Irvine *Irvine, California*

B.S. in Physics **June 2005**
University of Science and Technology of China *Hefei, China*

RESEARCH EXPERIENCE

Research Assistant in Statistics **2010 – 2014**
University of California, Irvine *Irvine, California*
Research Assistant in Statistics **2009**
University of California, Irvine *Irvine, California*
Research Assistant in Physics **2006 – 2008**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teacher Assistant in Statistics **2008 – 2013**
University of California, Irvine *Irvine, California*

Teacher Assistant in Physics **2005 – 2008**
University of California, Irvine *Irvine, California*

WORKING EXPERIENCE

Statistical Consultant **2012 – 2013**
Strategic Marketing Science Inc. *Newport Beach, California*

Intern Statistician **2013**
Allergan Pharmaceuticals, Inc. *Irvine, California*

PUBLICATIONS AND TALKS

Bayesian Inference for Assessing the Association Between Urinary Incontinence and Hormone Profiles During the Menopausal Transition **2013**
Contributed talk at JSM

Tuning transport properties of nanofluidic devices with local charge inversion **2009**
Journal of the American Chemical Society

Synthetic Nanopores as a Test Case for Ion Channel Theories: The Anomalous Mole Fraction Effect **2008**
Biophysical Journal

ABSTRACT OF THE DISSERTATION

Bayesian Cluster Analysis with Longitudinal Data

By

Yan He

Doctor of Philosophy in Statistics

University of California, Irvine, 2014

Professor Wesley O. Johnson, Chair

In the thesis, we focused on cluster analysis with longitudinal data. In the Study of Womens Health Across the Nation (SWAN), we are interested in identifying trending groups based on repeated hormone observations, and in the association between hormone trends and womens health during menopause. We proposed a Bayesian semi-parametric model with a Dirichlet Process (DP) that allows for irregular data and model-based clustering of women based on their hormone profiles. We identified distinct developmental trajectories for both E2 and FSH hormones, and discussed relationships between profiles and other factors. Urinary Incontinence (UI) is considered as a specific measure of womens health, and we extended our approach to jointly model UI and hormone outcomes in order to explore the association between UI status and the pattern of hormone changes over the menopausal transition.

In the study of Johnes disease in cattle, we developed a particular finite mixture model for diagnosis based on bivariate longitudinal outcomes. The subjects were clustered into three disease phases: uninfected and two categories of disease.

In addition, we discussed Label Switching issues in Bayesian finite and infinite mixture models, and proposed an ad-hoc method to address them to obtain inferences for the parameters on the component level.

Chapter 1

Introduction and Background Materials

In the Bayesian framework, parametric modeling of a data vector y assumes that y is characterized by a parametric distribution $F_\theta(y)$. The collection of parameters θ is modeled using a prior distribution $p(\theta)$. Thus the general form of a Bayesian parametric model is written as:

$$\begin{aligned} y|\theta &\sim F_\theta(y) \\ \theta &\sim p(\theta) \end{aligned} \qquad (1.1)$$

Inferences for the parameters can be obtained using the posterior density obtained via Bayes' rule, $f(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta} \propto f(y|\theta)p(\theta)$, and predictive inference can be obtained using the predictive density;

$$f(y^{new}|y) = \int f(y^{new}|\theta)p(\theta|y)d\theta \qquad (1.2)$$

where $y^{new} \perp y|\theta$.

In Bayesian parametric models, inferences will be sensitive to the choice of the probability model $F_\theta(y)$. Here, we would like to relax parametric assumptions.

A Bayesian non-parametric model generally involves an infinite-dimensional parameter space, which of course is a paradox in terminology. However, since we live in a finite dimensional world, non-parametric models are often approximated by using a large but finite number of parameters. We would argue that they are thus flexibly parametric. Bayesian non-parametric models, such as those based on the Dirichlet Process (DP), allow the number of parameters, in theory to be infinite, but in applications, the number of parameters grows to adapt to the data.

1.1 The Dirichlet Process

The DP (Ferguson 1973) [11] is a stochastic process that has been widely used for the Bayesian non-parametric modeling of data. In particular, we assume that data, say $(X_1, \ldots, X_n)$, are distributed according to an unknown distribution G. Since we use the Bayesian approach, we need a so-called prior distribution for G. This is a complicated matter, that has been studied in great detail starting with Ferguson (1973) and before. Here we describe the Dirichlet Process, a particular richly parametric probability model for G.

1.1.1 The Formal Definition of DP

Before defining the DP, we introduce the Dirichlet distribution. Let $Z_1, Z_2, ..., Z_k$ be independent random variables with $Z_j \overset{\perp}{\sim} \Gamma(\alpha_j, 1)$, where $\alpha_j \geq 0$ for all j, and $\alpha_j > 0$ for some j, and define $Y_j = \frac{Z_j}{\sum_{i=1}^{k} Z_i}$, for $j \in \{1, 2, ..., k\}$. Then the distribution of $(Y_1, Y_2, ..., Y_k)$ is defined as the Dirichlet distribution with parameter $(\alpha_1, \alpha_2, ..., \alpha_k)$, denoted by $(Y_1, Y_2, ..., Y_k) \sim \mathcal{D}(\alpha_1, \alpha_2, ..., \alpha_k)$.

The Dirichlet distribution is a multivariate generalization of the Beta distribution, which is often used as a prior distribution for multinomial data in Bayesian analysis. The distribution has mean $E(Y_j) = \frac{\alpha_j}{\alpha}$ and variance $Var(Y_j) = \frac{\alpha_j}{\alpha}(1 - \frac{\alpha_j}{\alpha})/(1 + \alpha)$ with $\alpha = \sum_{i=1}^{k} \alpha_i$.

Ferguson (1973) [11] formally extended the simple Dirichlet distribution to the DP, and established its existence using the Kolmogorov's Consistency Theorem. Briefly, let μ be a non-null finite measure on a measurable space $(\Omega, \mathcal{F})$. The random probability measure G is defined as a DP on $(\Omega, \mathcal{F})$ if for every finite k, and measurable partition $(A_1, ..., A_k)$ of Ω, the distribution of $(G(A_1), ..., G(A_k))$ is Dirichlet distributed as $\mathcal{D}(\mu(A_1), ..., \mu(A_k))$, denoted by $G \sim \mathcal{DP}(\mu)$. In practice, μ can be represented as $\mu = \alpha G_0$, where $\alpha > 0$ is a concentration parameter and G_0 is the so-called base distribution of the DP. We write the DP as

$$G \sim \mathcal{DP}(\alpha, G_0)$$

Since the Dirichlet distribution is a multivariate generalization of the Beta distribution, we have $G(A)|\alpha, G_0 \sim Beta\big(\,\alpha G_0(A), \alpha(1 - G_0(A))\,\big)$ for any $A \in \mathcal{F}$. We can see that the base distribution is the mean of the DP $(E(\,G(A)\,) = G_0(A))$, and the concentration parameter is related to the variance $\big(\,Var(\,G(A)\,) = \frac{G_0(A)(1-G_0(A))}{1+\alpha}\,\big)$. Figure 1.1 shows that the variance of the DP decrease when α increases.

1.1.2 Stick-breaking Process

Sethuraman (1994) [34] provided a constructive representation of the DP, named a stick-breaking process. Let $\{W_i : i = 1, 2, \ldots\}$ be an infinite set of independent and identically distributed $Beta(1, \alpha)$ variables, and $\{X_i : i = 1, 2, \ldots\}$ be a set of independent random

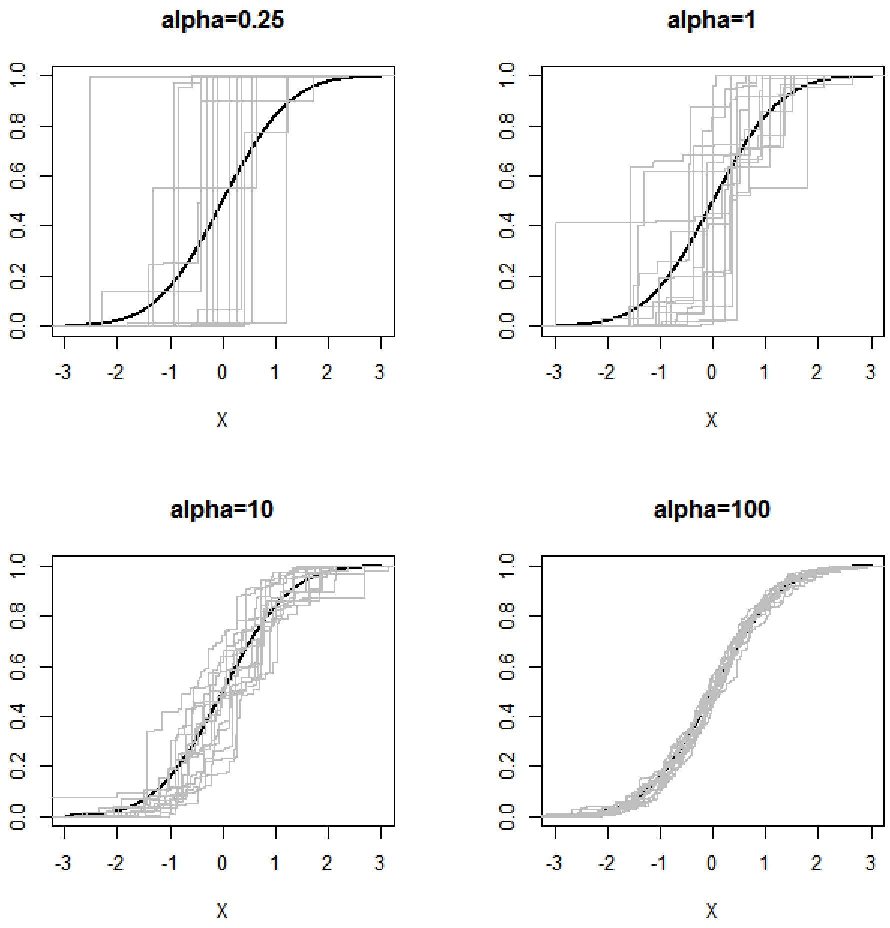


Figure 1.1: Plots of sample CDF's from $G \sim \mathcal{DP}(\alpha,\, G_0 = N(0,1))$ with four α values. In each plot, the black line is the CDF of the base distribution, $N(0,1)$, and the gray lines are the empirical CDF's of 15 realizations from G. The Sethuraman characterization in Equation (1.3) was used to obtain the samples based on truncation at 1000 terms in the sum.

samples drawn from the base distribution G_0. Also define $P_i = W_i \prod_{j=1}^{i-1}(1 - W_j)$. Then if $G \sim \mathcal{DP}(\alpha, G_0)$, G can be expressed as:

$$G(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{X_i}(\cdot) \tag{1.3}$$

where $\delta_{X_i}(\cdot)$ is the point mass at X_i.

We used the stick-breaking process to draw samples from $\mathcal{DP}\big(\alpha, G_0 = N(0,1)\big)$ in Figure 1.1. With this process, it is apparent that DP is discrete with probability one.

1.1.3 Polya Urn Scheme

Blackwell and MacQueen (1973) [4] provided an alternative approach to the DP by exploiting its connection with generalized Polya Urn schemes. In [4], they defined a Polya sequence and described the connection between Polya sequences with the DP.

A Polya sequence with parameter μ ($\mu = \alpha G_0$) is a sequence of random variables $\{X_i : i = 1, 2, \ldots\}$ taking values in Ω, that satisfy

$$\begin{aligned} P(X_1 \in A) &= G_0(A) \\ P(X_{i+1} \in A | X_1, \ldots, X_i) &= G_n(A) \end{aligned} \tag{1.4}$$

for any $A \subset \mathcal{F}$, where $G_n(\cdot) = \frac{\alpha}{\alpha+n} G_0(\cdot) + \frac{1}{\alpha+n} \sum_{i=1}^{n} \delta_{X_i}(\cdot)$, where $\delta_x(\cdot)$ denotes point mass at x. They proved if $\{X_n : n = 1, 2, \ldots\}$ is a Polya sequence with parameter αG_0, G_n converges to $G \sim \mathcal{DP}(\alpha, G_0)$ with probability one.

Ferguson also established that the Polya Urn scheme describes the marginal distribution of a random sample from the DP, G. The form shown in Equation (1.4) is extremely useful for posterior sampling using the Gibbs sampler in the context of a marginalized DP, because it can be used to derive full conditional distributions for parameters.

1.1.4 Chinese Restaurant Process

The Chinese restaurant process (CRP) gives a nice way to describe the Polya Urn scheme, which clearly shows the clustering capability of the DP. The process can be described as follows:

- Initially imagine an empty restaurant containing an infinite number of tables.
- The first person to enter sits down at a table (selects a cluster), and orders food from the menu for the table (selects parameters from the base distribution for the cluster). Then everyone else who joins the table shares the same food (parameters) with him/her.
- The second person to enter sits down at a table. With probability $\frac{\alpha}{1+\alpha}$ he/she sits down at a new table (selects a new cluster) and orders food for the table; with probability $\frac{1}{1+\alpha}$ he/she sits with the first person and shares the food (parameters) with him/her.
- When the $(n+1)^{th}$ person enters $(n \geq 2)$, he/she sits at a new table with probability $\frac{\alpha}{n+\alpha}$ and at table k with probability $\frac{n_k}{n+\alpha}$, where n_k is the number of customers currently sitting at table k.

We continue by letting G_0 describe the base distribution used to sample food dishes from the menu when new tables are selected and denoting $\{X_i : i = 1, 2, \ldots\}$ to be the food (parameters) for each person. The distribution of $\{X_i : i = 1, 2, \ldots\}$ is identical to the marginal distribution of say $\{X_i' : i = 1, 2, \ldots\}$, where $X_i'|G \sim G$ and $G \sim \mathcal{DP}(\alpha, G_0)$.

Customers are grouped into clusters (tables) with the Chinese restaurant process (or DP). Instead of fixing the number of clusters, the DP allows it to grow as more customers (data) come in. Thus a cluster analysis based on DP has an advantage over finite mixture models.

The Chinese restaurant process also implies the more customers (data points) there are at a table (cluster), the more likely it is that new customers (new data points) will join it. That is the so-called “the rich get richer” property of the DP.

1.2 Dirichlet Process Mixture

The DP has been established (Sethuraman 1994 [34], Ferguson 1973) as a potentially poor model for modeling the distribution of data. This had led to the development of the Dirichlet Process Mixture (DPM) (Lo 1984 [21], Escobar 1994 [8]). Both the DP and DPM involve placing distributions on distributions.

The application of the DP as a model for the distribution of data is limited since the DP is discrete with probability one as shown in Figure 1.1. It is not reasonable to model the distributions of responses like blood pressure and BMI as discrete. Instead, the DPM is an infinite DP weighted mixture of parametric densities, which provides a flexible but continuous model for the data. It is a random mixture since the mixing distribution, the DP, is random.

Escobar (1994) [8], Escobar and West (1995) [9] recognized the potential for DP mixture (DPM) of parametric distributions. We discuss this as follows.

1.2.1 The DPM model

Let $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ be the data, and $\theta = \{\theta_1, \theta_2, ..., \theta_n\}$ be a collection of corresponding parameters. The DPM model is expressed below:

$$\begin{aligned} y_i|\theta_i &\overset{\perp}{\sim} F(y_i|\theta_i) \\ \theta_i|G &\overset{i.i.d}{\sim} G \qquad i = 1, \ldots, n \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{1.5}$$

where $F(\cdot|\theta)$ is the parametric distribution selected for the observations, and the parameters, θ_i's, are modeled with the DP. Thus, using Equation (1.3), we obtain

$$F(y_i|G) = \sum_{j=1}^{\infty} P_j F(y_i|\tilde{\theta}_j) \tag{1.6}$$

where P_i's are defined as in Equation (1.3), and $\tilde{\theta}_j \overset{i.i.d}{\sim} G_0$.

Integrating out G, and using exchangeability of $(\theta_1, \ldots, \theta_n)$, we obtain the conditional form

$$\theta_i|\theta_{-i} \sim \frac{1}{\alpha + n - 1}\big[\sum_{j=1, j\neq i}^{n} \delta_{\theta_j}(\theta_i) + \alpha G_0(\theta_i)\big] \tag{1.7}$$

where $\theta_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n\}$. This result also corresponds to Equation (1.4) from the Polya Urn scheme.

With Bayes rule, the full conditional distribution can be written as:

$$\theta_i|\theta_{-i}, \mathbf{y} \sim c\Big[\sum_{j\neq i} f(y_i|\theta_j)\delta_{\theta_j}(\theta_i) + \alpha\Big(\int f(y_i|\theta)dG_0(\theta)\Big)H_i(\theta_i)\Big] \tag{1.8}$$

where H_i is the posterior distribution of θ based on single observation y_i and the prior G_0, and c is the normalizing constant. The pdf of H_i is $h(\theta|y_i) = \frac{f(y_i|\theta)G_0(d\theta)}{\int f(y_i|\theta)G_0(d\theta)}$. Equation (1.8) is used for posterior sampling using the Gibbs sampler for marginalized DP based models.

Escobar and West (1995) [9] examined the problem of univariate density estimation and modality assessment. They used a DPM model to analyze the data involving velocities of distant galaxies diverging from our own (Roeder 1990) [32]. The DPM for the velocity data was a DP mxiture of normal distributions with DP mixing on the mean and precision, $\{\mu, \tau\}$, of the normal distribution.

In their paper, the predictive density $f(y^{new}|\mathbf{y})$ was obtained using the following methods. Firstly, they assigned normal/inverse-gamma (gamma for precision τ_i) distributions to G_0 as $\mu_i|\tau_i \sim N(m, \frac{\lambda}{\tau_i})$ and $\tau_i \sim \Gamma(\frac{s}{2}, \frac{S}{2})$, which made the integral $\int f(y_i|\theta)dG_0(\theta)$ in Equation (1.8) tractable due to conditional conjugacy. Similar to Equation (1.7), the distribution of $\theta^{new}|\boldsymbol{\theta}$ is

$$\theta^{new}|\boldsymbol{\theta} \sim \frac{1}{\alpha+n}\Big[\sum_{j=1}^{n} \delta_{\theta_j}(\theta^{new}) + \alpha G_0(\theta^{new})\Big] \tag{1.9}$$

With $f(y^{new}|\theta^{new}) \sim N(\mu^{new}, \tau^{new})$, we obtain

$$\begin{aligned} f(y^{new}|\boldsymbol{\theta}) &= \int f(y^{new}|\theta^{new})p(\theta^{new}|\boldsymbol{\theta})d\theta^{new} \\ &\sim \frac{1}{\alpha+n}\Big[\sum_{j=1}^{n} N(y^{new}|\mu_j, \tau_j) + \alpha\, T_s(\frac{y_i - m}{\sqrt{(\lambda+1)S/s}})\Big] \end{aligned} \tag{1.10}$$

where T_s is the student-t distribution with degrees of freedom s. Using Equation (1.2), the predictive density is:

$$f(y^{new}|\mathbf{y}) = \int f(y^{new}|\boldsymbol{\theta})dp(\boldsymbol{\theta}|\mathbf{y}) \tag{1.11}$$

Direct evaluation of Equation (1.11) is computationally complicated. Generally, Monte Carlo approximation is used to approximate the density.

1.2.2 Gibbs Sampler for the DPM

In Bayesian non-parametric models with the DP, it is not possible to have a closed-form for the joint posterior distribution. As such, inferences are generally made numerically via Gibbs sampling.

Gibbs sampling is a method for constructing a Markov chain Monte Carlo sample from the joint posterior. This method is extremely useful for a DPM model since we have the full conditional distributions for each parameter in analytical form as seen in Equation (1.8).

In order to obtain posterior samples from $p(\theta|\mathbf{y})$, we implement the Gibbs sampler as follows:

1. We begin with the initial value $\theta^{(0)}$.
2. Draw posterior samples for θ using the Markov Chain Monte Carlo (MCMC) method. At the j^{th} MCMC iteration, for each $i \in \{1, \ldots, n\}$, sample θ_i^j from the conditional distribution $\theta_i|\mathbf{y}, \theta_1^j, \ldots, \theta_{i-1}^j, \theta_{i+1}^{j-1}, \ldots, \theta_n^{j-1}$ using Equation (1.8). Thus, sample each θ_i from its full conditional distribution, making use of the most recent values of θ_{-i}, and updating θ_i with its new value as soon as it has been sampled.
3. Repeat step 2 for $j = 1, 2, \ldots$ till the Markov chain converges.

In the Galaxy example from Escobar and West (1995) [9], Gibbs sampling was used for the posterior inference. We illustrate the sampling procedure for DPM models with this example. The model for the galaxy velocity data is:

$$\begin{aligned} y_i|\theta_i = \{\mu_i, \tau_i\} &\sim N(\mu_i, \frac{1}{\tau_i}), \\ \theta_i|G &\sim G, \\ G &\sim \mathcal{DP}(\alpha, G_0). \end{aligned} \tag{1.12}$$

where the parameter space θ is $\{\mu, \tau\}$. In addition, the base distribution G_0 was chosen to be normal/gamma model due to conjugacy;

$$\mu_i|\tau_i \sim N(m, \frac{\lambda}{\tau_i}), \quad \tau_i \sim \Gamma(\frac{s}{2}, \frac{S}{2}).$$

Let $f_n(\cdot|\mu, \tau)$, $f_t(\cdot|df)$ denote the density functions of $N(\mu, \tau^{-1})$ and $t(df)$ with degrees of freedom df, respectively. The full conditional distribution is:

$$\theta_i|\theta^{(-i)}, \mathbf{y}, \alpha \sim \sum_{j \neq i} q_{ij}\delta_{\theta_j}(\theta_i) + q_{i0}H_i(\theta_i), \tag{1.13}$$

where $q_{ij} \propto f_n(y_i|\mu_j, \tau_j)$, $q_{i0} \propto \alpha\, f_t(\frac{y_i - m}{\sqrt{(\lambda+1)S/s}}|s)$ subject to $\sum_{j \neq i} q_{ij} + q_{i0} = 1$. H_i is normal/gamma distribution with

$$\begin{aligned} \mu_i|\tau_i, \mathbf{y} &\sim N(\frac{\lambda y_i + m}{\lambda + 1}, \frac{\lambda}{(\lambda+1)\tau_i}), \\ \tau_i|\mathbf{y} &\sim \Gamma(\frac{s+1}{2}, \frac{S}{2} + \frac{(y_i - m)^2}{2(\lambda+1)}). \end{aligned}$$

Escobar and West proved that the Markov chain converges to the posterior distribution using results in Tierney (1994) [39]. In addition, the concentration parameter α of the DP is a critical smoothing parameter in the model, which is related to the number of components in the mixture model. It was regarded as a random variable and was assigned with a gamma prior distribution.

1.3 Sampling Schemes

Posterior Bayesian non-parametric models, a variety computational algorithms have been commonly used to solve issues and improve efficiency. For example, Metropolis-Hastings sampling is widely used for sampling parameters for which the full conditional is not analytically tractable. Reversible-jump MCMC has been used for sampling posterior distributions with varying dimensions. We will introduce the two algorithms in this section.

1.3.1 Reparameterization to the DP

Suppose we have $(\theta_1, \ldots, \theta_n)$ where

$$\theta_i | G \overset{i.i.d}{\sim} G, \quad G \sim DP(\alpha, G_0).$$

Then we can sample θ_i's using the Polya Urn Scheme in Equation (1.7). The properties of MCMC samples taken as in Equation (1.7) are not efficient (MacEachern and Muller 1998 [22]). So we must consider improved methods.

Let K be the number of distinct values in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, and let $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_K\}$ to denote the K distinct values in $\boldsymbol{\theta}$. We define a vector of cluster indicators $\mathbf{S} = \{s_i : i = 1, \ldots, n\}$ with each $s_i = j$ if $\theta_i = \phi_j$. Then knowing $\boldsymbol{\theta}$ is equivalent to knowing $(\mathbf{S}, \boldsymbol{\phi})$. MacEachern and Muller [22] realized that they could sample $\boldsymbol{\theta}$ more efficiently according to the Polya Urn Scheme by sampling $(\mathbf{S}, \boldsymbol{\phi})$.

Neal 2000 [24] presents a particular finite Dirichlet mixture that is useful for deriving full conditional distributions $(\mathbf{s}, \boldsymbol{\phi})$. The model considered is:

$$\begin{aligned}
y_i|s_i, \boldsymbol{\phi} &\overset{\perp}{\sim} F(y_i|\phi_{s_i}), \\
s_i|\pi &\overset{\perp}{\sim} Multinomial(\pi_1, \dots, \pi_L), \\
\phi_{s_i} &\sim G_0, \\
(\pi_1, \dots, \pi_L) &\sim Dirichlet(\frac{\alpha}{L}, \dots, \frac{\alpha}{L}). \qquad (1.14)
\end{aligned}$$

Ishwaran and Zarepour (2002) [16] proved that $\sum_{j=1}^{L} \pi_j \delta_{\phi_j}$ converges to $\mathcal{DP}(\alpha, G_0)$ in distribution as $L \to \infty$. Thus model (1.14) can be regarded as an approximation to the DPM model when $L \to \infty$. So we can use this representation to sample for DPM model using parameters $(\mathbf{s}, \boldsymbol{\phi})$ instead of $\boldsymbol{\theta}$. Compared to sampling using Equation (1.7), the Gibbs sampler based on this representation is far more efficient and achieves convergence much faster. We will introduce the full conditional distributions used for Gibbs sampling below.

By integrating out the mixing proportions π, we write the conditional distribution for s_i in the following form

$$P(s_i = s|\mathbf{s}_{-\mathbf{i}}) = \frac{n_{-i,s} + \alpha/L}{n - 1 + \alpha} \qquad (1.15)$$

where $\mathbf{S}_{-\mathbf{i}} = \{s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n\}$ and $n_{-i,s}$ is the number of s_j satisfying $s_j = s$ for all $j \neq i$.

As $L \to \infty$, the conditional distribution above becomes:

$$P(s_i = s|\mathbf{s_{-i}}) = \begin{cases} \frac{n_{-i,s}}{n-1+\alpha}, & \text{if } s \in \mathbf{s_{-i}} \\ \frac{\alpha}{n-1+\alpha}, & \text{if } s \text{ is a new label.} \end{cases} \tag{1.16}$$

Gibbs sampling for $\{s_i : i = 1, \ldots, n\}$ is based on the following conditional probabilities:

$$P(s_i = s|\mathbf{s_{-i}}, y_i, \boldsymbol{\phi}) = \begin{cases} b\frac{n_{-i,s}}{n-1+\alpha}F(y_i|\phi_s) & \text{if } s \in \mathbf{s_{-i}} \\ b\frac{\alpha}{n-1+\alpha}\int F(y_i|\phi)dG_0(\phi) & s \text{ is a new label.} \end{cases} \tag{1.17}$$

where b is the normalizing constant.

The full conditional distribution for $\{\phi_s : s = 1, \ldots, K\}$ is:

$$f(\phi_s|\mathbf{S}, y) \propto \Big(\prod_{i,s_i=s} F(y_i|\phi_s)\Big)dG_0(\phi_s) \tag{1.18}$$

In forthcoming data analyses, we will use this representation of the DPM model for posterior sampling and inferences. The details of the sampler will be introduced in chapter 2 using the model built for the hormone data from Study of Women's Health Across the Nation (SWAN).

1.3.2 Algorithm for Non-conjugate Prior in DPM

In the DPM given in model (1.5), Gibbs sampling is used to sample the cluster membership s_i for each individual using Equation (1.17). When ϕ is assigned with a conjugate prior, this step is simple since the integral $\int F(y_i|\phi)dG_0(\phi)$ can be computed analytically. When the prior is not conjugate, the integral has no analytical form and requires numerical approximation, which can be computationally challenging. In order to resolve the problem, we discuss Algorithm 8 proposed by Neal (2000) [24].

We retain the notation defined in Model (1.5) and Equation (1.17). Below is the algorithm used to sample the cluster membership s_i for each individual i:

- Let K_{-i} be the number of distinct elements of $\mathbf{s_{-i}}$. For simplicity, we let $\mathbf{s_{-i}}$ be labeled with values in $\{1, \ldots, K_{-i}\}$. Then the corresponding parameters are $\{\phi_1, \ldots, \phi_{K_{-i}}\}$. In addition, we need to specify a positive integer h in the procedure. We use $h = 3$ in our analysis.

- Check whether $s_i \in \mathbf{s_{-i}}$.

- If yes, draw h values from $\phi \sim G_0$ independently, and label them as $\{\phi_{K_{-i}+1}, \ldots, \phi_{K_{-i}+h}\}$.

- If no, let $\phi_{K_{-i}+1} = \phi_{s_i}$. Then draw $h - 1$ values from $\phi \sim G_0$ independently, and label them as $\{\phi_{K_{-i}+2}, \ldots, \phi_{K_{-i}+h}\}$.

- Draw a new value for s_i from $\{1, \ldots, K_{-i} + h\}$ using the following probabilities:

$$P(s_i = s|\mathbf{s_{-i}}, y_i, \phi_1, \ldots, \phi_{K_{-i}+h}) = \begin{cases} b\frac{n_{-i,s}}{n-1+\alpha}F(y_i|\phi_s) & \text{if } 1 \leq s \leq K_{-i} \\ b\frac{\alpha/h}{n-1+\alpha}F(y_i|\phi_s) & \text{if } K_{-i}+1 \leq s \leq K_{-i}+h \end{cases} \tag{1.19}$$

- Update ϕ_{s_i} with the ϕ value corresponding to the new s_i.

As $h \to \infty$, this algorithm approaches the behavior of the Polya Urn Scheme described at Equation (1.17), since the h values for ϕ_s drawn from G_0 effectively produce a Monte Carlo approximation to the integral $\int F(y_i|\phi)dG_0(\phi)$.

The equilibrium distribution of the Markov chain defined by this algorithm is exactly correct for any value of h. It does not require h to be large for a Monte Carlo approximation. For example, when $h = 1$, this algorithm resembles the "no gap" algorithm proposed by MacEachern and Muller (1998) [22].

1.3.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm is a Markov chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. It was first used by Metropolis (1953) [23] and extended to more general cases by Hastings (1970) [15]. This algorithm is well-known to the majority of Bayesian statisticians and has been discussed in detail in multiple books and articles, so we limit our introduction to the basic ideas in this section.

Considering the general Bayesian model (1.1), we are interested in the posterior distribution $f(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta}$. The M-H algorithm is particularly useful for posterior sampling in many cases because: (1) the distribution of $\theta|y$ is not required to be recognizable; and (2) the evaluation of normalizing factor $\int f(y|\theta)p(\theta)d\theta$, which is often extremely difficult in practice, is not needed.

The M-H algorithm draws posterior samples from $f(\theta|y)$ by constructing a Markov chain, $\{\theta_j : j = 0, 1, \ldots\}$, that has a stationary distribution $f(\theta|y)$. Let $q(\theta'|\theta)$ be a conditional density that is easily sampled, for example a normal distribution. The Markov chain can be constructed with the following steps:

1. Pick an arbitrary value for θ_0.

2. At each iteration $j(j \geq 0)$, generate a candidate parameter θ' from $q(\theta'|\theta_j)$.

3. Calculate the acceptance probability $\alpha(\theta_j, \theta') = min\{1, \frac{f(y|\theta')p(\theta')q(\theta_j|\theta')}{f(y|\theta_j)p(\theta_j)q(\theta'|\theta_j)}\}$.

4. Let $\theta_{j+1} = \theta'$ with probability $\alpha(\theta_j, \theta')$, and $\theta_{j+1} = \theta_j$ with probability $1 - \alpha(\theta_j, \theta')$.

5. Repeat step 2-4 till convergence.

1.3.4 Reversible Jump MCMC

The reversible-jump MCMC sampler (Green, 1995 [12]) provides a general framework for MCMC simulation in which the dimension of the parameter space can vary between iterates of the chain. The reversible jump sampler can be viewed as an extension of the Metropolis-Hastings algorithm onto more general state spaces.

Reversible-jump MCMC can be applied to Bayesian change-point problems, where the number and location of change-points are unknown. For example, Fan and Brooks (2000) [10] used it to model the shape of prehistoric tombs, where the curvature of the dome changes an unknown number of times. In our analysis in Chapter 5, we will use the method to model the serology scores collected from cows, where both whether and when the cow was infected with Johne's disease is unknown.

We now consider a general setting of the method. Suppose that for observed data y we have a collection of candidate models $\{\mathcal{M}_k : k = 1, \ldots, K\}$. The index k can be considered as an auxiliary model indicator variable. Each model $\mathcal{M}_k$ has an n_k-dimensional vector of unknown parameters, $\theta_k \in \mathbb{R}^{n_k}$, where n_k can be different for each $k = 1, \ldots, K$. Hence the joint posterior distribution is

$$f(k, \theta_k | y) \propto L(y|k, \theta_k) p(\theta_k|k) p(k)$$

where $L(y|k, \theta_k)$ is the likelihood under model k.

The MCMC sampler of the reversible-jump algorithm, which is targeted to have the posterior distribution as its stationary distribution, is constructed over the state space $\Theta = \{(k, \theta_k) : k = 1, \ldots, K; \theta_k \in \mathbb{R}^{n_k}\}$. The dimension of (k, θ_k) can vary over the state space.

The reversible-jump algorithm is regarded as an extension of the M-H algorithm since it also constructs a Markov chain with reversibility and resembles the M-H acceptance probability $\alpha(\theta, \theta') = min\{1, \frac{f(y|\theta')p(\theta')q(\theta|\theta')}{f(y|\theta)p(\theta)q(\theta'|\theta)}\}$. The difference is that there are two kinds of moves for the update of θ in the reversible-jump MCMC. One is a "within-model" move, which fixes model k and only updates θ_k with an appropriate MCMC scheme. The other is the a "between-models" move, which simultaneously updates the model indicator k and the parameters θ_k according to the general reversible proposal and acceptance mechanism.

In practice, the construction of proposal moves between different models is achieved via the concept of "dimension matching". Suppose that we are currently in state (k, θ_k) in model $\mathcal{M}_k$ and wish to propose a move to state $(k', \theta_{k'})$ in model $\mathcal{M}_{k'}$, which is of a higher dimension $(n_{k'} > n_k)$. In order to match dimensions of the two model states, we generate a random vector u of length $n_{k'} - n_k$ from a known density $q_{k \to k'}(u)$. The current state θ_k and u are then mapped to the new state $\theta_{k'} = g_{k \to k'}(\theta_k, u)$ through a one-to-one mapping function $g_{k \to k'} : \mathbb{R}^{n_k} \times \mathbb{R}^{n_{k'} - n_k} \longrightarrow \mathbb{R}^{n_{k'}}$. The acceptance probability of the proposal is then:

$$\alpha(\mathcal{M}_k, \mathcal{M}_{k'}) = min\{1, \frac{f(k', \theta_{k'}|y)q(k' \to k)}{f(k, \theta_k|y)q(k \to k')q_{k \to k'}(u)} \mid \frac{\partial g_{k \to k'}(\theta_k, u)}{\partial(\theta_k, u)} \mid\} \tag{1.20}$$

where $q_{k \to k'}(u)$ is the probability of proposing a move from model $\mathcal{M}_k$ to $\mathcal{M}_{k'}$, and the partial derivative is the determinant of the Jacobian matrix, that is used for the parameter transformation between (θ_k, u) and $\theta_{k'}$. We will use this acceptance probability in Equation (1.20) for our change-point model in Chapter 5.

In the example above, the proposal θ_k is deterministic for the reverse move from $\mathcal{M}_{k'}$ to $\mathcal{M}_k$. More generally, we can relax the condition by allowing longer length of the vector u, say d_u. In this case, non-deterministic reverse moves can be made by generating a random vector u'

of length $d_{u'}$ from $q_{k'\to k}(u')$, so that the dimension matching condition, $n_k + d_u = n_{k'} + d_{u'}$, is satified. Then the one-to-one mapping is $(\theta_{k'}, u') = g_{k\to k'}(\theta_k, u)$. The corresponding acceptance probability becomes:

$$\alpha(\mathcal{M}_k, \mathcal{M}_{k'}) = min\{1, \frac{f(k', \theta_{k'}|y)q(k' \to k)q_{k'\to k}(u')}{f(k, \theta_k|y)q(k \to k')q_{k\to k'}(u)} \mid \frac{\partial g_{k\to k'}(\theta_k, u)}{\partial(\theta_k, u)} \mid\} \tag{1.21}$$

1.4 Trajectory modeling

In longitudinal studies, we are interested in the developmental trajectory of some response, which describes the course of the behavior over time. Generally, longitudinal trends are non-linear.

1.4.1 Sigmoid Function

In longitudinal epidemiological studies, it is useful to characterize an individual's reaction to some stimulus like disease infection in order to provide early detection of disease. Considering some measure of symptoms which indicates the reaction, it intuitively should have a "S"-shape curve: the reaction to the infection should be mild at the initial stage and get stronger and stronger with the progressing of the disease. After some time, the measure of the reaction gradually levels at some equilibrium level. Of course, with some diseases, the infection marker may just increase without bound until death.

A sigmoid function, whose shape is like "S", is naturally suitable for fitting such a developmental trend. Figure 1.2 shows how a four-parameter sigmoid function looks. In our data analysis in Chapter 4, we use it to model the progression of serology scores collected from cows.

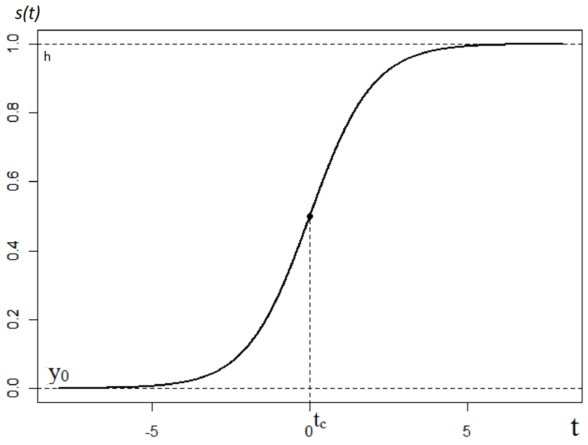


Figure 1.2: Four-parameter sigmoid function with $\{t_c = 0, y_0 = 0, h = 1, r = 1\}$

This four-parameter sigmoid function is mathematically defined as below:

$$s(t|t_c, y_0, h, r) = y_0 + \frac{h}{1 + e^{-r(t-t_c)}} \tag{1.22}$$

As the name implies, the function $s(t)$ has four parameters $\{t_c, y_0, h, r\}$, which determine the location and shape of the sigmoid curve.

The function in Figure 1.2 has the four parameter values $\{t_c = 0, y_0 = 0, h = 1, r = 1\}$. Among them, $\{t_c, y_0\}$ indicate the location of the curve. y_0 is the lower bound of $s(t)$ and t_c is the x-axis coordinate value of the half-height point. Note $s(t)$ is symmetric about the half-height point at coordinate $(t_c, y_0 + \frac{h}{2})$. $\{h, r\}$ are the shape parameters. As shown in Figure 5.2, h is the limiting range of $s(t)$ and r is related to the changing rate of the curve. When $r > 0$, the curve is increasing as shown in the figure. When $r < 0$, the curve is monotonically decreasing. In addition, the curve reaches its maximum changing rate at the half-height point, where $\frac{d\,s(t)}{dt}|_{t=t_c} = 4/r$.

It might be worth mentioning that the most general (flexible) sigmoid function has five parameters, and the one extra degree of freedom allows the curve to be asymmetric about the half-height point. The mathematical form of the function is

$$s(t|t_c, y_0, h, r_1, r_2) = y_0 + \frac{h}{1 + w(t)e^{-r_1(t-t_c)} + (1 - w(t))e^{-r_2(t-t_c)}}$$

where $w(t) = \frac{1}{1+e^{-v(t-t_c)}}$ and $v = \frac{2r_1r_2}{|r_1+r_2|}$. We did not choose it because: (1) model simplicity; (2) the data did not show significant asymmetric pattern.

1.4.2 Basis Functions

In mathematics, a basis function is an element of a particular basis for a function space. Every continuous function in the function space can be represented as a linear combination of basis functions, just as every vector in a vector space can be represented as a linear combination of basis vectors. In a longitudinal data analysis, it allows flexibility in the trajectory shape to fit the developmental trend with basis functions.

The most commonly used basis function is the polynomial basis $\{t^i : i = 0, 1, \ldots\}$ on $t \in \mathbb{R}$. In statistics, they have been used for polynomial regression, in which the relationship between the independent variable x and the dependent variable y is modeled. Polynomial regression fits a nonlinear relationship between the time t and the corresponding conditional mean of response y, denoted $E(y|t)$, with a linear combination of the basis functions $E(y|t) = \beta_0 + \beta_1 t + \ldots + \beta_p t^p$.

It is important to choose the function space, which is determined by the order p for the polynomials. The higher the value p, the more closely the trajectory will fit the data. When $p \to \infty$, $E(y|t)$ can represent any trajectory. However, it is not appropriate to use large number of basis functions in a data analysis due to over-fitting. So the choice of p should balance model fitting and model complexity. In our analysis, we chose basis functions up to cubic terms ($p = 3$).

In addition, simple polynomial basis is not the only choice for trajectory modeling. For example, Fourier series are commonly used to model the periodic signals. The Daubechies wavelet is suitable for modeling the data with abrupt changes or oscillations in a short time. Below we introduce two sets of basis functions which were used in our analysis.

1.4.3 Legendre Polynomials

In mathematics, Legendre functions are solutions to Legendre's differential equation,

$$\frac{d}{dt}\left[(1-t^2)\frac{d}{dt}x_k(t)\right] + k(k+1)x_k(t) = 0.$$

The solutions for $k = 0, 1, 2, \ldots$ form a polynomial sequence called the Legendre Polynomials. The functions are calculated with Bonnet's recursion formula:

$$x_0(t) = 1; \quad x_1(t) = t$$

$$(k+1)x_{k+1}(t) = (2k+1)t \cdot x_k(t) - k \cdot x_{k-1}(t), \quad k \geq 1$$

One important property of the Legendre polynomials is that they are orthogonal on the interval $(-1, 1)$:

$$\int_{-1}^{1} x_k(t)x_{k'}(t)\,dt = \frac{2}{2k+1}\delta_{kk'}$$

For data analysis, we prefer to using standardized basis functions by multiplying a constant $\sqrt{\frac{2k+1}{2}}$ times each $x_k(t)$. Figure 1.3 shows standardized Legendre basis functions up to degree 5.

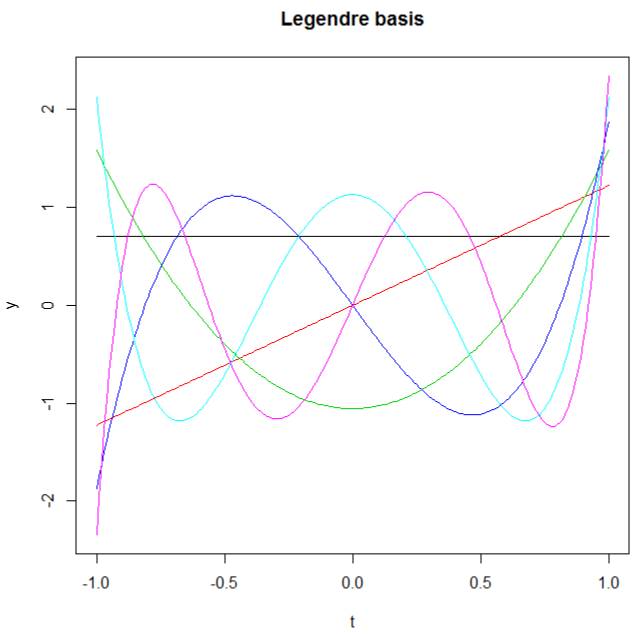


Figure 1.3: The first 6 Legendre basis functions

1.4.4 Orthogonal Polynomials

In a balanced longitudinal study, all observations are collected at a sequence of fixed time points. In such a case, the time t should be regarded as discrete instead of continuous, and we need to use a set of basis functions that are defined on discrete time points.

One choice is the set of orthogonal polynomials discussed by William J. Kennedy and James E. Gentle (1980) [18]. Let t be a vector of fixed time points with length n_t. Then orthogonal polynomials are calculated using the following recurrence relation:

$$\begin{aligned} x_{-1}(t) = 0; \quad x_0(t) = 1 \\ x_{k+1}(t) = (t - u_{k+1})x_k(t) - v_k x_{k-1}(t), \quad k \geq 0 \end{aligned}$$

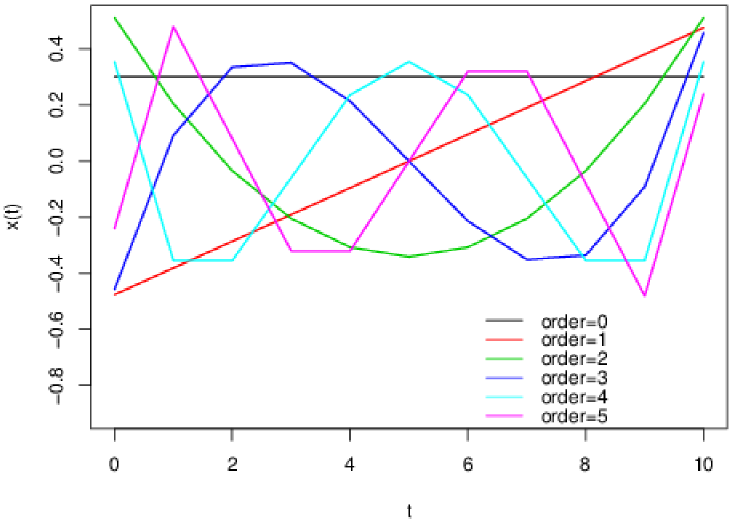


Figure 1.4: The first 6 orthogonal polynomial basis functions, which are defined on $\{t = 0, 1, \ldots, 10\}$

where,

$$u_{j+1} = \begin{cases} 0, & j = 0 \\ \frac{\sum_{i=1}^{n_t} t_i \left(x_j(t_i)\right)^2}{\sum_{i=1}^{n_t} \left(x_{j-1}(t_i)\right)^2}, & j > 0 \end{cases}, \quad v_j = \begin{cases} 0, & j = 0 \\ \frac{\sum_{i=1}^{n_t} \left(x_j(t_i)\right)^2}{\sum_{i=1}^{n_t} \left(x_{j-1}(t_i)\right)^2}, & j > 0 \end{cases}$$

Orthogonal polynomial basis functions also have the orthogonality as indicated in its name:

$$\begin{cases} \sum_{i=1}^{n_t} \left(x_k(t_i)\right)^2 = c_k \\ \sum_{i=1}^{n_t} x_k(t_i) x_{k'}(t_i) = 0, \quad k \neq k' \end{cases}$$

where $c_k \neq 0$ is a constant. We used the standardized orthogonal polynomial basis functions in our model, and Figure 1.4 shows the orthogonal polynomial basis functions up to order 5.

Chapter 2

Clustering Longitudinal Processes with Dirichlet Process Mixtures

2.1 Introduction

A longitudinal study refers to an investigation where participant outcomes and possibly treatment or exposure variables are collected at multiple follow-up times. Such studies play a prominent role in health, social, and behavioral sciences as well as in the biological sciences, economics, marketing and finance. In a longitudinal study, an interesting question is to identify trending groups (those with outcomes that start high and stay high, those that start low and stay low, those that start low and increase to high etc). For example, in marketing analytics, we would like to cluster customers based on their behavior trends, like spending habits in the past several months/years. Different business decisions like promotion strategies could be implemented targeting each "trend group". In epidemiological research, it is also interesting to group patients based on the profile of a series of biological responses like hormone levels for example. We can then look at individual factors that are associated with “trend group” membership.

We are motivated by data from the Study of Women’s Health Across the Nation (SWAN), which involves the collection of hormone data on women through the menopausal transition. Menopause is a universal female phenomenon defined by a specific event, the final menstrual period (FMP). Menopausal transition is a series of stages of variable length from pre-, early peri- and late peri- to post-menopause, each defined by changes in menstrual and hormonal patterns. The SWAN study is a multi-site longitudinal epidemiologic study designed to examine the health of women during these years. In the dataset we consider here, 11 years of E2 (Estradiol) and FSH (Follicle-stimulating hormone) values were collected annually from 928 women who experienced the menopausal transition. We are interested in clustering the women based on their hormone profiles through menopause. This particular data set consists of a rather small subset of the entire SWAN data, which was abstracted for the purpose of studying the relationship between hormone characteristics and the incidence of urinary incontinence (UI). Here we restrict ourselves to the hormone profile data only. In Chapter 4, we jointly model profiles and UI incidence.

In this data analysis, we use a Bayesian semi-parametric model with a Dirichlet Process Mixture (DPM) model involving curve shapes combined with mixed effects as a clustering mechanism to group women with similar hormone profiles. In Section 2, we review the existing clustering algorithms, point out the challenges we face in this analysis, and propose our solutions and strategies used to overcome the challenges. Then we specify our clustering model in detail in Section 3. In Section 4, we introduce the algorithms and techniques used to draw posterior samples. In Section 5, we use a simulation study to establish inference validity and compare our method with methods for existing clustering models. We illustrate our method using longitudinal hormone data from SWAN in Section 6.

2.2 Background

Our goal is to cluster women into groups based on their longitudinal hormone profiles. However, we face several challenges. In this section, we will discuss the existential difficulties.

We divide the content of the section into three parts: (i) introduction to the conventional clustering algorithms and their limitations in handling longitudinal data; (ii) discussion of over-fitting issues and methods for avoiding them; and (iii) a review of some existing clustering methods for longitudinal data, and our proposal for improving the clustering performance.

2.2.1 Conventional Clustering methods

Cluster analysis involves the task of grouping a set of objects in such a way that those in the same group (cluster) are more similar to each other than to those in other groups (clusters). This general statistical technique is used in many fields, including data mining, machine learning, pattern recognition, bioinformatics etc., as well as across disciplines. Some typical clustering methods include (i) the use of hierarchical models, where a hierarchy of clusters is built based on a measure of dissimilarity, (ii) centroid models (eg: K-means), where observations are partitioned into K clusters in which each observation belongs to the cluster with the "shortest" centroid distance to the center, and (iii) mixture models (eg. Gaussian mixtures), where allocation to clusters/mixtures is based on the posterior probability of cluster membership, which could be one of a pre-specified number of K clusters, etc.

However, these general clustering methods do not apply here, because the data structure is different from what we conventionally have for unsupervised learning. In a dataset for conventional cluster analysis, multiple features/covariates were observed once and only once for each individual. Then the individuals could be grouped by comparing their values of these features using well-known algorithms like K-means. Let us assume that we have vectors of data measured from r features on n individuals, say $y_i = (y_{i1}, y_{i2}, ..., y_{ir})'$ for each $i \in \{1, ..., n\}$. The observations y_{ij} and $y_{i'j}$ are comparable in the sense that they are observed values on the j^{th} feature for any two individuals i and i'. We anticipate that there may be, say K, groups of individuals who share common features. If we let s_i be a latent variable indicating the "cluster membership" of individual i, eg: $s_i = k$ if individual i belongs to cluster k, $k = 1, ..., K$, then we anticipate that $E(y_i \mid s_i = k) = (\mu_{k1}, \ldots, \mu_{kr})'$ for all $k = 1, \ldots, K$. The general goal is to decide which individuals belong to which cluster based on the observed data $y = \{y_1, ..., y_n\}$. This formulation presumes that there are no missing data, that observations are comparable across j, and that there are exactly the same number, r, of observations on each individual.

Longitudinal data analysis generally involves a scalar response that was measured repeatedly in time for each individual. Longitudinal data for individual i thus involves outcomes $y_i = (y_{i1}, y_{i2}, \ldots, y_{ir_i})'$ corresponding to times $t_i = (t_{i1}, t_{i2}, \ldots, t_{ir_i})'$. The observation vectors from different individuals are generally not comparable. Firstly, r_i is not necessarily equal to $r_{i'}$, which means the individuals have unequal numbers of observations. Secondly, the observations y_{ij} and $y_{i'j}$ may not be comparable since t_{ij} is generally not equal to $t_{ij'}$ for studies with an unbalanced design. Thirdly, missing data are very common in longitudinal studies. All these differences in data structures lead to inappropriate of application of algorithms like K-means to longitudinal data directly.

Moreover, we are interested in trajectory/trend patterns in time, thus the conventional methods mentioned above are far from optimal for clustering individuals by comparing the observation values. Functional data analysis is a form of longitudinal analysis that is used to model such trends. We base our cluster analysis on them.

Define y_i and t_i as above. Our goal is to provide a model that allows for a variety of shapes in mean responses. Let X_i denote an $r_i \times (p+1)$ design matrix of basis functions up to degree

p and let $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$ be a vector of regression coefficients corresponding to these basis functions for individual i. The mean response is modeled as a linear combination of basis functions, namely $E(y_i \mid X_i, \beta_i) = X_i\beta_i$, which allows considerable flexibility in terms of profile shapes.

Instead of clustering individuals by comparing the observed y_i's, cluster analysis will be based on regression coefficient vectors, β_i. In order to accomplish this, we consider a discrete hierarchical model for β_i where the β_i's are allowed to cluster. An obvious and clear advantage is that the study design does not have to be balanced, and missing data are also allowed.

2.2.2 Over-fitting in longitudinal analysis

Over-fitting is an issue requiring consideration for cluster analysis based on longitudinal profiles. Since clustering is based on β's, we have to estimate each individual coefficient vector β_i. That would be problematic when individual data are sparse, since estimation of β_i was only based on the single vector y_i. Over-fitting could result in fitted curves with peculiar shapes for these individuals, which would not demonstrate the real trends of these individual's responses, resulting in poor statistical inferences.

Moreover, the over-fitting issue may not be resolved by increasing the sample size in such a case. For example, considering a longitudinal data analysis for customers' historical spending behaviors, there is always a portion of customers who do not make purchases very often and thus we have a limited number of observations from them. Therefore, the over-fitting issue persists no matter how many more customers are added to the data. In this example, it would be absurd to remove customers because they did not contribute enough data. We will develop a method that can handle the over-fitting issue.

Mixture models have been used to classify individuals based on longitudinal data and while attentioning to minimize over-fitting. In this case, individual trajectories are assumed to be the same for subjects in the same mixture component, thus allowing one to borrow information from similar subjects. For example, Group-Based Trajectory Modeling (GBTM) by Nagin (1999, 2005) and Growth Mixture Modeling (GMM) by Muthen (2001) have been widely used for analyzing longitudinal outcome data in psychology, medicine and criminology. Both methods apply finite mixture models for clustering. Using these methods, individuals in the same cluster share the same mean trajectory. If we let $E(y_i \mid X_i, \beta_i) = X_i\beta_i$, then $\beta_i = \beta_j$ for any two individuals i and j that are clustered. So the observations for all the individuals in the cluster are used to estimate the cluster mean trajectory, which alleviates the over-fitting issue.

However, there are significant disadvantages in using these methods. Firstly, the methods assume independence between any two observations. This is problematic in many scenarios because the observations from the same individual could be highly correlated compared with observations from different individuals. Secondly, pre-specifying the number of clusters, K, is another criticism of the methods, since in many cases we have no knowledge about how many clusters there might be, especially in an exploratory analysis. The specification of K has been discussed extensively (McLachlan and Peel 2004, Muthen 2004, Nagin 2005, Nylund et al. 2007). AIC and BIC criteria have been used to determine K.

In order to overcome these deficiencies, we employ the Dirichlet Process (DP) which involves an infinite number of random mixtures that can be used to cluster individuals without having to specify K. For recent examples, Medvedovic and Sivaganesan (2002), Shahbaba and Johnson (2012) used DP models to cluster gene expression profiles. In the next section, we consider a DPM model for our analysis.

2.2.3 Dirichlet Process Mixture (DPM) for longitudinal data analysis

The DP is a random probability measure (RPM), meaning that it is unknown and uncertain and that its realizations are probability measures. It was formally defined by Ferguson (1973), and further characterized by Sethuraman (1994). Let G_0 be a non-null finite measure on a measurable space $(\Omega, \mathcal{F})$. A random probability measure, G, is defined as a Dirichlet process on $(\Omega, \mathcal{F})$ if for every finite $k = 1, 2, ...$, and measurable partition $(A_1, ..., A_k)$ of Ω, the marginal distribution of $(G(A_1), ..., G(A_k))$ is Dirichlet distributed, $\mathcal{D}(\alpha G_0(A_1), ..., \alpha G_0(A_k))$. We write

$$G \sim \mathcal{DP}(\alpha, G_0)$$

The base distribution G_0 is selected to be continuous, but the DP, G, is almost surely discrete, which can be seen using the stick-breaking representation by Sethuraman (1994). There is always a non-zero probability of two or more samples from G being tied. This property of the DP allows for clustering. If we model the distribution of the β_i's with a DP, individuals who share the same β value are in the same cluster. So in this way we can group the n individuals into a random number of K clusters ($K \leq n$).

The DP model has been used for clustering longitudinal data. Kleinman and Ibrahim (1998) [19] used a DP distribution for the distribution of unknown random effects, but without incorporating trajectories. Ray and Mallick (2004) [40] developed a nonparametric Bayesian wavelet based model to cluster functional data using the DP. Bigelow and Dunson (2005, 2006) [2] [3] developed a semiparametric Bayesian adaptive spline model to cluster pregnant women based on their reproductive hormone trajectories. In these models, the DP was assigned to trajectory parameters β, since the DP is almost surely discrete and can be used for clustering.

Their clustering models could be summarized as generalized versions of:

$$\begin{aligned} y_i|\beta_i, \tau_e &\sim Normal(X_i\beta_i, \tau_e^{-1}I_{r_i}), \\ \beta_i|G &\overset{iid}{\sim} G, \\ G &\sim DP(\alpha, G_0), \\ G_0 &= N(u_\beta, \Xi_\beta^{-1}), \end{aligned} \tag{2.1}$$

where X_i is the design matrix of basis functions at $t_i = (t_{i1}, t_{i2}, \ldots, t_{ir_i})'$. τ_e is the unkown precision and (u_β, Ξ_β) are considered known and fixed. Model (2.1) is actually a Dirichlet Process Mixture (DPM) model for the data, since the normal distributions are mixed using the DP on the distribution of the β_i's.

Model (2.1) is a simplified version of their models, but emphasizes how clustering will be accomplished using the DP. The actual models proposed in these papers are more complicated and sophisticated. Specifically, Wang, Ray and Mallick (2004) [40] proposed a wavelet based nonparametric model, which assigned a DP to the distribution of the β and τ_e jointly rather than simply assigning a parametric prior to τ_e. Bigelow and Dunson (2005) [2] used a semi-parametric model with an adaptive spline basis, which allows extra flexibility in the trajectory shape by considering varying numbers and locations of knots. However, model (2.1) points out one common feature of the cited models, which is that they all modeled the random effects $\boldsymbol{\beta} = \{\beta_i : i = 1, \ldots, n\}$ using a DP.

The most significant advantage of DPM models is that the number of clusters, say K, is random and model-based. It fits one of our purposes in exploratory data analysis, in which we generally have no idea how many clusters exist in truth.

The method has been used to cluster individuals based on longitudinal data. However, the model performance may not always be ideal, since it may tend to produce too many clusters. Here we cite a simulation result presented in Bigelow and Dunson (2005) [2] as an example. In the paper, Figure 1 in their paper shows the data used for their simulation, which were generated based on four true cluster mean trajectories (solid lines), and Figure 3 shows their statistical inference. Theoretically, we should expect 4 clusters since the true number of clusters is 4. In fact, 10 clusters were identified by post-processing the MCMC output of their model. Figure 1 managed to combine the 10 cluster mean curves (dash lines) manually corresponding to the four true trajectory shapes (solid lines). But this construction is generally not straightforward. It could be done in this study because the simulated true cluster trajectories are very different from each other. If we consider a dataset with unknown cluster information, the task is more difficult, especially for data in which trajectories from different clusters are not as distinct as those in Figure 3.

Figure 4 in Bigelow and Dunson (2005) [2] illustrates the difficulty of the task when the model is applied to real data. The plot shows how the women are clustered based on data from the North Carolina Early Pregnancy Study (EPS)(Wilcox et al., 1988), where data consist of daily progesterone measurements in women who are trying to become pregnant. The number of “true” clusters was unknown. Eventually, eight clusters were inferred using their method. But it is difficult to tell whether the estimated cluster mean trajectories were different in trend and it is difficult to combine them manually.

This questionable model performance is evidently caused by the individual variation within cluster in trajectory shapes, which is not accounted for in model (2.1). It is assumed that all the individuals in a cluster share precisely the same mean trajectory, which may not be sensible for longitudinal data. From our perspective, responses of two individuals in the same cluster, while sharing an overall trend, should be allowed to have their own variability about that trend. In other words, individual variation inside each cluster should be allowed whereby two individuals' trajectories from the same cluster should be more or less different from each other.

We propose to add trajectory based mixed effects in the model to account for the individual variation in trajectory shape within clusters. The response y is modeled with $N(X_i\beta_i + X_ib_i, \tau_e^{-1}I_{r_i})$. (Note β_i is the trajectory coefficient vector of the cluster to which individual i belongs. There is no identifiability issue between β_i and b_i since β_i is a parameter vector on the cluster level and b_i is on the individual level.) We believe this model will generate better clustering results because subjects with a similar overall trajectory will be clustered together in the model, and since perturbations within clusters with overall trajectory shape will be taken up by X_ib_i rather than adding new clusters.

We eventually chose to use the marginalized form of the mixed effect model with b_i integrated out. Since b_i is a vector on the individual level, and since inferences about it will depend greatly on the single observation, y_i, we are concerned that Markov chains used for fitting unmarginalized models may get stuck and thus have poor performance. Our experiences in fitting both forms of the model are consistent with this belief. In the marginalized form of the model, b_i is thus not directly estimated. That does not affect our analysis since we are not interested in inferences for b_i. Our question of interest is how the individuals are clustered, which is inferred from β. Therefore, the marginalized form of the model is identical to the non-marginalized form, and evidently has better MCMC properties.

2.3 The Model

In this section, we specify our model in detail, including choice of basis functions. Then we discuss specification of prior distributions for the corresponding model parameters.

2.3.1 Model specification

SWAN data consist of longitudinal hormone observations from n women. Our goal is to cluster them based on the trends in their hormone trajectories over time. The data consist of vectors $y_i = (y_{i1}, y_{i2}, \ldots, y_{ir_i})'$ corresponding to times $t_i = (t_{i1}, t_{i2}, \ldots, t_{ir_i})'$ for individuals, $i = 1, 2, \ldots, n$. We use orthogonal polynomial basis functions to fit the trajectory. Denote X_i as the $r_i \times (p+1)$ design matrix containing the basis functions up to p^{th} order for woman i. We use the following model for the data:

$$\begin{aligned} y_i|\beta_i &\sim N_{r_i}(X_i\beta_i, \tau_e^{-1}W_i) \\ \beta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= N_{p+1}(u_\beta, \Xi_\beta^{-1}) \end{aligned} \tag{2.2}$$

where $W_i = I_{r_i} + X_i\Gamma^{-1}X_i^T$ and

$$\Gamma = \begin{pmatrix} \gamma_0 & 0 & \cdots & 0 \\ 0 & \gamma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_p \end{pmatrix}.$$

This model for the response y is the integrated form of the mixed effects model as below:

$$\begin{aligned} y_i|\beta_i, b_i &\sim N(X_i\beta_i + X_ib_i, \tau_e^{-1}I_i) \\ b_i &\sim N(0, (\tau_e\Gamma)^{-1}). \end{aligned} \tag{2.3}$$

In the model, β_i is modeled with a DP, which allows for clustering of individuals. The vector of the mean trajectory corresponding to times t_i is $X(t_i)\beta_i$. If $\beta_i = \beta_{i'}$, any two individuals i and i' are in the same cluster.

The vector b_i of mixed-effect coefficients is integrated out and posterior sampling is based on the corresponding marginal likelihood. It is worth mentioning that the covariance structure for b_i is $(\tau_e\Gamma)^{-1}$. Conventionally, we model the mixed effect with $b_{ij} \sim N(0, \tau_{bj})$ for $j = 0, 1, \ldots, p$. Here we reparameterized its precision for calculation simplicity. We let $\gamma_j = \frac{\tau_b j}{\tau_e}$. γ_j can be interpreted as the precision ratio between mixed effect coefficient b_{ij} and white noise ϵ_i for all $j = 0, \ldots, p$. As previously mentioned, the mixed-effect $X(t_i)b_i$ is used to account for the variation of individual trajectories within cluster.

With the model above, we can write the likelihood. Let Θ denote the collection of parameters in the model. The likelihood is

$$\begin{aligned} L(\Theta) &= \prod_{i=1}^{n} f(y_i|\Theta_i) \\ &\propto \prod_{i=1}^{n} \tau_e^{\frac{r_i}{2}} |W_i|^{-\frac{1}{2}} e^{-\frac{\tau_e}{2}(y_i - X_i\beta_i)^T W_i^{-1}(y_i - X_i\beta_i)} \end{aligned} \tag{2.4}$$

2.3.2 The Dirichlet Process

As discussed in Section 1.1.2, Sethuramann (1994) constructed the stick-breaking process. Let $\theta_i \overset{iid}{\sim} G_0$ and $w_i \overset{iid}{\sim} Beta(1, \alpha)$, where $\theta_i \perp w_i$, for all $i = 1, 2, \ldots$; and define $P_i = w_i \prod_{j=1}^{i-1}(1 - w_j)$. Sethuramann defined a random probability measure $G(\cdot) = \sum_{i=1}^{\infty} P_i \, \delta_{\theta_i}(\cdot)$, and established that $G \sim \mathcal{DP}(\alpha, G_0)$. This representation makes clear that the DP is discrete with probability one.

Consider model (2.2). We write the density function below using the representation of the DP:

$$f(y_i|G) = \sum_{j=1}^{\infty} P_j f_N(y_i|X_i\beta_j, \tau_e^{-1}W_i)$$

where $\beta_j \overset{iid}{\sim} G_0$ and $f_N(\cdot|\mu, \Sigma)$ is the density function for a multivariate normal distribution with mean μ and covariance Σ. This expression makes clear the nature of the DPM.

Now consider the joint marginal distribution of $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_n\}$, which is complex. However, it is well-known that the marginal full conditional distribution for β_i given all other β values, $\boldsymbol{\beta}_{-i} = \{\beta_j : j = 1, \ldots, i-1, i+1, \ldots, n\}$, is characterized by the Polya Urn scheme (section 1.1.3). The conditional distribution has the following form using Equation (1.7):

$$\beta_i|\boldsymbol{\beta}_{-i} \sim c\Big(\sum_{j=1, j\neq i}^{n} \delta_{\beta_j}(\beta_i) + \alpha G_0(\beta_i)\Big)$$

where c is a normalizing constant.

It is thus possible to sample $\boldsymbol{\beta}$ using Gibbs sampling. This conditional distribution makes it possible to derive the full conditional distributions:

$$\beta_i|\boldsymbol{\beta}_{-i}, \mathbf{y} \sim c\Big[\sum_{j\neq i} f_N(y_i|\beta_j)\delta_{\beta_i}(\beta_j) + \alpha\Big(\int f(y_i|\beta)dG_0(\beta)\Big)H_i(\beta_i)\Big] \tag{2.5}$$

where H_i is the posterior distribution of β based on the single observation y_i and the prior G_0. The pdf of H_i is $h(\beta|y_i) = \frac{f_N(y_i|\beta)G_0(d\beta)}{\int f_N(y_i|\beta)G_0(d\beta)}$.

Those results will be important for obtaining Markov chain Monte Carlo (MCMC) approximation to the joint posterior distribution, which is discussed in Section 2.4.

We observe here that sampling from Equation (2.5) will be easy if the integral $\int f(y_i|\beta)dG_0(\beta)$ can be calculated analytically, and if the distribution H_i is easy to sample. Even if H_i is not a known distribution, it is possible to sample from it using, for example, the Metropolis-Hastings algorithm. However, a key impediment to sampling from Equation (2.5) has been the potential tractability of the integral $\int f(y_i|\beta)dG_0(\beta)$. A number of approaches have been considered, including Bush and MacEachern (1996) [6], MacEachern and Muller (1998) [22], Neal (2000) [24], Jain and Neal (2002) [17]. Neal (2000) summarized methods up to that point in time and also provided novel methods, in particular his Algorithm 8, which were currently used by many. In this chapter, we use a conditionally conjugate prior G_0 and the integral can be computed analytically. We discussed Neal's Algorithm 8 in Chapter 1, however we do not use it in this chapter.

Neal (2000) and others before him also realized that the convergence of the MCMC algorithm using Equation (2.5) could be slow, and proposed his Algorithm 2 to improve efficiency. The problem is that there are often groups of individuals who are associated with the same β value with high probability, due to discreteness of the DP. A change in the β value for such a group would occur only rarely, since the algorithm can not change the β value for more than one individual simultaneously. Accomplishing such a change requires passage through a low-probability intermediate state in which individuals in the group do not all have the same β value.

The main idea in improving the original algorithm was to re-sample $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_n\}$ at each iteration of a Monte Carlo sampler. It was noted that knowing $\boldsymbol{\beta}$ is equivalent to knowing $\mathbf{S} = \{s_1, \ldots, s_n\}$ and $\boldsymbol{\beta}^\star = \{\beta_1^\star, \ldots, \beta_K^\star\}$, where $\boldsymbol{\beta}^\star$ constitutes the distinct β values ($K \leq n$), and where s_i is the label identifying which value in $\boldsymbol{\beta}^\star$ corresponds to individual i. For example, let $n = 5$ and $K = 3$. If we have the label values $\mathbf{S} = \{1, 2, 2, 1, 3\}$, and the distinctive β values $\boldsymbol{\beta}^\star = \{1.2, 4.1, 5.0\}$ with each corresponding to the labels $\{1, 2, 3\}$ respectively, then we know $\boldsymbol{\beta} = \{1.2, 4.1, 4.1, 1.2, 5.0\}$.

The solution to the problem is to use the CRP representation of the DP, which has been discussed in Chapter 1. A Dirichlet distribution mixture model of order L has been established to be asymptotic to the $DP(\alpha, G_0)$ as $L \to \infty$. The following model is used as in Chapter 1 to obtain a nice form for sampling $(\int, \boldsymbol{\phi})$.

$$\begin{aligned} s_i|\mathbf{p} &\sim Multinomial(p_1, \ldots, p_L) \\ p_1, \ldots, p_L &\sim Dirichlet(\frac{\alpha}{L}, \ldots, \frac{\alpha}{L}) \\ \beta_s^\star &\sim G_0 \\ G_0 &= Normal(u_{\beta^\star}, \Xi_{\beta^\star}^{-1}) \end{aligned} \tag{2.6}$$

By integrating over the proportions $(p_1, \ldots, p_L)$, and letting $L \to \infty$, we obtain the conditional distribution for s_i in the following form:

$$P(s_i = s|\mathbf{S}_{-i}) = \begin{cases} \frac{n_{s,-i}}{n-1+\alpha}, & \text{if } s \in \mathbf{S}_{-i} \\ \\ \frac{\alpha}{n-1+\alpha}, & \text{otherwise} \end{cases} \tag{2.7}$$

where $\mathbf{S}_{-i} = \{s_j : j = 1, \ldots, i-1, i+1, \ldots, n\}$, and $n_{s,-i}$ is the number of s_j which satisfies $s_j = s$ for all $s_j \in \mathbf{S}_{-i}$.

In the DPM model, we have the full conditional distribution for the cluster label s_i for Gibbs sampling as below:

$$P(s_i = s|\mathbf{S}_{-i}, y_i, \boldsymbol{\beta}^\star) \propto \begin{cases} n_{s,-i}\, f(y_i|\beta_s^\star) & \text{if } s \in \mathbf{S}_{-i} \\ \alpha \int f(y_i|\beta^\star)\, dG_0(\beta^\star) & \text{if } s \notin \mathbf{S}_{-i} \end{cases} \tag{2.8}$$

With prior $\beta_s^\star \sim G_0$ in Equation (2.6), the full conditional distribution for $\beta_s^\star$ is obtained using:

$$f(\beta_s^\star|\mathbf{S},\mathbf{y}) \propto \Big(\prod_{i:s_i=s} f_N(y_i|\beta_s^\star)\Big)\, dG_0(\beta_s^\star) \tag{2.9}$$

In this subsection, we derived the full conditional distributions used in the Gibbs sampler for the parameters involved in the DP. They are based on Algorithm 2 in Neal (2000). In section 2.4, we will introduce the detailed posterior sampling procedure.

2.3.3 Choice of basis function

An important aspect of the model is the choice of basis functions, $X_i(t)$, which are used to model a potentially non-linear relationship between response y and time t. It is important to choose an appropriate basis and its function space, since they could affect the MCMC convergence significantly.

As discussed in Chapter 1, the choice of basis functions is also related to the study design. The SWAN study, for example, has a balanced design and the hormone observations were collected annually from all women. In the data analysis, we decided to use orthogonal polynomial basis functions up to cubic terms.

We chose to use orthogonal polynomials for balanced data and Legendre polynomials for unbalanced data in all data analyses in the chapter. Both types of the basis functions are orthogonal. We prefer such basis functions due to simplicity. Recall that the non-marginalized model (2.3) specified $b_i \sim N(0, \tau_e^{-1}\Gamma^{-1})$. Due to orthogonality of the basis functions, we have specified the covariance matrix $(\tau_e\Gamma)^{-1}$ to be diagonal. Otherwise, we would require more parameters to model the covariance structure. In addition, having orthogonal basis functions takes noise out of the estimation processes leading to more efficient MCMC numerical approximations.

2.3.4 Prior Specification

We complete the specification of our model by assigning priors to all parameters. Let $\boldsymbol{\gamma} = \{\gamma_j : j = 0, 1, \ldots, p\}$ and $\mathbf{S} = \{s_i : i = 1, \ldots, n\}$. The parameter space is $\Theta = \{\mathbf{S}, \boldsymbol{\beta}, \tau_e, \boldsymbol{\gamma}, \alpha\}$. Note we use $\boldsymbol{\beta} = \{\beta_1^\star, \ldots, \beta_K^\star\}$ to denote the collection of distinct β_i's instead of $\boldsymbol{\beta}^\star$ in order to simplify notation.

Both $\mathbf{S}$ and $\boldsymbol{\beta}$ are latent variables associated with the DP. We only need to assign priors to τ_e, γ and α. The gamma prior has been used for them since they are all positive,

$$\begin{aligned} \tau_e &\sim \Gamma(\frac{a_e}{2}, \frac{b_e}{2}) \\ \gamma_j &\sim \Gamma(\frac{a_\gamma}{2}, \frac{b_\gamma}{2}),\ j = 0, 1, \ldots, p \\ \alpha &\sim \Gamma(a_\alpha, b_\alpha) \\ G_0 &= Normal(u_\beta, \Xi_\beta^{-1}) \end{aligned}$$

We let $a_e = 2.02$ and $b_e = 0.02$. It is a diffuse prior since the distribution has mode $\frac{a_e - 2}{b_e} = 1$ and variance 10000. Figure 2.1 shows the probability density of this gamma distribution, which shows the prior assigns probability to a large range of τ_e. We also let $a_\alpha = a_{\gamma j} = 2.02$ and $b_\alpha = b_{\gamma j} = 0.02$ for all $j = 0, 1, \ldots, p$. In addition, we use $u_\beta = (0, 0, 0, 0)'$ and $\Xi_\beta = diag(0.01, 0.01, 0.01, 0.01)$.

On some occasions, we might be interested in assigning informative priors to $\{\gamma_j : j = 0, 1, \ldots, p\}$ since it represents the variation of individual trajectories within cluster. For example, if we let $\gamma_j \to \infty$ for all j, model (2.2) is simplified to model (2.1) because $W_i \to I_{r_i}$. Therefore, we can control this variation by assigning an appropriate prior to γ. We discuss this in detail in Appendix A1.

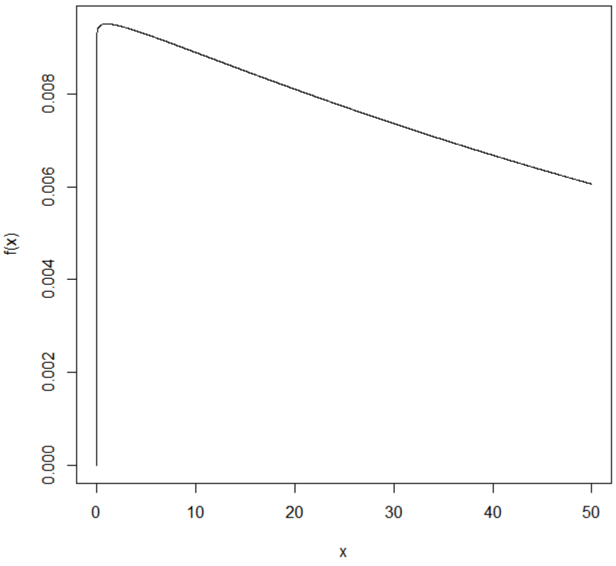


Figure 2.1: Probability density function for Gamma($\frac{2.02}{2}$, $\frac{0.02}{2}$)

2.4 Numerical Approximation to Posterior Distributions

This section presents algorithms used to draw posterior samples and make statistical inferences. We first list the full conditional distributions, and subsequently specify several algorithms used for Gibbs sampling. Then we briefly introduce the Label Switching issue and how we deal with it. It is a common issue in the inference for Bayesian mixture models. We will discuss it in detail in the next chapter. In the last subsection, we focus on statistical inferences.

2.4.1 Posterior Distribution

Recall the model approximation in 1.14. The model we built is

$$\begin{aligned}
y_i|s_i, \boldsymbol{\beta} &\sim N_{r_i}(X_i\beta_{s_i}, \tau_e^{-1}W_i) \\
s_i|\mathbf{p} &\sim Multinomial(p_1, \ldots, p_L) \\
p_1, \ldots, p_L &\sim Dirichlet(\frac{\alpha}{L}, \ldots, \frac{\alpha}{L}) \\
\beta_{s_i} &\sim G_0 \\
G_0 &= N_{p+1}(\mu_\beta, \Xi_\beta^{-1}) \qquad (2.10)
\end{aligned}$$

with $L \to \infty$, where $\boldsymbol{\beta}$ denotes the collection of clusters' β_s values.

The posterior distribution for parameters $\Theta = \{\tau_e, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{S}, \alpha\}$ could be derived from the likelihood and prior distributions introduced above using

$$P(\Theta|y) = \frac{P(y|\Theta)P(\Theta)}{P(y)} \propto L(\Theta)P(\Theta)$$

An MCMC sampling scheme is used to draw samples from posterior distribution. We employ a combination of Gibbs sampling (Gelfand and Smith 1990) and Metropolis within Gibbs sampling, plus we use a Gibbs sampler within two of the Gibbs sampling steps in order to easily sample the parameters α and $\{\gamma_j, j = 0, \ldots, p\}$.

We assigned a conditionally conjugate prior to τ_e, and the full conditional distribution is:

$$\begin{aligned}
\tau_e|else \sim\ & \Gamma\big[\frac{(a_e + n)}{2}, \qquad (2.11) \\
& \frac{\big(b_e + \sum_i (y_i - X(t_i)\beta_{s_i})^T \big(X(t_i)\Gamma^{-1}X(t_i)^T + I_{r_i}\big)^{-1}(y_i - X(t_i)\beta_{s_i})\big)}{2}\big]
\end{aligned}$$

We obtain the full conditionals for s_i and β_s as follow using Equation (2.8):

$$P(s_i = s | else) \propto \begin{cases} n_{s,-i} f(y_i | \beta_{s_i}), & \text{if } s \in \mathbf{S}_{-i} \\ \alpha H_i(y_i), & \text{if } s \notin \mathbf{S}_{-i} \end{cases}$$

$$\beta_s | else \sim N(\widetilde{u}_\beta, \widetilde{\Xi}_\beta^{-1}) \tag{2.12}$$

where,

$$\begin{aligned} H_i &= \int f(y_i | \beta, \Theta) dG_0(\beta) \\ &= N(y_i \mid X_i \mu_\beta, \frac{1}{\tau_e} W_i + X_i \Xi_\beta^{-1} X_i^T) \\ \widetilde{u}_\beta &= u_\beta + \big(\sum_{i:s_i=s} \tau_e X_i^T W_i^{-1} X_i + \Xi_\beta \big)^{-1} \big(\sum_{i:s_i=s} \tau_e X_i^T W_i^{-1} (y_i - X_i \mu_\beta) \big) \\ \widetilde{\Xi}_\beta &= \sum_{i:s_i=s} \tau_e X_i^T W_i^{-1} X_i + \Xi_\beta \\ W_i &= I_{r_i} + X_i \Gamma^{-1} X_i^T \end{aligned}$$

2.4.2 Gibbs within Gibbs Algorithm

The conditional distributions for α and $\gamma_j, j = 0, \ldots, p$ are unrecognizable. We use a method which we call “Gibbs-in-Gibbs” to draw posterior samples for them.

The idea of this so-called “Gibbs-in-Gibbs” sampler comes from the sampling method of the smoothing parameter α in the DP by Escobar and West (1995). Let the prior distribution of α be $\Gamma(a_\alpha, b_\alpha)$. We can sample it with two conditional distributions as below. By considering an auxiliary variable η, the following trick facilitates the posterior MC sampler;

$$\begin{aligned}
\eta|\alpha, K &\sim Beta(\alpha+1, n) \\
\alpha|\eta, K &\sim \pi_\eta Gamma(a_\alpha + K, b_\alpha - log(\eta)) \\
&\quad + (1-\pi_\eta) Gamma(a_\alpha + K - 1, b_\alpha - log(\eta))
\end{aligned} \tag{2.13}$$

where $\frac{\pi_\eta}{1-\pi_\eta} = \frac{a_\alpha + m - 1}{n(b_\alpha - log(\eta))}$. We tried several different values for (a_α, b_α), and found the clustering results were insensitive to the prior distribution of α over the range considered. In the analysis, we have used $a_\alpha = b_\alpha = 0.1$ and $a_\alpha = 2.02, b_\alpha = 0.02$ in the models.

Eacobar and West (1995) has used it to sample α. The full conditional distribution of α is $P(\alpha|K) \propto p(\alpha)\alpha^{K-1}(\alpha+n)\int_0^1 x^\alpha(1-x)^{n-1}\,\mathrm{d}x$. Even though it is not recognizable, it implies that $P(\alpha|K)$ is the marginal distribution from a joint for α and a continuous quantity η such that

$$P(\alpha, \eta|K) \propto p(\alpha)\alpha^{K-1}(\alpha+n)\eta^\alpha(1-\eta)^{n-1}$$

Hence, we could use an auxiliary variable η in the sampling procedure. We have two recognizable conditional posteriors $P(\alpha|\eta, K)$ and $P(\eta|\alpha, K)$ shown in Equation (2.13). At each Gibbs iteration, α is sampled in two steps using (2.13): (1) first sampling an η value from the beta distribution; (2) and then sampling the new α value from the mixture of gamma distributions.

$\boldsymbol{\gamma}$ could be sampled using the same idea as for α. Note our model (2.2) is a marginalized mixed model from (2.3). Taking the mixed term b_i as an auxiliary variable, we can write the probability density function for y_i in the following form, where $f_N(\cdot|\mu, \Sigma)$ is denoted as the probability function of multivariate normal distribution with mean vector μ and covariance matrix Σ;

$$\begin{aligned} f(y_i|\Theta) &= f_N(y_i|X_i\beta_{s_i}, \tau_e^{-1}W_i) \\ &= \int f_N(y_i|X_i(\beta_{s_i}+b_i), \tau_e^{-1}I_{r_i})\, f_N(b_i|0, \tau_e^{-1}\Gamma^{-1})\, db_i \end{aligned}$$

Then, we can write the posterior distribution of $\boldsymbol{\gamma}$ in a integral form as below.

$$P(\boldsymbol{\gamma}|else) \propto \Big(\prod_{i=1}^{n}\int f_N(y_i|X_i(\beta_{s_i}+b_i), \tau_e^{-1}I_{r_i})\, f_N(b_i|0, \tau_e^{-1}\Gamma^{-1})\, db_i\Big)\cdot p(\boldsymbol{\gamma})$$

Let B denote the mixed terms, $B = \{b_i : i = 1, \ldots, n\}$. Then $P(\boldsymbol{\gamma}|else)$ could be regarded as the marginal distribution of a joint of $\boldsymbol{\gamma}$ and B;

$$P(\boldsymbol{\gamma}, B|else) \propto \Big(\prod_{i=1}^{n} f_N(y_i|X_i(\beta_{s_i}+b_i), \tau_e^{-1}I_{r_i})\cdot f_N(b_i|0, \tau_e^{-1}\Gamma^{-1})\Big)\cdot p(\boldsymbol{\gamma})$$

With the mixed terms B as auxiliary variables, it becomes much easier to draw posterior samples for $\{\gamma_j, j = 0, \ldots, p\}$. Two conditional distributions could be obtained from the joint distribution $P(\boldsymbol{\gamma}, B|else)$;

$$\begin{aligned} b_i|y,\Theta &\sim N(\widetilde{\mu}_{b_i}, \widetilde{\Xi}_{b_i}^{-1}),\ i = 1, 2, \ldots, n \qquad (2.14) \\ \gamma_j|y, B, else &\sim \Gamma(\widetilde{a}_{\gamma j}, \widetilde{b}_{\gamma j}),\ \ j = 0, 1, \ldots, p \qquad (2.15) \end{aligned}$$

where

$$\begin{aligned}
\widetilde{\mu}_{b_i} &= (X_i^T X_i + \Gamma)^{-1} X_i^T (y_i - X_i \beta_{s_i}) \\
\widetilde{\Xi}_{b_i} &= \tau_e (X_i^T X_i + \Gamma) \\
\widetilde{a}_{\gamma j} &= \frac{1}{2}(a_{\gamma j} + n) \\
\widetilde{b}_{\gamma j} &= \frac{1}{2}(b_{\gamma j} + \tau_e \sum_{i=1}^{n} b_{ij}^2)
\end{aligned}$$

At each Gibbs iteration, we sample b_i from the normal distribution (2.14) for all $i = 1, \ldots, n$, and then γ_j from the gamma distribution (2.15) for all $j = 0, 1, \ldots, p$.

With all the full conditionals listed above, the Gibbs sampling algorithm for posterior simulation consists of nine steps:

1. Assign appropriate initial values to all the parameters $\Theta^0 = \{\tau_e^0, \beta_i^0, \alpha^0, \Gamma^0\}$. Then we obtain all s_i^0 and $\beta_{s_i}^0$ values based on β_i^0.

2. At iteration l, draw a new τ_e^l with Equation (2.11) using parameter values from previous step.

3. Sample cluster membership $\boldsymbol{s}^l$ with posterior distribution $s_i^l | y_i, \tau_e^l, \beta^{l-1}, \alpha^{l-1}, \Gamma^{l-1}$ given in Equation (2.4.1) for all $i = 1, \ldots, n$.

4. Count the number of clusters K^l.

5. Update β_s with $\beta_s^l \mid y, \boldsymbol{s}^l, \tau_e^l, \alpha^{l-1}, \Gamma^{l-1}$ using Equation (2.4.1) for all $s = i, \ldots, K^l$.

6. Obtain β_i^l with $\beta_i^l = \beta_{s=s_i}^l$.

7. Sample η^l and α^l using Equation (2.13).

8. Sample auxiliary parameter b_i^l from conditional posterior $b_i^l|y_i, \tau_e^l, \beta_i^l, \alpha^l, \Gamma^{l-1}$ using Equation (2.14) for all $i = 1, \ldots, n$.
9. Sample γ_j^l from $\gamma_j^l|y, \tau_e^l, \beta_i^l, \alpha^l, B^l$ using Equation (2.15).
10. Repeat Step 2-9 iteratively for $l = 1, \ldots, N^{MC}$ to reach convergence.

2.4.3 Statistical Inference

We are interested in making inferences about parameters $\Theta = \{\tau_e, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{S}, \alpha\}$, especially how the individuals are clustered. However, it is not easy to draw inferences directly from MCMC output using ergodic averaging. The main challenge in Bayesian analysis with mixtures is the non-identifiability of the mixture components, since the labels $\{s_i : i = 1, \ldots, n\}$ have no physical meaning in terms of the model. When there are K components in the model, we can find $K!$ equivalent labelings for the posterior distribution of β. That is, the posterior distribution is invariant to permutations in the labeling of the parameters. That can cause problems when we try to estimate parameters that relate to individual components of the mixture.

Due to identifiability issues mentioned above, we are not able to make inference directly from MCMC output for the parameters related to individual components of the mixture, which are $\mathbf{S} = \{s_i,\ i = 1, \ldots, n\}$ and $\boldsymbol{\beta} = \{\beta_s,\ s = 1, \ldots, K\}$. However, we still can draw inferences directly using ergodic averaging for τ_e, $\boldsymbol{\gamma} = \{\gamma_j,\ j = 0, 1, \ldots, p\}$ and α since they are not involved with the cluster information. We call them global parameters of the model.

The key goal is to identify the latent clusters from a large number of MCMC samples, say N_{mc}, taken from the joint posterior distribution. We use a hierarchical clustering algorithm that was proposed by Medvedovic and Sivaganesan (2002) and used by Bigelow and Dunson (2005). A "pairwise distance" measure r_{ij} is defined between any two individuals i and j as follows:

$$r_{ij} = \frac{\sum_{iter=1}^{N_{mc}} I(s_i^{iter} \neq s_j^{iter})}{N_{mc}}. \tag{2.16}$$

It is the proportion of iterations in which i and j were not classified in the same cluster. Then the minimum distance $r_{ij} = 0$ means the two individuals have the same mean trajectory and they were clustered together throughout the MCMC iterations, and the maximum $r_{ij} = 1$ means the two individuals have different trends and they were never in the same cluster.

A dendrogram could be made with the pairwise distances obtained above. A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by a hierarchical clustering method. It is created by the linkage function which specifies the dissimilarity of sets of individuals as a function of the pairwise distances in the sets. Some commonly used linkage functions include complete/maximum linkage, single/minimum linkage and average linkage. In our analysis, we chose complete linkage to make the dendrogram.

We require a threshold distance for the dendrogram to cluster the individuals based on the pairwise distances. Individuals with distance smaller than the threshold are allocated to the same cluster. Thus the threshold is the minimum distance allowed between two individuals in different clusters. A large threshold yields few clusters, and a small threshold leads to many clusters. The posterior distribution of K could be helpful in choosing the appropriate threshold. If the posterior distribution of cluster number K has the mode K_0, we could pick the threshold so that around K_0 clusters are obtained from the dendrogram.

It is worth mentioning that no method could be ideal, because of the impossibility of directly obtaining a result from the Markov Chain due to non-identifiability of the mixture components. Researchers have made efforts to deal with label switching, and have attempted to draw inferences about $\mathbf{S}$ directly from MCMC output. But that is difficult, especially for an infinite mixture model.

Inference about $\boldsymbol{\beta}$ is of great interest since each β_s provides the mean trajectory of cluster s for all $s = 1, \ldots, K$. But it can not be estimated easily with ergodic averaging due to the label switching issue. The difficulty lies in identifying the posterior samples for each cluster. In the next chapter, we focus on the label switching issue in Bayesian infinite mixture models and will present our method for constructing the posterior samples for $\boldsymbol{\beta}$. In this chapter, we only use the result obtained from the method when it is needed.

In longitudinal data analysis, we are often interested in predicting the individual's future response based on the trajectory trend. For example, in our hormone problem, it would be interesting to predict hormone values after menopause based on a woman's hormone responses collected up to some current time, say t_{ir_i} for woman i. Let $\widetilde{y}_i$ be the observations that are to be predicted for individual i, and let y_i be the observations already collected from this individual. Then the prediction is accomplished using the predictive density

$$f(\widetilde{y}_i|y_i) = \int f(\widetilde{y}_i|\Theta, y_i)p(\Theta|y)\, d\Theta$$

2.5 Simulations

In this section, we illustrate the validity of our clustering method by applying it to simulated datasets. We first simulate two datasets. One is balanced without missing values (denoted as SimData 1), and the second is unbalanced with missing values (SimData 2). We apply our model to both in order to check model validity. Then we compare the clustering performance of our clustering model (2.2) with that of model (2.1), which has an independence assumption, using a third simulated dataset (SimData 3). We finally discuss an MCMC convergence problem caused by over-fitting.

2.5.1 Model Validity with Balanced Complete Data

The simulated data are drawn from three clusters with different mean trajectories to show how the method works. One hundred individuals with 1100 observations in total were simulated from the three clusters of size 61, 28 and 11, respectively. The time scale for the longitudinal data was defined at 11 discrete points $\{-3, -2, \ldots, 7\}$. At each time point, $t_{ij}, j = -3, -2, \ldots, 7$, the response y_{ij} was generated for each individual $i, i = 1, \ldots, n$. This dataset is balanced and complete. For an individual i that belongs to group $s, s = 1, 2, 3$, the response y_i was generated using mixed effects model with $y_i = X_i\beta_s + Z_i b_i + \epsilon_i$, where random errors, ϵ_i, were generated as $N(0, 1)$. We generate the mean curve $X_i\beta_s$ for each cluster, s, using orthogonal basis functions up to order 3, and the true coefficients, β, for the three clusters are $(1.5, 4, -1, -0.5)'$, $(0.25, 2, -0.5, 0)'$ and $(0, 3, -3, -1.5)'$, respectively. The random effects within cluster are simulated with the mixed terms $Z_i b_i$; Z_i is the design matrix of orthogonal basis functions up to degree 5, and $b_i \sim N_6((0, 0, 0, 0, 0, 0)', diag((1.5, 2, 1.5, 1, 0.1, 0.1)^2))$. The true value of $\boldsymbol{\gamma}$ is $(0.44, 0.25, 0.44, 1, 100, 100)$. The combined simulated data are plotted in Figure 2.2.

Note that we used Z_i up to degree 5 to generate mixed effects instead of using $Z_i = X_i$ with order 3. We would like to "challenge" our model by generating extra variation that the model does not account for. We want to see whether it still performs well in such a case.

We ran model (2.10) in R for 20000 iterations on a desktop equipped with Intel Core2 Duo CPU E6550 @ 2.33GHz and 3GB RAM. It took us about 2 hours to get the output. Due to lack of identifiability, we are not able to make inferences directly from the MCMC output for the parameters on the cluster level, namely for $s_{i:i=1,\ldots,n}$ and $\beta_{s:s=1,\ldots,K}$. However, we still can make inferences directly for other parameters including τ_e and $\gamma_{j:j=0,1,\ldots,p}$. Table 2.1 shows the parameter estimates and 95% posterior probability intervals. Since the model only used degree 3 for the random effect within cluster, there is no estimate for γ_4 and γ_5. In addition,

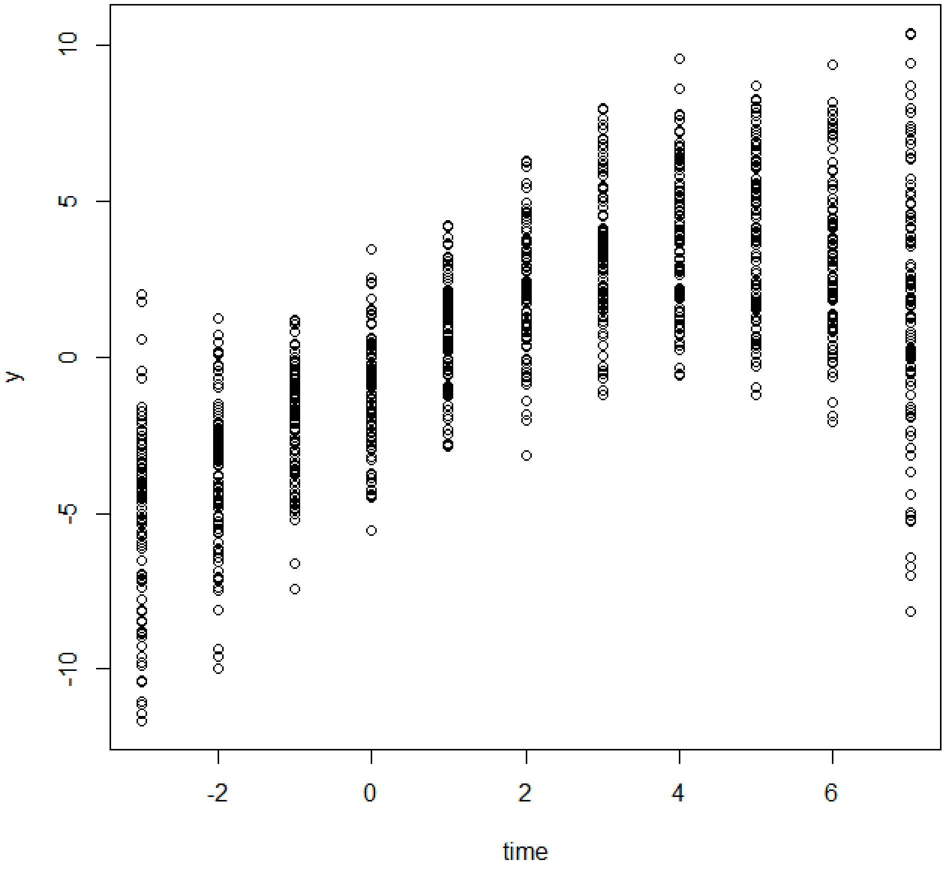


Figure 2.2: Sphagetti plot of the simulated data, SimData 1.

Table 2.1: Parameter estimates for simulated data, SimData 1.

				95% PI	
Parameter	Truth	Mean	Median	Lower	Upper
τ_e	1	1.0	1.0	0.95	1.06
γ_0	0.44	0.41	0.40	0.27	0.59
γ_1	0.25	0.22	0.22	0.14	0.32
γ_2	0.44	0.61	0.59	0.35	0.95
γ_3	1	0.84	0.81	0.53	1.31
γ_4	100				
γ_5	100				

Table 2.2: Posterior Distribution of K for SimData 1

Cluster Number K	3	4
Posterior Probability	0.979	0.021

the posterior mean and median of γ_2 and γ_3 are somewhat off from the truth even though they both are within 95% probability intervals.

Inference for the number of clusters K can also be used to assess the validity of the clustering result. We have three clusters in truth for the dataset. The posterior distribution of K in Table (2.2) shows the number is 3 with probability 0.979, which fits the truth very well. In addition, this posterior distribution of K is useful for our analysis using the hierarchical clustering method. Information about K helps in choosing the appropriate threshold.

We use hierarchical clustering with complete linkage to summarize the clustering information from the MCMC output. The dendrogram (2.3) was made based on the "pairwise distances" r_{ij} with complete linkage. The threshold was chosen to be 0.65 to obtain 3 clusters since $K = 3$ has the largest posterior probability. The dendrogram shows that the 3 clusters were well separated and identified with our model. We construct Table 2.3 to show the accuracy of our clustering outcomes; 93% of individuals are correctly clustered.

Figure (2.4) shows that our model can identify the three trajectory patterns in truth. The allocation of the individuals from our model (lower plot) fits the truth (upper) very well. Considering the similarity of the 3 cluster mean trajectories, the model performance is sat-

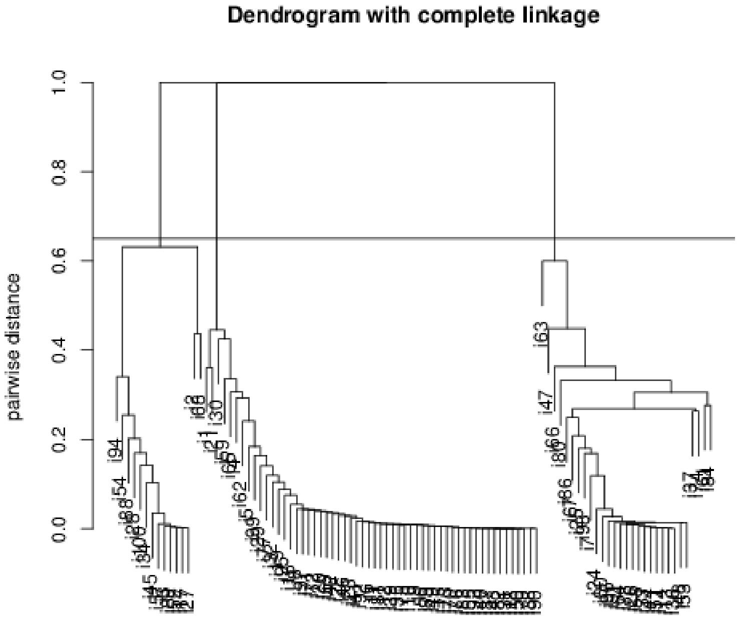


Figure 2.3: Dendrogram constructed according to hierarchical clustering model for SimData 1 using complete linkage with cutoff=0.65

Table 2.3: Classification Accuracy: SimData 1

		Model Classification		
		Cluster 1	Cluster 2	Cluster 3
Truth	Cluster 1	55	2	4
	Cluster 2	1	27	0
	Cluster 3	0	0	11

isfying. In addition, Figure (2.5) compares the estimated cluster mean trajectories with the truth. The three true trajectories are plotted as solid lines, and the fitted mean curves with 95% probability bands in dashed lines. The plots show the three trajectory trends were well estimated by the model.

2.5.2 Model Validity: Unbalanced Incomplete Data

In longitudinal studies, missing data are very common due to missed appointments, dropouts, and data analysis occurring before the study has been completed. For example, in the SWAN data, the hormone observations after FMP are missing for women who have not reached it yet at the time of data analysis. In addition, many longitudinal studies do not have balanced designs. In that case, the time scale has to be treated as continuous, and we have to use continuous basis functions to fit the data. In this simulation study, we are showing the model validity using Legendre basis functions applied to unbalanced data with missing values. The simulated data are shown in Figure (2.6).

The dataset was generated in two steps. We created a complete dataset with 100 individuals, in which each individual has 10 observations at different times for different individuals. The time of the observations ranges from -1 to 1. We generated individuals from 3 clusters with mean trajectory generated with Legendre basis functions up to degree 3. Their coefficients were $(2, 4.5, -1, -0.5)$, $(0, 2, -0.5, 0)$ and $(0, 2.5, -2, -2)$. We also generated individual variation within clusters and white noise in the data in a similar manner as above. The true values of the parameters were $\tau_e = 1$ and $\boldsymbol{\gamma} = (1, 0.44, 4, 1.78)$.

We then assign missing data to 50 individuals who were randomly selected out of the 100. The missingness was created as follows: for each selected individual i, a number j was randomly sampled from $\{2, 3, \ldots, 10\}$, and then all the observations on and after t_{ij} were set to be missing for the individual.

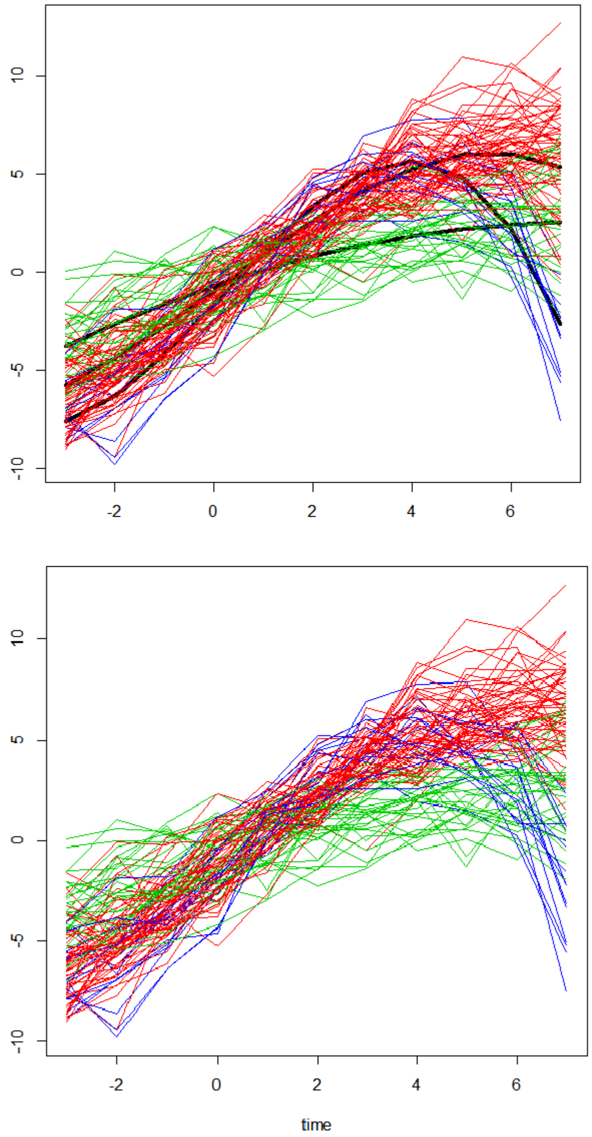


Figure 2.4: Sphagetti plots for SimData 1. Upper: true clustering with true mean trajectories using solid lines. The individuals are plotted with different colors indicating different cluster memberships. Lower: clustering of the individuals from the model.

Table 2.4: Parameter estimates for unbalanced data, SimData 2

				95% PI	
Parameter	Truth	Mean	Median	Lower	Upper
τ_e	1	1.09	1.09	0.95	1.24
γ_0	1	1.25	1.16	0.6	2.4
γ_1	0.44	0.42	0.4	0.27	0.65
γ_2	4	3.41	3.13	1.62	6.63
γ_3	1.78	2.04	1.9	1.16	3.6

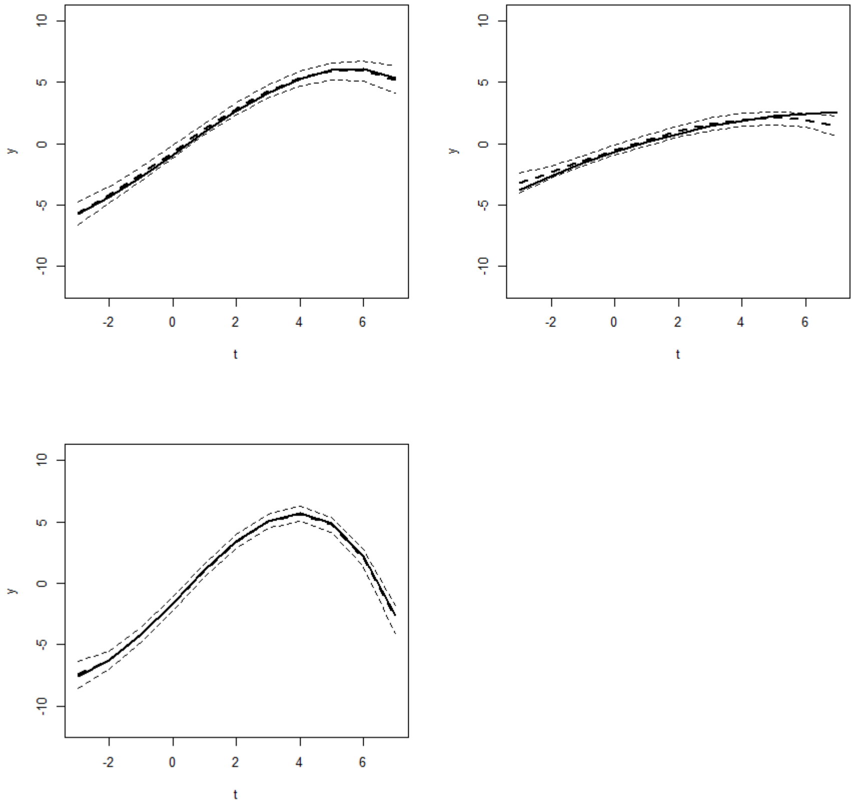


Figure 2.5: The mean trajectories of the 3 clusters for SimData. Solid: true curves; Dashed: the estimated cluster mean trajectories with 95% probability bands

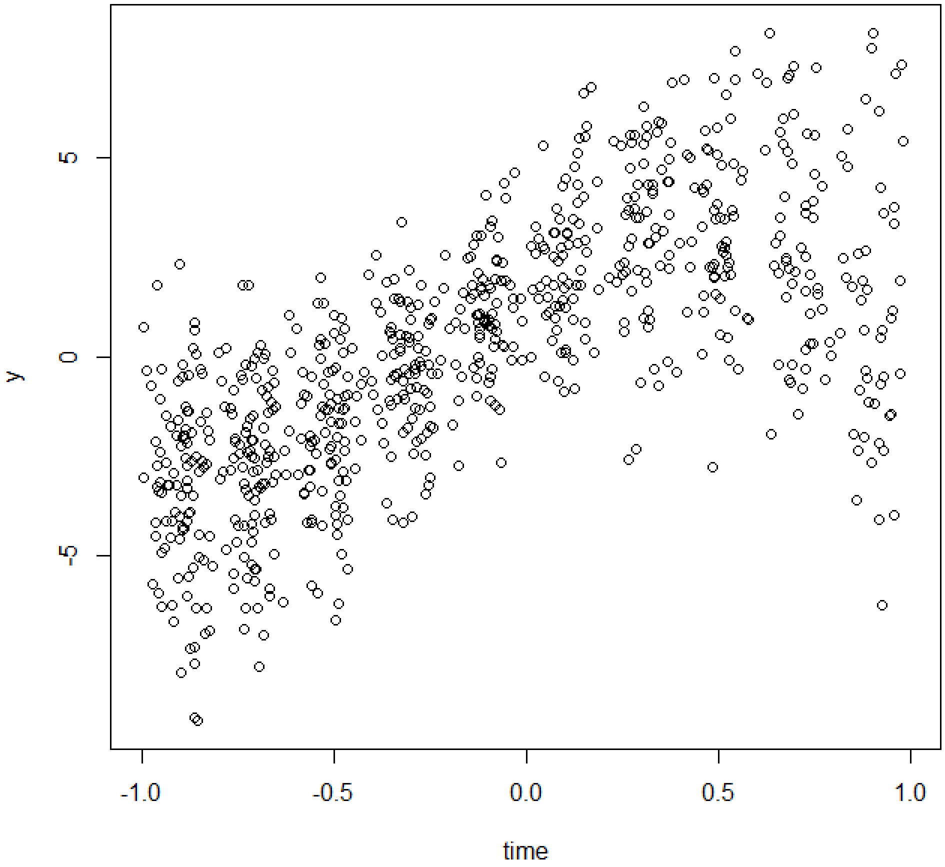


Figure 2.6: Scatterplot of unbalanced simulated data with missing values, SimData 2.

Table 2.5: Posterior Distribution of K for unbalanced data, SimData 2

Cluster Number K	3	4	5	6
Posterior Probability	0.888	0.102	0.009	0.001

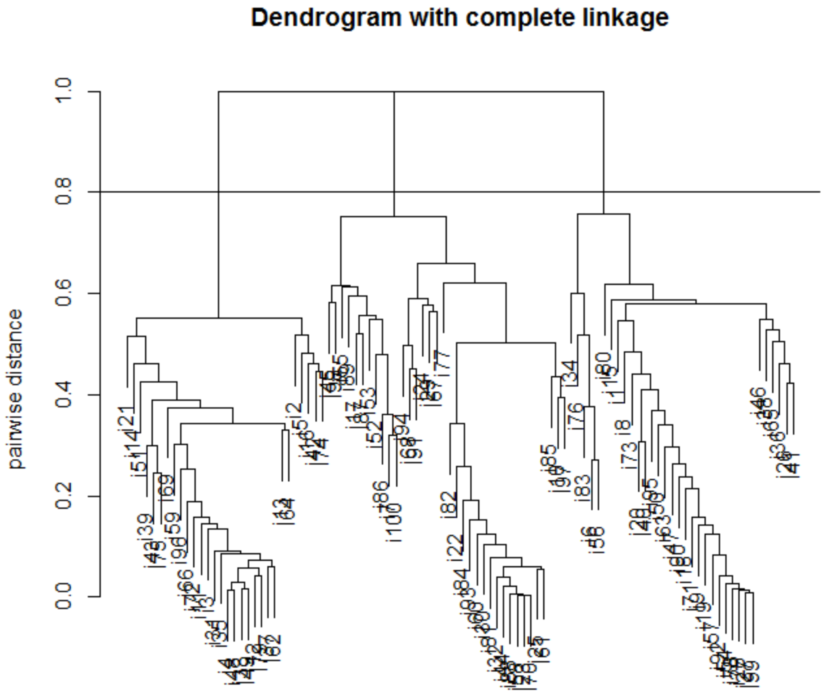


Figure 2.7: Dendrogram based on the hierarchical clustering model for SimData 2

Table 2.4 gives the posterior inferences for the global parameters, which shows reasonable model fitting since all the true parameter values are captured by 95% probability intervals. Then we obtain the clustering information for the individuals with the posterior distribution of cluster number K in Table 2.5 and the corresponding dendrogram in Figure 2.7. The clustering accuracy is presented in Table 2.6, and 75% of the sample has been correctly classified. Precision is not as good as it was for the balanced data. This is not surprising considering the extent of missing data, which was 23.9% . We found that the majority of mis-classified individuals had missing data. We also made the spaghetti plots in Figure 2.8 to show that our model still identified the three trajectory trends well.

Table 2.6: Classification Accuracy for SimData 2

		Model Classification		
		Cluster 1	Cluster 2	Cluster 3
Truth	Cluster 1	28	5	6
	Cluster 2	5	25	2
	Cluster 3	3	4	22

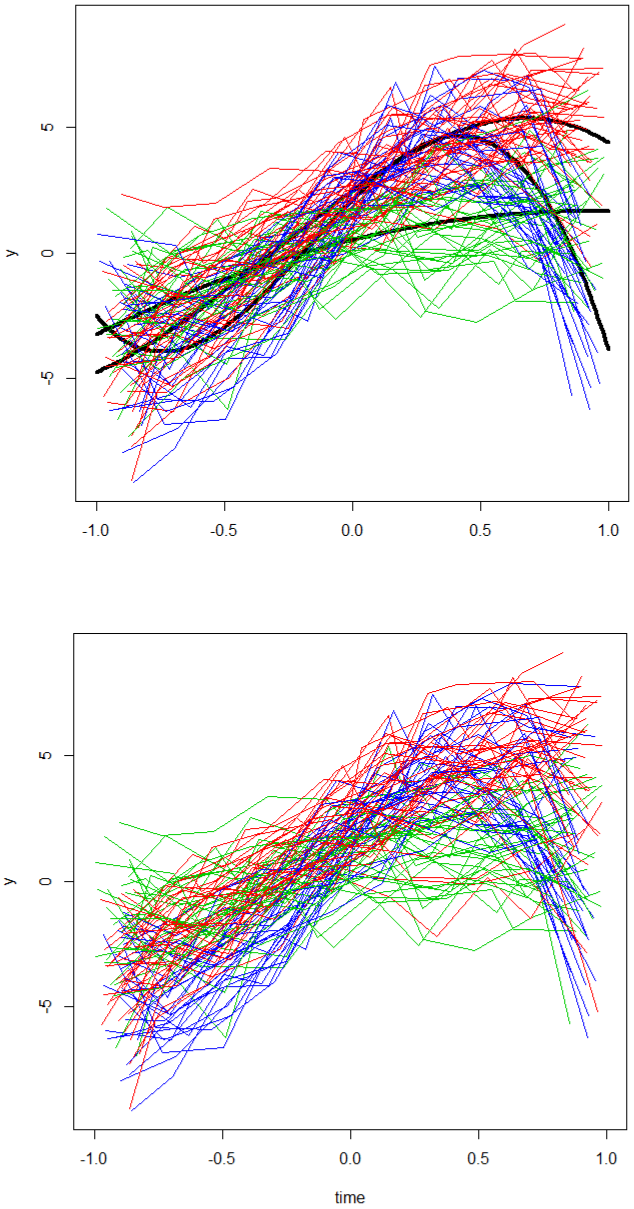


Figure 2.8: Sphagetti plots of the individuals with clustering for SimData 2. Upper: true clustering. The individuals are plotted with different color indicating different cluster membership. The black solid lines are the mean trajectories of the three clusters. Lower: clustering from the model.

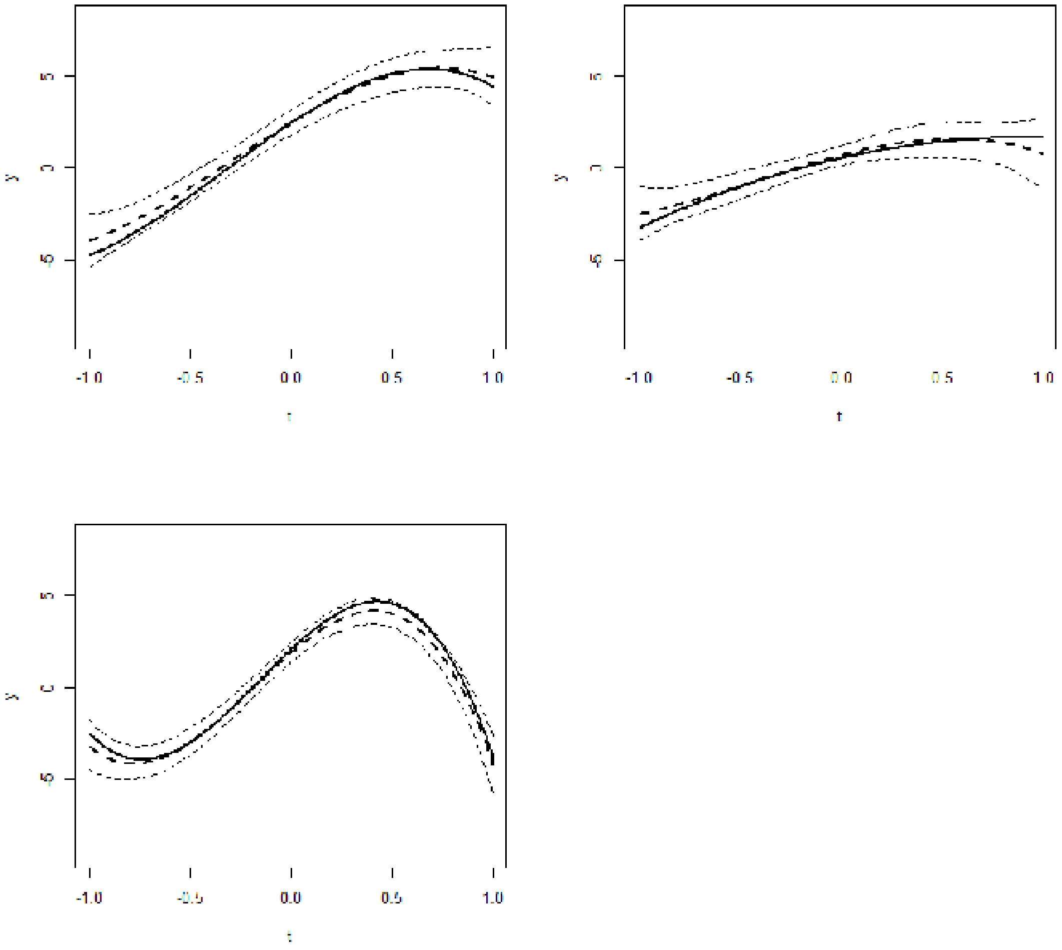


Figure 2.9: The mean trajectory of the 3 clusters for SimData 2. Solid line: true trajectory; Dashed line: the cluster mean estimate and 95% probability bands.

2.5.3 Model Comparison

Here we compare the model performance between our model (2.2) and model (2.1). Here we use another simulated dataset for the analysis. Since model (2.1) is a special case of our model (2.2) with $\gamma_j \to \infty$ for all $j = 0, 1, \ldots, p$. It makes it easy to apply the model by assigning large values to $\boldsymbol{\gamma}$. In this case, we assigned an strongly informative prior $Gamma(\frac{a_\gamma}{2}, \frac{b_\gamma}{2})$ to γ_j, with $a_\gamma = 10^6$ and $b_\gamma = 1$. The posterior estimates of γ_j's are all greater than 10000.

We generate a new dataset (denoted by SimData 3) for the comparison, because our clustering model (2.2) has a clear advantage over model (2.1) especially when the clusters are not easily identified. Figure 2.10(a) is a scatterplot of the simulated dataset, SimData 3. We can see the trend difference is hardly seen in this plot. This dataset emphasizes a difficulty in clustering when the trajectory trends are visually indistinguishable. Usually they are covered by the measurement errors and/or the individual trajectory variations. Figure 2.10(b) presents the three true cluster mean curves (black solid lines) together with the spaghetti plot of the individuals in each cluster separated with different colors.

The data were sampled from three tight clusters with 73, 78 and 49 individuals in each cluster shown in Figure 2.10(b). The dataset has 1671 observations in total for the 200 individuals, and the number of observations for each individual ranges from 2 to 10. The x-axis is time, which consists of randomly generated continuous values from -1 to 1. The y-axis represents a continuous time-variate response. The correlation within individuals was induced by assigning each individual a random term, $X(t_i)b_i$, with b_i drawn from

$$N_{p+1}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1.5^2 & 0 & 0 \\ 0 & 0 & 0.5^2 & 0 \\ 0 & 0 & 0 & 0.75^2 \end{pmatrix}\right)$$

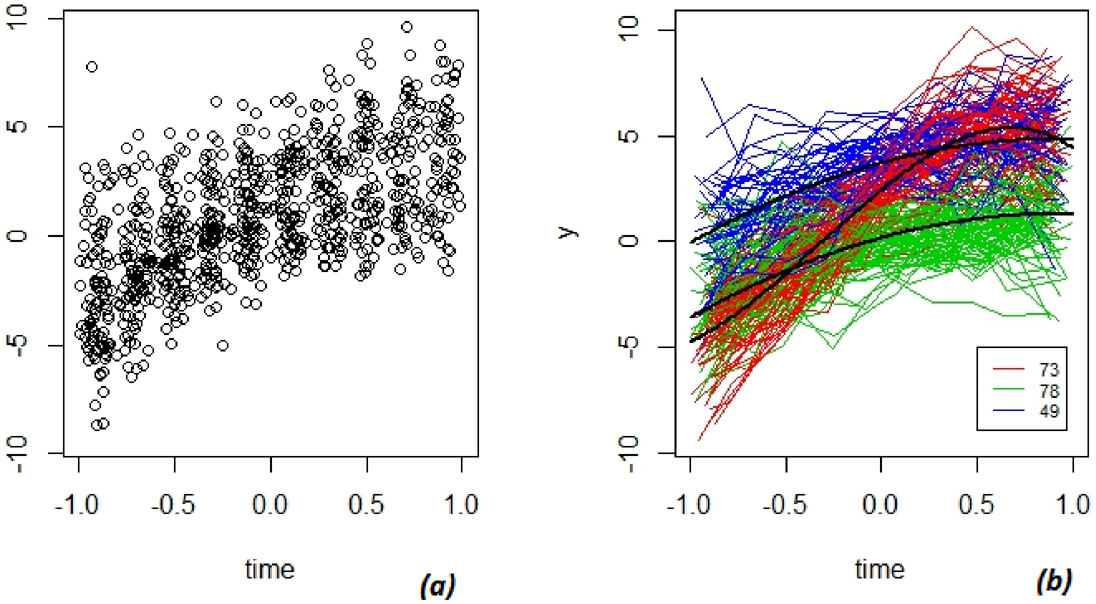


Figure 2.10: SimData 3. (a) is a scatterplot of all the observations; (b) gives the true cluster mean trajectories (solid black curves) and the spaghetti plot of the individuals in each cluster distinguished with three colors.

Table 2.7: Posterior Distribution of the Number of Clusters K using model (2.10) for SimData 3.

K	3	4	5	6
Posterior Prob	0.9283	0.0681	0.0035	0.0001

In addition, the white noise for each observation was drawn from $N(0, 0.75^2)$.

Table 2.7 lists the posterior probabilities of number of clusters K, and it shows that $K = 3$ has the highest posterior probability 92.83%. The predictive distribution of β^{new} in Figure 2.11 shows the location of the three clusters clearly.

We also obtain clustering results using model (2.1) which has no mixed effects, and compare them with the results from our model presented above.

Using model (2.1), we obtain posterior probabilities of number of the clusters in Table 2.8, and the predictive distribution of β^{new} in Figure 2.12. It is clear the model is not capable of identifying the three true trends in the simulated data, since the number of clusters K

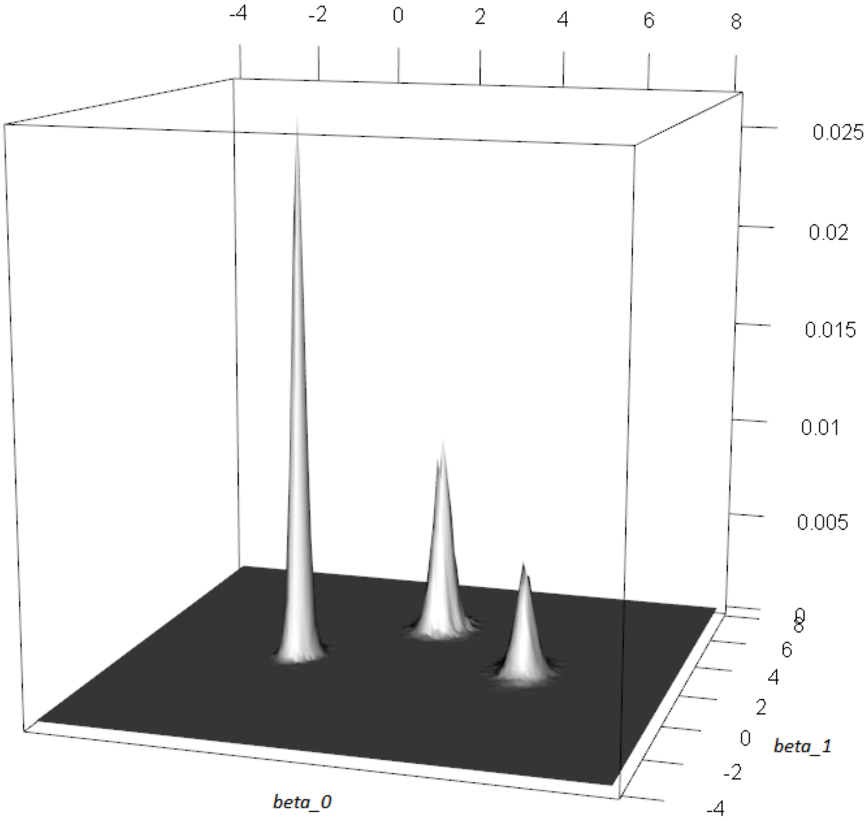


Figure 2.11: Predictive distribution for a new latent variable, β^{new}, for SimData 3. Note the plot was made with the first two elements of the β^{new} vector $\{\beta_0^{new}, \beta_1^{new}\}$ only, since we could not accommodate all 4 parameters in one plot.

Table 2.8: Posterior Distribution of the Number of Clusters K using model (2.1); SimData 3.

K	28-30	31-35	36-40	40-45
Posterior Prob	0.038	0.620	0.323	0.019

Table 2.9: Posterior Distribution of the number of clusters K using model (2.1) for SimData 1

Cluster Number K	4	5	6	7	8	9
Posterior Probability	0.018	0.263	0.313	0.312	0.090	0.003

ranges from 28 to 45 in the posterior samples. The predictive distribution also shows that the model produced too many clusters compared to the truth.

In order to confirm our findings, we apply model (2.1) to SimData 1 and SimData 2. The posterior distribution of cluster number K is listed in Table 2.9 for SimData 1. The model anticipates six to nine clusters with high posterior probabilities, which is considerably more than the truth.

We also applied the model to SimData 2. The clustering result is even messier. The posterior distribution of K in Table 2.10 shows there were too many clusters in the model compared with the truth.

We finally check the performance of our model applied to SimData 3. Table 2.11 shows the inferences for global parameters; the true parameter values are all captured by their 95% probability intervals.

The dendrogram in Figure 2.13 shows that the three clusters were clearly separated. A

Table 2.10: Posterior Distribution of the number of clusters K using model (2.1) for SimData 2

Cluster Number K	4	5	6	7	8
Posterior Probability	0.001	0.026	0.133	0.31	0.29
Cluster Number K	9	10	11	12	13
Posterior Probability	0.153	0.063	0.018	0.006	0.001

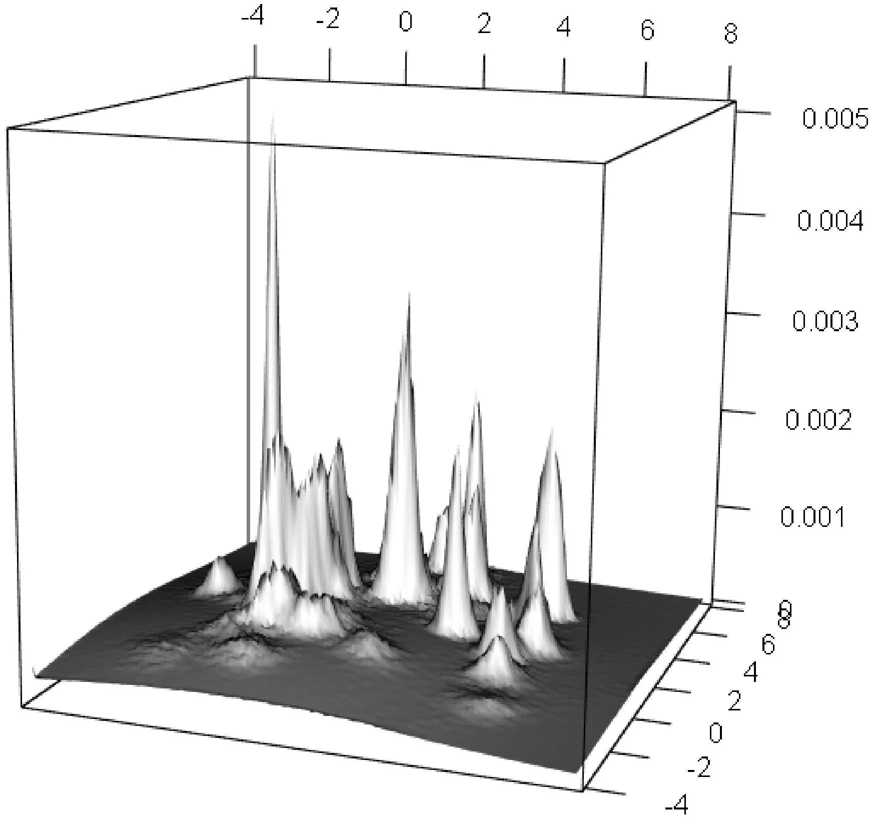


Figure 2.12: posterior predictive distribution for a new latent variable β^{new} using model (2.1) for SimData 3. Note that only $\{\beta_0^{new}, \beta_1^{new}\}$ is plotted.

Table 2.11: Global Parameters Estimates for SimData 3

Para-				95% PI	
meter	Truth	Mean	Median	Lower	Upper
τ_e	0.75	0.75	0.75	0.72	0.79
γ_0	0.56	0.65	0.64	0.44	0.91
γ_1	0.25	0.22	0.22	0.16	0.29
γ_2	2.25	2.87	2.80	1.92	4.21
γ_3	1	1.21	1.20	0.91	1.59
α		0.41	0.34	0.05	1.17

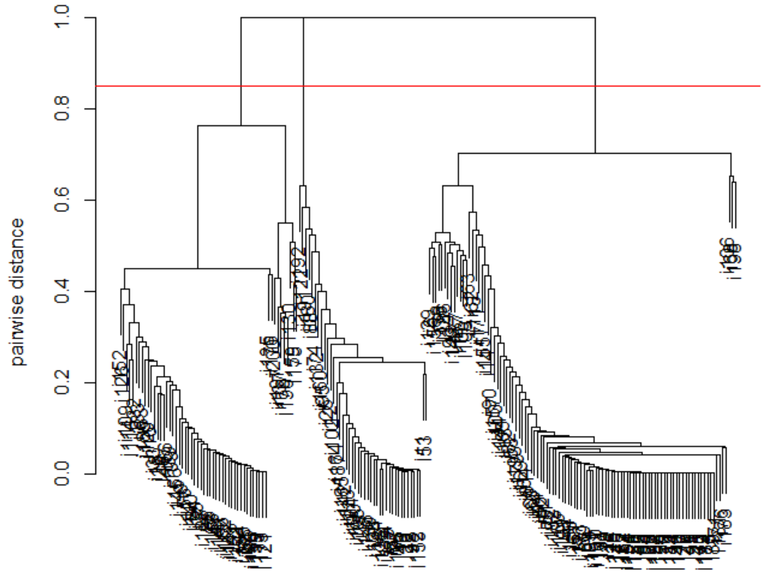


Figure 2.13: The Dendrogram for Hierarchical Clustering Algorithm with Complete Linkage for SimData 3. The red horizontal line is the threshold at which the clusters were created, and three clusters are apparent at the top of the tree.

Table 2.12: Clustering Accuracy for SimData 3

		Clustering Result		
		Cluster 1	Cluster 2	Cluster 3
	Cluster 1	53	16	4
Truth	Cluster 2	1	77	0
	Cluster 3	7	1	41

cutoff 0.85 of the mutual distance was chosen to obtain the clustering information of the individuals.

Table 2.12 shows the accuracy of the clustering result for the simulated data; 171 out 200 individuals were correctly clustered. Considering the close proximity of the three clusters, 85.5% accuracy rate is good. In addition, comparing Figure 2.14 and Figure 2.10b, it is clear that the three true clusters were identified by our model.

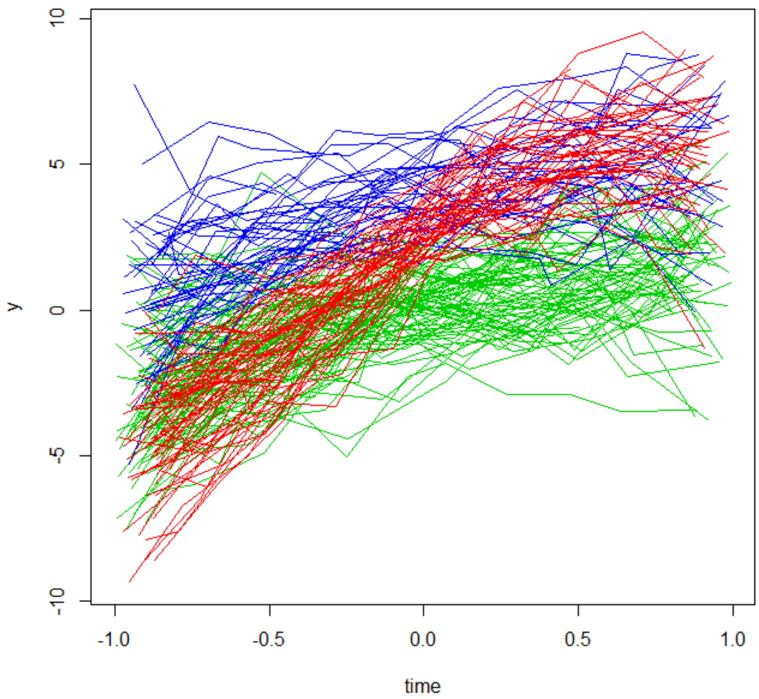


Figure 2.14: Clustering Results for SimData 3. Colors separate the individuals into different clusters.

2.5.4 Over-fitting

We have mentioned that over-fitting could be a major issue in estimating trajectory coefficients for individuals if we use individual coefficients for clustering. Here we apply such a model listed below to SimData 2 to highlight the issue. In this model, the DPM is applied to β_i, in which the individual curve is estimated based on y_i only. Since the data for some individuals are sparse in the unbalanced dataset, we believe the estimated β_i's could be bad leading to problematic clustering.

Table 2.13: Posterior Distribution of K for two Markov Chains using model (2.17) with SimData 2

	Chain 1		Chain 2	
Cluster Number K	4	5	6	7
Posterior Probability	0.999	0.001	0.992	0.008

The model is:

$$\begin{aligned}
y_i|\beta_i &\sim N_{r_i}(X_i\beta_i, \tau_e^{-1}I_{r_i}) \\
\beta_i|\mu_{\beta_i} &\sim N_{p+1}(\mu_{\beta_i}, (\tau_e\Gamma)^{-1})) \\
\mu_{\beta_i}|G &\overset{iid}{\sim} G \\
G &\sim DP(\alpha, G_0) \\
G_0 &= N_{p+1}(\mu_\beta, \Xi_\beta^{-1}) \qquad (2.17)
\end{aligned}$$

Using this model for SimData 2, we had convergence problems. We ran the model twice with different initial values for MCMC. The posterior distribution of K was different for the two chains shown in Table 2.13. In chain 1, there were 4 clusters in most iterations of the MCMC. But there were 6 in chain 2. Neither chain converged to the true number of clusters 3. In addition, inferences for other parameters were incorrect since all the 95% probability intervals failed to cover the true values used for simulating the data. For example, the true value of τ_e is 1, but its 95% PI is $(0.29, 0.35)$ for chain 1 and $(0.3, 0.36)$ for chain 2. All of the issues made it impossible to obtain clustering results based on this model for SimData 2.

Table 2.14: Global Parameter Estimates for the model applied to log $E2$ data.

			95% PI	
Parameter	Mean	Median	Lower	Upper
σ_e	0.67	0.67	0.66	0.70
γ_0	1.64	1.63	1.14	2.21
γ_1	16.67	11.63	3.45	69.40
γ_2	3.40	2.87	1.59	7.95
γ_3	31.59	23.73	11.27	93.06

2.6 SWAN Data Analysis

The Study of Women's Health Across the Nation (SWAN) is a multi-center, multi-ethnic, prospective study of the menopausal transition, which is designed to study women's health during this period. Hormone data were collected on women during this time period. In the dataset, 11 years of E2 (Estradiol) values were collected annually from 928 women. In addition, the time of Final Menstrual Period (FMP) was also recorded for each woman. In the analysis, we re-adjusted the time scale by setting $t_i = 0$ to be the year of FMP for woman i. As a result, the time scale ranges from -10 to 9, and hormone observations are missing in 9 out of the 20 years for each woman. Figure 2.15 shows the scatterplot of the log transformed E2 hormone data after time re-adjustment. Ultimately, we are interested in characterizing hormone profiles of women during menopause.

We applied our clustering model to the E2 hormone data using orthogonal polynomial basis functions for the trajectories. The analysis yields several interesting results. Inferences for the global parameters are listed in Table 2.14. Among the estimated $\gamma_j, j = 0, \ldots, p$ values, the estimate for $\gamma_0 = 1.64$ is the smallest and much lower than the other 3. This indicates the individual variation within cluster mostly lies on the intercept of the curve, since γ_j's are precision parameters. The trajectory trend does not have much variation within cluster.

We identified four clusters with different trajectory trends. Table 2.15 shows that the posterior probability for $K = 4$ is 89.1%.

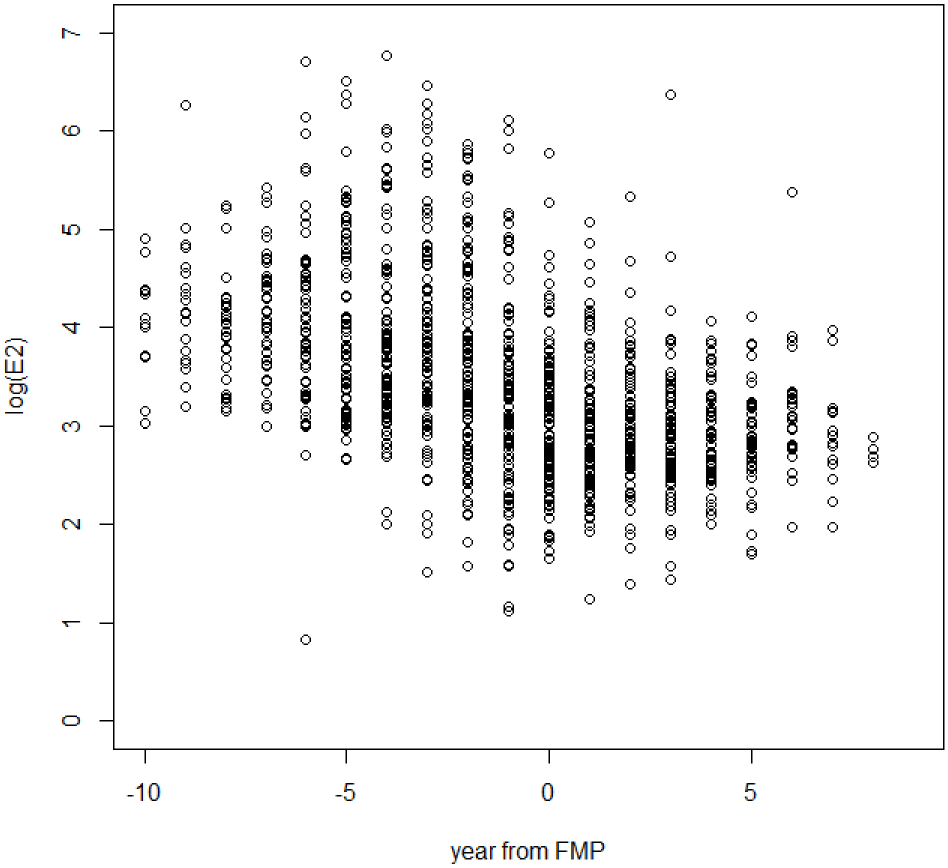


Figure 2.15: Scatterplot of $\log E2$ Data from SWAN. The time scale has been re-adjusted to $t_{FMP} = 0$.

Table 2.15: Posterior Distribution of the Number of Clusters K for $\log E2$ hormone data

Cluster Number m	3	4	5	6
Posterior Probability	0.046	0.891	0.062	0.002

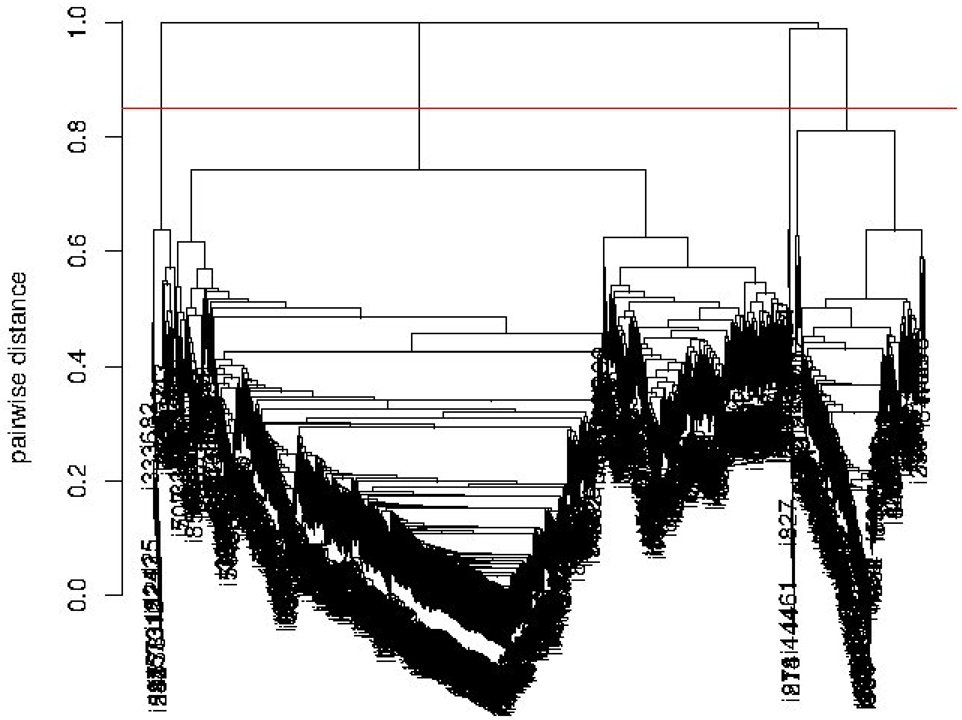


Figure 2.16: Dendrogram for Hierarchical Clustering Algorithm with Complete Linkage, based on $\log E2$ hormone data from SWAN.

We also obtained clustering results for women using the mutual distance defined in Equation (2.16). Figure 2.16 is the corresponding dendrogram for the hierarchical clustering model with complete linkage applied to the E2 data, where the 4 clusters were seen to be clearly separated. The spaghetti plots in Figure 2.17 display the women's $\log E2$ data in each of the 4 clusters. It is clear that the mean trajectory trends showed in the four plots are appreciably different from each other, and the women in the same cluster have similar trends. The model appears to cluster women appropriately based on their E2 profiles.

2.7 Model Expansion and Future Work

In this section, we discuss the potential future work that could be extended from our clustering model. In the first subsection, we consider different covariance structures to account for

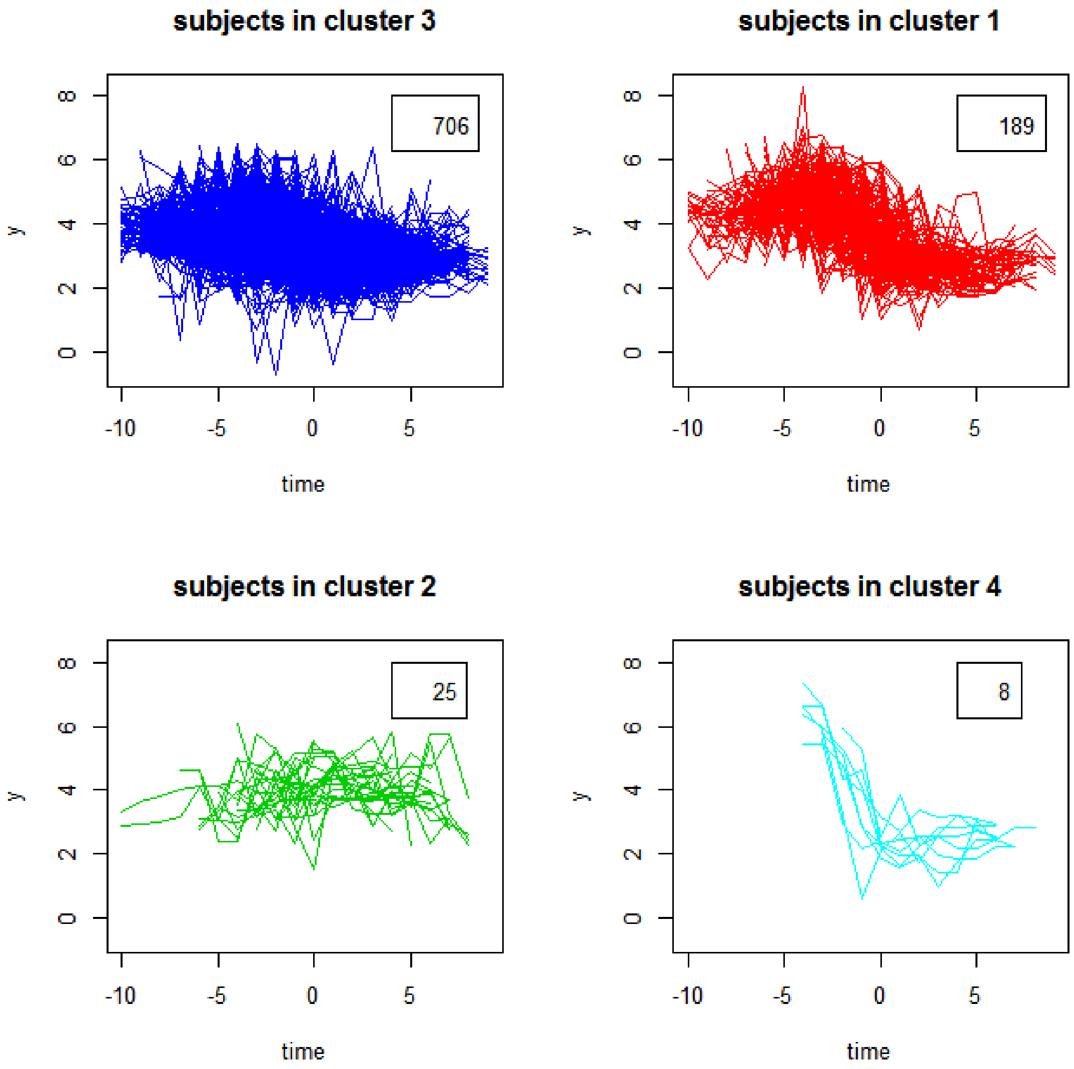


Figure 2.17: Spagetti plots for the women in each of the four clusters obtained based on their log $E2$ profiles. The plots are ordered according to cluster size instead of cluster label.

the individual variation instead of mixed effects. In the second, we consider a more flexible model that allows the order of the basis functions to change for different clusters.

2.7.1 Covariance Structure

A major positive feature of model (2.2) is the inclusion of mixed effects which are used to account for individual variation within clusters. The inclusion appreciably improves clustering result compared with model (2.1). However, if we compare the mathematical form of model (2.2) with model (2.1), they are not very different. The only difference is in the covariance structure for y_i. In model (2.2), the covariance has the form $\tau_e^{-1} W_i = I_{r_i} + X_i \Gamma^{-1} X_i^T$.

We can also use other covariance structures for the model. They may not have the same clear statistical meaning as random-effect terms, which allow for variation in individual trajectories. But it does account for the correlation between observations from the same individual. For example, we could specify a covariance structure with compound symmetry for y_i as follows:

$$Cov(y_i) = \tau^{-1} \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

where $var(y_{ij}) = \tau^{-1}$ and $cov(y_{ij}, y_{ij'}) = \rho\tau^{-1}, j \neq j'$. That is equivalent to having a mixed-effect model with intercept only.

There are many choices for the correlation structure. Autoregressive and continuous autoregressive are another two possible correlation structures. In addition, we can also use stochastic processes to model the serial correlation among the observations from the same individual, provided the process could be marginalized analytically.

2.7.2 Changing the Degree of the Basis Function

It is important to select an appropriate set of basis functions for cluster analysis. We have chosen the basis functions up to cubic terms ($p = 3$) for our model. However, this may not be the ideal choice.

We illustrate using clustering results for the $log(E_2)$ profiles from the SWAN data analysis as an example. In Figure 2.18, we show the four clusters of $log(E_2)$ profiles and the cluster mean trajectories (plotted in black). Even though the trends appear to be well described, the curve fitting could still be improved. Take cluster 3 (blue) for example, the mean trajectory indicates that $log(E_2)$ reaches its maximum around 6 six years before FMP and has an increasing trend after year 5. But in fact, the actual spaghetti plots indicate that the maximum $log(E_2)$ value is reached later than the estimate (about 3 or 4 years after FMP), and the $log(E_2)$ trend after FMP is almost flat instead of increasing.

This inaccurate curve fitting is caused by the limited number of basis functions. In this analysis, We only have 4 polynomials by choosing $p = 3$. The trajectory shape that is fitted with a linear combination of the 4 basis functions is highly restricted. For cluster 3, we need more basis functions to fit the $log(E_2)$ trend better.

On the other hand, the E2 profile is not equally complicated for each cluster. For example, the E2 profile of cluster 2 is almost flat, and a linear curve is good enough to describe the developmental trend. In this case, we do not want to use quadratic and higher order terms to model its trend. Therefore, a changing p for different clusters is desired to model cluster mean trajectories with different complexity at the same time.

We propose the limiting version of the model below to allow for varying p for different

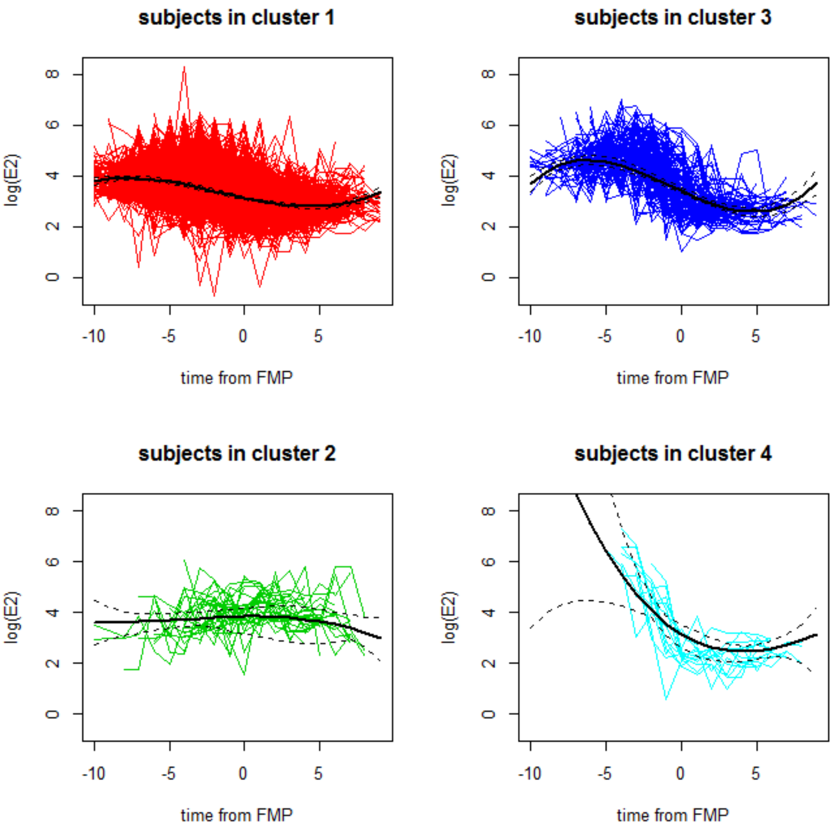


Figure 2.18: Spagetti plots of the women's log $E2$ data in each of the four clusters together with the estimated cluster mean trajectories.

clusters, which is a revision of model (2.10), namely:

$$
\begin{aligned}
y_i|s_i, \boldsymbol{\beta} &\sim N_{r_i}(X_i(p_{s_i})\beta_{s_i}, \tau_e^{-1}W_i) \\
s_i|\pi &\sim Multinomial(\pi_1, \ldots, \pi_L) \\
\pi_1, \ldots, \pi_L &\sim Dirichlet(\frac{\alpha}{L}, \ldots, \frac{\alpha}{L}) \\
\{\beta_{s_i}, p_{s_i}\} &\sim G_0 \\
G_0 &= N_{p+1}(\mu_\beta, \Xi_\beta^{-1}) \times Uniform(p_{min}, p_{max})
\end{aligned}
$$

In this model, we assign a DP to the joint distribution for (β, p) so that each cluster corresponds to a (β, p) pair. The base distribution we choose for p is a discrete uniform distribution. Based on some expert information about the trajectory, we select upper and lower boundaries, p_{min} and p_{max}, respectively, which are the lowest and highest order we would use in the model.

We still use the Gibbs sampler to draw posterior samples for the model parameters. The sampling algorithms are very similar to those for model (2.10) except that we need to sample p_{s_i} from its full conditional with this model. We use reversible-jump MCMC to sample it since the dimension of parameter space changes with p_{s_i}. We introduce the algorithm in Chapter 5 and omit details here.

Unfortunately, MCMC convergence was an issue. We believe the failure might be caused by too much flexibility in the model. We also could not rule out the possibility of a fatal error in the R coding. Our future work will focus on resolving the issue.

2.8 Conclusions

We have introduced a Bayesian semi-parametric model using the DP for the purpose of clustering individuals based on longitudinal data. Unlike other approaches, we integrate mixed effects into the clustering model to account for individual variation resulting in improved clustering performance. Furthermore, the method works with missing and unbalanced data. In a very small simulation study, we have demonstrated the validity of the model using both balanced and unbalanced data, and have shown the superiority of our model over others. The application of our method to longitudinal hormone data from SWAN successfully identified distinctive $log(E_2)$ developmental patterns.

Chapter 3

Label Switching in Bayesian Nonparametric Models

3.1 Introduction

Label Switching is a well-known issue arising in Bayesian analysis using mixture models. It describes the the invariance of the likelihood under relabelling of the mixture components, which is caused by non-identifiability of the mixture components. This leads to intractable inference on the mixture component level. It is inappropriate to make inferences about component-specific parameters from the MCMC samples using ergodic averaging of iterates. This label switching appears in Bayesian non-parametric models with the Dirichlet Process since it involves infinite mixtures. In chapter 2, we mentioned that label switching causes difficulty in inference for the cluster mean trajectory coefficients β_s. Due to the uncertain number of mixture components, there are additional label related issues in Dirichlet Process Mixture models (DPM). In this chapter, we discuss these issues for DPM models, and propose a method of dealing with them.

The structure of the chapter is as follows. In section 2, we describe the Label Switching issue in detail and focus on its special characteristics in DPM models. We introduce notation and illustrate the issue using the Galaxy Data (Roeder 1990 [32]), which consist of the velocities of distant galaxies diverging from our own, and were sampled from six conic sections of Corona Borealis. In section 3, we summarize several common relabeling strategies to remove label switching, and specifically discuss their applicability in non-parametric models. Then we propose our method in section 4, and demonstrate its validity and advantages.

3.2 Description of Label Switching

Let $\mathbf{y} = \{y_i : i = 1, \ldots, n\}$ denote a collection of independent responses. We model the distribution from which y_i is drawn as a finite K-component mixture of distributions;

$$p(y_i|\Theta) \;=\; \pi_1\, f(y_i|\phi_1, \omega) + \ldots + \pi_K\, f(y_i|\phi_K, \omega)$$

where $f(y_i|\phi, \omega)$ is a parametric probability density function. Define $\Theta = \{\boldsymbol{\pi}, \boldsymbol{\phi}, \omega\}$, where ω is a collection of global parameters, and where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_K\}$ are component-specific parameters. The components of mixing probabilities $\boldsymbol{\pi}$ sum to 1, and $\boldsymbol{\phi}$ is a vector or a list of size K, with the k^{th} element being the parameter or the vector of parameters for mixture component k. The likelihood function is

$$L(\Theta) \;=\; \prod_{i=1}^{n} \left[\pi_1\, f(y_i|\phi_1, \omega) + \ldots + \pi_K\, f(y_i|\phi_K, \omega)\right] \tag{3.1}$$

Label switching results from exchangeability of the component-specific parameters. For a permutation ν of $(1, \ldots, K)$, $\nu = (\nu(1), \ldots, \nu(K))$, we define the corresponding permutation of the parameter space by

$$\nu(\Theta) = \{(\pi_{\nu(1)}, \ldots, \pi_{\nu(K)}), (\phi_{\nu(1)}, \ldots, \phi_{\nu(K)}), \omega\}$$

Then $L(\Theta) = L(\nu(\Theta))$ for any permutation ν. Therefore, if $\hat{\Theta}$ is the MLE, $\nu(\hat{\Theta})$ is also the MLE for any ν out of the $K!$ permutations. That means the labels we assign to $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ are not identifiable unless we put additional restrictions on the model.

A hierarchical form of the mixture model is helpful for understanding the issue. Let $s_i = j$ if subject i in the data belongs to mixture component j. If $\mathbf{s} = \{s_i : i = 1, \ldots, n\}$ were known, the model would be identifiable. The vector $\mathbf{s}$ is regarded as latent, and the mixture model can be re-expressed as,

$$\begin{aligned} y_i|s_i, \boldsymbol{\phi}, \omega &\sim F(\phi_{s_i}, \omega) \\ s_i|\boldsymbol{\pi} &\sim Multinomial(\pi_1, \ldots, \pi_K) \end{aligned} \tag{3.2}$$

where $F(\phi_{s_i}, \omega)$ is the CDF corresponding to pdf $f(y_i|\phi_{s_i}, \omega)$ above. Note the label $\mathbf{s}$ is a vector of latent variables indicating cluster membership, which have no statistical meaning.

In Bayesian analysis, we often assign a symmetric prior distribution to $\boldsymbol{\pi}$, which is an exchangeable prior for all the elements of $\boldsymbol{\pi}$. Thus the posterior distribution for $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ is invariant to the permutation of the labels.

When $\mathbf{s}$ is observed, we make inferences about the component-specific parameters $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ using MCMC samples with the known labels. However, when $\mathbf{s}$ is latent, there is no guarantee that each label consistently marks the same component through the MCMC iterations. As a result, it is not appropriate to make inferences about the component-specific parameters using MCMC samples.

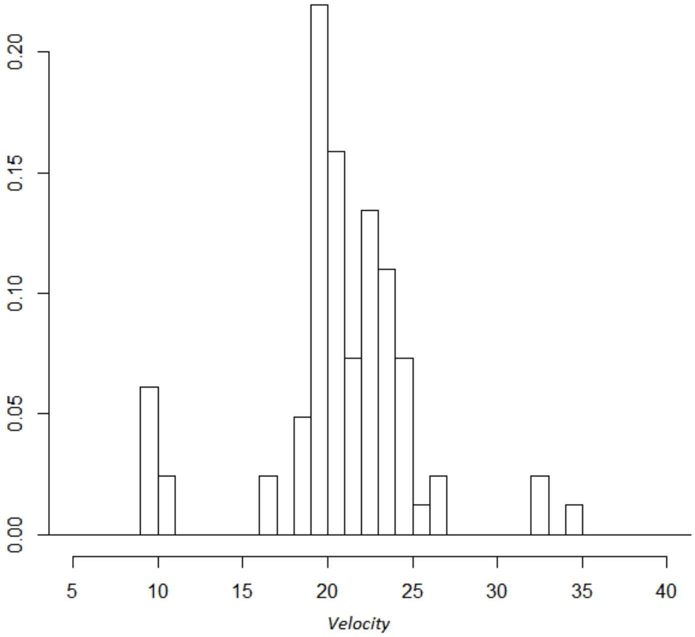


Figure 3.1: Histogram of the Galaxy Data

We use results presented in Stephens (2000) [35] to illustrate the problem. The Galaxy Data were also used in his example. Figure (3.1) gives a histogram of the data. Figure 2 in the paper illustrates the effect of label switching in the raw output of the Gibbs sampler when fitting a finite mixture model with 6 normal distributions fitted to the Galaxy Data. Part (a) of the figure gives the trace plots for the 6 component means using the labels. Observe that the component means are all bouncing up and down, and that is because each label did not mark the same component throughout the MCMC iterations. As a result, the marginal posterior density of component means are nonsensical as shown in Part (b).

3.3 Relabelling Algorithms for Finite Mixture Models

In order to make inferences for component-specific parameters, many relabeling algorithms have been proposed to process MCMC output to obtain the posterior samples for each mixture component in mixture models.

A common response to the label switching issue is to impose an identifiability constraint (IC) on the parameter space that can be satisfied by only one permutation of a parameter on the component level in $\boldsymbol{\phi}$. This breaks the symmetry of the prior (and thus of the posterior) distribution of the parameters and so might seem to resolve the label switching issue. Concerns about imposing an identifiability constraint have been discussed in Celeux, Hurn and Robert (2000) [**?**] and Stephens (1997 [36], 2000 [35]). The choice of constraint may be artificial, unless it arises from genuine knowledge or belief about the model. Moreover, IC does not apply to infinite mixture models, where the number of components is random. It is impossible to match the component-specific parameters between two MCMC iterations when they have a different number of components.

Stephens (1997 [36], 2000 [35]) proposed a relabeling algorithm to post process the MCMC output, in which a loss function was defined based on KullbackLiebler divergence for selecting permutations at each MCMC iteration. Stephens illustrated with examples that his method has better performance than IC in many scenarios. But this method still does not apply to infinite mixture models, because the algorithm uses the same idea as IC to "line up" the mixture components by switching the labels so that each label would consistently identify the same mixture component throughout all the MCMC iterations.

We use the Galaxy Data example to show that alignment of mixture components is precarious even for finite mixture models. We applied a finite Gaussian mixture model using Equation (3.2) with $K = 5$ components. Figure (3.2) shows two approximate predictive density iterates, plotted with different colors. In order to specify components, we use numbers 1-5 as labels in both density curves. We can see that it is impossible to "line up" the components with any label switching algorithm, because component 2 in red does not exist in blue, and component 4 in red looks like the combination of 3 and 4 in blue.

An assumption for the cited label switching algorithms ([36], [35], [43]) that align components is that the mixture components are consistent throughout MCMC iterations. However, this

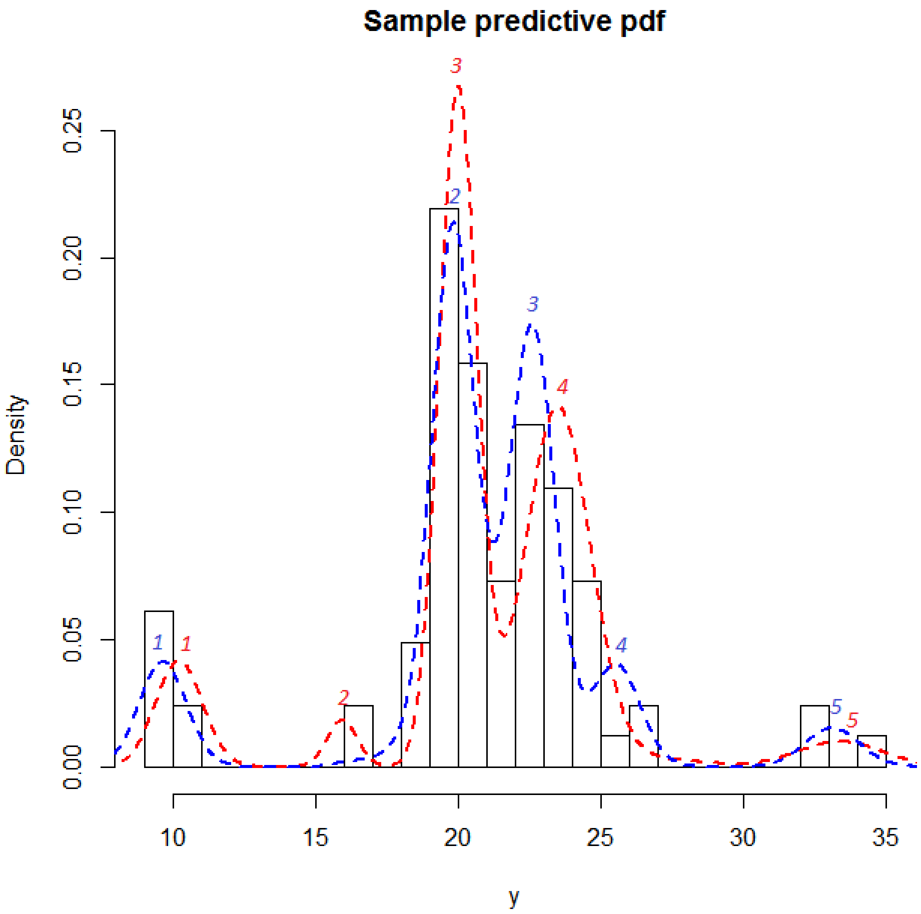


Figure 3.2: pdf iterates for the Galaxy Data

assumption fails unless components are defined beforehand. Absent this definition, components are determined by the random partitioning of subjects, which varies from one iteration to another. With different partitions of subjects, the components formed by the partitions might have different statistical meaning. There is no guarantee that the components obtained in each iteration are consistent, either in number or meaning. The five components in red are different from those in blue in Figure (3.2). It is impossible to align the components in these two iterations, or to attach meaning to one partition versus the other.

3.4 Partitioning of Subjects

The partitioning of subjects leads to the components constructed in the Markov Chain. The two iterations in Figure (3.2) indicate two very different partitionings of the subjects. If we use numbers $1, \ldots, n$ to label the Galaxy Data with increasing order, the partitioning is $\{\{1, \ldots, 7\}, \{8, 9\}, \{10, \ldots, 44\}, \{45, \ldots, 79\}, \{80, 81, 82\}\}$ for the red iteration, and $\{\{1, \ldots, 7, 8, 9\}, \{10, \ldots, 44\}, \{45, \ldots, 76\}, \{77, 78, 79\}, \{80, 81, 82\}\}$ for the blue. We can see that the constructed components are different in the two iterations due to the partitioning of the subjects.

However, different partitionings may represent the same structure of components. Consider another iteration from the MCMC output for the Galaxy Data, which we call the green iteration for convenience (even though it is not plotted in Figure (3.2)). Its partitioning is $\{\{1, \ldots, 7\}, \{8, 9\}, \{10, \ldots, 44, 45\}, \{46, \ldots, 79\}, \{80, 81, 82\}\}$, which is slightly different from the red. However, it indicates the same components as the red iteration.

Therefore, there are two different kinds of randomness to consider: (i) One is random component membership on the individual level. A subject could be randomly assigned into any mixture component with positive probability. Comparing the red and green iterations, subject No.45 changes its component membership from 4 to 3. During the MCMC iterations, an individual might change its component membership frequently, especially for the subjects who are near the boundary of two components. In such cases, the individuals do not affect the consistency of components. The Markov Chain retains the same set of components. (ii) The second is randomness on the component level. An example is the change from red to blue iteration. All subjects in component 2 in red merged into component 3 to form one component in blue. In such cases, the component structure changes.

In order to estimate the component-specific parameters, we must know which component configuration to select for making inferences and which iterations correspond to the same configuration. The target configuration would certainly be the one with highest posterior probability, or with at least high posterior probability. Our aim is to find iterations that consistently correspond to high likelihood configurations and use them as posterior samples for inference. Yao and Linsay (2009) [43] proposed an algorithm based on highest posterior density in post processing. They used the parameter values at each iteration as initial values for an EM algorithm applied to the finite mixture model to find which mode each iteration would converge to. Then the iterations converging to the global mode would be chosen for inference. They also show the advantages of the algorithm by comparing with relabeling algorithms including IC and Stephen's method.

One drawback of their algorithm is that it only applies to finite mixture models, which limits its usage. In finite mixture models, the number of clusters, K, is pre-determined and the partition of subjects is generally consistent throughout MCMC iterations. In Figure (3.2), if the Markov Chain jumps from red to blue, it requires all subjects in component 2 to change component membership to component 3, and component 4 splitting into two components, simultaneously. This will not happen frequently, especially if the chain has found the high likelihood configuration.

With infinite mixture models, there is more flexibility in the partitioning of subjects, since partitioning is random, as is the number of clusters formed. It is common that partitioning of subjects changes frequently on the component level between MCMC iterations. In the next section, we will focus on the label related issues in infinite mixture models.

3.4.1 DPM Models and Label Switching

The DPM model has been broadly used for clustering. We analyzed SWAN data in chapter 2, where our goal was to cluster women based on their hormone profiles. Using the notation from the previous section, the simplest form of DPM is,

$$
\begin{aligned}
y_i|\phi_i, \omega &\sim F_{\phi_i,\omega} \\
\phi_i|G &\overset{\perp}{\sim} G \quad i = 1, \ldots, n \\
G &\sim DP(\alpha, G_0) \\
\omega &\sim P(w)
\end{aligned}
\tag{3.3}
$$

where α is the concentration parameter and G_0 is the base distribution of the DP. The global parameters ω are assigned a prior $P(w)$.

Label Switching arises naturally in Bayesian non-parametric models. As mentioned in the previous chapter, a finite mixture model is asymptotically equivalent to the DPM model

Table 3.1: Posterior Distribution of the number of mixture components

Cluster Number K	3	4	5	6	7	8	9	10	10+
Posterior Probability	0.05	0.09	0.16	0.21	0.20	0.14	0.08	0.04	0.03

(3.3) by taking the limit as $L \to \infty$. This equivalent model was mentioned in Neal (2000),

$$\begin{aligned} y_i|s_i, \boldsymbol{\phi}, \omega &\overset{\perp}{\sim} F_{\phi_{s_i},\omega} \\ s_i|\pi &\overset{\perp}{\sim} Multinomial(\pi_1, \ldots, \pi_L) \\ \phi_{s_i} &\sim G_0 \\ (\pi_1, \ldots, \pi_L) &\sim Dirichlet(\frac{\alpha}{L}, \ldots, \frac{\alpha}{L}) \end{aligned} \tag{3.4}$$

The label switching issue applies to the DPM model above since it could be regarded as a mixture model. Moreover, since the actual number of clusters, K, is not pre-determined in the limit as $L \to \infty$, there are additional label related issues caused by inconsistent partitioning of the subjects, which will be discussed in the next section. Before we discuss these issues, we would like to point out an important difference between LDA models in Equation (3.4) and general finite mixture models.

In addition, in order to illustrate the label related issues in non-parametric models, we continue with the Galaxy Data as an example. We applied model (3.3) to the data and ran 20000 iterations with 10000 burn-in to obtain posterior samples. Figure (3.3) shows the predictive density for the data, and Table (3.1) gives posterior probabilities for the actual number of mixture components.

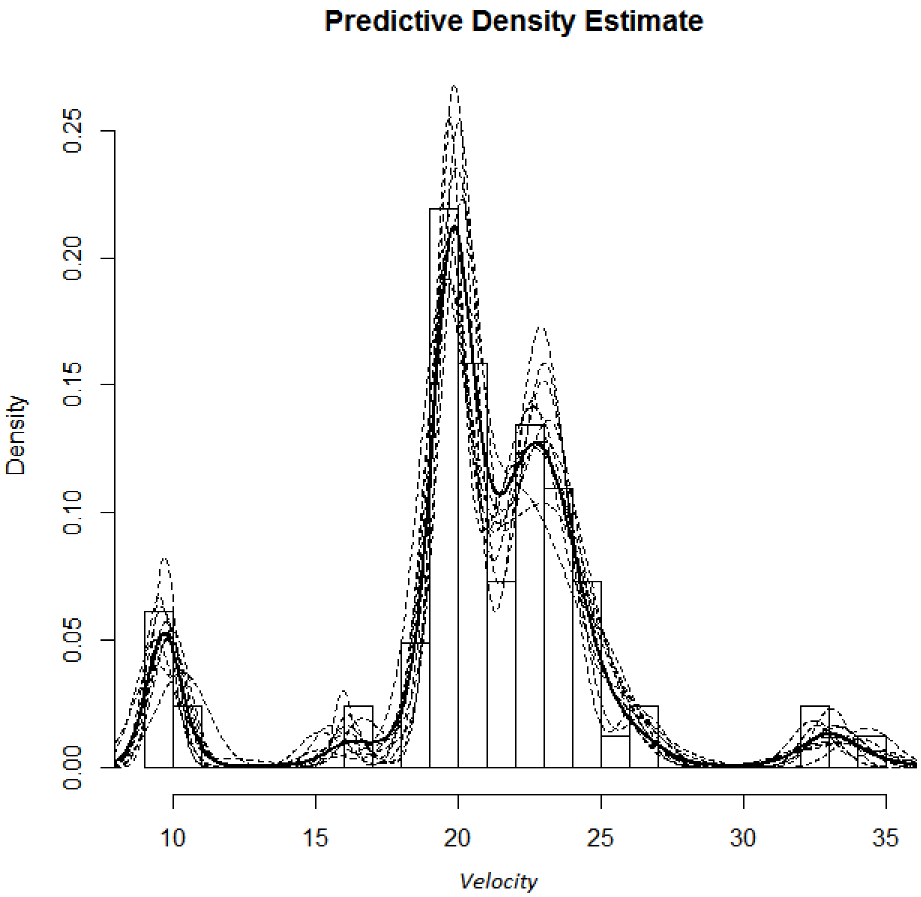


Figure 3.3: Histogram of the Galaxy Data, overlaid with the predictive density esimate based on a DPM model plotted with solid line. The dashed lines are 10 samples of the pdf from MCMC output.

3.4.2 Label Related Issues for DPM Models

As mentioned above, the number of clusters K in a DPM is a random variable and changes from one MCMC iteration to another. This flexibility makes it essential to consider the randomness on the component level, since it would happen frequently that one component splits into two or two components merge into one from one MCMC iteration to the next. This randomness would cause issues in making inferences for component-specific parameters. In this section, we examine the Gibbs sampler used for DPM models and discuss label related issues.

If we understand the DP sampling procedure using the Chinese Restaurant Process analogy, there could always be new tables (components) created at each iteration, and existing tables (components) removed from the chain. Taking model (3.4) for example, with Gibbs sampling, s_i is drawn from its full conditional distribution:

$$P(s_i = s|s_{-i}, y_i, \boldsymbol{\phi}, \omega) \propto \begin{cases} \frac{n_{-i,s}}{n-1+\alpha}\, f(y_i|\phi_s, \omega) & \text{if } s = s_j \text{ for some } j \neq i \\ \frac{\alpha}{n-1+\alpha} \int f(y_i|\phi, \omega)\, dG_0(\phi) & \text{if } s \text{ corresponds to a new label} \end{cases}$$

where $n_{-i,s}$ is the number of subjects associated with mixture component s except subject i, and $\sum_s n_{-i,s} = n-1$. Since K is random, there are always existing components removed from the chain and new components created. At each MCMC iteration, a component with only one subject, say individual i, would be removed from the chain in the next iteration, since s_i could be updated and moved to another existing table (component) with the probability shown above. On the other hand, when each s_i is sampled, there is probability proportional to $\frac{\alpha}{n-1+\alpha} \int f(y_i|\phi, \omega)\, dG_0(\phi)$ of creating a new mixture component for individual i. When a new component is created, we use a new value s to label it. Note the value s is uniquely used to mark this component. If the component is removed from the chain later, we also remove its label value s and do not recycle it for other components.

Now we require a definition of “cluster” and “mixture component” in our context, because we assign them different statistical meaning in DPM models. A cluster is a collection of subjects who have similar characteristics based on some criterion, and it is defined by the partitioning of subjects. A mixture component is a component distribution in the mixture model, and it is marked by a unique label. We separate the two concepts since the number of mixture components is allowed to be more than K in a DPM model. As discussed above, there are always existing components dropped from the chain and new components created. If we gave each new component a unique label, eventually the number of labels we obtained from the Markov Chain would be much larger than the number of clusters K. Table (3.1) shows the modal cluster number K is six with the highest posterior probability for the Galaxy Data. However, in the 10000 posterior samples after burn-in, there were 9262 mixture component labels in the chain.

That leads to the first label related issue in DPM models, which is that a cluster could be associated with multiple labels. Figure (3.4) shows the trace plot of cluster mean μ's for the Galaxy Data. Unlike Figure (**??**), we are not able to plot the trace of each component label, since there were too many (9262) mixture components in the Markov Chain after burn-in iterations. In Figure (3.4), we show the μ trace plots for all components in the same graph, and use different colors to indicate different components (labels). We can see that no component survives through all MCMC iterations. It is very common for an existing component to be substituted by a new component at some iteration. Taking the cluster with the lowest μ value for example, there are 7 subjects belonging to it. (Note we did not know which subjects belong to the cluster beforehand.). Table (3.2) shows the labels of the 7 subjects from iteration 7057 to 7062 in the MCMC output. We can see the cluster was labeled as component 6431 at first. At iteration 7059, a new component with label 6474 was generated, and it substituted for component 6431 to represent the cluster in the following two iterations.

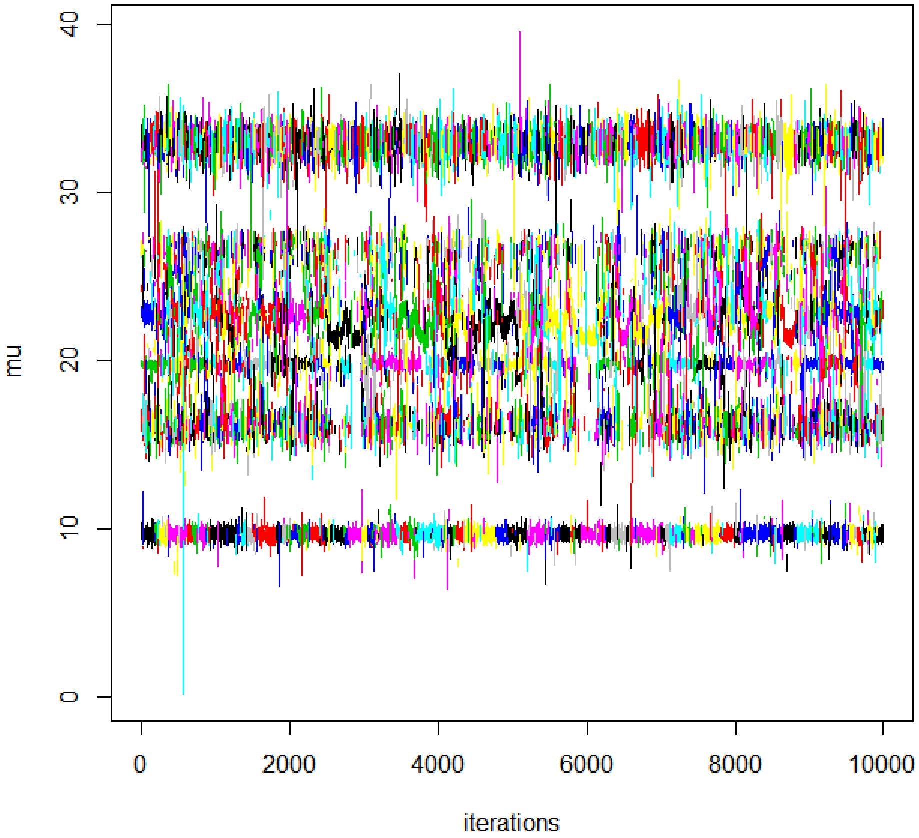


Figure 3.4: Trace plot of μ of all components from MCMC output. The color is used to indicate different component labels.

This issue is caused by non-identifiability of the mixture components, which is the same as label switching. But in DPM models, the change of a component's label does not require switching with another component. It could be substituted by a new component since K is not fixed. In Figure (3.4), the color of the trace for each cluster changed many times. Each change means an existing component for the cluster was substituted by a new component. So the cluster was represented by multiple components throughout the MCMC output. Apparently, that would lead to problematic inference about the component-specific parameters with ergodic averaging. For easy reference, we name this issue "label substitution".

A second issue is that two or more clusters might merge into one mixture component in some iterations, which we call "label merging". Figure (3.4) and Table (3.1) show $K = 6$ is appropriate for the Galaxy Data. But K is random, and the table also shows there are only

Table 3.2: Labels of the 7 subjects in Cluster 1 from Iteration 7057 to 7062

	Iteration					
Subject	7057	7058	7059	7060	7061	7062
sub 1	6431	6431	6431	6431	6474	6474
sub 2	6431	6431	6431	6431	6474	6474
sub 3	6431	6431	6431	6431	6474	6474
sub 4	6431	6431	6431	6474	6474	6474
sub 5	6431	6431	6474	6474	6474	6474
sub 6	6431	6431	6431	6431	6474	6474
sub 7	6431	6431	6475	6431	6474	6474

five or less mixture components in about 30% of MCMC iterations of the Markov Chain. In these iterations, two or more of the 6 clusters have merged. As a result, one component (label) would represent multiple clusters. We can see that the 4 clusters in the middle merged for a while around iteration 6000 in Figure (3.4).

A third issue is that a cluster might split into two or more mixture components in some iterations, which we call “label splitting”. Both “label merging” and “label splitting” issues are caused by the randomness on the mixture component level. The two issues appear when the Markov chain jumps.

A final issue is that a mixture component might not represent any existing cluster. In the Galaxy Data example, 9262 mixture components appeared in the MCMC output. Among these components, most of them did not exist for very long, because they were dropped from the Markov Chain quickly. We call them “transient components”. For example, component 6475 in Table (3.2) is a transient component, because it was created at iteration 7059 and get dropped in the next iteration. These components have no cluster information value and should be excluded from post-posterior inferences.

The problems discussed above are common for non-parametric models, and they all affect the inference of the component-specific parameters using ergodic averaging.

3.5 Our Solution to Construct Post-Posterior Samples

In this section, we propose a method to deal with the issues mentioned above, which works well for both finite mixture models and DPM models (infinite mixture models).

In order to construct posterior samples for post processing, we develop a method constructing what we call a reference partition of subjects. Since the clusters are formed by subjects, we could identify a reference partition of subjects to have the highest, or modal, posterior probability. This would surely not be easy and also, we might expect that there would be many possible partitions with high posterior probability. In the next section, we will introduce our method of finding a particular reference partition that appears to be quite useful.

We then use the reference partition to deal with label switching. As mentioned above, clusters are formed by the partitioning of subjects. Taking the Galaxy Data in Figure (3.4), for example, it appears that there are six distinctive clusters, which are defined by their μ values, and fundamentally by the subjects that constitute the corresponding clusters. Therefore, our idea is, if we can track the subjects associated with each cluster throughout the MCMC output instead of tracking the labels, we will be able to obtain the posterior samples for each cluster without considering how it is labeled.

3.5.1 Partitioning of Subjects

In order to determine a reference partition of subjects, we extract information from MCMC output. First, the number of partitions (clusters) can be informed with its posterior distribution. For example, we see the posterior distribution of the number of clusters K is concentrated around six for Galaxy Data since number 5 to 8 all have moderate posterior probability based on Table (3.1). In addition, the choice seems to be plausible from the trace plot for μ's in Figure (3.4).

In the MCMC output, we do not know how the label changes from one iteration to another for each cluster. But we do know how the subjects are clustered in any iteration. Subjects with the same label are associated with the same component, and vice versa. Therefore, we know the posterior probability of any two subjects being associated with the same component, which is the proportion of the MCMC iterations where they have the same component label. This probability could be used as a distance measure between any two subjects. Medvedovic and Sivaganesan (2002) proposed a mutual distance between any two subjects i and j based on the idea as follows:

$$r_{ij} = \frac{\sum_{iter=1}^{n_{mc}} I(s_i^{iter} \neq s_j^{iter})}{n_{mc}} \tag{3.5}$$

where n_{mc} is the number of MCMC iterations after burn-in. Then $r_{ij} = 0$ means the two subjects are clustered together throughout the MCMC iterations, and $r_{ij} = 1$ means the two are always associated with different components.

Based on the mutual distance obtained above, we can partition subjects using a hierachical clustering method that has been used by Biglow and Dunson (2005) [2]. We give algorithm details in the Appendix. Figure (3.5) is the dendrogram produced by the algorithm, using the Galaxy Data, which illustrates the arrangement of the clusters. The partition of the subjects can be obtained by choosing an appropriate cutoff value for the mutual distance. For the Galaxy Data, it seems to be reasonable to pick 0.6 as the cutoff value resulting in five clusters. So we choose $K = 5$ for the analysis, and the allocation of the subjects is shown in Figure (3.6). If we keep using numbers $1, \ldots, n$ to label the Galaxy Data with increasing order, the reference partition is $\{\{1, \ldots, 7\}, \{8, 9\}, \{10, \ldots, 44\}, \{45, \ldots, 79\}, \{80, 81, 82\}\}$.

However, finding the reference partition is not our ultimate goal in the analysis. It is merely a intermediate step used as a reference cluster configuration in order to construct posterior

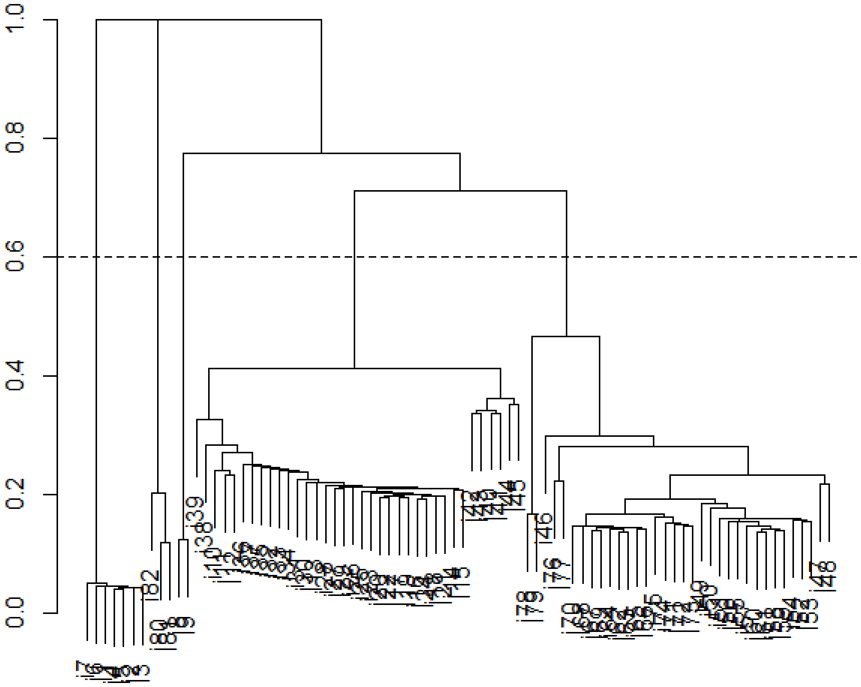


Figure 3.5: Dendrogram made with a hierarchical clustering method for the Galaxy Data

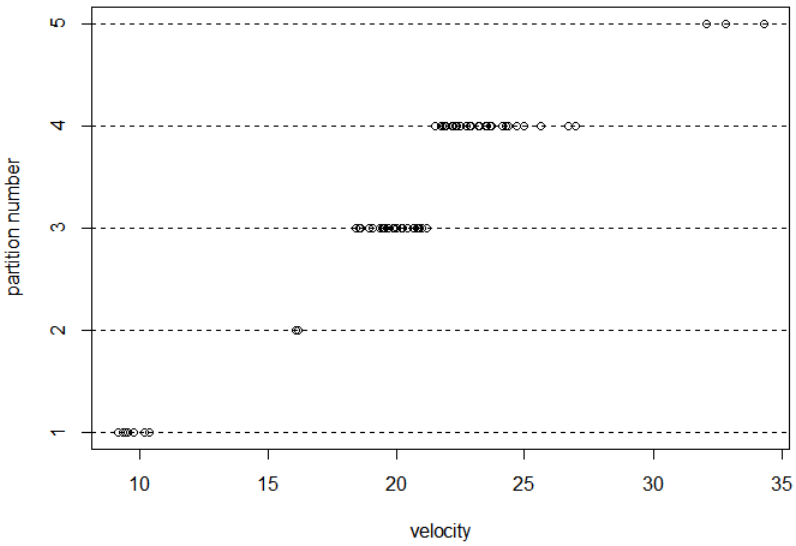


Figure 3.6: Reference partition of the Galaxy Data into five components using the Hierarchical Clustering method

samples for inference. In the next section, we introduce our algorithm to obtain the posterior samples for the parameters on the cluster level.

3.5.2 Posterior samples for component-specific parameters

In this subsection, we discuss how to decide which iterates of the MCMC output are consistent with the selected reference partition. We compare the observed partition at each MCMC iteration with the reference partition to select posterior samples as discussed below. Iterations with the same cluster configuration as the reference will be used for posterior inference.

The question is, how do we judge whether a given partition has the "same" cluster configuration as the reference or not. We have mentioned that the partitions in MCMC iterations are more or less different from each other, and of course they will be more or less different from reference partition as well. In the Galaxy Data example with the reference partition used there, the question is which MCMC iterations have clusters "similar" to those in the reference partition.

At this point, we need a definition for "existence of a cluster" within a partition, so that we know whether any of the five clusters exists in each iteration as in the reference configuration for the Galaxy example. At iteration 7060 in Table (3.2), the cluster was split into two with 2 subjects labeled with component 6474 and the other 5 labeled with 6431. In this case, it seems reasonable to think the cluster exists and component 6431 can represent the cluster because 5 out of the 7 subjects are labeled with it. But we do need a criterion to decide when a component could represent the cluster, because there is no clear boundary between randomness on component and individual levels. We adopted an ad-hoc criterion as follows: if the majority (greater than 50%) of subjects in a known reference cluster are grouped together in a MCMC iteration, the component corresponding to these subjects is regarded to be representative of the cluster at that iteration. Otherwise, we think the cluster does not exist due to “label splitting”, and the iteration should be removed from the posterior samples for the cluster.

We also need to remove iterations with “label merging”, in which a cluster merges with some other cluster(s) at some MCMC iterations to form a larger cluster. For example, in Figure (3.2) of the Galaxy Data example, clusters 3 and 4 in blue merged into cluster 4 in red. If we are interested in the partitioning represented by the blue iteration, in which clusters 3 and 4 are separated, then the red iteration should not be used as a posterior sample for making inferences. So these iterations should be removed from the posterior samples for the inference.

By comparing with the reference partition at each iteration, we naturally solve the “label switching” and “label substitution” issues. There is no need to consider the original labels from the MCMC output. We just need to consistently relabel the clusters that share 50% or more subjects within each cluster in the reference partition.

In addition, the method also eliminates the “transient components” from the analysis. Since we only select the components that correspond to the clusters existing in the reference partition at each iteration, the transient components would not be relabeled and naturally would be excluded from the post processing sample.

With all the label related issues resolved, we can obtain the posterior samples for each cluster using the post-processing algorithm below (Algorithm I). After processing, we select posterior samples that have the same cluster configuration as the reference partition and label each cluster consistently throughout MCMC iterations. We plot the reconstructed posterior samples in Figure 3.7.

Algorithm I:

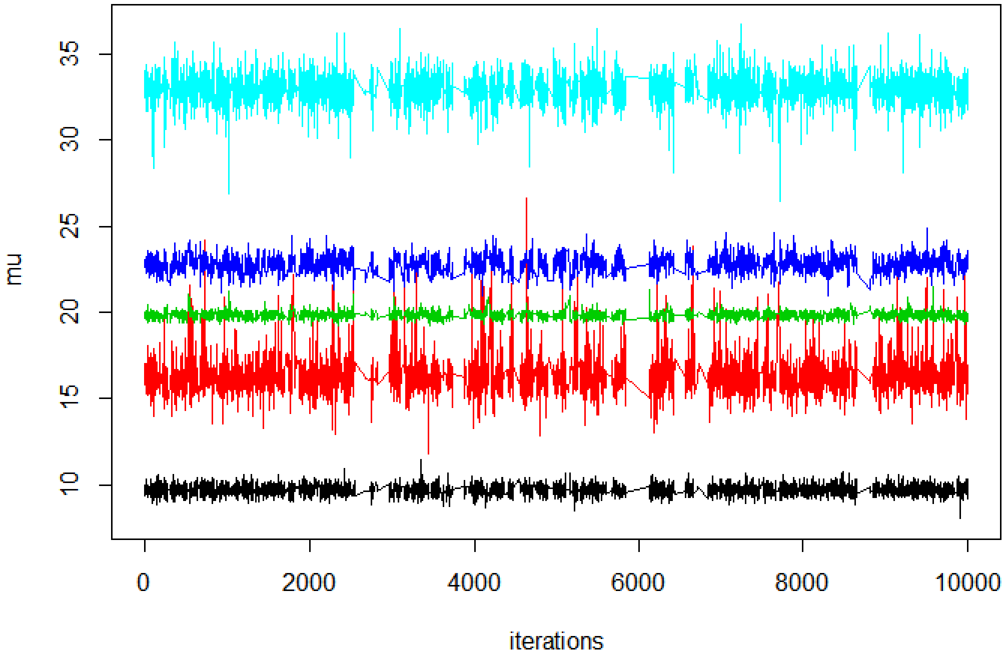


Figure 3.7: Trace plot of μ for six clusters after post-processing using Algorithm I. Each cluster is indicated with a unique color.

1. Begin with MCMC output, which has all the labeling information and parameters corresponding to each label.

2. Determine the number of clusters, K, using its posterior probability and trace plots of the parameters.

3. Calculate the mutual distance between subjects based on the labels, apply the hierarchical clustering method to identify the reference partition of the subjects and label the K clusters with numbers $1, \dots, K$. Use n_k, $k = 1, \dots, K$ to denote the number of subjects in each cluster. This creates the reference cluster.

4. At each iteration, determine if every member of cluster k in the reference configuration belongs to some cluster in the current iteration. Then if this holds for all K clusters in the reference configuration, keep the current iterate for further scrutiny. Otherwise, remove the current iteration from consideration. Thus if the current iterate is subject

to additional scrutiny, proceed as follows:

- For each cluster $k, k = 1, \ldots, K$ in the reference partition, determine whether 50% or more subjects in the reference cluster k are grouped together in a component of the current iteration. If not, this cluster k does not exist in the current iteration according to our criterion. Remove the current iteration from the posterior sample and proceed to the next iteration.
- If yes, locate the component, and check whether it has been relabeled. If yes, this cluster k merges together with some other cluster in the current iteration. Remove the current iteration and proceed to the next iteration.
- If not, cluster k exists in the current iteration, and it is not merged with another cluster. Relabel the located component with number k.
- Repeat the 3 substeps above for $k = 1, \ldots, K$ in each iteration.

5. At the end, the posterior samples for each cluster will be consistently labeled with k for $k = 1, \ldots, K$.

We illustrate the algorithm with a toy example. Assume the data has 10 observations labeled with $\{1, \ldots, 10\}$, and we obtain the reference partition below:

$$\big\{\{\text{cluster } k = 1 : \text{subjects } 1, 2, 3\}, \{k = 2 : 4, 5, 6\}, \{k = 3 : 7, 8, 9, 10\}\big\}$$

We then use two iterations in MCMC, which have partition $\big\{\{\text{label } 101 : 1, 2, 3\}, \{\text{label } 134 : 4, 5, 6\}, \{\text{label } 234 : 7\}, \{\text{label } 222 : 8, 9, 10\}\big\}$ and $\big\{\{\text{label } 501 : 1, 2, 3, 4, 5, 6\}, \{\text{label } 621 : 7, 8, 9, 10\}\big\}$, to show how the algorithm works. Note the component labels are different from the cluster numbers in the reference partition.

For the first iteration, we process it as follows:

- Start from $k = 1$, 50% or more of $\{\text{cluster } k = 1 : \text{subjects } 1, 2, 3\}$ are in the component with label 101. Relabel the component with k=1, so that we know it is a posterior sample for cluster 1.
- When $k = 2$, 50% or more of $\{k = 2 : 4, 5, 6\}$ are in the component with label 134. Relabel it with $k = 2$ since it has not been relabeled.
- When $k = 3$, we relabel component 222 with $k = 3$.
- Proceed to the next iteration.

We could see that component 234, which is a "transient component", is naturally dismissed from posterior inferences.

Considering the second iteration, we process it with the following steps:

- Start from $k = 1$, 50% or more of $\{\text{cluster } k = 1 : \text{subjects } 1, 2, 3\}$ are in the component with label 501. Relabel the component with $k = 1$.
- When $k = 2$, 50% or more of $\{k = 2 : 4, 5, 6\}$ are in the component with $k = 1$ (originally label 501). Since it has been relabeled, it means that clusters 1 and 2 in the reference partition are merged in this iteration. So we remove the iteration.
- Proceed to the next iteration.

The process for this iteration shows how we deal with "label merging" issue.

Robert (2010) [31] mentioned "When, given a target π, if an MCMC sampler that never visits more than 50% of the support of π, it can be argued that the sampler does not converge". Later we will use this concept as a method of validating the reference partition. In the Galaxy example, we kept 6082 iterations out 10000 for the posterior samples. In this case,

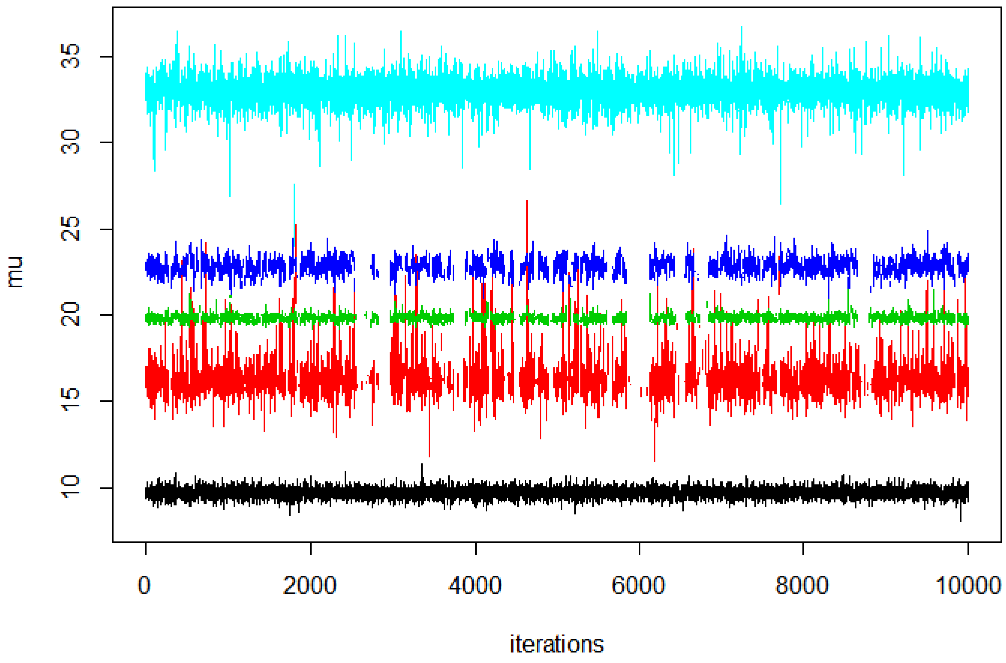


Figure 3.8: Trace plot of μ for the five clusters after post-processing using Algorithm II. Each cluster is indicated with a unique color.

we believe the Markov chain has converged and that reference partition is well chosen since more than 50% of the MCMC iterations share the same cluster structure.

In the trace plot Figure (3.4), we found the splitting and merging of components only happened within clusters 2, 3 and 4. Clusters 1 and 5 existed consistently and all the iterations could be used for inference for these two clusters. So we realized that the algorithm above was not the most efficient for selecting the posterior samples, since we removed many samples that could have been used for inference for clusters 1 and 5. Now suppose a cluster in the reference partition "exists" in an iteration. Then the component corresponding to it could be used as part of its posterior samples. With this idea, we updated the post-process algorithm as below (Algorithm II), and obtained the posterior samples plotted in Figure (3.8)

Algorithm II:

1. Begin with MCMC output that has all the labeling information and the parameters

corresponding to each label.

2. Determine the number of clusters, K, using its posterior probability and trace plots of the parameters.

3. Calculate the mutual distance between subjects based on the labels, and apply hierarchical clustering method to identify the ideal partition (reference partition) of the subjects. Use n_k, $k = 1, \dots, K$ to denote the number of subjects in each cluster. Thus create the reference partition.

4. Post-process each iteration t, $t = 1, \dots, N^{MC}$, of the MCMC output. At iteration t, we implement the following substeps:

    - For each cluster $k, k = 1, \dots, K$, of reference partition, determine whether there is a component having 50% or more of its subjects in cluster k. If yes, label the component with k and record the corresponding parameter values as ϕ_k^t. Otherwise, let ϕ_k^t be a missing value denoted by 'NA'. We then obtain a vector $\boldsymbol{\phi}^t = \{\phi_1^t, \dots, \phi_K^t\}$.
    - If any two non-missing values in $\boldsymbol{\phi}^t$ satisfying $\phi_s^t = \phi_{s'}^t$ for $s \neq s'$, set both ϕ_s^t and $\phi_{s'}^t$ to be 'NA', since the two clusters s and s' are merged together in the current iteration.

5. We eventually obtain the posterior sample for each cluster k: $\{\phi_k^t : t = 1, \dots, N^{MC}\}$, and use non-missing values of ϕ_k^t for posterior inferences in each cluster k.

Comparing Figures 3.7 and 3.8, we can see that Algorithm II efficiently retains more posterior samples for inference in clusters 1 and 5.

Table 3.3: Inference of cluster-specific parameters for the Galaxy Data

			95% PI	
Parameters	Mean	Median	lower	upper
μ_1	9.71	9.71	9.19	10.24
μ_2	16.34	16.20	14.83	19.29
μ_3	19.82	19.81	19.46	20.26
μ_4	22.89	22.90	21.99	23.73
μ_5	32.96	32.97	31.48	34.37
σ_1	0.65	0.62	0.41	1.06
σ_2	0.86	0.73	0.41	2.14
σ_3	0.71	0.69	0.47	1.05
σ_4	1.34	1.31	0.70	2.16
σ_5	1.01	0.92	0.52	2.07

3.6 Statistical Inference

The primary goal to deal with label related issues is to make inferences for the component-specific parameters. Table (3.3) shows inferences for component-specific parameters μ and σ.

We are also interested in how the subjects clustered. The reference partition for the subjects has given us some information, but it should not be over interpreted as discussed above. With the posterior samples, we are able to calculate the posterior probabilities that a subject i is associated with each cluster, $P(k_i|y_i, K = 5)$. Table (B.5) in the Appendix lists the posterior distribution for all 82 subjects in the Galaxy Data; $P(k_i = other|y_i, K = 5)$ is the probability that subject i is not associated with any of the six clusters, when it might be associated with a transient component.

3.7 Discussion

3.7.1 The Partitioning Method

As previously mentioned, there are multiple methods that could be used to obtain the reference partition. For finite mixture models, maximum a posteriori probability (MAP) estimate identifies a partition with largest posterior probability, which can be regarded as the reference partition. That method has been used by Robert (2008).

In DPM models, the MAP estimate may not work well due to the "transient components". Since the MAP estimate comes from one selected iteration of the Markov chain, it is possible that one or more transient components exist in that iteration. If the MAP estimate is used as the reference partition, it would be problematic to find posterior samples for the transient components. We thus prefer using the mutual distance defined by Medvedovic and Sivaganesan (2002) to identify the reference partition using the hierarchical clustering or K-medroid methods. The transient components would not appear in the reference partition since it is obtained with the clustering information of the whole Markov chain.

With the mutual distance r_{ij}, the two algorithms (hierarchical clustering and K-medroid) are both capable of identifying a reference partition. Generally we implement both and compare the two partitions to assure the choice of reference partition. In many cases, they produce similar reference partitions, and inferences for the component-specific parameters are consistent.

They also may produce different clustering results sometimes. In the next subsection, we use the Galaxy Data example to show the discrepancy of the clustering results based on the two algorithms, and illustrate how to choose the appropriate algorithm. Comparing the two algorithms, we think K-medroid is more understandable and interpretable. However, the reference partition was identified more accurately with the hierarchical clustering method in the Galaxy example.

3.7.2 The Reference Partition of Subjects

There are multiple methods that could be used to obtain the reference partition. For example, we have introduced the method using mutual distances r_{ij} and the hierarchical clustering model in a previous section. But other methods (K-medroids for example) could also be applied. The partitions obtained from each method might be somewhat different from the others. Readers might be concerned about the validity of the reference partition that is used to construct posterior samples.

We would like to re-emphasize that the reference partition is not our primary interest, and we just use it to identify a useful cluster configuration. Dunson (2010) [7] mentioned that "it is important to note that one should be very careful to avoid over-interpretation of the estimated partition. Even if one is able to identify the optimal partition from among the very high dimensional set of possible partitions, this partition may have extremely low posterior probability, and there may be a very large number of partitions having very similar posterior probability to the optimal partition".

In many cases, the reference partitions from different methods share the same cluster configuration, even though the reference partitions are not exactly the same. They differ from each other in the partitioning of a small portion of the subjects. We are still able to obtain consistent inference with different reference partitions in such cases.

However, different reference partitions could certainly lead to different inferences in some cases. For example, inferences would be surely different if reference partitions have different numbers of clusters, K. In the Galaxy example, there is ambiguity about K; $K = 6$ has the highest posterior probability in Table (B.5). If we choose $K = 6$ and the K-medroids method

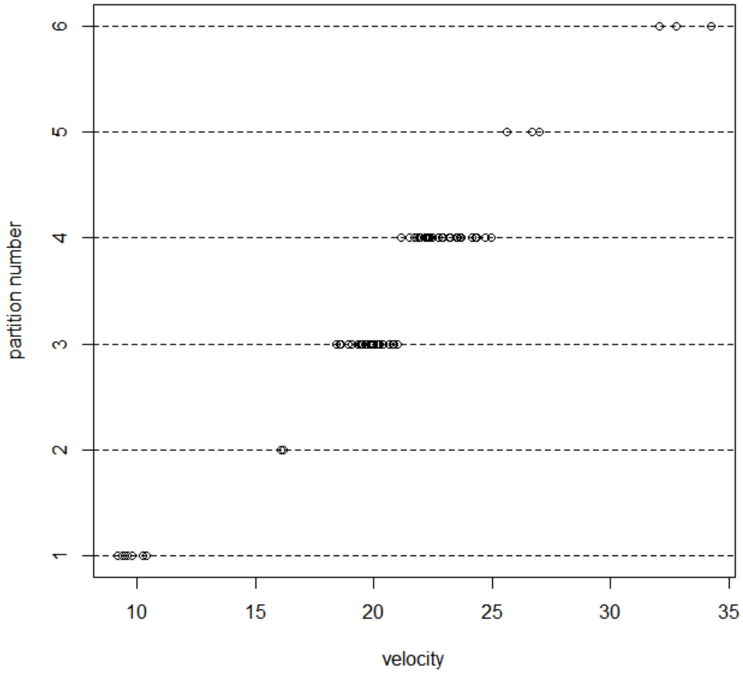


Figure 3.9: Partition of the Galaxy Data into six components using K-medroids

for obtaining the reference partition, we obtain the allocation of the subjects shown in Figure (3.9). In addition, the K-medroids algorithm is similar to K-means but can work with an arbitrary matrix of distances instead of only squared Euclidean distances. The details of the K-medroids algorithm are given in the Appendix.

The clustering result obtained using K-medroids is different from that based on the hierarchical clustering method, though both results look reasonable. The trace plot in Figure (3.4) indicates six clusters. Figure 3.3 shows the predictive density has only five modes. Figure (3.10) gives the reconstructed samples using the reference partition based on K-medroids, in which clusters 4 and 5 are combined in the partition from hierarchical clustering. It is significant that cluster 5 has small size (2 or 3) with observed values that are near to values for observations in cluster 4. Thus it is ambiguous whether the subjects in the two clusters should be separated or combined without knowing the truth.

In order to determine the appropriate reference partition, we compare the posterior samples

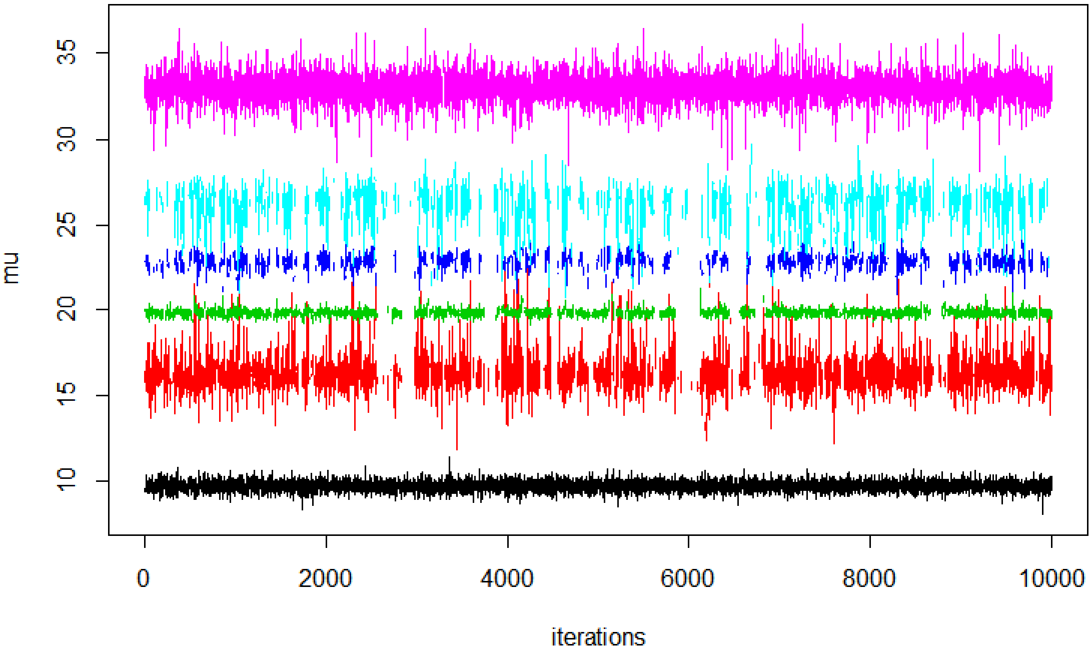


Figure 3.10: Trace plot of μ for the six clusters after post-processing with the reference partition from K-medroids

selected for inference. As mentioned before, "given a target π, a MCMC sampler should visit more than 50% of the support of π". By analogy, if we choose the right reference partition, we should have at least half of the iterations having the same cluster configuration as the reference. In the Galaxy Data example, only 3051 iterations were selected out of 10000 for inference using the reference partition based on the K-medroid method. On the other hand, 6082 iterations were chosen with the reference partition from hierarchical clustering method, which is clearly the better choice.

3.8 Conclusion

In this chapter, we introduced Label Switching and other issues caused by non-identifiability of the components in mixture models. It is an issue that hinders us from making inferences about parameters on the cluster level. We proposed an algorithm to post-process the MCMC output and reconstruct an appropriate subset of the posterior samples. Compared with other post-processing algorithms, which only apply to finite mixture models, our method appears to work for both finite and infinite mixture models, at least in cases considered. In addition, our method requires much less complicated computation and much less processing time compared to other methods. For example, Stephen's method requires the calculation of a loss function for all $K!$ combination of labels at each iteration. As for Yao and Lindsay's method, it applies an EM algorithm at each iteration by using the sample as the initial value. Both methods are computationally demanding, while our ad-hoc algorithm is capable of delivering reasonable results in several seconds.

Chapter 4

Association between UI and Hormone Profiles

4.1 Introduction

Menopause is a universal female phenomenon defined by a specific event, the final menstrual period (FMP). The menopausal transition is a series of stages of variable length from pre-, early peri- and late peri- to post-menopause defined by changes in menstrual and hormonal patterns. Irregularity of the menstrual cycle marks the start of the menopausal transition in most women. Commonly in the mid-40s, cycle length may initially shorten and then progressively lengthen with the approach of the FMP [27], [37]. This irregularity seems to correspond to changes in estrogen levels, which are predominantly the consequence of the decline of ovarian follicle numbers [5]. Ovarian aging may relate to clinical differences in menopause-specific symptoms and health outcomes such as sleep disturbances, changes in cognition, changes in bone mineral density, as well as initiation and progression of diseases of aging such as osteoarthritis.

Urinary Incontinence (UI), which is used as a measure of health status in our study, is a common, debilitating and costly problem, particularly in middle-aged and older women. It is associated with significant morbidity, such as decreased quality of life from social seclusion and psychological stress. Prevalence estimates in mid-life women range from about 5% for severe to 60% for mild incontinence [14], [33]. Personal and societal economic impact of incontinence is substantial with national cost estimates of up to $16 billion annually [42].

UI is thought possibly to be related to the menopausal reduction in endogenous estrogens, since cross-sectional epidemiological studies have found that the prevalence of incontinence is associated with the post-menopausal status [30], [38], [1]. However, it remains unclear whether and how the hormone profile affects the risk of developing UI.

We are motivated by data from the Study of Women's Health Across the Nation (SWAN), which involved the collection of hormone data on women through the menopausal transition. The study is a multi-site longitudinal epidemiologic study designed to examine the health of women during these years. We are interested in characterizing hormone profiles through menopause, and in how these changes might affect UI development.

Our approach is to classify women into groups based on their hormone profiles, and then study how the clustering is associated with women's health, in particular with regard to UI. Eventually, we found that certain distinct cluster shapes are related to women's UI.

In this chapter, we are focused on the analysis of SWAN data using the methods introduced in the past two chapters and their expansion to account for the UI data. We first introduce the SWAN data in Section 2. In Section 3, we apply the clustering model presented in Chapter 2 to both E2 and FSH data to identify different hormone trajectory patterns during menopause. We also discuss the factors that are related to hormone changes. In Section 4, we analyze longitudinal UI and hormone data with a Bayesian joint model to find the association between the UI incidence pattern and hormone profiles during menopause. In

Section 5, we focus on the question about whether the development of UI is associated with hormone changes.

4.2 The SWAN Data

In this study, we analyze a subset of SWAN data from 928 women enrolled in the Study of Womens Health Across the Nation (SWAN), who were followed through 11 annual follow-up visits. Eligibility criteria for entry into the SWAN cohort were age 42-52 years and self-identification as one of four racial/ethnic groups (African American, Chinese, Japanese, Caucasian). Exclusion criteria included inability to speak English, Spanish, Japanese, or Cantonese, no menstrual period in greater than 3 months before enrollment, hysterectomy and/or bilateral oophorectomy prior to enrollment, and current pregnancy, lactation, or hormone use.

In the study, annual samples of blood serum levels of estradiol (E2) and follicle-stimulating hormone (FSH) were collected. In addition, a self-administered questionnaire assessed incontinence at baseline and at each follow-up visit. Based on response to the question: "In the past year/since your last study visit, have you ever leaked even a small amount of urine involuntarily?", Frequency of incontinence was classified as "almost daily/daily" (daily), "several days per week" (weekly), "less than one day per week" (monthly), "less than once a month" or "none". We defined a binary UI response by letting "any incontinence" be defined as at least monthly occurrence during the year. We considered incontinence occurring less than once a month as clinically insignificant and thus combined this category with "no incontinence".

In the analysis, we adjusted the time scale by anchoring $t_i = 0$ at the year of FMP for woman i, so that the observations are comparable cross-sectionally. In the SWAN study, the menopausal status was assessed annually based on self-reported bleeding patterns. A woman was considered post-menopausal when she had no bleeding for at least 12 months. FMP was identified at the first visit when a woman became post menopausal. With the adjustment of time scale, t_i ranges from -10 to 9, which means the data were collected from 10 years before FMP to 9 years after FMP.

Time-invariant covariates include baseline age, race/ethnicity, BMI, marital status and education. A detailed summary of time-invariant covariates is given in Table 4.1. Among the 928 participating women, Caucasians comprised 53.0% of the sample, American Africans 21.3%, Chinese 12.1%, Japanese 13.6%, respectively. With the categorization criteria defined by the World Health Organization (WHO), 26.3% of women were overweight ($25 \leq BMI < 30$) at baseline, and 25.9% were obese ($BMI \geq 30$). In addition, 83.3% of women have college level or higher education, and 68.5% of them are married.

4.3 Cluster Analysis with Hormone Data

The menopausal transition corresponds to changes in estrogen levels. As the menopause transition progresses, E2 levels eventually decline significantly and remain low, while FSH levels increase and remain high [5], [29]. While this general sequence of events has been assumed to be consistent for most women, several small studies have observed that E2 may increase immediately prior to the final menstrual period [41], [13]. Moreover, the timing and magnitude of increase in FSH varies among individuals [28]; aggregation of FSH levels across individual woman may be misleading and consequently overlook factors related to FSH developmental patterns. The literature so far has not conclusively defined the variations of E2 and FSH trajectories or factors related to these variations.

In order to identify distinct E2 and FSH trajectories, we apply our clustering model to both

Table 4.1: Summary of the time-invariant covariates for the SWAN data

Covariates	Numerical Summary
Age at entry (years)	
mean(±SD)	52.5(±2.5)
Ethnicity	
Caucasion	492(53.0%)
American African	198(21.3%)
Chinese	112(12.1%)
Japanese	126(13.6%)
BMI (kg/m^2)	
Normal(< 25)	444(47.8%)
Overweight($25 - 30$)	244(26.3%)
Obese(> 30)	240(25.9%)
Education	
College or higher	773(83.3%)
High School or lower	151(16.3%)
Missing	4(0.4%)
Marital Status	
Married	636(68.5%)
Single	164(12.6%)
Separated, widow or divorced	117(17.7%)
Missing	11(1.2%)

E2 and FSH data to group the women based on their hormone profiles. We then show both clustering results in this section. We are able to identify three distinct developmental patterns for E2 and four for FSH. With the clustering results, we also try to ascertain which factor(s) might be related to the variation of the hormone trajectory patterns.

4.3.1 Clustering of E2 Profiles

Let the E2 response be a vector $y_i = (y_{i1}, y_{i2}, \ldots, y_{ir_i})'$ corresponding to times $t_i = (t_{i1}, t_{i2}, \ldots, t_{ir_i})'$ for woman i, $i = 1, 2, \ldots, n$. In the analysis, we use log-transformed E2 data as y_i since the values are all positive with large variation. The log-transformed E2 data are shown in Figure 4.1.

We retain the same notation used in Chapter 2. The clustering model is built as follows for the hormone data, including both E2 and FSH;

$$\begin{aligned} y_i|\beta_i &\sim N_{r_i}(X_i\beta_i, \tau_e^{-1}W_i) \\ \beta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= N_{p+1}(u_\beta, \Xi_\beta^{-1}) \end{aligned} \tag{4.1}$$

where X_i is the $r_i \times (p+1)$ design matrix containing the basis functions up to p^{th} order for

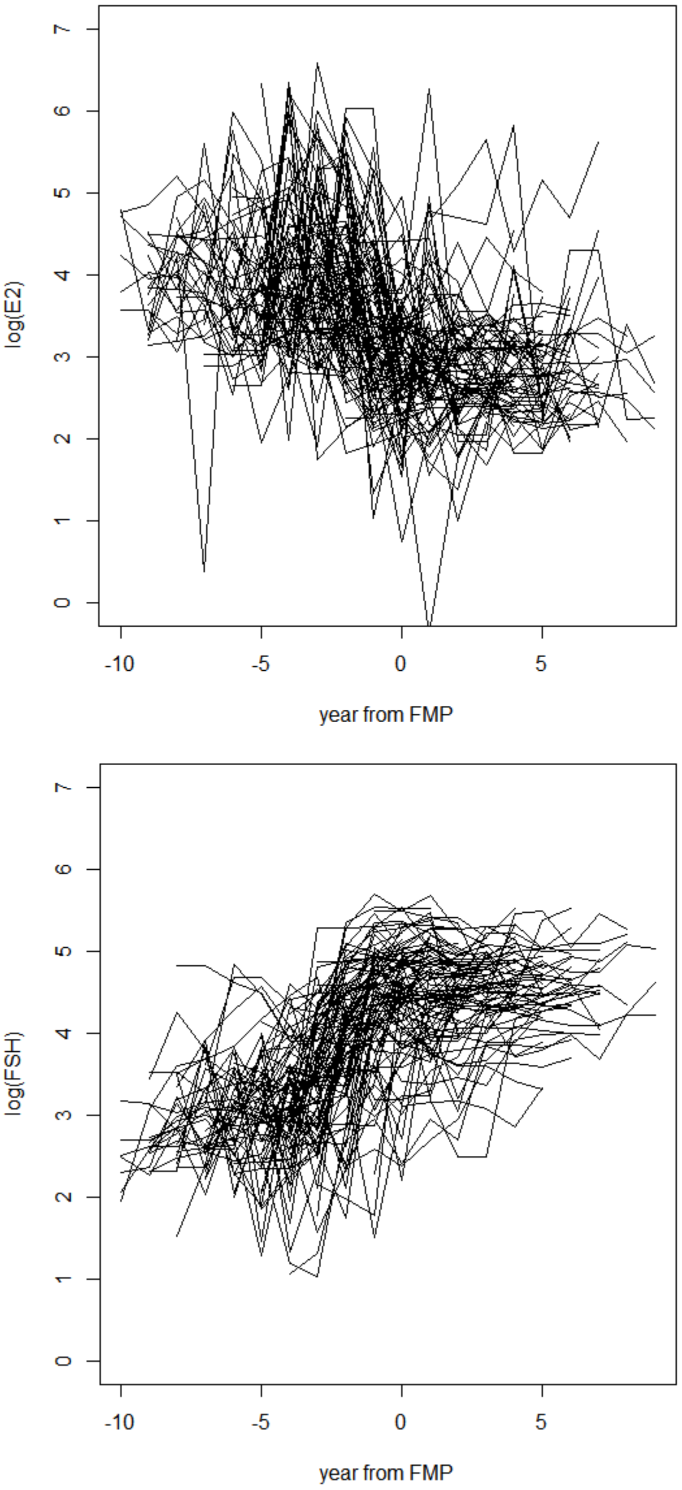


Figure 4.1: Spaghetti plots of $\log E2$ (Upper) and $\log FSH$ (Below) from 100 randomly selected women.

Table 4.2: Posterior Distribution of the Number of Clusters K for E2 hormone data

Cluster Number K	4	5	6	7	8
Posterior Probability	0.0002	0.755	0.220	0.024	0.0008

woman i, $W_i = I_{r_i} + X_i \Gamma^{-1} X_i^T$ and

$$\Gamma = \begin{pmatrix} \gamma_0 & 0 & \cdots & 0 \\ 0 & \gamma_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_p \end{pmatrix}.$$

We apply the clustering model to E2 hormone data, and it yields some interesting results. Here, we list the posterior probability distribution for the number of clusters, K, in Table 4.2. According to the methods introduced in section 2.4.3, we can ascertain how the women are clustered. In Figure 4.2, we show the spaghetti plots of the women in each cluster.

In Chapter 3, we introduced our algorithm to construct the posterior samples for parameters on the cluster level. Then we are able to make inferences about cluster mean trajectories. In Figure 4.2, we plot the five mean trajectories with 95% Probability bands using black lines.

The estimated mean trajectories for clusters 4 and 5 may not reflect overall trends well. The estimated curves are dramatic at early and late times for cluster 4 ($t_i \leq -6$ and $t_i \geq 5$) and at the early stage for cluster 5 ($t_i \leq -4$). Evidently, the issue is caused by extrapolation since there were no E2 observations in these time regions. We notice the estimated mean trajectories for clusters 4 and 5 are similar to that for cluster 3 over the ranges of observed values, as can be seen in Figure 4.3. We thus combine the three clusters in the following analysis, and tag the combined cluster as cluster 3. Evidently, our clustering method selected clusters 4 and 5 because of the missing data.

Eventually, we obtain three distinct E2 profiles shown in Figure 4.4. The three E2 trajectories

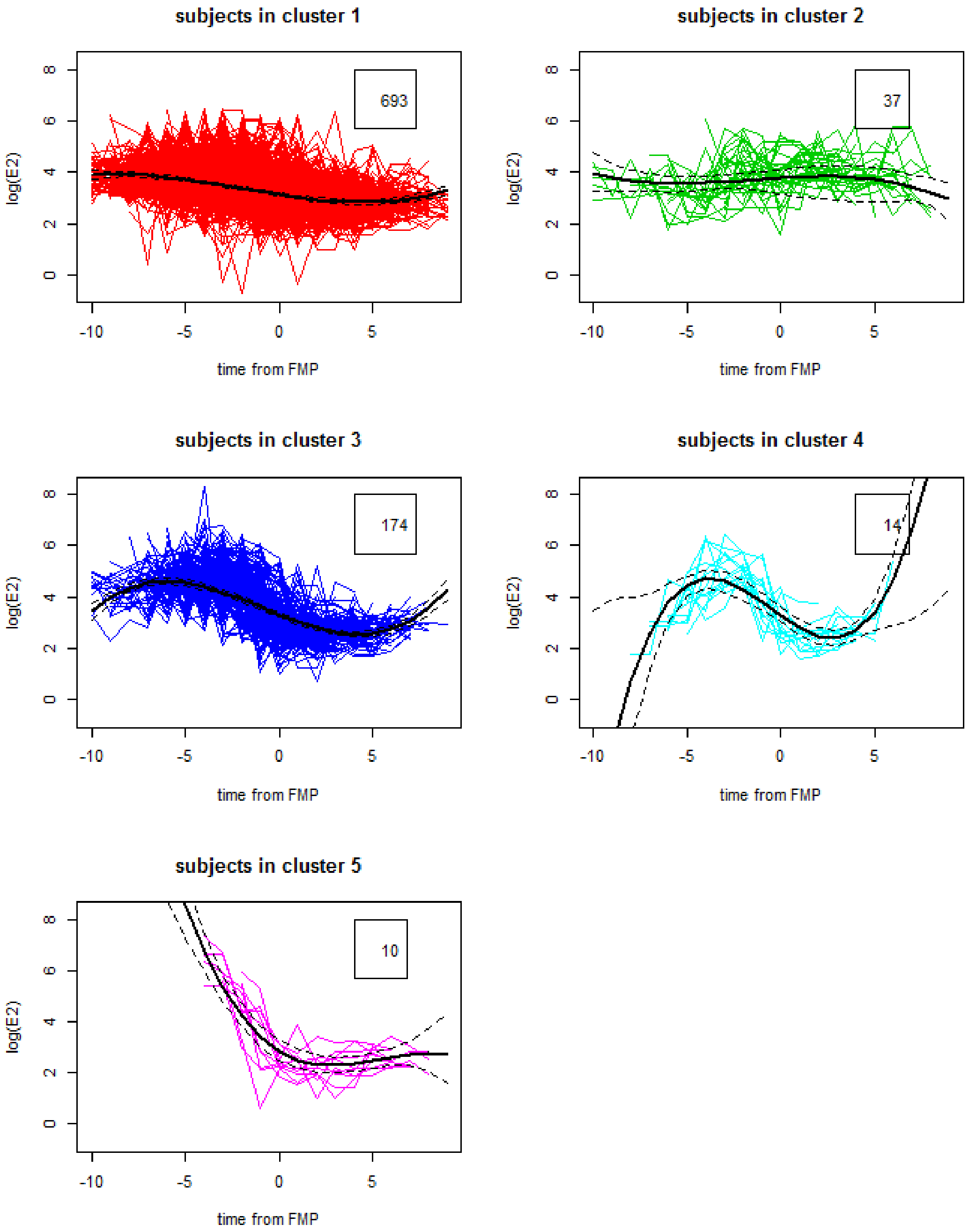


Figure 4.2: Spaghetti plots of the women in each of the five clusters obtained based on their E2 profiles.

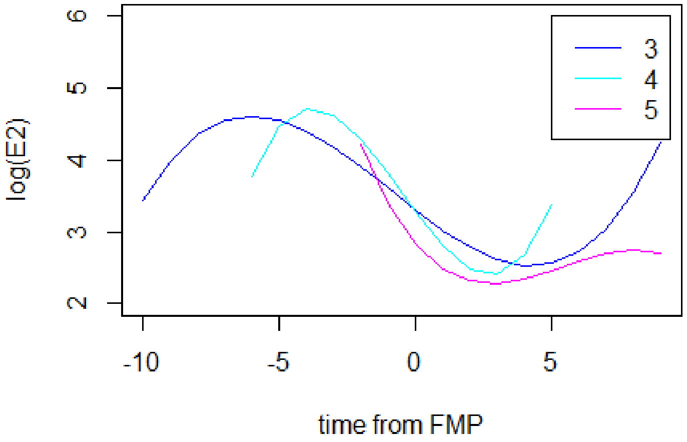


Figure 4.3: The estimated mean E2 trajectories for clusters 3, 4 and 5, which are separated with different colors. The trajectories are truncated to the regions with data support for clusters 4 and 5.

Table 4.3: Global Parameter Estimates for the model applied to FSH data.

Parameter	Mean	Median	95% PI Lower	95% PI Upper
σ_e	0.47	0.47	0.46	0.48
γ_0	4.2	4.1	3.1	5.9
γ_1	142.1	116.5	46.3	274.3
γ_2	12.6	10.17	6.3	32.9
γ_3	10.1	9.2	5.6	19.8

are summarized as: gradually decline (Cluster 1, 74.7% of the sample); almost flat (Cluster 2, 4.0%); rise then decline (Cluster 3, 21.3%).

4.3.2 Clustering of FSH profiles

The same analysis was applied to FSH profiles. We list the estimates of global parameters in Table 4.3. Then we select 12 clusters based on the posterior probability distribution for K in Table 4.4. Figure 4.5 shows the allocation of the 928 women and the 12 cluster mean trajectories.

Clustering identified several women with FSH observation outliers, who were located in

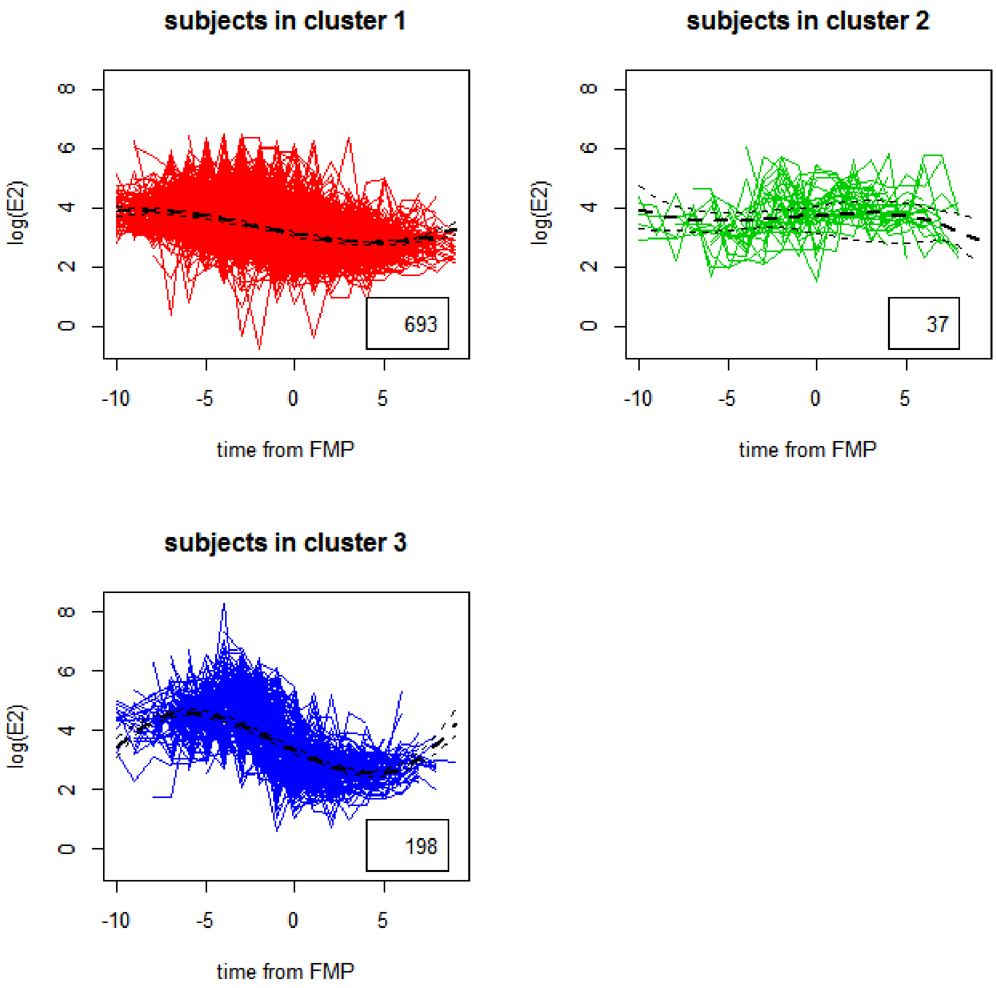


Figure 4.4: Spaghetti plots of the women in each of the three reconstructed clusters based on their E2 profiles.

Table 4.4: Posterior Distribution of the Number of Clusters K for FSH hormone data

Cluster Number K	10	11	12	13	14	15	16
Posterior Probability	0.007	0.257	0.594	0.128	0.013	0.0014	0.0002

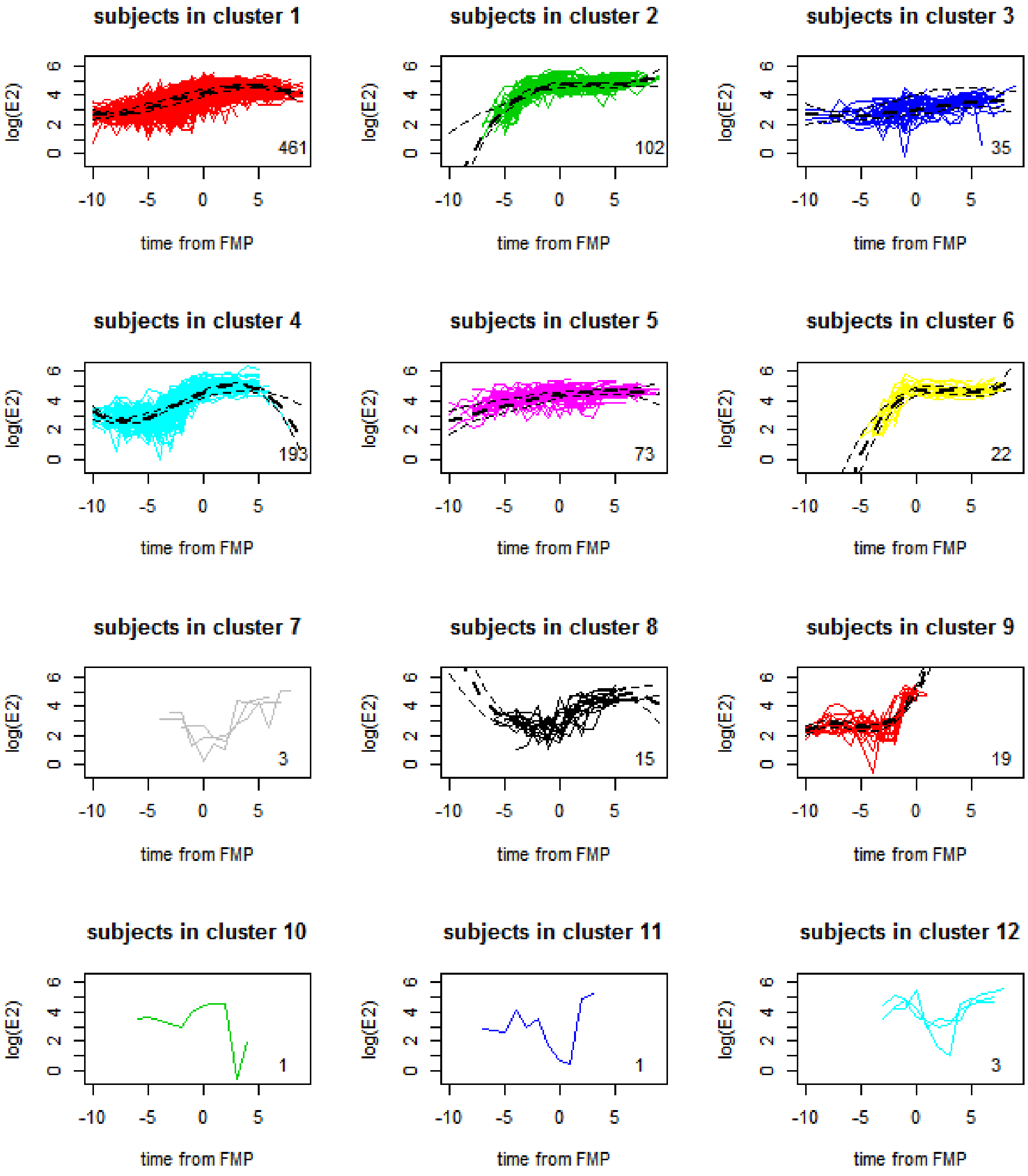


Figure 4.5: Spaghetti plots for the women in each of the 12 clusters obtained based on their FSH profiles.

clusters 7, 10, 11 and 12. These women have extremely low values of FSH observations (< 3 mIU/ml), which are unrealistically small for women during menopause. We thus exclude them in the following analysis.

In addition, similar to the clustering results for E2, missing data caused extrapolation problems for the estimation of cluster mean trajectories, and thus lead to separation of similar FSH patterns. We can see the FSH developmental patterns in clusters 1, 2, 4, 6 and 9 are all similar in the region over which FSH is observed. Therefore, we combine them into one cluster and relabel the clusters as shown in Figure 4.6

We obtain four distinctive FSH profiles during menopause with our clustering model, which we describe as: gradually rise (Cluster 1, 88.1% of the sample); early rise (Cluster 2, 7.5%); almost flat at low level (Cluster 3, 2.3%); late rise (Cluster 4, 2.1%). The typical FSH trend is rising gradually during menopause for most women (88.1% of the sample). The rise starts about 5 years before FMP and reaches the ceiling one to two years after FMP on average. A relatively small portion of the women (11.9% in total) were found to have different FSH profiles. FSH for women in cluster 2 rises to a high FSH level very early (about four years before FMP) and remains at the high level through remainder of menopause. On the contrary, the women in cluster 3 maintain a relatively low FSH level throughout menopause and only show a slight increasing trend after FMP. The FSH level for the women in cluster 4 increases during menopause, but the rise starts around FMP, which is much later than for cluster 1. Our clustering results show that the timing and magnitude of increase in FSH can vary among individuals.

4.3.3 The Factors Associated with Hormone Profiles

We are also interested in whether other covariates, including race/ethnicity and BMI, were related to the different hormone trajectories. Such information will help us understand

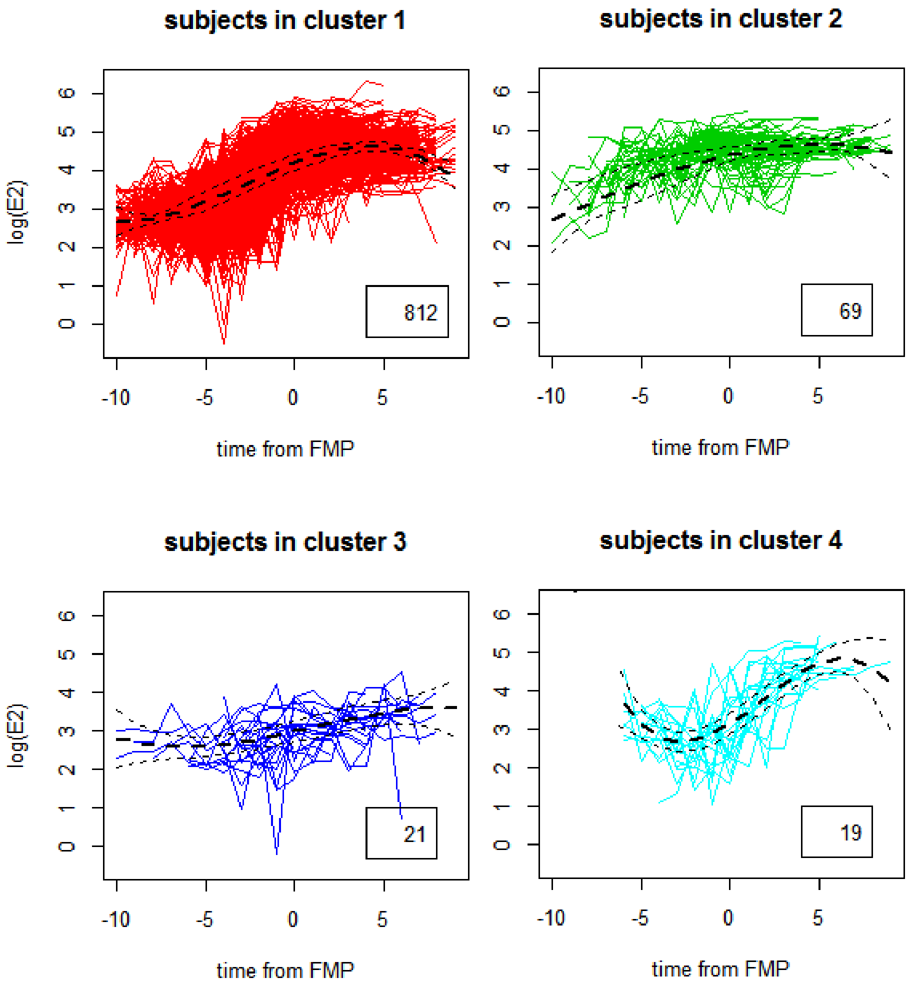


Figure 4.6: Spaghetti plots for the women in each of the four reconstructed clusters based on the FSH profiles. Cluster mean trajectories with 95% probability bands are plotted in black.

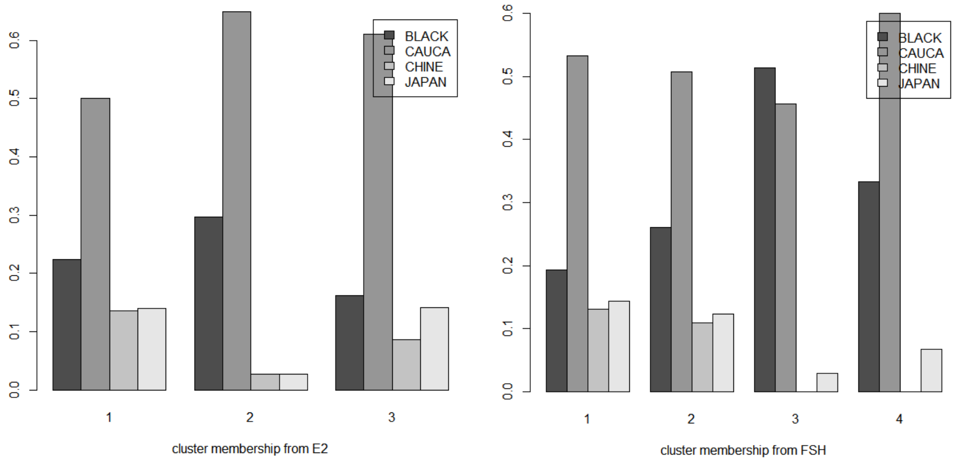


Figure 4.7: Conditional empirical distribution of race in each cluster based on hormone profiles. Left: E2; Right: FSH

differences in ovarian aging among different populations, which in turn may relate to clinical differences in menopausal-specific symptoms and health outcomes.

These differing hormone trajectories were strongly related to BMI and race/ethnicity but not smoking, physical activity, or demographic variables.

Figure 4.7 shows the conditional empirical distribution of ethnicity in each cluster for both E2 and FSH. We find some atypical hormone patterns tend to appear more in African Americans and Caucasians in the sample data. For E2, the typical trend is to drop during menopause for most women. The flat trend in cluster 2 is atypical. Among the 37 women in cluster 2, only 2 are Asian (Chinese/Japanese). The barplot also shows the proportion for Asians is extremely low in cluster 2. Regarding FSH, there are no Chinese in clusters 3 and 4, and the proportion for Japanese is also low.

We used a Chi-square test to test the independence of hormone based cluster membership (E2 and FSH) and ethnicity, and found a statistically significant relationship between them. The p-values are 0.006 and 0.001 for E2 and FSH, respectively. So the distribution of ethnicity within clusters varies from cluster to cluster.

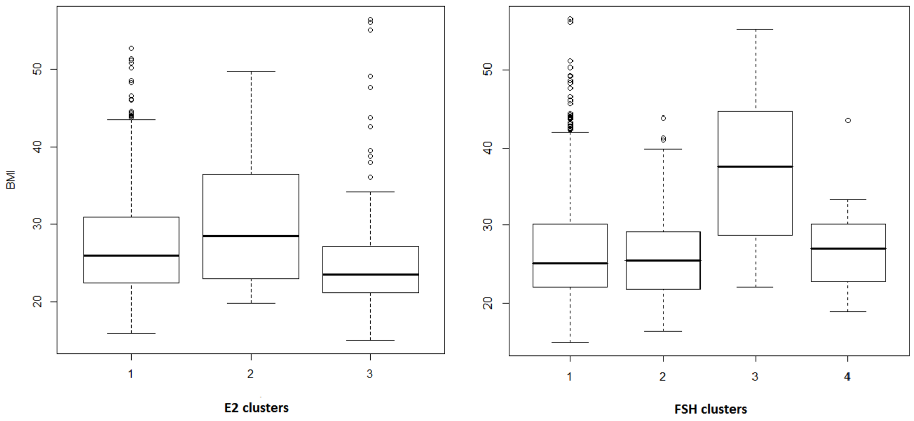


Figure 4.8: Boxplot of BMI in each hormone cluster. Left: E2 clusters; Right: FSH clusters

We found that women with the flat hormone pattern have higher BMI on average. Figure 4.8 shows that women with the flat E2 trend in cluster 2 have the highest median BMI among the 3 clusters, and cluster 3 in the FSH clustering also has significantly higher median BMI than the other three.

We categorized BMI using criteria defined by World Health Organization (WHO): Normal ($18.5 \leq \text{BMI} < 25$), overweight ($25 \leq \text{BMI} < 30$) and obese ($\text{BMI} \geq 30$). The Chi-square test shows a significant relationship between BMI and hormone clustering. The p-value is < 0.001 for E2 and 0.023 for FSH. We also constructed a barplot in Figure 4.9 to show the relationship between hormone patterns and BMI. It clearly shows that the obese group has higher proportion of women in cluster 2 for E2 and cluster 3 for FSH.

Another interesting finding is that the women with "typical" hormone patterns (cluster 1 in both E2 and FSH clusters) reaches FMP later in age than others on average. Figure 4.10 shows the median age at FMP in cluster 1 is higher for both E2 and FSH clusters.

In addition, we constructed the barplots in Figure 4.11 to show the relationship between hormone clustering and other factors including education and marital status. There is no significant association found for E2 clustering using Chi-square test for independence (p-value=0.3 for education and p-value=0.2 for marital status). For FSH clustering, the rela-

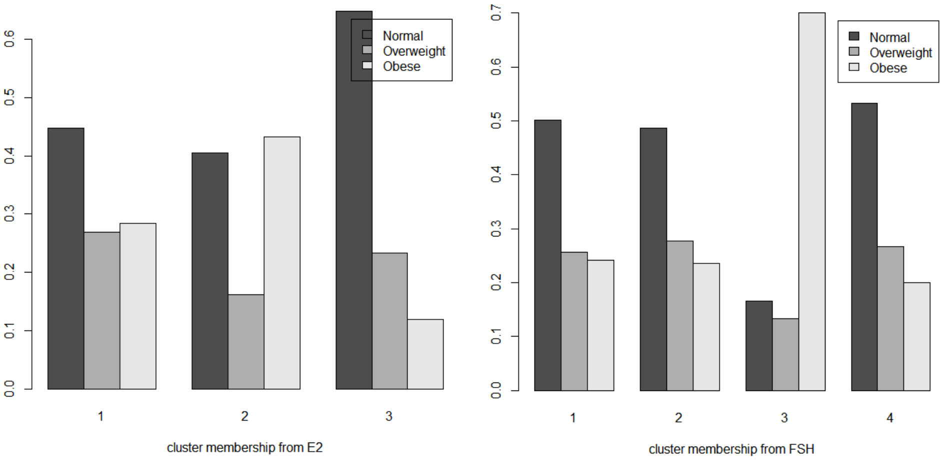


Figure 4.9: Conditional distribution of BMI categories in each cluster based on hormone profiles. Left: E2; Right: FSH

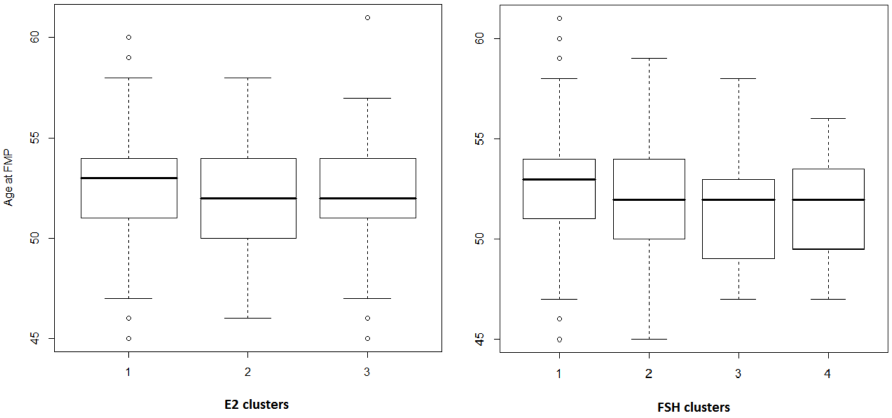


Figure 4.10: Boxplot of age at FMP in each hormone cluster. Left: E2 clusters; Right: FSH clusters

tionship is significant with both p-values equal to 0.01. The barplot shows the proportion of "seperated, widow or divorced" is higher and that of "single" is lower in cluster 3 compared with the other three clusters.

4.4 Joint Modeling of hormone profiles and UI

Our ultimate goal is to ascertain whether and how UI development is associated with hormone trends through menopause. In the SWAN data, women's UI status is measured annually. Let $u_i = (u_{i1}, \ldots, u_{ir_i})$ denote the vector of UI status for woman i corresponding to $(t_{i1}, \ldots, t_{ir_i})$. We summarize the proportion of UI incidence in each year in Figure 4.12, from which we could see the trend is women developing UI with aging.

There is a steep drop at year -9 and a quick increase after year 8. It is difficult to tell whether these are real or not. Because of relatively small numbers of observations in the pre- and post- stages of menopause, the effects could be purely random. We give the number of observations at each year in Figure 4.13, which shows the number drops quickly with time moving away from FMP. In addition, missing data also caused some issues with E2 clustering. In section 3, we found some women with similar hormone profiles were separated into different clusters due to missing data in these years.

Therefore, we decide to truncate the time scale to $(-8, 6)$ in the following analysis, which means we only keep the data collected in this time region. Since we are interested in the menstrual cycle around FMP, we do not lose much information with the truncation.

In addition, the baseline BMI value is missing for 10 women in the data. We removed them from the analysis because we consider the effect of BMI on UI incidence. We thus have 918 women with 9600 observations in the data.

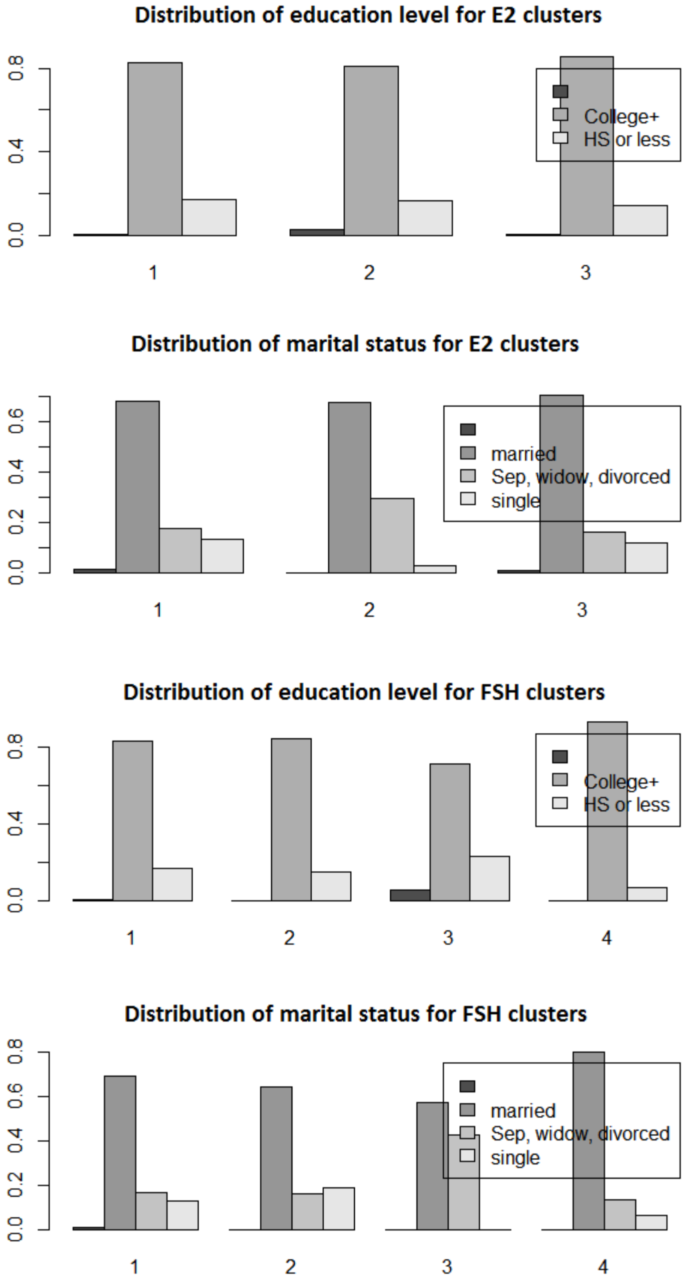


Figure 4.11: Conditional distribution of education and marital status in each cluster based on hormone profiles. The bar without label indicates missing data for education and marital status.

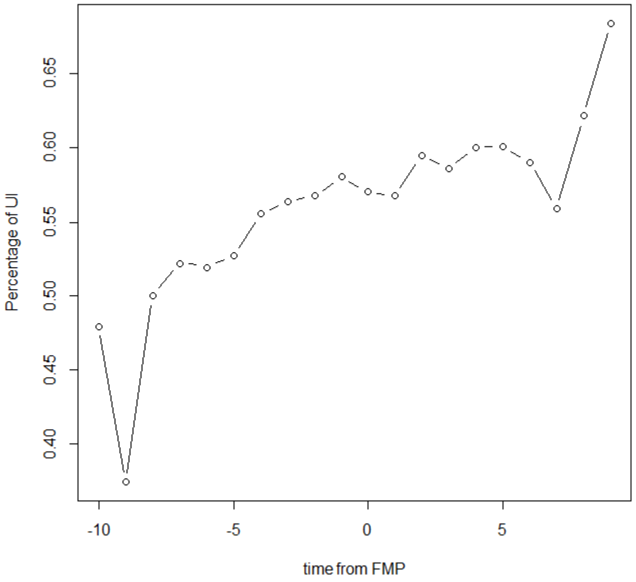


Figure 4.12: Sample proportion of UI incidence in each year

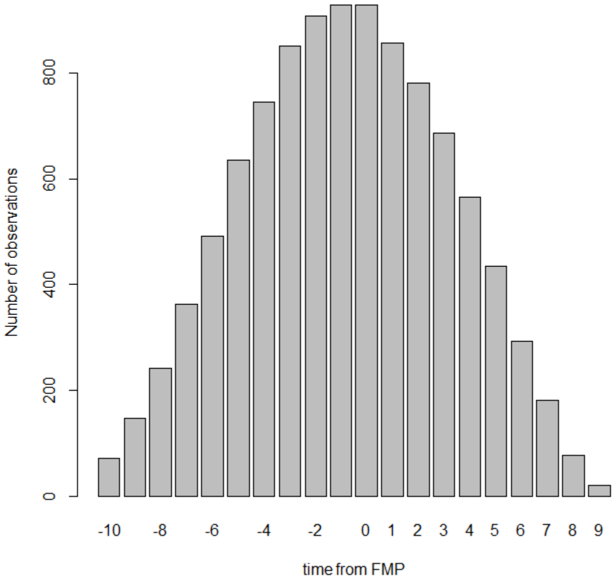


Figure 4.13: The barplot shows the number of observations collected each year with the adjusted time scale.

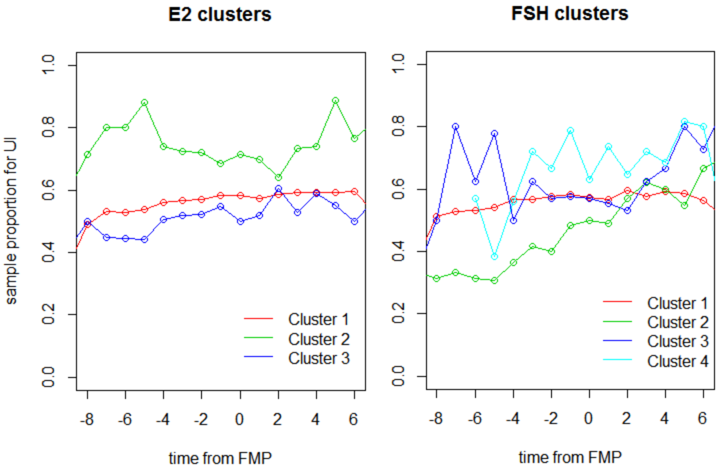


Figure 4.14: Sample proportion of UI incidence in each hormone cluster by time. Left: E2; Right: FSH

In Figure 4.14, we plot the sample proportion of UI incidence in each hormone cluster. The UI incidence is clearly related to E2 cluster membership, since the UI incidence pattern is different corresponding to different E2 clusters. The UI incidence rate for cluster 2 is constantly higher than for the other two clusters through menopause. In addition, the UI incidence patterns are also somewhat different corresponding to the four FSH clusters.

In this section, we propose a Bayesian semi-parametric approach to jointly model both UI and hormone data in order to study their association. We will introduce our method and present the analysis using E2 and UI data in the first subsection, and show the results using FSH and UI data in the second.

4.4.1 Association between E2 and UI

We are interested in the factors that are associated with UI incidence. We use logistic regression to model the UI response at each time period. Since the probability of UI incidence is clearly not constant with time, we use a linear combination of basis functions to fit the probability trend. In addition, we include time-invariant covariates including BMI and age at FMP in the model because they are considered to be related to UI incidence. We jointly model UI and the hormone trajectory information since we would like to know how the UI development is related to hormone profiles. We incorporate the logistic regression model with the clustering model to see how the UI developmental pattern relates to different hormone profiles.

The joint probability model for UI and hormone trajectory is:

$$\begin{aligned}
u_{ij}|p_{ij} &\sim Bern(p_{ij}) \text{ for all } j=1,\ldots,r_i \\
logit(p_{ij}) &= X_{ij}\lambda_i + z_i\omega \\
y_i|\beta_i &\sim N_{r_i}(X_i\beta_i, \tau_e^{-1}W_i) \\
(\beta_i,\lambda_i)|G &\overset{iid}{\sim} G \\
G &\sim DP(\alpha, G_0) \\
G_0 &= N_{p+1}(\mu_\beta, \Xi_\beta^{-1}) * N_{p+1}(\mu_\lambda, \Xi_\lambda^{-1}) \qquad (4.2)
\end{aligned}$$

where

- p_{ij} is the UI incidence rate at t_{ij} for woman i.
- X_i is the $r_i \times (p+1)$ design matrix containing basis functions up to the p^{th} order for woman i. Since the time scale is the same for UI and E2, we use the same orthogonal polynomial basis functions that we used for E2 clustering.
- X_{ij} is the j^{th} row of X_i, which is the basis function values at t_{ij} for woman i.
- λ_i is the coefficient vector, which models the UI incidence probability trend for woman i.
- The DP has been assigned to the joint distribution for β and λ, so that each hormone

Table 4.5: Global Parameter Estimates for the joint clustering model applied to E2 and UI data.

			95% PI	
Parameter	Mean	Median	Lower	Upper
σ_e	0.70	0.70	0.69	0.71
ω_{bmi}	0.35	0.35	0.21	0.50
ω_{age}	0.25	0.25	0.12	0.37
γ_0	7.30	7.24	6.02	8.91
γ_1	95.45	64.19	21.46	393.86
γ_2	13.47	13.09	9.81	19.36
γ_3	17.60	16.62	11.38	29.83
α	0.93	0.87	0.35	1.82

profile $X_i\beta_i$ corresponds to a UI development pattern, $X_i\lambda_i$, in the clustering result.

- $z_i = (z_{i1}, \ldots, z_{iq})$ is the vector of time-invariant covariates, including BMI and age at FMP.

- $\omega = (\omega_{bmi}, \omega_{age})$ is the vector of coefficients for time-invariant covariates BMI and age at FMP.

Part of the model is based on model (4.1) (same as 2.2), which was discussed in Chapter 2. Many prior distributions are identical for both models, and the sampling algorithms are also similar.

With the posterior MCMC samples, we obtain the inferences for the global parameters listed in Table 4.5. Both BMI and women's age at FMP are positively associated with UI incidence.

We summarize clustering results in Table 4.6, which shows the posterior probabilities for K; $K = 6$ and $K = 7$ have the highest probabilities, 0.35 and 0.42 respectively. We select six clusters for our analysis since $K = 6$ appears more reasonable from the Dendrogram in Figure 4.15. Even though $K = 7$ has higher posterior probability than $K = 6$, we do not think there are seven distinctive clusters. Iterations with $K = 7$ indicate a transient component in addition to the six stable clusters.

Table 4.6: Posterior Distribution of the Number of Clusters, K, for the joint E2 and UI model

Cluster Number m	6	7	8	9	10
Posterior Probability	0.350	0.423	0.183	0.039	0.005

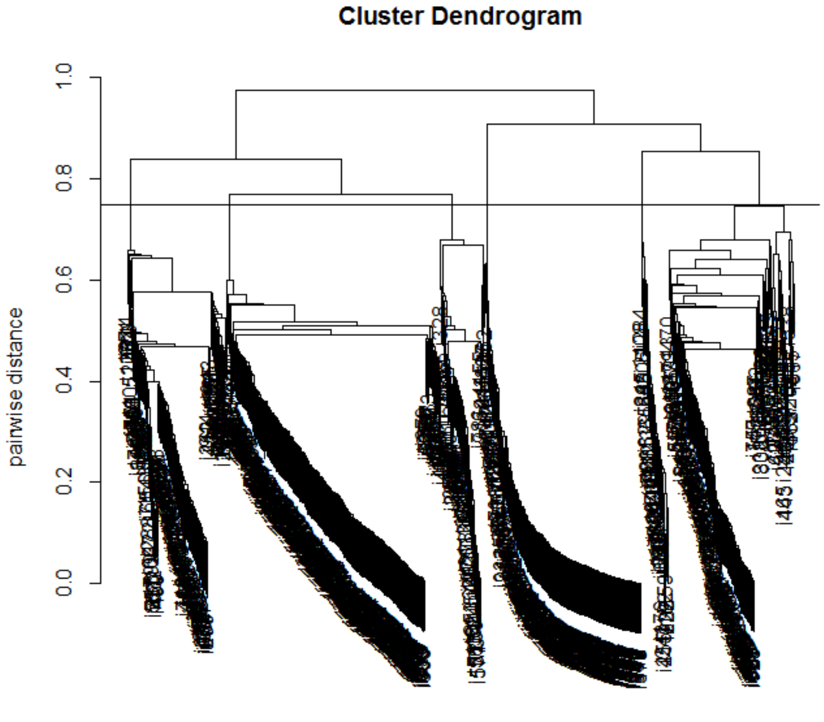
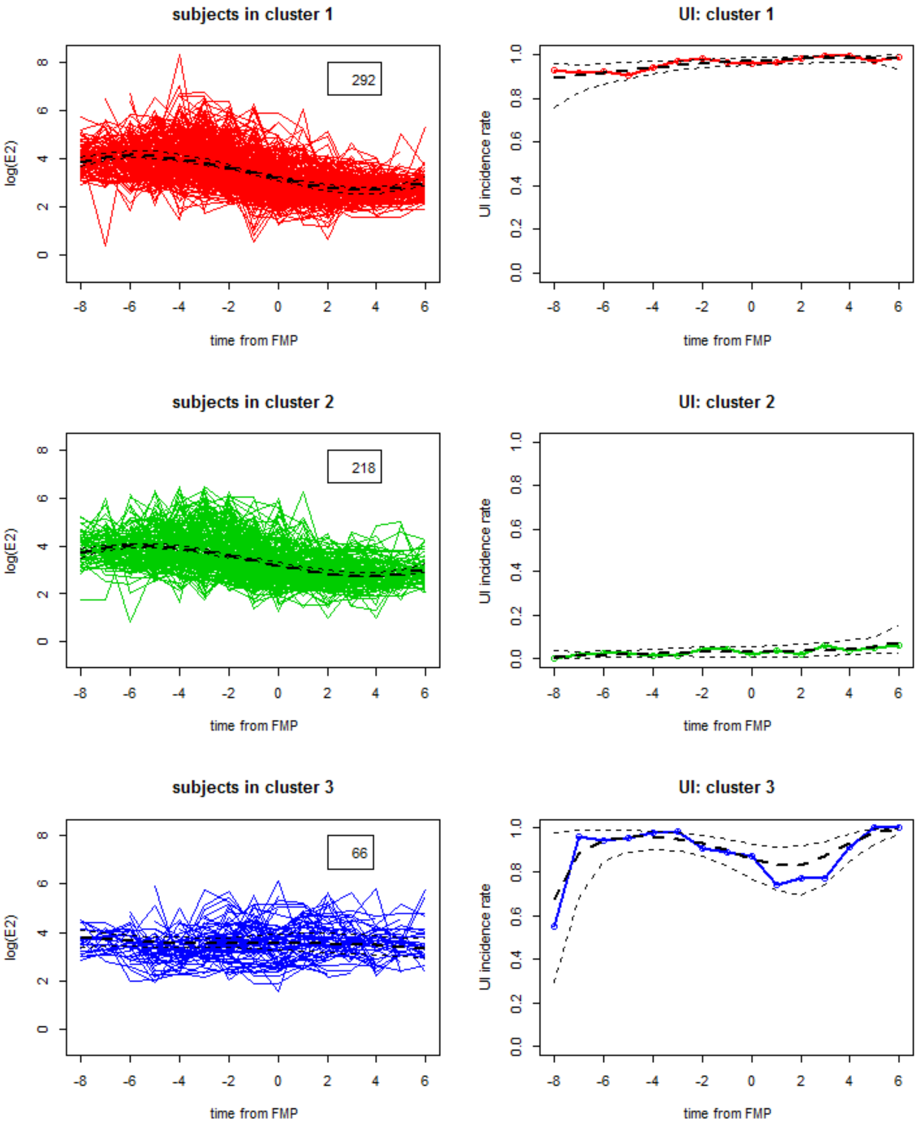


Figure 4.15: Dendrogram for Hierarchical Clustering Algorithm, using E2 and UI data from SWAN.


subjects in cluster 1
292
log(E2)
time from FMP
UI: cluster 1
UI incidence rate
time from FMP
subjects in cluster 2
218
log(E2)
time from FMP
UI: cluster 2
UI incidence rate
time from FMP
subjects in cluster 3
66
log(E2)
time from FMP
UI: cluster 3
UI incidence rate
time from FMP
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
-8
-6
-4
-2
0
2
4
6

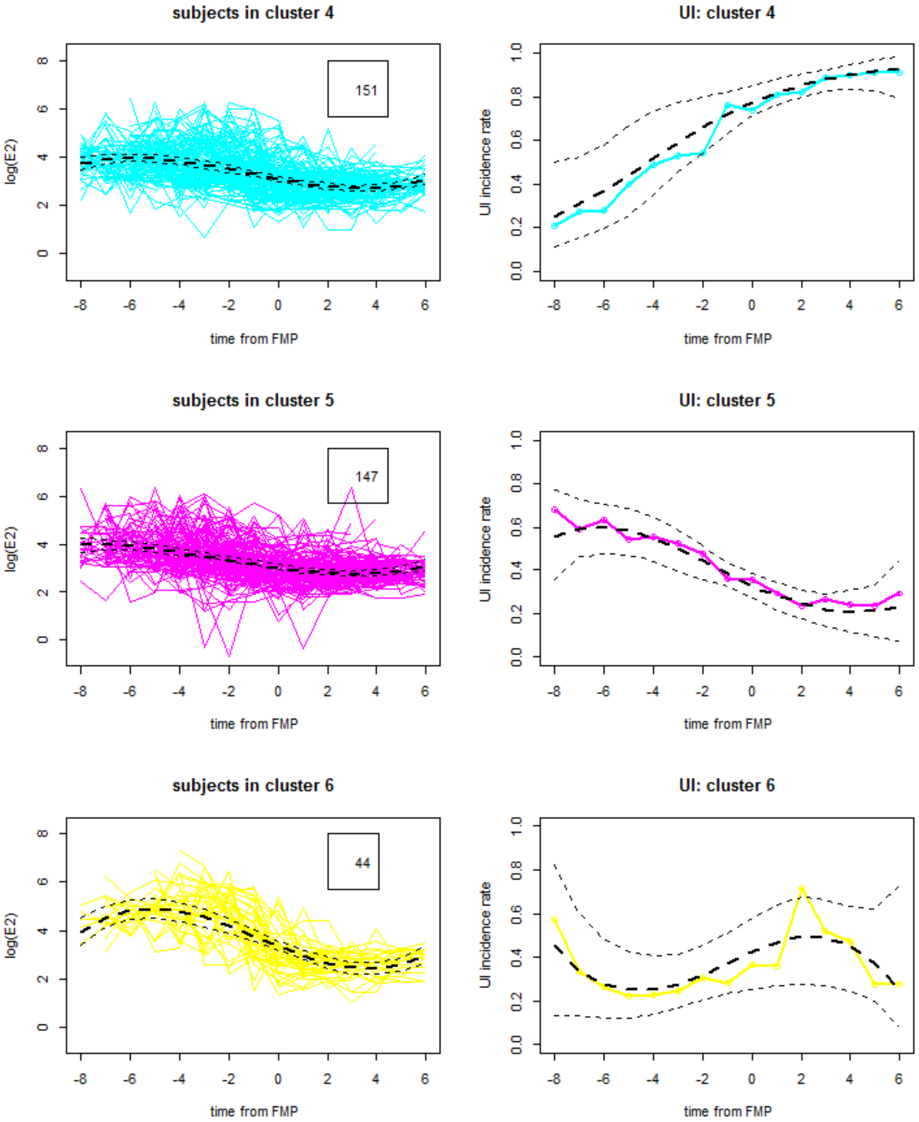


Figure 4.16: Six clusters have been identified from the joint modeling of E2 and UI, and each row has a pair of graphs that plot log $E2$ (left) and UI (right) patterns for each cluster. The estimated cluster mean trajectories with 95% probability bands are plotted in black. The sample data are plotted in different colors for different clusters. Spaghetti plots are made for E2, and the sample proportions are plotted for UI incidence.

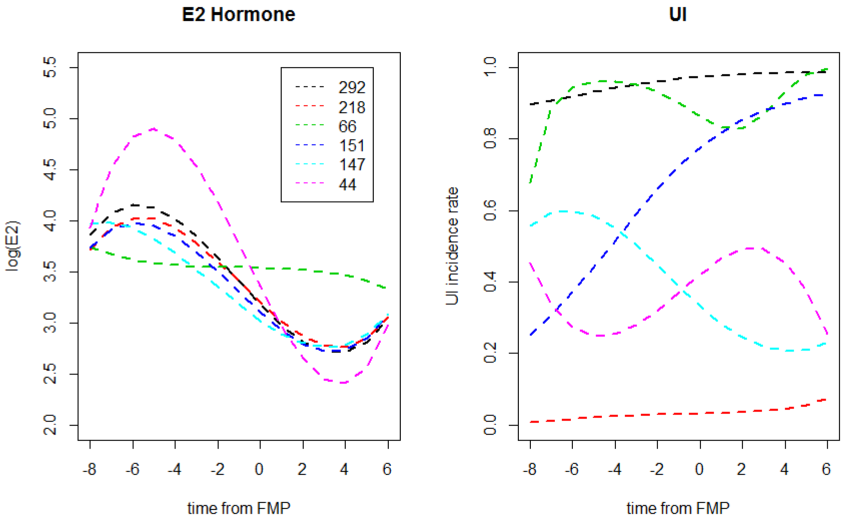


Figure 4.17: The estimated mean trajectories for the six clusters separated by different colors.

The cluster mean trajectories are estimated and plotted in Figure 4.16. With the joint model, we identify the same three E2 patterns that were found in Section 3. Cluster 3 is the “almost flat” pattern and cluster 6 is “rise then steep decline”. Clusters 1, 2, 4 and 5 are all “slight rise then gradual decline” pattern, and they are separated due to different UI patterns. We combine the estimated cluster mean trajectories together in Figure 4.17 to compare the six clusters.

It is interesting to see that the women with the “almost flat” pattern for E2 trajectory (7.1%, green line) tend to have a high UI incidence rate (≥ 0.8) through menopause. The “rise then steep decline” pattern (4.7%, magenta) corresponds to a low rate (≤ 0.5).

Most women (88.2%) are clustered with “slight rise then gradual decline” E2 pattern, and they correspond to four different UI developmental patterns. Some women had no UI and stayed healthy throughout Menopause (23.5%, red). Some women had UI at baseline and stayed affected (31.5%, black), some women had no UI at baseline and gradually developed it during menopause (16.3%, blue), and a portion of women who had UI at baseline showed UI symptoms decreasing(15.8%, light blue).

Table 4.7: Global Parameter Estimates for the joint clustering model applied to FSH and UI data.

Parameter	Mean	Median	95% PI	
			Lower	Upper
σ_e	0.484	0.484	0.476	0.492
ω_{bmi}	0.39	0.39	0.24	0.53
ω_{age}	0.22	0.22	0.09	0.35
γ_0	2.87	2.87	2.47	3.31
γ_1	16.42	15.0	9.61	31.13
γ_2	13.14	12.53	9.16	20.80
γ_3	5.15	5.11	4.31	6.17
α	0.97	0.91	0.37	1.88

Table 4.8: Posterior Distribution of the Number of Clusters, K, for FSH and UI joint model

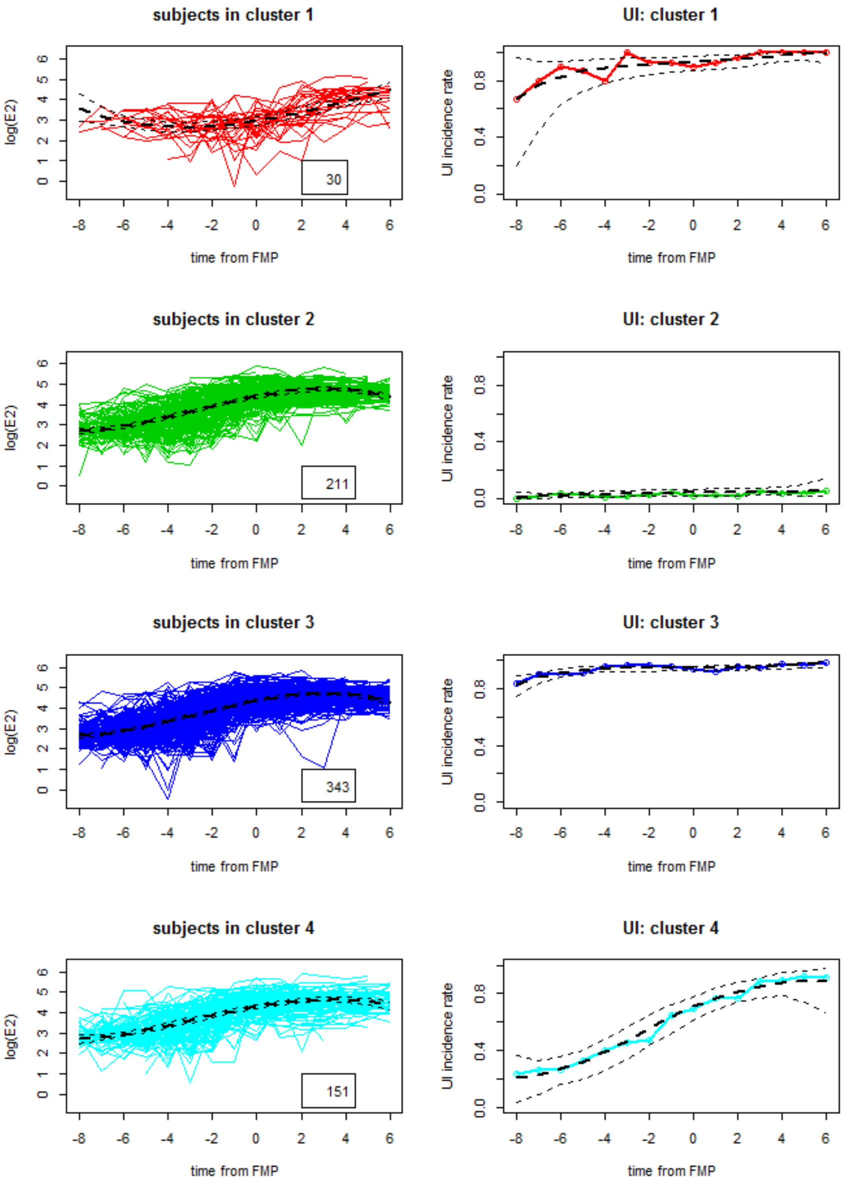
Cluster Number m	7	8	9
Posterior Probability	0.920	0.078	0.002

4.4.2 Association between FSH and UI

We used the same model to analyze the association between FSH and UI incidence. We obtained statistical inferences as follows. Table 4.7 gives estimates of global parameters with 95% probability intervals. The estimates of ω_{bmi} and ω_{age} are similar to results from the joint analysis of E2 and UI above. This is reasonable since they indicate the relationship between UI incidence and baseline BMI/age at FMP, in both analyses. However, the estimates are adjusted by inclusion of the effect of FSH in this model and E2 in the previous model.

Table 4.8 shows $K = 7$ has the dominating posterior probability 0.92, thus we select seven clusters.

The cluster mean trajectories are estimated and plotted in Figure 4.18. Cluster 7 only contains one woman, who is identified as an outlier. We exclude it from the following analysis. We combine the estimated cluster mean trajectories from the remaining six clusters together in a single plot, Figure 4.19, to compare them. We identify two distinctive FSH profiles, which correspond to different UI patterns. One is the "gradual rise" profile (95.2% of the


subjects in cluster 1
log(E2)
0
1
2
3
4
5
6
30
-8
-6
-4
-2
0
2
4
6
time from FMP
UI: cluster 1
UI incidence rate
0.0
0.4
0.8
time from FMP
subjects in cluster 2
log(E2)
211
time from FMP
UI: cluster 2
UI incidence rate
time from FMP
subjects in cluster 3
log(E2)
343
time from FMP
UI: cluster 3
UI incidence rate
time from FMP
subjects in cluster 4
log(E2)
151
time from FMP
UI: cluster 4
UI incidence rate
time from FMP

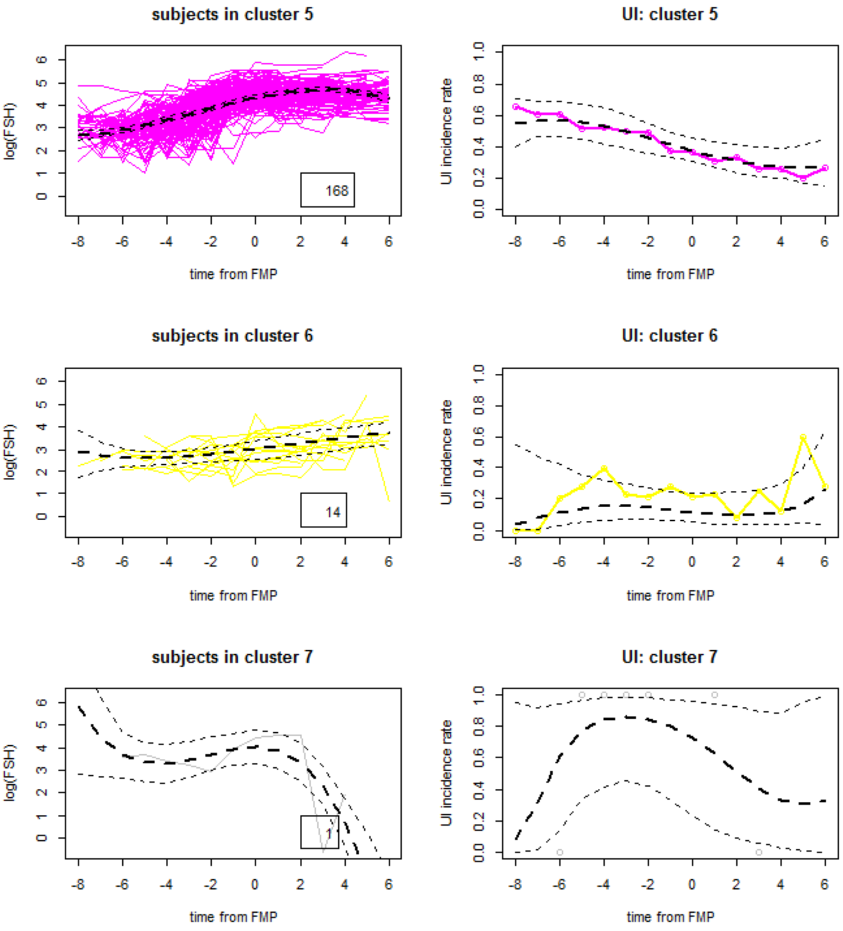


Figure 4.18: Seven clusters have been identified from the joint modeling of FSH and UI, and each row has a pair of graphs which plot the FSH (left) and UI (right) patterns for each cluster. The estimated cluster mean trajectories with 95% probability bands are plotted in black. And the sample data are plotted in different colors for different clusters. The spaghetti plots are made for FSH, and the sample proportions are plotted for UI incidence.

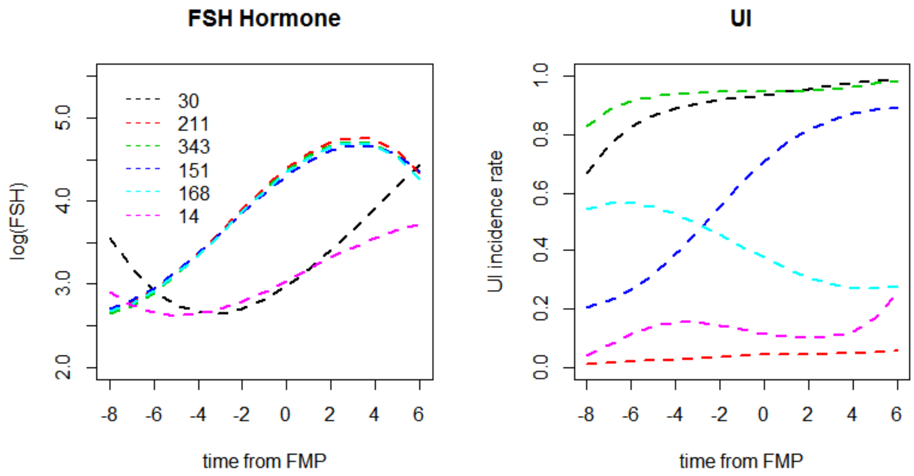


Figure 4.19: The estimated mean trajectories for the seven clusters separated by different colors.

women) in which FSH starts to increase in the pre- or early peri- menopause stage and reaches the maximum around FMP. Clusters 2, 3, 4 and 5 all have this profile. The other is a “late rise” profile (4.8%) in which FSH starts to increase in the late peri-menopause, and both the increase rate and magnitude are lower than for the “gradual rise” profile.

Similar to the E2 results, there are four different UI developmental patterns corresponding to the main “gradual rise” FSH profile: consistently affected by UI (37.4%, green line), consistently healthy (23.0%, red), gradually developed UI (16.5%, blue) and easing of UI symptoms (18.3%, light blue).

We identified 44 women having the “late rise” FSH profile. There are only two UI developmental patterns corresponding to this FSH profile: (i) consistently affected by UI (3.3%, black) and (ii) consistently healthy (1.5%, magenta). We do not see women with the FSH profile changing UI status from the clustering results. However, it is hard to say whether there are other UI patterns corresponding to this profile in reality due to the small sample size.

4.5 Analysis of hormone profiles with scalar UI response

One question of interest is to predict whether a woman who did not have UI at baseline would develop UI during menopause or not. This prediction might be improved by considering the relation between hormone patterns and the incidence of UI. In order to answer the question, we need to process the data first so that the new dataset fits our analysis.

Since we are interested whether a woman developed UI during menopause or not, we only keep the women who did not have UI at baseline in the sample. We have 456 women with 4794 observations left for the analysis. In addition, we create a scalar UI response denoted by v_i to indicate whether woman i developed UI. We let

$$v_i = \begin{cases} 0 & \text{if } u_{ij} = 0 \text{ for all } j, j = 0, \dots, r_i \\ 1 & \text{otherwise} \end{cases}$$

A common strategy for the analysis is to define several summaries of the hormone patterns, such as the rate of change and average value across a region. These summaries can then be plugged in as predictors in a generalized linear model (GLM) for the UI development response. Unfortunately, it is typically not clear how best to choose summaries of the function, and multicolinearity becomes a concern if many summaries are chosen.

Bigelow and Dunson (2009) [3] proposed a semi-parametric Bayesian approach for assessing the relationship between functional predictors and a scalar response, which allocates women to clusters that are defined in terms of the womans progesterone trajectory and risk of early pregnancy loss (EPL) for the North Carolina Early Pregnancy Study. This model builds the association between the scalar response EPL and the functional predictor progesterone trajectory by matching a probability of EPL with each progesterone pattern using the DP.

We use this joint modeling idea, and combine it with our clustering method to build a model for our SWAN data analysis.

Our joint model is built as follows:

$$
\begin{aligned}
v_i|p_i &\sim Bern(p_i) \\
logit(p_i) &= \pi_i + z_i\omega \\
y_i|\beta_i &\sim N_{r_i}(X_i\beta_i, \tau_e^{-1}W_i) \\
(\beta_i, \pi_i)|G &\overset{iid}{\sim} G \\
G &\sim DP(\alpha, G_0) \\
G_0 &= N_{p+1}(\mu_\beta, \Xi_\beta^{-1}) * N_{p+1}(\mu_\lambda, \Xi_\lambda^{-1}) \qquad (4.3)
\end{aligned}
$$

where both β and π are modeled with DP, so that each hormone profile corresponds to a UI incidence level in the clustering result.

This model is actually the simplified version of model (4.2) by assuming the UI incidence rate to be constant with time. So we can easily adapt the prior specification and sampling algorithm from model (4.2) to obtain results for this analysis.

We applied the model to scalar UI responses and E2 data to find the association between E2 profile and scalar UI response, adjusted by BMI and age to FMP. The inference shows that BMI is positively correlated with UI incidence, but age at FMP is not. Table 4.9 shows the detailed estimates of global parameters.

We identified two clusters of E2 profiles together with UI incidence rate. Figure 4.20 plots the two E2 profiles. Compared to the clustering results with all 928 women in Section 3, the flat E2 pattern is not showing up in this analysis. We have already found that the flat E2

Table 4.9: Global Parameter Estimates for the joint clustering model applied to FSH and scalar UI data.

			95% PI	
Parameter	Mean	Median	Lower	Upper
σ_e	0.72	0.72	0.70	0.73
λ_{bmi}	0.30	0.30	0.09	0.51
λ_{age}	-0.09	-0.09	-0.28	0.11
γ_0	9.7	9.5	7.0	14.1
γ_1	51.5	36.6	12.0	200.8
γ_2	38.7	20.4	10.5	243.1
γ_3	194.7	153.3	41.3	570.1
α	0.28	0.23	0.03	0.78

pattern is associated with high UI incidence through menopause. It makes sense that the pattern was not singled out because most of the women with it have UI at baseline and have been removed from the analysis.

The logarithm of the odds ratio for UI incidence between the two clusters is the difference $\pi_{s=1} - \pi_{s=2}$. Its posterior median is 0.06 with 95% probability interval $(-0.52, 0.61)$, which means the UI incidence rates of the two clusters are not statistically different. We obtained the two clusters mostly due to the difference of E2 profiles.

4.6 Conclusions

While our analysis is disappointing, we point out that it may not have been an ideal analysis to perform in the first place. Since all the women are observed over different times and are actually starting at different times with the study ending for them at different times, there is no single time period over which all women are exposed to the risk of UI. Thus some women may have much shorter exposure times than others and thus their probabilities of getting UI should be adjusted in some way to account for this fact. Since this is a dissertation and not a paper, we include the analysis to help our future thinking on the topic.

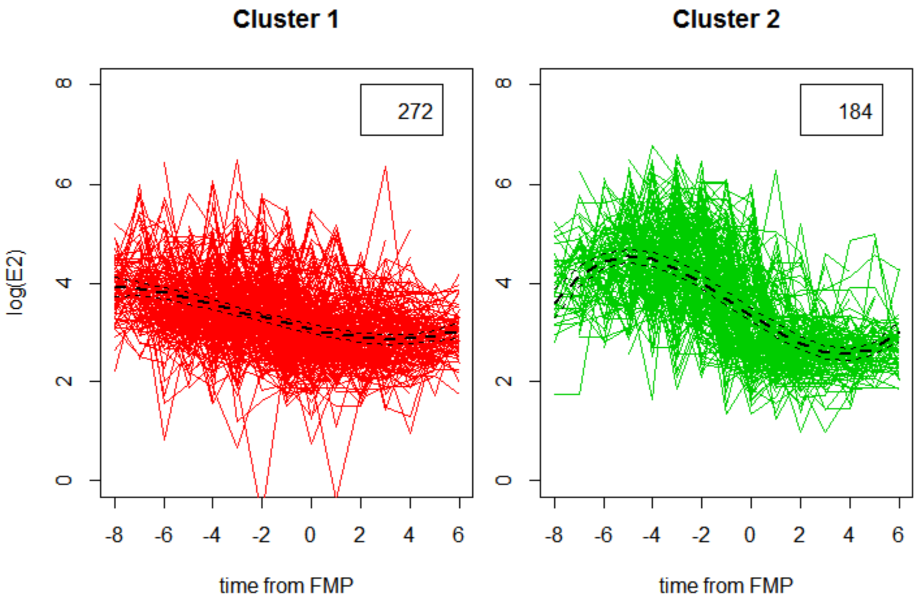


Figure 4.20: The spaghetti plots of the E2 data together with the estimated mean E2 trajectories for the two clusters.

Chapter 5

Joint Modeling of Bivariate Longitudinal Screening Data

Diagnostic screening involves testing humans or animals for the presence of disease or infection. For some diseases, a “gold-standard” test does not exist or is too invasive or expensive to use. Hence, the goals of diagnostic testing may include: quantifying the performance of an imperfect test, diagnosing individuals, and estimating disease prevalence in the absence of a perfect reference test.

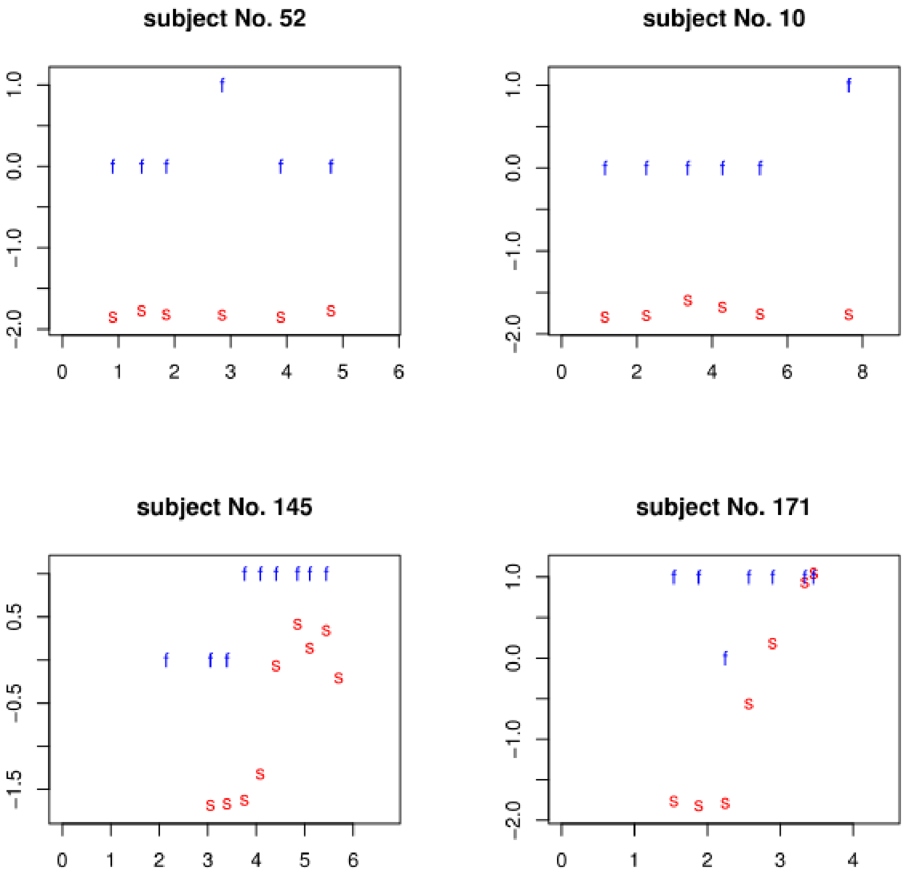
Our work is focused on developing a model for bi-variate longitudinal diagnostic outcomes in the no-gold standard case. We consider the situation where an imperfect binary test is repeatedly administered to each individual together with a continuous measure, which can be used to help the disease diagnosis. For infected individuals, we assume the existence of a change-point corresponding to time of infection and posit appropriate changes to model the responses thereafter.

In this chapter, we briefly introduce the data in section one. In section two, we propose a Bayesian hierarchical joint model for the longitudinal diagnostic outcomes. We specify prior distributions in section three and illustrate posterior sampling algorithms in section four. The reversible jump MCMC and Metropolis-within-Gibbs algorithms are also explained in that section. We check our model with simulated data and show results in section six. In section seven, we apply the model to the longitudinal screening data for Johne's disease and present statistical inferences. In section eight, we identify different developmental trajectories for serology scores using a Dirichlet Process Mixture model. Some concerns about the model are also discussed there.

5.1 Background

The motivating dataset for this study consists of joint longitudinal screening data for Johne's Disease (JD) in cattle. Johne's disease is a chronic bacterial infection caused by Mycobacterium avium subspp paratuberculosis (Map), which can cause weight loss, reduced milk production, edema and diarrhea. However, these signs may not manifest themselves for months to years after infection, if at all. Early diagnosis in this asymptomatic phase is desirable since infected cows may pass the infection on to herd mates whether or not they are exhibiting signs of infection.

Our data consist of records from 12 dairy herds known to be infected with JD [25]. Data were collected from 1984-2003, and tests were performed about every six months. There was, however, substantial deviation from this testing schedule with one-fourth of inter-test times below 4 months and one-fourth above 8.3 months. Herd size ranged from 50 to 160 milking cows (median = 60). Three hundred and sixty five cows from this study were included in our analysis; the number of observations for each cow ranged from 2 to 23 with a median of 6. At each screening time, both fecal culture (FC) and serum ELISA tests were performed, although for various reasons, either of these may be missing on a given test date. FC test results were categorized as positive if at least one Map bacterial colony formed in culture


subject No. 52

subject No. 10

subject No. 145

subject No. 171

Figure 5.1: Data plots for four selected cows, ‘f’ indicates the binary FC test outcome, and ‘s’ represents the ELISA test serology score.

and were categorized negative otherwise. ELISA tests measure antibody concentration on a continuous scale through the Optical Density (OD), which is standardized relative to positive and negative control sera on each test plate.

We selected 4 cows from the dataset to plot their serology scores and FC test results in Figure 5.1. Subject 52 represents cows that are likely uninfected since all but one FC results are negative and since all serology scores are consistently at a low level. The fact that there is one positive result among six FC observations indicates that the FC test may not be perfect. Subjects 145 and 171 are likely infected due to multiple positive FC outcomes and increasing serology scores.

Observe that there is evidently an appreciable delay before ELISA outcomes begin to increase after FC tests are positive. It is well-known that there is a lag after infection before a serological response is mounted; the lag is typically 10-17 months (Lepper, 1989 [20]). Our method will reflect this issue. A true positive FC outcome, on the other hand, can occur much sooner after infection.

The existence of the lag leads to a third infection state, which we call the "intermediate" stage. This state corresponds to cows that have been infected with JD, and which have not shown any serology reaction due to the lag, at the end of the study. For example, subject 10 in Figure 5.1 might be in this state since its last FC was positive, while its serology score does not show any reaction, as of the end of the study.

5.2 Model Specification

In this section, a finite mixture model with three latent states is presented to model FC outcomes and serology scores jointly. In addition, we specify the sigmoid function and its parameterization that is used to fit the serologic trend.

5.2.1 Joint Modeling of Serology and Fecal Culture Outcomes

We first define the three infection states mentioned above, since one primary goal of our analysis is to diagnose the disease status of all cows. We use a latent variable k_i to denote the infection state, and $k_i = 1, 2, 3$ represents the "uninfected", "intermediate" and "infected" states, respectively.

Let $S_i = \{S_{i1}, \ldots, S_{im_i}\}$ and $F_i = \{F_{i1}, \ldots, F_{im_i}\}$ be the serology and FC outcomes collected at times $t_i = \{t_{i1}, \ldots, t_{im_i}\}$, for the i^{th} individual. The dataset we have is $\mathbf{D} = \{(S_i, F_i) : i =$

$1, \ldots, n\}$. In this section, we propose a Bayesian hierarchical trans-dimensional model for the joint longitudinal diagnostic outcomes, where the word "trans-dimensional" means the dimension of parameter space in each state is allowed to vary. In this model, we separate the parameter space into global parameters and "cow-specific" parameters, which correspond to individual cows. We use Θ to denote the global parameters and $\phi_i^{k_i}$ to denote the "cow-specific" parameters. Note the dimension of $\phi_i^{k_i}$ is different for different k_i.

Consider the i^{th} cow in the data. If $k_i = 1$, we model the response jointly as below:

$$\begin{aligned} S_{ij}|(k_i = 1, \phi_i^1), \Theta &\overset{\perp}{\sim} \beta_{0i} + \epsilon_{ij} \\ F_{ij}|(k_i = 1, \phi_i^1), \Theta &\overset{\perp}{\sim} \text{Bern}(1 - sp) \end{aligned} \tag{5.1}$$

where S_{ij} and F_{ij} are regarded to be independent; β_{0i} is the base serology score of cow i, and $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \tau_e)$ is random error. There is only one "state-specific" effect in state one, β_{0i}. In addition, we let sp be the specificity of FC test and model F_{ij} with a Bernoulli distribution, since the test is not perfect.

If $k_i = 2$, cow i is in the "intermediate" state, where the cow is infected but due to a lag, there is no serologic reaction to the infection. The model for S_{ij} is the same as it was for state one, however the model for F_{ij} is different:

$$\begin{aligned} S_{ij} \mid (k_i = 2, \phi_i^2), \Theta &\overset{\perp}{\sim} \beta_{0i} + \epsilon_{ij} \\ F_{ij} \mid (k_i = 2, \phi_i^2), \Theta &\overset{\perp}{\sim} \begin{cases} \text{Bernoulli}(1 - sp) & \text{if } t < t_i^\star \\ \text{Bernoulli}(se) & \text{if } t \geq t_i^\star \end{cases} \end{aligned} \tag{5.2}$$

where $t_i^\star$ is the infection time for individual i, and $\phi_i^2 = \{\beta_{0i}, t_i^\star\}$. We let se be the sensitivity of FC test. Clearly, the probability of a positive FC outcome is different before and after the disease infection $t_i^\star$.

If $k_i = 3$, cow i is infected with JD and has shown serology increase. The model is constructed as below:

$$
\begin{aligned}
S_{ij} \mid (k_i = 3, \phi_i^3), \Theta &\overset{\perp}{\sim} \beta_{0i} + g(t_{ij}|\phi_i^3) + \epsilon_{ij} \\
F_{ij} \mid (k_i = 3, \phi_i^3), \Theta &\overset{\perp}{\sim} \begin{cases} \text{Bernoulli}(1 - sp) & \text{if } t < t_i^\star \\ \text{Bernoulli}(se) & \text{if } t \geq t_i^\star \end{cases}
\end{aligned}
\tag{5.3}
$$

where $g(t|\phi)$ is the function used to fit the serology score after infection since it increases after a lag from infection time.

In fact, this model could be regarded as a finite mixture model for cluster analysis, which is used to classify cows into the three infection states (clusters). With this mixture model, we are able to estimate whether and when a cow was infected with JD.

We mention that we have implemented cluster analysss using the Dirichlet Process when analyzing SWAN data in previous chapters. We prefer the DP there because clustering is model based. In this study, we know that there are only three infection states (clusters).

5.2.2 Sigmoid Function

In model (5.3), we use a function, $g(t|\phi)$, to fit the developmental trajectory for the serology scores, and there are numerous choices for the function. For example, Norris, Johnson and Gardner (2009) [25] used a linear function to fit the developmental trajectory. However, a linear function may be an oversimplification of the serology trend. For example, it may be expected that the trend would increase gradually and then, at some stage, level off; sigmoid

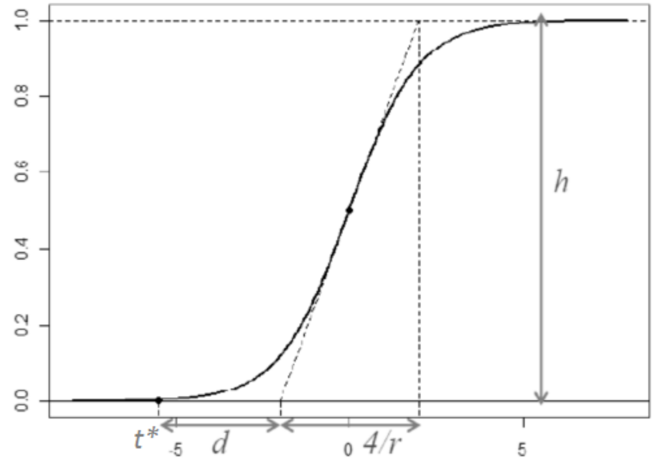


Figure 5.2: Four-parameter sigmoid function with reparameterization

functions come to mind. We choose a four-parameter sigmoid curve to fit serology scores after the infection time $t_i^\star$ in state 3. The function was briefly introduced in chapter one.

In our data analysis, we use a re-parameterized sigmoid function for the curve fitting compared with Equation (1.22), and a new set of parameters is $\{t^\star, d, h, r\}$. The mathematical form of the function is:

$$g(t|\phi = \{t^\star, d, h, r\}) = (\frac{h}{1 + e^{-r(t-t^\star-d-2/r)}} - \frac{h}{1 + e^{r(d+2/r)}})I(t \geq t^\star) \tag{5.4}$$

Figure 5.2 displays that the parameters have nice interpretations; $t^\star$ is the infection time, h is the upper bound of the maximum serology increase an individual could achieve ($max(g(t|\phi)) = h(1 - \frac{1}{1+e^{2+rd}}) \rightarrow h$), d is regarded as a kind of lag time of the serologic reaction, and r is related to the rate of increase in the function. When $r > 0$, the curve is increasing as shown in the figure. When $r < 0$, the curve is monotonically decreasing. In addition, the curve reaches its maximum changing rate, $hr/4$, at the half-height point of the sigmoid function, namely,

$$\frac{d\,g(t|\phi)}{d\,t}\, |_{t=t^\star+d+2/r} = \frac{hre^{-r(t-t^\star-d-2/r)}}{(1+e^{-r(t-t^\star-d-2/r)})^2}\, |_{t=t^\star+d+2/r} = hr/4$$

Thus considering the tangent line at the half-height point, the point of intersection on the x-axis is $t^\star + d$. The steeper is the slope at the half-height point, the flatter is the curve to the left of $t^\star + d$, indicating less serologic reaction up to that point in time, but with a strong reaction after that point in time. Thus instead of having a fixed lag time as was done in Norris et al (2009), we regard d in our sigmoid model as reflecting lag time, which is allowed to be different for all cows in stage 3. Since each cow in state 3 will be modeled to have its own sigmoid function, these parameters will be actually modeled as random effects.

5.2.3 Likelihood

With the models specified above, we can write the likelihood contribution for each cow, i, below:

$$
\begin{aligned}
f\big(S_i, F_i \mid (k_i, \phi_i^{k_i}), \Theta\big) &= f(S_i \mid (k_i, \phi_i^{k_i}), \Theta)\, f(F_i \mid (k_i, \phi_i^{k_i}), \Theta\big) \qquad (5.5)\\
\propto &\Big[(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2} sp^{1-F_{ij}} (1-sp)^{F_{ij}}) \Big]^{I(k_i=1)}\\
&\cdot \Big[\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2} \big) \big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}} (1-sp)^{F_{ij}} \big) \big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}} (1-se)^{1-F_{ij}} \big) \Big]^{I(k_i=2)}\\
&\cdot \Big[\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_{k_i=3})\right)^2} \big) \big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}} (1-sp)^{F_{ij}} \big) \big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}} (1-se)^{1-F_{ij}} \big) \Big]^{I(k_i=3)}
\end{aligned}
$$

where $g(t|\phi) = 0$ for $t < t_i^\star$.

Let $(\mathbf{k}, \boldsymbol{\phi}) = \{(k_i, \phi_i^{k_i}) : i = 1, \ldots, n\}$. Then we obtain the likelihood function:

$$
\begin{aligned}
L\big((\mathbf{k}, \boldsymbol{\phi}), \Theta\big) &= \prod_{i=1}^{n} f\big(S_i, F_i \mid (k_i, \phi_{k_i}), \Theta\big) \qquad (5.6) \\
&\propto \prod_{i:k_i=1} \Big[\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2} sp^{1-F_{ij}} (1-sp)^{F_{ij}}\big) \Big] \\
&\cdot \prod_{i:k_i=2} \Big[\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2} \big) \big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}} (1-sp)^{F_{ij}} \big) \big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}} (1-se)^{1-F_{ij}} \big) \Big] \\
&\cdot \prod_{i:k_i=3} \Big[\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2} \big) \big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}} (1-sp)^{F_{ij}} \big) \big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}} (1-se)^{1-F_{ij}} \big) \Big]
\end{aligned}
$$

5.3 Prior Specification

We specify prior distributions for parameters in this section, including both cow-specific and global parameters.

5.3.1 Cow-specific Effects

First, we assign a multinomial distribution to latent variable k_i:

$$k_i \sim \text{Multinomial}(q_1, q_2, q_3)$$

where q_k is the proportion of cows in state k for $k = 1, 2, 3$.

Then, we assign distributions to the cow-specific effects, ϕ_i^k, for each k value. In state 1 ($k = 1$), there is only one such parameter, β_{0i}. We let the base serology score $\beta_{0i} \overset{\perp}{\sim} N(\mu_{\beta_0}, \tau_{\beta_0})$. Since β_{0i} exists in all three states and has the same interpretation, we keep using the same distribution for β_{0i} in the three models corresponding to the three states; μ_{β_0} is the overall

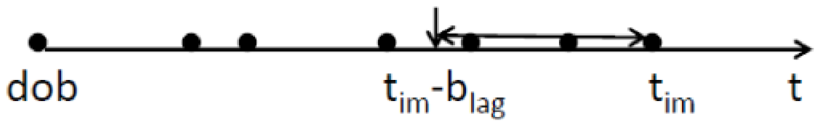


Figure 5.3: The interval for infection time in state 2

mean serology level before infection.

In state 2, $\phi_i^{k_i=2} = \{\beta_{0i}, t_i^\star\}$. We model uncertainty about $t_i^\star$ as Uniform$(t_{im_i} - b_{lag}, t_{im_i})$ as shown in Fig 5.3; t_{im_i} is the last screening time for cow i, and $b_{lag} = 17$ months is the upper bound of possible lag times. The serology trajectory is a horizontal line since the serology score has not started to rise in the screening window. Thus all cows in state 2 would be infected inside a b_{lag} length window prior to their last screening time.

The state-specific effects in state 3 include $\phi_i^{k_i=3} = \{\beta_{0i}, t_i^\star, h_i, d_i, r_i\}$. The modeled distribution for $t_i^\star$ is different from state 2. In state 3, we assign it a uniform distribution Uniform$(dob_i, t_{im_i} - a_{lag})$ as shown in Fig 5.4, where dob_i is the i^{th} cow's date of birth, and a_{lag} is the lower bound for the lag time, which is 4 months. The upper bound for the distribution is $t_{im_i} - a_{lag}$ instead of t_{im_i} because there is a lag for serologic reaction and we should not have observed if the cow was infected after $t_{im_i} - a_{lag}$, if they are in stage 3. The other three effects, $\{h_i, d_i, r_i\}$, determine the shape of the sigmoid function. We assign distributions as:

$$\begin{aligned} log(d_i) &\sim N(\mu_d, \tfrac{1}{\tau_d}) \\ log(h_i - c_h) &\sim N(\mu_h, \tfrac{1}{\tau_h}) \\ log(r_i) &\sim N(\mu_r, \tfrac{1}{\tau_r}) \end{aligned}$$

where c_h is a positive number, which forces $h_i > c_h$. This model in state 3 is the same as the model in state 2 when $h_i \to 0$. We choose the cutoff $c_h > 0$ since otherwise, it is difficult to discriminate between states 2 and 3 for some cows. In our analysis, we let $c_h = 0.2$.

We can write the pdf for the state-specific parameters $(k_i, \phi_i^{k_i})$ corresponding to each indi-

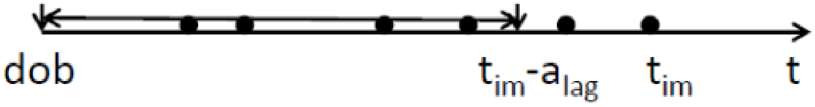


Figure 5.4: The interval for infection time in stage 3

vidual i:

$$\begin{aligned}
f\big((k_i,\phi_i^{k_i}) \mid \Theta\big) =\; & f(\phi_i^{k_i} \mid k_i,\Theta)\, f(k_i \mid \Theta) \\
\propto\; & \Big(q_1 \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2}\Big)^{I(k_i=1)} \Big(q_2 \frac{I_{(t_{im_i}-b_{lag},t_{im_i})}(t_i^\star)}{b_{lag}} \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2}\Big)^{I(k_i=2)} \\
& \Big(q_3 \frac{I_{(dob_i,t_{im_i}-a_{lag})}(t_i^\star)}{t_{im_i}-a_{lag}-dob_i} \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2} \frac{\tau_h^{\frac{1}{2}}}{h_i-c_h} e^{-\frac{\tau_h}{2}(\log(h_i-c_h)-\mu_1)^2} \\
& \frac{\tau_r^{\frac{1}{2}}}{r_i} e^{-\frac{\tau_r}{2}(\log r_i-\mu_2)^2} \frac{\tau_d^{\frac{1}{2}}}{d_i} e^{-\frac{\tau_d}{2}(\log d_i-\mu_3)^2}\Big)^{I(k_i=3)}
\end{aligned} \tag{5.7}$$

We obtain the joint density for all state parameters as follow:

$$\begin{aligned}
f\big((\mathbf{k},\boldsymbol{\phi}) \mid \Theta\big) =\; & \prod_{i=1}^{n} f(\phi_i^{k_i} \mid k_i,\Theta)\, f(k_i \mid \Theta) \\
\propto\; & \prod_{i:k_i=1} \Big(q_1 \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2}\Big) \cdot \prod_{i:k_i=2} \Big(q_2 \frac{I_{(t_{im_i}-b_{lag},t_{im_i})}(t_i^\star)}{b_{lag}} \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2}\Big) \\
& \prod_{i:k_i=3} \Big(q_3 \frac{I_{(dob_i,t_{im_i}-a_{lag})}(t_i^\star)}{t_{im_i}-a_{lag}-dob_i} \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2} \frac{\tau_h^{\frac{1}{2}}}{h_i-c_h} e^{-\frac{\tau_h}{2}(\log(h_i-c_h)-\mu_1)^2} \\
& \quad \frac{\tau_r^{\frac{1}{2}}}{r_i} e^{-\frac{\tau_r}{2}(\log r_i-\mu_2)^2} \frac{\tau_d^{\frac{1}{2}}}{d_i} e^{-\frac{\tau_d}{2}(\log d_i-\mu_3)^2}\Big)
\end{aligned} \tag{5.8}$$

5.3.2 Global Parameters

We now specify the prior distributions for the global parameters Θ, which are

$$\Theta = \{q_1, q_2, q_3, se, sp, \tau_e, \beta_0, \tau_{\beta_0}, \mu_h, \tau_h, \mu_r, \tau_r, \mu_d, \tau_d\}$$

where (q_1, q_2, q_3) are the proportions of cows in the 3 stages, respectively. We assign a Dirichlet prior as:

$$(q_1, q_2, q_3) \sim \text{Dirichlet}(a_{q1}, a_{q2}, a_{q3}) \tag{5.9}$$

where a_{q1}, a_{q2}, a_{q3} are pre-selected numbers. We let them be 1 in our analysis.

Now (se, sp) are the sensitivity and specificity of the FC test. We assign independent beta distributions to both of them;

$$\begin{aligned} se &\sim \text{Beta}(a_{se}, b_{se}) \\ sp &\sim \text{Beta}(a_{sp}, b_{sp}) \end{aligned} \tag{5.10}$$

where we let $a_{se} = b_{se} = a_{sp} = b_{sp} = 1$. Note better priors could be used for se and sp since they both are greater than 50%. For example, uniform$(0.5, 1)$ can be used as their prior.

The prior distributions for the rest of global parameters are listed below. Conventionally, we assigned normal priors to means μ and gamma priors to precisions τ. Namely,

$$\begin{aligned}
\tau_e &\sim \Gamma(a_{\tau_e}, b_{\tau_e}) & & \\
\mu_{\beta_0} &\sim N(u_{\beta_0}, v_{\beta_0}) & \tau_{\beta_0} &\sim \Gamma(a_{\beta_0}, b_{\beta_0}) \\
\mu_d &\sim N(u_d, v_d) & \tau_d &\sim \Gamma(a_d, b_d) \\
\mu_h &\sim N(u_h, v_h) & \tau_h &\sim \Gamma(a_h, b_h) \\
\mu_r &\sim N(u_r, v_r) & \tau_r &\sim \Gamma(a_r, b_r)
\end{aligned} \tag{5.11}$$

where $a., b., u., v.$ are all hyper-parameters. We use diffuse prior for precisions τ with $a_{\tau_e} = a_{\beta_0} = b_{\tau_e} = b_{\beta_0} = 0.001$, $a_d = a_h = a_r = 2$ and $b_d = b_h = b_r$=0.01. For the mean, we let $u_d = u_h = u_r = 0$, $v_d = v_h = v_r = 1$, $u_{\beta_0} = -2$ and $v_{\beta_0} = 1$ based on the expert elicited prior information. For example, we let $\mu_{\beta_0} \sim N(-2, 1)$ since the expert were at least 95% confident that μ_{β_0} is in the range $(-4, 0)$. In addition, we tried different hyper-parameter values for the prior distributions, and the posterior inferences are consistent with the choices.

5.4 Posterior Sampling

Combining the likelihood (Equation 5.6) and the prior distributions (Equation 5.8, 5.9, 5.10 and 5.11), we obtain the posterior distribution using Bayes rule:

$$f\big((\mathbf{k}, \boldsymbol{\phi}), \Theta | \mathcal{D}\big) \propto L\big((\mathbf{k}, \boldsymbol{\phi}), \Theta\big)\, f\big((\mathbf{k}, \boldsymbol{\phi}) \mid \Theta\big)\, f(\Theta) \tag{5.12}$$

5.4.1 Gibbs Sampling

Since the posterior distribution is not recognizable, a Gibbs sampler was employed for posterior sampling. The full conditional distributions for the majority of the parameters are recognizable due to conditional conjugacy, and we can easily draw samples from them. For example, the gamma distribution is commonly used as a conditionally conjugate prior for precisions τ. So we can obtain the full conditional distribution of τ_e easily as below:

$$\tau_e|else \sim \Gamma\Big(a_{\tau_e} + \tfrac{\sum_i m_i}{2}, b_{\tau_e} + \sum_{i:k_i=1,2} \sum_j \tfrac{1}{2}\big(S_{ij} - \beta_{0i})^2 + \sum_{i:k_i=3} \sum_j \tfrac{1}{2}(S_{ij} - \beta_{0i} - g(t_{ij}|\phi_{k_i=3})\big)^2\Big) \tag{5.13}$$

Another example is the prior distribution $(q_1, q_2, q_3) \sim \text{Dirichlet}(a_{q_1}, a_{q_2}, a_{q_3})$, which is conjugate to the Multinomial(q_1, q_2, q_3). The full conditional distribution for (q_1, q_2, q_3) is:

$$(q_1, q_2, q_3)|else \sim \text{Dirichlet}(n_1 + a_{q_1}, n_2 + a_{q_2}, n_3 + a_{q_3}) \tag{5.14}$$

where $n_k, k = 1, 2, 3$, is the number of cows in state k.

We list all the recognizable full conditional distributions below:

$$\begin{aligned}
se|else &\sim \text{Beta}\Big(a_{se} + \sum_{i:k_i=2,3} \sum_{j:t_{ij} \geq t_i^\star} F_{ij}, b_{se} + \sum_{i:k_i=2,3} \sum_{j:t_{ij} \geq t_i^\star} (1 - F_{ij})\Big) \\
sp|else &\sim \text{Beta}\Big(a_{sp} + \sum_{i:k_i=1} \sum_j (1 - F_{ij}) + \sum_{i:k_i=2,3} \sum_{j:t_{ij} < t_i^\star} (1 - F_{ij}\Big), \\
&\qquad b_{sp} + \sum_{i:k_i=1} \sum_j F_{ij} + \sum_{i:k_i=2,3} \sum_{j:t_{ij} < t_i^\star} F_{ij} \ ) \\
\mu_{\beta_0}|else &\sim N\Big(\frac{\tau_{\beta_0} \sum_i \beta_{0i} + v_{\beta_0}^{-1} u_{\beta_0}}{n\tau_{\beta_0} + v_{\beta_0}^{-1}}, (n\tau_{\beta_0} + v_{\beta_0}^{-1})^{-1}\Big) \\
\tau_{\beta_0}|else &\sim \Gamma\Big(a_{\beta_0} + \frac{n}{2}, b_{\beta_0} + \frac{\sum_i (\beta_{0i} - \mu_{\beta_0})^2}{2}\Big)
\end{aligned} \tag{5.15}$$

$$
\begin{aligned}
\mu_h|else &\sim N\Big(\frac{u_h v_h^{-1} + \tau_h \sum_{i:k_i=3}\log(h_i - c_h)}{n_3\tau_h + v_h^{-1}}, (n_3\tau_h + v_h^{-1})^{-1}\Big)\\
\tau_h|else &\sim \Gamma\Big(a_h + \frac{n_3}{2}, b_h + \frac{\sum_{i:k_i=3}(\log(h_i - c_h) - \mu_h)^2}{2}\Big)\\
\mu_r|else &\sim N\Big(\frac{u_r v_r^{-1} + \tau_r \sum_{i:k_i=3}\log r_i}{n_3\tau_r + v_r^{-1}}, (n_3\tau_r + v_r^{-1})^{-1}\Big)\\
\tau_r|else &\sim \Gamma\Big(a_r + \frac{n_3}{2}, b_r + \frac{\sum_{i:k_i=3}(\log r_i - \mu_r)^2}{2}\Big)\\
\mu_d|else &\sim N\Big(\frac{u_d v_d^{-1} + \tau_d \sum_{i:k_i=3}\log d_i}{n_3\tau_d + v_d^{-1}}, (n_3\tau_d + v_d^{-1})^{-1}\Big)\\
\tau_d|else &\sim \Gamma\Big(a_d + \frac{n_3}{2}, b_d + \frac{\sum_{i:k_i=3}(\log d_i - \mu_d)^2}{2}\Big)
\end{aligned}
$$

5.4.2 Reversible-Jump MCMC

There is difficulty in sampling from $(k_i, \phi_{k_i})|else$, since the three infection states have different dimension of parameter spaces. We use a reversible-jump MCMC algorithm to solve the problem, which has been introduced in chapter 1. According to Equation (1.20), we need to construct a transition kernel $q(k \to k')q_{k\to k'}(u)$ for the reversible jump algorithm.

We first let $q(k \to k') = \frac{1}{3}$ for any $k, k' \in \{1, 2, 3\}$, so that an individual has equal probability to move to one of the three infection states at each MCMC iteration. When $k' = k$, the individual stays in the same state as in the previous iteration, where the acceptance rate $\alpha(k, k' = k) = 1$.

We use the reversible jump algorithm for the jumps between different infection states k and k'. There are three different jumps we have to consider: $k = 1 \leftrightarrows k' = 2$, $k = 1 \leftrightarrows k' = 3$ and $k = 2 \leftrightarrows k' = 3$.

In the Appendix, we illustrate the reversible-jump MCMC sampler using a move from state

1 ($k = 1$) to state 3 ($k' = 3$) in detail. The acceptance rate for the move from state 1 to state 3 is

$$\alpha(k = 1, k' = 3) = min(1, \alpha_{13}) \tag{5.16}$$

where

$$\begin{aligned}\alpha_{13} \quad = \quad & \frac{q_3 \frac{1}{t_{im_i} - a_{lag} - dob_i}}{q_1 \psi(t_i^\star)} \cdot \frac{\Big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2} \left(S_{ij} - \beta_{0i} - g(t_{ij} | \phi_i^3) \right)^2} \Big)}{\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2} (S_{ij} - \beta_{0i})^2}\big)} \\ & \cdot \frac{\Big(\prod_{j: t_{ij} < t_i^\star} sp^{1-F_{ij}} (1 - sp)^{F_{ij}} \Big) \Big(\prod_{j: t_{ij} \geq t_i^\star} se^{F_{ij}} (1 - se)^{1-F_{ij}} \Big)}{\prod_{j=1}^{m_i} sp^{1-F_{ij}} (1 - sp)^{F_{ij}}}\end{aligned}$$

Similarly, we obtain the acceptance rate for the move from state 1 to state 2 and the move from state 2 to state 3:

$$\alpha(k = 1, k' = 2) = min(1, \alpha_{12}); \quad \alpha(k = 2, k' = 3) = min(1, \alpha_{23}) \tag{5.17}$$

where

$$
\begin{aligned}
\alpha_{12} \quad &= \quad \frac{q_2 \frac{1}{b_{lag}}}{q_1 \psi_2(t_i^\star)} \cdot \frac{\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2}\big)}{(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2})} \\
&\cdot \frac{\big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big)\big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\big)}{\prod_{j=1}^{m_i} sp^{1-F_{ij}}(1-sp)^{F_{ij}}} \\
\alpha_{23} \quad &= \quad \frac{q_3 \frac{1}{t_{im_i}-a_{lag}-dob_i}\psi_2(t_i^\star)}{q_2 \psi_3(t_i^{\star\prime})\frac{1}{b_{lag}}} \cdot \frac{\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2}\big)}{(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2})} \\
&\cdot \frac{\big(\prod_{j:t_{ij}<t_i^{\star\prime}} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big)\big(\prod_{j:t_{ij}\geq t_i^{\star\prime}} se^{F_{ij}}(1-se)^{1-F_{ij}}\big)}{\big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big)\big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\big)}
\end{aligned}
$$

Since the constructed Markov chain is reversible, we can easily get the acceptance rates for the remaining moves:

$$
\begin{aligned}
&\alpha(k=3,k'=1) = min(1,\alpha_{31}) \quad \text{where } \alpha_{31} = \alpha_{13}^{-1} \\
&\alpha(k=2,k'=1) = min(1,\alpha_{21}) \quad \text{where } \alpha_{21} = \alpha_{12}^{-1} \\
&\alpha(k=3,k'=2) = min(1,\alpha_{32}) \quad \text{where } \alpha_{32} = \alpha_{23}^{-1}
\end{aligned}
\tag{5.18}
$$

5.4.3 Within-State Moves

At each MCMC iteration, we resample $\phi_i^{k_i}$ within state k_i in order to accelerate the convergence of the Markov chain. Consider $\phi_i^{k_i=1} = \beta_{0i}$ first. The posterior sample can be easily drawn from the following distribution due to conjugacy,

$$\beta_{0i}|k_i = 1, else \sim N\big(\frac{\tau_e \sum_{j=1}^{m_i} S_{ij} + \tau_{\beta_0}\beta_0}{\tau_e m_i + \tau_{\beta_0}}, (\tau_e m_i + \tau_{\beta_0})^{-1}\big) \tag{5.19}$$

The cow-specific parameters are $\phi_i^{k_i=2} = \{\beta_{0i}, t_i^\star\}$ for state 2. The full conditional distribution for $\beta_{0i}|k_i = 2, else$ is the same as Equation (5.19), since the model for serology scores is the same for states 1 and 2,

$$\beta_{0i}|k_i = 2, else \sim N\big(\frac{\tau_e \sum_{j=1}^{m_i} S_{ij} + \tau_{\beta_0}\beta_0}{\tau_e m_i + \tau_{\beta_0}}, (\tau_e m_i + \tau_{\beta_0})^{-1}\big) \tag{5.20}$$

The full conditional distribution for $t_i^\star$ has the following form:

$$f(t_i^\star|k_i = 2, else) \propto \big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big)\big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\big)\frac{1}{b_{lag}} I_{(t_{im_i}-b_{lag}, t_{im_i})}(t_i^\star)$$

For $t_i^\star$, we use a Metropolis step.

We propose a new value $t_i^{\star'}$ from the interval $(t_{im_i} - b_{lag}, t_{im_i})$ using a truncated normal distribution with location at $t_i^\star$ and spread parameter 0.4 years. We use $f_{tn}(t_i^{\star'}|t_i^\star, 0.4, t_{im_i} - b_{lag}, t_{im_i})$ to denote the density function of the proposal distribution. Then the acceptance rate is:

$$\alpha(t_i^\star, t_i^{\star'}) = min(1, \frac{f(t_i^{\star'}|k_i = 2, else) f_{tn}(t_i^\star|t_i^{\star'}, 0.4, t_{im_i} - b_{lag}, t_{im_i})}{f(t_i^\star|k_i = 2, else) f_{tn}(t_i^{\star'}|t_i^\star, 0.4, t_{im_i} - b_{lag}, t_{im_i})}) \tag{5.21}$$

Finally consider posterior sampling for $\phi_i^{k_i=3} = \{\beta_{0i}, t_i^\star, h_i, d_i, r_i\}$. We can draw posterior samples for $\beta_{0i}|k_i = 3, else$ using the full conditional distribution:

$$\beta_{0i}|k_i = 3, else \sim N(\frac{\tau_e \sum_{j=1}^{m_i}(S_{ij} - g(t_{ij}|t_i^\star, h_i, d_i, r_i)) + \tau_{\beta_0}\beta_0}{\tau_e m_i + \tau_{\beta_0}}, (\tau_e m_i + \tau_{\beta_0})^{-1}) \tag{5.22}$$

We regard the four parameters $\{t_i^\star, h_i, d_i, r_i\}$ as a block, denoted by u_i, and we draw posterior samples for $u_i = \{t_i^\star, h_i, d_i, r_i\}$ simultaneously. The full conditional distribution for u_i has the following form:

$$\begin{aligned}
& f(u_i|k_i = 3, else) \propto f\big(S_i, F_i \mid (k_i, \phi_i^3 = (\beta_{i0}, u_i)), \Theta\big)\, f(u_i \mid k_i, \Theta) \\
& \propto \ \Big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2}\Big)\Big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\Big) \\
& \quad \Big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\Big) \cdot \Big(\frac{1}{t_{im_i} - a_{lag} - dob_i} I_{(dob_i, t_{im_i}-a_{lag})}(t_i^\star) \\
& \quad \cdot \frac{\tau_h^{\frac{1}{2}}}{h_i - c_h} e^{-\frac{\tau_h}{2}(\log(h_i-c_h)-\mu_1)^2} \frac{\tau_r^{\frac{1}{2}}}{r_i} e^{-\frac{\tau_r}{2}(\log r_i-\mu_2)^2} \frac{\tau_d^{\frac{1}{2}}}{d_i} e^{-\frac{\tau_d}{2}(\log d_i-\mu_3)^2}\Big)
\end{aligned} \tag{5.23}$$

It is evident that the distribution is not recognizable, and we employ a Metropolis step for sampling u_i. The algorithm is the same as sampling for $t_i^\star$ in state 2. In order to remove redundancy, we specify the detailed steps in the Appendix. The acceptance rate for the update of u_i is listed in Equation (C.40).

5.4.4 Sampling scheme

With all the full conditionals listed above, the Gibbs sampling algorithm for posterior simulation consists of nine steps:

1. Assign appropriate initial values to all the parameters including global parameters Θ^0 and cow-specific parameters $\{(k_i^0, \phi_i^{k_i 0})\}$.

2. At iteration l, draw a new τ_e^l using Equation (5.13) with the values from previous step.

3. Sample (q_1^l, q_2^l, q_3^l) with full conditional distribution given in Equation (5.14) using τ_e^l and other parameter values from previous steps.

4. Similarly, sample each global parameter in $\{se, sp, \beta_0, \tau_{\beta_0}, \mu_h, \tau_h, \mu_r, \tau_r, \mu_d, \tau_d\}$ successively with full conditional distributions listed in Equation (5.15) using the newest updated values for other parameters.

5. For each cow i, update its infection state, k_i, using the reversible jump MCMC algorithm,

   - For a proposed $(k_i', \phi_i^{k_i'})$, evaluate the acceptance rate $\alpha(k, k')$ using Equation (5.16), (5.17) or (5.18) depending on the (k_i, k_i') values.
   - Let $k_i^l = k_i'$ and $\phi_i^{k_i l} = \phi_i^{k_i'}$ with probability $\alpha(k, k')$, and let $k_i^l = k_i^{l-1}$ and $\phi_i^{k_i l} = \phi_i^{k_i l-1}$ with probability $1 - \alpha(k, k')$.

6. For each cow i, update $\phi_i^{k_i}$ within state.

   - If $k_i = 1$, draw β_{0i}^l using Equation (5.19)
   - If $k_i = 2$, draw β_{0i}^l using Equation (5.20), and use Metropolis step to sample $t_i^{\star l}$.
     - For a proposed $t_i^{\star'}$, evaluate the acceptance rate $\alpha(t_i^\star, t_i^{\star'})$ using Equation (5.21).
     - Let $t_i^{\star l} = t_i^{\star'}$ with probability $\alpha(t_i^\star, t_i^{\star'})$, and let $t_i^{\star l} = t_i^{\star l-1}$ with probability $1 - \alpha(t_i^\star, t_i^{\star'})$.
   - If $k_i = 3$, draw β_{0i}^l with Equation (5.22), and use Metropolis step to sample $u_i^l = \{t_i^{\star l}, h_i^l, d_i^l, r_i^l\}$.
     - For a proposed u', evaluate the acceptance rate $\alpha(u, u')$ using Equation (C.40).
     - Let $u^l = u'$ with probability $\alpha(u, u')$, and let $u^l = u^{l-1}$ with probability $1 - \alpha(u, u')$.

7. Repeat Step 2-6 iteratively for $l = 1, \ldots, N^{MC}$ to reach convergence.

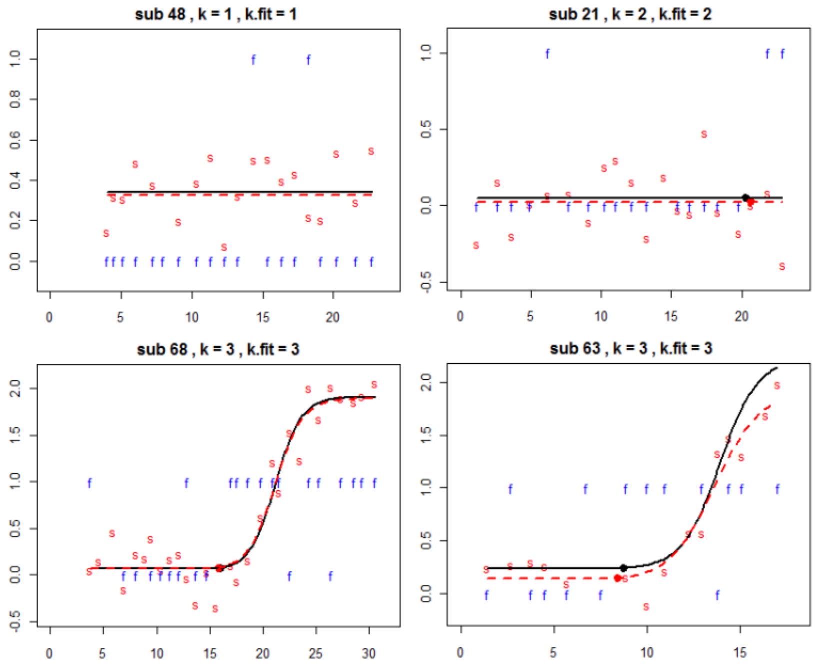


Figure 5.5: Plots of four selected subjects, ‘s’ and ‘F’ represent the observed serology score and FC test results. The solid black line is the true mean curve used to simulate the serology scores. The red dash line is the estimated mean curve.

5.5 Simulation

We use a simulation study to check model performance. The simulated data consist of 100 subjects with 30, 25 and 45 subjects in each of the three infection states 1, 2, 3, respectively. The data from four selected subjects are shown in Figure 5.5. For each subject i, $i = 1, \ldots, 100$, we generated 12 to 36 repeated observations on a time scale ranging from 0 to 40 years. The simulated data are unbalanced, but the screening was generated around annually. In addition, we give the true values of the parameters used for simulation in Table 5.1.

Figure 5.5 also shows that the estimated mean trajectories fit to the serology data were quite accurate. Table 5.1 gives posterior means, medians and 95% probability intervals together with the true values used to simulated the data.

One question of interest for the study is to diagnose the cows. We would like to know how the cows are classified into the three infection states. The classification is done using a 0-1

Table 5.1: Parameter estimates and 95% probability intervals for simulated data

Parameters	Truth	Posterior	Posterior	95% PI	
Θ		Mean	Median	lower	upper
q_1	0.3	0.291	0.29	0.184	0.404
q_2	0.25	0.252	0.249	0.147	0.368
q_3	0.45	0.457	0.456	0.36	0.557
se	0.7	0.714	0.714	0.677	0.751
sp	0.9	0.903	0.903	0.887	0.917
σ_e	0.2	0.199	0.199	0.193	0.206
β_0	0.1	0.105	0.105	0.076	0.134
σ_{β_0}	0.141	0.134	0.135	0.115	0.159
μ_d	1	1.007	1.008	0.84	1.169
σ_d	0.15	0.093	0.089	0.061	0.144
μ_h	0.55	0.527	0.527	0.485	0.568
σ_h	0.1	0.083	0.082	0.063	0.108
μ_r	-0.2	-0.234	-0.239	-0.333	-0.114
σ_r	0.1	0.1	0.097	0.066	0.152

Table 5.2: Classification of subjects in simulation.

		Fitted		
		1	2	3
	1	23	6	1
Truth	2	4	21	0
	3	0	1	44

loss function, where subjects are classified into the infection state with the highest posterior probability. We show classification results in Table 5.2. The model produced a reasonably accurate results, in which 88 out of 100 subjects were correctly classified.

The continuous response serology score is often used for disease diagnosis. It would be interesting to quantify its capability in diagnosing disease. We now consider diagnosis solely based on serology. In this case, classification of subjects as diseased or non-diseased is typically based on a dichotomized score. Subjects with serology scores above a specified cutoff, denoted c, are classified as diseased while those below are classified as non-diseased. We can calculate the sensitivity and specificity of the serology test corresponding to different cutoffs c. In addition, since the concentration of antibodies in an infected subject rise significantly over time after a lag, the sensitivity will be a function of time.

Let $S(t)$ be the serology score at time t for a new infected subject. We use $se(c,t)$ to denote the sensitivity of the serology test at t corresponding to cutoff c, and obtain the predicted $se(c,t)$ using the following equation:

$$
\begin{aligned}
\hat{se}(c,t) &= P(S_i(t) > c|\mathbf{D}) \\
&= \int p(\beta_0 + g(t|\phi^3) + \epsilon_t > c|k = 3, \phi^3, \Theta) f(\phi^3, \Theta|\mathbf{D}) \, d(\phi^3, \Theta)
\end{aligned}
\tag{5.24}
$$

We evaluate the integral using Monte Carlo approximation based on posterior samples.

The specificity of the serology test is assumed not to vary with t. Let $sp(c)$ denote the specificity at cutoff c. We can estimate it similarly as:

$$
\begin{aligned}
\hat{sp}(c) &= P(S(t) < c|\mathbf{D}) \\
&= \int p(\beta_0 + \epsilon_t < c|k = 1, \phi^1, \Theta) f(\phi^1, \Theta|\mathbf{D}) \, d(\phi^1, \Theta)
\end{aligned}
\tag{5.25}
$$

If we fix t and vary the cutoff c, we obtain a set of points, $\big(1 - \hat{sp}(c),\ \hat{se}(c,t)\big)$. The curve generated by these points is the estimated receiver-operating characteristic (ROC) curve at time t. By varying t over a grid of values, we obtain a family of ROC curves as shown in Figure 5.6. This gives us insight into how useful the serology score is in disease diagnosis and how the performance changes with time past infection t. In the simulation study, the generated lags (d_i in the model) averaged 2.7 years. The ROC curves correspond to it well.

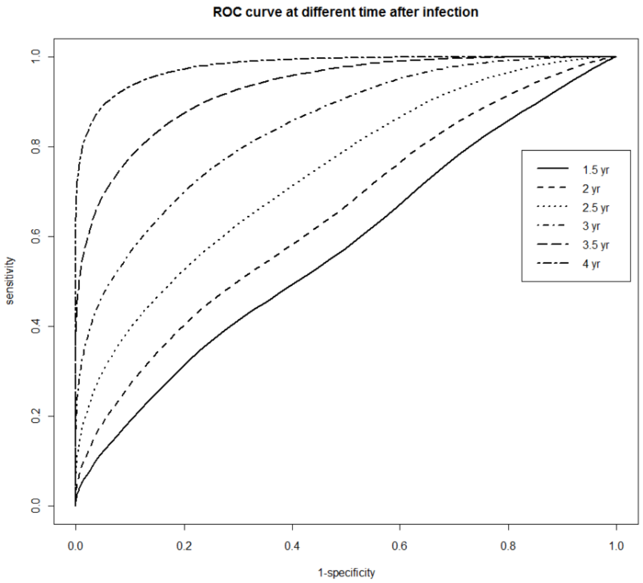


Figure 5.6: Estimated ROC curves from simulation.

5.6 Data Analysis for Johne's Disease

We apply our finite mixture model to the joint longitudinal screening data for JD. Table 5.3 provides inferences for the global parameters. First, we are interested in disease diagnosis. We obtain inferences for disease prevalence 0.47 (0.36,0.59) by combining the posterior proportion of states 2 and 3, since the cows in state 2 and 3 are infected. Among the infected cows, about 60% are still in the "intermediate" state, which indicates JD is likely spreading.

The performance of the FC test is also quantified. We can see the FC test has a high specificity 0.965 (0.954, 0.974), which is important for such a test. Of concern is having an infected cow pass the test and is not removed from the herd. Infected cows continue infecting other healthy cows. The estimated sensitivity of the FC test is 0.635 (0.582, 0.686), which is low. That is why we would like to include serology score in the model to improve the disease diagnosis.

We also show model fitting to 12 cows in Figure 5.7. The cows were selected to show the model fitting for the three infection states.

Table 5.3: Parameter estimates and 95% probability intervals for the cow data

Parameters	Posterior	Posterior	95% PI	
Θ	Mean	Median	lower	upper
q_1	0.533	0.534	0.467	0.601
q_2	0.282	0.281	0.231	0.335
q_3	0.185	0.184	0.125	0.249
se	0.635	0.635	0.582	0.686
sp	0.965	0.965	0.954	0.974
σ_e	0.105	0.105	0.101	0.109
β_0	-1.749	-1.749	-1.757	-1.741
σ_{β_0}	0.054	0.054	0.061	0.047
μ_d	-0.08	-0.08	-0.174	0.026
σ_d	0.164	0.16	0.106	0.246
μ_h	0.652	0.653	0.502	0.799
σ_h	0.535	0.53	0.424	0.698
μ_r	1.868	1.866	1.578	2.16
σ_r	1.21	1.184	0.881	1.685

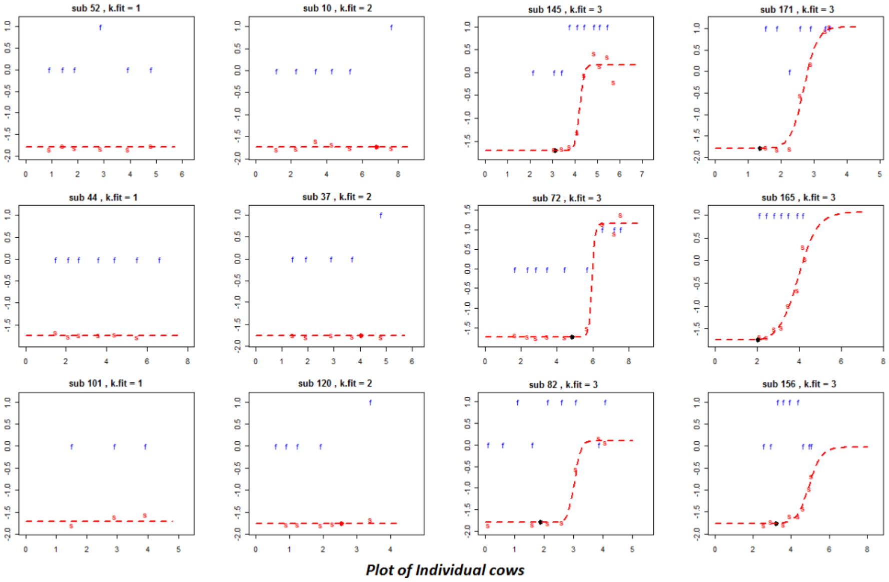


Figure 5.7: Model fitting shown with 12 selected cows. Column 1: three uninfected cows; column 2: three cows in state 2; column 3 and 4: six cows in state 3. Cows were predicted by the model to be in these states.

We can compare our parameter estimates with Norris (2009) [25] since our models have many similarities. The majority of the estimates are comparable. For example, the estimated mean baseline serology score is -1.749 for our model and -1.745 for hers. A significant difference comes from the estimated proportions of cows in the three states, which are $(0.53, 0.28, 0.19)$ for our model and $(0.46, 0.25, 0.29)$ for hers.

Our model differs from hers in modeling the serology trend after infection. We use a four-parameter sigmoid function to fit the serology data, and she used a linear function. That leads to different inferences for the parameters related to serologic reaction. Firstly, the lag time for serologic reaction is pre-determined and universal for all cows. However, it is model based and is different for each cow in our model, and the estimated mean lag time is $0.92\,(0.84, 1.03)$ years. Secondly, the estimated overall slope in her model is $1.09\,(0.10, 12.55)$, which has large variation. The corresponding parameter in our model is the maximum increasing rate $h_i r_i/4$, which is estimated to be $3.1\,(2.0, 4.4)$. Figure 5.7 shows that our model fits the serology data well with sigmoid function.

In addition, we give an ROC plot in Figure 5.8 to evaluate the serology test. Recall that the lag for serologic reaction ranges from 4 to 17 months (0.25 to 1.4 years). The serology test shows some effectiveness at $t = 0.8$ years after infection. When $t \geq 1.4$ years, most of the cows have developed serologic reaction, and the serology test is nearly perfect.

5.7 DPM Model

By examining the six infected cows in Figure 5.7, we see there are different developmental patterns for the serology score. For example, cow 145 has a steep rise in serology score, but it does not rise to a high level. On the contrary, cow 165 has lower increasing rate, but it eventually reaches a much higher serology level. Therefore, the normality assumption for

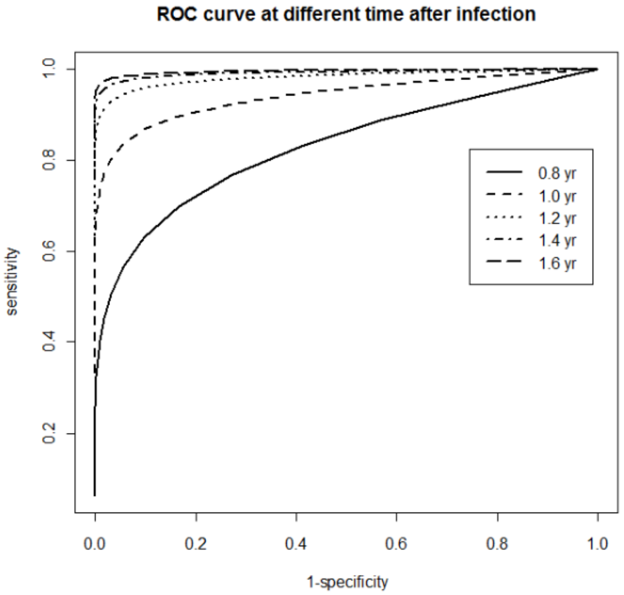


Figure 5.8: Estimated ROC curves based on the cow data.

$\log(h_i - c_h)$ and $\log r_i$ is questionable. In order to relax the assumption, we use a Dirichlet Process Mixture (DPM) for these two parameters in state 3.

In state 3, we model the two cow-specific parameters $\{h_i, r_i\}$ using a DPM as follows:

$$\begin{aligned} log(h_i - c_h)|\mu_h, \tau_h &\sim N(\mu_h, \frac{1}{\tau_h}) \\ log(r_i)|\mu_r, \tau_r &\sim N(\mu_r, \frac{1}{\tau_r}) \\ \{\mu_h, \tau_h, \mu_r, \tau_r\}|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{5.26}$$

where the base distribution G_0 includes:

$$\mu_h|\tau_h \sim N(u_h, \tfrac{\lambda_h}{\tau_h}) \qquad \tau_h \sim \Gamma(a_h/2, b_h/2)$$

$$\mu_r|\tau_r \sim N(u_r, \tfrac{\lambda_r}{\tau_r}) \qquad \tau_r \sim \Gamma(a_r/2, b_r/2)$$

Table 5.4: Posterior Distribution for the number of clusters

Cluster Number	3	4	5	6	7	8	9	≥10
Posterior Probability	0.01	0.23	0.29	0.21	0.12	0.08	0.04	0.02

We use a Gibbs sampler to draw posterior samples as in the parametric case. In previous chapters, we introduced the DPM model in detail and illustrated sampling schemes with examples.

We applied this model to the joint longitudinal screening data for JD, and were able to identify different developmental trends based on longitudinal serology scores. Table 5.4 shows the posterior distribution for the number of clusters. In our analysis, each cluster corresponds to a distinctive developmental pattern. It is clear that there is more than one increasing pattern.

Figure 5.9 shows density estimation for the two cow-specific parameters $\{h_i, r_i\}$. We can identify two significant modes: One mode has the highest density, which corresponds to a large $\log(r_i)$ value. Since the maximum rate of increase is $h_i r_i/4$ for the sigmoid curve, this mode corresponds to a high increasing rate. The other mode has much lower peak compared to the first one. It has a large $\log(h_i - c_h)$ value and a low $\log(r_i)$ value. This mode corresponds to a sigmoid curve reaching a high serology level with a moderate increasing rate. There seems to be another insignificant mode with low $\log(h_i - c_h)$ and $\log(r_i)$ values, which corresponds to a small bump in the figure. The sigmoid curve of this mode has a low serology level with low increasing rate. Our clustering also results correspond to those of Norris, Johnson and Gardner (2014) [26], where they modeled only slopes after infection plus lag with a DP. They identified two distinct slopes for serologic reaction.

However, we have some concern for this analysis, due to having convergence issues on multiple occasions when we ran the Gibbs sampler for the DPM model using the joint longitudinal screening data for JD. We believe the problem was caused by having a limited number of observations from some cows. This kind of issue was discussed in chapter 4, and is more

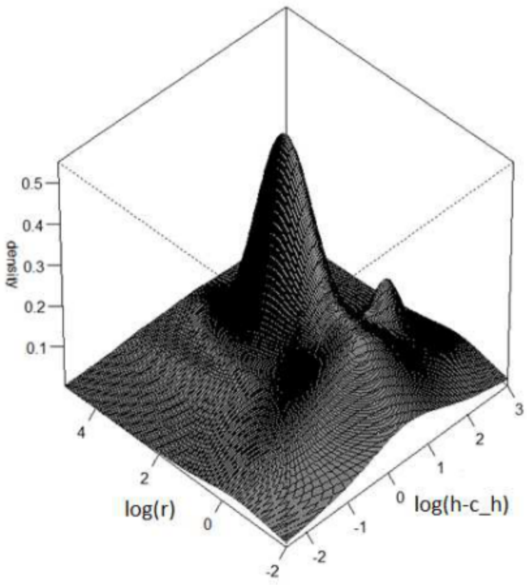


Figure 5.9: Predictive density estimation for $\{\log(h_i - c_h), \log r_i\}$

difficult here due to having latent states.

In the DPM model, the normality assumption has been relaxed for $\{h_i, r_i\}$. Their estimates depend solely on their serology score observations after infection time. By checking the data, we found some infected cows had a small number of observations after infection and did not fully develop serological reactions in the screening window. For these cows, it is impossible to obtain a reasonable estimate for $\{h_i, r_i\}$, thus leading to questionable clustering results for them.

5.8 Conclusion

In this chapter, we built a finite mixture model for the diagnosis of JD by fitting longitudinal FC outcomes and serology scores jointly. The parametric model was shown to work well using simulated data and it was illustrated using real JD data. As for the non-parametric model with the DPM, we identified different increasing pattern for serologic reaction. However, our method had convergence issues due to missing data for after infection serology scores. Our future work will focus on resolving that problem.

Bibliography

[1] G. Bachmann. Urogenital ageing: an old problem newly recognized. *Maturitas*, 22:S1–S5, 1995.

[2] J. Bigelow and D. B. Dunson. Semiparametric classification in hierarchical functional data analysis. *Duke University ISDS Discussion paper*, pages 05–18, 2005.

[3] J. L. Bigelow and D. B. Dunson. Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association*, 104(485), 2009.

[4] D. Blackwell and J. B. MacQueen. Ferguson distributions via pólya urn schemes. *The annals of statistics*, pages 353–355, 1973.

[5] H. Burger, G. Hale, D. Robertson, and L. Dennerstein. A review of hormonal changes during the menopausal transition: focus on findings from the melbourne women's midlife health project. *Human reproduction update*, 13(6):559–565, 2007.

[6] C. A. Bush and S. N. MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

[7] D. B. Dunson. Nonparametric bayes applications to biostatistics. *Bayesian nonparametrics*, 28:223–273, 2010.

[8] M. D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

[9] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[10] Y. Fan and S. Brooks. Bayesian modelling of prehistoric corbelled domes. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):339–354, 2000.

[11] T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[12] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[13] G. E. Hale, C. L. Hughes, H. G. Burger, D. M. Robertson, and I. S. Fraser. Atypical estradiol secretion and ovulation patterns caused by luteal out-of-phase (loop) events underlying irregular ovulatory menstrual cycles in the menopausal transition. *Menopause*, 16(1):50–59, 2009.

[14] Y. S. Hannestad, G. Rortveit, H. Sandvik, and S. Hunskaar. A community-based epidemiological survey of female urinary incontinence:: The norwegian epincont study. *Journal of clinical epidemiology*, 53(11):1150–1157, 2000.

[15] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[16] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

[17] S. Jain and R. M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

[18] W. J. Kennedy and J. E. Gentle. Statistical computing. *New York*, 1980.

[19] K. P. Kleinman and J. G. Ibrahim. A semiparametric bayesian approach to the random effects model. *Biometrics*, 54(3):921–938, 1998.

[20] A. Lepper, C. Wilks, M. Kotiw, J. Whitehead, and K. Swart. Sequential bacteriological observations in relation to cell-mediated and humoral antibody responses of cattle infected with mycobacterium paratuberculosis and maintained on normal or high iron intake. *Australian veterinary journal*, 66(2):50–55, 1989.

[21] A. Y. Lo et al. On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, 12(1):351–357, 1984.

[22] S. N. MacEachern and P. Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 2004.

[24] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[25] M. Norris, W. O. Johnson, and I. A. Gardner. Modeling bivariate longitudinal diagnostic outcome data in the absence of a gold standard. *Stat Interface*, 2:171–185, 2009.

[26] M. Norris, W. O. Johnson, and I. A. Gardner. Bayesian semi-parametric joint modeling of biomarker data with a latent changepoint: Assessing the temporal performance of enzyme-linked immunosorbent assay (elisa) testing for paratuberculosis. *Stat Interface*, 2014 in press.

[27] K. A. O'Connor, D. J. Holman, and J. W. Wood. Menstrual cycle variability and the perimenopause. *American Journal of Human Biology*, 13(4):465–478, 2001.

[28] J. C. Prior. Perimenopause: the complex endocrinology of the menopausal transition. *Endocrine Reviews*, 19(4):397–428, 1998.

[29] J. F. Randolph Jr, M. Sowers, I. V. Bondarenko, S. D. Harlow, J. L. Luborsky, and R. J. Little. Change in estradiol and follicle-stimulating hormone across the early menopausal transition: effects of ethnicity and age. *Journal of Clinical Endocrinology & Metabolism*, 89(4):1555–1561, 2004.

[30] H. Rekers, A. Drogendijk, H. Valkenburg, and F. Riphagen. The menopause, urinary incontinence and other symptoms of the genito-urinary tract. *Maturitas*, 15(2):101–111, 1992.

[31] C. Robert. Multimodality and label switching: a discussion. Workshop on mixtures, ICMS, 2010.

[32] K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

[33] C. M. Sampselle, S. D. Harlow, J. Skurnick, L. Brubaker, and I. Bondarenko. Urinary incontinence predictors and life impact in ethnically diverse perimenopausal women. *Obstetrics & Gynecology*, 100(6):1230–1238, 2002.

[34] J. Sethuraman. A constructive definition of dirichlet priors. Technical report, DTIC Document, 1991.

[35] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[36] M. Stephens and D. Phil. Bayesian methods for mixtures of normal distributions, 1997.

[37] J. R. Taffe and L. Dennerstein. Menstrual patterns leading to the final menstrual period. *Menopause*, 9(1):32–40, 2002.

[38] D. H. Thom and J. S. Brown. Reproductive and hormonal risk factors for urinary incontinence in later life: a review of the clinical and epidemiologic literature. *Journal of the American Geriatrics Society*, 46(11):1411–1417, 1998.

[39] L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[40] X. Wang, S. Ray, and B. K. Mallick. Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, 102(479):962–973, 2007.

[41] G. Weiss, J. H. Skurnick, L. T. Goldsmith, N. F. Santoro, and S. J. Park. Menopause and hypothalamic-pituitary sensitivity to estrogen. *Jama*, 292(24):2991–2996, 2004.

[42] L. Wilson, J. S. Brown, G. P. Shin, K.-O. Luc, and L. L. Subak. Annual direct cost of urinary incontinence. *Obstetrics & Gynecology*, 98(3):398–406, 2001.

[43] W. Yao and B. G. Lindsay. Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association*, 104(486), 2009.

Appendices

A Appendix A

A.1 Details about assigning prior to γ

The statistical meaning of γ is easier to understand with model (2.3) by considering:

$$\begin{aligned} Var(y_i|\beta_i) &= Var(X_i b_i + \epsilon_i) \\ &= \tau_e^{-1}(X_i \Gamma^{-1} X_i^T + I_{r_i}) \end{aligned}$$

The variation is contributed by two parts: the random error $\tau_e^{-1} I_{r_i}$ and the mixed effects within cluster part $\tau_e^{-1} X_i \Gamma^{-1} X_i^T$. Then the overall variation of the vector y_i is $\sum_{j=1}^{r_i} Var(y_{ij}|\beta_i)$, which consists of

$$trace(\tau_e^{-1} I_{r_i}) = r_i \tau_e^{-1}$$

and

$$trace(\tau_e^{-1} X_i \Gamma^{-1} X_i^T) = trace(\tau_e^{-1} X_i^T X_i \Gamma^{-1})$$

If orthogonal polynomial basis functions are used for balanced data, we can write $X_i = (x_{0i}, x_{1i}, \ldots, x_{pi})$ with x_k being an orthogonal basis function of degreee k in the defined discrete time space for all $k = 0, 1, \ldots, p$. And thus $X_i^T X_i = I_{p+1}$ and $trace(\tau_e^{-1} X_i \Gamma^{-1} X_i^T) = \tau_e^{-1} \sum_{j=0}^{p} \gamma_j^{-1}$. Taking the ratio of this to the total, we see that the mixed effect explains $\frac{\sum_{j=0}^{p} \gamma_j^{-1}}{\sum_{j=0}^{p} \gamma_j^{-1} + r_i} \times 100$ percent of variation corresponding to individual i.

The result gives us an idea about how to select a prior distribution for $\boldsymbol{\gamma}$. For example, assume basis functions up to cubic terms ($p = 3$) are used and the individual has $r_i = 10$ observations. Then 28.2% variation is explained by the mixed effect if we let the precision ratio $\gamma_j = 1$ for all $j = 0, 1, \ldots, p$. In our analysis, we assigned a diffuse Gamma prior to $\boldsymbol{\gamma}$:

$$\gamma_j \sim Gamma(\frac{a_\gamma}{2}, \frac{b_\gamma}{2}), \text{ for all } j = 0, 1, \ldots, p$$

where $a_\gamma = 2.02$ and $b_\gamma = 0.02$. It covers a wide range of γ values with sufficient probability.

Notice when $\gamma_j \to \infty$, $X_i \Gamma^{-1} X_i^T$ tends to 0. The contribution from mixed effect diminishes and our model (2.2) becomes model (2.1). In fact, model (2.1) could be regarded as the special case of our model (2.2) with mixed effects tending to 0. That is analogous to DP vs DPM, where DP is a special case of DPM with the standard deviation of each normal mixture goes to 0. We used a similar result when taking model (2.1) to be a special case of our model (2.1) when performing a simulation in Chapter 2. We used $Gamma(\frac{a_\gamma}{2}, \frac{b_\gamma}{2})$ with $a_\gamma = 10^6$ and $b_\gamma = 1$.

If Legendre basis functions are used for unbalanced data, $X_i^T X_i$ is not exactly equal to, but still close to I_{p+1}. The diagonal element is

$$\begin{aligned} x_{ki}^T x_{ki} &= \sum_{j=t_{i1}}^{t_{ir_i}} x_{ki}^2(t_j) \\ &= \frac{r_i}{2} \sum_{j=t_{i1}}^{t_{ir_i}} x_{ki}^2(t_j) \frac{2}{r_i} \\ &\approx \frac{r_i}{2} \int_{-1}^{1} x_k^2(t)\, dt \\ &= \frac{r_i}{2} \quad \text{for all } k = 0, 1, \ldots, p \end{aligned}$$

Similarly, the non-diagonal element is

$$x_{ki}^T x_{li} \approx \frac{r_i}{2} \int_{-1}^{1} x_k(t) x_l(t)\, dt = 0, \ \text{for all },\ 0 \leq k \neq l \leq p$$

Similar to the procedure above, we can still obtain a rough idea about the relative contribution of mixed effects variation in selecting an appropriate prior for $\boldsymbol{\gamma}$.

B Appendix B

B.1 K-medoids clustering method

The K-medoids algorithm is a clustering algorithm related to the K-means algorithm and the medoidshift algorithm. Medoids are representative individuals of a data set or a cluster with a data set whose average dissimilarity to all the individuals in the cluster is minimal. Medoids are similar in concept to means or centroids, but medoids are always members of the data set.

Both the K-means and K-medoids algorithms break the dataset up into groups, and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the commonly used K-means algorithm, K-medoids chooses data points as centers (medoids) and works with an arbitrary matrix of distances between data points instead of squared Euclidean distance. In our analysis, we do not know the coordinates of the subjects, and thus are not able to define the cluster centers. However, we do have the pairwise distances between any two subjects in the data set. Therefore, the K-mediods algorithm is applicable in our case.

K-medoid is a classical partitioning technique for clustering the data set of n subjects into K clusters, where K is pre-specified. In our case, K is picked based on the posterior distribution of the number of clusters.

The most common realization of K-medoid clustering is the Partitioning Around Medoids (PAM) algorithm. When applied to our analysis, the procedure is as follows:

1. Initialize: randomly select K of the n data points as the medoids.
2. Associate each data point with the closest medoid using pairwise distances defined in Equation (3.5).
3. For each medoid $k, k = 1, \ldots, K$ and each non-medoid subject i in the data, swap k and i and compute the total cost of the configuration.
4. Repeat step 2 to 4 until there is no change in the medoids.

B.2 Hierarchical Clustering Models

Hierarchical clustering is a method of cluster analysis in data mining, which seeks to build a hierarchy of clusters. Given a set of n subjects to be clustered, and an $n * n$ mutual distance (or similarity) matrix, the basic process of agglomerative hierarchical clustering is as follows:

1. Start by assigning each subject to a cluster, so that if you have n subjects, you now have n clusters, each containing just one subject. Let the distances between two clusters the same as the largest mutual distance between any pair of subjects from them.
2. Find the closest (with the shortest distance) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
3. Compute distances between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size n.

Hierarchical clustering arranges items in a hierarchy with a treelike structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram. Figure (fig:dendrogram.galaxy) is an example.

Step 3 can be done with different linkage functions. Some commonly used linkage functions include single-linkage, complete-linkage and average-linkage. The dendrogram and clustering of subjects are different with different linkage.

I personally prefer the K-medroid method over Hierarchical clustering becuase of the linkage function. Sometimes the clustering results are very different for two different linkage functions. And it is not evident how different linkage functions affect the clustering results.

B.3 Table of $P(k_i|y, K = 6)$

Table B.5: Posterior probability for clustering of the subjects in the Galaxy Data

Subjects	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Other
sub 1	0.97	0	0	0	0	0	0.03
sub 2	0.973	0	0	0	0	0	0.027
sub 3	0.974	0	0	0	0	0	0.026
sub 4	0.975	0	0	0	0	0	0.025
sub 5	0.974	0	0	0	0	0	0.026
sub 6	0.971	0	0	0	0	0	0.029
sub 7	0.965	0.001	0	0	0	0	0.034
sub 8	0	0.752	0.053	0.075	0.073	0.001	0.046
sub 9	0	0.746	0.054	0.076	0.073	0.001	0.05
sub 10	0	0.096	0.615	0.1	0.085	0	0.104
sub 11	0	0.087	0.636	0.094	0.08	0	0.102
sub 12	0	0.083	0.646	0.091	0.077	0	0.102
sub 13	0	0.066	0.692	0.077	0.064	0	0.102
sub 14	0	0.062	0.701	0.073	0.061	0	0.103
sub 15	0	0.061	0.7	0.075	0.061	0	0.104
sub 16	0	0.055	0.714	0.07	0.058	0	0.104
sub 17	0	0.057	0.714	0.07	0.057	0	0.102
sub 18	0	0.056	0.711	0.072	0.059	0	0.102
sub 19	0	0.056	0.712	0.069	0.057	0	0.106
sub 20	0	0.056	0.715	0.072	0.058	0	0.1
sub 21	0	0.054	0.711	0.074	0.058	0	0.103
sub 22	0	0.054	0.719	0.069	0.056	0	0.102

sub 23	0	0.055	0.712	0.072	0.059	0	0.102
sub 24	0	0.053	0.71	0.075	0.059	0	0.103
sub 25	0	0.054	0.709	0.076	0.061	0	0.099
sub 26	0	0.055	0.705	0.08	0.063	0	0.097
sub 27	0	0.056	0.705	0.079	0.061	0	0.099
sub 28	0	0.056	0.702	0.081	0.064	0	0.098
sub 29	0	0.056	0.707	0.082	0.062	0	0.094
sub 30	0	0.055	0.698	0.082	0.063	0	0.102
sub 31	0	0.055	0.695	0.087	0.067	0	0.097
sub 32	0	0.057	0.674	0.099	0.075	0	0.094
sub 33	0	0.058	0.676	0.1	0.075	0	0.09
sub 34	0	0.058	0.678	0.1	0.074	0	0.09
sub 35	0	0.058	0.671	0.102	0.077	0	0.093
sub 36	0	0.058	0.667	0.106	0.077	0	0.092
sub 37	0	0.059	0.667	0.106	0.077	0	0.092
sub 38	0	0.062	0.629	0.13	0.09	0	0.089
sub 39	0	0.069	0.565	0.17	0.11	0	0.086
sub 40	0	0.073	0.505	0.207	0.127	0	0.089
sub 41	0	0.074	0.494	0.216	0.131	0	0.084
sub 42	0	0.073	0.481	0.224	0.134	0	0.089
sub 43	0	0.073	0.475	0.231	0.135	0	0.086
sub 44	0	0.075	0.415	0.266	0.153	0	0.09
sub 45	0	0.079	0.348	0.314	0.168	0	0.092
sub 46	0	0.082	0.203	0.422	0.197	0	0.096
sub 47	0	0.08	0.144	0.47	0.208	0	0.098
sub 48	0	0.079	0.123	0.489	0.212	0	0.096
sub 49	0	0.079	0.11	0.499	0.214	0	0.098

sub 50	0	0.079	0.104	0.506	0.215	0	0.096
sub 51	0	0.077	0.081	0.522	0.217	0	0.102
sub 52	0	0.077	0.079	0.527	0.219	0	0.098
sub 53	0	0.078	0.078	0.526	0.221	0	0.096
sub 54	0	0.078	0.079	0.527	0.219	0	0.098
sub 55	0	0.076	0.075	0.532	0.221	0	0.096
sub 56	0	0.077	0.073	0.529	0.221	0	0.1
sub 57	0	0.076	0.066	0.538	0.224	0	0.096
sub 58	0	0.075	0.062	0.54	0.226	0	0.098
sub 59	0	0.074	0.062	0.542	0.226	0	0.095
sub 60	0	0.073	0.06	0.545	0.227	0	0.095
sub 61	0	0.073	0.06	0.541	0.228	0	0.097
sub 62	0	0.074	0.058	0.544	0.232	0	0.093
sub 63	0	0.073	0.056	0.544	0.235	0	0.092
sub 64	0	0.072	0.056	0.54	0.234	0	0.098
sub 65	0	0.072	0.054	0.534	0.241	0	0.097
sub 66	0	0.072	0.055	0.535	0.238	0	0.1
sub 67	0	0.072	0.056	0.533	0.24	0	0.099
sub 68	0	0.072	0.054	0.529	0.243	0	0.1
sub 69	0	0.073	0.055	0.531	0.245	0	0.097
sub 70	0	0.072	0.055	0.525	0.246	0	0.102
sub 71	0	0.072	0.053	0.509	0.261	0.001	0.104
sub 72	0	0.073	0.053	0.501	0.268	0	0.105
sub 73	0	0.072	0.053	0.502	0.268	0.001	0.105
sub 74	0	0.072	0.053	0.497	0.274	0	0.105
sub 75	0	0.073	0.052	0.473	0.294	0.001	0.107
sub 76	0	0.073	0.052	0.444	0.324	0.001	0.107

sub 77	0	0.073	0.051	0.355	0.441	0.001	0.078
sub 78	0	0.071	0.047	0.226	0.581	0.004	0.071
sub 79	0	0.071	0.046	0.211	0.571	0.005	0.096
sub 80	0	0.002	0	0	0.005	0.909	0.084
sub 81	0	0.001	0	0	0.003	0.949	0.047
sub 82	0	0.001	0	0	0.003	0.845	0.15

C Appendix C

C.1 Calculation of Accepatance Rate for Reversible Jump MCMC

In the reversible-jump MCMC sampler used in chapter 5, we have to consider six different moves including $k = 1 \leftrightarrows k' = 2$, $k = 1 \leftrightarrows k' = 3$ and $k = 2 \leftrightarrows k' = 3$, and calculate the acceptance rate for each move. The acceptance rate can be derived using Equation (1.21), which is also listed below:

$$\alpha(\mathcal{M}_k, \mathcal{M}_{k'}) = min\{1, \frac{f(k', \theta_{k'}|y)q(k' \to k)q_{k' \to k}(u')}{f(k, \theta_k|y)q(k \to k')q_{k \to k'}(u)} \mid \frac{\partial g_{k \to k'}(\theta_k, u)}{\partial(\theta_k, u)} \mid\}$$

We start to illustrate the reversible-jump MCMC sampler in detail using a move from state 1 ($k = 1$) to state 3 ($k' = 3$). Firstly, the full conditional distribution of $(k_i, \phi_i^{k_i})|else$ can be obtained with Equation (5.5) and (5.7) using

$$f\big((k_i, \phi_i^{k_i})|else\big) \propto f\big(S_i, F_i \mid (k_i, \phi_i^{k_i}), \Theta\big)\, f\big((k_i, \phi_i^{k_i}) \mid \Theta\big)$$

We then obtain $(k_i, \phi_i^{k_i})|else$ for both state 1 and state 3:

$$\begin{aligned}
f\big((k_i = 1, \phi_i^1)|else\big) &\propto \big(\textstyle\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big) q_1 \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2} \\
f\big((k_i = 3, \phi_i^3)|else\big) &\propto \big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2}\big)\big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big) \qquad \text{(C.27)} \\
&\big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\big) \Big(q_3 \frac{1}{t_{im_i}-a_{lag}-dob_i} \tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2} \\
&\frac{\tau_h^{\frac{1}{2}}}{h_i-c_h} e^{-\frac{\tau_h}{2}(\log(h_i-c_h)-\mu_1)^2} \frac{\tau_r^{\frac{1}{2}}}{r_i} e^{-\frac{\tau_r}{2}(\log r_i-\mu_2)^2} \frac{\tau_d^{\frac{1}{2}}}{d_i} e^{-\frac{\tau_d}{2}(\log d_i-\mu_3)^2}\Big)
\end{aligned}$$

The state-specific parameters are $\phi_i^1 = \beta_{0i}$ for state 1 and $\phi_i^3 = \{\beta_{0i}, t_i^\star, h_i, d_i, r_i\}$ for state 3. So we construct a random vector $u = \{t, u_1, u_2, u_3\}$ for state 1 to match the dimension of the parameter space in state 3. The proposal distribution $q_{1\to3}(u)$ includes:

$$\begin{aligned}
u_1 &\sim N(\mu_h, \tau_h^{-1}) \\
u_2 &\sim N(\mu_r, \tau_r^{-1}) \\
u_3 &\sim N(\mu_d, \tau_d^{-1})
\end{aligned}$$

where the three distributions are the same as the prior distribution for $\{h_i, d_i, r_i\}$. We generate t with the distribution below:

$$\psi_3(t) \propto Sp_F^{\sum_{j:t_{ij}<t} 1-F_{ij}} (1-Sp_F)^{\sum_{j:t_{ij}<t} F_{ij}} (1-Se_F)^{\sum_{j:t_{ij}\geq t} 1-F_{ij}} Se_F^{\sum_{j:t_{ij}\geq t} F_{ij}} I(dob_i \leq t \leq t_{im_i} - a_{lag})$$

We then have the proposal density function:

$$q_{1\to3}(u) \propto \psi_3(t) \tau_h^{\frac{1}{2}} e^{-\frac{\tau_h}{2}(u_1-\mu_1)^2} \tau_r^{\frac{1}{2}} e^{-\frac{\tau_r}{2}(u_2-\mu_2)^2} \tau_d^{\frac{1}{2}} e^{-\frac{\tau_d}{2}(u_3-\mu_3)^2} \qquad \text{(C.28)}$$

The one-to-one mapping between (β_{0i}, u) and $\{\beta_{0i}, t_i^\star, h_i, d_i, r_i\}$ has relation as: $t_i^\star = t, h_i = c_h + e^{u_1}, r_i = e^{u_2}$ and $d_i = e^{u_3}$.

Let J denote the Jacobian matrix in Equation (1.21). The determinant of the Jacobian matrix is:

$$|J| = |\frac{\partial(\beta_{0i}, t_i^\star, h_i, r_i, d_i)}{\partial(\beta_{0i}, t, u_1, u_2, u_3)}| = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & e^{u_1} & 0 & 0 \\ 0 & 0 & 0 & e^{u_2} & 0 \\ 0 & 0 & 0 & 0 & e^{u_3} \end{vmatrix} = e^{u_1+u_2+u_3} = r_i d_i (h_i - c_h) \quad \text{(C.29)}$$

Denote $\alpha_{13} = \frac{f((k_i=3,\phi_i^3)|else)q(k'=3\to k=1)}{f((k_i=1,\phi_i^1)|else)q(k=1\to k'=3)q_{1\to 3}(u)} \mid J \mid$, which is the ratio term in Equation (1.20) applied to our case. By plugging in the results from Equation (C.27), (C.1) and (C.29), we obtain:

$$\begin{aligned} \alpha_{13} &= \frac{f((k_i=3,\phi_i^3)|else)\cdot\frac{1}{3}\cdot e^{u_1+u_2+u_3}}{f((k_i=1,\phi_i^1)|else)\cdot\frac{1}{3}\cdot q_{1\to 3}(u)} \qquad \text{(C.30)} \\ &= \frac{q_3\frac{1}{t_{im_i}-a_{lag}-dob_i}}{q_1\psi_3(t_i^\star)} \cdot \frac{\Big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2}\Big)}{\left(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2}\right)} \\ &\cdot \frac{\Big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\Big)\Big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\Big)}{\prod_{j=1}^{m_i} sp^{1-F_{ij}}(1-sp)^{F_{ij}}} \end{aligned}$$

And the acceptance rate for the move from state 1 to state 3 is

$$\alpha(k=1,k'=3) = min(1,\alpha_{13}) \qquad \text{(C.31)}$$

Since the constructed Markov chain is reversible, we can easily get the acceptance rate for the move from state 3 to state 1:

$$\alpha(k=3,k'=1) = min(1,\alpha_{31}), \text{ where } \alpha_{31} = \alpha_{13}^{-1} \tag{C.32}$$

Similarly, we consider the move from state 1 to state 2, we have the full conditional distribution $(k_i, \phi_i^{k_i})|else$ for state 2 as follow:

$$\begin{aligned} f\big((k_i=2,\phi_i^2)|else\big) \;\propto\; & \Big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2}\Big)\Big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\Big) \\ & \cdot\Big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\Big)\Big(q_2\frac{1}{b_{lag}}\tau_{\beta_0}^{\frac{1}{2}} e^{-\frac{\tau_{\beta_0}}{2}(\beta_{0i}-\beta_0)^2}\Big) \end{aligned} \tag{C.33}$$

Since $\phi_i^2 = \{\beta_{0i}, t_i^\star\}$ for state 2, we need to create just one random variable $u = t$ for state 1 to match the parameter space in state 2. The proposal distribution $q_{1\to 2}(u)$ is:

$$\psi_2(t) \propto Sp_F^{\sum_{j:t_{ij}<t} 1-F_{ij}}(1-Sp_F)^{\sum_{j:t_{ij}<t} F_{ij}}(1-Se_F)^{\sum_{j:t_{ij}\geq t} 1-F_{ij}} Se_F^{\sum_{j:t_{ij}\geq t} F_{ij}} I(t_{im_i} - b_{lag} \leq t \leq t_{im_i})$$

We then have the proposal density function:

$$q_{1\to 2}(u) = \psi_2(t) \tag{C.34}$$

Note the proposal distribution for $t_i^\star$ is different for the move to state 3 ($\psi_3(t)$) and for the move to state 2 ($\psi_2(t)$). Even though $t_i^\star$ has the same interpretation in both states, which is the infection time, they are mathematically different because they were defined in different domains. In state 2, $t_i^\star$ ranges from $t_{im_i} - b_{lag}$ to t_{im_i}. But its domain is $(dob_i, t_{im_i} - a_{lag})$ in state 3.

We then obtain the acceptance rate as:

$$\alpha(k=1,k'=2) = min(1,\alpha_{12}) \tag{C.35}$$

where

$$\begin{aligned}\alpha_{12} \quad = \quad & \frac{q_2\frac{1}{b_{lag}}}{q_1\psi_2(t_i^\star)} \cdot \frac{\big(\prod\limits_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2}\big)}{(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2})} \\ & \cdot \frac{\big(\prod\limits_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\big)\big(\prod\limits_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\big)}{\prod_{j=1}^{m_i} sp^{1-F_{ij}}(1-sp)^{F_{ij}}}\end{aligned}$$

Consider the move from state 2 to state 3, the state-specific parameters are $\phi_i^2 = \{\beta_{0i}, t_i^\star\}$ for state 1 and $\phi_i^3 = \{\beta_{0i}, t_i^{\star'}, h_i, d_i, r_i\}$ for state 3. Here we use $t_i^{\star'}$ to denote the infection time in state 3 in order to differentiate it from $t_i^\star$ in state 2. We construct a random vector $u = \{t, u_1, u_2, u_3\}$ for state 2 and a random variable $u' = t'$ from state 3.

The generation of u is the same as in the move between state 1 and 3. So the one-to-one mapping still has the relation: $t_i^{\star'} = t, h_i = c_h + e^{u_1}, r_i = e^{u_2}$ and $d_i = e^{u_3}$. The variable $u' = t'$ is generated from $\psi_2(t)$, which is used to match $t_i^\star$: $t' = t_i^\star$. Therefore, the proposal distribution $q_{2\to3}(u) = q_{1\to3}(u)$, and the proposal $q_{3\to2}(u') = \psi_2(t')$.

We plug in the results to Equation (1.21), and obtain the acceptance rate:

$$\alpha(k=2,k'=3) = min(1,\alpha_{23}) \tag{C.36}$$

where

$$\begin{aligned}\alpha_{23} \;&=\; \frac{q_3 \frac{1}{t_{im_i}-a_{lag}-dob_i}\psi_2(t_i^\star)}{q_2\psi_3(t_i^{\star\prime})\frac{1}{b_{lag}}} \cdot \frac{\Big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2}\Big)}{\big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}(S_{ij}-\beta_{0i})^2}\big)} \\ &\quad\cdot \frac{\Big(\prod_{j:t_{ij}<t_i^{\star\prime}} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\Big)\Big(\prod_{j:t_{ij}\geq t_i^{\star\prime}} se^{F_{ij}}(1-se)^{1-F_{ij}}\Big)}{\Big(\prod_{j:t_{ij}<t_i^{\star}} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\Big)\Big(\prod_{j:t_{ij}\geq t_i^{\star}} se^{F_{ij}}(1-se)^{1-F_{ij}}\Big)}\end{aligned}$$

Using reversibility of the Markov chain, we also have

$$\alpha(k=2,k'=1) = min(1,\alpha_{21}) \quad \text{where } \alpha_{21} = \alpha_{12}^{-1} \tag{C.37}$$

$$\alpha(k=3,k'=2) = min(1,\alpha_{32}) \quad \text{where } \alpha_{32} = \alpha_{23}^{-1} \tag{C.38}$$

C.2 Metropolis-within-Gibbs

In state 3, the cow-specific parameters include $\phi_i^{k_i=3} = \{\beta_{0i}, t_i^\star, h_i, d_i, r_i\}$. We regard the four parameters $\{t_i^\star, h_i, d_i, r_i\}$ as a block, denoted by u_i, and we draw posterior samples for $u_i = \{t_i^\star, h_i, d_i, r_i\}$ simultaneously. The full conditional distribution for u_i has the following form:

$$\begin{aligned}&f(u_i|k_i=3, else) \propto f\big(S_i, F_i \mid (k_i, \phi_i^3 = (\beta_{i0}, u_i)), \Theta\big)\, f(u_i \mid k_i, \Theta) \\ &\;\propto\; \Big(\prod_{j=1}^{m_i} \tau_e^{\frac{1}{2}} e^{-\frac{\tau_e}{2}\left(S_{ij}-\beta_{0i}-g(t_{ij}|\phi_i^3)\right)^2}\Big)\Big(\prod_{j:t_{ij}<t_i^\star} sp^{1-F_{ij}}(1-sp)^{F_{ij}}\Big)\Big(\prod_{j:t_{ij}\geq t_i^\star} se^{F_{ij}}(1-se)^{1-F_{ij}}\Big) \\ &\quad\cdot\left(\frac{1}{t_{im_i}-a_{lag}-dob_i} \cdot \frac{\tau_h^{\frac{1}{2}}}{h_i-c_h} e^{-\frac{\tau_h}{2}(\log(h_i-c_h)-\mu_1)^2} \frac{\tau_r^{\frac{1}{2}}}{r_i} e^{-\frac{\tau_r}{2}(\log r_i-\mu_2)^2} \frac{\tau_d^{\frac{1}{2}}}{d_i} e^{-\frac{\tau_d}{2}(\log d_i-\mu_3)^2}\right)\end{aligned}$$

It is apparent that the distribution is not recognizable, and we employ Metropolis-within-Gibbs algorithm for sampling u. We choose the transition kernel below to propose a new step u_i':

$$\begin{pmatrix} \log(h_i' - c_h) \\ \log r_i' \\ d_i' \\ logit(\frac{t_i^{\star\prime} - dob_i}{t_{im_i} - a_{lag} - dob_i}) \end{pmatrix} \sim N_4 \left(\begin{pmatrix} \log(h_i - c_h) \\ \log r_i \\ d_i \\ logit(\frac{t_i^{\star} - dob_i}{t_{im_i} - a_{lag} - dob_i}) \end{pmatrix}, \hat{\Sigma} = \begin{pmatrix} 0.05 & 0 & 0 & 0 \\ 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix} \right)$$

where $\hat{\Sigma}$ controls the step size of the Markov chain, which can be adjusted in order to attain a good acceptance rate.

Let $f_n(\cdot)$ be the density function for the transition kernel, which is a multivariate normal distribution. We can obtain the density function for the propsal u_i':

$$\begin{aligned} q(u_i \to u_i') &= f_n\big((\log(h_i' - c_h), \log r_i', d_i', logit(\frac{t_i^{\star\prime} - dob_i}{t_{im_i} - a_{lag} - dob_i}))|u_i, \hat{\Sigma}\big) \\ &\quad \cdot |\frac{\partial\big(\log(h_i' - c_h), \log r_i', d_i', logit(\frac{t_i^{\star\prime} - dob_i}{t_{im_i} - a_{lag} - dob_i})\big)}{\partial(t_i^\star, h_i, d_i, r_i)}| \end{aligned}$$

where the determinant of the Jacobian matrix can be simplified:

$$|\frac{\partial\big(\log(h_i' - c_h), \log r_i', d_i', logit(\frac{t_i^{\star\prime} - dob_i}{t_{im_i} - a_{lag} - dob_i})\big)}{\partial(t_i^\star, h_i, d_i, r_i)}| = \begin{vmatrix} \frac{1}{h_i' - c_h} & 0 & 0 & 0 \\ 0 & \frac{1}{r_i'} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{t_{im_i} - a_{lag} - dob_i}{(t_i^{\star\prime} - dob_i)(t_{im_i} - a_{lag} - t_i^{\star\prime})} \end{vmatrix}$$

We then can obtain the ratio of the proposal distribution:

$$\frac{q(u_i' \rightarrow u_i)}{q(u_i \rightarrow u_i')} = \frac{r_i'(h_i' - c_h)(t_i^{\star'} - dob_i)(t_{im_i} - a_{lag} - t_i^{\star'})}{r_i(h_i - c_h)(t_i^{\star} - dob_i)(t_{im_i} - a_{lag} - t_i^{\star})} \tag{C.39}$$

Using the results from Equation (5.23) and (C.39), we can calculate the acceptance rate of the move:

$$\alpha(u, u') = min(1, \frac{f(u_i'|k_i = 3, else)q(u_i' \rightarrow u_i)}{f(u_i|k_i = 3, else)q(u_i \rightarrow u_i')}) \tag{C.40}$$