Semiparametric Latent Variable Models with Bayesian P-splines

LU, Zhaohua

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Statistics

The Chinese University of Hong Kong
Jun 2010

Thesis/Assessment Committee

Professor LEE Sik-Yum (Chair)

Professor SONG Xin-Yuan (Thesis Supervisor)

Professor CHAN Ping-Shing (Committee Member)

Professor TANG Man-Lai (External Examiner)

Abstract of thesis entitled:

Semiparametric Latent Variable Models with Bayesian P-splines

Submitted by LU, Zhaohua

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in Jun 2010

In medical, behavioral, and social-psychological sciences, latent variable models are useful in handling variables that cannot be directly measured by a single observed variable, but instead are assessed through a number of observed variables. Traditional latent variable models are usually based on parametric assumptions on both relations between outcome and explanatory latent variables, and error distributions. In this thesis, semiparametric models with Bayesian P-splines are developed to relax these rigid assumptions.

In the second part of the thesis, a latent variable model is proposed to relax the first assumption, in which unknown additive functions of latent variables in the structural equation are modeled by Bayesian P-splines. The estimation of nonparametric functions is based on powerful Markov chain Monte Carlo (MCMC) algorithm with block update scheme. A simulation study shows that the proposed method can handle much wider situation than traditional models. The proposed semiparametric latent variable model is applied to a study on osteoporosis prevention and control. Some interesting functional relations, which may be overlooked by traditional parametric latent variable models, are revealed.

In the third part of the thesis, a transformation model is devel-

oped to relax the second assumption, which usually assumes the normality of observed variables and random errors. In our proposed model, the nonnormal response variables are transformed to normal by unknown functions modeled with Bayesian P-splines. This semiparametric transformation model is shown to be applicable to a wide range of statistical analysis. The model is applied to a study on the intervention treatment of polydrug use in which the traditional model assumption is violated because many observed variables exhibit serious departure from normality.

In the fourth part of the thesis, the methodology developed in the third part is further extended to a varying coefficient model with latent variables. Varying coefficient model is a class of flexible semiparametric models in which the effects of covariates are modeled dynamically by unspecified smooth functions. A transformation varying coefficient model can handle arbitrarily distributed dynamic data. A simulation study shows that our proposed method performs well in the analysis of this complex model.

In the last part of the thesis, we propose a finite mixture of varying coefficient models to analyze dynamic data with heterogeneity. A simulation study demonstrates that our proposed method can explore possible existence of different groups in a dynamic data, where in each group the dynamic influences of covariates on the response variables have different patterns. The proposed method is applied to a longitudinal study concerning the effectiveness of heroin treatment. Distinct patterns of heroin use and treatment effect in different patient groups are identified.

摘要

在医学，行为学和社会心理科学中，潜变量模型常用于处理不能直接由单个观测变量测量的变量。这样的潜变量需要通过一系列的观测变量去测量。传统的潜变量模型通常会基于参数假设。参数假设包括：1. 潜变量中的因变量和自变量之间的关系可以用一个已知的函数反映；2. 随机误差服从正态分布。在这篇论文中，基于贝叶斯惩罚样条的半参数模型将用于放松这些参数的假设。

论文的第二部分提出可以放松第一种假设的潜变量模型。这种潜变量模型中的结构方程包含可加的潜变量的未知函数。这些未知函数可以用贝叶斯惩罚样条来逼近。我们用带有块更新（block-update）的马尔科夫链蒙特卡洛算法来估计这些非参数函数。这种模型被应用到一个关于预防骨质疏松症的研究中。

在论文的第三部分，我们提出非参数变换模型放松第二种假设。在大多数模型中，因变量的误差分布都被假定为正态。我们提出的模型只假设因变量在一个未知的变换之后满足一般模型中的正态假设。这种变换模型可以广泛应用于非正态数据的统计分析，并在对治疗滥用多种药物的研究中得到应用。

在论文的第四部分，第三部分提出的非参数变换方法被推广到带潜变量的变系数模型中。变系数模型是一类灵活的半参数模型。自变量的系数可以是一个动态变化的未知函数。这种变系数变换模型可用于动态的非正态数据分析。

在论文的最后一部分，我们提出另一种统计模型来处理动态的非正态数据。在总体存在异质性的情况下，我们提出一个变系数的有限混合模型。模型能够分析由若干个子总体组成的数据，其中每一个子总体可用一个不同的变系数模型进行研究。这个方法被应用于一个关于海洛因治疗效果的研究。

Acknowledgement

I am greatly indebted to my supervisor, Professor Xin-Yuan Song for her supervision, patience, and generous encouragement during the course of this study program. I am very grateful to Professor Sik-Yum Lee for his instructions and help. I also thank Professor Ping-Shing CHAN and Professor Man-Lai Tang for their constructive comments. And I would like to express my gratitude to Professor Xin-Yuan Song, Professor Sik-Yum Lee, and Professor Ming-Gao Gu for their help to the beginning of my career.

Finally, I would like to dedicate this thesis to my parents for their support and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Bayesian P-splines

Studying the relations between variables has always been an important topic in statistics literature. One useful model is the linear regression model (Searle, 1971), which represents the relation between a response (outcome) variable y and a set of indicator (explanatory) variables $\mathbf{x} = (x_1, \ldots, x_q)^T$ through a linear equation:

$$y = \mathbf{x}^T \boldsymbol{\alpha} + \epsilon, \tag{1.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$ is a vector of regression coefficients and ϵ is a random error. Despite its simplicity and ease of use, the model is too restrictive because its linearity assumption may not be true in many practical applications. Nonlinear regression model (Seber and Wild, 2003) was introduced to relax such a rigid assumption, in which the relation between response and indicator variables can be expressed as:

$$y = g(\mathbf{x}, \boldsymbol{\alpha}) + \epsilon, \tag{1.2}$$

where g is a known nonlinear function which comes from theoretical derivation corresponding to some specific application backgrounds. For example, a logistic function can be used to model the growth pattern of soybean plants (Pinheiro and Bates, 2000). Although

modeling with nonlinear function $g(\mathbf{x}, \boldsymbol{\alpha})$ may be supported by certain background theories, whether the model $g(\mathbf{x}, \boldsymbol{\alpha})$ fits the data well must be checked further after model fitting. Besides, such theoretical justifications do not always exist. To tackle these problems, nonparametric and semiparametric modeling techniques were developed for building a more flexible model, where $g(\mathbf{x}, \boldsymbol{\alpha})$ is assumed to be unknown and to be determined by data.

There can be little doubt that nonparametric modeling with various smoothing techniques has a very respectable place in statistics. Approaches for nonparametric modeling include, but are not limited to smoothing splines (Green and Silverman, 1994), kernel methods with local polynomials (Fan and Gijbels, 1996; Fan and Zhang, 1999), regression splines (Hastie and Tibshirani, 1990), and penalized splines (Ruppert et al., 2003). We follow a Bayesian approach because Bayesian analysis has several nice properties. First, useful prior information can be incorporated to produce more accurate estimates. Second, more involved models can be analyzed by the Bayesian approach due to its related powerful sampling-based tools, such as the Markov chain Monte Carlo (MCMC) techniques (Geman and Geman, 1984; Chen et al., 2000; Liu, 2008). Third, as compared with other approaches, the standard error estimates of unknown parameters and model comparison statistics can be easily obtained using the MCMC samples produced in the estimation procedure. Finally, Bayesian approach does not rely on the large-sample asymptotic theory, thereby produces more reliable results even in situations with small sample size (Scheines et al., 1999; Dunson, 2000; Lee and Song, 2004). More recently, due to these nice features of the Bayesian approach, the development of Bayesian methodology for the estimation of smooth functions has become both extensive and fruitful; see, DiMatteo et al. (2001),

Biller and Fahrmeir (2001), Behseta et al. (2005), Panagiotelis and Smith (2008), among others. In particular, Bayesian P-splines (Berry et al., 2002; Lang and Brezger, 2004), which is a Bayesian approach for penalized splines, is appealing because it can flexibly fit unknown smooth function using a large number of basis functions with a simple penalty on differences between coefficients of adjacent basis functions, and simultaneously estimate smooth functions and smoothing parameters. In this thesis, we further apply the Bayesian P-splines to latent variable models (LVMs), transformation models, varying coefficient models, and mixture varying coefficient models, which all contain unknown functions that need to be estimated.

1.2 Latent Variable Models (LVMs)

LVMs generally refer to models with latent variables that are unobserved or cannot be measured by a single observed variable. LVMs have been attracting much attention in the literature (Bollen, 1989; Bentler, 1995; Jöreskog and Sörbom, 1996; Shi and Lee, 2000; Song and Lee, 2005; Lee, 2007, among others). Latent variables may be included in the model for several reasons. First, latent variables can be used to characterize the correlation among observed variables. For example, in the analysis of longitudinal data with repeated measures using mixed effects models (see for example, Lindstrom and Bates, 1988), an individual specific latent variable (random effect) is used to model the correlation among the repeated measures of the same individual at different time points. Second, latent variables can stand for some attributes that are hard to be measured directly, for example, fear, ambition and satisfactory.

Structural equation models (SEMs) are important members in

LVMs. One main goal of SEMs is to study the relations among latent variables. Therefore, SEMs are widely used in behavioral, educational, medical, psychological, and social researches. Classical SEMs such as LISREL (Jöreskog and Sörbom, 1996) and EQS (Bentler, 1995) are based on linear structural equations, in which the relations among latent variables are expressed similarly to (1.1). Researchers later realized that adding nonlinear terms of latent variables to structural equations can improve the model fitting (Kenny and Judd, 1984; Jaccard and Wan, 1995; Schumacker and Marcoulides, 1998; Lee and Zhu, 2002; Song and Lee, 2002; Lee and Song, 2003a). In these articles, the relations among latent variables are expressed similarly to (1.2), in which the nonlinear terms were limited to quadratic and interaction terms of latent variables. In this thesis, we relax the parametric assumption of $g(\cdot)$ in (1.2) and allow this unknown function to be estimated from data. Different from the traditional nonparametric regression modeling, our semiparametric LVMs incorporate latent variables in both sides of equation (1.2).

Besides assuming parametric relations between variables, another assumption usually made by traditional LVMs is the normality of observed variables and random errors. However, this assumption is usually violated in real applications. Transformation models, in which a transformation of data is fitted to the conceived model instead of the data themsevles, are one approach for alleviating the departure of data from the model assumption. In this thesis, we develop a Bayesian semiparametric transformation method which can be applied to a wide range of statistical models including LVMs.

Along with the aforementioned rigid assumptions, traditional LVMs have some other limitations. First, most of the traditional

LVMs are developed to study static data with constant regression coefficients. However, varying regression coefficients, which reflect dynamic influences of covariates on response variables, are of great interest in many studies. Second, traditional LVMs usually assume data coming from a homogeneous population. The statistical analysis based on this assumption cannot accommodate the heterogeneous characteristic existing in commonly encountered real life data. Thus, there is a need to develop more complex LVMs to analyze these kinds of data. In this thesis, we further develop transformation varying coefficient LVMs and mixture varying coefficient LVMs to handle highly nonnormal dynamic data and heterogeneous dynamic data.

1.3 Outline of the Thesis

In this thesis, we apply the Bayesian P-splines method to the models with unknown smooth functions. In Chapter 2, we develop a LVM with a nonparametric structural equation, in which outcome latent variables are influenced by unknown smooth functions of explanatory latent variables and covariates. A simulation study demonstrates that the proposed estimation procedure performs satisfactorily. The method is applied to a study concerning the effects of sexual hormones on the bone mineral density of older men. Some interesting findings are obtained, which are hard to be revealed by traditional parametric LVMs.

Transformation models, which can be dated back to Box and Cox (1964), were developed to alleviate the deviations of data from model assumptions, such as skewness, multimodal, heterogeneity, etc. Ever since this influential paper, a large number of works have been done on transformation models, most of which

considered parametric transformation functions. In Chapter 3, we propose a Bayesian semiparametric transformation model. Nonparametric transformation functions are modeled with Bayesian P-splines. The transformed variables can be fitted to a general nonlinear mixed model, which includes but is not limited to, linear or nonlinear regression models, mixed effect models, factor analysis models, and other LVMs as its special cases. Markov chain Monte Carlo (MCMC) algorithms are implemented to estimate transformation functions and unknown quantities in the model. The performance of the developed methodology is demonstrated by a simulation study. The application to a real study on polydrug use is presented.

In Chapter 4, we study methods to eliminate violations of model assumptions in the varying coefficient model (Hastie and Tibshirani, 1993; Hoover et al., 1998; Fan and Zhang, 2008). Bayesian varying coefficient models are usually based on the normal assumption (Biller and Fahrmeir, 2001; Lang and Brezger, 2004). However, response variables in practical studies may disobey the normality assumption. To alleviate such a violation, we extend the Bayesian semiparametric transformation method to varying coefficient models. A simulation study shows that the proposed method can estimate the model parameters accurately when the nonparametric transformations are applied to non-normal data, and neglecting the violation of the normal assumption or transforming the data by inappropriate parametric transformations may produce misleading results.

In Chapter 5, we develop a finite mixture of varying coefficient models with unknown number of mixture components. Finite mixture approach is one of the approaches to model heterogenous data with non-normal distributions (see for example, Titterington et al.,

1985; McLachlan and Peel, 2000; Cai and Song, 2010). In our model, the unknown number of mixture components is determined by a modified Deviance Information Criterion (DIC, see Spiegelhalter et al., 2002; Celeux et al., 2006; Cai et al., 2010). A Simulation study shows that the estimates of parameters are accurate and the modified DIC can select the right number of components. The model is applied to a longitudinal study of treatment effect on the control of heroin use. Three distinct patterns of treatment effect are identified.

Chapter 2

Semiparametric Latent Variable Models

2.1 Introduction

In the behavioral, biomedical, and social-psychological sciences, it is common to encounter latent variables that cannot be accurately measured by an observed variable, but instead are assessed through many observed variables. Latent variable models (LVMs) (Bollen, 1989; Jöreskog and Sörbom, 1996; Bentler, 1995; Lee, 2007) are useful methods in the assessment of interrelationships among observed and latent variables. Basically, LVMs are formulated by a measurement equation, which is a confirmatory factor analysis model for grouping correlated observed variables to "measure" their corresponding latent variables, and a regression type structural equation with latent variables for examining the effects of explanatory latent variables on outcome latent variables of interest. Because the major objective of LVMs is the analysis of latent variables, the structural equation with latent variables plays the most important role. In traditional LVMs, the structural equation is linear. It has recently become recognized that interaction or quadratic terms of explanatory latent variables should be included in relating or predicting

outcome latent variables (Kenny and Judd, 1984; Schumacker and Marcoulides, 1998). As a result, nonlinear LVMs with nonlinear terms of explanatory latent variables have been developed (see Lee and Zhu, 2002; Song and Lee, 2002; Lee and Song, 2003a; Moustaki, 2003). Although such nonlinear LVMs have been found to be useful, their structural equations are parametric and hence may be too restrictive to correctly reflect the reality. It is thus necessary to consider more general structural equations for revealing the true functional relations among outcome and explanatory latent variables, and covariates.

There can be little doubt that nonparametric modeling with various smoothing techniques has a very respectable place in statistics. Approaches to nonparametric modeling include, but are not limited to smoothing splines (Green and Silverman, 1994), kernel methods with local polynomials (Fan and Gijbels, 1996; Fan and Zhang, 1999), regression splines (Hastie and Tibshirani, 1990), and penalized splines (Ruppert et al., 2003). We follow a Bayesian approach because Bayesian analysis have certain nice properties. First, useful prior information can be incorporated to produce more accurate estimates. Second, more involved models can be estimated with the Bayesian approach due to the powerful sampling-based tools, such as the Markov chain Monte Carlo (MCMC) techniques (Geman and Geman, 1984; Chen et al., 2000; Liu, 2008). Third, inference, such as calculating the standard errors of parameter estimates and model comparison statistics, can be done comparatively easily with the MCMC samples produced in the estimation procedure. Finally, Bayesian approach does not rely on the large-sample asymptotic theory, thereby produces more reliable results even in situations with small sample size (Scheines et al., 1999; Dunson, 2000; Lee and Song, 2004). More recently, due to these nice features of the

Bayesian approach, the development of Bayesian methodology for the estimation of smooth functions has become both extensive and fruitful; see, DiMatteo et al. (2001), Biller and Fahrmeir (2001), Behseta et al. (2005), Panagiotelis and Smith (2008), among others. In particular, Bayesian P-splines (Berry et al., 2002; Lang and Brezger, 2004), which is a Bayesian approach for penalized splines, is appealing because it can flexibly fit unknown smooth function using a large number of basis functions with a simple penalty on differences between coefficients of adjacent basis functions, and simultaneously estimate smooth functions and smoothing parameters. Based on a specific latent variable model without any explanatory latent variable in the structural equation, Fahrmeir and Raach (2007) applied Bayesian P-splines in developing semiparametric methods for analyzing smooth functions of observed covariates and spatial effects in their structural equation. As their focus was on observed covariates and spatial effects, relations among latent variables were not involved in their model, and nonparametric smooth functions of the important latent variables were not accommodated in the crucial structural equation. Hence, the existing nonparametric methods cannot be applied to analyze functional relations among latent variables. To extend the applicability of LVMs, this chapter aims to develop a novel semiparametric LVM, in which the important structural equation is formulated via unspecified smooth functions of latent variables, along with covariates if applicable. The structural equation in our proposed semiparametric LVMs can be regarded as a generalization of the ordinary nonparametric regression model with the new inclusion of unknown smooth functions of latent variables. The Bayesian P-splines approach, together with a MCMC algorithm will be developed to estimate smooth functions, unknown parameters, and latent vari-

ables in the model.

The remainder of this chapter is organized as follows. Section 2.2 defines a nonparametric LVM, in which the structural equation is formulated by a series of unspecified smooth functions of covariates and explanatory latent variables. Section 2.3 introduces the Bayesian P-splines for modeling the unspecified smooth functions. The MCMC sampling and the related computational issues are presented in Section 2.4. In Section 2.5, a simulation study is given to demonstrate the performance of the proposed method, and the methodology is applied to an illustrative example on a study of the bone mineral density in older men. Section 2.6 ends the chapter with a conclusion. Some technical details are provided in Appendix A.

2.2 Model Description

Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{ip})^T$ be a random vector of observed variables and $\boldsymbol{\omega}_i = (\omega_{i1}, \cdots, \omega_{iq})^T$ be a random vector of latent variables for n observations, $p > q$. The relationship between $\mathbf{y}_i$ and $\boldsymbol{\omega}_i$ is given by the following measurement equation:

$$\mathbf{y}_i = \mathbf{A}\mathbf{c}_i + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \tag{2.1}$$

where $\mathbf{c}_i$ is a vector of fixed covariates, $\mathbf{A}$ is a matrix of coefficients, $\boldsymbol{\Lambda}$ is a factor loading matrix, and $\boldsymbol{\epsilon}_i$ is a residual random vector which is independent of $\boldsymbol{\omega}_i$ and has a distribution $N[\mathbf{0}, \boldsymbol{\Psi}]$, in which $\boldsymbol{\Psi}$ is a diagonal matrix. It is worth noting that the measurement equation is regarded as confirmatory with pre-specified number of latent variables and structure of $\boldsymbol{\Lambda}$. These specifications are chosen on the basis of the nature of observed variables and/or prior knowledge from experts, and serve the purpose in identifying the model. As an example, see (2.21) in Section 2.5.2 for a

non-overlapping structure of $\mathbf{\Lambda}$ in the analysis of the BMD data set.

Based on the objective of the substantive study, the latent variables in $\boldsymbol{\omega}$ are distinguished into outcome and explanatory latent variables, then the functional effects of explanatory latent variables on the outcome latent variables are assessed by a nonparametric structural equation. For this purpose, we partition $\boldsymbol{\omega}_i$ to $(\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$, where $\boldsymbol{\eta}_i(q_1 \times 1)$ and $\boldsymbol{\xi}_i(q_2 \times 1)$ are outcome and explanatory latent vectors, respectively. It is assumed that the distribution of $\boldsymbol{\xi}_i$ is $N[\mathbf{0}, \mathbf{\Phi}]$ with a non-diagonal covariance matrix $\mathbf{\Phi}$. To assess the functional effects of $\boldsymbol{\xi}$ on $\boldsymbol{\eta}$, the following nonparametric structural equation is proposed. For an arbitrary element η_{ih} in $\boldsymbol{\eta}_i$, $i = 1, \cdots, n$, and $h = 1, \cdots, q_1$,

$$\eta_{ih} = g_{h1}(x_{i1}) + \cdots + g_{hD}(x_{iD}) + f_{h1}(\xi_{i1}) + \cdots + f_{hq_2}(\xi_{iq_2}) + \delta_{ih}, \quad (2.2)$$

where $x_{i1}, \cdots, x_{iD}$ are fixed covariates that can be directly observed such as age and weight, $g_{h1}, \cdots, g_{hD}, f_{h1}, \cdots, f_{hq_2}$ are unspecified smooth functions with continuous second order derivatives, and δ_{ih} is the residual error with a distribution $N[0, \psi_{\delta h}]$ and is independent of ξ's and δ_{ij}, for $j \neq h$. For notational simplicity, we suppress the subscript h in (2.2) by assuming $q_1 = 1$ in the following sections. An extension to the case with $q_1 > 1$ is straightforward.

Fahrmeir and Raach (2007) (F & R) developed a semiparametric LVM which has a similar measurement equation as equation (2.1) but a different structural equation as follows:

$$\omega_{ir} = g_{r1}(x_{i1}) + \cdots + g_{rd}(x_{id}) + g_{r,spat}(e_i) + \boldsymbol{\gamma}_r^T \mathbf{u}_i + \delta_{ir}, \quad (2.3)$$

where ω_{ir} is the r-th component of $\boldsymbol{\omega}_i$, $g_{r1}, \cdots, g_{rd}$ are nonparametric smooth functions of fixed covariates $x_1, \cdots, x_d$, $g_{r,spat}$ is a spatial effect of the location e_i, $\mathbf{u}$ is another vector of fixed covariates, $\boldsymbol{\gamma}_r$ is a vector of regression coefficients, and δ_{ir} is the random error with

variance 1.0 and is independent of δ_{ik} for $k \neq r$. Comparing (2.3) with our structural equation (2.2), we observe substantial differences between our approach and the F & R approach. First, their approach focused on the effects of fixed covariates and a spatial effect of a given location, involved no partition of $\boldsymbol{\omega}$ into outcome latent vector $\boldsymbol{\eta}$ and explanatory latent vector $\boldsymbol{\xi}$, and their structural equation did not involve explanatory latent variables $\xi_{i1}, \cdots, \xi_{iq_2}$. As a result, neither parametric nor nonparametric effects related to latent variables can be assessed by their approach. Second, the covariance matrix of their latent vector $\boldsymbol{\omega}$ was assumed to be an identity matrix. In our model, the covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ can be non-diagonal. The corrected structure of latent variables in our approach leads to a more complicated computing algorithm because: (i) it requires more sampling steps to cope with the additional parameters involved in the non-diagonal covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ (see steps (b4), (b5), and (b7) in Section 2.4.2), and (ii) more complex MCMC techniques are needed to draw samples in our sampling scheme (see steps (a) and (b7) in Section 2.4.2 and the implementation of the MCMC algorithm in Appendix A). As latent variables in practical applications are usually correlated, the accommodation of non-diagonal covariance matrix is important. Third, due to the presence of explanatory latent variables in our nonparametric structural equation, some extra computational difficulties are encountered in estimating unknown smooth functions of latent variables. In Section 2.3.2, we propose some novel methodologies for solving these difficulties.

2.3 The Bayesian P-splines

2.3.1 General formulation

In analyzing the proposed semiparametric LVM defined by (2.1) and (2.2), modeling the smooth nonparametric function is an important issue. Because of the nice features of Bayesian P-splines (see Lang and Brezger, 2004) and the advantages of the Bayesian approach in analyzing LVMs (see Dunson, 2000; Lee, 2007), we will concentrate on using the Bayesian P-splines approach. To present the basic ideas, we first consider the following simple case:

$$\eta_i = f(\xi_i) + \delta_i, \tag{2.4}$$

and then extend it to the more general case as defined by (2.2). We take the common assumption that $f(\xi_i)$ has a continuous second order derivative, and can be modeled by a sum of B-spline (De Boor, 2001) basis determined by a series of knots in the domain of ξ_i as follows:

$$f(\xi_i) = \sum_{k=1}^{K} \beta_k B_k(\xi_i), \tag{2.5}$$

where K is the number of splines determined by the number of knots, β_k is an unknown parameter, and $B_k(\cdot)$ is a B-spline of appropriate order. A common choice is the cubic B-spline. In practice, a choice of K in the range of 10 to 60 provides flexibility of fitting. To prevent overfitting due to the use of a relatively large number of knots, Eilers and Marx (1996) proposed using a difference penalty on coefficients of adjacent B-splines. More specifically, we shall minimize

$$\sum_{i=1}^{n}(\eta_i - \sum_{k=1}^{K} \beta_k B_k(\xi_i))^2 + \lambda \sum_{k=m+1}^{K} (\triangle^m \beta_k)^2, \tag{2.6}$$

where λ is a smoothing parameter to control the penalty, and $\triangle^m \beta_k$ is the difference penalty defined in a recursive manner as follows: $\triangle^m \beta_k = \triangle^{m-1}\beta_k - \triangle^{m-1}\beta_{k-1}$ (for example, $\triangle\beta_k = \beta_k - \beta_{k-1}$, $\triangle^2\beta_k = \triangle\beta_k - \triangle\beta_{k-1} = \beta_k - 2\beta_{k-1} + \beta_{k-2}$). It can be shown that (2.6) can be expressed as

$$\sum_{i=1}^{n}(\eta_i - \sum_{k=1}^{K} \beta_k B_k(\xi_i))^2 + \lambda \boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta}, \tag{2.7}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)^T$, and $\mathbf{M}$ is the penalty matrix that can be obtained with the specified difference penalty.

In the Bayesian framework (Lang and Brezger, 2004), coefficients β_k are treated as random and the difference penalty in (2.7) is replaced by its stochastic analogues: $\beta_k = \beta_{k-1} + u_k$, where u_k is i.i.d. as $N[0, \tau_\beta]$. With this model framework, the amount of smoothness is controlled by the additional variance parameter τ_β, which corresponds to the inverse of the smoothing parameter λ in (2.7). Consequently, τ_β can be regarded as a new smoothing parameter. Let $K^* = rank(\mathbf{M})$, and the complete-data log-likelihood based on (2.5) is proportional to

$$-\frac{1}{2}\left[\frac{1}{\psi_\delta}\sum_{i=1}^{n}\left\{\eta_i - \sum_{k=1}^{K}\beta_k B_k(\xi_i)\right\}^2 + \frac{1}{\tau_\beta}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}\right] - \frac{n}{2}ln(\psi_\delta) - \frac{K^*}{2}ln(\tau_\beta). \tag{2.8}$$

The above approach can be applied to the more general situation defined by (2.2). By using the sum of B-spline basis in (2.5) to model each of the smooth nonparametric functions, the structural equation (2.2) is reformulated as:

$$\eta_i = \sum_{d=1}^{D}\sum_{k=1}^{K_{bd}} b_{dk} B_{dk}^x(x_{id}) + \sum_{j=1}^{q_2}\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(\xi_{ij}) + \delta_i, \tag{2.9}$$

and the corresponding complete-data log-likelihood is proportional to

$$-\frac{1}{2}\Bigg[\frac{1}{\psi_\delta}\sum_{i=1}^{n}\Bigg\{\eta_i-\sum_{d=1}^{D}\sum_{k=1}^{K_{bd}}b_{dk}B_{dk}^{x}(x_{id})-\sum_{j=1}^{q_2}\sum_{k=1}^{K_j}\beta_{jk}B_{jk}(\xi_{ij})\Bigg\}^2+\sum_{d=1}^{D}\frac{1}{\tau_{bd}}\mathbf{b}_d^T\mathbf{M}_{bd}\mathbf{b}_d+\frac{1}{\tau_{\beta j}}\boldsymbol{\beta}_j^T\,\mathbf{M}_{\beta j}\boldsymbol{\beta}_j\Bigg]-\sum_{j=1}^{q_2}\frac{n}{2}ln(\psi_\delta)-\sum_{d=1}^{D}\frac{K_{bd}^*}{2}ln(\tau_{bd})-\sum_{j=1}^{q_2}\frac{K_j^*}{2}ln(\tau_{\beta j}),\tag{2.10}$$

where $b_{dk}, B_{dk}^x, \beta_{jk}, B_{jk}$, $\mathbf{b}_d, \boldsymbol{\beta}_j, \mathbf{M}_{bd}, \mathbf{M}_{\beta j}$, K_{bd}, K_{bd}^*, K_j, K_j^*, τ_{bd}, and $\tau_{\beta j}$ are defined similarly as in (2.7) and (2.8).

2.3.2 Modeling nonparametric functions of latent variables

The modeling of nonparametric functions of latent variables is a main challenge in this study. Compared with the existing literature, some additional difficulties will be encountered in the current analysis that involves the nonparametric smooth functions of latent variables.

First, traditional B-splines are defined in finite intervals. This induces some difficulties since observations of the latent variables obtained in MCMC iterations may be outside these intervals. To tackle this problem, we consider the B-spline basis for natural cubic splines: $N_{jk}(\xi_{ij})$, $1 \le k \le K_j, 1 \le j \le q_2$. Suppose $(\kappa_{j1}, \kappa_{j2}, \ldots, \kappa_{j,K_j-1}, \kappa_{jK_j})$ are the knots and $(\kappa_{j1}, \kappa_{jK_j})$ are the boundary knots. Each $N_{jk}(\xi_{ij})$ is a piecewise cubic spline which smoothly joins at knots. This property is similar to B-splines inside the boundaries. Moreover, in $(-\infty, \kappa_{j1})$ and (κ_{jK_j}, ∞), $N_{jk}(\xi_{ij})$ is linear and smoothly joined at κ_{j1} and κ_{jK_j} with the part inside the

boundaries. Hence, $\sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij})$ is used to model the unknown smooth function $f_j(\xi_{ij})$. Even if some ξ_{ij} generated in a MCMC iteration exceed the boundaries, $\sum_{j=1}^{q_2} \sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij})$ is still well defined. Thus, the formulation in (2.9) is modified as

$$\eta_i = \sum_{d=1}^{D} \sum_{k=1}^{K_{bd}} b_{dk} B_{dk}^{x}(x_{id}) + \sum_{j=1}^{q_2} \sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij}) + \delta_i. \qquad (2.11)$$

The second difficulty is caused by the unknown scales of latent variables, which makes it impossible to determine the range of a latent variable and the positions of the knots beforehand. To solve this problem, for each ξ_{ij} we choose the product of a free scale parameter s_j and fixed quantiles of $N[0,1]$ as knots. For example, if the quantiles are $(\kappa_1, \kappa_2, \ldots, \kappa_{K_j})$, then the knots for constructing $\{N_{jk}(\xi_{ij})\}_{1 \le k \le K_j}$ are $(s_j\kappa_1, s_j\kappa_2, \ldots, s_j\kappa_{K_j})$. Now, as $\{N_{jk}(\xi_{ij})\}_{1 \le k \le K_j}$ depends on the scale parameter s_j, the formulation in (2.11) should be further modified by

$$\eta_i = \sum_{d=1}^{D} \sum_{k=1}^{K_{bd}} b_{dk} B_{dk}^{x}(x_{id}) + \sum_{j=1}^{q_2} \sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij}|s_j) + \delta_i. \qquad (2.12)$$

The third difficulty is related to the fact that each function $f_j(\xi_{ij})$, $j = 1, \cdots, q_2$, is not identified up to a constant. Inspired by Panagiotelis and Smith (2008) in a simpler context of semiparametric regression, we solve this problem by imposing a constraint on $\boldsymbol{\beta}_j$. Denote $\mathbf{f}_j = (f_j(\xi_{1j}), f_j(\xi_{2j}), \ldots, f_j(\xi_{nj}))^T$ and $\mathbf{1} = (1, \ldots, 1)^T$. In each MCMC iteration, we practically set $\mathbf{1}^T\mathbf{f}_j = 0$. It can be shown that this is equivalent to $\mathbf{1}^T\mathbf{N}_j\boldsymbol{\beta}_j = 0$, where $\mathbf{N}_j = [N_{jk}(\xi_{ij}|s_j)]_{n \times K_j}$ is a matrix in which the elements are the B-spline basis of natural cubic splines. Let $\mathbf{Q}_j = \mathbf{1}^T\mathbf{N}_j$, then the above constraint can be written as $\mathbf{Q}_j\boldsymbol{\beta}_j = 0$. A similar constraint will also be imposed on $\mathbf{b}_d$ to identify nonparametric functions $g_d(x_{id})$, $d = 1, \cdots, D$. Let

$\mathbf{B}_{bd} = [B^x_{dk}(x_{id})]_{n\times K_{bd}}$, and $\mathbf{Q}_{bd} = \mathbf{1}^T\mathbf{B}_{bd}$, then the constraint on $\mathbf{b}_d$ can be written as $\mathbf{Q}_{bd}\mathbf{b}_d = 0$.

2.4 Bayesian Estimation and MCMC Algorithm

2.4.1 Prior distributions

In Bayesian methods, parameters are treated as random. An important issue is to specify appropriate prior distributions of the parameters.

First, we hope to control the scale parameter s_j such that it is not too small or too large. If the scale of s_j is too small, many generated samples of ξ_{ij} may fall outside the boundary knots $(s_j\kappa_1, s_j\kappa_{K_j})$. As discussed in Section 2.3.2, natural cubic splines are defined with linear functions outside the boundary knots. As a result, the nonlinear function of ξ_{ij}, $f_j(\xi_{ij})$, would be estimated by a linear function at a large range of ξ_{ij}, leading to a poor estimation of $f_j(\xi_{ij})$. If the scale of s_j is too large, the generated sample of ξ_{ij} would vary within a small part of a large range, leading to many spline knots being wasted. Consequently, the remaining knots may not be enough to accurately estimate the unknown function of ξ_{ij}. For this purpose, inspired by the (2.14) and (2.17), we propose the prior distribution of $\mathbf{s}$ as follows:

$$p(\mathbf{s}|\tau_{s1}, \cdots, \tau_{sq_2}) \propto \prod_{j=1}^{q_2} \frac{1}{(2\pi\tau_{sj})^{K_j/2}} \exp\Big[-\frac{1}{2\tau_{sj}}\sum_{k=1}^{K_j}\{\ln(|s_j\kappa_k|)\}^2\Big], \tag{2.13}$$

where κ_k is the k-th quantile of $N[0,1]$. The prior distribution in (2.13) is equivalent to penalizing $\{s_1, \ldots, s_{q_2}\}$ with the penalty

$$\sum_{j=1}^{q_2}\sum_{k=1}^{K_j}[\ln(|s_j\kappa_k|)]^2/2\tau_{sj}.$$

Based on (2.13), we have

$$\log p(\mathbf{s}|\cdot) \propto -\frac{1}{2\psi_\delta}\Big[\sum_{i=1}^{n}\{\eta_i - \sum_{d=1}^{D}\sum_{k=1}^{K_{bd}} b_{dk}B_{dk}^{x}(x_{id}) - \sum_{j=1}^{q_2}\sum_{k=1}^{K_j}\beta_{jk}N_{jk}(\xi_{ij}|s_j)\}^2\Big] - \frac{1}{2}\sum_{j=1}^{q_2}\sum_{k=1}^{K_j}\frac{\{\ln(|s_j\kappa_k|)\}^2}{\tau_{sj}},$$

which consists of a quadratic term and a penalty term. A too small or too large s_j results in a large penalty and a small $\log p(\mathbf{s}|\cdot)$. Hence, for a given τ_{sj}, it is less likely to draw a too small or too large scale parameter s_j under prior distribution (2.13). Here, τ_{sj} is an unknown parameter for determining the amount of penalty put on s_j, and it plays the similar role as $\tau_{\beta j}$ and τ_{bd} in the prior distributions specified in (2.14) and (2.17) below. As we never know the true scales of latent variables, this unknown parameter can help to capture the information from the data and thus automatically update the amount of penalty on s_j in the MCMC iterations.

Second, to identify the unspecified smooth function $f_j(\xi_{ij})$, the identifiability constraint $\mathbf{Q}_j\boldsymbol{\beta}_j = 0$ should be imposed on $\boldsymbol{\beta}_j$. Under this constraint, the prior distribution of the unknown parameter $\boldsymbol{\beta}_j$ is assigned as

$$\boldsymbol{\beta}_j|\tau_{\beta j} \propto \exp\{-\frac{1}{2\tau_{\beta j}}\boldsymbol{\beta}_j^T\mathbf{M}_{\beta j}\boldsymbol{\beta}_j\}I(\mathbf{Q}_j\boldsymbol{\beta}_j = 0), \tag{2.14}$$

with appropriate penalty matrix $\mathbf{M}_{\beta j}$, which is a linearly constrained Gaussian density. The prior distribution given in (2.14) is a conjugate type because the posterior distribution of $\boldsymbol{\beta}_j$ given others is still linearly constrained Gaussian with following density:

$$N(\boldsymbol{\beta}_j^*, \boldsymbol{\Sigma}_j^*)I(\mathbf{Q}_j\boldsymbol{\beta}_j = 0), \tag{2.15}$$

where $\boldsymbol{\Sigma}_j^* = (\mathbf{N}_j^T\mathbf{N}_j/\psi_\delta + \mathbf{M}_{\beta j}/\tau_{\beta j})^{-1}$, $\boldsymbol{\beta}_j^* = \boldsymbol{\Sigma}_j^*\mathbf{N}_j^T\boldsymbol{\eta}^*/\psi_\delta$, and $\boldsymbol{\eta}^* =$

$(\eta_1^*, \ldots, \eta_n^*)^T$, in which

$$\eta_i^* = \eta_i - \sum_{d=1}^{D} \sum_{k=1}^{K_{bd}} b_{dk} B_{dk}^x(x_{id}) - \sum_{l \neq j} \sum_{k=1}^{K_l} \beta_{lk} N_{lk}(\xi_{il}|s_l).$$

According to Panagiotelis and Smith (2008), sampling an observation $\boldsymbol{\beta}_j$ from (2.15) is equivalent to sampling an observation $\boldsymbol{\beta}_j^{(\text{new})}$ from $N(\boldsymbol{\beta}_j^*, \boldsymbol{\Sigma}_j^*)$, then $\boldsymbol{\beta}_j^{(\text{new})}$ is transformed to $\boldsymbol{\beta}_j$ by

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(\text{new})} - \boldsymbol{\Sigma}_j^* \mathbf{Q}_j^T (\mathbf{Q}_j \boldsymbol{\Sigma}_j^* \mathbf{Q}_j^T)^{-1} \mathbf{Q}_j \boldsymbol{\beta}_j^{(\text{new})}. \tag{2.16}$$

Similarly, under the identifiability constraint $\mathbf{Q}_{bd}\mathbf{b}_d = 0$, the prior distribution of the unknown parameter $\mathbf{b}_d$ is assigned as

$$\mathbf{b}_d|\tau_{bd} \propto \exp\{-\frac{1}{2\tau_{bd}} \mathbf{b}_d^T \mathbf{M}_{bd} \mathbf{b}_d\} I(\mathbf{Q}_{bd}\mathbf{b}_d = 0). \tag{2.17}$$

The posterior distribution of $\mathbf{b}_d$ given others is

$$N(\mathbf{b}_d^*, \boldsymbol{\Sigma}_{bd}^*) I(\mathbf{Q}_{bd}\mathbf{b}_d = 0), \tag{2.18}$$

where $\boldsymbol{\Sigma}_{bd}^* = (\mathbf{B}_{bd}^T \mathbf{B}_{bd}/\psi_\delta + \mathbf{M}_{bd}/\tau_{bd})^{-1}$, $\mathbf{b}_d^* = \boldsymbol{\Sigma}_{bd}^* \mathbf{B}_{bd}^T \boldsymbol{\eta}_x^*/\psi_\delta$, and $\boldsymbol{\eta}_x^* = (\eta_{x1}^*, \ldots, \eta_{xn}^*)^T$, in which

$$\eta_{xi}^* = \eta_i - \sum_{l \neq d} \sum_{k=1}^{K_{bl}} b_{lk} B_{lk}^x(x_{il}) - \sum_{j=1}^{q_2} \sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij}|s_j).$$

Moreover, in a full Bayesian analysis, the inverse smooth parameters τ_{bd}, $\tau_{\beta j}$, and τ_{sj} are also treated as random and estimated simultaneously with the unknown $\mathbf{b}_d$, $\boldsymbol{\beta}_j$, and s_j. According to a common practice in the literature (see for example, Lang and Brezger, 2004; Fahrmeir and Raach, 2007, and references therein), the highly dispersed (but proper) inverse gamma priors are assigned to these parameters. For $d = 1, \cdots, D$, and $j = 1, \cdots, q_2$:

$$\tau_{bd} \sim IG(\alpha_{b0}, \beta_{b0}),\ \tau_{\beta j} \sim IG(\alpha_{\beta 0}, \beta_{\beta 0}),\ \tau_{sj} \sim IG(\alpha_{\tau 0}, \beta_{\tau 0}) C_j^{-1}(\tau_{sj}), \tag{2.19}$$

where $\alpha_{b0}, \beta_{b0}, \alpha_{\beta 0}, \beta_{\beta 0}, \alpha_{\tau 0}$, and $\beta_{\tau 0}$ are hyperparameters whose values are preassigned, and $\prod_{j=1}^{q_2} C_j(\tau_{sj})$ is the normalizing constant in (2.13). The common choices for these hyperparameters are $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0} = 1$, β_{b0}, $\beta_{\beta 0}$, and $\beta_{\tau 0}$ are small. Throughout this chapter we set $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0} = 1$ and $\beta_{b0} = \beta_{\beta 0} = \beta_{\tau 0} = 0.005$.

Finally, for structural parameters such as $\mathbf{A}, \mathbf{\Lambda}, \mathbf{\Psi}, \psi_\delta$, and $\mathbf{\Phi}$, the following conjugate prior distributions are assigned according to the common practice in LVMs (see Lee, 2007). For $j = 1, \cdots, p$,

$$\begin{aligned} &\mathbf{A}_j \sim N(\mathbf{A}_{j0}, \mathbf{\Sigma}_{aj0}), && \mathbf{\Lambda}_j \sim N(\mathbf{\Lambda}_{j0}, \psi_j \mathbf{\Sigma}_{j0}), \\ &\psi_\delta^{-1} \sim \text{gamma}(\alpha_{\delta 0}, \beta_{\delta 0}), && \psi_j^{-1} \sim \text{gamma}(\alpha_{j0}, \beta_{j0}), \\ &\mathbf{\Phi}^{-1} \sim \text{Wishart}(\mathbf{R}_0, \rho_0), \end{aligned}$$

where $\mathbf{A}_j^T$ and $\mathbf{\Lambda}_j^T$ are the j-th row of $\mathbf{A}$ and $\mathbf{\Lambda}$, respectively; ψ_j is the j-th diagonal element of $\mathbf{\Psi}$, $\mathbf{A}_{j0}$, $\mathbf{\Lambda}_{j0}$, α_{0j}, β_{0j}, $\alpha_{\delta 0}$, $\beta_{\delta 0}$, ρ_0, and positive definite matrices $\mathbf{\Sigma}_{aj0}$, $\mathbf{\Sigma}_{j0}$, and $\mathbf{R}_0$ are hyperparameters whose values are assumed to be given by the prior information.

2.4.2 Posterior inference via MCMC sampling

The Bayesian estimate of $\boldsymbol{\theta}$ is obtained by observations drawn from $p(\mathbf{\Omega}, \boldsymbol{\theta}|\mathbf{Y})$ with some MCMC tools such as the Gibbs sampler (Geman and Geman, 1984) and the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). Let $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)$, $\mathbf{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n)$, $\mathbf{\Omega}_1 = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$, $\mathbf{\Omega}_2 = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$, $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_D\}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{q_2}\}$, $\mathbf{s} = \{s_1, \ldots, s_{q_2}\}$, $\boldsymbol{\tau}_b = \{\tau_{b1}, \ldots, \tau_{bD}\}$, $\boldsymbol{\tau}_\beta = \{\tau_{\beta 1}, \ldots, \tau_{\beta q_2}\}$, $\boldsymbol{\tau}_s = \{\tau_{s1}, \ldots, \tau_{sq_2}\}$, and $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{\Lambda}, \mathbf{\Psi}, \psi_\delta, \mathbf{\Phi}, \boldsymbol{\tau}_b, \boldsymbol{\tau}_\beta, \boldsymbol{\tau}_s, \mathbf{b}, \boldsymbol{\beta}, \mathbf{s}\}$. The computing algorithm in our method is implemented as follows: At the t-th iteration with current values $\{\mathbf{\Omega}^{(t)}, \boldsymbol{\theta}^{(t)}\}$:

(a) drawing $\mathbf{\Omega}^{(t+1)}$ from $p(\mathbf{\Omega}|\mathbf{Y}, \mathbf{A}^{(t)}, \mathbf{\Lambda}^{(t)}, \mathbf{\Psi}^{(t)}, \psi_\delta^{(t)}, \mathbf{\Phi}^{(t)}, \mathbf{b}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)})$;

(b) drawing $\boldsymbol{\theta}^{(t+1)}$ from $p(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\Omega}^{(t+1)})$. Due to its complexity, step (b) is further decomposed into:

(b1) draw $\mathbf{A}^{(t+1)}$ from $p(\mathbf{A}|\mathbf{Y}, \boldsymbol{\Omega}^{(t+1)}, \boldsymbol{\Lambda}^{(t)}, \boldsymbol{\Psi}^{(t)})$;

(b2) draw $(\boldsymbol{\Lambda}^{(t+1)}, \boldsymbol{\Psi}^{(t+1)})$ from $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}|\mathbf{Y}, \boldsymbol{\Omega}^{(t+1)}, \mathbf{A}^{(t+1)})$;

(b3) draw $\psi_\delta^{(t+1)}$ from $p(\psi_\delta|\boldsymbol{\Omega}^{(t+1)}, \mathbf{b}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)})$;

(b4) draw $\boldsymbol{\Phi}^{(t+1)}$ from $p(\boldsymbol{\Phi}|\boldsymbol{\Omega}_2^{(t+1)})$;

(b5) draw $\boldsymbol{\tau}_b^{(t+1)}$ from $p(\boldsymbol{\tau}_b|\mathbf{b}^{(t)})$, $\boldsymbol{\tau}_\beta^{(t+1)}$ from $p(\boldsymbol{\tau}_\beta|\boldsymbol{\beta}^{(t)})$, and $\boldsymbol{\tau}_s^{(t+1)}$ from $p(\boldsymbol{\tau}_s|\mathbf{s}^{(t)})$;

(b6) draw $\mathbf{b}^{(t+1)}$ from $p(\mathbf{b}|\boldsymbol{\Omega}^{(t+1)}, \psi_\delta^{(t+1)}, \boldsymbol{\tau}_b^{(t+1)}, \boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)})$;

(b7) draw $(\boldsymbol{\beta}^{(t+1)}, \mathbf{s}^{(t+1)})$ from $p(\boldsymbol{\beta}, \mathbf{s}|\boldsymbol{\Omega}^{(t+1)}, \psi_\delta^{(t+1)}, \boldsymbol{\tau}_\beta^{(t+1)}, \boldsymbol{\tau}_s^{(t+1)}, \mathbf{b}^{(t+1)})$.

The conditional distributions involved in steps (b1)-(b5) are normal, inverse gamma, and inverse Wishart distributions, respectively. Simulating observations from them is standard (see Lee, 2007). However, the conditional distributions involved in steps (a) and (b7) are nonstandard and need to be derived. The MH algorithm is used to simulate observations from these non-standard distributions. All the full conditional distributions, and the implementation of the MH algorithm are given in Appendix A.

2.5 Numerical Studies

2.5.1 A simulation study

The main purpose of this simulation study is to demonstrate the empirical performance of the proposed approach. The data set is simulated on the basis of the model defined by (2.1) and (2.2) with $p = 12, q = 4, q_1 = 1, q_2 = 3$, and $D = 1$. For $i = 1, \cdots, 300$, the covariate c_i is fixed at 1.0 such that $\mathbf{A}^T = (a_1, \cdots, a_{12})$ is a vector

of intercepts, the covariate x_i is independently drawn from $N[0,1]$, and the latent vector $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \xi_{i3})^T$ is drawn from $N[\mathbf{0}, \boldsymbol{\Phi}]$. The structure of the loading matrix is

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 1 & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \lambda_{52} & \lambda_{62} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{83} & \lambda_{93} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{11,4} & \lambda_{12,4} \end{bmatrix},$$

where the one's and the zero's are fixed parameters in achieving an identified model, and the λ's are unknown parameters. The nonparametric structural equation is

$$\eta_i = g(x_i) + f_1(\xi_{i1}) + f_2(\xi_{i2}) + f_3(\xi_{i3}) + \delta_i, \tag{2.20}$$

with $g(x_i) = (x_i/2)^3$, $f_1(\xi_{i1}) = \sin(1.5\xi_{i1}) - \xi_{i1}$, $f_2(\xi_{i2}) = 1.65 - \exp(\xi_{i2})$, and $f_3(\xi_{i3}) = -0.5 + \exp(2\xi_{i3})/[1 + \exp(2\xi_{i3})]$. The true population values are taken as $a_j = 0.5, \psi_j = \psi_\delta = 0.3, j = 1, \cdots, 12$, $\lambda_{21} = \lambda_{31} = \lambda_{52} = \lambda_{62} = \lambda_{83} = \lambda_{93} = \lambda_{11,4} = \lambda_{12,4} = 0.8$, and $\{\phi_{11}, \phi_{12}, \phi_{13}, \phi_{22}, \phi_{23}, \phi_{33}\}$ in $\boldsymbol{\Phi}$ are $\{1.0, 0.2, 0.2, 1.0, 0.2, 1.0\}$.

In the simulation study, the conjugate priors specified in Section 2.4.1 with the following hyperparameters were used. The elements in $\mathbf{A}_{j0}$ and $\boldsymbol{\Lambda}_{j0}$ were taken as zeros, and $\boldsymbol{\Sigma}_{aj0}$ and $\boldsymbol{\Sigma}_{j0}$ were taken as identity matrices with appropriate dimensions. $\alpha_{j0} = \alpha_{\delta 0} = 9$, $\beta_{j0} = \beta_{\delta 0} = 4$, $\rho_0 = 7$, $\mathbf{R}_0 = 3\mathbf{I}_3$, where $\mathbf{I}_3$ is a three-dimensional identity matrix, $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0} = 1$, and $\beta_{b0} = \beta_{\beta 0} = \beta_{\tau 0} = 0.005$.

A total of 20 equidistant knots were used to construct cubic P-splines of covariate x_i. The 20 knots based on quantiles of $N(0,1)$ multiplied by the scale parameter s_j were adopted for constructing B-spline basis for natural cubic splines of latent variables ξ_{i1}, ξ_{i2}, and ξ_{i3}. The second order random walk penalties were used for the Bayesian P-splines to estimate the unknown smooth functions. On

the basis of 100 replications, the bias (BIAS) and the root mean squares (RMS) between the Bayesian estimates and the true population values of the parameters were computed. The main results are presented in Table 2.1 (Column 'Semipara LVM'), and some less important parameters (a_j and ψ_j, $j = 1, \cdots, 12$) are not reported to save space. We observed that the 'BIAS' and 'RMS' were small, which indicates that the Bayesian estimates of unknown parameters are accurate. Based on 100 replications, the average of the pointwise posterior means of nonparametric functions, together with the 5%- and 95%-pointwise quantiles are presented in Figure 2.1. Compared with their true functions (represented by solid curves), the estimated curves correctly capture the true functional relations among latent and observed variables.

In order to compare the empirical performance of the proposed semiparametric LVM with some parametric LVMs, the data sets in the 100 replications were re-analyzed on the basis of M_1: a simple linear LVM and M_2: a non-linear parametric LVM where the exact forms of functions g, f_1, f_2, and f_3 are known. The estimates of unknown parameters and unknown smooth functions are respectively presented in Table 2.1 (Columns 'Linear LVM' and 'Non-linear LVM') and Figure 2.2, which show that (a) the results obtained from our approach are close to those obtained under M_2. Note that the b_1 and γ's in our approach were estimated by fitting true parametric functions to $\hat{g}(x_i)$ and $\hat{f}_j(\xi_{ij})$. (b) As the misspecification in M_1 only focuses on the structural equation, the performance of parameter estimates associated with the measurement equation in M_1 is similar to that in M_2 and our approach. However, the results associated with the structural equation in M_1 are very misleading (see Figure 2.2 and the bold-faced parameter estimates in Table 2.1). In particular, the large estimated variance

of the residual in the structural equation, $\hat{\psi}_\delta$, reveals inadequacy of a linear LVM in capturing unknown functional relations among latent variables and covariates.

The above analysis was repeated for two different choices of (1, 0.05) and (0.001, 0.001) for $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0}$ and $\beta_{b0} = \beta_{\beta 0} = \beta_{\tau 0}$, and some perturbations of the hyperparamters in the prior distributions of structural parameters. The sensitivity analysis revealed that the Bayesian results are robust to different choices of $\{(\alpha_{b0}, \beta_{b0}), (\alpha_{\beta 0}, \beta_{\beta 0}), (\alpha_{\tau 0}, \beta_{\tau 0})\}$ and the hyperparameters related to structural parameters. It took about 20 minutes in a PC with Intel Core2 3.00GHz CPU and 2G ram for completing the computation for each replication in the simulation study.

2.5.2 Application: A study on osteoporosis prevention and control of older men

The proposed methodology was applied to a partial study on osteoporosis prevention and control which concerns the influence of serum concentration of sex hormones, their precursors and metabolites on the bone mineral density in older men. A total of 1446 Chinese men aged 65 years and above were recruited using a combination of private solicitation and public advertising from community centers and public housing estates. The primary objective of this study is to investigate the functional relations among bone mineral density (BMD) and its correlated determinants including 'Estrogen', 'Androgen', 'Precursors', and 'Metabolites'. We notice from prior medical knowledge that BMD and its correlated determinants are latent constructs that cannot be measured by a single observed variable. For instance, 'BMD' is formed by observed variables 'Spine BMD' and 'Hip BMD' because the bone mineral densities are measured at both spine and hip. How 'Estro-

gen', 'Androgen', 'Precursors', and 'Metabolites' influence BMD functionally is a main interest of the medical research, and the related study will be beneficial to osteoporosis prevention and control. Due to the complex nature of the above mentioned latent variables, we expect that the traditional LVMs with parametric structural equations are inadequate to provide accurate analysis of the true functional relations among latent variables. Hence, the proposed semiparametric LVM is necessary. The following observed variables were selected in establishing a model to achieve the objective: spine BMD, hip BMD, estrone (E1), estrone sulphate (E1-S), estradiol (E2), testosterone (TESTO), 5-Androstenediol (5-DIOL), dihydrotestosterone (DHT), androstenedione (4-DIO NE), dehydroepiandrosterone (DHEA), DHEA sulphate (DHEA-S), androsterone (ADT), ADT glucuronide (ADT-G), 3a-diol-3G (3G), and 3d-diol-17G (17G). Moreover, weight and age were also included as covariates. All the above continuous measurements were standardized.

Based on the medical meaning of the observed variables, we identified five latent variables through the measurement equation (2.1). More specifically, spine BMD and hip BMD were grouped into a latent variable named 'BMD'; similarly, {E1, E1 -S, E2}, {TESTO, 5-DIOL, DHT}, {4-DIONE, DHEA, DHEA -S}, and {ADT, ADT-G, 3G, 3G-17G} were respectively grouped into four latent variables which could be interpreted as 'Estrogen', 'Androgen', 'Precursors', and 'Metabolites', respectively. Hence, we considered a measurement equation with $\mathbf{A} = \mathbf{0}$, a 5 by 1 random vector $\boldsymbol{\omega}$ containing the above latent variables, and following non-

overlapping factor loading matrix,

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 1 & \lambda_{21} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{42} & \lambda_{52} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{73} & \lambda_{83} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{10,4} & \lambda_{11,4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{13,5} & \lambda_{14,5} & \lambda_{15,5} \end{bmatrix}. \tag{2.21}$$

The one's and zero's are fixed parameters for identifying the model according to the common practice of LVMs (see Lee, 2007). The λ_{ij}'s represent unknown factor loadings and reflect the associations between latent variables and the relative observed variables. According to the main objective of this study, which is to investigate the functional effects of 'Estrogen', 'Androgen', 'Precursors', and 'Metabolites' on 'BMD', we define the outcome latent variable $\boldsymbol{\eta}$ to be BMD, and define the vector of explanatory latent variables $\boldsymbol{\xi}$ as $(\xi_1, \xi_2, \xi_3, \xi_4)^T$ =(Estrogen, Androgen, Precursors, Metabolites)^T. Finally, to take also into account the effects of fixed covariates weight and age, the nonparametric structural equation of the proposed model was defined by:

$$\eta_i = g_1(x_{i1}) + g_2(x_{i2}) + f_1(\xi_{i1}) + f_2(\xi_{i2}) + f_3(\xi_{i3}) + f_4(\xi_{i4}) + \delta_i, \tag{2.22}$$

where x_{i1} and x_{i2} are weight and age, respectively. The weight effect and age effect were modeled by cubic P-splines, and the effects of the explanatory latent variables on BMD were modeled by natural cubic P-splines, which are combinations of B-spline basis for natural cubic splines and difference penalties. A first order random walk penalty and 20 knots were used in the analysis. As specified in Section 2.4.1, the conjugate priors were used as the prior distributions for most of the unknown parameters.

The diffuse hyperpriors $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0} = 1$ and $\beta_{b0} = \beta_{\beta 0} = \beta_{\tau 0} = 0.005$ were used for the inverse smoothing parameters in $\boldsymbol{\tau}_b, \boldsymbol{\tau}_\beta$, and $\boldsymbol{\tau}_s$. To obtain some prior knowledge for the structural parameters, we conducted a preliminary analysis for the current data set by using a traditional LVM, which has been defined by measurement equation (2.1) and a linear structural equation $\eta_i = b_1 x_{i1} + b_2 x_{i2} + \gamma_1 \xi_{i1} + \gamma_2 \xi_{i2} + \gamma_3 \xi_{i3} + \gamma_4 \xi_{i4} + \delta_i$. The standard package LISREL 8 (Jöreskog and Sörbom, 1996) produced the maximum likelihood (ML) estimates of unknown parameters for the parametric LVM. The hyperparameters in the prior distributions of structural parameters were taken as: $\boldsymbol{\Lambda}_{j0} = \hat{\boldsymbol{\Lambda}}_j$, $\alpha_{j0} = \alpha_{\delta 0} = 9$, $\beta_{j0} = (\alpha_{j0} - 1)\hat{\psi}_j$, $\beta_{\delta 0} = (\alpha_{\delta 0} - 1)\hat{\psi}_\delta$, $\rho_0 = 8$, and $\mathbf{R}_0 = (\rho_0 - q_2 - 1)\hat{\boldsymbol{\Phi}}$, where $\hat{\boldsymbol{\Lambda}}_j, \hat{\psi}_j, \hat{\psi}_\delta$, and $\hat{\boldsymbol{\Phi}}$ were the ML estimates obtained via the parametric LVM.

After checking the convergence, we found that the MCMC algorithm converged within 10,000 iterations. After discarding 10,000 burn-in iterations, 20,000 observations generated by the MCMC algorithm were used to obtain the Bayesian results. The estimates of factor loadings and their standard error estimates are reported in Figure 2.3. For clarity, the less important structural parameters in $\boldsymbol{\Psi}, \psi_\delta$, and $\boldsymbol{\Phi}$ are not reported. The pointwise posterior means of unknown smooth functions, together with the 10%- and 90%-pointwise quantiles are depicted in Figure 2.4. We observed that most of the fitted curves were neither linear nor quadratic. This fact provides verification that traditional parametric LVMs with linear and/or quadratic terms of latent variables cannot correctly reflect the true functional relations among latent and observed variables, and would give misleading conclusion if the data were analyzed through a parametric approach.

The specific interpretation of the functional relations are given

as follows. (i) Weight had a positive effect on BMD. Roughly speaking, this effect rose linearly. The increasing rate slowed down when subjects were overweight, say with weights exceeding 80kg. (ii) The effect of age on BMD was basically negative. This negative effect was less significant for 65-75 years old men, but became increasingly significant when the subjects were over 75 years old. (iii) The influence of Estrogen on BMD rose with Estrogen score, indicating that the subjects with a higher level of Estrogen would have had a higher level of BMD and thus a lower risk of osteoporotic fractures. (iv) The influence of Androgen on BMD exhibited a nonlinear pattern. It changed direction from positive to negative, indicating a positive effect for subjects with low Androgen scores and a negative effect for those with moderate or high Androgen scores. Therefore, some insights about the influence of Androgen on BMD might be achieved: although increasing Androgen might have had a positive impact on BMD for those with a low level of Androgen, for most of subjects, however, controlling Androgen level could have helped improving BMD and thus preventing the development of osteoporotic fractures. (v) The influence of Precursors on BMD was hardly significant when Precursors level was low, and it became more and more significant with the increase of Precursors level. Since Precursors play an important role in producing sex hormones, for older men with relatively high Precursors level, controlling Precursors might have helped controlling Androgen and thus improving BMD. (vi) The influence of Metabolites on BMD presented a composite effect of an increasing trend and a sinusoidal shape for periodic pattern, which resulted in an overall increase trend but with significant changes in both tails of the curve. If we partitioned the subjects into three types with respectively low, middle, and high levels of Metabolites, the estimated

curve revealed that the effect of Metabolites on BMD was clearly nonlinear and had completely different patterns for different types of subjects. Therefore, different treatments should be taken to improve BMD thus prevent the development of osteoporotic fractures for older men with low, middle, and high levels of Metabolites. The above insights obtained from nonlinear curves cannot be achieved by parametric LVMs.

To assess the sensitivity of the Bayesian results to inputs of hyperparameters in the prior distributions, the above analysis was repeated with some *ad hoc* perturbations of the current prior inputs. In particular, two different choices of $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0} = 1$, $\beta_{b0} = \beta_{\beta 0} = \beta_{\tau 0} = 0.05$ and $\alpha_{b0} = \alpha_{\beta 0} = \alpha_{\tau 0} = 0.001$, $\beta_{b0} = \beta_{\beta 0} = \beta_{\tau 0} = 0.001$ were used. As close Bayesian estimates of unknown parameters and similar estimated curves of unknown smooth functions were obtained, the Bayesian results are not very sensitive to different prior inputs. The program is written in R. It took about 60 minutes in a PC with Intel Core2 3.00GHz CPU and 2G ram to complete all numerical results in this example.

2.6 Conclusion

In this chapter, a nonparametric LVM is proposed to assess the functional relations among latent and observed variables. Different from traditional LVMs, the proposed model formulates the structural equation in a nonparametric way by introducing a series of unspecified smooth functions. The Bayesian P-splines incorporating MCMC techniques are employed to conduct the analysis. Some additional difficulties, such as unfixed ranges and un-predetermined scales of latent variables in each MCMC iteration, are encountered in the assessment of the functional relations among latent vari-

ables. Hence, a modified Bayesian P-splines approach is introduced to solve the problems. Results obtained from a simulation study demonstrate that the empirical performance is satisfactory. The proposed approach is also applied to study the influences of serum concentrations of sex hormones, their precursors and metabolites on bone mineral density in older men. This study will be helpful in a long-term project in osteoporosis prevention and control.

Table 2.1: The comparison of parameter estimates under three models in the simulation study $n = 300$, number of replications $= 100$

		Semipara LVM		Linear LVM		Non-linear LVM	
Para	TRUE	BIAS	RMS	BIAS	RMS	BIAS	RMS
λ_{21}	0.80	-0.001	0.018	0.005	0.021	0.001	0.019
λ_{31}	0.80	-0.001	0.019	0.003	0.020	-0.001	0.019
λ_{52}	0.80	0.022	0.044	0.022	0.054	0.013	0.049
λ_{62}	0.80	0.021	0.048	0.021	0.045	0.009	0.037
λ_{83}	0.80	0.022	0.050	0.015	0.049	0.012	0.046
λ_{93}	0.80	0.023	0.048	0.008	0.053	0.007	0.051
$\lambda_{11,4}$	0.80	0.007	0.049	0.016	0.049	0.014	0.048
$\lambda_{12,4}$	0.80	0.016	0.046	0.017	0.047	0.014	0.046
ϕ_{11}	1.00	-0.057	0.110	-0.069	.0126	-0.049	0.111
ϕ_{12}	0.20	-0.022	0.070	-0.003	0.054	-0.004	0.054
ϕ_{13}	0.20	-0.010	0.064	-0.019	0.068	-0.017	0.067
ϕ_{22}	1.00	-0.037	0.113	-0.037	0.122	-0.029	0.117
ϕ_{23}	0.20	0.003	0.058	-0.006	0.073	-0.003	0.074
ϕ_{33}	1.00	-0.025	0.116	-0.043	0.111	-0.040	0.111
ψ_δ	0.30	0.020	0.040	**2.080**	**2.814**	0.036	0.056
b_1	1.00	0.003	0.151	**-0.708**	**0.720**	-0.019	0.130
γ_1	1.00	0.022	0.128	**-1.511**	**1.523**	0.027	0.121
γ_2	1.00	0.030	0.088	**0.695**	**0.736**	0.041	0.103
γ_3	1.00	0.023	0.186	**-0.719**	**0.731**	-0.029	0.163

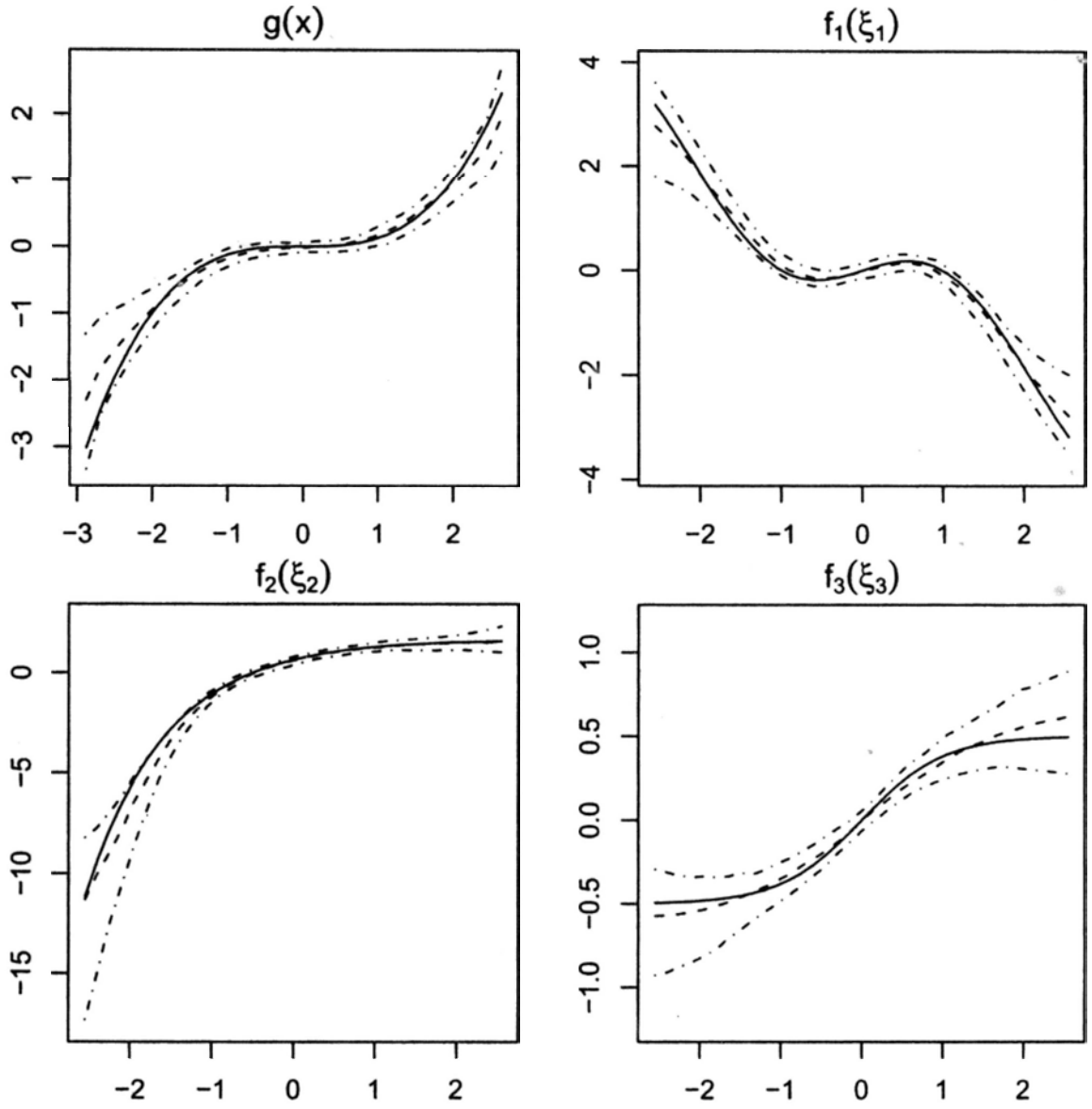

Figure 2.1: Estimates of the unknown smooth functions in simulation study. The solid curves represent the true curves, the dashed and dot-dashed curves respectively represent the estimated average, and the 5%- and 95%-quantiles of the pointwise posterior means on the basis of 100 replications.

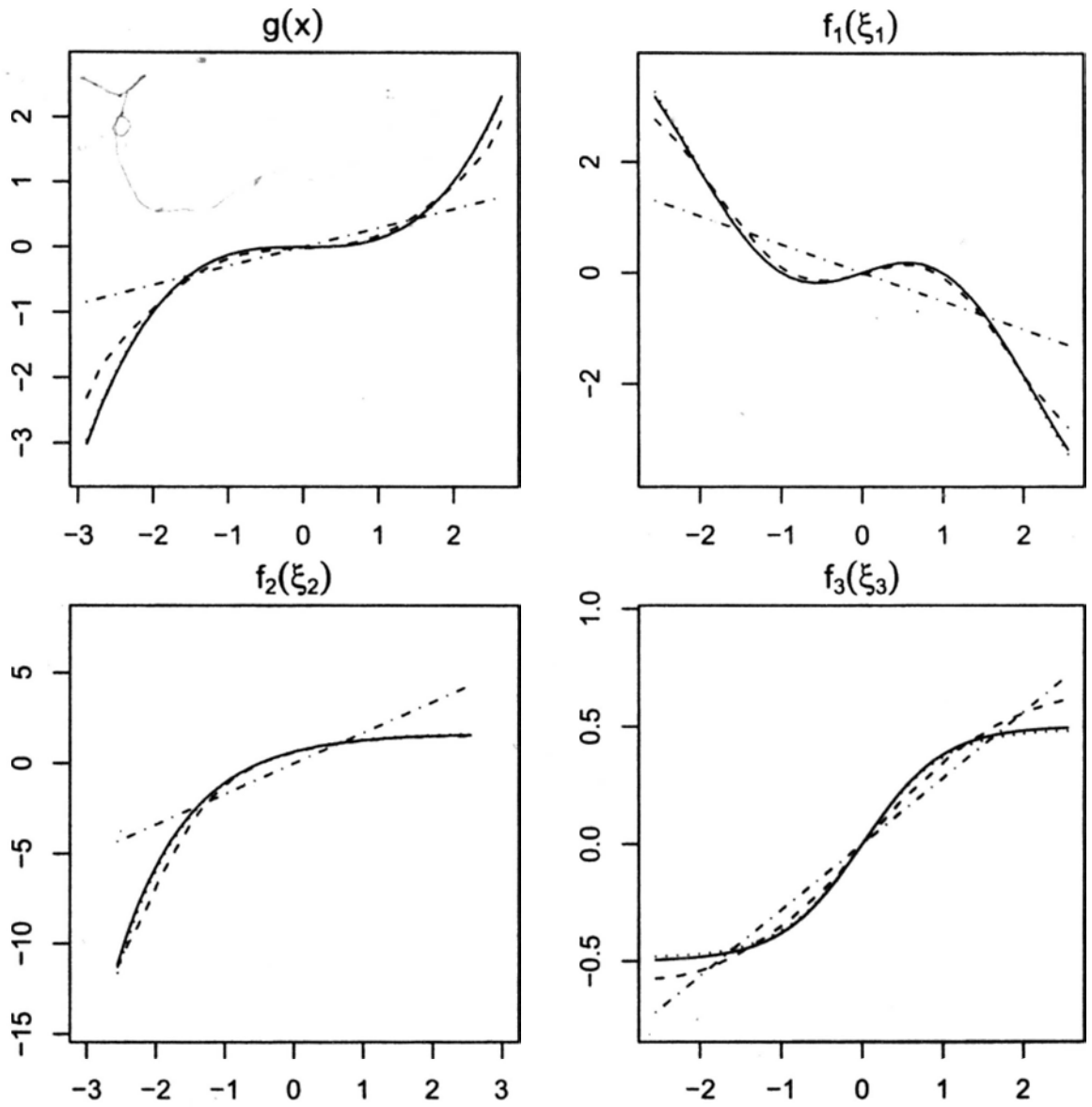

Figure 2.2: Estimates of the unknown smooth functions in the simulation study. The dashed curves are estimated with our semiparametric LVM. The dot-dash curves are estimated with a linear LVM. The dotted curves are estimated with a non-linear parametric LVM when the exact forms of g, f_1, f_2, and f_3 are know. The solid curves are true curves. All the estimated curves are obtained on the basis of 100 replications.

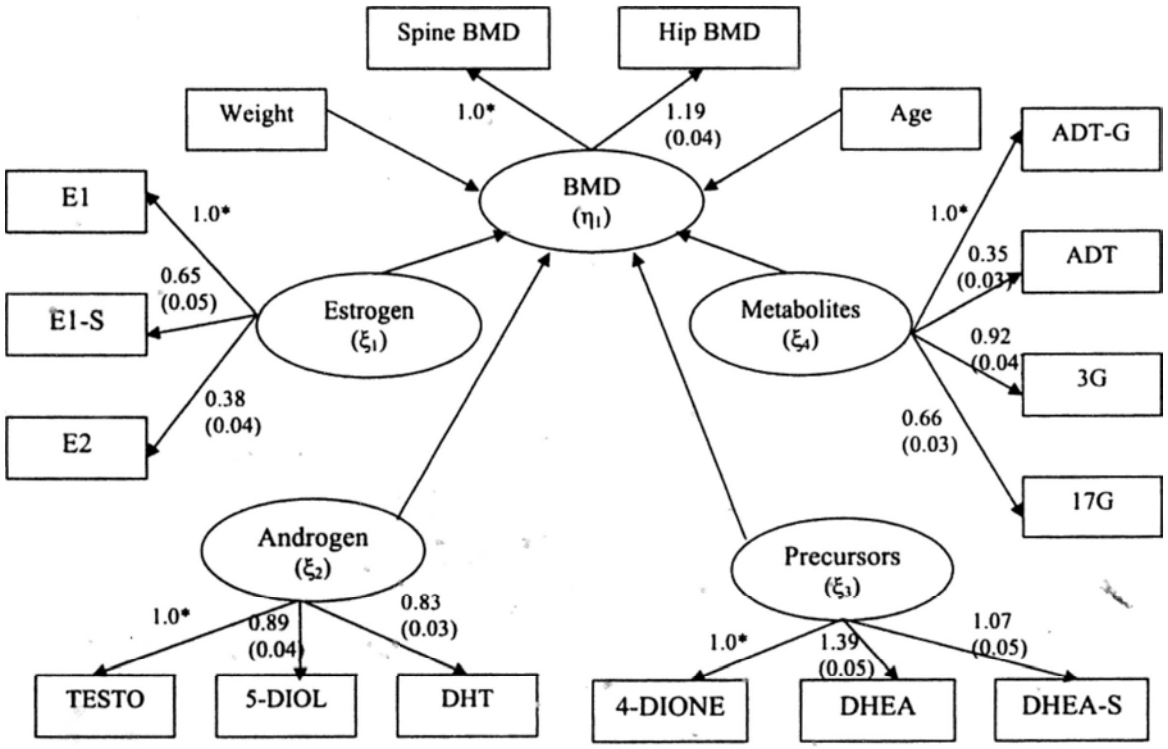

Figure 2.3: The path diagram, together with the estimated factor loadings and their standard error estimates (in parentheses), of the proposed LVM in the analysis of BMD data.

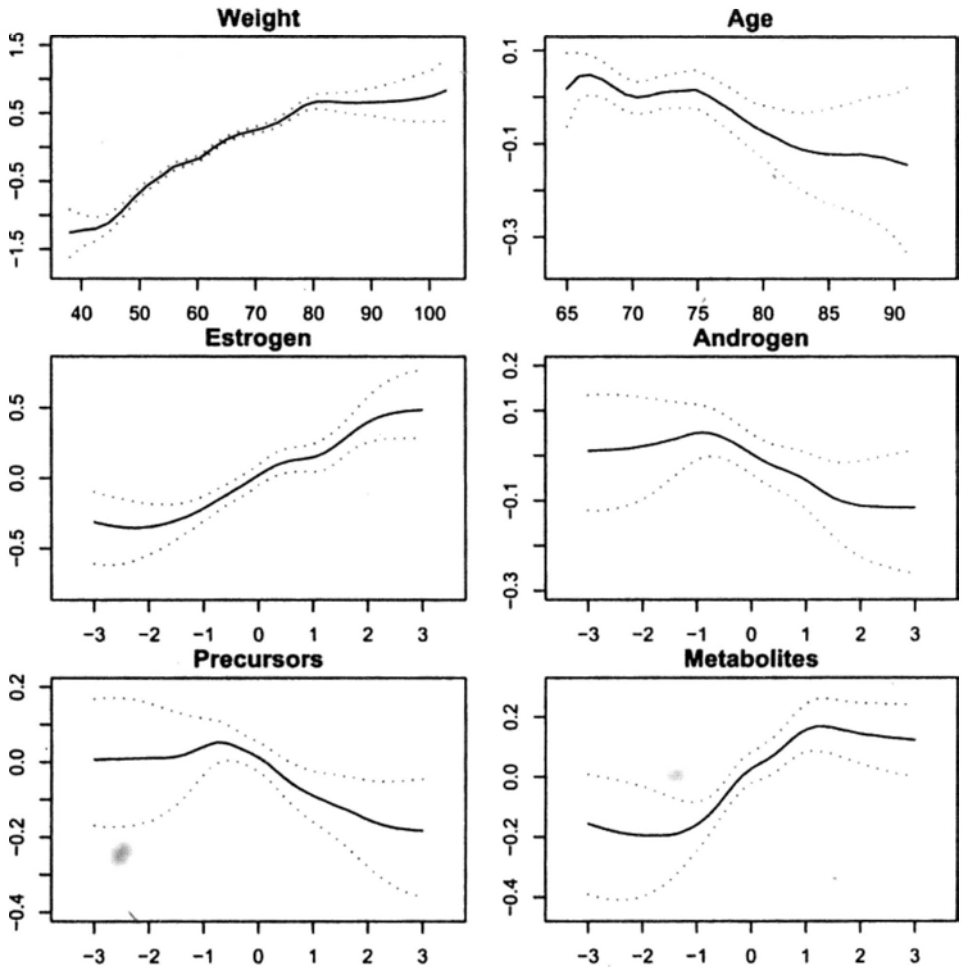

Figure 2.4: Estimates of the unknown smooth functions in the real example. The solid curves represent the estimated pointwise posterior mean curves, while the dotted curves represent the 10%- and 90%-pointwise quantiles.

Chapter 3

Semiparametric Transformation Models

3.1 Introduction

The semiparametric transformation models proposed in this chapter are motivated from a study of the dropout behavior in a treatment program which aimed at preventing polydrug use in five California counties in 2004. Two main interests of this study are to investigate how the history of drug use and convicted crime affects retention in the drug treatment, and how this retention will affect future drug use. One obstacle in analyzing the data set related to the study is that the distributions of most variables are extremely non-normal, such as U-shaped and highly skewed distributions. For example, due to the fact that drug users either did not get addict to drugs or used them very frequently, the distribution of the variable that measures patients' drug use in the past 30 days at intake was U-shaped. Another example is the variable that records the number of arrests in the lifetime of the drug users before entering this treatment. As only a few subjects were arrested frequently, the corresponding distribution was highly skewed.

To introduce the transformation models, we consider the fol-

lowing statistical model $y_i = g(\mathbf{x}_i, \boldsymbol{\gamma}, \mathbf{b}_i) + \epsilon_i$, $i = 1, \ldots, n$, where y_i is a response variable; g is a function of a vector of fixed covariates $\mathbf{x}_i$, a vector of unknown parameters $\boldsymbol{\gamma}$, and a vector of random effects $\mathbf{b}_i$; and ϵ_i is the random error, which is independent of $\mathbf{b}_i$. The distribution of the response variable is usually assumed to be normal. One common approach in handling the violation of this normal assumption is using a transformation model $f(y_i) = \tilde{y}_i = g(\mathbf{x}_i, \boldsymbol{\gamma}, \mathbf{b}_i) + \epsilon_i$, in which the response variable y_i in the model is transformed to $\tilde{y}_i$. The transformation function $f(\cdot)$ was first modeled by parametric functions in the sense that these functions could be specifically determined by a small number of parameters. (Box and Cox, 1964) provided a family of power transformations and estimated it through both maximum likelihood (ML) and Bayesian approaches. Since then, various extensions and other parametric transformation families have been proposed under the ML framework; for example, see John and Draper (1980), Bickel and Doksum (1981), David (1993), Lipsitz et al. (2000), Foster et al. (2001), and references therein. With the rapid development of statistical computing, the Bayesian parametric transformation also received much attention in the literature; for example, see Pericchi (1981), Smith and Kohn (1996), Hoeting et al. (2002), and Yin and Ibrahim (2006).

As parametric transformations may not provide good solutions for many situations, it is natural to consider transformation models with unspecified smooth functions. This nonparametric modeling has been developed rapidly in a non-Bayesian framework; for example, see Ramsay (1988), Tibshirani (1988), Nychka and Ruppert (1995), Wang and Ruppert (1995), He and Shen (1997), and Ramsay (1998), and references therein. Although the development of the nonparametric transformation under the non-Bayesian frame-

work has been fruitful, research with a Bayesian framework is limited. As far as we know, the only contribution is from Mallick and Walker (2003), whose work focused on a regression model without random effects. In their model, the nonparametric function in modeling the response variable was based on a composite function of a mixture of incomplete beta functions with unknown weights but with a pre-determined base function. The number of incomplete beta functions was also unknown; hence, the dimension of the parameter space was unknown. Estimation was based on reversible jump Markov chain Monte Carlo (MCMC) algorithm. Moreover, because their model and the MCMC algorithm were specifically designed for survival analysis, the domain of each response variable was limited to $(0, +\infty)$.

There are other classes of models related to the nonparametric transformation model. One class is the generalized linear model (GLM) based on unknown link functions, which can be treated as transformation functions of the means of the response variables. Using notations above, the model of this class in the case of a normal response variable can be expressed as $E(y_i|\mathbf{b}_i) = f(g(\mathbf{x}_i, \boldsymbol{\gamma}, \mathbf{b}_i))$, which differentiates itself from the transformation models described above because it transforms the means of the response variables instead of response variables themselves. Mallick and Gelfand (1994) used the method similar to Mallick and Walker (2003) to study the unknown link function in GLM, which was applied by Mallick and Gelfand (1996) to study errors-in-variables models. Muggeo and Ferrara (2008) used P-splines to model unknown link functions in GLM by likelihood-based method. In these papers, $g(\mathbf{x}_i, \boldsymbol{\gamma}, \mathbf{b}_i)$ are all linear as $\mathbf{x}_i^T\boldsymbol{\gamma}$ and without $\mathbf{b}_i$. Another class of model is the curve registration studied by Telesca and Inoue (2008) and the reference therein. The model in Telesca and

Inoue (2008) can be expressed as $y_i(x) = g_i(x, \boldsymbol{\gamma}, \mathbf{b}_i) + \epsilon_i$ where x is a scalar, $y_i(x)$ is the i-th curve at x. A transformation $h(x)$ on x was applied through $g_i(h^{-1}(x), \boldsymbol{\gamma}, \mathbf{b}_i)$ and $h(x)$ was modeled by Bayesian P-splines. The model was specifically designed for functional data $y_i(x)$ and the transformation was applied through x. How this methodology can be applied to common data y_i is not clear.

In this chapter, we consider a general nonlinear mixed effect model with random effects, which subsumes the regression model, mixed effect model, and factor analysis model as its special cases. As our proposed model involves random effects, and the domain of each response variable is $(-\infty, \infty)$, the estimation method developed in Mallick and Walker (2003) cannot be applied. Inspired by the recently developed efficient Bayesian methods for analyzing general nonparametric functions, such as DiMatteo et al. (2001), Biller and Fahrmeir (2001), Lang and Brezger (2004), Brezger and Steiner (2008) among others, we develop our MCMC algorithm with Bayesian P-splines in estimating the transformation functions, unknown parameters, and random effects in the proposed model. Hence, our objectives and estimation method are substantially different from those given in Mallick and Walker (2003).

The chapter is organized as follows. Section 3.2 defines a semiparametric transformation nonlinear mixed model, and introduces Bayesian P-splines with monotonic constraints. The Markov chain Monte Carlo (MCMC) sampling and the related computational issues are presented in Section 3.3. In Section 3.4, a simulation study demonstrates that the proposed methodology is effective in handling highly non-normal variables in the context of a nonlinear mixed effect model, and the method is applied to a real study about polydrug use. Setion 3.5 gives a conclusion. Some details of

the real study are provided in Appendix B.

3.2 Model Description

3.2.1 General model formulation

For $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip_i})^T$ be a random vector of observed variables measured in each of the n independently distributed observations. In practice, for each subject i, y_{ij} can represent one of the repeated measurements at p_i different time points, or the measurement corresponding to the j-th item in a questionnaire. Let f_j be an unspecified smooth transformation function for y_{ij}, and $\tilde{y}_{ij} = f_j(y_{ij})$. Like parametric transformations, the unknown functions f_j are assumed to be strictly monotonic increasing. A semiparametric transformation nonlinear mixed model is defined as follows:

$$f_j(y_{ij}) = \tilde{y}_{ij} = g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) + \epsilon_{ij}, \tag{3.1}$$

where $\mathbf{x}_{ij}$ is a vector of covariates, $\boldsymbol{\gamma}$ is a vector of unknown parameters, $\mathbf{b}_i$ is a vector of random effects, and ϵ_{ij} is a random error independently distributed as $N(0, \sigma_j^2)$. It is assumed that $\mathbf{b}_i$ is independent of ϵ_{ij} and follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Phi})$. The $g(\cdot)$ in model (3.1) is a specified general function that can be used to subsume a wide range of statistical models such as linear or nonlinear regression models, random effect models, multilevel models, factor analysis models, and some other latent variable models.

3.2.2 Bayesian P-splines and prior distributions

The transformation functions f_j introduced in (3.1) are strictly monotonic increasing smooth functions. These functions are un-

specified and have to be determined by the data. Modeling such nonparametric transformation functions to make the distribution of $\tilde{y}_{ij}$ close to normal is an important issue. Bayesian P-splines (Lang and Brezger, 2004), which is a Bayesian analog to the P-splines method proposed by Eilers and Marx (1996), has been found to be useful in modeling unspecified smooth functions. We will develop an efficient approach in modeling nonparametric transformation functions through some modifications of the Bayesian P-splines.

The idea of P-splines is that the unknown smooth function $f_j(y_{ij})$ is approximated by the following sum of B-splines $B_{jk}(y_{ij})$, $\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij})$ (De Boor, 2001), where K_j is the number of splines determined by the number of knots in the support of y_{ij}, and $\beta_{j1}, \cdots, \beta_{jK_j}$ are unknown coefficients. With $f_j(y_{ij})$ approximated by P-splines, model (3.1) can be rewritten as

$$\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) = g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) + \epsilon_{ij}. \tag{3.2}$$

The complete data likelihood corresponding to model (3.2) is as follows:

$$\begin{aligned} &\prod_{i=1}^{n} \prod_{j=1}^{p_i} \Bigg[\frac{1}{\sqrt{2\pi}\sigma_j} \exp\Big\{ -\frac{1}{2\sigma_j^2} \Big(\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) - \\ &g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) \Big)^2 \Big\} \sum_{k=1}^{K_j} \beta_{jk} B'_{jk}(y_{ij}) \Bigg] \frac{1}{(\sqrt{2\pi})^{nq} |\boldsymbol{\Phi}|^{n/2}} \times \\ &\exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} \mathbf{b}_i^T \boldsymbol{\Phi}^{-1} \mathbf{b}_i \Big\}, \end{aligned} \tag{3.3}$$

where q is the dimension of $\mathbf{b}_i$, $\sum_{k=1}^{K_j} \beta_{jk} B'_{jk}(y_{ij})$ is the Jacobian of the transformation, and $B'_{jk}(y_{ij})$ is the first derivative of $B_{jk}(y_{ij})$.

The flexibility of this approach in modeling f_j is achieved by a large number (from 10 to 60) of $B_{jk}(y_{ij})$, while the smoothness of

f_j is controlled by penalizing very large differences between coefficients of adjacent B-splines. In a Bayesian framework, this penalization is conveniently incorporated to the coefficients β_{jk} through the first- or the second-order random walk priors. The first- and second-order random walks are defined as:

$$\beta_{jk} = \beta_{j,k-1} + u_{jk}, \quad \text{and} \quad \beta_{jk} = 2\beta_{j,k-1} - \beta_{j,k-2} + u_{jk}, \tag{3.4}$$

with $u_{jk} \sim N(0, \tau_j^2)$, and a diffuse prior $\beta_{j1} \propto$ constant for the first-order random walk, and $\beta_{j1} \propto$ constant and $\beta_{j2} \propto$ constant for the second-order random walk. The variance τ_j^2 can be viewed as an inverse smoothing parameter, which determines the smoothness of the resulting function f_j.

Let $\boldsymbol{\beta}_j = (\beta_{j1}, \cdots, \beta_{jK_j})^T$, and the resulting prior distribution of $\boldsymbol{\beta}_j$ is

$$p(\boldsymbol{\beta}_j) \propto \left(\frac{1}{\sqrt{2\pi}\tau_j}\right)^{K_j - d} \exp\left\{ -\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j^T \mathbf{M} \boldsymbol{\beta}_j \right\}, \tag{3.5}$$

where d is the order of the random walk, $\mathbf{M} = (\mathbf{D}_{d-1} \times \cdots \times \mathbf{D}_0)^T (\mathbf{D}_{d-1} \times \cdots \times \mathbf{D}_0)$, and $\mathbf{D}_l$ is a $(K_j - l - 1) \times (K_j - l)$ matrix:

$$\mathbf{D}_l = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}, \quad l = 0, \cdots, d-1. \tag{3.6}$$

To ensure $\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij})$ to be strictly monotonic increasing, the constraint $\beta_{j1} < \cdots < \beta_{jK_j}$ is needed. This constraint can be incorporated in the analysis by assigning the following prior distribution of $\boldsymbol{\beta}_j$:

$$\left(\frac{1}{\sqrt{2\pi}\tau_j}\right)^{K_j - d} \exp\left\{ -\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j^T \mathbf{M} \boldsymbol{\beta}_j \right\} I(\beta_{j1} < \cdots < \beta_{jK_j}), \tag{3.7}$$

where $I(\cdot)$ is an indicator function. Moreover, we realize that the model may not be identified in the following situations. First, the coefficients in $\boldsymbol{\gamma}$ are linear, for example, $g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) = \boldsymbol{\gamma}_x \mathbf{x}_{ij} + \boldsymbol{\gamma}_b \mathbf{b}_i$, where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_x, \boldsymbol{\gamma}_b)$. As $(\boldsymbol{\beta}_j, \boldsymbol{\gamma}, \sigma_j)$ and $(c\boldsymbol{\beta}_j, c\boldsymbol{\gamma}, c\sigma_j)$ yield the same likelihood (3.3) for an arbitrary constant c under this case, the model is not identified. Hence, σ_j^2 is fixed at 1.0 for identification purposes. Second, an intercept, say μ, exists in the function $g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i)$. As $\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) - \mu = \sum_{k=1}^{K_j} (\beta_{jk} + c) B_{jk}(y_{ij}) - (\mu + c)$, which results in a non-identified model because both (β_{jk}, μ) and $(\beta_{jk} + c, \mu + c)$ yield the same likelihood (3.3). To achieve identification, we impose further constraints on $\boldsymbol{\beta}_j$ as follows. Let $\mathbf{B}_{ij} = (B_{j1}(y_{ij}), \ldots, B_{jK_j}(y_{ij}))^T$, $\mathbf{B}_j = (\mathbf{B}_{1j}, \ldots, \mathbf{B}_{n_j,j})^T$, in which n_j is the number of observations at the j-th measurement, and $\mathbf{1} = (1, \ldots, 1)^T$. To achieve identification, we set $\sum_{i=1}^{n_j} \sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) = 0$, which is equivalent to $\mathbf{1}^T \mathbf{B}_j \boldsymbol{\beta}_j = 0$. Let $\mathbf{Q}_j = \mathbf{1}^T \mathbf{B}_j$. The constraint becomes $\mathbf{Q}_j \boldsymbol{\beta}_j = 0$, and it can be incorporated by adding an additional constraint to the prior distribution of $\boldsymbol{\beta}_j$ as follows:

$$C(\tau_j) \exp\left\{ -\frac{1}{2\tau_j^2} \boldsymbol{\beta}_j^T \mathbf{M} \boldsymbol{\beta}_j \right\} I(\beta_{j1} < \cdots < \beta_{jK_j}, \mathbf{Q}_j \boldsymbol{\beta}_j = 0), \quad (3.8)$$

where $C(\tau_j) = (1/\sqrt{2\pi}\tau_j)^{K_j - d}$. In a full Bayesian analysis, the inverse smoothing parameters τ_j^2 are treated as random. According to a common practice,

$$\tau_j^{-2} \sim Gamma(\alpha_{j1}, \alpha_{j2}), \quad (3.9)$$

where α_{j1}, and α_{j2} are specified hyperparameters. In this chapter we use $\alpha_{j1} = 1$ and $\alpha_{j2} = 0.005$ to obtain a highly dispersed (but proper) gamma prior of τ_j^{-2}.

For the parameters involved in the righthand side of model (3.1) or (3.2), the following conjugate prior distributions are assigned.

$$\boldsymbol{\gamma} \sim N(\boldsymbol{\gamma}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Phi}^{-1} \sim Wishart(\mathbf{R}_0, r_0), \quad (3.10)$$

where γ_0, r_0, and positive definite matrices Σ_0 and $\mathbf{R}_0$ are hyperparameters whose values are assumed to be given by the prior information.

3.3 Estimation of Nonparametric Transformation

Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{ip_i})^T$, $\mathbf{Y} = \{\mathbf{y}_i; i = 1, \cdots, n\}$, $\mathbf{b} = \{\mathbf{b}_i; i = 1, \cdots, n\}$, and $\boldsymbol{\theta}$ be a vector of all unknown parameters in the nonparametric transformation model (3.2). The Gibbs sampler (Geman and Geman, 1984) is used to draw observations from the joint posterior distribution of unknown quantities, $p(\boldsymbol{\theta}, \mathbf{b}|\mathbf{Y})$, for Bayesian estimation. The related full conditional distributions corresponding to the components of $(\boldsymbol{\theta}, \mathbf{b})$ for implementing the Gibbs sampler are presented as follows.

3.3.1 Full conditional distributions

(a) *Full conditional distribution of* $\boldsymbol{\beta}_j$

Let $\boldsymbol{\theta}_{-\boldsymbol{\beta}_j}$ be the subvector of $\boldsymbol{\theta}$ excluding $\boldsymbol{\beta}_j$, and n_j be the number of observations at the j-th measurement. The full conditional distribution of $\boldsymbol{\beta}_j$ is given by

$$
\begin{aligned}
& p(\boldsymbol{\beta}_j|\mathbf{Y}, \mathbf{b}, \boldsymbol{\theta}_{-\boldsymbol{\beta}_j}) \\
\propto & \prod_{i=1}^{n_j} \Bigg[\exp\Big\{ -\frac{1}{2}\Big(\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) - g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i)\Big)^2 \Big\} \\
& \times \sum_{k=1}^{K_j} \beta_{jk} B'_{jk}(y_{ij}) \Bigg] \exp\Big\{ -\frac{1}{2\tau_j^2} \boldsymbol{\beta}_j^T \mathbf{M} \boldsymbol{\beta}_j \Big\} \\
& \times I(\beta_{j1} < \cdots < \beta_{jK_j}, \mathbf{Q}_j \boldsymbol{\beta}_j = 0). \qquad (3.11)
\end{aligned}
$$

In some circumstances, the same transformation $f = \sum_{k=1}^{K} \beta_k B_k(\cdot)$ is applied to each y_{ij} for $j = 1, \cdots, p_i$, resulting in common transformation parameters τ^2, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)^T$. A typical example is transforming the repeated measures of subject i at different time points $j = 1, \cdots, p_i$. Identical transformations should be applied to all y_{ij} because the meanings of the measures are the same. In this case, the full conditional distribution of $\boldsymbol{\beta}$ is obtained by pooling all y_{ij} together as follows:

$$
\begin{aligned}
& p(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{b}, \boldsymbol{\theta}_{-\boldsymbol{\beta}}) \\
\propto & \prod_{i=1}^{n} \prod_{j=1}^{p_i} \Big[\exp\Big\{ -\frac{1}{2} \Big(\sum_{k=1}^{K} \beta_k B_k(y_{ij}) - g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) \Big)^2 \Big\} \\
& \times \sum_{k=1}^{K} \beta_k B_k'(y_{ij}) \Big] \exp\Big\{ -\frac{1}{2\tau^2} \boldsymbol{\beta}^T \mathbf{M} \boldsymbol{\beta} \Big\} \\
& \times I(\beta_1 < \cdots < \beta_K, \mathbf{Q}\boldsymbol{\beta} = 0),
\end{aligned} \tag{3.12}
$$

where $\mathbf{Q} = \mathbf{1}^T(\mathbf{B}_1^T, \ldots, \mathbf{B}_p^T)^T$, $p = \max_i(p_i)$, and $\boldsymbol{\theta}_{-\boldsymbol{\beta}}$ is similarly defined as $\boldsymbol{\theta}_{-\boldsymbol{\beta}_j}$ in (3.11). These full conditional distributions are defined on a truncated and degenerated space. The algorithm used to draw samples from these distributions are given in the next two subsections.

(b) *Full conditional distributions of* $\mathbf{b}_i$, $\boldsymbol{\gamma}$, $\boldsymbol{\Phi}$, *and* τ_j^2

$$
\begin{aligned}
& p(\mathbf{b}_i|\mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\Phi}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \\
\propto & \exp\Big\{ -\sum_{j=1}^{p_i} \frac{1}{2} \Big(\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) - g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) \Big)^2 \Big\} \\
& \times \exp\Big\{ -\frac{1}{2} \mathbf{b}_i^T \boldsymbol{\Phi}^{-1} \mathbf{b}_i \Big\},
\end{aligned} \tag{3.13}
$$

$$
\begin{aligned}
& p(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{b}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) \\
\propto\ & \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{p_i} \Big(\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(y_{ij}) - g(\mathbf{x}_{ij}, \boldsymbol{\gamma}, \mathbf{b}_i) \Big)^2 \Big\} \\
& \times \exp\Big\{ -\frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) \Big\}, \qquad (3.14)
\end{aligned}
$$

$$
p(\boldsymbol{\Phi}^{-1}|\mathbf{b}) \stackrel{D}{=} Wishart\Big(\sum_{i=1}^{n} \mathbf{b}_i \mathbf{b}_i^T + \mathbf{R}_0,\ n + r_0 \Big), \qquad (3.15)
$$

$$
p(\tau_j^{-2}|\boldsymbol{\beta}_j) \stackrel{D}{=} Gamma\Big[\alpha_{j1} + \frac{K_j - d}{2},\ \alpha_{j2} + \frac{1}{2} \boldsymbol{\beta}_j^T \mathbf{M} \boldsymbol{\beta}_j \Big]. \qquad (3.16)
$$

For $\mathbf{b}_i$ and $\boldsymbol{\gamma}$, the associated full conditional distributions are non-standard. Given $\boldsymbol{\beta}_j$, MCMC sampling schemes, such as Metropolis-Hastings (MH) (Metropolis et al., 1953; Hastings, 1970) algorithm, can be implemented similarly to models without transformation.

3.3.2 The Random-Ray algorithm

For notational simplicity, we temporarily suppress the subscript j in this and in the next subsections. In the implementation of the Gibbs sampler for the Bayesian estimation, simulating $\boldsymbol{\beta}$ from (3.11) or (3.12) is challenging because the corresponding full conditional distribution is irregular, and the parameter space is truncated and degenerated. Let $S = \{\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K); \beta_1 < \ldots < \beta_K, \mathbf{Q}\boldsymbol{\beta} = 0\}$ be the constrained parameter space of $\boldsymbol{\beta}$. It is difficult to apply the conventional MH algorithm with a multivariate normal proposal distribution in sampling $\boldsymbol{\beta}$ in S because it is very hard to generate a candidate in an irregular space. Moreover, the $\boldsymbol{\beta}$-dependent normalizing constant in (3.11) or (3.12) cannot be calculated analytically. Even for a $\boldsymbol{\beta}$ with moderate dimension, it is very time consuming to obtain an accurate approximation. Chen and Schmeiser (1993) proposed the Hit-and-Run (HR) algorithm,

which is suitable for sampling observations in a constrained parameter space. Based on a current state $\boldsymbol{\beta}^{(t)}$, the HR algorithm randomly generates a direction vector $\mathbf{e}$ and a scalar r, providing a candidate $\boldsymbol{\beta}^{(t)} + r\mathbf{e}$. Consequently, sampling in a multivariate constrained space is converted to sampling in a univariate truncated space of r. Liu et al. (2000) proposed a Random-Ray (RR) algorithm, which combines the ideas of the HR algorithm and the Multiple-Try Metropolis (MTM) algorithm (Liu et al., 2000). The MTM algorithm can increase the acceptance rate without narrowing down the Metropolis jumps but at the expense of generating multiple candidates from the proposal distribution in each iteration.

The RR algorithm in generating $\boldsymbol{\beta}$ from the full conditional distribution $p(\boldsymbol{\beta}|\cdot)$ in (3.11) or (3.12) at the current state $\boldsymbol{\beta}^{(t)}$ is described as follows:

(a) Randomly generate a directional unit vector $\mathbf{e}$ (see the following Section 3.3.3).

(b) Draw $\mathbf{w}_1, \cdots, \mathbf{w}_m$ along the direction $\mathbf{e}$ by generating scalars r_{1k} from a univariate proposal distribution $T_e(\boldsymbol{\beta}^{(t)}, \mathbf{w})$, and calculating $\mathbf{w}_k = \boldsymbol{\beta}^{(t)} + r_{1k}\mathbf{e}, \quad k = 1, \cdots, m.$

(c) Choose $\mathbf{w}^*$ from candidates $\mathbf{w}_1, \cdots, \mathbf{w}_m$ with the following probabilities:

$$p(\mathbf{w}^* = \mathbf{w}_k) \propto p(\mathbf{w}_k|\cdot)T_e(\mathbf{w}_k, \boldsymbol{\beta}^{(t)}), \qquad k = 1, \cdots, m.$$

(d) Draw $\mathbf{v}_1, \cdots, \mathbf{v}_{m-1}$ along the direction $\mathbf{e}$ by generating r_{2k} from $T_e(\mathbf{w}^*, \mathbf{v})$, and calculating $\mathbf{v}_k = \mathbf{w}^* + r_{2k}\mathbf{e}$, for $k = 1, \ldots, m-1$. Let $\mathbf{v}_m = \boldsymbol{\beta}^{(t)}$, and r_{2m} be the scalar such that

$\boldsymbol{\beta}^{(t)} = \mathbf{w}^* + r_{2m}\mathbf{e}$. Compute the generalized Metropolis ratio

$$R^{(t)} = \min\left\{1, \frac{\sum_{k=1}^{m} p(\mathbf{w}_k|\cdot)T_e(\mathbf{w}_k, \boldsymbol{\beta}^{(t)})}{\sum_{k=1}^{m} p(\mathbf{v}_k|\cdot)T_e(\mathbf{v}_k, \mathbf{w}^*)}\right\}. \tag{3.17}$$

(e) Generate $v_1 \sim Uniform(0,1)$, and let $\boldsymbol{\beta}^{(t+1)} = \mathbf{w}^*$ if $v_1 \leq R^{(t)}$; $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ if $v_1 > R^{(t)}$.

In steps (a)-(e), m is the number of multiple candidates. For $l = 1, 2$, $k = 1, \ldots, m$, the truncated normal distribution $N(0, \sigma_r^2)I(l_{lk}^{(t)}, u_{lk}^{(t)})$ is used as the proposal distribution of r_{lk}, where the truncation range $(l_{lk}^{(t)}, u_{lk}^{(t)})$ is determined by $\mathbf{e}$ (see Section 3.3.3). Therefore, all $(l_{lk}^{(t)}, u_{lk}^{(t)})$ have to be updated in each MCMC iteration. In most situations studied in this chapter, σ_r^2 ranging from 4 to 8 and m ranging from 5 to 10 produce acceptance rates between 0.4 and 0.6. In our semiparametric transformation model with Bayesian P-splines, generating $\mathbf{e}$ with the RR algorithm is crucial because it affects the efficiency of the sampling scheme. A modified Random-Ray algorithm is proposed.

3.3.3 Modified Random-Ray algorithm

Let $\boldsymbol{\beta} + r\mathbf{e}$ be a candidate sample generated along the direction $\mathbf{e} = (e_1, \cdots, e_K)$ based on current state $\boldsymbol{\beta}$. To ensure that all new samples generated by the RR algorithm are in the constrained parameter space S, the following two conditions should be satisfied:

(1) $\mathbf{Qe} = 0$. Since $\mathbf{Q}\boldsymbol{\beta} = 0$, we have $\mathbf{Qe} = 0$ iff $\mathbf{Q}(\boldsymbol{\beta} + r\mathbf{e}) = 0$.

(2) $\beta_{k+1} + re_{k+1} > \beta_k + re_k$, for $k = 1, \ldots, K-1$. We can determine the range of r as follows. For $k = 1, \cdots, K-1$,

$$r \in \begin{cases} [-(\beta_{k+1}-\beta_k)/(e_{k+1}-e_k), \infty], & e_{k+1} > e_k \\ [-\infty, -(\beta_{k+1}-\beta_k)/(e_{k+1}-e_k)], & e_{k+1} < e_k. \end{cases}$$

Let $I_l = \{k|e_{k+1} > e_k\}$ and $I_u = \{k|e_{k+1} < e_k\}$. The domain of r is the intersection of these $K-1$ intervals:

$$\Big[\max\Big\{-\infty,\ \max_{k\in I_l}(-\frac{\beta_{k+1}-\beta_k}{e_{k+1}-e_k})\Big\},\ \min\Big\{\min_{k\in I_u}(-\frac{\beta_{k+1}-\beta_k}{e_{k+1}-e_k}),\ \infty\Big\}\Big]. \tag{3.18}$$

As the directional vector $\mathbf{e}$ affects the range of r, it also affects the efficiency of the algorithm. One approach (Chen and Schmeiser, 1993; Liu et al., 2000) is to generate $\tilde{e}_k \sim \text{Uniform}(-1,1)$, $k = 1,\ldots,K$, and let $e_k = \tilde{e}_k/\sqrt{\sum_k \tilde{e}_k^2}$. However, in the transformation model that we considered, this sampling scheme ignores the feature of $p(\boldsymbol{\beta}|\cdot)$, thereby leading to an inefficient exploration in the parameter space S. Hence, we modify the conventional sampling of $\mathbf{e}$ by drawing $\mathbf{e}$ from $N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}))I(\mathbf{Qe}=0)$ with

$$\begin{aligned} \boldsymbol{\Sigma}(\boldsymbol{\beta}) &= \partial^2\{-\log p(\boldsymbol{\beta}|\cdot)\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' \\ &= \mathbf{B}^T\mathbf{B} + \mathbf{M}/\tau^2 + (\mathbf{B}')^T\mathbf{D}(\boldsymbol{\beta})\mathbf{B}', \end{aligned}$$

in which

$$\begin{aligned} &\mathbf{B}' = (\mathbf{B}'_1,\ldots,\mathbf{B}'_n)^T, \\ &\mathbf{B}'_i = (B'_1(y_i),\ldots,B'_K(y_i))^T, \\ &\mathbf{D}(\boldsymbol{\beta}) = diag(1/(\boldsymbol{\beta}^T\mathbf{B}'_1)^2,\ldots,1/(\boldsymbol{\beta}^T\mathbf{B}'_n)^2). \end{aligned}$$

According to our numerical experience, this modified RR (MRR) algorithm produces better efficiency in exploring the constrained

parameter space of $\boldsymbol{\beta}$. In the MRR algorithm, generating the directional vector $\mathbf{e}$ from $N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}))I(\mathbf{Qe} = 0)$ in step (a) is efficient by using the algorithm in Rue (2004), because $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ is a band matrix. As $\boldsymbol{\beta}$ is involved in $\boldsymbol{\Sigma}(\boldsymbol{\beta})$, we implement step (a) in the MRR algorithm with the following two phases to simplify the computation of (3.17):

(a1) In the burn-in phase, we use $\boldsymbol{\Sigma} = \mathbf{B}^T\mathbf{B} + \mathbf{M}/\tau^2$ as a rough approximation of $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ in generating $\mathbf{e}$.

(a2) After the burn-in phase, we use $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ as an approximation of $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ in generating $\mathbf{e}$, where $\boldsymbol{\beta}_0$ is the mean of $\boldsymbol{\beta}$ obtained with the burn-in samples.

Steps (a1) and (a2) take into account the information obtained from the initial stage to achieve a better approximation, and avoid the calculation of $\boldsymbol{\Sigma}(\mathbf{w}_k)$ and $\boldsymbol{\Sigma}(\mathbf{v}_k)$, $k = 1, \ldots, m$ in $R^{(t)}$ (see (3.17)) at each iteration. Therefore, considerable computing time is saved.

3.4 Numerical Studies

3.4.1 A simulation study

To study the empirical performance of the proposed semiparametric transformation model, we conduct a simulation study based on the following nonlinear mixed model given in Pinheiro and Bates (2000):

$$f(y_{ij}) = \tilde{y}_{ij} = (\gamma_1 + b_{i1}) + (b_{i2} + \gamma_2)x_{ij}^3 \exp\{-(b_{i3} + \gamma_3)x_{ij}\} + \epsilon_{ij}, \tag{3.19}$$

where $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3})^T \sim N(\mathbf{0}, \boldsymbol{\Phi})$, for $i = 1, \ldots, 100$ are subject specific random effects, $\epsilon_{ij} \sim N(0, \sigma^2)$, for $i = 1, \ldots, 100$

and $j = 1, \ldots, 14$, and $\mathbf{b}_i$ and ϵ_{ij} are independent. This model involves the linear random effects $\{b_{i1}, b_{i2}\}$, nonlinear random effect b_{i3}, and fixed effects $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T$. Three typical non-normal distributions of y_{ij}, including highly skewed, U-shaped, and bimodal distributions, are considered in our simulation study. To produce non-normal data y_{ij}, we simulate $\tilde{y}_{ij}$ based on equation (3.19), and then take the inverse transformation f^{-1} for $\tilde{y}_{ij}$ such that $y_{ij} = f^{-1}(\tilde{y}_{ij})$. The three settings of non-normality are as follows: (1) $f^{-1}(\tilde{y}) = \exp(\frac{\tilde{y}-55}{10})/(1 + \exp(\frac{\tilde{y}-55}{10}))$, resulting in a highly skewed y_{ij}; (2) $f^{-1}(\tilde{y}) = 15\text{argtan}(\log(\tilde{y}/30))$, resulting in a non-symmetrically U-shaped y_{ij}; and (3) $f^{-1}(\tilde{y}) = \tilde{y}/7 + \sin((\tilde{y} - 20)/10)$, resulting in a bimodal y_{ij}. The histograms of the simulated data under these settings are depicted in the upper part of Figure 3.1. Let $x_{i1} = -1$, $x_{ij} = 2(j - 2)$ for $j = 2, \ldots, 8$, and $x_{ij} = 12 + 3(j - 8)$ for $j = 9, \ldots, 14$ (see Pinheiro and Bates, 2000). The true population values of the unknown parameters are taken as $\boldsymbol{\gamma}^T = (23.810, 4.762, -0.55)$ and $\boldsymbol{\Phi} = (1.306, -0.192, -0.082; -0.192, 0.580, 0.063; -0.082, 0.063, 0.010\ )$, and $\sigma^2 = 1$ is fixed for identification purposes.

In each of the above settings, identical transformation $f(\cdot)$ is applied to y_{ij} for $j = 1, \cdots, 14$. The P-splines with $K = 20$ is used to approximate the transformation function $f(\cdot)$. When conducting the Bayesian analysis, the prior inputs in (3.10) were taken as: $\boldsymbol{\gamma}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = 10^{-4}\mathbf{I}$, $r_0 = 5$, and $\mathbf{R}_0 = \boldsymbol{\Phi}$. The estimates of the unknown transformation function and unknown parameters were obtained by the Bayesian P-splines approach based on 100 replications. In each replication, 5,000 burn-in samples were discarded, and 45,000 samples were acquired as posterior samples. The estimated transformation functions, together with the 95% pointwise credible intervals, are depicted in the lower part of Fig-

ure 3.19, which indicates that the proposed approach accurately estimates the unknown transformation functions. The bias (BIAS) and the root mean square (RMS) between the true values of the parameters and their estimates are reported in Tables 3.1 and 3.2, respectively. The obtained results show that the proposed method performs well for the response variables with the above-mentioned highly non-normal distributions. Note that although the P-splines can estimate the shape of $f(\cdot)$ under constraint $\mathbf{Q}\boldsymbol{\beta} = 0$, the estimated $\sum_{k=1}^{K} \beta_k B_k(y_{ij})$ deviates from $f(y_{ij})$ roughly with a constant shift. Thus, the estimated γ_1, which represents the overall mean, is meaningless in the transformation model and is not presented in Table 3.1.

To study the sensitivity of the Bayesian results to the prior inputs, the simulated data sets were reanalyzed by using two different prior settings: (I) $\boldsymbol{\gamma}_0 = 2\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $r_0 = 5$, $\mathbf{R}_0 = 2\boldsymbol{\Phi}$, $\alpha_1 = 1$, and $\alpha_2 = 0.05$; and (II) $\boldsymbol{\gamma}_0 = 0.5\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $r_0 = 5$, $\mathbf{R}_0 = 0.5\boldsymbol{\Phi}$, $\alpha_1 = 0.001$, and $\alpha_2 = 0.001$. The estimated unknown parameters and nonparametric transformation functions are close to those reported in Table 3.1 and 3.2, and Figure 3.1, indicating that the Bayesian results obtained by our method are not very sensitive to the prior inputs under the given sample sizes and model settings. It took 30 minutes to produce the Bayesian estimation for one replication using a PC with Core 2 6300@1.86 GHz and 1G RAM.

It should be noted that the non-Bayesian transformation models and their associative statistical methods (see references in Section 3.1), and the Bayesian transformation model considered in Mallick and Walker (2003) cannot be applied to analyze the nonlinear mixed model defined in equation (3.19). To compare our method with other possible existing methods, the simulated data were reanalyzed by the following conventional transformations. The first one

is the Box-Cox transformation, which is the most popular parametric transformation method. We considered the Box-Cox transformation with the index parameter λ in $\{-1, -1/2, 0, 1/2, 1\}$, which are commonly used and associated with the reciprocal, reciprocal of square root, logarithm, square root, and identity transformation, respectively. The identity transformation with $\lambda = 1$ is equivalent to ignoring the non-normality and simply fitting the data to the nonlinear model in the right hand side of (3.1). The second one is to discretize the non-normal response variable y_{ij} into a categorical variable via threshold specification (see Cowles, 1996). This is a commonly used ad hoc method in dealing with highly non-normal data in practical applications. The estimated unknown parameters produced by the above-mentioned two conventional transformation methods are presented in Tables 3.1 and 3.2 for comparison. The results show that the performance of our proposed method with the Bayesian P-splines is much better than those associated with the Box-Cox transformation and the discretization method.

3.4.2 Application: A study on the intervention treatment of preventing polydrug use

The proposed semiparametric transformation model with Bayesian P-splines is applied to a real study concerned on Proposition 36 (Prop36, Evans et al., 2009) initiated by California voters. The proposition directs drug offenders to a community-based drug treatment to reduce drug abuse using proven and effective treatment strategies. Objectives in this study include examining the reason why court mandated offenders dropout of the drug treatment, and comparing their characteristics, treatment experiences, perceptions, and outcomes with treatment completers (see Evans et al., 2009). The entire data set was obtained based on a number of

self-reported and administrative questionnaire items on drug treatment dropouts, drug-related crime histories, and drug use histories. Also, information about services and tests received was collected from the participants at intake, three-month, and 12-month follow-up interviews. In our analysis, we employ a latent variable model to investigate how the drug use severity and convicted crime history affect the retention in the drug treatment and how this retention will affect the behavior of the future drug use. Variables related to the following items are included in the analysis: "drug problems in past 30 days at intake (Drgplm30, y_1)," "drug use in past 30 days at intake (Drgday30, y_2)," "number of kinds of drugs used in past 30 days at intake (DrgN30, y_3)," "number of incarcerations in lifetime at intake (Incar, y_4)," "number of arrests in lifetime at intake (ArrN, y_5)," "age of first arrest (y_6)," "treatment retention (y_7)," and "primary drug use in past 30 days at 12 month interview (M12drg30, y_8)." The first three observed variables (y_1, y_2, y_3) reveal the characteristics of patients' drug severity. These variables were grouped into a latent factor "drug severity, b_1." The next three observed variables (y_4, y_5, y_6), which reflect the patients' crime history, were grouped into a latent factor "crime history, b_2." A confirmatory factor analysis (CFA) model was proposed to group the observed variables into latent factors. Three observed variables, including "services received in past 3 months at a 3 month interview (Sericem, x_1)," "number of drug tests in past 3 months at the 3 month interview (DrugtestTX, x_2)," and "number of drug tests by criminal justice in past 3 months at the 3 month interview (DrugtestCJ, x_3)," were considered as covariates because they were expected to affect treatment retention according to prior medical knowledge. The sample size of the data set was 1,028, and all variables were treated as continuous. When looking

at the histograms of the variables (see Figure 3.2), we found that most distributions were far away from normal. Specifically, y_1 and y_2 were non-symmetrically U-shaped, y_4 and y_7 were bimodal, and the rest were highly right skewed. For these highly non-normal variables, the conventional parametric and discretizing transformations were not expected to work well. Therefore, we applied the proposed methodology to analyze the data. As the patterns of non-normality and the meaning of the variables are very different for y_{ij} at distinct j, component specific transformations $f_j(\cdot)$ are applied to y_{ij} for each j.

In formulating an appropriate model to analyze the data, "treatment retention, y_7" plays a special role because both the covariates (x_1, x_2, x_3) and latent factors (b_1, b_2) influence y_7, and y_7 itself further influences the drug use at the 12-month interview (y_8). Therefore, y_7 is not only in the sets of the response variables but also in the sets of the covariates in the model. For $i = 1, \cdots, 1028$,

$$\mathbf{f}(\mathbf{y}_i) = \tilde{\mathbf{y}}_i = \boldsymbol{\gamma}_x \mathbf{x}_i + \boldsymbol{\gamma}_b \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{3.20}$$

where $\tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, \cdots, \tilde{y}_{i8})^T$, $\mathbf{f}(\mathbf{y}_i) = (f_1(y_{i1}), \cdots, f_8(y_{i8}))^T$, $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, y_{i7})^T$, $\boldsymbol{\gamma}_x$ and $\boldsymbol{\gamma}_b$ are the matrices of unknown regression coefficients corresponding to fixed effects $\mathbf{x}_i$ and random effects (latent factors) $\mathbf{b}_i$, respectively; $\mathbf{b}_i = (b_{i1}, b_{i2})^T \sim N(\mathbf{0}, \boldsymbol{\Phi})$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{i8})^T \sim N(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix for identifying the model. Based on the prior knowledge of experts in formulating the relations among response variables and latent factors (see the path diagram in Figure 3.3), the structures of $\boldsymbol{\gamma}_x$

and $\boldsymbol{\gamma}_b$ in (3.20) are prespecified as follows:

$$\boldsymbol{\gamma}_x = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \gamma_{x1} & \gamma_{x2} & \gamma_{x3} & 0 \\ 0 & 0 & 0 & \gamma_{x4} \end{bmatrix},$$

$$\boldsymbol{\gamma}_b = \begin{bmatrix} 1 & \gamma_{b1} & \gamma_{b2} & 0 & 0 & 0 & \gamma_{b5} & 0 \\ 0 & 0 & 0 & 1 & \gamma_{b3} & \gamma_{b4} & \gamma_{b6} & 0 \end{bmatrix}^T,$$

where the $\mathbf{0}$'s in $\boldsymbol{\gamma}_x$ indicate 6×1 vectors of zeros. All the zeros and ones in $\boldsymbol{\gamma}_x$ and $\boldsymbol{\gamma}_b$ were fixed for defining an identified CFA model (see Shi and Lee, 2000) as given in Figure 3.3. Hence, the unknown parameters are $\boldsymbol{\gamma} = (\gamma_{x1}, \gamma_{x2}, \gamma_{x3}, \gamma_{x4}, \gamma_{b1}, \gamma_{b2}, \gamma_{b3}, \gamma_{b4}, \gamma_{b5}, \gamma_{b6})^T$, and $\{\phi_{11}, \phi_{12}, \phi_{22}\}$ in $\boldsymbol{\Phi}$. We encounter the following problem in applying the nonparametric transformation to each y_{ij}. Although the y_{ij}s vary within a wide range, most of them take integer values (e.g., number of days or number of arrests), which results in ties. As the transformations are one to one, it is impossible to transform the variables with many ties to normal. To solve this problem, small random noises generated from N(0,0.01) were added to break the ties, which still maintained the orders of the patients for every variables. In this analysis, we take $K_4 = 40$, $K_5 = 60$, and $K_j = 30$ for the rest j as the number of splines for approximating the unknown transformation functions f_j. The prior distributions of $\boldsymbol{\gamma}$ and $\boldsymbol{\Phi}$ were taken as those in (3.10) with the hyperparameter inputs: $\boldsymbol{\gamma}_0 = \hat{\boldsymbol{\gamma}}$, $r_0 = 6$, and $\mathbf{R}_0 = (r_0 - q - 1)\hat{\boldsymbol{\Phi}}$, where q is the dimension of $\mathbf{b}_i$, and $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\Phi}}$ are the ML estimates obtained via the discretization method.

After checking the convergence, we found that the MCMC algorithm converged within 30,000 iterations. To be conservative, 50,000 generated samples were used to obtain the Bayesian results

after discarding 50,000 burn-in iterations. The estimates of factor loadings, regression coefficients, and their standard error estimates are reported in Figure 3.3. The estimates of the pointwise posterior means of the unknown transformation functions, together with the 5%- and 95%-pointwise quantiles, are depicted in Figure 3.4. For most curves, the estimated credible intervals formed by the pointwise quantiles are narrow, indicating good estimates of the unknown transformation functions with the Bayesian P-splines approach. To investigate the sensitivity of the Bayesian results to the prior inputs, the above analysis was repeated with some perturbations of the current prior input. In particular, two different choices of $\alpha_\tau = 1$, $\beta_\tau = 0.05$, and $\alpha_\tau = 0.001$, $\beta_\tau = 0.001$ were used. We obtained close Bayesian estimates of unknown parameters and similar estimated curves of unknown smooth functions. The program was written in R. The computational time for completing the above analysis was about 230 min using a PC with Intel Core2 1.86 GHz CPU 1G RAM.

The interpretation of the results are given as follows. First, all the factor loadings are significant, indicating strong associations between latent factors and the corresponding observed variables. The loading on "age of first arrest" is negative, indicating that patients who were involved in crime earlier in their lives tend to have more serious crime history, which may be reflected by large numbers of incarcerations and arrests. Second, both "drug severity" and "crime history" have negative impacts on "retention," indicating that patients who have more serious "drug severity" and worse "crime history" tend to have less treatment retention. Therefore, some enforcement actions might be necessary for those patients. For example, more attention should be given to patients who were young criminals or who were serious drug users. Third, the covari-

ates that measure services and tests in the three-month interview give a significant indication of retention measurement. x_1 is the mean value of a questionnaire about the treatment with 42 questions at the three-month interview, which reflects the attitude of the patients towards the treatment after the first three months. x_2 and x_3 are numbers of tests the patients received in the first three months, which reflect how much attention is given to monitoring the treatment process. Patients who received more attention were likely to stay longer in the treatment. Furthermore, the results from the three-month interview can serve as an indicator to monitor and improve the treatment process. To decrease the number of dropouts from the treatment, special attention should be given to the patients who filled the questionnaire with low values and received few drug tests. Finally, "retention" has negative impact on drug use in the 12-month follow-up interview, which indicates that longer treatment retention leads to less drug use at 12 months. This finding indicates a positive effect of the Proposition 36 treatment program in reducing drug abuse.

3.5 Conclusion

In this chapter, a semiparametric transformation model is proposed to analyze data involving highly non-normal variables. Different from traditional transformation methods such as the Box-Cox transformation, the current model formulates the unknown transformation function through Bayesian P-splines. To solve the difficulties encountered in the development of the proposed methodology, a modified constrained Bayesian P-splines approach incorporated with powerful MCMC techniques is employed. A simulation study demonstrates that the proposed model and methodology

perform satisfactorily with several commonly encountered highly non-normal distributions. The novel model and methodology are applied to analyze a data set related to polydrug use intervention, in which the observed variables are extremely non-normal, such as U-shaped and highly skewed. Some interesting findings are obtained.

Table 3.1: The Bayesian estimates of fix effect coefficients in 100 replications

	Par		P-spline	Dis	λ in Box-Cox transfomation				
					-1	-0.5	0	0.5	1
Skewed	γ_2	Bias	-0.055	-4.279	-2.883	-3.612	-4.162	-4.138	-3.989
		Rms	0.230	4.282	2.886	3.613	4.162	4.138	3.989
	γ_3	Bias	-0.002	-0.091	0.684	0.420	0.069	-0.010	0.013
		Rms	0.010	0.100	0.685	0.423	0.072	0.015	0.016
U-shaped	γ_2	Bias	-0.056	-2.171	-1.673	-2.790	-4.003	-4.140	-4.134
		Rms	0.239	2.183	1.702	2.793	4.004	4.140	4.134
	γ_3	Bias	-0.002	0.004	0.797	0.526	0.056	-0.059	0.070
		Rms	0.010	0.011	0.798	0.527	0.063	0.060	0.071
Bimodal	γ_2	Bias	-0.090	-1.681	-1.547	-2.489	-3.042	-3.482	-3.422
		Rms	0.249	1.704	2.463	2.495	3.044	3.483	3.422
	γ_3	Bias	-0.003	0.022	0.628	0.494	0.282	0.076	-0.006
		Rms	0.010	0.032	0.657	0.496	0.285	0.078	0.013

Note 1: "Dis" denotes "Discretize".

Note 2: Under transformation, the estimated $\tilde{y}_{ij}$ is close the $\tilde{y}_{ij}$ generated in the simulation with a constant shift. Therefore, overall mean γ_1 is meaningless and is not reported here.

Table 3.2: The Bayesian estimates of diagonal elements of covariance matrix of random effects in 100 replications

	Par		P-spline	Dis	λ in Box-Cox transfomation				
					-1	-0.5	0	0.5	1
Skewed	σ_{11}	Bias	-0.0073	-0.559	0.973	1.348	2.250	2.251	0.988
		Rms	0.191	0.666	0.977	1.352	2.250	2.252	0.989
	σ_{22}	Bias	-0.0046	-0.384	1.845	0.796	0.924	0.954	-0.456
		Rms	0.077	0.402	1.870	0.799	0.924	0.954	0.456
	σ_{33}	Bias	0.00006	-0.010	0.010	-0.005	-0.010	-0.010	0.007
		Rms	0.002	0.010	0.011	0.007	0.010	0.010	0.007
U-shaped	σ_{11}	Bias	-0.010	-0.090	-1.172	-1.121	-0.963	1.031	-1.132
		Rms	0.197	0.204	1.172	1.122	0.964	1.031	1.132
	σ_{22}	Bias	0.002	-0.205	2.585	0.213	-0.520	-0.537	-0.547
		Rms	0.099	0.212	2.900	0.341	0.520	0.537	0.547
	σ_{33}	Bias	0.0001	-0.002	0.028	0.006	-0.009	-0.009	-0.008
		Rms	0.0002	0.003	0.029	0.007	0.009	0.009	0.008
Bimodal	σ_{11}	Bias	-0.024	-0.152	-1.214	-1.223	-1.185	-1.052	-0.877
		Rms	0.198	0.222	1.216	1.223	1.185	1.053	0.879
	σ_{22}	Bias	0.0004	0.539	212.3	0.987	0.254	-0.276	-0.414
		Rms	0.097	1.507	579.1	1.039	0.279	0.282	0.415
	σ_{33}	Bias	0.0002	-0.003	0.043	0.059	0.035	0.008	0.0002
		Rms	0.002	0.010	0.051	0.062	0.009	0.009	0.002

Note: "Dis" denotes "Discretize".

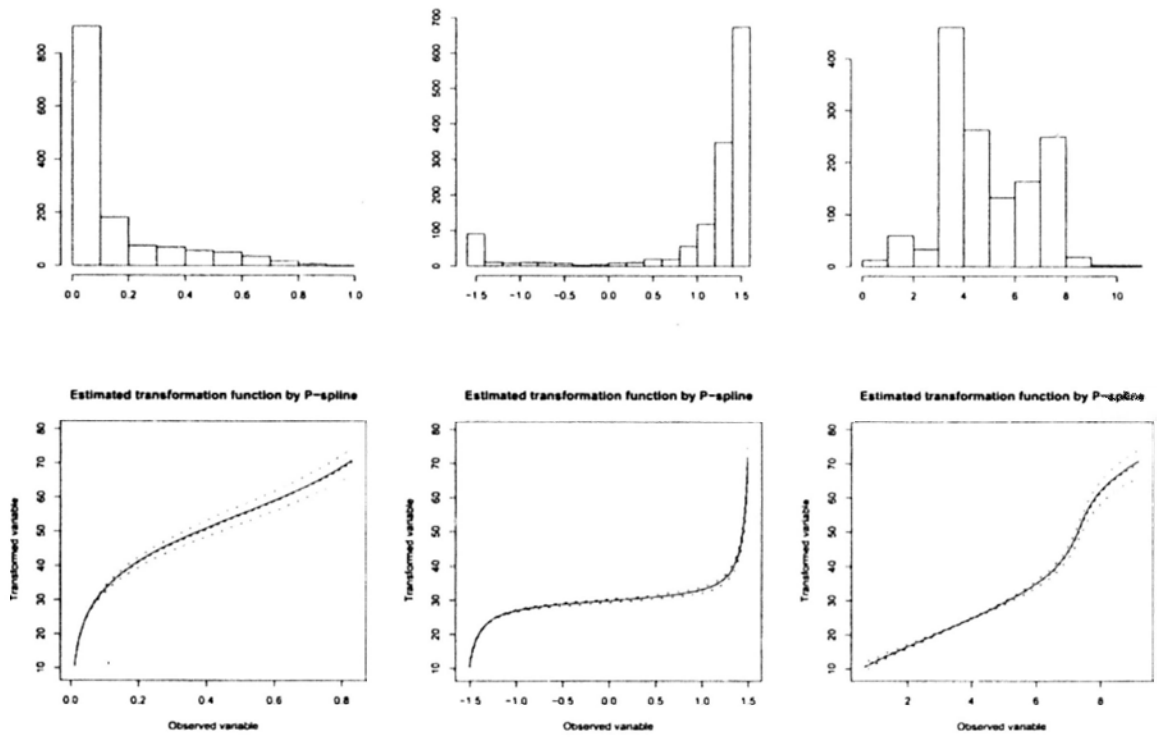

Figure 3.1: Upper part contains histograms of y_{ij} randomly selected from 100 replications in the three situations. Lower left to right are estimates of the transformation function f in three situations: Highly Skewed, U-shaped, Bimodal. The solid lines are the true transformation curves; dashed lines are estimated mean curves; dot-dash lines form the estimated point-wise 95% credible intervals.

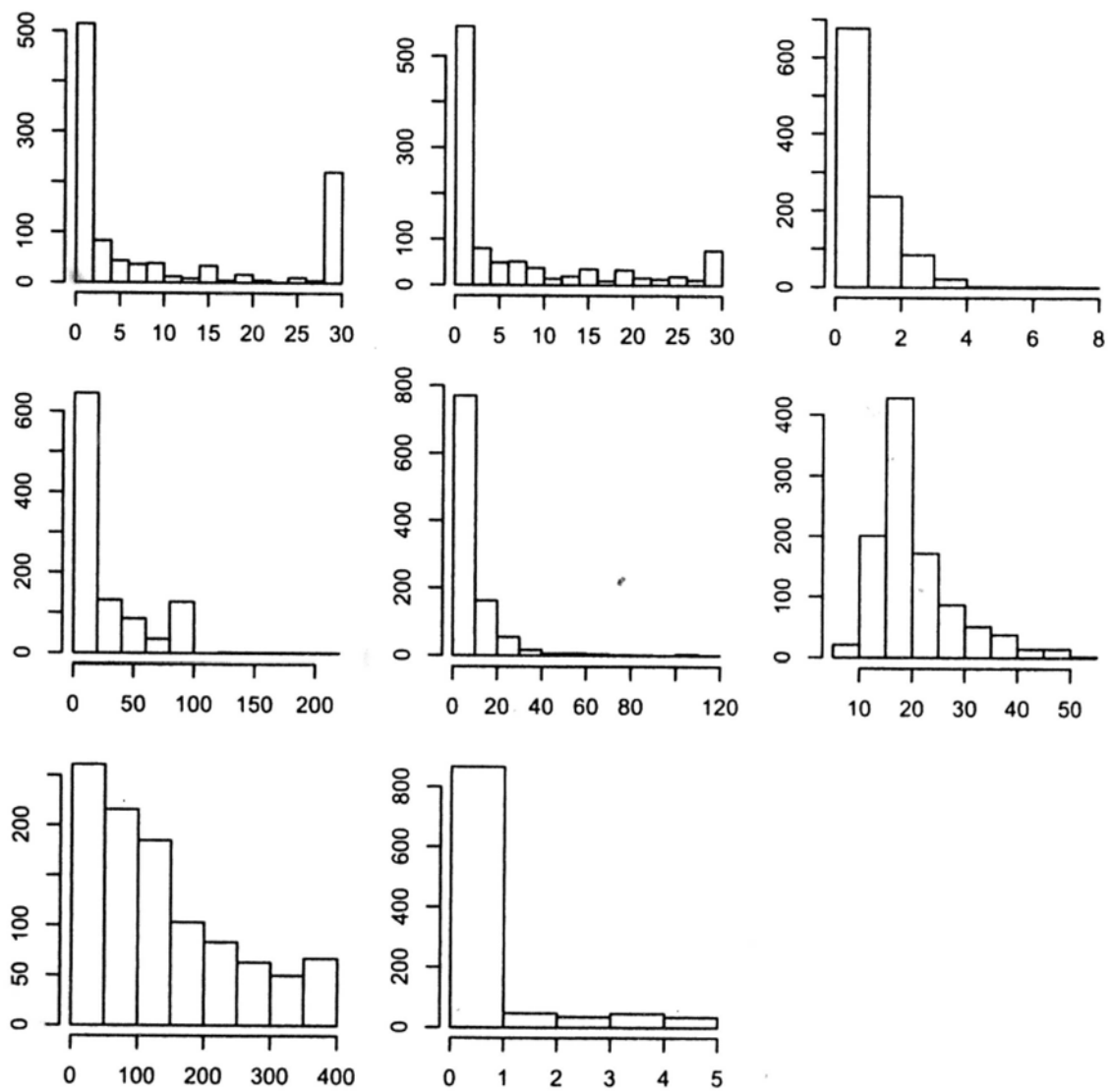

Figure 3.2: Histogram of response variables in real example. First row from left to right: y_1, y_2, and y_3; Second row from left to right: y_4, y_5, and y_6; Third row from left to right: y_7, y_8.

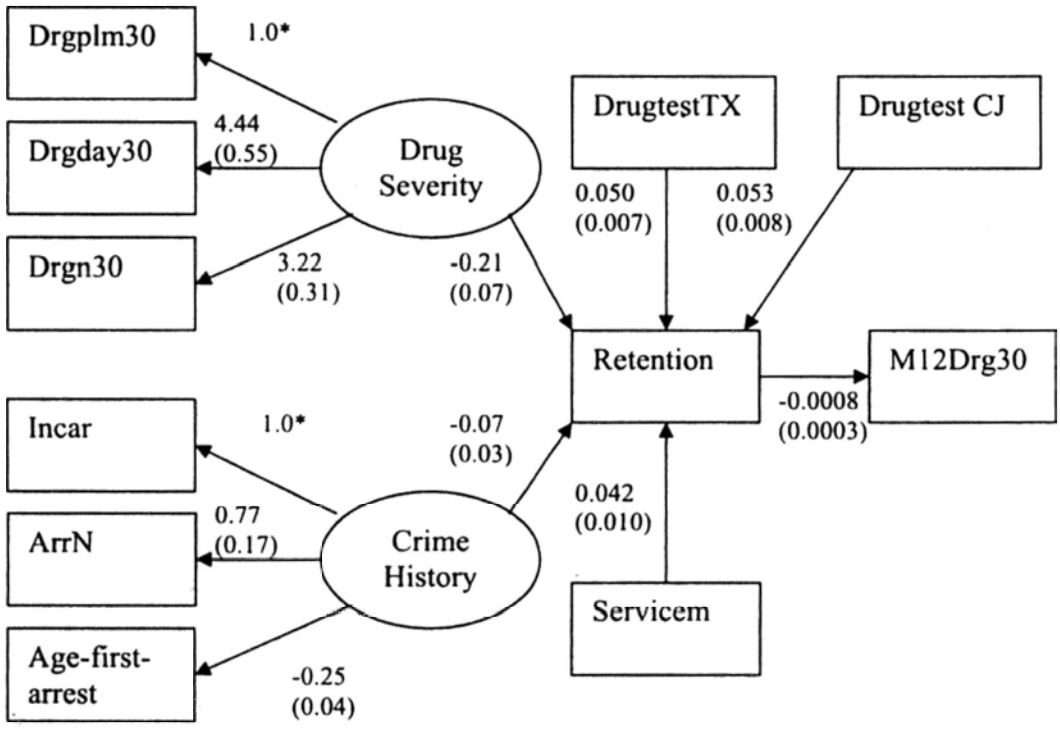

Figure 3.3: The path diagram, together with the estimated regression coefficients and their standard error estimates (in parentheses) of polydrug use data analyzed by the Bayesian P-splines transformation model.

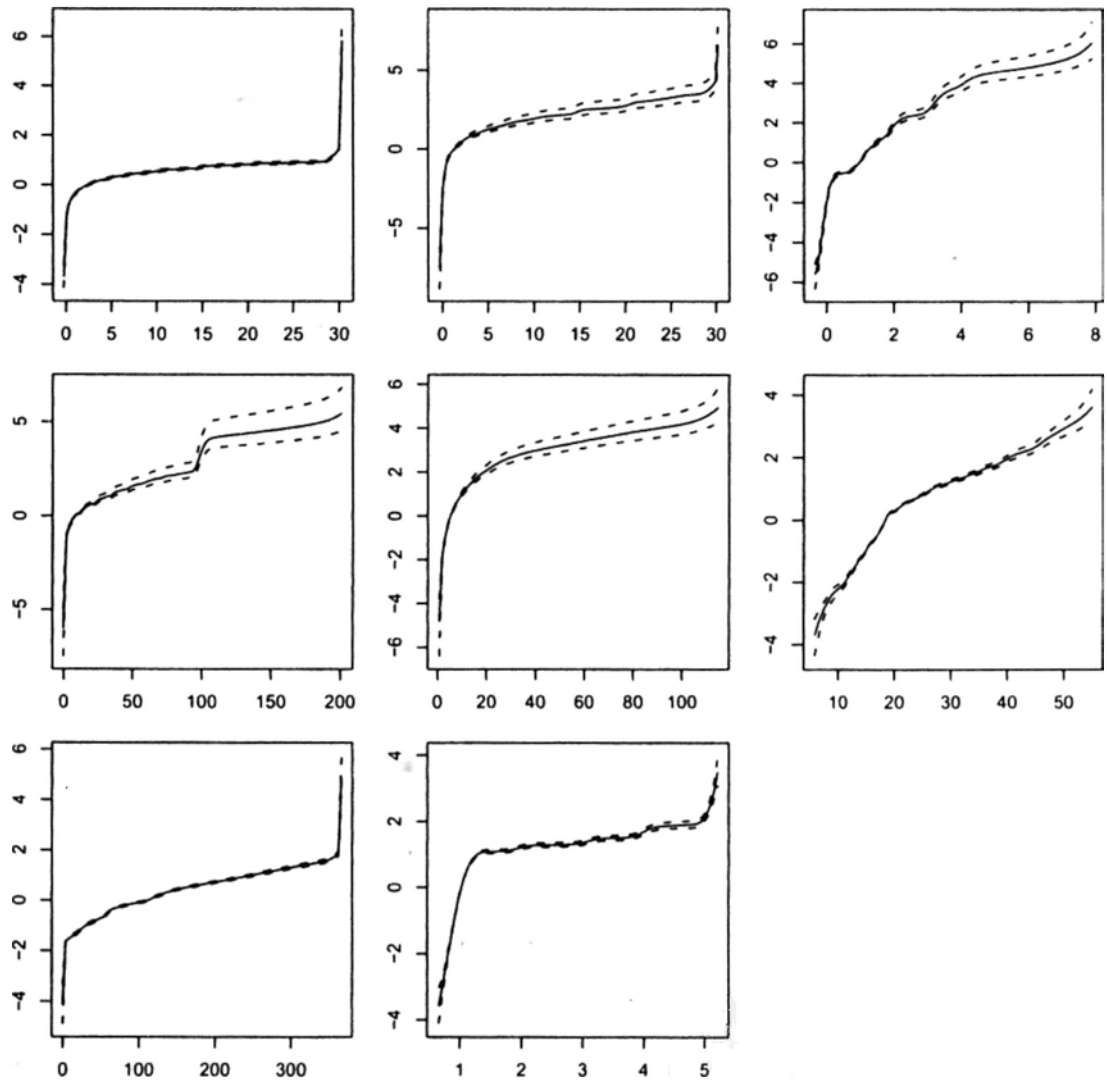

Figure 3.4: Estimates of the unknown transformation functions in the real example. The solid curves represent the estimated pointwise posterior mean curves, while the dashed curves represent the 5%- and 95%-pointwise quantiles.

Chapter 4

Transformation Varying Coefficient Models

4.1 Introduction

Varying coefficient models (see Hastie and Tibshirani, 1993; Hoover et al., 1998; Fan and Zhang, 1999; Chiang et al., 2001; Eubank et al., 2004, among others) are useful and flexible nonparametric regression models in analyzing effects of covariates dynamically according to certain modifiers, e.g., time or location. More recently, Bayesian varying coefficient models (see Biller and Fahrmeir, 2001; Lang and Brezger, 2004) have also attracted much attention, because the Bayesian approach based on MCMC sampling can be more easily applied to complicated models (Gelfand et al., 2003; Dunson et al., 2003; Hennerfeind et al., 2006; Haneuse et al., 2008; Su and Hogan, 2010). In most of varying coefficient models, distributions of error terms are usually assumed to be standard, e.g., normal distribution (see Biller and Fahrmeir, 2001; Lang and Brezger, 2004; Rodrigues and Assunção, 2008), and Gaussian time process (see Su and Hogan, 2010). Although varying coefficient models are very flexible nonparametric models, the above assumptions may be violated in real applications, and thus the validity of estimation

would be undermined. One common approach to alleviate such violation is to transform the observed variables.

In this chapter, we consider a Bayesian semivarying coefficient model with a nonparametric transformation and random effects. Inspired by the recently developed efficient Bayesian methods in the analysis of general nonparametric functions, such as DiMatteo et al. (2001), Biller and Fahrmeir (2001), Berry et al. (2002), Lang and Brezger (2004), Brezger and Steiner (2008), and Song and Lu (2010), among others, we develop our MCMC algorithm to estimate the transformation functions, the varying coefficient functions, unknown parameters, random effects, and smoothing parameters in the proposed model.

The chapter is organized as follows. Section 4.2 defines a semivarying coefficient model with a nonparametric transformation and random effects. The unknown functions, including nonparametric transformation functions and varying coefficient functions, are modeled with Bayesian P-splines. The MCMC sampling scheme and the related computational issues are given in Section 4.3. Section 4.4 presents a simulation study to demonstrate the efficiency of the proposed methodology for handling highly non-normal variables in the context of a semivarying coefficient model with random effects. Section 4.5 concludes the chapter with a discussion.

4.2 Model Description

4.2.1 General model specification

For $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip_i})^T$ be a random vector of observed variables measured in each of n independent observations. In practice, for each subject i, y_{ij} can represent one of the repeated measurements at p_i different time points. Let f be an unspecified

smooth transformation function for y_{ij}, and $\tilde{y}_{ij} = f(y_{ij})$. Similar to parametric transformations, the unknown function f is assumed to be strictly monotonic increasing. A nonparametric transformation semivarying coefficient model is defined as follows:

$$f(y_{ij}) = \tilde{y}_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\gamma}(\mathbf{u}_{ij}) + \mathbf{z}_{ij}^T \mathbf{b}_i + \mathbf{w}_{ij}^T \boldsymbol{\alpha} + \epsilon_{ij}, \tag{4.1}$$

where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijq_1})^T$ and $\mathbf{u}_{ij} = (u_{ij1}, \ldots, u_{ijq_1})^T$ are vectors of fixed covariates, $\mathbf{z}_{ij}$ is a $q_2 \times 1$ vector of covariates, and ϵ_{ij} is a random error which is independently distributed as $N(0, \sigma^2)$. It is assumed that $\mathbf{b}_i$ is independent of ϵ_{ij} and follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Phi})$. $\mathbf{w}_{ij}$ and $\boldsymbol{\alpha}$ are $q_3 \times 1$ vectors of covariates and fixed coefficients, respectively. $\boldsymbol{\gamma}(\mathbf{u}_{ij}) = (\gamma_1(u_{ij1}), \ldots, \gamma_{q_1}(u_{ijq_1}))^T$ is a vector of coefficient functions on $\mathbf{x}_{ij}$. All $\gamma_l(\cdot)$ are assumed to be unknown smooth functions, which result in varying coefficients of x_{ijl} according to u_{ijl}, $l = 1, \ldots, q_1$.

4.2.2 Modeling unknown smooth functions

For notational simplicity, we assume $q_1 = 1$. An extension to the case with $q_1 > 1$ is straightforward. Under this assumption, the model (4.1) is simplified to

$$f(y_{ij}) = \tilde{y}_{ij} = x_{ij}\gamma(u_{ij}) + \mathbf{z}_{ij}^T \mathbf{b}_i + \mathbf{w}_{ij}^T \boldsymbol{\alpha} + \epsilon_{ij}, \tag{4.2}$$

where $f(\cdot)$ and $\gamma(\cdot)$ are assumed to be unspecified smooth functions, and to be determined by data. In addition, the transformation function $f(\cdot)$ is assumed to be strictly monotonic increasing. Flexible f can eliminate serious departure of y_{ij} from normality. We adopt the different versions of Bayesian P-splines introduced in Chapter 2 and Chapter 3 to model $\gamma(\cdot)$ and $f(\cdot)$, respectively. Specifically, $f(y_{ij})$ and $\gamma(u_{ij})$ are approximated by the following

sums of B-splines $G_k(y_{ij})$ and $B_k(u_{ij})$ (De Boor, 2001):

$$f(y_{ij}) = \sum_{k=1}^{K_y} \nu_k G_k(y_{ij}), \qquad \gamma(u_{ij}) = \sum_{k=1}^{K} \beta_k B_k(u_{ij}), \tag{4.3}$$

where K_y and K are the numbers of splines determined by the numbers of knots in the supports of y_{ij} and u_{ij}. They are usually set in advance between 10 to 60 in order to provide the model with enough flexibility. Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{K_y})$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, $\mathbf{G}_{ij} = (G_1(y_{ij}), \ldots, G_{K_y}(y_{ij}))^T$, and $\mathbf{B}_{ij} = (x_{ij}B_1(u_{ij}), \ldots, x_{ij}B_K(u_{ij}))^T$. With $f(y_{ij})$ and $\gamma(u_{ij})$ approximated by (4.3), model (4.2) can be expressed as:

$$\mathbf{G}_{ij}^T\boldsymbol{\nu} = \mathbf{B}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i + \mathbf{w}_{ij}^T\boldsymbol{\alpha} + \epsilon_{ij}. \tag{4.4}$$

The complete-data likelihood corresponding to model (4.4) is

$$\begin{aligned}
&\frac{1}{(\sqrt{2\pi})^{\tilde{n}q_2}|\boldsymbol{\Phi}|^{n/2}} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n} \mathbf{b}_i^T\boldsymbol{\Phi}^{-1}\mathbf{b}_i \Big\} \times \\
&\prod_{i=1}^{n}\prod_{j=1}^{p_i}\Big[\sum_{k=1}^{K_y} \nu_k G_k'(y_{ij}) \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{1}{2\sigma^2}\big(\mathbf{G}_{ij}^T\boldsymbol{\nu} - \\
&\mathbf{B}_{ij}^T\boldsymbol{\beta} - \mathbf{z}_{ij}^T\mathbf{b}_i - \mathbf{w}_{ij}^T\boldsymbol{\alpha}\big)^2 \Big\}\Big],
\end{aligned} \tag{4.5}$$

where $G_k'(y_{ij})$ is the first derivative of $G_k(y_{ij})$, and $\sum_{k=1}^{K_y} \nu_k\, G_k'(y_{ij})$ is the Jacobian of the transformation.

4.2.3 Prior distributions

The following prior distribution is assigned to $\boldsymbol{\beta}$ to prevent over-fitting induced by large K:

$$\begin{aligned}
p(\boldsymbol{\beta}) &= \prod_{k=d+1}^{K} \Big[\frac{1}{\sqrt{2\pi}\tau} \exp\Big\{ -\frac{1}{2\tau^2}(\beta_k - \mu_k)^2 \Big\}\Big] \\
&= \Big(\frac{1}{\sqrt{2\pi}\tau}\Big)^{K-d} \exp\Big\{ -\frac{1}{2\tau^2}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta} \Big\},
\end{aligned} \tag{4.6}$$

where d is the order of the random walk, $\mu_k = \beta_{k-1}$ and $\mu_k = 2\beta_{k-1} - \beta_{k-2}$ are the first- and second-order random walks, respectively, and $\mathbf{M} = (\mathbf{D}_{d-1} \times \cdots \times \mathbf{D}_0)^T(\mathbf{D}_{d-1} \times \cdots \times \mathbf{D}_0)$, where $\mathbf{D}_l$, $l = 0, \ldots, d-1$, are defined as in (3.6).

Several constraints have to be imposed in order to identify the model. First, to ensure $f(y_{ij}) = \sum_{k=1}^{K_y} \nu_k G_k(y_{ij})$ to be strictly monotonic increasing, the constraint $\nu_1 < \cdots < \nu_{K_y}$ is needed. This constraint can be incorporated in the analysis by assigning the prior distribution of $\boldsymbol{\nu}$ as follows:

$$\left(\frac{1}{\sqrt{2\pi}\tau_y}\right)^{K_y - d_y} \exp\left\{-\frac{1}{2\tau_y^2}\boldsymbol{\nu}^T\mathbf{M}_y\boldsymbol{\nu}\right\} I(\nu_1 < \cdots < \nu_{K_y}), \quad (4.7)$$

where $I(\cdot)$ is an indicator function, and $\mathbf{M}_y$ and d_y are similarly defined as $\mathbf{M}$ and d in (4.6), respectively. Second, the coefficients in $\boldsymbol{\beta}$, $\mathbf{b}_i$, and $\boldsymbol{\alpha}$ are linear in model (4.4). As $(\boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\alpha}, \mathbf{b}_i^T, \sigma, \boldsymbol{\Phi})$ and $(c\boldsymbol{\beta}, c\boldsymbol{\nu}, c\boldsymbol{\alpha}, , c\mathbf{b}_i^T, c\sigma, c^2\boldsymbol{\Phi})$ yield the same likelihood (4.5) for an arbitrary constant c under this case, the model is not identified. Hence, σ^2 is fixed at 1.0 for identification purpose. Third, an intercept, say α_1, exists in $\mathbf{w}_{ij}^T\boldsymbol{\alpha}$. As $\sum_{k=1}^{K_y} G_k(y_{ij}) = 1$,

$$\sum_{k=1}^{K_y} \nu_k G_k(y_{ij}) - \alpha_1 = \sum_{k=1}^{K_y} (\nu_k + c) G_k(y_{ij}) - (\alpha_1 + c),$$

which results in a non-identified model because both $(\boldsymbol{\nu}, \alpha_1)$ and $(\boldsymbol{\nu} + c, \alpha_1 + c)$ yield the same likelihood (4.5). To achieve identification, we impose another constraint on $\boldsymbol{\nu}$:

$$\sum_{i=1}^{n}\sum_{j=1}^{p_i} \mathbf{G}_{ij}^T\boldsymbol{\nu} = 0.$$

Let $\mathbf{G}_i = (\mathbf{G}_{i1}, \ldots, \mathbf{G}_{ip_i})^T$, $\mathbf{G} = (\mathbf{G}_1^T, \ldots, \mathbf{G}_n^T)^T$, and $\mathbf{1} = (1, \ldots, 1)^T$. The third identification constraint is equivalent to $\mathbf{1}^T\mathbf{G}\boldsymbol{\nu} = 0$. Let

$\mathbf{Q}_y = \mathbf{1}^T\mathbf{G}$. This constraint becomes $\mathbf{Q}_y\boldsymbol{\nu} = 0$, which can be incorporated in the analysis by modifying the prior distribution of $\boldsymbol{\nu}$ in (4.7) as follows:

$$\left(\frac{1}{\sqrt{2\pi}\tau_y}\right)^{K_y-d_y} \exp\left\{-\frac{1}{2\tau_y^2}\boldsymbol{\nu}^T\mathbf{M}_y\boldsymbol{\nu}\right\} I(\boldsymbol{\nu} \in C_y), \tag{4.8}$$

where $C_y = \{\boldsymbol{\nu} | \nu_1 < \cdots < \nu_{K_y}, \mathbf{Q}_y\boldsymbol{\nu} = 0\}$.

In a full Bayesian analysis, the inverse smoothing parameters τ_y^2 and τ^2 are treated as random. According to a common practice, we assign the following prior distributions for τ_y^2 and τ^2:

$$p(\tau_y^{-2}) \stackrel{D}{=} Gamma(a_1, a_2), \quad p(\tau^{-2}) \stackrel{D}{=} Gamma(a_1, a_2), \tag{4.9}$$

where a_1 and a_2 are specified hyperparameters. In this chapter we use $a_1 = 1$ and $a_2 = 0.005$ to obtain a highly dispersed (but proper) gamma prior of τ_y^{-2} and τ^{-2}.

For the parameters involved in the righthand side of model (4.2), the following conjugate prior distributions are assigned.

$$\boldsymbol{\alpha} \stackrel{D}{=} N(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Phi}^{-1} \stackrel{D}{=} Wishart(\mathbf{R}_0, r_0), \tag{4.10}$$

where $\boldsymbol{\alpha}_0$, r_0, and positive definite matrices $\boldsymbol{\Sigma}_0$ and $\mathbf{R}_0$ are hyperparameters whose values are assumed to be given by the prior information.

4.3 Estimation of Nonparametric Transformation and Varying Coefficient Functions

Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{ip_i})^T$, $\mathbf{Y} = \{\mathbf{y}_i; i = 1, \cdots, n\}$, $\mathbf{b} = \{\mathbf{b}_i; i = 1, \cdots, n\}$, and $\boldsymbol{\theta}$ be a vector of all unknown parameters. We use the Gibbs sampler (Geman and Geman, 1984) to draw observations from the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{b}|\mathbf{Y})$ for the Bayesian estimation. The related full conditional distributions are presented as follows.

4.3.1 Full conditional distributions

(a) *Full conditional distributions of* $\boldsymbol{\nu}$ *and* $\boldsymbol{\beta}$

Let $\boldsymbol{\theta}_{-\boldsymbol{\nu}}$ be the subvector of $\boldsymbol{\theta}$ excluding $\boldsymbol{\nu}$. The full conditional distribution of $\boldsymbol{\nu}$ is given by

$$\begin{aligned} & p(\boldsymbol{\nu}|\mathbf{Y}, \mathbf{b}, \boldsymbol{\theta}_{-\boldsymbol{\nu}}) \\ \propto & \prod_{i=1}^{n}\prod_{j=1}^{p_i}\Bigg[\exp\Big\{-\frac{1}{2}\big(\mathbf{G}_{ij}^T\boldsymbol{\nu} - \mathbf{B}_{ij}^T\boldsymbol{\beta} - \mathbf{z}_{ij}^T\mathbf{b}_i - \mathbf{w}_{ij}^T\boldsymbol{\alpha}\big)^2\Big\} \times \\ & \sum_{k=1}^{K_y}\nu_k G_k'(y_{ij})\Bigg]\exp\Big\{-\frac{1}{2\tau_y^2}\boldsymbol{\nu}^T\mathbf{M}_y\boldsymbol{\nu}\Big\}I(\boldsymbol{\nu}\in C_y). \end{aligned} \tag{4.11}$$

The full conditional distribution in (4.11) is defined on a truncated and degenerated space. The algorithm used to draw samples from this distribution is analog to that described in Section 3.3.2 except the following modifications: (i) the target distribution in (3.17) is replaced by (4.11), and (ii) the covariance matrix of the proposal distribution of the directional vector in Section 3.3.3 is calculated by differentiating the logarithm of (4.11) with respect to $\boldsymbol{\nu}$ twice.

Let $\mathbf{B}_i = (\mathbf{B}_{i1}, \ldots, \mathbf{B}_{ip_i})^T$, $\mathbf{B} = (\mathbf{B}_1^T, \ldots, \mathbf{B}_n^T)^T$. The full conditional distributions of $\boldsymbol{\beta}$ is:

$$p(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{b}, \boldsymbol{\theta}_{-\boldsymbol{\beta}}) \overset{D}{=} N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_\beta), \tag{4.12}$$

where $\boldsymbol{\Sigma}_\beta = (\mathbf{B}^T\mathbf{B} + \mathbf{M}/\tau^2)^{-1}$, $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_\beta(\mathbf{B}^T\mathbf{y}^*)$, $\mathbf{y}^* = (\mathbf{y}_1^{*T}, \ldots, \mathbf{y}_n^{*T})^T$, $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{ip_i}^*)^T$, and $y_{ij}^* = \mathbf{G}_{ij}^T\boldsymbol{\nu} - \mathbf{z}_{ij}^T\mathbf{b}_i - \mathbf{w}_{ij}^T\boldsymbol{\alpha}$. Sampling observations from this multivariate normal distribution can be done efficiently using the algorithm in Rue (2004).

(b) *Full conditional distributions of smoothing parameters* τ_y^2 *and* τ^2.

Based on the prior distributions given in (4.9),

$$p(\tau_y^{-2}|\boldsymbol{\nu}) \overset{D}{=} Gamma\Big[a_1 + \tfrac{K_y - d_y}{2},\ a_2 + \tfrac{1}{2}\boldsymbol{\nu}^T\mathbf{M}_y\boldsymbol{\nu}\Big],$$
$$p(\tau^{-2}|\boldsymbol{\beta}) \overset{D}{=} Gamma\Big[a_1 + \tfrac{K-d}{2},\ a_2 + \tfrac{1}{2}\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}\Big]. \quad (4.13)$$

(c) *Full conditional distributions of* $\mathbf{b}_i$, $\boldsymbol{\alpha}$, *and* $\boldsymbol{\Phi}$

$$p(\mathbf{b}_i|\mathbf{Y}, \boldsymbol{\beta}_y, \boldsymbol{\beta}, \boldsymbol{\alpha}) \overset{D}{=} N(\mathbf{b}_i^*, \boldsymbol{\Sigma}_{bi}^*), \quad (4.14)$$

$$p(\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{b}, \boldsymbol{\beta}_y, \boldsymbol{\beta}) \overset{D}{=} N(\boldsymbol{\alpha}^*, \boldsymbol{\Sigma}_{\alpha}^*), \quad (4.15)$$

$$p(\boldsymbol{\Phi}^{-1}|\mathbf{b}) \overset{D}{=} Wishart\Big(\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T + \mathbf{R}_0,\ n + r_0\Big), \quad (4.16)$$

where $\boldsymbol{\Sigma}_{bi}^* = (\mathbf{Z}_i^T\mathbf{Z}_i + \boldsymbol{\Phi}^{-1})^{-1}$, $\mathbf{b}_i^* = \boldsymbol{\Sigma}_{bi}^*(\mathbf{Z}_i^T\mathbf{y}_{bi}^*)$, $\boldsymbol{\Sigma}_{\alpha}^* = (\mathbf{W}\mathbf{W}^T + \boldsymbol{\Sigma}_0^{-1})^{-1}$, $\boldsymbol{\alpha}^* = \boldsymbol{\Sigma}_{\alpha}^*(\mathbf{W}^T\mathbf{y}_{\alpha}^* + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\alpha}_0)$, $\mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{ip_i})^T$, $\mathbf{W} = (w_{11}, \ldots, w_{1p_1}, \ldots, w_{n1}, \ldots, w_{np_n})^T$, $\mathbf{y}_{bi}^* = (y_{bi1}^*, \ldots, y_{bip_i}^*)^T$, $\mathbf{y}_{\alpha}^* = (y_{\alpha 11}^*, \ldots, y_{\alpha 1p_1}^*, \ldots, y_{\alpha n1}^*, \ldots, y_{\alpha np_n}^*)^T$, $y_{bij}^* = \mathbf{G}_{ij}^T\boldsymbol{\nu} - \mathbf{B}_{ij}^T\boldsymbol{\beta} - \mathbf{w}_{ij}^T\boldsymbol{\alpha}$, and $y_{\alpha ij}^* = \mathbf{G}_{ij}^T\boldsymbol{\nu} - \mathbf{B}_{ij}^T\boldsymbol{\beta} - \mathbf{z}_{ij}^T\mathbf{b}_i$.

4.4 Numerical Studies

4.4.1 A simulation study

To study the empirical performance of the proposed semivarying coefficient model with nonparametric transformation, we conduct a simulation study based on the following model:

$$f(y_{ij}) = \tilde{y}_{ij} = x_{ij1}\gamma_1(u_{ij1}) + x_{ij2}\gamma_2(u_{ij2}) + \mathbf{z}_{ij}^T\mathbf{b}_i + \mathbf{w}_{ij}^T\boldsymbol{\alpha}, \quad (4.17)$$

where $\mathbf{b}_i = (b_{i1}, b_{i2})^T \sim N(\mathbf{0}, \boldsymbol{\Phi})$, $i = 1, \ldots, 100$ are subject specific random effects, and $\epsilon_{ij} \sim N(0, 1)$, $i = 1, \ldots, 100$, $j = 1, \ldots, 16$, which are independent of $\mathbf{b}_i$. Three typical non-normal distributions of y_{ij}, including highly skewed, U-shaped, and bimodal distributions, are considered in our simulation study. To produce

non-normal data y_{ij}, we simulate $\tilde{y}_{ij}$ based on equation (4.17), and then take the inverse transformation $f^{-1}(\cdot)$ for $\tilde{y}_{ij}$ such that $y_{ij} = f^{-1}(\tilde{y}_{ij})$. The $f(\cdot)$ corresponding to the three settings of non-normality are as follows: (1) $f^{-1}(v) = -\exp(-0.2v)$, resulting in a highly skewed y_{ij}; (2) $f^{-1}(v) = \exp(0.8(v-0.5))/[1+\exp\{0.8(v-0.5)\}]$, resulting in a non-symmetrically U-shaped y_{ij}; and (3) $f^{-1}(v) = 1.2v + \sin(v)$, resulting in a bimodal y_{ij}. The histograms of the simulated data under these three settings are depicted in the lower part of Figure 4.1. The varying coefficient functions for the three settings are given as follows: (1) for skewed y_{ij}, $\gamma_1(u) = 0.5\exp((u-8)/4) - 2$, $\gamma_2(u) = (0.18(u-8))^2$; (2) for U-shaped y_{ij}, $\gamma_1(u) = \sin(0.5(u-8))$, $\gamma_2(u) = (0.15(u-8))^3$; and (3) for bimodal y_{ij}, $\gamma_1(u) = 1.25\phi(u-8)$, $\gamma_2(u) = (0.25(u-8))$, where $\phi(\cdot)$ is the standard normal probability density function. We fix z_{1ij} and w_{1ij} at 1 to make b_{i1} and α_1 to be random and fixed intercepts, respectively. The mean of b_{i1} is 0, and b_{i1} and α_1 are identified. z_{2ij} are generated from the standard normal distribution. x_{1ij} and x_{2ij}, w_{2ij} and w_{3ij} are independently generated from the uniform distribution on $(-2, 2)$, and u_{1ij} and u_{2ij} are independently generated from the uniform distribution on $(0, 16)$. The true population values of the unknown parameters are taken as $\boldsymbol{\alpha}^T = (0, 1, 1)$ and $\boldsymbol{\Phi} = (1.0, 0.3; 0.3, 1.0)$; $\sigma^2 = 1$ is fixed for identification purposes.

The P-splines with 25 equal-distant knots in the support of y_{ij}, u_{ij1}, and u_{ij2} are used to approximate the transformation function $f(\cdot)$ and the varying coefficient functions $\gamma_1(u_{ij1})$ and $\gamma_2(u_{ij2})$. The prior inputs in (4.10) are taken as: $\boldsymbol{\alpha}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $r_0 = 5$, and $\mathbf{R}_0 = \boldsymbol{\Phi}$. The Bayesian estimates of unknown transformation function, varying coefficient functions, and unknown parameters are obtained based on 100 replications. In each replication, 2,000 burn-in samples are discarded, and 7,000 samples are acquired as

posterior samples. The estimated transformation functions and the varying coefficient functions, together with their 95% pointwise credible intervals, are respectively depicted in the upper part of Figure 4.1, and Figure 4.2, which indicate that the proposed approach accurately estimates the unknown transformation functions and varying coefficient functions. The bias (BIAS) and the root mean square (RMS) between the true values of the parameters and their estimates are reported in Tables 4.1 and 4.2 under the column "P-spline", respectively. The obtained results show that the proposed method performs well when the response variables are highly non-normal. Note that although the P-splines can estimate the shape of $f(\cdot)$ under the constraint $\mathbf{Q}_y\boldsymbol{\nu} = 0$, the estimated $\mathbf{G}_{ij}\boldsymbol{\nu} = 0$ deviates from $f(y_{ij})$ roughly with a constant shift. Thus, the estimated α_1, which represents the overall mean, is meaningless in the transformation model and is not presented in Table 4.1.

To study the sensitivity of the Bayesian results to the prior inputs, the simulated data sets are reanalyzed by using two different prior settings: (I) $\boldsymbol{\alpha}_0 = \mathbf{1}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $r_0 = 5$, $\mathbf{R}_0 = 2\boldsymbol{\Phi}$, $a_1 = 1$, and $a_2 = 0.05$; and (II) $\boldsymbol{\alpha}_0 = -\mathbf{1}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$, $r_0 = 5$, $\mathbf{R}_0 = 0.5\boldsymbol{\Phi}$, $a_1 = 0.001$, and $a_2 = 0.001$. The estimated unknown parameters, varying coefficient functions and nonparametric transformation functions are close to those reported in Tables 4.1 and 4.2, and Figures 4.1 and 4.2. Therefore, the Bayesian results obtained by our method are not sensitive to the considered prior inputs under the given sample sizes and model settings.

To compare our method with other possible existing methods, the simulated data are reanalyzed by the following conventional transformations. We consider the Box-Cox transformation with the index parameter λ in $\{-1, -1/2, 0, 1/2, 1\}$, which are commonly used and are associated with the reciprocal, reciprocal of

square root, logarithm, square root, and identity transformation, respectively. The identity transformation with $\lambda = 1$ is equivalent to ignoring the non-normality and simply fitting the data to the semivarying coefficient model defined in the right hand side of equation (4.17). The estimated unknown parameters obtained with the Box-Cox transformation method are also reported in Tables 4.1 and 4.2 for comparison. The results demonstrate that the performance of our proposed method is much better than those associated with the Box-Cox transformations. The program is written in R. It takes about 15 minutes to produce all the Bayesian estimates for one replication using a PC with Core 2 8400@3.00 GHz and 2G RAM.

4.5 Conclusion

In this chapter, a semivarying coefficient model with nonparametric transformation is proposed to analyze data with repeatedly measured highly non-normal variables. The unknown transformation functions and unknown varying coefficient functions are approximated with Bayesian P-splines. The nonparametric transformation considered in our model is different from the traditional parametric transformation methods such as the Box-Cox transformation. A modified Random-Ray algorithm together with other MCMC methods are proposed to solve difficulties in developing our methodology. A simulation study demonstrates that the proposed method performs satisfactorily in handling repeatedly measured and highly non-normal data.

Table 4.1: The Bayesian estimates of fix effect coefficients based on 100 replications

	Par		P-spline	λ in Box-Cox transfomation				
				-1	-0.5	0	0.5	1
Skewed	α_2	Bias	-0.0041	-1.026	-1.06	-0.757	-0.55	-0.402
		Rms	0.034	1.026	1.061	0.759	0.551	0.404
	α_3	Bias	0.0071	-1.023	-1.058	-0.759	-0.055	0.402
		Rms	0.028	1.023	1.058	0.761	0.552	0.403
U-shaped	α_2	Bias	0.0003	-1.356	-1.550	-0.232	-0.170	-0.191
		Rms	0.033	1.357	1.550	0.234	0.173	0.194
	α_3	Bias	0.0035	-1.357	-1.551	-0.231	-0.169	-0.19
		Rms	0.027	1.357	1.551	0.233	0.171	0.192
Bimodal	α_2	Bias	-0.0036	-1.027	-1.102	-0.524	-0.21	-0.131
		Rms	0.033	1.028	1.102	0.525	0.212	0.135
	α_3	Bias	-0.0067	-1.029	-1.103	-0.523	-0.208	-0.129
		Rms	0.027	1.029	1.103	0.524	0.21	0.132

Note: Under transformation, the estimated $\tilde{y}_{ij}$ is close the $\tilde{y}_{ij}$ generated in the simulation with a constant shift. Therefore, overall mean α_1 is meaningless and is not reported here.

Table 4.2: The Bayesian estimates of diagonal elements of the covariance matrix of random effects based on 100 replications

	Par		P-spline	λ in Box-Cox transfomation				
				-1	-0.5	0	0.5	1
Skewed	ϕ_{11}	Bias	0.011	-0.945	-0.693	-0.498	0.201	-0.544
		Rms	0.139	0.946	0.697	0.507	0.268	0.548
	ϕ_{12}	Bias	0.009	-0.259	-0.194	-0.135	0.06	-0.147
		Rms	0.104	0.286	0.227	0.17	0.107	0.162
	ϕ_{22}	Bias	0.012	-0.292	-0.554	-0.366	0.283	-0.482
		Rms	0.143	0.862	0.65	0.492	0.391	0.493
U-shaped	ϕ_{11}	Bias	0.0014	-0.809	-0.654	-0.378	0.371	0.107
		Rms	0.137	0.811	0.657	0.388	0.382	0.138
	ϕ_{12}	Bias	-0.0025	-0.214	-0.185	-0.129	-0.064	-0.012
		Rms	0.01	0.223	0.194	0.144	0.087	0.057
	ϕ_{22}	Bias	-0.0189	-0.751	0.615	-0.402	0.306	0.036
		Rms	0.014	0.754	0.062	0.412	0.319	0.081
Bimodal	ϕ_{11}	Bias	-0.0097	-0.956	-0.721	-0.541	-0.164	-0.238
		Rms	0.139	0.975	0.723	0.548	0.2	0.265
	ϕ_{12}	Bias	0.0007	-0.272	-0.213	-0.152	-0.05	-0.071
		Rms	0.103	0.281	0.225	0.166	0.088	0.106
	ϕ_{22}	Bias	0.0097	-0.875	-0.646	-0.46	-0.126	-0.253
		Rms	0.144	0.888	0.662	0.482	0.172	0.276

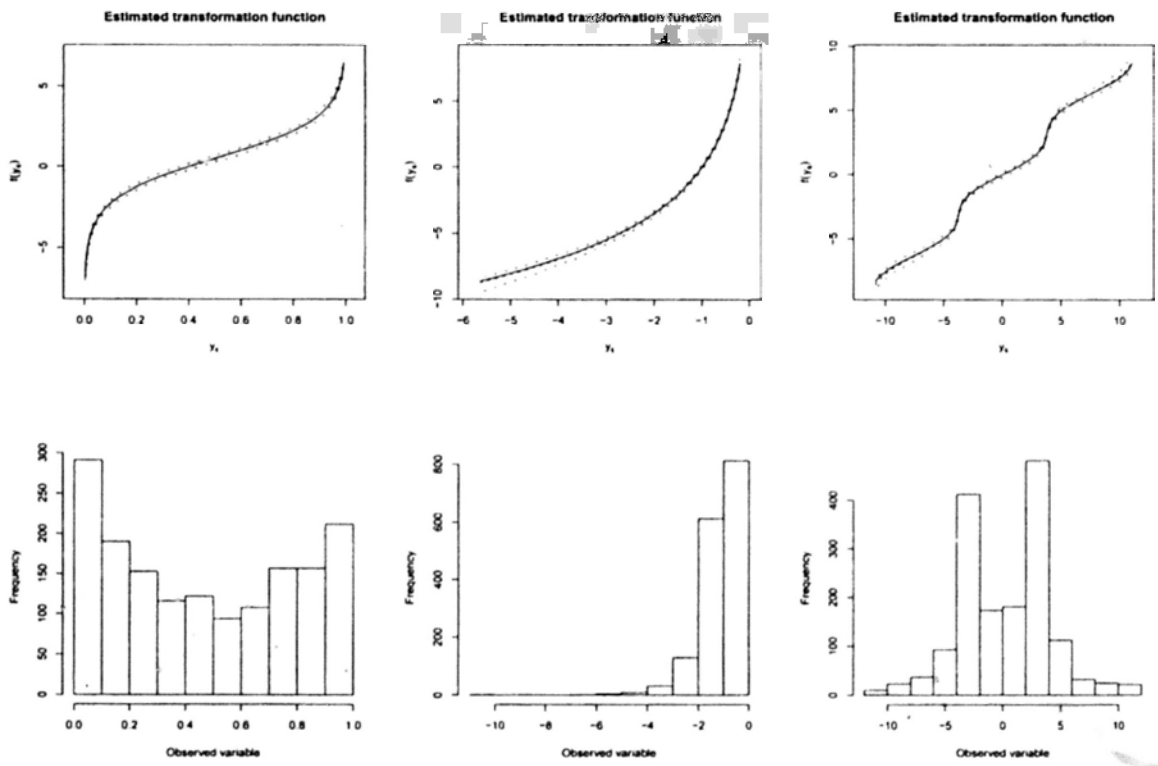

Figure 4.1: The first three graphs in the upper part are the estimates of the unknown transformation functions $f()$ in the simulation. The solid curves represent the underlying true curves. The estimates of the pointwise posterior mean curves are depicted by dashed lines. The dot-dash curves represent the 2.5%- and 97.5%-pointwise quantiles based on 100 replications. The lower part are the histogram of y_{ij} applied to the corresponding transformation.

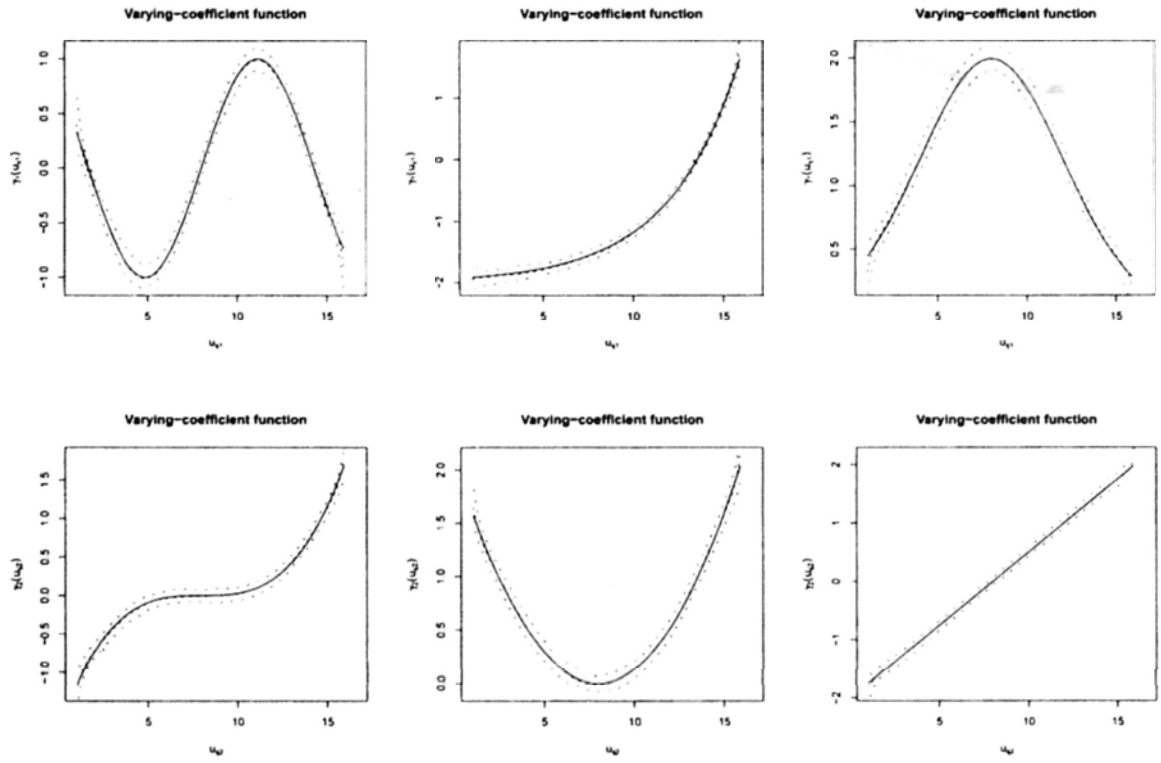

Figure 4.2: Upper and lower graphs are estimated $\gamma_1()$ and $\gamma_2()$, from U-shape, Skewed, and Bimodal situations, respectively. The solid curves represent the underlying true curves. The estimates of the pointwise posterior mean curves are depicted by dashed lines. The dot-dash curves represent the 2.5%- and 97.5%-pointwise quantiles based on 100 replications.

Chapter 5

Finite Mixture Varying Coefficient Models

5.1 Introduction

In this chapter, a finite mixture of varying coefficient models is motivated from a longitudinal study concerning compulsory treatment for heroin patients carried out by the California Civil Addict Program. How treatment affects the level of heroin use in different patient groups is of great interest. In this longitudinal study, assuming the effect of treatment to be invariant with respect to time is irrational. Moreover, it will be more informative if the dynamic treatment effect in each patient group can be estimated. As discussed in Chapter 4, varying coefficient models will be a good candidate for modeling this kind of time dependent effects. Another important concern of this study is to distinguish patient groups with different patterns of heroin use. To accommodate the possible existence of heterogeneity in this longitudinal data, we propose to use a finite mixture of varying coefficient models.

Heterogeneous population, which is composed of several subpopulations (groups), is inevitable in many fields such as economics, psychology, education, and sociology. The analysis of treating

heterogeneous population as homogenous will bury the valuable information of each subpopulation (group). And the estimates that ignore different characteristics of subpopulations (groups) are misleading and hard to interpret. As the group membership of each independent observation is unknown, methods for multiple group problems (situations with known group membership) can not be applied directly. As a result, mixture modeling (Redner and Walker, 1984; Titterington et al., 1985; McLachlan and Peel, 2000), in which the group memberships of observations are treated as unknown quantities and are estimated together with parameters, has received much attention in both statistics and other disciplines.

Finite mixture models have been studied extensively using different estimation procedures, including the method of moments (Lindsay and Basak, 1993), the maximum likelihood method (Hathaway, 1985; Yung, 1997; Lee and Song, 2003b) and the Bayesian method (Diebolt and Robert, 1994; Richardson and Green, 1997; Lee and Song, 2003c; Lee, 2007). Determining the number of mixture components is also an important issue. In a Bayesian framework, Richardson and Green (1997) developed a full Bayesian approach based on the reversible jump Markov chain Monte Carlo (MCMC) algorithm (Green, 1995). Lee and Song (2003c) proposed a model selection procedure for selecting the number of components in a finite mixture of structural equation models (SEMs) by using Bayes factor together with path sampling method. Recently, Cai et al. (2010) developed a finite mixture of SEMs with nonignorable missing responses and covariates. They determined the number of mixture components by a model selection criterion, namely modified Deviance Information Criterion (DIC, Spiegelhalter et al., 2002; Celeux et al., 2006). In this chapter, we use a Bayesian approach to analyze the finite mixture of varying coefficient models.

The number of components is determined by the modified DIC because it can avoid complicated varying dimensional MCMC algorithm and its computational burden is light given the simulated MCMC samples in the estimation procedure.

This chapter is organized as follows. Section 5.2 defines the proposed mixture varying coefficient models. The prior distributions of unknown parameters in this model are also discussed. Section 5.3 develops the full conditional distributions required in producing the estimation and determining the number of components involved in the mixture model. In Section 5.4, a simulation study is conducted to demonstrate the performance of our proposed methodology. The method is applied to the motivated example and some interesting findings are obtained. A conclusion is given in Section 5.5.

5.2 Model Description

5.2.1 General model specification

For $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip_i})^T$ be a random vector of observed variables measured in each of the n independent observations, and $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijq_1})^T$ and $\mathbf{u}_{ij} = (u_{ij1}, \ldots, u_{ijq_1})^T$ be vectors of fixed covariates. In practice, for each subject i, y_{ij} can represent one of the repeated measurements at p_i different time points. In the proposed mixture varying coefficient models, y_{ij} is assumed to have the following probability density function:

$$p(y_{ij}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R) = \sum_{r=1}^{R} \varphi_{ir} p_r(y_{ij}|\boldsymbol{\theta}_r), \tag{5.1}$$

for $j = 1, \ldots, p_i$, $i = 1, \ldots, n$. R is a given integer representing the number of components used to model the density of y_{ij}. $\boldsymbol{\theta}_r$ is a vector of parameters which characterize the density of the rth

component of the mixture. φ_{ir} is the probability of the ith observation in the rth component, and $\sum_{r=1}^{R} \varphi_{ir} = 1$, for $i = 1, \ldots, n$. According to a common practice in mixture modeling, it is useful to introduce a latent allocation variable ζ_i. ζ_i is assumed to take integer value $1, \ldots, R$ with probability $p(\zeta_i = r) = \varphi_{ir}$, $r = 1, \ldots, R$. $\zeta_i = r$ indicates that y_{ij} is from the rth component. The probability density function of y_{ij} given ζ_i is

$$p(y_{ij}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R, \zeta_i = r) = p_r(y_{ij}|\boldsymbol{\theta}_r). \tag{5.2}$$

Conditional on $\zeta_i = r$, y_{ij} is modeled by a varying coefficient model which is defined as follows:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\gamma}_r(\mathbf{u}_{ij}) + \mathbf{z}_{ij}^T \mathbf{b}_i + \mathbf{w}_{ij}^T \boldsymbol{\alpha}_r + \epsilon_{ij}, \tag{5.3}$$

where $\mathbf{z}_{ij}$ is a $q_2 \times 1$ vector of covariates, and ϵ_{ij} is a random error independently distributed as $N(0, \sigma_r^2)$. It is assumed that $\mathbf{b}_i$ is independent of ϵ_{ij} and follows a multivariate normal distribution $N(\boldsymbol{\mu}_r, \boldsymbol{\Phi}_r)$. $\mathbf{w}_{ij}$ and $\boldsymbol{\alpha}_r$ are $q_3 \times 1$ vectors of covariates and coefficients, respectively. Let $\boldsymbol{\gamma}_r(\mathbf{u}_{ij}) = (\gamma_{1r}(u_{ij1}), \ldots, \gamma_{q_1 r}(u_{ijq_1}))^T$ be a vector of functional coefficients related to $\mathbf{x}_{ij}$, and $(u_{ij1}, \ldots, u_{ijq_1})^T$ be a vector of covariates. They are usually called modifiers, such as time or location, etc. Elements of $\boldsymbol{\gamma}_r()$ are assumed to be unknown smooth functions, which allow the coefficients of $\mathbf{x}_{ij}$ to vary according to $\mathbf{u}_{ij}$.

For notational simplicity, we assume $q_1 = 1$. An extension to the case with $q_1 > 1$ is straightforward. Under this assumption, the model (5.3) is simplified as:

$$y_{ij} = x_{ij} \gamma_r(u_{ij}) + \mathbf{z}_{ij}^T \mathbf{b}_i + \mathbf{w}_{ij}^T \boldsymbol{\alpha}_r + \epsilon_{ij}. \tag{5.4}$$

We use the Bayesian P-splines to model $\gamma_r(u_{ij})$. The $\gamma_r(u_{ij})$ is

approximated by the following sum of B-splines $B_k(u_{ij})$:

$$\gamma_r(u_{ij}) = \sum_{k=1}^{K} \beta_{rk} B_k(u_{ij}), \tag{5.5}$$

where K is the number of splines determined by the number of knots in the support of u_{ij}. Let $\boldsymbol{\beta}_r = (\beta_{r1}, \ldots, \beta_{rK})^T$ be a vector of unknown coefficients in the rth component. K is set in advance between 10 and 60 to ensure that the approximation is adequate. Denote $\mathbf{B}_{ij} = (x_{ij}B_1(u_{ij}), \ldots, x_{ij}B_K(u_{ij}))^T$. With $\gamma_r(u_{ij})$ approximated by (5.5), model (5.4) can be expressed as:

$$y_{ij} = \mathbf{B}_{ij}^T \boldsymbol{\beta}_r + \mathbf{z}_{ij}^T \mathbf{b}_i + \mathbf{w}_{ij}^T \boldsymbol{\alpha}_r + \epsilon_{ij}. \tag{5.6}$$

Let $\boldsymbol{\varphi}_i = (\varphi_{i1}, \ldots, \varphi_{iR})$, $i = 1, \ldots, n$. $\boldsymbol{\varphi}_i$ is modeled by the following multiple logistic regression model:

$$p(\zeta_i = r) = \varphi_{ir} = \frac{\exp(\boldsymbol{\vartheta}_r^T \mathbf{v}_i)}{\sum_{r=1}^{R} \exp(\boldsymbol{\vartheta}_r^T \mathbf{v}_i)}, \tag{5.7}$$

where $\boldsymbol{\vartheta}_r$ and $\mathbf{v}_i$ are vectors of coefficients and covariates, respectively. The probability of each observation in certain component differs for certain reasons characterized by the covariates in $\mathbf{v}_i$. $\boldsymbol{\vartheta}_R$ is set to be $\mathbf{0}$ for model identification according to a common practice.

5.2.2 Prior distribution

Using a common prior specification in Bayesian P-splines, the prior distribution of $\boldsymbol{\beta}_r$ is:

$$\begin{aligned} p(\boldsymbol{\beta}_r) &= \prod_{k=d+1}^{K} \frac{1}{\sqrt{2\pi}\tau_r} \exp\left\{ -\frac{1}{2\tau_r^2}(\beta_{rk} - \beta_{rk0})^2 \right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\tau_r}\right)^{K-d} \exp\left\{ -\frac{1}{2\tau_r^2} \boldsymbol{\beta}_r^T \mathbf{M} \boldsymbol{\beta}_r \right\}, \end{aligned} \tag{5.8}$$

where d is the order of the random walk, $\beta_{rk0} = \beta_{r,k-1}$ and $\beta_{rk0} = 2\beta_{r,k-1} - \beta_{r,k-2}$ are the first- and second-order random walks, respectively. $\mathbf{M} = (D_{d-1} \times \cdots \times D_0)^T (D_{d_l-1} \times \cdots \times D_0)$, in which D_l, $l = 0, \ldots, d-1$, are defined as in (3.6).

In a full Bayesian analysis, the inverse smoothing parameters τ_r^2 are treated as random. According to the similar reason as given before, for $r = 1, \ldots, R$,

$$\tau_r^{-2} \sim Gamma(a_{1r}, a_{2r}), \tag{5.9}$$

where a_{1r} and a_{2r} are specified hyperparameters and we use $a_{1r} = 1$ and $a_{2r} = 0.005$ to obtain a highly dispersed (but proper) gamma prior of τ_r^{-2}.

For $\boldsymbol{\vartheta}_r$ in (5.7), $r = 1, \ldots, R-1$, we assign a multivariate normal prior:

$$\boldsymbol{\vartheta}_r \sim N(\boldsymbol{\vartheta}_{0r}, \boldsymbol{\Sigma}_{\vartheta 0r}). \tag{5.10}$$

For the parameters involved in the right hand side of model (5.6), the following conjugate prior distributions are assigned. For $r = 1, \ldots, R$,

$$\begin{aligned} \boldsymbol{\alpha}_r &\sim N(\boldsymbol{\alpha}_{0r}, \boldsymbol{\Sigma}_{\alpha 0r}), \quad & \boldsymbol{\Phi}_r^{-1} &\sim Wishart(\mathbf{R}_{0r}, \rho_{0r}), \\ \boldsymbol{\mu}_r &\sim N(\boldsymbol{\mu}_{0r}, \boldsymbol{\Sigma}_{\mu 0r}), \quad & \sigma_r^{-2} &\sim Gamma(a_{\sigma 1r}, a_{\sigma 2r}), \end{aligned} \tag{5.11}$$

where $\boldsymbol{\vartheta}_{0r}$, $\boldsymbol{\alpha}_{0r}$, ρ_{0r}, $\boldsymbol{\mu}_{0r}$, $a_{\sigma 1r}$, $a_{\sigma 2r}$, and positive definite matrices $\boldsymbol{\Sigma}_{\vartheta 0r}$, $\boldsymbol{\Sigma}_{\alpha 0r}$, $\mathbf{R}_{0r}$, and $\boldsymbol{\Sigma}_{\mu 0r}$ are hyperparameters whose values are assumed to be given by the prior information.

5.3 Estimation and Model Selection of Mixture Varying Coefficient Models

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip_i})^T$, $\mathbf{Y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$, $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)^T$, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^T$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$, where $\boldsymbol{\theta}_r$ is a vector that

contains all unknown parameters in $\boldsymbol{\beta}_r$, $\boldsymbol{\alpha}_r$, $\boldsymbol{\mu}_r$, $\boldsymbol{\Phi}_r$, σ_r^2, and $\boldsymbol{\vartheta}_r$. The complete-data likelihood of observed response variables $\mathbf{Y}$ and unobserved latent variables $\mathbf{b}$ and $\boldsymbol{\zeta}$ is

$$
\begin{aligned}
& p(\mathbf{Y}, \mathbf{b}, \boldsymbol{\zeta}|\boldsymbol{\theta}) \\
= \; & p(\mathbf{Y}|\mathbf{b}, \boldsymbol{\zeta}, \boldsymbol{\theta}) p(\mathbf{b}|\boldsymbol{\zeta}, \boldsymbol{\theta}) p(\boldsymbol{\zeta}|\boldsymbol{\theta}) \\
= \; & \prod_{i=1}^{n} \Bigg[\frac{1}{(\sqrt{2\pi})^{q_2} |\boldsymbol{\Phi}_{\zeta_i}|^{1/2}} \exp\Big\{ -\frac{1}{2} (\mathbf{b}_i - \boldsymbol{\mu}_{\zeta_i})^T \boldsymbol{\Phi}_{\zeta_i}^{-1} (\mathbf{b}_i - \boldsymbol{\mu}_{\zeta_i}) \Big\} \\
& \frac{\exp(\boldsymbol{\vartheta}_{\zeta_i}^T \mathbf{v}_i)}{\sum_{r=1}^{R} \exp(\boldsymbol{\vartheta}_r^T \mathbf{v}_i)} \prod_{j=1}^{p_i} \Big\{ \frac{1}{\sqrt{2\pi\sigma_{\zeta_i}^2}} \exp\Big(-\frac{1}{2\sigma_{\zeta_i}^2} (y_{ij} - \\
& \mathbf{B}_{ij}^T \boldsymbol{\beta}_{\zeta_i} - \mathbf{z}_{ij}^T \mathbf{b}_i - \mathbf{w}_{ij}^T \boldsymbol{\alpha}_{\zeta_i})^2 \Big) \Big\} \Bigg]. \qquad (5.12)
\end{aligned}
$$

We use the Gibbs sampler (Geman and Geman, 1984) to draw observations from the joint posterior distribution of all unknown quantities, $p(\boldsymbol{\theta}, \mathbf{b}, \boldsymbol{\zeta}|\mathbf{Y})$, for Bayesian estimation. The full conditional distributions in implementing the Gibbs sampler are presented as follows.

5.3.1 Full conditional distributions

(a) *Full conditional distributions of $\boldsymbol{\beta}_r$ and τ_r*

Let $\mathbf{B}_i = (\mathbf{B}_{i1}, \ldots, \mathbf{B}_{in_i})^T$, and $\mathbf{B} = (\mathbf{B}_1^T, \ldots, \mathbf{B}_n^T)^T$. Denote $\tilde{\mathbf{B}}_r$ be the submatrix of $\mathbf{B}$ which only retains $\mathbf{B}_i$ with $\zeta_i = r$. Let $\boldsymbol{\beta}_{-r} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{r-1}, \boldsymbol{\beta}_{r+1}, \ldots, \boldsymbol{\beta}_R)$, $y_{ij}^* = y_{ij} - \mathbf{z}_{ij}^T \mathbf{b}_i - \mathbf{w}_{ij}^T \boldsymbol{\alpha}_r$, $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{in_i}^*)^T$, $\mathbf{y}^* = (\mathbf{y}_1^{*T}, \ldots, \mathbf{y}_n^{*T})^T$, and $\mathbf{y}_{(r)}^*$ be a subvector which retains the elements $\mathbf{y}_i^*$ in $\mathbf{y}^*$ if $\zeta_i = r$. The full conditional distributions of $\boldsymbol{\beta}_r$, $r = 1, \ldots, R$, are:

$$
p(\boldsymbol{\beta}_r|\mathbf{Y}, \mathbf{b}, \boldsymbol{\zeta}, \boldsymbol{\beta}_{-r}, \boldsymbol{\alpha}_r, \sigma^2) \overset{D}{=} N(\boldsymbol{\beta}_r^*, \boldsymbol{\Sigma}_{\beta r}), \qquad (5.13)
$$

where $\boldsymbol{\Sigma}_{\beta r} = (\tilde{\mathbf{B}}_r^T \tilde{\mathbf{B}}_r / \sigma_r^2 + \mathbf{M}/\tau_r^2)^{-1}$ and $\boldsymbol{\beta}_r^* = \boldsymbol{\Sigma}_{\beta r} (\tilde{\mathbf{B}}_r^T \mathbf{y}_{(r)}^* / \sigma_r^2)$. Sampling from this multivariate normal distribution can be done

efficiently using the algorithm in Rue (2004) because $\mathbf{\Sigma}_{\beta r}$ is a block diagonal matrix.

For $l = 0, \cdots, q_1$,

$$p(\tau_r^{-2}|\boldsymbol{\beta}_r) \stackrel{D}{=} Gamma\Big[\alpha_{1r} + \frac{K-d}{2},\ \alpha_{2r} + \frac{1}{2}\boldsymbol{\beta}_r^T \mathbf{M} \boldsymbol{\beta}_r\Big]. \quad (5.14)$$

(b) *Full conditional distributions of* $\boldsymbol{\alpha}_r$, σ_r^2, $\boldsymbol{\mu}_r$ *and* $\mathbf{\Phi}_r$.

For $r = 1, \ldots, R$, denote $\mathbf{W} = (\mathbf{w}_{11}, \ldots, \mathbf{w}_{1p_1}, \ldots, \mathbf{w}_{n1}, \ldots, \mathbf{w}_{np_n})^T$, $y_{\sigma ij} = y_{ij} - \mathbf{B}_{ij}^T \boldsymbol{\beta}_r - \mathbf{z}_{ij}^T \mathbf{b}_i - \mathbf{w}_{ij}^T \boldsymbol{\alpha}_r$, $y_{\alpha ij} = y_{ij} - \mathbf{B}_{ij}^T \boldsymbol{\beta}_r - \mathbf{z}_{ij}^T \mathbf{b}_i$, $\mathbf{y}_\sigma = (y_{\sigma 11}, \ldots, y_{\sigma 1p_1}, \ldots, y_{\sigma n1}, \ldots, y_{\sigma np_n})^T$, and $\mathbf{y}_\alpha = (y_{\alpha 11}, \ldots, y_{\alpha 1p_1}, \ldots, y_{\alpha n1}, \ldots, y_{\alpha np_n})^T$. Let $\mathbf{W}_r$ and $\mathbf{b}_{(r)}$ be submatrices of $\mathbf{W}$ and $\mathbf{b}$ without $\mathbf{w}_{ij}$ and $\mathbf{b}_i$ if $\zeta_i \neq r$, respectively. Define $\mathbf{y}_{\sigma r}$, $\mathbf{y}_{\alpha r}$, and $\mathbf{Y}_r$ be subvectors of $\mathbf{y}_\sigma$, $\mathbf{y}_\alpha$, and $\mathbf{Y}$, which only retain $y_{\sigma ij}$, $y_{\alpha ij}$, and y_{ij} if $\zeta_i = r$, respectively. Let $S_r = \{i|\zeta_i = r\}$, and n_r^* be the number of integers in S_r, $r = 1, \ldots, R$.

$$p(\sigma_r^{-2}|\mathbf{Y}_r, \mathbf{b}_{(r)}, \boldsymbol{\beta}_r, \boldsymbol{\alpha}_r) \stackrel{D}{=} Gamma(a_{\sigma 1r}^*, a_{\sigma 2r}^*), \quad (5.15)$$

$$p(\boldsymbol{\alpha}_r|\mathbf{Y}_r, \mathbf{b}_{(r)}, \boldsymbol{\beta}_r, \sigma_r^2) \stackrel{D}{=} N(\boldsymbol{\alpha}_r^*, \mathbf{\Sigma}_{\alpha r}^*), \quad (5.16)$$

$$p(\boldsymbol{\mu}_r|\mathbf{b}_{(r)}, \mathbf{\Phi}_r) \stackrel{D}{=} N(\boldsymbol{\mu}_r^*, \mathbf{\Sigma}_{\mu r}^*), \quad (5.17)$$

$$p(\mathbf{\Phi}_r^{-1}|\mathbf{b}_{(r)}, \boldsymbol{\mu}_r) \stackrel{D}{=} Wishart\Big(\mathbf{R}_r^*,\ n_r^* + \rho_{0r}\Big), \quad (5.18)$$

where $a_{\sigma 1r}^* = a_{\sigma 1r} + \frac{1}{2}\sum_{i \in S_r} n_i$, $a_{\sigma 2r}^* = a_{\sigma 2r} + \frac{1}{2}\mathbf{y}_{\sigma r}^T \mathbf{y}_{\sigma r}$, $\mathbf{\Sigma}_{\alpha r}^* = (\mathbf{W}_r \mathbf{W}_r^T + \mathbf{\Sigma}_{\alpha 0r}^{-1})^{-1}$, $\boldsymbol{\alpha}_r^* = \mathbf{\Sigma}_{\alpha r}^*(\mathbf{W}_r^T \mathbf{y}_{\alpha r} + \mathbf{\Sigma}_{\alpha 0r}^{-1} \boldsymbol{\alpha}_{0r})$, $\mathbf{\Sigma}_{\mu r}^* = (n_r^* \mathbf{\Phi}_r^{-1} + \mathbf{\Sigma}_{\mu 0r}^{-1})^{-1}$, and $\boldsymbol{\mu}_r^* = \mathbf{\Sigma}_{\mu r}^*(\mathbf{\Phi}_r^{-1} \sum_{i \in S_r} \mathbf{b}_i + \mathbf{\Sigma}_{\mu 0r}^{-1} \boldsymbol{\mu}_{0r})$. $\mathbf{R}_r^* = \sum_{i \in S_r} (\mathbf{b}_i - \boldsymbol{\mu}_r)(\mathbf{b}_i - \boldsymbol{\mu}_r)^T + \mathbf{R}_{0r}$.

(c) *Full conditional distribution of* $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1^T, \ldots, \boldsymbol{\vartheta}_{R-1}^T)^T$.

$$p(\boldsymbol{\vartheta}|\boldsymbol{\zeta}) \propto p(\boldsymbol{\zeta}|\boldsymbol{\vartheta}) p(\boldsymbol{\vartheta})$$

$$= \Big[\frac{\prod_{i=1}^n \exp(\boldsymbol{\vartheta}_{\zeta_i}^T \mathbf{v}_i)}{\{\sum_{r=1}^R \exp(\boldsymbol{\vartheta}_r^T \mathbf{v}_i)\}^n}\Big] \prod_{r=1}^{R-1} \exp\Big\{ -\frac{1}{2}(\boldsymbol{\vartheta}_r - \boldsymbol{\vartheta}_{0r})^T \mathbf{\Sigma}_{\vartheta 0r}^{-1}(\boldsymbol{\vartheta}_r - \boldsymbol{\vartheta}_{0r})\Big\}. \quad (5.19)$$

The full conditional distribution of $\boldsymbol{\vartheta}$ is nonstandard. MCMC sampling schemes, such as the Metropolis-Hastings algorithm (MH) (Metropolis et al., 1953; Hastings, 1970), can be used to draw the posterior observations from this nonstandard distribution.

(d) *Full conditional distributions of* $\mathbf{b}_i$ *and* ζ_i

For $i = 1, \ldots, n$,

$$p(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta}, \zeta_i = r) \stackrel{D}{=} N(\mathbf{b}_i^*, \boldsymbol{\Sigma}_{bi}^*), \tag{5.20}$$

where $\boldsymbol{\Sigma}_{bi}^* = (\mathbf{Z}_i^T\mathbf{Z}_i + \boldsymbol{\Phi}_r^{-1})^{-1}$, $\mathbf{b}_i^* = \boldsymbol{\Sigma}_{bi}^*(\mathbf{Z}_i^T\mathbf{y}_{bi})$, $\mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{ip_i})^T$, $\mathbf{y}_{bi} = (y_{bi1}, \ldots, y_{bip_i})^T$, and $y_{bij} = y_{ij} - \mathbf{B}_{ij}^T\boldsymbol{\beta}_r - \mathbf{w}_{ij}^T\boldsymbol{\alpha}_r$.

For $r = 1, \ldots, R$. the full conditional probability mass function of ζ_i is:

$$p(\zeta_i = r|\mathbf{y}_i, \mathbf{b}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{y}_i, \mathbf{b}_i, \zeta_i = r|\boldsymbol{\theta})}{p(\mathbf{y}_i, \mathbf{b}_i)}. \tag{5.21}$$

The probability in the denominator is independent of ζ_i. To find $p(\zeta_i = r|\mathbf{y}_i, \mathbf{b}_i, \boldsymbol{\theta})$, it is sufficient to calculate

$$\begin{aligned} & p(\mathbf{y}_i, \mathbf{b}_i, \zeta_i = r|\boldsymbol{\theta}) \qquad (5.22) \\ = \; & p(\mathbf{y}_i|\mathbf{b}_i, \zeta_i = r, \boldsymbol{\theta})p(\mathbf{b}_i|\zeta_i = r, \boldsymbol{\theta})p(\zeta_i = r|\boldsymbol{\theta}) \\ = \; & \frac{1}{(\sqrt{2\pi})^{q_2}|\boldsymbol{\Phi}_r|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{b}_i - \boldsymbol{\mu}_r)^T\boldsymbol{\Phi}_r^{-1}(\mathbf{b}_i - \boldsymbol{\mu}_r)\Big\} \\ & \frac{\exp(\boldsymbol{\vartheta}_r^T\mathbf{v}_i)}{\sum_{r=1}^R \exp(\boldsymbol{\vartheta}_r^T\mathbf{v}_i)} \prod_{j=1}^{p_i} \left[\frac{1}{\sqrt{2\pi\sigma_r^2}} \exp\left\{ -\frac{1}{2\sigma_r^2}(y_{ij} - \right.\right. \\ & \left.\left. \mathbf{B}_{ij}^T\boldsymbol{\beta}_r - \mathbf{z}_{ij}^T\mathbf{b}_i - \mathbf{w}_{ij}^T\boldsymbol{\alpha}_r)^2 \right\}\right]. \end{aligned}$$

Hence, (5.21) can be calculated and ζ_i can be drawn from a multinomial distribution with known probability mass function.

The allocation variable ζ_i is of interest for classification purpose, which is estimated by

$$\hat{\zeta}_i = argmax_r\{\hat{\varphi}_{ir}\}, \tag{5.23}$$

where $\hat{\varphi}_{ir} = \sum_{t=1}^{J} I(\zeta_i^{(t)} = r)/J$, J is the number of MCMC samples after burn-in period, and $\zeta_i^{(t)}$ is the sample of ζ_i at the tth MCMC iteration.

5.3.2 Identification issue

In finite mixture models, it is well known that under symmetric prior of parameters in different components, the label switching problem has to be solved for identification purpose. The label switching is caused by the fact that the likelihood in mixture models (e.g. (5.1)) is invariant with a permutation of the group labels $1, \ldots, R$, so is the case for the posterior distribution under symmetric prior. This will lead to $R!$ subspaces with identical posterior distribution, each of which corresponds to a different way in labeling the groups. The resulting posterior distribution will be multimodal, which is challenging in sampling with MCMC algorithm. The general idea to solve this problem is to put constraints on the parameters in order to get samples from only one subspace out of $R!$ candidates. In the literature, several solutions have been proposed to solve the label switching problem, see Celeux et al. (2000), Stephens (2000), Frühwirth-Schnatter (2001), Jasra et al. (2005), among others. Inspired by Lee (2007), we use the random permutation sampler by Frühwirth-Schnatter (2001). The idea of random permutation sampler is as follows. Each sample is first drawn from the posterior distribution in the unconstrained parameter space. The simulated samples are relabeled according to some predetermined parameter constraints described later. The relabeled samples can then be used for MCMC inference. The corresponding MCMC algorithm using the full conditional distributions in the unconstrained space is as follows. Given latent variables and parameters at the tth iteration of the MCMC chain, $\{\boldsymbol{\theta}^{(t)}, \mathbf{b}^{(t)}, \boldsymbol{\zeta}^{(t)}\}$,

the sample at the $(t+1)$th iteration is generated as follows:

(a) Draw $\boldsymbol{\theta}^{(t+1)}$ from $p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{b}^{(t)}, \boldsymbol{\zeta}^{(t)})$ according to (5.13) - (5.19);

(a) Draw $\mathbf{b}^{(t+1)}$ from $p(\mathbf{b}|\mathbf{Y}, \boldsymbol{\theta}^{(t+1)}, \boldsymbol{\zeta}^{(t)})$ according to (5.20);

(a) Draw $\boldsymbol{\zeta}^{(t+1)}$ from $p(\boldsymbol{\zeta}|\mathbf{Y}, \boldsymbol{\theta}^{(t+1)}, \mathbf{b}^{(t+1)})$ according to (5.21);

(d) Let S^* be a permutation of $1, \ldots, R$ corresponding to a predetermined identification constraint. Relabeling $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$ such that $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{S^*(r)}$, where $S^*(r)$ is the rth element in S^*. For $i = 1, \ldots, n$, let $\zeta_i = r$ if $\zeta_i = S^*(r)$ before relabeling.

As described in Lee (2007), we use the permutation sampler in two stages. At the first stage, the sampling scheme based on steps (a) - (c) is implemented in an unconstrained space. This procedure produces samples that explore the whole unconstrained parameter space and jump between the different labeling subspaces in a balanced fashion. The resulting samples are used to find an appropriate identification constraint used in the permutation sampler. At the second stage, we use the permutation sampler to produce the posterior samples from a specific subspace determined by the chosen identification constraint at the first stage. Specifically, S^* is determined based on the results obtained with steps (a) - (c) in an unconstrained space at stage 1, and then is used to perform the permutation sampler with step (d) at stage 2.

5.3.3 Selecting the number of components with a modified Deviance Information Criterion

There are several methods to determine the number of components in finite mixture models. Richardson and Green (1997) proposed to treat the number of components as random and use the reversible

jump MCMC algorithm to sample from the joint posterior distribution of the number of components and other unknown quantities in the model. Lee and Song (2003b) compared mixture models with different fixed number of components, and selected an appropriate number by using Bayes factor and path sampling procedure. We use a modified Deviance Information Criterion (DIC) to determine the number of components in the mixture of varying coefficient models. DIC was proposed by Spiegelhalter et al. (2002), which is a generalization of Akaike Information Criterion (AIC) because it aims to seek for an appropriate model by finding the balance between the measure of goodness-of-fit and model complexity under a Bayesian framework. It can be estimated with the MCMC samples, which are the by-products in the estimation procedure. Thus, the extra computational burden in calculating DIC is light. As noted by existing literature (see for example Spiegelhalter et al., 2003; Cai et al., 2010), applying DIC directly in selecting mixture models is problematic. We adopt a modified DIC proposed by Celeux et al. (2006) to overcome this problem.

In terms of our proposed model, the modified DIC can be expressed as

$$-4E_{\boldsymbol{\theta},\mathbf{b},\boldsymbol{\zeta}}[\log p(\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}|\boldsymbol{\theta})] + 2E_{\mathbf{b},\boldsymbol{\zeta}}[\log p(\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}|\boldsymbol{\theta}^*)|\mathbf{Y}], \quad (5.24)$$

where $p(\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}|\boldsymbol{\theta})$ is the complete-data likelihood defined in (5.12), and $\boldsymbol{\theta}^* = E_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}]$. The first expectation in (5.24) can be estimated as follows:

$$E_{\boldsymbol{\theta},\mathbf{b},\boldsymbol{\zeta}}[\log p(\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}|\boldsymbol{\theta})] \approx \frac{1}{J}\sum_{t=1}^{J} \log p(\mathbf{Y},\mathbf{b}^{(t)},\boldsymbol{\zeta}^{(t)}|\boldsymbol{\theta}^{(t)}).$$

To estimate $E_{\mathbf{b},\boldsymbol{\zeta}}[\log p(\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}|\boldsymbol{\theta}^*)|\mathbf{Y}]$, at the tth iteration, we generate J_1 extra samples $\boldsymbol{\theta}^{(t,1)},\ldots,\boldsymbol{\theta}^{(t,J_1)}$ from $p(\boldsymbol{\theta}|\mathbf{Y},\mathbf{b}^{(t)},\boldsymbol{\zeta}^{(t)})$ by repeating (a) in Section 5.3.2 J_1 times. Let $\boldsymbol{\theta}_*^{(t)} = \sum_{t_1=1}^{J_1}\boldsymbol{\theta}^{(t,t_1)}/J_1$.

The second expectation in (5.24) can be estimated as follows:

$$E_{\mathbf{b},\boldsymbol{\zeta}}[\log p(\mathbf{Y},\mathbf{b},\boldsymbol{\zeta}|\boldsymbol{\theta}^*)|\mathbf{Y}] \approx \frac{1}{J}\sum_{t=1}^{J}\log p(\mathbf{Y},\mathbf{b}^{(t)},\boldsymbol{\zeta}^{(t)}|\boldsymbol{\theta}_*^{(t)}).$$

The model with the smallest DIC value will be selected.

5.4 Numerical Studies

5.4.1 A simulation study

We conduct a simulation study to evaluate the empirical performance of the proposed mixture varying coefficient models. Conditional on $\zeta_i = r$, y_{ij} is defined as follows:

$$y_{ij} = x_{ij1}\gamma_{1r}(u_{ij1}) + x_{ij2}\gamma_{2r}(u_{ij2}) + z_{ij}b_i + w_{ij1}\alpha_{1r} + w_{ij2}\alpha_{2r} + \epsilon_{ij}, \tag{5.25}$$

where x_{ij1}, x_{ij2}, u_{ij1}, and u_{ij2} are generated independently from U(-1,1), which is the uniform distribution on $[-1,1]$. z_{ij} is fixed at 1 for all i and j, and $b_i \sim N(\mu_r, \phi_r)$ is a subject-specific random intercept, which is independent of $\epsilon_{ij} \sim N(0,\sigma_r^2)$. w_{i11} and w_{i12} are generated from U(-1,1), and let $w_{i11} = \ldots = w_{in_i1}$ and $w_{i12} = \ldots = w_{in_i2}$ for all i, which are subject-invariant covariates. The latent group variable ζ_i is generated from the multinomial distribution with probabilities specified by the multiple logistic regression model:

$$p(\zeta_i = r) = \varphi_{ir} = \frac{\exp(\vartheta_{1r}v_{i1} + \vartheta_{2r}v_{i2})}{\sum_{r=1}^{R}\exp(\vartheta_{1r}v_{i1} + \vartheta_{2r}v_{i2})}, \tag{5.26}$$

where v_{i1} and v_{i2} are generated independently from U(-1,1).

The simulation is conducted in two scenarios. In the first scenario, the mixture is composed of two components. The true values of the unknown parameters are taken as $\mu_1 = -1$, $\mu_2 = 1$,

$\alpha_{11} = 0.6$, $\alpha_{21} = 0.6$, $\alpha_{12} = -0.6$, $\alpha_{22} = -0.6$, $\phi_1 = \phi_2 = 0.5$, $\sigma_1^2 = \sigma_2^2 = 0.3$, $\vartheta_{11} = 1$, $\vartheta_{21} = 1$, $\vartheta_{12} = 0$, and $\vartheta_{22} = 0$. The proportion of each component is around 50%. The underlying varying coefficients are defined as:

$$\gamma_{11}(u) = -4\phi(2(u+0.5)), \quad \gamma_{21}(u) = 1.5\sin(2.5u),$$
$$\gamma_{12}(u) = \exp(u), \quad \gamma_{22}(u) = \cos(2u) + 0.5u,$$

where $\phi(\cdot)$ is the probability density function of standard normal distribution.

In the second scenario, the mixture is composed of three components. The true values of the unknown parameters are taken as $\mu_1 = -2$, $\mu_2 = 0$, $\mu_3 = 2$, $\alpha_{11} = 0.6$, $\alpha_{21} = 0.6$, $\alpha_{12} = -0.6$, $\alpha_{22} = -0.6$, $\alpha_{13} = 0.6$, $\alpha_{23} = -0.6$, $\phi_1 = \phi_2 = \phi_3 = 0.5$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.3$, $\vartheta_{11} = 1$, $\vartheta_{21} = -1$, $\vartheta_{12} = -1$, $\vartheta_{22} = 1$, $\vartheta_{13} = 0$, and $\vartheta_{23} = 0$. The proportions of components are about 40%, 40%, 20% for $r = 1, 2, 3$, respectively. The underlying varying coefficients are defined as:

$$\gamma_{11}(u) = -4\phi(2(u+0.5)), \quad \gamma_{21}(u) = 1.5\sin(2.5u),$$
$$\gamma_{12}(u) = \exp(u), \quad \gamma_{22}(u) = \cos(2u) + 0.5u,$$
$$\gamma_{13}(u) = 1.5u^3, \quad \gamma_{23}(u) = 1.5u^2.$$

Each scenario is studied with two different sample sizes. In the first scenario, $n = 400$ and $n = 800$ are used, and in the second scenario $n = 500$ and $n = 1000$ are used. P-splines in (5.5) with $K_l = 20$ are used to approximate the $\gamma_{lr}(u_{ijl})$, $l = 1, 2$. The prior distributions in (5.9) – (5.11) are taken as invariant with r as follows. (I): for all r, $a_{1r} = 1$, $a_{2r} = 0.005$, $\boldsymbol{\vartheta}_{0r} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\vartheta 0r} = 0.0001\mathbf{I}$, $\boldsymbol{\alpha}_{0r} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\alpha 0r} = 10^{-4}\mathbf{I}$, $\mathbf{R}_{0r} = 0.5$, $\rho_{0r} = 3$, $\boldsymbol{\mu}_{0r} = 0$, $\boldsymbol{\Sigma}_{\mu 0r} = 10^{-4}\mathbf{I}$, $a_{\sigma 1r} = 3$, and $a_{\sigma 2r} = 1$, where $\mathbf{I}$ denotes the identity matrix with appropriate dimension. After checking convergence with well separated starting values, we find the algorithm converges after 2,000

MCMC iterations. The estimates of varying coefficient functions and unknown parameters are obtained based on 100 replications. In each replication, 2,000 burn-in samples were discarded, and 4,000 samples were acquired for posterior inference.

In first scenario with sample sizes $n = 400$ and $n = 800$, the estimated varying coefficient functions and the 95% pointwise credible intervals are depicted in Figures 5.1 and 5.2, respectively. The bias (BIAS) and the root mean square (RMS) between the Bayesian estimates and the true values of parameters are reported in Table 5.1. The obtained results show that the estimates of unknown parameters are accurate and the estimated varying coefficient functions are close to the underlying true functions. The performance of the estimation improves as sample size increases. Results of the second scenario are reported in Figures 5.3 and 5.4, and Table 5.2, which give the same conclusion.

To demonstrate that the modified DIC can select the right number of mixture components, the simulated data in the first scenario with $n = 400$ are reanalyzed with 1- and 3-component mixture models defined by (5.1) and (5.25). Let M_k denote the mixture model with k components, $k = 1, 2, 3$. The modified DIC values of M_1 to M_3 are calculated. The results are reported in the left part of Table 5.4. The modified DIC value of M_2 is the smallest in all 100 replications, indicating that the true model M_2 can be consistently selected by the modified DIC in this scenario. The simulated data in the second scenario with $n = 500$ are reanalyzed by M_1 and M_2. The modified DIC values of M_1 to M_3 are calculated. The results are reported in the right part of Table 5.4. The true model M_3 is also consistently selected in all 100 replications in the second scenario.

To study the sensitivity of the Bayesian results to the prior

inputs, the simulated data in the first scenario with $n = 400$ are reanalyzed with two additional different prior settings: (II) $a_{1r} = 1$, $a_{2r} = 0.05$, $\boldsymbol{\vartheta}_{0r} = \mathbf{21}$, $\boldsymbol{\Sigma}_{\vartheta 0r} = \mathbf{I}$, $\boldsymbol{\alpha}_{0r} = \mathbf{51}$, $\boldsymbol{\Sigma}_{\alpha 0r} = \mathbf{I}$, $\rho_{0r} = 4$, $\mathbf{R}_{0r} = 1$, $\boldsymbol{\mu}_{0r} = \mathbf{21}$, $\boldsymbol{\Sigma}_{\mu 0r} = \mathbf{I}$, $a_{\sigma 1r} = 9$, and $a_{\sigma 2r} = 8$; (III) $a_{1r} = 0.001$, $a_{2r} = 0.001$, $\boldsymbol{\vartheta}_{0r} = -\mathbf{21}$, $\boldsymbol{\Sigma}_{\vartheta 0r} = \mathbf{I}$, $\boldsymbol{\alpha}_{0r} = -\mathbf{51}$, $\boldsymbol{\Sigma}_{\alpha 0r} = \mathbf{I}$, $\rho_{0r} = 7$, $\mathbf{R}_{0r} = 0.2$, $\boldsymbol{\mu}_{0r} = -\mathbf{21}$, $\boldsymbol{\Sigma}_{\mu 0r} = \mathbf{I}$, $a_{\sigma 1r} = 9$, and $a_{\sigma 2r} = 2$; where $\mathbf{I}$ is an identity matrix and $\mathbf{1}$ is a vector of 1 with appropriate dimensions. The estimates of unknown parameters are reported in Table 5.3, and the estimates of varying coefficient functions are close to those reported in Figure 5.1. Hence, the Bayesian results obtained by our method are not very sensitive to the prior inputs under the given sample sizes and model settings.

The program is written in R. In a PC with Intel Core 2 8400@3.00 GHz and 2G RAM, it takes about 70 minutes to produce the Bayesian estimation and the modified DIC value in one replication of the first scenario with 400 samples.

5.4.2 Application: A longitudinal study of the treatment effect on the control of heroin use

In this section, we applied the mixture varying coefficient models to the motivated example concerning treatment effect on heroin use control (Hser et al., 2007). The sample consists of 437 patients originally admitted to the California Civil Addict Program, which was established to give compulsory drug treatment for narcotics-dependent criminal offenders committed under court order. Each patient was followed up at the first 16 years after his/her onset of heroin use. Self reported longitudinal data were acquired to study subsequent, long-term outcomes of interests. Whether the compulsory treatment had any effect on the control of heroin use and how the effect evolves are of particular interest. Another important

concern of the study is to distinguish different patient groups with different patterns of heroin use and treatment effects.

The data include time variant variables: "Mean days of heroin use per month (y_{ij})," and "Months in treatment per year (x_{ij})" in each of the 16 years, $i = 1, \ldots, 437$, $j = 1, \ldots, p_i$, where $p_i \leq 16$. The data are unbalance because some patients were incarcerated for 12 months in certain years. During these years, the patients were completely separated from heroin. Therefore, the mechanism of heroin use in these years were different from those when the patients had free access to heroin, and we exclude them from the data. The data also include time invariant variables: "age at the first treatment (w_{i1})" and "age at the first heroin use (w_{i2})," $i = 1, \ldots, 437$. The proposed model was applied to find the varying coefficients of treatment over time in possible groups of heroin users. In the rth group,

$$y_{ij} = \gamma_{1r}(u_{ij}) + x_{ij}\gamma_{2r}(u_{ij}) + w_{i1}\alpha_{1r} + w_{i2}\alpha_{2r} + b_i + \epsilon_{ij}, \quad (5.27)$$

where u_{ij} stands for the number of years from the last incarceration for 12 months in one year, $\gamma_{1r}(u_{ij})$ is the trend of overall mean of heroin use, and $\gamma_{2r}(u_{ij})$ represents the time-varying effect of treatment on heroin control in the rth group. $b_i \sim N(\mu_r, \phi_r)$ is a subject-specific random effect and is independent of $\epsilon_{ij} \sim N(0, \sigma_r^2)$. 10 equal-distant knots in $[1, 16]$ were used to construct the B-spline basis for estimating $\gamma_{1r}(u_{ij})$ and $\gamma_{2r}(u_{ij})$. As $\gamma_{1r}(\cdot)$ and μ_r are not identifiable, we use the method discussed in Section 2.4.1 to solve this problem.

Latent allocation variables ζ_i is assumed to follow a multinomial distribution with probability $(\varphi_{i1}, \ldots, \varphi_{iR})$:

$$p(\zeta_i = r) = \varphi_{ir} = \frac{\exp(\vartheta_{1r}v_{i1} + \vartheta_{2r}v_{i2} + \vartheta_{3r})}{\sum_{r=1}^{R} \exp(\vartheta_{1r}v_{i1} + \vartheta_{2r}v_{i2} + \vartheta_{3r})}, \quad (5.28)$$

where v_{i1} and v_{i2} are "age at the first treatment" and "age at the first heroin use." Before conducting the analysis, y_{ij}, x_{ij1}, w_{i1}, w_{i2}, v_{i1}, and v_{i2} were standardized.

We fitted the data with three models, 1-component (M_1), 2-component (M_2), and 3-component (M_3) mixture varying coefficient models, and compared their modified DIC values. Although a 4-component mixture model (M_4) was also fitted, the estimation exhibits serious label switching problem even under the identifiability constraint determined by the permutation sampler. Hence, we only considered the candidates M_1 to M_3. The prior inputs in (5.9) – (5.11) were taken as invariant with r: for all r, $a_{1r} = 1$, $a_{2r} = 0.005$, $\boldsymbol{\vartheta}_{0r} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\vartheta 0r} = 0.0001\mathbf{I}$, $\boldsymbol{\alpha}_{0r} = \mathbf{0}$, $\boldsymbol{\Sigma}_{\alpha 0r} = 10^{-4}\mathbf{I}$, $\mathbf{R}_{0r} = 0.15$, $\rho_{0r} = 3$, $\boldsymbol{\mu}_{0r} = 0$, $\boldsymbol{\Sigma}_{\mu 0r} = 10^{-4}\mathbf{I}$, $a_{\sigma 1r} = 3$, and $a_{\sigma 2r} = 1$. The permutation sampler was used to find a suitable identification constraint. The constraint which can separate the components well is $\mu_r > \mu_{r-1}$. After discarding 3500 burn-in iterations, 3500 samples were acquired to obtain the Bayesian estimates of unknown parameters, unknown coefficient functions, and the modified DIC values. $J_1 = 10$ was used to calculate the modified DIC. $DIC_{M_1} = 15913$, $DIC_{M_2} = 15444$, and $DIC_{M_3} = 15021$, indicating that a 3-component model is selected. The estimates of unknown parameters, together with their standard errors (SE) estimates under the selected model are reported in Table 5.5. The estimated varying coefficient functions and the corresponding 95% pointwise credible intervals are shown in Figure 5.5. The upper part of Figure 5.5 includes the trends of overall means of heroin use in each group, which are obtained by adjusting the scale of the estimates of $\gamma_{1r}(\cdot) + \mu_r$ to that of y_{ij} before standardization.

The data were reanalyzed with some perturbations of J_1 and the hyperparameters in the prior distributions. Under different

prior inputs, the modified DIC values with $J_1 = 20$ consistently select M_3, and the estimates of unknown parameters and varying coefficient functions in M_3 are close to those reported in Table 5.5 and Figure 5.5, Thus, the Bayesian results in this study are robust to the prior inputs.

The interpretation for the estimates of time-varying and time-invarying coefficients under the selected model is as follows: (1) The estimated trends of overall means depicted in the upper part of Figure 5.5 show the patterns of heroin use in each group. Group 1 (13%) is composed of serious addicted patients who used heroin frequently (more than 20 days per month) in the entire study, Group 2 (47%) consists of patients that used heroin with increasing frequency, and Group 3 (40%) is formed by patients who became addicted in the first two years and decreased steadily afterwards. (2) The lower part of Figure 5.5 shows different patterns of treatment effect in different groups. The treatment effects in Group 1 and Group 2 are positive in controlling heroin use. Also, there is a positive correlation between the severity of addict and the treatment effect, which indicates that the proposed treatment is especially useful for serious heroin addicts. For Group 3 in which the patients decreased heroin use continuously, the treatment effect had a reverse pattern, indicating that the proposed treatment may not be suitable for the less heroin addicts. Other possible treatments are desirable for Group 3 patients. (3) "age at the first heroin use" is negatively associated with the probability of Group 2 (ϑ_{22}), indicating that the younger the patients use heroin, the more likely they addict themselves to heroin with an increasing trend. Therefore, we should pay particular attention to those who involved in heroin early, and provide them prompt and effective treatments.

5.5 Conclusion

In this chapter, a finite mixture of varying coefficient models is analyzed. The unknown varying coefficient functions are modeled with Bayesian P-splines. The MCMC algorithm is used to obtain the Bayesian estimates of unknown parameters and varying coefficient functions. The modified DIC is used to determine the number of components in the mixture model. A simulation study shows that the proposed method can estimate the varying coefficient functions and unknown parameters accurately, and the DIC can correctly determine the number of components in the mixture model. The model is applied to a longitudinal study concerning treatment effect on the control of heroin use, where the data exhibit heterogeneity. Distinct patterns of heroin use and treatment effect in different patient groups were identified. The results show that the proposed mixture varying coefficient models is particularly useful in the analysis of heterogeneous data with dynamic effects.

Table 5.1: The Bayesian estimates of parameters in M_2 based on 100 replications

	n=400		n=800	
Parameters	Bias	RMS	Bias	RMS
α_{11}	0.015	0.093	0.004	0.056
α_{21}	0.014	0.104	0.012	0.060
b_{01}	0.012	0.050	0.002	0.037
ϕ_1	-0.006	0.059	-0.002	0.049
σ_1^2	0.003	0.016	0.004	0.013
ϑ_{11}	0.013	0.194	0.013	0.171
ϑ_{21}	-0.020	0.192	-0.026	0.163
α_{12}	-0.008	0.080	-0.007	0.066
α_{22}	-0.008	0.099	-0.012	0.076
b_{02}	0.002	0.062	0.002	0.037
ϕ_2	-0.005	0.071	-0.001	0.042
σ_2^2	0.003	0.015	0.006	0.014

Table 5.2: The Bayesian estimates of parameters in M_3 based on 100 replications

n=500			n=1000		
Parameters	Bias	RMS	Parameters	Bias	RMS
α_{11}	0.001	0.096	α_{11}	0.008	0.070
α_{21}	0.012	0.117	α_{21}	0.002	0.070
b_{01}	-0.000	0.063	b_{01}	0.000	0.040
ϕ_1	0.006	0.070	ϕ_1	0.006	0.044
σ_1^2	-0.000	0.018	σ_1^2	0.002	0.011
ϑ_{11}	0.028	0.187	ϑ_{11}	0.024	0.168
ϑ_{21}	0.015	0.194	ϑ_{21}	-0.001	0.163
α_{12}	0.000	0.110	α_{12}	-0.011	0.065
α_{22}	-0.010	0.103	α_{22}	0.004	0.078
b_{02}	-0.016	0.070	b_{02}	-0.004	0.041
ϕ_2	0.001	0.073	ϕ_2	0.007	0.049
σ_2^2	0.002	0.017	σ_2^2	0.001	0.012
ϑ_{12}	0.014	0.260	ϑ_{12}	0.009	0.168
ϑ_{22}	0.071	0.249	ϑ_{22}	-0.005	0.151
α_{13}	0.000	0.117	α_{13}	-0.007	0.088
α_{23}	-0.021	0.127	α_{23}	-0.011	0.097
b_{03}	0.002	0.073	b_{03}	-0.004	0.046
ϕ_3	0.001	0.084	ϕ_3	0.005	0.059
σ_3^2	-0.002	0.020	σ_3^2	0.000	0.015

Table 5.3: The sensitivity analysis of the Bayesian estimates in M_2 with $n = 400$ based on 100 replications

	Prior (I)		Prior (II)		Prior (III)	
Parameters	Bias	RMS	Bias	RMS	Bias	RMS
α_{11}	0.015	0.093	0.058	0.110	-0.027	0.093
α_{21}	0.014	0.104	0.061	0.118	-0.027	0.108
b_{01}	0.012	0.050	0.023	0.055	-0.005	0.049
ϕ_1	-0.006	0.059	-0.008	0.059	-0.020	0.062
σ_1^2	0.003	0.016	0.015	0.022	0.000	0.017
ϑ_{11}	0.013	0.194	0.145	0.229	-0.018	0.191
ϑ_{21}	-0.020	0.192	0.104	0.194	-0.075	0.199
α_{12}	-0.008	0.080	0.031	0.090	-0.053	0.096
α_{22}	-0.008	0.099	0.027	0.098	-0.054	0.110
b_{02}	0.002	0.062	0.006	0.062	0.008	0.063
ϕ_2	-0.005	0.071	-0.006	0.072	-0.015	0.074
σ_2^2	0.003	0.015	0.016	0.022	0.000	0.014

Table 5.4: The estimates of the modified DIC in two scenarios of the simulation study.

	Scenario 1 M_2, n=400		Scenario 2 M_3, n=500	
	DIC	SD	DIC	SD
M_1	6442.93	69.74	8219.78	86.37
M_2	4651.19	81.10	6934.56	115.7
M_3	4888.06	77.90	6131.62	91.91

Table 5.5: The Bayesian estimates of parameters in heroin use control study with the selected model M_3

Component 1			Component 2			Component 3		
Para	Est	SE	Para	Est	SE	Para	Est	SE
α_{11}	0.008	0.041	α_{12}	-0.087	0.042	α_{13}	-0.109	0.038
α_{21}	0.003	0.032	α_{22}	0.031	0.053	α_{23}	-0.012	0.043
ϑ_{11}	-0.211	0.184	ϑ_{12}	-0.075	0.131	ϑ_{13}	0	-
ϑ_{21}	0.018	0.170	ϑ_{22}	-0.704	0.163	ϑ_{23}	0	-
ϑ_{31}	-1.203	0.160	ϑ_{32}	0.096	0.160	ϑ_{33}	0	-
b_{01}	0.825	0.041	b_{02}	0.075	0.045	b_{03}	-0.197	0.045
ϕ_{111}	0.046	0.013	ϕ_{112}	0.186	0.028	ϕ_{113}	0.203	0.028
σ_1^2	0.214	0.017	σ_2^2	0.625	0.025	σ_3^2	0.603	0.026

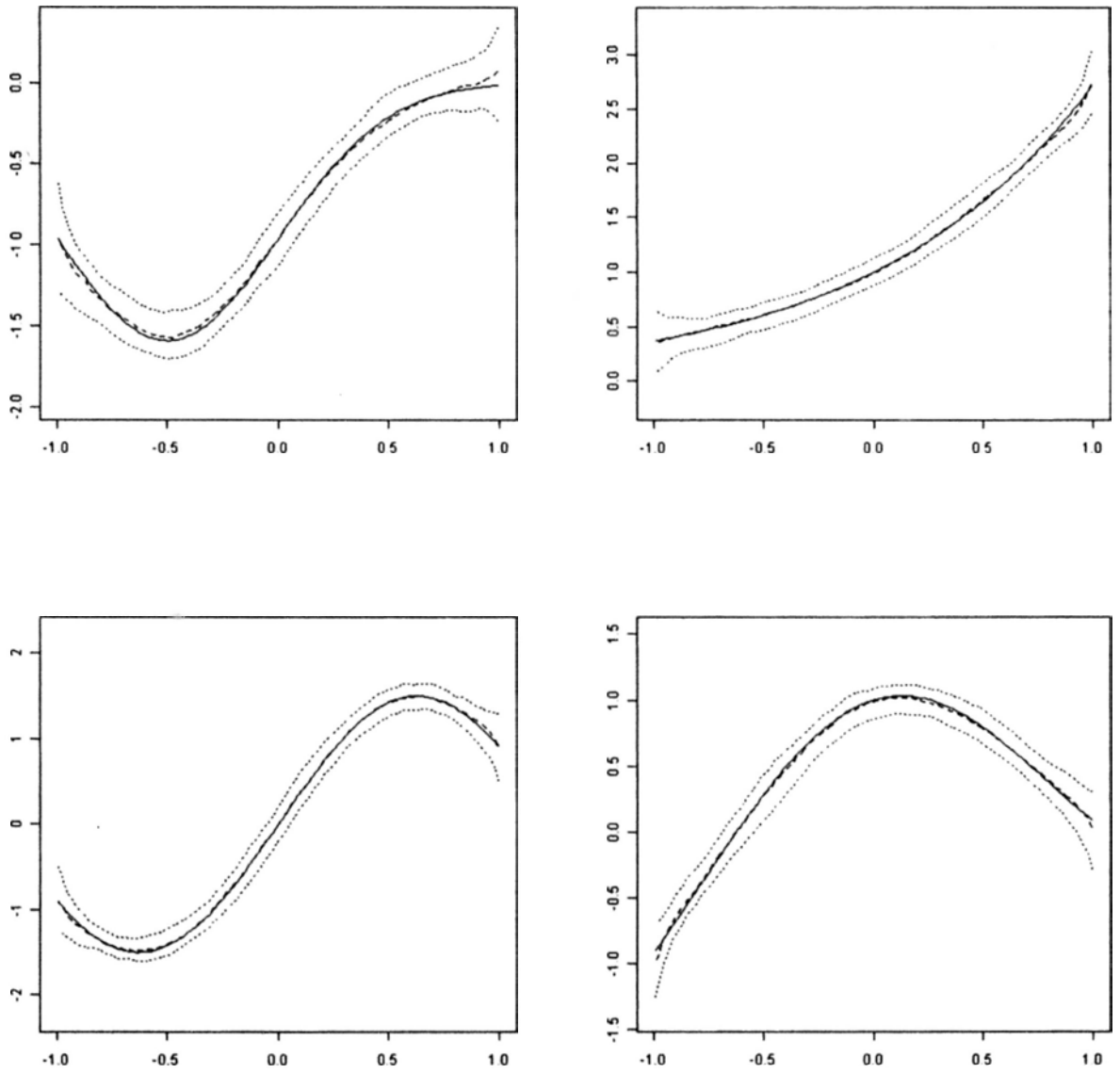

Figure 5.1: Estimated varying coefficient functions based on 100 replications in M_2 with sample size 400. From left to right, top to bottom, the figure is composed of $\gamma_{11}(\cdot)$, $\gamma_{12}(\cdot)$, $\gamma_{21}(\cdot)$, and $\gamma_{22}(\cdot)$. The solid curves represent the true curves, and the dashed and dotted curves are the pointwise median and 2.5% and 97.5 % quantiles of the varying coefficient functions.

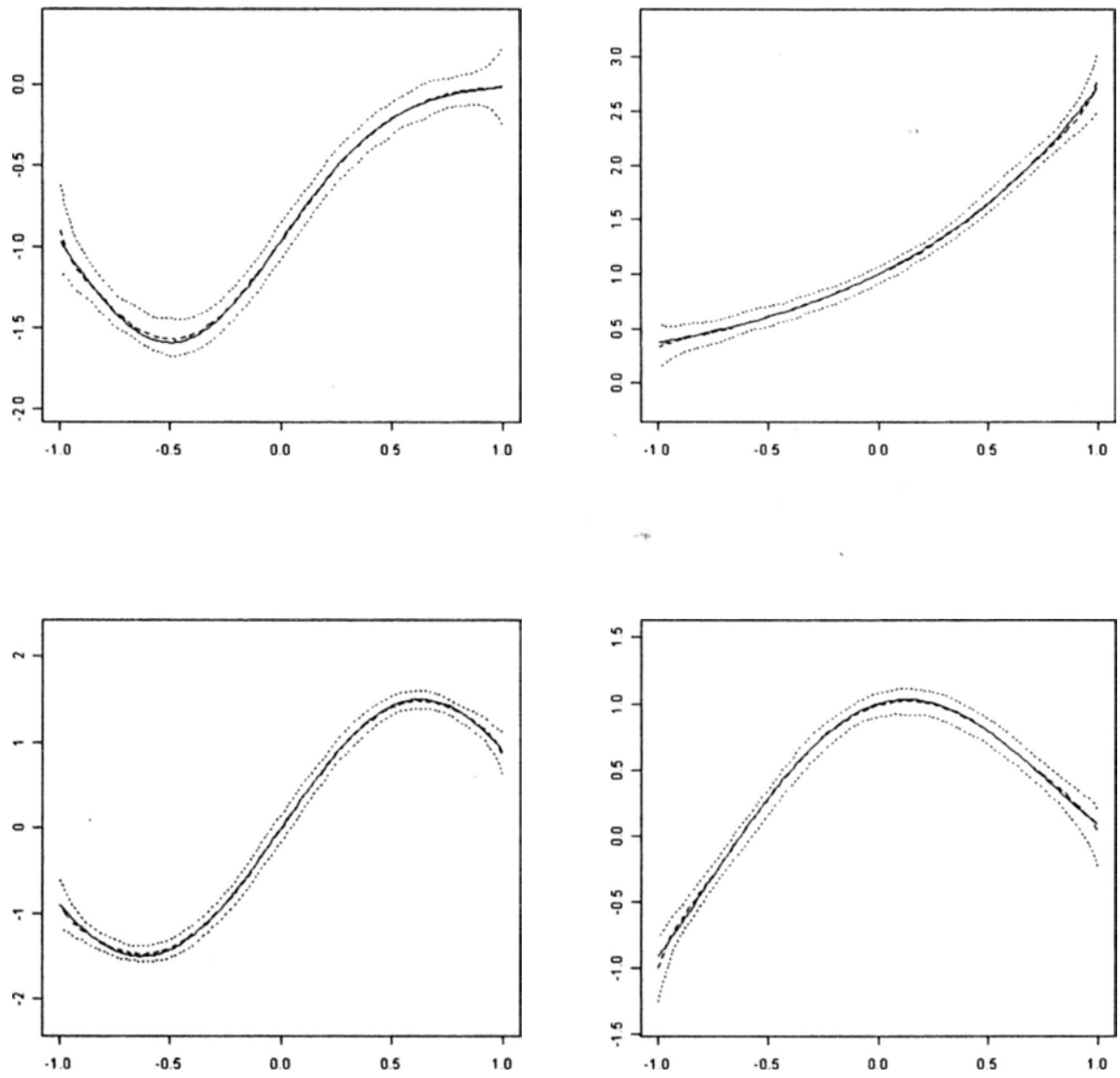

Figure 5.2: Estimated varying coefficient functions based on 100 replications in M_2 with sample size 800. From left to right, top to bottom, the figure is composed of $\gamma_{11}(\cdot)$, $\gamma_{12}(\cdot)$, $\gamma_{21}(\cdot)$, and $\gamma_{22}(\cdot)$. The solid curves represent the true curves, and the dashed and dotted curves are the pointwise median and 2.5% and 97.5 % quantiles of the varying coefficient functions.

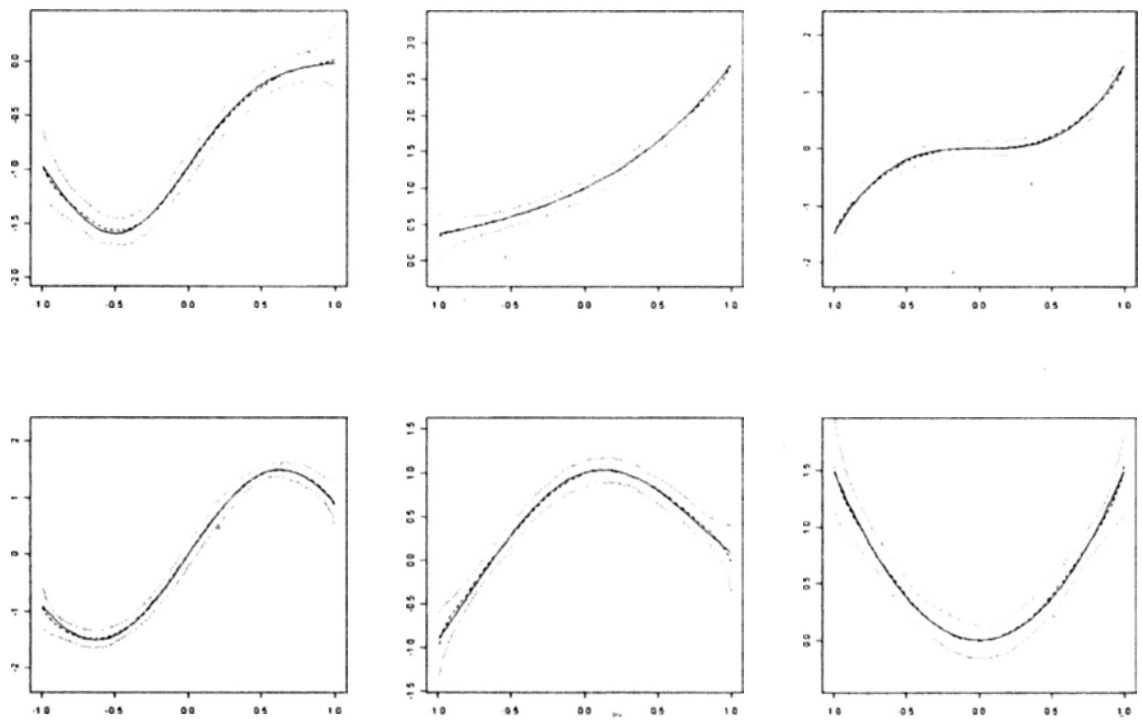

Figure 5.3: Estimated varying coefficient functions based on 100 replications in M_3 with sample size 500. From left to right, top to bottom, the figure is composed of $\gamma_{11}(\cdot)$, $\gamma_{12}(\cdot)$, $\gamma_{13}(\cdot)$, $\gamma_{21}(\cdot)$, $\gamma_{22}(\cdot)$, and $\gamma_{23}(\cdot)$. The solid curves represent the true curves, and the dashed and dotted curves are the pointwise median and 2.5% and 97.5 % quantiles of the varying coefficient functions.

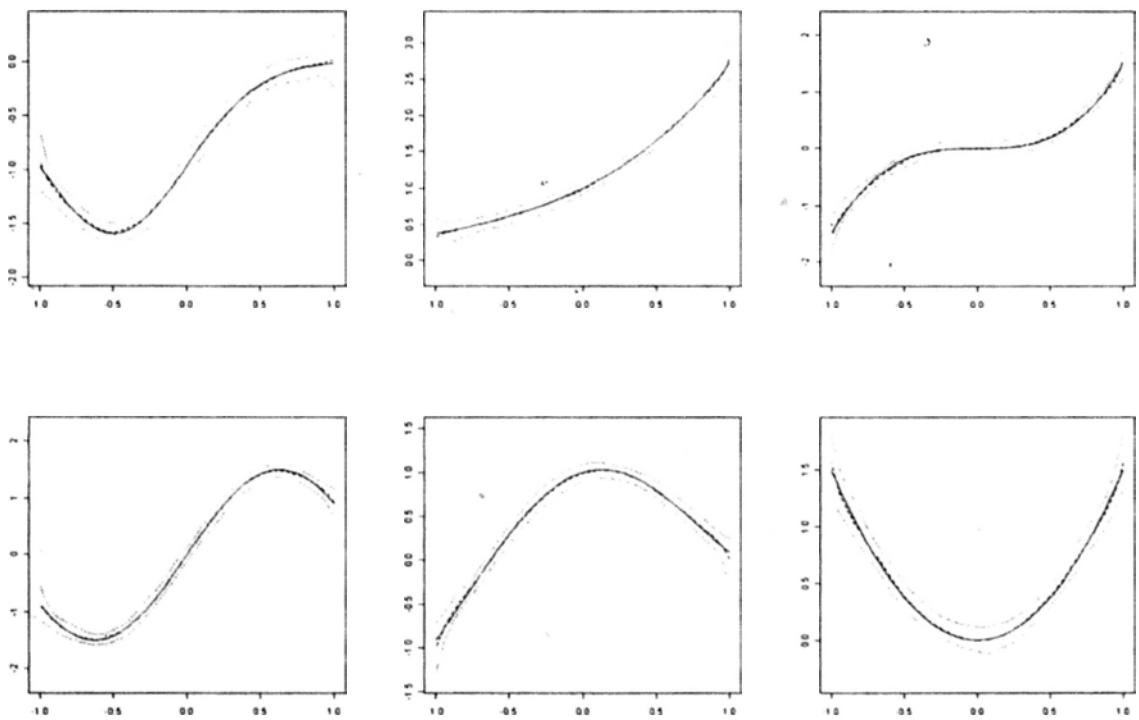

Figure 5.4: Estimated varying coefficient functions based on 100 replications in M_3 with sample size 1000. From left to right, top to bottom, the figure is composed of $\gamma_{11}(\cdot)$, $\gamma_{12}(\cdot)$, $\gamma_{13}(\cdot)$, $\gamma_{21}(\cdot)$, $\gamma_{22}(\cdot)$, and $\gamma_{23}(\cdot)$. The solid curves represent the true curves, and the dashed and dotted curves are the pointwise median and 2.5% and 97.5 % quantiles of the varying coefficient functions.

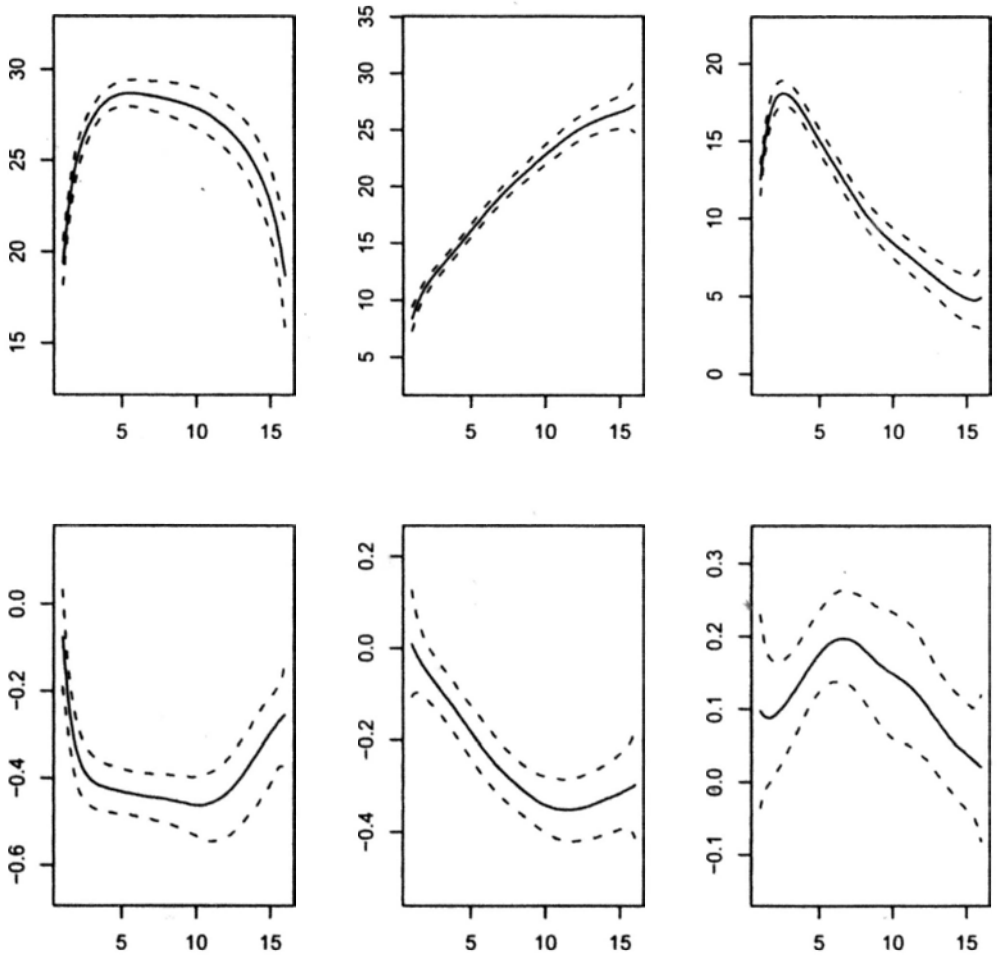

Figure 5.5: Estimated varying coefficient functions with M_3 in the real example. The upper part contains $\gamma_{11}(\cdot)$, $\gamma_{12}(\cdot)$, and $\gamma_{13}(\cdot)$, and the lower part includes $\gamma_{21}(\cdot)$, $\gamma_{22}(\cdot)$, and $\gamma_{23}(\cdot)$. The solid curves and dashed curves are the estimated pointwise median, and 2.5% and 97.5 % quantiles of the varying coefficient functions.

Appendix A

Technical Details of MCMC Sampler

The full conditional distributions involved in steps (a) and (b1)-(b7) (see Section 2.4.2), and the implementation of the MH algorithm are given below.

(A) Full conditional distribution of Ω in step (a).

$$\begin{aligned} & P(\boldsymbol{\Omega}|\mathbf{Y},\boldsymbol{\theta}) \\ = & \prod_{i=1}^{n} P(\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta}) \propto \prod_{i=1}^{n} P(\mathbf{y}_i|\boldsymbol{\omega}_i,\boldsymbol{\theta}) P(\eta_i|\boldsymbol{\xi}_i,\boldsymbol{\theta}) P(\boldsymbol{\xi}_i|\boldsymbol{\theta}) \\ \propto & \prod_{i=1}^{n} \Big[\{ \prod_{j=1}^{p} P(y_{ij}|\boldsymbol{\omega}_i,\mathbf{A}_j,\boldsymbol{\Lambda}_j,\psi_j) \} P(\eta_i|\boldsymbol{\xi}_i,\psi_\delta,\mathbf{b},\boldsymbol{\beta},\mathbf{s}) p(\boldsymbol{\xi}_i|\boldsymbol{\Phi}) \Big]. \end{aligned} \tag{A.1}$$

So

$$\begin{aligned} & P(\boldsymbol{\omega}_i|\mathbf{y}_i,\boldsymbol{\theta}) \\ \propto & \exp\{-\sum_{j=1}^{p} \frac{1}{2\psi_j}(y_{ij} - \mathbf{A}_j^T\mathbf{c}_i - \boldsymbol{\Lambda}_j^T\boldsymbol{\omega}_i)^2\} \exp(-\frac{1}{2}\boldsymbol{\xi}_i^T\boldsymbol{\Phi}^{-1}\boldsymbol{\xi}_i) \\ \times \exp & \Big[-\frac{1}{2\psi_\delta} \Big\{ \eta_i - \sum_{d=1}^{D}\sum_{k=1}^{K_{bd}} b_{dk} B_{dk}^x(x_{id}) - \sum_{j=1}^{q_2}\sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij}|s_j) \Big\}^2 \Big]. \end{aligned} \tag{A.2}$$

The conditional distribution given in (A.2) is nonstandard and complex. Hence, the MH algorithm is used to generate observations from this conditional distribution. For the target density $p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta})$, we choose $N[\boldsymbol{\mu}_\omega, \sigma_\omega^2 \boldsymbol{\Sigma}_\omega]$ as the proposal distribution, where

$$\boldsymbol{\Sigma}_\omega = \left\{ \begin{pmatrix} \psi_\delta^{-1} & -\psi_\delta^{-1}\boldsymbol{\beta}\boldsymbol{\Delta} \\ -\psi_\delta^{-1}(\boldsymbol{\beta}\boldsymbol{\Delta})^T & \boldsymbol{\Phi}^{-1} + \psi_\delta^{-1}\boldsymbol{\Delta}^T\boldsymbol{\beta}^T\boldsymbol{\beta}\boldsymbol{\Delta} \end{pmatrix} + \boldsymbol{\Lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \right\}^{-1},$$

where $\boldsymbol{\beta}\boldsymbol{\Delta}$ is the first derivative of $\sum_{j=1}^{q_2}\sum_{k=1}^{K_j}\beta_{jk}N_{jk}(\xi_{ij}|s_j)$ with respect to $\boldsymbol{\xi}_i$ at $\boldsymbol{\xi}_i = \mathbf{0}$. Let $q(\cdot|\boldsymbol{\mu}_\omega, \sigma_\omega, \boldsymbol{\Sigma}_\omega)$ be the proposal density corresponding to $N[\boldsymbol{\mu}_\omega, \sigma_\omega^2\boldsymbol{\Sigma}_\omega]$, the MH algorithm is implemented as follows: At the t-th iteration with a current value $\boldsymbol{\omega}_i^{(t)}$, a new candidate $\boldsymbol{\omega}_i$ is generated from $q(\cdot|\boldsymbol{\omega}_i^{(t)}, \sigma_\omega, \boldsymbol{\Sigma}_\omega)$, and accepted with the probability

$$\min\left\{1, \frac{p(\boldsymbol{\omega}_i|\mathbf{y}_i, \boldsymbol{\theta})}{p(\boldsymbol{\omega}_i^{(t)}|\mathbf{y}_i, \boldsymbol{\theta})}\right\}.$$

The variance σ_ω^2 can be chosen such that the average acceptance rate is approximately 0.45.

(B) Full conditional distributions of structural parameters $\mathbf{A}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, ψ_δ, and $\boldsymbol{\Phi}$ in steps (b1)-(b4).

Let $\mathbf{A}_j^T$ and $\boldsymbol{\Lambda}_j^T$ be the j-th rows of $\mathbf{A}$ and $\boldsymbol{\Lambda}$, respectively; and ψ_j be the j-th diagonal element of $\boldsymbol{\Psi}$. Let $y_{ij}^* = y_{ij} - \boldsymbol{\Lambda}_j^T\boldsymbol{\omega}_i$, $\mathbf{y}_j^* = (y_{1j}^*, \ldots, y_{nj}^*)^T$, $\tilde{y}_{ij} = y_{ij} - \mathbf{A}_j^T\mathbf{c}_i$, $\tilde{\mathbf{y}}_j = (\tilde{y}_{1j}, \ldots, \tilde{y}_{nj})^T$, $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$, and $\boldsymbol{\Omega}_2 = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$. The full conditional distribu-

tions of $\mathbf{A}_j$, $\mathbf{\Lambda}_j$, ψ_j, ψ_δ, and $\mathbf{\Phi}$ are given as follows:

$$\begin{aligned}
&p(\mathbf{A}_j|\cdot) \sim N(\boldsymbol{\mu}_{aj}, \mathbf{\Sigma}_{aj}), \qquad p(\mathbf{\Lambda}_j|\cdot) \sim N(\boldsymbol{\mu}_{\lambda j}, \mathbf{\Sigma}_{\lambda j}),\\
&p(\psi_\delta|\cdot) \sim IG[\alpha_{\delta 0} + n/2,\ \beta_\delta],\ p(\psi_j|\cdot) \sim IG[\alpha_{j0} + n/2,\ \beta_{\lambda j}],\\
&p(\mathbf{\Phi}|\cdot) \sim IW[\mathbf{\Omega}_2\mathbf{\Omega}_2^T + \mathbf{R}_0,\ n + \rho_0],
\end{aligned} \tag{A.3}$$

where $\mathbf{\Sigma}_{aj} = (\psi_j^{-1}\mathbf{C}\mathbf{C}^T + \mathbf{\Sigma}_{aj0}^{-1})^{-1}$, $\boldsymbol{\mu}_{aj} = \mathbf{\Sigma}_{aj}(\mathbf{\Sigma}_{aj0}^{-1}\mathbf{A}_{j0} + \psi_j^{-1}\mathbf{C}\mathbf{y}_j^*)$, $\mathbf{\Sigma}_{\lambda j} = (\mathbf{\Omega}\mathbf{\Omega}^T + \mathbf{\Sigma}_{j0}^{-1})^{-1}$, $\boldsymbol{\mu}_{\lambda j} = \mathbf{\Sigma}_{\lambda j}(\mathbf{\Sigma}_{j0}^{-1}\mathbf{\Lambda}_{j0} + \mathbf{\Omega}\tilde{\mathbf{y}}_j)$, $\beta_{\lambda j} = \beta_{j0} + (\tilde{\mathbf{y}}_j^T\tilde{\mathbf{y}}_j - \boldsymbol{\mu}_{\lambda j}^T\mathbf{\Sigma}_{\lambda j}^{-1}\boldsymbol{\mu}_{\lambda j} + \mathbf{\Lambda}_{j0}^T\mathbf{\Sigma}_{j0}^{-1}\mathbf{\Lambda}_{j0})/2$, $\beta_\delta = \beta_{\delta 0} + \sum_{i=1}^{n}\{\eta_i - \sum_{d=1}^{D}\sum_{k=1}^{K_{bd}} b_{dk}B_{dk}^x(x_{id}) - \sum_{j=1}^{q_2}\sum_{k=1}^{K_j}\beta_{jk}N_{jk}(\xi_{ij}|s_j)\}^2/2$, $IG[\ ]$ and $IW[\ ]$ denote the inverted gamma distribution and inverted Wishart distribution, respectively.

(C) Full conditional distributions of $\boldsymbol{\tau}_b$, $\boldsymbol{\tau}_\beta$ and $\boldsymbol{\tau}_s$ in step (b5).

It can be shown that the full conditional distributions of $\boldsymbol{\tau}_b$, $\boldsymbol{\tau}_\beta$, and $\boldsymbol{\tau}_s$ are:

$$\begin{aligned}
&p(\tau_{\beta j}|\cdot) \sim IG[\alpha_{\beta 0} + (K_j - m)/2,\ \beta_{\beta 0} + \boldsymbol{\beta}_j^T\mathbf{M}_{\beta j}\boldsymbol{\beta}_j/2],\\
&p(\tau_{bd}|\cdot) \sim IG[\alpha_{b0} + (K_{bd} - m)/2,\ \beta_{b0} + \mathbf{b}_d^T\mathbf{M}_{bd}\mathbf{b}_d/2],\\
&p(\tau_{sj}|\cdot) \sim IG(\alpha_{\tau 0} + K_j/2,\ \beta_{\tau 0} + \sum_{k=1}^{K_j}\{\ln(|s_j\kappa_{jk}|)\}^2/2),
\end{aligned} \tag{A.4}$$

for $j = 1, \ldots, q_2$, $d = 1, \ldots, D$.

(D) Full conditional distribution of $\mathbf{b}_d$ in step (b6).

The full distribution of $\mathbf{b}_d$ is

$$N(\mathbf{b}_d^*, \mathbf{\Sigma}_{bd}^*)I(\mathbf{Q}_{bd}\mathbf{b}_d = 0), \tag{A.5}$$

where $\mathbf{\Sigma}_{bd}^* = (\mathbf{B}_{bd}^T\mathbf{B}_{bd}/\psi_\delta + \mathbf{M}_{bd}/\tau_{bd})^{-1}$, $\mathbf{b}_d^* = \mathbf{\Sigma}_{bd}^*\mathbf{B}_{bd}^T\boldsymbol{\eta}_x^*/\psi_\delta$, and $\boldsymbol{\eta}_x^* = (\eta_{x1}^*, \ldots, \eta_{xn}^*)^T$ with

$$\eta_{xi}^* = \eta_i - \sum_{l\neq d}\sum_{k=1}^{K_{bl}} b_{lk}B_{lk}^x(x_{il}) - \sum_{j=1}^{q_2}\sum_{k=1}^{K_j}\beta_{jk}N_{jk}(\xi_{ij}|s_j).$$

According to the discussion in Section 2.4.1, sampling an observation $\mathbf{b}_d$ from $N(\mathbf{b}_d^*, \mathbf{\Sigma}_{bd}^*)I(\mathbf{Q}_{bd}\mathbf{b}_d = 0)$ is equivalent to sampling an observation $\mathbf{b}_d^{(\text{new})}$ from $N(\mathbf{b}_d^*, \mathbf{\Sigma}_{bd}^*)$, then $\mathbf{b}_d^{(\text{new})}$ is transformed to $\mathbf{b}_d$ by

$$\mathbf{b}_d = \mathbf{b}_d^{(\text{new})} - \mathbf{\Sigma}_{bd}^* \mathbf{Q}_{bd}^T (\mathbf{Q}_{bd} \mathbf{\Sigma}_{bd}^* \mathbf{Q}_{bd}^T)^{-1} \mathbf{Q}_{bd} \mathbf{b}_d^{(\text{new})}.$$

(E) Full conditional distribution of $(\boldsymbol{\beta}_j, s_j)$ in step (b7).

For $j = 1, \cdots, q_2$, let $\boldsymbol{\beta}_{-j}$ and $\mathbf{s}_{-j}$ denote subvectors of $\boldsymbol{\beta}$ and $\mathbf{s}$ that exclude $\boldsymbol{\beta}_j$ and s_j, respectively. The conditional distribution of $(\boldsymbol{\beta}_j, s_j)$ is

$$\begin{aligned}
& p(\boldsymbol{\beta}_j, s_j | \mathbf{\Omega}, \psi_\delta, \tau_{\beta j}, \tau_{sj}, \mathbf{b}, \boldsymbol{\beta}_{-j}, \mathbf{s}_{-j}) \\
\propto\ & p(\boldsymbol{\beta}_j | \mathbf{\Omega}, \psi_\delta, \tau_{\beta j}, \mathbf{b}, \boldsymbol{\beta}_{-j}, \mathbf{s}) p(s_j | \tau_{sj}) \\
\propto\ & \exp\Bigg[- \frac{1}{2\psi_\delta} \sum_{i=1}^{n} \{\eta_i^* - \sum_{k=1}^{K_j} \beta_{jk} N_{jk}(\xi_{ij} | s_j)\}^2 - \frac{1}{2\tau_{\beta j}} \boldsymbol{\beta}_j^T \mathbf{M}_{\beta j} \boldsymbol{\beta}_j - \\
& \frac{1}{2\tau_{sj}} \sum_{k=1}^{K_j} \{ln(|s_j \kappa_k|)\}^2 \Bigg]. \qquad \text{(A.6)}
\end{aligned}$$

The reason for updating $(\boldsymbol{\beta}_j, s_j)$ in a block is given below. It is noticed from $f_j(\xi_{ij}) = \sum_{k=1}^{K_j} \beta_{jk}\ N_{jk}(\xi_{ij} | s_j)$ that the highly correlated $\boldsymbol{\beta}_j$ and s_j are used to model the unknown smooth function f_j together. Updating them within a block will accelerate MCMC convergence.

As the conditional distribution in (A.6) is nonstandard, the MH algorithm is used to simulate observations from it. Inspired by Rue (2004), given current $s_j^{(t-1)}$, we first generate a new $s_j^{(t)}$ using the random walk defined by $s_j^{(t)} = f * s_j^{(t-1)}$, with $p(f) \propto 1 + 1/f$, $f \in [1/C, C]$ $(C > 1)$, in which C is a tuning parameter. It can be shown that

$$q(s_j^{(t)} | s_j^{(t-1)}) \propto \frac{1}{s_j^{(t)}} + \frac{1}{s_j^{(t-1)}}\ . \qquad \text{(A.7)}$$

The random walk defined in (A.7) is symmetric in that $q(s_j^{(t)}|s_j^{(t-1)}) = q(s_j^{(t-1)}|s_j^{(t)})$. With this proposal density, a new sample $\{\boldsymbol{\beta}_j^{(t)}, s_j^{(t)}\}$ is accepted with probability $\min\{1, R_j\}$, and

$$R_j = \frac{p(s_j^{(t)}|\cdot)}{p(s_j^{(t-1)}|\cdot)},$$

where $p(s_j^{(t)}|\cdot)$ is the marginal distribution of full conditional distribution $p(\boldsymbol{\beta}_j^{(t)}, s_j^{(t)}|\cdot)$, which can be derived analytically as follows:

$$\begin{aligned}
& p(s_j|\cdot) \\
\propto\ & \exp[-\frac{1}{2\tau_{sj}}\sum_{k=1}^{K_j}\{ln(|s_j\kappa_k|)\}^2] \times \\
& \int \exp\left[-\frac{1}{2}\left\{\boldsymbol{\beta}_j^T(\frac{1}{\psi_\delta}\mathbf{N}_j^T\mathbf{N}_j + \frac{1}{\tau_{\beta j}}\mathbf{M}_{\beta j})\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j^T(\frac{\mathbf{N}_j^T\boldsymbol{\eta}^*}{\psi_\delta})\right\}\right] d\boldsymbol{\beta}_j \\
\propto\ & |\boldsymbol{\Sigma}_j^*|^{1/2}\exp\left[\frac{1}{2}\boldsymbol{\beta}_j^{*T}\boldsymbol{\Sigma}_j^{*-1}\boldsymbol{\beta}_j^* - \frac{1}{2\tau_{sj}}\sum_{k=1}^{K_j}\{ln(|s_j\kappa_k|)\}^2\right]. \qquad \text{(A.8)}
\end{aligned}$$

Since R_j depends only on $s_j^{(t)}$ and $s_j^{(t-1)}$ and not on $\boldsymbol{\beta}_j$, a new observation of $\boldsymbol{\beta}_j$ is generated from $N(\boldsymbol{\beta}_j^*, \boldsymbol{\Sigma}_j^*)I(\mathbf{Q}_j\boldsymbol{\beta}_j = 0)$ when $s_j^{(t)}$ is accepted. As discussed in Section 2.4.1, sampling an observation $\boldsymbol{\beta}_j$ from $N(\boldsymbol{\beta}_j^*, \boldsymbol{\Sigma}_j^*)I(\mathbf{Q}_j\boldsymbol{\beta}_j = 0)$ is equivalent to sampling an observation $\boldsymbol{\beta}_j^{(\text{new})}$ from $N(\boldsymbol{\beta}_j^*, \boldsymbol{\Sigma}_j^*)$, then $\boldsymbol{\beta}_j^{(\text{new})}$ is transformed to $\boldsymbol{\beta}_j$ by

$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(\text{new})} - \boldsymbol{\Sigma}_j^*\mathbf{Q}_j^T(\mathbf{Q}_j\boldsymbol{\Sigma}_j^*\mathbf{Q}_j^T)^{-1}\mathbf{Q}_j\boldsymbol{\beta}_j^{(\text{new})}.$$

Appendix B

A Description of the Polydrug Use Data

y_1 (Drgplm30): Drug problems in past 30 days at intake, which ranges from 0 to 30.

y_2 (Drgday30): Drug use in past 30 days at intake, which ranges from 0 to 30.

y_3 (DrgN30): The number of kinds of drugs used in past 30 days at intake, which ranges from 1 to 8.

y_4 (Incar): The number of incarcerations in lifetime at intake, which ranges from 0 to 216.

y_5 (ArrN): The number of arrests in lifetime at intake, which ranges from 1 to 115.

y_6 (Agefirstarrest): The age of first arrest, which ranges from 6 to 57.

y_7 (Retent): Days of stay in treatment or retention, which ranges from 0 to 365.

y_8 (M12drg30): Primary drug use in past 30 days at 12 month interview, which ranges from 1 to 5.

x1 (Servicem): Services received in past 3 months at TSI 3 month interview.

x2 (DrugtestTX): The number of drug tests by TX in past 3 months at TSI 3 month inter- view, which ranges from 0 to 36.

x3 (DrugtestCJ): The number of drug tests by criminal justice in past 3 months at TSI 3 month interview, which ranges from 0 to 12.

Bibliography

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika*, 92, 419–434.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97, 160–169.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296–311.

Biller, C. and Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, 1, 195–211.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Box, G. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211–246.

Brezger, A. and Steiner, W. (2008). Monotonic regression based on Bayesian P-splines: an application to estimating price response functions from store-level scanner data. *Journal of Bussiness & Economic Statistics*, 26, 90–104.

Cai, J. H., Song, X. Y., and Hser, Y. I. (2010). A Bayesian analysis of mixture structural equation models with nonignorable missing responses and covariates. *Statistics in Medicine*, to appear.

Cai, J. H. and Song, X. Y. (2010). Bayesian analysis of mixtures in structural equation models with non-ignorable missing data. *The British Journal of Mathematical and Statistical Psychology*, to appear.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1, 651–674.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and Inferential Difficulties with Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95, 957–970.

Chen, M. and Schmeiser, B. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *Journal of Computational and Graphical Statistics*, 2, 251–272.

Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation.* New York: Springer-Verlag.

Chiang, C. T, Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96, 605–619.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6, 11–20.

David, M. R. (1993). On the beta transformation family. *Technometrics*, 35, 72–81.

De Boor, C. (2001). *A practical guide to splines.* New York: Springer-Verlag, revised edition.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56, 363–375.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve fitting with free-knot splines. *Biometrika*, 88, 1055–1071.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, 62, 355–366.

Dunson, D. B., Chulada, P., and Arbes Jr, S. J. (2003). Bayesian modeling of time-varying and waning exposure effects. *Biometrics*, 59, 83–91.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalities. *Statistical Science*, 11, 89–121.

Eubank, R. L., Huang, C., Maldonado, Y. M., Wang, N., Wang, S., and Buchanan, R. J. (2004). Smoothing spline estimation in varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 66, 653–667.

Evans, E., Li, L., and Hser, Y. I. (2009). Client and program factors associated with dropout from court mandated drug treatment. *Evaluation and Program Planning*, 32, 204–212.

Fahrmeir, L. and Raach, A. (2007). A Bayesian semiparametric latent variable model for mixed responses. *Psychometrika*, 72, 327–346.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 27, 1491–1518.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, 1, 179–195.

Foster, A. M., Tian, L., and Wei, L. J. (2001). Estimation for the Box-Cox transformation model without assuming parametric error. *Journal of the American Statistical Association*, 96, 1097–1101.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96, 194–209.

Gelfand, A. E., Kim, H. J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98, 387–396.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysisand Machine Intelligence*, 6, 721–741.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* London: Chapman and Hall.

Haneuse, S. J., Rudser, K. D., and Gillen, D. L. (2008). The separation of time scales in Bayesian survival modeling of the time-varying effect of a time-dependent exposure. *Biostatistics*, 9, 400–410.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* London: Chapman and Hall.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55, 757–796.

Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their application. *Biometrika*, 57, 97–100.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13, 795–800.

He, X. and Shen, L. (1997). Linear regression after spline transformation. *Biometrika*, 84, 474–481.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, 101, 1065–1075.

Hoeting, J. A., Raftery, A. E., and Madigen, D. (2002). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics*, 11, 485–507.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85, 809–822.

Hser, Y. I., Huang, D., Chou, C. P., and Anglin, M. D. (2007). Trajectories of heroin addiction: growth mixture modeling results based on a 33-year follow-up study. *Evaluation Review*, 31, 548–563.

Jaccard, J. and Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological bulletin*, 117, 348–357.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20, 50–67.

John, J. A. and Draper, N. R. (1980). An alternative family of transformations. *Journal of the Royal Statistical Society, Series C*, 29, 190–197.

Jöreskog, K. G. and Sörbom, D. (1996). *LISREL 8 user's reference guide*. United States: Scientific Software.

Kenny, D. A. and Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201–210.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183–212.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach*. United Kingdom, Chichester: Wiley.

Lee, S. Y. and Song, X. Y. (2003a). Model comparison of nonlinear structural equation models with fixed covariates. *Psychometrika*, 68, 27–47.

Lee, S. Y. and Song, X. Y. (2003b). Maximum likelihood estimation and model comparison for mixtures of structural equation models with ignorable missing data. *Journal of Classification*, 20, 221–255.

Lee, S. Y. and Song, X. Y. (2003c). Bayesian model selection for mixtures of structural equation models with an unknown number of components. *British Journal of Mathematical and Statistical Psychology*, 56, 145–165.

Lee, S. Y. and Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39, 653–686.

Lee, S.Y. and Zhu, H.T. (2002). Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika*, 67, 189–210.

Lindsay, B. G. and Basak, P. (1993). Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association*, 88, 468–476.

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

Lipsitz, S. R., Ibrahim, J., and Molenberghs, G. (2000). Using a Box-Cox transformation in analysis of longitudinal data with incomplete responses. *Journal of the Royal Statistical Society, Series C*, 49, 287–296.

Liu, J. S., Liang, F., and Wong, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95, 121–133.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing.* New York: Springer-Verlag.

Mallick, B. K. and Gelfand, A. (1994). Generalized linear models with unknown link functions. *Biometrika*, 81, 237–245.

Mallick, B. K. and Gelfand, A. (1996). Semiparametric errors-in-variables models A Bayesian approach. *Journal of Statistical Planning and Inference*, 52, 307–321.

Mallick, B. K. and Walker, S. (2003). A Bayesian semiparametric transformation model incorporating frailties. *Journal of Statistical Planning and Inference*, 112, 159–174.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models.* United States: Wiley-Interscience.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21, 1087–1091.

Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, 56, 337–357.

Muggeo, V. M. R. and Ferrara, G. (2008). Fitting generalized linear models with unspecified link function: A P-spline approach. *Computational Statistics and Data Analysis*, 52, 2529–2537.

Nychka, D. and Ruppert, D. (1995). Nonparametric transformations for both sides of a regression model. *Journal of the Royal Statistical Society, Series B*, 57, 519–532.

Panagiotelis, A. and Smith, M. (2008). Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. *Journal of Econometrics*, 143, 291–316.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. *Biometrika*, 68, 35–43.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. New York: Springer-Verlag.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3, 425–461.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the American Statistical Association*, 60, 365–375.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26, 195–239.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* , 59, 731–792.

Rodrigues, A. and Assunção, R. (2008). Propriety of posterior in Bayesian space varying parameter models with normal data. *Statistics & Probability Letters*, 78, 2408–2411.

Rue, H. (2004). *Gaussian Markov Random Fields*. Boca Raton: Chapman and Hall/CRC.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression.* United Kingdom: Cambridge University Press.

Scheines, R., Hoijtink, H., and Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64, 37–52.

Schumacker, R. E. and Marcoulides, G. A. (Eds) (1998). *Interaction and nonlinear effects in structural equation modeling.* New Jersey: Lawrence Erlbaum Associates, Inc.

Searle, S. R. (1971). *Linear models.* New York: Wiley.

Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear regression.* New Jersey, Hoboken: John Wiley & Sons.

Shi, J. Q. and Lee, S. Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, Series B*, 62, 77–87.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317–343.

Song, X. Y. and Lee, S. Y. (2002). A Bayesian approach for multigroup nonlinear factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 9, 523–553.

Song, X. Y. and Lee, S. Y. (2005). A multivariate probit latent variable model for analyzing dichotomous responses. *Statistica Sinica*, 15, 645–664.

Song, X. Y. and Lu, Z. H. (2010). Semiparametric Latent Variable Models With Bayesian P-splines . *Journal of Computational and Graphical Statistics* , to appear.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society, Series B*, 64, 583–639.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS User Manual 1.4*. MRC Biostatistics Unit, Cambridge, England.

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

Su, L. and Hogan, J. W. (2010). Varying-coefficient models for longitudinal processes with continuous-time informative dropout. *Biostatistics*, 11, 93–110.

Telesca, D. and Inoue, L. Y. T. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association*, 103, 328–339.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, 83, 394–405.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: John Wiley & Sons.

Wang, N. and Ruppert, D. (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. *Journal of the American Statistical Association*, 90, 522–534.

Yin, G. and Ibrahim, J. (2006). Bayesian transformation hazard models. In J. Rojo, ed., *Optimality: The Second Erich L.*

Lehmann Symposium. IMS Lecture Notes-Monograph Series Volume 49, 170–182. Beachwood, OH: Institute of Mathematical Statistics.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62, 297–330.