BAYESIAN VARYING-COEFFICIENT MODEL WITH MISSING DATA

HUANG ZHIPENG

NATIONAL UNIVERSITY OF SINGAPORE

2013

BAYESIAN VARYING-COEFFICIENT MODEL WITH MISSING DATA

HUANG ZHIPENG

(B.Sc. University of Science and Technology of China)

SUPERVISED BY

A/P LI JIALIANG & A/P DAVID JOHN NOTT

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2013

ACKNOWLEDGEMENTS

I am so grateful that I have Associate Professor Li Jia-Liang as my supervisor and Associate Professor David John Nott as my co-supervisor. They are truly great mentors in statistics. I would like to thank them for their guidance, encouragement, time, and endless patience. Next, I would like to thank Dr. Feng Lei for his help in my real data analysis. I also thank all my friends who helped me to make life easier as a graduate student. I wish to express my gratitude to the university and the department for supporting me through NUS Graduate Research Scholarship. Finally, I will thank my family for their love and support.

CONTENTS

SUMMARY

Motivated by Singapore Longitudinal Aging Study (SLAS), we propose a Bayesian approach for the estimation of semiparametric varying-coefficient models for longitudinal normal and cross-sectional binary responses. These models have proved to be more flexible than simple parametric regression models, and our Bayesian solution eases the computation complexity of these models. We also consider adapting all kinds of familiar statistical strategies to address the missing data issue in SLAS. Our simulation results indicate that Bayesian imputation approach performs better than complete-case and available-case approaches, especially under small sample designs, and may provide more useful results in practice. In the real data analysis for SLAS, the results from Bayesian imputation are similar to available-case analysis, differing from those with complete-case analysis.

LIST Of NOTATIONS

$\mathbf{1}_n$	$n \times 1$ vector with all elements equal to 1
$\mathbb{R}^p$	p-dimensional Euclidean space
M^T	transpose of a matrix or vector M
x_+	maximum of x and 0, where $x \in \mathbb{R}$
$\mathbf{I}_n$	n-dimensional identity matrix
$I(\cdot)$	indicator function
$g^{-1}(\cdot)$	inverse of function $g(\cdot)$
$min(a, b)$	minimum of a and b, where $a, b \in \mathbb{R}$
$diag(A, B)$	block diagonal matrix, where A, B are square matrices

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Review of nonparametric methods & Bayesian inference

Currently, nonparametric approaches are used more and more besides parametric approaches in fitting regression model. The most popular parametric methods of inference, for example estimation, hypotheses testing and confidence interval, are based on Fisher's maximum likelihood (e.g. Aldrich (1997)). The maximum

likelihood estimation usually achieves the optimal efficiency of estimation as described by its variance property. However, if the specified parametric model is wrong or far away from the true model, the results of parametric estimation can be very misleading. On the other hand, nonparametric models make only basic assumptions, such as independence among the observations and finity of the variance of the data, or existence of r-th derivative of the density function $f(x)$ of the data, where r is a positive integer and the form of $f(x)$ is never specifically assumed. Thus nonparametric approaches achieve more widely applicable and stable results and the models are robust. From the view point of nonparametric, all parametric models are too rigid. Besides, there are situations when a workable parametric model is hard to establish, for instance, in biased sampling.

Nonparametric methods can be classified as classical nonparametric methods which are based on signs and ranks developed in 1940s $\sim$ 1970s and modern nonparametric methods which involve (i) smoothing methods and (ii) the jackknife, Bootstrap (e.g. Efron and Gong (1983) & Shao and Tu (1995)) and other re-sampling methods. These methods are called modern because they were developed after the wide spread of modern computer power. Smoothing methods contain kernel smoothing, regression splines, smoothing splines, penalty splines and others. Regression splines is an important smoothing method which uses basis technique to approximate the curves or functions to be estimated and the truncated power basis is a commonly used regression spline basis. By using quadratic

or cubic or even higher order truncated power basis, the nonparametric curves or functions to be estimated can be approximated by parametric model. Then parametric approaches can be employed for the estimation.

Varying-coefficient partial linear model is a mixture of parametric linear model and nonparametric linear model as part of the coefficients are parametric and part of the coefficients are nonparametric, say varying-coefficients. Because of this, it is referred to as semi-parametric linear regression model. These varying-coefficients can be approximated using regression splines described above and thus the semi-parametric model is approximated by parametric model.

Over the past 30 years there has been a great deal of interest and activity in the general area of nonparametric smoothing in statistics. Different kinds of smoothing methods are proposed, such as kernel smoothing which contains Nadaraya Watson estimator, local linear regression, local polynomial smoothing and others, regression splines, e.g. Eubank (1999) & Wu and Zhang (2006), smoothing splines, e.g. Green et al. (1994) & Wu and Zhang (2006) and penalized splines, e.g. Eilers and Marx (1996), Hastie (1996), Lang and Brezger (2004) & Wu and Zhang (2006). This area is developing rapidly, but more future works are still needed because all the proposals mentioned above have their limitations though they are all suitable for some particular cases. For example, Hastie (1996) described a method for constructing a family of low rank penalized scatter-plot smoothers, the so called

pseudosplines which had a shrinking behavior that was similar to that of smoothing splines; however, if too small a rank was chosen, the family of pseudo-splines would be limited to fits of total rank which may be insufficient.

As the fast development of nonparametric smoothing, a special form of nonparametric model is explored. Hastie and Tibshirani (1993) first explored varying-coefficient models: a class of regression and generalized regression models in which the coefficients are allowed to vary as smooth functions of other variables. Subsequently, this topic has become more and more popular, e.g. Fan et al. (2003), Eubank et al. (2004), Wang et al. (2008) & Lu et al. (2009). Besides, the so called varying-coefficient partial linear model has also been explored since then. This model is a mixture of parametric linear model and nonparametric linear model as part of the coefficients are parametric and part of the coefficients are nonparametric, say varying-coefficients. It is referred to as semi-parametric linear regression model because of this. The estimation of semi-parametric linear regression models are studied intensively, e.g. Lin and Carroll (2001), Ruppert et al. (2003), Li and Wong (2009), Li and Palta (2009) & Li et al. (2009).

In parametric inference, parameters can be considered as some fixed unknown values to be estimated, which are typical frequentist inferences. However, from the view of Bayesian inference, parameters are random variables which have distributions. The purpose of inference is to calculate and interpret the conditional

posterior distributions of the parameters given the observed data. Thus, for inference about the statistical models, statisticians can be divided into two schools: frequentist and Bayesian. In the following review, we will focus on the Bayesian inference.

Bayesian inference has developed rapidly and been more and more popular in recent decades due to the rapid development of modern computer power. It is competent for many relatively complicated models which are hard to treat from the view of frequentist inference. An overview of Bayesian inference can be found in any Bayesian textbook, e.g. Gelman et al. (2004). One of the important components of Bayesian simulation is the selection of the prior. If the prior is conjugate, then the simulation usually will be simplified. For variance parameters, *inverse gamma* distribution is commonly chosen as the prior as it is usually conjugate, e.g. Ruppert et al. (2003). Gelman (2006) constructed a new folded-non-central-t family of conditionally conjugate priors for hierarchical standard deviation parameters and considered non-informative and weakly informative priors in this family. His proposal increases the choice of prior selection and overcomes the serious problems that might occur when the commonly used inverse-gamma prior for variance parameters is used. Other important concerns about Bayesian inference are the outcome and convergence of the Monte Carle simulation. The commonly used

Bayesian simulation algorithms, e.g. the Gibbs sampler, the algorithm of Metropolis and similar iterative simulation methods, are potentially very helpful for summarizing multivariate distributions. Used naively, however, iterative simulation can give misleading answers. Based on this, Gelman and Rubin (1992) recommended using several independent sequences of interative simulation for Bayesian posterior distributions, with starting points sampled from an over-dispersed distribution. Besides, Brooks and Gelman (1998) generalized the method proposed by Gelman and Rubin (1992) for monitoring the convergence of iterative simulations by comparing between and within variances of multiple chains, in order to obtain a family of tests for convergence. However, as the authors pointed out, although multiple-chain-based diagnostics are safer than single-chain-based diagnostics, they can still be highly dependent upon the starting points of the simulations. When employing Bayesian method for estimation of generalized regression model, a problem usually occurs that the posteriors of the concerned parameters are non-conjugated which makes the Bayesian simulation complicated. The problem was partially solved when Holmes and Held (2006) proposed using Bayesian auxiliary variable Models for binary and multinomial regression. Their approaches were ideally suited to automated Markov chain Monte Carlo simulation as the algorithms they proposed are fully automatic with no-user set parameters and no necessary Metropolis-Hastings accept/reject steps which might cause the simulation converge slowly when the reject rate is high. However, as the number of

parameters increases, it may be too time-consuming.

Bayesian treatment of semiparametric and nonparametric regression models has developed rapidly in recent decades, e.g. Biller and Fahrmeir (2001), Fahrmeir et al. (2004), Lambert and Eilers (2005), Brezger and Lang (2006), Wang et al. (2013). Among them, Biller and Fahrmeir (2001) proposed Bayesian varying-coefficient models using adaptive regression splines. They presented a full Bayesian B-spline basis function approach with adaptive knot selection, and used reversible jump Markov chain Monte Carlo sampling to estimate the number and location of knots and B-spline coefficients for each of the unknown regression functions. However, as the authors pointed out, they didn't consider the situation involving random effects for longitudinal data or missing data.

1.2 Review of longitudinal data & missing data

Longitudinal data study has grown tremendously over the past two decades, especially in the clinical trials. Varying-coefficient models can be employed to analyze longitudinal data by adding random effects to the models. The models are particularly appealing in longitudinal studies as they allow us to inspect the extent to which covariates affect responses over time, e.g. Hoover et al. (1998)

& Fan and Zhang (2000). Besides, when carrying out longitudinal analysis where subjects are repeatedly measured over time, it is highly possible that some of the measurements are missing. For example, in a clinical trial, the patients are supposed to take several times of scheduled medical tests over a special period of time; however, some of them may quit in midway after the first several tests, and some of them may lose contact for some time then appear again, etc. Thus it is necessary to deal with the missing values, especially when the missing rate is considerable. Fortunately, the statistical analysis of data with missing values has flourished since the early 1970s, spurred by advances in computer technology that made previously laborious numerical calculations a simple matter (Little and Rubin (2002)). Since then, various methodologies and algorithms were proposed for handling missing data problems, such as *Weighting Procedures*, *Imputation-Based Procedures* etc.

There are several kinds of missing-data patterns. According to Little and Rubin (2002), there are mainly three types of missing data mechanisms with respect to how the missing values are related to the observed values: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) and *Non-Missing at Random* (NMAR). If subjects who have missing data are a random subset of the complete sample of subjects, missing data are called MCAR (Rubin (1976)). Under this condition, most simple techniques for handling missing data, including complete case and available case analysis, will give unbiased results (Greenland and Finkle

(1995)). If the probability that an observation is missing depends on information that is not observed, such as the value of the observation itself, missing data are called NMAR (Rubin (1976)). In this case, valuable information is lost from the data and there is no universal method of handling the missing data properly, (e.g. Greenland and Finkle (1995), Little (1992), Rubin (1976) & Rubin (2009)). Mostly, missing data are neither MCAR nor NMAR (Booth (2000)). Instead, the probability that an observation is missing commonly depends on information for that subject that is present, i.e., reason for missingness is based on other observed variables, in other words, the probability that an individual value is missing depends only on the observed variables but not on the missing ones. This type of missing data is called MAR, because missing data can indeed be considered random conditional on these other observed variables that determined their missingness (Rubin (1976)). Under MAR, a complete case or available case analysis is no longer based on a random sample from the source population and selection bias likely occurs. Generally, when missing data are MAR, all simple techniques for handling missing data, i.e. complete case and available case analysis and overall mean imputation, give biased results. However, more sophisticated techniques like single and multiple imputations give unbiased results when missing data are MAR, (e.g. Greenland and Finkle (1995), Little (1992), Rubin (1976) & Rubin (2009)).

Besides, according to Little and Rubin (2002), methods on the analysis of partially missing data can be grouped into the following four categories, which are

not mutually exclusive: *Procedures Based on Completely Recorded Units*, *Weighting Procedures*, *Imputation-Based Procedures* and *Model-Based Procedures*. In our research, we will focus on Imputation-Based Procedures, which means that the missing values are filled in and the resultantly completed data are analyzed by standard methods. For valid inferences to result, modifications to the standard analyzes are required to allow for the differing status of the real and the imputed values.

Imputations are means or draws from a predictive distribution of the missing values which require a method of creating a predictive distribution for the imputation based on the observed data. There are two generic approaches to generating this distribution: *Explicit modeling* and *Implicit modeling*. In this study, we will focus on Explicit modeling, that is the predictive distribution is based on a formal statistical model (e.g. normal), hence the assumptions are explicit. It include *mean imputation*, *regression imputation*, *stochastic regression imputation* and *Bayesian imputation* (*Data augmentation*, Tanner and Wong (1987)) among others.

Regression imputation replaces missing values by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present. Stochastic regression imputation replaces missing values by values predicted by regression imputation plus residuals, drawn to reflect uncertainty in the predicted values. With normal linear regression models, the residual will naturally be normal with zero mean and

variance equal to the residual variance in the regression. With a binary outcome, as in logistic regression, the predicted value is a probability of 1 versus 0, thus the imputed valued is a 1 or 0 drawn with that probability.

To describe Bayesian imputation, we assume Y^{obs} is the vector of observed responses, Y^{mis} is the vector of missing responses and θ is the vector of parameters to be estimated; besides, we assume the predictors are all observed. Bayesian imputation method treats Y^{mis} as a vector of variables and contains two step: the imputation step and the proposal step. Roughly speaking, in the imputation step, we draw a sample of Y^{mis} from the conditional density of Y^{mis} given Y^{obs} and θ; in the proposal step, we draw a sample of θ from the conditional density of θ given Y^{obs} and Y^{mis}. The details will be given in the main body of this thesis when we come to it.

If the estimated distribution results based on the observed subjects in the study sample would be identical to the 'true' underlying distribution in the population, the single imputation procedure would be equivalent to direct replacement of the true values of the missing data. However, this will seldom be the case, but the estimated distribution can certainly be an unbiased estimate of the population distribution. Therefore, the associations under study estimated after missing data have been imputed by single imputation are unbiased. Doing so, however, one analyzes the completed data set as if all data were indeed observed. Because this

was not the case, the single imputation procedure commonly results in an underestimation of the standard errors, i.e. overestimation of the precision of the study associations, (e.g. Greenland and Finkle (1995), Rubin (2009),& Vach (1994)). Thus, we should take into account the imprecision caused by the fact that the distribution of the variables with missing values is estimated to obtain correct estimates of the standard errors. According to Rubin (2009) & Schafer (2010), this can be done by creating not a single imputed data set, but multiple imputed data sets in which different imputations are based on a random draw from different estimated underlying distribution, such as Bayesian imputation described above.

1.3 Focus of this thesis

Although frequentist and Baysian estimation procedures for semiparametric varying-coefficient model have been abundant in the literature, there is a relative lack of estimation procedures for this type of model involving longitudinal or missing data. This thesis is to implement a general Bayesian procedure to fit the semiparametric varying-coefficient model for cross-sectional normal response variable and binary response variable, and also for missing data which is more and

more popular and commonly occurs in practice now. Specially the nonparametric components are approximated with a functional basis expansion and Bayesian spline techniques are introduced to facilitate the computation (Lang and Brezger (2004); Nott and Li (2010)). This study will also consider fitting longitudinal normal data using varying-coefficient mixed model which adds random effect to varying-coefficient model. The results of this study may provide an alternative method for fitting varying-coefficient model, especially when the model involves binary response variable or missing data which is relatively complicated. This study may also provide an alternative method for fitting varying-coefficient mixed model using random effect for longitudinal data. For the situation of missing data, this thesis will only focus on the case when the response variable is longitudinal normal and simple binary; the case when the response variable is longitudinal binary will not be considered because it is too time-consuming for estimation. Besides, this thesis will concentrate on the case of MAR which is the most common case in reality. Moreover, in regression analysis, we assumed the predictors are all observed while only some of the responses are missing in our study although the case of missing data in covariates is also encountered often, e.g. White and Carlin (2010). Also, in the analysis of missing data in this thesis, we will ignore single imputation methods and implement Bayesian imputation methods and then compare the estimates with those got from complete case or available case analysis.

In Chapter 2, we will describe Bayesian estimation of varying-coefficient model

for normal response variable, with respect to cross-sectional data, longitudinal data and longitudinal data involving missing value. In Chapter 3, we will carry on similar processes for cross-sectional binary response variable and cross-sectional binary response variable involving missing value. Chapters 2, 3 will both contain the introduction of the model, fitting of the model followed by simulations to assess the performs of estimations respectively. In Chapter 4, we will apply the methodology described in previous chapters to analyze the real data from Singapore Longitudinal Aging Study (SLAS). Discussion and Conclusion will be provided in Chapter 5.

CHAPTER 2

Varying-coefficient model for normal response

2.1 Varying-coefficient model

2.1.1 Statistical model

Let $Y \in \mathbb{R}$ be a response variable and $\{U \in \mathbb{R}, X \in \mathbb{R}^p, Z \in \mathbb{R}^q\}$ be the covariates. The varying coefficient model assumes the following structure:

$$Y = \alpha^T(U)X + \beta^T Z + \epsilon, \tag{2.1}$$

where ϵ is normal and independent of (U, X, Z) with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$; $\beta = [\beta_1, \ldots, \beta_q]^T$ is a vector of q-dimensional fixed coefficient parameters for covariates Z and $\alpha(.) = [\alpha_1(.), \ldots, \alpha_p(.)]^T$ are unknown varying-coefficients for covariates X. All varying coefficients are assumed to be smooth functions with continuous second derivatives.

Suppose we have a random sample of size n, $\{(U_i, X_{i1}, \ldots, X_{ip}, Z_{i1}, \ldots, Z_{iq}, Y_i), i = 1, \ldots, n\}$ from model (2.1).

Firstly, to tackle the infinite dimensional functions $\alpha_k(u),\ k = 1, \ldots, p$, we consider using the cubic truncated power basis

$$\phi(u) = [1, u, u^2, u^3, (u - \tau_1)_+^3, \ldots, (u - \tau_K)_+^3]^T$$

for approximation and denote the corresponding coefficient vector to be

$$\gamma_k = [\gamma_{0k}, \gamma_{1k}, \gamma_{2k}, \gamma_{3k}, \gamma_{4k}, \ldots, \gamma_{K+3,k}]^T,\ k = 1, \ldots, p,$$

where $\tau_l,\ \ l = 1, \ldots, K$ are the knots and K is number of knots.

Under the basis expansion, Model (2.1) can be rewritten as

$$\begin{aligned} Y_i &= \sum_{k=1}^{p} X_{ik}\phi^T(U_i)\gamma_k + Z_i^T\beta + \epsilon_i, \\ &= C_i^T\gamma + Z_i^T\beta + \epsilon_i, \quad i = 1, \ldots, n, \end{aligned} \tag{2.2}$$

where $C_i = [X_{i1}\phi^T(U_i), \ldots, X_{ip}\phi^T(U_i)]^T,\ \gamma = [\gamma_1^T, \ldots, \gamma_p^T]^T,\ Z_i = [Z_{i1}, \ldots, Z_{iq}]^T$; We may further express the above model in a matrix form as

$$Y = C\gamma + Z\beta + \epsilon, \tag{2.3}$$

where $Y = [Y_1, \ldots, Y_n]^T$, $C = (C_1, \ldots, C_n)^T$, $Z = (Z_1, \ldots, Z_n)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$.

2.1.2 Bayesian inference

2.1.2.1 The prior

We consider a Baysian approach to fit Model (2.3). Thus we need to decide the distribution of Y given the parameters and specify prior distributions on model parameters.

Here we assume

$$Y|\gamma, \beta, C, Z \sim MN(C\gamma + Z\beta,\ \sigma_\epsilon^2 \mathbf{I}_n),$$

where $\mathbf{I}_n$ is the n-dimensional unit matrix and $MN(\mu,\ \sigma^2\mathbf{I}_n)$ is $multi - normal$ distribution with mean μ and variance matrix $\sigma^2\mathbf{I}_n$.

Following the idea of Ruppert et al. (2003), we assume the priors for $\beta,\ \gamma,\ \sigma_\epsilon^2$ independently as following which will achieve conditionally-conjugate prior:

let $\beta \sim MN(0,\ \sigma_\beta^2\mathbf{I}_q)$ with σ_β^2 so large that, for all intents and purpose, the normal distribution is uniform on the range of β.

Let $\gamma \sim MN(0,\ V)$ where $V = diag(V_1, \ldots, V_p)$ and $V_k = diag(\sigma_\gamma^2\mathbf{I}_4,\ \sigma_k^2\mathbf{I}_K),\ k = 1, \ldots, p$. Similarly to σ_β^2, σ_γ^2 are large enough to obtain noninformative prior. For

simplification, we set $\sigma_\gamma^2 = \sigma_\beta^2 = 10^8$ here.

For the purpose of shrinkage, we assume the priors on hyperpriors $\sigma_k^2,\ k = 1, \ldots, p$ are *inverse gamma* with parameters A_γ and B_γ–denoted $IG(A_\gamma, B_\gamma)$ independently. Thus its density is

$$\pi(\sigma_k^2) \quad = \quad \frac{B_\gamma^{A_\gamma}}{\Gamma(A_\gamma)}(\sigma_k^2)^{-(A_\gamma+1)} \exp(-\frac{B_\gamma}{\sigma_\gamma^2}), \tag{2.4}$$

where $\Gamma(\cdot)$ is the Gamma function. We may write this as $\sigma_k^2 \sim IG(A_\gamma,\ B_\gamma)$.

Further, we assume that the prior $\sigma_\epsilon^2 \sim IG(A_\epsilon,\ B_\epsilon)$. Here A_ϵ, B_ϵ, A_γ, B_γ are *hyperparameters* that determine the priors and must be chosen by us. These hyperparamaters must be strictly positive in order for the prior and hyperpriors to be proper. If $A_\epsilon > 1$ then the mean of this random variable is finite and equals to $B_\epsilon/(A_\epsilon - 1)$; if $A_\epsilon > 2$ then its variance is finite and equals to $B_\epsilon^2/((A_\epsilon - 1)^2(A_\epsilon - 2))$. If A_ϵ and B_ϵ were zero, then $\pi(\sigma_\epsilon^2)$ would be proportional to the improper prior $1/\sigma_\epsilon^2$, which is equivalent to $log(\sigma_\epsilon)$ having an improper uniform prior. Therefore, choosing A_ϵ and B_ϵ both close to zero (say, both equal to 0.001) gives an essentially noninformative, but proper prior. The same reasoning applies to A_γ and B_γ.

The model we have constructed is a hierarchical Bayes model, where the random variables are arranged in a hierarchy such that distributions at each level are determined by the random variables in the previous levels. At the bottom of the hierarchy are the known hyperparameters. At the next level are the fixed effects parameters and variance components whose distributions are determined by the

hyperparameters. At the level above this are β, γ and ϵ, whose distributions are determined by the variance components. The top level contains the data, Y.

2.1.2.2 The posterior

Denote the parameter space by Θ and parameter in Θ by θ where $\theta = [\gamma^T, \beta^T, \sigma_1^2, \ldots, \sigma_p^2, \sigma_\epsilon^2]^T$. Denote the prior distribution of θ by $\pi(\theta)$ and the conditional distribution of sample Y by $p(Y|\theta)$ where $Y = [Y_1, \ldots, Y_n]^T$. By Bayes Theorem we get the posterior of θ :

$$\pi(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{\int_\Theta p(Y|\theta)\pi(\theta)d\theta} \propto p(Y|\theta)\pi(\theta);$$

that is

$$\begin{aligned}
\pi(\theta|Y) &\propto p(Y|\theta)\pi(\gamma|\sigma_1^2, \ldots, \sigma_p^2)\pi(\sigma_1^2), \ldots, \pi(\sigma_p^2)\pi(\beta)\pi(\sigma_\epsilon^2) \\
&= (\frac{1}{\sqrt{2\pi}\sigma_\epsilon})^n \exp(-\frac{\|Y - C\gamma - Z\beta\|^2}{2\sigma_\epsilon^2}) \\
&\quad \times \frac{1}{(2\pi)^{p(K+4)/2}|V|^{1/2}} \exp(-\frac{1}{2}\gamma^T V^{-1}\gamma) \\
&\quad \times \prod_{k=1}^{p} \frac{B_\gamma^{A_\gamma}}{\Gamma(A_\gamma)} (\sigma_k^2)^{-(A_\gamma+1)} \exp(-\frac{B_\gamma}{\sigma_k^2}) \\
&\quad \times (\frac{1}{\sqrt{2\pi}\sigma_\beta})^q \exp(-\frac{\|\beta\|^2}{2\sigma_\beta^2}) \\
&\quad \times \frac{B_\epsilon^{A_\epsilon}}{\Gamma(A_\epsilon)} (\sigma_\epsilon^2)^{-(A_\epsilon+1)} \exp(-\frac{B_\epsilon}{\sigma_\epsilon^2}). \qquad (2.5)
\end{aligned}$$

If we isolate the part of (2.5) that depends on $(\gamma,\ \beta)$ then we see that the conditional posterior of $(\gamma,\ \beta)$ given $(\sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2)$–that is, the complete conditional–is proportional to

$$\exp(-\frac{\|Y-C\gamma-Z\beta\|^2}{2\sigma_\epsilon^2}-\frac{\gamma^TV^{-1}\gamma}{2}-\frac{\|\beta\|^2}{2\sigma_\beta^2}). \tag{2.6}$$

The term in parentheses in (2.6) is a nonnegative quadratic function of $(\gamma,\ \beta)$ and so (2.6) is proportional to a multivariate normal density. By the usual technique of "completing the square", it may be shown that

$$\pi(\gamma,\ \beta|\sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2,Y)\ \ \sim\ \ MN(\mu_{\gamma,\beta},\ \Sigma_{\gamma,\beta}). \tag{2.7}$$

Here $\mu_{\gamma,\beta}=\frac{1}{\sigma_\epsilon^2}(\frac{D^TD}{\sigma_\epsilon^2}+E^TE+\frac{F^TF}{\sigma_\beta^2})^{-1}D^TY$, $\Sigma_{\gamma,\beta}=(\frac{D^TD}{\sigma_\epsilon^2}+E^TE+\frac{F^TF}{\sigma_\beta^2})^{-1}$, $D=[C,Z]$, $E=[V^{-1/2},\ \mathbf{0}_{q\times q}]$ and $F=[\mathbf{0}_{p(K+4)\times p(K+4)},\ \mathbf{I}_q]$ where $\mathbf{0}_{q\times q}$ is $q\times q$-dimension zero matrix. The $p(K+4)\times p(K+4)$-dimension zero matrix corresponds to the p smooth unknown varying-coefficients, the cubic truncated power basis and the K knots. Thus, as part of the MCMC chain, one generates $(\gamma,\ \beta)$ from the current values of $(\sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2)$ according to the multivariate normal distribution, with mean and covariance matrix given by (2.7).

The complete conditional for $\sigma_k^2,\ k=1,\ldots,p$ is proportional to

$$(\sigma_k^2)^{-(K/2+A_\gamma+1)}\exp(-\frac{1}{\sigma_k^2}(\frac{1}{2}(\gamma_{4,k}^2+\ldots+\gamma_{(K+3),k}^2)+B_\gamma)). \tag{2.8}$$

Therefore, comparing (2.8) to (2.4) shows that

$$\pi(\sigma_k^2|Y,\gamma,\beta,\sigma_1^2,\ldots,\sigma_{k-1}^2,\sigma_{k+1}^2,\ldots,\sigma_p^2,\sigma_\epsilon^2)\sim IG(A_\gamma+\frac{1}{2}K,\ B_\gamma+\frac{1}{2}(\gamma_{4,k}^2+\ldots+\gamma_{K+3,k}^2)).$$

By the same reasoning,

$$\pi(\sigma_\epsilon^2|Y,\gamma,\beta,\sigma_1^2,\ldots,\sigma_p^2) \;\sim\; IG(A_\epsilon+\frac{1}{2}n,\; B_\epsilon+\frac{1}{2}\|Y-C\gamma-Z\beta\|^2). \quad (2.9)$$

2.1.2.3 Simulation algorithm

To sample from the posterior, we iterate S times through the following four steps.

Step 1 Sample $(\gamma,\ \beta)$ from the multivariate normal distribution: $MN(\mu_{\gamma,\beta},\ \Sigma_{\gamma,\beta})$;

Step 2 Sample $\sigma_k^2,\ k=1,\ldots,p$ from $IG(A_\gamma+\frac{1}{2}K,\ B_\gamma+\frac{1}{2}(\gamma_{4,k}^2+\ldots+\gamma_{K+3,k}^2))$ respectively;

Step 3 Sample σ_ϵ^2 from $IG(A_\epsilon+\frac{1}{2}n,\ B_\epsilon+\frac{1}{2}\|Y-C\gamma-Z\beta\|^2)$;

Step 4 Return to Step 1 and iterate until converge.

In Step 1, an alternative method to sampling is considered based on the following fact:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}) \Rightarrow (X_1|X_2=a) \sim N(\bar{\mu},\ \bar{\Sigma}),$$

where $\bar{\mu}=\mu_1+\Sigma_{21}\Sigma_{22}^{-1}(a-\mu_2)$ and $\bar{\Sigma}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Base on this fact, we may sample each component of $(\gamma,\ \beta)$ from univariate normal distribution conditional

on all the other components of $(\gamma,\ \beta)$ and on $\sigma_1^2, \ldots, \sigma_p^2, \sigma_\epsilon^2$.

In Step 2 and Step 3, notice that, given $(\gamma,\ \beta,\ Y)$, it follows that $\sigma_1^2, \ldots, \sigma_p^2$ and σ_ϵ^2 are mutually independent. Therefore, the net effect of Step 2 and Step 3 is to sample $(\sigma_1^2, \ldots, \sigma_p^2, \sigma_\epsilon^2)$ from its conditional distribution given $(\gamma,\ \beta,\ Y)$. Also, because of this independence, interchanging the order of Step 2 and Step 3 and the order in Step 2 inbetween $\sigma_1^2, \ldots, \sigma_p^2$ has no effect on the algorithm.

2.1.2.4 Convergence of MCMC simulation

Indeed, the above simulation process in Section 2.1.2.3 is a typical Gibbs sampler which is a particular Markov chain algorithm and can be viewed as a special case of the Metropolis-Hastings (M-H) algorithm.

Metropolis-Hastings algorithm is a MCMC methods based on random walk. The most important point of the M-H algorithm is the ratio of ratios blew:

$$r \quad = \quad \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}, \tag{2.10}$$

where $p(\theta|y)$ is the posterior, $J_t(\theta^*|\theta^{t-1})$ is the proposal distribution, θ^{t-1} is the simulation value of θ at time (step) $t-1$ and θ^* is a proposal from the proposal distribution at time (step) t. We accept $\theta^t = \theta^*$ with probability $min(r, 1)$ and $\theta^t = \theta^{t-1}$ otherwise.

Gibbs sampling can be viewed as M-H algorithm in the following way. We

first define iteration t to consist of a series of d steps, with step j of iteration t corresponding to an update of the subvector θ_j conditional on all the other elements of θ, that is θ_{-j}. Then the jumping distribution $J_{j,t}(\cdot|\cdot)$ at step j of iteration t only jumps along the jth subvector, and does so with the conditional posterior density of θ_j given θ_{-j}^{t-1}:

$$J_{j,t}^{Gibbs}(\theta^*|\theta^{t-1}) = \begin{cases} p(\theta_j^*|\theta_{-j}^{t-1}, y) & \text{if } \theta_{-j}^* = \theta_{-j}^{t-1} \\ 0 & \text{otherwise.} \end{cases}$$

The only possible jumps are to parameter vectors θ^* that match θ^{t-1} on all components other than the jth. Under this jumping distribution, the ratio at the jth step of iteration t can be proved to be r=1, thus every jump is accepted.

The proof that the simulation sequence of iterations from M-H algorithm converges to the target distribution contains two steps:

(1) The simulation sequence is Markov chain with a unique stationary distribution,

(2) The stationary distribution equals this target posterior distribution.

The first step of the proof holds if the Markov Chain is irreducible, aperiodic and not transient. Except for trivial exceptions, the latter two conditions hold for a random walk on any proper distribution, and irreducibility holds as long as the random walk has a positive probability of eventually reaching any state from any other state; that is, the jumping distribution J_t must eventually be able to jump

to all states with positive probability, which is satisfied in our simulation.

To see that the target distribution is the stationary distribution of the Markov chain generated by the M-H algorithm, consider starting the algorithm at time $t-1$ with a draw θ^{t-1} from the target distribution $p(\theta|y)$. Now consider any two such points θ_a and θ_b, drawn from $p(\theta|y)$ and labeled so that $p(\theta_b|y)J_t(\theta_a|\theta_b) \geq p(\theta_a|y)J_t(\theta_b|\theta_a)$. The unconditional probability density of a transition from θ_a to θ_b is

$$p(\theta^{t-1} = \theta_a,\ \theta^t = \theta_b) = p(\theta_a|y)J_t(\theta_b|\theta_a),$$

where the acceptance probability is 1 because of our labeling of a and b, and the unconditional probability density of a transition from θ_b to θ_a is, from (2.10),

$$\begin{aligned} p(\theta^{t-1} = \theta_b,\ \theta^t = \theta_a) &= p(\theta_b|y)J_t(\theta_a|\theta_b)\frac{p(\theta_a|y)/J_t(\theta_a|\theta_b)}{p(\theta_b|y)/J_t(\theta_b|\theta_a)} \\ &= p(\theta_a|y)J_t(\theta_b|\theta_a), \end{aligned}$$

which is the same as the probability of a transition from θ_a to θ_b. Since their joint distribution is symmetric, θ^t and θ^{t-1} have the same marginal distributions, and so $p(\theta|y)$ is the stationary distribution of the Markov chain of θ. For more detailed theoretical concerns, see Gelman et al. (2004).

2.1.3 Simulation

We conduct a simulation study to assess the performance of varying coefficient model (2.1). We generate data from the following model:

$$\begin{aligned} Y_i &= \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_{1i} + \alpha_{\mathbf{2}}(U_i)X_{2i} \\ &\quad +\beta_{\mathbf{1}}Z_{1i} + \beta_{\mathbf{2}}Z_{2i} + \epsilon_i, \quad i = 1, \ldots, n, \end{aligned} \tag{2.11}$$

where U is from a uniform distribution $U(0,\ 1)$, $X_1,\ X_2,\ Z_1,\ Z_2$ are generated from the standard normal distribution $N(0,\ 1)$, and ϵ is from the normal distribution $N(0,\ \sigma_\epsilon^2 = 0.2^2)$. The variables $U,\ X_1,\ X_2,\ Z_1,\ Z_2,\ \epsilon$ are all mutually independent. We set sample size $n = 200$ & 400. The coefficients are $\alpha_0(u) = 1.5(1+\exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$, $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$, $\beta_1 = 0.2$ and $\beta_2 = 0.3$.

For approximation of the varying coefficient functions $\alpha_0,\ \alpha_1,\ \alpha_2$, we decide the knots of the cubic truncated power basis by using the $(l+1)/(K+2)$ sample quantiles of the observed predictors U, where $l = 1, \ldots,\ K$ and $K = min(n/4, 30)$=30 here.

We implement the MCMC simulation using R software. It takes about 50s and 80s to run a MCMC simulation for $n = 200$ and 400 respectively on a PC with Intel (R) Core (TM) i7 3.1 GHz processor. We use a burnin of size 2000, followed by 3000 retained iterations. From the graphical results we can conclude the convergence of the chains. The results after 500 simulations are given in Figure 2.1 (on page

26), Figure 2.2 (on page 27) and Table 2.1 (on page 28). Figure 2.3 (on page 29) and Figure 2.4 (on page 30) show the estimations of varying-coefficients arbitrarily from one of 500 simulations using Model (2.11) for $n = 200$ and 400 respectively.

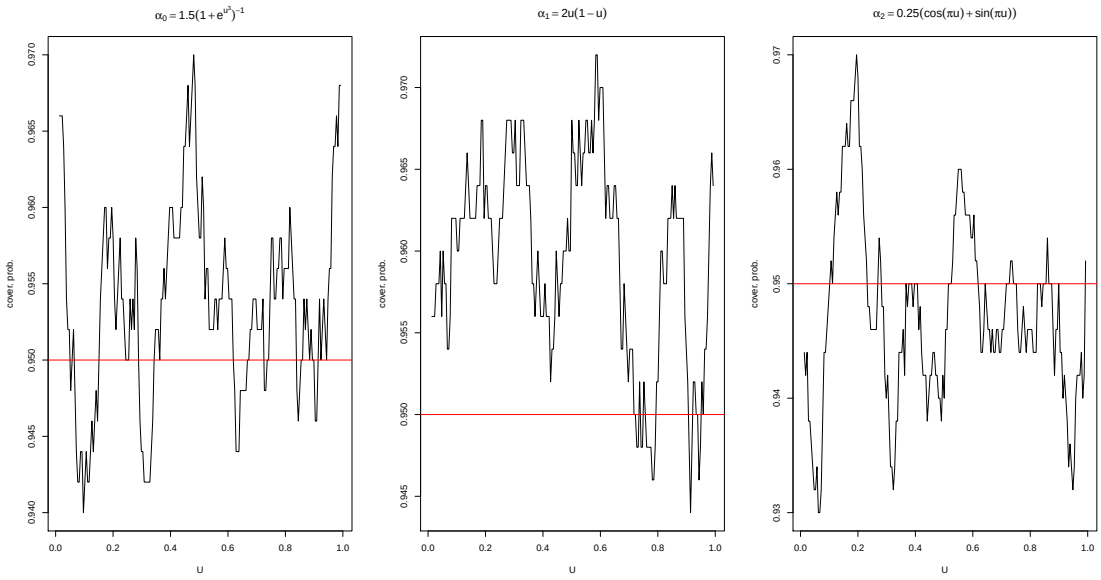

Figure 2.1 The pointwise 95% coverage probabilities for $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$ based on 500 simulations using Model (2.11), $n = 200$. The horizonal line is $y = 0.95$.

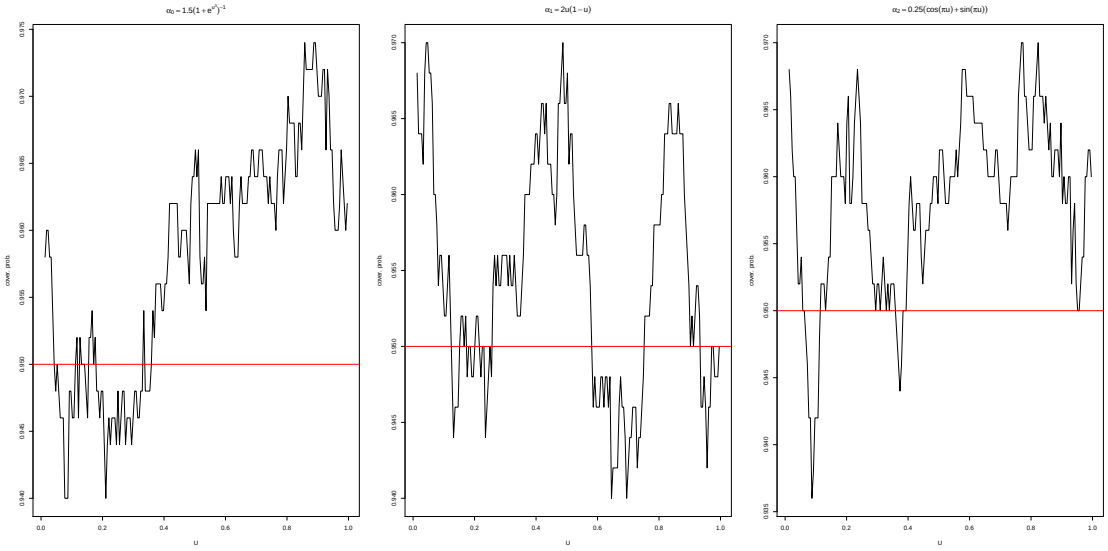

Figure 2.2 The pointwise 95% coverage probabilities for $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$ based on 500 simulations using Model (2.11), $n = 400$. The horizonal line is $y = 0.95$.

	$n = 200$			$n = 400$		
	CP	LCI	MISE	CP	LCI	MISE
$\alpha_0(\cdot)$	0.954	0.123	0.001	0.958	0.085	4.90×10^{-4}
$\alpha_1(\cdot)$	0.960	0.127	0.001	0.955	0.086	4.92×10^{-4}
$\alpha_2(\cdot)$	0.949	0.127	0.001	0.958	0.087	4.97×10^{-4}
β_1	0.948	0.058		0.948	0.040	
β_2	0.950	0.058		0.966	0.040	
σ_ϵ	0.952	0.041		0.952	0.028	

Table 2.1 Summary of 500 simulations using Model (2.11) and $n = 200$ & 400. CP is the coverage probability of 95% credible intervals, LCI is the mean length of 95% credible intervals, and MISE is the mean integrated squared error.

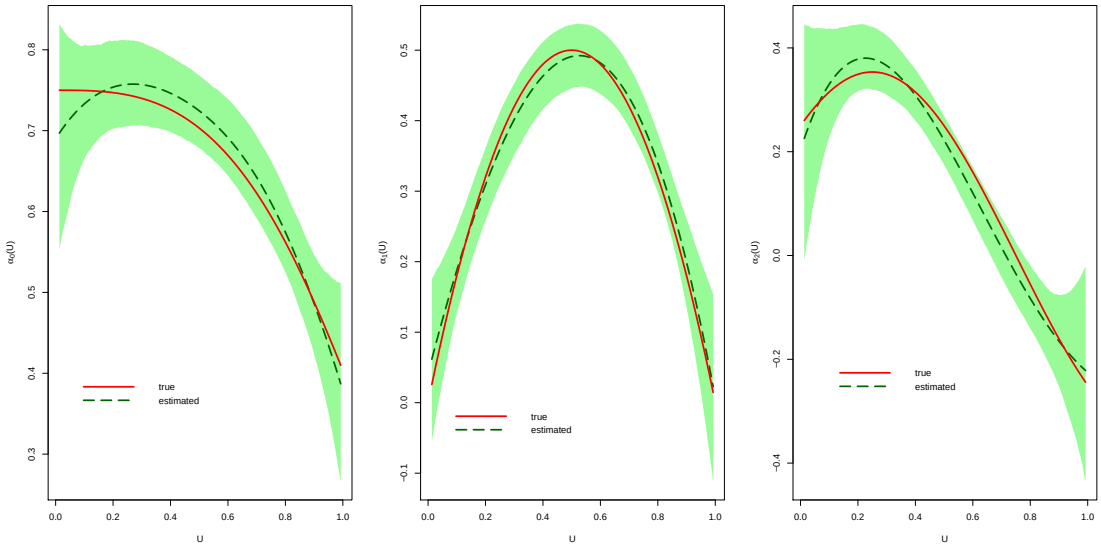

Figure 2.3 Estimation of $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u)+\cos(\pi u))$ arbitrarily from one of 500 simulations using Model (2.11), $n = 200$. The solid curves are the true functions while the dash curves are the estimations. The shaded regions correspond to point-wise 95% credible intervals.

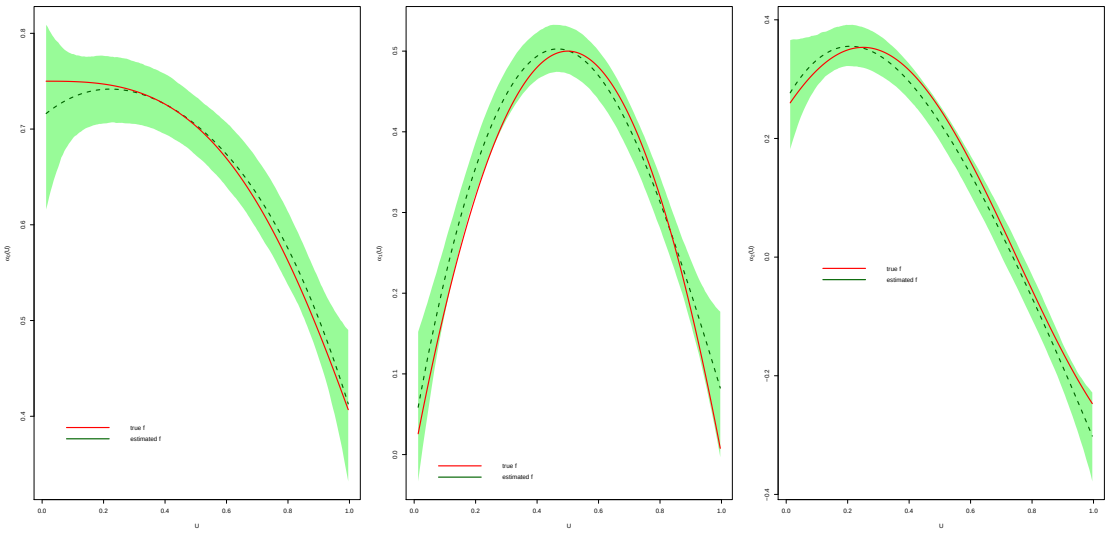

Figure 2.4 Estimation of $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u)+\cos(\pi u))$ arbitrarily from one of 500 simulations using Model (2.11), $n = 400$. The solid curves are the true functions while the dash curves are the estimations. The shaded regions correspond to point-wise 95% credible intervals.

2.2 Varying-coefficient mixed effects model

2.2.1 Statistical model

In a longitudinal study, individuals are measured repeatedly over time. Suppose we have a cohort of N subjects. For ith subject, denote the outcome variable by

Y_{ij} and the covariate vector $[X_{ij}^T, Z_{ij}^T] = [X_{ij1}, \ldots, X_{ijp}, Z_{ij1}, \ldots, Z_{ijq}]$ collected at time occasion $j = 1, \ldots, n_i$, where n_i is the total number of observations for ith subject. Denote the index variable to be U_i for ith subject, $i = 1, \ldots, N$.

We consider a semiparametric varying coefficient mixed model with random effect which can provide flexible functional coefficient estimates and meaningful interpretation. The model assumes the following dependence structure:

$$\begin{aligned} Y_{ij} &= X_{ij}^T \alpha(U_i) + Z_{ij}^T \beta \\ &\quad + b_i + \epsilon_{ij} \qquad i = 1, \ldots, N, \quad j = 1, \ldots, n_i, \end{aligned} \tag{2.12}$$

where ϵ_{ij} is normal and independent of $[U_i, X_{ij}^T, Z_{ij}^T, b_i]$ with $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma_\epsilon^2$; b_i is the random effect for ith subject which is normal and independent among subjects with $E(b_i) = 0$ and $Var(b_i) = \sigma_b^2$; $\beta = [\beta_1, \ldots, \beta_q]^T$ is a vector of q-dimensional fixed coefficient parameters for covariates Z_{ij}^T and $\alpha(\cdot) = [\alpha_1(\cdot), \ldots, \alpha_p(\cdot)]^T$ are unknown varying-coefficients for covariates X_{ij}^T. All varying coefficients are assumed to be smooth functions with continuous second derivatives.

Again, to tackle the infinite dimensional functions, we consider using the cubic truncated power basis

$$\phi(u) = [1, u, u^2, u^3, (u - \tau_1)_+^3, \ldots, (u - \tau_K)_+^3]^T$$

for approximation and denote the corresponding coefficient vector to be

$$\gamma_k = [\gamma_{k0}, \gamma_{k1}, \gamma_{k2}, \gamma_{k3}, \gamma_{k4}, \ldots, \gamma_{k,K+3}]^T, \ k = 1, \ldots, p.$$

Under the basis expansion, Model (2.12) can be re-written as

$$\begin{aligned} Y_{ij} &= X_{ij1}\phi^T(U_i)\gamma_1 + \ldots + X_{ijp}\phi^T(U_i)\gamma_p \\ &\quad + Z_{ij}^T\beta + b_i + \epsilon_{ij}, \end{aligned} \tag{2.13}$$

We may further express the above model in a matrix form as

$$\mathbb{Y} = \mathbb{C}\gamma + \mathbb{Z}\beta + \mathbb{L}b + \epsilon, \tag{2.14}$$

where $\mathbb{Y} = [Y_1^T, \ldots, Y_N^T]^T$, $Y_i = [Y_{i1}, \ldots, Y_{in_i}]^T, \quad i = 1, \ldots, N$;

$\mathbb{C} = [C_1^T, \ldots, C_N^T]^T$, $\gamma = [\gamma_1^T, \ldots, \gamma_p^T]^T$; $C_i = [C_{i1}^T, \ldots, C_{in_i}^T]^T, \ i = 1, \ldots, N$,

$C_{ij} = [X_{ij1}\phi^T(U_i), \ldots, X_{ijp}\phi^T(U_i)], \ i = 1, \ldots, N, \quad j = 1, \ldots, n_i$;

$\mathbb{Z} = [Z_1^T, \ldots, Z_N^T]^T$, $Z_i = [Z_{i1}, \ldots, Z_{in_i}]^T, \quad i = 1, \ldots, N$;

$\mathbb{L} = diag(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_N})$, $b = [b_1, \ldots, b_N]^T$;

$\epsilon = [\epsilon_1^T, \ldots, \epsilon_N^T]^T$, $\epsilon_i = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^T, \quad i = 1, \ldots, N$;

where $\mathbf{1}_{n_i}$ is a n_i dimension column vector with all elements equal to 1 and $\mathbb{L}$ is $\mathscr{N} \times N$-dimension block diagonal matrix with $\mathscr{N} = \sum_{i=1}^{N} n_i$ which is the total number of observations.

2.2.2 Bayesian inference

2.2.2.1 The prior

According to the model assumption, we have

$$\mathbb{Y}|\gamma, \beta, \mathbb{C}, \mathbb{Z}, \mathbb{L} \sim MN(\mathbb{C}\gamma + \mathbb{Z}\beta + \mathbb{L}b,\ \sigma_\epsilon^2 \mathbf{I}_\mathscr{N}),$$

where $\mathbf{I}_\mathscr{N}$ is $\mathscr{N}$-dimension unit matrix. The random effects b follow normal distribution independently, that is $b \sim MN(0,\ \sigma_b^2 \mathbf{I}_N)$.

We assume the prior for the parameters similarly to Section 2.1.2.1 as follow:

$\beta \sim MN(0,\ \sigma_\beta^2 \mathbf{I}_q)$, where $\mathbf{I}_q$ is q-dimension identical matrix and σ_β^2 so large that, for all intents and purpose, the normal distribution is uniform on the range of β.

Let $\gamma \sim MN(0,\ V)$ where $V = diag(V_1, \ldots, V_p)$ and $V_k = diag(\sigma_\gamma^2 \mathbf{I}_4,\ \sigma_k^2 \mathbf{I}_K)$, $k = 1, \ldots, p$. Similarly to σ_β^2, σ_γ^2 are large enough to obtain noninformative prior. For simplification, we set $\sigma_\gamma^2 = \sigma_\beta^2 = 10^8$ here.

For the purpose of shrinkage, we assume the priors on hyperpriors $\sigma_k^2,\ k = 1, \ldots, p$ are *inverse gamma* with parameters A_γ and B_γ independently, that is $\sigma_k^2 \sim IG(A_\gamma,\ B_\gamma),\ k = 1, \ldots, p$ where A_γ=B_γ=0.001. Further, we assume that the prior $\sigma_\epsilon^2 \sim IG(A_\epsilon,\ B_\epsilon)$ and the hyperprior $\sigma_b^2 \sim IG(A_b,\ B_b)$, where A_ϵ=B_ϵ=A_b=B_b=0.001.

2.2.2.2 The posterior

The parameter vector is now $\theta = [\gamma^T, \beta^T, b^T, \sigma_1^2, \ldots, \sigma_p^2, \sigma_b^2, \sigma_\epsilon^2]^T$, where $b = [b_1, \ldots, b_N]^T$. Similarly discussed as Section 2.1.2.2, we may derive the posterior of θ to be

$$\begin{aligned}
\pi(\theta|\mathbb{Y}) &\propto p(\mathbb{Y}|\theta)\pi(\gamma|\sigma_1^2, \ldots, \sigma_p^2)\pi(\beta)\pi(b|\sigma_b)\pi(\sigma_\epsilon^2)\pi(\sigma_b^2)\prod_{k=1}^{p}\pi(\sigma_k^2) \\
&= (\frac{1}{\sqrt{2\pi}\sigma_\epsilon})^{\mathscr{N}} \exp(-\frac{\|\mathbb{Y} - \mathbb{C}\gamma - \mathbb{Z}\beta - \mathbb{L}b\|^2}{2\sigma_\epsilon^2}) \\
&\quad \times \frac{1}{(2\pi)^{p(K+4)/2}|V|^{1/2}} \exp(-\frac{1}{2}\gamma^T V^{-1}\gamma) \\
&\quad \times (\frac{1}{\sqrt{2\pi}\sigma_\beta})^q \exp(-\frac{\|\beta\|^2}{2\sigma_\beta^2}) \\
&\quad \times (\frac{1}{\sqrt{2\pi}\sigma_b})^N \exp(-\frac{\|b\|^2}{2\sigma_b^2}) \\
&\quad \times \prod_{k=1}^{p} \frac{B_\gamma^{A_\gamma}}{\Gamma(A_\gamma)}(\sigma_k^2)^{-(A_\gamma+1)} \exp(-\frac{B_\gamma}{\sigma_k^2}) \\
&\quad \times \frac{B_\epsilon^{A_\epsilon}}{\Gamma(A_\epsilon)}(\sigma_\epsilon^2)^{-(A_\epsilon+1)} \exp(-\frac{B_\epsilon}{\sigma_\epsilon^2}) \\
&\quad \times \frac{B_b^{A_b}}{\Gamma(A_b)}(\sigma_b^2)^{-(A_b+1)} \exp(-\frac{B_b}{\sigma_b^2}). \qquad (2.15)
\end{aligned}$$

If we isolate the part of (2.15) that depends on ($\gamma,\ \beta,\ b$), we notice that the conditional posterior of ($\gamma,\ \beta,\ b$) given ($\sigma_1^2, \ldots, \sigma_p^2,\ \sigma_\epsilon^2,\ \sigma_b^2$), that is, the complete conditional–is proportional to

$$\exp(-\frac{\|\mathbb{Y} - \mathbb{C}\gamma - \mathbb{Z}\beta - \mathbb{L}b\|^2}{2\sigma_\epsilon^2} - \frac{\gamma^T V^{-1}\gamma}{2} - \frac{\|\beta\|^2}{2\sigma_\beta^2} - \frac{\|b\|^2}{2\sigma_b^2}). \qquad (2.16)$$

The term in parentheses in (2.16) is a nonnegative quadratic function of ($\gamma,\ \beta,\ b$)

and so (2.16) is proportional to a multivariate normal density. Using the usual technique of "completing the square", we can shown that

$$\pi(\gamma,\ \beta,\ b|\sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2,\sigma_b^2,\mathbb{Y}) \ \sim\ MN(\mu_{\gamma,\beta,b},\ \Sigma_{\gamma,\beta,b}). \tag{2.17}$$

Here $\mu_{\gamma,\beta,b}=\frac{1}{\sigma_\epsilon^2}\Sigma_{\gamma,\beta,b}\mathbb{D}^T\mathbb{Y}$, $\Sigma_{\gamma,\beta,b}=(\frac{\mathbb{D}^T\mathbb{D}}{\sigma_\epsilon^2}+\mathbb{E}^T\mathbb{E}+\frac{\mathbb{F}^T\mathbb{F}}{\sigma_\beta^2}+\frac{\mathbb{G}^T\mathbb{G}}{\sigma_b^2})^{-1}$, $\mathbb{D}=[\mathbb{C},\mathbb{Z},\mathbb{L}]$, $\mathbb{E}=[V^{-1/2},\mathbf{0}_{q\times q},\mathbf{0}_{N\times N}]$, $\mathbb{F}=[\mathbf{0}_{p(K+4)\times p(K+4)},\mathbf{I}_q,\mathbf{0}_{N\times N}]$ and $\mathbb{G}=[\mathbf{0}_{p(K+4)\times p(K+4)},\mathbf{0}_{q\times q},\mathbf{I}_N]$ where $\mathbf{0}_{q\times q}$ is $q\times q$-dimension zero matrix. The $p(K+4)\times p(K+4)$-dimension zero matrix corresponds to the p smooth unknown varying-coefficients, the cubic truncated power basis and the K knots. The $N\times N$-dimension zero matrix corresponds to the N random effects. Thus, as part of the MCMC chain, one generates $(\gamma,\ \beta,\ b)$ from the current values of $(\sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2,\sigma_b^2)$ according to the multivariate normal distribution, with mean and covariance matrix given by (2.17).

However, the generation of $(\gamma,\ \beta,\ b)$ from (2.17) may be too time-consuming as it involves the calculation of covariance matrix inversion, especially when N is considerable large. Fortunately, the matrix $\mathbb{L}$ is a sparse matrix with a special structure such that $\mathbb{L}^T\mathbb{L}$=$diag(n_1,\ldots,n_N)$. Thus, we can overcome the above problem by generating $(\gamma,\ \beta)$ and b respectively. Denote η=$(\gamma,\ \beta)^T$, by matrix algebra, it can be shown that

$$\begin{aligned}\pi(\gamma,\ \beta|b,\ \sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2,\sigma_b^2,\mathbb{Y}) \ &\sim\ MN(\mu_{\gamma,\beta}^{(1)},\ \Sigma_{\gamma,\beta}^{(1)}),\\ \pi(b|\gamma,\ \beta,\ \sigma_1^2,\ldots,\sigma_p^2,\sigma_\epsilon^2,\sigma_b^2,\mathbb{Y}) \ &\sim\ MN(\mu_b^{(1)},\ \Sigma_b^{(1)}).\end{aligned} \tag{2.18}$$

Here $\mu_{\gamma,\beta}^{(1)}=\frac{1}{\sigma_\epsilon^2}\Sigma_{\gamma,\beta}^{(1)}\mathbb{D}_{(1)}^T(\mathbb{Y}-\mathbb{L}b)$, $\Sigma_{\gamma,\beta}^{(1)}=(\frac{\mathbb{D}_{(1)}^T\mathbb{D}_{(1)}}{\sigma_\epsilon^2}+\mathbb{E}_{(1)}^T\mathbb{E}_{(1)}+\frac{\mathbb{F}_{(1)}^T\mathbb{F}_{(1)}}{\sigma_\beta^2})^{-1}$, $\mu_b^{(1)}=$

$\frac{1}{\sigma_\epsilon^2}\Sigma_b^{(1)}\mathbb{L}^T(\mathbb{Y}-\mathbb{D}_{(1)}\eta)$, $\Sigma_b^{(1)} = diag((\frac{n_1}{\sigma_\epsilon^2}+\frac{1}{\sigma_b^2})^{-1},\ldots,(\frac{n_N}{\sigma_\epsilon^2}+\frac{1}{\sigma_b^2})^{-1})$, $\mathbb{D}_{(1)} = [\mathbb{C},\mathbb{Z}]$, $\mathbb{E}_{(1)} = [V^{-1/2}, \mathbf{0}_{q\times q}]$ and $\mathbb{F}_{(1)} = [\mathbf{0}_{p(K+4)\times p(K+4)}, \mathbf{I}_q]$. Compared to (2.17), it will be faster to generate $(\gamma,\ \beta)$ and b by (2.18) as the dimension of $\Sigma_{\gamma,\beta}^{(1)}$ will not increase as the size of the sample increases and $\Sigma_b^{(1)}$ is a diagonal matrix.

The complete conditional for $\sigma_k^2,\ k=1,\ldots,p$ is proportional to

$$(\sigma_k^2)^{-(K/2+A_\gamma+1)}\exp(-\frac{1}{\sigma_k^2}(\frac{1}{2}(\gamma_{4,k}^2+\ldots+\gamma_{(K+3),k}^2)+B_\gamma)); \tag{2.19}$$

The complete conditional for σ_ϵ^2 is proportional to

$$(\sigma_\epsilon^2)^{-(\mathcal{N}/2+A_\epsilon+1)}\exp(-\frac{1}{\sigma_\epsilon^2}(\frac{1}{2}\|\mathbb{Y}-\mathbb{C}\gamma-\mathbb{Z}\beta-\mathbb{L}b\|^2+B_\epsilon)); \tag{2.20}$$

The complete conditional for σ_b^2 is proportional to

$$(\sigma_b^2)^{-(N/2+A_b+1)}\exp(-\frac{1}{\sigma_b^2}(\frac{1}{2}\|b\|^2+B_b)); \tag{2.21}$$

which implies that the complete conditional for $\sigma_k^2, k=1,\ldots,p,\ \sigma_\epsilon^2,\ \sigma_b^2$ all follow the Inverse-Gamma distribution.

2.2.2.3 Simulation algorithm

To sample from the posterior, we iterate S times through the following six steps.

Step 1 Sample $(\gamma,\ \beta)$ from the multivariate normal distribution: $MN(\mu_{\gamma,\beta}^{(1)},\ \Sigma_{\gamma,\beta}^{(1)})$;

Step 2 Sample b from the multivariate normal distribution: $MN(\mu_b^{(1)},\ \Sigma_b^{(1)})$;

Step 3 Sample $\sigma_k^2,\ k = 1, \ldots, p$ from $IG(A_\gamma + \frac{1}{2}K,\ B_\gamma + \frac{1}{2}(\gamma_{4,k}^2 + \ldots + \gamma_{K+3,k}^2))$ respectively;

Step 4 Sample σ_ϵ^2 from $IG(A_\epsilon + \frac{1}{2}\mathcal{N},\ B_\epsilon + \frac{1}{2}\|\mathbb{Y} - \mathbb{C}\gamma - \mathbb{Z}\beta - \mathbb{L}b\|^2)$;

Step 5 Sample σ_b^2 from $IG(A_b + \frac{1}{2}N,\ B_b + \frac{1}{2}\|b\|^2)$;

Step 6 Return to Step 1 and iterate until converge.

The above algorithm is a typical Gibbs Sampler. $(\gamma,\ \beta)$ and b can be sampled directly from multivariate normal distribution. $\sigma_k^2,\ k = 1, \ldots, p$, σ_ϵ^2 and σ_b^2 can be sampled directly from Inverse-Gamma distribution respectively.

2.2.3 Simulation

We conduct simulation study to assess the performance of varying coefficient mixed model (2.12). We generate data from the following model:

$$\begin{aligned} Y_{ij} &= \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_{1ij} + \alpha_{\mathbf{2}}(U_i)X_{2ij} \\ &\quad + \beta_{\mathbf{1}}Z_{1ij} + \beta_{\mathbf{2}}Z_{2ij} + b_i + \epsilon_{ij}, \qquad i = 1, \ldots, N,\ \ j = 1, \ldots, n_i, \end{aligned} \tag{2.22}$$

where U is from $U(0,\ 1)$, $X_1,\ X_2,\ Z_1,\ Z_2$ are generated from the standard normal distribution $N(0,\ 1)$, b is from $N(0,\ \sigma_b^2 = 0.2^2)$, and ϵ is from $N(0,\ \sigma_\epsilon^2 = 0.1^2)$. The variables $U,\ X_1,\ X_2,\ Z_1,\ Z_2,\ b,\ \epsilon$ are all mutually independent. We set sample size (subjects) $N = 200$ & 400, and n_i is randomly generated from 4 to 8 for ith subject. The total numbers of measurements are $\mathscr{N} = 1202$ and 2382 for $N = 200$ and 400 respectively. The coefficients are $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$, $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$, β_1=0.2 and β_2=0.3. For approximation of the varying coefficient functions $\alpha_0,\ \alpha_1,\ \alpha_2$, we decide the knots of the cubic truncated power basis by using the $(l+1)/(K+2)$ sample quantiles of the observed predictors U, where $l = 1, \ldots,\ K$ and $K = min(N/4, 30)$=30 here.

We implement the MCMC simulation using R software. We use a burnin of size 3000, followed by 3000 retained iterations. It takes about 285s and 1900s to run a MCMC simulation for $N = 200$ and 400 respectively on a PC with Intel (R) Core (TM) i7 3.1 GHz processor. From the graphical results we can conclude that the convergence is plausible. The results after 500 simulations are given in Figure 2.5 (on page 39), Figure 2.6 (on page 40) and Table 2.2 (on page 41). Figure 2.7 (on page 42) and Figure 2.8 (on page 43) show the estimations of varying-coefficients arbitrarily from one of 500 simulations using Model (2.22) for $N = 200$ and 400 respectively.

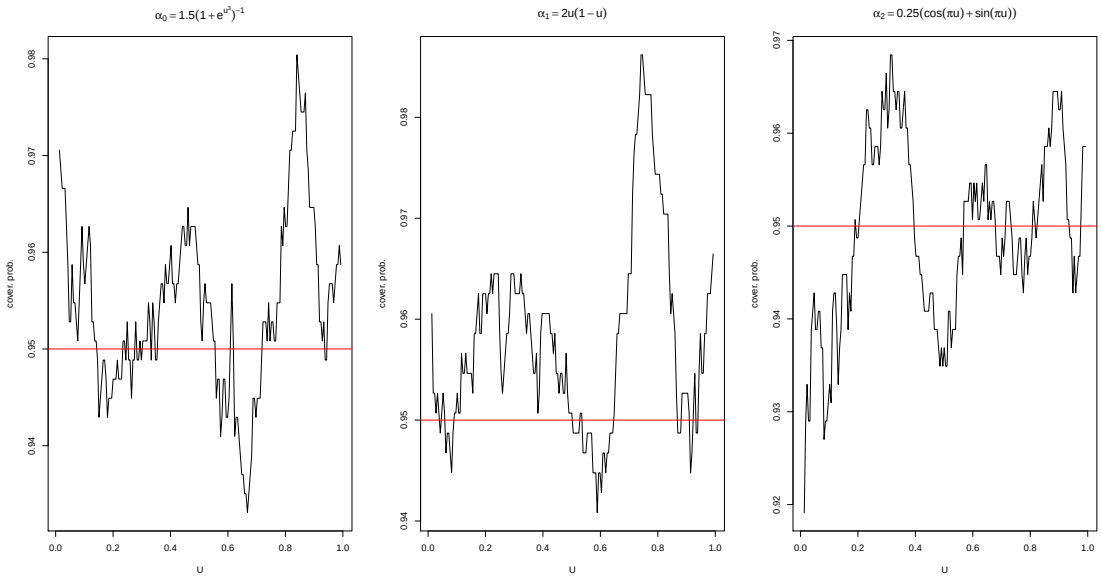

Figure 2.5 The pointwise 95% coverage probabilities for $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$ based on 500 simulations using Model (2.22), $N = 200$. The horizonal line is $y = 0.95$

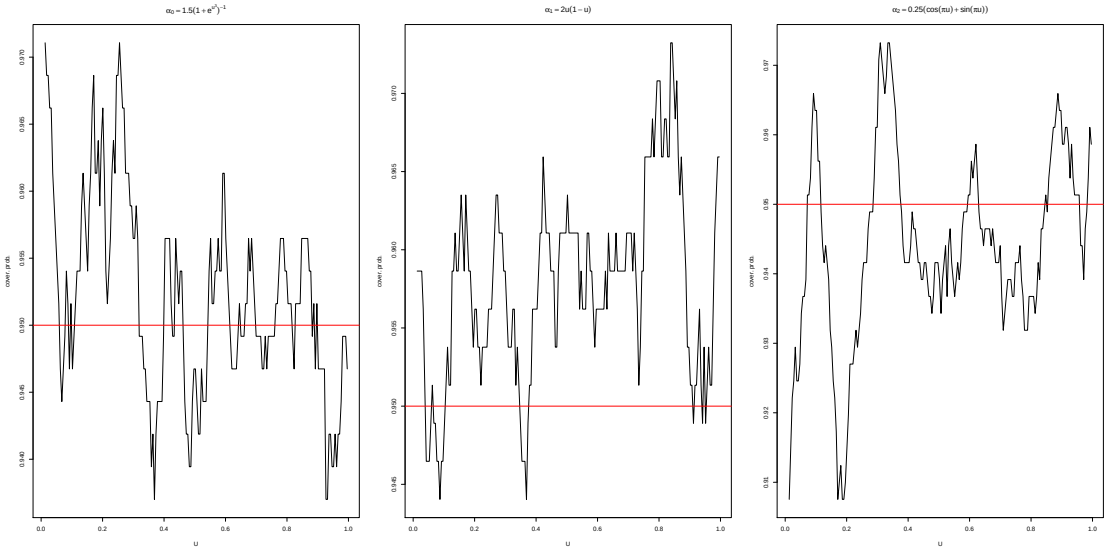

Figure 2.6 The pointwise 95% coverage probabilities for $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$ based on 500 simulations using Model (2.22), $N = 400$. The horizonal line is $y = 0.95$

	$N = 200$			$N = 400$		
	CP	LCI	MISE	CP	LCI	MISE
$\alpha_0(\cdot)$	0.952	0.119	9.89×10^{-4}	0.952	0.084	4.98×10^{-4}
$\alpha_1(\cdot)$	0.958	0.027	4.83×10^{-5}	0.958	0.019	2.41×10^{-5}
$\alpha_2(\cdot)$	0.949	0.028	5.49×10^{-5}	0.944	0.020	3.03×10^{-5}
β_1	0.956	0.012		0.952	0.009	
β_2	0.948	0.012		0.948	0.009	
σ_ϵ	0.956	0.009		0.960	0.006	
σ_b	0.970	0.042		0.954	0.029	

Table 2.2 Summary of 500 simulations using Model (2.22) and $N = 200$ & 400. CP is the coverage probability of 95% credible intervals, LCI is the mean length of 95% credible intervals, and MISE is the mean integrated squared error.

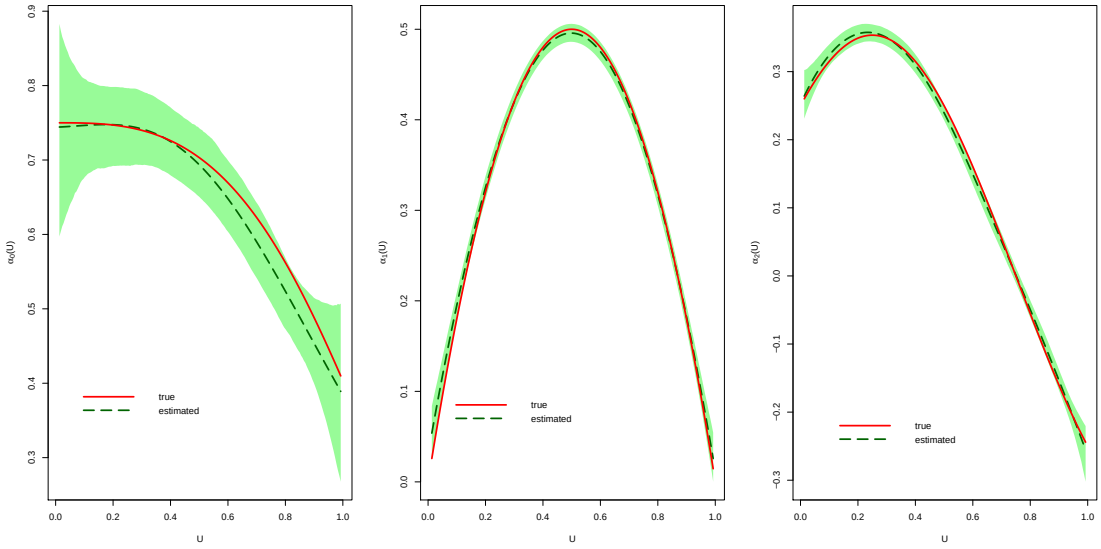

Figure 2.7 Estimation of $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u)+\cos(\pi u))$ arbitrarily from one of 500 simulations using Model (2.22), $N = 200$. The solid curves are the true functions while the dash curves are the estimations. The shaded regions correspond to point-wise 95% credible intervals.

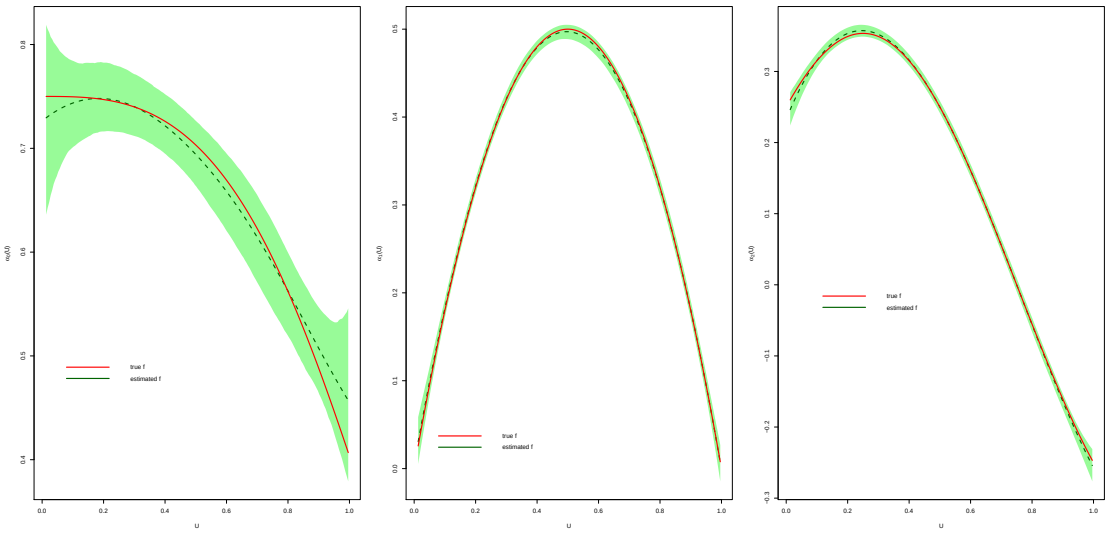

Figure 2.8 Estimation of $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1 - u)$ and $\alpha_2(u) = 0.25(\sin(\pi u)+\cos(\pi u))$ arbitrarily from one of 500 simulations using Model (2.22), $N = 400$. The solid curves are the true functions while the dash curves are the estimations. The shaded regions correspond to point-wise 95% credible intervals.

2.3 Missing data

2.3.1 Statistical model

As mentioned previously, in this study, we only consider the situation of Missing at Random (MAR). Moreover, we assumed the predictors are all observed while

only some of the responses are missing. Also, the observed response is denoted as Y^{obs} and the missing response is denoted as Y^{mis}. More specially, in a longitudinal data, Y_{ij}^{obs} means the jth measurement of ith subject and it is observed while Y_{ij}^{mis} means the jth measurement of ith subject and it is missing. All the other denotation is the same to that in Section 2.2.1.

As in Section 2.2.1, we consider using the semiparametric varying coefficient mixed model with random effect to fit the normal longitudinal data involving missing observations. We will implement three widely-used approaches to estimate the model, namely, (i) Complete-Case analysis (CC); (ii) Available-Case analysis (AC); and (iii) Bayesian imputation (BI) method (or Data augmentation).

Complete-Case analysis confines attention to cases where all the variable are present. In our context, this means that only those subjects with all the required measurements observed will be retained. In other words, ith subject will be kept only if all the n_i measurements of responses are observed, $i = 1, \ldots, N$, where N is the number of subjects and n_i is the required measurements for subject i. The model is quite similar to Model (2.12):

$$\begin{aligned} \text{CC:} \quad Y_{ij}^{obs} &= X_{ij}^T \alpha(U_i) + Z_{ij}^T \beta \\ &\quad + b_i + \epsilon_{ij} \qquad i = 1, \ldots, N^{CC}, \quad j = 1, \ldots, n_i, \end{aligned} \tag{2.23}$$

where ϵ_{ij} is normal and independent of $[U_i, X_{ij}^T, Z_{ij}^T, b_i]$ with $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma_\epsilon^2$; b_i is the random effect for ith subject which is normal and independent among

subjects with $E(b_i) = 0$ and $Var(b_i) = \sigma_b^2$; $\beta = [\beta_1, \ldots, \beta_q]^T$ is a vector of q-dimensional fixed coefficients and $\alpha(\cdot) = [\alpha_1(\cdot), \ldots, \alpha_p(\cdot)]^T$ are varying-coefficients which are assumed to be smooth with continuous second derivatives. N^{CC} is the number of subjects with all the required measurements observed and n_i is the required measurements for subject i.

Available-Case analysis includes all the observations (or measurements) where the variable of interest is present. In our context, Y_{ij}^{obs} will be included in the analysis even some of ith subject's measurements are missing, $i = 1, \ldots, N$. Besides, we assume that all the N subjects have their first measurement of response observed. This makes sense in real life, e.g. clinical trials. The model is quite similar to Model (2.23):

$$\begin{aligned} \text{AC:} \quad Y_{ij}^{obs} &= X_{ij}^T \alpha(U_i) + Z_{ij}^T \beta \\ &\quad + b_i + \epsilon_{ij} \qquad i = 1, \ldots, N, \quad j = 1, \ldots, n_i^{obs}, \end{aligned} \tag{2.24}$$

where Y_{ij}^{obs} means the jth measurement of subject i is available, N is the total number of subjects and n_i^{obs} is the set of all the observed measurements of ith subject. The other denotations are the same to Model (2.23).

Bayesian imputation is a missing value imputation procedure. Our estimation of the longitudinal regression model involving missing observations by BI mainly bases on the iterative procedure to be described later and the following formula

and model:

$$\begin{aligned} \text{BI:} \quad \hat{Y}_{ij}^{mis} &= X_{ij}^T\hat{\alpha}(U_i) + Z_{ij}^T\hat{\beta} \\ &\quad +\hat{b}_i + \hat{\epsilon}_{ij} \quad i = 1,\ldots,N, \;\; j = 1,\ldots,n_i^{mis}, \end{aligned} \tag{2.25}$$

$$\begin{aligned} Y_{ij} &= X_{ij}^T\alpha(U_i) + Z_{ij}^T\beta \\ &\quad +b_i + \epsilon_{ij} \qquad i = 1,\ldots,N, \;\; j = 1,\ldots,n_i. \end{aligned} \tag{2.26}$$

In Formula (2.25), n_i^{mis} is the set of all the missing measurements of ith subject; $\hat{\epsilon}_{ij}$ is randomly drew from $N(0,\ \hat{\sigma}_\epsilon^2)$; $\hat{\alpha}(U)$, $\hat{\beta}$, $\hat{b}_i$ and $\hat{\sigma}_\epsilon$ are the estimated varying-coefficients, constant coefficients, random effect and standard deviation of the random errors from Model (2.26) respectively; $\hat{Y}_{ij}^{mis}$ is the imputed missing measurement. In Model (2.26), Y_{ij} could be Y_{ij}^{obs} if the measurement is observed or the imputed value $\hat{Y}_{ij}^{mis}$ from Formula (2.25) if the original measurement is missing. n_i the set of all the measurements supposed to be recorded for ith subject.

The iterative procedure is summarized as follows: denote θ as the vector of all the parameters to be estimated, carrying out an AC analysis by fitting Model (2.24) to obtain the estimates for θ based on Y^{obs} and set these estimates as the initial draw $\theta_{(0)}$. Given a value $\theta_{(t)}$ of θ drawn at iteration t:

Step 1 Draw (or impute) $Y_{(t+1)}^{mis}$ from the distribution $p(Y^{mis}|Y^{obs},\theta_{(t)})$ with respect to Formula (2.25).

Step 2 Draw (or propose) $\theta_{(t+1)}$ from the distribution $p(\theta|Y^{obs}, Y^{mis}_{(t+1)})$ with respect to Model (2.26).

Step 3 The procedure terminates when the estimation of the regression coefficients converge.

The detail of the simulation algorithm will be given later.

Then for the above three estimation process, we consider using the cubic truncated power basis to approximate the varying-coefficient functions $\alpha_i(u)$, $i = 1, \ldots, p$, similarly to that in Section 2.2.1. The detail is almost the same to that in Section 2.2.1 and is ignored here. To be clear, we rewrite the model after the process of regression splines approximation for the varying-coefficient functions. That is

$$\mathbb{Y} = \mathbb{C}\gamma + \mathbb{Z}\beta + \mathbb{L}b + \epsilon, \tag{2.27}$$

where $\mathbb{C}$, $\mathbb{Z}$, $\mathbb{L}$, γ, β, b, ϵ is the same to that in Model (2.14), $\mathbb{Y}$ represents Y^{obs} w.r.t. Model (2.23) & Model (2.24), and represents both Y^{obs} and $\hat{Y}^{mis}$ w.r.t. Model (2.26).

Besides the BI method described above, there exist many alternative imputation programs such as mean imputation, last observation carry forward and frequentist regression imputation. We choose the BI method because it can be easily adapted to our Bayesian estimation procedure. From our limited numerical studies other

imputation methods have similar or even inferior performance to BI.

2.3.2 Bayesian inference

2.3.2.1 The prior and the posterior

Similarly to Section 2.2.2.1, we assume

$$\mathbb{Y}|\gamma, \beta, \mathbb{C}, \mathbb{Z}, \mathbb{L} \sim MN(\mathbb{C}\gamma + \mathbb{Z}\beta + \mathbb{L}b,\ \sigma_\epsilon^2 \mathbf{I}_{\mathscr{N}}),$$

where $\mathbf{I}_{\mathscr{N}}$ is $\mathscr{N}$-dimension unit matrix. For complete-case method, $\mathscr{N}$ is the total number of observed measurements from all the subjects with all required measurements available; for available-case method, $\mathscr{N}$ is the total number of observed measurements from all the subjects, no matter whether all required measurements of the subjects is available or not; for Bayesian imputation method, at iteration i, $\mathscr{N}$ is the total number of observed measurements supposed to be observed from all the subjects.

The assumptions for the priors of $\gamma,\ \beta,\ b$ and the priors of the hyperpriors $\sigma_k^2,\ k = 1, \ldots, p,\ \sigma_\beta^2,\ \sigma_b^2,\ \sigma_\epsilon^2$ are the same to that in Section 2.2.2.1 which is ignored here.

To point out, for Bayesian imputation method, we treat the missing measurements from all the subjects as parameters, so it is necessary to specify the priors

for them: $Y^{mis} \sim MN(\mathbb{C}^{mis}\gamma + \mathbb{Z}^{mis}\beta + \mathbb{L}^{mis}b,\ \sigma_\epsilon^2\mathbf{I}_{N^{mis}})$, where N^{mis} is the total number of missing measurements from all the subjects, $\mathbb{C}^{mis}$, $\mathbb{Z}^{mis}$, $\mathbb{L}^{mis}$ are corresponding to Y^{mis}.

The derivation of the posterior is similar to that in Section 2.2.2.2, so we ignore it and come directly to the simulation algorithm in below.

2.3.2.2 Simulation algorithm

For Complete-Case method and Available-Case method, the simulation algorithm is the same to that in Section 2.2.2.3, so we ignore it here. The only difference is that, for Complete-Case method, $\mathbb{Y}$ and $\mathscr{N}$ represent the observed responses Y^{obs} and the number of observed measurements respectively, given that the observed responses is from subjects that have all the required measurements available; for Available-Case method, $\mathbb{Y}$ and $\mathscr{N}$ represent all the observed responses Y^{obs} and the total number of all observed measurements respectively.

For Bayesian imputation method, the sampling algorithm of the posterior is as the following seven steps:

Step 1 Sample $(\gamma,\ \beta)$ from the multivariate normal distribution: $MN(\mu^{(1)}_{\gamma,\beta},\ \Sigma^{(1)}_{\gamma,\beta})$;

Step 2 Sample b from the multivariate normal distribution: $MN(\mu^{(1)}_{b},\ \Sigma^{(1)}_{b})$;

Step 3 Sample $\sigma_k^2,\ k = 1, \ldots, p$ from $IG(A_\gamma + \frac{1}{2}K,\ B_\gamma + \frac{1}{2}(\gamma_{4,k}^2 + \ldots + \gamma_{K+3,k}^2))$ respectively;

Step 4 Sample σ_ϵ^2 from $IG(A_\epsilon + \frac{1}{2}\mathcal{N},\ B_\epsilon + \frac{1}{2}\|\mathbb{Y} - \mathbb{C}\gamma - \mathbb{Z}\beta - \mathbb{L}b\|^2)$;

Step 5 Sample σ_b^2 from $IG(A_b + \frac{1}{2}N,\ B_b + \frac{1}{2}\|b\|^2)$;

Step 6 Sample Y^{mis} from $MN(\mathbb{C}^{mis}\gamma + \mathbb{Z}^{mis}\beta + \mathbb{L}^{mis}b,\ \sigma_\epsilon^2\mathbf{I}_{N^{mis}})$;

Step 7 Return to Step 1 and iterate until converge;

where all the notations from Step 1 to Step 4 are corresponding to the combination of Y^{obs} and $\hat{Y}^{mis}$ estimated from Bayes imputation; and $\mathbb{C}^{mis},\ \mathbb{Z}^{mis},\ \mathbb{L}^{mis},\ N^{mis}$ are corresponding to Y^{mis}.

2.3.3 Simulation

We conduct a simulation study to assess the performances of the three approaches: (i) Complete-Case (CC), (ii) Available-Case (AC) and Bayesian imputation (BI) method in fitting varying coefficient mixed models with random effect for longitudinal data involving missing data with normal response variables mentioned in Section 2.3.1. We generate data from the following model:

$$Y_{ij} \quad = \quad \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_{1ij} + \alpha_{\mathbf{2}}(U_i)X_{2ij} + \beta_{\mathbf{1}}Z_{1ij} + \beta_{\mathbf{2}}Z_{2ij}$$

$$+b_i + \epsilon_{ij}, \qquad i = 1, \ldots, N, \quad j = 1, \ldots, n_i, \tag{2.28}$$

where U is from a uniform distribution $U(0,\ 1)$, $X_1,\ X_2,\ Z_1,\ Z_2$ are generated from the standard normal distribution $N(0,\ 1)$, b is from the normal distribution $N(0,\ \sigma_b^2 = 0.2^2)$, and ϵ is from $N(0,\ \sigma_\epsilon^2 = 0.1^2)$. The variables $U,\ X_1,\ X_2,\ Z_1,\ Z_2,\ b,\ \epsilon$ are all mutually independent. We set sample size N=200 and n_i=n=4 for all subjects. The coefficients are $\alpha_0(u) = 1.5(1 + \exp(U^3))^{-1}$, $\alpha_1(u) = 2u(1 - u)$, $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$, β_1=0.2 and β_2=0.3

Next, for each $Y_{ij},\ j > 1$, we generate a missing value indicator $M_{ij} = I(p_{ij} \geq c)$ where the missing probability $p_{ij} = \{1 + \exp(Y_{i1} - 0.5X_{1ij} - 0.3X_{2ij} - 0.2Z_{1ij} - 0.3Z_{2ij} - 0.3)\}^{-1}$ and c is a cut-off to control the missing rate. The observed data set in each simulation only involve Y_{ij} where its corresponding $M_{ij} = 0$. Our simulation setting thus satisfies the missing at random (MAR) condition but not the missing completely at random (MCAR) condition.

We implement the MCMC simulations using R software. We use a burnin of size 2000, followed by 3000 retained iterations. It takes about an average of 13s, 122s and 150s to run a MCMC simulation Using CC, AC and BI methods on a PC with Intel (R) Core (TM) i7 3.1 GHz processor respectively. From the graphical results we can conclude the convergence of the chains. Figure 2.9 (on page 52) shows the trace plots of MCMC chains arbitrarily from one of the simulations by BI method. The results after 500 simulations are given in Table 2.3 (on page 53).

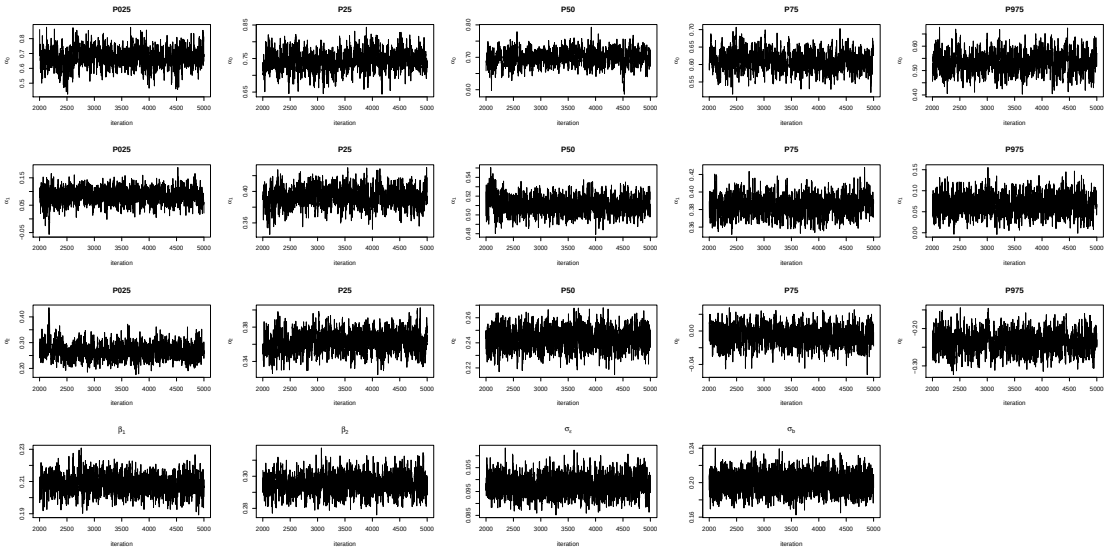

Figure 2.9 Trace plot of MCMC chains for the five percentiles (2.5th, 25th, 50th, 75th, and 97.5th) of $\alpha_0(U)$, $\alpha_1(U)$, $\alpha_2(U)$; β_1, β_2, σ_ϵ and σ_b by BI w.r.t. Model (2.28), $N = 200$.

	CC	AC	BI
CP for $\alpha_0(\cdot)$	0.403	0.943	0.952
CP for $\alpha_1(\cdot)$	0.766	0.948	0.959
CP for $\alpha_2(\cdot)$	0.811	0.942	0.949
LCI for $\alpha_0(\cdot)$	0.143	0.138	0.137
LCI for $\alpha_1(\cdot)$	0.070	0.055	0.054
LCI for $\alpha_2(\cdot)$	0.070	0.055	0.055
MISE for $\alpha_0(\cdot)$	1.10×10^{-2}	1.14×10^{-3}	1.12×10^{-3}
MISE for $\alpha_1(\cdot)$	4.07×10^{-4}	1.71×10^{-4}	1.71×10^{-4}
MISE for $\alpha_2(\cdot)$	4.24×10^{-4}	1.87×10^{-4}	1.87×10^{-4}
CP for β_1	0.908	0.940	0.952
CP for β_2	0.912	0.942	0.950
CP for σ_ϵ	0.934	0.944	0.954
CP for σ_b	0.130	0.944	0.948
LCI for β_1	0.031	0.023	0.022
LCI for β_2	0.031	0.023	0.022
LCI for σ_ϵ	0.022	0.016	0.016
LCI for σ_b	0.063	0.045	0.045

Table 2.3 Summary of 500 simulations using three missing value approaches (CC is complete case analysis, AC is available case analysis and BI is Bayesian imputation) and $N = 200$. The average missing rate is 0.3747. CP is the coverage probability of 95% credible intervals, LCI is the mean length of 95% credible intervals, and MISE is the mean integrated squared error.

The results from CC method is unacceptable with large bias and loss of precision. The bias can be observed from the low coverage probabilities for the varying and constant coefficients. Loss of precision is reflected in the relatively wide mean lengths of the 95% credible intervals for the varying and constant coefficients and for the variance parameters, and the relatively large MISEs for varying-coefficients. The results from AC analysis is better than those from CC method by increasing the coverage probabilities, decreasing the mean lengths of the 95% credible intervals and the MISEs; thus results in less bias estimation and higher precision. In general, BI method improves the estimation for the model compared to the other two and is the most recommended approach among the three.

We increase the number of subjects to N=400 and repeat the simulations in Section 2.3.3.1. It takes about an average of 60s, 650s and 800s to run a MCMC simulation with a burnin of size 3000 and 3000 retained iterations using CC, AC and BI methods on a PC with Intel (R) Core (TM) i7 3.1 GHz processor respectively. The convergence is plausible from the graphical results.Figure 2.10 (on page 55) shows the trace plots of MCMC chains arbitrarily from one of the simulations by BI method. The results after 500 simulations are given in Table 2.4 (on page 56).

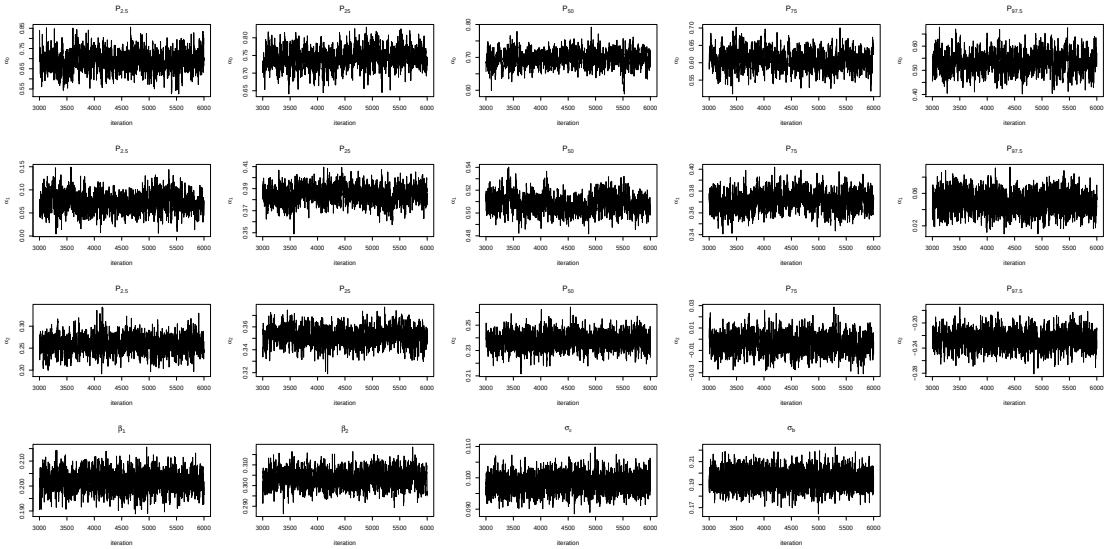

Figure 2.10 Trace plot of MCMC chains for the five percentiles (2.5th, 25th, 50th, 75th, and 97.5th) of $\alpha_0(U)$, $\alpha_1(U)$, $\alpha_2(U)$; β_1, β_2, σ_ϵ and σ_b by BI w.r.t. Model (2.28), $N = 400$.

	CC	AC	BI
CP for $\alpha_0(\cdot)$	0.267	0.942	0.951
CP for $\alpha_1(\cdot)$	0.798	0.949	0.960
CP for $\alpha_2(\cdot)$	0.872	0.939	0.942
LCI for $\alpha_0(\cdot)$	0.096	0.097	0.096
LCI for $\alpha_1(\cdot)$	0.046	0.038	0.036
LCI for $\alpha_2(\cdot)$	0.047	0.038	0.037
MISE for $\alpha_0(\cdot)$	8.21×10^{-3}	5.30×10^{-4}	5.25×10^{-4}
MISE for $\alpha_1(\cdot)$	1.87×10^{-4}	7.89×10^{-5}	7.86×10^{-5}
MISE for $\alpha_2(\cdot)$	1.87×10^{-4}	9.19×10^{-5}	9.19×10^{-5}
CP for β_1	0.914	0.944	0.946
CP for β_2	0.918	0.940	0.946
CP for σ_ϵ	0.914	0.938	0.940
CP for σ_b	0.012	0.938	0.944
LCI for β_1	0.021	0.016	0.015
LCI for β_2	0.021	0.016	0.015
LCI for σ_ϵ	0.015	0.011	0.011
LCI for σ_b	0.044	0.032	0.031

Table 2.4 Summary of 500 simulations using three missing value approaches (CC is complete case analysis, AC is available case analysis and BI is Bayesian imputation) and $N = 400$. The average missing rate is 0.3751. CP is the coverage probability of 95% credible intervals, LCI is the mean length of 95% credible intervals, and MISE is the mean integrated squared error.

Table 2.4 (on page 56) shows a similar result to that in Table 2.3 (on page 53). CC and AC both improve as sample size grows. BI analysis still performs the best among the three methods.

CHAPTER 3

Varying-coefficient model for binary response

3.1 Model & estimation

3.1.1 Statistical model

When the response variable is binary, Model (2.1) is unsuitable. We consider a varying-coefficient model for binary response variable. Perhaps the most well-know

regression model is the logistic regression which we extend as follows:

$$\begin{aligned} Y &\sim Bernoulli(g^{-1}(\eta)) \\ \eta &= \alpha^T(U)X + \beta^T Z, \end{aligned} \tag{3.1}$$

where β and $\alpha(\cdot)$ play the same role as those in Model (2.1) and the link function $g(\pi)=\log\frac{\pi}{1-\pi}$ is considered in the study. Other link functions for binary outcome can be similarly constructed.

Suppose we have a random sample of size $n, \{(U_i, X_{i1}, \ldots, X_{ip}, Z_{i1}, \ldots, Z_{iq}, Y_i), i = 1, \ldots, n\}$ from Model (3.1), where Y_k are assumed to be independent of each other.

As in Model (2.1), to tackle the infinite dimensional functions, we consider using the cubic truncated power basis

$$\phi(u) = [1, u, u^2, u^3, (u-\tau_1)_+^3, \ldots, (u-\tau_K)_+^3]^T$$

for approximation and denote the corresponding coefficient vector to be

$$\gamma_k = [\gamma_{k0}, \gamma_{k1}, \gamma_{k2}, \gamma_{k3}, \gamma_{k4}, \ldots, \gamma_{k,K+3}]^T, \ k = 1, \ldots, p,$$

where $\tau_l, \ \ l = 1, \ldots, K$ are the knots and K is number of knots.

Under the basis expansion, Model (3.1) can be re-written as

$$\begin{aligned} \eta_i &= \sum_{k=1}^{p} X_{ik}\phi^T(U_i)\gamma_k + Z_i^T\beta, \\ &= C_i^T\gamma + Z_i^T\beta, \quad i = 1, \ldots, n, \end{aligned} \tag{3.2}$$

where $C_i = [X_{i1}\phi^T(U_i), \ldots, X_{ip}\phi^T(U_i)]^T,\ \gamma = [\gamma_1^T, \ldots, \gamma_p^T]^T$,

$Z_i = [Z_{i1}, \ldots, Z_{iq}]^T$; The above model in matrix form is

$$\eta \ = \ C\gamma + Z\beta,$$

where $\eta = [\eta_1, \ldots, \eta_n]^T$, $C = (C_1, \ldots, C_n)^T$, and $Z = (Z_1, \ldots, Z_n)^T$.

Then, Model (3.1) is rewritten as follows:

$$\begin{aligned} Y_i &\sim Bernouilli(g^{-1}(\eta_i)) \\ \eta_i &= C_i^T\gamma + Z_i^T\beta, \quad i = 1, \ldots, n. \end{aligned} \tag{3.3}$$

3.1.2 Bayesian inference

3.1.2.1 The prior

According to model assumption, we have

$$Y_i|\gamma, \beta, C, Z \sim Bernouilli(\frac{exp(C_i^T\gamma + Z_i^T\beta)}{1 + exp(C_i^T\gamma + Z_i^T\beta)}).$$

We assume the priors for $\beta,\ \gamma$ similarly to that in Section 2.1.2.1 as follow:

Let $\beta \sim MN(0,\ \sigma_\beta^2\mathbf{I}_q)$ where $\mathbf{I}_q$ is q-dimension identical matrix and σ_β^2 so large that, for all intents and purpose, the normal distribution is uniform on the range of β.

Let $\gamma \sim MN(0,\ V)$ where $V = diag(V_1, \ldots, V_p)$ and $V_k = diag(\sigma_\gamma^2\mathbf{I}_4,\ \sigma_k^2\mathbf{I}_K),\ k =$

$1, \ldots, p$. Similarly to σ_β^2, σ_γ^2 are large enough to obtain noninformative prior. For simplification, we set $\sigma_\gamma^2 = \sigma_\beta^2 = 10^8$ here.

For the purpose of shrinkage, we assume the priors on hyperpriors $\sigma_k^2,\ k = 1, \ldots, p$ are *inverse gamma* with parameters A_γ and B_γ independently, that is $\sigma_k^2 \sim IG(A_\gamma,\ B_\gamma),\ k = 1, \ldots, p$ where A_γ=B_γ=0.001.

3.1.2.2 The posterior

Denote the parameter space by Θ and parameter in Θ by θ where $\theta = [\gamma^T, \beta^T, \sigma_1^2, \ldots, \sigma_p^2]^T$. Denote the prior distribution of θ by $\pi(\theta)$ and the conditional distribution of sample Y by $p(Y|\theta)$ where $Y = [Y_1, \ldots, Y_n]^T$. Then the posterior of θ:

$$\pi(\theta|Y) \propto p(Y|\theta)\pi(\theta).$$

By independency, $p(Y|\theta) = \prod_{k=1}^n p(Y_k|\theta)$, where

$$p(Y_i|\theta) = \pi_i^{Y_i}(1-\pi_i)^{(1-Y_i)} = e^{Y_i(C_i^T\gamma + Z_i^T\beta)}(1 + e^{C_i^T\gamma + Z_i^T\beta})^{-1};$$

By independency, $\pi(\theta) = \pi(\gamma)\pi(\beta)\prod_{k=1}^p \pi(\sigma_k^2)$, where

$$\pi(\gamma) = \frac{1}{(2\pi)^{(p+1)(K+4)/2}|V|^{1/2}} \exp(-\frac{1}{2}\gamma^T V^{-1}\gamma),$$

$$\pi(\beta) = (\frac{1}{\sqrt{2\pi}\sigma_\beta})^q \exp(-\frac{\|\beta\|^2}{2\sigma_\beta^2}),$$

$$\pi(\sigma_k^2) = \frac{B_\gamma^{A_\gamma}}{\Gamma(A_\gamma)}(\sigma_k^2)^{-(A_\gamma+1)}\exp(-\frac{B_\gamma}{\sigma_k^2}).$$

Then

$$\begin{aligned}\pi(\theta|Y) \quad \propto \quad & e^{\sum_{i=1}^n Y_i(C_i^T\gamma+Z_i^T\beta)} \times \prod_{i=1}^n (1+e^{C_i^T\gamma+Z_i^T\beta})^{-1} \\ & \times \frac{1}{|V|^{1/2}}\exp(-\frac{1}{2}\gamma^T V^{-1}\gamma) \\ & \times \exp(-\frac{\|\beta\|^2}{2\sigma_\beta^2}) \\ & \times \prod_{k=1}^p (\sigma_k^2)^{-(A_\gamma+1)}\exp(-\frac{B_\gamma}{\sigma_k^2}). \end{aligned} \tag{3.4}$$

From (3.4), we can see the conditional posterior of γ given $(\beta, \sigma_1^2, \ldots, \sigma_p^2)$–that is, the complete conditional–is proportional to

$$e^{\sum_{i=1}^n Y_i(C_i^T\gamma+Z_i^T\beta)} \prod_{i=1}^n (1+e^{C_i^T\gamma+Z_i^T\beta})^{-1}\exp(-\frac{1}{2}\gamma^T V^{-1}\gamma). \tag{3.5}$$

The complete conditional for β is proportional to

$$e^{\sum_{i=1}^n Y_i(C_i^T\gamma+Z_i^T\beta)} \prod_{i=1}^n (1+e^{C_i^T\gamma+Z_i^T\beta})^{-1}\exp(-\frac{\|\beta\|^2}{2\sigma_\beta^2}). \tag{3.6}$$

The complete conditional for $\sigma_k^2,\ k = 1, \ldots, p$ is proportional to

$$(\sigma_k^2)^{-(K/2+A_\gamma+1)}\exp(-\frac{1}{\sigma_k^2}(\frac{1}{2}(\gamma_{4,k}^2 + \ldots + \gamma_{(K+3),k}^2) + B_\gamma)), \tag{3.7}$$

which implies that the complete conditionals for $\sigma_k^2,\ k = 1, \ldots, p$ follow the Inverse-Gamma distribution.

3.1.2.3 Simulation algorithm

To sample from the posterior, we iterate S times through the following four steps.

Step 1 Sample γ from the multivariate distribution

$$f(\gamma) = R_\gamma e^{\sum_{i=1}^{n} Y_i(C_i^T\gamma + Z_i^T\beta)} \prod_{i=1}^{n} (1 + e^{C_i^T\gamma + Z_i^T\beta})^{-1} \exp(-\frac{1}{2}\gamma^T V^{-1}\gamma),$$

where R_γ is the normalizing factor to make the integral of $f(\gamma)$=1;

Step 2 Sample β from the multivariate distribution

$$f(\beta) = R_\beta e^{\sum_{i=1}^{n} Y_i(C_i^T\gamma + Z_i^T\beta)} \prod_{i=1}^{n} (1 + e^{C_i^T\gamma + Z_i^T\beta})^{-1} \exp(-\frac{\|\beta\|^2}{2\sigma_\beta^2}),$$

where R_β is the normalizing factor to make the integral of $f(\beta)$=1;

Step 3 Sample $\sigma_k^2,\ k = 1, \ldots, p$ from $IG(A_\gamma + \frac{1}{2}K, B_\gamma + \frac{1}{2}(\gamma_{4,k}^2 + \ldots + \gamma_{K+3,k}^2))$ respectively;

Step 4 Return to Step 1 and iterate until converge.

The above three steps process is a typical Gibbs sampler which is a particular Markov chain simulation algorithm.

In Step 3 we can sample $\sigma_k^2,\ k = 1, \ldots, p$ directly from Inverse-Gamma distribution. However, Step 1 and Step 2 both involve complicated conditional posterior

distributions more than 1 dimension.

Conventionally, we adopt multidimensional M-H algorithm in both Step 1 and Step 2, that is updating γ in block and β in block respectively. However, the M-H algorithm introduces an accept-reject stage which makes the simulation slow, especially when the proposal distribution is not suitable. Besides, a potential problem lurks in that there is strong posterior correlation between γ and β, which is likely to cause poor mixing and slow convergence in the simulation chain.

In the next section, we propose a new model using auxiliary variables which overcome the above problems.

3.1.3 Data augmentation

3.1.3.1 Auxiliary variable model

In this section, we propose a new model by adding auxiliary variables which make the Gibbs sampler computation simple and convergence accelerated because we can then use conjugate prior for the parameters in the model, thus directly sampling the parameters from the conditional posterior instead of using M-H algorithm is available according to Holmes and Held (2006). The idea of adding variables is also called *data augmentation*. We consider the following hierarchical

data generation mechanism:

$$\begin{aligned} Y &= \begin{cases} 1 & \text{if } \omega > 0 \\ 0 & \text{otherwise} \end{cases} \\ \omega &= \alpha^T(U)X + \beta^T Z + \epsilon \\ \epsilon &\sim N(0,\ \lambda) \\ \lambda &= (2\psi)^2 \\ \psi &\sim KS, \end{aligned} \tag{3.8}$$

where β and $\alpha(\cdot)$ are the same to that in model (3.1). ψ follows the Kolmogorov-Smirnov (KS) distribution. For more detailed information about KS distribution, see Devroye (1986).

Similarly to Section 3.1.1, suppose we have a random sample of size n, $\{(U_i, X_{i1}, \ldots, X_{ip}, Z_{i1}, \ldots, Z_{iq}, Y_i), i = 1, \ldots, n\}$ from Model (3.8), where Y_i are assumed to be independent of each other.

By the same procedure in Section 3.1.1 using truncated power basis and matrix algebra, Model (3.8) is re-written as follows:

$$\begin{aligned} Y_i &= \begin{cases} 1 & \text{if } \omega_i > 0 \\ 0 & \text{otherwise} \end{cases} \\ \omega_i &= C_i^T\gamma + Z_i^T\beta + \epsilon_i \\ \epsilon_i &\sim N(0,\ \lambda_i) \\ \lambda_i &= (2\psi_i)^2 \end{aligned}$$

$$\psi_i \sim KS, \quad i = 1, \ldots, n, \tag{3.9}$$

where γ, C_i and Z_i are the same to those in Model (3.2) and $\psi_i,\ i = 1, \ldots, n$ are independent random variables following the KS distribution. In this case, ϵ_i has a scale mixture of normal form with a marginal logistic distribution (see Andrews and Mallows (1974)) so that the marginal likelihood $L(\gamma, \beta | Y_1, \ldots, Y_n)$ for Model (3.9) and Model (3.3) are equivalent.

3.1.3.2 The prior

We use the same prior for β, γ and hyperprior for $\sigma_k^2,\ k = 1, \ldots, p$ to that in Section 3.1.2.1. That is:

Let $\beta \sim MN(0,\ \sigma_\beta^2 \mathbf{I}_q)$ where $\sigma_\beta^2 = 10^8$; $\gamma \sim MN(0,\ V)$ where $V = diag(V_1, \ldots, V_p)$, $V_k = diag(\sigma_\gamma^2 \mathbf{I}_4,\ \sigma_k^2 \mathbf{I}_K),\ k = 1, \ldots, p$ and $\sigma_\gamma^2 = 10^8$ here; for the purpose of shrinkage, $\sigma_k^2 \sim IG(A_\gamma,\ B_\gamma),\ k = 1, \ldots, p$ where A_γ=B_γ=0.001.

3.1.3.3 The posterior

We denote $\omega = [\omega_1, \ldots, \omega_n]^T$, $\lambda = [\lambda_1, \ldots, \lambda_n]^T$, $Y = [Y_1, \ldots, Y_n]^T$, $C = [C_1, \ldots, C_n]^T$ and $Z = [Z_1, \ldots, Z_n]^T$.

Under the aforementioned prior and model specification, we can easily derive posterior full conditionals. Specifically, the full conditional distribution of (γ, β) given $(\sigma_1^2, \ldots, \sigma_p^2, \omega, \lambda)$ is still normal,

$$\begin{aligned} \gamma,\ \beta|\sigma_1^2, \ldots, \sigma_p^2, \omega, \lambda &\sim MN(\mu_{\gamma,\beta},\ \Sigma_{\gamma,\beta}) \\ \mu_{\gamma,\beta} &= \Sigma_{\gamma,\beta} D^T \Lambda^{-1} \omega \\ \Sigma_{\gamma,\beta} &= (D^T \Lambda^{-1} D + E^T E + \frac{F^T F}{\sigma_\beta^2})^{-1} \\ \Lambda &= diag(\lambda_1, \ldots, \lambda_n), \end{aligned} \tag{3.10}$$

where $D = [C, Z]$, $E = [V^{-1/2}, \mathbf{0}_{q\times q}]$, $F = [\mathbf{0}_{p(K+4)\times p(K+4)}, \mathbf{I}_q]$ and $\mathbf{0}_{q\times q}$ is $q \times q$-dimension zero matrix.

The full conditional for $\sigma_k^2,\ k = 1, \ldots, p$ are still *inverse gamma*,

$$\sigma_k^2|\gamma_k \sim IG((A_\gamma + \frac{1}{2}K),\ B_\gamma + \frac{1}{2}(\gamma_{4,k}^2 + \ldots + \gamma_{(K+3),k}^2)). \tag{3.11}$$

The full conditional for each element $\omega_i,\ i = 1, \ldots, n$ is truncated normal,

$$\omega_i|\gamma, \beta, Y_i, \lambda_i \propto \begin{cases} N(C_i^T\gamma + Z_i^T\beta,\ \lambda_i) I(\omega_i > 0) & \text{if } Y_i = 1 \\ N(C_i^T\gamma + Z_i^T\beta,\ \lambda_i) I(\omega_i \leq 0) & \text{otherwise.} \end{cases} \tag{3.12}$$

The conditional distribution $\pi(\lambda_i|\omega_i, \gamma, \beta)$ does not have a standard form. However, it can be generated using rejection sampling as outlined in Holmes and Held (2006).

3.1.3.4 Simulation algorithm

We can implement automatic sampling from the posterior using iterative updates according to the above specification, that is, $(\gamma, \beta|\sigma_1^2, \ldots, \sigma_p^2, \omega, \lambda)$ followed by $(\sigma_k^2|\gamma_k),\ k = 1, \ldots, p$, followed by $(\omega|\gamma, \beta, Y, \lambda)$ and then $(\lambda|\omega, \gamma, \beta)$. The sampling scheme will be slower due to the strong posterior correlation between (γ, β) and ω, as can be seen in Model (3.9).

According to Holmes and Held (2006), there are two options to improve matters through joint updating. On the one hand, we can propose to update (γ, β) and ω jointly making use of the factorization,

$$\pi(\gamma, \beta, \omega|Y, \lambda, \sigma_1^2, \ldots, \sigma_p^2) = \pi(\omega|Y, \lambda)\pi(\gamma, \beta|\omega, \lambda, \sigma_1^2, \ldots, \sigma_p^2),$$

followed by an update to $\lambda|\omega, \gamma, \beta$ and then $\sigma_k^2|\gamma_k,\ k = 1, \ldots, p$. On the other hand we can update jointly (ω, λ) given γ, β,

$$\pi(\omega, \lambda|\gamma, \beta, Y) = \pi(\omega|\gamma, \beta, Y)\pi(\lambda|\omega, \gamma, \beta),$$

followed by an update to $\gamma, \beta|\omega, \lambda, \sigma_1^2, \ldots, \sigma_p^2$ and then $\sigma_k^2|\gamma_k,\ k = 1, \ldots, p$. In this latter case the marginal densities for the $\omega_i's,\ i = 1, \ldots, n$ are independent truncated logistic distributions,

$$\omega_i|\gamma, \beta, Y_i \ \propto \ \begin{cases} Logistic(C_i^T\gamma + Z_i^T\beta,\ 1)I(\omega_i > 0) & \text{if } Y_i = 1 \\ Logistic(C_i^T\gamma + Z_i^T\beta,\ 1)I(\omega_i \leq 0) & \text{otherwise,} \end{cases} \tag{3.13}$$

where $Logistic(a,\ b)$ denotes the density function of the logistic distribution with mean a and scale parameter b (see Devroye (1986)). As showed in Holmes and Held (2006), this latter approach has an advantage that sampling from the truncated logistic distribution can be done efficiently by the inversion method, because both the distribution function and its inverse have simple analytic form. In our simulation, we adopt the latter approach. We outline the algorithm below:

To sample from the posterior, we iterate S times through the following five steps.

Step 1 Sample $(\gamma,\ \beta)$ from the multivariate normal distribution: $MN(\mu_{\gamma,\beta},\ \Sigma_{\gamma,\beta})$;

Step 2 Sample $\sigma_k^2,\ k = 1, \ldots, p$ from $IG(A_\gamma + \frac{1}{2}K,\ B_\gamma + \frac{1}{2}(\gamma_{4,k}^2 + \ldots + \gamma_{K+3,k}^2))$ respectively;

Step 3 Sample $\omega_i,\ i = 1, \ldots, n$ according to (3.13);

Step 4 Sample $\lambda_i,\ i = 1, \ldots, n$ using reject sampling outlined in Holmes and Held (2006);

Step 5 Return to Step 1 and iterate until converge.

3.1.4 Simulation

We conduct simulation study to assess the performance of binary varying coefficient model (3.9).

Our simulation contain two parts: (i) generate data using Model (3.1); (ii) fit the data using Model (3.9).

In **Part i**, we generated data from the following model:

$$\begin{aligned} Y_i &\sim Bernoulli(g^{-1}(\eta_i)) \\ \eta_i &= \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_{1i} + \alpha_{\mathbf{2}}(U_i)X_{2i} + \beta_{\mathbf{1}}Z_{1i} + \beta_{\mathbf{2}}Z_{2i}, \quad i = 1, \ldots, n, \end{aligned} \tag{3.14}$$

where U is from $U(0,\ 1)$, $X_1,\ X_2,\ Z_1,\ Z_2$ are all generated from $N(0,\ 1)$. $U,\ X_1,\ X_2,\ Z_1,\ Z_2$ are all mutually independent. The link function is $g(\pi)$=$\log\frac{\pi}{1-\pi}$. The sample size is $n = 200$ and 400. The coefficients are $\alpha_0(u) = 1.5(1+\exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$, $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$, $\beta_1 = 0.2$ and $\beta_2 = 0.3$.

In **Part ii**, we fit the data using Model (3.9). For approximation of the varying coefficient functions $\alpha_0,\ \alpha_1,\ \alpha_2$, we decide the knots of the cubic truncated power basis by using the $(l+1)/(K+2)$ sample quantiles of the observed predictors U, where $l = 1, \ldots,\ K$ and $K = min(n/4, 30)$=30 here.

We implement the MCMC simulation using R software. We use a burnin of size 2000, followed by 3000 retained iterations. It takes about 160s and 355s to run a MCMC simulation for $n = 200$ and 400 respectively on a PC with Intel (R) Core

(TM) i7 3.1 GHz processor. From the graphical results we can conclude that the convergence is plausible. The results after 500 simulations are given in Figure 3.1 (on page 71), Figure 3.2 (on page 72) and Table 3.1 (on page 72).

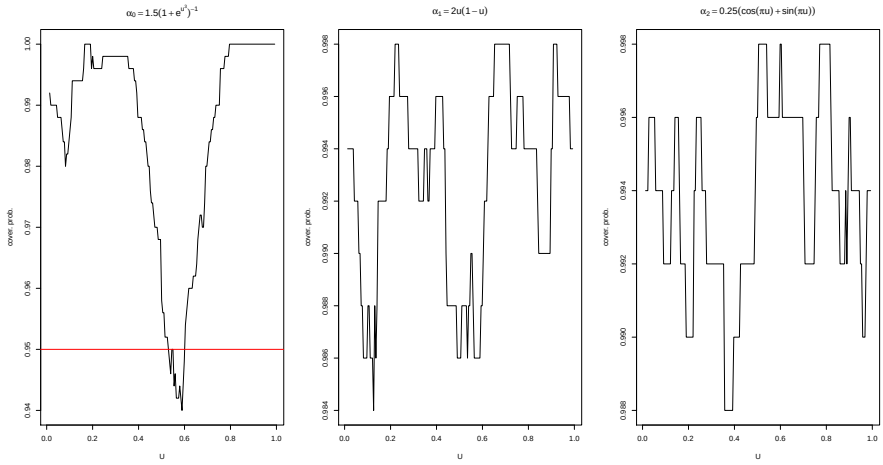

Figure 3.1 The pointwise 95% coverage probabilities for $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$ based on 500 simulations using Model (3.14), $n = 200$.

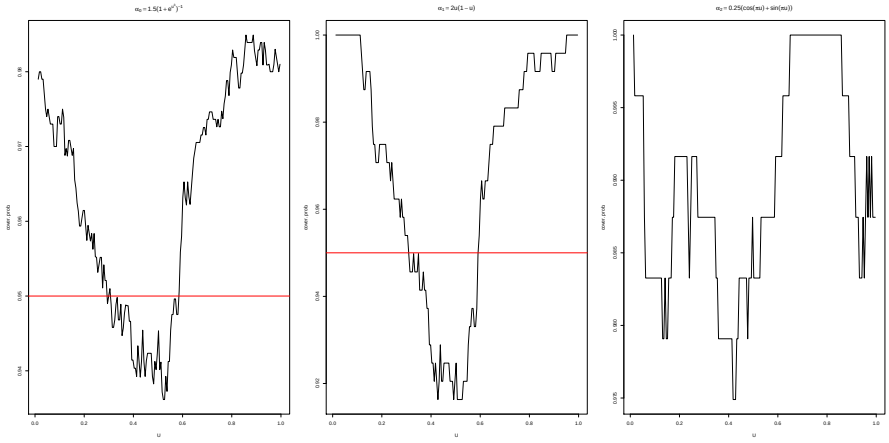

Figure 3.2 The pointwise 95% coverage probabilities for $\alpha_0(u) = 1.5(1 + \exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$ and $\alpha_2(u) = 0.25(\sin(\pi u) + \cos(\pi u))$ based on 500 simulations using Model (3.14), $n = 400$.

	$n = 200$			$n = 400$		
	CP	LCI	MISE	CP	LCI	MISE
$\alpha_0(\cdot)$	0.986	1.414	0.125	0.965	0.860	0.126
$\alpha_1(\cdot)$	0.993	1.528	0.107	0.969	0.904	0.044
$\alpha_2(\cdot)$	0.994	1.519	0.100	0.990	0.896	0.033
β_1	0.992	0.665		0.974	0.428	
β_2	0.992	0.670		0.966	0.430	

Table 3.1 Summary of the 500 simulations using Model (3.14) and $n = 200$ & 400. CP is the coverage probability of 95% credible intervals, LCI is the mean length of 95% credible intervals, and MISE is the mean integrated squared error.

3.2 Missing data

3.2.1 Statistical model

In this section, we consider the binary varying-coefficient model with respect to the situation when some of the binary responses are missing at random (MAR). Similarly to Section 2.3.1, we assumed the predictors are all observed while some of the responses are missing. The observed response is denoted as Y^{obs} and the missing response is denoted as Y^{mis}. All the other denotations are the same to that in Section 3.1.1.

As in Section 3.1.1, we consider fitting the binary semiparametric varying-coefficient model involving missing responses.

For Complete-Case (CC) method when all the missing data are deleted, the model is similar to Model (3.1):

$$\begin{aligned} \text{CC:} \quad Y_i^{obs} &\sim Bernoulli(g^{-1}(\eta_i^{obs})) \\ \eta_i^{obs} &= \alpha^T(U_i)X_i + \beta^T Z_i, \quad i = 1, \ldots, n^{obs}, \end{aligned} \tag{3.15}$$

where β and $\alpha(\cdot)$ are the same to those in Model (3.1) and the link function $g(\pi)$=$\log\frac{\pi}{1-\pi}$. n^{obs} is the number of observed data.

In the simple regression data context, Available-Case analysis is exactly identical to Complete-Case analysis which confines attention to cases where all the

variable are present. We ignore it here.

For Bayesian imputation method (BI), the estimation process is based on Model (3.1) and an iterative procedure where the missing response Y^{mis} is treated as parameters, which is quite similar to that in Section 2.3.1. The detail of the simulation algorithm will be given later.

3.2.2 Data augmentation

For the above two estimation process (CC & BI), we adopt Model (3.8) proposed in Section 3.1.3.1 to avoid the problem described in Section 3.1.2.3 and make the Gibbs Sampler simpler and convergence accelerated. We apply regression splines technique and matrix algebra similar to Section 3.1.3.1 too.

Then for CC method, we propose the following model instead of Model (3.15):

$$\begin{aligned} Y_i^{obs} &= \begin{cases} 1 & \text{if } \omega_i > 0 \\ 0 & \text{otherwise} \end{cases} \\ \omega_i &= C_i^T\gamma + Z_i^T\beta + \epsilon_i \\ \epsilon_i &\sim N(0,\ \lambda_i) \\ \lambda_i &= (2\psi_i)^2 \\ \psi_i &\sim KS, \quad i = 1, \ldots, n^{obs}, \end{aligned} \tag{3.16}$$

where all the denotations are the same to that in Model (3.9) and n^{obs} is the number of observed data.

For BI method, the estimation process is based on Model (3.9) and an iteration procedure similar to that in Section 2.3.1.

3.2.3 Bayesian inference

3.2.3.1 The prior and the posterior

The assumptions for the priors of γ , β and the hyperpriors $\sigma_k^2,\ k = 1, \ldots, p$ are the same to that in Section 3.1.3.2 which is ignored here.

The derivation of the posterior for CC method is similar to that in Section 3.1.3.3, so we ignore it here and go directly to the simulation algorithm below.

3.2.3.2 Simulation of the posterior

For CC method, the simulation algorithm is the same to that in Section 3.1.3.4, so we ignore here.

For BI imputation method, the detailed sampling algorithm involved six steps

is outlined as below:

Step 1 Sample $(\gamma,\ \beta)$ from the multivariate normal distribution: $MN(\mu_{\gamma,\beta},\ \Sigma_{\gamma,\beta})$;

Step 2 Sample $\sigma_i^2,\ i = 1, \ldots, p$ from $IG(A_\gamma + \frac{1}{2}K,\ B_\gamma + \frac{1}{2}(\gamma_{4,i}^2 + \ldots + \gamma_{K+3,i}^2))$ respectively;

Step 3 Sample $\omega_i,\ i = 1, \ldots, n$ according to (3.13);

Step 4 Sample $\lambda_i,\ i = 1, \ldots, n$ using reject sampling outlined in Holmes and Held (2006);

Step 5 calculate $\eta^{mis} = C^{mis}\gamma + Z^{mis}\beta$ and sample Y^{mis} from $Bernoulli(g^{-1}(\eta^{mis}))$;

Step 6 Return to Step 1 and iterate until converge;

where all the denotations are similar to Section 3.1.3.4 except η^{mis} and Y^{mis}. n is the total size of the sample, including observed and missing. All the random draws from Step 1 to Step 4 are with respect to the combination of Y^{obs} and Y^{mis} where Y^{mis} is drew from Step 5. C^{mis} and Z^{mis} is similar to that in Model (3.3) with respect to Y^{mis} and $g(\pi)$=$\log\frac{\pi}{1-\pi}$.

3.2.4 Simulation

We conduct simulation study to assess the performance of the two method (CC and BI) in fitting binary varying-coefficient model involving missing data. We generate data from the following model:

$$\begin{aligned} Y_i &\sim Bernoulli(g^{-1}(\eta_i)) \\ \eta_i &= \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_{1i} + \alpha_{\mathbf{2}}(U_i)X_{2i} + \beta_{\mathbf{1}}Z_{1i} + \beta_{\mathbf{2}}Z_{2i}, \quad i = 1, \ldots, n, \end{aligned} \tag{3.17}$$

where U is generated from $U(0,\ 1)$ and $X_1,\ X_2,\ Z_1,\ Z_2$ are generated from $N(0,\ 1)$. $U,\ X_1,\ X_2,\ Z_1,\ Z_2$ are all mutually independent. The coefficients are set to be $\alpha_0(u) = 1.5(1+\exp(u^3))^{-1}$, $\alpha_1(u) = 2u(1-u)$, $\alpha_2(u) = 0.25(\sin(\pi u)+\cos(\pi u))$, $\beta_1 = 0.2$ and $\beta_2 = 0.3$. The missing value indicator is set to be $M_i = I(p_i \geq c),\ i = 1, \ldots, n$, where p_i is calculated by $\{1+\exp(-[.3+.5X_{1i}+.3X_{2i}+.2Z_{1i}+.3Z_{2i}])\}^{-1}$ and c is a preset constant to control the missing rate. The observed data set in each simulation only involve Y_i where its corresponding $M_i = 0$. Thus our simulation setting satisfies MAR condition but not MCAR condition. The sample size is $n = 200$ or 400.

We implement the MCMC simulation using R software. We use a burnin of size 2000 followed by 3000 retained iterations. It takes about an average of 65s & 170s ($n = 200$) and 160s & 360s ($n = 400$) to run a MCMC simulation using CC and BI methods on a PC with Intel (R) Core (TM) i7 3.1 GHz processor respectively.

From graphical results we can conclude the convergence of the chains. Figure 3.3 (on page 78) shows the trace plots of MCMC chains arbitrarily from one of the simulations by BI method with $n = 400$. The results after 500 simulations for both $n = 200$ & 400 are given in Table 3.2 (on page 79).

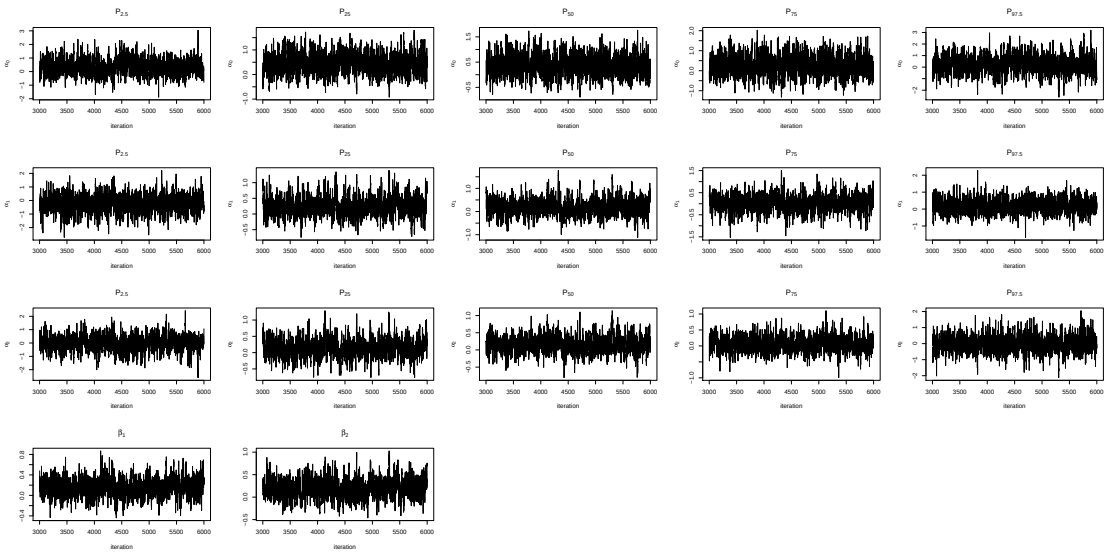

Figure 3.3 Trace plot of MCMC chains for the five percentiles (2.5th, 25th, 50th, 75th, and 97.5th) of $\alpha_0(U)$, $\alpha_1(U)$, $\alpha_2(U)$; β_1 and β_2 by BI w.r.t. Model (3.17), $n = 400$.

	CC	BI	CC	BI
Sample Size	n=200	n=200	n=400	n=400
CP for $\alpha_0(\cdot)$	0.963	0.983	0.971	0.992
CP for $\alpha_1(\cdot)$	0.962	0.991	0.980	0.998
CP for $\alpha_2(\cdot)$	0.966	0.991	0.985	0.998
LCI for $\alpha_0(\cdot)$	4.474	2.907	2.081	2.040
LCI for $\alpha_1(\cdot)$	4.516	2.270	2.028	1.495
LCI for $\alpha_2(\cdot)$	4.236	1.986	1.896	1.255
MISE for $\alpha_0(\cdot)$	5.580	0.349	0.267	0.189
MISE for $\alpha_1(\cdot)$	4.235	0.179	0.218	0.062
MISE for $\alpha_2(\cdot)$	4.404	0.149	0.203	0.034
CP for β_1	0.973	0.990	0.988	0.998
CP for β_2	0.980	0.990	0.990	0.998
LCI for β_1	1.394	1.079	0.799	0.766
LCI for β_2	1.581	1.381	0.906	0.898

Table 3.2 Summary of 500 simulations using two missing value approaches (CC is complete case analysis and BI is Bayesian imputation). The average missing rate is 0.5068. CP is the coverage probability of 95% credible intervals, LCI is the mean length of 95% credible intervals, and MISE is the mean integrated squared error.

From Table 3.2 (on page 79), we can observe that BI method outperforms CC method overall. Both methods get unbiased estimations while BI analysis gets higher coverage probabilities for the parameters. However, compared to CC analysis, BI analysis improves the precision of the estimation by reducing the mean lengths of the 95% credible intervals and the MISEs a lot, especially when sample size is small.

CHAPTER 4

Real data analysis

4.1 Background of the data

Dementia will be a major global public health issue in the coming decades, especially in Chinese aging populations. Previous studies have consistently reported sharply increased prevalence and incidence rate of dementia with increasing age in the Chinese elderly, (e.g. Sahadevan et al. (1997), Zhang et al. (2005), Feng et al. (2012)). Singapore Longitudinal Aging Study (SLAS) is a recent prospective study in Singapore which enrolled 2501 Chinese elder adults aged ≥ 55 from September 2003 to December 2005, two rounds of follow-up assessments were conducted from

March 2005 to September 2007 and from August 2007 to December 2008 correspondingly. This observational study attempts to investigate various epidemiology questions for dementia patients.

The primary outcome variable is the Mini-Mental State Examination (MMSE), which measures global cognitive functioning in domains of memory, attention, language, praxis, and visuospatial ability. The MMSE score ranged from 0 to 30 with higher values indicating better cognitive functioning. Cognitive impairment may be defined by an MMSE score ≤ 23. Participants in this study were examined and their MMSE values were recorded at the baseline and two follow-up visits. The longitudinal observations of MMSE may be used to cast insights on the natural disease progression for aging adults.

A secondary outcome variable in this study is the Clinical Dementia Rating (CDR) score. CDR is a numeric scale used to quantify the severity of dementia symptoms. A CDR global score of 0 indicates no dementia, whereas CDR global score of 0.5, 1, 2 or 3 indicate very mild, mild, moderate or severe dementia respectively. The CDR was administered by trained researchers with necessary medical background in SLAS. In this study, we used CDR = 0.5 to define early cognitive impairment. Therefore, CDR could be treated as a binary cross-sectional variable with two categories, CDR ≥ 0.5 and CDR < 0.5, which was computed at baseline and used to assess the prevalence of dementia in the study sample.

In addition, information on age, gender, education level, alcohol consumption,

cigarette consumption, amount of social and physical activities were collected by research nurses through face-to-face interview, and data on the possible related diseases such as hypertension, heart attack and diabetes etc. were collected from the patients' records.

4.2 Pretreatment of the data

As mentioned before, there are 2501 participants in total involved in the study. However, there are missing values for almost all the variables, especially for the responses. For baseline MMSE, there is no missing value; for 1st-MMSE, the missing rate is 840/2501=33.6%; for 2nd-MMSE, the missing rate is 1315/2501=52.6%. The reasons of missing contain: some people quit after the baseline MMSE or 1st-MMSE, some people couldn't be connected after baseline MMSE or 1st-MMSE, etc. For CDR, the missing rate reaches 2030/2501=81.2%. The high missing rate of CDR may partially due to the fact that most of the participants need not to be administered CDR. The missing cases are relatively much fewer in the predictors, the sum of missing cases throughout all the predictors are 343, so the missing rate for predictors is 343/2501=13.7%.

Here our objectives are:

(1) to impute the large amount of missing values for 1st-MMSE, 2nd-MMSE respectively with Bayesian imputation method, treat baseline MMSE, 1st-MMSE and 2nd-MMSE as longitudinal data and assess the varying-coefficient mixed effects model in fitting longitudinal data involving missing data;

(2) to impute the large amount of missing value CDR and assess the varying-coefficient model in fitting binary data involving missing data.

For these reasons, before the estimation, we delete the cases where predictors are missing which results in a renewed data as following :

categorical variable	category	meaning	dummy variable	percentage
Apolipoprotein E carrier	0	no		82.8%
	1	yes	apoe	17.2%
sex	1	male	sex	36.2%
	2	female		73.8%
education level	0	7+		48.1%
	1	0-6 yr	edu	51.9%
hypertension history	0	no		46.0%
	1	yes	hpt	54.0%
diabetes history	0	no		83.5%
	1	yes	dia	16.5%
stroke history	0	no, or not sure		96.3%
	1	yes	str	3.7%
heart failure	0	no		94.7%
or attack history	1	yes	heart	5.3%
social activities score	1	1-5	soc1	31.7%
	2	6-8	soc2	37.1%
	3	9 or more		31.2%
physical activities score	1	0-1	phy1	30.9%
	2	2-3	phy2	32.1%
	3	3 or more		37.0%
continuous variable	mean	s.d.	range	
age (years)	65.75	7.64	(52.83, 97.57)	
1st follow-up time (years from baseline)			(0.57, 3.96)	
2nd follow-up time (years from baseline)			(2.54, 5.31)	

Table 4.1 A summary of the predictors after pretreatment of the real data.

The pretreatment and results for fitting MMSEs and CDR will be a little different which will be specified in the corresponding sections below.

After the pretreatment, we provide a general Bayesian procedure to fit semiparametric varying-coefficient regression model for longitudinal normal (MMSEs) and cross-sectional binary (CDR) response. Specifically, the nonparametric components are approximated with a cubic truncated power basis expansion and Bayesian spline technique described in previous reports are used in the estimation.

4.3 Varying-coefficient mixed effects model for MMSEs

The baseline MMSE, 1st-MMSE and 2nd-MMSE are measured for each subject during the period of the study as well as the times when measured. Thus the can be treated as longitudinal data. For simplification, we denote baseline MMSE, 1st-MMSE and 2nd-MMSE as Y_0, Y_1 and Y_2 respectively. We assume that the responses follow the normal distribution and there is a random effect b for each subject among the three measurements Y_0, Y_1 and Y_2. The predictors are all measured one time when Y_0 was measured except the measured times which are recorded when MMSEs were measured. We are particularly interested in the effects of Apolipoprotein E

carrier (*apoe*) and model its coefficient using a varying-coefficient function of *age*. Besides, we also assume *age* has a varying effect on the responses. The other predictors variables are adjusted in the model with constant coefficients. For those constant coefficient predictors with k categories, we create dummy variables for their first $k-1$ categories respectively.

Then we propose the model as below:

$$\begin{aligned} Y_{ij} &= \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_i + \sum_{k=1}^{10} \beta_{\mathbf{k}} Z_{ik} + \beta_{\mathbf{11}} t_{ij} \\ &\quad + b_i + \epsilon_{ij}, \qquad i = 1, \ldots, N, \quad j = 1, \ldots, 3, \end{aligned} \tag{4.1}$$

where Y_{ij} is the 3 MMSEs; U is *age*; X is *apoe*; $\alpha_0(U)$ and $\alpha_1(U)$ are smoothing functions with continuous second derivatives; $\beta_k,\ k = 1, \ldots, 11$ are constants; t is the time distance between a pre-set time point and the baseline MMSE time or 1st-MMSE time or 2nd-MMSE time, where the pre-set time point is set as the baseline MMSE time for each subject, thus t_{i1}=0 for all the subjects. t is measured in years; b_i is the random effect within each subject and follows a normal distribution with zero mean and variance σ_b^2; ϵ_{ij} is the random error following a normal distribution with mean 0 and variance σ_ϵ^2 and Z is a 10-dimension-vector which contains constant-coefficient continuous predictors and dummy variables for categorical predictors. Because each subject has 3 measurements of MMSEEs, j takes the values of 1, 2 or 3.

One of our objectives in fitting this model is to impute the large amount of

missing values for 1st-MMSE, 2nd-MMSE respectively with Bayesian imputation method then fit the varying-coefficient mixed model for longitudinal data with respect to the cases where all predictors are observed.

Then one problem occurs in Model (4.1): among all the predictors, only the 3 MMSEs measured times can reflect the difference among the 3 MMSEs for each subject as they are recorded synchronously when baseline MMSE, 1st-MMSE or 2nd-MMSE are observed; however, when the respective 1st-MMSE or 2nd-MMSE is not observed, the measured time will not be recorded too. This problem prevents us from implementing the missing MMSEs and fit the model to satisfy our goal.

To overcome this problem, we propose the following treatment: impute the missing 1st-MMSE measured times and 2nd-MMSE measured times using the respect medians of all the observed 1st-MMSE measured times and 2nd-MMSE measured times.

Then we carry on the pretreatment mentioned in Section 4.2. Thus after deleting the data where any predictor is missing, the number of subjects under analysis is N=2256 with some of them have all the 3 measurements observed and some of them have 1st or 2nd-MMSE missing.

Then we use complete-case analysis (CC), available-case analysis (AC) and Bayes imputation method (BI) to fit the data using Model (4.1) by implementing the cubic truncated power basis expansion and Bayesian procedure described in Chapter 2. Then we compare their estimation results. For CC analysis, the number

of subjects is N^{CC}=891 and the total number of measurements is $\mathscr{N}^{CC}$=2673; for AC analysis, the corresponding numbers are N^{AC}=2256 and $\mathscr{N}^{AC}$=4842 respectively; for BI analysis, the corresponding numbers are N^{BI}=2256 and $\mathscr{N}^{BI}$=6768 respectively.

We implement the MCMC simulation using R software. We use a burnin of size 3000, followed by 3000 retained iterations. They take us about 1081s, 26.6h and 27.1h to run the MCMC simulation on a PC with Intel (R) Core (TM) i7 3.1 GHz processor respectively. The convergence of the simulations is plausible after we check the MCMC chains. Figure 4.1 (on page 90) and Figure 4.2 (on page 91) show some of the trace plots of MCMC chains by BI analysis. The nonparametric and parametric estimation results are given in Figure 4.3 (on page 92) and Table 4.2 (on page 93) respectively.

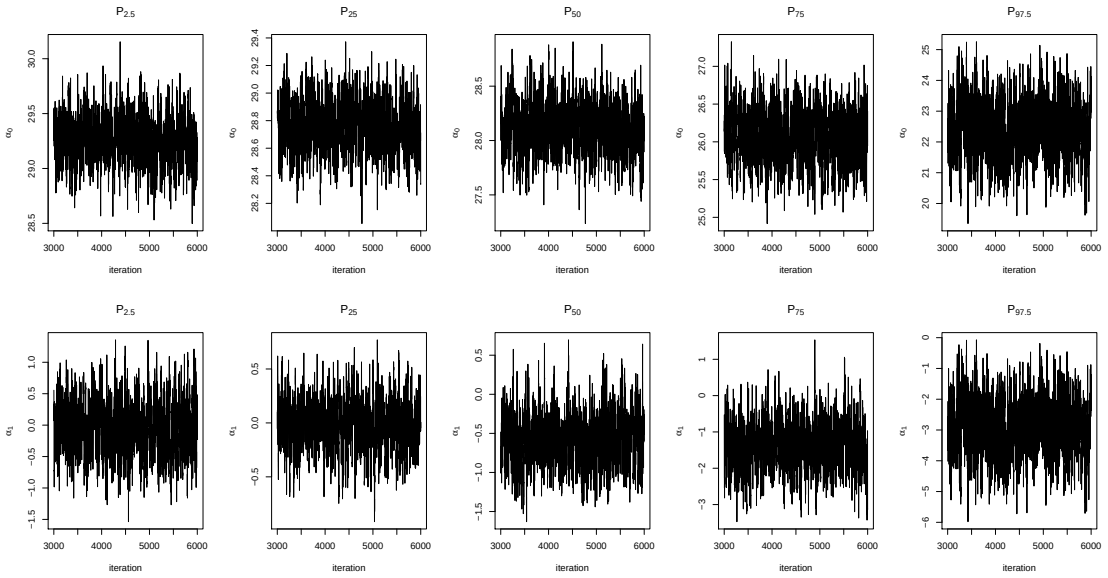

Figure 4.1 Trace plot of MCMC chains for the five percentiles (2.5th, 25th, 50th, 75th and 97.5th) of $\alpha_0(U)$ and $\alpha_1(U)$ by BI analysis w.r.t Model (4.1); The upper five plots are respective to $\alpha_0(U)$ and the lower five plots are respective to $\alpha_1(U)$.

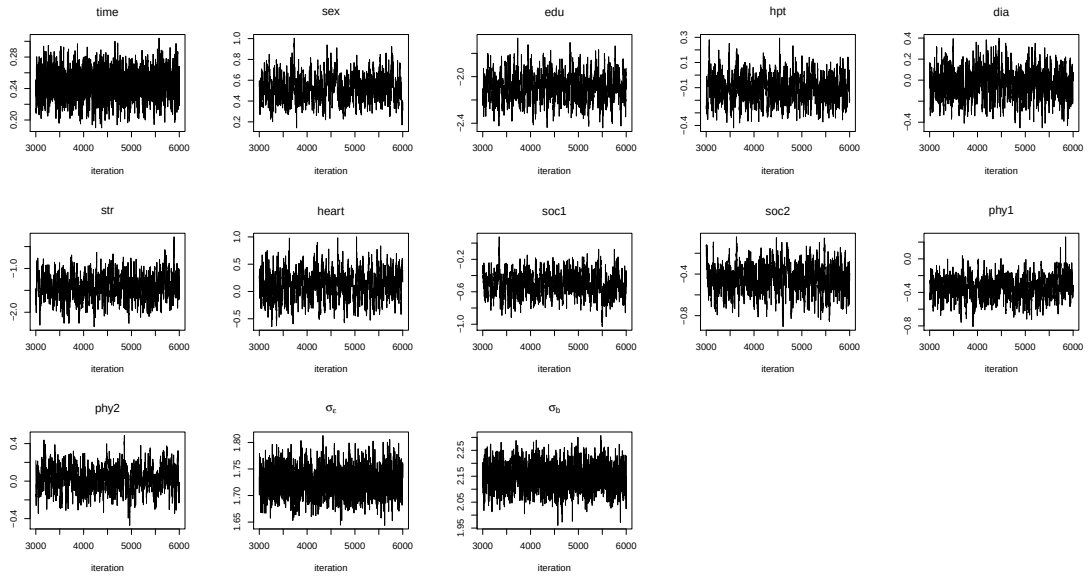

Figure 4.2 Trace plot of MCMC chains for the constant coefficients and variance parameters by BI analysis w.r.t Model (4.1).

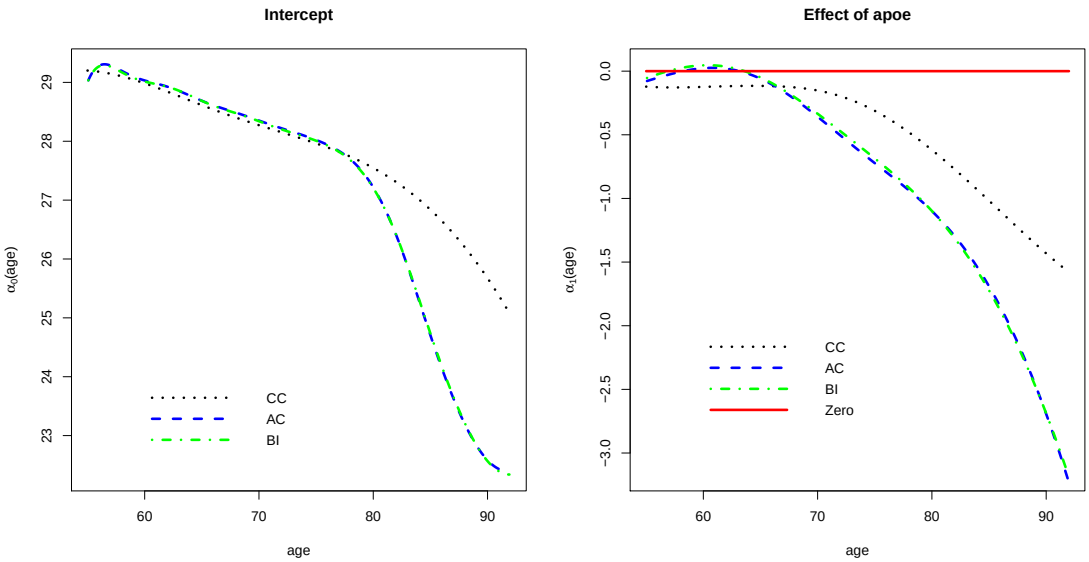

Figure 4.3 Comparison of the estimations of varying-coefficient functions for MMSEs using Model (4.1) by CC, AC and BI.

The results from available-case analysis and Bayesian imputation method are rather similar, differing from those under complete-case analysis. The intercept is a decreasing function of age, indicating older subjects tend to have lower MMSE. The regression coefficients of Apolipoprotein E are also decreasing with age, indicating that the carrier of this gene trait tend to have a lower MMSE and older carriers tend to decrease their MMSE more than younger carriers.

Variable	CC	AC	BI
time	0.21 (0.17, 0.25)	0.24 (0.21, 0.28)	0.25 (0.21, 0.28)
sex	0.44 (0.14, 0.72)	0.51 (0.29, 0.75)	0.53 (0.30, 0.78)
edu	-1.64 (-1.92, -1.36)	-2.09 (-2.32, -1.87)	-2.09 (-2.31, -1.85)
hpt	-0.09 (-0.34, 0.20)	-0.11 (-0.32, 0.10)	-0.11 (-0.30, 0.10)
dia	-0.18 (-0.55, 0.17)	-0.01 (-0.31, 0.28)	-0.03 (-0.29, 0.24)
str	-0.52 (-1.21, 0.17)	-1.42 (-1.93, -0.86)	-1.42 (-1.95, -0.88)
heart	0.02 (-0.56, 0.57)	0.10 (-0.36, 0.56)	0.12 (-0.36, 0.63)
soc1	-0.48 (-0.83, -0.14)	-0.53 (-0.81, -0.24)	-0.51 (-0.77, -0.24)
soc2	-0.36 (-0.66, -0.06)	-0.45 (-0.71, -0.21)	-0.45 (-0.72, -0.19)
phy1	-0.24 (-0.56, 0.08)	-0.36 (-0.61, -0.11)	-0.35 (-0.60, -0.09)
phy2	0.19 (-0.08, 0.48)	0.02 (-0.21, 0.26)	0.02 (-0.23, 0.27)
σ_ϵ	1.63 (1.58, 1.69)	1.73 (1.68, 1.77)	1.73 (1.68, 1.78)
σ_b	1.60 (1.51, 1.71)	2.15 (2.05, 2.25)	2.15 (2.06, 2.24)

Table 4.2 Comparison of the estimated posterior means and 95% credible intervals of constant-coefficients and variance parameters for MMSE response using Model (4.1) by CC, AC and BI.

Table 4.2 (on page 93) shows that the results by AC and BI analysis are quite similar which are a little different from that by CC analysis. From CC, AC and

BI analysis, we all conclude that time, sex, education level and the two level of social activities scores are significant. However, the coefficient for stroke produces a credible interval (-1.21, 0.17) under complete-case analysis, indicating stroke is not significantly associated with MMSE; the results under available case analysis and Bayesian imputation method both claim a significant negative effects of stroke with credible intervals excluding zero. The first level of physical activities score is also estimated non-significant under complete case analysis and significant under available case and Bayesian imputation.

4.4 Varying-coefficient model for CDR

As mentioned in Section 4.1, a Clinical Dementia Rating (CDR) global score of 0 indicates no dementia, whereas CDR global score of 0.5, 1, 2 or indicate very mild, moderate or severe dementia respectively. As the number of CDR ≥ 1 (1, 2 and 3) cases are too few, in this section we combine them with CDR $= 0.5$ group, thus CDR could be treated as binary variable with two categories, CDR < 0.5 and CDR ≥ 0.5. We denote the response CDR as Y and assume that Y follows the Bernoulli distribution. Similarly to Section 4.3, we model the coefficient of *apoe* using a varying-coefficient function of *age*. *age* also has a varying

effect on CDR. The other predictors have constant coefficients. For those constant coefficient predictors with k categories, we create dummy variables for their first $k-1$ categories respectively. Among these constant predictors, we should point out that MMSEs have strong effect on CDR according to dementia experts (e.g. Morris (1993), Feng et al. (2012)); so beside those constant predictors in Model (4.1) (excluding MMSEs measured times t), we add the mean of the 3 MMSEs scores, the difference between 1st-MMSE and baseline MMSE, and the difference between 2nd-MMSE and 1st-MMSE.

Then we propose the following model:

$$\begin{aligned} Y_i &\sim Bernoulli(g^{-1}(\eta_i)) \\ \eta_i &= \alpha_{\mathbf{0}}(U_i) + \alpha_{\mathbf{1}}(U_i)X_i + \sum_{k=1}^{13} \beta_{\mathbf{k}} Z_{ik} \quad i = 1, \ldots, n, \end{aligned} \tag{4.2}$$

where Y_i is CDR; U is *age*; X is *apoe*; Z is a 13-dimension-vector which contains constant-coefficient continuous predictors and dummy variables for categorical predictors and the link function $g(\eta) = \log\frac{\eta}{1-\eta}$.

Before we fit the model, we carry on the pretreatment mentioned in Section 4.2. Thus after deleting the data where any predictor is missing, n=264 for CC estimation and n=891 for BI estimation. As mentioned before, the highly missing measurements of CDR is partially due to the fact that most of the participants need not to be administered CDR.

Then we use CC and BI methods to fit the data using Model (4.2) by implementing the cubic truncated power basis expansion, Bayesian procedure and data augmentation described in Chapter 3.

We implement the MCMC simulation using R software. We use a burnin of size 3000, followed by 3000 retained iterations. They take us about 262s and 1262s to run the MCMC simulation on a PC with Intel (R) Core (TM) i7 3.1 GHz processor respectively. The convergence of the simulation is plausible after we check the MCMC chains. Figure 4.4 (on page 97) and Figure 4.5 (on page 98) show some of the trace plots of MCMC chains by BI analysis. The nonparametric and parametric estimation results are given in Figure 4.6 (on page 99) and Table 4.3 (on page 100) respectively.

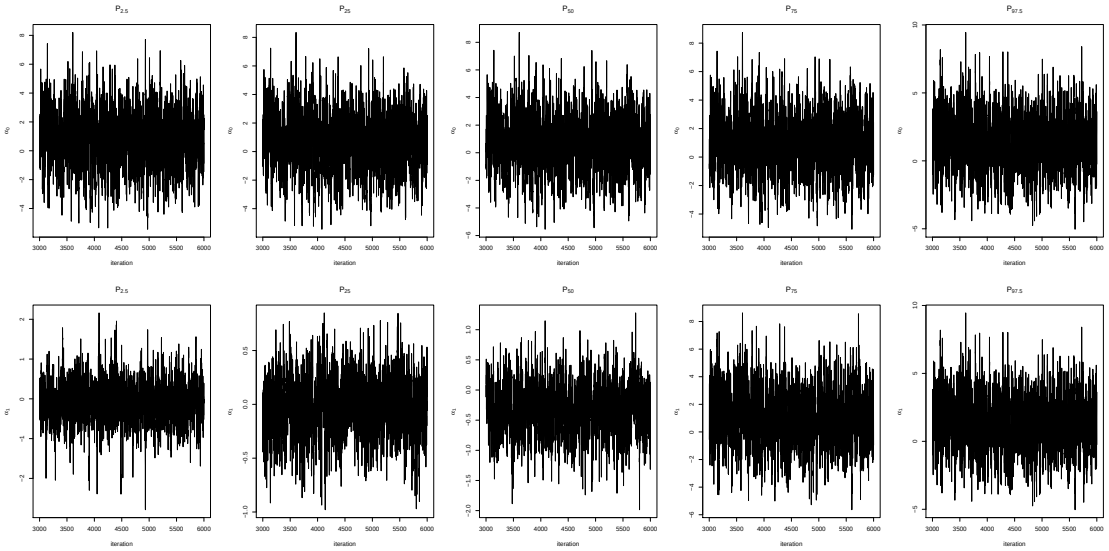

Figure 4.4 Trace plot of MCMC chains for the five percentiles (2.5th, 25th, 50th, 75th and 97.5th) of $\alpha_0(U)$ and $\alpha_1(U)$ by BI analysis w.r.t Model (4.2); The upper five plots are respective to $\alpha_0(U)$ and the lower five plots are respective to $\alpha_1(U)$.

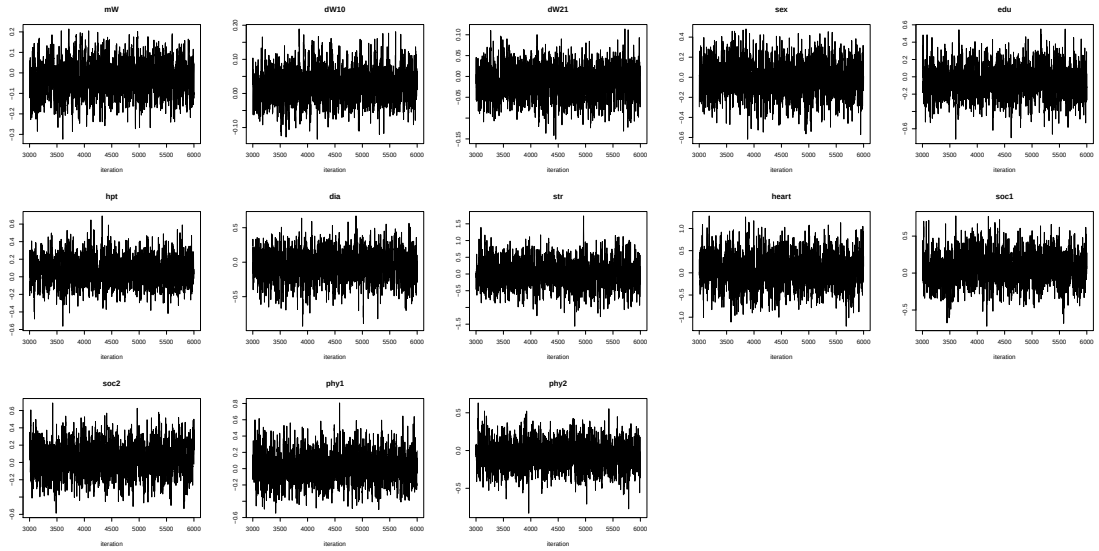

Figure 4.5 Trace plot of MCMC chains for the constant coefficients by BI analysis w.r.t Model (4.2).

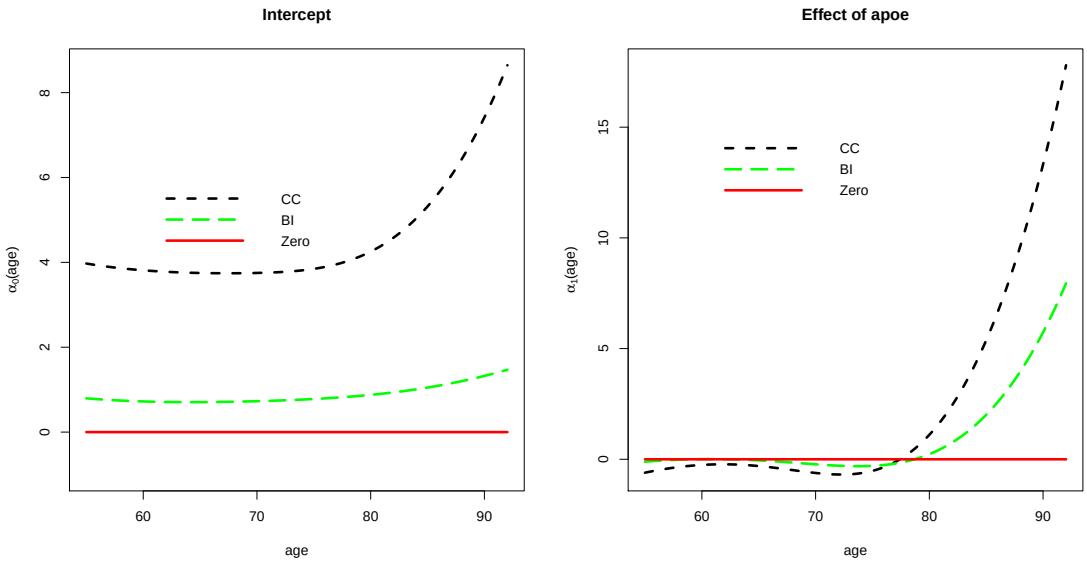

Figure 4.6 Comparion of the estimations of varying-coefficient functions $\alpha_0(U)$ and $\alpha_1(U)$ for CDR using Model (4.2) by CC and BI, while CDR scores are divided into two subsets: CDR $= 0$ and ≥ 0.5.

The estimation results from Bayesian imputation method differs from those with only complete cases. The estimated intercept function indicates that the log-odds of dementia remains constant before age 80 and then rapidly climbs up. The estimated coefficient function for Apolipoprotein E also remains roughly zero before age 78 and then rapidly climbs up, indicating a stronger and stronger positive effect for older subjects. In this case, complete-case analysis produces a much higher log-odds as patients become older than Bayesian imputation method.

Variable	CC	BI
mW	-0.15 (-0.30, -0.02)	-0.03 (-0.20, 0.14)
dW10	0.02 (-0.09, 0.13)	0.03 (-0.06, 0.12)
dW21	-0.03 (-0.13, 0.08)	-0.02 (-0.09, 0.06)
sex	0.05 (-0.60, 0.68)	-0.01 (-0.33, 0.31)
edu	-0.22 (-0.96, 0.51)	-0.05 (-0.39, 0.29)
hpt	0.31 (-0.26, 0.87)	0.07 (-0.21, 0.39)
dia	-0.10 (-0.85, 0.65)	-0.05 (-0.48, 0.34)
str	0.13 (-1.06, 1.41)	0.01 (-0.76, 0.77)
heart	0.23 (-0.90, 1.41)	0.06 (-0.64, 0.78)
soc1	0.09 (-0.69, 0.83)	0.06 (-0.33, 0.48)
soc2	0.14 (-0.50, 0.82)	0.04 (-0.31, 0.39)
phy1	0.09 (-0.57, 0.73)	0.02 (-0.33, 0.40)
phy2	-0.18 (-0.85, 0.44)	-0.07 (-0.40, 0.28)

Table 4.3 Comparison of the estimated posterior means and 95% credible intervals of constant-coefficients for CDR using Model (4.2) by CC and BI, while CDR scores are divided into two subsets: CDR=0 and ≥ 0.5.

In the above table, mW, dW10 and dW21 are the estimated coefficients of the mean of the 3 MMSEs scores, the difference between the 1st-MMSE and baseline

MMSE and the difference between the 2nd-MMSE and 1st-MMSE respectively. From the above table, we conclude that only the mean of the 3 MMSEs scores (mW) is significant (-0.15 (-0.29, -0.02)) by complete-case analysis, while none of constant coefficient predictors is significant by Bayesian imputation method.

CHAPTER 5

Conclusion

In Chapter 2, we have proposed fitting the varying-coefficient model for cross-sectional normal response variables by using splines and Bayesian techniques. For normal longitudinal data, we have proposed fitting the varying-coefficient mixed model which adds random effects to the varying-coefficient model. We achieved in fitting both models, which could be seen from our simulation studies by checking the results, especially from the coverage probabilities. We have demonstrated that the model successfully explains the random error within each subject among the multiple measures by adding a random effect and that the model is quite easy to estimate under Bayesian context. For longitudinal normal response data, we have

also considered the situation when missing responses are involved. We have compared the estimation results when adapting different approaches to fit the model under Bayesian context.

In Chapter 3, we have proposed fitting the varying-coefficient model by splines and Bayesian methods for cross-sectional binary response variables. The fitting of the model was executed using data augmentation approach by adding auxiliary variables, and turned out to be good when checked by simulations. We have shown that the method of using data augmentation approach leads to direct sampling from the conditional distribution and avoids the Metropolis-Hastings accept/reject steps which are commonly encountered under Bayesian binary regression, thus making Bayesian estimation process easy to implement. For cross-sectional binary response data involving missing value, we have also compared the estimation results when using different approaches to fit the model under Bayesian context.

In Chapter 4, we have analyzed the real data by implementing the methodology described in Chapter 2 and 3. The result is reasonable from the medical experts' view, e.g. Feng et al. (2012).

The proposal of this study has provided an alternative method for fitting varying-coefficient model, especially when the model involves binary response variable or missing data which is relatively complicated and provided an alternative method for fitting varying-coefficient mixed effects model for longitudinal data.

This study did not consider the situation when binary longitudinal response

variables are involved. This is because the estimation processes are too time-consuming in both cases under our proposed model.

For future works, one could consider fitting the varying-coefficient mixed effects model for binary longitudinal data which is a direct extension of our works. One could also extend the work to more general area which contains generalized regression model. One could also consider fitting the varying-coefficient model involving miss data under NMAR pattern.

Bibliography

[1]. Aldrich, J. (1997), 'R. A. Fisher and the making of maximum likelihood 1912-1922', *Statistical Science* **12**(3), 162–176.

[2]. Andrews, D. F. and Mallows, C. L. (1974), 'Scale mixtures of normal distributions', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **36**(1), 99–102.

[3]. Biller, C. and Fahrmeir, L. (2001), 'Bayesian varying-coefficient models using adaptive regression splines', *Statistical Modelling* **1**(3), 195–211.

[4]. Booth, D. E. (2000), 'Analysis of incomplete multivariate data', *Technometrics* **42**(2), 213–214.

[5]. Brezger, A. and Lang, S. (2006), 'Generalized structured additive regression based on bayesian p-splines', *Computational Statistics and Data Analysis* **50**(4), 967–991.

[6]. Brooks, S. P. and Gelman, A. (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**(4), 434–455.

[7]. Cheng, M., Zhang, W. and Chen, L. (2009), 'Statistical estimation in generalized multiparameter likelihood models', *Journal of the American Statistical Association* **104**(487), 1179–1191.

[8]. Cottet, R., Kohn, R. J. and Nott, D. J. (2008), 'Variable selection and model averaging in semiparametric overdispersed generalized linear models', *Journal of the American Statistical Association* **103**(482), 661–671.

[9]. Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer.

[10]. Efron, B. and Gong, G. (1983), 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *The American Statistician* **37**(1), 36–48.

[11]. Eilers, P. H. and Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical Science* **11**(2), 89–102.

[12]. Eubank, R., Huang, C., Maldonado, Y. M., Wang, N., Wang, S. and Buchanan, R. (2004), 'Smoothing spline estimation in varying-coefficient models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(3), 653–667.

[13]. Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, Vol. 157, CRC Press.

[14]. Fahrmeir, L., Kneib, T. and Lang, S. (2004), 'Penalized structured additive regression for space-time data: a bayesian perspective', *Statistica Sinica* **14**(3), 731–762.

[15]. Fan, J., Yao, Q. and Cai, Z. (2003), 'Adaptive varying-coefficient linear models', *Journal of the Royal Statistical Society: series B (Statistical Methodology)* **65**(1), 57–80.

[16]. Fan, J. and Zhang, J.-T. (2000), 'Two-step estimation of functional linear models with applications to longitudinal data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2), 303–322.

[17]. Fan, J. and Zhang, W. (1999), 'Statistical estimation in varying coefficient models', *The Annals of Statistics* **27**(5), 1491–1518.

[18]. Feng, L., Chong, M., Lim, W. and Ng, T. (2012), 'The modified mini-mental state examination test: normative data for singapore chinese older adults and its

performance in detecting early cognitive impairment', *Singapore Medical Journal* **53**(7), 458–462.

[19]. Ferrari, C., Xu, W., Wang, H., Winblad, B., Sorbi, S., Qiu, C. and Fratiglioni, L. (2013), 'How can elderly apolipoprotein e $\varepsilon 4$ carriers remain free from dementia?', *Neurobiology of Aging* **34**(1), 13–21.

[20]. Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)', *Bayesian Analysis* **1**(3), 515–534.

[21]. Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC.

[22]. Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.

[23]. Green, P. J., Silverman, B. W., Silverman, B. W. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*, Chapman & Hall London.

[24]. Greenland, S. and Finkle, W. D. (1995), 'A critical look at methods for handling missing covariates in epidemiologic regression analyses', *American Journal of Epidemiology* **142**(12), 1255–1264.

[25]. Haan, M. N., Shemanski, L., Jagust, W. J., Manolio, T. A. and Kuller, L. (1999), 'The role of apoe $\varepsilon 4$ in modulating effects of other risk factors for cognitive decline in elderly persons', *Journal of the American Medical Association* **282**(1), 40–46.

[26]. Hastie, T. (1996), 'Pseudosplines', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **58**(2), 379–396.

[27]. Hastie, T. and Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **55**(4), 757–796.

[28]. Holmes, C. C. and Held, L. (2006), 'Bayesian auxiliary variable models for binary and multinomial regression', *Bayesian Analysis* **1**(1), 145–168.

[29]. 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**(4), 809–822.

[30]. Lambert, P. and Eilers, P. H. (2005), 'Bayesian proportional hazards model with time-varying regression coefficients: a penalized poisson regression approach', *Statistics in Medicine* **24**(24), 3977–3989.

[31]. Lang, S. and Brezger, A. (2004), 'Bayesian p-splines', *Journal of Computational and Graphical Statistics* **13**(1), 183–212.

[32]. Li, J. and Palta, M. (2009), 'Bandwidth selection through cross-validation for

semi-parametric varying-coefficient partially linear models', *Journal of Statistical Computation and Simulation* **79**(11), 1277–1286.

[33]. Li, J. and Wong, W. K. (2009), 'A semi-parametric analysis for identifying scleroderma patients responsive to an anti-fibrotic agent', *Contemporary Clinical Trials* **30**(2), 105–113.

[34]. Li, J., Xia, Y., Palta, M. and Shankar, A. (2009), 'Impact of unknown covariance structures in semiparametric models for longitudinal data: An application to wisconsin diabetes data', *Computational Statistics and Data Analysis* **53**(12), 4186–4197.

[35]. Lin, X. and Carroll, R. J. (2001), 'Semiparametric regression for clustered data using generalized estimating equations', *Journal of the American Statistical Association* **96**(455), 1045–1056.

[36]. Little, R. J. (1992), 'Regression with missing x's: a review', *Journal of the American Statistical Association* **87**(420), 1227–1237.

[37]. Little, R. J. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Wiley.

[38]. Lu, Z., Steinskog, D. J., Tjøstheim, D. and Yao, Q. (2009), 'Adaptively varying-coefficient spatiotemporal models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(4), 859–880.

[39]. Marley, J. K. and Wand, M. P. (2010), 'Non-standard semiparametric regression via brugs', *Journal of Statistical Software* **37**(5), 1–30.

[40]. Morris, J. C. (1993), 'The clinical dementia rating (cdr): current version and scoring rules', *Neurology* **43**(11), 2412–2414.

[41]. Nott, D. J. and Li, J. (2010), 'A sign based loss approach to model selection in nonparametric regression', *Statistics and Computing* **20**(4), 485–498.

[42]. Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

[43]. Rubin, D. B. (2009), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

[44]. Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

[45]. Sahadevan, S., Tan, N. J., Tan, T. and Tan, S. (1997), 'Cognitive testing of elderly chinese people in singapore: influence of education and age on normative scores', *Age and Ageing* **26**(6), 481–486.

[46]. Schafer, J. L. (2010), *Analysis of Incomplete Multivariate Data*, CRC Press.

[47]. Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag New York.

[48]. Shi, J. Q., Wang, B., Will, E. J. and West, R. M. (2012), 'Mixed-effects gaussian process functional regression models with application to dose-response curve prediction', *Statistics in Medicine* **31**(26), 3165–3177.

[49]. Tanner, M. A. and Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association* **82**(398), 528–540.

[50]. Vach, W. (1994), *Logistic Regression with Missing Values in the Covariates*, Springer-Verlag New York.

[51]. Wang, L., Li, H. and Huang, J. Z. (2008), 'Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements', *Journal of the American Statistical Association* **103**(484), 1556–1569.

[52]. Wang, X., Chen, M.-H. and Yan, J. (2013), 'Bayesian dynamic regression models for interval censored survival data with application to children dental health', *Lifetime data analysis* pp. 1–20.

[53]. White, I. R. and Carlin, J. B. (2010), 'Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values', *Statistics in Medicine* **29**(28), 2920–2931.

[54]. Wu, H. and Zhang, J.-T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*, Wiley-Interscience.

[55]. Xia, Y., Zhang, W. and Tong, H. (2004), 'Efficient estimation for semivarying-coefficient models', *Biometrika* **91**(3), 661–681.

[56]. Zhang, M., Katzman, R., Salmon, D., Jin, H., Cai, G., Wang, Z., Qu, G., Grant, I., Yu, E., Levy, P. et al. (1990), 'The prevalence of dementia and alzheimer's disease in shanghai, china: impact of age, gender, and education', *Annals of Neurology* **27**(4), 428–437.

[57]. Zhang, W. (2011), 'Identification of the constant components in generalised semi-varying coefficient models by cross–validation', *Statistica Sinica* **21**(4), 1913–1929.

[58]. Zhang, Z., Zahner, G. E., Román, G. C., Liu, J., Hong, Z., Qu, Q., Liu, X., Zhang, X., Zhou, B., Wu, C. et al. (2005), 'Dementia subtypes in china: prevalence in beijing, xian, shanghai, and chengdu', *Archives of Neurology* **62**(3), 447.