Bayesian Analysis (2016) **11**, Number 4, pp. 1137–1163

Bayesian Semiparametric Inference on Functional Relationships in Linear Mixed Models

Seonghyun Jeong[*] and Taeyoung Park[†]

Abstract. Regression models with varying coefficients changing over certain underlying covariates offer great flexibility in capturing a functional relationship between the response and other covariates. This article extends such regression models to include random effects and to account for correlation and heteroscedasticity in error terms, and proposes an efficient new data-driven method to estimate varying regression coefficients via reparameterization and partial collapse. The proposed methodology is illustrated with a simulated study and longitudinal data from a study of soybean growth.

MSC 2010 subject classifications: Primary 62F15; secondary 62J99.

Keywords: longitudinal data, random effects, model selection, partial collapse, panel data, reparameterization.

1 Introduction

Nonparametric statistical methods are well known to reduce modeling bias and increase the flexibility and applicability of parametric statistical models, but their use can be hampered by the curse of dimensionality. Many studies have been conducted to overcome the problem by imposing semiparametric structures to nonparametric models. An additive model (Breiman and Friedman, 1985; Buja et al., 1989), a partially linear model (Engle et al., 1986), and a varying-coefficient model (Cleveland et al., 1991; Hastie and Tibshirani, 1993) are among examples that represent such attempts. Although these models were motivated by different ideas and concepts, they can be viewed as special cases of a so-called partially linear varying-coefficient model (Ahmad et al., 2005; Fan and Huang, 2005), which includes both fixed effects and varying regression coefficients that change over some underlying covariates such as time, temperature, and geographical locations.

The partially linear varying-coefficient model is motivated by applications rather than simply a desire to extend a mathematical model (Fan and Zhang, 2008; Park et al., 2015). In many applications, there is often interest in assessing a relationship between a response variable and a certain covariate, which can be confounded with another underlying covariate called an effect modifier. Then one may consider a model in which the main effects of both covariates on the response are added with their interaction term. The interaction term, however, fails to account for nonlinear confounding and is

[*] Department of Applied Statistics, Yonsei University, Seoul 03722, Korea
[†] Department of Applied Statistics, Yonsei University, Seoul 03722, Korea, tpark@yonsei.ac.kr

DOI: 10.1214/15-BA987

not even easily interpretable when both covariates are continuous. In such cases, the varying-coefficient model allows the effect of a certain covariate on a response variable to vary with an effect modifier, which permits a nonlinear interaction between the certain covariate and effect modifier in an easily interpretable manner. For example, in studying the effects of risk factors on infant mortality in China, there may be need to account for a potential effect modifier. That is, it is well known that China has undergone a dramatic change since 1949, so that it is not reasonable to assume the constant effects of risk factors on infant mortality over time (Cheng et al., 2009). In this case, the varying-coefficient model can include the main effects of risk factors on infant mortality and allow the corresponding regression coefficients to smoothly vary as a function of time; see also our longitudinal data example in Section 5.

Different nonparametric approaches have been proposed to approximate an unknown smooth function of a varying regression coefficient. Estimation of the varying-coefficient using a smoothing spline has been studied in Hastie and Tibshirani (1993) and Hoover et al. (1998), while a kernel smoothing method is used in Fan and Zhang (1999), Wu and Chiang (2000), and Cai (2002). These approaches, however, require estimating the optimal value of a tuning parameter, that is, either a smoothing parameter or a bandwidth, which is a continuous variable. An approximate solution to the optimization based on cross-validation is then derived by discretizing the support of the tuning parameter. There also has been an issue of dealing with multiple tuning parameters. In the presence of multiple varying-coefficients, local polynomial fitting based on a single tuning parameter is not possible to achieve plausible smoothing for all varying-coefficients simultaneously. The curse of dimensionality is then inevitably incurred as the dimension of tuning parameters increases to improve model flexibility with multiple varying-coefficients.

In addition to the smoothing spline and kernel smoothing method, a regression spline, defined as a linear combination of spline basis terms such as B-spline or radial basis functions, has been also used to model a varying regression coefficient, where the basis terms are defined by a set of knots (Huang et al., 2002, 2004). Because the shape and smoothness of an unknown function modeled as a regression spline is determined by knot placement, the estimation of a varying regression coefficient based on a regression spline typically uses the number of *equally spaced* knots as a tuning parameter. In this way, a regression spline no longer requires approximation to the support of the tuning parameter, and cross-validation can be used to determine the optimal number of knots. However, methods based on cross-validation are again hampered by the curse of dimensionality, and simultaneous inference on the number and location of knots becomes a challenging task with the regression spline. To simultaneously determine an *unknown* number of knots and their *unknown* locations from data, we develop a data-driven method based on the notion of Bayesian free-knot splines (Smith and Kohn, 1996; Denison et al., 1998; DiMatteo et al., 2001; Kohn et al., 2001; Park et al., 2012). By doing so, we avoid biases in the estimates of fixed effects whose inference depends on the unknown number and location of knots in spline basis terms. In the context of a varying regression coefficient, Biller and Fahrmeir (2001) proposed a method based on the reversible-jump Markov chain Monte Carlo by treating the number and location of knots as random variables. The method, however, yields results that are sensitive to the choice of a prior distribution on the number of knots and includes only varying-coefficients without fixed or random effects, as compared to our proposed method.

In a (partially linear) varying-coefficient model, error terms are typically assumed to follow a zero-mean Gaussian process with *known* covariance structure. Such restriction is, however, sometimes inappropriate because it fails to fully account for correlated and heteroscedastic errors that are commonly present in longitudinal or spatial data. In this article, we thus generalize the existing partially linear varying-coefficient model to allow for an *unknown* covariance matrix of errors as well as random effects. This extension allows a parsimonious and flexible modeling framework for the marginal covariance structure of the Gaussian process. In this paper, we thus develop an efficient data-driven method for estimating varying regression coefficients with the unknown number and location of knots in spline basis terms, while simultaneously accounting for correlation and heteroscedasticity in a covariance structure as well as random effects.

The rest of this article is organized as follows. In Section 2, we develop a partially linear varying-coefficient mixed models with correlated and heteroscedastic errors based on radial basis functions and covariance matrix decomposition. Section 3 proposes an efficient posterior simulation method via reparameterization and partial collapse. In Section 4, a simulation study is conducted to illustrate the performance of the proposed method. Section 5 applies the methodology to the soybean growth data, and Section 6 concludes with discussion.

2 Modeling procedures

2.1 Model description

Let Y_{ij} be the jth observation within group i and let n_i be the total number of observations for group i. A general form of a partially linear varying-coefficient mixed model is given by

$$Y_{ij} = \mathbf{w}_{ij}^\top \boldsymbol{\alpha}(\mathbf{u}_{ij}) + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\xi}_i + \epsilon_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, n_i, \tag{1}$$

where $\mathbf{w}_{ij} = (w_{ij1}, \dots, w_{ijp_\mathbf{w}})^\top$, $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp_\mathbf{x}})^\top$, and $\mathbf{z}_{ij} = (z_{ij1}, \dots, z_{ijp_\mathbf{z}})^\top$ are $p_\mathbf{w} \times 1$, $p_\mathbf{x} \times 1$, and $p_\mathbf{z} \times 1$ vectors of covariates, respectively, $\mathbf{u}_{ij} = (u_{ij1}, \dots, u_{ijp_\mathbf{w}})^\top$ is a $p_\mathbf{w} \times 1$ vector of underlying effect modifiers that change the effects of $\mathbf{w}_{ij}$ on Y_{ij} in a nonparametric way, $\boldsymbol{\alpha}(\mathbf{u}_{ij}) = (\alpha_1(u_{ij1}), \dots, \alpha_{p_\mathbf{w}}(u_{ijp_\mathbf{w}}))^\top$ is a $p_\mathbf{w} \times 1$ vector of varying-coefficients to $\mathbf{w}_{ij}$, $\boldsymbol{\beta}$ is a $p_\mathbf{x} \times 1$ vector of fixed effects, $\boldsymbol{\xi}_i$ is a $p_\mathbf{z} \times 1$ vector of random effects following an independent and identical multivariate normal distribution, $\boldsymbol{\xi}_i \overset{\text{iid}}{\sim} \mathrm{N}_{p_\mathbf{z}}(\mathbf{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi}$ a $p_\mathbf{z} \times p_\mathbf{z}$ positive definite covariance matrix, and ϵ_{ij}'s are error terms following a normal distribution with a certain covariance structure which is assumed independent across groups, that is, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})^\top \overset{\text{ind}}{\sim} \mathrm{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$, with $\boldsymbol{\Sigma}_i$ being an $n_i \times n_i$ positive definite matrix, and $\boldsymbol{\xi}_i \perp \boldsymbol{\epsilon}_i$. We refer to the covariance matrix $\boldsymbol{\Sigma}_i$ as a within-group covariance matrix. The within-group covariance is often assumed to be $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{R}_i$ with a known positive definite matrix $\mathbf{R}_i$ (Laird and Ware, 1982; van Dyk, 2000; Park and Min, 2014). Such restriction do not flexibly model a marginal covariance matrix, causing bias in the estimation of a mean response. We thus use a within-group covariance structure that is known up to some variance parameters. Specifically, we define $\mathbf{H}_i(\boldsymbol{\eta}) = \mathrm{diag}(\exp(\mathbf{v}_{i1}^\top \boldsymbol{\eta}/2), \ldots, \exp(\mathbf{v}_{in_i}^\top \boldsymbol{\eta}/2))$ and

$$\mathbf{P}_i(\rho) = \begin{pmatrix} 1 & \rho^{\|\mathbf{t}_{i1}-\mathbf{t}_{i2}\|} & \cdots & \rho^{\|\mathbf{t}_{i1}-\mathbf{t}_{in_i}\|} \\ \rho^{\|\mathbf{t}_{i1}-\mathbf{t}_{i2}\|} & 1 & \cdots & \rho^{\|\mathbf{t}_{i2}-\mathbf{t}_{in_i}\|} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{\|\mathbf{t}_{i1}-\mathbf{t}_{in_i}\|} & \rho^{\|\mathbf{t}_{i2}-\mathbf{t}_{in_i}\|} & \cdots & 1 \end{pmatrix},$$

where $\|\cdot\|$ denotes the L^2-norm, $\mathbf{v}_{ij} = (v_{ij1}, \ldots, v_{ijp_{\mathbf{v}}})^\top$ and $\mathbf{t}_{ij} = (t_{ij1}, \ldots, t_{ijp_{\mathbf{t}}})^\top$ are $p_{\mathbf{v}} \times 1$ and $p_{\mathbf{t}} \times 1$ vectors of covariates, respectively, $\boldsymbol{\eta}$ is a $p_{\mathbf{v}} \times 1$ vector of parameters representing heteroscedasticity related to the covariates $\mathbf{v}_{ij}$, and ρ is a parameter representing non-negative correlation. The covariance matrix can then be decomposed into $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{H}_i(\boldsymbol{\eta}) \mathbf{P}_i(\rho) \mathbf{H}_i(\boldsymbol{\eta})$ in terms of variance parameters (σ^2, $\boldsymbol{\eta}$, ρ), so that the conditional covariance structure model is written as

$$\begin{aligned} \log \mathrm{Var}(Y_{ij}|\boldsymbol{\xi}_i) &= \log(\sigma^2) + \mathbf{v}_{ij}^\top \boldsymbol{\eta}, \\ \mathrm{Corr}(Y_{ij}, Y_{ik}|\boldsymbol{\xi}_i) &= \rho^{\|\mathbf{t}_{ij}-\mathbf{t}_{ik}\|}, \end{aligned}$$

where $\log(\sigma^2)$ is considered an intercept of the model. The correlation matrix $\mathbf{P}_i(\rho)$ implies that different observations within the same group become more closely correlated as the Euclidean distance of their $\mathbf{t}_{ij}$'s decreases. This formulation of the correlation matrix is useful in a temporal or spatial setting if $\mathbf{t}_{ij}$ represents time or geographical locations. In particular, when $\mathbf{t}_{ij}$ denotes the univariate time of observation, $\mathbf{P}_i(\rho)$ is reduced to a serial correlation matrix of continuous-time AR(1) models in Jones and Boadi-Boateng (1991). When $\mathbf{P}_i(\rho)$ is an $n_i \times n_i$ identity matrix, $\boldsymbol{\Sigma}_i$ can be viewed as a covariance matrix specified in heteroscedastic linear regression models (Chan et al., 2006; Leslie et al., 2007).

Note that, for identifiability, the sets of covariates $\mathbf{w}_{ij}$ and $\mathbf{x}_{ij}$ should be disjoint because the varying-coefficients $\boldsymbol{\alpha}(\mathbf{u}_{ij})$ include constant effects. In practice, it is of interest to partition the covariates into two groups, depending on whether the effect of each covariate is varying or constant. Without any prior knowledge on covariates with constant effects, it is recommended that all covariates be assigned to $\mathbf{w}_{ij}$, assuming interactions with $\mathbf{u}_{ij}$. Then our proposed data-driven method in Section 2.2 can distinguish covariates with constant effects from those with varying effects via Bayesian knot selection, circumventing the problem of overfitting; refer to Section 4.3.

If the effects of covariates $\mathbf{w}_{ij}$ on Y_{ij} are believed to vary with a single effect modifier, we let $u_{ijk} = u_{ij}^*$ for $k = 1, \ldots, p_{\mathbf{w}}$ and the model in (1) is rewritten as

$$Y_{ij} = \mathbf{w}_{ij}^\top \boldsymbol{\alpha}(u_{ij}^*) + \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\xi}_i + \epsilon_{ij}, \tag{2}$$

where $\boldsymbol{\alpha}(u_{ij}^*) = (\alpha_1(u_{ij}^*), \ldots, \alpha_{p_{\mathbf{w}}}(u_{ij}^*))^\top$. When the common effect modifier u_{ij}^* represents the time of observation, the model in (2) is reduced to a time-varying-coefficient model that has been widely studied in the literature (Hoover et al., 1998; Cai et al., 2000; Huang et al., 2002; Liang et al., 2003; Huang et al., 2004; Cai, 2007; Li et al., 2011). Unlike the existing time-varying-coefficient model, the model in (2) simultaneously accounts for both the random effects and non-spherical within-group errors to flexibly model marginal errors in a longitudinal setting. If we let $\mathbf{v}_{ij} = \mathbf{t}_{ij} = u^*_{ij}$ for the within-group errors and u^*_{ij} represents the time of observation, the model in (2) can capture serial correlation and monotone heteroscedasticity over time.

2.2 Approximation with a regression spline

To closely approximate an unknown smooth function of a varying-coefficient, we propose to express $\alpha_k(\cdot)$ as a linear combination of spline basis terms. In the regression spline, the number of knots for the basis terms controls the shape and smoothness of a fitted curve. In this paper, we develop a data-driven method to determine the unknown number and location of knots while simultaneously fitting the other model parameters. Because the number of knots is assumed unknown, the dimension of the spline basis terms is not fixed but varying. To deal with the varying-dimensional spline basis terms, we consider a set of potential basis terms and let data suggest a set of plausible ones. In particular, we use the $(L_k + 2) \times 1$ vector of potential radial basis functions defined by

$$\mathbf{b}_k(u) = \left\{1,\ u,\ \left|\frac{u-\tau_{k1}}{c_k}\right|^2 \log\left(\left|\frac{u-\tau_{k1}}{c_k}\right|^2\right), \dots, \left|\frac{u-\tau_{kL_k}}{c_k}\right|^2 \log\left(\left|\frac{u-\tau_{kL_k}}{c_k}\right|^2\right)\right\}^\top, \quad (3)$$

where the abscissae, $\tau_{k1}, \dots, \tau_{kL_k}$, are candidates for knot locations which lie in the range of the kth underlying effect modifier, and c_k denotes a predetermined scale factor that is set to the sample standard deviation of the kth underlying effect modifier in our study. Then we introduce the $(L_k + 1) \times 1$ vector of latent knot indicator variables $\boldsymbol{\gamma}_k = (\gamma_{k0}, \gamma_{k1}, \dots, \gamma_{kL_k})^\top$, where $\gamma_{km} = 1$ if the $(m + 2)$th term in (3) is used as a basis function and 0 otherwise, for $m = 0, 1, \dots, L_k$. Note that the first term in (3) corresponds to a constant basis function that is always included in the model. If the effect of a covariate is truly constant, all knot indicator variables will become zero and overfitting can be avoided by using only the first term in (3); see Section 4.3 for details.

One issue with our model specification is the choice of candidates for knot locations. The knot candidates are typically specified by the equally spaced interior points or observed order statistics of an underlying effect modifier. When observed data are locally sparse, it is more stable to use the observed order statistics. While the number L_k of knot candidates is desirable to be large enough to capture an unknown functional structure, spline estimates with too many knot candidates can yield a high local variance. As a guideline, it is recommended to use 20 to 30 knot candidates if an underlying effect modifier is continuous. If an underlying effect modifier is ordinal, the number of knot candidates must be smaller than the number of distinct values.

For notational simplicity, we define $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \dots, \boldsymbol{\gamma}_{p_\mathbf{w}}^\top)^\top$, $J_k(\boldsymbol{\gamma}_k) = \sum_{m=0}^{L_k} \gamma_{km}$, and $J(\boldsymbol{\gamma}) = \sum_{k=1}^{p_\mathbf{w}} J_k(\boldsymbol{\gamma}_k)$. Letting $\mathbf{b}_{\boldsymbol{\gamma}_k}(\cdot)$ denote a $(J_k(\boldsymbol{\gamma}_k)+1) \times 1$ vector of basis functions determined by $\boldsymbol{\gamma}_k$ and letting $\boldsymbol{\phi}_{\boldsymbol{\gamma}_k}$ denote a $(J_k(\boldsymbol{\gamma}_k)+1) \times 1$ vector of the corresponding coefficients, each varying-coefficient $\alpha_k(\cdot)$ can be approximated as

$$\alpha_k(U_k) \approx \mathbf{b}_{\boldsymbol{\gamma}_k}^\top(U_k)\boldsymbol{\phi}_{\boldsymbol{\gamma}_k},$$

where U_k denotes the kth underlying effect modifier. The model in (1) is then expressed in a matrix form as

$$\mathbf{Y}_i = \mathbf{W}_i^\star \mathbf{B}_{\boldsymbol{\gamma},i}\boldsymbol{\phi}_{\boldsymbol{\gamma}} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \tag{4}$$

where $\mathbf{Y}_i$ is an $n_i \times 1$ vector of observations for group i, $\mathbf{W}_i^\star = \bigoplus_{j=1}^{n_i} \mathbf{w}_{ij}^\top$ is an $n_i \times n_i p_\mathbf{w}$ design matrix, $\mathbf{B}_{\boldsymbol{\gamma},i} = (\bigoplus_{k=1}^{p_\mathbf{w}} \mathbf{b}_{\boldsymbol{\gamma}_k}(u_{i1k}), \ldots, \bigoplus_{k=1}^{p_\mathbf{w}} \mathbf{b}_{\boldsymbol{\gamma}_k}(u_{in_ik}))^\top$ is an $n_i p_\mathbf{w} \times (J(\boldsymbol{\gamma}) + p_\mathbf{w})$ matrix of radial basis functions, $\boldsymbol{\phi}_{\boldsymbol{\gamma}} = (\boldsymbol{\phi}_{\boldsymbol{\gamma}_1}^\top, \ldots, \boldsymbol{\phi}_{\boldsymbol{\gamma}_p}^\top)^\top$ is a $(J(\boldsymbol{\gamma}) + p_\mathbf{w}) \times 1$ vector of the corresponding coefficients, and $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^\top$ and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i})^\top$ are $n_i \times p_\mathbf{x}$ and $n_i \times p_\mathbf{z}$ design matrices, respectively. The notation $\bigoplus$ denotes the direct sum of matrices or vectors. Then the model in (4) is equivalent to linear mixed models with $(\boldsymbol{\phi}_{\boldsymbol{\gamma}}, \boldsymbol{\beta})$ treated as fixed effects, except that the dimension of $\boldsymbol{\phi}_{\boldsymbol{\gamma}}$ is varying. As a result, the marginal covariance matrix is computed as $\mathrm{Cov}(\mathbf{Y}_i) = \sigma^2\mathbf{H}_i(\boldsymbol{\eta})\mathbf{P}_i(\rho)\mathbf{H}_i(\boldsymbol{\eta}) + \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^\top$, which flexibly models marginal variation accounting for covariate effects. When $\boldsymbol{\gamma}$ is given, we consider adopting the method of partial collapse after reparameterization for efficient Bayesian inference on the linear mixed models; see Section 3.

In practice, the most popular choice of basis functions in the context of a regression spline model for a varying-coefficient is a B-spline basis due to its numerically stable computation as compared to polynomial spline bases. However, we employ radial basis functions to circumvent the drawbacks of the B-spline bases arising in a Bayesian knot selection procedure. Specifically, when B-spline bases are used for the regression spline, the addition or deletion of a knot does not simply add or delete a basis term, but forms new basis terms by altering some neighboring bases (Thompson and Rosen, 2008). Because knot-selection does not correspond to the selection of basis functions and it is required to modify some columns of $\mathbf{W}_i^\star \mathbf{B}_{\boldsymbol{\gamma},i}$ every iteration, the use of B-splines makes an iterative algorithm inefficient and time-consuming. In the case of radial basis functions, however, the selection of knots corresponds to the selection of basis terms because $\mathbf{b}_{\boldsymbol{\gamma}_k}(\cdot)$ is a subvector of $\mathbf{b}_k(\cdot)$, and thus the design matrix $\mathbf{W}_i^\star \mathbf{B}_{\boldsymbol{\gamma},i}$ consists of selected columns of $\mathbf{W}_i^\star \mathbf{B}_i$, where $\mathbf{B}_i = (\bigoplus_{k=1}^{p_\mathbf{w}} \mathbf{b}_k(u_{i1k}), \ldots, \bigoplus_{k=1}^{p_\mathbf{w}} \mathbf{b}_k(u_{in_ik}))^\top$. This difference makes an iterative algorithm with radial basis functions computationally much faster than with B-spline bases.

3 Bayesian semiparametric inference

3.1 Prior specification

The indicator variables $\gamma_{k0}, \ldots, \gamma_{kL_k}$ for the kth varying-coefficient are assigned independent and identical Bernoulli prior distributions,

$$\gamma_{km}|\pi \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi),\ k = 1, \ldots, p_\mathbf{w},\ m = 0, \ldots, L_k,$$

where the hyper-parameter π has a beta prior distribution,

$$\pi \ \sim \ \mathrm{Beta}(a, b),$$

where a and b are considered known. We assign a multivariate normal prior distribution to the fixed effect $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\boldsymbol{\phi}_{\boldsymbol{\gamma}}^\top, \boldsymbol{\beta}^\top)^\top$,

$$\boldsymbol{\theta}_{\boldsymbol{\gamma}} | (\boldsymbol{\gamma}, \kappa, \rho, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\Lambda}) \ \sim \ \mathrm{N}_{J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}} \left(\mathbf{0}, \kappa\sigma^2 \left[\sum\nolimits_{i=1}^{N} \mathbf{C}_{\boldsymbol{\gamma},i}^\top \boldsymbol{\Omega}_i(\rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \mathbf{C}_{\boldsymbol{\gamma},i} \right]^{-1} \right),$$

where $\mathbf{C}_{\boldsymbol{\gamma},i} = [\mathbf{W}_i^\star \mathbf{B}_{\boldsymbol{\gamma},i}, \mathbf{X}_i]$ is an $n_i \times (J(\boldsymbol{\gamma}) + p_{\mathbf{w}} + p_{\mathbf{x}})$ design matrix, $\boldsymbol{\Omega}_i(\rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) = (\mathbf{H}_i(\boldsymbol{\eta})\mathbf{P}_i(\rho)\mathbf{H}_i(\boldsymbol{\eta}) + \mathbf{Z}_i\boldsymbol{\Lambda}\mathbf{Z}_i^\top)^{-1}$ is an $n_i \times n_i$ weight matrix, κ is considered as a dispersion factor, and $\boldsymbol{\Lambda}$ is a part of the covariance matrix of random effects such that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Lambda}$; see Section 3.2 for details. This prior distribution can be viewed as Zellner's g-prior (Zellner, 1986) for the linear mixed models, which facilitates posterior computation by reducing and simplifying some terms in the sampling scheme and has an invariance property with respect to the scale of covariates and observations; refer to Kohn et al. (2001) for the fixed-effects models with independent and identically distributed errors. To get minimal prior information on $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ with $\boldsymbol{\gamma}$ fixed, κ needs to be sufficiently large. In the limiting case when $\kappa \to \infty$, however, the prior specification does not have vague prior information because $p(\gamma_{km} = 1 | \boldsymbol{\gamma}_{-km}, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})$ goes to 0 for all k and m; refer to Step 1 in Section 3.2. Thus, it is difficult to choose a fixed value of κ representing vague prior information. Some authors refer to this problem as Bartlett's paradox of g-priors (Clyde and George, 2004; Liang et al., 2008). To circumvent such difficulty, we treat κ as a random variable by assigning a scaled inverse chi-squared prior distribution,

$$\kappa \ \sim \ qN/\chi_q^2,$$

where N is the number of groups and q is a constant fixed in advance. Considering the parameter space of ρ, we assign a beta prior distribution to ρ,

$$\rho \ \sim \ \mathrm{Beta}(c, d),$$

where c and d are fixed constants. To construct a Metropolis–Hastings proposal kernel using the method in Gamerman (1997), the parameters representing heteroscedasticity $\boldsymbol{\eta}$ are assigned a multivariate normal prior distribution,

$$\boldsymbol{\eta} \ \sim \ \mathrm{N}_{p_{\mathbf{v}}}(\mathbf{m}, \mathbf{S}),$$

with fixed constants $\mathbf{m}$ and $\mathbf{S}$; see Appendix A. To complete a Bayesian specification of the proposed model, we choose the following conjugate prior distributions,

$$\begin{aligned} \sigma^2 \ &\sim \ rs^2/\chi_r^2, \\ \boldsymbol{\Lambda}|\sigma^2 \ &\sim \ \mathrm{IW}(g, \mathbf{D}/\sigma^2), \end{aligned}$$

for the remaining model parameters, where IW stands for an inverse Wishart distribution. In our study, the constants a, b, q, c, d, $\mathbf{m}$, $\mathbf{S}$, r, s^2, g, and $\mathbf{D}$ are fixed in advance to make the corresponding prior distribution diffuse.

3.2 Efficient posterior computation

For efficient posterior computation, we consider (partially) collapsing some model parameters out of the proposed model (Liu, 1994; van Dyk and Park, 2008). First, the knot-indicator probability π is completely collapsed out of the model without complicating posterior computation. Thus, we aim to generate samples from the target posterior distribution $p(\boldsymbol{\gamma}, \kappa, \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \rho, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\xi}, \boldsymbol{\Lambda} | \mathbf{Y})$, where $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_N^\top)^\top$ is a collection of random effects and $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_N^\top)^\top$ is a collection of observations. Second, to further facilitate posterior computation via the method of partial collapse, we reparameterize the covariance matrix of random effects such that $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Lambda}$ (Park and Min, 2014). The utility of this reparameterization was also noted in the context of an EM-type algorithm for linear mixed models (van Dyk, 2000). The reparameterization allows $(\boldsymbol{\xi}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \sigma^2)$ to be partially collapsed out of the model. Using the three basic tools developed by van Dyk and Park (2008), the posterior sampling scheme based on the method of partial collapse is constructed as follows.

Step 1. Draw γ_{km} from $p(\gamma_{km} | \boldsymbol{\gamma}_{-km}, \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})$, which is Bernoulli,

$$
\begin{aligned}
&\gamma_{km} | (\boldsymbol{\gamma}_{-km}, \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y}) \\
&\quad \sim \text{Bernoulli}\left(\frac{p(\gamma_{km} = 1, \boldsymbol{\gamma}_{-km} | \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})}{p(\gamma_{km} = 1, \boldsymbol{\gamma}_{-km} | \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y}) + p(\gamma_{km} = 0, \boldsymbol{\gamma}_{-km} | \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})} \right),
\end{aligned}
$$

for $k = 1, \ldots, p_{\mathbf{w}}$ and $m = 0, \ldots, L_k$, where $\boldsymbol{\gamma}_{-km} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_{k-1}^\top, \boldsymbol{\gamma}_{k\backslash m}^\top, \boldsymbol{\gamma}_{k+1}^\top, \ldots, \boldsymbol{\gamma}_{p_{\mathbf{w}}}^\top)^\top$ with $\boldsymbol{\gamma}_{k\backslash m} = (\gamma_{k0}, \ldots, \gamma_{k,m-1}, \gamma_{k,m+1}, \ldots, \gamma_{kL_k})^\top$, and the conditional distribution of $\boldsymbol{\gamma}$ given $(\kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})$ is given by

$$
\begin{aligned}
p(\boldsymbol{\gamma} | \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y}) \;\propto\; & \Gamma\Big(a + J(\boldsymbol{\gamma})\Big) \Gamma\Big(b + \sum\nolimits_{k=1}^{p_{\mathbf{w}}} L_k + p_{\mathbf{w}} - J(\boldsymbol{\gamma})\Big) \\
& \times (\kappa + 1)^{-(J(\boldsymbol{\gamma}) + p_{\mathbf{w}} + p_{\mathbf{x}})/2} \, \boldsymbol{\Xi}(\boldsymbol{\gamma}, \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda})^{-(r + \sum_{i=1}^N n_i + g p_{\mathbf{z}})/2}
\end{aligned}
$$

with

$$
\begin{aligned}
\boldsymbol{\Xi}(\boldsymbol{\gamma}, \kappa, \rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \;=\; & r s^2 + \operatorname{tr}(\mathbf{D} \boldsymbol{\Lambda}^{-1}) + \sum\nolimits_{i=1}^N \mathbf{Y}_i^\top \boldsymbol{\Omega}_i(\rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \mathbf{Y}_i \\
& - \frac{\kappa}{\kappa + 1} \bigg\{ \Big[\sum\nolimits_{i=1}^N \mathbf{C}_{\boldsymbol{\gamma}, i}^\top \boldsymbol{\Omega}_i(\rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \mathbf{Y}_i \Big]^\top \\
& \qquad \times \Big[\sum\nolimits_{i=1}^N \mathbf{C}_{\boldsymbol{\gamma}, i}^\top \boldsymbol{\Omega}_i(\rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \mathbf{C}_{\boldsymbol{\gamma}, i} \Big]^{-1} \Big[\sum\nolimits_{i=1}^N \mathbf{C}_{\boldsymbol{\gamma}, i}^\top \boldsymbol{\Omega}_i(\rho, \boldsymbol{\eta}, \boldsymbol{\Lambda}) \mathbf{Y}_i \Big] \bigg\}.
\end{aligned}
$$

Step 2. Draw $\varphi = \operatorname{logit}(\rho) = \log\{\rho/(1-\rho)\}$ from $p(\varphi | \boldsymbol{\gamma}, \kappa, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})$ and set $\rho = \operatorname{logit}^{-1}(\varphi)$, where the conditional distribution of φ given $(\boldsymbol{\gamma}, \kappa, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y})$ is proportional to

$$
\begin{aligned}
p(\varphi | \boldsymbol{\gamma}, \kappa, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \mathbf{Y}) \;\propto\; & \left[\operatorname{logit}^{-1}(\varphi)\right]^c \left[1 - \operatorname{logit}^{-1}(\varphi)\right]^d \\
& \times \boldsymbol{\Xi}(\boldsymbol{\gamma}, \kappa, \operatorname{logit}^{-1}(\varphi), \boldsymbol{\eta}, \boldsymbol{\Lambda})^{-(r + \sum_{i=1}^N n_i + g p_{\mathbf{z}})/2}
\end{aligned}
$$

$$\times \prod_{i=1}^{N}\Big(\big|\mathbf{\Lambda}-\mathbf{\Lambda}\mathbf{Z}_i^\top\mathbf{\Omega}_i(\mathrm{logit}^{-1}(\varphi),\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{Z}_i\mathbf{\Lambda}\big|^{1/2} \\ \times \big|\mathbf{H}_i(\boldsymbol{\eta})\mathbf{P}_i(\mathrm{logit}^{-1}(\varphi))\mathbf{H}_i(\boldsymbol{\eta})\big|^{-1/2}\Big).$$

The logit transformation makes the algorithm more stable because the support of ρ is bounded by 0 and 1. To simulate φ, we use a Metropolized independent sampler with a normal proposal distribution that is tuned to match the mode and curvature of $p(\varphi|\boldsymbol{\gamma},\kappa,\boldsymbol{\eta},\mathbf{\Lambda},\mathbf{Y})$.

Step 3. Draw σ^2 from $p(\sigma^2|\boldsymbol{\gamma},\kappa,\rho,\boldsymbol{\eta},\mathbf{\Lambda},\mathbf{Y})$, which is scaled inverse chi-squared,

$$\sigma^2|(\boldsymbol{\gamma},\kappa,\rho,\boldsymbol{\eta},\mathbf{\Lambda},\mathbf{Y}) \;\sim\; \mathbf{\Xi}(\boldsymbol{\gamma},\kappa,\rho,\boldsymbol{\eta},\mathbf{\Lambda})\Big/\chi^2_{r+\sum_{i=1}^N n_i+gp_{\mathbf{z}}}.$$

Step 4. Draw $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ from $p(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma},\kappa,\rho,\boldsymbol{\eta},\sigma^2,\mathbf{\Lambda},\mathbf{Y})$, which is multivariate normal,

$$\boldsymbol{\theta}_{\boldsymbol{\gamma}}|(\boldsymbol{\gamma},\kappa,\rho,\boldsymbol{\eta},\sigma^2,\mathbf{\Lambda},\mathbf{Y}) \\ \sim\; \mathrm{N}_{J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}}\left(\frac{\kappa}{\kappa+1}\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma},\rho,\boldsymbol{\eta},\mathbf{\Lambda}),\frac{\kappa\sigma^2}{\kappa+1}\Big[\sum_{i=1}^{N}\mathbf{C}_{\boldsymbol{\gamma},i}^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}\Big]^{-1}\right),$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma},\rho,\boldsymbol{\eta},\mathbf{\Lambda})=[\sum_{i=1}^N\mathbf{C}_{\boldsymbol{\gamma},i}^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}]^{-1}[\sum_{i=1}^N\mathbf{C}_{\boldsymbol{\gamma},i}^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{Y}_i]$.

Step 5. Draw $\boldsymbol{\xi}$ from $p(\boldsymbol{\xi}|\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\mathbf{\Lambda},\mathbf{Y})$, which is a product of N independent multivariate normal distributions,

$$\boldsymbol{\xi}_i|(\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\mathbf{\Lambda},\mathbf{Y}) \\ \overset{\text{ind}}{\sim}\; \mathrm{N}_{p_{\mathbf{z}}}\Big(\hat{\boldsymbol{\xi}}_i(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\mathbf{\Lambda}),\sigma^2\big[\mathbf{\Lambda}-\mathbf{\Lambda}\mathbf{Z}_i^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{Z}_i\mathbf{\Lambda}\big]\Big),\; i=1,\ldots,N,$$

with $\hat{\boldsymbol{\xi}}_i(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\mathbf{\Lambda})=\mathbf{\Lambda}\mathbf{Z}_i^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})(\mathbf{Y}_i-\mathbf{C}_{\boldsymbol{\gamma},i}\boldsymbol{\theta}_{\boldsymbol{\gamma}})$.

Step 6. Draw κ from $p(\kappa|\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi},\mathbf{\Lambda},\mathbf{Y})$, which is scaled inverse chi-squared,

$$\kappa|(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi},\mathbf{\Lambda},\mathbf{Y}) \\ \sim\; \left(qN+\frac{1}{\sigma^2}\boldsymbol{\theta}_{\boldsymbol{\gamma}}^\top\Big[\sum_{i=1}^{N}\mathbf{C}_{\boldsymbol{\gamma},i}^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}\Big]\boldsymbol{\theta}_{\boldsymbol{\gamma}}\right)\Big/\chi^2_{q+J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}}.$$

Step 7. Draw $\mathbf{\Lambda}$ from $p(\mathbf{\Lambda}|\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi},\mathbf{Y})$, where the conditional distribution of $\mathbf{\Lambda}$ given $(\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi},\mathbf{Y})$ is proportional to

$$p(\mathbf{\Lambda}|\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi},\mathbf{Y})\propto\mathcal{N}_{J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}}\left(\boldsymbol{\theta}_{\boldsymbol{\gamma}};\mathbf{0},\kappa\sigma^2\Big[\sum_{i=1}^{N}\mathbf{C}_{\boldsymbol{\gamma},i}^\top\mathbf{\Omega}_i(\rho,\boldsymbol{\eta},\mathbf{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}\Big]^{-1}\right) \\ \times\mathcal{IW}\Big(\mathbf{\Lambda};g+N,\Big[\mathbf{D}+\sum_{i=1}^{N}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\Big]\Big/\sigma^2\Big), \tag{5}$$

where $\mathcal{N}_d(\,\cdot\,;\boldsymbol{\mu},\mathbf{\Sigma})$ denotes the probability density function of a d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, and $\mathcal{IW}(\,\cdot\,;\nu,\mathbf{\Psi})$

denotes the probability density function of an inverse Wishart distribution with ν degrees of freedom and scale matrix $\boldsymbol{\Psi}$. To simulate $\boldsymbol{\Lambda}$, we use a Metropolized independent sampler with an inverse Wishart proposal distribution given by

$$\mathcal{IW}\left(\boldsymbol{\Lambda}; g+N, \left[\mathbf{D}+\sum\nolimits_{i=1}^{N}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top\right]\Big/\sigma^2\right);$$

see Appendix A for the construction of the proposal distribution.

Step 8. Draw $\boldsymbol{\eta}$ from $p(\boldsymbol{\eta}|\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\sigma^2,\boldsymbol{\xi},\boldsymbol{\Lambda},\mathbf{Y})$, which is proportional to

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{\gamma},\kappa,\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\sigma^2,\boldsymbol{\xi},\boldsymbol{\Lambda},\mathbf{Y}) \propto\ & \mathcal{N}_{J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}}\left(\boldsymbol{\theta}_{\boldsymbol{\gamma}};\mathbf{0},\kappa\sigma^2\left[\sum\nolimits_{i=1}^{N}\mathbf{C}_{\boldsymbol{\gamma},i}^\top\boldsymbol{\Omega}_i(\rho,\boldsymbol{\eta},\boldsymbol{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}\right]^{-1}\right)\\ & \times\mathcal{N}_{p_{\mathbf{v}}}(\boldsymbol{\eta};\mathbf{m},\mathbf{S})\\ & \times\prod\nolimits_{i=1}^{N}\mathcal{N}_{n_i}\left(\mathbf{Y}_i;\mathbf{C}_{\boldsymbol{\gamma},i}\boldsymbol{\theta}_{\boldsymbol{\gamma}}+\mathbf{Z}_i\boldsymbol{\xi}_i,\sigma^2\mathbf{H}_i(\boldsymbol{\eta})\mathbf{P}_i(\rho)\mathbf{H}_i(\boldsymbol{\eta})\right). \end{aligned} \tag{6}$$

To simulate $\boldsymbol{\eta}$, we devise a Metropolis–Hastings algorithm with an efficient proposal distribution that is multivariate normal with mean vector $\boldsymbol{\zeta}(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta}^{(t)},\sigma^2,\boldsymbol{\xi})$ and covariance matrix $\boldsymbol{\Delta}(\rho)$, where

$$\begin{aligned} \boldsymbol{\zeta}(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi}) &= \boldsymbol{\Delta}(\rho)\left(\mathbf{S}^{-1}\mathbf{m}+\frac{1}{2}\sum\nolimits_{i=1}^{N}\mathbf{V}_i^\top\left[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)\right]^{-1}\boldsymbol{\nu}_i(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi})\right),\\ \boldsymbol{\nu}_i(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi}) &= \left(\mathbf{v}_{i1}^\top\boldsymbol{\eta}-1+\frac{e_{i1}^2(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\boldsymbol{\xi})}{\sigma^2\exp(\mathbf{v}_{i1}^\top\boldsymbol{\eta})},\cdots,\mathbf{v}_{in_i}^\top\boldsymbol{\eta}-1+\frac{e_{in_i}^2(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\boldsymbol{\xi})}{\sigma^2\exp(\mathbf{v}_{in_i}^\top\boldsymbol{\eta})}\right)^\top,\\ \boldsymbol{\Delta}(\rho) &= \left(\mathbf{S}^{-1}+\frac{1}{2}\sum\nolimits_{i=1}^{N}\mathbf{V}_i^\top\left[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)\right]^{-1}\mathbf{V}_i\right)^{-1}, \end{aligned}$$

$\mathbf{V}_i=(\mathbf{v}_{i1},\ldots,\mathbf{v}_{in_i})^\top$ is an $n_i\times p_{\mathbf{v}}$ design matrix, $\circ$ denotes the Schur product, and $e_{ij}^2(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\boldsymbol{\xi})$ is a squared residual calculated as the square of the jth element of the vector $(\mathbf{Y}_i-\mathbf{C}_{\boldsymbol{\gamma},i}\boldsymbol{\theta}_{\boldsymbol{\gamma}}-\mathbf{Z}_i\boldsymbol{\xi}_i)$ in each iteration; see Appendix A for details.

Note that the resulting posterior sampling scheme consists of a functionally incompatible set of conditional distributions, so that changing the order of the sampling steps may upset the target stationary distribution of the corresponding Markov transition kernel (van Dyk and Park, 2008; Park and van Dyk, 2009). Thus, great care must be taken to implement the resulting posterior sampling scheme.

4 Simulation study

In this section, we assess the performance and validity of the proposed method by conducting a simulation study. Here we consider three scenarios of varying-coefficients to illustrate the issues of flexibility and robustness.

4.1 Varying-coefficients from spline bases

Under the first scenario, varying-coefficients are generated from a regression spline with the radial basis functions in (3). The true functions for varying-coefficients are given by $\alpha_1(U_1) = \mathbf{b}_1^\top(U_1)\boldsymbol{\phi}_1$ and $\alpha_2(U_2) = \mathbf{b}_2^\top(U_2)\boldsymbol{\phi}_2$, for $0 < U_1, U_2 < 1$, where $\mathbf{b}_k(\cdot)$ is the radial basis functions of the kth varying-coefficient with the following internal knots,

$$\begin{aligned}\boldsymbol{\tau}_1^* &= (0.2,\, 0.5,\, 0.8)^\top, \\ \boldsymbol{\tau}_2^* &= (0.05,\, 0.2,\, 0.35,\, 0.5,\, 0.65,\, 0.8,\, 0.95)^\top,\end{aligned}$$

and the corresponding coefficients are

$$\begin{aligned}\boldsymbol{\phi}_1 &= (-1.5,\, 0,\, 0.4,\, -1.5,\, 0.4)^\top, \\ \boldsymbol{\phi}_2 &= (-15,\, 0,\, 3.8,\, -6,\, 2.7,\, -4.5,\, 2.7,\, -6,\, 3.8)^\top.\end{aligned}$$

Along with these varying-coefficients, the true values of two fixed effects are given by $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (-3, 3)^\top$, the true values of parameters for two random effects are given by $\sigma^2 = 3$ and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{12} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 5 & 7 \\ 7 & 10 \end{pmatrix},$$

and the true values of parameters representing correlation and heteroscedasticity are given by $\rho = 0.3$ and $\boldsymbol{\eta} = (\eta_1, \eta_2)^\top = (2, -2)^\top$. That is, a positive correlation exists among observations in each group and the random effects are strongly and positively correlated. The number n_i of observations for each group is randomly generated from Poisson(10) + 1, the number N of groups is set to 100, the values of underlying effect modifiers $\mathbf{u}_{ij} = (u_{ij1}, u_{ij2})^\top$ are randomly drawn from a uniform distribution between 0 and 1, and all values of known covariates $\mathbf{w}_{ij} = (w_{ij1}, w_{ij2})^\top$, $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2})^\top$, $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2})^\top$, $\mathbf{v}_{ij} = (v_{ij1}, v_{ij2})^\top$, and $\mathbf{t}_{ij} = (t_{ij1}, t_{ij2})^\top$ are generated from independent standard normal distributions for all i and j. For each varying-coefficient, candidates of knot locations are chosen to be 30 equally spaced interior points which lie in the range of the corresponding underlying effect modifier. Given the test data generated under the simulation settings, we run the proposed sampling scheme for 10000 iterations with two over-dispersed starting values. The convergence of the sampling scheme is then assessed by computing the $\hat{R}^{1/2}$ statistic (Gelman and Rubin, 1992) for all model parameters of fixed dimension, and our posterior inference is based on the collection of second halves from the two chains.

Posterior estimates of two varying-coefficients modeled with regression splines are presented in Figure 1. The top two panels of Figure 1 show an estimated posterior inclusion probabilities of the linear basis term and knot locations for each of the two varying-coefficients. The true locations of knots are correctly specified by the high posterior inclusion probabilities of knot locations in the vicinity of true ones. The estimated inclusion probability of the linear basis term in each varying-coefficient is close to zero because its true value equals zero. The bottom two panels of Figure 1 present an estimated trajectory of each varying-coefficient as a function of the corresponding underly-

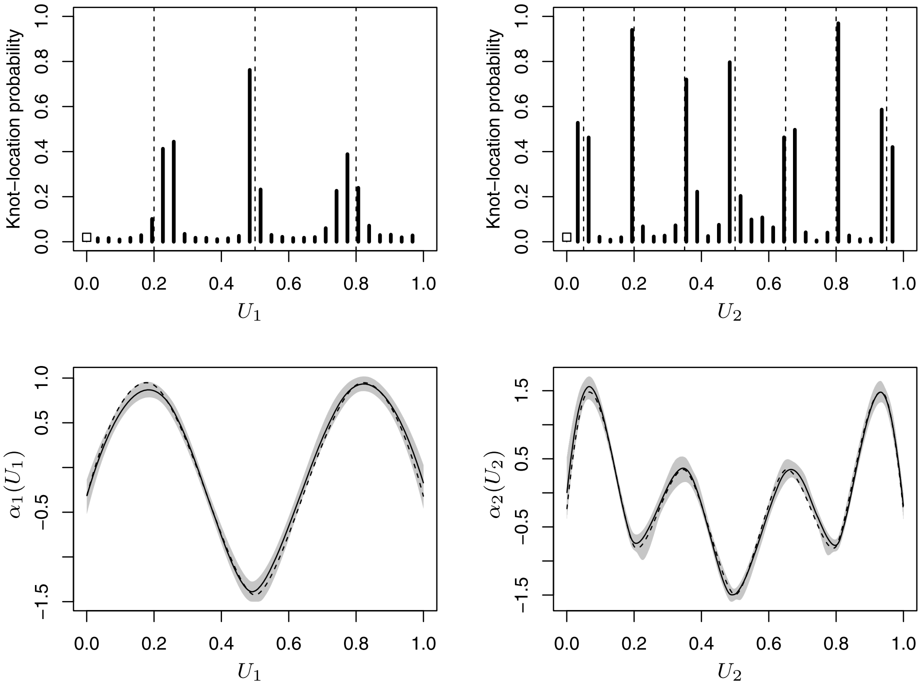

Figure 1: Posterior inclusion probabilities of knot locations and estimated functions of varying-coefficients under the first scenario. The top two panels show the posterior inclusion probabilities of the linear basis term (open square) and knot locations (solid lines) with true knot locations (vertical dashed lines). The bottom two panels show the true curves of varying-coefficients (dashed lines), point-wise posterior medians of the estimated curves (solid lines), and point-wise 95% intervals (gray regions).

ing effect modifier. The true trajectory of each varying-coefficient is well estimated by its point-wise posterior medians and reasonably covered by its 95% point-wise posterior intervals. Figure 2 shows the marginal posterior distributions of model parameters other than ones related to varying-coefficients, along with their autocorrelation plots. To numerically evaluate the performance of the proposed sampling scheme, the top right corner of each autocorrelation plot in Figure 2 presents the estimate of the integrated autocorrelation time (IAT), defined by

$$\text{IAT} = 1 + 2\sum\nolimits_{k=1}^{\infty} \upsilon_k,$$

where υ_k is an autocorrelation at lag k. The autocorrelation plots along with the estimates of the IAT illustrate the quick convergence characteristics of the proposed sampling scheme. In Figure 2, the true values of the model parameters used to simulate the test data are well covered by the corresponding marginal posterior distributions, thereby illustrating the validity of the proposed method.

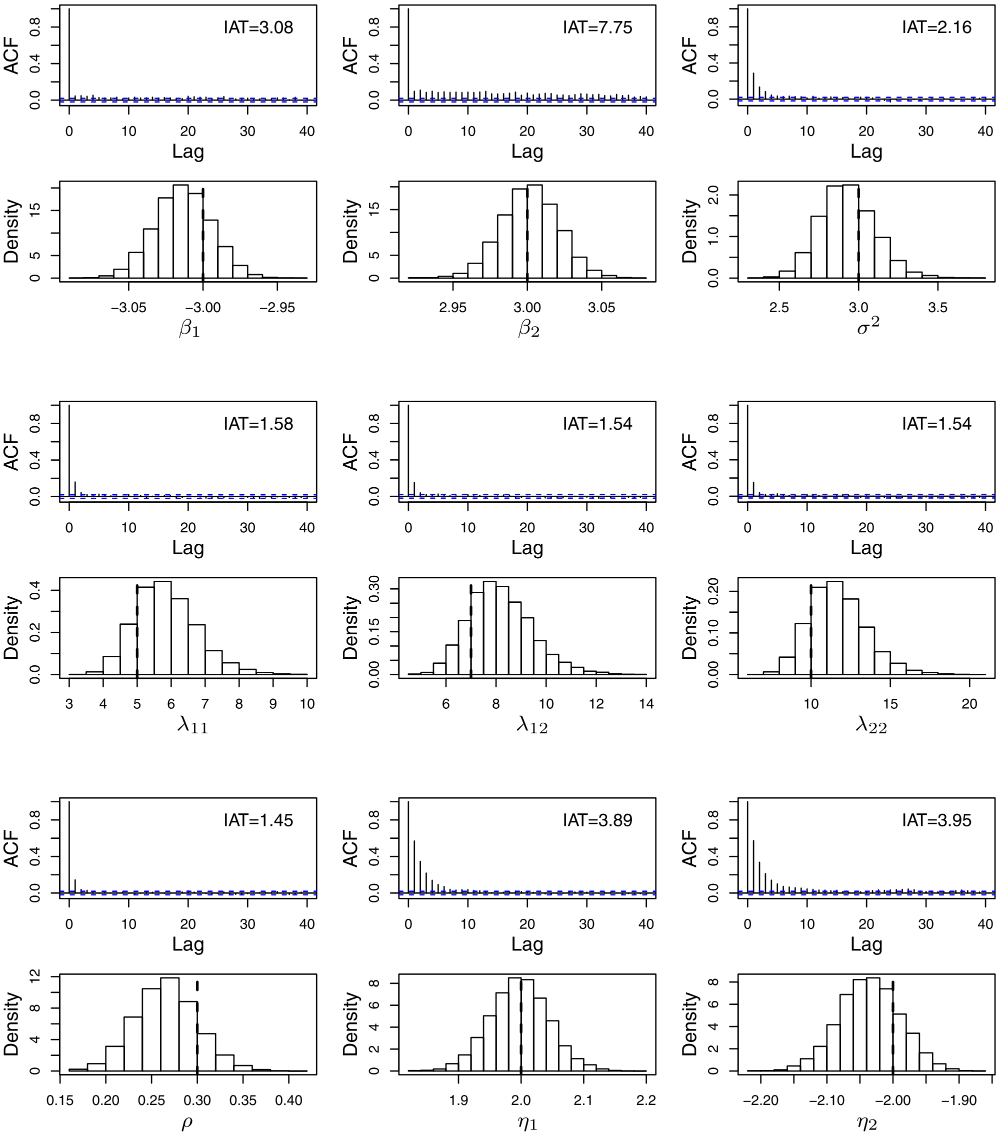

Figure 2: Autocorrelation plots and marginal posterior distributions for model parameters other than ones related to varying-coefficients. The vertical dashed lines in the marginal posterior distributions represent the true values of the parameters.

4.2 Varying-coefficients with different parametric structures

To assess the flexibility of the proposed method, we now suppose that each varying-coefficient is a unique parametric function rather than a regression spline with the radial

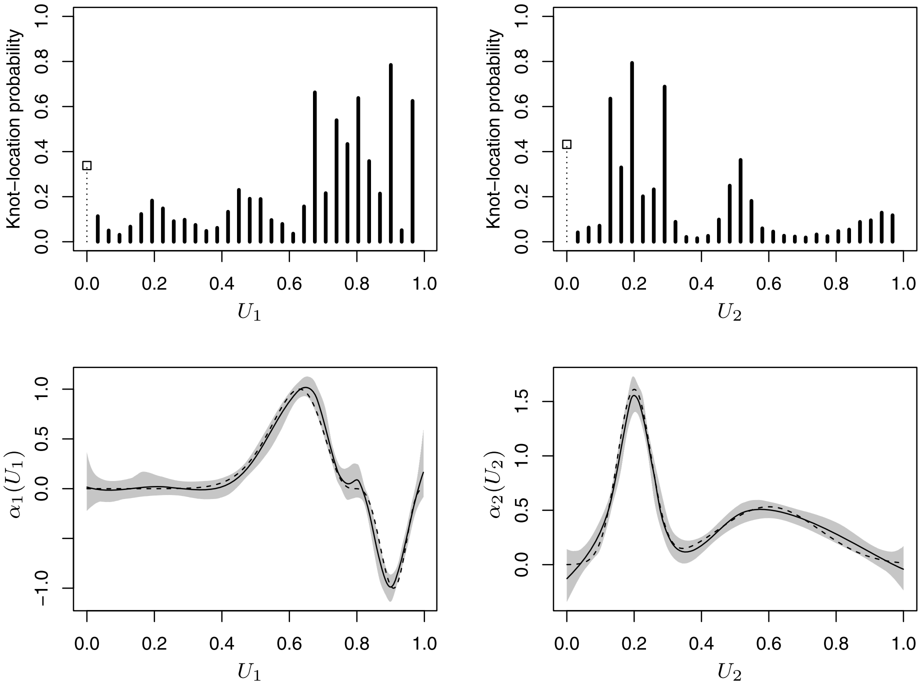

Figure 3: Posterior inclusion probabilities of knot locations and estimated functions of varying-coefficients under the second scenario. The top two panels show the posterior inclusion probabilities of the linear basis term (open square with a dotted line) and knot locations (solid lines). The bottom two panels show the true curves of varying-coefficients (dashed lines), point-wise posterior medians of the estimated curves (solid lines), and point-wise 95% intervals (gray regions).

basis functions. Under the second scenario, the true functions for varying-coefficients correspond to

$$\begin{aligned}\alpha_1(U_1) &= \sin^3(2\pi {U_1}^3),\\ \alpha_2(U_2) &= \mathcal{N}_1(U_2; 0.2, 0.05^2)/5 + \mathcal{N}_1(U_2; 0.6, 0.15^2)/5,\end{aligned}$$

for $0 < U_1, U_2 < 1$. Other simulation settings are identical to the first scenario in Section 4.1.

The posterior inclusion probabilities and the corresponding posterior estimates of varying-coefficients are presented in Figure 3. The top two panels of Figure 3 illustrate that more knots are required to estimate the wiggly parts of the true functions for varying-coefficients, and the proposed method can adapt to bumps in the true curves. Because of not being based on cross-validation with equally spaced fixed knots, the proposed method can avoid overfitting and determine the number of knots and their locations with uncertainty from data. In the bottom two panels of Figure 3, the true functions for varying-coefficients are correctly estimated by the point-wise posterior medians of regression splines and well covered by their point-wise 95% posterior intervals. While the true functions for varying-coefficients are well estimated by the free-knot regression splines, the marginal posterior distributions of other model parameters also correctly estimate their true values, as in Figure 2.

4.3 Robustness against overfitting

A disadvantage of existing methods based on cross-validation is their susceptibility to overfitting. Under the third scenario, we thus consider a misspecified model which may overfit data, and conduct a simulation study to illustrate that the proposed method is not so susceptible to overfitting and provides reliable results. Specifically, test data are generated from a simple linear mixed model with

$$Y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \boldsymbol{\xi}_i + \epsilon_{ij}, \tag{7}$$

where ϵ_{ij}'s are independent and identically distributed as $\mathrm{N}_1(0, \sigma^2)$, assuming neither correlation nor heteroscedasticity, i.e., $\rho = 0$ and $\boldsymbol{\eta} = \mathbf{0}$. The true values of four fixed effects are given by $\boldsymbol{\beta} = (0, 0, -3, 3)^\top$, the true value of the within-group variance is set to $\sigma^2 = 3$, and the true value of $\boldsymbol{\Lambda}$ is set as in Section 4.1. The number n_i of observations for each group and the number N of groups are identical to the simulation settings in Section 4.1. The values of known covariates, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij4})^\top$ and $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2})^\top$, are generated in the same way as in Section 4.1.

The test data generated from a simple model in (7) are fit with a complex model given by

$$Y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\alpha}(\mathbf{u}_{ij}) + \mathbf{z}_{ij}^\top \boldsymbol{\xi}_i + \epsilon_{ij}, \tag{8}$$

where the effects of $\mathbf{x}_{ij}$ on Y_{ij} are assumed to vary smoothly with the underlying effect modifiers $\mathbf{u}_{ij}$ and ϵ_{ij}'s are correlated and heteroscedastic errors. The values of additional known covariates, $\mathbf{u}_{ij} = (u_{ij1}, \ldots, u_{ij4})^\top$, $\mathbf{v}_{ij} = (v_{ij1}, v_{ij2})^\top$, and $\mathbf{t}_{ij} = (t_{ij1}, t_{ij2})^\top$, are generated in the same way as in Section 4.1.

Figure 4 shows the posterior inclusion probabilities of knot locations and the posterior estimates of varying-coefficients. The top panels of Figure 4 illustrate that our data-driven method used to fit the complex model in (8) tends to include only the constant basis term because the posterior knot inclusion probabilities of the other basis terms are all close to zero. That is, if the effect of a covariate is truly constant, the possibility of overfitting is eliminated by data-driven Bayesian knot selection. The bottom panels of Figure 4 show that the varying-coefficients are estimated as constant and the true values of constant effects are well covered by the posterior point-wise 95% intervals.

While the test data are generated with uncorrelated errors having the same variance, the fitted model in (8) assumes correlated and heteroscedastic errors with the possibility of overfitting. As shown in Figure 5, however, the marginal posterior distributions of parameters associated with correlation and heteroscedasticity suggest neither correlation nor heteroscedasticity, avoiding the possibility of misleading the results.

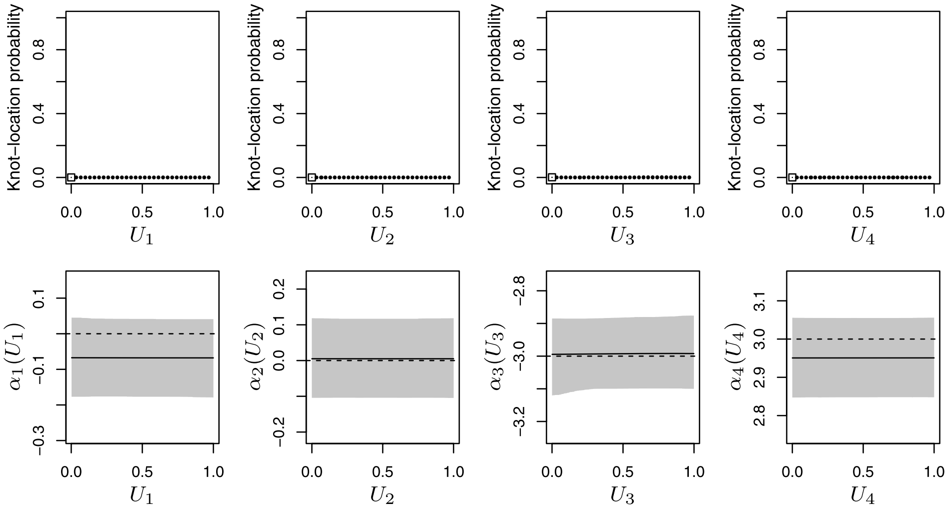

Figure 4: Posterior inclusion probabilities of knot locations and estimated functions of varying-coefficients under the third scenario. The top panels illustrate that only the constant basis term tends to be used in the model if the effect of a covariate is truly constant. The bottom panels show the estimated effect (solid lines) of a covariate is constant and the corresponding point-wise 95% intervals (gray regions) cover the true value (dashed lines) reasonably well.

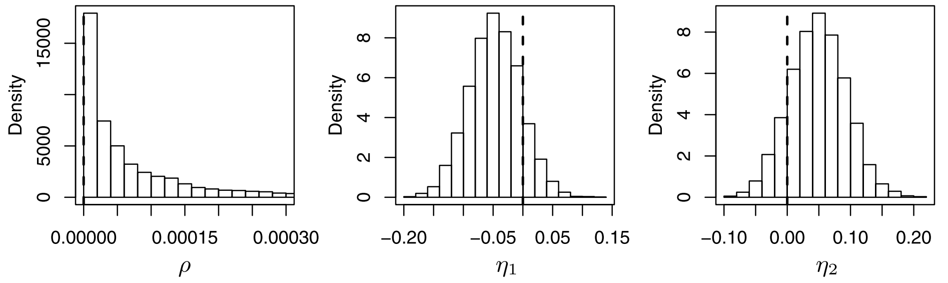

Figure 5: Marginal posterior distributions for selected model parameters associated with correlation and heteroscedasticity. The vertical dashed lines represent the true values of the parameters.

5 Application to soybean data

5.1 Study background and model construction

We illustrate the proposed methodology using longitudinal data from a study of soybean growth (Davidian and Giltinan, 1993, 1995). The soybean data were collected to compare the growth characteristics of soybeans with two genotypes: plant introduction

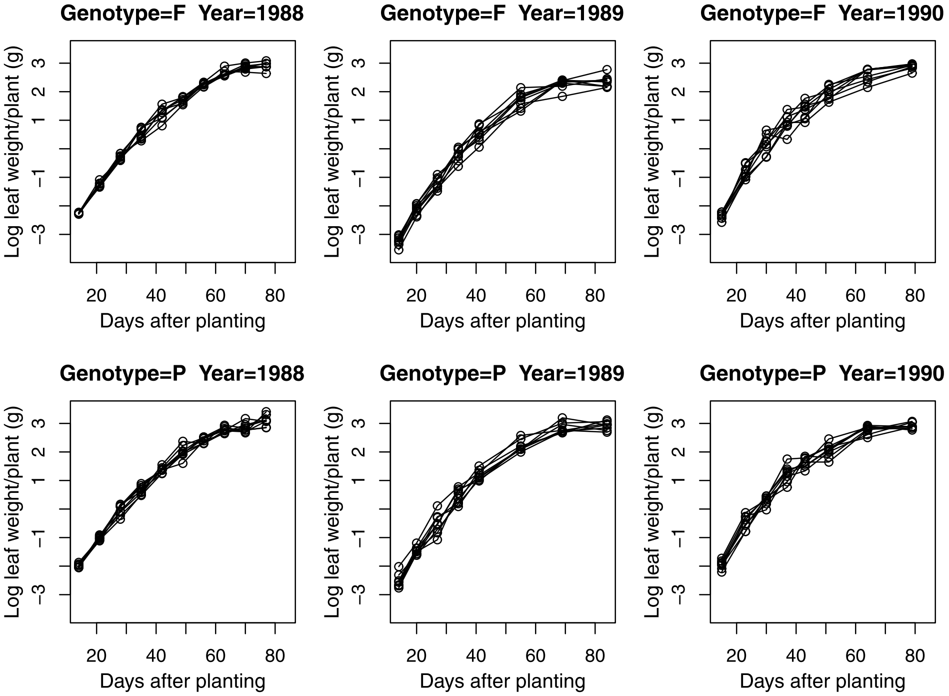

Figure 6: Trajectories of the log-average leaf weight per plant over days after planting.

#416937, an experimental strain, and Forrest, a commercial variety. These two genotypes are labeled as P and F, respectively. The experiment was conducted in three consecutive years (1988–1990). In each year, eight plots were planted with each of the two genotypes. Beginning from two weeks after planting, six plants were randomly selected from each plot roughly weekly and the leaves of the plants were aggregated and weighed, calculating an average leaf weight per plant. The trajectories of observations for the two genotypes are presented in Figure 6. Davidian and Giltinan (1995) used a logistic growth curve model to fit the soybean data. Here the same data are analyzed with the proposed model that flexibly models a functional relationship between the log-average leaf weight and genotype as a function of days after planting, while simultaneously accounting for random effects and non-spherical within-plot errors.

Figure 6 shows that the log-average leaf weight per plant varies among planting years, which suggests that the effect of planting year on soybean growth should be accounted for. This discrepancy may reflect different meteorological conditions: unusually dry weather in 1988, moist climate in 1989, and relatively normal conditions in 1990. Figure 6 also shows that the log-average leaf weight per plant has different variation for each combination of genotype and year, and that there is a plot-specific effect in the observed trajectories. To account for the variation among observations, we consider a random intercept for the plot-specific effect and allow within-plot variation to vary with days after planting, genotypes, and planting years. In addition, it is reasonable to assume the presence of serial correlation among observations in such longitudinal data obtained from repeated measurements over time, so that we account for serial correlation in the within-plot covariance matrix. The genotype–year interaction effects are assumed to be constant based on our pilot study showing that there is no varying effect on the interactions; refer to Section 4.3. The proposed model is then given by

$$\begin{aligned} Y_{ij} &= \alpha_1(D_{ij}) + \alpha_2(D_{ij})G_i + \alpha_3(D_{ij})I_{1989,i} + \alpha_4(D_{ij})I_{1990,i} \\ &+ \beta_1 G_i I_{1989,i} + \beta_2 G_i I_{1990,i} + \xi_i + \epsilon_{ij}, \end{aligned} \tag{9}$$

where Y_{ij} is the jth observation for the log-average leaf weight of the ith plot, D_{ij} denotes days after planting, G_i denotes a genotype labeled as 1 if a plant is P and 0 otherwise, $I_{t,i}$ is a dummy variable of year t, and ξ_i is a random effect following a normal distribution, $\mathrm{N}_1(0, \sigma^2\lambda)$. The within-plot covariance structure assumes that the log of a within-plot variance is linear in all covariates and a within-plot serial correlation is subject to a difference in days after planting, i.e.,

$$\begin{aligned} \log \mathrm{Var}(Y_{ij}|\xi_i) &= \log(\sigma^2) + \eta_1 D_{ij} + \eta_2 G_i + \eta_3 I_{1989,i} + \eta_4 I_{1990,i}, \\ \mathrm{Corr}(Y_{ij}, Y_{ik}|\xi_i) &= \rho^{|D_{ij} - D_{ik}|/7}, \end{aligned}$$

respectively. The observations were measured at approximately weekly intervals, so that we divide a difference in time intervals by 7 to make ρ represent a within-plot correlation coefficient between two consecutive observations. The model in (9) corresponds to the proposed model in (1), where

$$\begin{aligned} \mathbf{w}_{ij} &= (1, G_i, I_{1989,i}, I_{1990,i})^\top, \\ \mathbf{x}_{ij} &= (G_i I_{1989,i}, G_i I_{1990,i})^\top, \\ \mathbf{z}_{ij} &= 1, \\ \mathbf{u}_{ij} &= (D_{ij}, D_{ij}, D_{ij}, D_{ij})^\top, \\ \mathbf{v}_{ij} &= (D_{ij}, G_i, I_{1989,i}, I_{1990,i})^\top \\ \mathbf{t}_{ij} &= D_{ij}/7. \end{aligned}$$

As compared to a nonparametric model with no assumption on the structure, the proposed model in (9) can avoid the curse of dimensionality and the estimated functions for varying-coefficients are easy to interpret; refer to the case of the varying-coefficient model (Fan and Zhang, 1999).

5.2 Bayesian semiparametric inference and results

We ran the proposed sampling scheme in Section 3.2 with 15 knot candidates chosen by the observed order statistics of days after planting. Figure 7 presents the posterior estimates and point-wise 95% posterior intervals of varying-coefficients. The top-left panel of Figure 7 shows that the baseline log-average leaf weight per plant increases as a non-linear function of days after planting. In the top-right panel of Figure 7, the effect of genotype on the log-average leaf weight per plant is also nonlinear in days after planting and significantly positive, implying that the log-average leaf weight of plant

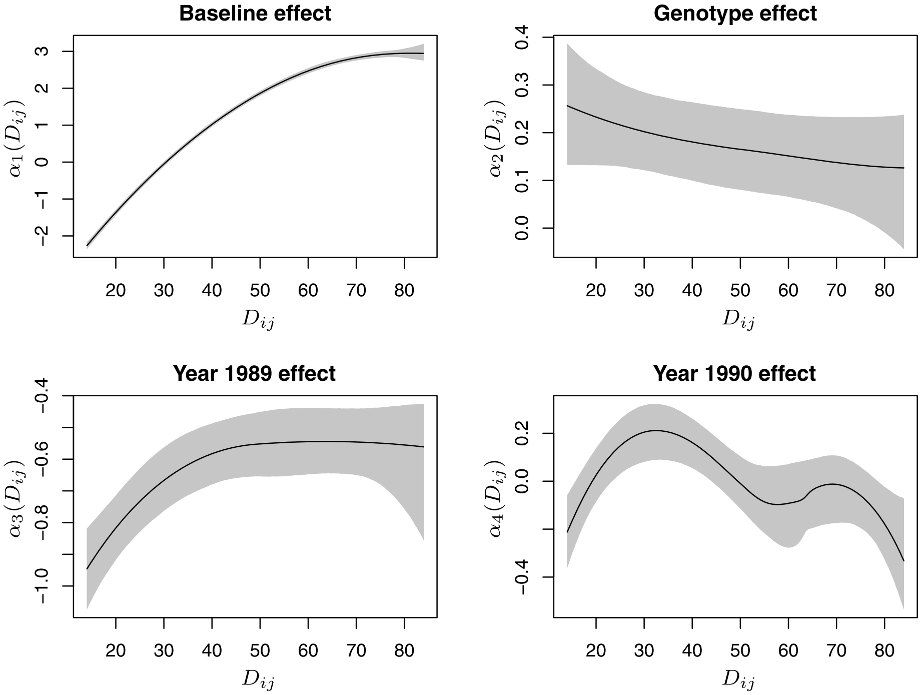

Figure 7: Posterior estimates of varying-coefficients for the soybean data.

introduction #416937 is significantly larger than that of Forest as days after planting progress. The bottom two panels of Figure 7 show the effects of the year 1989 and 1990 on the log-average leaf weight per plant, as compared to the year 1988. While the effect of the year 1989 is significantly different from that of the year 1988, the year effect on the log-average leaf weight per plant is not significantly different between the year 1988 and 1990.

Posterior summary statistics of model parameters other than ones related to varying-coefficients are presented in Table 1. The genotype-year interaction effects are estimated as 0.4420 with a 95% posterior interval of (0.3208, 0.5612) for the year 1989 and 0.0437 with a 95% posterior interval of (−0.0806, 0.1673) for the year 1990. As compared to the year 1988, the genotype-year interaction effect is thus significant for the year 1989 but not for the year 1990. The variance λ of a random effect has a posterior estimate of 0.1979 with a 95% posterior interval of (0.0795, 0.4713), reflecting significant plot-specific effects. The serial correlation parameter ρ is estimated as 0.0977 with a 95% posterior interval of (0.0178, 0.2136), which implies that there exists significant positive serial correlation over time in within-plot variation. While both 95% posterior intervals for η_1 and η_2 include zero, the year-specific heteroscedasticity parameters, η_3 and η_4, are estimated as 0.8159 and 0.9373 with 95% posterior intervals of (0.4585, 1.1673) and (0.5798, 1.2926), respectively. Thus, there exists significant heteroscedasticity across years but neither with genotype nor with days after planting.

Parameter	Posterior mean	Posterior median	2.5 percentile	97.5 percentile
β_1	0.4420	0.4427	0.3208	0.5612
β_2	0.0437	0.0440	-0.0806	0.1673
σ^2	0.0175	0.0169	0.0112	0.0270
λ	0.1979	0.1711	0.0795	0.4713
ρ	0.0977	0.0925	0.0178	0.2136
η_1	0.0005	0.0006	-0.0076	0.0084
η_2	0.1128	0.1131	-0.1778	0.4134
η_3	0.8159	0.8163	0.4585	1.1673
η_4	0.9373	0.9357	0.5798	1.2926

Table 1: Posterior summary statistics of selected model parameters.

5.3 Model diagnostics

To investigate the fit of the proposed model, we make a posterior predictive assessment that checks the ability of the fitted model to predict data to which the model is fit (Meng, 1994; Park et al., 2008). In the posterior predictive checks, we compare observed data to a posterior predictive distribution and base our comparison on six test statistics, $T_i(\mathbf{Y})$ for $i = 1, \ldots, 6$: the (1) minimum, (2) first quartile, (3) median, (4) third quartile, (5) maximum, and (6) mean. To do so, we generate M replicated data sets, denoted by $\{\tilde{\mathbf{Y}}^{(m)}, m = 1, \ldots, M\}$, from a posterior predictive distribution and compute the test statistics for each data set. The posterior predictive p-value (ppp-value) is calculated by the proportion of these M replicated data sets for which each test statistic equals or exceeds the corresponding observed test statistic, i.e., $T_i(\tilde{\mathbf{Y}}^{(m)}) \geq T_i(\mathbf{Y})$ for $m = 1, \ldots, M$ and $i = 1, \ldots, 6$.

Figure 8 displays a histogram of the values of $T_i(\tilde{\mathbf{Y}}^{(m)})$ for $m = 1, \ldots, M$ and $i = 1, \ldots, 6$, with the observed value, $T_i(\mathbf{Y})$, shown by vertical dashed lines. As shown in Figure 8, the observed data look plausible under the posterior predictive distribution for all six test statistics, and the corresponding ppp-values fall into a reasonable range, i.e., between 0.05 and 0.95. Thus, the posterior predictive assessment implies the proposed model in (9) provides an adequate fit under the model's assumptions.

6 Discussion

This paper proposes a partially linear varying-coefficient mixed model in which errors are correlated and heteroscedastic, and develops a novel data-driven method for efficient Bayesian semi-parametric inference on the proposed model by means of reparameterization and partial collapse. When making inferences on the varying effects of covariates on the mean response, existing methods fail to fully account for correlated and heteroscedastic errors that are commonly present in longitudinal or spatial data. The soybean growth application presents one typical example of this scenario. By not only generalizing the error terms but also allowing for the data-adaptive selection of the number and location of knots in a regression spline, our proposed Bayesian method provides a unified and flexible approach to inference on a linear mixed model that in-

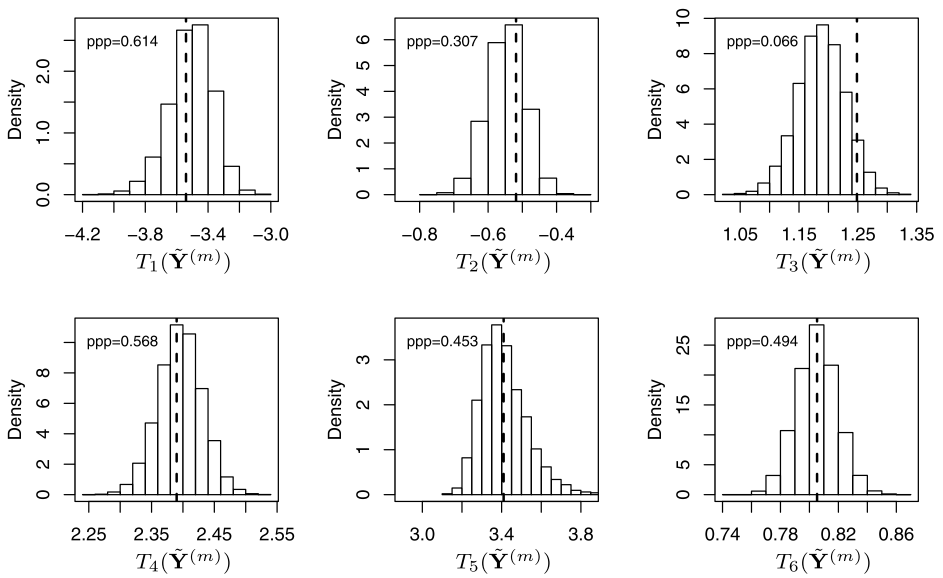

Figure 8: Posterior predictive checks. The vertical dashed lines represent the values of observed test statistics.

volves varying-coefficients. The efficiency of the posterior sampling scheme used to infer the proposed model is achieved by adopting the method of partial collapse without complicating posterior updates via reparameterization.

A generalization to random effects in the proposed model can be an interesting area for future research, allowing random effects to vary over some underlying effect modifiers. It is possible that varying covariate effects across different groups can be group-specific. This idea is conceptually related to the so-called semiparametric stochastic mixed models (Zeger and Diggle, 1994; Zhang et al., 1998), though previous methods are based on either kernel or spline smoothing under known covariance structure. In future research, the proposed method can be also generalized to analyze count data or aggregate data within a generalized linear mixed model framework (Park and Jeong, 2015). Lastly, each varying-coefficient in our proposed model can vary as a function of a single underlying effect modifier, but this restriction may be relaxed by employing an additive structure or multidimensional smoothing methods (Lee et al., 2012; Park et al., 2015; Ma and Song, 2015).

Appendix A

Efficient construction of a proposal distribution for Λ

In order to construct an efficient proposal distribution for a Metropolized independent sampler in Step 7, we approximate the target conditional distribution of $\mathbf{\Lambda}$ in (5). To do so, we make the following assumptions:

Assumption 1. *The marginal posterior distribution of the model parameters* $(\boldsymbol{\gamma}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \rho, \boldsymbol{\eta}, \sigma^2, \boldsymbol{\Lambda})$ *converges to a degenerate distribution at its mode* $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\theta}}_{\widetilde{\boldsymbol{\gamma}}}, \widetilde{\rho}, \widetilde{\boldsymbol{\eta}}, \widetilde{\sigma}^2, \widetilde{\boldsymbol{\Lambda}})$ *as* N *goes to infinity.*

Assumption 2. *The matrix* $N^{-1}\sum_{i=1}^{N}\mathbf{C}_{\boldsymbol{\gamma},i}^{\top}\boldsymbol{\Omega}_i(\rho,\boldsymbol{\eta},\boldsymbol{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}$ *converges to a random matrix as* N *goes to infinity.*

Assumption 3. *The matrix* $N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\xi}}_i(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\boldsymbol{\Lambda})\hat{\boldsymbol{\xi}}_i(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\boldsymbol{\Lambda})^{\top}$ *converges to a random matrix as* N *goes to infinity.*

If Assumptions 1 and 2 hold, it can be seen that, for large N, the marginal distribution of $\kappa' = \kappa/N$ is approximately scaled inverse chi-squared,

$$\kappa'|\mathbf{Y} \overset{\text{app}}{\sim} \left(q + \frac{1}{\widetilde{\sigma}^2}\widetilde{\boldsymbol{\theta}}_{\widetilde{\boldsymbol{\gamma}}}^{\top}\mathbf{G}(\widetilde{\boldsymbol{\gamma}},\widetilde{\rho},\widetilde{\boldsymbol{\eta}},\widetilde{\boldsymbol{\Lambda}})\widetilde{\boldsymbol{\theta}}_{\widetilde{\boldsymbol{\gamma}}}\right)\Big/\chi^2_{q+J(\widetilde{\boldsymbol{\gamma}})+p_{\mathbf{w}}+p_{\mathbf{x}}},$$

where $\mathbf{G}(\boldsymbol{\gamma},\rho,\boldsymbol{\eta},\boldsymbol{\Lambda}) = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\mathbf{C}_{\boldsymbol{\gamma},i}^{\top}\boldsymbol{\Omega}_i(\rho,\boldsymbol{\eta},\boldsymbol{\Lambda})\mathbf{C}_{\boldsymbol{\gamma},i}$, and that the marginal distribution of $\boldsymbol{\xi}_i$ is approximately multivariate normal,

$$\boldsymbol{\xi}_i|\mathbf{Y} \overset{\text{app}}{\sim} \mathrm{N}_{p_{\mathbf{z}}}\left(\hat{\boldsymbol{\xi}}_i(\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\theta}}_{\widetilde{\boldsymbol{\gamma}}},\widetilde{\rho},\widetilde{\boldsymbol{\eta}},\widetilde{\boldsymbol{\Lambda}}),\widetilde{\sigma}^2\left[\widetilde{\boldsymbol{\Lambda}} - \widetilde{\boldsymbol{\Lambda}}\mathbf{Z}_i^{\top}\boldsymbol{\Omega}_i(\widetilde{\rho},\widetilde{\boldsymbol{\eta}},\widetilde{\boldsymbol{\Lambda}})\mathbf{Z}_i\widetilde{\boldsymbol{\Lambda}}\right]\right).$$

Thus the approximate marginal modes of κ' and $\boldsymbol{\xi}_i$, denoted by $\widetilde{\kappa}'$ and $\widetilde{\boldsymbol{\xi}}_i$, respectively, tend to be independent of N as $N \to \infty$. If Assumption 1 holds and the chain converges to the stationary distribution, the target conditional distribution in (5) is approximately proportional to

$$\mathcal{N}_{J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}}\left(\widetilde{\boldsymbol{\theta}}_{\widetilde{\boldsymbol{\gamma}}};\mathbf{0},\widetilde{\kappa}'\widetilde{\sigma}^2\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{C}_{\widetilde{\boldsymbol{\gamma}},i}^{\top}\boldsymbol{\Omega}_i(\widetilde{\rho},\widetilde{\boldsymbol{\eta}},\boldsymbol{\Lambda})\mathbf{C}_{\widetilde{\boldsymbol{\gamma}},i}\right]^{-1}\right)$$
$$\times\,\mathcal{IW}\left(\boldsymbol{\Lambda}; g+N, \left[\mathbf{D} + \sum_{i=1}^{N}\widetilde{\boldsymbol{\xi}}_i\widetilde{\boldsymbol{\xi}}_i^{\top}\right]\Big/\widetilde{\sigma}^2\right), \quad (10)$$

for large N.

If Assumption 2 holds, the first term in (10) converges to a function of $\boldsymbol{\Lambda}$ that is independent of N, so that its dispersion is likely to remain constant as $N \to \infty$. By contrast, the dispersion of the second term in (10) decreases as $N \to \infty$, when Assumption 3 holds. With large N and under stationarity, this result implies that the target conditional distribution in (5) is mostly affected by the second term rather than the first term in (5). Thus, we construct an efficient proposal distribution of the Metropolized independent sampler for $\boldsymbol{\Lambda}$ from the second term in (5).

Efficient construction of a proposal distribution for $\boldsymbol{\eta}$

In order to construct an efficient proposal distribution of a Metropolis–Hastings algorithm in Step 8, we again approximate the target conditional distribution in (6). To do so, we make the following proposition.

Proposition 1. *When other parameters are assumed known and $\mathbf{P}_i(\rho)$ is the true correlation matrix, $\sqrt{N}(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}})$ asymptotically follows a multivariate normal distribution with zero mean vector and covariance matrix*

$$\lim_{N\to\infty} 2\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{V}_i^\top\left[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)\right]^{-1}\mathbf{V}_i\right)^{-1},$$

under mild regularity conditions, where $\hat{\boldsymbol{\eta}}$ is the maximum likelihood estimator of $\boldsymbol{\eta}$ computed by an iterative procedure,

$$\hat{\boldsymbol{\eta}}^{(\ell+1)} = \boldsymbol{\Delta}^*(\rho)\Big(\frac{1}{2}\sum\nolimits_{i=1}^{N}\mathbf{V}_i^\top\left[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)\right]^{-1}\boldsymbol{\nu}_i^\top(\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\hat{\boldsymbol{\eta}}^{(\ell)},\sigma^2,\boldsymbol{\xi})\Big),$$

where $\boldsymbol{\Delta}^*(\rho) = 2(\sum_{i=1}^{N}\mathbf{V}_i^\top[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)]^{-1}\mathbf{V}_i)^{-1}$ and $\hat{\boldsymbol{\eta}}^{(\ell)}$ denotes the ℓth iterate of $\boldsymbol{\eta}$ in the procedure.

Proof. Given fixed parameters, $\epsilon_{ij}^2/[\sigma^2\exp(\mathbf{v}_{ij}^\top\boldsymbol{\eta})]$ follows a chi-squared distribution with one degree of freedom and $\mathrm{corr}(\epsilon_{ij}^2,\epsilon_{ik}^2) = \rho^{2\|\mathbf{t}_{ij}-\mathbf{t}_{ik}\|}$ for $j\neq k$. Thus ϵ_{ij}^2 follows a correlated gamma distribution with mean $\sigma^2\exp(\mathbf{v}_{ij}^\top\boldsymbol{\eta})$, so that $\boldsymbol{\eta}$ can be viewed as the coefficients of a generalized linear model with correlated gamma family and log-link. The proof is then straightforward by the theory of a generalized linear model with correlated data in Liang and Zeger (1986). □

If Assumption 1 holds, Proposition 1 implies that, for large N, the target conditional distribution in (6) is approximately proportional to

$$\mathcal{N}_{J(\boldsymbol{\gamma})+p_{\mathbf{w}}+p_{\mathbf{x}}}\Big(\widetilde{\boldsymbol{\theta}}_{\widetilde{\gamma}};\mathbf{0},\widetilde{\kappa}'\widetilde{\sigma}^2\Big[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{C}_{\widetilde{\boldsymbol{\gamma}},i}^\top\boldsymbol{\Omega}_i(\widetilde{\rho},\boldsymbol{\eta},\widetilde{\boldsymbol{\Lambda}})\mathbf{C}_{\widetilde{\boldsymbol{\gamma}},i}\Big]^{-1}\Big)\mathcal{N}_{p_{\mathbf{v}}}\big(\boldsymbol{\eta};\boldsymbol{\zeta}^*(\widetilde{\rho}),\boldsymbol{\Delta}(\widetilde{\rho})\big), \quad (11)$$

where $\boldsymbol{\zeta}^*(\rho) = \boldsymbol{\Delta}(\rho)(\mathbf{S}^{-1}\mathbf{m}+(1/2)\sum_{i=1}^{N}\mathbf{V}_i^\top[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)]^{-1}\mathbf{V}_i\hat{\boldsymbol{\eta}})$. To further approximate this function, we make the following additional assumption.

Assumption 4. *The matrix $N^{-1}\sum_{i=1}^{N}\mathbf{V}_i^\top[\mathbf{P}_i(\rho)\circ\mathbf{P}_i(\rho)]^{-1}\mathbf{V}_i$ converges to a random matrix as N goes to infinity.*

When Assumptions 2 and 4 hold, the dispersion of the first term in (11) is likely to remain constant as $N\to\infty$. By constrast, the dispersion of the second term in (11) decreases as $N\to\infty$. With large N and under stationarity, this result implies that the target conditional distribution in (6) is mostly affected by the last two terms rather than the first term in (6). Thus, we construct an efficient proposal distribution of the Metropolis–Hastings algorithm for $\boldsymbol{\eta}$ based on the likelihood function $p(\mathbf{Y}|\boldsymbol{\gamma},\boldsymbol{\theta}_{\boldsymbol{\gamma}},\rho,\boldsymbol{\eta},\sigma^2,\boldsymbol{\xi})$ and the prior distribution $p(\boldsymbol{\eta})$ in (6).

Under an independence assumption, Gamerman (1997) developed an efficient Metropolis–Hastings proposal distribution for a generalized linear model based on an iterative weighted least square procedure for the maximum likelihood estimator and its approximate normality. This method was applied to a heteroscedastic regression model in Chan et al. (2006). While the previous methods are restricted to an independent case, we extend the results to correlated data. The proposal distribution in Step 8 is then constructed by combining the prior distribution $p(\boldsymbol{\eta})$ and the asymptotic result in Proposition 1.

References

Ahmad, I., Leelahanon, S., and Li, Q. (2005). "Efficient estimation of a semiparametric partially linear varying coefficient model." *The Annals of Statistics*, 33(1): 258–283. MR2157803. doi: http://dx.doi.org/10.1214/009053604000000931. 1137

Biller, C. and Fahrmeir, L. (2001). "Bayesian varying-coefficient models using adaptive regression splines." *Statistical Modelling*, 1(3): 195–211. 1138

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association*, 80(391): 580–598. MR0803258. 1137

Buja, A., Hastie, T., and Tibshirani, R. (1989). "Linear smoothers and additive models." *The Annals of Statistics*, 17(2): 453–510. MR0994249. doi: http://dx.doi.org/10.1214/aos/1176347115. 1137

Cai, Z. (2002). "Two-step likelihood estimation procedure for varying-coefficient models." *Journal of Multivariate Analysis*, 82(1): 189–209. MR1918620. doi: http://dx.doi.org/10.1006/jmva.2001.2013. 1138

Cai, Z. (2007). "Trending time-varying coefficient time series models with serially correlated errors." *Journal of Econometrics*, 136(1): 163–188. MR2328589. doi: http://dx.doi.org/10.1016/j.jeconom.2005.08.004. 1140

Cai, Z., Fan, J., and Li, R. (2000). "Efficient estimation and inferences for varying-coefficient models." *Journal of the American Statistical Association*, 95(451): 888–902. MR1804446. doi: http://dx.doi.org/10.2307/2669472. 1140

Chan, D., Kohn, R., Nott, D., and Kirby, C. (2006). "Locally adaptive semiparametric estimation of the mean and variance functions in regression models." *Journal of Computational and Graphical Statistics*, 15(4): 915–936. MR2273484. doi: http://dx.doi.org/10.1198/106186006X157441. 1140, 1159

Cheng, M.-Y., Zhang, W., and Chen, L.-H. (2009). "Statistical estimation in generalized multiparameter likelihood models." *Journal of the American Statistical Association*, 104(487). MR2750243. doi: http://dx.doi.org/10.1198/jasa.2009.tm08430. 1138

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1991). "Local regression models." In: Chambers, J. M. and Hastie, T. J. (eds.), *Statistical Models in S*, 309–376. Pacific Grove: Wadsworth & Brooks/Cole. 1137

Clyde, M. and George, E. I. (2004). "Model uncertainty." *Statistical Science*, 81–94. MR2082148. doi: http://dx.doi.org/10.1214/088342304000000035. 1143

Davidian, M. and Giltinan, D. M. (1993). "Some general estimation methods for nonlinear mixed-effects model." *Journal of Biopharmaceutical Statistics*, 3(1): 23–55. 1152

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, New York. 1152, 1153

Denison, D., Mallick, B., and Smith, A. (1998). “Automatic Bayesian curve fitting.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2): 333–350. MR1616029. doi: http://dx.doi.org/10.1111/1467-9868.00128. 1138

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). “Bayesian curve-fitting with free-knot splines.” *Biometrika*, 88(4): 1055–1071. MR1872219. doi: http://dx.doi.org/10.1093/biomet/88.4.1055. 1138

Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986). “Semiparametric estimates of the relation between weather and electricity sales.” *Journal of the American Statistical Association*, 81(394): 310–320. 1137

Fan, J. and Huang, T. (2005). “Profile likelihood inferences on semiparametric varying-coefficient partially linear models.” *Bernoulli*, 11(6): 1031–1057. MR2189080. doi: http://dx.doi.org/10.3150/bj/1137421639. 1137

Fan, J. and Zhang, W. (1999). “Statistical estimation in varying coefficient models.” *The Annals of Statistics*, 27(5): 1491–1518. MR1742497. doi: http://dx.doi.org/10.1214/aos/1017939139. 1138, 1154

Fan, J. and Zhang, W. (2008). “Statistical methods with varying coefficient models.” *Statistics and Its Interface*, 1(1): 179–195. MR2425354. doi: http://dx.doi.org/10.4310/SII.2008.v1.n1.a15. 1137

Gamerman, D. (1997). “Sampling from the posterior distribution in generalized linear mixed models.” *Statistics and Computing*, 7(1): 57–68. 1143, 1159

Gelman, A. and Rubin, D. B. (1992). “Inference from iterative simulations using multiple sequences (with discussion).” *Statistical Science*, 7: 457–472. MR1196276. 1147

Hastie, T. and Tibshirani, R. (1993). “Varying-coefficient models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55(4): 757–796. MR1229881. 1137, 1138

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. MR1666699. doi: http://dx.doi.org/10.1093/biomet/85.4.809. 1138, 1140

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). “Varying-coefficient models and basis function approximations for the analysis of repeated measurements.” *Biometrika*, 89(1): 111–128. MR1888349. doi: http://dx.doi.org/10.1093/biomet/89.1.111. 1138, 1140

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). “Polynomial spline estimation and inference for varying coefficient models with longitudinal data.” *Statistica Sinica*, 14(3): 763–788. MR2087972. 1138, 1140

Jones, R. H. and Boadi-Boateng, F. (1991). “Unequally spaced longitudinal data with AR(1) serial correlation.” *Biometrics*, 47: 161–175. 1140

Kohn, R., Smith, M., and Chan, D. (2001). “Nonparametric regression using linear combinations of basis functions.” *Statistics and Computing*, 11(4): 313–322. MR1863502. doi: http://dx.doi.org/10.1023/A:1011916902934. 1138, 1143

Laird, N. M. and Ware, J. H. (1982). "Random-effects models for longitudinal data." *Biometrics*, 38(4): 963–974. 1139

Lee, Y. K., Mammen, E., and Park, B. U. (2012). "Flexible generalized varying coefficient regression models." *The Annals of Statistics*, 40(3): 1906–1933. MR3015048. doi: http://dx.doi.org/10.1214/12-AOS1026. 1157

Leslie, D. S., Kohn, R., and Nott, D. J. (2007). "A general approach to heteroscedastic linear regression." *Statistics and Computing*, 17(2): 131–146. MR2380642. doi: http://dx.doi.org/10.1007/s11222-006-9013-8. 1140

Li, D., Chen, J., and Gao, J. (2011). "Non-parametric time-varying coefficient panel data models with fixed effects." *The Econometrics Journal*, 14(3): 387–408. MR2853233. doi: http://dx.doi.org/10.1111/j.1368-423X.2011.00350.x. 1140

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481). MR2420243. doi: http://dx.doi.org/10.1198/016214507000001337. 1143

Liang, H., Wu, H., and Carroll, R. J. (2003). "The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient models with measurement error." *Biostatistics*, 4(2): 297–312. 1140

Liang, K.-Y. and Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1): 13–22. MR0836430. doi: http://dx.doi.org/10.1093/biomet/73.1.13. 1159

Liu, J. S. (1994). "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem." *Journal of the American Statistical Association*, 89(427): 958–966. MR1294740. 1144

Ma, S. and Song, P. X.-K. (2015). "Varying index coefficient models." *Journal of the American Statistical Association*, 110(509): 341–356. MR3338507. doi: http://dx.doi.org/10.1080/01621459.2014.903185. 1157

Meng, X.-L. (1994). "Posterior predictive p-values." *The Annals of Statistics*, 1142–1160. MR1311969. doi: http://dx.doi.org/10.1214/aos/1176325622. 1156

Park, B. U., Mammen, E., Lee, Y. K., and Lee, E. R. (2015). "Varying coefficient regression models: a review and new developments." *International Statistical Review*, 83(1): 36–64. MR3341079. doi: http://dx.doi.org/10.1111/insr.12029. 1137, 1157

Park, T. and Jeong, S. (2015). "Efficient Bayesian analysis of multivariate aggregate choices." *Journal of Statistical Computation and Simulation*, 85(16): 3352–3366. MR3384053. doi: http://dx.doi.org/10.1080/00949655.2014.973411. 1157

Park, T., Krafty, R. T., and Sanchez, A. I. (2012). "Bayesian semi-parametric analysis of Poisson change-point regression models: application to policy-making." *Journal of Applied Statistics*, 39(10): 2285–2298. MR2968025. doi: http://dx.doi.org/10.1080/02664763.2012.709227. 1138

Park, T. and Min, S. (2014). "Partially collapsed Gibbs sampling for linear mixed-effects models." *Communications in Statistics – Simulation and Computation*, doi: http://dx.doi.org/10.1080/03610918.2013.857687. 1139, 1144

Park, T. and van Dyk, D. A. (2009). "Partially collapsed Gibbs samplers: illustrations and applications." *Journal of Computational and Graphical Statistics*, 18(2): 283–305. MR2749833. doi: http://dx.doi.org/10.1198/jcgs.2009.08108. 1146

Park, T., van Dyk, D. A., and Siemiginowska, A. (2008). "Searching for Narrow Emission Lines in X-ray Spectra: Computation and Methods." *The Astrophysical Journal*, 688(2): 807–825. 1156

Smith, M. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection." *Journal of Econometrics*, 75(2): 317–343. 1138

Thompson, W. K. and Rosen, O. (2008). "A Bayesian model for sparse functional data." *Biometrics*, 64(1): 54–63. MR2422819. doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00829.x. 1142

van Dyk, D. A. (2000). "Fitting mixed-effects models using efficient EM-type algorithms." *Journal of Computational and Graphical Statistics*, 9(1): 78–98. MR1826277. doi: http://dx.doi.org/10.2307/1390614. 1139, 1144

van Dyk, D. A. and Park, T. (2008). "Partially collapsed Gibbs samplers: theory and methods." *Journal of the American Statistical Association*, 193(482): 790–796. MR2524010. doi: http://dx.doi.org/10.1198/016214508000000409. 1144, 1146

Wu, C. O. and Chiang, C.-T. (2000). "Kernel smoothing on varying coefficient models with longitudinal dependent variable." *Statistica Sinica*, 10(2): 433–456. MR1769751. 1138

Zeger, S. L. and Diggle, P. J. (1994). "Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters." *Biometrics*, 50(3): 689–699. 1157

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243. MR0881437. 1143

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). "Semiparametric stochastic mixed models for longitudinal data." *Journal of the American Statistical Association*, 93(442): 710–719. MR1631369. doi: http://dx.doi.org/10.2307/2670121. 1157

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2056049). We also thank the editors and referees for their many helpful comments.