Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Functional clustering methods for binary longitudinal data with temporal heterogeneity

Jinwon Sohn[a], Seonghyun Jeong[b,c,1], Young Min Cho[d], Taeyoung Park[b,c,*]

[a] *Department of Statistics, Purdue University, IN 47907, USA*
[b] *Department of Applied Statistics, Yonsei University, Seoul 03722, Korea*
[c] *Department of Statistics and Data Science, Yonsei University, Seoul 03722, Korea*
[d] *Department of Computer and Information Science, University of Pennsylvania, PA 19104, USA*

ARTICLE INFO

Article history:
Received 22 June 2021
Received in revised form 8 April 2023
Accepted 15 April 2023
Available online 25 April 2023

Keywords:
Longitudinal data
Probit mixed models
Varying-coefficients
Partial collapsed Gibbs sampler
Dirichlet process

ABSTRACT

In the analysis of binary longitudinal data, it is of interest to model a dynamic relationship between a response and covariates as a function of time, while also investigating similar patterns of time-dependent interactions. We present a novel generalized varying-coefficient model that accounts for within-subject variability and simultaneously clusters varying-coefficient functions, without restricting the number of clusters nor overfitting the data. In the analysis of a heterogeneous series of binary data, the model extracts population-level fixed effects, cluster-level varying effects, and subject-level random effects. Various simulation studies show the validity and utility of the proposed method to correctly specify cluster-specific varying-coefficients when the number of clusters is unknown. The proposed method is applied to a heterogeneous series of binary data in the German Socioeconomic Panel (GSOEP) study, where we identify three major clusters demonstrating the different varying effects of socioeconomic predictors as a function of age on the working status.

1. Introduction

Mixed-effects models are commonly used in binary longitudinal studies in the social, behavioral, and health sciences. These models' popularity stems from their ability to capture longitudinal effects generated by repeated-measurement processes. To be more specific, random effects are introduced into linear models with only fixed effects to reflect the correlation between observations on the same subject. This extension also avoids some of the technical issues that can arise during the analysis of variance. For example, Stiratelli et al. (1984) showed that mixed-effects models have advantages over Markov models when dealing with a series of binary data because they are better at interpreting the effects of covariates and circumventing some of the difficult issues caused by unbalanced design or missing values. Other approaches to dealing with serial effects in longitudinal data provide practical recommendations by combining the mixed-effects models with other statistical methods (Varin and Czado, 2009; Guerra et al., 2012).

A varying-coefficient model has been shown to be extremely effective for modeling of time-varying effects in longitudinal studies (Hastie and Tibshirani, 1993; Hoover et al., 1998; Wu et al., 1998; Lang and Brezger, 2004; Sun and Wu, 2005; Fan and Zhang, 2008; Lu and Zhang, 2009; Jeong and Park, 2016; Jeong et al., 2017; Park and Jeong, 2018). Such varying-

[*] Corresponding author at: Department of Applied Statistics, Yonsei University, Seoul 03722, Korea.
E-mail addresses: sohn24@purdue.edu (J. Sohn), sjeong@yonsei.ac.kr (S. Jeong), jch0@seas.upenn.edu (Y.M. Cho), tpark@yonsei.ac.kr (T. Park).
[1] Equally contributed with the first author.

https://doi.org/10.1016/j.csda.2023.107766
0167-9473/© 2023 Elsevier B.V. All rights reserved.

coefficient functions can be easily modeled by Bayesian methods, e.g., Bayesian P-splines (Lang and Brezger, 2004), series priors (Shen and Ghosal, 2015), Gaussian process priors (Neal, 1998), Bayesian wavelets (Chipman et al., 1997), and free-knot splines and adaptive knot selection (Smith and Kohn, 1996; DiMatteo et al., 2001). The main advantage of using the Bayesian approaches is that uncertainty quantification is naturally performed with credible sets obtained by Markov chain Monte Carlo (MCMC). The method of free-knot splines and adaptive knot selection, in particular, exhibits natural local adaptation to spatially inhomogeneous smoothness (Smith and Kohn, 1996; Kohn et al., 2001; Ruppert et al., 2003; Kang and Jeong, 2023).

Traditional varying-coefficient mixed models focus on exploring common varying coefficients shared across all subjects. However, because there are often various sources of heterogeneity among subjects, particularly in longitudinal studies, such a common structure of varying coefficients may be oversimplified, leading to incorrect conclusions. To uncover the heterogeneity of the population, many model-based clustering approaches have been proposed from both frequentist and Bayesian perspectives (Lenk and DeSarbo, 2000; James and Sugar, 2003; Heard et al., 2006; Shi and Wang, 2008; Aßmann and Boysen-Hogrefe, 2011; Coffey et al., 2014; Berrettini et al., 2022). For example, James and Sugar (2003) suggested modeling individual basis coefficients by random effects with the mean indexed by the cluster. Heard et al. (2006) developed a hierarchical Bayesian model that avoids MCMC using their particular model formulation. Coffey et al. (2014) developed a clustering method for longitudinal gene profiles via penalized splines. More recently, Berrettini et al. (2022) devised a semi-parametric mixture model with mixture weights and conditional means that are modeled as nonlinear functions of covariates. Although these frameworks clearly offer inferential advantages in the presence of heterogeneity, most require determining the appropriate number of clusters and additional evaluation steps decoupled from parameter estimation such as cross validation.

In contrast, the Bayesian nonparametric framework naturally chooses the required number of clusters in a data-driven way by using stochastic process priors that randomly partition a sample space to be clustered. These priors include the Dirichlet process (DP) (Ferguson, 1973), the two-parameter Poisson-Dirichlet process (Pitman and Yor, 1997), and the generalized stick-breaking process (Ishwaran and James, 2001), to name a few. As they are basically infinite-dimensional priors, they have become essential clustering tools for modeling an infinite number of clusters in various areas (Ishwaran and James, 2001; Teh et al., 2006; Lau and Green, 2007; Wallach et al., 2010; Canale and Dunson, 2011; Yerebakan et al., 2014; Kyung, 2015). In the context of nonparametric regression, Müller et al. (1996) employed the DP prior to jointly partition the support of response and predictor variables, which performs locally weighted regression estimation in terms of Bayesian predictive inference. Gelfand et al. (2005) incorporates the dependent DP prior (MacEachern et al., 2001) into the Gaussian process prior for spatial analysis. Ray and Mallick (2006) studies the Bayesian wavelet regression model where the DP prior has a base measure that expedites sparsity of the wavelet coefficients. Petrone et al. (2009) further considered the local heterogeneity in a subgroup of curves by proposing a hybrid Dirichlet prior that overcomes the global heterogeneity. Chib and Greenberg (2010) used the DP prior for modeling an error distribution while approximating nonlinear components via cubic splines with a smoothness prior regularizing difference of spline coefficients at the knot locations. Rodriguez and Dunson (2014) employed the generalized DP prior to cluster curves smoothed by the free-knot spline method. Suarez and Ghosal (2016) assigned the DP prior on each wavelet coefficient independently not jointly on the set of coefficients with the sparsity structure used in Ray and Mallick (2006). Margaritella et al. (2021) applied the DP prior to the clustering of functional principle scores, which improves the curve and correlation reconstruction.

While functional clustering for continuous response has emerged, there have been relatively fewer works for longitudinal binary responses with nonparametric components. Kuss et al. (2006) considered a parametric logistic model where the cluster allocation is assumed to follow a multinomial distribution. Similarly, Aßmann and Boysen-Hogrefe (2011) designed a Bayesian probit regression model with the multinomial label allocation. On the one hand, Hannah et al. (2011) proposed a DP mixture model for generalized linear models in the spirit of Müller et al. (1996), with the restriction of a linear relationship within each cluster. More recently, Zhu et al. (2021) devised a model-free clustering method for binary longitudinal data using a pairwise penalty to nearby clusters, but their model does not account for any functional effects. We find that none of the aforementioned studies deal with both model-based clustering on a mixed-effects model (especially with the DP prior) and nonparametric function estimation with guaranteed smoothness.

Our contribution is threefold. First, we propose a flexible framework for simultaneously modeling population-level fixed effects, cluster-level varying effects, and subject-level random effects in the analysis of binary longitudinal data. The proposed model is a probit varying-coefficient mixed model that flexibly and adaptively identifies different subpopulations having their own varying-coefficient functions that can be either constant, linear, or nonlinear. Second, we devise new prior distributions for the posterior analysis and effective functional clustering of the proposed model. In particular, it is well known that the measurement scale of data must be considered in choosing the base measure for the DP prior (Gelman et al., 2015). We carefully design a prior distribution with a reasonable scale so that it can start a new cluster well within a sampler to account for an infinite number of clusters, while achieving suitable smoothing for function estimation with spatial adaptation. Third, we construct a partially collapsed Gibbs (PCG) sampler to cover the varying-dimensional parameter issue of a standard Gibbs sampler and facilitate posterior computation via the method of partial collapse (van Dyk and Park, 2008; Park and van Dyk, 2009). To maintain a transition kernel, a PCG sampler, unlike a standard Gibbs sampler, requires its steps to be performed in a specific order. We thus develop a PCG sampler that can be used in the fitting of the proposed model.

The remainder of this paper is organized as follows. In Section 2, we describe the probit varying-coefficient mixed model and discuss how the DP prior constructs model-based clustering. Section 3 specifies prior distributions and constructs efficient sampling steps based on the method of partial collapse. In Section 4, simulation studies are presented to validate the proposed method. Section 5 applies the proposed model to the GSOEP data, and Section 6 discusses the results. Appendix A contains a detailed description of the proposed method, while Appendix B describes how to install the R package for the proposed method. The R package is currently available on the first author's github[2] to demonstrate that all of the results in Sections 4 and 5 can be reproduced.

2. Probit varying-coefficient mixed models for functional clustering

Let Y_{ij} represent a binary response observed at time t_{ij} for observation j on subject i, where $i = 1, \dots, N$ and $j = 1, \dots, n_i$. The outcome of the response Y_{ij} can be expressed as an indicator function of the sign of a latent variable L_{ij}, i.e.,

$$Y_{ij} = I(L_{ij} > 0),$$

where the latent variable is introduced for computational convenience but can be interpreted as a utility difference between choosing $Y_{ij} = 1$ or 0.

In longitudinal studies, a relationship between the latent variable L_{ij} and the available covariates is commonly specified by a generalized linear mixed model to account for between-subject variability. Specifically, the generalized linear mixed model is expressed as

$$L_{ij} = \mathbf{X}_i^{(j)}\boldsymbol{\beta} + \mathbf{Z}_i^{(j)}\mathbf{b}_i + \epsilon_{ij}, \tag{1}$$

where $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iq})$ and $\mathbf{Z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{ir})$ are $n_i \times q$, and $n_i \times r$ design matrices for subject i, $\mathbf{X}_i^{(j)}$ and $\mathbf{Z}_i^{(j)}$ denote the jth row vectors of $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively, $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed effects, $\mathbf{b}_i$ is a $r \times 1$ vector of random effects for subject i, and ϵ_{ij} is an underlying error term that is assumed to follow a logistic or normal distribution in the logit or probit model, respectively. Note that the model in (1) involves multiple random effects on binary longitudinal data; for linear mixed models with multiple random effects, see Vines et al. (1996); Meng and van Dyk (1998); van Dyk (2000); Kim et al. (2013); Park and Min (2016).

In the presence of within-subject correlation over time, Jeong et al. (2017) extends the generalized linear mixed model in (1) to incorporate varying-coefficients $\boldsymbol{\alpha}(t)$ that vary over time t, i.e.,

$$L_{ij} = \mathbf{W}_i^{(j)}\boldsymbol{\alpha}(t_{ij}) + \mathbf{X}_i^{(j)}\boldsymbol{\beta} + \mathbf{Z}_i^{(j)}\mathbf{b}_i + \epsilon_{ij}, \tag{2}$$

where $\mathbf{W}_i = (\mathbf{w}_{i1}, \dots, \mathbf{w}_{ip})$ is $n_i \times p$ design matrices for subject i, $\mathbf{W}_i^{(j)}$ denotes the jth row vector of $\mathbf{W}_i$, $\boldsymbol{\alpha}(t) = (\alpha_1(t), \dots, \alpha_p(t))^\top$ is a $p \times 1$ vector of unknown smooth functions that vary over time t, and t_{ij} is the jth time of the ith subject. Note that $\boldsymbol{\alpha}(t_{ij})$ is a vector of real values of $\boldsymbol{\alpha}(t)$ evaluated at t_{ij}. The lth time-varying function in the vector $\boldsymbol{\alpha}(t)$, i.e., $\boldsymbol{\alpha}_l(t)$ can be modeled with regression splines that use a linear combination of basis functions, e.g.,

$$\mathbf{B}_l(t) = \left(1, t, |t - \omega_{l1}|^3, |t - \omega_{l2}|^3, \dots, \left|t - \omega_{lM_l}\right|^3\right)^\top, \tag{3}$$

where $\boldsymbol{\omega}_l = (\omega_{l1}, \dots, \omega_{lM_l})$ is an ordered sequence of knot-candidate locations within the range of observed time points, for $l = 1, \dots, p$. The amount of smoothness for the lth regression spline is controlled by the number M_l and locations $\boldsymbol{\omega}_l$.

To account for heterogeneity among the subjects while borrowing strength across the different subjects, we consider allocating each subject to its own cluster with different functions of varying-coefficients. To do so, we represent the set of functions $\boldsymbol{\alpha}(t)$ in (2) as the subject-level varying-coefficients, i.e., $\boldsymbol{\alpha}_i(t) = (\alpha_{i1}(t), \dots, \alpha_{ip}(t))^\top$. Each of the unknown subject-specific varying-coefficient functions is assumed to fall in the linear span of a set of its own basis functions according to basis selection, i.e.,

$$\alpha_{il}(t) \approx (\mathbf{B}_l(t) \odot \boldsymbol{\gamma}_{il})^\top \boldsymbol{\phi}_{il},$$

where $\odot$ denotes element-wise multiplication of vectors in accordance with values of $\boldsymbol{\gamma}_{il}$, $\boldsymbol{\gamma}_{il} = (1, \gamma_{il0}, \gamma_{il1}, \dots, \gamma_{ilM_l})^\top$ denotes an $(M_l + 2) \times 1$ vector of indicator variables for basis inclusion, $\gamma_{ilm} = 1$ represents that the $(m+2)$th element in $\boldsymbol{\gamma}_{il}$ is used as a basis function, and $\boldsymbol{\phi}_{il}$ denotes an $(M_l + 2) \times 1$ vector of basis coefficients corresponding to the lth varying-coefficient for subject i. The first element in $\boldsymbol{\gamma}_{il}$ equals one, so the constant basis function in (3) always remains in the model. When the corresponding covariate has no interaction with time, we have $\gamma_{ilm} = 0$ for $m = 0, \dots, M_l$, and the model in (2) is reduced to the generalized linear mixed model in (1). That is, when the true varying-coefficient function is constant, our model can estimate it as a constant function by choosing $\gamma_{ilm} = 0$ for $m = 0, \dots, M_l$, reducing modeling

[2] https://github.com/Jwsohn612/fvcc.

bias and avoiding the possibility of overfitting. When the true varying-coefficient function is nonlinear, selecting appropriate knots allows the estimated function to adapt to the true one's curvature. That is, we use data to adjust the spatially inhomogeneous smoothness of a varying-coefficient function, so that more knots are used in a high-curvature region and fewer knots in a low-curvature region. This implies that we do not need to a priori determine whether a varying-coefficient function is constant, linear, or nonlinear (Jeong and Park, 2016; Jeong et al., 2017, 2022). The value of M_l for the knot-candidates is not crucial as long as it is large enough to capture the global and local characteristics of a function. Following the literature (e.g., Kohn et al., 2001), we recommend using 20 to 30 knot-candidates chosen by the sample quantiles of the time variable t. If the time variable is repeatedly observed at some points in time, M_l should not be larger than the number of the non-duplicated values for t.

Next, the individualized vector of functions $\boldsymbol{\alpha}_i(t)$ is given the DP prior, which induces functional clustering with respect to the functions. Section 3.1 describes how the model leverages the DP prior to cluster varying-coefficients in detail. Let $C_i = k$ denote that subject i belongs to cluster k sharing identical basis functions for varying-coefficients, for $i = 1, \dots, N$ and $k = 1, \dots, K$. Then through the DP prior, we have $\boldsymbol{\alpha}_i(t) = \boldsymbol{\alpha}^*_k(t)$, which implies $\boldsymbol{\gamma}_{il} = \boldsymbol{\gamma}^*_{kl}$ and $\boldsymbol{\phi}_{il} = \boldsymbol{\phi}^*_{kl}$ as well. To be specific,

$$\alpha_{il}(t) \;=\; \alpha^*_{kl}(t) \;\approx\; (\mathbf{B}_l(t) \odot \boldsymbol{\gamma}^*_{kl})^\top \boldsymbol{\phi}^*_{kl}, \quad l = 1, \dots, p, \tag{4}$$

for the ith subject who is allocated to cluster k, having $C_i = k$. Thus, for the ith subject, the model in (2) can be represented in a matrix form,

$$\mathbf{L}_i \;=\; \sum_{l=1}^{p} \left(\mathbf{w}_{il} \odot \alpha^*_{C_i l}(t_{ij})|_{j=1}^{n_i} \right) + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{5}$$

where $\alpha^*_{C_i l}(t_{ij})|_{j=1}^{n_i} = \left(\alpha^*_{C_i l}(t_{i1}), \dots, \alpha^*_{C_i l}(t_{in_i}) \right)^\top$ is a $n_i \times 1$ vector of real values, $\mathbf{L}_i$ is an $n_i \times 1$ vector of latent variables, and $\boldsymbol{\epsilon}_i$ is a $n_i \times 1$ error vector. This representation implies that the clustering process is implemented with information about only dynamic covariates.

To express (5) with the approximation in (4), we define an $n_i \times (M_l + 2)$ matrix $\mathbf{B}^*_{il} = \left(\oplus_{j=1}^{n_i} (\mathbf{B}_l(t_{ij}) \odot \boldsymbol{\gamma}^*_{C_i l}) \right)^\top$ where $\oplus$ represents the direct sum of vectors or matrices, for the lth covariate of the ith subject. It is obvious that this matrix can have zero column vectors when the corresponding elements of $\boldsymbol{\gamma}^*_{C_i l}$ are zero. By removing the columns of 0's, we can obtain an $n_i \times |\boldsymbol{\gamma}^*_{C_i l}|$ submatrix of $\mathbf{B}^*_{il}$, which is denoted by $\mathbf{B}^\star_{il}$, where $|\boldsymbol{\gamma}^*_{C_i l}| = \sum_m \gamma^*_{C_i lm}$. Then the model in (5) can be written as

$$\mathbf{L}_i \;=\; \mathbf{W}^\star_{i(\boldsymbol{\gamma}^*_{C_i})} \boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}^*_{C_i})} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{6}$$

where $\boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}^*_{C_i})}$ is a vector of cluster-level basis coefficients whose size is the sum of all elements of $\boldsymbol{\gamma}^*_{C_i}$. It is

$$\boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}^*_{C_i})} \;=\; \left(\boldsymbol{\phi}^*_{\boldsymbol{\gamma}^*_{C_i 1}}, \dots, \boldsymbol{\phi}^*_{\boldsymbol{\gamma}^*_{C_i p}} \right)^\top \;\in\; \mathbb{R}^{|\boldsymbol{\gamma}^*_{C_i}| \times 1},$$

where $|\boldsymbol{\gamma}^*_{C_i}| = \sum_{l,m} \gamma^*_{C_i lm}$ and $\boldsymbol{\phi}^*_{\boldsymbol{\gamma}^*_{C_i l}}$ is a $|\boldsymbol{\gamma}^*_{C_i l}| \times 1$ subvector of $\boldsymbol{\phi}^*_{C_i l}$ whose elements correspond to nonzero columns of $\mathbf{B}^*_{il}$. Then, the design matrix $\mathbf{W}^\star_{i(\boldsymbol{\gamma}^*_{C_i})}$ is constructed by multiplying each set of selected basis terms to each column of $\mathbf{W}_i$, i.e.,

$$\mathbf{W}^\star_{i(\boldsymbol{\gamma}^*_{C_i})} \;=\; \left[\mathbf{w}_{i1} \mathbf{1}^\top_{|\boldsymbol{\gamma}^*_{C_i 1}|} \odot \mathbf{B}^\star_{i1}, \dots, \mathbf{w}_{ip} \mathbf{1}^\top_{|\boldsymbol{\gamma}^*_{C_i p}|} \odot \mathbf{B}^\star_{ip} \right] \;\in\; \mathbb{R}^{n_i \times |\boldsymbol{\gamma}^*_{C_i}|}.$$

where $\mathbf{1}_{|\boldsymbol{\gamma}^*_{C_i l}|}$ is a vector of ones in $\mathbb{R}^{|\boldsymbol{\gamma}^*_{C_i l}| \times 1}$.

3. Bayesian analysis

3.1. Dirichlet process prior

In this paper, the DP is used as a prior distribution to cluster $\boldsymbol{\alpha}_i(t)$, and this clustering procedure is equivalent to clustering the set of $(\boldsymbol{\phi}_i, \boldsymbol{\gamma}_i)$, where $\boldsymbol{\phi}_i = (\boldsymbol{\phi}_{i1}, \dots, \boldsymbol{\phi}_{ip})^\top$ and $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}, \dots, \boldsymbol{\gamma}_{ip})^\top$, as implied in (4). This process assigning the DP prior to the set of $(\boldsymbol{\phi}_i, \boldsymbol{\gamma}_i)$ is expressed as

$$\begin{aligned} (\boldsymbol{\phi}_i, \boldsymbol{\gamma}_i)|\mathcal{H} &\overset{\text{iid}}{\sim} \mathcal{H}, \quad i = 1, \dots, N, \\ \mathcal{H} &\sim \text{DP}(\nu, \mathcal{H}_0), \end{aligned}$$

where $\nu > 0$ and $\mathcal{H}_0$ is a base distribution that randomly generates cluster-level parameters for $(\boldsymbol{\phi}_i, \boldsymbol{\gamma}_i)$. As another representation of the DP, it is worthwhile to look at the stick-breaking process (Sethuraman, 1994; Ishwaran and James, 2001) that allows the truncation of the summation in the DP after a large K component, i.e.,

$$\mathcal{H}(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{(\boldsymbol{\phi}_k^*, \boldsymbol{\gamma}_k^*)}(\cdot) \approx \sum_{k=1}^{K} \pi_k \delta_{(\boldsymbol{\phi}_k^*, \boldsymbol{\gamma}_k^*)}(\cdot), \quad (\boldsymbol{\phi}_k^*, \boldsymbol{\gamma}_k^*) \overset{\text{iid}}{\sim} \mathcal{H}_0, \tag{7}$$

where $\boldsymbol{\phi}_k^* = (\boldsymbol{\phi}_{k1}^*, \ldots, \boldsymbol{\phi}_{kp}^*)^\top$ and $\boldsymbol{\gamma}_k^* = (\boldsymbol{\gamma}_{k1}^*, \ldots, \boldsymbol{\gamma}_{kp}^*)^\top$ are the parameters for cluster k, $\delta_{(\boldsymbol{\phi}_k^*, \boldsymbol{\gamma}_k^*)}(\cdot)$ is a Dirac measure at $(\boldsymbol{\phi}_k^*, \boldsymbol{\gamma}_k^*)$, π_k is the probability mass at atom $(\boldsymbol{\phi}_k^*, \boldsymbol{\gamma}_k^*)$, and K is a finite truncation for the maximum number of clusters. The equation in (7) implies that the model in (6) explores latent subpopulations by limiting the maximum number of subpopulations to K, not to infinity; see Ishwaran and James (2001) for theoretical arguments. Meanwhile, the set of cluster-level parameters, $\boldsymbol{\phi}_k^*$ and $\boldsymbol{\gamma}_k^*$, is drawn from the base distribution $\mathcal{H}_0$, and the random weight π_k derives from a set of random variables that each follows a beta distribution, i.e.,

$$\pi_k = \pi_k(\mathbf{V}) = V_k \prod_{\ell<k} (1 - V_\ell), \quad V_k \overset{\text{ind}}{\sim} \text{Beta}(1, \nu), \quad k = 1, \ldots, K-1,$$

where $\mathbf{V} = \{V_1, \ldots, V_K\}$ and $V_K = 1$, which guarantees the sum of all random weights is equal to one. Then, we can write $P(C_i = k|\mathbf{V}) = \pi_k(\mathbf{V})$. The specification of ν may affect clustering performance. As ν goes to 0, the concentration toward the existing clusters gets stronger by decreasing the probability that a vector $(\boldsymbol{\phi}_i, \boldsymbol{\gamma}_i)$ forms a new cluster. In this work, we set $\nu = 1$ by default so that every subject has the equal probability for shaping a new cluster.

3.2. Prior specification for the submodel parameters

This section discusses the specification of prior distributions of each model component. The set of indicator variables $\boldsymbol{\gamma}_{kl}^*$ has a beta-binomial prior distribution, i.e.,

$$p(\boldsymbol{\gamma}_{kl}^*) \propto B(|\boldsymbol{\gamma}_{kl}^*| + a, M_l + 1 - |\boldsymbol{\gamma}_{kl}^*| + b), \tag{8}$$

for $k = 1, \ldots, K$ and $l = 1, \ldots, p$, where $B(\cdot, \cdot)$ denotes the beta function. If $a = b = 1$, this prior distribution allocates equal probabilities for the number of active knots (Scott and Berger, 2010). This choice has been shown work successfully for function estimation with knot selection (Jeong and Park, 2016; Jeong et al., 2017).

For the basis coefficients $\boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}_k^*)}$ that have varying dimension in each iteration, we consider the following mixture of g-priors,

$$\begin{aligned}
\boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}_k^*)} | (\boldsymbol{\gamma}_k^*, \tau_k) &\overset{\text{ind}}{\sim} \text{N}_{|\boldsymbol{\gamma}_k^*|}\left(\mathbf{0}, \tau_k \mathbf{R}^{-1}_{k(\boldsymbol{\gamma}_k^*)}\right), \\
\mathbf{R}_{k(\boldsymbol{\gamma}_k^*)} &= \sum_{i=1}^{N} \mathbf{W}^{\star\top}_{i(\boldsymbol{\gamma}_k^*)} \mathbf{W}^\star_{i(\boldsymbol{\gamma}_k^*)}, \quad k = 1, \ldots, K, \\
\tau_k &\overset{\text{iid}}{\sim} \text{IG}(1/2, N/2), \quad k = 1 \ldots, K.
\end{aligned} \tag{9}$$

By characterizing the scale parameter with the total number of subjects, the prior in (9) corresponds to a Zellner-Siow prior, which is a multivariate Cauchy prior marginally for $\boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}_k^*)}$ (Liang et al., 2008). Therefore, the base measure $\mathcal{H}_0$ is constructed by combining a beta-binomial distribution and a multivariate Cauchy distribution.

The prior in (9) has several desirable properties. First, the prior distribution is invariant to linear transformations of the design matrix (Zellner, 1986). This means that the posterior distribution of the varying-coefficients is not affected by linear transformations of the basis functions in (3). More importantly, the prior in (9) utilizes the population-level covariance, which is determined by assuming that all individuals belong to the kth cluster. This specific structure enhances the convergence efficiency of MCMC because the cluster-level basis coefficients $\boldsymbol{\phi}^\star_{(\boldsymbol{\gamma}_k^*)}$ of empty clusters are sampled by taking advantage of the summed information of all subjects. As a result, the prior can naturally begin a new cluster to which a few subjects may belong. Indeed, it is well known that choosing a reasonable scale for the base measure $\mathcal{H}_0$ is very important in using the DP prior (Gelman et al., 2015). In this regard, our prior construction in (9) has a clear advantage over the related studies, which use the DP prior but do not fully account the concern of scale (Ray and Mallick, 2006; Rodriguez and Dunson, 2014). Furthermore, having the right scale may be difficult with other penalty priors. For example, the Bayesian P-spline is a widely used Bayesian approach for nonparametric regression (Lang and Brezger, 2004). Since its covariance structure plays an important role in smoothing, however, the prior distribution cannot be simply modified to have a reasonable scale for the DP prior.

The remaining specification of priors for the fixed-dimensional parameter is standard. As for the fixed effects $\boldsymbol{\beta}$, we assign a multivariate normal distribution whose covariance matrix is positive definite, i.e.,

Algorithm 1: One iteration of the PCG sampler.

```
Initialize: $(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{b}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L})$
for $k = 1, 2, \ldots, K$ do
    for $l = 1, 2, \ldots, p$ do
        Step 1: Draw $\boldsymbol{\gamma}^*_{kl}$ from $p(\boldsymbol{\gamma}^*_{kl} | \boldsymbol{\gamma}^*_{-kl}, \boldsymbol{\beta}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$
for $k = 1, 2, \ldots, K$ do
    Step 2: Draw $V_k$ from $p(V_k | \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{C}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$
for $k = 1, 2, \ldots, K$ do
    Step 3: Draw $\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_k)}$ from $p(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_k)} | \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$
Step 4: Draw $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta} | \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$,
for $i = 1, 2, \ldots, N$ do
    Step 5: Draw $\mathbf{b}_i$ from $p(\mathbf{b}_i | \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$
for $k = 1, 2, \ldots, K$ do
    Step 6: Draw $\tau_k$ from $p(\tau_k | \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{b}, \mathbf{C}, \mathbf{V}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$
Step 7: Draw $\boldsymbol{\Psi}$ from $p(\boldsymbol{\Psi} | \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{b}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \mathbf{L}, \mathbf{Y})$,
for $i = 1, 2, \ldots, N$ do
    for $j = 1, 2, \ldots, n_i$ do
        Step 8: Draw $L_{ij}$ from $p(L_{ij} | \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{b}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{Y})$
for $i = 1, 2, \ldots, N$ do
    Step 9: Draw $C_i$ from $p(C_i | \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{b}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$
```

$$\boldsymbol{\beta} \sim \mathrm{N}_q(\mathbf{0}, \mathbf{P}).$$

For practical purposes, $\mathbf{P}$ can be chosen as a diagonal matrix with large diagonal entries. For random effects, a multivariate normal distribution is used to generate the effects,

$$\mathbf{b}_i | \boldsymbol{\Psi} \overset{\text{iid}}{\sim} \mathrm{N}_r(\mathbf{0}, \boldsymbol{\Psi}), \quad i = 1, \ldots, N,$$

where $\boldsymbol{\Psi}$ is the covariance matrix of the random effects and has an inverse-Wishart prior,

$$\boldsymbol{\Psi} \sim \mathrm{IW}(u, \mathbf{D}).$$

In our study, u and $\mathbf{D}$ are fixed in advance to make the prior distribution diffuse.

3.3. Partially collapsed Gibbs sampler

Given the prior distributions in Section 3.2, we propose a sampling algorithm used to simulate the target posterior distribution,

$$p(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{b}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L} | \mathbf{Y}), \tag{10}$$

where $\boldsymbol{\gamma}^* = \{\boldsymbol{\gamma}^*_1, \ldots, \boldsymbol{\gamma}^*_K\}$, $\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)} = \{\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_1)}, \ldots, \boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_K)}\}$, $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_N\}$, $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_K\}$, $\mathbf{C} = \{C_1, \ldots, C_N\}$, $\mathbf{L} = \{\mathbf{L}_1, \ldots, \mathbf{L}_N\}$, and $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ denoting $\mathbf{Y}_i$ as a set of binary responses for subject i. To simulate the target posterior distribution in (10), a standard Gibbs sampler based on its full conditional distributions cannot be implemented because the dimension $|\boldsymbol{\gamma}^*_k| \times 1$ of $\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}$ depends on another model component $\boldsymbol{\gamma}^*$. In such a varying-dimensional problem, PCG sampling avoids the need of jumping between spaces of different dimensions through marginalization, permutation, and trimming, thereby making it implementable with the expectation of faster convergence; refer to Section 4 in Park and van Dyk (2009). In this study, we consider marginalizing the random effects $\mathbf{b}$ and the basis coefficients $\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}$ in (10), thereby producing the following marginal distributions,

$$p(\boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L} | \mathbf{Y}), \tag{11}$$

$$p(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)}, \boldsymbol{\gamma}^*, \boldsymbol{\beta}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L} | \mathbf{Y}). \tag{12}$$

One iteration of the PCG sampler is shown in Algorithm 1. Steps 1 and 2 are marginalized by using (11), while Steps 3 and 4 are marginalized by using (12). To maintain the target stationary distribution of Algorithm 1, the sampling steps are permuted in a specific order. Trimming is used to remove the redundant samples of components. For more applications of the PCG sampling including other varying dimensional cases, refer to Park and van Dyk (2009); Jeong and Park (2016); Jeong et al. (2017); Park and Jeong (2018); Park et al. (2019). Because the target stationary distribution of Algorithm 1 is maintained in a specific order, the change of the order of sampling steps may not guarantee the stationarity of a Markov chain, and care must be taken not to change the sampling order; refer to van Dyk and Park (2008). The details of Algorithm 1 are given in Appendix A.

4. A simulation study

In this section, we validate the robustness and sensitivity of the proposed method through extensive simulation studies. All simulation results are based on 300 replicated datasets.

4.1. Simulation setting

Throughout the simulation, we mainly consider three groups of varying coefficients, and the total number N of subjects and the number of subjects in each cluster will be specified later based on simulation setups. The values of an underlying effect modifier t, $\{t_{ij} : i = 1, \ldots, N, j = 1, \ldots, n_i\}$, are randomly generated from a uniform distribution between 0 and 1, and the values of known covariates for subject i, i.e., $\mathbf{W}_i$, $\mathbf{X}_i$, and $\mathbf{Z}_i$, are independently generated from a standard normal distribution, except that the first column of both $\mathbf{W}_i$ and $\mathbf{Z}_i$ is set to a column vector of 1's. Within the range of t between 0 and 1, $\alpha_{kl}(t)$ denotes a varying-coefficient function of the lth covariate in cluster k. The varying coefficients for three clusters are constant, linear, or nonlinear, as described below:

$$\begin{aligned}
\alpha_{11}(t) &= 2\exp\{-200(t-0.2)^2\} + \exp\{-10(t-0.6)^2\}, \\
\alpha_{12}(t) &= \sin(2\pi t^3), \\
\alpha_{21}(t) &= \sin\{8(t-0.5)\} + 1.5\exp\{-400(t-0.5)^2\}, \\
\alpha_{22}(t) &= 2t, \\
\alpha_{31}(t) &= -2t, \\
\alpha_{32}(t) &= 0.
\end{aligned}$$

The true values of the other model parameters are set to $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (1, -1)^\top$ and

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix} = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.8 \end{pmatrix}.$$

The latent response L_{ij} of the probit varying-coefficient mixed model is drawn from

$$L_{ij} \sim \mathrm{N}\left(\mathbf{W}_i^{(j)} \boldsymbol{\alpha}_{C_i}(t_{ij}) + \mathbf{X}_i^{(j)} \boldsymbol{\beta} + \mathbf{Z}_i^{(j)} \mathbf{b}_i, 1\right), \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i,$$

where $\boldsymbol{\alpha}_{C_i}(t_{ij}) = (\alpha_{C_i 1}(t_{ij}), \alpha_{C_i 2}(t_{ij}))^\top$ and it is used to generate a series of binary data such that $Y_{ij} = I(L_{ij} > 0)$ for observation j on subject i.

4.2. Performance of the proposed method

In this section, we demonstrate the performance of the proposed method under various simulation setups. We consider three different scenarios: Scenario I, where each cluster has 400 subjects ($N = 1200$), Scenario II, where each cluster has 200 subjects ($N = 600$), and Scenario III, where the three clusters have 600, 400, and 200 subjects, respectively ($N = 1200$). For each scenario, three different values of the concentration parameter are also considered to examine the robustness of the DP prior: $\nu \in \{0.1, 1, 10\}$. The proposed method is applied to each combination of the scenarios for the sample size and the concentration parameter with 300 replications of the datasets. We run 20,000 iterations of the PCG sampler, discarding the first half of the draws as burn-in and using the second half for our posterior analysis.

The side-by-side boxplots in Fig. 1 illustrate clustering performance for the 300 replicated datasets in terms of precision, recall, and F1-score. The clustering labels of all subjects are chosen by the posterior modes. As expected, clusters with larger sample sizes perform better in clustering. The metrics show similar results across different values of the concentration parameter ν, indicating that clustering performance is robust to the specification of the hyperparameter for the DP prior. We calculate coverage probabilities of the pointwise 95% credible bands with the 300 replicates, where the bands are specified by the 2.5% and 97.5% posterior quantiles. Fig. 2 shows that the coverage probabilities are close to the nominal value of 0.95, which validates the uncertainty quantification through the posterior distribution. The coverage probabilities are consistent for different values of ν, further supporting the robustness of our proposed method against the hyperparameter specification. With 300 replicated datasets, Table 1 shows the root-mean-square errors (RMSE) of the posterior median and coverage probabilities of the 95% credible intervals for the fixed-dimensional parameters. The results indicate that the fixed-dimensional parameters are also insensitive to the hyperparameter specification.

To assess the efficiency of our proposed PCG sampler, we calculate the multivariate effective sample size (ESS) (Vats et al., 2019) for all simulation settings, as shown in Fig. 3. The second half of the chain with 20,000 iterations is used to calculate the multivariate ESS, along with the running time of the sampler on a server equipped with CentOS7 and two Sky Lake CPUs @ 2.60GHz. The target parameters have a dimension of 185, including two varying-coefficient functions

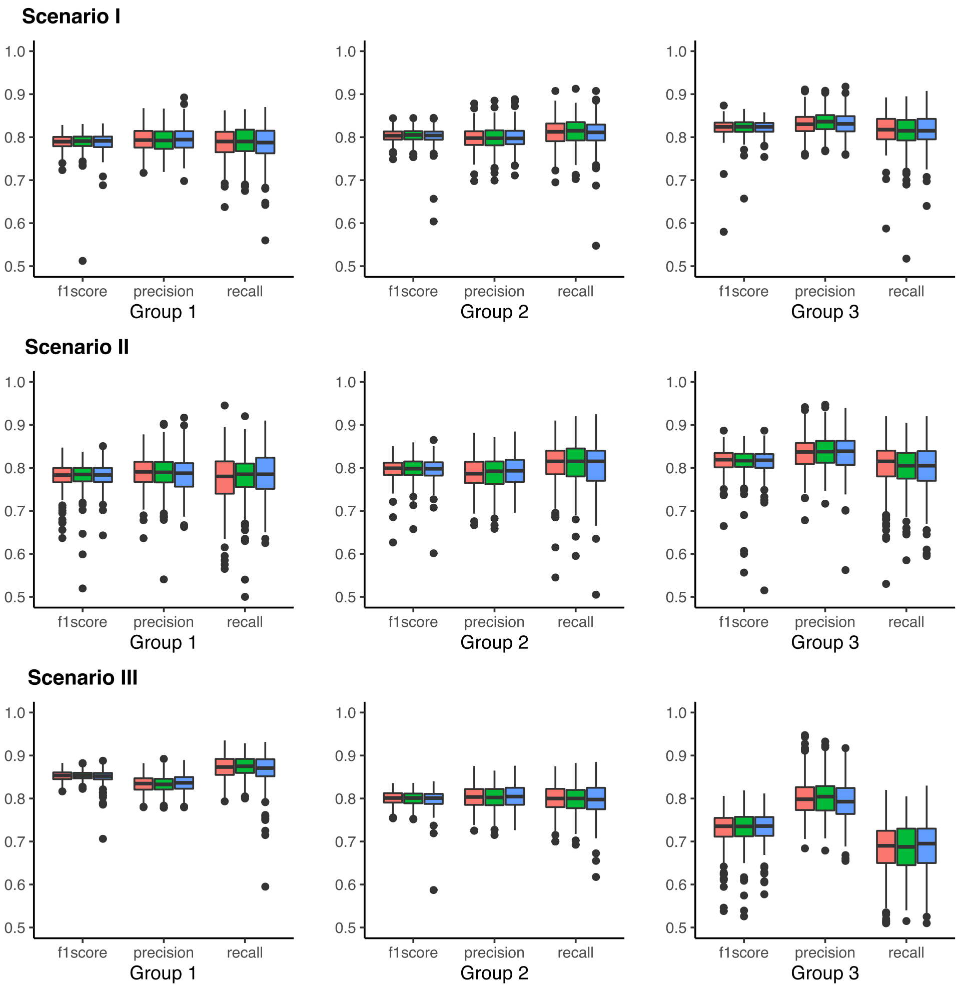

Fig. 1. Performance measures for the clustering procedure obtained from 300 replicated datasets with $\nu = 0.1$ (red), $\nu = 1$ (green), and $\nu = 10$ (blue). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 1
RMSE of the posterior median and coverage probabilities of the 95% credible intervals for the fixed-dimensional parameters obtained from 300 replicated datasets.

		RMSE			95% coverage		
	Parameter	$\nu = 0.1$	$\nu = 1$	$\nu = 10$	$\nu = 0.1$	$\nu = 1$	$\nu = 10$
Scenario I	β_1	0.029	0.030	0.031	0.930	0.930	0.930
	β_2	0.032	0.033	0.033	0.913	0.920	0.909
	ψ_{11}	0.070	0.072	0.072	0.937	0.916	0.937
	ψ_{12}	0.042	0.043	0.043	0.955	0.944	0.948
	ψ_{22}	0.073	0.074	0.075	0.937	0.934	0.941
Scenario II	β_1	0.042	0.046	0.044	0.962	0.919	0.920
	β_2	0.041	0.047	0.046	0.941	0.912	0.934
	ψ_{11}	0.119	0.124	0.121	0.882	0.908	0.885
	ψ_{12}	0.067	0.069	0.069	0.948	0.951	0.944
	ψ_{22}	0.102	0.110	0.107	0.948	0.930	0.948
Scenario III	β_1	0.029	0.030	0.031	0.946	0.932	0.917
	β_2	0.032	0.031	0.034	0.907	0.929	0.897
	ψ_{11}	0.063	0.064	0.064	0.929	0.943	0.921
	ψ_{12}	0.041	0.041	0.042	0.946	0.957	0.952
	ψ_{22}	0.077	0.077	0.079	0.946	0.954	0.941

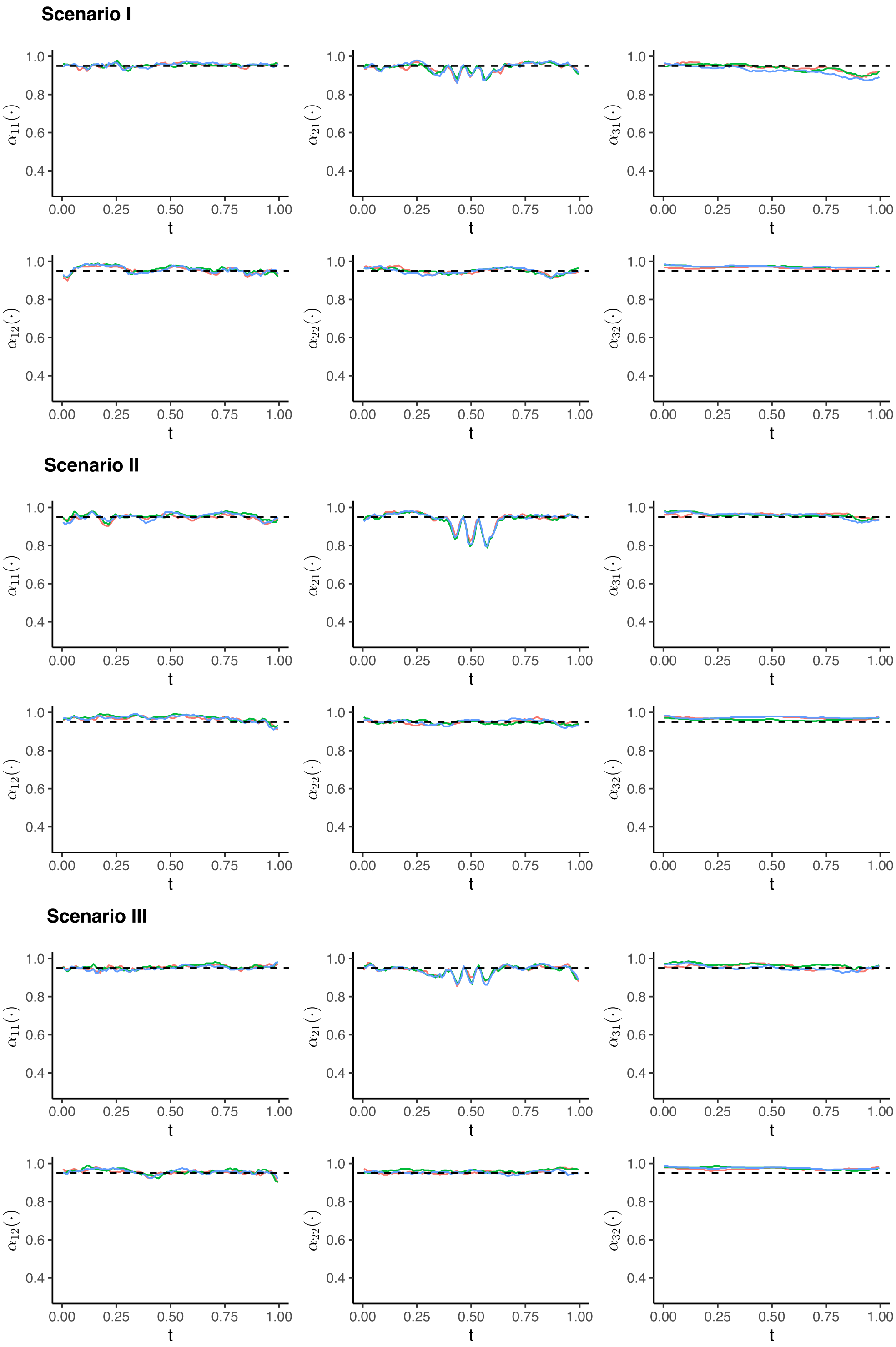

Fig. 2. Coverage probabilities of the pointwise 95% credible bands for the varying coefficients obtained from 300 replicated datasets with $\nu = 0.1$ (red), $\nu = 1$ (green), and $\nu = 10$ (blue).

with 30 knots for each of three clusters and five fixed-dimensional parameters. As shown in Fig. 3, the multivariate ESS is approximately 4,000–5,000 out of 10,000 iterations, implying that the proposed sampling algorithm exhibits reasonable convergence characteristics. Fig. 3 also shows the multivariate ESS divided by time (seconds), providing the number of independent draws obtained per unit time. The results demonstrate that roughly 1.5 to 2.5 independent draws are obtained

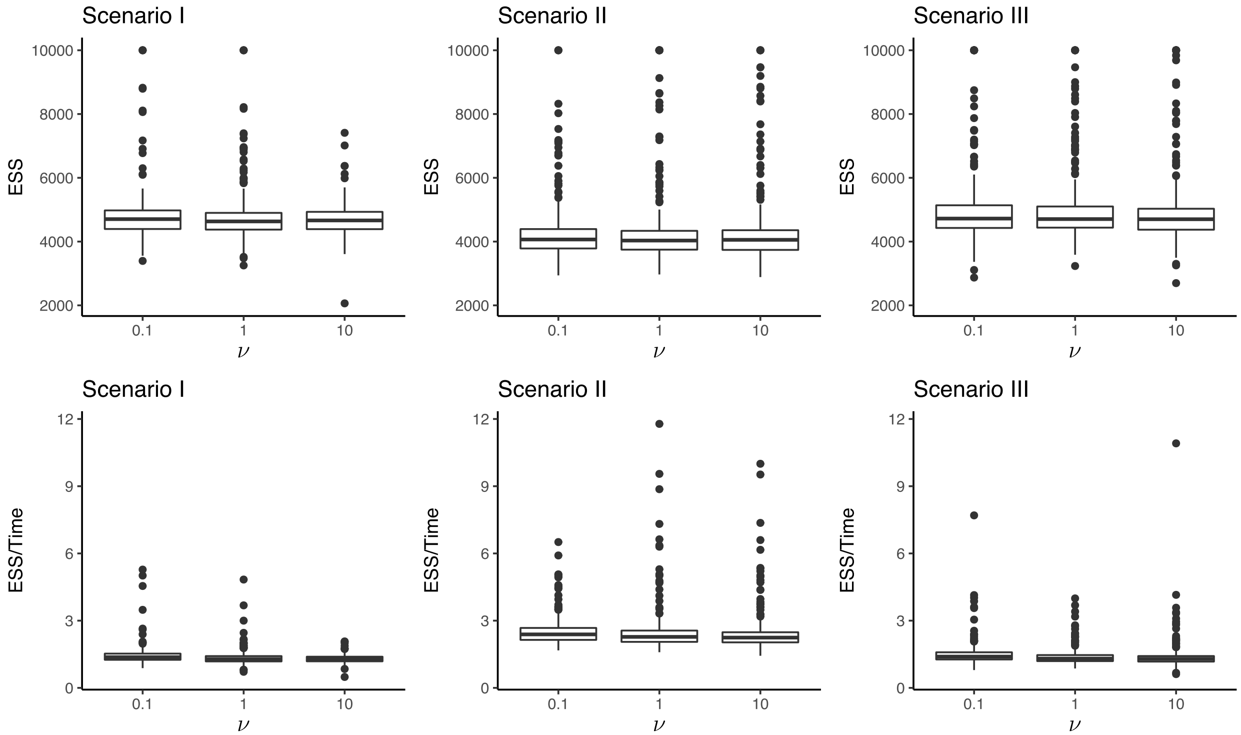

Fig. 3. Estimates of the multivariate ESS out of 10,000 PCG iterations and the multivariate ESS per second. The estimates of the multivariate ESS are replaced by 10,000 if they are larger than 10,000.

per second. Furthermore, the concentration hyperparameter has no significant influence on the multivariate ESS, indicating the robustness of our proposed method to the specification of the hyperparameter.

Fig. 4 overlays the pointwise posterior medians of the varying coefficients of the 300 replicated datasets using the default concentration parameter value of $\nu = 1$. Our results show that the estimated posterior medians become closer to the true functions as sample sizes increase. We also examined the posterior medians obtained with alternative values of ν, specifically $\nu = 0.1$ and $\nu = 10$, but found that the results were similar and therefore omitted here.

4.3. Effects of ignoring subpopulation

To evaluate the necessity of subpopulation modeling, we compared our proposed method with two competing approaches that do not account for functional clustering: Jeong et al. (2017) and the `mgcv` package (Wood, 2017). Jeong et al. (2017) is a fully Bayesian method for estimating probit varying-coefficient mixed models with a homogeneous population using free-knot splines (Smith and Kohn, 1996). The `mgcv` package uses penalized quasi-likelihood to estimate the same model. To avoid redundancy, we compared the two competing methods with our proposed method using replicated datasets under Scenario I specified in Section 4.2

In Fig. 5, we present the estimates of varying coefficients obtained by two competing methods that do not account for functional clustering. The estimated trends appear to be the average of the varying coefficients across the clusters in Scenario I, resulting in significant estimation bias for the cluster-specific effects. Furthermore, ignoring subpopulations leads to significant estimation bias for the fixed-dimensional parameters, as demonstrated in Table 2, despite the fact that they are common across all clusters. This suggests that accounting for subpopulation effects is crucial even for parameters that are shared among clusters.

4.4. Homogeneous population

We also investigate the performance of the proposed method in a homogeneous population with a single true cluster. To generate the simulation datasets, we set all subjects to have the same values of $\alpha_{11}(t)$ and $\alpha_{12}(t)$, as described in Section 4.1. Specifically, we set $\boldsymbol{\alpha}_{C_i}(t_{ij}) = (\alpha_{11}(t_{ij}), \alpha_{12}(t_{ij}))^\top$ for all i, resulting in a single cluster. The number of subjects is set to $N = 300$, and all other simulation settings are identical to those specified in Section 4.1.

Fig. 6 displays the pointwise posterior medians of the varying coefficients using the proposed method and two competitors under the homogeneous population assumption. The proposed method shows a reasonably small estimation bias compared to Jeong et al. (2017) despite some deviation for the incorrectly clustered subjects, implying its applicability

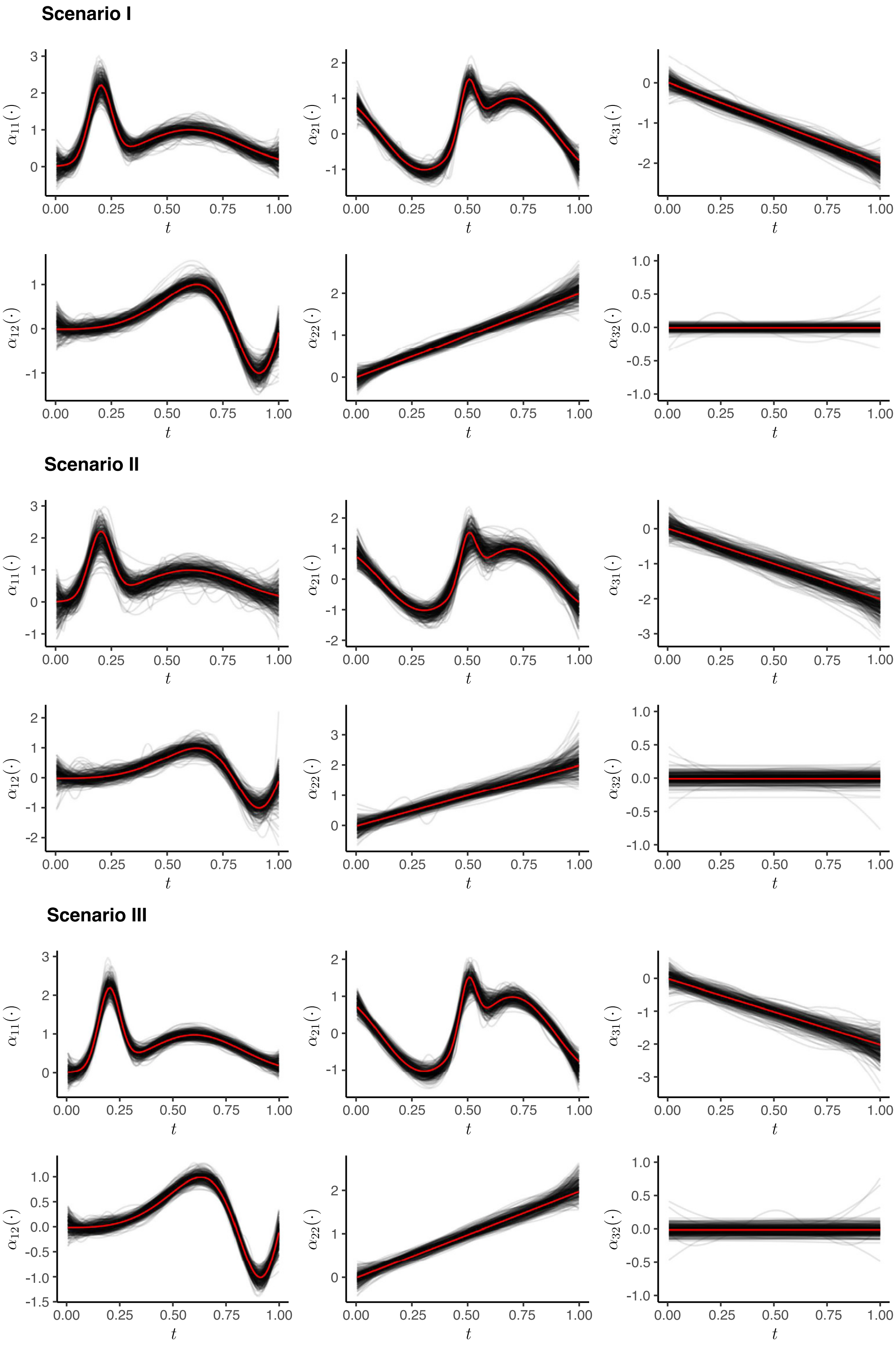

Fig. 4. Pointwise posterior medians of the varying coefficients for 300 replicated datasets (gray solid lines) and the true varying-coefficient functions (red solid lines).

without knowing the population structure. In contrast, mgcv yields larger estimation bias than the other two methods. Fig. 7 presents the coverage probabilities of the pointwise 95% credible or confidence bands of the varying coefficients, further demonstrating the worse performance of mgcv. Table 3 summarizes the results of the fixed-dimensional parameters, indicating that the RMSEs of the proposed method are slightly larger than those for Jeong et al. (2017). Considering

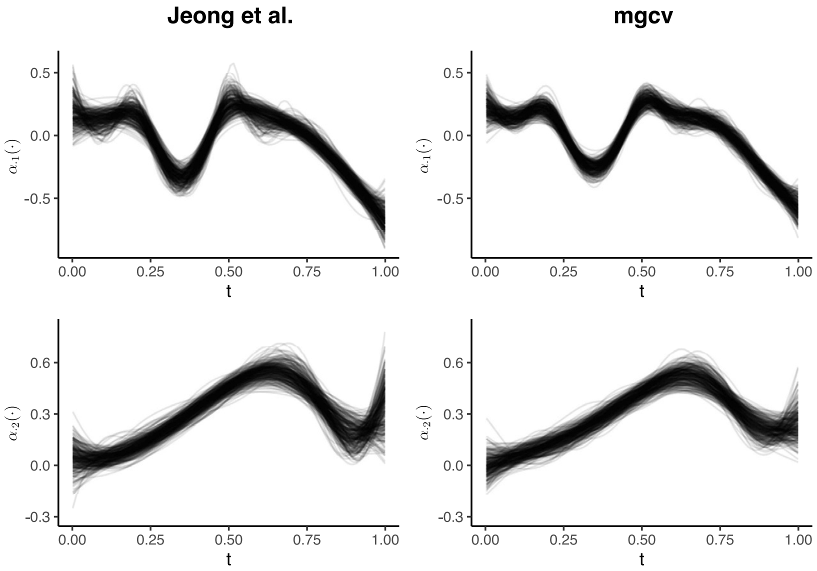

Fig. 5. Estimates of the varying coefficients obtained from 300 replicated datasets. The estimates are the pointwise posterior medians for Jeong et al. (2017) and the penalized quasi-likelihood estimates for mgcv.

Table 2
RMSE of the estimates and coverage probabilities of the 95% intervals for the fixed-dimensional parameters obtained from 300 replicated datasets. The coverage probabilities are obtained by the 95% credible intervals for Jeong et al. (2017) and the 95% confidence intervals approximated with standard errors for mgcv. The mgcv package does not provide standard errors for the covariance of random effects.

	Proposed method ($\nu = 1$)		Jeong et al. (2017)		mgcv	
Parameter	RMSE	95% coverage	RMSE	95% coverage	RMSE	95% coverage
β_1	0.030	0.930	0.205	0.000	0.190	0.000
β_2	0.033	0.920	0.204	0.000	0.190	0.000
ψ_{11}	0.072	0.916	0.220	0.000	0.199	–
ψ_{12}	0.043	0.944	0.097	0.170	0.099	–
ψ_{22}	0.074	0.934	0.299	0.000	0.351	–

Table 3
RMSE of the estimates and coverage probabilities of the 95% intervals for the fixed-dimensional parameters obtained from 300 replicated datasets. The coverage probabilities are obtained by the 95% credible intervals for Jeong et al. (2017) and the 95% confidence intervals approximated with standard errors for mgcv. The mgcv package does not provide standard errors for the covariance of random effects.

	Proposed method ($\nu = 1$)		Jeong et al. (2017)		mgcv	
Parameter	RMSE	95% coverage	RMSE	95% coverage	RMSE	95% coverage
β_1	0.055	0.954	0.048	0.960	0.058	0.916
β_2	0.055	0.944	0.050	0.953	0.058	0.923
ψ_{11}	0.111	0.923	0.089	0.950	0.678	-
ψ_{12}	0.073	0.958	0.066	0.960	0.093	-
ψ_{22}	0.143	0.944	0.135	0.943	0.148	-

the flexibility of the proposed method in accounting for potentially heterogeneous populations, however, it can be deemed more useful than Jeong et al. (2017).

5. Application to binary longitudinal data

5.1. Data description and modeling procedure

In this section, we consider the German Socioeconomic Panel (GSOEP) data (Riphahn et al., 2003). The dataset consists of repeated observations from 7,293 subjects in Germany for the years 1984–1988, 1991, and 1994. The response variable of interest is working status (employed=1; otherwise=0) and covariates consist of A_{ij} (age), M_{ij} (marital status; married=1, otherwise=0), K_{ij} (children under the age of 16 in the household; yes=1, no=0), H_{ij} (degree of handicap; 0 to 100 in percent), and S_{ij} (personal health satisfaction; 0 to 10). We confine our samples to 893 subjects under the age of 53 with

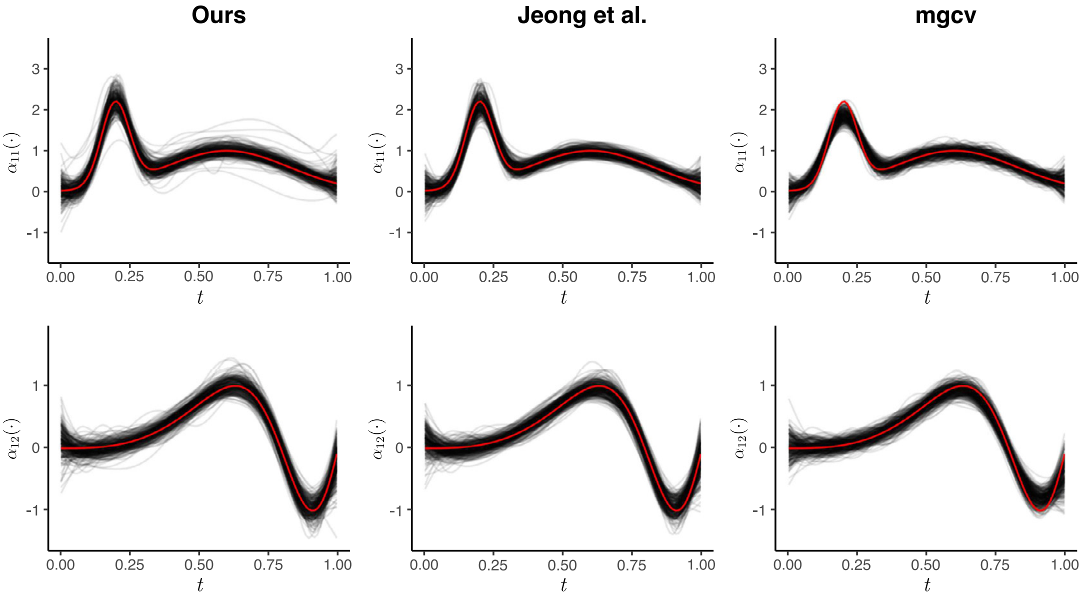

Fig. 6. Estimates of varying coefficients of 300 replications (solid gray lines) and the true functions (solid red lines). The estimates are the pointwise posterior medians for the proposed method and Jeong et al. (2017) and the penalized quasi-likelihood estimates for mgcv.

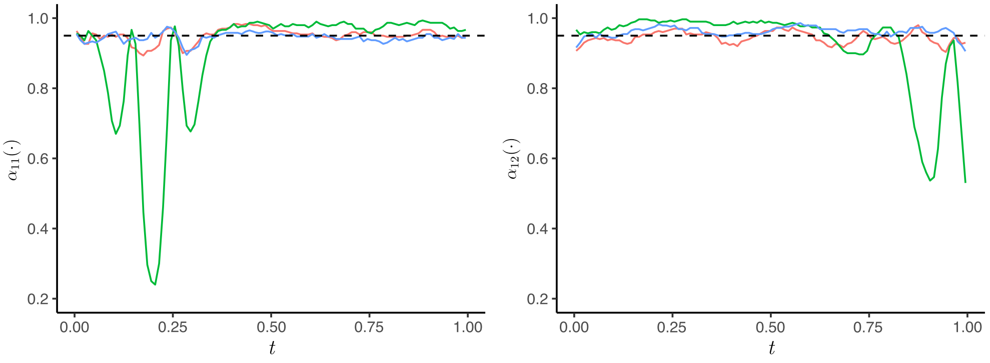

Fig. 7. Coverage probabilities of the 95% intervals for the varying coefficients obtained from 300 replicated datasets: the proposed method (red), Jeong et al. (2017) (blue), and mgcv (green). The coverage probabilities are obtained by the pointwise 95% credible bands for the proposed method and Jeong et al. (2017) and the pointwise 95% confidence bands approximated with standard errors for mgcv.

Abitur degrees in order to examine the varying effect of having young children in the household on working status as a function of age for people with secondary education.

According to our preliminary analysis, which assumes varying effects for all covariates, we have decided to treat handicap, personal health satisfaction, and marital status as constant effects in subsequent analyses. As a result, we aim to model the varying effects of the intercept and the presence of children under the age of 16, while treating the remaining covariates as fixed effects. Our target model is then given by

$$L_{ij} = \alpha^*_{C_i0}(A_{ij}) + \alpha^*_{C_iK}(A_{ij})K_{ij} + \beta_M M_{ij} + \beta_S S_{ij} + \beta_H H_{ij} + b_i + \epsilon_{ij}.$$

The reduced model complexity based on the preliminary analysis leads to faster computation and greater stability.

Similar to Section 4.3, we compare the proposed method with the model in Jeong et al. (2017) that ignores heterogeneity among subjects; mgcv is not considered because Jeong et al. (2017) outperforms it (see Section 4.4). The corresponding simple model is given by

$$L_{ij} = \alpha_0(A_{ij}) + \alpha_K(A_{ij})K_{ij} + \beta_M M_{ij} + \beta_S S_{ij} + \beta_H H_{ij} + b_i + \epsilon_{ij}.$$

As shown in the next section, the simple model fails to account for heterogeneity among samples and can be obtained by pooling the results of the proposed target model.

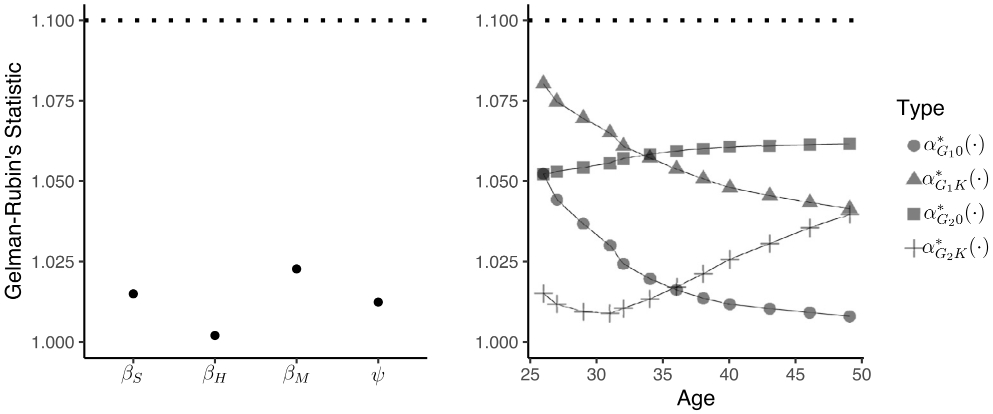

Fig. 8. The $R^{1/2}$ statistics for the fixed-dimensional parameters (left) and the fixed points of the varying coefficients (right).

Table 4
Characteristics of the clustered groups.

Covariate	Overall	Group 1	Group 2
Female	35.1%	33.3%	89.1%
Married	57.7%	55.9%	71.2%
White-collar workers	38.7%	41.1%	20.8%
Civil servants	25.8%	29.0%	4.0%
University	57.2%	58.4%	42.0%

5.2. Analysis and results

We ran the proposed PCG sampler with three over-dispersed initial values. Fig. 8 shows the convergence characteristics of the sampler by using the $R^{1/2}$ diagnostic for the fixed-dimensional parameters and the fixed points of the varying coefficients (Gelman and Rubin, 1992). Because all $R^{1/2}$ statistics are below 1.1, we combine the second halves of three chains each with 150,000 iterations through a label switching algorithm. Then, after thinning every 50th sample, our posterior inference is based on 4,500 mixed samples.

According to our posterior analysis, there are two main groups and a few minor groups. The group membership is determined by the posterior modes of the cluster labels. The interpretation of the analysis focuses on the two main groups. The first largest cluster, Group 1, accounts for 89.4% of subjects, the second largest, Group 2, accounts for 7.2%, and the remaining clusters account for 3.4%. Some characteristics of the two major groups are summarized in Table 4, demonstrating that these groups are made up of heterogeneous subjects. Specifically, Group 1 has a much lower proportion of females who tend to be more responsible for parenting than Group 2. In addition, Group 1 has a higher proportion of white-collar workers, civil servants, and university graduates with high job security than Group 2. Such difference in characteristics results in the different posterior estimates of varying-coefficient functions, as shown in Fig. 9.

Fig. 9 shows the posterior summaries of the varying-coefficient functions resulting from functional clustering. The first row of Fig. 9 corresponds to the group-level varying-intercept functions. The intercept function of Group 1 is significantly positive and keeps increasing up to early 50s, implying that a posterior probability of being employed becomes higher as one tends to be old while holding all covariates constant. In contrast, Group 2 has a slightly positive but constant intercept function in all ages. The second row of Fig. 9 shows the varying-coefficient function for having children below age 16 in the household. The 95% pointwise posterior intervals for Group 1 includes 0, which implies that having children below age 16 in the household does not significantly affect the probability of employment. Unlike Group 1, the existence of young children in Group 2's household significantly decreases the probability of employment until his/her mid 40s. The probability is further decreased when the employee's age tends to be younger. This is due in part to the fact that Group 2 has a higher proportion of females than Group 1, and females were more responsible for child care in the late twentieth century.

Table 5 shows the posterior summaries of fixed-dimensional parameters, where $\text{Var}(b_i) = \psi$ represents the variance of a random effect, and β_H, β_S, and β_M represent the coefficients of fixed effects, H_{ij}, S_{ij}, and M_{ij}, respectively. Based on the fact that the 95% posterior intervals include zero, we decide that the fixed effects have no significant influence on the employment status when the varying effects of having young children in the household for heterogeneous subpopulations are accounted for in the model. When the heterogeneous subpopulation assumption is ignored, the results obtained by Jeong et al. (2017) appear to be similar, but with narrower 95% credible intervals. This is due to the fact that Jeong et al. (2017) does not fully account for the variability of heterogeneous subpopulations.

Our proposed method identifies two major subpopulations with different characteristics, as shown in Table 4, and these subpopulations show different age-varying effects of having young children in a household on working status, as illus-

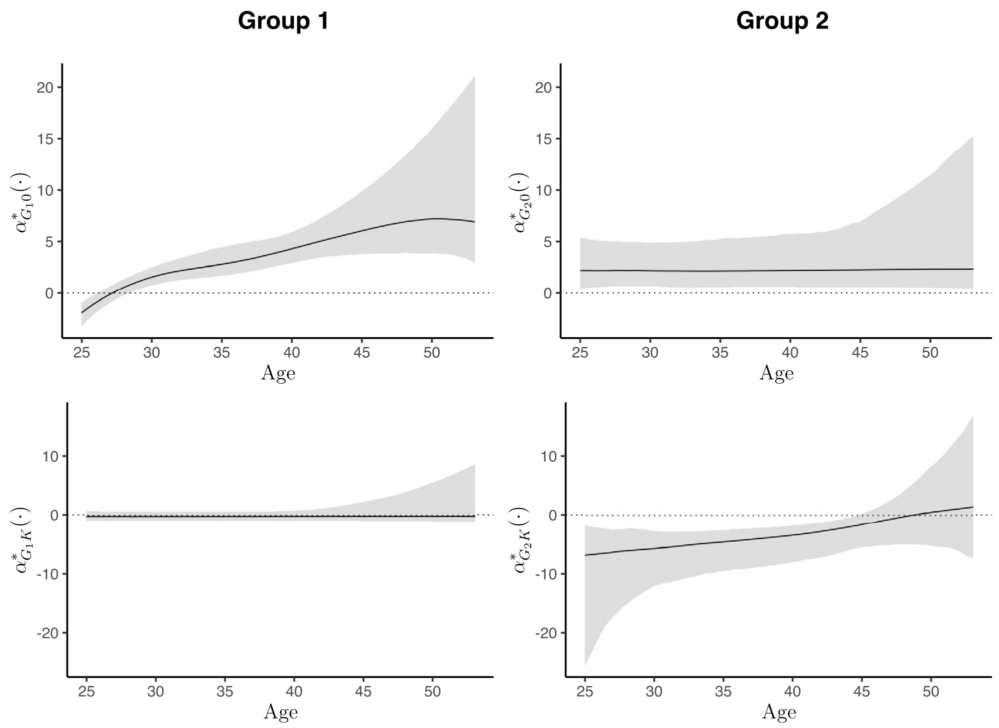

Fig. 9. Posterior summaries of the varying coefficients for two major groups. Solid lines represent the pointwise posterior medians of each varying coefficient function and gray regions correspond to pointwise 95% posterior intervals.

Table 5
Posterior summaries of the fixed-dimensional parameters.

	Proposed method			Jeong et al. (2017)		
Parameter	Mean	Median	95% interval	Mean	Median	95% interval
ψ	3.539	3.357	$(1.573, 6.334)$	3.949	3.893	$(3.031, 5.106)$
β_S	0.016	0.016	$(-0.065, 0.097)$	0.006	0.006	$(-0.051, 0.063)$
β_H	-0.002	-0.002	$(-0.023, 0.020)$	-0.010	-0.010	$(-0.026, 0.006)$
β_M	0.292	0.284	$(-0.178, 0.793)$	0.299	0.297	$(-0.041, 0.647)$

trated in Fig. 9. When the heterogeneous subpopulation assumption is ignored, however, the single population model proposed by Jeong et al. (2017) estimates the varying-coefficient functions applied to the entire population, as shown in the first column of Fig. 10. In the presence of heterogeneous subpopulations, such an approach would fail to separate subpopulations with different characteristics, leading to erroneous conclusions. This is confirmed by producing the pooled varying-coefficient functions estimated by the proposed method, as shown in the second column of Fig. 10. These findings demonstrate the proposed model's validity and utility in accounting for a heterogeneous population.

6. Discussion

In this paper, we propose a novel model-based functional clustering method for analyzing a heterogeneous series of binary data. Our proposed method models the varying effects of covariates on a series of binary responses as a function of an effect modifier, while accounting for heterogeneity among subjects using functional clustering and random effects. The proposed model estimates population-level fixed effects, cluster-level varying effects, and subject-level random effects. Even when the number of clusters is unknown, our proposed method accurately estimates cluster-specific varying coefficients with appropriate smoothness using a free-knot spline prior. We use the DP prior for functional clustering, which avoids specifying the exact number of clusters in advance. To perform posterior inference, we carefully develop a PCG sampler by specifying appropriate prior distributions and marginalizing some model components.

We suggest that there are several directions for future research, such as extending the clustering methodology to other generalized semi-parametric models using partitioning priors. Furthermore, a Gaussian process prior may be used instead of a free-knot spline for the functional clustering of varying coefficients because it may easily achieve the right scale for the base measure of the DP prior by employing a suitable covariance kernel.

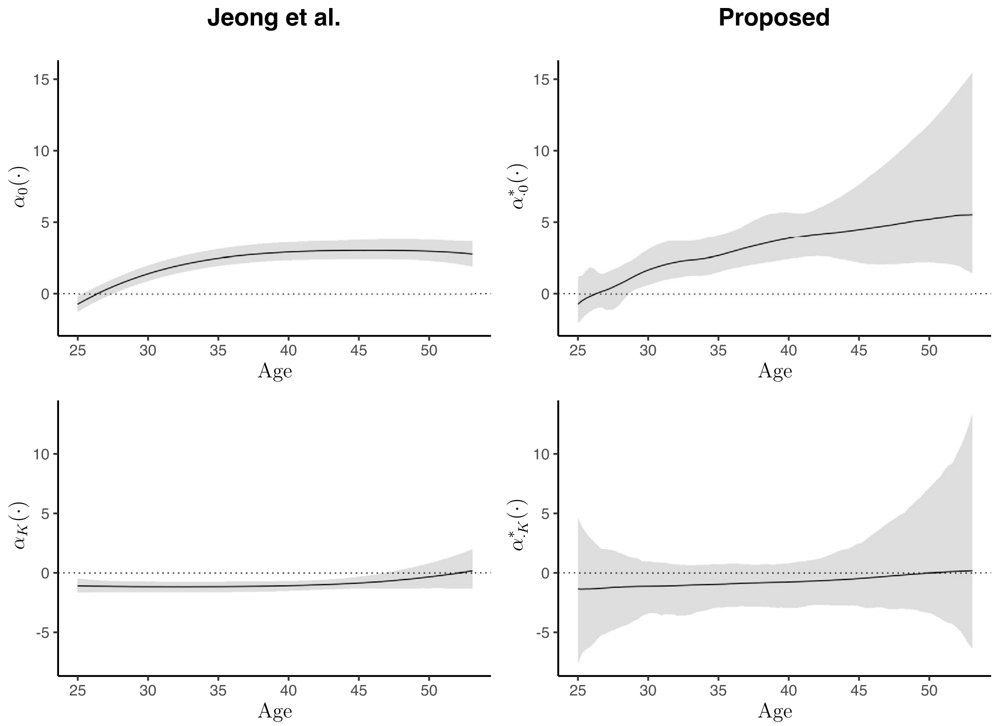

Fig. 10. The first column shows the posterior summaries of the varying coefficients obtained by Jeong et al. (2017), and the second column is with respect to the proposed method pooled by the weights corresponding to cluster assignments, with posterior medians (solid lines) and pointwise 95% posterior intervals (gray areas).

Acknowledgements

S. Jeong's research was supported by the Yonsei University Research Fund of 2022-22-0097. T. Park's research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03033536) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2C1A01005949).

Appendix A. Details of Algorithm 1

In this section, we describe the details of Algorithm 1. Let $\mathcal{C}_d$ denote a set of clusters containing at least one subject and let $\mathcal{C}_d^c$ denote a set of clusters containing no subject, in the dth sampling iteration. For dth iteration:

Step 1: Draw $\boldsymbol{\gamma}^*_{kl}$ from $p(\boldsymbol{\gamma}^*_{kl}|\boldsymbol{\gamma}^*_{-kl}, \boldsymbol{\beta}, \mathbf{C}, \mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\Psi}, \mathbf{L}, \mathbf{Y})$ that is a Bernoulli with success probability

$$\frac{f(\boldsymbol{\gamma}^*_{kl}=1, \boldsymbol{\gamma}^*_{-kl})}{f(\boldsymbol{\gamma}^*_{kl}=1, \boldsymbol{\gamma}^*_{-kl}) + f(\boldsymbol{\gamma}^*_{kl}=0, \boldsymbol{\gamma}^*_{-kl})},\ k \in \mathcal{C}_d,\ l = 1, \ldots, p,$$

where $\boldsymbol{\gamma}^*_{-kl}$ denotes all latent indicator variables except $\boldsymbol{\gamma}^*_{kl}$ in $\boldsymbol{\gamma}^*_k$,

$$\begin{aligned} f(\boldsymbol{\gamma}^*_{kl}, \boldsymbol{\gamma}^*_{-kl}) &= B\left(|\boldsymbol{\gamma}^*_{kl}| + a, M_l + 1 - |\boldsymbol{\gamma}^*_{kl}| + b\right) \times \det\left(\tau_k \mathbf{R}^{-1}_{k(\boldsymbol{\gamma}^*_k)} \Xi_{k(\boldsymbol{\gamma}^*_k, \mathbf{C}, \boldsymbol{\Psi})} + \mathbf{I}_{|\boldsymbol{\gamma}^*_k|}\right)^{-1/2} \\ &\quad \times \exp\left\{\frac{1}{2} \xi^\top_{k(\boldsymbol{\gamma}^*_k, \mathbf{C}, \boldsymbol{\Psi}, \mathbf{L}, \boldsymbol{\beta})} \left(\Xi_{k(\boldsymbol{\gamma}^*_k, \mathbf{C}, \boldsymbol{\Psi})} + \tau_k^{-1} \mathbf{R}_{k(\boldsymbol{\gamma}^*_k)}\right)^{-1} \xi_{k(\boldsymbol{\gamma}^*_k, \mathbf{C}, \boldsymbol{\Psi}, \mathbf{L}, \boldsymbol{\beta})}\right\}, \\ \Xi_{k(\boldsymbol{\gamma}^*_k, \mathbf{C}, \boldsymbol{\Psi})} &= \sum_{i:C_i=k} \mathbf{W}^{\star\top}_{i(\boldsymbol{\gamma}^*_k)} \mathbf{T}_i^{-1} \mathbf{W}^{\star}_{i(\boldsymbol{\gamma}^*_k)}, \\ \xi_{k(\boldsymbol{\gamma}^*_k, \mathbf{C}, \boldsymbol{\Psi}, \mathbf{L}, \boldsymbol{\beta})} &= \sum_{i:C_i=k} \mathbf{W}^{\star\top}_{i(\boldsymbol{\gamma}^*_k)} \mathbf{T}_i^{-1} (\mathbf{L}_i - \mathbf{X}_i \boldsymbol{\beta}), \\ \mathbf{T}_i &= \mathbf{I}_{n_i} + \mathbf{Z}_i \boldsymbol{\Psi} \mathbf{Z}_i^\top, \end{aligned}$$

and $\mathbf{I}_{|\boldsymbol{\gamma}_k^*|}$ and $\mathbf{I}_{n_i}$ are identity matrices whose sizes of each dimension are $|\boldsymbol{\gamma}_k^*|$ and n_i respectively. In the case of $k \in \mathcal{C}_d^c$, $\boldsymbol{\gamma}_{kl}^*$ is drawn from its prior distribution because $\Xi_{k(\boldsymbol{\gamma}_k^*,\mathbf{C},\boldsymbol{\Psi})}$ and $\xi_{k(\boldsymbol{\gamma}_k^*,\mathbf{C},\boldsymbol{\Psi},\mathbf{L},\boldsymbol{\beta})}$ do not exist.

Step 2: Draw V_k from $p(V_k|\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})$ that is a beta, i.e.,

$$V_k|(\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y}) \sim \text{Beta}\left(1+m_k, \nu + \sum_{h=k+1}^{K} m_h\right), \quad k=1,\ldots,K-1,$$

where $m_k = \sum_{i=1}^N I(C_i = k)$.

Step 3: Draw $\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}_k^*)}$ from $p(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}_k^*)}|\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})$ that is a multivariate normal distribution, i.e.,

$$\begin{aligned}&\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}_k^*)}|(\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})\\ &\quad\sim \mathrm{N}_{|\boldsymbol{\gamma}_k^*|}\left(\left(\Xi_{k(\boldsymbol{\gamma}_k^*,\mathbf{C},\boldsymbol{\Psi})} + \tau_k^{-1}\mathbf{R}_{k(\boldsymbol{\gamma}_k^*)}\right)^{-1}\xi_{k(\boldsymbol{\gamma}_k^*,\mathbf{C},\boldsymbol{\Psi},\mathbf{L},\boldsymbol{\beta})}, \left(\Xi_{k(\boldsymbol{\gamma}_k^*,\mathbf{C},\boldsymbol{\Psi})} + \tau_k^{-1}\mathbf{R}_{k(\boldsymbol{\gamma}_k^*)}\right)^{-1}\right), \ k \in \mathcal{C}_d,\end{aligned}$$

and

$$\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}_k^*)}|(\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y}) \sim \mathrm{N}_{|\boldsymbol{\gamma}_k^*|}\left(\mathbf{0}, \tau_k\mathbf{R}_{k(\boldsymbol{\gamma}_k^*)}^{-1}\right), \quad k \in \mathcal{C}_d^c.$$

Step 4: Draw $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})$ that is multivariate normal, i.e.,

$$\boldsymbol{\beta}|(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y}) \sim \mathrm{N}_q\left(\boldsymbol{\Delta}^{-1}\sum_{i=1}^N \mathbf{X}_i^\top\mathbf{T}_i^{-1}\left(\mathbf{L}_i - \mathbf{W}^{\star}_{i(\boldsymbol{\gamma}^*_{C_i})}\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_{C_i})}\right), \boldsymbol{\Delta}^{-1}\right),$$

where $\boldsymbol{\Delta} = \mathbf{P}^{-1} + \sum_{i=1}^N \mathbf{X}_i^\top\mathbf{T}_i^{-1}\mathbf{X}_i$.

Step 5: Draw $\mathbf{b}_i$ from $p(\mathbf{b}_i|\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})$ that is multivariate normal, i.e.,

$$\mathbf{b}_i|(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L}_i,\mathbf{Y}) \sim \mathrm{N}_r\left(\mathbf{U}_{i(\boldsymbol{\Psi},\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_{C_i})},\boldsymbol{\gamma}^*_{C_i},\mathbf{L}_i,\boldsymbol{\beta})}, \mathbf{A}_{(\boldsymbol{\Psi})}\right), \ i=1,\ldots N,$$

where

$$\begin{aligned}\mathbf{U}_{i(\boldsymbol{\Psi},\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_{C_i})},\boldsymbol{\gamma}^*_{C_i},\mathbf{L}_i,\boldsymbol{\beta})} &= \boldsymbol{\Psi}\mathbf{Z}_i^\top\mathbf{T}_i^{-1}(\mathbf{L}_i - \mathbf{W}^{\star}_{i(\boldsymbol{\gamma}^*_{C_i})}\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_{C_i})} - \mathbf{X}_i\boldsymbol{\beta}),\\ \mathbf{A}_{(\boldsymbol{\Psi})} &= \boldsymbol{\Psi} - \boldsymbol{\Psi}\mathbf{Z}_i^\top\mathbf{T}_i^{-1}\mathbf{Z}_i\boldsymbol{\Psi}.\end{aligned}$$

Step 6: Draw τ_k from $p(\tau_k|\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{C},\mathbf{V},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})$ that is an inverse gamma, i.e.,

$$\tau_k|(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{C},\mathbf{V},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y}) \sim \mathrm{IG}\left(\frac{1+|\boldsymbol{\gamma}_k^*|}{2}, \frac{N + \boldsymbol{\phi}^{\star\top}_{(\boldsymbol{\gamma}_k^*)}\mathbf{R}_{k(\boldsymbol{\gamma}_k^*)}\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}_k^*)}}{2}\right), \ k=1,\ldots,K.$$

Step 7: Draw $\boldsymbol{\Psi}$ from $p(\boldsymbol{\Psi}|\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\mathbf{L},\mathbf{Y})$ that is an inverse Wishart,

$$\boldsymbol{\Psi} \sim \mathrm{IW}\left(u+N, \mathbf{D} + \sum_{i=1}^N \mathbf{b}_i\mathbf{b}_i^\top\right).$$

Step 8: Draw L_{ij} from $p(L_{ij}|\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{Y})$ that is truncated normal, i.e.,

$$L_{ij}|(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{C},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{Y}) \sim \begin{cases}\mathrm{TN}_{(-\infty,0]}(\boldsymbol{\mu}^{(j)}_{C_i}, 1) & \text{if } Y_{ij}=0\\ \mathrm{TN}_{(0,\infty)}(\boldsymbol{\mu}^{(j)}_{C_i}, 1) & \text{if } Y_{ij}=1\end{cases}, \ i=1,\ldots,N, \ j=1,\ldots,n_i,$$

where $\boldsymbol{\mu}^{(j)}_{C_i}$ denotes the jth element of $\boldsymbol{\mu}_{C_i} = \mathbf{W}^{\star}_{i(\boldsymbol{\gamma}^*_{C_i})}\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*_{C_i})} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$.

Step 9: Draw C_i from $p(C_i|\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y})$ that has a discrete distribution with probabilities

$$P(C_i = k|(\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}^*)},\boldsymbol{\gamma}^*,\boldsymbol{\beta},\mathbf{b},\mathbf{V},\boldsymbol{\tau},\boldsymbol{\Psi},\mathbf{L},\mathbf{Y}) \propto \frac{\pi_k(\mathbf{V})\mathrm{N}_{n_i}\left(\mathbf{L}_i;\boldsymbol{\mu}_k, \mathbf{I}_{n_i}\right)}{\sum_{k=1}^K \pi_k(\mathbf{V})\mathrm{N}_{n_i}\left(\mathbf{L}_i;\boldsymbol{\mu}_k, \mathbf{I}_{n_i}\right)}, \ k=1,\ldots,K,$$

where $\boldsymbol{\mu}_k = \mathbf{W}^{\star}_{i(\boldsymbol{\gamma}_k^*)}\boldsymbol{\phi}^{\star}_{(\boldsymbol{\gamma}_k^*)} + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$.

Appendix B. R package fvcc

We provide an R package called fvcc for the proposed model. The package can be installed with the devtools package available in CRAN as follows.

```
devtools::install_github("jwsohn612/fvcc")
library(fvcc)
help(fvcc)
```

The main function fvcc contains a code script for reproducing the simulation results in Section 4.

References

Aßmann, C., Boysen-Hogrefe, J., 2011. A Bayesian approach to model-based clustering for binary panel probit models. Comput. Stat. Data Anal. 55, 261–279. https://doi.org/10.1016/j.csda.2010.04.016.

Berrettini, M., Galimberti, G., Ranciati, S., 2022. Semiparametric finite mixture of regression models with Bayesian P-splines. Adv. Data Anal. Classif., 1–31.

Canale, A., Dunson, D.B., 2011. Bayesian kernel mixtures for counts. J. Am. Stat. Assoc. 106, 1528–1539.

Chib, S., Greenberg, E., 2010. Additive cubic spline regression with Dirichlet process mixture errors. Econom. J. 156, 322–336.

Chipman, H.A., Kolaczyk, E.D., McCulloch, R.E., 1997. Adaptive Bayesian wavelet shrinkage. J. Am. Stat. Assoc. 92, 1413–1421.

Coffey, N., Hinde, J., Holian, E., 2014. Clustering longitudinal profiles using P-splines and mixed effects models applied to time-course gene expression data. Comput. Stat. Data Anal. 71, 14–29.

DiMatteo, I., Genovese, C.R., Kass, R.E., 2001. Bayesian curve-fitting with free-knot splines. Biometrika 88, 1055–1071. https://doi.org/10.1093/biomet/88.4.1055.

Fan, J., Zhang, W., 2008. Statistical methods with varying coefficient models. Stat. Interface 1, 179. https://doi.org/10.4310/sii.2008.v1.n1.a15.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Stat., 209–230.

Gelfand, A.E., Kottas, A., MacEachern, S.N., 2005. Bayesian nonparametric spatial modeling with Dirichlet process mixing. J. Am. Stat. Assoc. 100, 1021–1035.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–472. https://doi.org/10.1214/ss/1177011136.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2015. Bayesian Data Analysis. CRC Press etc.

Guerra, M.W., Shults, J., Amsterdam, J., Ten-Have, T., 2012. The analysis of binary longitudinal data with time-dependent covariates. Stat. Med. 31, 931–948. https://doi.org/10.1002/sim.4465.

Hannah, L.A., Blei, D.M., Powell, W.B., 2011. Dirichlet process mixtures of generalized linear models. J. Mach. Learn. Res. 12.

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. J. R. Stat. Soc., Ser. B, Methodol., 757–796. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x.

Heard, N.A., Holmes, C.C., Stephens, D.A., 2006. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. J. Am. Stat. Assoc. 101, 18–29.

Hoover, D.R., Rice, J.A., Wu, C.O., Yang, L.-P., 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85, 809–822. https://doi.org/10.1093/biomet/85.4.809.

Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. 96, 161–173. https://doi.org/10.1198/016214501750332758.

James, G.M., Sugar, C.A., 2003. Clustering for sparsely sampled functional data. J. Am. Stat. Assoc. 98, 397–408.

Jeong, S., Park, T., 2016. Bayesian semiparametric inference on functional relationships in linear mixed models. Bayesian Anal. 11, 1137–1163. https://doi.org/10.1214/15-BA987.

Jeong, S., Park, M., Park, T., 2017. Analysis of binary longitudinal data with time-varying effects. Comput. Stat. Data Anal. 112, 145–153. https://doi.org/10.1016/j.csda.2017.03.007.

Jeong, S., Park, T., van Dyk, D.A., 2022. Bayesian model selection in additive partial linear models via locally adaptive splines. J. Comput. Graph. Stat. 31, 324–336.

Kang, G., Jeong, S., 2023. Model selection-based estimation for generalized additive models using mixtures of g-priors: towards systematization. arXiv preprint. arXiv:2301.10468.

Kim, Y., Choi, Y.-K., Emery, S., 2013. Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. Am. Stat. 67, 171–182. https://doi.org/10.1080/00031305.2013.817357.

Kohn, R., Smith, M., Chan, D., 2001. Nonparametric regression using linear combinations of basis functions. Stat. Comput. 11, 313–322. https://doi.org/10.1023/A:1011916902934.

Kuss, O., Gromann, C., Diepgen, T.L., 2006. Model-based clustering of binary longitudinal atopic dermatitis disease histories by latent class mixture models. Biom. J. 48, 105–116.

Kyung, M., 2015. Dirichlet process mixtures of linear mixed regressions. Commun. Stat. Appl. Methods 22, 625–637. https://doi.org/10.5351/CSAM.2015.22.6.625.

Lang, S., Brezger, A., 2004. Bayesian P-splines. J. Comput. Graph. Stat. 13, 183–212. https://doi.org/10.1198/1061860043010.

Lau, J.W., Green, P.J., 2007. Bayesian model-based clustering procedures. J. Comput. Graph. Stat. 16, 526–558.

Lenk, P.J., DeSarbo, W.S., 2000. Bayesian inference for finite mixtures of generalized linear models with random effects. Psychometrika 65, 93–119. https://doi.org/10.1007/BF02294188.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2008. Mixtures of g priors for Bayesian variable selection. J. Am. Stat. Assoc. 103, 410–423. https://doi.org/10.1198/016214507000001337.

Lu, Y., Zhang, R., 2009. Smoothing spline estimation of generalised varying-coefficient mixed model. J. Nonparametr. Stat. 21, 815–825. https://doi.org/10.1080/10485250903151078.

MacEachern, S.N., Kottas, A., Gelfand, A.E., 2001. Spatial nonparametric Bayesian models. In: Proceedings of the 2001 Joint Statistical Meetings, vol. 3, p. 14.

Margaritella, N., Inácio, V., King, R., 2021. Parameter clustering in Bayesian functional principal component analysis of neuroscientific data. Stat. Med. 40, 167–184.

Meng, X.-L., van Dyk, D., 1998. Fast em-type implementations for mixed effects models. J. R. Stat. Soc., Ser. B, Stat. Methodol. 60, 559–578. https://doi.org/10.1111/1467-9868.00140.

Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83, 67–79.

Neal, R.M., 1998. Regression and classification using gaussian process priors. Bayesian Stat. 6, 475.

Park, H., Park, T., Lee, Y.-S., 2019. Partially collapsed Gibbs sampling for latent Dirichlet allocation. Expert Syst. Appl. 131, 208–218. https://doi.org/10.1016/j.eswa.2019.04.028.

Park, T., Jeong, S., 2018. Analysis of Poisson varying-coefficient models with autoregression. Statistics 52, 34–49. https://doi.org/10.1080/02331888.2017.1353514.

Park, T., Min, S., 2016. Partially collapsed Gibbs sampling for linear mixed-effects models. Commun. Stat., Simul. Comput. 45, 165–180.

Park, T., van Dyk, D.A., 2009. Partially collapsed Gibbs samplers: illustrations and applications. J. Comput. Graph. Stat. 18, 283–305. https://doi.org/10.1198/jcgs.2009.08108.

Petrone, S., Guindani, M., Gelfand, A.E., 2009. Hybrid Dirichlet mixture models for functional data. J. R. Stat. Soc., Ser. B, Stat. Methodol. 71, 755–782.

Pitman, J., Yor, M., 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab., 855–900.

Ray, S., Mallick, B., 2006. Functional clustering by Bayesian wavelet methods. J. R. Stat. Soc., Ser. B, Stat. Methodol. 68, 305–332.

Riphahn, R.T., Wambach, A., Million, A., 2003. Incentive effects in the demand for health care: a bivariate panel count data estimation. J. Appl. Econom. 18, 387–405. https://doi.org/10.1002/jae.680.

Rodriguez, A., Dunson, D.B., 2014. Functional clustering in nested designs: modeling variability in reproductive epidemiology studies.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression, vol. 12. Cambridge University Press.

Scott, J.G., Berger, J.O., 2010. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann. Stat. 38, 2587–2619. https://doi.org/10.1214/10-AOS792.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Stat. Sin., 639–650.

Shen, W., Ghosal, S., 2015. Adaptive Bayesian procedures using random series priors. Scand. J. Stat. 42, 1194–1213.

Shi, J.Q., Wang, B., 2008. Curve prediction and clustering with mixtures of gaussian process functional regression models. Stat. Comput. 18, 267–283.

Smith, M., Kohn, R., 1996. Nonparametric regression using Bayesian variable selection. J. Econ. 75, 317–343. https://doi.org/10.1016/0304-4076(95)01763-1.

Stiratelli, R., Laird, N., Ware, J.H., 1984. Random-effects models for serial observations with binary response. Biometrics, 961–971. https://doi.org/10.2307/2531147.

Suarez, A.J., Ghosal, S., 2016. Bayesian clustering of functional data using local features.

Sun, Y., Wu, H., 2005. Semiparametric time-varying coefficients regression model for longitudinal data. Scand. J. Stat. 32, 21–47. https://doi.org/10.1111/j.1467-9469.2005.00413.x.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. J. Am. Stat. Assoc. 101, 1566–1581. https://doi.org/10.1198/016214506000000302.

van Dyk, D.A., 2000. Fitting mixed-effects models using efficient em-type algorithms. J. Comput. Graph. Stat. 9, 78–98. https://doi.org/10.1080/10618600.2000.10474867.

van Dyk, D.A., Park, T., 2008. Partially collapsed Gibbs samplers: theory and methods. J. Am. Stat. Assoc. 103, 790–796. https://doi.org/10.1198/016214508000000409.

Varin, C., Czado, C., 2009. A mixed autoregressive probit model for ordinal longitudinal data. Biostatistics 11, 127–138. https://doi.org/10.1093/biostatistics/kxp042.

Vats, D., Flegal, J.M., Jones, G.L., 2019. Multivariate output analysis for Markov chain Monte Carlo. Biometrika 106, 321–337.

Vines, S., Gilks, W., Wild, P., 1996. Fitting Bayesian multiple random effects models. Stat. Comput. 6, 337–346. https://doi.org/10.1007/BF00143554.

Wallach, H., Jensen, S., Dicker, L., Heller, K., 2010. An alternative prior process for nonparametric Bayesian clustering. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.

Wood, S., 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall/CRC.

Wu, C.O., Chiang, C.-T., Hoover, D.R., 1998. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. J. Am. Stat. Assoc. 93, 1388–1402. https://doi.org/10.1080/01621459.1998.10473800.

Yerebakan, H.Z., Rajwa, B., Dundar, M., 2014. The Infinite Mixture of Infinite Gaussian Mixtures. Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc.

Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti 6, pp. 233–243.

Zhu, X., Tang, X., Qu, A., 2021. Longitudinal clustering for heterogeneous binary data. Stat. Sin. 31, 603–624.