Published in final edited form as:
Stat Med. 2019 April 30; 38(9): 1634–1650. doi:10.1002/sim.8051.

Bayesian adaptive group lasso with semiparametric hidden Markov models

Kai Kang[1], **Xinyuan Song**[1,2], **X. Joan Hu**[3], and **Hongtu Zhu**[4,5]

[1]Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong

[2]Shenzhen Research Institute, The Chinese University of Hong Kong, Shatin, Hong Kong

[3]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada

[4]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[5]Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

Abstract

This paper presents a Bayesian adaptive group least absolute shrinkage and selection operator method to conduct simultaneous model selection and estimation under semiparametric hidden Markov models. We specify the conditional regression model and the transition probability model in the hidden Markov model into additive nonparametric functions of covariates. A basis expansion is adopted to approximate the nonparametric functions. We introduce multivariate conditional Laplace priors to impose adaptive penalties on regression coefficients and different groups of basis expansions under the Bayesian framework. An efficient Markov chain Monte Carlo algorithm is then proposed to identify the nonexistent, constant, linear, and nonlinear forms of covariate effects in both conditional and transition models. The empirical performance of the proposed methodology is evaluated via simulation studies. We apply the proposed model to analyze a real data set that was collected from the Alzheimer's Disease Neuroimaging Initiative study. The analysis identifies important risk factors on cognitive decline and the transition from cognitive normal to Alzheimer's disease.

Keywords

linear basis expansion; Markov chain Monte Carlo; simultaneous model selection and estimation

1 | INTRODUCTION

Hidden Markov models (HMMs) have been widely used in the medical, behavioral, social, environmental, and psychological sciences where longitudinal data are frequently collected.[1-6] Basically, HMMs are designed to have two parts: a transition model to investigate the

Correspondence Xinyuan Song, Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong., xysong@sta.cuhk.edu.hk.

effects of covariates on the dynamic transition process of hidden states and a conditional regression model to examine state-specific covariate effects on the response of interest. In these two parts, the effect of a covariate on the response or on the transition process can be nonexistent, constant, linear, or nonlinear. Identifying the specific forms of such covariate effects is useful not only in achieving a parsimonious model but also in obtaining enhanced parameter estimation and attractive interpretations.

Conventional studies on HMMs have focused on a parametric framework, wherein the forms of covariate effects on responses and/or on transition probabilities are prespecified. However, one fundamental issue overlooked by these parametric HMMs is that the complex relationships among variables are seldom known a priori, and the parametric form is thus too restrictive to correctly reflect the reality. Several nonparametric approaches have been investigated recently to relax the parametric assumption of HMMs. Yau et al[7] developed a Bayesian nonparametric HMM, where the sampling distribution of the observations at each state was assumed unknown and modeled via a mixture of Dirichlet processes. Although their method did not rely on the distributional assumption of the observed process, it cannot reveal the functional effects of potential explanatory variables on the outcome of interest. Song et al[8] considered Bayesian P-splines for describing the nonparametric relation among latent variables in HMMs, but they did not consider the model selection problem.

Model selection is an important issue beyond estimation in the application of HMMs. Classical model selection methods are mainly developed on the basis of a pairwise comparison through common model selection criteria, such as the Akaike information criterion and the Bayesian information criterion. However, such pairwise-based procedure usually becomes increasingly computationally demanding when the search dimension is high. An appealing alternative is to adopt least absolute shrinkage and selection operator (lasso)–type variable selection techniques. Choi et al[9] applied lasso to correlated HMMs to detect the important parameters in transition models. Städler and Mukherjee[10] introduced L_1 penalization to obtain a sparse HMM with state-specific graphical models. However, the preceding studies consider only parametric HMMs. Recently, some variants of lasso, such as group lasso, adaptive lasso, and adaptive group lasso, have been developed to manage group variables and address the issue of lasso and group lasso possibly suffering from appreciable bias. Owing to the computational efficiency and stability of the Bayesian approach, the Bayesian analogs of lasso and its variants have been proposed.[11,12] However, the available Bayesian lasso-type methods are all developed in the context of cross-sectional models without between-state transitions, thereby making them inapplicable to the proposed semiparametric HMMs.

In this paper, we propose a Bayesian adaptive group lasso (BaGlasso) procedure to conduct simultaneous model selection and estimation for semiparametric HMMs. With the use of basis expansion and appropriate penalties, the non-parametric relationships that subsume nonexistent, constant, linear, and nonlinear relationships between covariates and the response can be automatically identified. The proposed procedure has the following appealing features: first, the group effects and additional correlation within the basis expansion are well addressed by the group lasso, thus ensuring estimation accuracy. Second, adaptive penalties imposed on different groups of coefficients enable us to achieve an efficient variable selection. Finally, the proposed procedure avoids tedious pairwise comparisons among competing models with different combinations of covariates in the conditional and transition models. This entirely data-driven feature not only relaxes the dependence on experts' knowledge in empirical studies but also reduces the computational burden. To the best of our knowledge, this study is the first to introduce Bayesian lasso-type procedure into semiparametric HMMs.

The proposed method is motivated by a real study conducted by the Alzheimer's Disease Neuroimaging Initiative (ADNI). A set of biomarkers, namely, gender, age, educational levels, marital status, hippocampal volume, and apolipoprotein E (APOE)-ϵ4, is collected across several time points in this data set. The purpose of this study is to detect the potential risk factors of Alzheimer's disease (AD) from two perspectives. First, considering that the pathology of AD usually evolves from cognitive normal (CN) to mild cognitive impairment (MCI) to dementia, characterizing the disease pathology, identifying hidden states that correspond to the diagnosed stages of cognitive decline, and examining the potential risk factors of the neurodegenerative transition are of scientific interest and practical value. Given that the effects of biomarkers on the pathology from one state to another may vary across nonexistent, constant, linear, and nonlinear ones, allowing their forms to be unspecified and introducing penalties to penalize unimportant effects can reveal the patterns of the effects to the greatest extent. Previous studies[13] pointed out that the relationships between some biomarkers and cognitive decline are variant across different states. Therefore, identifying the significant state-specific risk factors of cognitive decline and investigating the subtle forms of their effects are of great interest. However, existing relevant research either restricts the examination of the above relationships under a parametric framework or emphasizes only estimation. The proposed methodology enables us to perfectly accommodate all the aforementioned features and provide new insights into the prevention of AD.

The rest of this paper is organized as follows. Section 2 introduces the semiparametric HMM and discusses the associated identifiability issue. Section 3 illustrates the statistical inference of the proposed model. Specifically, BaGlasso for simultaneous variable selection and parameter estimation as well as the deviance information criterion (DIC) for the determination of the number of hidden states are presented. Section 4 investigates the empirical performance of the proposed method via simulation studies. Section 5 presents an application of the proposed method to the aforementioned ADNI study. Several important biomarkers are detected to have significant functional effects on patients' cognitive decline across neurodegenerative states and/or on transition probabilities. The extension of the model is discussed in Section 6.

2 | MODEL DESCRIPTION

2.1 | Semiparametric HMMs

Let y_{it} with subject $i = 1, \dots, n$ at $t = 1, \dots, T$ be the observation process. $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iT})'$, the hidden-state sequence, is commonly assumed to follow a first-order Markov chain taking values in a finite set $\{1, \dots, S\}$. Given the hidden state Z_{it}, the conditional semiparametric regression model is formulated as follows:

$$[y_{it} \mid Z_{it} = s] = \mu_s + \boldsymbol{\alpha}_s'\mathbf{c}_{it} + \sum_{j=1}^{q} f_{sj}(x_{itj}) + \delta_{it}, \quad (1)$$

where $\mathbf{c}_{it} = (c_{it1}, \&, c_{itp})'$ and $\mathbf{x}_{it} = (x_{it1}, \&, x_{itq})'$ are a $p \times 1$ vector of discrete covariates and a $q \times 1$ vector of continuous covariates, respectively; intercept μ_s, fixed effects $\boldsymbol{\alpha}_s' = (\alpha_{s1}, \ldots, \alpha_{sp})$, and unknown smoothing function $f_{sj}(\cdot)$s are all defined as state-specific to address the heterogeneity underlying the observations; δ_{it} is a random residual independent of y_{it}; and $[\delta_{it}|Z_{it} = s] \sim N[0, \psi_s]$.

In addition to the observable process, the hidden process, $\mathbf{Z}_i$, is formulated as follows: let p_{itus} denote the transition probability from state $Z_{i,t-1} = u$ at occasion $t - 1$ to state $Z_{it} = s$ at occasion t for individual i. Then, we have

$$p_{itus} = P(Z_{it} = s \mid Z_{i1}, Z_{i2}, \ldots, Z_{i,t-1} = u) = P(Z_{it} = s \mid Z_{i,t-1} = u). \quad (2)$$

Notably, model (2) is guaranteed by the assumed property of Markov chain. A common setting for the initial distribution of Z_{i1} is the multinomial distribution with probability $(\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_S)'$, such that $\boldsymbol{\pi}_s \geq 0$ and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iT})'$. Thus, the hidden-state sequence $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iT})'$ is fully specified by the initial and transition probabilities.

Considering that the hidden states usually have natural ranking information in empirical studies, we assume the hidden states $\{1, \ldots, S\}$ to be ordered and consider a continuation-ratio logit model[14] as follows: for $t = 2, \ldots, T$, $s = 1, \ldots, S - 1$, and $u = 1, \ldots, S$, we have

$$\log\left(\frac{P(Z_{it} = s \mid Z_{i,t-1} = u)}{P(Z_{it} > s \mid Z_{i,t-1} = u)}\right) = \log\left(\frac{p_{itus}}{p_{itu,s+1} + \cdots + p_{ituS}}\right) = \zeta_{us} + \widetilde{\boldsymbol{\alpha}}'\mathbf{c}_{it} + \sum_{j=1}^{q} g_j(x_{itj}), \quad (3)$$

where the left-hand side is the log odds of transition to state s rather than to a state that is higher than s given $Z_{i,t-1} = u$, ζ_{us} is a transition-specific intercept, $\mathbf{c}_{it} = (c_{it1}, \ldots, c_{itp})'$ and $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itq})'$ are the covariate vectors defined in (1), $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_p)'$ is a $p \times 1$ vector of fixed effect, and $g_j(\cdot)$s are unknown smoothing functions. Let $\vartheta_{itus} = P(Z_{it} = s|Z_{it} \geq s, Z_{i,t-1} = u)$. Then, the continuation-ratio logits in (3) can be rewritten as

$$
\begin{aligned}
&\log\left(\frac{P(Z_{it}=s \mid Z_{i,t-1}=u)}{P(Z_{it}>s \mid Z_{i,t-1}=u)}\right)\\
&=\log\left(\frac{P(Z_{it}=s, Z_{i,t-1}=u)}{P(Z_{it}\geq s, Z_{i,t-1}=u)-P(Z_{it}=s, Z_{i,t-1}=u)}\right)\\
&=\log\left(\frac{P(Z_{it}=s, Z_{i,t-1}=u)\big/P(Z_{it}\geq s, Z_{i,t-1}=u)}{1-P(Z_{it}=s, Z_{i,t-1}=u)\big/P(Z_{it}\geq s, Z_{i,t-1}=u)}\right)\\
&=\log\left(\frac{P(Z_{it}=s \mid Z_{it}\geq s, Z_{i,t-1}=u)}{1-P(Z_{it}=s \mid Z_{it}\geq s, Z_{i,t-1}=u)}\right)\\
&=\log\left(\frac{\vartheta_{itus}}{1-\vartheta_{itus}}\right).
\end{aligned}
$$

Thus, the continuation-ratio logit (3) can be rewritten as a conventional logistic regression model as follows:

$$\text{logit}(\vartheta_{itus}) = \zeta_{us} + \tilde{\boldsymbol{\alpha}}'\mathbf{c}_{it} + \sum_{j=1}^{q} g_j(x_{itj}), \quad (4)$$

where logit(ϑ_{itus}) is the log odds of $Z_{it} = s$ given $Z_{it} \geq s$ and $Z_{i,t-1} = u$. In model (3) or (4), $\tilde{\boldsymbol{\alpha}}$ and $g_j(-)$s are assumed to be independent of u and s. This proportional odds assumption is compulsory in modeling an ordinal variable because it ensures the that $P(Z_{it} < 1) < P(Z_{it} < 2) < \cdots < P(Z_{it} < S)$ for ordered states $1 < 2 < \cdots < S$.[14,15] Moreover, the proportional odds assumption avoids a tedious inference, in which every possible transition of origination and destination elicits a set of parameters, and it, in turn, greatly reduces the complexity and enhances the interpretability of the transition model.

2.2 | Nonparametric modeling

We use linear basis expansion to estimate the nonparametric functions $f_{sj}(\cdot)$ and $g_j(\cdot)$ in (1) and (3). Given that $g_j(\cdot)$ can be regarded as a special case (without a state-specific setting) of $f_{sj}(\cdot)$ we describe only the modeling of $f_{sj}(\cdot)$ in this section. Specifically, $f_{sj}(x_{itj})$ can be approximated as follows:

$$f_{sj}(x_{itj}) = \sum_{m=1}^{M_j} \beta_{sjm} h_m(x_{itj}) = \boldsymbol{\beta}'_{sj}\mathbf{h}_{itj}, \quad (5)$$

where $h_m(\cdot)$s are basis functions, such as piecewise polynomials or natural cubic splines,[16] $h_{itj} = (h_1(x_{itj}), \ldots, h_{M_j}(x_{itj}))'$, and M_j is the number of basis functions that are used to estimate the jth unknown smoothing function. For notational simplicity, M_j is set to be invariant to states. An extension to relax this assumption is straightforward.

An important issue regarding the model selection of (1) and (3) is whether a functional effect, eg, $f_{sj}(\cdot)$, truly exists or not. In this study, we utilize a norm $\|\cdot\|$ to quantify the magnitude of nonparametric function f_{sj}. Let $\mathbf{x}_{sj}$ and $\mathbf{H}_{sj}$ denote the submatrix of $\mathbf{x}_\mathrm{j} = (x_{11j}, \ldots, x_{nTj})'$ and $\mathbf{H}_j$, respectively, with the rows corresponding to $Z_{it} \neq s$ deleted, where $\mathbf{H}_j$ is formed by

$$\mathbf{H}_j = \begin{pmatrix} \mathbf{h}'(x_{11j}) \\ \vdots \\ \mathbf{h}'(x_{nTj}) \end{pmatrix} = \begin{pmatrix} h_1(x_{11j}) & \cdots & h_{M_j}(x_{11j}) \\ \vdots & \ddots & \vdots \\ h_1(x_{nTj}) & \cdots & h_{M_j}(x_{nTj}) \end{pmatrix}_{nT \times M_j} . \quad (6)$$

The norm of f_{sj}, $\|f_{sj}\|$, is defined as $\sqrt{E(f_{sj}^2(\mathbf{x}_{sj}))}$. Then, $f_{sj} = 0$ is equivalent to $\|f_{sj}\| = 0$. On the basis of (5), $\|f_{sj}$ can be approximated by $\left\|\boldsymbol{\beta}_{sj}\right\|_{\mathrm{G}_{sj}} = (\boldsymbol{\beta}'_{sj}\mathbf{G}_{sj}\boldsymbol{\beta}_{sj})^{1/2}$ with positive definite matrix $\mathbf{G}_{sj} = \mathbf{H}'_{sj}\mathbf{H}_{sj} / n_s$, where n_s is the number of subjects staying in state s. Denote $\left\|\hat{f}_{sj}\right\|$ as the estimator of $\|f_{sj}\|$. In the model selection procedure, if $\left\|\hat{f}_{sj}\right\| = 0$, then $f_{sj} = 0$. The nonparametric function $g_j(\mathrm{x}_{itj})$ can be similarly approximated by

$$g_j(x_{itj}) = \sum_{m=1}^{M_j} \tilde{\beta}_{jm} h_m(x_{itj}) = \tilde{\boldsymbol{\beta}}'_j \mathbf{h}_{itj}, \quad (7)$$

where $\tilde{\boldsymbol{\beta}}_{jm}$, $h_m(\cdot)$, h_{itj}, M_j, and $\tilde{\boldsymbol{\beta}}_j$ are defined in the same manner as those in (5). Likewise, $\|g_j\|$ can be approximated by $\left\|g_j\right\|_{\tilde{\mathrm{G}}_j} = (\tilde{\boldsymbol{\beta}}'_j\tilde{\mathbf{G}}_j\tilde{\boldsymbol{\beta}}_j)^{1/2}$, where $\tilde{\mathbf{G}}_j = \mathbf{H}'_j\mathbf{H}_j / (n \times (T-1))$.

Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{iT})'$, $\mathbf{Y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_n)'$, $\mathbf{d}_{it} = (\mathbf{c}'_{it}, \mathbf{x}'_{it})'$, $\mathbf{D}_i = (\mathbf{d}'_{i1}, \ldots, \mathbf{d}'_{iT})'$, $\mathbf{D} = (\mathbf{D}'_1, \ldots, \mathbf{D}'_n)'$, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iT})'$, $\mathbf{Z} = (\mathbf{Z}'_1, \ldots, \mathbf{Z}'_n)'$, and $\boldsymbol{\theta}$ be the vector that includes all the unknown parameters. With the linear basis expansion, the complete-data log-likelihood function is given by

$$\log p(\mathbf{Y}, \mathbf{D}, \mathbf{Z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\log p(\mathbf{y}_i \mid \mathbf{D}_i, \mathbf{Z}_i, \boldsymbol{\theta}) + \log p(\mathbf{Z}_i \mid \mathbf{D}_i, \boldsymbol{\theta})\right]$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \log p(y_{it} \mid \mathbf{d}_{it}, Z_{it} = s, \boldsymbol{\theta}) + \sum_{i=1}^{n} \sum_{t=2}^{T} \log p(Z_{it} = s \mid Z_{i,t-1} = u, \mathbf{d}_{it}, \boldsymbol{\theta}) + \sum_{i=1}^{n} \log p(Z_{i1} = s \mid \boldsymbol{\theta})$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \sum_{t=1}^{T} \left[\log(2\pi\Psi_s) + (y_{it} - \eta_{it})^2 / \Psi_s\right] + \sum_{i=1}^{n} \sum_{t=2}^{T} \log(p_{itus}) + \sum_{i=1}^{n} \log(p_{i10s}),$$

(8)

where

$$
\begin{aligned}
\eta_{it} &= \mu_s + \boldsymbol{\alpha}_s'\mathbf{c}_{it} + \sum_{j=1}^{q} \boldsymbol{\beta}_{sj}'\mathbf{h}_{itj}, \qquad p_{i10s} = \pi_s, \qquad s = 1, \ldots, S, \\
p_{itu1} &= \frac{\exp\{a_{itu1}\}}{1+\exp\{a_{itu1}\}}, \qquad p_{i1uS} = \prod_{j=1}^{S-1} \frac{1}{1+\exp\{a_{ituj}\}}, \\
p_{itus} &= \frac{\exp\{a_{itus}\}}{1+\exp\{a_{itus}\}} \prod_{j=1}^{s-1} \frac{1}{1+\exp\{a_{ituj}\}}, \qquad s = 2, \ldots, S-1,
\end{aligned} \tag{9}
$$

with $a_{itus} = \zeta_{us} + \tilde{\boldsymbol{\alpha}}'\mathbf{c}_{it} + \sum_{j=1}^{q} \tilde{\boldsymbol{\beta}}_j'\mathbf{h}_{itj}$.

2.3 | Related issues

The proposed model is not identifiable because of the following two model indeterminacies. First, the basis functions involved in basis expansion may contain constant parts. When applying such constant basis functions in every $f_{sj}(\cdot)$ and/or $g_j(\cdot)$, each unknown function is not identifiable up to a constant. To address this issue, we need to impose the following constraints on the unknown functions to enforce their integrations in the ranges of predictors to zero[17,18]:

$$
\int_{\chi_j} f_{sj}(x)dx = 0, \quad \text{for } s = 1, \ldots, S, \quad j = 1, \ldots, q, \tag{10}
$$

where χ_j is the domain of $\mathbf{x}_j$. Second, the label switching problem, which is caused by the invariance of the likelihood function to a random permutation of the state labels, arises and leads to a multimodal posterior under a symmetric prior specification. We address this issue by imposing constraint $\mu_1 < \cdots < \mu_S$ on posterior samples.

3 | BAYESIAN ANALYSIS

3.1 | Adaptive group lasso penalties

We explain the key idea of the adaptive group lasso penalties in the context of a simple linear regression model: $\mathbf{y} = \mu\mathbf{1}_\mathrm{n} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}$, where $\mathbf{y}$ is the response vector, μ is an intercept, $\mathbf{1}_n$ is an n-dimensional vector of all elements being 1, $\mathbf{X}$ is a standardized design matrix, $\boldsymbol{\delta}$ is the vector of residuals, $\boldsymbol{\delta} \sim N(\mathbf{0}, \psi\mathbf{I}_n)$, and $\mathbf{I}_n$ is an n-dimensional identity matrix. Tibshirani[19] first introduced the lasso procedure for simultaneous model selection and parameter estimation of the above linear regression. The lasso estimator of $\boldsymbol{\beta}$ can be expressed as

$$
\operatorname{argmin}_{\boldsymbol{\beta}}\left\{(\mathbf{y} - \mu\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mu\mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}) + \gamma \sum_{h=1}^{p} |\beta_h|\right\}, \tag{11}
$$

where $\gamma \geq 0$ can be regarded as an L_1-penalty that automatically shrinks unimportant covariate effects to 0. Given that the covariates in $\mathbf{X}$ are standardized to the same scale, the magnitudes of the coefficients in $\boldsymbol{\beta}$ can represent the significance of predictors. If some elements of $\boldsymbol{\beta}$ are close to 0, then the corresponding covariates are unimportant and can be removed from the model.

However, when simply applying lasso to the proposed semiparametric HMMs, at least two problems exist. First, lasso is originally designed for the selection of individual variables. Yuan and Lin[20] showed that lasso tends to select more factors than necessary in the presence of group variables. Moreover, the pairwise correlations among group variables jeopardize the model selection accuracy of the lasso estimator.[21] In this study, high correlations exist among the basis functions $h_m(x_{it})$s in the conditional and transition models because they can be viewed as different transformations of x_{it}. Consequently, the linear basis expansion involves group variables and should not be treated separately. Second, lasso applies the same tuning parameter γ to different regression coefficients, thereby introducing the same amount of shrinkage to different covariate effects. This inflexible setting may add considerable bias to the resulting estimates.[22,23]

To address the aforementioned issues, Yuan and Lin[20] proposed group lasso to perform model selection among group variables. Wang and Leng[24] further developed adaptive group lasso to assign different tuning parameters to different groups of regression coefficients. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_S')'$, $\boldsymbol{\beta}_s' = (\boldsymbol{\beta}_{s1}', \ldots, \boldsymbol{\beta}_{sq}')'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_S')'$, $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1', \ldots, \widetilde{\boldsymbol{\beta}}_q')'$, and $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}', \widetilde{\boldsymbol{\alpha}}', \boldsymbol{\beta}', \widetilde{\boldsymbol{\beta}}')'$. On the basis of the proposed model defined in (1)–(7), the adaptive group lasso estimator can be formulated as

$$\arg\min_{\boldsymbol{\theta}^*} \left\{ \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \eta_{it})'(y_{it} - \eta_{it}) - \sum_{i=1}^{n} \sum_{t=2}^{T} \log(p_{itus}) - P(\boldsymbol{\theta}^*) \right\}, \quad (12)$$

where η_η is the mean of y_{it}, p_{itus} is the transition probability defined in (2) and (9), and

$$P(\boldsymbol{\theta}^*) = \sum_{s=1}^{S} \sum_{h=1}^{p} \gamma_{\alpha sh} \mid \alpha_{sh} \mid + \sum_{h=1}^{p} \widetilde{\gamma}_{\alpha h} \mid \widetilde{\alpha}_h \mid + \sum_{s=1}^{S} \sum_{j=1}^{q} \gamma_{\boldsymbol{\beta} sj} \left\| \boldsymbol{\beta}_{sj} \right\|_{\mathrm{G}_{sj}} + \sum_{j=1}^{q} \widetilde{\gamma}_{\boldsymbol{\beta} j} \left\| \widetilde{\boldsymbol{\beta}}_j \right\|_{\widetilde{\mathrm{G}}_j},$$

(13)

in which α_{sh}, $\widetilde{\alpha}_h$, $\boldsymbol{\beta}_{sj}$, and $\widetilde{\boldsymbol{\beta}}_j$ are coefficients of fixed effects and basis functions in the conditional and transition models; γ_{ash}, $\widetilde{\gamma}_{\alpha h}$, $\gamma_{\boldsymbol{\beta} sj}$ and $\widetilde{\gamma}_{\boldsymbol{\beta} j}$ are the corresponding tuning parameters; and the norms $\|\boldsymbol{\beta}_{sj}\|_{\mathbf{G}_{sl}}$ and $\left\|\widetilde{\boldsymbol{\beta}}_j\right\|_{\widetilde{\mathrm{G}}_l}$ are defined in Section 2.2. Notably, the coefficients of discrete covariates, namely, α_{sh} and $\widetilde{\alpha}_h$, are simply assigned adaptive penalties, whereas the coefficients of unknown smooth functions $\boldsymbol{\beta}_{sj}$ and $\widetilde{\boldsymbol{\beta}}_j$, which have groupwise features, are assigned adaptive group lasso penalties. The initial probabilities p_{i10s} are excluded from (13) because they are independent of $\widetilde{\alpha}_h$ and $\widetilde{\boldsymbol{\beta}}_j$.

Yuan and Lin[20] argued that the penalty function in (13) is intermediate between the L_1-penalty used in lasso and the L_2-penalty used in ridge regression. Therefore, the adaptive group lasso not only has the same advantages of lasso in model selection but also alleviates the problem caused by the existence of high pairwise correlation among basis functions. Furthermore, with the use of different tuning parameters $\gamma_{\boldsymbol{\beta}sj}$ and $\widetilde{\gamma}_{\boldsymbol{\beta}j}$, the adaptive group lasso automatically imposes large penalties on groups of unimportant coefficients to efficiently shrink them to 0. Moreover, the penalty terms $\|\boldsymbol{\beta}_{sj}\|_{\mathbf{G}_{sj}}$ and $\left\|\widetilde{\boldsymbol{\beta}}_j\right\|_{\widetilde{\mathrm{G}}_j}$ can be regarded as the scaled version of the groupwise prediction penalty suggested by Buhlmann and Van De Geer.[25] With the great power of adaptive group lasso, the estimation of all unknown parameters and the structure detection for important functional covariate effects on the observed response and on the hidden-state process can be simultaneously and efficiently obtained.

3.2 | BaGlasso and prior specification

Under the Bayesian framework, the adaptive group lasso procedure can be implemented by introducing a multivariate conditional Laplace prior to the regression coefficients in $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}', \widetilde{\boldsymbol{\alpha}}', \boldsymbol{\beta}', \widetilde{\boldsymbol{\beta}}')'$ as follows:

$$p(\boldsymbol{\theta}^* \mid \Psi, \sigma^2) \propto \exp\left(-\sum_{h=1}^{p}\left(\frac{\gamma_{\alpha sh}}{\sqrt{\Psi_s}} \mid \alpha_{sh} \mid + \frac{\widetilde{\gamma}_{\alpha h}}{\sqrt{\sigma^2}} \mid \widetilde{\alpha}_h \mid\right) - \sum_{j=1}^{q}\left(\frac{\gamma_{\beta sj}}{\sqrt{\Psi_s}}\left\|\boldsymbol{\beta}_{sj}\right\|_{\mathrm{G}_{sj}} + \frac{\widetilde{\gamma}_{\beta j}}{\sqrt{\sigma^2}}\left\|\widetilde{\boldsymbol{\beta}}\right\|_{\widetilde{\mathrm{G}}_j}\right)\right), \tag{14}$$

where $\psi = (\psi_1, \dots, \psi_S)'$. This conditional Laplace prior can be represented as a scale mixture of normals with an exponential mixing density, leading to a hierarchical representation of the full model as follows: for $i = 1, \dots, n$, $t = 1, \dots, T$, $s = 1, \dots, S$, $h = 1, \dots, p$, and $j = 1, \dots, q$, we have

$$
\begin{aligned}
& y_{it} \mid Z_{it} = s, \mu_s, \mathbf{c}_{it}, \boldsymbol{\alpha}_s, \boldsymbol{\beta}_s, \Psi_s \sim N(\eta_{it}, \Psi_s), \\
& \boldsymbol{\alpha}_s \mid \Psi_s, \tau^2_{\alpha s1}, \ldots, \tau^2_{\alpha sp} \overset{\text{ind}}{\sim} N_p(\mathbf{0}, \Psi_s, \Sigma_{\alpha s}), \quad \Sigma_{\alpha s} = \operatorname{diag}(\tau_{\alpha s1}, \ldots, \tau_{\alpha sp}) \\
& \widetilde{\boldsymbol{\alpha}} \mid \sigma^2, \widetilde{\tau}^2_{\alpha 1}, \ldots, \widetilde{\tau}^2_{\alpha p} \sim N_p(\mathbf{0}, \sigma^2 \widetilde{\Sigma}_\alpha), \qquad \widetilde{\Sigma}_\alpha = \operatorname{diag}(\widetilde{\tau}_{\alpha 1}, \ldots, \widetilde{\tau}_{\alpha p}) \\
& \boldsymbol{\beta}_{sj} \mid \Psi_s, \tau^2_{\boldsymbol{\beta} sj} \overset{\text{ind}}{\sim} N_{M_j}\left(\mathbf{0}, \Psi_s \tau^2_{\boldsymbol{\beta} sj} \mathbf{G}^{-1}_{sj}\right), \quad \widetilde{\boldsymbol{\beta}}_j \mid \sigma^2, \widetilde{\tau}^2_{\boldsymbol{\beta} j} \overset{\text{ind}}{\sim} N_{M_j}\left(\mathbf{0}, \sigma^2 \widetilde{\tau}^2_{\boldsymbol{\beta} j} \widetilde{\mathbf{G}}^{-1}_j\right), \\
& \tau^2_{\alpha sh} \overset{\text{ind}}{\sim} \operatorname{Gamma}\left(1, \frac{\gamma^2_{\alpha sh}}{2}\right), \quad \widetilde{\tau}^2_{\alpha h} \overset{\text{ind}}{\sim} \operatorname{Gamma}\left(1, \frac{\widetilde{\gamma}^2_{\alpha h}}{2}\right) \\
& \tau^2_{\beta sj} \overset{\text{ind}}{\sim} \operatorname{Gamma}\left(\frac{M_j + 1}{2}, \frac{\gamma^2_{\beta sj}}{2}\right), \quad \widetilde{\tau}^2_{\beta j} \overset{\text{ind}}{\sim} \operatorname{Gamma}\left(\frac{M_j + 1}{2}, \frac{\widetilde{\gamma}^2_{\beta j}}{2}\right) \\
& \Psi_s^{-1} \overset{\text{ind}}{\sim} \operatorname{Gamma}(\alpha_{\Psi s0}, \beta_{\Psi s0}), \quad \sigma^{-2} \sim \operatorname{Gamma}(\alpha_{\sigma 0}, \beta_{\sigma 0},
\end{aligned}
\tag{15}
$$

where $\overset{\text{ind}}{\sim}$ represents “independently distributed according to” and η_{it} is defined in (9). For the tuning parameters $\gamma_{\alpha sh}$, $\widetilde{\gamma}_{\alpha h}$, $\gamma_{\boldsymbol{\beta} sj}$, and $\widetilde{\gamma}_{\boldsymbol{\beta} j}$, we assign gamma priors as follows:

$$
\begin{aligned}
& p(\gamma^2_{\alpha sh}) \overset{\text{ind}}{\sim} \operatorname{Gamma}(\alpha_{\alpha sh0}, \beta_{\alpha sh0}), \quad p(\widetilde{\gamma}^2_{\alpha h}) \overset{\text{ind}}{\sim} \operatorname{Gamma}(\widetilde{\alpha}_{\alpha h0}, \widetilde{\beta}_{\alpha h0}), \\
& p(\gamma^2_{\beta sj}) \overset{\text{ind}}{\sim} \operatorname{Gamma}(\alpha_{\beta sj0}, \beta_{\beta sj0}), \quad p(\widetilde{\gamma}^2_{\beta j}) \overset{\text{ind}}{\sim} \operatorname{Gamma}(\widetilde{\alpha}_{\beta j0}, \widetilde{\beta}_{\beta j0}),
\end{aligned}
\tag{16}
$$

where $\alpha_{\alpha sh0}$, $\widetilde{\alpha}_{\alpha h0}$, $\alpha_{\boldsymbol{\beta} sj0}$, $\widetilde{\alpha}_{\boldsymbol{\beta} j0}$, $\beta_{\alpha sh0}$, $\widetilde{\boldsymbol{\beta}}_{\alpha h0}$, $\beta_{\boldsymbol{\beta} sj0}$, and $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\beta} j0}$ are hyperparameters with prespecified values. We follow a common practice in the literature[11,12] to set $\alpha_{\alpha sh0} = \widetilde{\alpha}_{\alpha h0} = \alpha_{\boldsymbol{\beta} sj0} = \widetilde{\alpha}_{\boldsymbol{\beta} j0} = 1$, $\boldsymbol{\beta}_{\alpha sh0} = \widetilde{\boldsymbol{\beta}}_{\alpha h0} = 0.1$, and $\boldsymbol{\beta}_{\boldsymbol{\beta} sj0} = \widetilde{\boldsymbol{\beta}}_{\boldsymbol{\beta} j0} = 0.01$ to obtain relatively dispersed gamma priors. The key idea of BaGlasso is to properly update the tuning parameters by using the data, thereby automatically imposing large penalties on unimportant coefficients. This target can be naturally achieved by introducing dispersed priors with small hyperparameters $\beta_{\alpha sh0}$, $\widetilde{\boldsymbol{\beta}}_{\alpha h0}$, $\beta_{\boldsymbol{\beta} sj0}$, and $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\beta} j0}$. We explain this regularization procedure further through the posterior distribution of the tuning parameters as follows:

$$
\begin{aligned}
&p\left(\tau_{\alpha sh}^{-2} \,\middle|\, \cdot\right) \sim \mathrm{In}-\mathrm{Gaussian}\left(\sqrt{\frac{\gamma_{\alpha sh}^{2}\Psi_{s}}{|\,\alpha_{sh}\,|^{2}}}, \gamma_{\alpha sh}^{2}\right), \qquad p\left(\widetilde{\tau}_{\alpha h}^{-2} \,\middle|\, \cdot\right) \sim \mathrm{In}-\mathrm{Gaussian}\left(\sqrt{\frac{\widetilde{\gamma}_{\alpha sh}^{2}\sigma^{2}}{|\,\widetilde{\alpha}_{sh}\,|^{2}}}, \widetilde{\gamma}_{\alpha sh}^{2}\right),\\
&p\left(\tau_{\beta sj}^{-2} \,\middle|\, \cdot\right) \sim \mathrm{In}-\mathrm{Gaussian}\left(\sqrt{\frac{\gamma_{\beta sj}^{2}\Psi_{s}}{\|\boldsymbol{\beta}_{sj}\|_{\mathrm{G}_{sj}}}}, \gamma_{\beta sj}^{2}\right), \qquad p\left(\widetilde{\tau}_{\beta j}^{-2} \,\middle|\, \cdot\right) \sim \mathrm{In}-\mathrm{Gaussian}\left(\sqrt{\frac{\widetilde{\gamma}_{\beta sj}^{2}\sigma^{2}}{\|\widetilde{\boldsymbol{\beta}}_{j}\|_{\widetilde{\mathrm{G}}_{j}}}}, \widetilde{\gamma}_{\beta j}^{2}\right),\\
&p\left(\gamma_{\alpha sh}^{2} \,\middle|\, \cdot\right) \sim \mathrm{Gamma}\left(\alpha_{\alpha sh0}+1, \beta_{\alpha sh0}+\frac{\tau_{\alpha sh}^{2}}{2}\right), \qquad p\left(\widetilde{\gamma}_{\alpha h}^{2} \,\middle|\, \cdot\right) \sim \mathrm{Gamma}\left(\widetilde{\alpha}_{\alpha h0}+1, \widetilde{\beta}_{\alpha h0}+\frac{\widetilde{\tau}_{\alpha h}^{2}}{2}\right),\\
&p\left(\gamma_{\beta sj}^{2} \,\middle|\, \cdot\right) \sim \mathrm{Gamma}\left(\alpha_{\beta sj0}+\frac{M_{j}+1}{2}, \beta_{\beta sj0}+\frac{\tau_{\beta sj}^{2}}{2}\right),\\
&p\left(\gamma_{\beta j}^{2} \,\middle|\, \cdot\right) \sim \mathrm{Gamma}\left(\widetilde{\alpha}_{\beta j0}+\frac{M_{j}+1}{2}, \widetilde{\beta}_{\beta j0}+\frac{\tau_{\beta j}^{2}}{2}\right),
\end{aligned}
\tag{17}
$$

where "In-Gaussian(·)" denotes the inverse Gaussian distribution. We omit the tedious subscripts and use generic terms τ and γ to simplify notations below. On the basis of (17), if the coefficients are significant, then τ^2 tends to be large. As a result, the corresponding tuning parameter γ is dominated by τ^2, leading γ to be mostly data driven. If the coefficients are insignificant, then τ^2 tends to be small. Consequently, the corresponding tuning parameter γ is dominated by the dispersed prior information, leading to a large value of γ. Thus, the degree of dispersion of the gamma priors in (16) determines the amount of penalties imposed on unimportant predictors. This rationale explains why we assign higher dispersed priors to $\gamma_{\boldsymbol{\beta} sj}^{2}$ and $\widetilde{\gamma}_{\boldsymbol{\beta} j}^{2}$ than to $\gamma_{\alpha sh}^{2}$ and $\widetilde{\alpha}_{\alpha h}^{2}$ because the coefficients of the nonlinear parts of basis functions are more difficult to shrink to 0 than those of the linear parts.

To conduct a full Bayesian analysis, we specify appropriate prior distributions for other unknown parameters, such as μ_s, $\boldsymbol{\pi}_s$, and ζ_{us}. For $u = 1, \ldots, S$ and $s = 1, \ldots, S$, the following Gaussian priors are considered:

$$
p(\mu_s) \overset{\text{ind}}{\sim} N\left(\mu_{s0}, \sigma_{\mu s0}^{2}\right), \qquad p(\pi_s) \overset{\text{ind}}{\sim} N\left(\pi_{s0}, \sigma_{\pi s0}^{2}\right), \qquad p(\zeta_{us}) \overset{\text{ind}}{\sim} N\left(\zeta_{us0}, \sigma_{\zeta us0}^{2}\right), \tag{18}
$$

where μ_{s0}, $\sigma_{\mu s0}^{2}$, $\boldsymbol{\pi}_{s0}$, $\sigma_{\pi s0}^{2}$, ζ_{us0}, and $\sigma_{\zeta us0}^{2}$ are hyperparameters with preassigned values.

3.3 | Posterior inference

The Bayesian estimate of $\boldsymbol{\theta}$ can be obtained through the mean or mode of the posterior samples drawn from $p(\boldsymbol{\theta}|\mathbf{Y})$. However, directly sampling from $p(\boldsymbol{\theta}|\mathbf{Y})$ is intractable because of the existence of latent states. To address this issue, we adopt the data augmentation technique to work on $p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y})$ and utilize the Gibbs sampler to simulate each of the unknowns from its full conditional distribution iteratively. Owing to the nonlinearity of the continuation-logit transition model, the full conditional distributions related to the transition model have complex forms. Thus, Markov chain Monte Carlo (MCMC) methods, such as the forward filtering and backward sampling algorithm[26] and the Metropolis-Hastings algorithm,[27,28] are used to sample from them. The details of the full conditional distributions are provided in the Appendix.

For nonparametric functions involved in (1) and (3), as suggested by Li et al,[29] a functional effect of a covariate is detected as significant and included in the regression if at least one of its coefficients of the basis expansion has a two-sided 95% credible interval estimate that does not cover zero. The latent state Z_{it}, which usually has actual meaning in empirical studies, is also of great interest for scientists. By using posterior samples, we can estimate the hidden state as follows:

$$\hat{Z}_{it} = \arg \max_{s \in \{1, \ldots, S\}} P(Z_{it} = s \mid \mathrm{y}_i, \boldsymbol{\theta}) \approx \arg \max_{s \in \{1, \ldots, S\}} \frac{1}{L} \sum_{l=1}^{L} \mathrm{I}\left(Z_{it}^{(l)} = s\right), \quad (19)$$

where $Z_{it}^{(l)}$ denotes the latent allocation of y_{it} at the lth iteration, and $\frac{1}{L}\sum_{l=1}^{M} \mathrm{I}(Z_{it}^{(l)} = s)$ is the posterior mean of the latent allocations of y_{it} drawn from the MCMC iterations.

3.4 | Determination of the number of hidden states

In the analysis of HMMs, the number of hidden states, S, is usually determined a priori. We use a modified DIC, which was developed by Celeux et al,[30] for model comparison in the presence of incomplete data, to determine the number of hidden states of the proposed model. The modified DIC is defined as follows:

$$\mathrm{DIC} = \overline{D(\boldsymbol{\theta})} + p_D, \quad (20)$$

where $\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta}, \mathrm{Z}}[-2 \log p(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}) \mid \mathbf{Y}]$ is the posterior mean deviance to reflect the goodness of fit of the model, p_D is the effective number of parameters to penalize an overcomplex model, and $p_D = E_{\theta,Z}[-2\log p(\mathbf{Y}, \mathbf{Z}|\theta)|\mathbf{Y}] + 2E_{\mathrm{Z}}[\log p(\mathbf{Y}, \mathbf{Z})|E_{\theta}[\theta|\mathbf{Y}, \mathbf{Z}])|\mathbf{Y}]$. The expectations involved in (20) can be approximated by averaging the posterior samples collected through the MCMC algorithm.[30,31] The model with the smallest value of DIC is selected.

4 | SIMULATION STUDY

This section contains two simulations: Simulation 1 assesses the empirical performance of the proposed BaGlasso for simultaneous estimation and variable selection in the context of semiparametric HMMs, and Simulation 2 examines the performance of the DIC in determining the number of hidden states in semiparametric HMMs.

4.1 | Simulation 1

We consider 100 simulated data sets, each consisting of $n = 700$ subjects and $T = 9$ time points. For each data set, observations are generated from a two-state semiparametric HMM with a continuous response y_{it}, two discrete covariates $\mathbf{c}_{it} = (c_{it1}, c_{it2})^r$ (p = 2), and three continuous covariates $x_{it} = (x_{it1}, x_{it2}, x_{it3})'$ ($q = 3$). For $i = 1, \ldots, 700$ and $t = 1, \ldots, 9$, c_{it1} and c_{it2} are independently generated from the Bernoulli distribution with a probability of success of 0.5, and x_{it1}, x_{it2}, and x_{it3} are generated from $U(-1, 1)$, $N(0, 1)$, and $N(\sqrt{t}, 1)$, respectively, and they are standardized to the same scale beforehand. Here, x_{it1} and x_{it2} are set as time-invariant covariates, whereas x_{it3} is set as a time-variant one. The conditional regression model is defined as follows:

$$[y_{it} \mid Z_{it} = s] = \mu_s + \alpha_{s1} c_{it1} + \alpha_{s2} c_{it2} + f_{s1}(x_{it1}) + f_{s2}(x_{it2}) + f_{s3}(x_{it3}) + \delta_{it}, \quad (21)$$

where $f_{11}(x_{it1}) = 0$, $f_{12}(x_{it2}) = \sin(1.5x_{it2}) + x_{it2} - 0.6$, $f_{13}(x_{it1}) = -0.8x_{it3}$, $f_{21}(x_{it1}) = 2.08 - \exp(x_{it1})$, $f_{22}(x_{it_2}) = 0$, and $f_{23}(x_{it3}) = -0.105 + \cos(2x_{it3}) + 0.5x_{it3}$.

The transition model is defined as

$$\operatorname{logit}(\vartheta_{itus}) = \zeta_{us} + \tilde{\alpha}_1 c_{it1} + \tilde{\alpha}_2 c_{it2} + g_1(x_{it1}) + g_2(x_{it2}) + g_3(x_{it3}), \quad (22)$$

where $g_1(x_{it1}) = -\log(2 + x_{it1})/(2 - x_{it1})$, $g_2(x_{it2}) = 1.5x_{it2}$, and $g_3(x_{it3}) = 0$. The true population values of the unknown parameters are set as $\boldsymbol{\mu} = (\mu_1, \mu_2)' = (-1, 1)'$, $\boldsymbol{\pi} = (\pi_1, \pi_2)' = (0.5, 0.5)'$, $\zeta_{11} = \zeta_{21} = 0.5$, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})' = (0, 0.5)'$, $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22})' = (-0.5, 0)-$, $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \tilde{\alpha}_2)' = (-1, 0)'$, and $\boldsymbol{\psi} = (\psi_1, \psi_2)' = (0.36, 0.16)'$.

In this study, we use a simple version of natural cubic splines derived from a truncated power series basis function[16] to approximate the nonparametric functions: $h_{j1}(x_{itj}) = 1$, $h_{j2}(x_{itj}) = x_{itj}$, and $h_{j,m+2} = u_{jm}(x_{itj}) - u_{j,M_j-1}(x_{itj})$ for $m = 1, \ldots, M_j - 2$, where $u_{j,m}(x_{itj}) = [(x_{itj} - \kappa_j M_j)_+^3 - (x_{itj} - \kappa_{jm})_+^3] \big/ (\kappa_j M_j - \kappa_{jm})$, and κ_{jm}, $m = 1, \ldots, M_j$, are the knots taken in the range of x_{itj}. The prior inputs in (15), (16), and (18) are assigned as follows: $\mu_{s0} = \zeta_{us0} = \pi_{s0} = 0$, $\sigma^2_{\mu s0} = \sigma^2_{\zeta us0} = \sigma^2_{\pi 0} = 1$, $a_{\psi s0} = a_{\sigma 0} = 9$, $\beta_{\psi s0} = \beta_{\sigma 0} = 4$, $\alpha_{\alpha sh0} = \tilde{\alpha}_{\alpha h0} = \alpha_{\beta sj0} = \tilde{\alpha}_{\beta j0} = 1$, $\boldsymbol{\beta}_{\alpha sh0} = \tilde{\boldsymbol{\beta}}_{\alpha h0} = 0.1$, and $\beta_{\beta sj0} = \tilde{\beta}_{\beta j0} = 0.01$. For each x_{itj}, $M_j = 10$ knots are used. We impose the constraint $\mu_1 < \mu_2$ in each MCMC iteration to avoid label switching and check the convergence of the algorithm using the estimated potential scale reduction (EPSR) proposed by Gelman et al.[32] The MCMC algorithm converges within 5000 iterations. Thus, we collect posterior samples with a size of 20 000 with the first 10 000 as burn-in iterations. The performance of Bayesian estimates is assessed through the bias (BIAS) and the root-mean-square error (RMSE) between the Bayesian estimates and the true population values of the parameters.

Table 1 summarizes the estimation results on the basis of the 100 data sets. The BIAS and RMSE for most of the parameters are close to zero, indicating a satisfactory performance of

Bayesian estimation regarding the parametric part. Figure 1 depicts the averages of the pointwise posterior means of the nonparametric functions, along with their 2.5% and 97.5% pointwise quantiles. Three nonexistent functions are successfully shrunk to almost zero by the proposed BaGlasso procedure. The posterior means of other nonzero nonparametric functions are close to their true curves, and all the ranges of the 2.5% and 97.5% pointwise quantiles are small, indicating that the estimated nonparametric curves can correctly recover the complex functional relationships between the response and covariates. Moreover, the average of the correct classification rates calculated from (19) is approximately 95%, implying the good performance of the proposed method in identifying the hidden states of the observations.

To reveal the sensitivity of Bayesian estimates to the input of prior distributions, we disturb the prior input as follows: $\mu_{s0} = \zeta_{us0} = \boldsymbol{\pi}_{s0} = 2$, $\sigma^2_{\mu s0} = \sigma^2_{\zeta us0} = \sigma^2_{\pi s0} = 2$, $\boldsymbol{\alpha}_{\psi s0} = 3$, $\boldsymbol{\beta}_{\psi s0} = 2$, $\alpha_{\alpha sh0} = \widetilde{\alpha}_{\alpha h0} = \alpha_{\boldsymbol{\beta} sj0} = \widetilde{\alpha}_{\boldsymbol{\beta} j0} = 1$, $\boldsymbol{\beta}_{\alpha sh0} = \widetilde{\boldsymbol{\beta}}_{\alpha h0} = 0.5$, and $\beta_{\boldsymbol{\beta} sj0} = \widetilde{\beta}_{\boldsymbol{\beta} j0} = 0.01$. The Bayesian results obtained under the disturbed prior are similar and not reported.

Notably, this simulation study contains five covariates in the conditional and transition models, which result in a large number ($2^{2\times5}$) of competing models with various combinations of covariates in both models. Traditional Bayesian model selection statistics, such as the Bayes factor and the DIC, are extremely time consuming in performing variable selection because they compare these competing models in a pairwise basis. By contrast, the proposed BaGlasso procedure automatically selects important predictors and avoids the tedious pairwise comparison, thereby greatly reducing the computational time. In this simulation study, the computing time for simultaneous variable selection and parameter estimation in each replication is 48 minutes using a PC Intel Core i7-6700 3.40-GHz CPU and 16 G of RAM.

4.2 | Simulation 2

To examine the performance of the DIC in determining the number of hidden states of a semiparametric HMM, we consider five competing models M_1, M_2, M_3, M_4, and M_5, where M_s is a model defined by (1)–(3) with $S = s$, $s = 1, \dots, 5$. Here, M_4 is the true model, whereas M_1, M_2, M_3, and M_5 are models with incorrect numbers of hidden states. To mimic the scenario of the ADNI data set in the subsequent real example, we generate 100 data sets from (1)–(3) with $S = 4$, $n = 633$, $T = 4$,$p = 4$, and $q = 3$. For $i = 1, \dots, 633$ and $t = 1, \dots, 4$, c_{it1} to c_{it4} are independently generated from the Bernoulli distribution with a probability of success of 0.5, and x_{it1}, x_{it2}, and x_{it3} are generated from $U(-1, 1)$, $N(0, 1)$, and $N(\sqrt{t}, 1)$, respectively, and they are standardized prior to analysis. The true functions are set as $f_{11}(x_{it1}) = 0$, $f_{12}(x_{it2}) = \sin(1.5x_{it2}) + x_{it2} - 0.6$, $f_{13}(x_{it1}) = -0.8x_{it3}$, $f_{21}(x_{it1}) = 2.08 - \exp(x_{it1})$, $f_{22}(x_{it2}) = 0$, $f_{23}(x_{it3}) = -0.105 + \cos(2x_{it3}) + 0.5x_{it3}$, $f_{31}(x_{it1}) = 0.5x_{it1}$, $f_{32}(x_{it2}) = 0$, $f_{33}(x_{it1}) = -x_{it3}$, $f_{41}(x_{it1}) = 2x_{it1}$, $f_{42}(x_{it2}) = 1.5x_{it2}$, $f_{43}(x_{it3}) = 0$, $g_1(x_{it1}) = -\log(2 + x_{it1})/(2 - x_{it1})$, $g_2(x_{it2}) = 1.5x_{it2}$, and $g_3(x_{it3}) = 0$. The true population values of the unknown parameters are set as $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)' = (-4, -2, 2, 4)'$, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3, \boldsymbol{\pi}_4)' = (0.25, 0.25, 0.25, 0.25)'$, $\zeta_{11} = \zeta_{21} = \zeta_{31} = \zeta_{41} = -1$, $\zeta_{12} = \zeta_{22} = \zeta_{32} = \zeta_{42} = 0$, $\zeta_{13} = \zeta_{23} = \zeta_{33} = \zeta_{43} = 1$, $\boldsymbol{a}1 = (a_{11}, a_{12}, a_{13}, a_{14})' = (1, 0, 0.5, 1)'$, $\boldsymbol{a}_2 = (a_{21}, a_{22}, a_{23}, a_{24})' = (0.5, -0.5, 0, -1)'$, $\boldsymbol{a}_3 = (a_{31}, a_{32}, a_{33}, a_{34})' = (0.5, -1, 1, 0)'$, $\boldsymbol{a}_4 = (a_{41}, a_{42}, a_{43}, a_{44})' = (0.5, 1,$

$-0.5, 0)'$, $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4)' = (-1, 0.5, 0, 1)'$, and $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3, \psi_4)' = (0.16, 0.16, 0.16, 0.16)'$. The prior distributions and other settings are specified in the same manner as in Simulation 1. On the basis of the 100 simulated data sets, the means and standard deviations of the DIC values for M_1 to M_5 are reported in Table 2, which suggests that the true model M_4 is consistently selected in each of the 100 replications.

The computer code for conducting the preceding analyses is written in R and is freely available at http://www.sta.cuhk.edu.hk/xysong/codes/BaGLassoHMMs.

5 | ADNI STUDY

To demonstrate the empirical utility of our proposed method, we conduct real data analysis on the basis of the ADNI study. The data set collected imaging, genetic, clinical, and cognitive data from participants under CN controls and participants with mild cognitive impairment or AD. ADNI-1 was first conducted in 2004, and several extensions, namely, ADNI-GO, ADNI-2, and ADNI-3, followed afterward. In this study, we focused on 633 participants collected from ADNI-1 and included their clinical and genetic variables at four time points, namely, baseline, 6 months, 12 months, and 24 months. Functional Assessment Questionnaire (FAQ), a widely used assessment of abilities to function independently in daily life, was used as a response variable (y_{it}) to reflect cognitive decline over time. Patients with higher FAQ scores have lower cognitive abilities. Three continuous covariates, namely, the logarithm of the ratio of hippocampal volume over whole brain (x_{it1}), age at baseline (x_{it2}), and years of education (x_{it3}), were considered. Moreover, we included a genetic variable, APOE-ϵ4 (c_{it1} and c_{it2}), which was coded as 0, 1, and 2, denoting the number of APOE-ϵ4 alleles. Other discrete demographic characteristics, such as gender (c_{it3}, 0 = female; 1 = male) and marital status (c_{it4}, 0 = has been married; 1 = has not been married), were also included. The three continuous variables, namely, FAQ score, hippocampus, and age, were standardized prior to analysis. The main objective of this study is to examine the complex effects of potential risk factors on the transition of neurodegenerative states and on the cognitive decline of participants across different states.

We first determined the number of hidden states. We considered five competing models M_k, $k = 1, \dots, 5$, where M_k represents a semiparametric HMM defined in (1)–(3) with k states. We used natural cubic splines for $\mathbf{h}_{itj}$ and $M_j = 10$ in approximating the unknown smoothing functions. The hyperparameters were assigned in the same manner as those in the simulation study, and the identifiability constraint $\mu_1 < \cdots < \mu_5$ was taken to avoid label switching. We generated several MCMC chains with different initial values to monitor the convergence of the MCMC algorithm. The EPSR plot depicted in Figure 2 indicated that the MCMC algorithm converged within 10 000 iterations. Therefore, we collected 10 000 observations after discarding 10 000 burn-in iterations to calculate the DIC values of the competing models.

The values of $\overline{D(\boldsymbol{\theta})}$, p_D, and DIC corresponding to M_1 to M_4 are reported in Table 3. When fitting the data to M_5, the MCMC algorithm broke down after several iterations. After carefully checking the results, we found that one of the states included only fewer than six subjects after several iterations. This phenomenon implies the nonexistence of such a state and the inapplicability of the five-state model in this study. On the basis of the results in Table 3, the four-state model M_4 with the smallest DIC was selected. Then, we used the proposed BaGlasso procedure to conduct a simultaneous estimation and variable selection under M_4. Results are presented in Table 4 (parametric part) and Figure 3 (nonparametric part), in which only significant functional effects are reported.

We obtain the following observations: first, intercepts μ_1, μ_2, μ_3, and μ_4 were ranked in ascending order. Patients in state 1 had the lowest mean score of FAQ, whereas those in state 4 received the highest mean score. That is, patients' cognitive ability reflected by independent functioning in daily life steadily deteriorated from state 1 to state 4. According to the existing literature,[33] state 1 to state 4 can be explained as CN, early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD, respectively.

Second, BaGlasso selected six significant functional effects across the states. The effect of hippocampus on the FAQ score exhibits a descending trend in all the states. Specifically, in the CN state, participants with a greater hippocampal volume tend to have slightly better memory. This result is consistent with the common sense that the hippocampus helps consolidate outside information from short-term memory to long-term memory. In EMCI and LMCI states, the magnitude of the functional effect of the hippocampus on FAQ becomes increasingly large, confirming that atrophy in hippocampal volume continuously impairs patients' cognitive ability during the progression from EMCI to LMCI. Published medical reports[34-36] also revealed the similar result that the loss of hippocampal volume greatly affects dementia. In the AD state, preventing the loss of hippocampal volume is still beneficial to postpone cognitive decline, but this effect is significant only in a small range of hippocampal volume. The effect of age on FAQ is nonsignificant in the first three states, implying that age influences cognitive function mainly in the AD state. Relatively younger AD patients (around 75 years old) have better functional independence in daily life compared with elder ones. This age effect was also revealed by previous research.[37,38] The effect of educational level on FAQ is likewise significant only in the AD state. Such effect becomes large when educational level is high, indicating that patients with higher educational levels tend to experience more pronounced cognitive decline compared with patients with lower educational levels. This finding is in line with the existing literature.[39,40]

Third, for the parametric part, gender has a negative effect on FAQ in the LMCI and AD states, implying that women suffer more serious cognitive decline than men in the late progression period of AD. This result agrees with existing medical reports.[41-43]

Fourth, in the transition model, the functional effect of the hippocampus exhibits an ascending trend with the growth of hippocampal volume. In the progression of AD, patients with larger hippocampal volumes are more likely to remain in the current state rather than transit to a worse one compared with those with smaller hippocampal volumes. By contrast, patients with APOE-ϵ4 alleles are more likely to transit to a worse state rather than remain in the current one. Thus, APOE-ϵ4 alleles are important risk factors for the development of AD. This result is consistent with the existing finding.[44] However, the estimates of other covariates, such as age, educational level, gender, and marital status, were shrunk to nearly zero by BaGlasso, implying that conditional on hippocampus and APOE-ϵ4, the direct effects of age, educational level, gender, and marital status on the transition probability are weak.

For comparison, we reanalyzed the ADNI data set using a parametric HMM as follows:

$$
\begin{aligned}
[y_{it} \mid Z_{it} = s] &= \mu_s + \alpha_{s1} c_{it1} + \alpha_{s2} c_{it2} + \alpha_{s3} c_{it3} + \alpha_{s4} c_{it4} + \beta_{s1} x_{it1} + \beta_{s2} x_{it2} + \beta_{s3} x_{it3} + \delta_{it}, \\
\operatorname{logit}(\vartheta_{itus}) &= \zeta_{us} + \tilde{\alpha}_1 c_{it1} + \tilde{\alpha}_2 c_{it2} + \tilde{\alpha}_3 c_{it3} + \tilde{\alpha}_4 c_{it4} + \tilde{\beta}_1 x_{it1} + \tilde{\beta}_2 x_{it2} + \tilde{\beta}_3 x_{it3} \,.
\end{aligned}
$$

The Bayesian adaptive lasso procedure was used to perform estimation. Table 5 presents the results of parameters β_{sj} and $\tilde{\beta}_j$. The results of μ_s, ζ_{us}, α_{sh}, and $\tilde{\alpha}_h$ are similar to those in Table 4 and not reported. Several differences exist between the results obtained using the parametric and semiparametric HMMs. First, the parametric model shows a negative constant effect of the hippocampus on FAQ in the CN, EMCI, and LMCI states, whereas the semiparametric model reveals that these negative effects have a descending trend. Second, the parametric model indicates that the effects of the hippocampus, age, and educational level on FAQ are all insignificant in the AD state, whereas the semiparametric model reveals that these effects are actually significant in certain covariate ranges. Finally, the parametric model shows that the effect of age on FAQ is negative in the NC and EMCI states but positive in the LMCI state. This diverse effect is hard to interpret and probably caused by overlooking the subtle structure of the age effect in the parametric model.

6 | CONCLUSION

In this paper, we have introduced a BaGlasso procedure to conduct simultaneous variable selection and parameter estimation in the context of semiparametric HMMs. We developed a full Bayesian approach, along with efficient MCMC methods and the basis expansion technique, to implement the procedure and estimate nonparametric functions. The methodology was demonstrated by a simulation study and an application to the analysis of the ADNI data set. In the proposed model, covariates are allowed to affect both responses and transition probabilities. This feature enables the model to cope with general situations where certain covariates simultaneously influence the two stochastic processes in various ways. An alternative method of including covariates in HMMs is to use an exclusion restriction to split the overall set of covariates into two groups: one contains covariates affecting only the responses, and the other contains covariates affecting the hidden-state transition. However, determining such an exclusion restriction may be subjective and difficult to justify in practice, which, in turn, elicits model selection issues.

This study can be extended in several directions. First, in approximating nonparametric functions, we considered only a simple version of natural cubic splines. Highly sophisticated smoothing techniques, such as splines and local polynomial kernel methods, may be used to enhance the performance of estimation and variable selection. Second, we simply used a single indicator, FAQ, to reflect cognitive ability in the ADNI data analysis. A comprehensive way to characterize cognitive function is to account for other relevant tests, such as the Alzheimer's Disease Assessment Scale and the Mini-Mental State Examination. Grouping such highly correlated but different perspectives into an integrated latent variable through factor analysis can improve the analytic power and interpretability of the model. Finally, our model framework includes only binary and continuous variables. Given that ordered and unordered categorical data are frequently encountered in medical, social, and psychological sciences, generalizing the existing framework to accommodate a wide variety of data types is of great interest.

ACKNOWLEDGEMENTS

The work of Xinyuan Song was supported by the Research Grants Council of Hong Kong under grant 14303017, The Chinese University of Hong Kong under direct grants, and the National Natural Science Foundation of China under grant 11471277. The work of Joan Hu was supported by the Canadian Institutes of Health Research under grant RN120660 and the Natural Sciences and Engineering Research Council of Canada under grant 177430. The authors are thankful to the Editor, the Associate Editor, and three anonymous reviewers for their valuable comments and suggestions.

Funding information

Research Grants Council of Hong Kong, Grant/Award Number: 14303017; National Natural Science Foundation of China, Grant/Award Number: 11471277; Canadian Institutes of Health Research, Grant/Award Number: RN120660; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: DAS 177430

APPENDIX A

FULL CONDITIONAL DISTRIBUTIONS

A.1 | Full conditional distributions of Z_{it}

Let $\mathbf{y}_{\mathrm{i}} = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{d}_{it} = (\mathbf{c}'_{it}, \mathbf{x}'_{it})'$, and $\mathbf{D}_i = (\mathbf{d}'_{i1}, \ldots, \mathbf{d}'_{iT})'$. Then, we have

$$
\begin{aligned}
& p(Z_{it} \mid \cdot\,) \propto p(\mathbf{y}_i, \mathbf{D}_i, Z_{it} \mid \boldsymbol{\theta}) \\
& = p(y_{i1}, \ldots, y_{it}, \mathbf{d}_{i1}, \ldots, \mathbf{d}_{it}, Z_{it} \mid \boldsymbol{\theta}) \times p(y_{i,t+1}, \ldots, y_{iT}, \mathbf{d}_{i,t+1}, \ldots, \mathbf{d}_{iT} \mid Z_{it}, \boldsymbol{\theta}) \\
& = q_{it}(\mathbf{y}_i, \mathbf{D}_i, Z_{it} \mid \boldsymbol{\theta}) \times \overline{q}_{it}(\mathbf{y}_i, \mathbf{D}_i \mid Z_{it}, \boldsymbol{\theta})\,.
\end{aligned}
$$

We first initialize $q_{i1}(\mathbf{y}_i, \mathbf{D}_i, Z_{it}|\boldsymbol{\theta}) = p(y_{i1}, \mathbf{d}_{i1}, Z_{it}|\boldsymbol{\theta}) = p(y_{i1}|\mathbf{d}_{i1}, Z_{i1}, \boldsymbol{\theta})p(Z_{i1}|\boldsymbol{\theta})$ and calculate $q_{it}(\mathbf{y}_i, \mathbf{D}_i, Z_{it}|\boldsymbol{\theta})$ for $t = 2, \ldots, T$, in a recursion manner as follows:

$$
\begin{aligned}
& q_{it}(\mathbf{y}_i, \mathbf{D}_i, Z_{it} \mid \boldsymbol{\theta}) = q_{it}(y_{i1}, \ldots, y_{it}, \mathbf{d}_{i1}, \ldots, \mathbf{d}_{iT}, Z_{it} \mid \boldsymbol{\theta}) \\
= & \sum_{u=1}^{S} p(y_{i1}, \ldots, y_{it}, \mathbf{d}_{i1}, \ldots, \mathbf{d}_{iT}, Z_{it}, Z_{i,t-1} = u \mid \boldsymbol{\theta}) \\
= & \left. \sum_{u=1}^{S} p(y_{i1}, \ldots, y_{it}, \mathbf{d}_{i1}, \ldots, \mathbf{d}_{iT} Z_{i,t-1} = u \mid \boldsymbol{\theta}) \times p(Z_{it} \mid Z_{i,t-1} = u, \mathbf{d}_{it}, \boldsymbol{\theta}) \times p(y_{it} \mid \mathbf{d}_{it}, Z_{it}, \boldsymbol{\theta}) \right] \\
= & \sum_{u=1}^{S} \left[q_{i,t-1}(\mathbf{y}_i, \mathbf{D}_i, Z_{i,t-1} = u \mid \boldsymbol{\theta}) \times p(Z_{i,t-1} = u, \mathbf{d}_{it}, \boldsymbol{\theta}) \times p(y_{it} \mid \mathbf{d}_{it}, Z_{it} \boldsymbol{\theta}) \right],
\end{aligned}
\tag{A1}
$$

where $p(Z_{it}|Z_{i,t-1} = u, \mathbf{d}_{it}, \boldsymbol{\theta})$ and $p(y_{it}, \mathbf{d}_{it}|Z_{it}, w_{i1}, \boldsymbol{\theta})$ can be calculated on the basis of (8).

Similarly, we initialize $\overline{q}_{iT}(\mathbf{y}_i, \mathbf{D}_i \mid Z_{iT}, \boldsymbol{\theta}) = 1$ and calculate $\overline{q}_{it}(\mathbf{y}_i, \mathbf{D}_i \mid Z_{it}, \boldsymbol{\theta})$ for $t = T - 1, -, 1$ as follows:

$$\begin{aligned}
\overline{q}_{it}\ (\mathbf{y}_i, \mathbf{D}_i \mid Z_{it}, \boldsymbol{\theta}) &= p(y_{i,t+1}, \ldots, y_{iT}, \mathbf{d}_{i,t+1}, \ldots, \mathbf{d}_{iT} \mid Z_{it}, \boldsymbol{\theta}) \\
&= \sum_{u=1}^{S} p(y_{i,t+1}, \ldots, y_{iT}, \mathbf{d}_{i,t+1}, \ldots, \mathbf{d}_{iT}, Z_{i,t+1} = u \mid Z_{it}, \boldsymbol{\theta}) \\
&= \sum_{u=1}^{S} \Big[p(y_{i,t+1}, \ldots, y_{iT}, \mathbf{d}_{i,t+1}, \ldots, \mathbf{d}_{iT} \mid Z_{i,t+1} = u, \boldsymbol{\theta}) \times p(Z_{i,t+1} = u \mid Z_{it}, \mathbf{d}_{i,t+1}, \boldsymbol{\theta}) \\
&\qquad \times p(y_{i,t+1} \mid \mathbf{d}_{i,t+1}, Z_{i,t+1} = u, \boldsymbol{\theta})\Big] \\
&= \sum_{u=1}^{S} \Big[\overline{q}_{i,t+1}(\mathbf{y}_i, \mathbf{D}_i \mid Z_{i,t+1}{=}u, \boldsymbol{\theta}) \times p(Z_{i,t+1} = u \mid Z_{it}, \mathbf{d}_{i,t+1}, \boldsymbol{\theta}) \times p(y_{i,t+1} \mid \mathbf{d}_{i,t+1}, Z_{i,t-1} = u, \boldsymbol{\theta})\Big]
\end{aligned}$$

(A2)

Thus, Z_{it} can be directly generated from (A1) when all $q_{it}(\cdot)$ and $\overline{q}_{it}(\cdot)$S defined in (A1) and (A2) are well calculated.

A.2 | Full conditional distributions of $\boldsymbol{\mu_s}$, $\boldsymbol{\alpha_s}$, and $\boldsymbol{\psi_s}$

$$[\mu_s \mid \ \cdot\ ] \sim N\Big[\mu_s^*, \sigma_{\mu s}^*\Big], \quad [\boldsymbol{\alpha}_s \mid \ \cdot\ ] \sim N\Big[\boldsymbol{\alpha}_s^*, \Sigma_{\alpha s}^*\Big], \quad \Big[\Psi_s^{-1} \mid \ \cdot\Big] \sim \text{Gamma}\Big[\alpha_{\Psi s}^*, \beta_{\Psi s}^*\Big] \quad \text{(A3)}$$

In the above equation, $\boldsymbol{\alpha}_{\Psi s}^* = (n_s + p + \sum_{j=1}^{q} M_j) \Big/ 2 + \widetilde{\alpha}_{s0}$, $\sigma_{\mu s}^* = (n_s \Psi_s^{-1} + \sigma_{\mu s0}^{-1})^{-1}$, and

$$\beta_{\Psi s}^* = \widetilde{\beta}_{s0} + \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} \mathrm{I}(Z_{it} = s)\left(y_{it} - \mu_s - \boldsymbol{\alpha}_s'\mathbf{c}_{it} - \sum_{j=1}^{q} \boldsymbol{\beta}_{sj}'\mathbf{h}_{itj}\right)^2 + \sum_{j=1}^{q} \frac{\left\|\boldsymbol{\beta}_{sj}\right\|_{\mathrm{G}_{sj}}^2}{\tau_{\boldsymbol{\beta} sj}^2} + \sum_{h=1}^{p} + \frac{\mid \alpha_{sh} \mid^2}{\tau_{\alpha sh}^2}\right],$$

$$\Sigma_{\alpha s}^* = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{c}_{it}\mathbf{c}_{it}'\Psi_s^{-1}\mathrm{I}(Z_{it} = s) + \mathbf{D}_{\alpha s}^{-1}\right)^{-1}, \quad \mathbf{D}_{\boldsymbol{\alpha} s} = \mathrm{diag}(\tau_{\boldsymbol{\alpha} s1}^2, \ldots, \tau_{\boldsymbol{\alpha} sp}^2),$$

$$\mu_s^* = \sigma_{\mu s}^*\left[\Psi_s^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathrm{I}(Z_{it} = s)\left(y_{it} - \boldsymbol{\alpha}_s'\mathbf{c}_{it} - \sum_{j=1}^{q} \boldsymbol{\beta}_{sj}'\mathbf{h}_{itj}\right) + \sigma_{\mu s0}^{-1}\mu_{s0}\right],$$

$$\boldsymbol{\alpha}_s = \Sigma_s^*\left[\Psi_s^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathrm{I}(Z_{it} = s)\mathbf{c}_{it}\left(y_{it} - \mu_s - \sum_{j=1}^{q} \boldsymbol{\beta}_{sj}'\mathbf{h}_{itj}\right) + \Sigma_{\alpha s0}^{-1}\boldsymbol{\alpha}_{s0}\right].$$

A.3 | Full conditional distributions of $\boldsymbol{\beta_{sj}}$

$$[\boldsymbol{\beta}_{sj} \mid \ \cdot\ ] \sim N\Big[\boldsymbol{\beta}_{sj}^*, \Sigma_{sj}^*\Big] \quad \mathrm{I}(\mathbf{1}_{n_s}'\mathbf{H}_{sj}\boldsymbol{\beta}_{sj} = 0) \quad \text{(A4)}$$

In the above equation, $\mathit{\Sigma}_{sj}^{*} = \mathit{\Psi}_{s}(\mathbf{H}'_{sj}\mathbf{H}_{sj} + \tau_{\boldsymbol{\beta}sj}^{-1}\mathbf{G}_{sj})^{-1}$, $\boldsymbol{\beta}_{sj}^{*} = \mathit{\Psi}_{s}^{-1}\Sigma_{sj}^{*}\mathbf{H}'_{sj}\mathbf{y}_{s}^{*}$, and $\mathbf{y}_{s}^{*} = \{y_{its}^{*}\}$ is an $n_s \times 1$ vector with

$$y_{its}^{*} = y_{it} - \mu_{s} - \boldsymbol{\alpha}'_{s}\mathbf{c}_{it} - \sum_{l \neq j, l = 1}^{q} \boldsymbol{\beta}'_{sl}\mathbf{h}_{itl}, \qquad \text{for} \quad Z_{it} = s\,.$$

A.4 | Full conditional distributions of π_s, ζ_{us}, and $\widetilde{\boldsymbol{\alpha}}$

$$\begin{aligned} p(\pi_s \mid \cdot\,) &\propto \exp\left\{\sum_{u=s}^{S}\sum_{i=1}^{n}\log(p_{i10u}) \times \mathrm{I}(Z_{i1} = u) - \frac{(\pi_s - \pi_{s0})^2}{2\sigma_{\pi0}^2}\right\} \\ p(\zeta_{us} \mid \cdot\,) &\propto \exp\left\{\sum_{\nu=s}^{S}\sum_{i=1}^{n}\sum_{t=2}^{T}\log(p_{itu\nu}) \times \mathrm{I}(Z_{it} = \nu, Z_{i,t-1} = u) - \frac{(\zeta_{us} - \zeta_{us0})^2}{2\sigma_{\zeta us0}^2}\right\} \\ p(\widetilde{\boldsymbol{\alpha}} \mid \cdot\,) &\propto \exp\left\{\sum_{i=1}^{n}\sum_{t=2}^{T}\log(p_{itus}) \times \mathrm{I}(Z_{it} = s, Z_{i,t-1} = u) - \frac{1}{2}(\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}_0)'\widetilde{\mathbf{D}}_{\alpha}^{-1}(\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}_0)\right\} \end{aligned} \tag{A5}$$

In the above equation, $\widetilde{\mathbf{D}}_{\boldsymbol{\alpha}} = \sigma^2\mathrm{diag}(\widetilde{\tau}_{\boldsymbol{\alpha}1}^2, \ldots, \widetilde{\tau}_{\boldsymbol{\alpha}p}^2)$, and p_{itu0} and p_{itus} can be calculated on the basis of (9).

A.5 | Full conditional distributions of $\widetilde{\boldsymbol{\beta}}_j$

$$p(\widetilde{\boldsymbol{\beta}}_j \mid \cdot\,) \propto \exp\left\{\sum_{i=1}^{n}\sum_{t=2}^{T}\log(p_{itus}) \times \mathrm{I}(Z_{it} = s, Z_{i,t-1} = u) - \frac{1}{2}\left(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\beta}_{j0}\right)'\widetilde{\mathbf{D}}_{\beta j}^{-1}\left(\widetilde{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_{j0}\right)\right\}$$

(A6)

The above equation is with the constraint $\mathbf{1}'_{n(T-1)}\mathbf{H}_j\widetilde{\boldsymbol{\beta}}_j = 0$, where $\widetilde{\mathbf{D}}_{\boldsymbol{\beta}j} = \sigma^2\widetilde{\tau}_{\boldsymbol{\beta}j}^2\widetilde{\mathbf{G}}_j^{-1}$, and p_{itus} can be calculated on the basis of (9).

Notably, the full conditional distributions in (A5) and (A6) are not familiar probability distributions. Therefore, the Metropolis-Hastings algorithm is used to sample from them. Besides, the full conditional distributions in (A4) and (A6) involve constraints, and the procedure for sampling from them can be found in the work of Song and Lu.[18]

REFERENCES

1. Bartolucci F, Farcomeni A. A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. J Am Stat Assoc. 2009;104:816–831.
2. Chow SM, Grimm KJ, Filteau G, Dolan CV, McArdle JJ. Regime-switching bivariate dual change score model. Multivar Behav Res. 2013;48:463–502.

3. Vermunt JK, Langeheine R, Bockenholt U. Discrete-time discrete-state Latent Markov models with time-constant and time-varying covariates. J Educ Behav Stat. 1999;24:179–207.
4. Schmittmann VD, Dolan CV, van der Maas HL, Neale MC. Discrete latent Markov models for normally distributed response data. Multivar Behav Res. 2005;40:461–488.
5. Scott SL, James GM, Sugar CA. Hidden Markov models for longitudinal comparisons. J Am Stat Assoc. 2005;100:359–369.
6. Bartolucci F, Farcomeni A, Pennoni F. Latent Markov Models for Longitudinal Data. Boca Raton, FL: Chapman & Hall/CRC; 2012.
7. Yau C, Papaspiliopoulos O, Roberts GO, Holmes CC. Bayesian nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. J Royal Stat Soc: Ser B (Stat Methodol). 2011;73:37–57.
8. Song X, Kang K, Ouyang M, Jiang X, Cai J. Bayesian analysis of semiparametric hidden Markov models with latent variables. Struct Equ Model: Multidiscip J. 2018;25:1–20.
9. Choi H, Fermin D, Nesvizhskii AI, Ghosh D, Qin ZS. Sparsely correlated hidden Markov models with application to genome-wide location studies. Bioinformatics. 2013;29:533–541. [PubMed: 23325620]
10. Städler N, Mukheijee S. Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. Ann Appl Stat. 2013;7:2157–2179.
11. Guo R, Zhu H, Chow SM, Ibrahim JG. Bayesian lasso for semiparametric structural equation models. Biometrics. 2012;68:567–577. [PubMed: 22376150]
12. Feng XN, Wang GC, Wang YF, Song XY. Structure detection of semiparametric structural equation models with Bayesian adaptive group lasso. Statist Med. 2015;34:1527–1547.
13. Kang K, Cai J, Song X, Zhu H. Bayesian hidden Markov models for delineating the pathology of Alzheimer's disease. Stat Methods Med Res. 2018. Online first.
14. Agresti A. Categorical Data Analysis. Hoboken, NJ: John Wiley & Sons; 2002.
15. Song X, Xia Y, Zhu H. Hidden Markov latent variable models with multivariate longitudinal data. Biometrics. 2017;73:313–323. [PubMed: 27148857]
16. Hastie T, Tibshirani R, Friedman JH. Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd ed. New York, NY: Springer; 2009.
17. Panagiotelis A, Smith M. Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. J Econom. 2008;143:291–316.
18. Song XY, Lu ZH. Semiparametric latent variable models with Bayesian P-splines. J Comput Graph Stat. 2010;19:590–608.
19. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc: Ser B (Stat Methodol). 1996;58:267–288.
20. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J Royal Stat Soc: Ser B (Stat Methodol). 2006; 68:49–67.
21. Kyung M, Gill J, Ghosh M, Casella G. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. 2010;5:369–411.
22. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
23. Wang H, Li G, Tsai CL. Regression coefficient and autoregressive order shrinkage and selection via the lasso. J Royal Stat Soc: Ser B (Stat Methodol). 2007;69:63–78.
24. Wang H, Leng C. A note on adaptive group lasso. Comput Stat Data Anal. 2008;52:5277–5286.
25. Bühlmann P, Van De Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. New York, NY: Springer Science and Business Media; 2011.
26. Cappé O, Moulines E, Rydén T. Inference in Hidden Markov Models. New York, NY: Springer; 2005.
27. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machine. J Chem Phys. 1953;21:1087–1092.
28. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57:97–109.

29. Li J, Wang Z, Li R, Wu R. Bayesian group lasso for nonparametric varying coefficient models with application to functional genome-wide association studies. Ann Appl Stat. 2015;9:640–664. [PubMed: 26478762]

30. Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. Bayesian Anal. 2006;1:651–673.

31. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J Royal Stat Soc: Ser B (Stat Methodol). 2002;64:583–639.

32. Gelman A, Roberts GO, Gilks WR. Efficient Metropolis jumping rules In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, eds. Bayesian Statistics. Vol. 5 Oxford, UK: Oxford University Press; 1996:599–607.

33. Kantarci K, Gunter JL, Tosakulwong N, et al. Focal hemosiderin deposits and I^{2}-amyloid load in the ADNI cohort. Alzheimer's Dement. 2013;9:S116–S123. [PubMed: 23375568]

34. Kesslak JP, Nalcioglu O, Cotman CW. Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer's disease. Neurology. 1991;41:51. [PubMed: 1985296]

35. Jack CR, Petersen RC, O'Brien PC, Tangalos EG. MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. Neurology. 1992;42:183. [PubMed: 1734300]

36. Dickerson BC, Wolk D. Biomarker-based prediction of progression in MCI: comparison of AD-signature and hippocampal volume with spinal fluid amyloid-β and tau. Front Aging Neurosci. 2013;5:55. [PubMed: 24130528]

37. Gao S, Hendrie HC, Hall KS, Hui S. The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta-analysis. Arch Gen Psychiatry. 1998;55:809–815. [PubMed: 9736007]

38. Lindsay J, Laurin D, Verreault R, et al. Risk factors for Alzheimer's disease: a prospective analysis from the Canadian Study of Health and Aging. Am J Epidemiol. 2002;156:445–453. [PubMed: 12196314]

39. Bruandet A, Richard F, Bombois S, et al. Cognitive decline and survival in Alzheimer's disease according to education level. Dement Geriatr Cogn Disord. 2008;25:74–80. [PubMed: 18042993]

40. Stern Y, Albert S, Tang MX, Tsai WY. Rate of memory decline in AD is related to education and occupation. Neurology. 1999;53:1942. [PubMed: 10599762]

41. Vina J, Lloret A. Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-β peptide. J Alzheimer's Dis. 2010;20:S527–S533. [PubMed: 20442496]

42. Heun R, Kockler M. Gender differences in the cognitive impairment in Alzheimer's disease. Arch Women's Ment Health. 2002;4: 129–137.

43. Mazure CM, Swendsen J. Sex differences in Alzheimer's disease and other dementias. Lancet Neurol. 2016;15:451–452. [PubMed: 26987699]

44. Lee E, Zhu H, Kong D, Wang Y, Giovanello KS, Ibrahim JG. BFLCRM: a Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease. Ann Appl Stat. 2015;9:2153–2178. [PubMed: 26900412]

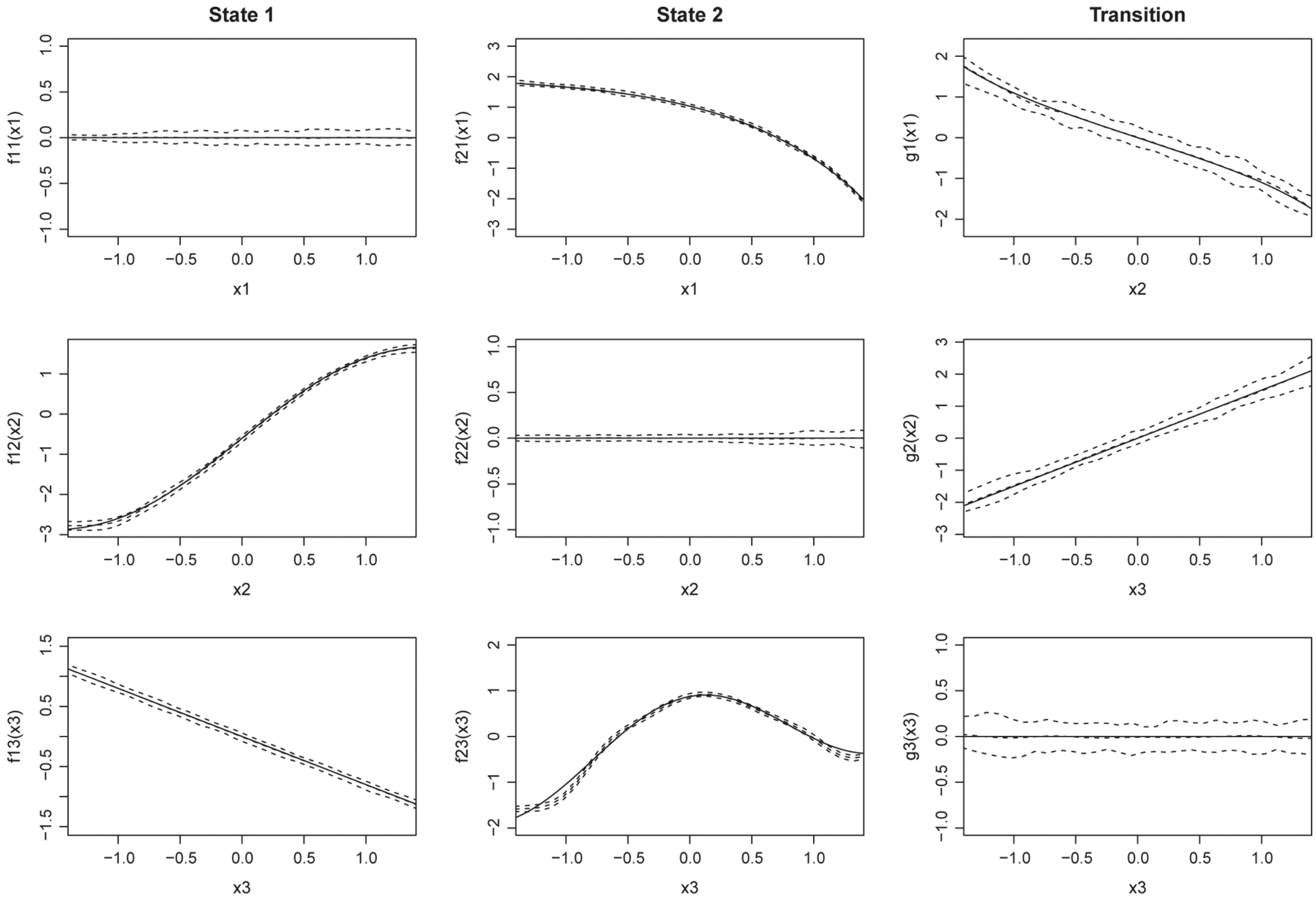

FIGURE 1.

Estimates of the unknown smooth functions in the simulation study. The solid curves represent the true curves, and the dashed curves represent the estimated posterior means and the 2.5% and 97.5% pointwise quantiles on the basis of 100 replications

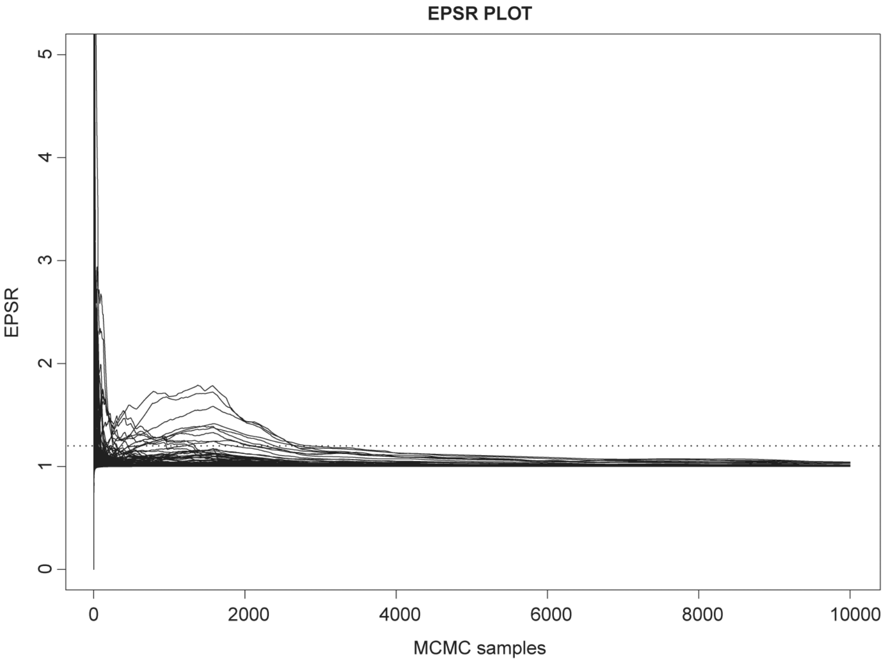

FIGURE 2.

Plot of estimated potential scale reduction (EPSR) values for the parameters in the ADNI-1 (Alzheimer's Disease Neuroimaging Initiative) data analysis. The horizontal dotted line is for EPSR = 1.2. MCMC, Markov chain Monte Carlo

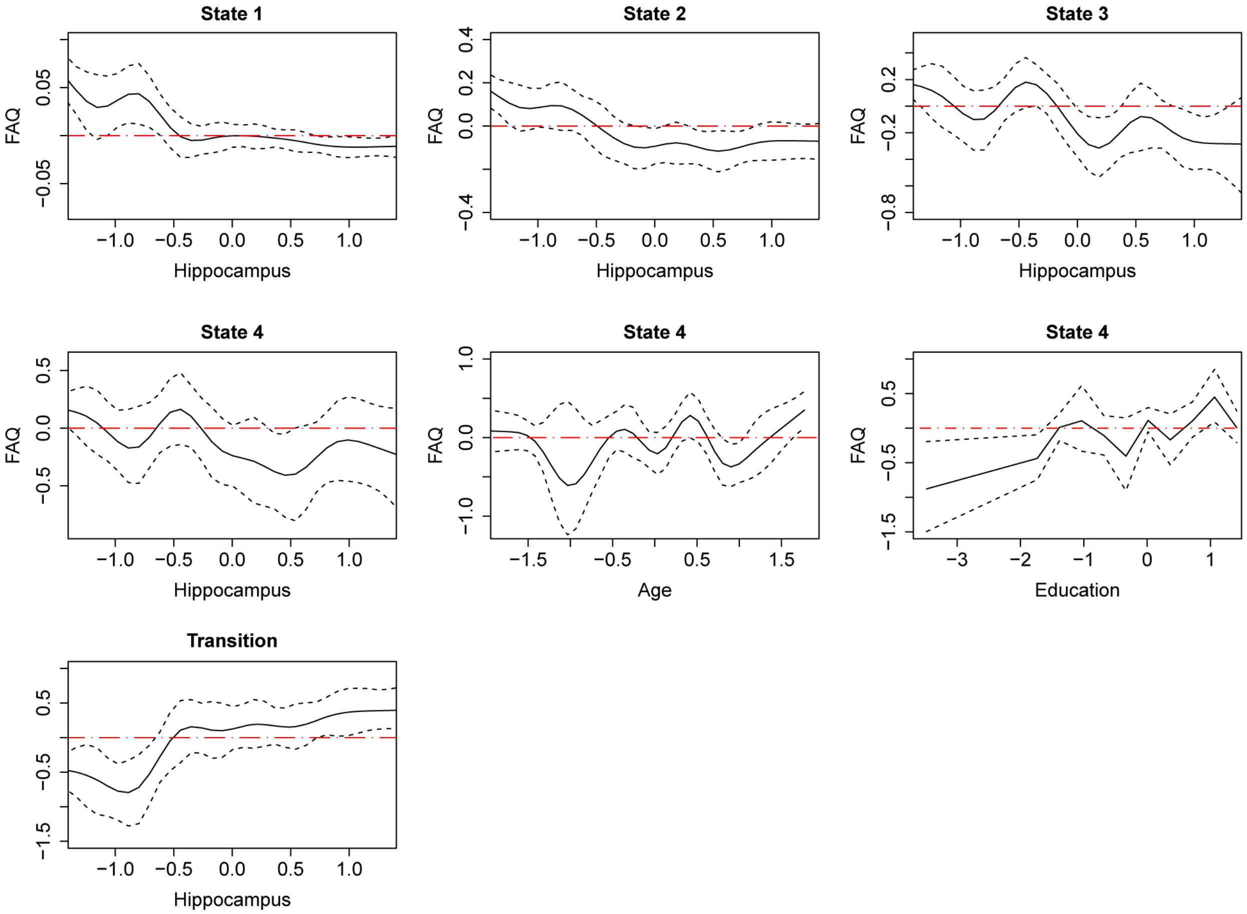

FIGURE 3.

ADNI-1 (Alzheimer’s Disease Neuroimaging Initiative) data analysis results: the estimates of significant unknown smooth functions at the corresponding states. The solid curves represent the pointwise mean curves, and the dashed curves represent the 2.5% and 97.5% pointwise quantiles. Line $y = 0$ is denoted in each picture by a red dot-dash line to illustrate the range of significant effects for each risk factor. FAQ, Functional Assessment Questionnaire [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1

Bayesian estimates of the parameters in the simulation study

Parameters in the Conditional Regression Model							
State 1				State 2			
Par	**True**	**Est**	**RMSE**	**Par**	**True**	**Est**	**RMSE**
μ_1	−1.0	−0.969	0.041	μ_2	1.0	1.006	0.033
a_{11}	0.0	−0.000	0.025	a_{21}	−0.5	−0.499	0.015
a_{12}	0.5	0.501	0.023	a_{22}	0.0	0.001	0.015
ψ_1	0.36	0.392	0.034	ψ_2	0.16	0.191	0.032

Parameters in the Probability Transition Model							
Par	**True**	**Est**	**RMSE**	**Par**	**True**	**Est**	**RMSE**
$\tilde{\alpha}_1$	−1.0	−0.985	0.080	$\tilde{\alpha}_2$	0.0	−0.000	0.055
π_1	0.5	0.528	0.036	π_2	0.5	0.472	0.036
ζ_{11}	0.5	0.501	0.152	ζ_{21}	0.5	0.504	0.152

Abbreviation: RMSE, root-mean-square error.

TABLE 2

Summary of deviance information criterion (DIC) values in the simulation study

Competing Model	DIC (mean)	DIC (std)	No. of Selections
M_1	12 018	79	0
M_2	10 912	92	0
M_3	10 124	461	0
M_4	8988	128	100
M_5	10 052	158	0

Note: No. of selections represents the number of times that the DIC value of M_S ($s = 1, \ldots, 5$) is the smanest among all competing models in 100 replications.

TABLE 3

Summary of deviance information criterion (DIC) values in the analysis of the Alzheimer's Disease Neuroimaging Initiative data set

Competing Model	$\overline{D(\theta)}$	$\boldsymbol{p_D}$	**DIC**
M_1	6294	35	6329
M_2	1434	69	1503
M_3	1016	97	1113
M_4	972	126	1098

TABLE 4

Estimation results in the ADNI-1 (Alzheimer's Disease Neuroimaging Initiative) data analysis: parametric part

Parameters in the Conditional Regression Model

State 1			State 2			State 3			State 4		
Par	**Est**	**SE**	**Par**	**Est**	**SE**	**Par**	**Est**	**SE**	**Par**	**Est**	**SE**
μ_1	−0.608	0.005	μ_2	−0.200	0.032	μ_3	0.948	0.075	μ_4	2.466	0.127
α_{11}	0.000	0.005	α_{21}	0.059	0.040	α_{31}	0.113	0.082	α_{41}	0.256	0.151
α_{12}	0.015	0.013	α_{22}	0.012	0.040	α_{32}	0.068	0.086	α_{42}	0.120	0.143
α_{13}	0.003	0.005	α_{23}	0.019	0.031	α_{33}	−0.303	0.107	α_{43}	−0.427	0.157
α_{14}	0.003	0.005	α_{24}	0.008	0.030	α_{34}	−0.047	0.073	α_{44}	−0.115	0.143
ψ_1	0.009	0.000	ψ_2	0.073	0.008	ψ_3	0.173	0.020	ψ_4	0.437	0.052

Parameters in the Transition Model

Par	**Est**	**SE**	**Par**	**Est**	**SE**	**Par**	**Est**	**SE**	**Par**	**Est**	**SE**
$\tilde{\alpha}_1$	−0.386	0.174	$\tilde{\alpha}_2$	−0.821	0.253	$\tilde{\alpha}_3$	0.012	0.078	$\tilde{\alpha}_4$	−0.150	0.132
π_1	0.592	0.022	π_2	0.198	0.022	π_3	0.149	0.018	π_4	0.060	0.014
ζ_{11}	2.513	0.165	ζ_{21}	−1.459	0.246	ζ_{31}	−3.278	0.451	ζ_{41}	−3.343	0.500
ζ_{12}	2.395	0.418	ζ_{22}	1.498	0.253	ζ_{32}	−1.674	0.331	ζ_{42}	−3.320	0.498
ζ_{13}	1.405	0.740	ζ_{23}	2.840	0.447	ζ_{33}	1.657	0.279	ζ_{43}	−2.017	0.426

TABLE 5

Estimation results of the parametric hidden Markov model in the ADNI-1 (Alzheimer's Disease Neuroimaging Initiative) data analysis

Parameters in the Conditional Regression Model											
	State 1			State 2			State 3			State 4	
Par	Est	SE	Par	Est	SE	Par	Est	SE	Par	Est	SE
β_{11}	−0.022	0.004	β_{21}	−0.122	0.023	β_{31}	−0.155	0.039	β_{41}	−0.127	0.065
β_{12}	−0.006	0.003	β_{22}	−0.008	0.017	β_{32}	0.070	0.034	β_{42}	0.088	0.055
β_{13}	−0.004	0.003	β_{23}	−0.014	0.018	β_{33}	0.030	0.029	β_{43}	0.025	0.051

Parameters in the Transition Model											
Par	Est	SE	Par	Est	SE	Par	Est	SE	Par	Est	SE
$\tilde{\beta}_1$	0.351	0.042	$\tilde{\beta}_2$	−0.033	0.034	$\tilde{\beta}_3$	0.004	0.023			