Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Bayesian Heterogeneous Hidden Markov Models with an Unknown Number of States

Yudan Zou, Yiqi Lin & Xinyuan Song

To cite this article: Yudan Zou, Yiqi Lin & Xinyuan Song (10 Aug 2023): Bayesian Heterogeneous Hidden Markov Models with an Unknown Number of States, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2023.2231055

To link to this article: https://doi.org/10.1080/10618600.2023.2231055

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 00, NO. 0, 1–10
https://doi.org/10.1080/10618600.2023.2231055

Bayesian Heterogeneous Hidden Markov Models with an Unknown Number of States

Yudan Zou, Yiqi Lin, and Xinyuan Song

Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong

ABSTRACT

Hidden Markov models (HMMs) are valuable tools for analyzing longitudinal data due to their capability to describe dynamic heterogeneity. Conventional HMMs typically assume that the number of hidden states (i.e., the order of HMMs) is known or predetermined through criterion-based methods. However, prior knowledge about the order is often unavailable, and a pairwise comparison using criterion-based methods becomes increasingly tedious and computationally demanding when the model space enlarges. A few studies have considered simultaneously performing order selection and parameter estimation under the frequentist framework. Still, they focused only on homogeneous HMMs and thus cannot accommodate situations where potential covariates affect the between-state transition. This study proposes a Bayesian double-penalized (BDP) procedure to conduct a simultaneous order selection and parameter estimation for heterogeneous HMMs. We develop a novel Markov chain Monte Carlo algorithm coupled with an efficient adjust-bound reversible jump scheme to address the challenges in updating the order. Simulation studies show that the proposed BDP procedure considerably outperforms the commonly used criterion-based methods. An application to the Alzheimer's Disease Neuroimaging Initiative study further confirms the utility of the proposed method. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2022
Accepted June 2023

KEYWORDS
Bayesian method; Double penalization; Dynamic heterogeneity; Longitudinal data; Order selection

1. Introduction

Hidden Markov Models (HMMs) have broad applications in medical, behavioral, social, and psychological sciences, wherein heterogeneous longitudinal data are frequently collected and analyzed. HMMs consist of two parts: a transition model to characterize the dynamic transition process between hidden states and a conditional regression (emission) model to examine state-specific covariate effects on the response of interest.

Conventional HMMs typically assume that the number of hidden states (i.e., order of HMM) is known or predetermined through criterion-based methods, such as the Akaike's information criterion (AIC, Akaike 1974) and Bayesian information criterion (BIC, Schwarz 1978). However, despite their successful applications in many substantive studies (e.g., Celeux and Durand 2008; Ip et al. 2013; Song et al. 2017), these criterion-based methods conduct pairwise comparisons among candidate models, which could become increasingly tedious and computationally intensive when the model space is ample. Moreover, these procedures perform estimation in two stages: choosing the order in the first stage and estimating the parameter of the selected model in the second stage, and thus may not be as effective as single-stage approaches.

Penalization methods are valuable alternatives to their criterion-based counterparts in estimating HMMs with unknown order. Two types of over-fitting exist in performing estimation for HMMs. The first type arises when some hidden states are almost empty and thus leading to near-zero mixing probabilities. The second type appears when two or more states have similar emission densities resulting in nearly identical parameter values. Chen and Khalili (2008) pointed out the necessity of preventing the second type of overfitting induced by similar-density components in finite mixture models and suggested a double penalization procedure to avoid the two types of overfitting simultaneously. Ye et al. (2019) extended their method to a finite mixture of varying coefficient models. Manole and Khalili (2021) developed the Group-Sort-Fuse (GSF) procedure for order selection and parameter estimation in multidimensional finite mixture models. For HMMs, Mackay (2002) proposed a single penalization on small state proportions and obtained a consistent order estimate of HMMs. Hung et al. (2013) introduced the double penalized method to non-regression Gaussian HMMs. Zhou et al. (2020) considered continuous-time HMMs and proposed a modified penalized maximum likelihood estimation approach. Lin and Song (2022) extended the GSF procedure and adapted the double penalization idea into regression-based HMMs. Apart from the frequentist penalization methods, Liu and Song (2020) also developed a Bayesian approach by regarding the order of HMMs as a random variable and updating it with other parameters using the reversible jump Markov chain Monte Carlo (MCMC) algorithm (Green 1995). Nevertheless, the available methods focused mainly on finite mixture models or homogeneous HMMs. Moreover, all the existing double

CONTACT Xinyuan Song *xysong@sta.cuhk.edu.hk* Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

penalized procedures are developed under the frequentist framework. However, unlike its frequentist counterpart, the Bayesian approach typically converts the penalization problem to introducing appropriate priors to relevant parameters and updates the penalty through posterior, making the penalization data-driven and easy to implement. Unfortunately, Bayesian double penalization procedures have never been considered in the literature.

This study aims to fill the gap and proposes a novel Bayesian double penalized (BDP) procedure for the simultaneous order selection and parameter estimation of heterogeneous HMMs. The procedure includes two penalties. The first is a lower bound imposed on the summation of mixing proportions to prevent states with near-zero initial probabilities. The second is a least absolute shrinkage and selection operator (lasso)-type penalty introduced to the distance between regression coefficients to avoid states with nearly identical parameters. We develop a hybrid MCMC algorithm that integrates the data augmentation, Gibbs sampler, forward filtering backward sampling (FFBS, Baum et al. 1970), and the Metropolis-Hastings (MH) algorithm. In particular, we offer an efficient adjust-bound reversible jump (ABRJ) sampling scheme to address the challenges of updating the order in implementing the MCMC algorithm. Simulation studies in Section 5 demonstrate that the proposed BDP procedure considerably outperforms the commonly used AIC and BIC in order selection accuracy and two existing one-stage methods in state allocation accuracy. In addition, by setting a sizable upper bound of the order (e.g., 200), the proposed method allows sufficient flexibility in estimating the order of HMMs and thus can accommodate the case where many states exist. Last but not least, the BDP procedure accomplishes order selection and parameter estimation in a single stage. By contrast, criterion-based approaches perform pairwise comparison and parameter estimation on a two-stage basis, and the related computational burden dramatically increases when the candidate model space enlarges. Therefore, the proposed BDP procedure is also superior to the criterion-based methods in terms of computational efficiency.

The rest of this article is organized as follows. Section 2 describes the model and related identifiability issues. Sections 3 and 4 present the BDP procedure and specific sampling schemes. Section 5 evaluates the empirical performance of the proposed method through simulation studies, and Section 6 reports an application to the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Section 7 concludes the article. The technical details are provided in the supplementary material.

2. Model

Let $\boldsymbol{Y} = (\boldsymbol{y_1}', \ldots, \boldsymbol{y_n}')'$, where $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{iT})'$, and y_{it} is the response of subject i at time t; $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_n')'$, where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{iT}')'$, and $\boldsymbol{x}_{it}$ is the covariate vector of subject i at time t; $\boldsymbol{D} = (\boldsymbol{D}_1', \ldots, \boldsymbol{D}_n')'$, where $\boldsymbol{D}_i = (\boldsymbol{d}_{i1}', \ldots, \boldsymbol{d}_{iT}')'$, and $\boldsymbol{d}_{it}$ is another covariate vector of subject i at time t, and the elements of $\boldsymbol{d}_{it}$ can be distinct or overlapped with those of $\boldsymbol{x}_{it}$; $\boldsymbol{Z} = (\boldsymbol{Z_1}', \ldots, \boldsymbol{Z_n}')'$, where $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{iT})'$, and Z_{it} is the hidden state of subject i at occasion t, which follows a first-order Markov chain and takes the values of $\{1, \ldots, K\}$. Given the hidden state Z_{it}, y_{it} is assumed to be independent for all i and t, and is formulated through the conditional regression model as follows:

$$[y_{it}|Z_{it} = s] = \boldsymbol{\beta}_s'\boldsymbol{x}_{it} + \delta_{it}, \tag{1}$$

where $\boldsymbol{x}_{it} = (1, x_{it1}, \ldots, x_{it,p-1})'$ be the $p \times 1$ vector of covariates, $\boldsymbol{\beta}_s$ is the $p \times 1$ vector of state-specific regression coefficients, δ_{it} is the residual term independent of $\boldsymbol{x}_{it}$, and $[\delta_{it}|Z_{it} = s] \sim N(0, \psi_s)$.

Given that hidden states typically have a natural ranking and real meanings in most practical situations, we assume that hidden states $\{1, \ldots, K\}$ are ordered. The hidden transition process is then formulated by $Z_{i1} \sim$ multinomial$(\pi_1, \ldots, \pi_K)$ such that $0 \le \pi_s \le 1$ and $\sum_{s=1}^{K} \pi_s = 1$, and a continuation-ratio logit model (Agresti 2003) as follows: for $t = 2, \ldots, T, s = 1, \ldots K-1, u = 1, \ldots, K$:

$$\log\left(\frac{P_{itus}}{P_{itu,s+1} + \cdots + P_{ituK}}\right) = \zeta_{us} + \boldsymbol{\alpha}'\boldsymbol{d}_{it}, \tag{2}$$

where $P_{itus} = P(Z_{it} = s|Z_{i,t-1} = u)$, ζ_{us} is a transition-specific intercept, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'$ is a $q \times 1$ vector of regression coefficients. Let $\vartheta_{itus} = P(Z_{it} = s|Z_{it} \ge s, Z_{i,t-1} = u)$. Then, we can easily check that $\log\left(\frac{P_{itus}}{P_{itu,s+1}+\cdots+P_{ituK}}\right) = \log\left(\frac{P(Z_{it}=s|Z_{i,t-1}=u)}{P(Z_{it}>s|Z_{i,t-1}=u)}\right) =$ logit(ϑ_{itus}), which is the log conditional odds of transitioning to the sth state instead of a higher state given $Z_{i,t-1} = u$. Therefore, the transition model (2) can be equivalently rewritten as logit$(\vartheta_{itus}) = \zeta_{us} + \boldsymbol{\alpha}'\boldsymbol{d}_{it}$.

Equations (1) and (2) define a heterogeneous HMM, under which some time-variant or baseline covariates affect the between-state transition. Let $\boldsymbol{\theta}$ be the vector containing all the regression and variance parameters in the proposed model. Then, the complete-data log-likelihood function of the proposed model is given by

$$\begin{aligned}
\log\ p[\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{Z}|\boldsymbol{\theta}] &= \sum_{i=1}^{n}[\log\ p(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}) + \log\ p(\boldsymbol{Z}_i|\boldsymbol{D}_i, \boldsymbol{\theta})] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T}\log\ p(y_{it}|\boldsymbol{x}_{it}, Z_{it}, \boldsymbol{\theta}) + \sum_{i=1}^{n}\sum_{t=2}^{T}\log\ p(Z_{it}|Z_{i,t-1}, \boldsymbol{d}_{it}, \boldsymbol{\theta}) \\
&\quad + \sum_{i=1}^{n}\log\ p(Z_{i1}|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T}\log\ p(y_{it}|\boldsymbol{x}_{it}, Z_{it}, \boldsymbol{\theta}) + \sum_{i=1}^{n}\sum_{t=2}^{T}\log(P_{itZ_{i,t-1}Z_{it}}) \\
&\quad + \sum_{i=1}^{n}\log(P_{i10Z_{i1}})
\end{aligned} \tag{3}$$

where

$$\begin{aligned}
&P_{i10s} = \pi_s, \quad s = 1, \ldots, K, \\
&P_{itu1} = \frac{\exp(a_{itu1})}{1 + \exp(a_{itu1})}, \qquad P_{ituK} = \prod_{j=1}^{K-1}\frac{1}{1 + \exp(a_{ituj})}, \\
&P_{itus} = \frac{\exp(a_{itus})}{1 + \exp(a_{itus})}\prod_{j=1}^{s-1}\frac{1}{1 + \exp(a_{ituj})}, \quad s = 2, \ldots, K-1
\end{aligned} \tag{4}$$

with $a_{itus} = \zeta_{us} + \boldsymbol{\alpha}'\boldsymbol{d}_{it}$, for $t = 2, \ldots, T$, $u = 1, \ldots, K$, $s = 1, \ldots, K-1$.

The proposed model is unidentifiable due to the label switching problem. Label switching arises because a random permutation of state labels does not change the likelihood function, which leads to a multi-modal posterior under a symmetric prior distribution. We address the problem by introducing the cluster ordering procedure proposed by Zhou et al. (2020) to sort the multidimensional parameters in the conditional regression model (1), which satisfies the atom property mentioned in Manole and Khalili (2021). We define the cluster ordering procedure in the context of the proposed model as follows.

Definition 2.1. A cluster ordering procedure is a mapping α_β: $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K\} \to \{\boldsymbol{\beta}_{(1)}, \ldots, \boldsymbol{\beta}_{(K)}\}$, such that

$$\begin{aligned} \boldsymbol{\beta}_{(1)} &= \underset{\boldsymbol{\beta}_j : j=1,\ldots,K}{\arg\max} ||\boldsymbol{\beta}_j||_2 \\ \boldsymbol{\beta}_{(k)} &= \underset{\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_{(i)}, i=1,\ldots k-1}{\arg\min} ||\boldsymbol{\beta}_j - \boldsymbol{\beta}_{(k-1)}||_2, \qquad k = 2, \ldots, K, \end{aligned} \tag{5}$$

where $||\cdot||_2$ denotes the L_2 norm.

The above cluster ordering procedure incorporates the ideas from Zhou et al. (2020). The difference is that they number the parameter with the smallest L_2 norm as state one while we take the largest as the first. This procedure guarantees that the state labels are uniquely determined and induces a set of differences $\boldsymbol{\eta}_1 = \boldsymbol{\beta}_{(1)}$, and $\boldsymbol{\eta}_k = \boldsymbol{\beta}_{(k)} - \boldsymbol{\beta}_{(k-1)}$ for $k = 2, \ldots, K$. Recall that the hidden states are assumed ordered; we can then rewrite the conditional regression model (1) as follows:

$$[y_{it}|Z_{it} = s] = \sum_{k=1}^{s} (\boldsymbol{\eta}_k' \boldsymbol{x}_{it}) + \delta_{it}. \tag{6}$$

Hence, by constructing the bijective mapping between $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$, the complete-data log-likelihood function can be formulated as

$$\sum_{i=1}^{n}\sum_{t=1}^{T} \Big(y_{it} - \sum_{k=1}^{s} \boldsymbol{\eta}_k' \boldsymbol{x}_{it}\Big)^2 - \sum_{i=1}^{n}\sum_{t=2}^{T} \log(P_{itus}) - \sum_{i=1}^{n} \log(P_{i10s}), \tag{7}$$

which facilitates the double penalization in the next section.

3. Bayesian Double Penalized (BDP) Procedure

In analyzing HMMs, we must tackle two types of over-fitting: nearly empty states and redundant states. To prevent the first type of overfitting, we impose a lower bound on π_k to ensure the existence of a proper partition and avoid nearly empty states or near-zero mixing proportions. Under the Bayesian framework, the lower-bound penalization is implemented by assigning a symmetric Dirichlet prior distribution to $\boldsymbol{\pi}$, denoted as $(\pi_1, \ldots, \pi_K) \sim \text{Dir}(c_K, \ldots, c_K)$, where the concentration parameter $c_K = c\frac{n}{K}$ and $c > 0$ is a preassigned constant. With such a prior specification, we have the following proposition:

Proposition 3.1. Suppose $Z_i \sim \text{categorical}(\pi_1, \ldots, \pi_K)$, $i = 1, \ldots, n$, with $0 \leq \pi_s \leq 1$, $\sum_{s=1}^{K} \pi_s = 1$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \text{Dir}(c_K, \ldots, c_K)$, where $c_K = c\frac{n}{K}$, $c > 0$ is a constant. Then, we have

$$E(\pi_s|\boldsymbol{Z}) \geq \frac{c}{c+1}\frac{1}{K}, \quad s = 1, \ldots, K. \tag{8}$$

The derivation of Proposition 3.1 is provided in the supplementary material. This proposition ensures that the conditional mean of each element of $\boldsymbol{\pi}$ is lower bounded by $\frac{c}{c+1}\frac{1}{K}$, thereby preventing near-zero probabilities or nearly empty states. The constant c can be determined according to the degree of penalty required for specific problems. The lower bound of $E(\pi_s|\boldsymbol{Z})$ is close to $\frac{1}{K}$ when c increases while it tapers off when c approaches zero.

To address the second type of overfitting, we impose penalization on the norm of the discrepancy between different coefficient vectors. Manole and Khalili (2021) pointed out that the ordering procedure considerably outperforms the naive approach that penalizes all pairwise differences between $\boldsymbol{\beta}_k$ when many hidden states exist. Therefore, instead of naively penalizing all $\binom{K}{2}$ pairwise differences between $\boldsymbol{\beta}_k, k = 1, \ldots, K$, we only penalize the L_2-norm of $K-1$ consecutive differences $\boldsymbol{\eta}_k$. Notably, $\boldsymbol{\eta}_k = \boldsymbol{\beta}_{(k)} - \boldsymbol{\beta}_{(k-1)}$; if $\|\boldsymbol{\eta}_k\|_2 = 0$, then $\boldsymbol{\beta}_{(k)} = \boldsymbol{\beta}_{(k-1)}$, implying that states k and $k-1$ are redundant. Thus, we penalize $\|\boldsymbol{\eta}_k\|_2$ for preventing redundant states. Park and Casella (2008) introduced the Bayesian lasso to achieve shrinkage on regression coefficients by assigning them a conditional Laplace prior. Based on this idea, we modify the Bayesian lasso by introducing the conditional Laplace prior to $||\boldsymbol{\eta}_k||_2$ to achieve shrinkage on the entire vector $\boldsymbol{\eta}_k$ as follows:

$$P(\boldsymbol{\eta}_k|\psi_k) = \frac{\gamma_k}{2\sqrt{\psi_k}} \exp\Big(-\frac{\gamma_k}{\sqrt{\psi_k}}||\boldsymbol{\eta}_k||_2\Big), \quad k = 2, \ldots, K. \tag{9}$$

Then, the proposed model can be formulated through the following hierarchical representation: for $s = 1, \ldots, K$,

$$\begin{aligned} &[y_{it}|Z_{it} = s, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_s, \psi_s] \sim N\left(\sum_{k=1}^{s} \left(\boldsymbol{\eta}_k' \boldsymbol{x}_{it}\right), \psi_s\right), \\ &[\boldsymbol{\eta}_s|\psi_s, \tau_s^2] \sim N(\boldsymbol{0}, \psi_s\tau_s^2\boldsymbol{I}_p), \quad \psi_s^{-1} \overset{\text{ind}}{\sim} \text{Gamma}(a_{\psi s0}, b_{\psi s0}), \\ &\tau_s^2 \overset{\text{ind}}{\sim} \text{Gamma}\Big(\frac{p+1}{2}, \frac{\gamma_s^2}{2}\Big), \quad \gamma_k^2 \overset{\text{ind}}{\sim} \text{Gamma}(a_{\gamma k0}, b_{\gamma k0}), \\ &\quad k = 2, \cdots, K, \end{aligned} \tag{10}$$

where $\boldsymbol{0}$ is the vector of zero elements, $\boldsymbol{I}_p$ is the p-dimensional identify matrix, $a_{\psi s0}$, $b_{\psi s0}$, $a_{\gamma k0}$, and $b_{\gamma k0}$, are hyperparameters whose values are prespecified.

Proposition 3.2. Under the hierarchical model (10), the conditional prior distribution of $\boldsymbol{\eta}_k$ has the form of (9).

The derivation of Proposition 3.2 is provided in Appendix 1 of supplementary material. Notably, (9) modifies the conditional Laplace prior proposed by Park and Casella (2008) and plays a similar role to the Bayesian lasso penalty to achieve shrinkage on $\boldsymbol{\eta}_k$. Moreover, it introduces a state-specific tuning parameter γ_k to each $\|\boldsymbol{\eta}_k\|_2$ and penalizes the L_2 norm of the entire vector

$\boldsymbol{\eta}_k$ rather than its elements. Therefore, the penalty in (9) is adaptive and group-wise, denoted as a modified adaptive group lasso (MAGlasso) penalty. The MAGlasso procedure aims to update the tuning parameters by exploiting the data, thereby automatically imposing large penalties on unimportant coefficients. This target can be naturally achieved by introducing dispersed priors with small hyperparameters $a_{\gamma k0}$ and $b_{\gamma k0}$. The degree of dispersion of the gamma priors determines the magnitudes of penalties imposed on unimportant components. Typically, setting $a_{\gamma k0}$ to a positive integer (e.g., 1) and $b_{\gamma k0}$ to a small value (e.g., 0.1 or 0.01) can induce a dispersed gamma prior. With this prior specification, we can derive the posterior distribution of the tuning parameters, which have the following forms:

$$[\tau_s^{-2}|\cdot] \sim \text{Inverse-Gaussian}\left\{\sqrt{\frac{\gamma_s^2\psi_s}{||\boldsymbol{\eta}_s||_2^2}}, \gamma_s^2\right\}, \tag{11}$$

$$[\gamma_s^2|\cdot] \sim \text{Gamma}\left(a_{\gamma s0} + \frac{p+1}{2}, b_{\gamma s0} + \frac{\tau_s^2}{2}\right). \tag{12}$$

If $||\boldsymbol{\eta}_s||_2$ is significant, τ_s^2 tends to be large based on (11). Then, the corresponding γ_s is dominated by τ_s^2 based on (12). On the contrary, if $||\boldsymbol{\eta}_s||_2$ is insignificant, τ_s^2 tends to be small, and the related γ_s is then dominated by the dispersed prior.

Considering that the Bayesian lasso does not shrink coefficients precisely to zero, we need a criterion to quantify the closeness of $||\boldsymbol{\eta}_s||_2$ to a zero vector. Based on the specification of (9), we can show that $P(\boldsymbol{\eta}_s|\boldsymbol{Y},,\boldsymbol{\theta}) \sim N(\boldsymbol{\eta}_s^*, \boldsymbol{\Sigma}_s^*)$, where $\boldsymbol{\eta}_s^*$ and $\boldsymbol{\Sigma}_s^*$ are provided in Appendix 2 of supplementary material. Therefore, the squared Mahalanobis distance $d_s^2 = (\boldsymbol{\eta}-\boldsymbol{\eta}_s^*)'\boldsymbol{\Sigma}_s^{*^{-1}}(\boldsymbol{\eta}-\boldsymbol{\eta}_s^*) \sim \chi_p^2$ determines a hyper-ellipse density contour centered at $\boldsymbol{\eta}_s^*$. In this study, we adopt the 95% highest posterior credible region (HPCR) criterion (Harper and Hooker 1976) or equivalently the smallest region covering 95% of posterior probability mass. $\boldsymbol{\eta}_s$ is regarded as redundant if its 95% HPCR covers $\mathbf{0}$. Alternatively, we can transform the decision rule to a direct comparison of the squared Mahalanobis distance between $\mathbf{0}$ and $\boldsymbol{\eta}_s^*$ with a critical value of χ_p^2, that is, if $\boldsymbol{\eta}_s^{*'}\boldsymbol{\Sigma}_s^{*^{-1}}\boldsymbol{\eta}_s^* \le \chi_{p,0.05}^2$, then $\boldsymbol{\eta}_s$ is redundant; otherwise, it is significant.

4. Posterior Sampling

4.1. Tuning Parameters and Other Prior Specification

This section introduces how to tune the parameters in the prior distribution to facilitate double penalization and sets a proper prior for the parameters not discussed in Section 3.

As discussed in Section 3, we assign a Dirichlet prior to $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ to prevent empty states as follows:

$$(\pi_1, \ldots, \pi_K) \sim \text{Dir}(c_K, \ldots, c_K), \quad c_K = c\frac{n}{K}, \tag{13}$$

where c is a hyperparameter. The constant c can be determined according to the degree of penalty required for specific problems. Typically, c around 0.5 can effectively prevent near-zero π_s. Alternatively, one can regard c as another tuning parameter and update it in the MCMC algorithm. However, our numerical results show that this data-driving method increases computational complexity but performs similarly to the approach that fixes c in the interval of $(0, 1)$. Moreover, setting a moderate value in $(0, 1)$ can avoid an extreme (too small or too large) penalty on the mixing probabilities. Based on our extensive simulation study, a value around 0.5 performs satisfactorily.

Furthermore, we introduce the conditional Laplace prior (9) to $\|\boldsymbol{\eta}_k\|_2$ to prevent redundant states with almost identical parameter values. The proposed model can then be formulated through the hierarchical representation (10). For the tuning parameter γ_k involved in (9) or (10), we assign the following dispersed gamma prior:

$$\gamma_k^2 \overset{\text{ind}}{\sim} \text{Gamma}(a_{\gamma k0}, b_{\gamma k0}), \quad k = 2, \ldots, K, \tag{14}$$

where $a_{\gamma k0}$ and $b_{\gamma k0}$ are hyperparameters whose values are prespecified to achieve a highly dispersed prior. The degree of dispersion determines the magnitude of the penalty on unimportant regression parameters. In this study, we follow the common practice (Guo et al. 2012; Kang et al. 2019) to set $a_{\gamma k0} = 1$ and $b_{\gamma k0} = 0.01$.

For other parameters involved in the transition model (2), we assign conjugate priors:

$$\zeta_{us} \overset{\text{ind}}{\sim} N(\zeta_{us0}, \sigma_{us0}^2), \quad \boldsymbol{\alpha}_k \overset{\text{ind}}{\sim} N(\boldsymbol{\alpha}_{k0}, \boldsymbol{\Sigma}_{\alpha k0}),\ k = 1, \ldots, q, \tag{15}$$

where $\zeta_{us0}, \sigma_{us0}^2, \boldsymbol{\alpha}_{k0}$, and $\boldsymbol{\Sigma}_{\alpha k0}^2$ are hyperparameters with prespecified values. The common practice is to set ζ_{us0} and the elements of $\boldsymbol{\alpha}_{k0}$ to zero and assign σ_{us0}^2 and the diagonal elements of $\boldsymbol{\Sigma}_{\alpha k0}$ to large values to induce vague priors if the preliminary information about ζ_{us} and $\boldsymbol{\alpha}_k$ is unavailable.

4.2. MCMC Algorithm

Unlike conventional HMMs that prespecify K, this study regards K as another unknown parameter and updates it with other model parameters in $\boldsymbol{\theta}$. The Bayesian estimate of $(\boldsymbol{\theta}, K)$ can be obtained through the mean of the posterior samples drawn from $P(\boldsymbol{\theta}, K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$. However, $P(\boldsymbol{\theta}, K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$ involves unknown hidden states, leading to intractable sampling from $P(\boldsymbol{\theta}, K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$. Using the data augmentation technique, we instead work on $P(\boldsymbol{Z}, \boldsymbol{\theta}, K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$. However, the joint posterior distribution $P(\boldsymbol{Z}, \boldsymbol{\theta}, K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D})$ is still complex. Thus, the Gibbs sampler is employed to iteratively update each component through sampling from its full conditional distribution as follows: (a) update hidden states by sampling $\boldsymbol{Z}$ from $P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{\theta}, K)$, (b) update the model parameters by sampling $\boldsymbol{\theta}$ from $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{Z}, K)$, and (c) update the order K by sampling from $P(K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\theta})$. Owing to the transitioning features of hidden states and nonlinearity of the transition model (2), steps (a) and (b) require MCMC techniques, such as the FFBS and MH algorithms. The full conditional distributions involved in steps (a) and (b) are derived in Appendix 2 of supplementary material. Step (c) is the so-called ABRJ step, which allows K to be updated at each MCMC iteration as follows.

Let $(K_{\min}, K_{\max})$ be the lower and upper bounds of K, and $(K_{\min}^{(0)}, K_{\max}^{(0)})$ and $(K_{\min}^{(j)}, K_{\max}^{(j)})$ be their values at the initial stage and jth iteration of the MCMC algorithm. Typically, we set $K_{\min}^{(0)} = 2$ and $K_{\max}^{(0)}$ to a relatively large positive integer (e.g.,

100 or 200) to allow sufficient flexibility in updating K. At the $(j+1)$th iteration, $K^{(j+1)}$ can remain unchanged, increase, or decrease by 1. To update $K^{(j)}$, we first locate a state s_*, such that $s_* = \text{argmin}_{s=1,\ldots,K^{(j)}} ||\boldsymbol{\eta}_s^{(j)}||_2$. Then, we calculate $d^2_{s_*} = \boldsymbol{\eta}_{s_*}^{(j)'} \boldsymbol{\Sigma}_{s_*}^{(j)^{-1}} \boldsymbol{\eta}_{s_*}^{(j)}$ and compare $d^2_{s_*}$ with $\chi^2_{p,0.05}$. If $d^2_{s_*} \le \chi^2_{p,0.05}$, we regard this component as redundant and update $K^{(j)}$ downward to $K^{(j+1)} = \max(K^{(j)} - 1, K^{(j)}_{\min})$. Meanwhile, we adjust $K^{(j)}_{\max}$ as $K^{(j+1)}_{\max} = \min(K^{(j)}_{\max}, K^{(j)})$. If $d^2_{s_*} > \chi^2_{p,0.05}$, we regard this component as necessary. Then, we jump $K^{(j)}$ upward to $K^{(j+1)} = \min(K^{(j)}+1, K^{(j)}_{\max}-1)$. If $d^2_{s_*} > \chi^2_{p,0.05}$ but $K^{(j)} = K^{(j)}_{\max}-1$, then $K^{(j)}$ remains unchanged, that is, $K^{(j+1)} = K^{(j)}$. Figure 1 shows the strategy of updating K in the ABRJ step. A pseudocode for implementing the MCMC algorithm is given below.

It should be noted that the two penalties in the suggested strategy target potential overfitting difficulties without providing adequate solutions for underfitting concerns. As a result, the initial upper bound should be a relative large number and the effect of order selection occurs when jumping in an upside-down manner.

5. Simulation Study

This section includes three simulations to demonstrate the effectiveness of the proposed algorithm in order selection and parameter estimation under various scenarios. Simulation 1 evaluates the performance of the proposed method in the case of $K = 2$, Simulation 2 assesses estimation performance for HMMs with higher orders, and Simulation 3 focuses on order selection and compares the proposed method with AIC and BIC.

5.1. Simulation 1

This simulation considers a 2-state HMM with $p = 4$ and $q = 1$. Two sample sizes, $(n, T) = (50, 4), (200, 4)$, are considered. In each setting, 100 datasets are generated from the following model:

$$\begin{aligned} [y_{it}|Z_{it} = s] &= \boldsymbol{\beta}_s' \boldsymbol{x}_{it} + \delta_{it}, \\ \text{logit}(\vartheta_{itus}) &= \zeta_{us} + \alpha d_{it}, \end{aligned} \tag{16}$$

where $\boldsymbol{x}_{it} = (1, x_{it1}, x_{it2}, x_{it3})'$, $x_{it1} \overset{\text{ind}}{\sim} N(0,1)$, $x_{it2} \overset{\text{ind}}{\sim} U(-1,1)$, $U(-1,1)$ denotes the uniform distribution in $(-1,1)$, $x_{it3} \overset{\text{ind}}{\sim}$

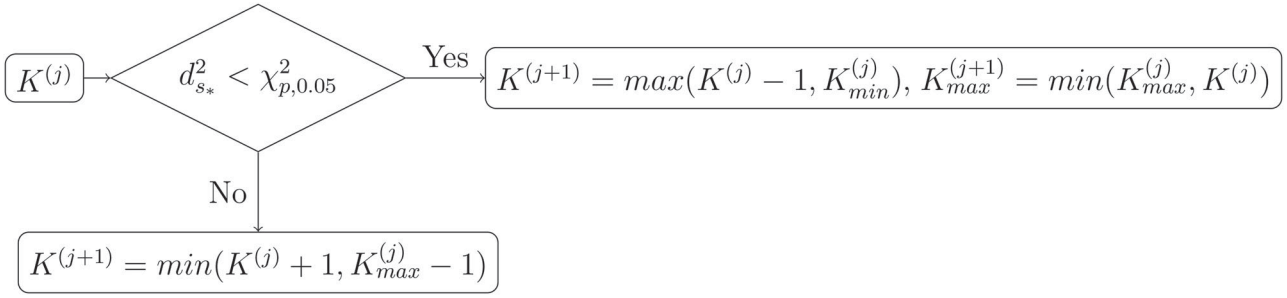

Figure 1. Strategy of updating K in the ABRJ step (c).

Algorithm 1 MCMC algorithm for the estimation of heterogeneous HMMs

Data: $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, J, K^{(0)}_{\min}, K^{(0)}_{\max}$ ▷ J denotes the total number of iterations

$K^{(0)} = K^{(0)}_{\min}$
for $j = 1$ to J **do**
  Update $\boldsymbol{Z}^{(j)}$ by sampling from $P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\theta}^{(j)}, K^{(j)})$ ▷ FFBS algorithm
  Update $\boldsymbol{\theta}^{(j)}$ by sampling from $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$ ▷ see details in Appendix B
  $s_* = \text{argmin}_{s=1,\ldots,K^{(j)}} ||\boldsymbol{\eta}_s^{(j)}||_2$
  $\boldsymbol{\eta}_{s_*}^{(j)} = E(\boldsymbol{\eta}_{s_*}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$ ▷ posterior mean vector
  $\boldsymbol{\Sigma}_{s_*}^{(j)} = \text{var}(\boldsymbol{\eta}_{s_*}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$ ▷ posterior covariance matrix
  $d^2_{s_*} = \boldsymbol{\eta}_{s_*}^{(j)'} \boldsymbol{\Sigma}_{s_*}^{(j)^{-1}} \boldsymbol{\eta}_{s_*}^{(j)}$
  if $d^2_{s_*} < \chi^2_{p,0.05}$ **then**
    $K^{(j+1)} = \max(K^{(j)} - 1, K^{(j)}_{\min})$
    $K^{(j+1)}_{\max} = \min(K^{(j)}, K^{(j)}_{\max})$
  else if $d^2_{s_*} \ge \chi^2_{p,0.05}$ **then**
    $K^{(j+1)} = \min(K^{(j)}_{\max} - 1, K^{(j)} + 1)$
    if $K^{(j)} = K^{(j)}_{\max} - 1$ **then**
      $K^{(j+1)} = K^{(j)}$
    end if
  end if
  $j = j + 1$
end for

Bernoulli(0.6), and $d_{it} \overset{\text{ind}}{\sim} N(0,1)$. The true population values of the parameters are set as follows: $\boldsymbol{\beta}_1 = (2,2,1,1)'$, $\boldsymbol{\beta}_2 = (0,1,2,-1)'$, $\psi_1 = \psi_2 = 0.25$, $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\zeta} = (\zeta_{11}, \zeta_{21}) = (-2,2)'$, and $\alpha = -1$.

The hyperparameters of the prior distributions in (13)–(15) are specified as follows (Prior I): $a_{\psi s0} = 9$, $b_{\psi s0} = 4$, $a_{\gamma k0} = 1$, $b_{\gamma k0} = 0.1$, $c = 0.5$, $\alpha_{k0} = \zeta_{us0} = 0$, and $\sigma^2_{\alpha k0} = \sigma^2_{us0} = 1$. In implementing the MCMC algorithm, we impose a constraint described in Definition 2.1 (i.e., $\boldsymbol{\beta}_{(1)} > \boldsymbol{\beta}_{(2)}$) to each MCMC iteration to avoid label switching. Moreover, we set $K^{(0)}_{\min} = 2$ and $K^{(0)}_{\max} = 200$, which provides an extensive range for K. The algorithm's convergence is checked through the trace plots of the parameters. Figure 2(a) presents the trace plots of three MCMC chains of K starting from different initial values in an arbitrarily selected replication. The three MCMC chains of K mix rapidly and converge to the true value $K_0 = 2$ within a few iterations. Figure S1 of supplementary material presents the trace plots of three MCMC chains for other randomly selected parameters. Both figures indicate a fast convergence of the MCMC algorithm. To be conservative, we collect 10,000 posterior samples, discard the first 3000 iterations as burn-in, and calculate the bias and root mean square error (RMS) between the parameter estimates and their true values based on the remaining 7000 posterior samples corresponding to the selected order. Table 1 presents the estimation result. The bias and RMS are close to zero, and the performance improves when the sample size increases.

To reveal the sensitivity of Bayesian estimates to the prior input, we disturb the hyperparameters as follows (Prior II): $a_{\psi s0} = 13$, $b_{\psi s0} = 6$, $a_{\gamma k0} = 1$, $b_{\gamma k0} = 0.01$, $c = 0.3$, $\alpha_{k0} = \zeta_{us0} = 2$, and $\sigma^2_{\alpha k0} = \sigma^2_{us0} = 100$. Table S1 of Supplementary Material reports the obtained results. The parameter estimates perform similarly to those in Table 1, indicating that the proposed Bayesian estimation is insensitive to the disturbed prior considered.

Furthermore, we check the sensitivity of Bayesian estimation to the misspecification of the distribution of δ_{it} by considering two nonnormal cases: (1) $\delta_{it} \sim U(-1,1)$ and (2) $\delta_{it} \sim 0.4N(1,1) + 0.6N(-1,1)$. We simulate 100 datasets from the proposed model with $n = 200$ and δ_{it} drawn from case (1) or (2). The hyperparameters and other settings are the same as

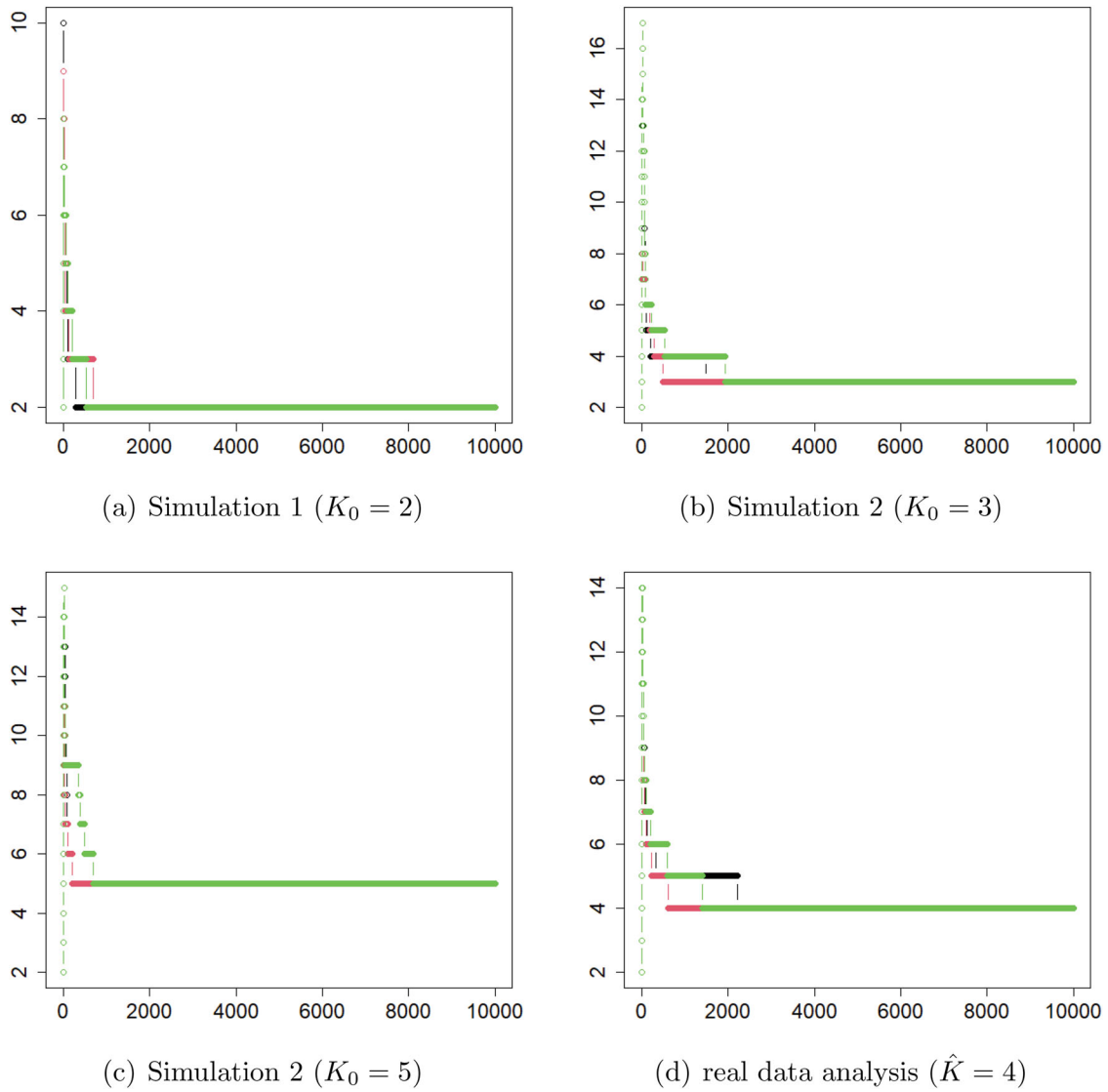

Figure 2. Trace plots of three MCMC chains of K in simulations and the ADNI study.

Table 1. Parameter estimates under Prior I in Simulation 1: $K_0 = 2$.

$n = 50$						$n = 200$					
State 1			State 2			State 1			State 2		
Par	Bias	RMS	Par	Bias	RMS	Par	Bias	RMS	Par	Bias	RMS
Parameters in the conditional regression model											
β_{11}	0.034	0.049	β_{12}	−0.026	0.069	β_{11}	0.007	0.030	β_{12}	−0.006	0.040
β_{21}	−0.002	0.032	β_{22}	−0.001	0.035	β_{21}	−0.003	0.035	β_{22}	0.014	0.029
β_{31}	0.020	0.055	β_{32}	−0.018	0.051	β_{31}	0.001	0.019	β_{32}	−0.017	0.026
β_{41}	−0.020	0.083	β_{42}	0.045	0.070	β_{41}	−0.022	0.050	β_{42}	0.015	0.042
ψ_1	0.020	0.042	ψ_2	0.038	0.058	ψ_1	0.009	0.016	ψ_2	0.008	0.022
Parameters in the transition model											
ζ_{11}	0.030	0.168	ζ_{21}	−0.040	0.155	ζ_{11}	−0.008	0.096	ζ_{21}	−0.028	0.115
π_1	−0.007	0.052	π_2	0.007	0.052	π_1	−0.001	0.020	π_2	0.001	0.020
α_1	0.020	0.107				α_1	0.010	0.085			

in Simulation 1. Table S2 of Supplementary Material presents the estimation results obtained under the two nonnormal cases. Except for the variance of δ_{it} and some parameters involved in the transition model, most parameter estimates are robust to the violation of the normality assumption of δ_{it}. Therefore, the impact of misspecifying the distribution of δ_{it} is mainly on estimating its variance.

5.2. Simulation 2

This simulation examines estimation performance for higher-order models and a model with unordered states. We first consider a 3-state HMM. Covariates $\boldsymbol{x}_{it}$ is the same as in Simulation 1. For simplicity, we set $\boldsymbol{d}_{it} = \boldsymbol{x}_{it}^*$, where $\boldsymbol{x}_{it}^*$ is the subvector of $\boldsymbol{x}_{it}$ excluding 1. Two sample sizes, $(n, T) = (200, 6)$ and $(400, 6)$, are considered. The true population values of the parameters are set as $\boldsymbol{\beta}_1 = (3, 3, 3, 3)'$, $\boldsymbol{\beta}_2 = (0, 1, 2, 2)'$, $\boldsymbol{\beta}_3 = (-4, 2, 1, 1)'$, $\boldsymbol{\psi} = (0.25, 0.25, 0.25)'$, $\boldsymbol{\pi} = (0.3, 0.4, 0.3)'$, $\boldsymbol{\zeta}_1 = (-1, -1, -1)'$, $\boldsymbol{\zeta}_2 = (1, 1, 1)'$, $\boldsymbol{\alpha} = (1, -1, -1)'$. The prior specification and simulation settings are similar to Simulation 1, except that the hyperparameters for $\boldsymbol{\alpha}$ are set as $\boldsymbol{\alpha}_{k0} = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_{\alpha k0} = \boldsymbol{I}_3$. Figure 2(b) presents the trace plots of three MCMC chains of K starting from different initial values in an arbitrarily selected replication. Again, the MCMC chains mix and converge to the true value of $K_0 = 3$ rapidly. The trace plots of other parameters (Figure S2 of supplementary material) suggest that the algorithm converges within 3000 iterations. Therefore, we discard 3000 burn-in and use the remaining 7000 posterior samples to obtain the Bayesian estimates of the parameters. Table S3 of Supplementary Material shows the estimation results based on 100 replications under $(n, T) = (200, 6)$, indicating that the proposed method performs satisfactorily in bias and RMS. The results under $(n, T) = (400, 6)$ are further improved and not reported.

Next, we further increase the order to $K_0 = 5$. We consider a 5-state HMM with covariates $\boldsymbol{x}_{it} = (1, x_{it1}, x_{it2})'$, where $x_{it1} \overset{\text{ind}}{\sim} N(0, 1)$ and $x_{it2} \overset{\text{ind}}{\sim} U(-1, 1)$, and $\boldsymbol{d}_{it} = \boldsymbol{x}_{it}^*$. The other model setup is the same as above. The true population values of unknown parameters are set as $\boldsymbol{\beta}_1 = (5, 5, 5)'$, $\boldsymbol{\beta}_2 = (3, 4, 3)'$, $\boldsymbol{\beta}_3 = (0, 3, 4)'$, $\boldsymbol{\beta}_4 = (-2, 4, 3)'$, $\boldsymbol{\beta}_5 = (-5, 5, 2)'$, $\boldsymbol{\psi} = (0.25, 0.25, 0.25, 0.25, 0.25)'$, $\boldsymbol{\pi} = (0.2, 0.2, 0.2, 0.2, 0.2)'$, $\boldsymbol{\zeta}_1 = (-2, -2, -2, -2, -2)'$, $\boldsymbol{\zeta}_2 = (-1, -1, -1, -1, -1)'$, $\boldsymbol{\zeta}_3 = (1, 1, 1, 1, 1)'$, $\boldsymbol{\zeta}_4 = (2, 2, 2, 2, 2)'$, $\boldsymbol{\alpha} = (2, 1)'$. The prior and simulation settings are similar to those given above. Figure 2(c) presents the trace plots of three MCMC chains of K starting from different initial values in an arbitrarily selected replication, showing that the iterative K quickly converges to its true value $K_0 = 5$. The trace plots of other parameters (Figure S3 of supplementary material) suggest that the MCMC chains mix well within 4000 iterations. Thus, we discard 4000 burn-in and use the remaining 6000 posterior samples to obtain the Bayesian estimates of the parameters. Table S4 of supplementary material presents the estimation results under $(n, T) = (200, 6)$, indicating that the proposed method performs satisfactorily in order selection and parameter estimation when K_0 increases to 5. The results under a larger size of $(n, T) = (400, 6)$ are better and not reported.

To accommodate the scenario when a hidden state may represent a political belief or something unordered, we adapt the transition model (2) to a multinomial logit model and conduct the analysis. Table S5 of supplementary material presents the estimation results. Although the bias and RMS are slightly larger than those in Table 1 due to a more complicated transition model, they still show the satisfactory performance of the proposed method. Moreover, Figure S5 indicates that three MCMC chains of K starting from different initial values mix rapidly and converge to the true value $K_0 = 2$.

5.3. Simulation 3

This simulation assesses the performance of order selection. Considering that no existing methods can simultaneously estimate the order and model parameters of heterogeneous HMMs, we compare the proposed BDP procedure with criterion-based approaches, AIC and BIC, in order selection accuracy.

Table 2 presents the proportions of correct order selections calculated based on 100 replications with $K_0 = \{3, 4, 5\}$ and $n = \{100, 200, 400\}$. The results show that the proposed BDP procedure consistently outperforms AIC and BIC in all the scenarios considered. In general, the performance of the three methods improves when the sample size increases but declines when K_0 increases. In particular, AIC and BIC perform poorly when $K_0 = 5$ regardless of the sample size; their correct selection proportions are below or around 0.5. By contrast, our proposed method performs much better, and its correct

Table 2. Proportions of correct order selections among 100 replications.

		$n = 100$			$n = 200$			$n = 400$		
K_0	$\hat{K}$	AIC	BIC	BDP	AIC	BIC	BDP	AIC	BIC	BDP
3	2	0	0	0.13	0	0	0.10	0	0	0
	3	0.70	0.76	0.87	0.69	0.80	0.82	0.78	0.98	1
	4	0.18	0.22	0	0.25	0.20	0.08	0.20	0.02	0
	5	0.12	0.02	0	0.06	0	0	0.02	0	0
4	2	0	0	0.11	0	0	0.02	0	0	0
	3	0	0	0.16	0	0	0.21	0	0.12	0.09
	4	0.45	0.66	0.73	0.55	0.60	0.77	0.67	0.67	0.85
	5	0.30	0.31	0	0.30	0.25	0	0.21	0.09	0.06
	6	0.25	0.03	0	0.15	0.15	0	0.12	0.12	0
5	3	0	0	0.16	0	0	0.04	0	0	0.13
	4	0.24	0.28	0.23	0.18	0.21	0.15	0.07	0.14	0.06
	5	0.36	0.40	0.61	0.47	0.50	0.77	0.55	0.59	0.81
	6	0.32	0.28	0	0.29	0.26	0.04	0.21	0.10	0
	7	0.08	0.04	0	0.06	0.03	0	0.17	0.17	0

selection proportion attains 0.81 when $n = 400$. This comparison confirms that our procedure achieves higher accuracy than the conventional criterion-based approaches. In addition, unlike the two-stage methods that perform order selection and parameter estimation in two stages, our proposed method accomplishes the two tasks in a single stage. Moreover, our procedure guarantees that sampling only occurs when the states are necessary and retained. Hence, its advantages in computational efficiency over existing ones become increasingly pronounced when the candidate model space enlarges.

Per an anonymous referee's suggestion, we also compare the BDP procedure with two existing one-stage methods offered by Lin and Song (2022) and Liu and Song (2020). Since these two available methods assumed that the between-state transition is homogeneous, the heterogeneity is ignored when applying them to the generated 100 datasets in Simulation 1. In addition, we generate 100 datasets from a homogeneous model by setting $\boldsymbol{\alpha} = \mathbf{0}$ in the transition model (2). Table S6 of supplementary material presents the comparison results, from which we have two findings. First, the three approaches perform comparably in order selection, and our method performs better than Lin's, especially in the heterogeneous case, but slightly worse than Liu's. Second, our method significantly outperforms the other two in estimating hidden states, especially for heterogeneous data. This result is anticipated, given that the other two approaches disregard heterogeneity. Finally, it is worth mentioning that despite the excellent performance of Liu's method in order selection, it performs the poorest in state allocation accuracy, likely due to its constant switching between states and unstable allocation of each individual. Therefore, the comparison results demonstrate the advantages of the proposed method in handling both homogeneous and heterogeneous scenarios.

6. Real Data Analysis

In this section, we applied the proposed method to the dataset extracted from the ADNI study to demonstrate the practical utility of the proposed method. ADNI is a longitudinal multicenter study that began in 2004, collecting various participants' imaging and clinical assessments. More information is referred to the official website: *www.adni-info.org*.

We focused on 616 subjects collected from the ADNI study with four follow-up visits, namely, the baseline, six months, 12 months, and 24 months. Alzheimer Disease Assessment Scale-Cognitive 13 (ADAS13), which reflects cognitive impairment in AD assessment, is treated as response y_{it}. Generally, a high ADAS13 score indicates low cognitive ability. In addition, some clinical and genetic variables were considered as covariates. One is a time-variant continuous variable, x_{it1}: the logarithm of the ratio of hippocampal volume over the whole brain volume (HIP). Other covariates include APOE-ϵ4, coded as 0, 1, 2, denoting the number of APOE-ϵ4 alleles and represented by x_{it2} ($x_{it2} = 1$ if carrying one allele and 0 otherwise) and x_{it3} ($x_{it3} = 1$ if carrying two alleles and 0 otherwise), patients' age at baseline, x_{it4}, and patients' gender, x_{it5} ($x_{it5} = 1$ if female). In this study, we assume The main goal of this study is to simultaneously identify the number of hidden states and the state-specific relationship between ADAS 13 and its important risk factors.

The prior specification and other settings are similar to the simulation study. We imposed constraint described in Definition 2.1 to each MCMC iteration to avoid label switching. The trace plots of K shown in Figure 2(d) indicte that the MCMC chains of K from different initial values quickly converge to $K = 4$, suggesting a 4-state HMM for the data. Figure S4 of Supplementary Material presents the trace plots of other parameters involved in the selected model. The MCMC chains mixed well within 5000 iterations. Thus, we discarded 5000 burn-in iterations and used the remaining 5000 posterior samples to obtain the parameter estimates. Table 3 presents the parameter estimates of the selected 4-state HMM. Based on the results, we have the following observations.

First, the state-specific intercept β_{1s} exhibits an ascending trend. Patients have the lowest ADAS mean score in state 1 and highest mean score in state 4. According to the existing literature (Kantarci et al. 2013), states 1 to 4 can be interpreted as CN, early mild cognitive impairment (MCI), late MCI, and AD accordingly.

Second, HIP (β_{2s}) exerts an adverse effect on ADAS13, implying that a sizable hippocampal volume is associated with a low ADAS13 score and thus high cognitive ability. Moreover, the magnitude of the HIP effect on ADAS13 increases from CN to AD, implying that hippocampal atrophy continuously impairs patients' cognitive ability during AD progression. The published medical reports (e.g., Dickerson and Wolk 2013) also revealed that the loss of hippocampal volume significantly affects AD.

Third, the effects of APEP-ϵ4 (β_{3s} and β_{4s}) on ADAS13 are positive, suggesting that carrying APOE-ϵ4 increases AD risk, and such impact becomes increasingly pronounced with the disease progression. This finding is in line with the medical report (Risacher et al. 2015) that APOE-ϵ4 is a crucial biomarker of AD. Furthermore, the magnitude of β_{4s} is larger than β_{3s} for $s = 1, \ldots, 4$, implying that carrying two alleles, in general, impairs cognitive function more significantly than carrying only one allele. Besides, patients' age and gender do not substantially affect ADAS13 when controlling hippocampal volume and APOE-ϵ4. An exception lies in $\beta_{64} = 0.399(0.101)$, which suggests that females suffer more severe cognitive decline than males in the late AD progression period. This result again agrees with the existing literature (e.g., Via et al. 2010; Kang et al. 2019).

Table 3. Parameter estimation results for ADNI study.

Parameters in the conditional regression model				
Parameters	State 1 Est(sd)	State 2 Est(sd)	State3 Est(sd)	State 4 Est(sd)
Intercept	−0.803(0.037)	−0.191(0.059)	0.521(0.089)	1.559(0.114)
HIP	−0.164(0.036)	−0.281(0.048)	−0.301(0.042)	−0.331(0.090)
1 APOE-ϵ4	0.039(0.035)	0.135(0.056)	0.198(0.098)	0.253(0.116)
2 APOE-ϵ4	0.263(0.073)	0.542(0.136)	1.138(0.096)	1.906(0.202)
Age	−0.070(0.094)	0.035(0.037)	0.061(0.048)	0.037(0.062)
Female	−0.029(0.031)	0.046(0.051)	0.093(0.068)	0.399(0.101)
ψ	0.094(0.008)	0.101(0.007)	0.136(0.013)	0.421(0.049)

Parameters in the transition model					
Parameters	Est(sd)	Parameters	Est(sd)	Parameters	Est(sd)
Intercept_{11}	2.863(0.255)	Intercept_{12}	3.130(0.806)	Intercept_{13}	−0.762(0.852)
Intercept_{21}	−2.217(0.234)	Intercept_{22}	3.652(0.462)	Intercept_{23}	2.089(0.511)
Intercept_{31}	−3.867(0.450)	Intercept_{32}	−1.701(0.320)	Intercept_{33}	3.941(0.542)
Intercept_{41}	−3.537(0.479)	Intercept_{42}	−3.343(0.431)	Intercept_{43}	−2.271(0.441)
Probability_1	0.322(0.023)	Probability_2	0.314(0.019)	Probability_3	0.224(0.017)
Probability_4	0.140(0.013)	HIP	−0.139(0.099)	1 APOE-ϵ4	−0.538(0.229)
2 APOE-ϵ4	−0.742(0.350)	Age	−0.019(0.104)	Female	0.090(0.106)

Lastly, the transition pattern described by $\boldsymbol{\zeta}$ exhibits a banding structure. That is, patients are likely to transit between adjacent states. Moreover, α_2 and α_3 are significant and negative, implying that the transition pattern between hidden states exhibits heterogeneity. APOE-ϵ4 allele carriers are more likely to transit to a worse state rather than remain in the current one than noncarriers; carrying two alleles induces a higher risk of transitioning to a worse state than carrying one allele. This result is consistent with the existing finding (Eunjee et al. 2015) that APOE-ϵ4 alleles increase the risk of developing AD. However, other covariates, such as age and HIP, do not significantly affect the between-state transition given APOE-ϵ4. This result implies that conditional on APOE-ϵ4, the direct effects of age and hippocampal volume on the transition probability are weak.

7. Discussion

In this study, we have proposed a double penalized method to perform order selection and parameter estimation for heterogeneous HMMs under the Bayesian framework. In addition, we have developed a novel MCMC algorithm with an ABRJ sampling scheme to facilitate a joint updating of the order and model parameters. Multiple simulation studies and an application to the ANDI dataset demonstrate the superiority of the proposed method over existing ones and its utility in realistic settings. Furthermore, the proposed model can cope with general situations where specific covariates simultaneously influence the emission and transition processes.

The present work can be extended in several directions. First, we use a single indicator to represent a response or predictor. For example, we adopt ADAS13 to reflect cognitive ability in the ADNI data analysis. While in practice, multiple tests can be used to examine cognitive impairment, and their scores can be integrated into a univariate latent construct through factor analysis. Such an extension can accommodate latent responses or covariates, reduce model dimensionality, and improve interpretability. Second, our conditional regression model only accommodates a continuous variable. Given that complex data types are frequently encountered in medical, social, and psychological sciences, generalizing our method to incorporate multivariate, functional, or image variables can considerably enhance model capability. Third, this study mainly focuses on order selection. However, variable selection is potentially interesting in the presence of high-dimensional variables. Thus, we can consider additional penalties for simultaneous order and variable selection. Finally, the two penalties in the proposed method aim to tackle the possible overfitting problems without effective strategies to address underfitting issues. Therefore, we design the jumping in an upside-down direction. Given this design, developing a penalization method to prevent overfitting and underfitting simultaneously is of great interest. However, these possible extensions may raise new theoretical and computational challenges.

Supplementary Materials

Supplementary Material: The supplemental files include the Appendix which gives the proof of the propositions in Section 3, full conditional distributions in Section 4, and additional numerical results in Sections 5 and 6. (BHMM Supp.pdf)

R code: The supplemental files for this article include R programs which can be used to replicate the simulation study included in the article. Please read file README contained in the zip file for more details. (program package.zip)

Acknowledgments

The authors are thankful to the editor, the associate editor, and two anonymous reviewers for their valuable comments and suggestions, which have helped improve the article substantially.

Disclosure Statement

The authors report there are no competing interest to declare.

Funding

This research was fully supported by GRF Grants (14302519, 14302220) from Research Grant Council of the Hong Kong Special Administration Region.

References

Agresti, A. (2003), *Categorical Data Analysis*, Hoboken: Wiley. [2]

Akaike, H. (1974), "New Look at Statistical-Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [1]

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, 41, 164–171. [2]

Celeux, G., and Durand, J.-B. (2008), "Selecting Hidden Markov Model State Number with Cross-Validated Likelihood," *Computational Statistics*, 23, 541–564. [1]

Chen, J., and Khalili, A. (2008), "Order Selection in Finite Mixture Models with a Nonsmooth Penalty," *Journal of the American Statistical Association*, 103, 1674–1683. [1]

Dickerson, B., and Wolk, D. (2013), "Biomarker-based Prediction of Progression in MCI: Comparison of AD-Signature and Hippocampal Volume with Spinal Fluid Amyloid-β and Tau," *Front Aging Neurosci*, 5, 55. [8]

Eunjee, L., Hongtu, Z., Dehan, K., Wang, Y., Giovanello, K. S., Ibrahim, J. G., et al. (2015), "BFLCRM: A Bayesian Functional Linear Cox Regression Model for Predicting Time to Conversion to Alzheimer's Disease," *The Annals of Applied Statistics*, 9, 2153–2178. [9]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [1]

Guo, R., Zhu, H., Chow, S.-M., and Ibrahim, J. G. (2012), "Bayesian Lasso for Semiparametric Structural Equation Models," *Biometrics*, 68, 567–577. [4]

Harper, W., and Hooker, C. (1976), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Dordrecht: Springer. [4]

Hung, Y., Wang, Y., Zarnitsyna, V., Zhu, C., and Wu, C.-F. (2013), "Hidden Markov Models with Applications in Cell Adhesion Experiments," *Journal of the American Statistical Association*, 108, 1469–1479. [1]

Ip, E., Zhang, Q., Rejeski, J., Harris, T., and Kritchevsky, S. (2013), "Partially Ordered Mixed Hidden Markov Model for the Disablement Process of Older Adults," *Journal of the American Statistical Association*, 108, 370–384. [1]

Kang, K., Song, X., Hu, X. J., and Zhu, H. (2019), "Bayesian Adaptive Group Lasso with Semiparametric Hidden Markov Models," *Statistics in Medicine*, 38, 1634–1650. [4,8]

Kantarci, K., Gunter, J., Tosakulwong, N., Weigand, S. D., Senjem, M. S., Petersen, R. C., et al. (2013), "Focal Hemosiderin Deposits and β-amyloid Load in the adni Cohort," *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, 9, S116–S123. [8]

Lin, Y., and Song, X. (2022), "Order Selection for Regression-based Hidden Markov Model," *Journal of Multivariate Analysis* (to appear). [1,8]

Liu, H., and Song, X. (2020), "Bayesian Analysis of Hidden Markov Structural Equation Models with an Unknown Number of Hidden States," *Econometrics and Statistics*, 18, 29–43. [1,8]

Mackay, R. (2002), "Estimating the Order of a Hidden Markov Model," *Canadian Journal of Statistics*, 30, 573–589. [1]

Manole, T., and Khalili, A. (2021), "Estimating the Number of Components in Finite Mixture Models via the Group-Sort-Fuse Procedure," *The Annals of Statistics*, 49, 3043–3069. [1,3]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [3]

Risacher, S. L., Kim, S., Nho, K., Foroud, T., Shen, L., Petersen, R. C., et al. (2015), "Apoe Effect on Alzheimer's Disease Biomarkers in Older Adults with Significant Memory Concern," *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, 11, 1417–1429. [8]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [1]

Song, X., Xia, Y., and Zhu, H. (2017), "Hidden Markov Latent Variable Models with Multivariate Longitudinal Data," *Biometrics*, 73, 313–323. [1]

Via, J., and Lloret, A. (2010), "Why Women have more Alzheimer's Disease than Men: Gender and Mitochondrial Toxicity of Amyloid-β Peptide," *Journal of Alzheimer's Disease*, 20, 527–533. [8]

Ye, M., Lu, Z., Li, Y., and Song, X. (2019), "Finite Mixture of Varying Coefficient Model: Estimation and Component Selection," *Journal of Multivariate Analysis*, 171, 452–474. [1]

Zhou, J., Song, X., and Sun, L. (2020), "Continuous Time Hidden Markov model for Longitudinal Data," *Journal of Multivariate Analysis*, 179, 104646. [1,3]