Bayesian Order Selection in Heterogeneous Hidden Markov Models

Yudan Zou, Yiqi Lin, Xinyuan Song

Department of Statistics, CUHK

December, 2022

Outline

Introduction

Hidden Markov Models (HMMs) consist of two parts: a conditional regression (emission) model to examine state-specific covariate effects on the response of interest and a transition model to characterize the dynamic transition process between hidden states, which can be generally formulated as equation (1) and (2).

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

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

Background

Determine the number of hidden states (i.e., order of HMM):

- known or predetermined,
- criterion-based methods, such as the AIC and BIC,
- the reversible jump Markov chain Monte Carlo algorithm(Green,1995)
- penalization procedure (Khalili, 2008; Hung, 2013; Song, 2020; Lin, 2022).

However, these methods either could become increasingly tedious and computationally intensive when the model space is ample, or not directly applicable to the discrete-time heterogeneous HMMs.

Notation

Conditional Regression Model

$$[y_{it}|Z_{it} = s] = \boldsymbol{\beta}'_s \boldsymbol{x}_{it} + \delta_{it}, \tag{3}$$

Transition Model

$$\begin{aligned} Z_{i1} &\sim categorical(\pi_1, ..., \pi_K), \\ \log\left(\frac{P_{itus}}{P_{itu,s+1} + ... + P_{ituK}}\right) &= \zeta_{us} + \boldsymbol{\alpha}'\boldsymbol{d}_{it}. \end{aligned} \tag{4}$$

- y_{it} is the response of subject i at time t,
- $\boldsymbol{x}_{it}$ and $\boldsymbol{d}_{it}$ are two covariate vectors (distinct or overlapped),
- Z_{it} is the hidden states that takes the values of $\{1, ..., K\}$,
- $P_{itus} = P(Z_{it} = s|Z_{i1}, Z_{i2}, ..., Z_{i,t-1} = u) = P(Z_{it} = s|Z_{i,t-1} = u)$,
- $\boldsymbol{\beta}_s$ is the $p \times 1$ vector of state-specific regression coefficients,
- δ_{it} is the residual term independent of $\boldsymbol{x}_{it}$,
- ζ_{us} is a transition-specific intercept,
- $\boldsymbol{\alpha}$ is a $q \times 1$ vector of regression coefficients.

Identification

To address the label switching problem, we introduce the cluster ordering procedure to sort the multidimensional parameters in the conditional regression model as follows:

Definition

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

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

Transformation

The cluster ordering 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 $k = 2, \ldots, K$. Hence, we rewrite (3) as follows:

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

Hence, the complete-data log-likelihood of $\boldsymbol{\theta}$ (all parameters) is of the form:

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

Motivations

2 types of over-fitting in performing parameter estimation based on (7):

- close emission densities,
- near-zero mixing probabilities.

Double penalize framework in preceding developments focused either on finite mixture models or homogeneous HMMs and are thus not directly applicable to the present heterogeneous HMMs. In addition, the previous studies are mainly developed in the frequentist framework. Their Bayesian versions have never been considered in the literature.

This study considers a novel Bayesian double penalized method for simultaneous order selection and parameter estimation for heterogeneous HMMs.

Bayesian double penalized (BDP) procedure

Specifically, the double penalty is introduced as follows:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\sum_{k=1}^{s}\boldsymbol{\eta}_k'\boldsymbol{x}_{it})^2-\sum_{i=1}^{n}\sum_{t=2}^{T}\log(P_{itus})-\sum_{i=1}^{n}\log(P_{i10s})-P(\boldsymbol{\theta}) \quad (8)$$

where

$$P(\boldsymbol{\theta})=(1-c_K)\sum_{k=1}^{K}\log(\pi_k)+\sum_{k=2}^{K}\gamma_k||\boldsymbol{\eta}_k||_2, \quad (9)$$

in which c_K and $\{\gamma_2,\ldots,\gamma_K\}$ are tuning parameters and $||\cdot||_2$ denotes the L_2 norm.

BDP procedure

Proposition

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

$$E(\pi_s|\boldsymbol{Z}) \geq \frac{c}{c+1}\frac{1}{K}, \quad s = 1, \ldots, K. \tag{10}$$

This proposition ensures that the first penalty prevents near-zero probabilities and thus nearly empty states. The constant c can be determined according to the degree of penalty required for specific problems. Based on our extensive simulations, c takes 0.5 works effectively.

BDP procedure

Second, we use the MAGlasso to tackle the second type of overfitting problems. The basic idea of the Bayesian lasso is to penalize $\boldsymbol{\eta}_k$ by imposing a conditional Laplace prior on $\boldsymbol{\eta}_k$ as follows:

$$P(\boldsymbol{\eta}_k|\psi_k) = \frac{\gamma_k}{2\sqrt{\psi_k}} \exp\big(- \frac{\gamma_k}{\sqrt{\psi_k}}||\boldsymbol{\eta}_k||_2\big), \quad k = 2, \ldots, K. \tag{11}$$

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

$$\begin{aligned}
&[y_{it}|Z_{it} = s, \boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_s, \psi_s] \sim N(\sum_{k=1}^{s}(\boldsymbol{\eta}_k' \boldsymbol{x}_{it}), \psi_s), \\
&[\boldsymbol{\eta}_s|\psi_s, \tau_s^2] \sim N(\mathbf{0}, \psi_s \tau_s^2 \boldsymbol{I}_p), \quad \psi_s^{-1} \overset{ind}{\sim} Gamma(a_{\psi s0}, b_{\psi s0}), \\
&\tau_s^2 \overset{ind}{\sim} Gamma(\frac{p+1}{2}, \frac{\gamma_s^2}{2}), \quad \gamma_k^2 \overset{ind}{\sim} Gamma(a_{\gamma k0}, b_{\gamma k0}), \; k = 2, \cdots, K,
\end{aligned} \tag{12}$$

BDP procedure

Proposition

Under the hierarchical model (12), *the conditional prior distribution of* $\boldsymbol{\eta}_k$, $k = 2, \ldots, K$ *are in the form of* (11).

The MAGlasso procedure aims to update the tuning parameters by exploiting the data, thereby automatically imposing large penalties on unimportant coefficients. With this prior specification, the posterior distribution of the tuning parameters have the following forms:

$$[\tau_s^{-2}|\cdot] \sim \text{Inverse-Gaussian}\left\{\sqrt{\frac{\gamma_s^2\psi_s}{||\boldsymbol{\eta}_s||_2^2}}, \gamma_s^2\right\}, \tag{13}$$

$$[\gamma_s^2|\cdot] \sim Gamma(a_{\gamma s0} + \frac{p+1}{2}, b_{\gamma s0} + \frac{\tau_s^2}{2}). \tag{14}$$

BDP procedure

Considering that the Bayesian lasso does not shrink coefficients to precisely zero, we adopt the 95% highest posterior credible region (HPCR) criterion to test whether a component is significant. Based on the specification of (11), we can show that $P(\boldsymbol{\eta}_s|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{Z}, \boldsymbol{\theta}) \sim N(\boldsymbol{\eta}_s^*, \boldsymbol{\Sigma}_s^*)$, and 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 $\boldsymbol{\eta}_s^*$.

Thus, $\boldsymbol{\eta}_s$ 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, i.e.,

- redundant, if ${\boldsymbol{\eta}_s^*}' {\boldsymbol{\Sigma}_s^*}^{-1} \boldsymbol{\eta}_s^* \leq \chi_{p,0.05}^2$
- necessary, otherwise

Bayesian analysis

Prior Specification:

$$
\begin{aligned}
&[y_{it}|Z_{it}=s,\boldsymbol{\eta}_1,...,\boldsymbol{\eta}_s,\psi_s]\sim N(\sum_{k=1}^{s}(\boldsymbol{\eta}_k'\boldsymbol{x}_{it}),\psi_s),\\
&[\boldsymbol{\eta}_s|\psi_s,\tau_s^2]\sim N(\mathbf{0},\psi_s\tau_s^2\boldsymbol{I}_p),\\
&\tau_s^2\overset{ind}{\sim} Gamma(\frac{p+1}{2},\frac{\gamma_s^2}{2}),\\
&\psi_s^{-1}\overset{ind}{\sim} Gamma(a_{\psi s0},b_{\psi s0}),\\
&\gamma_k^2\overset{ind}{\sim} Gamma(a_{\gamma k0},b_{\gamma k0}),\\
&(\pi_1,...,\pi_K)\sim Dir(c_K,...,c_K),\quad c_K=c\frac{n}{K},\\
&\zeta_{us}\overset{ind}{\sim} N(\zeta_{us0},\sigma_{us0}^2),\\
&\boldsymbol{\alpha}_k\overset{ind}{\sim} N(\boldsymbol{\alpha}_{k0},\boldsymbol{\Sigma}_{\alpha k0}),\ k=1,\cdots,q.
\end{aligned}
\tag{15}
$$

Bayesian analysis

Posterior Distribution:

1. Full conditional distribution of Z_{it}

$$\begin{aligned} p[Z_{it}|\cdot] &\propto p(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i, Z_{it}|\boldsymbol{\theta}) \\ &= p(y_{i1}, ..., y_{it}, \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{it}, \boldsymbol{d}_{i1}, ..., \boldsymbol{d}_{it}, Z_{it}|\boldsymbol{\theta}) \\ &\times p(y_{i,t+1}, ..., y_{iT}, \boldsymbol{x}_{i,t+1}, ..., \boldsymbol{x}_{iT}, \boldsymbol{d}_{i,t+1}, ..., \boldsymbol{d}_{iT}|\boldsymbol{\theta}, Z_{it}) \\ &\doteq q_{it}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i, Z_{it}|\boldsymbol{\theta}) \times \tilde{q}_{it}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i|\boldsymbol{\theta}, Z_{it}). \end{aligned} \tag{16}$$

By forward filtering and backward sampling scheme, we first initialize $q_{i1}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i, Z_{it}|\boldsymbol{\theta}) = p(y_{i1}, \boldsymbol{x}_{i1}, \boldsymbol{d}_{i1}, Z_{i1}|\boldsymbol{\theta}) = p(y_{i1}, \boldsymbol{x}_{i1}, \boldsymbol{d}_{i1}|\boldsymbol{\theta}, Z_{i1}) \times p(Z_{i1}|\boldsymbol{\theta})$ and calculate $q_{it}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i, Z_{it}|\boldsymbol{\theta})$ for $t = 2, ..., T$ in a recursion manner as follows:

$$\begin{aligned} & q_{it}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i, Z_{it}|\boldsymbol{\theta}) \\ =& \sum_{u=1}^{K} p(y_{i1}, ..., y_{i,t-1}, \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{i,t-1}, \boldsymbol{d}_{i1}, ..., \boldsymbol{d}_{i,t-1}, Z_{i,t-1} = u|\boldsymbol{\theta}) \\ & \times p(Z_{it}|Z_{i,t-1} = u, \boldsymbol{d}_{it}, \boldsymbol{\theta}) p(y_{it}|\boldsymbol{x}_{it}, Z_{it}, \boldsymbol{\theta}) \\ =& \sum_{u=1}^{K} [q_{i,t-1}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i, Z_{i,t-1} = u|\boldsymbol{\theta}) p(Z_{it}|Z_{i,t-1} = u, \boldsymbol{d}_{it}, \boldsymbol{\theta}) p(y_{it}|\boldsymbol{x}_{it}, Z_{it}, \boldsymbol{\theta})]. \end{aligned} \tag{17}$$

Bayesian analysis

Similarly, we initialize $\tilde{q}_{iT}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i|\boldsymbol{\theta}, Z_{iT}) = 1$ and calculate $\tilde{q_{it}}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i|\boldsymbol{\theta}, Z_{it})$ for $t = T-1, ..., 1$ in a recursion manner as follows:

$$\begin{aligned}
&\tilde{q}_{it}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i|Z_{it}, \boldsymbol{\theta}) \\
=&\sum_{u=1}^{K} p(y_{i,t+2}, ..., y_{iT}, \boldsymbol{x}_{i,t+2}, ..., \boldsymbol{x}_{iT}|Z_{i,t+1} = u, \boldsymbol{\theta}), \\
&\times p(Z_{i,t+1} = u|Z_{it}, \boldsymbol{d}_{i,t+1}, \boldsymbol{\theta}) p(y_{i,t+1}|\boldsymbol{x}_{i,t+1}, Z_{i,t+1} = u, \boldsymbol{\theta}) \\
=&\sum_{u=1}^{K} [\tilde{q}_{i,t+1}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{D}_i|Z_{i,t+1} = u, \boldsymbol{\theta}) p(Z_{i,t+1} = u|Z_{it}, \boldsymbol{d}_{i,t+1}, \boldsymbol{\theta}) \\
&\times p(y_{i,t+1}|\boldsymbol{x}_{i,t+1}, Z_{i,t+1} = u, \boldsymbol{\theta})].
\end{aligned}$$

2. Full conditional distribution of $\boldsymbol{\eta}_s$, ψ_s^{-1}

$$[\boldsymbol{\eta}_s|\cdot] \sim N(\boldsymbol{\eta}_s^*, \boldsymbol{\Sigma}_s^*), \qquad [\psi_s^{-1}|\cdot] \sim Gamma(a_{\psi s}^*, b_{\psi s}^*), \tag{18}$$

where $s_{\psi s}^* = a_{\psi s0} + (n_s + p)/2$, and

$$\boldsymbol{\Sigma}_s^* = (\sum_{i=1}^{n} \sum_{t=1}^{T} \psi_s^{-1} \boldsymbol{x}_{it} \boldsymbol{x}_{it}' \times I(Z_{it} = s) + \psi_s^{-1} \tau_s^{-2} I_p)^{-1},$$

$$\boldsymbol{\eta}_s^* = \boldsymbol{\Sigma}_s^* [\sum_{i=1}^{n} \sum_{t=1}^{T} \psi_s^{-1} \boldsymbol{x}_{it} (y_{it} - \sum_{m=1}^{s-1} (\boldsymbol{\eta}_m' \boldsymbol{x}_{it})) \times I(Z_{it} = s)],$$

$$b_{\psi s}^* = b_{\psi s0} + \frac{1}{2} [\sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \sum_{k=1}^{s} (\boldsymbol{\eta}_k' \boldsymbol{x}_{it}))^2 \times I(Z_{it} = s) + \tau_s^{-2} \boldsymbol{\eta}_s' \boldsymbol{\eta}_s].$$

Bayesian analysis

3. Full conditional distribution of π_s, ζ_{us}, $\boldsymbol{\alpha}$

$$[\boldsymbol{\pi}|\cdot] \sim Dir(c_K + \sum_{i=1}^{n} I(Z_{i1} = 1), \ldots, c_K + \sum_{i=1}^{n} I(Z_{i1} = K)),$$

$$p\,[\zeta_{us}|\cdot] \propto exp\left\{ \sum_{v=s}^{K} \sum_{i=1}^{n} \sum_{t=2}^{T} log(P_{ituv} \times I(Z_{it} = v, Z_{i,t-1} = u) - \frac{(\zeta_{us} - \zeta_{us0})^2}{2\sigma^2_{\zeta us0}} \right\},$$

$$p\,[\boldsymbol{\alpha}|\cdot] \propto exp\left\{ \sum_{i=1}^{n} \sum_{t=2}^{T} log(P_{itus} \times I(Z_{it} = s, Z_{i,t-1} = u) - \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)'\boldsymbol{\Sigma}_{\alpha}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) \right\},$$

MCMC algorithm

In this study, data augmentation technique and 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 $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{Z}, K)$.

(c) Update the order by sampling K from $P(K|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\theta})$.

Owing to the features of hidden states and nonlinearity of the transition model (2), steps (a) and (b) require some MCMC methods, such as the FFBS and MH algorithms. Step (c) is an adjust-bound reversible jump (ABRJ) step, which allows the number of hidden states, K, to be updated at each MCMC iteration.

MCMC algorithm

Let $(K_{min}^{(j)}, K_{max}^{(j)})$ be the lower and upper bounds of K at the jth iteration of the MCMC algorithm. Typically, we set $K_{min}^{(0)} = 2$ and $K_{max}^{(0)}$ relatively large for sufficient flexibility. At the $(j+1)$th iteration, $K^{(j+1)}$ can remain unchanged, increase, or decrease by 1. To update $K^{(j)}$, we first locate the state s_*, such that $s_* = argmin_{s=1,\ldots K^{(j)}} ||\boldsymbol{\eta}_s^{(j)}||_2$. Then, we calculate $d_{s_*}^2 = \boldsymbol{\eta}_{s_*}^{(j)'} \boldsymbol{\Sigma}_{s_*}^{(j)^{-1}} \boldsymbol{\eta}_{s_*}^{(j)}$ and compare $d_{s_*}^2$ with $\chi^2_{p,0.05}$. If $d_{s_*}^2 \leq \chi^2_{p,0.05}$, we regard this component as redundant and update $K^{(j)}$ downward to $K^{(j+1)} = max(K^{(j)} - 1, K_{min}^{(j)})$. Meanwhile, we adjust $K_{max}^{(j)}$ as $K_{max}^{(j+1)} = min(K_{max}^{(j)}, K^{(j)})$. If $d_{s_*}^2 > \chi^2_{p,0.05}$, we regard this component as necessary. Then, we jump $K^{(j)}$ upward to $K^{(j+1)} = min(K^{(j)} + 1, K_{max}^{(j)} - 1)$. If $d_{s_*}^2 > \chi^2_{p,0.05}$ but $K^{(j)} = K_{max}^{(j)} - 1$, then $K^{(j)}$ remains unchanged, i.e., $K^{(j+1)} = K^{(j)}$.

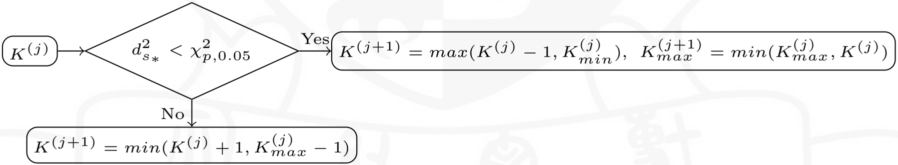

Figure 1: Strategy of updating K in the ABRJ step (c).

MCMC algorithm

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
  $s_* = argmin_{s=1,\dots,K^{(j)}} ||\boldsymbol{\eta}^{(j)}_s||_2$
  $\boldsymbol{\eta}^{(j)}_{s_*} = E(\boldsymbol{\eta}_{s_*}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$ ▷ posterior mean vector
  $\boldsymbol{\Sigma}^{(j)}_{s_*} = Var(\boldsymbol{\eta}_{s_*}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, K^{(j)})$ ▷ posterior covariance matrix
  $d^2_{s_*} = \boldsymbol{\eta}^{(j)\prime}_{s_*} \boldsymbol{\Sigma}^{(j)\,-1}_{s_*} \boldsymbol{\eta}^{(j)}_{s_*}$
  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 $d^2_{s_*} \geq \chi^2_{p,0.05}$ **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

Simulation results

Considers a 2-state heterogeneous HMM formulated by (19) 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:

$$\begin{aligned} [y_{it} | Z_{it} = s] &= \boldsymbol{\beta}_s' \boldsymbol{x}_{it} + \delta_{it}, \\ \mathrm{logit}(\vartheta_{itus}) &= \zeta_{us} + \alpha d_{it}, \end{aligned} \tag{19}$$

where $\boldsymbol{x}_{it} = (1, x_{it1}, x_{it2}, x_{it3})'$, $x_{it1} \overset{ind}{\sim} N(0,1)$, $x_{it2} \overset{ind}{\sim} U(-1,1)$, $x_{it3} \overset{ind}{\sim} Bernoulli(0.6)$, and $d_{it} \overset{ind}{\sim} N(0,1)$. The true values of the parameters are: $\boldsymbol{\beta_1} = (2, 2, 1, 1)'$, $\boldsymbol{\beta_2} = (0, 1, 2, -1)'$, $\psi_1 = \psi_2 = 0.25$, $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\zeta} = (\zeta_{11}, \zeta_{21}) = (-2, 2)'$, and $\alpha = -1$. The hyperparameters(Prior I) of the prior distributions are: $a_{\psi s0} = 9$, $b_{\psi s0} = 4$, $a_{\gamma k0} = 1$, $b_{\gamma k0} = 0.1$, $c = 0.5$, $\alpha_{k0} = \zeta_{us0} = 0$, and $\sigma^2_{\alpha k0} = \sigma^2_{us0} = 1$.

Simulation results

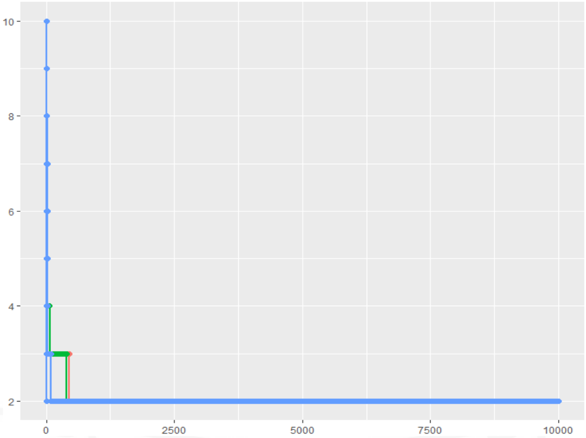

Figure 2: Trace plots of three MCMC chains of K in simulations ($K_0 = 2$)

Simulation results

Table 2: 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			

Other simulation study

Sensitivity analysis

- Disturbed hyperparameters (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 $\sigma^2_{\alpha k0} = \sigma^2_{us0} = 100$;
- Disturbed residual terms δ_{it}:
  one is uniform distributed $\delta_{it} \sim U(-1, 1)$, another one is mixture normal type as $\delta_{it} \sim 0.4N(1, 1) + 0.6N(-1, 1)$, while with other hyperparameters setting the same as Prior I.

Higher order settings

- 3-state HMM with $(n, T) = (200, 6), (400, 6)$;
- 5-state HMM with $(n, T) = (200, 6), (400, 6)$.

Performance comparison with AIC, BIC

Table 3: model selection proportion

		$n = 100$			$n = 200$			$n = 400$		
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

ADNI application

In this section, we applied the proposed method to the dataset extracted from the ADNI-I, ADNI-II, and ADNI-Go studies to demonstrate the practical utility of the proposed method. (`www.adni-info.org.`)

- sample size: $n = 616, T = 4$;
- y_{it}: ADAS13 (represents cognitive impairment in AD assessment);
- x_{it1}: HIP (the logarithm of the ratio of hippocampal volume over the whole brain volume)
- x_{it2}, x_{it3}: the number of apolipoprotein APOE-ϵ4 alleles
- x_{it4}: patients' age at baseline
- x_{it5}: patients' gender ($x_{it5} = 1$ if female)

ADNI application

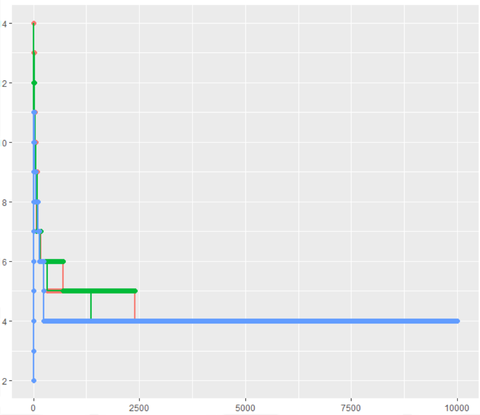

Figure 3: Trace plots of three MCMC chains of K in the ADNI study.

ADNI application

Table 4: Parameter estimation result for ADNI

State 1		State 2		State 3		State 4	
Par	Est(sd)	Par	Est(sd)	Par	Est(sd)	Par	Est(sd)
Parameters in the conditional regression model							
β_{11}	$-0.803(0.037)$	β_{12}	$-0.191(0.059)$	β_{13}	$0.521(0.089)$	β_{14}	$1.559(0.114)$
β_{21}	$-0.164(0.036)$	β_{22}	$-0.281(0.048)$	β_{23}	$-0.301(0.042)$	β_{24}	$-0.331(0.090)$
β_{31}	$0.039(0.035)$	β_{32}	$0.135(0.056)$	β_{33}	$0.198(0.098)$	β_{34}	$0.253(0.116)$
β_{41}	$0.263(0.073)$	β_{42}	$0.542(0.136)$	β_{43}	$1.138(0.096)$	β_{44}	$1.906(0.202)$
β_{51}	$-0.070(0.094)$	β_{52}	$0.035(0.037)$	β_{53}	$0.061(0.048)$	β_{54}	$0.037(0.062)$
β_{61}	$-0.029(0.031)$	β_{62}	$0.046(0.051)$	β_{63}	$0.093(0.068)$	β_{64}	$0.399(0.101)$
ψ_1	$0.094(0.008)$	ψ_2	$0.101(0.007)$	ψ_3	$0.136(0.013)$	ψ_4	$0.421(0.049)$
Parameters in the transition model							
ζ_{11}	$2.863(0.255)$	ζ_{21}	$-2.217(0.234)$	ζ_{31}	$-3.867(0.450)$	ζ_{41}	$-3.537(0.479)$
ζ_{12}	$3.130(0.806)$	ζ_{22}	$3.652(0.462)$	ζ_{32}	$-1.701(0.320)$	ζ_{42}	$-3.343(0.431)$
ζ_{13}	$-0.762(0.852)$	ζ_{23}	$2.089(0.511)$	ζ_{33}	$3.941(0.542)$	ζ_{43}	$-2.271(0.441)$
π_1	$0.322(0.023)$	π_2	$0.314(0.019)$	π_3	$0.224(0.017)$	π_4	$0.140(0.013)$
α_1	$-0.139(0.099)$	α_2	$-0.538(0.229)$	α_3	$-0.742(0.350)$	α_4	$-0.019(0.104)$
α_5	$0.090(0.106)$						

ADNI application

Conclusions:

- β_{1s} exhibits a descending trend, states 1 to 4 can be explained as CN, early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and AD accordingly.
- HIP (β_{2s}) exhibits adverse effects on ADAS13, and the atrophy in the hippocampus continuously impairs patients' cognitive ability during the progression of AD.
- 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.
- The transition pattern described by $\boldsymbol{\zeta}$ exhibits a banding structure, i.e., patients are likely to transit between adjacent states.
- α_2 and α_3 are significant and negative, implying heterogeneity in the transition pattern. APOE-ϵ4 allele carriers are more likely to transit to a worse state rather than remain in the current one than noncarriers.

Possible extensions

1. Introduce latent variable as response or covariates.
2. Additional penalty for variable selection.
3. Accommodate complicated data structure, such as image variable.

However, these possible extensions may raise new theoretical and computational challenges.

Thanks for Your Attention

References

[1] Alan Agresti. *Categorical data analysis*. English. John Wiley & Sons, 2003. ISBN: 9780470463635.

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

[3] Rachel MacKay Altman. "Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 201–210.

[4] Francesco Bartolucci and Alessio Farcomeni. "A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 816–831.

References

[5] Leonard E Baum et al. “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains”. In: *The Annals of Mathematical Statistics* 41.1 (1970), pp. 164–171.

[6] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. New York, NY: Springer, 2005.

[7] Gilles Celeux and Jean-Baptiste Durand. “Selecting hidden Markov model state number with cross-validated likelihood”. In: *Computational Statistics* 23 (2008), pp. 541–564.

[8] Gilles Celeux et al. “Deviance Information Criteria for Missing Data Models”. In: *Bayesian Analysis* 1 (2006), pp. 651–674.

[9] Jiahua Chen and Abbas Khalili. “Order Selection in Finite Mixture Models With a Nonsmooth Penalty”. In: *Journal of the American Statistical Association* 103 (2008), pp. 1674–1683.

References

[10] Bradford Dickerson and David Wolk. "Biomarker-based prediction of progression in MCI: comparison of AD-signature and hippocampal volume with spinal fluid amyloid-β and tau.". In: *Front Aging Neurosci* 5 (2013), p. 55.

[11] Lee Eunjee, Zhu Hongtu, Kong Dehan, et al. "BFLCRM: A Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease". In: *The Annals of Applied Statistics* 9 (2015), pp. 2153–2178.

[12] Peter J. Green. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination". In: *Biometrika* 82.4 (1995), pp. 711–732.

[13] Ruixin Guo et al. "Bayesian lasso for semiparametric structural equation models". In: *Biometrics* 68.2 (2012), pp. 567–577.

References

[14] William Harper and Cliff Hooker. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Jan. 1976. ISBN: 978-90-277-0621-8.

[15] Ying Hung et al. "Hidden Markov Models With Applications in Cell Adhesion Experiments". In: *Journal of the American Statistical Association* 108 (2013), pp. 1469–1479.

[16] Edward Ip et al. "Partially ordered mixed hidden Markov model for the disablement process of older adults". In: *Journal of the American Statistical Association* 108.502 (2013), pp. 370–384.

[17] K. Kang et al. "Bayesian adaptive group lasso with semiparametric hidden Markov models". In: *Statistics in medicine* 38 (2019), pp. 1634–1650.

References

[18] Kejal Kantarci, Jeffrey Gunter, et al. "Focal hemosiderin deposits and β-amyloid load in the ADNI cohort". In: *Alzheimer's & dementia : the journal of the Alzheimer's Association* 9 (2013), S116–S123.

[19] Sik-Yum Lee and X.Y. Song. "Bayesian model selection for mixtures of structural equation models with an unknown number of components". In: *The British journal of mathematical and statistical psychology* 56 (2003), pp. 145–165.

[20] Y. Lin and X. Song. "Order selection for regression-based hidden Markov model". In: *Journal of Multivariate Analysis* to appear (2022).

[21] H.F. Liu and X.Y. Song. "Bayesian analysis of hidden Markov structural equation models with an unknown number of hidden states". In: *Econometrics and Statistics* 18 (2020), pp. 29–43.

References

[22] Rachel Mackay. "Estimating the order of a hidden Markov model". In: *Canadian Journal of Statistics* 30 (2002), pp. 573–589.

[23] Tudor Manole and Abbas Khalili. "Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure". In: *The Annals of Statistics* 49.6 (2021), pp. 3043–3069.

[24] Tudor Manole and Abbas Khalili. "Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure". In: *The Annals of Statistics* 49 (2021).

[25] Antonello Maruotti. "Mixed hidden Markov models for longitudinal data: an overview". In: *International Statistical Review* 79.3 (2011), pp. 427–454.

References

[26] Trevor Park and George Casella. "The Bayesian lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686.

[27] Martyn Plummer. "Penalized loss functions for Bayesian model comparison". In: *Biostatistics* 9 (2008), pp. 523–539.

[28] Shannon L Risacher et al. "APOE effect on Alzheimer's disease biomarkers in older adults with significant memory concern". In: *Alzheimer's & dementia : the journal of the Alzheimer's Association* 11,12 (2015), pp. 1417–1429.

[29] Christian Robert, Tobias Rydén, and D. Titterington. ""Bayesian Inference in Hidden Markov Models through Jump Markov Chain Monte Carlo"". In: *Journal of the Royal Statistical Society Series B* 62 (2000), pp. 57–75.

References

[30] Gideon Schwarz. "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6 (1978), pp. 461–464.

[31] Steven Scott, Gareth James, and Alexander Sugar. "Hidden Markov Models for Longitudinal Comparisons". In: *Journal of the American Statistical Association* 100 (2005), pp. 359–369.

[32] Steven L Scott, Gareth M James, and Catherine A Sugar. "Hidden Markov models for longitudinal comparisons". In: *Journal of the American Statistical Association* 100.470 (2005), pp. 359–369.

[33] X.Y. Song, Y.M. Xia, and H.T. Zhu. "Hidden Markov latent variable models with multivariate longitudinal data". In: *Biometrics* 73 (2017), pp. 313–323.

References

[34] Dorra Trabelsi et al. “An Unsupervised Approach for Automatic Activity Recognition Based on Hidden Markov Model Regression”. In: *Automation Science and Engineering, IEEE Transactions on* 10 (2013), pp. 829–835.

[35] Via et al. “Why Women Have More Alzheimer’s Disease Than Men: Gender and Mitochondrial Toxicity of Amyloid-β Peptide.”. In: *Journal of Alzheimer’s Disease* 20 (2010), pp. 527–533.

[36] Mao Ye et al. “Finite mixture of varying coefficient model: Estimation and component selection”. In: *Journal of Multivariate Analysis* 171 (2019), pp. 452–474.

[37] J. Zhou, X. Song, and L. Sun. “Continuous time hidden Markov model for longitudinal data”. In: *Journal of Multivariate Analysis* 179 (2020), p. 104646.