Chapter 3

Bayesian Analysis for Hidden Markov Factor Analysis Models

Yemao Xia, Xiaoqian Zeng and Niansheng Tang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72837

Abstract

The purpose of this chapter is to provide an introduction to Bayesian approach within a general framework and develop a Bayesian procedure for analyzing multivariate longitudinal data within the hidden Markov factor analysis framework.

Keywords: hidden Markov factor analysis model, Markov chain Monte Carlo sampling, cocaine use

1. Introduction

The Bayesian approach is now well recognized in the statistics literature as an attractive approach to analyzing a wide variety of models [1], and there is rich literature on this issue. Here, we are not going to present a full coverage on the general Bayesian theory, and readers may refer to excellent books, for example [2, 3], for more details for this general statistical method. This chapter provides an introduction to the Bayesian approach within a general framework and develops a specific Bayesian procedure for analyzing multivariate longitudinal data within the hidden Markov factor analysis framework. We begin with the basic ideas of the Bayesian approach and then describe the model under consideration in the second section. The following section considers Bayesian inferences including parameter estimation, model selection, and posterior density estimates. The final section demonstrates the practical value of proposed methodology to cocaine use data to get some Bayesian results. Some technical details are given in the Appendix.

Consider a data set Y with the probability model $p(Y|\theta)$ where θ is a univariate or multivariate population parameters vector, which quantifies the uncertainty of data. In the statistical literature,

$p(Y|\theta)$ is called *likelihood* or sampling distribution and often represented as $L(\theta)$. From the frequency statistics point of view, statistical inferences are carried out based on $L(\theta)$. In this case, θ, though unknown, is treated as fixed. Unlike the frequency statistical inferences, the Bayesian approach for data analysis assumes that θ is random and has a distribution $\pi(\theta)$. This distribution, which represents the knowledge about θ, is referred to as prior distribution or *prior*. When data are available, the information on θ is summarized within the posterior distribution or *posterior*, a conditional distribution θ given data, i.e.,

$$p(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{p(Y)} \propto p(Y|\theta)p(\theta) \tag{1}$$

where $p(Y) = \int p(Y|\theta)\pi(\theta)d\theta$ is the marginal distribution of Y. The right-hand-side term in (1) omits the factor $p(Y)$ since given Y it is a known constant. In Bayes literature, $p(Y|\theta)p(\theta)$ is also termed the unnormalized posterior. Analogous to the role of likelihood in frequency statistical inferences, posterior is the starting point of Bayesian inferences.

Selecting proper priors for parameters is fundamental to Bayesian analysis. Basically, there are two kinds of prior distributions, namely, the noninformative prior distributions and the informative prior distributions. Noninformative prior distributions associate with situations when the prior distributions have no population basis. They are used when we have little prior information on θ and desire that the prior distributions play a minimal role in the posterior distribution distribution. Informative prior distribution represents the distribution of possible parameter values, from which the parameter θ has been drawn. We may have prior knowledge about this distribution, either from closed related data or from the subjective knowledge of experts. A commonly used informative prior distribution in the general Bayesian approach to statistical problems is the conjugate prior distribution, a prior ensuring that the posterior distribution follows the same parametric form as the prior distribution [1, 3].

A potential difficulty underlying Bayesian inferences is the statistical computation when posterior distribution takes on the complicated form. This is particularly true in the situation where latent variables or other unobservable quantities are involved in the model, as discussed in this chapter. In such cases, statistical inferences usually recur to simulation-based methods. Among various sampling methods, Markov chains Monte Carlo methods (MCMC) provide powerful tools for simulating observations from posterior. The key to Markov chain simulation is to create a Markov sequence whose stationary distribution is a specified posterior $p(\theta|Y)$. Posterior inferences are carried out based on these simulated observations. There are many ways of constructing these Markov chains, but all of them, including the Gibbs sampler [4, 5], are special cases of the general framework of Metropolis et al. [6] and Hastings [7]. However, we do not intend to pursue this issue here, and details on simulation-based methods can be referenced to [2, 3, 8, 9].

In what follows, as an illustration, we will develop a Bayesian analysis procedure for multivariate data under longitudinal setting. Multivariate longitudinal or clustered data occur when multiple items are measured repeatedly over periods of time or across occasions. Under such setting, the primary interest is inference about the dependence of the multiple measurements and the temporal correlation resulting from the repeated measures on the same items. But more often, particular interest also focuses on exploring the potential heterogeneity of data and investigating its transition pattern over time. In these cases, hidden Markov latent variable model (HMLVM) [10–13] provides a feasible and unified framework to address these issues. HMLVM assumes that the overall model constitutes the observed process and the underlying hidden state process. The state process, as the convention in the classic HMM (see for example, [14–17]), is an univariate discrete process, which follows a first-order Markov chain, while the observed process, conditional on the state sequence, is an independent process with emission distribution specified via LVMs [18]. Hence, in this regard, HMLVM provides a unified way of describing the correlation of multiple items, temporal dependence, and heterogeneity among the data simultaneously. However, the current existing developments cited beforehand focus on the maximum likelihood analysis in which statistical inferences heavily depend on the asymptotic properties. As an illustration of Bayesian inferences on practical problems, in this chapter, we develop a Bayesian procedure to analyze cocaine use data within the hidden Markov factor analysis model framework. Compared to ML, a basic nice feature of a Bayesian approach is its flexibility to utilize useful prior information for achieving better results. Additionally, simulation-based Bayesian methods depend less on asymptotic theory and hence have the potential to produce more reliable results even with small samples.

2. Model description

2.1. Hidden Markov factor analysis model

Consider a set of multivariate longitudinal observations formed by p-dimensional observed vectors $y_{it} = \left(y_{it1}, \ldots, y_{itp}\right)^{\intercal}$, which are recorded on p items over periods of length T: $t = 1, \cdots, T$ across N subjects: $i = 1, \cdots, N$. In the field of multivariate analysis, interest mainly focuses on exploring item dependence since measurements may be highly correlated arising from the multicollinearity problem. But more often, interest also concentrates on the heterogeneity resulting from the situation where the population of y_{it} constitutes more than one component. This is particularly true in the situation where the data illustrate extreme behaviors such as multimodal and/or skewed characteristics. In these cases, a finite mixture factor analysis model (FMFAM) can provide a powerful tool to address these issues. Typically, FMFAM assumes that conditioning on an univariate discrete value state variable z_{it} and an m-dimensional ($m < p$) continuous latent factor vector $\boldsymbol{\omega}_{it}$, $\mathbf{y}_{it}$ are independent and distributed with a p-dimensional multivariate normal distribution, and meanwhile, given z_{it}, $\boldsymbol{\omega}_{it}$ also follows an m-dimensional normal distribution, that is,

$$\begin{cases} \left(\mathbf{y}_{it} | \boldsymbol{\omega}_{it}, z_{it} = r\right) \sim \mathcal{N}_p\left(\boldsymbol{\mu}_r + \boldsymbol{\Lambda}_r \boldsymbol{\omega}_{it},\ \ \boldsymbol{\Psi}_{\varepsilon r}\right) \\ \left(\boldsymbol{\omega}_{it} | z_{it} = r\right) \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Phi}_r) \end{cases} \tag{2}$$

where $\boldsymbol{\mu}_r = \left(\mu_{r1}, \ldots, \mu_{rp}\right)^{\intercal}$ is a p-dimensional intercept vector, which represents the baseline level of $\mathbf{y}_{it}$, $\boldsymbol{\Lambda}_r = \left(\boldsymbol{\Lambda}_{r1}^{\intercal}, \ldots, \boldsymbol{\Lambda}_{rp}^{\intercal}\right)^{\intercal}$ is a $p \times m$ factor loading matrix, $\boldsymbol{\Psi}_{\epsilon r} = \mathrm{diag}\left\{\Psi_{\epsilon kr1}, \cdots, \Psi_{\epsilon krp}\right\}$ is a $p \times p$ diagonal matrix with the jth diagonal element $\Psi_{\epsilon krj} > 0$, and $\boldsymbol{\Phi}_r$ is an $m \times m$ positive definite matrix.

Formulation given in (2) has two basic features: one is to characterize heterogeneity of population of $\mathbf{y}_{it}$ at the occasion level and the other is to establish the dependence among the multiple measurements. The heterogeneous population is specified via state-specific parameters contained in the model while the dependence between different measurements is identified via sharing the common factors in the manner of liner combinations. In particular, apart from explaining the idiosyncratic part of measurements, latent factors also characterize the association between any two measurements. As a matter of fact, one can show that the correlation coefficient between y_{itj} and y_{itk} at state z_{it} is given by

$$\mathrm{Corr}\left(y_{itj}, y_{itk} | z_{it} = r\right) = \frac{\sum_{\ell=1}^{m}\sum_{h=1}^{m}\lambda_{rj\ell}\lambda_{rkh}\Phi_{rlh}}{\sqrt{\sum_{\ell=1}^{m}\sum_{h=1}^{m}\lambda_{rj\ell}\lambda_{rjh}\Phi_{r\ell h}+\Psi_{\epsilon rj}}\sqrt{\sum_{\ell=1}^{m}\sum_{h=1}^{m}\lambda_{rk\ell}\lambda_{rkh}\Phi_{r\ell h}+\Psi_{\epsilon rk}}} \tag{3}$$

in which λ_{rjk} is the (j,k)th element of $\mathbf{\Lambda}_r$ and $\Phi_{r,hk}$ is the (h,k)th element in $\mathbf{\Phi}$, respectively. The strength of correlation is identified by the factor loadings and covariance of factors together. In the case when $\boldsymbol{\omega}_{it}$ degenerates to zero (i.e., $\mathbf{\Phi} = \mathbf{0}$) or $\mathbf{\Lambda} = \mathbf{0}$, the association among items disappears and model (2) reduces to p-independent mean-variance models within cluster r. Hence, latent factors play a dominant role in characterizing association of multiple items. Note that, in actual applications, latent factors, though *unobservable*, often have their own physical interpretations. In psychology, for example, latent factors are often used to identify concepts such as treatment, temper, and anxiety, which are important within the framework of theoretical models. The measurements are just proxies for these unobserved concepts of interest. We will provide further interpretations in the real example.

The primary reason for collecting information on multiple occasions for each subject is that it allows investigation of change and/or temporal dependence over time within the subject. There exist various constructs for characterizing dynamic characteristics. A commonly used method is to construct proper dynamic structures for latent factors and establish dynamic factor models, see for example, [19–21]. An alternative choice we adopt here is specifying the joint distribution for state sequences. Following the common routine (see, for example, [22, 23]), we assume that each individual state sequence $\mathbf{z}_i = (z_{i1}, \cdots, z_{iT})$ satisfies the following first-order hidden Markov model

$$p(z_i) = p(z_{i1})\prod_{t=2}^{T} p(z_{it}|z_{i,t-1}) \tag{4}$$

where $p(z_{i1})$ and $p(z_{it}|z_{i,t-1})$ are, respectively, the initial distribution and transition probability given by

$$P(z_{i1} = r|) = \delta_r, \quad P(z_{it} = s|z_{i,t-1} = r) = Q_{rs} \ \ (r,s = 1, \cdots, S) \tag{5}$$

where S is a positive integer, $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_S)$ is an $S \times 1$ vector satisfying $\delta_r \geq 0$ and $\sum_{r=1}^{S}\delta_r = 1.0$, and $\mathbf{Q} = (Q_{rs})$ is an $S \times S$ transition matrix with the (r,s)th entry being Q_{rs}, that is, $Q_{rs} \geq 0$ and

$\sum_{s=1}^{S} Q_{rs} = 1.0$ for $r = 1, \cdots, S$. Modeling state sequences into (5) allows us to explore the transition pattern of individuals across occasions exactly. For example, in the cocaine use data analysis, z_{it} is often identified with the latent state of patient i at time t, then Q_{rs} specifies how individual i being in state r transfers to state s on two successive occasions. Surely, we can relax the time-homogeneous assumption of transition probabilities by including relevant covariates to interpret the inhomogeneous transition behavior among observation data (see, for example, [12, 13, 16]) but at the expense of computational burden.

The current model defined in (2)–(5) provides a comprehensive framework for modeling the multivariate longitudinal data with the latent variables. It accommodates the dynamic behavior of observed sequences, heterogeneity of observed data at the occasion level, and dependence among the multiple items simultaneously. In particular, it makes sense to measure effects of latent factors on the manifest variables quantitatively.

Let $\mathbf{Y}$ be the collection of all observations, and $\boldsymbol{\Omega}$ be the set of corresponding factors. Denote $\mathbf{Z} = \{z_{it} : 1 \le i \le N, 1 \le t \le T\}$ be set of state variables. It follows from Eqs. (2), (4) and (5) that the joint sampling distribution of $\mathbf{Y}, \boldsymbol{\Omega}$, and $\mathbf{Z}$ is given by

$$\begin{aligned} p(\mathbf{Y}, \boldsymbol{\Omega}, \mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{Q}) &= \prod_{i=1}^{N} p(\mathbf{y}_{i1}, \boldsymbol{\omega}_{it} | z_{i1}, \boldsymbol{\theta}) p(z_{i1} | \boldsymbol{\delta}) \prod_{t=2}^{T} p(\mathbf{y}_{it}, \boldsymbol{\omega}_{it} | z_{it}, \boldsymbol{\theta}) p(z_{it} | z_{i,t-1}, \boldsymbol{Q}) \\ &\propto \prod_{i=1}^{N} \prod_{t=1}^{T} \left(\frac{1}{\left| \boldsymbol{\Psi}_{\epsilon z_{it}} \right|^{1/2}} \exp\left\{ -\frac{1}{2} \mathrm{tr} \boldsymbol{\Psi}_{\epsilon z_{it}} \left(\mathbf{y}_{it} - \boldsymbol{\mu}_{z_{it}} - \boldsymbol{\Lambda}_{z_{it}}^{\intercal} \boldsymbol{\omega}_{it} \right)^{\otimes 2} \right\} \right. \\ &\left. \times \frac{1}{\left| \boldsymbol{\Phi}_{z_{it}} \right|^{1/2}} \exp\left\{ -\frac{1}{2} \mathrm{tr} \boldsymbol{\Phi}_{z_{it}}^{-1} \boldsymbol{\omega}_{it}^{\otimes 2} \right\} \right) \times \prod_{i=1}^{N} \prod_{t=1}^{T} \left(\prod_{r=1}^{S} \delta_r^{I\{z_{i1}=r\}} \prod_{s=1}^{S} Q_{rs}^{I\{z_{i,t-1}=r, z_{it}=s\}} \right) \end{aligned} \tag{6}$$

where $\boldsymbol{\theta}$ is formed by free parameters in $\boldsymbol{\mu}_r, \boldsymbol{\Lambda}_r, \boldsymbol{\Psi}_r$, and $\boldsymbol{\Phi}_r$. Here, we write $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\intercal}$ and denote $I(A)$ the indicator function of a set A. The observed data likelihood is then achieved by taking integration of $p(\mathbf{Y}, \boldsymbol{\Omega}, \mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{Q})$ over $\boldsymbol{\Omega}$ and $\mathbf{Z}$, which involves high-dimensional integrations.

3. Posterior inferences

3.1. Prior specifications

Let $\boldsymbol{\mu} = \{\boldsymbol{\mu}_r\}$, $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_r\}$, $\boldsymbol{\Psi}_\epsilon = \{\boldsymbol{\Psi}_{\epsilon r}\}$, and $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_{kr}\}$. For the Bayesian analysis, we need to assign priors to the unknown parameters involved for completing model specification. Since $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, and $\boldsymbol{Q}$ are involved in different submodels, it is natural to assume that $\boldsymbol{\theta}, \boldsymbol{\delta}$, and $\boldsymbol{Q}$ are mutually independent and the components contained in $\boldsymbol{\theta}$ are also mutually independent, that is,

$$p(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}) = p(\boldsymbol{\mu}) p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon) p(\boldsymbol{\Phi}) p(\boldsymbol{\delta}) p(\mathbf{Q}). \tag{7}$$

For the convenience of conjugacy, we assume that the parameters are drawn from the following commonly used conjugate types prior distributions (see for example [24]).

$$p(\boldsymbol{\mu}) = \prod_{r=1}^{S} p(\boldsymbol{\mu}_r) \overset{D}{=} \prod_{r=1}^{S} \mathcal{N}_p(\boldsymbol{\mu}_{0r}, \boldsymbol{\Sigma}_{0r}), \boldsymbol{\Phi} \sim \prod_{r=1}^{S} p(\boldsymbol{\Phi}_r) \overset{D}{=} \prod_{r=1}^{S} \mathcal{W}_m^{-1}(\rho_{0r}, \mathbf{R}_{0r}^{-1}),$$

$$p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon) = \prod_{r=1}^{S} p(\boldsymbol{\Lambda}_r | \boldsymbol{\Psi}_{\epsilon r}) \times p(\boldsymbol{\Psi}_{\epsilon r}) \overset{D}{=} \prod_{r=1}^{S} \prod_{j=1}^{p} \mathcal{N}_m\left(\boldsymbol{\Lambda}_{0rj}, \psi_{\epsilon rj} \mathbf{H}_{\epsilon 0rj}\right) \cdot \mathcal{G}a^{-1}\left(\alpha_{\epsilon 0rj}, \beta_{\epsilon 0rj}\right), \tag{8}$$

$$\boldsymbol{\delta} | \delta_0 \sim \mathcal{D}ir_S(\gamma_0, \ldots, \gamma_0), p(\mathbf{Q}) = \prod_{r=1}^{S} p(\mathbf{Q}_r) \overset{D}{=} \prod_{r=1}^{S} \mathcal{D}ir_S(\nu_0, \cdots, \nu_0)$$

where '$\mathcal{G}a^{-1}(a, b)$' denotes the inverse Gamma distributions with shape $a > 0$ and scale $b > 0$ and '$\mathcal{W}_{m_2}^{-1}(\rho_{0r}, \mathbf{R}_{0r}^{-1})$' represents the q-dimensional inverse Wishart distribution with ρ_{0r} degrees of freedom and $(m_2 \times m_2)$ scale matrix $\mathbf{R}_{0r}$; $\mathbf{Q}_r$ is the rth row vector of $\mathbf{Q}$. The scalars $\alpha_{\epsilon 0rj}$, $\beta_{\epsilon 0rj}$, ρ_{0r}, γ_0, ν_0, the vectors $\boldsymbol{\mu}_{0r}$, $\boldsymbol{\Lambda}_{0rj}$, and the matrices $\mathbf{R}_{0r}$ and $\mathbf{H}_{\epsilon 0rj}$ are assumed to be known. Thus, standard conjugate priors were specified for all parametric components in the model. The conjugate type prior distributions are sufficiently flexible in most applications, and for situations with a reasonable amount of data available, the hyperparameters scarcely affect the analysis. It should be noted that although Eq. (8) allows different hyperparameters for different latent states, in practice, we choose identical priors for all s. Details of hyperparameter choices are discussed later when we present the empirical results.

3.2. Gibbs sampling scheme and posterior analysis

Combining the sampling distribution for the observable $\mathbf{y}_{it}$'s and the prior distribution specified in (8) yields the joint posterior distribution of $\{\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}\}$ given by

$$p(\boldsymbol{\theta},\ \ \boldsymbol{\delta}, \mathbf{Q} | \mathbf{Y}) \propto p(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}) p(\boldsymbol{\theta}) p(\boldsymbol{\delta}) p(\mathbf{Q}) \tag{9}$$

where we ignore the normalization constant $p(\mathbf{Y})$. However, due to the latent factors and state variables present, the computation of $p(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q})$ is intractable since it involves high-dimensional integrals. Consequently, no closed form can be available for the posterior $p(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q} | \mathbf{Y})$. This problem can be addressed via the data augmentation idea in Tanner and Wong [25]. Data augmentation technique treats the latent quantities $\{\boldsymbol{\Omega}, \mathbf{Z}\}$ as the hypothetical missing data and augments them with the observed data to form complete data. The posterior analysis is now carried out based on the joint distribution $p(\boldsymbol{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q} | \mathbf{Y})$, which is proportional to $p(\mathbf{Y}, \boldsymbol{\Omega}, \mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}) p(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q})$, the product of likelihood of complete data and priors. Compared to the intractable observed data likelihood, the complete data likelihood has nice hierarchical structure based on conditional independent assumptions in (2) and (4) and hence is relatively easy to analyze. However, $p(\boldsymbol{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q} | \mathbf{Y})$ is still not in closed form and is thus difficult to deal with analytically. In this regard, simulation-based methods can be used to generate observations to carry out posterior analysis. In view of the multiple components involved, the usual independent sampling methods are not feasible. Note that, on the basis of complete data, the full conditional distributions of $\mathbf{\Omega}$, $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, and Q have closed forms. This provides a solid foundation for Markov chain Monte Carlo methods. Markov chain Monte Carlo sampling does not draw observations from $p(\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}|\mathbf{Y})$ directly. On the contrary, it generates observations from the full conditionals of each component alternatively, thus forming the dependent sample, i.e., Markov chains. Specifically, as pointed out in the introduction, we use Gibbs sampler [4, 5] to draw observations from this target distribution. Obviously, the sampling scheme in the Gibbs sampler includes two types of moves: updating the components involved in the factor analysis model and updating the components related to the hidden Markov model. We propose using the following Gibbs sampler which iteratively simulates from the conditional distributions, where variables are removed from the conditioning set either by explicit integration or by conditional independence. The steps involved in the Gibbs sampler are

Step a: Generate $\mathbf{Z}$ from $p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}, \mathbf{\Omega}, \mathbf{Y})$

Step b: Generate $\mathbf{\Omega}$ from $p(\mathbf{\Omega}|\mathbf{Z}, \boldsymbol{\theta}, \mathbf{Y})$

Step c: Generate $\{\boldsymbol{\mu}, \mathbf{\Lambda}, \mathbf{\Psi}_\epsilon\}$ from $p(\boldsymbol{\mu}, \mathbf{\Lambda}, \mathbf{\Psi}_\epsilon|\mathbf{Z}, \mathbf{\Omega}, \mathbf{Y})$

Step d: Generate $\mathbf{\Phi}$ from $p(\mathbf{\Phi}|\mathbf{Z}, \mathbf{\Omega})$

Step e: Generate $\boldsymbol{\delta}$ from $p(\boldsymbol{\delta}|\mathbf{Z})$

Step f: Generate $\mathbf{Q}$ from $p(\mathbf{Q}|\mathbf{Z})$

Under mild conditions and similar to [4] (see also, for example, [26]), one can show that for sufficiently large b, say B_0, the joint distribution of $\left\{\mathbf{\Omega}^{(b)}, \mathbf{Z}^{(b)}, \boldsymbol{\theta}^{(b)}, \boldsymbol{\delta}^{(b)}, \mathbf{Q}^{(b)}\right\}$ converges at an exponential rate to the desired posterior distribution $p(\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}|\mathbf{Y})$. Hence, $p(\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}|\mathbf{Y})$ can be approximated by the empirical distribution of $\left\{\mathbf{\Omega}^{(b)}, \mathbf{Z}^{(b)}, \boldsymbol{\theta}^{(b)}, \boldsymbol{\delta}^{(b)}, \mathbf{Q}^{(b)}\right\} : b = B_0 + 1, \cdots, B_0 + B\}$ where B is chosen to give sufficient precision to the empirical distribution. The convergence of the Gibbs sampler can be monitored by the 'estimated potential scale reduction (EPSR)' values as suggested by Gelman and Rubin [27] or by plotting the traces of estimates against iterations under different starting values.

Simulated observations obtained from the posterior can be used for statistical inferences via straightforward analysis procedures. For brevity, let $\left\{\boldsymbol{\theta}^{(b)}, \boldsymbol{\delta}^{(b)}, \mathbf{Q}^{(b)}, \mathbf{\Omega}^{(b)}, \mathbf{Z}^{(b)}\right\}$ be the random observations generated by the Gibbs sampler from $p(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}, \mathbf{\Omega}, \mathbf{Z}|\mathbf{Y})$. The joint Bayesian estimate of $\boldsymbol{\theta}$ and $\mathbf{\Omega}$ can be obtained easily via the corresponding sample means of the generated observations as follows:

$$\widehat{\boldsymbol{\theta}} = (B-1)^{-1}\sum_{b=1}^{B}\boldsymbol{\theta}^{(b)}, \widehat{\mathbf{\Omega}} = (B-1)^{-1}\sum_{b=1}^{B}\mathbf{\Omega}^{(b)}, \widehat{\mathbf{Z}} = (B-1)^{-1}\sum_{b=1}^{B}\mathbf{Z}^{(b)}. \tag{10}$$

Clearly, these Bayesian estimates are consistent estimates of the corresponding posterior means, see [26]. The consistent estimates of covariance matrix of estimates can be obtained as follows:

$$\mathrm{Cov}\widehat{(\boldsymbol{\theta}|\mathbf{Y})} = (B-1)^{-1}\sum_{b=1}^{B}\left(\boldsymbol{\theta}^{(b)} - \widehat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}^{(b)} - \widehat{\boldsymbol{\theta}}\right)^{\intercal} \tag{11}$$

$$\mathrm{Cov}\widehat{(\boldsymbol{\Omega}|\mathbf{Y})} = (B-1)^{-1}\sum_{b=1}^{B}\left(\boldsymbol{\Omega}^{(b)} - \widehat{\boldsymbol{\Omega}}\right)\left(\boldsymbol{\Omega}^{(b)} - \widehat{\boldsymbol{\Omega}}\right)^{\intercal} \tag{12}$$

Hence, the standard error estimates can be obtained conveniently by the Gibbs sampler algorithm. Other statistical inferences about $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ such as deriving the confidence intervals and statistics for hypothesis testing can be achieved based on the simulative observations as well (see, for example, [28, 29]).

One important statistical inference beyond estimation is on testing of various hypotheses about the model. In the field of hidden Markov modeling, determining the proper number of states may be the first step towards data analysis. Too many states may overfit the observations, meaning that it can fit the training data accurately but may not be a good model for underlying data-generating process. On the other hand, too few states may not be flexible enough to approximate the underlying model. In the context of Bayesian model selection, Bayes factor (BF, [30]) is a popular choice for model comparison. BF is defined as the ratio of the marginal likelihoods of data under two competing models. However, the computation of BF is difficult since it often involves the high-dimensional integrations. It has also been shown that BF is sensitive to the choice of priors and will become infeasible when improper priors are used. A simple and more convenient alternative is the L_ν-measures [31–34] which is based on the posterior predictive density. It has been shown [34] that this approach is conceptually and computationally simple and is useful in model checking for wide varieties of complicated situations. Moreover, the required computation is a by-product of the common Bayesian simulation procedures such as the Gibbs sampler or its related algorithms. Specifically, let $\mathbf{Y}^{\mathrm{rep}}$ denotes future values of $\mathbf{Y}$ in a replicate experiment, that is, $\mathbf{Y}^{\mathrm{rep}}$ has the same sampling density as that of $\mathbf{Y}$. The posterior predictive distribution $p(\mathbf{Y}^{\mathrm{rep}}|\mathbf{Y})$ is defined as

$$p(\mathbf{Y}^{\mathrm{rep}}|\mathbf{Y}) = \int p(\mathbf{Y}^{\mathrm{rep}}|\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q})p(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}|\mathbf{Y})d\boldsymbol{\theta}d\boldsymbol{\delta}d\mathbf{Q} \tag{13}$$

Naturally, if the posited model under consideration is the true model in the sense that from which the data are generated, then $\mathbf{Y}^{\mathrm{rep}}$ would behave like data $\mathbf{Y}$ and its squared biases and covariances should be small. With this notion in mind, Ibrahim, Chen, and Sinha [34] proposed an L statistics to assess the fitness of posited models to the data by weighting the squared biases and covariance, which can be interpreted as a trade-off between them. Here, we extend it to the multivariate longitudinal setting. Let $\mathbf{Y}^{\mathrm{rep}} = \left(\mathbf{y}_1^{\mathrm{rep}\intercal}, \cdots, y_N^{\mathrm{rep}\intercal}\right)^{\intercal}$ be a collection set of future responses in our proposal. For some $0 \leq \nu < 1$, we consider the following multivariate version of L_ν-measure:

$$L_\nu(\mathbf{Y}) = \sum_{i=1}^{N}\mathrm{tr}\left[\mathbb{C}ov\left(\mathbf{y}_i^{\mathrm{rep}}|\mathbf{Y}\right)\right] + \nu\sum_{i=1}^{N}\mathrm{tr}\left[\left\{\mathbb{E}\left(\mathbf{y}_i^{\mathrm{rep}}|\mathbf{Y}\right) - \mathbf{y}_i\right\}\left\{\mathbb{E}\left(\mathbf{y}_i^{\mathrm{rep}}|\mathbf{Y}\right) - \mathbf{y}_i\right\}^{\intercal}\right] \tag{14}$$

where the expectation is taken with respect to the posterior predictive distribution. Clearly, small values of the L_ν-measure indicate that the model gives predictions close to the observed values, and the variability in the predictions is low as well. Hence, the model with the smallest L_ν-measure is selected from a collection of competing models. It has been shown that L_ν-measure with $\nu = 0.5$ has nice theoretical properties [34]. Thus, this value of ν will be used in our empirical illustrations.

4. Cocaine use data analysis

In this section, a small portion of cocaine use data is analyzed to illustrate the practical value of the proposed methodology. The original data are collected from 321 cocaine use patients who were admitted in 1988–1989 to the West Los Angeles Veterans Affairs Medical Center. The whole data constitute 68 measurements of 17 items, which were recorded at four time points: at baseline, 1 year after the treatment, 2 years after the treatment, and 12 years after the treatment in 2002–2003. These measurements cover the information on the cocaine use, treatment received, psychological problems, social status, employments, and so on. As an illustration, three variables are selected to conduct data analysis: 'y_1 : days of cocaine use per month at intake (CC)', 'y_2 : times per month in formal treatment (FT)', and 'y_3 : months in formal treatment (MFT)', which, respectively, represent the severity of cocaine use and the levels of treatment received by a patient. Since these variables were measured in 0–120 points scale, to unify the scales, we take logarithms and standardize them. Among them, some measurements are missing. The missing proportion is about 8.4%. For brevity, we assume that the missing is missing at random [35]. A distinct characteristic underlying data are nonnormal and heavy tailed. **Figure 1** gives the plots of histograms and the posterior predictive density estimates (see below) of logarithms of CC, FT, and MFT (with missing data removed) on four occasions. The histograms illustrate that the distributions of selected variables are deviated from normality in terms of multimodality and skewness. The skewness and kurtosis of CC on four occasions are $\{-1.631, 5.031\}$, $\{-0.847, 3.354\}$, $\{0.328, 1.476\}$, and $\{-0.473, 2.467\}$, respectively. Data set also demonstrates dynamic characteristics. The distribution of CC, for instance, is skewed to the left at baseline and moves to the right gradually on the following two occasions and becomes right-skewed eventually. This implies that a single factor analysis model may not be appropriate to fit the data at each time point.

In this analysis, one of the objectives is to explore the effects of latent factors on the observed variables and assess the dependence among latent factors. Based on the nature of the problem under consideration, it is natural to group the single variable 'CC' to reflect one latent factor 'cocaine use' (η) and to group 'FT' and 'MFT' to represent another latent factor 'treatment' (ξ). Let $\mathbf{y}_{it} = (y_{it1}, y_{it2}, y_{it3})^\intercal$ and $\boldsymbol{\omega}_{it} = (\eta_{it}, \xi_{it})^\intercal$. To be convenient for interpretation and computation, $\boldsymbol{\Phi}_r$ and $\boldsymbol{\Lambda}_r$ are restricted to be invariant across states but leave the baseline level $\boldsymbol{\mu}_r$ varying with r. Further, the following non-overlapped structure for factor loading matrix is considered

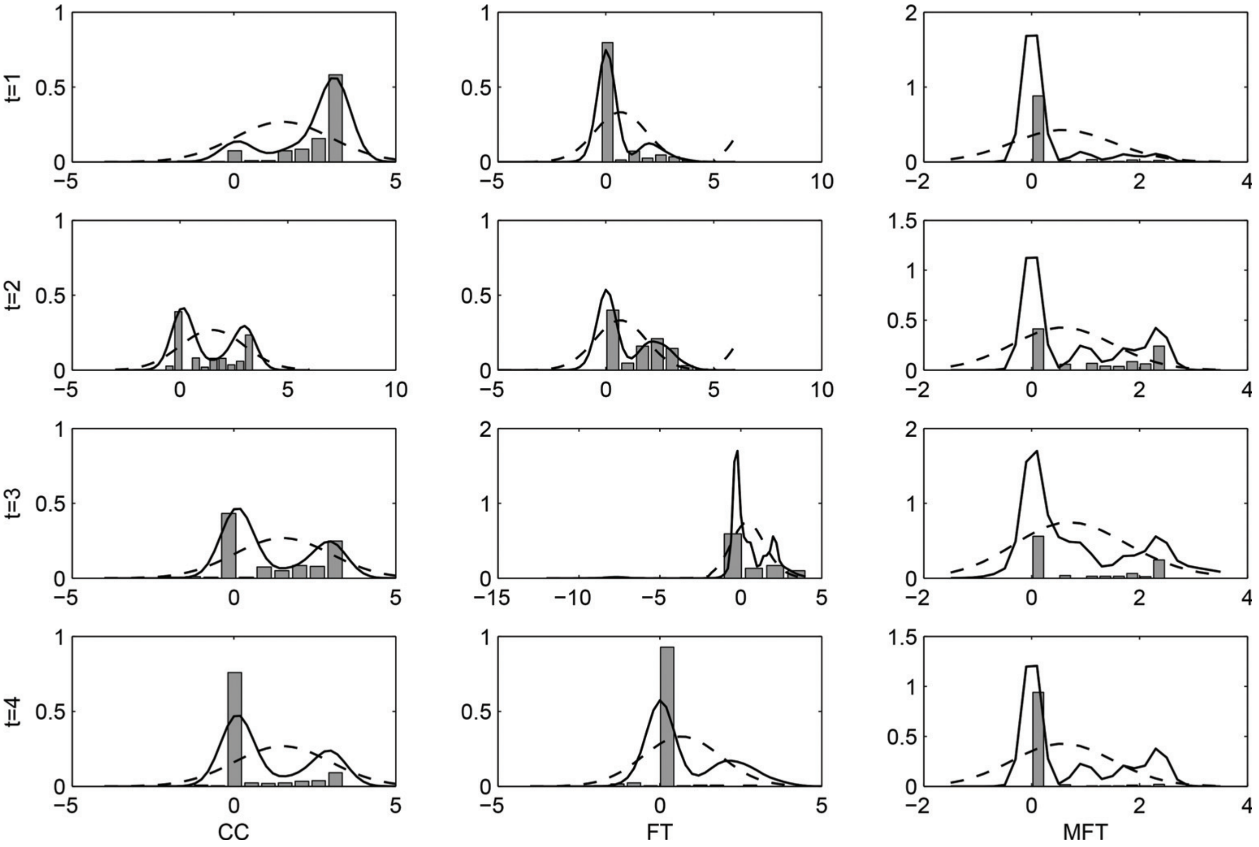

Figure 1. Plots of histograms and posterior predictive density estimates of 'CC', 'FT' and 'MFT' under FA model and hidden Markov CFA model with seven states in the cocaine use data analysis: the dashed lines denote CFA and the solid lines represent the hidden Markov FA.

$$\boldsymbol{\Lambda}^{\top} = \begin{pmatrix} 1^{*} & 0^{*} & 0^{*} \\ 0^{*} & 1^{*} & \Lambda_{32} \end{pmatrix} \tag{15}$$

where parameters with an asterisk are treated as fixed for identification. Note that fixing $\Lambda_{11} = 1$ indicates that η is identified with CC. This is similar to that in Λ_{22}. Hence, in this case, Φ_{12} in $\boldsymbol{\Phi}$ measures the magnitude of dependence of ξ on η.

Data set is fitted to the proposed models with 10 different transition models: $S = 1, \cdots, 10$. Although these state spaces are in nested forms, the corresponding models are not, since one cannot be reduced to another by constraining parameters in the interior of parameter space. This indicates that chi-square distribution may not be suitable for the classic likelihood ratio test statistic. We use L-measure to implement model selection. Obviously, if $\mathbb{S}_1$ is taken, then the proposed model reduces to common factor analysis model (CFA, [18]).

The following inputs are taken for the super-parameters involved in the prior distributions (8): for $r = 1, \cdots, S$, $\mu_{0rj} = \min\left\{y_{itj}\right\} + r/S$, $\boldsymbol{\Sigma}_{0r} = \mathbf{S}_{yy}/S$, where $\mathbf{S}_{yy}$ is the sample covariance matrix of data. The entries in $\boldsymbol{\Lambda}_0$ are set to be zeros, $\rho_0 = 10.0$, $\mathbf{R}_0^{-1} = 7.0 \times \mathbf{I}_2$, which leads to the mean of $\boldsymbol{\Phi}$ equal to $\mathbf{I}_2$, $\mathbf{H}_{\epsilon 0} = \mathbf{I}_3$, $\alpha_{\epsilon 0j} = 9.0$, $\beta_{\epsilon 0j} = 8.0$, $\nu_0 = \gamma_0 = 0.1$. Note that these values are the standard inputs in the latent variable analysis (see [24]). We also took other values for these inputs and found that the resulting estimates are scarcely affected.

We implement the proposed algorithm given in Section 3 to conduct Bayesian analysis. Let $\mathbf{Y}_{obs}$ be the collection of observed data and $\mathbf{Y}_{mis}$ be the set of missing data. Due to the missing data, we need to draw $\mathbf{Y}_{mis}$ from $p(\mathbf{Y}_{mis}|\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{Y}_{obs})$ in MCMC sampling. This can be implemented easily since conditioning on $\mathbf{\Omega}$, $\mathbf{Z}$, and $\boldsymbol{\theta}$, $p(\mathbf{Y}_{mis}|\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{Y}_{obs})$, independent of $\mathbf{Y}_{obs}$, has the normal distribution. Hence, drawing $\mathbf{Y}_{mis}$ is rather straightforward and fast. To obtain some idea about the number of the Gibbs sampler iterations in getting convergence, we conducted a few test runs as a pilot study and found that in all these runs, the Gibbs sampler converged in about 1000–2000 iterations, where the EPSR values [27] are less than 1.2. So, for all cases under consideration, we collect 3000 random observations after initial 2000 iterations being removed for posterior analysis.

We calculate the values of $L_{0.5}$ under each fitting. For computation, we use simulation-based method by drawing predictive values $\mathbf{Y}_{obs}^{rep}$ from $p\left(\mathbf{Y}_{obs}^{rep}|\mathbf{Y}_{obs}\right)$, where $\mathbf{Y}_{obs}^{rep}$ is the hypothetical replication of $\mathbf{Y}_{obs}$. Note that $p\left(\mathbf{Y}_{obs}^{rep}|\mathbf{Y}_{obs}\right) = \int p\left(\mathbf{Y}_{obs}^{rep}|\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}\right) p(\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}|\mathbf{Y}_{obs})\ d\mathbf{\Omega} d\mathbf{Z} d\boldsymbol{\theta}$. Hence, drawing $\mathbf{Y}_{obs}^{rep}$ is rather easy when $\mathbf{\Omega}$, $\mathbf{Z}$, and $\boldsymbol{\theta}$ are available. Given that we have M simulations from the posterior of $\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}$ via MCMC sampling discussed before, we just draw one $\mathbf{Y}_{obs}^{rep}$ from $p\left(\mathbf{Y}_{obs}^{rep}|\mathbf{\Omega}, \mathbf{Z}, \boldsymbol{\theta}\right)$ for each $\mathbf{\Omega}, \mathbf{Z}$, and $\boldsymbol{\theta}$ and obtain M simulations in the end for $\mathbf{Y}_{obs}^{rep}$. Based on these simulated observations, L_v measures can be estimated consistently via sample means. We draw 3000 observations after convergence of MCMC algorithm for calculating $L_{0.5}$ and the results are reported in **Table 1**.

Examination of **Table 1** indicates that the proposed model with six to eight latent states seems to give better fits to the data. Furthermore, we calculate the posterior predictive density estimates of $y_{tj}(t = 1, \cdots, 4, j = 1, \cdots, 3)$ under one state and seven states, respectively (see **Figure 1**). It can be seen clearly that our proposed method is successful in capturing the skewness and modes of data while factor analysis model fails. For the computation details, we choose 60–100 equally spaced grids in the interval $\left[\min\left\{y_{obs,itj}\right\} - 1.0, \max\left\{y_{obs,itj}\right\} + 1.0\right]$ and collect 3000 simulated observations from the Gibbs sampler at each point after removing initial 2000 iteration as burn-ins.

Table 2 presents the summary of Bayesian estimates of unknown parameters and their standard errors using the formula given in (11) with $S = 1$ (denoted by FA) and $S = 7$ (denoted by HMFA). For comparison, maximum likelihood estimates of unknown parameters with their standard deviations under HMFA are also presented in **Table 2**. The maximum likelihood

Model	$L_{0.5}$	**Model**	$L_{0.5}$
$S = 1$	2322.447	$S = 6$	590.448
$S = 2$	2107.514	$S = 7$	572.172
$S = 3$	1030.264	$S = 8$	597.843
$S = 4$	941.230	$S = 9$	932.763
$S = 5$	839.726	$S = 10$	1030.264

Table 1. Summary of $L_{0.5}$ under competing models in the analysis of cocaine use data.

	FA		ML		HMFA	
Para.	**Est.**	**SD**	**Est.**	**SD**	**Est.**	**SD**
Λ_{32}	0.001	0.014	0.196	0.029	0.752	0.045
$\Psi_{\varepsilon 1}$	1.443	0.315	0.559	0.297	0.432	0.049
$\Psi_{\varepsilon 2}$	0.439	0.056	0.204	0.039	0.339	0.034
$\Psi_{\varepsilon 3}$	0.305	0.030	0.008	NAN	0.025	0.001
Φ_{11}	0.770	0.315	0.510	0.132	0.346	0.049
Φ_{12}	−0.018	0.018	−0.053	0.041	−0.182	0.052
Φ_{22}	1.007	0.080	0.312	0.053	0.219	0.033

Table 2. Summary statistics for Bayesian and ML estimates in the cocaine use data analysis.

analysis is conducted via MCECM algorithm [36] and the standard error estimates are calculated via Louis formula [37].

Based on **Table 2**, we can find the following facts: First, three estimates of Λ_{32} give the positive effects of latent factor ξ on the 'MFT'. This is not surprising since ξ is related to the treatment level of a patient received. But there are obvious differences in magnitudes among the three methods. For FA and HMFA, the former gives $\widehat{\Lambda}_{32} = 0.001$ associated with standard deviation 0.014, while the latter gives $\widehat{\Lambda}_{32} = 0.752$ with standard deviation 0.045. This reflects that the heterogeneity of data affects the estimates $\widehat{\Lambda}_{32}$ seriously. Compared to the previous two methods, ML method produces that $\widehat{\Lambda}_{32} = 0.196$ with SD = 0.029, which are in between them. Second, the estimates of variance parameters $\Psi_{\epsilon j}$ under $S = 1$ are larger than those under $S = 7$. This indicates that factor analysis model accommodates heavy tails of data at the expense of variance inflation. Further investigations on the estimates of Φ_{jj} under FA and HMFA also reveal the same phenomenon as that of $\Psi_{\epsilon j}$. However, we observe that the ML estimate of $\Psi_{\epsilon 3}$, the unique variance corresponding to the third item, is equal to 0.008 with SD = NAN, an illogical number, which is very close to an improper Heywood case. As pointed out by Lee [18], Heywood cases in the ML estimation can be avoided by imposing an inequality constraint on $\Psi_{\epsilon 3}$ with a penalty function. In the Bayesian approach, the conjugate prior distribution of $\Psi_{\epsilon 3}^{-1}$ specified $\Psi_{\epsilon 3}$ in a region of positive values and hence has a similar effect as adding a penalty function. Hence, no Heywood cases are found in the Bayesian solution because of the penalty function induced by the prior distribution on $\Psi_{\epsilon 3}^{-1}$. Third, three estimates give the negative correlation between η and ξ, which is consistent with the fact that the improvement of treatment will decrease the intensity of cocaine use, thus leading to a decrease of cocaine use in days. ML estimates for Φ_{jk} are very close to those under HMFA. However, the estimate of Φ_{12} under $S = 1$ is -0.018, which is quite different from -0.182 for $S = 7$. Furthermore, the coefficients of correlation of ξ and η under $S = 1$ and $S = 7$ are -0.0204 and -0.6612, respectively. The former suggests that ξ and η are approximately independent while the latter implies stronger dependence between them.

Moreover, we computed the posterior probabilities $P(z_t = r|\mathbf{Y}_{obs})$ for $r = 1, \cdots, 7$ and $t = 1, \cdots, 4$ under $S = 7$ based on 10,000 simulated observations drawn from $p(\mathbf{Z}|\mathbf{Y}_{obs})$ and found that the transition path corresponding to the maximum posterior probability is $7 \rightarrow 1 \rightarrow 1 \rightarrow 1$. This implies that latent state of the patient being in is extremely serious at baseline and becomes moderate in the subsequent treatments. This also reflects a positive effect of intervention on the patient's latent state. Note that unlike the common Viterbi algorithm in exploring the optimal transition path of states in ML analysis, calculating posterior probability $P(z_t = r|\mathbf{Y}_{obs})$ within Bayesian framework is a by-product of the estimation procedure. This voids the complex computation of marginal likelihood of the observed data and hence is very fast.

5. Discussion

This chapter reviews Bayesian inferences within a general framework and proposes a Bayesian procedure for analyzing hidden Markov factor analysis model under multivariate longitudinal setting. Compared to ML method, the pragmatic advantage of Bayesian framework is its flexibility and generality for coping with very complex problems. When good prior information can be available, results obtained from Bayesian method are more reliable and accurate than that under ML. With increased access to computation advances in simulation-based approaches, in particular the MCMC methodology, Bayesian inferences provide enormous scope for realistic statistical modeling.

Although we concentrate our attention on applications of the hidden Markov factor analysis model, the methodology developed in this chapter can be extended to the case where the LVM is nonlinear. Another possible extension is to consider a dynamic LVM, wherein model parameters vary over time. These extensions will raise theoretical and computational challenges and certainly require further investigation.

Acknowledgements

The authors are grateful to the editor's valuable suggestions and comments which have greatly improved the manuscript. Xia's work was fully supported by grant from National Nature Science Foundation of China (No. 11471161) and Tang's work was supported by National Science Fund for Distinguished Young Scholars of China (No. 11225103).

A. Appendix. Full conditionals

(a) $p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{Q}, \boldsymbol{\Omega}, \mathbf{Y})$

Let $\boldsymbol{\omega}_i$ denote the sequence of latent factors across T occasions for individual i. To draw state variables $\mathbf{Z}$ from $p(\mathbf{Z}|\boldsymbol{\delta}, \mathbf{Q}, \boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{Y})$, we first notice that

$$p(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\Omega}, \mathbf{Y}) = \prod_{i=1}^{N} p\left(\mathbf{z}_i|\boldsymbol{\omega}_i, \boldsymbol{\delta}, \mathbf{Q}, \boldsymbol{\theta}, \mathbf{y}_i\right) \tag{16}$$

Hence, drawing $\mathbf{Z}$ can be accomplished via single-component method by drawing $\mathbf{z}_i$ independently from $p\left(\mathbf{z}_i|\boldsymbol{\omega}_i, \boldsymbol{\delta}, \mathbf{Q}, \boldsymbol{\theta}, \mathbf{y}_i\right)$. Furthermore, notice that the sequences $\left\{\mathbf{y}_i, \boldsymbol{\omega}_i, \mathbf{z}_i\right\}$ are still the one-order Markov sequences. Hence, we can simulate $\mathbf{z}_i$ through a well-known forward filtering-backward sampling algorithm (see, for example, [38]). For notation clarity, we suppress $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, and $\mathbf{Q}$ in the following derivations.

Forward filtering-backward sampling (FFBS) consists of first forward filtering (FF) and then backward sampling (BS). The forward filtering step recursively updates

$$\alpha_{i,\,t|t} = p\left(z_{it}|\,\boldsymbol{\omega}_{i,\,1:t}, \mathbf{y}_{i,\,1:t}\right), \quad t = 1, \ldots, T. \tag{17}$$

Here $\mathbf{y}_{i,\,1:t}$ represents the set of observations of subject i up to time t and so are $\boldsymbol{\omega}_{i,\,1:t}$ and $\mathbf{z}_{i,\,1:t}$. The backward sampling is to draw $\mathbf{z}_i$ from the joint distribution of the states given the data using

$$p\left(\mathbf{z}_{i,\,1:T}|\,\boldsymbol{\omega}_{i,\,1:T}, \mathbf{y}_{i,\,1:T}\right) = p\left(z_{iT}|\,\boldsymbol{\omega}_{i,\,1:T}, \mathbf{y}_{i,\,1:T}\right) \ldots p\left(z_{i1}|\mathbf{z}_{i,\,2:T}, \boldsymbol{\omega}_{i,\,1:T}, \mathbf{y}_{i,\,1:T}\right). \tag{18}$$

That is, we first draw the last state given all the data and then work backwards in time drawing each state conditional on all the subsequent ones.

To implement forward filtering, let

$$\alpha_{it}(r) = \mathbb{P}\left(\mathbf{y}_{i,\,1:t}, \boldsymbol{\omega}_{i,\,1:t}, z_{it} = r\right), \quad t = 1, \cdots, T \tag{19}$$

Obviously, $\alpha_{i1}(r) = \delta_r p\left(\mathbf{y}_{i1}, \boldsymbol{\omega}_{i1}|z_{i1} = r\right)$. Moreover, it can be shown that

$$\alpha_{it}(r) = \left(\sum_{s=1}^{S} \alpha_{it-1}(s) q_{sr}\right) p\left(\mathbf{y}_{it}, \boldsymbol{\omega}_{it}|z_{it} = r\right), \quad t = 2, \cdots, T \tag{20}$$

The outputs $\{\alpha_{it}\}_{t=1}^{T}$ from recursive Eq. (20) can be used to calculate the posterior probability

$$\alpha_{i,\,t|t}(r) = \mathbb{P}\left(z_{it} = r|\,\boldsymbol{\omega}_{i,\,1:t}, \mathbf{y}_{i,\,1:t}\right) = \frac{\alpha_{it}(r)}{\sum\limits_{s=1}^{S} \alpha_{it}(s)} \tag{21}$$

which leads to the forward filtering (FF) iteration.

The backward sampling step depends on the observation that

$$
\begin{aligned}
&p\left(z_{it}|\mathbf{z}_{i,t+1:T},\boldsymbol{\omega}_{i,1:T},\mathbf{y}_{i,1:T}\right)\propto p\left(z_{it},\mathbf{z}_{i,t+1:T},\boldsymbol{\omega}_{i,1:T},\mathbf{y}_{i,1:T}\right)\\
&=p\left(z_{it},\boldsymbol{\omega}_{i,1:t},\mathbf{y}_{i,1:t}\right)p\left(\mathbf{z}_{i,t+1:T},\boldsymbol{\omega}_{i,t+1:T},\mathbf{y}_{i,t+1:T}|z_{it},\boldsymbol{\omega}_{i,1:t},\mathbf{y}_{i,1:t}\right)\\
&=p\left(z_{it},\boldsymbol{\omega}_{i,1:t},\mathbf{y}_{i,1:t}\right)p\left(\mathbf{z}_{i,t+1:T}|z_{it},\boldsymbol{\omega}_{i,1:t},\mathbf{y}_{i,1:t}\right)p\left(\boldsymbol{\omega}_{i,t+1:T},\mathbf{y}_{i,t+1:T}|\mathbf{z}_{i,t+1:T},z_{it},\boldsymbol{\omega}_{i,1:t},\mathbf{y}_{i,1:t}\right)\\
&=p\left(z_{it},\boldsymbol{\omega}_{i,1:t},\mathbf{y}_{i,1:t}\right)p(\mathbf{z}_{i,t+1:T}|z_{it},)p\left(\boldsymbol{\omega}_{i,t+1:T},\mathbf{y}_{i,t+1:T}|\mathbf{z}_{i,t+1:T}\right)
\end{aligned}
\tag{22}
$$

The last equation holds since given z_{it}, $\left\{\mathbf{y}_{i,t:T},\boldsymbol{\omega}_{i,t:T},\mathbf{z}_{i,t+1:T}\right\}$ does not depend on the previous values due to the Markov Chain characteristics of $\left\{\mathbf{y}_{it},\boldsymbol{\omega}_{it},\mathbf{z}_{it}\right\}$. This leads to

$$
\mathbb{P}\left(z_{it}=r|\mathbf{z}_{i,t+1:T},\mathbf{y}_{i,1:T},\boldsymbol{\omega}_{i,1:T},\right)=\frac{\alpha_{i,t|t}(r)q_{rz_{i,t+1}}}{\sum_{s=1}^{S}\alpha_{i,t|t}(s)q_{sz_{i,t+1}}}\qquad t=T-1,\cdots,1. \tag{23}
$$

Hence, FFBS algorithm for drawing $\mathbf{z}_i$ is implemented by

Algorithm:

i. *running the recursion α_{it} and stored the conditional probabilities $\alpha_{i,t|t}$ for $t=1,\ldots,T$;*

ii. *sampling z_{iT} from the filtered conditional probability $\alpha_{i,T|T}$;*

iii. *for $t=T-1,\cdots,1$, sampling z_{it} from the conditional probability*

$$
\mathbb{P}\left(z_{it}=r|\,\boldsymbol{\omega}_{i,1:T},\mathbf{y}_{i,1:T},\mathbf{z}_{i,t+1:T}\right). \tag{24}
$$

(b) $p(\boldsymbol{\Omega}|\mathbf{Z},\boldsymbol{\theta},\mathbf{Y})$

To draw $\boldsymbol{\Omega}$, we first note that

$$
p(\boldsymbol{\Omega}|\mathbf{Z},\boldsymbol{\theta},\mathbf{Y})=\prod_{i=1}^{N}\prod_{t=1}^{T}p\left(\boldsymbol{\omega}_{it}|z_{it},\boldsymbol{\theta},\mathbf{y}_{it}\right) \tag{25}
$$

in which

$$
p\left(\boldsymbol{\omega}_{it}|z_{it},\boldsymbol{\theta},\mathbf{y}_{it}\right)\propto\exp\left\{-\frac{1}{2}\left(\mathbf{y}_{it}-\boldsymbol{\mu}_r-\boldsymbol{\Lambda}_r\boldsymbol{\omega}_{it}\right)^{\intercal}\boldsymbol{\Psi}_{\epsilon r}^{-1}\left(\mathbf{y}_{it}-\boldsymbol{\mu}_{kr}-\boldsymbol{\Lambda}_r\boldsymbol{\omega}_{it}\right)-\frac{1}{2}\boldsymbol{\omega}_{it}^{\intercal}\boldsymbol{\Phi}_r^{-1}\boldsymbol{\omega}_{it}\right\} \tag{26}
$$

with $r=z_{it}$. Hence, similar to that in drawing $\mathbf{Z}$, updating $\boldsymbol{\Omega}$ can be achieved by drawing $\boldsymbol{\omega}_{it}$ independently from $p\left(\boldsymbol{\omega}_{it}|z_{it},\boldsymbol{\theta},\mathbf{y}_{it}\right)$ for $i=1,\cdots,N$ and $t=1,\cdots,T$. It can be shown that

$$
p\left(\boldsymbol{\omega}_{it}|z_{it}=r,\boldsymbol{\theta},\mathbf{y}_{it}\right)\overset{D}{=}\mathcal{N}_m\left(\widehat{\mathbf{m}}_{it},\widehat{\boldsymbol{\Sigma}}_r\right). \tag{27}
$$

in which

$$\widehat{\mathbf{m}}_{it} = \widehat{\boldsymbol{\Sigma}}_r \boldsymbol{\Lambda}_r^{\intercal} \boldsymbol{\Psi}_{\epsilon r}^{-1} \big(\mathbf{y}_{it} - \boldsymbol{\mu}_r\big), \;\; \widehat{\boldsymbol{\Sigma}}_r = \big(\boldsymbol{\Lambda}_r^{\intercal} \boldsymbol{\Psi}_{\epsilon r}^{-1} \boldsymbol{\Lambda}_r + \boldsymbol{\Phi}_r^{-1}\big)^{-1}. \tag{28}$$

(c) $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon | \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y})$

To draw $\{\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon\}$, we can first draw $\boldsymbol{\mu}$ from $p(\boldsymbol{\mu} | \boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y})$ and then draw $\{\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon\}$ from $p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon | \boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y})$. For this end, let $\widehat{n}^{(r)} = \#\{z_{it} = r\}$ be the size of cluster z_{it}, and let $w_{itr} = I\{z_{it} = r\}$. Denote

$$\begin{aligned} \overline{\mathbf{Y}}^{(r)} &= \sum_{i=1}^{N}\sum_{t=1}^{T} w_{itr}\mathbf{y}_{it}/\widehat{n}^{(r)}, \quad \overline{\boldsymbol{\Omega}}^{(r)} = \sum_{i=1}^{N}\sum_{t=1}^{T} w_{itr}\,\boldsymbol{\omega}_{it}/\widehat{n}^{(r)}, \\ \mathbf{S}_{yy}^{(r)} &= \sum_{i=1}^{N}\sum_{t=1}^{T} w_{itr}\mathbf{y}_{it}\mathbf{y}_{it}^{\intercal}/\widehat{n}^{(r)}, \;\; \mathbf{S}_{\omega y}^{(r)} = \sum_{i=1}^{N}\sum_{t=1}^{T} w_{itr}\,\boldsymbol{\omega}_{it}\mathbf{y}_{it}^{\intercal}/\widehat{n}^{(r)}, \end{aligned} \tag{29}$$

be the sample means and covariance matrices of $\mathbf{Y}$ and $\boldsymbol{\Omega}$ within the rth cluster, respectively. By some algebra calculations, it can be shown that

$$\begin{aligned} p(\boldsymbol{\mu} | \boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon, \mathbf{K}, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}) &= \prod_{r=1}^{S} p\big(\boldsymbol{\mu}_r | \boldsymbol{\Lambda}_r, \boldsymbol{\Psi}_{\epsilon r}, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}\big), \;\; \text{and} \\ p(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_\epsilon | \boldsymbol{\mu}, \mathbf{K}, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}) &= \prod_{r=1}^{S} p\big(\boldsymbol{\Lambda}_r, \boldsymbol{\Psi}_{\epsilon r} | \boldsymbol{\mu}_r, \boldsymbol{\Omega}, \mathbf{Z}, \mathbf{Y}\big), \end{aligned} \tag{30}$$

where

$$\begin{aligned} p\big(\boldsymbol{\mu}_r | \boldsymbol{\Lambda}_r, \boldsymbol{\Psi}_{\epsilon r}, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}\big) &= \mathcal{N}_p\Big(\widehat{\mathbf{a}}_{\mu r}, \widehat{\boldsymbol{\Sigma}}_{\mu r}\Big), \\ p\big(\boldsymbol{\Lambda}_r, \boldsymbol{\Psi}_{\epsilon r} | \boldsymbol{\mu}_r, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}\big) &= \prod_{j=1}^{p} p(\Psi_{\epsilon rj} | \boldsymbol{\mu}_r, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}) p\big(\boldsymbol{\Lambda}_{rj} | \Psi_{\epsilon rj} | \boldsymbol{\mu}_r, \mathbf{Z}, \boldsymbol{\Omega}, \mathbf{Y}\big) \\ &\overset{D}{=} \prod_{j=1}^{p} \mathcal{G}a^{-1}\Big(\widehat{\alpha}_{\epsilon rj}, \widehat{\beta}_{\epsilon rj}\Big)\mathcal{N}_m\Big(\widehat{\boldsymbol{\Lambda}}_{rj}, \Psi_{\epsilon rj}\widehat{\mathbf{H}}_{rj}\Big), \end{aligned} \tag{31}$$

with

$$\begin{aligned} &\widehat{\mathbf{a}}_{\mu r} = \widehat{\boldsymbol{\Sigma}}_{\mu r}\Big[\boldsymbol{\Sigma}_{0r}^{-1}\boldsymbol{\mu}_{0r} + \widehat{n}^{(r)}\boldsymbol{\Psi}_{\epsilon r}^{-1}\Big(\overline{\mathbf{Y}}^{(r)} - \boldsymbol{\Lambda}_r\overline{\boldsymbol{\Omega}}^{(r)}\Big)\Big], \widehat{\boldsymbol{\Sigma}}_{\mu r} = \Big(\boldsymbol{\Sigma}_{0r}^{-1} + \widehat{n}^{(r)}\boldsymbol{\Psi}_{\epsilon r}^{-1}\Big)^{-1}, \\ &\widehat{\boldsymbol{\Lambda}}_{rj} = \widehat{\mathbf{H}}_{rj}\Big[\mathbf{H}_{\epsilon 0rj}^{-1}\boldsymbol{\Lambda}_{0rj} + \widehat{n}^{(r)}\Big(\mathbf{S}_{\omega y(j)}^{(r)} - \mu_{rj}\overline{\boldsymbol{\Omega}}^{(r)}\Big)\Big], \widehat{\mathbf{H}}_{rj}^{-1} = \mathbf{H}_{0rj}^{-1} + \widehat{n}^{(r)}\mathbf{S}_{\omega\omega}^{(r)}, \\ &\widehat{\alpha}_{\epsilon rj} = \alpha_{\epsilon 0rj} + \widehat{n}^{(r)}/2, \\ &\widehat{\beta}_{\epsilon rj} = \beta_{\epsilon 0rj} + \Big\{\boldsymbol{\Lambda}_{0rj}^{\intercal} H_{0rj}^{-1}\boldsymbol{\Lambda}_{0rj} + \widehat{n}^{(r)}\Big(S_{yy(j,j)}^{(r)} - 2\mu_{rj}\overline{y}_j^{(r)} + \mu_{rj}^2\Big) - \widehat{\boldsymbol{\Lambda}}_{rj}^{\intercal}\widehat{\mathbf{H}}_{rj}^{-1}\widehat{\boldsymbol{\Lambda}}_{rj}\Big\}/2, \end{aligned} \tag{32}$$

in which $\overline{y}^{(r)}_{(j)}$ is the jth element in $\overline{\mathbf{Y}}^{(r)}$, $S^{(r)}_{yy(j,j)}$ is the jth main diagonal element of $\mathbf{S}^{(r)}_{yy}$, and $\mathbf{S}^{(r)}_{\omega y(j)}$ is the jth column vector of $\mathbf{S}^{(r)}_{\omega y}$.

(d) $p(\boldsymbol{\Phi}|\boldsymbol{\Omega},\mathbf{Z})$

From the prior distribution of $\boldsymbol{\Phi}_r^{-1}$ and the distribution of $\boldsymbol{\Omega}$, it can be shown that

$$p(\boldsymbol{\Phi}_r|\boldsymbol{\Omega},\mathbf{Z})\propto|\boldsymbol{\Phi}_r|^{\left(\widehat{n}^{(r)}+\rho_{0r}+m+1\right)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\boldsymbol{\Phi}^{-1}\Big(\widehat{n}^{(r)}\mathbf{S}^{(r)}_{\omega\omega}+\mathbf{R}_0^{-1}\Big\}\right. \tag{33}$$

where $\widehat{n}^{(r)}$ and $\mathbf{S}_{\omega\omega}$ are given in (c). Hence, $p(\boldsymbol{\Phi}_r|\boldsymbol{\Omega},\mathbf{Z})$ is the m-dimensional inverse Wishart distribution $\mathcal{W}_m^{-1}\left(\widehat{n}^{(r)}+\rho_{0r},\widehat{n}^{(r)}\mathbf{S}^{(r)}_{\omega\omega}+\mathbf{R}_0^{-1}\right)$. It can be shown from exactly the same reasoning as before that drawing $\boldsymbol{\Phi}$ can be achieved by drawing $\boldsymbol{\Phi}_r$ from $p(\boldsymbol{\Phi}_r|\boldsymbol{\Omega},\mathbf{Z})$ independently.

(e) $p(\boldsymbol{\delta}|\mathbf{Z})$ and (f) $p(\mathbf{Q}|\mathbf{Z})$

It can be verified directly that

$$\begin{gathered}p(\boldsymbol{\delta}|\mathbf{Z})=p(\boldsymbol{\delta}_k|\mathbf{Z})\quad\text{and}\\ p(\boldsymbol{\delta}|\mathbf{Z})\overset{D}{=}\mathcal{D}ir_S\left(\gamma_0+\widehat{n}_{11},\dots,\gamma_0+\widehat{n}_{1S}\right)\end{gathered} \tag{34}$$

in which $\widehat{n}_{1r}=\sum_{i=1}^{N}I\{z_{i1}=r\}$. Similarly, it can be shown that

$$\begin{aligned}p(\mathbf{Q}|\mathbf{Z})&=\prod_{r=1}^{S}p(\mathbf{Q}_r|\mathbf{Z}),\\ p(\mathbf{Q}_r|\mathbf{Z})&\overset{D}{=}\prod_{r=1}^{S}\mathcal{D}ir_S(\nu_0+\widehat{n}_{r1},\dots,\nu_0+\widehat{n}_{rS}).\end{aligned} \tag{35}$$

in which $\widehat{n}_{rs}=\sum_{i=1}^{N}\sum_{t=2}^{T}I\{z_{it-1}=r,z_{it}=s\}$.

Author details

Yemao Xia[1]*, Xiaoqian Zeng[2] and Niansheng Tang[3]

*Address all correspondence to: ymxia@njfu.edu.cn

1 Department of Applied Mathematics, Nanjing Forestry University, Nanjing, China

2 School of Economics, Lanzhou University of Finance and Economics, Lanzhou, China

3 School of Mathematics and Statistics, Yunnan University, Kunming, China

References

[1] Berger JO. Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag; 1985. DOI: 10.1007/978-1-4757-4286-2

[2] Box GEP, Tiao GC. Bayesian Inference in Statistical Analysis. Reading, MA: Addison-Wesley; 1973. DOI: 10.1002/9781118033197

[3] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. London: Chapman & Hall Ltd; 1995

[4] Geman S, Geman D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984;(6):721-741. DOI: 10.1109/TPAMI.1984.4767596

[5] Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association. 1990;**85**:398-409. DOI: 10.1080/01621459.1990.10476213

[6] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machine. Journal of Chemical Physics. 1953;**21**:1087-1091

[7] Hastings WK. Monte Carlo sampling methods using Markov chains and their application. Biometrika. 1970;**57**(1):97-109. DOI: 10.1093/biomet/57.1.97

[8] Robert CR, Casella G. Monte Carlo Statistical Methods. New York, Inc.: Springer-Verlag; 1999. DOI: 10.1007/978-1-4757-3071-5

[9] Ross SM. Simulations. Amsterdam: Academic Press/Elsevier, Inc.; 2013. DOI: 10.1016/B978-0-12-375686-2.00001-7

[10] Schmittmann VD, Dolan CV, Han LJ, van der Maas, Neale CM. Discrete latent Markov models for normally distributed response data. Multivariate Behavioral Research. 2005;**40**(4):461-488. DOI: 10.1207/s15327906mbr4004_4

[11] Xia YM, Gou JW, Liu YA. Semi-parametric Bayesian analysis for factor analysis model mixed with hidden Markov model. Applied Mathematics A Journal of Chinese Universities, Series A. 2015;**30**(1):17-30

[12] Song XY, Xia YM, Zhu HT. Hidden Markov latent variable models with multivariate longitudinal data. Biometrics. 2017;**73**(1):313-323. DOI: 10.1111/biom.12536

[13] Xia YM, Tang NS, Gou JW. Generalized linear latent model for multivariate longitudinal measurements mixed with hidden Markov model. Journal of Multivariate Analysis. 2017;**152**:259-275. DOI: 10.1016/j.jmva.2016.09.001

[14] Wiggings LM. Panel Analysis: Latent Probability Models for Attitude and Behavior Processes. San Francisco, CA: Elsevier Scientific; 1973

[15] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;**77**(2):257-284. DOI: 10.1109/5.18626

[16] Altman RM. Mixed hidden Markov models: An extension of the hidden Markov mode to the longitudinal data setting. Journal of the American Statistical Association. 2007;**102** (477):201-210. DOI: 10.1198/016214506000001086

[17] Maruotti A. Mixed hidden Markov models for longitudinal data: An overview. International Statistical Review. 2011;**79**(3):427-454. DOI: 10.1111/j.1751-5823.2011.00160.x

[18] Lee SY. Structural Equation Modelling: A Bayesian Approach. New York: John Wiley & Sons; 2007

[19] Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. Journal of the American Statistical Association. 2003;**98**(463):555-563. DOI: 10.1198/016214503000000387

[20] Zhang ZY, Hamaker EL, Nesselroade JR. Comparisons of four methods for estimating a dynamic factor model. Structural Equation Modeling: A Multidisciplinary Journal. 2008;(3, 377):377-402. DOI: 10.1080/10705510802154281

[21] Chow SY, Tang NS, Yuan Y, Song XY, Zhu HT. Bayesian estimation of semiparametric nonlinear dynamic factor analysis model using the Dirichlet prior. British Journal of Mathematical and Statistical Psychology. 2011;**64**:69-106

[22] Ebbes P, Grewal R, DeSarbo WS. Modeling strategic group dynamics: A hidden Markov approach. Quantitative Marketing and Economics. 2010;**8**:241-274

[23] Marruotti A. Robust fitting of hidden Markov regression models under a longitudinal data. Journal of Statistical Computation and Simulation. 2014;**84**:1728-1747

[24] Zhu HT, Lee SY. A Bayesian analysis of finite mixtures in the LISREL model. Psychometrika. 2001;**66**(1):133-152. DOI: 10.1007/BF02295737

[25] Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association. 1987;**82**(398):528-550. DOI: 10.1080/01621459.1987.10478458

[26] Geyer CJ. Practical Markov chain Monte Carlo. Statistical Science. 1992;**7**(4):473-511. DOI: 10.1214/ss/1177011137

[27] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with discussion). Statistical Science. 1992;**7**(4):457-511. DOI: 10.1214/ss/1177011136

[28] Besage J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic system. Statistical Science. 1995;**10**(1):3-66. DOI: 10.1214/ss/1177010123

[29] Gelman A. Inference and monitoring convergence. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov Chain Monte Carlo in Practice. London: Chapman and Hall; 1996. pp. 131-140

[30] Kass RE, Raftery AE. Bayes factor (with discussion). Journal of the American Statistical Association. 1995;**90**(430):773-795. DOI: 10.1080/01621459.1995.10476572

[31] Geisser S, Eddy W. A predictive approach to model selection. Journal of the American Statistical Association. 1979;**74**(365):1537-1160. DOI: 10.1080/01621459.1979.10481632

[32] Laud PW, Ibrahim JG. Predictive model selection. Journal of the Royal Statistical Society, Series B. 1995;**57**(1):247-262. DOI: 10.2307/2346098

[33] Gelfand AE, Ghosh SK. Model choice: A minimum posterior predictive loss approach. Biometrika. 1998;**85**(1):1C13. DOI: 10.1093/biomet/85.1.1

[34] Ibrahim JG, Chen MH, Sinha D. Criterion based methods for Bayesian model assessment. Statistica Sinica. 2001;**11**:419-443

[35] Little RJA, Rubin DB. Statistical Analysis with Missing Data. New York: Wiley; 1987. DOI: 10.1002/9781119013563

[36] Wei GCG, Tanner MAA. Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association. 1990; **85**:699-704

[37] Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B. 1982;**44**:226-233

[38] Cap.ṕe O, Moulines E, Rydén T. Inference in Hidden Markov Models. New York: Springer Verlag; 2005