mathematics

MDPI

Article

Variational Bayesian Variable Selection for High-Dimensional Hidden Markov Models

Yao Zhai †, Wei Liu †, Yunzhi Jin * and Yanqing Zhang

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Department of Statistics, Yunnan University, Kunming 650091, China; zhaiyao@itc.ynu.edu.cn (Y.Z.); liuwei_clsa@stu.ynu.edu.cn (W.L.); zhangyanqing@ynu.edu.cn (Y.Z.)

* Correspondence: yzjin@ynu.edu.cn

† These authors contributed equally to this work.

Abstract: The Hidden Markov Model (HMM) is a crucial probabilistic modeling technique for sequence data processing and statistical learning that has been extensively utilized in various engineering applications. Traditionally, the EM algorithm is employed to fit HMMs, but currently, academics and professionals exhibit augmenting enthusiasm in Bayesian inference. In the Bayesian context, Markov Chain Monte Carlo (MCMC) methods are commonly used for inferring HMMs, but they can be computationally demanding for high-dimensional covariate data. As a rapid substitute, variational approximation has become a noteworthy and effective approximate inference approach, particularly in recent years, for representation learning in deep generative models. However, there has been limited exploration of variational inference for HMMs with high-dimensional covariates. In this article, we develop a mean-field Variational Bayesian method with the double-exponential shrinkage prior to fit high-dimensional HMMs whose hidden states are of discrete types. The proposed method offers the advantage of fitting the model and investigating specific factors that impact the response variable changes simultaneously. In addition, since the proposed method is based on the Variational Bayesian framework, the proposed method can avoid huge memory and intensive computational cost typical of traditional Bayesian methods. In the simulation studies, we demonstrate that the proposed method can quickly and accurately estimate the posterior distributions of the parameters with good performance. We analyzed the Beijing Multi-Site Air-Quality data and predicted the PM2.5 values via the fitted HMMs.

Keywords: Hidden Markov Models; high-dimensional data; shrinkage prior; variational inference

MSC: 62F15; 65K10; 62M05

Citation: Zhai, Y.; Liu, W.; Jin, Y.; Zhang, Y. Variational Bayesian Variable Selection for High-Dimensional Hidden Markov Models. *Mathematics* **2024**, *12*, 995. https://doi.org/10.3390/math12070995

Academic Editor: Diana Mindrila

Received: 15 February 2024
Revised: 22 March 2024
Accepted: 24 March 2024
Published: 27 March 2024

1. Introduction

Hidden Markov Models (HMMs) are a statistical model used to describe the evolution of observable events that depend on internal factors or states, which are not directly observable and called hidden states. Each hidden state can transition to another hidden state, including itself, with a certain probability, while we cannot observe them directly, we infer their presence and transitions between them based on observable outputs. HMMs have been widely used in various applications, including speech recognition, bioinformatics, natural language processing, and financial markets. In practice, HMMs often face high-dimensional issues, that is, a large number of covariates (or high-dimensional covariates) and multiple states result in high-dimensional parameters existing in the HMMs. The high-dimensional issue may result in the overfitting for the HMMs. The challenge then becomes identifying important variables or parameters in different hidden states. Thus, efficient parameter estimation and hidden Markov chain recovering are significant for the high-dimensional HMMs.

Currently, there have been many methods for estimating parameter estimation and recovering hidden Markov chains, including recursive algorithms [1–4] and traditional Bayesian methods [5,6]. Bayesian inference [7,8] is a versatile framework that utilizes sophisticated hierarchical data models for learning and consistently quantifies uncertainty in unknown parameters through the posterior distribution [9]. However, computing the posterior is demanding even for moderately intricate models and frequently necessitates approximation. Moreover, traditional Bayesian methods (e.g., Markov chain Monte Carlo (MCMC)) for the HMMs is often considered a black-box method by many statisticians due to its reliance on simulations to produce computable results, which is generally inefficient and unnecessary. Traditional Markov Chain Monte Carlo (MCMC) methods may also exhibit slow convergence and extended running times, as documented in prior studies [10–12].

The Variational Bayesian approach is an alternative to traditional MCMC algorithm in high-dimensional issue. Variational inference (VI) based on Bayesian method [13–16] can approximate posterior distributions quickly [17], since VI uses the Kullback–Leibler (KL) divergence to measure the difference between the variational posterior and the true posterior, and transforms the statistical inference problem into a mathematical optimization problem by minimizing the KL divergence. Therefore, the variational approach is as close as possible to the true posterior distribution according to the KL divergence. Wang and Blei [17] have proved that the variational posterior is consistent to the true posterior distribution. Moreover, there have been many efficient optimization algorithms to approximate complex probability densities such as coordinate-ascent (CAVI) [18] and gradient-based methods [15,19]. Currently, there exist many VI research studies for the HMMs [20–24]. For example, MacKay [20] was the pioneering proponent of the application of variational methods to HMMs, with a focus only on cases with discrete observations. Despite the limited comprehension of the state-removal phenomenon, which is that removing certain states from an HMM for simplifying the HMM while preserving its essential statistical properties does not significantly affect the the ability of the HMM to represent the underly stochastic process, variational methods are gaining popularity for HMMs within the machine learning community. C. A. McGrory [21] extended the deviance information criterion for Bayesian model selection within the Hidden Markov Model framework, utilizing a variational approximation. Nicholas J. Foti [22] devised an SVI algorithm to learn HMM parameters in settings of time-dependent data. Since VI can be seen as a special instance of the EM algorithm [23], Gruhl [23] integrates both approaches and uses the multivariate Gaussian output distribution of VI to train the HMM. Ding [24] employed variational inference techniques to investigate nonparametric Bayesian Hidden Markov Models built on Dirichlet processes. These processes enable an unbounded number of hidden states and adopt an infinite number of Gaussian components to handle continuous observations. However, variational inference has not been fully explored in HMMs, especially in HMMs with a high-dimensional covariate. It is important to note that research frequently entails using high-dimensional covariate datasets in real-world applications. When high-dimensional HMMs contain a large number of parameters, there is a possibility of overfitting the given data set. The challenge then becomes identifying which covariates have a substantial impact on the interpretation of observations and state shifts in each hidden state of the model, and which covariates have a negligible effect. This process is important for improving the predictive accuracy of the model, reducing the risk of overfitting, and improving the interpretability of the model.

In this article, we develop a Variational Bayesian method for variable selection. We utilize the double-exponential shrinkage prior [25–28] as the prior of coefficients in each hidden state models, to screen vital variables that affect each hidden state and obtain hidden Markov regression models. We use mean-field variational inference to identify variational densities for approximating complex posterior densities, via minimizing the difference between the approximate probability density and the actual posterior probability density. Moreover, we adopt the Monte Carlo Co-ordinate Ascent VI (MC-CAVI) [29] algorithm to compute the necessary expectations within the CAVI. Since the variational inference is a fast

alternative to the MCMC method and can avoid large memory and intensive computational cost compared to traditional Bayesian methods, the proposed approach inherits the good properties of variational inference, and can quickly and accurately estimate the posterior distributions and the unknown parameters. In the simulation studies and real data analysis, the proposed method outperforms the common methods in term of variable selection and prediction.

The main contributions of this article are as follows: First, the proposed method can perform variable selection for high-dimensional HMMs, and offer the advantage of fitting the model and investigating specific factors that impact the response variable changes simultaneously. Since the proposed method uses double-exponential shrinkage prior, which has the feature of being able to select important variables, the proposed method can simultaneously select important variables to the response variable and estimate the corresponding parameters. Second, since the proposed method is based on the Variational Bayesian framework, the proposed method can avoid huge memory and intensive computational cost of the traditional Bayesian methods, especially for the high-dimensional issue. Finally, we demonstrate that the proposed method can quickly and accurately estimate the posterior distributions of the parameters with good performance in the simulation studies. Moreover, we analyze Beijing Multi-Site Air-Quality data and predict the PM2.5 values well via the fitted HMMs.

The rest of the article is organized as follows: Section 2 introduce the Hidden Markov Model with high-dimensional covariate and shrinkage priors in the Bayesian inference. In Section 3, we propose an efficient Variational Bayesian estimation method with the double-exponential shrinkage prior for variable selection of the high-dimensional HMMs (HDVBHMM). In Section 4, we conduct simulation studies to investigate the finite sample performances of the proposed method. In Section 5, Beijing Multi-Site Air-Quality data are analyzed and the efficiency of the proposed method is verified. Section 6 concludes our work. Technical details are presented in the Appendix A.

2. Model and Notation

2.1. Hidden Markov Model

In this section, we first introduce Hidden Markov Models (HMMs). The HMMs are a type of doubly stochastic process that occurs over discrete time intervals and includes observations y_t and latent states z_t. In a traditional Hidden Markov Model without covariates, the observation y_t depends only on the current potential state z_t. The conditional distribution of the observation y_t when given the potential state $z_t = k$ can be expressed as:

$$y_t \mid z_t = k \sim F_k(\theta_k),$$

where $F_k(\theta_k)$ denotes a certain family of distributions, such as the normal distribution $N(\mu_k, \sigma_k^2)$. Extended HMM models can include covariates $x_t \in R^p$. That is, the set of observations is $y = (y_1, \ldots, y_T)$ and $x = (x_1, \ldots, x_T)$. Specifically, for $y = (y_1, \ldots, y_T)$, $z = (z_1, \ldots, z_T)$ and $x = (x_1, \ldots, x_T)$, the model expression is as follows:

$$y_t \mid x_t, z_t = k, \beta, \sigma^2 \sim N\left(x_t^\top \beta_k, \sigma^2\right) \text{ for } t = 1, \ldots, T,$$

where the symbol N represents the normal distribution, σ^2 denotes the variance of y_t, and $\beta = (\beta_1, \ldots, \beta_K)^\top$ is the coefficient of the covariate at all hidden states. In the article, we consider the high-dimensional issue of the covariate. We denote the dummy variable corresponding to z_t as the $v_t = (v_{t1}, \ldots, v_{tK})^\top$, where $v_{tk} = 1$ and other elements being zero if $z_t = k$. Thus,

$$P\left(y_t \mid x_t, \beta, \sigma^2\right) = \prod_{k=1}^{K} P\left(y_t \mid x_t, z_t = k, \beta, \sigma^2\right)^{v_{tk}} = \prod_{k=1}^{K} P\left(y_t \mid x_t, \beta_k, \sigma^2\right)^{v_{tk}}.$$

In hidden Markov chains, each hidden state z_t is independent from $z_1, \ldots, z_{t-2}$ and $z_{t+1}, \ldots, z_T$ conditionally on z_{t-1}. Therefore, we can assume that the probability distribution of z_1 is given by $z_1 \sim P(z_1|\pi) = (\pi_1, \pi_2, \ldots, \pi_K)^\top$, where $\sum_{k=1}^K \pi_k = 1$ and $\pi_k > 0$. The conditional probability of z_t given z_{t-1} is assumed as:

$$P(z_t \mid z_{t-1}, A) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{v_{t-1,j}\, v_{tk}},$$

where A is the transition matrix with elements A_{ij} for $i, j = 1, \ldots, K$, $\sum_{j=1}^K A_{ij} = 1$ and $A_{ij} > 0$, and A_{ij} represents the probability of transitioning from state i to state j. Thus, the joint distribution is as follows:

$$\begin{aligned} & P(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}, \pi, A, \beta, \sigma^2) \\ = \;& P(z_1|\pi) \prod_{t=2}^{T} P(z_t \mid z_{t-1}, A) \prod_{t=1}^{T} P(y_t \mid x_t, \beta, \sigma^2) \\ = \;& \left(\prod_{k=1}^{K} \pi_k^{v_{1k}}\right) \left(\prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{v_{t-1,j}\, v_{tk}}\right) \left(\prod_{t=1}^{T} \prod_{k=1}^{K} P(y_t \mid x_t, \beta, \sigma^2, z_t = k)^{v_{tk}}\right). \end{aligned} \tag{1}$$

2.2. Prior Selection in the HMMs

To make Variational Bayesian inference, we require specifying the prior of the parameters π, A, β and σ^2. Based on the characteristics of $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$, $\sum_{k=1}^K \pi_k = 1$ and $\pi_k > 0$, Dirichlet distribution is applied to the prior distribution of π as follows:

$$\pi \sim \mathrm{Dir}\left(\alpha^{(\pi)}\right), \tag{2}$$

where $\alpha^{(\pi)} = \left(\alpha_1^{(\pi)}, \ldots, \alpha_K^{(\pi)}\right)^\top$, $\sum_{k=1}^K \alpha_k^{(\pi)} = 1$, and $\alpha_k^{(\pi)} > 0$. In the model, A denotes the transition matrix of the hidden state $\boldsymbol{z}$ and can be expressed as follows:

$$A = \begin{pmatrix} A_{11} & \ldots & A_{1K} \\ \vdots & \ddots & \vdots \\ A_{K1} & \ldots & A_{KK} \end{pmatrix}.$$

Like Nicholas [22], we specify the prior of the jth row of the transition matrix A as:

$$A_j \sim \mathrm{Dir}\left(\alpha_j^{(A)}\right) \;\text{ for }\; j = 1, \ldots, K, \tag{3}$$

where $\alpha_j^{(A)} = \left(\alpha_{j1}^{(A)}, \alpha_{j2}^{(A)}, \ldots, \alpha_{jK}^{(A)}\right)^\top$, $\sum_{k=1}^K \alpha_{jk}^{(A)} = 1$, and $\alpha_{jk}^{(A)} > 0$. Since σ^2 is variance of the y, we specify the prior of the σ^2 as

$$\sigma^2 \sim f\left(\sigma^2\right) = \frac{1}{\sigma^2}. \tag{4}$$

In a high-dimensional and sparse issue, we consider the double-exponential shrinkage prior [25,27] as the prior of β, defined as follows:

$$\begin{aligned} & \beta_k \mid \sigma^2, \tau_1, \ldots, \tau_p^2 \sim N_p\left(0, \sigma^2 D_\tau\right), \\ & D_\tau = \mathrm{diag}\left(\tau_1^2, \ldots, \tau_p^2\right), \\ & \tau_m^2 \sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right) \;\text{ for }\; m = 1, 2, \ldots, p, \\ & \lambda^2 \sim \Gamma(r, \delta), \end{aligned} \tag{5}$$

where Γ represents gamma distribution and $\mathrm{Exp}(\cdot)$ represents the exponential distribution. The above prior can select important variables of the HMMs in each hidden state. Bayesian approaches can be used to solve the parameter estimation question with the above prior information. However, in high-dimensional data, the traditional Bayesian methods (e.g., MCMC) require huge memory and intensive computational cost. The Variational Bayesian approach is an alternative to the traditional MCMC algorithm in high-dimensional issue. Next, we introduce the proposed Variational Bayesian inference for high-dimensional HMMs.

3. Variational Bayesian Inference for the HMMS

3.1. Mean Field Variational

Mean-field Variational Bayesian inference is a prevalent approach in variational inference, and aims to identify an approximate density by minimizing the difference between the approximate probability density and the actual posterior probability density, while being bounded by the Kullback–Leibler divergence. In this subsection, we proposed the mean-field variational inference for HMMS with the high-dimensional covariates.

Let $\boldsymbol{D}$ be an observed data set, $\boldsymbol{D} = \{\boldsymbol{y}, \boldsymbol{x}\}$ with response set $\boldsymbol{y} = \{y_i \mid i = 1, \ldots, n\}$ and covariate set $\boldsymbol{x} = \{x_i \mid i = 1, \ldots, n\}$, and $\boldsymbol{\theta} = \{\pi, A, \beta, \sigma^2, \tau_m^2, \lambda^2\}$. The $\boldsymbol{\theta}$ and $\boldsymbol{z}$ include all parameters in the HMMs. We focus on the posterior distribution of parameters $\boldsymbol{\theta}$ and the hidden state z_t. Assume that there is an approximate density family $\boldsymbol{Q}$ containing possible densities over the parameters $\boldsymbol{\theta}, \boldsymbol{z}$. Minimizing the Kullback–Leibler (KL) divergence between the member of the family $q(\boldsymbol{\theta}, \boldsymbol{z})$ and the true posterior $P(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{D})$ is to obtain the optimal density approximation of the true posterior, with variational inference prioritizing optimization rather than sampling. That is,

$$q^*(\boldsymbol{\theta}, \boldsymbol{z}) = \arg \min_{q(\boldsymbol{\theta}, \boldsymbol{z}) \in \boldsymbol{Q}} \mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{z}) \| P(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{D})),$$

where the KL-divergence is:

$$\mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{z}) \| P(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{D})) = \int q(\boldsymbol{\theta}, \boldsymbol{z}) \log \left\{ \frac{q(\boldsymbol{\theta}, \boldsymbol{z})}{P(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{D})} \right\} d(\boldsymbol{\theta}, \boldsymbol{z}).$$

The KL-divergence can be further written as:

$$\begin{aligned} &\mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{z}) \| P(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{D})) \\ =&E_q[\log q(\boldsymbol{\theta}, \boldsymbol{z})] - E_q[\log P(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{D})] \\ =&E_q[\log q(\boldsymbol{\theta}, \boldsymbol{z})] - E_q[\log P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{D})] + \log P(\boldsymbol{D}), \end{aligned}$$

where $\log P(\boldsymbol{D})$ is a constant, E_q denotes the expected value of $\boldsymbol{\theta}$ and z drawn from the distribution q. Thus, minimizing the KL divergence is equivalent to maximizing the following evidence lower bound (ELBO):

$$\mathrm{ELBO}(q) = E_q[\log P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{D})] - E_q[\log q(\boldsymbol{\theta}, \boldsymbol{z})]. \tag{6}$$

From another perspective, the ELBO comprises the negative KL divergence and $\log P(\boldsymbol{D})$.

According to the mean-field variational framework [30,31], the parameters are assumed to be posterior independent of each other and to be controlled by a separate factor in the variational density. In the HMMs, $q(\boldsymbol{\theta}, \boldsymbol{z})$ is decomposed as:

$$q(\boldsymbol{\theta}, \boldsymbol{z}) = q(\pi) q(A) q\left(\sigma^2\right) q\left(\tau_m^2\right) q\left(\lambda^2\right) \prod_{t=1}^{T} q(z_t). \tag{7}$$

Each parameter $\boldsymbol{\theta}_i$ and latent state $\boldsymbol{z}$ is governed by its own variational factor. The forms of $q(\boldsymbol{\theta}_i)$ and $q(\boldsymbol{z})$ are unknown, but the form of the hypothesized factorization is determined. In the optimization process, the optimal solutions of these variational factors $q(\boldsymbol{\theta_i})$ and $q(\boldsymbol{z})$ are obtained by maximizing the ELBO of Equation (6) by the coordinate ascent method. Based on the consistency of the Variational Bayesian [17], the variational densities over the mean-field family are still consistent to the posterior densities, even though the mean field approximating family can be a brutal approximation. More generally, one can consider structured variational distributions involving partial factorizations that correspond to tractable substructures of parameters [32]. In this article, we only consider the mean field framework. To express the variational posterior formula concisely, we define $\phi = \{\boldsymbol{\theta}, \boldsymbol{z}\}$ and rewrite $q(\boldsymbol{\theta}, \boldsymbol{z})$ as $q(\phi)$.

3.2. The Coordinate Ascent Algorithm for Optimizing the ELBO

Based on the variational density decomposition, we can obtain each factor of the variational density via maximizing the ELBO. Let $q_i(\phi_i)$ for $i = 1, 2, \ldots, b$ be the ith factor of the variational density in .The common approaches to maximize the ELBO mainly include a Coordinate Ascent Variational Inference (CAVI) and a gradient-based approach [33]. The CAVI approach sequentially optimizes each factor of the variational density of the mean field to obtain a local maximizer for the ELBO, while keeping the others fixed. Based on the CAVI approach, we can obtain the optimal variational density $q_i^*(\phi_i)$ as follows:

$$q_i^*(\phi_i) \propto \exp\{E_{-i}[\log P(\phi_{i_-}, \phi_i, \phi_{i_+}, \boldsymbol{D})]\}, \tag{8}$$

where i_- (or i_+) refers to the ordered indexes that are less than (or greater than) i. Let $\phi_{-i} := (\phi_{i_-}, \phi_{i_+})$. The vector ϕ_{-i} represents the vector ϕ with the ith component ϕ_i removed. The E_{-i} denotes the expectation with respect to ϕ_{-i}.

Based on the joint distribution (1), the priors (2)–(5) and Formula (8), we can derive all variational posteriors (see Appendix A for details). The variational posterior of the π is:

$$q^*(\pi) \sim \text{Dir}\left(\alpha_{(\pi)}\right), \tag{9}$$

where $\alpha_{(\pi)} = E(z_1) + \alpha^{(\pi)}$. The variational posterior of the A_j is:

$$q^*(A_j) \sim \text{Dir}\left(\alpha_{(A_j)}\right) \text{ for } j = 1, \ldots, K, \tag{10}$$

where $\alpha_{(A_j)} = \sum_{t=2}^{T} E(v_{t-1,j} v_{tk}) + \alpha_{jk}^{(A)}$. The variational posterior of the β_k is:

$$q^*(\beta_k) \sim N_p(\beta_k; \mu_k, \Sigma_k), \tag{11}$$

where

$$\begin{aligned} \Sigma_k &= \left(E\left(\tfrac{1}{\sigma_2}\right) \sum_{t=1}^{T} E(v_{tk}) x_t x_t^\top + E\left(\tfrac{1}{\sigma^2}\right) E\left(D_\tau^{-1}\right)\right)^{-1}, \\ \mu_k &= \Sigma_k \left(E\left(\tfrac{1}{\sigma^2}\right) \sum_{t=1}^{T} y_t \cdot E(v_{tk}) \cdot x_t\right). \end{aligned}$$

The variational posterior of the σ^2 is:

$$q^*\left(\sigma^2\right) \sim \text{ Inverse-Gamma } \left(\alpha_{(\sigma^2)}, \beta_{(\sigma^2)}\right), \tag{12}$$

where $\alpha_{(\sigma^2)} = \frac{T}{2}$ and $\beta_{(\sigma^2)} = \frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K} E(v_{tk})\left[\left(y_t - x_t^\top \mu_k\right)^2 + x_t^\top \Sigma_k x_t\right]$. The variational posterior of the τ_m^2 is:

$$q^*\left(\tau_m^2\right) \sim \text{Generalized-Inverse–Gaussian}\ (C_{\tau_m}, a_{\tau_m}, b_{\tau_m}), \tag{13}$$

where $a_{\tau_m} = E(\lambda^2)$, $b_{\tau_m} = E(1/\sigma_2)\sum_{k=1}^{K} E(\beta_{km}^2)$, and $C_{\tau_m} = 1 - K/2$. The variational posterior of the λ^2 is:

$$q^*\left(\lambda^2\right) \sim \Gamma\left(\alpha_{(\lambda^2)}, \beta_{(\lambda^2)}\right), \tag{14}$$

where $\alpha_{(\lambda^2)} = p + r$ and $\beta_{(\lambda^2)} = \delta + \frac{1}{2}\sum_{m=1}^{p} E\left(\tau_m^2\right)$.

Based on the dependencies of hidden states, we divide the posterior of z into three parts. The variational posterior of the z_1 is:

$$q^*(z_1) \sim \text{Mult}\left(P_{(z_1)}\right), \tag{15}$$

where the Mult represents multinomial distribution, $P_{(z_1)} = \left(P_{(z_1)1}, \ldots, P_{(z_1)K}\right)^\top$ and

$$P_{(z_1)k} = \exp\{E[\log \pi_k]\}\exp\left\{E\left[\log P\left(y_1 \mid x_1, \beta_k, \sigma^2\right)\right]\right\}\prod_{j=1}^{K}\exp\left\{E\left[v_{2j}\right]E\left[\log A_{kj}\right]\right\}.$$

The variational posterior of the z_t for $t = 2, \ldots, T-1$ is:

$$q^*(z_t) \sim \text{Mult}\left(P_{(z_t)}\right) \ \text{ for } \ t = 2, \ldots, T-1, \tag{16}$$

where $P_{(z_t)} = \left(P_{(z_t)1}, \ldots, P_{(z_t)K}\right)^\top$ and

$$\begin{aligned} P_{(z_t)k} \ = \ & \exp\left(E\left[\log P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right]\right) \cdot \textstyle\prod_{j=1}^{K}\exp\left\{E\left[\log A_{jk}\right]E\left(v_{t-1,j}\right)\right\} \\ & \cdot \textstyle\prod_{j=1}^{K}\exp\left\{E\left[\log A_{kj}\right]E\left(v_{t+1,j}\right)\right\}. \end{aligned}$$

The variational posterior of the z_T is:

$$q^*(z_T) \sim \text{Mult}\left(P_{(z_T)}\right), \tag{17}$$

where $P_{(z_T)} = \left(P_{(z_T)1}, \ldots, P_{(z_T)K}\right)^\top$ and

$$P_{(z_T)k} = \exp\left\{E\left[\log P\left(y_T \mid x_T, \beta_k, \sigma^2\right)\right]\right\} \cdot \prod_{j=1}^{K}\exp\left\{E\left[\log A_{jk}\right]E(v_{T-1}, j)\right\}.$$

The expectation $E\left[\log P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right]$ in the above variational posteriors (15)–(17) is expressed as follows:

$$E\left[\log P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}E[\log(\sigma^2)] - \frac{1}{2}E\left(\frac{1}{\sigma^2}\right)\left[\left(y_t - x_t^\top \mu_k\right)^2 + x_t^\top \Sigma_k x_t\right].$$

Note that the expectation part of some parameter posterior formulas is difficult to derive analytically. One feasible method is to use Monte Carlo (MC) sampling to approximate the expectation part that cannot be derived analytically, that is, the Monte Carlo Coordinate Ascent VI (MC-CAVI) [29] algorithm. The MC-CAVI recursion approaches have been proved to be convergent to the maximizer of the ELBO with arbitrarily high probability under regularity conditions. In the article, we also use MC-CAVI to obtain the intractable expectations.

3.3. Implementation

Assume that the expectations $E_{-i}[\log P(\phi_{i-}, \phi_i, \phi_{i+}, \boldsymbol{D})]$ for $i \in I$ within an index set I can be analytically obtained across all updates of the variational density $q^*(\phi)$, and cannot be analytically obtained for $i \notin I$. For the MC-CAVI method, intractable integrals can be approximated using the MC methods if $i \notin I$. Specifically, for $i \notin I$, the samples with the sample size $N \geq 1$ are drawn from the current $q^*_{-i}(\phi_{-i})$ to obtain the expectation estimations as follows:

$$\widehat{E}_{-i}[\log P(\phi_{i-}, \phi_i, \phi_{i+}, \boldsymbol{D})] = \frac{\sum_{n=1}^{N} \log P\left(\phi_{i-}^{(n)}, \phi_i^{(n)}, \phi_{i+}^{(n)}, \boldsymbol{D}\right)}{N}.$$

The Algorithm 1 summarizes the implementation of MC-CAVI, where the $q_{i,k}(\phi_i)$ denotes the density of the ith density factor after it has undergone the kth updates, and $q_{-i,k}(\phi_{-i})$ refers to the density of all density factors except the ith factors after the kth updates to the factors preceding the ith factor and the $k-1$ updates to the blocks following it.

Algorithm 1 Main iteration steps of MC-CAVI

Necessary: Number of iteration cycles T.
Necessary: Quantity of Monte Carlo samples denoted as N.
Necessary: $\mathbb{E}_{-i}\left[\log P\left(\phi_{i_-}, \phi_i, \phi_{i_+}, \boldsymbol{D}\right)\right]$ in closed form for $i \in \mathcal{I}$.
1. Initialize $q_{i,0}(\phi_i)$ for $i = 1, \ldots, b$.
2. for $k = 1 \ldots T$:
3. for $i = 1 \ldots b$:
4. If $i \in \mathcal{I}$:
5. Set $q_{i,k}(\phi_i) \propto \exp\left\{\mathbb{E}_{-i,k}\left[\log P\left(\phi_{i_-}, \phi_i, \phi_{i_+}, x, y\right)\right]\right\}$;
6. If $i \notin \mathcal{I}$:
7. Obtain N samples $\left(\phi_{i_-,k}^{(n)}, \phi_{i_+,k-1}^{(n)}\right)$ from $q_{-i,k}(\theta_{-i})$ for $n = 1, 2, \ldots, N$;
8. Set $q_{i,k}(\phi_i) \propto \exp\left\{\frac{\sum_{n=1}^{N} \log p\left(\phi_{i_-,k}^{(n)}, \phi_i, \phi_{i_+,k-1}^{(n)}, \boldsymbol{D}\right)}{N}\right\}$;
9. end
10. end.

Combining with the MC-CAVI algorithm, we can summarize the implementation algorithm for variational posteriors for all parameters as follows in Algorithm 2. Based on the Algorithm 2, we can adopt the variational posterior means of the parameters as the estimators.

Algorithm 2 Variational Bayesian Algorithm for the high-dimensional HMMs

Data Input: $\{(x_t, y_t)\}, t = 1, \ldots, T$;
Hyperparameter Input: $\alpha^{(\pi)}, r > 0, \delta > 0$, and $\alpha_{jk}^{(A)}$ for $k, j = 1, \ldots K$;
Initialize: $\alpha_{(\pi)}$, $\alpha_{(A_j)}$, $\alpha_{(\sigma^2)}$, $\beta_{(\sigma^2)}$, $\beta_{(\lambda^2)}$, Σ_k and μ_k for $k = 1, \ldots K$,
a_{τ_m} and b_{τ_m} for $m = 1, \ldots, p$, iteration-index $\ell = 1$, a sufficiently small $\epsilon = 10^{-6}$
and a maximum iteration times $M = 1000$;
While the absolute change of the iterated ELBO $|L^\ell - L^{\ell-1}| > \epsilon$ and $\ell < M$ **do**:
Update $\alpha_{(\pi)}$ and $q^*(\pi)$ according to Equation (9);
Estimate $E[\log \pi_k]$ by the MC method;
for $j = 1, \ldots, K$ **:**
Update $\alpha_{(A_j)}$ and $q^*(A_j)$ according to Equation (10);
Estimate $E\left[\log A_{jk}\right]$ by the MC method;
end
for $k = 1, \ldots, K$ **:**
Update Σ_k, μ_k and $q^*(\beta_k)$ according to Equation (11);
end
Update $\alpha_{(\sigma^2)}$, $\beta(\sigma^2)$ and $q^*(\sigma^2)$ according to Equation (12);
Estimate $E\left[\log(\sigma^2)\right]$ by the MC method;
for $m = 1, \ldots, p$ **:**
Update a_{τ_m}, b_{τ_m} and $q^*(\tau_m^2)$ according to Equation (13);
end
Update $\beta_{(\lambda^2)}$ and $q^*(\lambda^2)$ according to Equation (14);
Update $P_{(z_1)}$ and $q^*(z_1)$ according to Equation (15);
Update $P_{(z_t)}$ and $q^*(z_t)$ according to Equation (16);
Update $P_{(z_T)}$ and $q^*(z_T)$ according to Equation (17);
Compute the ELBO using the formula (6), denoted as L^ℓ,
and the absolute change of the iterated ELBO $|L^\ell - L^{\ell-1}|$;
$\ell \to \ell + 1$;
Output: the variational densities $q^*(\pi)$, $q^*(A_j)$ for $j = 1, \ldots, K$, $q^*(\beta_k)$ for $k = 1, \ldots, K$,
$q^*(\sigma^2)$, $q^*(\lambda^2)$, $q^*(\tau_m^2)$ for $m = 1, \ldots, p$, and $q^*(z_t)$ for $t = 1, \ldots, T$;
and the posterior modes of parameters β_k for $k = 1, \ldots, K$.

4. Simulation Studies

In this section, we carry out simulation studies to investigate the finite sample performances of the proposed method, denoted as HDVBHMM. To evaluate the prediction performance, we compare the proposed method with some commonly used and popular methods, including Back Propagation Neural Network (BP), Long Short-Term Memory (LSTM), and Random Forest. The experimental code can be found via the github link (https://github.com/LiuWei-hub/VBHDHMM, accessed on 23 March 2024).

We consider the dataset $\{x_t, y_t : t = 1, \ldots, T\}$, where T is the number of the discrete time intervals, the covariate x_t is generated from the Gaussian distribution $N_p(0, 2I_p)$, and $y_t = x_t^\top \beta_{z_t} + \varepsilon_t$, in which the random error $\varepsilon_t \sim N(0, \sigma^2)$, and z_t is hidden state. Here, the initial hidden state z_1 is generated from Mult(π), where $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$. For $t = 2, \ldots, T$, the hidden state z_t is generated from Mult(A_j), where $A_j = (A_{1i}, A_{2i}, \ldots A_{Ki})$ and $A_{jk} = P(z_t = k \mid z_{t-1} = j)$. We set the number of hidden states $K = 3$, $\sigma = 0.4$, and $(\pi_1, \pi_2, \ldots, \pi_K) = (0.6, 0.3, 0.1)^\top$.

To assess the predictive performance, we use the samples in the last m time intervals as the testing set and the samples in the first $T - m$ time intervals as the training set. In addition, we use four criteria: (1) the mean absolute percentage error MAPE $= \frac{100\%}{m} \sum_{t=1}^{m} \left| \frac{\hat{y}_t - y_t}{y_t} \right|$, where y_t is the true value and $\hat{y}_t$ represents the predicted value; (2) the root mean square error RMSE $= \sqrt{\frac{1}{m} \sum_{t=1}^{m} (y_t - \hat{y}_t)^2}$; (3) the mean absolute error MAE $= \frac{1}{m} \sum_{t=1}^{m} |y_t - \hat{y}_t|$;

and (4) $R^2 = 1 - \frac{\sum_{t=1}^{m}(\hat{y}_t - y_t)^2}{\sum_{t=1}^{m}(\bar{y} - y_t)^2}$, where $\bar{y}$ represents the sample mean, $\sum_{t=1}^{m}(\hat{y}_t - y_t)^2$ is the error caused by the prediction, and $\sum_{t=1}^{m}(\bar{y} - y_t)^2$ is the error caused by the mean. The smaller the MAPE, RMSE and MAE values are, the better the performance of the method is. The larger the R^2 is, the better the performance of the method is. To evaluate the performance of the parameter estimation, we use two criteria: (1) the root mean square error loss RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta\right)^2}$, where n is the number of repeated experiments, $\hat{\theta}_i$ is the estimated value of the parameter obtained in the ith experiment, and θ is the true parameter value; and (2) Bias$(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta$. The RMSE and Bias values closer to zero imply better performance for the method. We repeat 10 simulation examples and calculate the average values of the above metrics for each method.

4.1. Experiment 1

In experiment 1 , we consider different dimensions p = 20, 30 and 40. In addition, the state transition matrix A is set as follows:

$$A = \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0.4 & 0.1 \end{pmatrix}.$$

Due to $K = 3$, we have three regression coefficients β_1, β_2, β_3. We set the coefficient as follows:

$$\beta = (\beta_1, \beta_2, \beta_3)^\top = \begin{pmatrix} 0.5 & 1 & 1.5 \\ -2 & -2 & -1.5 \\ 2 & 1.5 & 1 \\ -1 & -1.5 & -2 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}_{p\times 3}$$

where the first four rows are nonzero and other elements are zero. We set the number of the discrete time intervals $T = 300$ and the sample size in the testing set $m = 10$. In addition, the hyperparameters r, δ in the HDVBHMM method are set to 1. The results are shown in Tables 1 and 2.

In Table 1, the smaller MAPE, RMSE, and MAE index values, the better the algorithm performance. The larger the R^2 index, the better the algorithm performance. Bold indicates the optimal result in each scenario. It is clear that our method is optimal in all cases (bold), especially for $p = 20$, $p = 30$, and $p = 40$. In the small sample case, the prediction performance of the LSTM method decreases significantly as the dimensionality of the covariates increases. The prediction performance of the Random Forest and BP methods is not stable with increasing covariate dimensions. Although the performance of our method decreases as the covariate dimension increases, it is still significantly better than the other methods. Table 2 shows the RMSE and Bias of the estimated values of β and A. From Table 2, we can see that the proposed method performs well. Two metrics are small when the covariate dimension is 20 and 30. When the dimension is increased to 40, the value of the RMSE index increases, but it is still within the acceptable range.

Table 1. Average values of four metrics of all approaches with standard deviation in each parenthesis based on 10 simulations under $T = 300$.

p	Method	Estimate Performance			
		MAPE	**RMSE**	**MAE**	R^2
$p = 20$	LSTM	0.957 (1.109)	1.183 (0.132)	1.416 (0.316)	0.871 (0.057)
	BP	0.948 (1.018)	1.210 (0.175)	1.492 (0.436)	0.826 (0.117)
	Random Forest	1.215 (1.844)	1.430 (0.194)	2.081 (0.531)	0.741 (0.119)
	HDVBHMM	**0.467** (0.325)	**1.008** (0.152)	**1.038** (0.329)	**0.887** (0.082)
$p = 30$	LSTM	0.949 (0.485)	1.354 (0.266)	1.898 (0.740)	0.789 (0.144)
	BP	1.312 (0.685)	1.524 (0.198)	2.358 (0.615)	0.659 (0.169)
	Random Forest	1.081 (0.641)	1.568 (0.223)	2.505 (0.717)	0.595 (0.252)
	HDVBHMM	**0.876** (0.919)	**1.186** (0.244)	**1.461** (0.629)	**0.861** (0.073)
$p = 40$	LSTM	1.471 (1.121)	1.404 (2.471)	2.026 (0.716)	0.763 (0.131)
	BP	1.555 (1.218)	1.427 (0.156)	2.060 (0.412)	0.772 (0.117)
	Random Forest	1.210 (0.718)	1.457 (0.231)	2.173 (0.736)	0.754 (0.115)
	HDVBHMM	**1.023** (0.759)	**1.155** (0.264)	**1.398** (0.619)	**0.822** (0.202)

In the Long-Term and Short-Term Memory methods, the learning rate is lr = 0.001, the number of training cycles (Epochs) is set to 50, and the size of the hidden layer is set to 10. The hidden layer of the BP method consists of 20 neurons, and the maximum number of iterations is set to 10,000. In the random forest regression model, the number of trees is set to 100. The bold results are the optimal ones among four methods.

Table 2. Average values of the RMSE and Bias of A and β based on 10 replications in Experiment 1.

p	Parameter	Estimate Performance	
		RMSE	**Bias**
$p = 20$	β	0.001	0.001
	A	0.002	0.001
$p = 30$	β	0.002	0.001
	A	0.005	0.001
$p = 40$	β	0.004	0.001
	A	0.011	0.001

To better illustrate the performance of parameter estimation, Figure 1 shows box plots of the estimator values of A, β_1, β_2, β_3 under $p = 30$, where the horizontal coordinate is the index of the variables and the vertical coordinate is the values of estimators. The corresponding figures on $p = 20$ and $p = 40$ are shown in Appendix A.2. For the estimators β_1, β_2 and β_3, we can see the first four elements are estimated close to the true value, and the remaining values are estimated clear to zero; This implies that the proposed method can achieve good variable selection performance. In addition, all elements of the state increment matrix A are estimated close to the true values, which also confirms the good performance of our method.

In addition, to further verify that the algorithm is sensitive to the choice of hyperparameters r, δ, we conduct experiments on data with a covariate dimension of 30. Consider the following three experiments, the first with $r = 0.5, \delta = 0.5$; the second with $r = 1.0, \delta = 1.0$; and the third with $r = 1.5, \delta = 1.5$. The experimental results show that the estimation results are not sensitive to the choice of the two hyperparameters r and δ. The images of the Gamma distributions for the three different hyperparameter settings are very similar in shape. This similarity may contribute to the reason why, for a certain range of variations in r and δ values, the model's performance may not show sensitivity to these hyperparameters. We show the results in Appendix A.4.

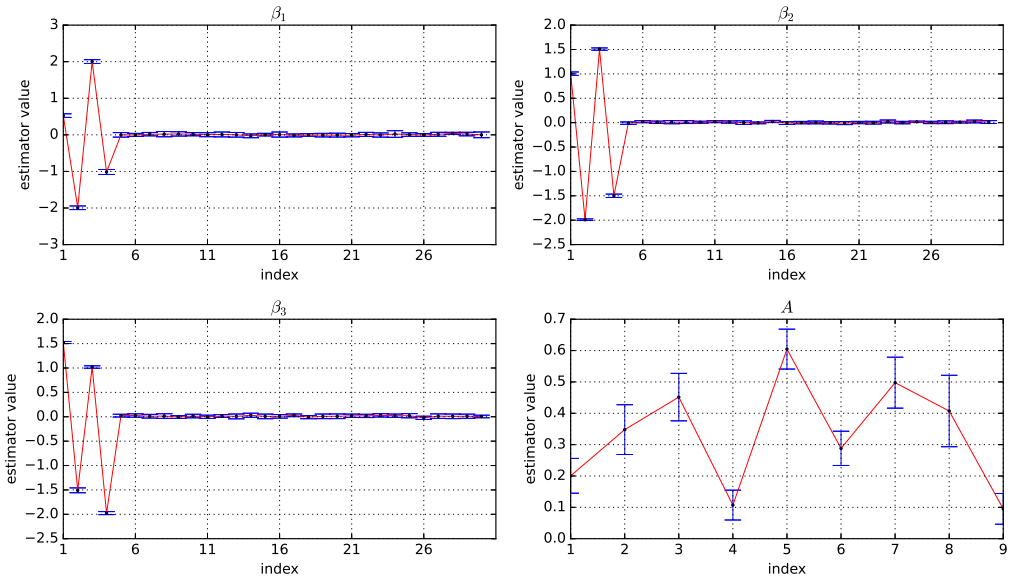

Figure 1. Box plots of the estimator values of A, β_1, β_2, β_3 based on 10 experiments under $p = 30$ and $T = 300$. The horizontal coordinate is the index of the variables and the vertical coordinate is the value of the estimators.

4.2. Experiment 2

In experiment 2, we consider the higher dimension cases: p = 60, 90, 120. We set the same A as experiment 1 and the coefficient as follows:

$$\beta = (\beta_1, \beta_2, \beta_3)^\top = \begin{pmatrix} 0.5 & 1 & 1.5 \\ -2 & -2 & -1.5 \\ 2 & 1.5 & 1 \\ -1 & -1.5 & -2 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}_{p\times 3}$$

where the first four rows are nonzero and other elements are zero. We set the number of discrete time intervals $T = 600$ and the sample size in the testing set $m = 10$. In addition, the hyperparameters r and δ in the HDVBHMM method are set to 1. The results are shown in Tables 3 and 4.

As can be seen in Table 4, when the covariate dimensions are increased and the sample size reaches 600, our method still performs well among the four methods. It should be noted that when the covariate dimension is 90 and 120, the MAPE metric of the random forest method is slightly smaller than our method. In addition, as the covariate dimension increases from p = 60 to P = 120, the performance of the LSTM method decreases significantly, which is the worst performance among the four methods. This shows that LSTM does not perform well on such small-sample high-dimensional datasets. As the dimensionality of the covariates increases, although the BP and Random Forest methods show better prediction performance than the LSTM method, they are also poorer than the prediction performance of the HDVBHMM method. Overall, our method outperforms the other three methods in terms of prediction performance as the dimensionality increases, suggesting that our method performs better on small-sample high-dimensional datasets.

Figure 2 shows box plots of the estimator values of A, β_1, β_2, β_3 under $p = 90$. The corresponding figures on $p = 60$ and $p = 120$ are shown in Appendix A.3. From Figure 2, we can see that the regression coefficients β_1, β_2, and β_3 are accurately estimated. the first four elements are estimated close to the true value, and the remaining values are estimated clear to zero. It implies that the proposed method can successfully achieve

variable screening even as the covariate dimension increases. In addition, all elements of the state increment matrix A are estimated close to the true values, which also confirms the good performance of the proposed method.

Table 3. Average values of four metrics of all approaches with standard deviation in each parenthesis based on 10 simulations under $T = 600$.

p	Method	Estimate Performance			
		MAPE	**RMSE**	**MAE**	R^2
$p = 60$	LSTM	1.608 (1.826)	1.524 (0.104)	2.332 (0.304)	0.722 (0.126)
	BP	1.511 (1.326)	1.513 (0.209)	2.328 (0.662)	0.725 (0.143)
	Random Forest	1.236 (1.748)	1.459 (0.175)	2.159 (0.525)	0.728 (0.151)
	HDVBHMM	**0.851** (0.871)	**1.091** (0.218)	**1.235** (0.489)	**0.884** (0.078)
$p = 90$	LSTM	1.690 (2.046)	1.725 (0.152)	2.998 (0.542)	0.523 (0.282)
	BP	2.780 (5.010)	1.606 (0.225)	2.626 (0.706)	0.636 (0.239)
	RF	**0.718** (0.365)	1.374 (0.246)	1.942 (0.694)	0.797 (0.091)
	HDVBHMM	0.878 (0.887)	**1.156** (0.249)	**1.392** (0.617)	**0.862** (0.135)
$p = 120$	LSTM	1.941 (1.839)	1.884 (0.372)	3.677 (1.389)	0.463 (0.323)
	BP	1.235 (1.023)	1.651 (0.274)	2.792 (0.854)	0.684 (0.181)
	Random Forest	**0.832** (0.679)	1.571 (0.193)	2.502 (0.621)	0.718 (0.154)
	HDVBHMM	0.910 (0.717)	**1.321** (0.253)	**1.804** (0.718)	**0.763** (0.393)

In the Long-Term and Short-Term Memory methods, the learning rate is lr = 0.001, the number of training cycles (Epochs) is set to 50, and the size of the hidden layer is set to 10. The hidden layer of the BP method consists of 20 neurons, and the maximum number of iterations is set to 10,000. In the random forest regression model, the number of trees is set to 100. The bold results are the optimal ones among four methods.

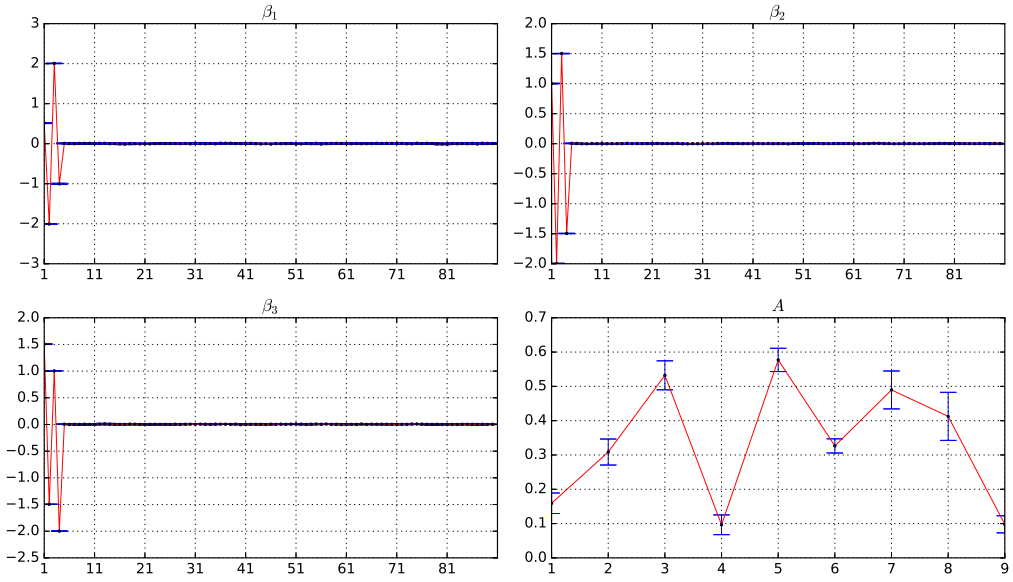

Figure 2. Box plots of the estimator values of A, β_1, β_2, β_3 based on 10 experiments under $p = 90$ and $T = 600$. The horizontal coordinate is the index of the variables and the vertical coordinate is the value of the estimators.

Table 4. Average values of the RMSE and Bias of A and β based on 10 replications in Experiment 2.

p	Parameter	Estimate Performance	
		RMSE	Bias
60	β	0.001	0.001
	A	0.002	0.001
90	β	0.001	0.001
	A	0.005	0.001
120	β	0.015	0.007
	A	0.034	0.009

5. Application to Real Datasets

In this section, we focus on Beijing Multi-Site Air-Quality data, which include 6 major air pollutants and 6 related meteorological variables at multiple locations in Beijing. These air-quality measurements are created by the Beijing Municipal Environmental Monitoring Center. In addition, meteorological data at each air quality location are paired with the nearest weather station provided by the China Meteorological Administration. The data span from 1 March 2013 to 28 February 2017. In our study, we consider PM2.5 concentration as response variable, and PM10 concentration, SO_2 concentration, NO_2 concentration, CO concentration, O_3 concentration, Temperature (TEMP), Pressure (PRES), Dew point temperature (DEWP), Precipitation (RAIN), and Wind speed (WSPM) as covariates; that is, $p = 10$. In order to study the performance on small sample datasets, we delete the missing values in the data and select the data samples in the first 200 time intervals from the Shunyi observation point in Beijing in 2017 for analysis. To assess the predictive performance, we use the first 140 samples as the training set, and the remaining 60 samples as the testing set. We compare the proposed method with the BP neural network, LSTM and Random Forest method similar to Section 4.

One of the main challenges in implementing the HMM is to determine the optimal number of hidden states. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two common model selection techniques, which select the best model by balancing the fitting accuracy and complexity of the model. In selecting the number of hidden states for a Hidden Markov Model, both AIC and BIC evaluate multiple models containing different numbers of states and select an optimal model that balances fitting accuracy and complexity. Multiple HMMs are trained separately using different numbers of hidden states, then the AIC or BIC values are calculated for each model, and finally the model with the smallest AIC or BIC value is selected [34]. Similar to the work of Dofadar et al. [34], we use AIC and BIC to select the number of the hidden states. The AIC equation used in this study is given by $AIC = 2k - 2L$, where k is the number of free parameters in the model and L is the log probability value. The formula for k used in this research is $k = n^2 + 2n - 1$, where n is the current value of the hidden state. The BIC equation used in this study is expressed as $BIC = \ln(T)k - 2L$, where T is the total number of observations. To find the best number of hidden states, we calculate AIC and BIC values based on the different numbers of hidden states: 2, 3, 4, and 5. The results are shown in Figure 3. Figure 3 shows that when the number of hidden states is 3, the AIC and BIC values are the smallest, indicating that choosing the number of hidden states as 3 is the closest to the real model. Therefore, we set the the number of hidden states $K = 3$.

Similar to Section 4, we calculate MAPE, RMSE, MAE and R^2 to evaluate the predictive performance. Since the time series data are positively skewed, MAE and MASE are the best evaluation metrics for evaluating the model performance [35]. The results are shown in Table 5. From Table 5, we can see that the MAPE and MAE of the proposed HDVBHMM method are smaller than ones of other methods, and the R^2 value of the proposed method is larger than one of other methods, indicating that the performance of the HDVBHMM method is better than other methods. Among the other three competing methods, the MAPE and MAE values of the Random Forest method are lowest among

those of the three competing methods, but its MAPE and MAE are still much larger than ones of the proposed method. The BP method is the worst performing among four methods with MAE = 27.570 and MAPE = 1.025.

To better illustrate the predictive performance, Figure 4 shows the true data and predicted values via four methods on the testing set. From Figure 4, we can see that in the first 30 time points, the proposed method fits the true values very well. In the second set of 30 time points, as the prediction time period increases, the predicted values exhibit a slight error, but they are still better than those of other methods. Overall, the prediction accuracy of the proposed method is much better than ones of other methods in term of both short and long time periods.

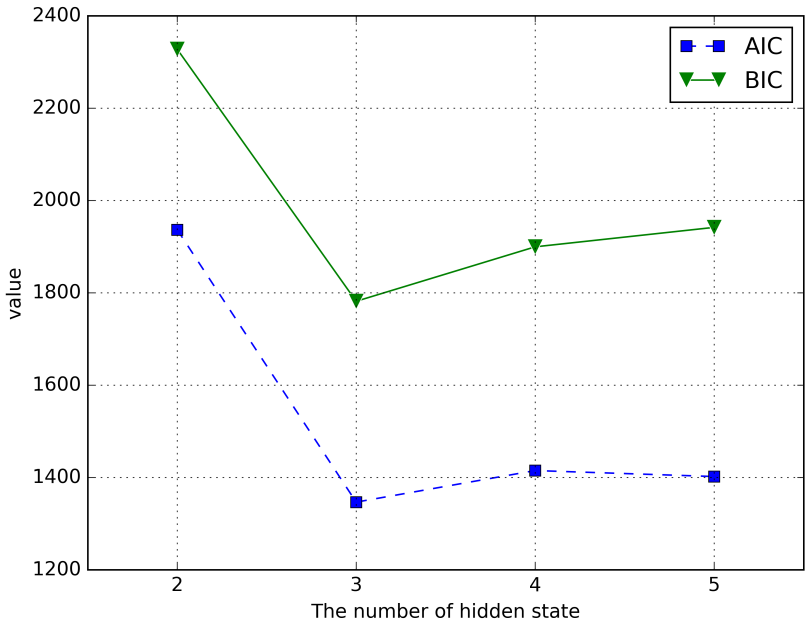

Figure 3. AIC and BIC values when the number of hidden states is 2, 3, 4, and 5 on the real dataset.

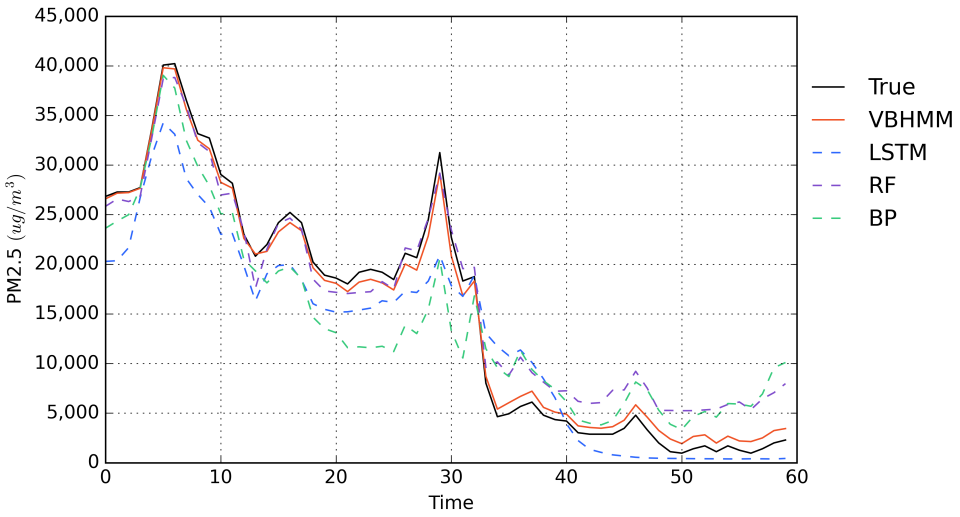

Figure 4. Comparison of observed hourly PM2.5 emissions (test set) with PM2.5 emissions predicted by four methods.

Table 5. Prediction Performance of Four Methods on the testing data.

Method	MAPE	RMSE	MAE	R^2
Random Forest	1.043	4.039	16.318	0.941
BP	1.025	5.250	27.570	0.852
LSTM	0.462	4.924	24.249	0.876
HDVBHMM	**0.317**	**2.249**	**5.058**	**0.993**

In the Long-Term and Short-Term Memory methods, learning rate is lr = 0.001, the number of training cycles (Epochs) is set to 50, and the size of the hidden layer is set to 40. the BP method contains 12 neuron hidden layers and the maximum number of iterations is set to 10,000. The hyperparameters hyperparameters r, δ in the HDVBHMM method were set to 1.0. The bold results are the optimal ones among four methods.

The estimated values of β corresponding to the three states are shown in Table 6. From Table 6, we can see that PM10, SO_2, TEMP (temperature), DEWP (dew point temperature), RAIN (precipitation), and WSPM (wind speed) have the greatest influence on PM2.5 emissions in state 1. PM10, SO_2, TEMP, and DEWP are the four factors that have a negative effect on the presence of PM2.5 emissions in the area, and as these four factors increase, PM2.5 emissions will decrease; meanwhile RAIN and WSPM have a positive effect on the presence of PM2.5 in the area. Rainfall and high wind speed may have increased PM2.5 concentrations through physical effects (such as windblown dust). The prediction formula of the PM2.5 in State 1 is as follows:

$$\begin{aligned} \mathrm{PM2.5}_1 = &-0.833\mathrm{PM10} - 0.349\mathrm{SO}_2 + 0.003\mathrm{CO} - 0.03\mathrm{NO}_2 - 0.09\mathrm{O3} - 2.625\mathrm{TEMP} \\ &- 0.017\mathrm{PRES} - 3.552\mathrm{DEWP} + 1.524\mathrm{RAIN} + 3.547\mathrm{WSPM}. \end{aligned}$$

Table 6. Estimates of the regression coefficients β for each hidden state.

State	PM10	SO_2	NO_2	CO	O_3	TEMP	PRES	DEWP	RAIN	WSPM
State 1	−0.833	−0.349	−0.030	0.003	−0.090	−2.625	−0.017	−3.552	1.524	3.547
State 2	0.965	−0.230	−0.189	−0.003	0.004	−2.154	−0.010	0.745	−4.227	3.891
State 3	−0.303	1.080	1.462	0.000	−0.279	18.55	−0.123	−0.127	9.057	19.217

In addition, PM10, Sulfur Dioxide, Nitrogen Dioxide, TEMP (temperature), DEWP (dew point temperature), RAIN (precipitation), and WSPM (wind speed) have the largest effect on PM2.5 in State 2. The results showed that in state 2, some chemical reactions led to the depletion of gases such as SO_2 and NO_2, which reduced the production of PM2.5, and rainfall also reduced the production of PM2.5. The high wind speed led to an increase in PM2.5 concentration, probably because the wind speed increased the diffusion and transport of particulate matter. The prediction formula of the PM2.5 in this State is as follows:

$$\begin{aligned} \mathrm{PM2.5}_2 = &0.965\mathrm{PM10} - 0.23\mathrm{SO}_2 - 0.003\mathrm{CO} - 0.189\mathrm{NO}_2 - 0.04\mathrm{O3} - 2.154\mathrm{TEMP} \\ &- 0.01\mathrm{PRES} + 0.745\mathrm{DEWP} - 4.227\mathrm{RAIN} + 3.891\mathrm{WSPM}. \end{aligned}$$

PM10, SO_2, NO_2, O_3, TEMP (temperature), PRES (pressure), DEWP (dew point temperature), RAIN (precipitation), and WSPM (wind speed) have the greatest impact on PM2.5 in state 3. It is worth noting that the increase in variables such as SO_2 and NO_2 leads to an increase in PM2.5 concentration. In addition, the significant positive coefficients for temperature indicate that higher temperatures promote the formation of PM2.5, which may be related to the acceleration of certain chemical reactions by high temperatures. The increase in SO_2 and NO_2 may promote the formation of secondary particulate matter, which in turn increases the PM2.5 concentration. Wind speed increases particulate dispersion,

and rainfall may also promote the formation of secondary particulate matter from some soluble substances. The prediction formula of the PM2.5 in this State is as follows:

$$\begin{aligned}\mathrm{PM2.5}_3 = &-0.303\mathrm{PM10} + 1.08\mathrm{SO}_2 + 1.462\mathrm{NO}_2 - 0.279\mathrm{O3} + 18.55\mathrm{TEMP} \\ &- 0.123\mathrm{PRES} - 0.127\mathrm{DEWP} + 9.057\mathrm{RAIN} + 19.217\mathrm{WSPM}.\end{aligned}$$

In summary, the regression coefficients for the three states reflect the effects of different environmental factors on PM2.5 concentrations. The positive and negative signs and magnitudes of these coefficients can provide scenarios on how to manage and predict PM2.5 concentrations by controlling these environmental factors under different environmental conditions. In particular, the fact that temperature, rainfall and wind speed have different effects on PM2.5 concentrations in different states suggests that PM2.5 management needs to take into account complex meteorological conditions and interactions between air pollutants.

6. Conclusions

In this paper, the variable selection for high-dimensional HMMs is studied based on the variational inference. We develop a Variational Bayesian method with the double-exponential shrinkage prior for variable selection. The proposed method can quickly and accurately estimate the posterior distributions and the unknown parameters. In the simulation studies and real data analysis, the proposed method outperforms the common methods in term of variable selection and prediction. In the Beijing Multi-Site Air-Quality analysis, we select the optimal number of the hidden stats based on the AIC and BIC methods, and fit the HMMs of the response variable PM2.5. In the current research work, we investigate variational inference for linear HMMs with high dimensional covariates; that is, the mean of the response variable is linear with respect to the high dimensional covariates. Many of the relationships between variables in practical applications may be not linear, so variational inference for nonlinear HMMs is worth studying. In addition, it is assumed that the variances in observations are the same in different hidden states in this study, but in practical applications, heteroskedasticity may be more in line with real-world data characteristics. For that reason, the heteroskedasticity issue for HMMs is also worth exploring deeply. Moreover, Ivan Gorynin's work [36] verifies that the Pairwise Markov Model (PMM) outperforms the traditional HMM in terms of accuracy when the observed variable y is highly autocorrelated or when the hidden chain is not Markovian. Unlike the HMM, which assumes that the hidden chain z is Markovian, the PMM assumes that (z, y) is Markovian. Since hidden chains are not necessarily Markovian in the PMM, it is more general than the HMM. Parameter estimation of PMM models is done using Variational Bayesian methods in the work of Katherine Morales [37]. However, the effect of including the covariate x on the target variable y was not considered in their work. Therefore, as an extension of the proposed method, which replaces the HMM with the PMM, the inclusion of high-dimensional covariates in the PMM may yield more accurate predictions.

Author Contributions: Conceptualization, Y.Z. (Yao Zhai), W.L. and Y.J.; methodology, Y.Z. (Yao Zhai), W.L. and Y.Z. (Yanqing Zhang); software, W.L., Y.Z. (Yao Zhai) and Y.J.; validation, Y.Z. (Yao Zhai), W.L., Y.J. and Y.Z. (Yanqing Zhang); formal analysis, Y.Z. (Yao Zhai) and Y.J.; investigation, Y.Z. (Yao Zhai), W.L. and Y.J.; resources, Y.Z. (Yao Zhai) and Y.J.; data curation, Y.Z. (Yao Zhai); writing—original draft preparation, Y.Z.; writing—review and editing, Y.J. and Y.Z. (Yanqing Zhang); visualization, Y.J.; supervision, Y.J.; project administration, Y.J.; funding acquisition, Y.J. and Y.Z. (Yanqing Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (No. 102022YFA1003701), the National Natural Science Foundation of China (No. 12271472, 12231017, 12001479, 11871420), the Natural Science Foundation of Yunnan Province of China (No. 202101AU070073 and 202201AT070101), and Yunnan University Graduate Student Research and Innovation Fund Project Grant (No. KC-22221108).

Data Availability Statement: The research data are available on the website https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data (accessed on 23 March 2024).

Acknowledgments: We would like to thank the action editors and referees for insightful comments and suggestions which improve the article significantly.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. Variational Posterior of Parameters

We derive the optimal variational densities based on Formula (8). The complete likelihood function of the model is:

$$\begin{aligned} & P\left(Y, z \mid X, \pi, A, \beta, \sigma^2\right) \\ & = P(z_1|\pi) \prod_{t=2}^{T} P(z_t \mid z_{t-1}, A) \prod_{t=1}^{T} P\left(y_t \mid x_t, z_t, \beta, \sigma^2\right) \\ & = \left(\prod_{k=1}^{K} \pi_k^{v_{1k}}\right) \left(\prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{v_{t-1,j} v_{tk}}\right) \left(\prod_{t=1}^{T} \prod_{k=1}^{K} P\left(y_t \mid x_t, \beta_k, \sigma^2\right)^{v_{tk}}\right). \end{aligned}$$

We derive the conditional posterior distribution of A as:

$$\begin{aligned} & P(A \mid \cdot) \propto P(Y, z \mid X, \pi, A, \beta, \sigma^2) P(A) \\ & \propto P(z_1|\pi) \left[\prod_{t=2}^{T} P(z_t \mid z_{t-1}, A)\right] \left[\prod_{t=1}^{T} P(y_t \mid x_t, z_t, \beta, \tau)\right] P(A) \\ & \propto \left(\prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{v_{t-1,j} \cdot v_{tk}}\right) \left(\prod_{j=1}^{K} \prod_{k=1}^{K} A_{jk}^{\alpha_{jk}^{(A)}-1}\right) \\ & \propto \prod_{j=1}^{K} \left(\prod_{k=1}^{K} A_{jk}^{\sum_{t=2}^{T} v_{t-1,j} \cdot v_{tk} + \alpha_{jk}^{(A)} - 1}\right). \end{aligned}$$

So $P\left(A_j \mid \cdot\right) \sim D_{ir}\left(\sum_{t=2}^{T} v_{t-1,j} \cdot v_{tk} + \alpha_{jk}^{(A)}\right)$ for $j = 1, \ldots K$.
According to Equation (10), The variational posterior distribution of A_j is given by:

$$\begin{aligned} q^*\left(A_j\right) & \propto \exp\left\{E\left[\log P\left(A_j \mid \cdot\right)\right]\right\} \\ & \propto \exp\left\{E\left[\sum_{k=1}^{K} \left(\sum_{t=2}^{T} v_{t-1,j} \cdot v_{tk} + \alpha_{jk}^{(A)} - 1\right) \log A_{jk}\right]\right\} \\ & \propto \prod_{k=1}^{K} A_{jk}^{\sum_{t=2}^{T} E\left(v_{t-1,j} v_{tk}\right) + \alpha_{jk}^{(A)} - 1}. \end{aligned}$$

So $q^*\left(A_j\right) \sim \mathrm{Dir}\left(\sum_{t=2}^{T} E\left(v_{t-1,j} v_{tk}\right) + \alpha_{jk}^{(A)}\right)$.

Similarly, we derive the conditional posterior distribution of σ^2 as:

$$\begin{aligned} P\left(\sigma^2 \mid \cdot\right) &\propto \prod_{t=1}^{T} P\left(y_t \mid \sigma^2, x_t, z_t, \beta\right) P\left(\sigma^2\right) \\ &\propto \prod_{t=1}^{T}\prod_{k=1}^{K}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{v_{tk}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_t - x_t^{\top}\beta_k\right)^2\right\}^{v_{tk}} \sigma^{-2} \\ &\propto \left(\sigma^2\right)^{-\left(\frac{T}{2}+1\right)} \exp\left\{-\frac{\sum_{t=1}^{T}\sum_{k=1}^{K} v_{tk}\left(y_t - x_t^{\top}\beta_k\right)^2}{2\sigma^2}\right\} \\ &\sim \text{IGamma}\left(\frac{T}{2}, \frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K} v_{tk}\left(y_t - x_t^{\top}\beta_k\right)^2\right). \end{aligned}$$

The variational posterior distribution of σ^2 is given by:

$$\begin{aligned} q^*\left(\sigma^2 \mid \cdot\right) &\propto \exp\left\{E\left[-\left(\frac{T}{2}+1\right)\cdot\log\left(\sigma^2\right) + \left(-\frac{\sum_{t=1}^{T}\sum_{k=1}^{K} v_{tk}\left(y_t - x_t^{T}\beta_k\right)^2}{2\sigma^2}\right)\right]\right\} \\ &\propto \left(\sigma^2\right)^{-\left(\frac{T}{2}+1\right)} \exp\left\{-\frac{\sum_{t=1}^{T}\sum_{k=1}^{K} E(v_{tk})\cdot E\left(y_t - x_t^{\top}\beta_k\right)^2}{2\sigma^2}\right\} \\ &\sim \text{IGamma}\left(\frac{T}{2}, \frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K} E(v_{tk}) E\left(y_t - x_t^{\top}\beta_k\right)^2\right). \end{aligned}$$

We derive the conditional posterior distribution of λ^2 as:

$$\begin{aligned} P\left(\lambda^2 \mid \cdot\right) &\propto P\left(\tau_1^2, \ldots, \tau_p^2 \mid \lambda^2\right) P\left(\lambda^2\right) \\ &\propto \prod_{m=1}^{p} \frac{\lambda^2}{2} \exp\left\{-\frac{\lambda^2\tau_m^2}{2}\right\} \left(\lambda^2\right)^{r-1} \exp\left(-\delta\lambda^2\right) \\ &\propto \left(\lambda^2\right)^{p+r-1} \cdot \exp\left\{-\left(\delta + \frac{1}{2}\sum_{m=1}^{p}\tau_m^2\right)\lambda^2\right\} \\ &\sim \Gamma\left(p+r, \delta + \frac{1}{2}\sum_{m=1}^{p}\tau_m^2\right). \end{aligned}$$

The variational posterior distribution of λ^2 is given by:

$$\begin{aligned} q^*\left(\lambda^2 \mid \cdot\right) &\propto \exp\left\{E\left[\log P\left(\lambda^2 \mid \cdot\right)\right]\right\} \\ &\propto \exp\left\{E\left[(p+r-1)\cdot\log\lambda^2 + \left\{-\left(\delta + \frac{1}{2}\sum_{m=1}^{p}\tau_m^2\right)\lambda^2\right\}\right]\right\} \\ &\propto \left(\lambda^2\right)^{p+r-1} \cdot \exp\left\{-\left(\delta + \frac{1}{2}\sum_{m=1}^{p} E\left(\tau_m^2\right)\right)\lambda^2\right\} \\ &\sim \Gamma\left(p+r, \delta + \frac{1}{2}\sum_{m=1}^{p} E\left(\tau_m^2\right)\right). \end{aligned}$$

We derive the conditional posterior distribution of τ_m^2. Note that since the variational posterior of τ_m^2 is difficult to obtain, we derive the variational posterior of $\frac{1}{\tau_m^2}$ as:

$$\begin{aligned}
P\left(\tau_m^2 \mid \cdot\right) &\propto P\left(\beta_{1m}, \beta_{2m}, \ldots \beta_{Km} \mid \tau_m^2\right) P\left(\tau_m^2\right) \\
&\propto \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}\sigma\tau_m} \exp\left\{-\frac{1}{2\tau_m^2\sigma^2}\beta_{km}^2\right\} \exp\left\{-\frac{\lambda^2\tau_m^2}{2}\right\} \\
&\propto \frac{1}{\tau_m^K} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{k=1}^{K}\beta_{km}^2\sigma^{-2}}{\tau_m^2} + \lambda^2\tau_m^2\right)\right\} \\
&\propto (\tau_m^2)^{-K/2} \exp\left(-\frac{1}{2}\left(\frac{\sum_{k=1}^{K}\beta_{km}^2}{\sigma^2}(\tau_m^2)^{-1} + \lambda^2\tau_m^2\right)\right).
\end{aligned}$$

The variational posterior distribution of τ_m^2 is given by:

$$\begin{aligned}
q^*\left(\tau_m^2\right) &\propto \exp\left\{E\left[\log P\left(\tau_m^2 \mid \cdot\right)\right]\right\} \\
&\propto \exp\left\{E\left[-\frac{K}{2}\log\tau_m^2 - \frac{1}{2}\left(\frac{\sum_{k=1}^{K}\beta_{km}^2}{\sigma^2}(\tau_m^2)^{-1} + \lambda^2\tau_m^2\right)\right]\right\} \\
&\propto (\tau_m^2)^{-\frac{K}{2}} \cdot \exp\left\{-\frac{1}{2}\left(E(\frac{1}{\sigma^2})\sum_{k=1}^{K}E(\beta_{km}^2)(\tau_m^2)^{-1} + E(\lambda^2)\tau_m^2\right)\right\} \\
&\sim \text{Generalized-Inverse–Gaussian}(C_{\tau_m}, a_{\tau_m}, b_{\tau_m}),
\end{aligned}$$

where $a_{\tau_m} = E(\lambda^2)$, $b_{\tau_m} = E(1/\sigma_2)\sum_{k=1}^{K}E(\beta_{km}^2)$, and $C_{\tau_m} = 1 - K/2$.

We derive the conditional posterior distribution of β as:

$$\begin{aligned}
P(\beta \mid \cdot) &\propto P(Y, z \mid X, \pi, A, \beta, \sigma^2)P(\beta) \\
&\propto \prod_{t=1}^{T} P\left(y_t \mid x_t, z_t, \beta, \sigma^2\right) P(\beta) \\
&\propto \left(\prod_{t=1}^{T}\prod_{k=1}^{K} P\left(y_t \mid x_t, \beta_k, \sigma^2\right)^{v_{tk}}\right)\left(\prod_{k=1}^{K} P\left(\beta_k \mid \sigma^2, \tau_1^2, \ldots \tau_p^2\right)\right) \\
&\propto \prod_{k=1}^{K}\left[\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}v_{tk}\left(y_t - x_t^\top\beta_k\right)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}\beta_k^\top D_\tau^{-1}\beta_k\right\}\right] \\
&\propto \prod_{k=1}^{K}\left[\exp\left\{-\frac{1}{2}\left(\beta_k^\top\left(\frac{1}{\sigma^2}\sum_{t=1}^{T}v_{tk}x_t x_t^\top + \frac{1}{\sigma^2}D_\tau^{-1}\right)\beta_k - 2\frac{1}{\sigma^2}\sum_{t=1}^{T}v_{tk}y_t\beta_k^\top x_t\right)\right\}\right] \\
&\sim \prod_{k=1}^{K} N_p(\beta_k, \mu_k, \Sigma_k),
\end{aligned}$$

$$\text{where } \mu_k = \left(\frac{1}{\sigma^2}\sum_{t=1}^{T}v_{tk}x_t x_t^\top + \frac{1}{\sigma^2}D_\tau^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\sum_{t=1}^{T}y_t v_{tk}x_t\right),$$

$$\Sigma_k = \left(\frac{1}{\sigma^2}\sum_{t=1}^{T}v_{tk}x_t x_t^\top + \frac{1}{\sigma^2}D_\tau^{-1}\right)^{-1}.$$

The variational posterior distribution of β is given by:

$$\begin{aligned}
q^*(\beta) &\propto \exp\{E[\log P(\beta \mid \cdot)]\} \\
&\propto \exp\left\{E\left[\sum_{k=1}^{K}\left\{-\frac{1}{2}\left(\beta_k^\top\left(\frac{1}{\sigma^2}\sum_{t=1}^{T} v_{tk}x_t x_t^\top + \frac{1}{\sigma^2}D_\tau^{-1}\right)\beta_k - 2\frac{1}{\sigma^2}\sum_{t=1}^{T} v_{tk}y_t\beta_k^T x_t\right\}\right]\right\} \\
&\propto \prod_{k=1}^{K}\left[\exp\left\{-\frac{1}{2}\left(\beta_k^\top\left(E\left(\frac{1}{\sigma^2}\right)\sum_{t=1}^{T} E(v_{tk})x_t x_t^\top + E\left(\frac{1}{\sigma^2}\right)\cdot E\left(D_\tau^{-1}\right)\right)\beta_k\right.\right.\right. \\
&\left.\left.\left. - E\left(\frac{2}{\sigma^2}\right)\sum_{t=1}^{T} E(v_{tk})y_t\beta^\top x_t\right)\right\}\right] \\
&\sim \prod_{k=1}^{K} N_p(\beta_k; \mu_k, \Sigma_k),
\end{aligned}$$

where

$$\begin{aligned}
\mu_k &= \left(E\left(\frac{1}{\sigma^2}\right)\sum_{t=1}^{T} E(v_{tk})x_t x_t^\top + E\left(\frac{1}{\sigma^2}\right)E\left(D_\tau^{-1}\right)\right)^{-1}\left(E\left(\frac{1}{\sigma^2}\right)\sum_{t=1}^{T} y_t\cdot E(v_{tk})\cdot x_t\right), \\
\Sigma_k &= \left(E\left(\frac{1}{\sigma_2}\right)\sum_{t=1}^{T} E(v_{tk})x_t x_t^\top + E\left(\frac{1}{\sigma^2}\right)E\left(D_\tau^{-1}\right)\right)^{-1}.
\end{aligned}$$

We derive the conditional posterior distribution of π:

$$\begin{aligned}
P(\pi \mid \cdot) &\propto P\left(Y, z \mid X, \pi, A, \beta, \sigma^2\right)P(\pi) \\
&\propto P(z_1|\pi)P(\pi) \\
&\propto \prod_{k=1}^{K}\pi_k^{v_{1k}}\cdot\prod_{k=1}^{K}\pi_k^{\alpha_k^{(\pi)}-1} \\
&\sim \mathrm{Dir}\left(z_1 + \alpha^{(\pi)}\right).
\end{aligned}$$

The variational posterior distribution of π is given by:

$$\begin{aligned}
q^*(\pi) &\propto \exp\{E[\log P(\pi|\cdot)]\} \\
&\propto \exp\left\{E\left[\sum_{k=1}^{K}\left(v_{1k} + \alpha_k^{(\pi)} - 1\right)\log\pi_k\right]\right\} \\
&\sim \mathrm{Dir}\left(E(z_1) + \alpha^{(\pi)}\right).
\end{aligned}$$

Finally, we derive the variational posterior of z. Based on the dependencies of hidden states, we divide the variational posterior of z into the following three parts.

We derive the conditional posterior distribution of z_1 as:

$$\begin{aligned}
P(z_1 \mid \cdot) &\propto P\left(Y, z \mid X, \pi, A, \beta, \sigma^2\right) \\
&\propto P(z_1 \mid \pi)\left[\prod_{t=2}^{T} P(z_t \mid z_{t-1}, A)\right]\left[\prod_{t=1}^{T} P\left(y_t \mid x_t, z_t, \beta, \sigma^2\right)\right] \\
&\propto \left(\prod_{k=1}^{K}\pi_k^{v_{1k}}\right)\left(\prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{v_{1j}v_{2k}}\right)\left(\prod_{k=1}^{K} P\left(y_1 \mid x_1, \beta_k, \sigma^2\right)^{v_{1k}}\right) \\
&\propto \prod_{k=1}^{K}\left[\pi_k P\left(y_1 \mid x_1, \beta_k, \sigma^2\right)\prod_{j=1}^{K} A_{kj}^{v_{2j}}\right]^{v_{1k}} \\
&\sim \ \mathrm{Mult}\left(\pi_k \cdot P\left(y_1 \mid x_1, \beta_k, \sigma^2\right)\prod_{j=1}^{K} A_{kj}^{v_{2j}}\right).
\end{aligned}$$

The variational posterior distribution of z_1 is given by:

$$\begin{aligned}
q^*(z_1) &\propto \exp\{E[\log P(z_1 \mid \cdot)]\} \\
&\propto \exp\left\{\sum_{k=1}^{K} v_{1k} E\left[\log \pi_k + \log P\left(y_1 \mid x_1, \beta_k, \sigma^2\right) + \sum_{j=1}^{K} v_{2j} \log A_{kj}\right]\right\} \\
&\propto \prod_{k=1}^{K}\left[\exp\{E[\log \pi_k]\} \exp\left\{E\left[\log P\left(y_1 \mid x_1, \beta_k, \sigma^2\right)\right]\right\} \cdot \prod_{j=1}^{K} \exp\left\{E\left[\log A_{kj}\right] \cdot E(v_{2j})\right\}\right]^{v_{1k}} \\
&\sim \text{Mult}\left(\exp\{E[\log \pi_k]\} \exp\left\{E\left[\log P\left(y_1 \mid x_1, \beta_k, \sigma^2\right)\right]\right\} \prod_{j=1}^{K} \exp\left\{E\left[v_{2j}\right] E\left[\log A_{kj}\right]\right\}\right).
\end{aligned}$$

We derive the conditional posterior distribution of z_t for $t = 2, \ldots, T-1$ as:

$$\begin{aligned}
P(z_t \mid \cdot) &\propto P\left(Y, z \mid X, \pi, A, \beta, \sigma^2\right) \\
&\propto P(z_1 \mid \pi)\left[\prod_{t=1}^{T}\prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{v_{t-1,j} v_{t,k}}\right]\left[\prod_{t=1}^{T}\prod_{k=1}^{K} P\left(y_t \mid x_t, \beta_k, \sigma^2\right)^{v_{tk}}\right] \\
&\propto \prod_{k=1}^{K}\left[\prod_{j=1}^{K} A_{jk}^{v_{t-1,j}} A_{kj}^{v_{t+1},j} P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right]^{v_{tk}} \\
&\sim \text{Mult}\left(\prod_{j=1}^{k} A_{j_k}^{v_{t-1,j}} A_{kj}^{v_{t+1,j}} P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right).
\end{aligned}$$

The variational posterior distribution of z_t for $t = 2, \ldots, T-1$ is given by:

$$\begin{aligned}
q^*(z_t) &\propto \exp\{E[\log P(z_t \mid \cdot)]\} \\
&\propto \exp\left\{\sum_{k=1}^{K} v_{tk} \cdot E\left[\log P\left(y_t \mid x_t, \beta_k, \sigma^2\right) + \sum_{j=1}^{K} v_{t-1,j} \log A_{jk} + \sum_{j=1}^{K} v_{t+1,j} \log A_{kj}\right]\right\} \\
&\propto \prod_{k=1}^{K}\left[\exp\left\{E\left[\log P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right]\right\} \prod_{j=1}^{K} \exp\left\{E\left[\log A_{jk}\right] E\left(v_{t-1,j}\right)\right\}\right. \\
&\left. \cdot \prod_{j=1}^{K} \exp\left\{E\left[\log A_{kj}\right] E\left(v_{t+1,j}\right)\right\}\right]^{v_{tk}} \\
&\sim \text{Mult}\left(\exp\left(E\left[\log P\left(y_t \mid x_t, \beta_k, \sigma^2\right)\right]\right) \cdot \prod_{j=1}^{K} \exp\left\{E\left[\log A_{jk}\right] E\left(v_{t-1,j}\right)\right\}\right. \\
&\left. \cdot \prod_{j=1}^{K} \exp\left\{E\left[\log A_{kj}\right] E\left(v_{t+1,j}\right)\right\}\right).
\end{aligned}$$

We derive the conditional posterior distribution of z_T as:

$$\begin{aligned}
P(z_T \mid \cdot) &\propto \prod_{k=1}^{K}\left(\prod_{j=1}^{K} A_{jk}^{v_{T-1,j}} P\left(y_T \mid x_T, \beta_k, \sigma^2\right)\right)^{v_{Tk}} \\
&\sim \text{Mult}\left(\prod_{j=1}^{K} A_{jk}^{v_{T-1,j}} P\left(y_T \mid x_T, \beta_k, \sigma^2\right)\right).
\end{aligned}$$

The variational posterior distribution of z_T is given by:

$$
\begin{aligned}
q^*(z_T) &\propto \exp\{E[\log P(z_T \mid \cdot)]\} \\
&\propto \exp\left\{\sum_{k=1}^{K} v_{Tk} E\left[\log P\left(y_T \mid x_T, \beta_k, \sigma^2\right) + \sum_{j=1}^{K} v_{T-1,j} \log A_{jk}\right]\right\} \\
&\propto \prod_{k=1}^{K}\left[\exp\left\{E\left[\log\left(y_T \mid x_T, \beta_k, \sigma^2\right)\right]\right\} \cdot \prod_{j=1}^{K} \exp\left\{E\left[\log A_{jk}\right] E\left(v_{T-1,j}\right)\right\}\right]^{v_{Tk}} \\
&\sim \text{Mult}\left(\exp\left\{E\left[\log P\left(y_T \mid x_T, \beta_k, \sigma^2\right)\right]\right\} \cdot \prod_{j=1}^{K} \exp\left\{E\left[\log A_{jk}\right] E\left(v_{T-1,j}\right)\right\}\right).
\end{aligned}
$$

Appendix A.2. Box Plots of the Estimator Values Based on 10 Experiments under $p = 20, 40$ *and* $T = 200$

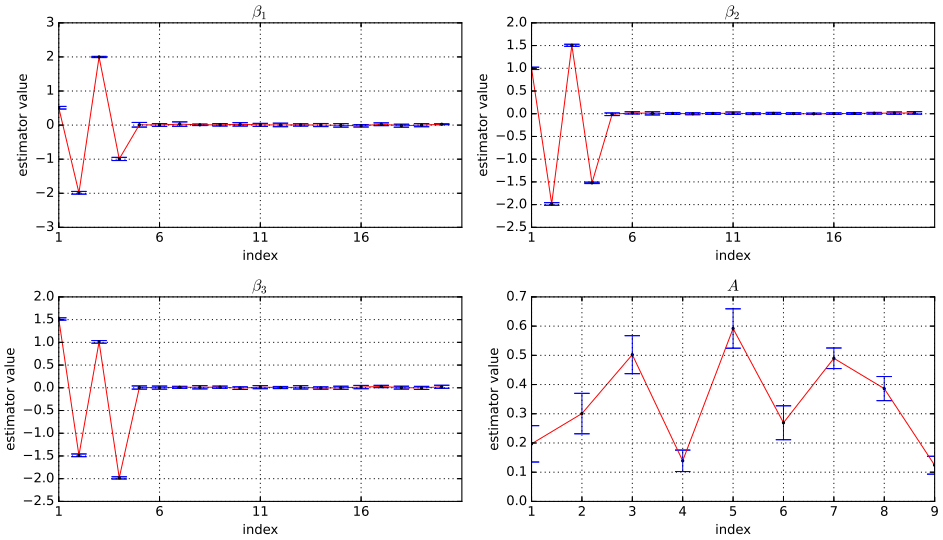

Figure A1. Box plots of the estimator values of A, β_1, β_2, and β_3 based on 50 experiments under $p = 20$ and $T = 200$. The horizontal coordinate is the index of the variables and the vertical coordinate is the values of the estimators.

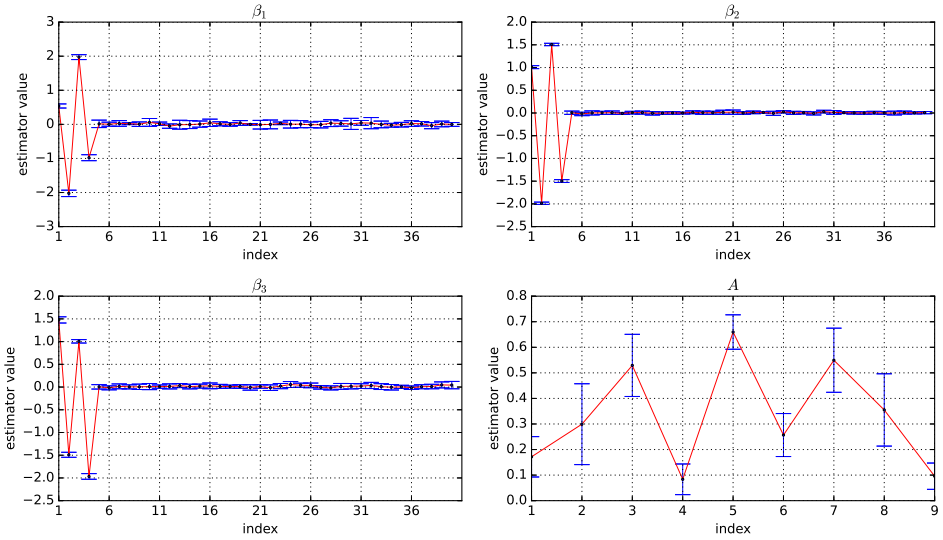

Figure A2. Box plots of the estimator values of A, β_1, β_2, and β_3 based on 50 experiments under $p = 40$ and $T = 200$. The horizontal coordinate is the index of the variables and the vertical coordinate is the values of the estimators.

Appendix A.3. Box Plots of the Estimator Values Based on 50 Experiments under $p = 60, 120$ and $T = 600$

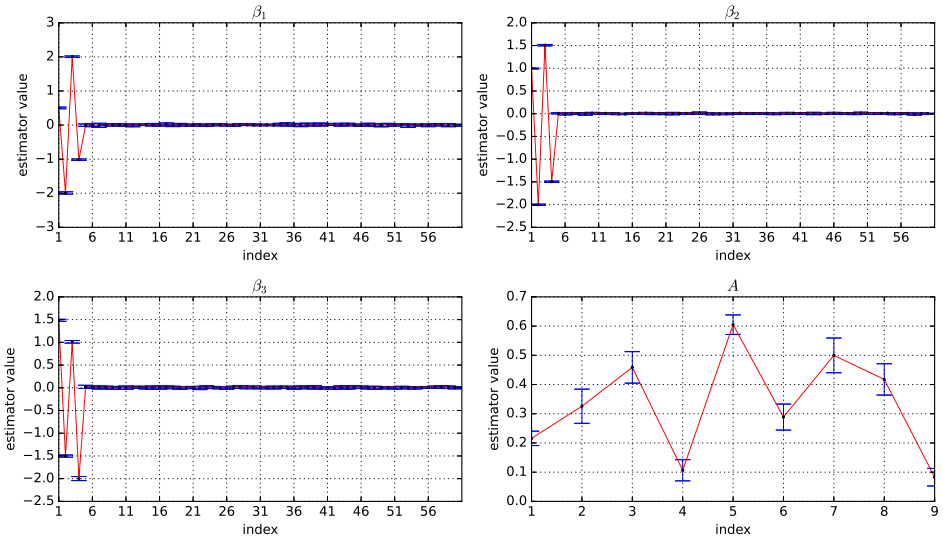

Figure A3. Box plots of the estimator values of A, β_1, β_2, and β_3 based on 10 experiments under $p = 60$ and $T = 600$. The horizontal coordinate is the index of the variables and the vertical coordinate is the values of the estimators.

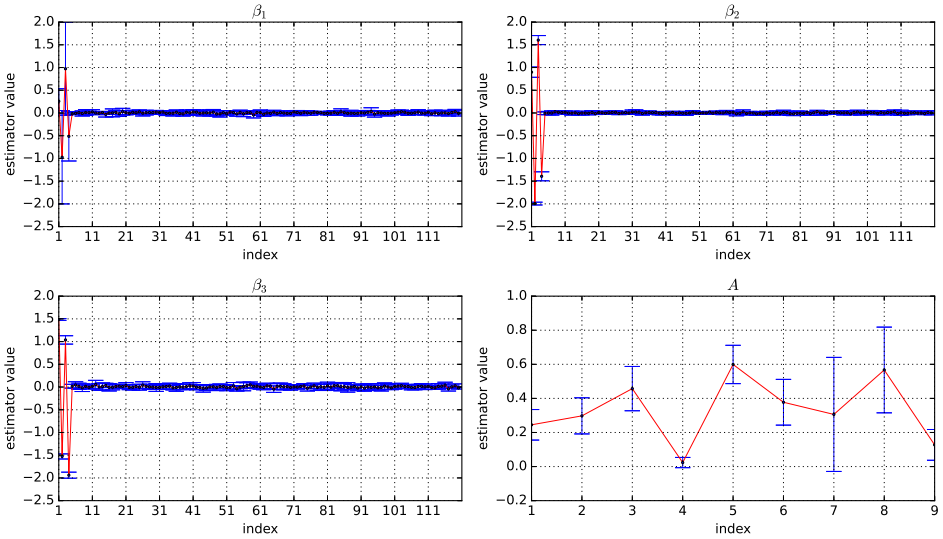

Figure A4. Box plots of the estimator values of A, β_1, β_2, and β_3 based on 10 experiments under $p = 120$ and $T = 600$. The horizontal coordinate is the index of the variables and the vertical coordinate is the values of the estimators.

Appendix A.4. Sensitivity Analysis Results for Different Hyperparameter Settings

To further understand whether the algorithm is sensitive to the choice of hyperparameters r, δ, we conduct experiments on simulated data with a sample size of 300 and a covariate dimension of 30 similar to Section 4. Consider the following experiments with three different hyperparameter settings, the first $r = 0.5, \delta = 0.5$; the second $r = 1.0, \delta = 1.0$; and the third $r = 1.5$, $\delta = 1.5$.

Table A1. The hyperparameters were set to $r = 0.5, \delta = 0.5$; $r = 1.0, \delta = 1.0$; and $r = 1.5$, $\delta = 1.5$ to compute the mean of the four metrics, with standard deviations in parentheses, based on 10 simulations under the conditions of $T = 300$, $p = 30$.

p	Method	Estimate Performance			
		MAPE	**RMSE**	**MAE**	R^2
$p = 30$	$r = 0.5, \delta = 0.5$	0.8766 (0.9195)	1.1861 (0.2441)	1.4605 (0.6291)	0.8617 (0.0739)
	$r = 1.0, \delta = 1.0$	0.8766 (0.9195)	1.1861 (0.2441)	1.4606 (0.6291)	0.8616 (0.0739)
	$r = 1.5, \delta = 1.5$	0.8766 (0.9195)	1.1861 (0.2440)	1.4606 (0.6290)	0.8616 (0.0739)

The experimental results show that the estimation results are not sensitive to the choice of the two hyperparameters r and δ. The images of the Gamma distributions for the three different hyperparameter settings are very similar in shape. It implies that the performances of the proposed method are not sensitive to these hyperparameters for a certain range of variations in r and δ values.

References

1. Baum, L.E.; Petrie, T.; Soules, G.; Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **1970**, *41*, 164–171. [CrossRef]
2. Forney, G.D. The viterbi algorithm. *Proc. IEEE* **1973**, *61*, 268–278. [CrossRef]
3. LeGland, F.; Mével, L. Recursive Estimation in Hidden Markov Models. In Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, USA, 12 December 1997; Volume 4, pp. 3468–3473.
4. Ford, J.J.; Moore, J.B. Adaptive estimation of HMM transition probabilities. *IEEE Trans. Signal Process.* **1998**, *46*, 1374–1385. [CrossRef]
5. Djuric, P.M.; Chun, J.H. An MCMC sampling approach to estimation of nonstationary hidden Markov models. *IEEE Trans. Signal Process.* **2002**, *50*, 1113–1123. [CrossRef]
6. Ma, Y.A.; Foti, N.J.; Fox, E.B. Stochastic gradient MCMC methods for Hidden Markov Models. In Proceedings of the International Conference on Machine Learning Research, Sydney, Australia, 6–11 August 2017; pp. 2265–2274.
7. Dellaportas, P.; Roberts, G.O. An introduction to MCMC. In *Spatial Statistics and Computational Methods*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 1–41.
8. Neal, R.M. MCMC using Hamiltonian dynamics. *Handb. Markov Chain. Monte Carlo* **2011**, *2*, 2.
9. Box, G.E.; Tiao, G.C. *Bayesian Inference in Statistical Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
10. Scott, S.L. Bayesian methods for Hidden Markov Models: Recursive computing in the 21st century. *J. Am. Stat. Assoc.* **2002**, *97*, 337–351. [CrossRef]
11. Rydén, T. EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Anal.* **2008**, *3*, 659–688. [CrossRef]
12. Brooks, S.P.; Roberts, G.O. Convergence assessment techniques for Markov chain Monte Carlo. *Stat. Comput.* **1998**, *8*, 319–335. [CrossRef]
13. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233. [CrossRef]
14. Tzikas, D.G.; Likas, A.C.; Galatsanos, N.P. The variational approximation for Bayesian inference. *IEEE Signal Process. Mag.* **2008**, *25*, 131–146. [CrossRef]
15. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**.
16. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
17. Wang, Y.; Blei, D.M. Frequentist Consistency of Variational Bayes. *J. Am. Stat. Assoc.* **2019**, *114*, 1147–1161. [CrossRef]
18. Han, W.; Yang, Y. Statistical inference in mean-field Variational Bayes. *arXiv* **2019**, arXiv:1911.01525.
19. Ranganath, R.; Gerrish, S.; Blei, D. Black box variational inference. In Proceedings of the Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014; pp. 814–822.
20. MacKay, D.J. *Ensemble Learning for Hidden Markov Models*; Technical Report; Cavendish Laboratory, University of Cambridge: Cambridge, UK, 1997.
21. McGrory, C.A.; Titterington, D. Variational Bayesian analysis for Hidden Markov Models. *Aust. N. Z. J. Stat.* **2009**, *51*, 227–244. [CrossRef]

22. Foti, N.; Xu, J.; Laird, D.; Fox, E. Stochastic variational inference for Hidden Markov Models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2014; pp. 1–9.
23. Gruhl, C.; Sick, B. Variational Bayesian inference for Hidden Markov Models with multivariate Gaussian output distributions. *arXiv* **2016**, arXiv:1605.08618.
24. Ding, N.; Ou, Z. Variational nonparametric Bayesian Hidden Markov Model. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 2098–2101.
25. Park, T.; Casella, G. The bayesian lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
26. Meinshausen, N. Relaxed lasso. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393. [CrossRef]
27. Hans, C. Bayesian lasso regression. *Biometrika* **2009**, *96*, 835–845. [CrossRef]
28. Ranstam, J.; Cook, J. LASSO regression. *J. Br. Surg.* **2018**, *105*, 1348. [CrossRef]
29. Ye, L.; Beskos, A.; De Iorio, M.; Hao, J. Monte Carlo co-ordinate ascent variational inference. *Stat. Comput.* **2020**, *30*, 887–905. [CrossRef]
30. Jaakkola, T.S. Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*; The MIT Press: Cambridge, MA, USA, 2000; pp. 129–159.
31. Jaakkola T.S., Jordan, M. Bayesian parameter estimation via variational methods. *Stat. Comput.* **2000**, *10*, 25–37. [CrossRef]
32. Tran, M.-N.; Nott, D.J.; Kuk, A.Y.C.; Kohn, R. Parallel Variational Bayes for Large Datasets with an Application to Generalized Linear Mixed Models. *J. Comput. Graph. Stat.* **2016**, *25*, 626–646. [CrossRef]
33. Winn, J.; Bishop, C.M.; Jaakkola, T. Variational message passing. *J. Mach. Learn. Res.* **2005**, *6*, 661–694.
34. Dofadar, D.F.; Khan, R.H.; Alam, M.G.R. COVID-19 Confirmed Cases and Deaths Prediction in Bangladesh Using Hidden Markov Model. In Proceedings of the 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), Paris, France, 21 January 2022; pp. 1–4.
35. Shoko, C.; Sigauke, C. Short-term forecasting of COVID-19 using support vector regression: An application using Zimbabwean data. *Am. J. Infect. Control.* **2023**, *51*, 1095–1107. [CrossRef]
36. Gorynin, I.; Gangloff, H.; Monfrini, E.; Pieczynski, W. Assessing the segmentation performance of pairwise and triplet Markov models. *Signal Process.* **2018**, *145*, 183–192. [CrossRef]
37. Morales, K.; Petetin, Y. Variational Bayesian inference for pairwise Markov models. In Proceedings of the 2021 IEEE Statistical Signal Processing Workshop (SSP), Rio de Janeiro, Brazil, 11–14 July 2021; pp. 251–255.