

Article

Bayesian Inference of Recurrent Switching Linear Dynamical Systems with Higher-Order Dependence

Houxiang Wang ¹ and Jiaqing Chen ^{1,2,*}¹ School of Science, Wuhan University of Technology, Wuhan 430071, China; 273212@whut.edu.cn² Hubei Longzhong Laboratory, Xiangyang 441100, China

* Correspondence: 9523@whut.edu.cn

Abstract: Many complicated dynamical events may be broken down into simpler pieces and efficiently described by a system that shifts among a variety of conditionally dynamical modes. Building on switching linear dynamical systems, we develop a new model that extends the switching linear dynamical systems for better discovering these dynamical modes. In the proposed model, the linear dynamics of latent variables can be described by a higher-order vector autoregressive process, which makes it feasible to evaluate the higher-order dependency relationships in the dynamics. In addition, the transition of switching states is determined by a stick-breaking logistic regression, overcoming the limitation of a restricted geometric state duration and recovering the symmetric dependency between the switching states and the latent variables from asymmetric relationships. Furthermore, logistic regression evidence potentials can appear as conditionally Gaussian potentials by utilizing the Pólya-gamma augmentation strategy. Filtering and smoothing algorithms and Bayesian inference for parameter learning in the proposed model are presented. The utility and versatility of the proposed model are demonstrated on synthetic data and public functional magnetic resonance imaging data. Our model improves the current methods for learning the switching linear dynamical modes, which will facilitate the identification and assessment of the dynamics of complex systems.

Keywords: dynamic systems; state space methods; Kalman filter; Bayesian inference



Citation: Wang, H.; Chen, J. Bayesian Inference of Recurrent Switching Linear Dynamical Systems with Higher-Order Dependence. *Symmetry* **2024**, *16*, 474. <https://doi.org/10.3390/sym16040474>

Academic Editor: Wiesław Leonski

Received: 6 March 2024

Revised: 7 April 2024

Accepted: 10 April 2024

Published: 13 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Complex systems frequently exhibit multiple levels of abstraction in their descriptions [1]. For example, a computer program can be characterized by the collection of functions it calls, the sequence in which it carries out statements or the assembly instructions it sends to the CPU. This assertion is true for a multitude of natural systems. Brain activity can be classified based on either broad psychological states or the activation of individual ion channels. The necessary amount of specificity may differ based on the particular task being performed [2]. Through the identification of these behavioral units and their interdependence, we can obtain a deeper understanding of the intricate mechanisms that give rise to complex natural occurrences. Furthermore, modern machine learning offers powerful tools to help model the dynamics of complex systems. The toolbox has recently been improved to incorporate more versatile elements, such as Gaussian processes [3] and neural networks [4], into probabilistic time series models.

Time series analysis encompasses several methodologies and models, such as stationary process models, spectral models, state space models, and non-linear models [5]. Among them, the state space methods are regarded as more versatile and adept at addressing a broader range of problems compared to the other models. Hidden Markov models (HMM) and switching linear dynamical systems (SLDS) are two well-known state space models. In numerous real-world time series situations, the present condition of a dynamic system is connected to its condition in prior time intervals. Both HMM and SLDS can address

such problems and have been widely used in several problem fields. Examples include human motion [6,7], computer vision [8], speech recognition [9], econometrics [10], machine learning [11] and neuroscience [12,13].

HMM is a discrete Markov process with doubly embedded stochastic models, namely, (i) an unobservable hidden process characterized by a Markov chain and (ii) an observation process determined by hidden states. HMM-based models have garnered significant interest and demonstrated their utility in the signal processing community for accurately simulating the intricate temporal progression of signals. More precisely, HMM have a lengthy track record in signal processing, particularly in the field of speech processing where they have achieved notable success [14]. The SLDS allows for the modeling of nonlinear time series data by dividing it into subsequences controlled by linear dynamics. Indeed, the generative model of SLDS is slightly different from HMM, where there is an additional set of latent variables between the switching states and the observations. In addition, a SLDS differs from a HMM by choosing from a collection of linear Gaussian dynamics that evolve continuously, instead of a standard Gaussian mixture density as in HMM. The SLDS can be viewed as an expansion of the HMM, where each HMM state, or mode, is linked to a linear dynamical process. One further benefit of SLDS is its ability to handle time series data with high dimensionality. From a generative model perspective, it can be claimed that high-dimensional time series can be cheaply represented by a dynamical process specified on a low-dimensional manifold. This representation benefits from the model structure of SLDS. Nevertheless, employing HMM directly on time series with high dimensionality is likely to result in overfitting. [15]. Therefore, SLDS offers the potential for increased descriptive capability [1].

Although SLDS offers advanced predictive models, the dynamics from the data are only stated under certain conditions, which poses some shortcomings of the generative model and may limit its application. First, the continuously evolving linear Gaussian dynamics of latent variables is essentially equivalent to a first-order switching vector autoregressive (VAR) process, i.e., VAR(1) process, which may not be suitable for the case when there are higher-order dependency relationships in the time series. Higher-order autoregressive interactions are more common in time series data [16,17]. Furthermore, determining an appropriate autoregressive order must trade off the model likelihood and complexity, constituting a non-trivial task in the applications [18]. Indeed, a restricted VAR(1) process involved in the linear dynamics of latent variables limits the ability of SLDS to recover the higher-order dependency relationships in dynamical phenomena.

Second, the switching state transition in SLDS follows the Markov assumption, which means that the duration time (i.e., time interval spent in a certain state) is restricted to a geometric distribution [19]. Thus, more weight is given to shorter consecutive time periods within a certain state, which means states switch frequently. This may not be appropriate when the system spends a long time in one state [20]. In addition, the state transitions are based only on the previous state and are unrelated to the latent variables. If a discrete switch happens when the system enters a particular area of the state space, the SLDS cannot recognize this relationship. While hidden semi-Markov model (HSMM) based on the semi-Markov assumption is an extension of the HMM that alleviates this issue, it should be noted that state transitions in HSMM still have no associations with the previous observations [20].

To address the mentioned limitations, an extension of SLDS, the recurrent switching linear dynamical systems with higher-order dependence (HO-rSLDS), is presented. The HO-rSLDS provides a new method that improves the modeling of dynamics from complex systems, by enabling a higher-order dependence in the linear Gaussian dynamics of latent variables and allowing the switching state transition probabilities to hinge on the preceding values of the latent variables. Specifically, to address the first limitation, we enhance the generative process of latent variables by directly making the current latent variable depend on the state values at the previous several time steps, rather than only the previous one time step. Therefore, the proposed model is feasible for capturing the higher-order

autoregressive relationships in the data. To address the second limitation, we utilize the stick-breaking logistic regression to determine the switching state transition [21]. In addition, we leverage a Pólya-gamma augmentation approach [22] to improve block inference updating. This enhancement transforms specific logistic regression evidence potentials into conditionally Gaussian potentials, facilitating efficient Bayesian inference procedures. In this way, the transition of switching states is directly associated with the previous switching state and latent variable. Moreover, since the switching states and the latent variables are mutually interdependent in HO-rSLDS, the symmetric dependency among them is recovered, enhancing the generative process of both switching states and latent variables.

Similar to SLDS, the exact inference in an HO-rSLDS is intractable, which impedes efficient model estimation and parameter learning [8]. In this study, the HO-rSLDS is learned through variational inference, which is an approximate scheme to maximize the evidence lower bound, i.e., minimize the Kullback–Leibler divergence between a restricted family of functions of the model parameters and the actual posterior distribution [23]. Moreover, we propose message-passing algorithms about switching states and latent variables in HO-rSLDS to evaluate sufficient statistics for further parameter updates.

The primary contributions of the novel approach suggested in this study are outlined below:

1. HO-rSLDS assumes the linear dynamics of latent variables can be described by a higher-order VAR process, which makes it feasible to dig out and evaluate the long-term dependency relationships from dynamical phenomena;
2. Stick-breaking logistic regression is applied to determine the switching state transition in HO-rSLDS. By this means, the transition probabilities are time-varying and can be adjusted according to the previous latent variable, which overcomes the limitation of restricted geometric state duration time by Markov assumption and recovers the symmetric dependency between the switching states and the latent variable;
3. The Pólya-gamma augmentation strategy permits efficient Bayesian inference algorithms. In addition, we propose message-passing algorithms, including a forward Kalman filter and backward Kalman smoother for HO-rSLDS, which facilitates the parameter update in variational inference.

The rest of this study is organized as follows. In Section 2.3, we discuss the related models. In Section 3, we present the model learning algorithms in detail. In Section 4, we demonstrate the performance of the proposed HO-rSLDS on simulated data. Our method is applied to public functional magnetic resonance imaging (fMRI) data as well, and the results are described in Section 5. Finally, we wrap up our conclusions in Section 6.

2. Methods

2.1. HMM and HSMM

HMMs are mathematical models that analyze dynamic systems by using random processes with hidden states. A HMM is constructed using a Markov chain and is composed of two layers: a layer representing hidden states and a layer representing observations. The observed sequences are allocated to their respective hidden states according to the observation probability distribution. In the context of HMM, the observations are solely influenced by the current state and are not influenced by any preceding states [20]. The entire state sequence can be presented as $z_{1:T} = \{z_t\}_{t=1}^T$, which can be viewed as a sequence over a finite set \mathcal{Z} with cardinality $|\mathcal{Z}|$. The state transition probability from state i to j is defined as $\pi_{ij} = p(z_{t+1} = j | z_t = i)$. The distribution of observations $y_t \in \mathcal{R}^n$ given a specific state is denoted by $p(y_t | z_t, \theta_j)$, where θ_j denotes the emission parameters of state j . Above all, the HMM can be described as:

$$z_t | z_{t-1} \sim \pi_{z_{t-1}}, \quad (1)$$

$$y_t | z_t \sim F(\theta_{z_t}), \quad (2)$$

where $F(\cdot)$ denotes emission distribution.

Compared to HMM, the HSMM formalism improves upon the standard HMM by incorporating a random state length time, chosen from a state-specific distribution, into the generative process. The state remains constant until the period elapses, at which juncture there is a Markov transition to a fresh state. By using this approach, the distribution of state durations is not limited to a geometric form [24,25]. The random variable D_t represents the length of time that a state is entered at time t . The associated probability mass function is denoted as $p(D_t|z_t = j)$ and the HSMM can be described as

$$\nu_s | \nu_{s-1} \sim \pi_{\nu_{s-1}}, \quad (3)$$

$$D_s \sim g(\omega_s), \quad (4)$$

$$z_{t_s^1:t_s^{D_s+1}} = \nu_s, \quad (5)$$

$$y_{t_s^1:t_s^{D_s+1}} | z_{t_s^1:t_s^{D_s+1}} \sim F(\theta_{\nu_s}, D_s), \quad (6)$$

where ν_s indexes the state shared by state segment s . A graphical illustration of HMM and HSMM is shown in Figure 1. HMM or HSMM typically uses a Gaussian mixture as the emission distribution, which fails to capture potential dependencies in observed time series. An effective approach to tackle this problem is to employ the VAR emission model, which describes the behavior of time series through linear historical interactions among the observed time series. In this way, the emission model in HMM or HSMM can be specifically expressed as:

$$y_t | z_t = k \sim \mathcal{N}\left(\sum_{l=1}^p W_{k,l} y_{t-l}, \Sigma_k\right), \quad (7)$$

where p denotes the maximum lag order, $W_{k,l}$ are $n \times n$ dimensional matrices representing the k -th state autoregressive coefficient matrices for lag l and the covariance matrix is Σ_k . Subsequently, we refer to the H(S)MM with VAR(p) emission model as the AR(p)-H(S)MM.

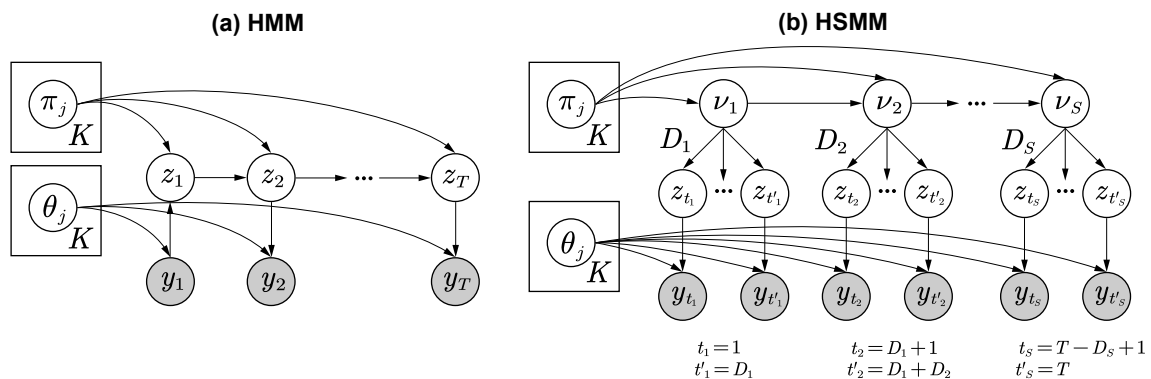


Figure 1. Graphical model of HMM (a) and HSMM (b). An HMM typically consists of two layers: a hidden state layer based on a Markov chain and an observation layer dependent on the current state. An HSMM enhances the generative process of the standard HMM with a random state duration time, which is drawn from some state-specific distribution when the state is entered. K denotes the cardinality of the hidden state space, which can be a finite or infinite value.

2.2. SLDS

A SLDS Markov process differs from a hidden Markov model (HMM) by choosing from a collection of linear Gaussian dynamics that evolve continuously, instead of a standard Gaussian mixture density as in HMM. Therefore, SLDS offers the potential for increased descriptive capability.

Similar to HMM, at each time $t = 1, 2, \dots, T$ there is a discrete switching state $z_t \in \{1, 2, \dots, K\}$ that evolves according to Markov dynamics:

$$z_{t+1}|z_t, \{\pi_k\}_{k=1}^K \sim \pi_{z_t}, \quad (8)$$

$\{\pi_k\}_{k=1}^K$ constitutes the Markov transition matrix and $\pi_k \in [0, 1]^K$ is its k -th row. The z_1 is assumed to be sampled from an initial distribution π_0 parameterized by prior parameter α_0 , i.e., $z_1|\alpha_0 \sim \pi_0(\alpha_0)$. Furthermore, a continuous latent variable $x_t \in \mathbb{R}^m$ follows conditionally linear dynamics depend on the state values with at the previous time step, and the switching state z_t determines the linear dynamical system at time t by:

$$x_{t+1} = A_{z_{t+1}}^{(0)} x_t + v_t, \quad v_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, Q_{z_{t+1}}^{(0)}), \quad (9)$$

where $A_{z_{t+1}}^{(0)} \in \mathbb{R}^{m \times m}$ denotes the coefficient matrix and $Q_{z_{t+1}}^{(0)} \in \mathbb{R}^{m \times m}$ is covariance matrix. At each time t , a linear Gaussian observation $y_t \in \mathbb{R}^n$ is produced from the associated latent continuous variable:

$$y_t = C^{(0)} x_t + w_t, \quad w_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, S^{(0)}), \quad (10)$$

where $C^{(0)} \in \mathbb{R}^{n \times m}$, $S^{(0)} \in \mathbb{R}^{n \times n}$. Usually, n is much larger than m , such that the dependence structure of high-dimensional observation data can be evaluated based on the low-dimensional latent variable space. A graphical illustration of SLDS is given in Figure 2.

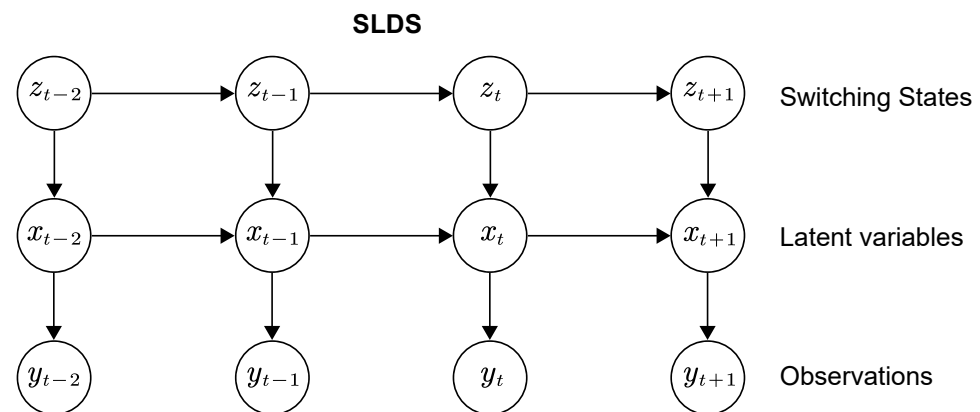


Figure 2. The graphical model of the SLDS, which consists of switching states (z_t), latent variables (x_t) and observations (y_t). Note that we assume it does not contain a link from z_t to y_t .

In classical form, however, there are some shortcomings in the generative model of SLDS, which may limit its application. First, the continuously-evolving linear Gaussian dynamics of x_t is essentially equivalent to a first-order switching vector autoregressive (VAR) process, i.e., the VAR(1) process, which may not be suitable for the case when there are higher-order dependency relationships in x_t or y_t . Substitute Equation (9) into Equation (10) and assume $C^{(0)\top} C^{(0)} = I$ and w_t approximately zero, one can immediately obtain the conclusion that SLDS describes the dependency relationships by a VAR(1) process since $y_t = C^{(0)} A_{z_t}^{(0)} C^{(0)\top} y_{t-1} + C^{(0)} v_t$. Second, the state transition of z_t follows the Markov assumption, which means that the duration time is restricted to a geometric distribution [19]. Thus, more weight is given to shorter consecutive time periods within a certain state, which means states switch frequently. This may not be appropriate when the system spends a long time in one state [20]. In addition, the state z_{t+1} transitions based only on the previous state z_t , and $z_{t+1}|z_t$ is unrelated to the latent variable x_t . If a discrete switch happens when the system enters a particular area of the state space, the SLDS will not be able to recognize this relationship. To address these issues, the proposed

HO-rSLDS enhances the generative process of x_t and the switching state transition, which is demonstrated in the following subsection.

2.3. Recurrent Switching Linear Dynamic System with Higher-Order Dependence

In this section, we describe the proposed HO-rSLDS comprising two main components that are switching linear dynamic systems with higher-order dependence (HO-SLDS, without recurrence) and the stick-breaking logistic regression as well as the Pólya-gamma augmentation for HO-SLDS. Each component is demonstrated below.

2.3.1. HO-SLDS

The switching state transition in HO-SLDS is the same as that in SLDS. However, the continuous latent variable $x_t \in \mathcal{R}^m$ follows conditionally linear dynamics depending on the state values at the previous p time steps, rather than only the previous one time step, and the switching state z_t determines the linear dynamical system at time t by:

$$x_{t+1} = A_{z_{t+1}} \bar{x}_t + v_t, \quad v_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, Q_{z_{t+1}}), \quad (11)$$

where $\bar{x}_t = [x_{t-1}^\top, \dots, x_{t-p}^\top]^\top \in \mathcal{R}^{pm}$, $A_{z_{t+1}} \in \mathcal{R}^{m \times pm}$ and $Q_{z_{t+1}} \in \mathcal{R}^{m \times m}$. Indeed, the evolution of the latent variable dynamics in x_t can be considered as a VAR(p) process. The initial \bar{x}_0 given the initial discrete state z_0 is supposed to be normally distributed with mean μ_0 and covariance Σ_0 , i.e., $\bar{x}_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$. Similarly, at each time t , a linear Gaussian observation $y_t \in \mathcal{R}^n$ is produced from the associated latent continuous variable:

$$y_t = C_{z_t} x_t + w_t, \quad w_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, S_{z_t}), \quad (12)$$

where $C_{z_t} \in \mathcal{R}^{n \times m}$, $S_{z_t} \in \mathcal{R}^{n \times n}$. Note that here we assume C_{z_t} and S_{z_t} are related to z_t . Usually, n is much larger than m , such that the dependence structure of high-dimensional observation data can be evaluated based on the low-dimensional latent variable space. The system parameters consist of the discrete Markov transition matrix and the collection of linear dynamical system matrices, denoted as:

$$\theta = \pi_0 \cup \{\mu_0, \Sigma_0, \pi_k, A_k, Q_k, C_k, S_k\}_{k=1}^K. \quad (13)$$

2.3.2. Stick-Breaking Logistic Regression and Pólya-Gamma Augmentation for HO-SLDS

Another component included in HO-rSLDS is a stick-breaking logistic regression. We utilize a Pólya-gamma augmentation technique to improve block inference updates [21]. This approach allows certain logistic regression evidence potentials to transform to conditionally Gaussian potentials in an augmented distribution, which facilitates the Bayesian inference algorithms.

Consider a logistic regression model with regressors $x \in \mathcal{R}^m$ that maps to a categorical distribution over the switching state $z \in \{1, 2, \dots, K\}$, denoted as:

$$z|x \sim \pi_{\text{SB}}(v), \quad v = Rx + r, \quad (14)$$

where $R \in \mathcal{R}^{K-1 \times M}$ represents a weight matrix and $r \in \mathcal{R}^{K-1}$ denotes a bias vector. We use a stick-breaking link function $\pi_{\text{SB}} : \mathcal{R}^{K-1} \rightarrow [0, 1]^K$, which transforms a real vector into a normalized probability vector based on the stick-breaking process:

$$\pi_{\text{SB}}(v) = \left(\pi_{\text{SB}}(v)^{(1)}, \dots, \pi_{\text{SB}}(v)^{(K)} \right), \quad (15)$$

$$\pi_{\text{SB}}^{(k)} = \sigma(v_k) \prod_{j < k} \sigma(-v_j), \quad (16)$$

for $k = 1, \dots, K-1$ and $\pi_{\text{SB}}^{(K)} = \prod_{k=1}^{K-1} \sigma(-v_k)$ where v_k represents the k th component of v and $\sigma(v) = (1 + e^{-v})^{-1}$ is the logistic function. The probability mass function $p(z|x)$ is given by

$$p(z|x) = \prod_{k=1}^{K-1} \sigma(v_k)^{\mathbb{I}(z=k)} \sigma(-v_k)^{\mathbb{I}(z>k)}, \quad (17)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The posterior density $p(x|z)$ is non-Gaussian and does not allow straightforward Bayesian inference for this regression model due to the likelihood $p(z|x)$ not aligning with a Gaussian prior density $p(x)$. To solve this problem, we introduce Pólya-gamma auxiliary variables $\omega = \{\omega_k\}_{k=1}^K$ so that the conditional density $p(x|z, \omega)$ becomes Gaussian [22]. Specifically, the conditional density of ω_k is distributed according to a Pólya-Gamma distribution. A random variable X has a Pólya-Gamma distribution with parameters $b > 0$ and $c \in \mathcal{R}$, denoted as $X \sim \text{PG}(b, c)$, if

$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)}, \quad (18)$$

where the $g_k \sim \text{Gamma}(b, 1)$ are independent Gamma random variables, and where $\stackrel{D}{=}$ indicates equality in distribution. In particular, by choosing $\omega_k|x, z \sim \text{PG}(\mathbb{I}(z > k), v_k)$, we have the following:

$$x|z, \omega \sim \mathcal{N}(\Omega^{-1}\kappa, \Omega^{-1}), \quad (19)$$

where the mean vector $\Omega^{-1}\kappa$ and covariance matrix Ω^{-1} are calculated based on the following:

$$\Omega = \text{diag}(\omega), \quad (20)$$

$$\kappa_k = \mathbb{I}(z = k) - \frac{1}{2}\mathbb{I}(z > k). \quad (21)$$

Thus allowing efficient block updates while maintaining Gaussian posterior distribution $p(x|z)$ in Bayesian inference.

The HO-rSLDS splits the latent space into K sections, with each section following its own linear dynamics by including the recurrence (Equation (17)) in the transition density of z_t , which can be described as follows:

$$z_{t+1}|z_t, x_t, R_{z_t}, r_{z_t} \sim \pi_{\text{SB}}(v_t), \quad (22)$$

$$v_t = R_{z_t}x_t + r_{z_t}. \quad (23)$$

Moreover, we place Gaussian priors on R_{z_t} and r_{z_t} . In this way, the probability mass function $p(z_{t+1}|x_t)$ can be expressed as:

$$\begin{aligned} p(z_{t+1}|x_t) &= \prod_{k=1}^{K-1} \sigma(v_k)^{\mathbb{I}(z_{t+1}=k)} \sigma(-v_k)^{\mathbb{I}(z_{t+1}>k)} \\ &= \prod_{k=1}^{K-1} \frac{(e^{v_{t,k}})^{\mathbb{I}(z_{t+1}=k)}}{(1 + e^{v_{t,k}})^{\mathbb{I}(z_{t+1}>k)}}. \end{aligned} \quad (24)$$

The Pólya-gamma augmentation focuses on exactly the above densities, utilizing the following integral identity:

$$\frac{(e^v)^a}{(1 + e^v)^b} = 2^{-b} e^{\kappa v} \int_0^{\infty} e^{-\omega v^2} p_{\text{PG}}(\omega|b, 0) d\omega, \quad (25)$$

where $\kappa = a - b/2$ and $p_{\text{PG}}(\omega|b, 0)$ refers to the density of the Pólya-gamma distribution, $\text{PG}(b, 0)$, which is independent of ν [22]. Combining Equations (24) and (25), $p(z_{t+1}|x_t)$ can be expressed as the marginalization of a component in the probability distribution $p(z_{t+1}|x_t, \omega_t)$, where $\omega_t \in \mathbb{R}^{K-1}$ represents a vector of auxiliary variables. Dependent on ν_t , there is

$$p(z_{t+1}, x_t, \omega_t) \propto \prod_{k=1}^{K-1} \exp \left\{ -\frac{1}{2} \omega_{t,k} \nu_{t,k}^2 + \kappa_{t+1,k} \nu_{t,k} \right\}, \quad (26)$$

where $\kappa_{t+1,k} = \mathbb{I}(z_{t+1} = k) - \frac{1}{2} \mathbb{I}(z_{t+1} > k)$. Hence,

$$p(z_{t+1}, x_t, \omega_t) \propto \mathcal{N}(\nu_t | \Omega_t^{-1} \kappa_{t+1}, \Omega_t^{-1}), \quad (27)$$

where $\Omega_t = \text{diag}(\omega_t)$ and $\kappa_{t+1} = (\kappa_{t+1,1}, \dots, \kappa_{t+1,K-1})$. After augmentation, the conditional density on x_t becomes virtually Gaussian, simplifying the evaluation of the integrals needed for message passing during the Bayesian inference. Finally, we summarize the proposed HO-rSLDS as

$$\begin{aligned} z_1 | \theta &\sim \pi_0(\alpha_0), \\ \bar{x}_0 | \theta &\sim \mathcal{N}(\bar{x}_0 | \mu_0, \Sigma_0), \\ z_{t+1} | z_t, x_t, \theta &\sim \pi_{\text{SB}}(\nu_t), \quad \nu_t = R_{z_t} x_t + r_{z_t}, \\ x_{t+1} | \bar{x}_t, z_{t+1}, \theta &\sim \mathcal{N}(x_{t+1} | A_{z_{t+1}} \bar{x}_t, Q_{z_{t+1}}), \\ y_t | x_t, z_t, \theta &\sim \mathcal{N}(y_t | C_{z_t} x_t, S_{z_t}), \\ \theta &= \{\alpha_0, \mu_0, \Sigma_0\} \cup \{A_k, Q_k, C_k, S_k, R_k, r_k\}_{k=1}^K. \end{aligned} \quad (28)$$

To learn the HO-rSLDS utilizing Bayesian inference, we assign conjugate priors to each component of the model parameters θ , given by

$$\begin{aligned} \pi_0 | \alpha_0 &\sim \text{Dir}(\alpha_0), \\ (A_k, Q_k) | \eta &\sim \text{MNIW}(\eta), \\ (C_k, S_k) | \lambda &\sim \text{MNIW}(\lambda), \\ R_k | \zeta &\sim \prod_{i=1}^{K-1} \mathcal{N}(R_{k,i} | \zeta_i), \\ r_k | \rho &\sim \mathcal{N}(r_k | \rho), \end{aligned} \quad (29)$$

where Dir and MNIW denote Dirichlet and matrix normal inverse Wishart distribution, respectively, $R_{k,i}$ denotes the i th row of R_k . A graphical illustration of HO-rSLDS is given in Figure 3.

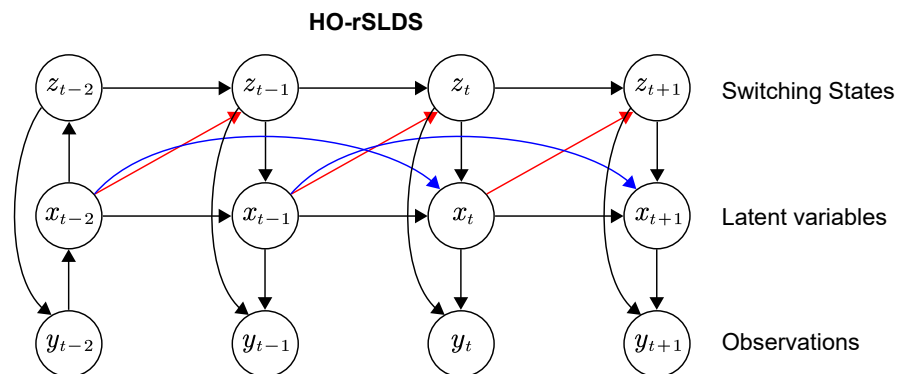


Figure 3. The graphical model of the HO-rSLDS, where there is higher-order dependence involved in x_t (blue) and the recurrent dependence from x_{t-1} to z_t (red).

3. Variational Inference of HO-rSLDS

Now we describe the model learning inference of HO-rSLDS, which is fulfilled by variational inference. In variational inference, a parameter is inferred by maximizing the evidence's lower bound, hence minimizing the Kullback–Leibler divergence as well as the variational free energy. Specifically, we denote $\mathcal{Y} := y_{1:T}$, $\Theta := \{z_{1:T}, (\bar{x}_0, x_{1:T}), \omega_{1:T}, \theta\}$, then the following decomposition holds:

$$\ln p(\mathcal{Y}) = \mathcal{L}(q(\Theta)) + \text{KL}(q(\Theta) || p(\Theta|\mathcal{Y})), \quad (30)$$

where

$$\mathcal{L}(q(\Theta)) = \int q(\Theta) \log \frac{p(\mathcal{Y}, \Theta)}{q(\Theta)} d\Theta, \quad (31)$$

$$\text{KL}(q(\Theta) || p(\Theta|\mathcal{Y})) = - \int q(\Theta) \log \frac{p(\Theta|\mathcal{Y})}{q(\Theta)} d\Theta. \quad (32)$$

The exact evaluation of $p(z_{1:T}, x_{0:T}, \omega_{1:T} | y_{1:T})$ where $x_{0:T} := \bar{x}_0 \cup x_{1:T}$ in HO-rSLDS is intractable. In this study, we leverage an approximate variational inference method to assess the approximate posterior $q(z_{1:T}, x_{0:T}, \omega_{1:T} | y_{1:T})$ in the proposed HO-rSLDS. In order to obtain a tractable inference, a structured mean field approximation is adopted on the generative model, given by

$$q(z_{1:T}, x_{0:T}, \omega_{1:T}, \theta | y_{1:T}) = q(z_{1:T})q(x_{0:T})q(\omega_{1:T})q(\theta). \quad (33)$$

Moreover, $q(\omega_{1:T})$ further factorizes as:

$$q(\omega_{1:T}) = \prod_{t=1}^T \prod_{k=1}^{K-1} q(\omega_{t,k}). \quad (34)$$

Given the factorized forms in Equations (33) and (34), we can obtain the optimal solution of the factorized variational posteriors by maximizing the evidence lower bound in Equation (31), i.e., $q(z_{1:T})$, $q(x_{0:T})$, $q(\omega_{1:T})$ and $q(\theta)$, given by

$$\ln q(z_{1:T}) \propto \mathbb{E}_{-q(z_{1:T})} \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, \theta, y_{1:T}), \quad (35)$$

$$\ln q(x_{0:T}) \propto \mathbb{E}_{-q(x_{0:T})} \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, \theta, y_{1:T}), \quad (36)$$

$$\ln q(\omega_{1:T}) \propto \mathbb{E}_{-q(\omega_{1:T})} \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, \theta, y_{1:T}), \quad (37)$$

$$\ln q(\theta) \propto \mathbb{E}_{-q(\theta)} \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, \theta, y_{1:T}), \quad (38)$$

where $\mathbb{E}_{-q(z_{1:T})}(\cdot) := \mathbb{E}_{q(x_{0:T})q(\omega_{1:T})q(\theta)}(\cdot)$. Hence the joint log-likelihood is

$$\ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, \theta, y_{1:T}) = \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, y_{1:T} | \theta) + \ln p(\theta). \quad (39)$$

The conditional log-likelihood $\ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, y_{1:T} | \theta)$ can be written as

$$\begin{aligned} & \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, y_{1:T} | \theta) = \\ & \ln p(z_1 | \theta) + \ln p(\bar{x}_0 | \theta) + \sum_{t=1}^T [\ln p(z_t | x_{t-1}, z_{t-1}, \omega_{t-1}, \theta) \mathbb{I}(t > 1) + \ln(x_t | \bar{x}_{t-1}, z_t, \theta)] \\ & + \sum_{t=1}^T p(y_t | z_t, x_t, \theta) \end{aligned} \quad (40)$$

$$\begin{aligned} & \propto \ln p(z_1 | \theta) - \frac{1}{2} (\bar{x}_0 - \mu_0)^\top (\Sigma_0)^{-1} (\bar{x}_0 - \mu_0) \\ & - \frac{1}{2} \sum_{t=1}^T \left[(v_{t-1} - \Omega_{t-1}^{-1} \kappa_t)^\top \Omega_{t-1} (v_{t-1} - \Omega_{t-1}^{-1} \kappa_t) \mathbb{I}(t > 1) \right. \\ & \quad \left. + (x_t - A_{z_t} \bar{x}_{t-1})^\top Q_{z_t}^{-1} (x_t - A_{z_t} \bar{x}_{t-1}) \right] \\ & - \frac{1}{2} \sum_{t=1}^T (y_t - C_{z_t} x_t)^\top S_{z_t}^{-1} (y_t - C_{z_t} x_t). \end{aligned} \quad (41)$$

The exact optimal solutions for $q(z_{1:T})$ and $q(x_{0:T})$ will not be explicitly evaluated; rather, we use message passing algorithms to obtain the sufficient statistics with respect to these factors. In the following subsections, we present the involved message passing algorithms and the detailed derivations of update formula for each factor.

3.1. Update for $q(x_{0:T})$

Since $\mathbb{E}_{-q(x_{0:T})} \ln p(\theta)$ is not a function of $x_{0:T}$, Equation (36) can be written as

$$\begin{aligned} \ln q(x_{0:T}) &\propto \mathbb{E}_{-q(x_{0:T})} \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, y_{1:T} | \theta), \\ &\propto \mathbb{E}_{-q(x_{0:T})} \left[-\frac{1}{2} (\bar{x}_0 - \mu_0)^\top (\Sigma_0)^{-1} (\bar{x}_0 - \mu^0) \right] \\ &\quad + \mathbb{E}_{-q(x_{0:T})} \left[-\frac{1}{2} \sum_{t=2}^T (\nu_{t-1} - \Omega_{t-1}^{-1} \kappa_t)^\top \Omega_{t-1} (\nu_{t-1} - \Omega_{t-1}^{-1} \kappa_t) \right] \\ &\quad + \mathbb{E}_{-q(x_{0:T})} \left[-\frac{1}{2} \sum_{t=1}^T (x_t - A_{z_t} \bar{x}_{t-1})^\top Q_{z_t}^{-1} (x_t - A_{z_t} \bar{x}_{t-1}) \right] \\ &\quad + \mathbb{E}_{-q(x_{0:T})} \left[-\frac{1}{2} \sum_{t=1}^T (y_t - C_{z_t} x_t)^\top S_{z_t}^{-1} (y_t - C_{z_t} x_t) \right]. \end{aligned} \quad (42)$$

From Equation (42) it can be observed that the structure of the variational posterior $\ln q(x_{0:T})$ resembles the joint log-likelihood function of a time-varying linear dynamical system with higher-order dependence (HO-LDS). Thus, the $\ln q(x_{0:T})$ can be re-expressed as

$$\begin{aligned} \ln q(x_{0:T}) &\propto -\frac{1}{2} (\bar{x}_0 - \hat{\mu}_0)^\top (\hat{\Sigma}_0)^{-1} (\bar{x}_0 - \hat{\mu}_0) \\ &\quad - \frac{1}{2} \sum_{t=1}^T (x_t - \hat{A}_t \bar{x}_{t-1})^\top \hat{Q}_t^{-1} (x_t - \hat{A}_t \bar{x}_{t-1}) \\ &\quad - \frac{1}{2} \sum_{t=1}^T (y_t - \hat{C}_t x_t)^\top \hat{S}_t^{-1} (y_t - \hat{C}_t x_t), \end{aligned} \quad (43)$$

and the set of variational parameters involved in Equation (43) are defined as

$$\hat{\theta}_{x_{0:T}} := \left\{ \hat{\mu}_0, \hat{\Sigma}_0, \{\hat{A}_t\}_{t=1}^T, \{\hat{Q}_t\}_{t=1}^T, \{\hat{C}_t\}_{t=1}^T, \{\hat{S}_t\}_{t=1}^T \right\}. \quad (44)$$

Each component in the $\hat{\theta}_{x_{0:T}}$ can be precisely determined by matching the coefficients of each term in Equations (42) and (43). Algorithm 1 outlines an efficient scheme for this purpose.

Algorithm 1: Obtaining $\hat{\theta}_{x_{0:T}}$

```

1 for  $t = T : 1$  do
2    $\hat{S}_t^{-1} \leftarrow \mathbb{E}_{-q(x_{0:T})} [S_{z_t}^{-1}]$ ;
3    $\hat{C}_t \leftarrow \hat{S}_t \mathbb{E}_{-q(x_{0:T})} [S_{z_t}^{-1} C_{z_t}]$ ;
4    $\hat{Q}_t^{-1} \leftarrow$ 
      
$$\begin{cases} \mathbb{E}_{-q(x_{0:T})} [Q_{z_t}^{-1} + C_{z_t}^\top S_{z_t}^{-1} C_{z_t}] - \hat{C}_t^\top \hat{S}_t^{-1} \hat{C}_t, & t = T \\ \mathbb{E}_{-q(x_{0:T})} [Q_{z_t}^{-1} + \sum_{j=1}^p A_{z_{t+j}}^\top(j) Q_{z_{t+j}}^{-1} A_{z_{t+j}}(j) + C_{z_t}^\top S_{z_t}^{-1} C_{z_t} + R_{z_t}^\top \Omega_t R_{z_t}] \\ \quad - \sum_{j=1}^p \hat{A}_{t+j}^\top(j) \hat{Q}_{t+j}^{-1} \hat{A}_{t+j}(j) - \hat{C}_t^\top \hat{S}_t^{-1} \hat{C}_t, & t < T \end{cases};$$

5    $\hat{A}_t \leftarrow \hat{Q}_t \mathbb{E}_{-q(x_{0:T})} [Q_{z_t}^{-1} A_{z_t}]$ ;
6 end
7  $\hat{\Sigma}_0^{-1} \leftarrow \mathbb{E}_{-q(x_{0:T})} [\Sigma^0 + A_{z_1}^\top Q_{z_1} A_{z_1}] - \hat{A}_1^\top \hat{Q}_1 \hat{A}_1$ ;
8  $\hat{\mu}_0 \leftarrow \hat{\Sigma}_0 \mathbb{E}_{-q(x_{0:T})} [(\Sigma^0)^{-1} \mu^0]$ ;
Output:  $\hat{\theta}_{x_{0:T}}$ 

```

Remark 1. $A_{z_t} \bar{x}_{t-1} = \sum_{j=1}^p A_{z_t}(j) x_{t-j}$ where $A_{z_t}(j)$ is defined as the submatrix of A_{z_t} associated with x_{t-j} in $A_{z_t} \bar{x}_{t-1}$.

Thus, we perform message passing on a time-varying HO-LDS parameterized by $\hat{\theta}_{x_{0:T}}$ to obtain the posterior distribution $q(x_{0:T})$ and sufficient statistics used in the update for the other factors. To be concrete, the message passing comprises the generalized Kalman filter and Kalman smoother for the time-varying HO-LDS, which are demonstrated by Theorems 1 and 2. To prove the theorems, we need the following lemmas [26]:

Lemma 1. If random variables X and Y follow the Gaussian probability distributions:

$$X \sim \mathcal{N}(m, P), \quad (45)$$

$$Y|X \sim \mathcal{N}(HX + u, R), \quad (46)$$

then the joint distribution of X and Y , as well as the marginal distribution of Y , are provided by

$$\begin{aligned} \begin{pmatrix} X \\ Y \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} m \\ Hm + u \end{pmatrix}, \begin{pmatrix} P & PH^\top \\ HP & HPH^\top + R \end{pmatrix}\right), \\ Y &\sim \mathcal{N}(Hm + u, HPH^\top + R). \end{aligned} \quad (47)$$

Lemma 2. If the random variables X and Y follow the joint Gaussian probability distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} A & C \\ C^\top & B \end{pmatrix}\right), \quad (48)$$

then the marginal and conditional distributions of X and Y are as follows:

$$\begin{aligned} X &\sim \mathcal{N}(a, A), \\ Y &\sim \mathcal{N}(b, B), \\ X|Y &\sim \mathcal{N}(a + CB^{-1}(Y - b), A - CB^{-1}C^\top), \\ Y|X &\sim \mathcal{N}(b + C^\top A^{-1}(X - a), B - C^\top A^{-1}C). \end{aligned} \quad (49)$$

Now we demonstrate Theorems 1 and 2 in turn, which provides the Kalman filter and smoother equations about $q(x_{0:T})$ in the time-varying HO-LDS.

Theorem 1. The Kalman filter equations for the time-varying HO-LDS (Equation (43)) parameterized by $\hat{\theta}_{x_{0:T}}$ can be evaluated in closed form, which results in Gaussian distributions:

$$p(x_t|y_{1:t-1}) = \mathcal{N}(x_t|\mu_t^{(p)}, \Sigma_t^{(p)}), \quad (50)$$

$$p(x_t|y_{1:t}) = \mathcal{N}(x_t|\mu_t^{(u)}, \Sigma_t^{(u)}), \quad (51)$$

$$p(\bar{x}_t|y_{1:t}) = \mathcal{N}(\bar{x}_t|\mu_t^{(m)}, \Sigma_t^{(m)}). \quad (52)$$

The distribution parameters can be calculated using the following Kalman filter prediction, update and merging steps detailed in the following proof. The recursion begins with the prior mean μ_0^m and covariance Σ_0^m (i.e., $\hat{\mu}_0$ and $\hat{\Sigma}_0$).

Proof. The Gaussian filter distributions (Equations (50)–(52)) can be obtained by the following steps:

1. Prediction step. By Lemma 1, the joint distribution of x_t and \bar{x}_{t-1} given $y_{1:t-1}$ is

$$\begin{aligned} p(x_t, \bar{x}_{t-1} | y_{1:t-1}) &= p(x_t | \bar{x}_{t-1}) p(\bar{x}_{t-1} | y_{1:t-1}) \\ &= \mathcal{N}(x_t | \hat{A}_t \bar{x}_{t-1}, \hat{Q}_t) \mathcal{N}(\bar{x}_{t-1} | \mu_{t-1}^{(m)}, \Sigma_{t-1}^{(m)}) \\ &= \mathcal{N}\left(\begin{pmatrix} \bar{x}_{t-1} \\ x_t \end{pmatrix} \middle| \mu^{(1)}, \Sigma^{(1)}\right), \end{aligned} \quad (53)$$

where

$$\mu^{(1)} = \begin{pmatrix} \mu_{t-1}^{(m)} \\ \hat{A}_t \mu_{t-1}^{(m)} \end{pmatrix}, \quad \Sigma^{(1)} = \begin{pmatrix} \Sigma_{t-1}^{(m)} & \Sigma_{t-1}^{(m)} \hat{A}_t^\top \\ \hat{A}_t \Sigma_{t-1}^{(m)} & \hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t \end{pmatrix}, \quad (54)$$

and by Lemma 2 the marginal distribution of x_t is given by

$$p(x_t | y_{1:t-1}) = \mathcal{N}(x_t | \mu_t^{(p)}, \Sigma_t^{(p)}), \quad (55)$$

where

$$\mu_t^{(p)} = \hat{A}_t \mu_{t-1}^{(m)}, \quad \Sigma_t^{(p)} = \hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t. \quad (56)$$

2. Update step. By Lemma 1, the joint distribution of y_t and x_t given $y_{1:t-1}$ is

$$\begin{aligned} p(x_t, y_t | y_{1:t-1}) &= p(y_t | x_t) p(x_t | y_{1:t-1}) \\ &= \mathcal{N}(y_t | \hat{C}_t x_t, \hat{S}_t) \mathcal{N}(x_t | \mu_t^{(p)}, \Sigma_t^{(p)}) \\ &= \mathcal{N}\left(\begin{pmatrix} x_t \\ y_t \end{pmatrix} \middle| \mu^{(2)}, \Sigma^{(2)}\right), \end{aligned} \quad (57)$$

where

$$\mu^{(2)} = \begin{pmatrix} \mu_t^{(p)} \\ \hat{C}_t \mu_t^{(p)} \end{pmatrix}, \quad \Sigma^{(2)} = \begin{pmatrix} \Sigma_t^{(p)} & \Sigma_t^{(p)} \hat{C}_t^\top \\ \hat{C}_t \Sigma_t^{(p)} & \hat{C}_t \Sigma_t^{(p)} \hat{C}_t^\top + \hat{S}_t \end{pmatrix}. \quad (58)$$

By Lemma 2 the conditional distribution of x_t given $y_{1:t}$ is

$$\begin{aligned} p(x_t | y_t, y_{1:t-1}) &= p(x_t | y_{1:t}) \\ &= \mathcal{N}(x_t | \mu^{(u)}, \Sigma^{(u)}), \end{aligned} \quad (59)$$

where

$$\mu^{(u)} = \mu_t^{(p)} + \Sigma_t^{(p)} \hat{C}_t^\top (\hat{C}_t \Sigma_t^{(p)} \hat{C}_t^\top + \hat{S}_t)^{-1} (y_t - \hat{C}_t \mu_t^{(p)}), \quad (60)$$

$$\Sigma^{(u)} = \Sigma_t^{(p)} - \Sigma_t^{(p)} \hat{C}_t^\top (\hat{C}_t \Sigma_t^{(p)} \hat{C}_t^\top + \hat{S}_t)^{-1} \hat{C}_t \Sigma_t^{(p)}. \quad (61)$$

3. Merging step. By Lemma 1, the joint distribution of \bar{x}_{t-1}, x_t, y_t given $y_{1:t-1}$ is

$$\begin{aligned} p(\bar{x}_{t-1}, x_t, y_t | y_{1:t-1}) &= p(y_t | x_t) p(x_t, \bar{x}_{t-1} | y_{1:t-1}) \\ &= \mathcal{N}(\hat{C}_t x_t, \hat{S}_t) \mathcal{N}\left(\begin{pmatrix} \bar{x}_{t-1} \\ x_t \end{pmatrix} \middle| \mu^{(1)}, \Sigma^{(1)}\right) \\ &= \mathcal{N}\left(\begin{pmatrix} \bar{x}_{t-1} \\ x_t \\ y_t \end{pmatrix} \middle| \mu^{(3)}, \Sigma^{(3)}\right), \end{aligned} \quad (62)$$

where

$$\mu^{(3)} = \begin{pmatrix} \mu_1^{(3)} \\ \mu_2^{(3)} \\ \mu_3^{(3)} \end{pmatrix} = \begin{pmatrix} \mu_{t-1}^{(m)} \\ \hat{A}_t \mu_{t-1}^{(m)} \\ \hat{C}_t \hat{A}_t \mu_{t-1}^{(m)} \end{pmatrix}, \quad (63)$$

$$\begin{aligned} \Sigma^{(3)} &= \begin{pmatrix} \Sigma_{11}^{(3)} & \Sigma_{12}^{(3)} & \Sigma_{13}^{(3)} \\ \Sigma_{21}^{(3)} & \Sigma_{22}^{(3)} & \Sigma_{23}^{(3)} \\ \Sigma_{31}^{(3)} & \Sigma_{32}^{(3)} & \Sigma_{33}^{(3)} \end{pmatrix}, \\ &= \begin{pmatrix} \Sigma_{t-1}^{(m)} & \Sigma_{t-1}^{(m)} \hat{A}_t^\top & \Sigma_{t-1}^{(m)} \hat{A}_t^\top \hat{C}_t^\top \\ \hat{A}_t \Sigma_{t-1}^{(m)} & \hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t & (\hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t) \hat{C}_t^\top \\ \hat{C}_t \hat{A}_t \Sigma_{t-1}^{(m)} & \hat{C}_t (\hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t) & \hat{C}_t (\hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t) \hat{C}_t^\top + \hat{S}_t \end{pmatrix}. \end{aligned} \quad (64)$$

By Lemma 2, the conditional distribution of \bar{x}_{t-1} and x_t given $y_{1:t}$ is

$$\begin{aligned} p(\bar{x}_{t-1}, x_t | y_{1:t}) &= p(\bar{x}_{t-1}, x_t | y_{1:t-1}) \\ &= \mathcal{N}(\bar{x}_{t-1}, x_t | \mu^{(4)}, \Sigma^{(4)}), \end{aligned} \quad (65)$$

where

$$\mu^{(4)} = \begin{pmatrix} \mu_1^{(3)} \\ \mu_2^{(3)} \end{pmatrix} + \begin{pmatrix} \Sigma_{13}^{(3)} \\ \Sigma_{23}^{(3)} \end{pmatrix} (\Sigma_{33}^{(3)})^{-1} (y_t - \mu_3^{(3)}), \quad (66)$$

$$\Sigma^{(4)} = \Sigma_{11}^{(3)} - \begin{pmatrix} \Sigma_{13}^{(3)} \\ \Sigma_{23}^{(3)} \end{pmatrix} (\Sigma_{33}^{(3)})^{-1} \begin{pmatrix} \Sigma_{13}^{(3)} & \Sigma_{23}^{(3)} \end{pmatrix}. \quad (67)$$

Finally, the marginal distribution of \bar{x}_t is

$$\begin{aligned} p(\bar{x}_t | y_{1:t}) &= \mathcal{N}(\bar{x}_t | \mu_t^{(m)}, \Sigma_t^{(m)}) \\ &= \mathcal{N}(\bar{x}_t | \mu^{(4)}(\bar{x}_t), \Sigma^{(4)}(\bar{x}_t)), \end{aligned} \quad (68)$$

where $\mu^{(4)}(\bar{x}_t)$ and $\Sigma^{(4)}(\bar{x}_t)$ are marginal mean vector and covariance matrix of \bar{x}_t . \square

The Kalman filter for the time-varying HO-LDS in Theorem 1 calculates the estimates based on measurements collected up to and including time step t . After obtaining the filtering posterior state distributions, Theorem 2 outlines the results for calculating the marginal posterior distributions for each time step based on all data up to the final time step T , which is described as the following:

Theorem 2. The Kalman smoother equations for the time-varying HO-LDS (Equation (43)) parameterized by $\hat{\theta}_{x_{0:T}}$ can be evaluated in closed form, which results Gaussian distributions:

$$p(x_t | y_{1:T}) = \mathcal{N}(x_t | \mu_t^{(s_1)}, \Sigma_t^{(s_1)}), \quad (69)$$

$$p(\bar{x}_t | y_{1:T}) = \mathcal{N}(\bar{x}_t | \mu_t^{(s_2)}, \Sigma_t^{(s_2)}), \quad (70)$$

$$p(\bar{x}_{t-1}, x_t | y_{1:T}) = \mathcal{N}((\bar{x}_{t-1}, x_t) | \mu_t^{(s_3)}, \Sigma_t^{(s_3)}). \quad (71)$$

The smoothing distribution and filtering distribution of the last time step are identical, allowing for the recursive computation of the smoothing distribution for all time steps by beginning from the last step and moving backward.

Proof. Firstly by Lemma 1, the joint distribution of \bar{x}_{t-1} and x_t given $y_{1:t-1}$ is

$$\begin{aligned} p(x_t, \bar{x}_{t-1} | y_{1:t-1}) &= p(x_t | \bar{x}_{t-1}) p(\bar{x}_{t-1} | y_{1:t-1}) \\ &= \mathcal{N}(x_t | \hat{A}_t \bar{x}_{t-1}, \hat{Q}_t) \mathcal{N}(\bar{x}_{t-1} | \mu_{t-1}^{(m)}, \Sigma_{t-1}^{(m)}) \\ &= \mathcal{N}\left(\begin{pmatrix} \bar{x}_{t-1} \\ x_t \end{pmatrix} \middle| \mu^{(1)}, \Sigma^{(1)}\right), \end{aligned} \quad (72)$$

where

$$\mu^{(1)} = \begin{pmatrix} \mu_{t-1}^{(m)} \\ \hat{A}_t \mu_{t-1}^{(m)} \end{pmatrix}, \quad \Sigma^{(1)} = \begin{pmatrix} \Sigma_{t-1}^{(m)} & \Sigma_{t-1}^{(m)} \hat{A}_t^\top \\ \hat{A}_t \Sigma_{t-1}^{(m)} & \hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t \end{pmatrix}. \quad (73)$$

According to the Markov property of the states, we obtain

$$p(\bar{x}_{t-1} | x_t, y_{1:T}) = p(\bar{x}_{t-1} | x_t, y_{t-1}). \quad (74)$$

Therefore, we have the conditional distribution as

$$\begin{aligned} p(\bar{x}_{t-1} | x_t, y_{1:T}) &= p(\bar{x}_{t-1} | x_t, y_{t-1}), \\ &= \mathcal{N}(\bar{x}_{t-1} | \tilde{\mu}^{(2)}, \tilde{\Sigma}^{(2)}), \end{aligned} \quad (75)$$

where

$$\begin{aligned} G_t &= \Sigma_{t-1}^{(m)} \hat{A}_t^\top (\hat{A}_t \Sigma_{t-1}^{(m)} \hat{A}_t^\top + \hat{Q}_t)^{-1}, \\ \tilde{\mu}^{(2)} &= \mu_{t-1}^{(m)} + G_t (x_t - \hat{A}_t \mu_{t-1}^{(m)}), \\ \tilde{\Sigma}^{(2)} &= \Sigma_{t-1}^{(m)} - G_t \hat{A}_t \Sigma_{t-1}^{(m)}. \end{aligned} \quad (76)$$

Then, the joint distribution of \bar{x}_{t-1} and x_t given all the data is

$$\begin{aligned} p(\bar{x}_{t-1}, x_t | y_{1:T}) &= p(\bar{x}_{t-1} | x_t, y_{1:T}) p(x_t | y_{1:T}) \\ &= \mathcal{N}(\bar{x}_{t-1} | \tilde{\mu}^{(2)}, \tilde{\Sigma}^{(2)}) \mathcal{N}(x_t | \mu_t^{(s_1)}, \Sigma_t^{(s_1)}) \\ &= \mathcal{N}\left(\begin{pmatrix} \bar{x}_{t-1} \\ x_t \end{pmatrix} \middle| \mu_t^{(s_3)}, \Sigma_t^{(s_3)}\right), \end{aligned} \quad (77)$$

where

$$\mu_t^{(s_3)} = \begin{pmatrix} G_t \mu_t^{(s_1)} + \mu_{t-1}^{(m)} - G_t \hat{A}_t \mu_{t-1}^{(m)} \\ \mu_t^{(s_1)} \end{pmatrix}, \quad (78)$$

$$\Sigma_t^{s_3} = \begin{pmatrix} G_t \Sigma_t^{(s_1)} G_t^\top + \tilde{\Sigma}^{(2)} & \Sigma_t^{(s_1)} G_t^\top \\ G_t \Sigma_t^{(s_1)} & \Sigma_t^{(s_1)} \end{pmatrix}. \quad (79)$$

Thus, the marginal distribution of \bar{x}_{t-1} is given as

$$p(\bar{x}_{t-1} | y_{1:T}) = \mathcal{N}(\bar{x}_{t-1} | \mu_t^{(s_2)}, \Sigma_t^{(s_2)}), \quad (80)$$

where

$$\mu_t^{(s_2)} = G_t \mu_t^{(s_1)} + \mu_{t-1}^{(m)} - G_t \hat{A}_t \mu_{t-1}^{(m)}, \quad (81)$$

$$\Sigma_t^{(s_2)} = G_t \Sigma_t^{(s_1)} G_t^\top + \tilde{\Sigma}^{(2)}. \quad (82)$$

The $p(x_{t-1}|y_{1:T})$ will be directly obtained from $p(\bar{x}_{t-1}|y_{1:T})$. Therefore, we have derived the concrete expressions of the Gaussian smoother distributions (Equations (69)–(71)). \square

Thus, we can obtain the posterior distributions of $q(x_{0:T}|y_{1:T})$ as well as the expectations involved in the update for the other factors, which are $\mathbb{E}_{q(x_{0:T})}(x_t)$, $\mathbb{E}_{q(x_{0:T})}(x_t x_t^\top)$ and $\mathbb{E}_{q(x_{0:T})}(\bar{x}_t x_{t+1}^\top)$.

3.2. Update for $q(\omega_{1:T})$

$q(\omega_{1:T})$ can be updated using a closed-form expression by the Pólya-gamma augmentation. Equations (37) and (41) results in the following update for $q(\omega_{1:T}) = \prod_{t=1}^T \prod_{k=1}^{K-1} q(\omega_{t,k})$:

$$\begin{aligned} \ln q(\omega_{t,k}) &\propto \mathbb{E}_{-q(\omega_{1:T})} \ln p(z_{t+1}, \omega_t | x_t) \\ &\propto \mathbb{E}_{-q(\omega_{1:T})} \ln p(z_{t+1} | \omega_t, x_t) + \mathbb{E}_{-q(\omega_{1:T})} \ln \text{PG}(\omega_{t,k} | \mathbb{I}(z_{t+1} > k), 0). \end{aligned} \quad (83)$$

Since $\mathbb{E}_{-q(\omega_{1:T})} \ln p(z_{t+1} | \omega_t, x_t) \propto -\frac{1}{2} \mathbb{E}_{-q(\omega_{1:T})}(\nu_{t,k}^2) \omega_{t,k}$, we have

$$q(\omega_{t,k}) = \text{PG}(\omega_{t,k} | \mathbb{E}_{q(z_{1:T})} \mathbb{I}(z_{t+1} > k), \mathbb{E}_{-q(\omega_{1:T})}(\nu_{t,k}^2)). \quad (84)$$

Moreover, the expectation of the posterior Pólya-gamma distribution is available in closed form [22]:

$$\mathbb{E}(\omega_{t,k}) = \frac{\mathbb{E}_{q(z_{1:T})} \mathbb{I}(z_{t+1} > k)}{2 \mathbb{E}_{q(z_{1:T})} q(x_{1:T}) q(\theta) (\nu_{t,k}^2)} \tanh \left[\frac{1}{2} \mathbb{E}_{q(z_{1:T})} q(x_{1:T}) q(\theta) (\nu_{t,k}^2) \right]. \quad (85)$$

3.3. Update for $q(z_{1:T})$

Since $\mathbb{E}_{-q(z_{1:T})} \ln p(\theta)$ is not a function of $z_{1:T}$, Equation (35) can be written as

$$\begin{aligned} \ln q(z_{1:T}) &\propto \mathbb{E}_{-q(z_{1:T})} \ln p(z_{1:T}, x_{0:T}, \omega_{1:T}, y_{1:T} | \theta) \\ &\propto \mathbb{E}_{-q(z_{1:T})} \left[\ln p(z_1 | \theta) + \sum_{t=2}^T \ln p(z_t | x_{t-1}, z_{t-1}, \omega_{t-1}, \theta) + \sum_{t=1}^T \ln p(x_t | z_t, \theta) \right]. \end{aligned} \quad (86)$$

From Equation (86) it is evident that the resulting factor $q(z_{1:T})$ has the form of an HMM parameterized by $q(x_{1:T}, \omega_{1:T}, \theta)$, and the expected sufficient statistics required for updating the other factors can be obtained by message passing algorithms. We define $\ln \tilde{\pi}_{z_1} := \mathbb{E}_{-q(z_{1:T})} \ln p(z_1 | \theta)$, $\ln \tilde{a}_{z_t z_{t+1}} := \mathbb{E}_{-q(z_{1:T})} \ln p(z_{t+1} | z_t, x_t, \omega_t, \theta)$, $\ln \tilde{b}_{z_t}(x_t) := \mathbb{E}_{-q(z_{1:T})} \ln p(x_t | z_t, \theta)$ and the set of the HMM parameters is denoted as $\hat{\theta}_{z_{1:T}}$. Thus, we have

$$\ln q(z_{1:T}) \propto \ln \tilde{\pi}_{z_1} + \sum_{t=2}^T \ln \tilde{a}_{z_{t-1} z_t}(t-1) + \sum_{t=1}^T \ln \tilde{b}_{z_t}(x_t). \quad (87)$$

Note that the densities $p(z_1 | \theta)$, $p(z_{t+1} | z_t, x_t, \omega_t, \theta)$, $p(x_t | z_t, \theta)$ belong to the exponential family, these expectations can be evaluated in closed form. The computation of the HMM parameters by Equation (86) is summarized in Algorithm 2.

We apply a standard forward-backward algorithm for an HMM where the forward recursion is:

$$\alpha_t(i) := p(z_t = i | x_{1:t}) \propto \tilde{b}_i(x_t) \sum_{j=1}^K \alpha_{t-1}(j), \quad (88)$$

where $\alpha_1(i) \propto \tilde{\pi}_i \tilde{b}_i(x_1)$ and the backward recursion is:

$$\beta_t(i) := p(x_{t+1:T} | z_t = i) \propto \sum_{j=1}^K \beta_t(j) \tilde{b}_j(x_t) \tilde{a}_{ij}(t). \quad (89)$$

Algorithm 2: Obtaining $\hat{\theta}_{z_{1:T}}$

```

1 for  $k = 1 : K$  do
2    $\tilde{\pi}_k \propto \exp \left\{ \mathbb{E}_{q(z_1)} (\ln p(z_1 = k | \theta)) \right\}$ 
3 end
4 for  $i, j = 1 : K$  do
5   for  $t = 1 : T$  do
6      $\tilde{a}_{ij}(t) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^K \mathbb{E}_{q(\omega_t)} \ln \omega_{t,i} - \frac{1}{2} \mathbb{E}_{q(\omega_t, x_t, \theta)} (x_t^\top R_i^\top \Omega_t R_i x_t) - \right.$ 
7        $\mathbb{E}_{q(\omega_t, x_t, \theta)} (x_t^\top R_i^\top \omega_t r_i - x_t^\top R_i^\top \kappa(z_{t+1} = j)) -$ 
8        $\left. \frac{1}{2} \mathbb{E}_{q(\omega_t, \theta)} \left[ (r_i - \Omega_t^{-1} \kappa(z_{t+1} = j))^\top \Omega_t (r_i - \Omega_t^{-1} \kappa(z_{t+1} = j)) \right] \right\}$ 
9   end
10 end
11 for  $k = 1 : K$  do
12   for  $t = 1 : T$  do
13      $\tilde{b}_k(x_t) \propto \exp \left\{ \mathbb{E}_{q(x_{1:T}, \theta)} \left[ -\frac{1}{2} \ln |Q_k| - \frac{1}{2} (x_t - A_k \bar{x}_{t-1})^\top Q_k (x_t - A_k \bar{x}_{t-1}) \right] \right\}$ 
14   end
15 end

```

Then we can obtain the sufficient statistics used in the update for the other factors using the following equations:

$$\gamma_t(i) := p(z_t = i | x_{1:T}) \propto \alpha_t(i) \beta_t(i), \quad (90)$$

$$\zeta_t(i, j) := p(z_t = i, z_{t+1} = j | x_{1:T}) \propto \alpha_t(i) \tilde{a}_{ij}(t) \tilde{b}_j(x_{t+1}) \beta_{t+1}(j). \quad (91)$$

3.4. Update for $q(\theta)$

Given the sufficient statistics of $q(z_{1:T})$ and $q(x_{1:T})$ obtained by message passing, the conjugate posterior distribution of the model parameters can be updated in closed form by evaluating Equation (38). For each parameter, we present the optimized variational distribution in the following:

- Update $q(\pi_0)$. Simplifying Equation (38) and extracting the terms related to π_0 , we have

$$\ln q(\pi_0) \propto \mathbb{E}_{q(-\theta)} \ln p(z_1 | \theta) + \ln p(\pi_0), \quad (92)$$

which results in

$$q(\pi_0) \sim \text{Dir}(\pi_0 | \alpha'_0), \quad (93)$$

where $\alpha'_0 = [\alpha'_{0,1}, \dots, \alpha'_{0,k}, \dots, \alpha'_{0,K}] \in \mathbb{R}^K$ and

$$\alpha'_{0,k} = \alpha_{0,k} + \gamma_1(k), \quad (94)$$

$\alpha_{0,k}$ are prior parameters of $p(\pi_0)$.

- Update (μ_0, Σ_0) . The posterior $q(\mu_0, \Sigma_0)$ can be obtained by the expectations of $q(x_{0:T})$, given by

$$\mu_0 \leftarrow \mathbb{E}_{q(x_{0:T})}(\bar{x}_0), \quad (95)$$

$$\Sigma_0 \leftarrow \mathbb{E}_{q(x_{0:T})}(\bar{x}_0 \bar{x}_0^\top) - \mu_0(\mu_0)^\top. \quad (96)$$

- Update $q(A_k, Q_k)$. Simplifying Equation (38) and extracting the terms related to (A_k, Q_k) , we have

$$\begin{aligned} \ln q(A_k, Q_k) \propto \\ \mathbb{E}_{-q(\theta)} \left[-\frac{\sum_{t=1}^T \gamma_t(k)}{2} \ln |Q_k| - \frac{1}{2} \sum_{t=1}^T \gamma_t(k) (x_t - A_k \bar{x}_{t-1})^\top Q_k^{-1} (x_t - A_k \bar{x}_{t-1}) \right] \\ + \ln p(A_k, Q_k), \end{aligned} \quad (97)$$

which results in

$$q(A_k, Q_k) \sim \text{MNIW}(A_k, Q_k | \eta'_k), \quad (98)$$

where $\eta'_k := \{M_k^{(\eta')}, \Phi_k^{(\eta')}, \iota_k^{(\eta')}, \Delta_k^{(\eta')}\}$,

$$\begin{aligned} q(A_k, Q_k) \propto |Q_k|^{-\frac{\iota_k^{(\eta')} + (p+1)m+1}{2}} \exp \left\{ -\frac{1}{2} \text{Tr}(\Delta_k^{(\eta')} Q_k^{-1}) \right. \\ \left. - \frac{1}{2} \text{Tr} \left[(\Phi_k^{(\eta')})^{-1} (A_k - M_k^{(\eta')})^\top Q_k^{-1} (A_k - M_k^{(\eta')}) \right] \right\}. \end{aligned} \quad (99)$$

The variational posterior parameters can be optimized as

$$\begin{aligned} \Phi_k^{(\eta')} &= \left((\Phi_k^{(\eta)})^{-1} + \sum_{t=1}^T \mathbb{E}_{q(x_{0:T})}(\gamma_t(k) \bar{x}_{t-1} \bar{x}_{t-1}^\top) \right)^{-1}, \\ M_k^{(\eta')} &= \left(M_k^{(\eta)} (\Phi_k^{(\eta)})^{-1} + \sum_{t=1}^T \mathbb{E}_{q(x_{0:T})}(\gamma_t(k) x_t \bar{x}_{t-1}^\top) \right) \Phi_k^{(\eta')}, \\ \Delta_k^{(\eta')} &= \Delta_k^{(\eta)} + \sum_{t=1}^T \mathbb{E}_{q(x_{0:T})}(\gamma_t(k) x_t x_t^\top) + M_k^{(\eta)} (\Phi_k^{(\eta)})^{-1} (M_k^{(\eta)})^\top - M_k^{(\eta')} (\Phi_k^{(\eta')})^{-1} (M_k^{(\eta')})^\top, \\ \iota_k^{(\eta')} &= \iota_k^{(\eta)} + \sum_{t=1}^T \gamma_t(k), \end{aligned} \quad (100)$$

where $M_k^{(\eta)}, \Phi_k^{(\eta)}, \iota_k^{(\eta)}, \Delta_k^{(\eta)}$ are prior parameters of $p(A_k, Q_k)$.

- Update $q(C_k, S_k)$. Simplifying Equation (38) and extracting the terms related to (C_k, S_k) , we have

$$\begin{aligned} \ln q(C_k, S_k) \propto \\ \mathbb{E}_{-q(\theta)} \left[-\frac{\sum_{t=1}^T \gamma_t(k)}{2} \ln |S_k| - \frac{1}{2} \sum_{t=1}^T (y_t - C_k x_t)^\top S_k^{-1} (y_t - C_k x_t) \right] \\ + \ln p(C_k, S_k), \end{aligned} \quad (101)$$

which results in

$$q(C_k, S_k) \sim \text{MNIW}(C_k, S_k | \lambda'_k), \quad (102)$$

where $\lambda'_k := \{M_k^{(\lambda')}, \Phi_k^{(\lambda')}, l_k^{(\lambda')}, \Delta_k^{(\lambda')}\}$,

$$q(C_k, S_k) \propto |S_k|^{-\frac{l_k^{(\lambda')} + (p+1)m+1}{2}} \exp \left\{ -\frac{1}{2} \text{Tr}(\Delta_k^{(\lambda')} S_k^{-1}) - \frac{1}{2} \text{Tr} \left[(\Phi_k^{(\lambda')})^{-1} (C_k - M_k^{(\lambda')})^\top Q_k^{-1} (C_k - M_k^{(\lambda')}) \right] \right\}. \quad (103)$$

The variational posterior parameters can be optimized as

$$\begin{aligned} \Phi_k^{(\lambda')} &= \left((\Phi_k^{(\lambda)})^{-1} + \sum_{t=1}^T \mathbb{E}_{q(x_{0:T})} (\gamma_t(k) x_t x_t^\top) \right)^{-1}, \\ M_k^{(\lambda')} &= \left(M_k^{(\lambda)} (\Phi_k^{(\lambda)})^{-1} + \sum_{t=1}^T \mathbb{E}_{q(x_{0:T})} (\gamma_t(k) y_t x_t^\top) \right) \Phi_k^{(\lambda')}, \\ \Delta_k^{(\lambda')} &= \Delta_k^{(\lambda)} + \sum_{t=1}^T \gamma_t(k) y_t y_t^\top + M_k^{(\lambda)} (\Phi_k^{(\lambda)})^{-1} (M_k^{(\lambda)})^\top - M_k^{(\lambda')} (\Phi_k^{(\lambda')})^{-1} (M_k^{(\lambda')})^\top, \\ l_k^{(\lambda')} &= l_k^{(\lambda)} + \sum_{t=1}^T \gamma_t(k), \end{aligned} \quad (104)$$

where $M_k^{(\lambda)}, \Phi_k^{(\lambda)}, l_k^{(\lambda)}, \Delta_k^{(\lambda)}$ are prior parameters of $p(C_k, S_k)$.

- Update $q(R_k)$. Simplifying Equation (38) and extracting the terms related to R_k , we have

$$\begin{aligned} \ln q(R_k) &\propto \\ &\mathbb{E}_{-q(\theta)} \left[-\frac{1}{2} \sum_{t=1}^{T-1} \gamma_t(k) (R_k x_t + r_k - \Omega_t^{-1} \kappa_{t+1})^\top \Omega_t (R_k x_t + r_k - \Omega_t^{-1} \kappa_{t+1}) \right] \\ &+ \sum_{i=1}^{K-1} \ln p(R_{k,i}), \end{aligned} \quad (105)$$

which results in

$$q(R_k) \sim \prod_{i=1}^{K-1} \mathcal{N}(R_{k,i} | \zeta'_{k,i}), \quad (106)$$

where $\zeta'_{k,i} := \{\mu_{k,i}^{(\zeta')}, \Sigma_{k,i}^{(\zeta')}\}$, which can be optimized as

$$\Sigma_{k,i}^{(\zeta')} = \left[\left(\Sigma_{k,i}^{(\zeta)} \right)^{-1} + \sum_{t=1}^T \mathbb{E}_{q(\omega_{1:T}, x_{0:T})} \left(x_t \gamma_t(k) \Omega_{t,i} x_t^\top \right) \right]^{-1}, \quad (107)$$

$$\mu_{k,i}^{(\zeta')} = \Sigma_{k,i}^{(\zeta')} \left[\sum_{t=1}^{T-1} \mathbb{E}_{q(\omega_{1:T}) q(x_{0:T})} \left(\gamma_t(k) \Omega_{t,i} H_{t,k,i} x_t \right) + \left(\Sigma_{k,i}^{(\zeta)} \right)^{-1} \mu_{k,i}^{(\zeta)} \right], \quad (108)$$

$$H_{t,k} = [H_{t,k,1}, \dots, H_{t,k,i}, \dots, H_{t,k,K-1}]^\top = \mathbb{E}_{q(r_k) q(\omega_{1:T})} \left(r_k - \Omega_t^{-1} \kappa_{t+1} \right), \quad (109)$$

and $\mu_{k,i}^{(\zeta)}, \Sigma_{k,i}^{(\zeta)}$ are prior parameters of $p(R_{k,i})$. To derive Equation (107), just note that $R_k^\top \Omega_t R_k = \sum_{i=1}^{K-1} R_{k,i} \Omega_{k,i} R_{k,i}^\top$ and use the matrix trace.

- Update $q(r_k)$. Simplifying Equation (38) and extracting the terms related to r_k , we have

$$\begin{aligned} \ln q(r_k) &\propto \\ &\mathbb{E}_{-q(\theta)} \left[-\frac{1}{2} \sum_{t=1}^{T-1} \gamma_t(k) (R_k x_t + r_k - \Omega_t^{-1} \kappa_{t+1})^\top \Omega_t (R_k x_t + r_k - \Omega_t^{-1} \kappa_{t+1}) \right] \\ &+ \ln p(r_k), \end{aligned} \quad (110)$$

which results in

$$q(r_k) \sim \mathcal{N}(r_k | \rho'_k), \quad (111)$$

where $\rho'_k := \{\mu_k^{(\rho')}, \Sigma_k^{(\rho')}\}$, which can be optimized as

$$\Sigma_k^{(\rho')} = \left[\left(\Sigma_k^{(\rho)} \right)^{-1} + \sum_{t=1}^{T-1} \mathbb{E}_{q(\omega_{1:T})} (\gamma_t(k) \Omega_t) \right]^{-1}, \quad (112)$$

$$\mu_k^{(\rho')} = \Sigma_k^{(\rho')} \left[\left(\Sigma_k^{(\rho)} \right)^{-1} \mu_k^{(\rho)} + \sum_{t=1}^{T-1} \mathbb{E}_{q(R_k)q(x_{0:T})q(\omega_{1:T})} \left(\gamma_t(k) \Omega_t (R_k x_t - \Omega_t^{-1} \kappa_{t+1}) \right) \right], \quad (113)$$

where $\mu_k^{(\rho)}, \Sigma_k^{(\rho)}$ are prior parameters of $p(r_k)$.

3.5. Initialization

To obtain a reliable variational inference, we initialized the model parameters and latent states with reasonable values by using the following initialization procedure:

1. Probabilistic Principal Component Analysis (PPCA) [27] is conducted on the data, to initialize the continuous latent variables, $x_{0:T}$ and the parameters C ;
2. Fit an AR(p)-HMM to initialize the switching states, $z_{1:T}$ and the parameters, $\{A_k, Q_k\}$. The autoregressive order, p , is determined by AIC, BIC, and HQ criterion;
3. To alleviate the possible and undesirable dependence on ordering that arises from the stick-breaking formulation during the inference, we adopt a strategy of greedily fitting a decision list [21] to identify the most suitable permutation of the switching states for the stick-breaking process. Specifically, we start by performing a greedy search on permutations by creating a decision list based on $(x_t, z_t), z_{t+1}$ pairs, given by

$$z_{t+1} = \begin{cases} o_1 & \text{if } p_1 \\ o_2 & \text{if } \neg p_1 \cap p_2 \\ \vdots & \\ o_K & \text{if } \bigcap_{i=1}^{K-1} (\neg p_i) \end{cases}, \quad (114)$$

where (o_1, \dots, o_K) is a permutation of $(1, \dots, K)$, and p_1, \dots, p_K are predicates that rely on (x_t, z_t) and provide a true or false. In our framework, these predicates are defined by logistic functions:

$$p_j = \sigma(R_{0,j}^\top x_t) > \varepsilon, \quad (115)$$

where $\varepsilon \in (0, 1)$ should be predetermined. To determine o_1 and r_1 , we used the maximum a posteriori estimate of the model for each of the K potential states. For the k th logistic regression, the inputs are $x_{0:T}$ and the outputs are $y_t = \mathbb{I}(z_{t+1} = k)$. We chose the logistic regression model with the largest likelihood as the first output. Then we excluded time points where $z_{t+1} = o_1$ from the data and proceeded to $K - 1$ logistic regressions to predict the subsequent output, o_2 , and so on. After cycling through all K results, we obtained the permutation of $z_{1:T}$. In addition, the $\{R_{0,j}\}_{j=1}^{K-1}$ served as an initialization of the recurrence weights (R).

4. Numerical Experiments

To investigate the performance of the proposed HO-rSLDS, we generated synthetic datasets through the following steps:

1. Benchmark model settings. We take AR-HSMM as a benchmark model to generate synthetic time series. Specifically, the dimensionality of the synthetic time series is $n = 5$, and the total length is $T = 1000$. The state number is $K = 3$, and the autoregressive order p takes values $\{1, 2, 3\}$.
2. Generating (A_k, Q_k) . (A_k, Q_k) corresponds to the emission parameters involved in the VAR process of latent variables x_t (Equation (11)). The autoregressive coefficient matrices A_k are generated with 50% sparsity (defined as the proportion of the zero elements), i.e., 50% of the elements in A_k are 0. The non-zero elements of A_k are generated to be positive or negative with equal probability, and their absolute value is sampled uniformly between 0.2 and 0.5. The covariance matrices Q_k are generated as $Q_k = P_k^\top \text{diag}(\sigma_1, \dots, \sigma_d) P_k$, with P_k being an orthogonal matrix and σ_i assumed positive. The matrix P_k is constructed by orthogonalizing a random matrix whose entries are simulated from a standard Normal, while each σ_i is uniformly sampled in the interval $[1, 3]$. Additionally, a rejection step is done to check that the sampled (A_k, Q_k) constituted a stable VAR process.
3. Generating switching state sequence $z_{1:T}$. To generate $z_{1:T}$, we set the state transition probabilities π_{jk} to 0.5, for $j \neq k = 1, 2, 3$. Note that in an HSMM, the self-transition probability is 0. Furthermore, we simulate the state duration time in two cases. In the first case, the state duration time is sampled from a geometric distribution (the between-state transition probabilities are set to 0.1). In the second case, the state duration time is directly sampled from $\{1, \dots, 10\}$ with specified sampling probabilities. Since the second case does not correspond to a parametric distribution, we denote these two cases as geometric and nonparametric, respectively. In addition, the initial state is uniformly sampled from $\{1, \dots, 10\}$.
4. Generating (C_k, S_k) . (C_k, S_k) corresponds to the emission parameters generating the Gaussian observations y_t (Equation (12)). We generate (C_k, S_k) in the same manner as (A_k, Q_k) .
5. Repeat the above procedure 100 times to obtain 100 synthetic data by setting different random seeds when generating (A_k, Q_k, C_k, S_k) .

Figure 4 displays all dimensions and the true state sequence of one synthetic dataset.

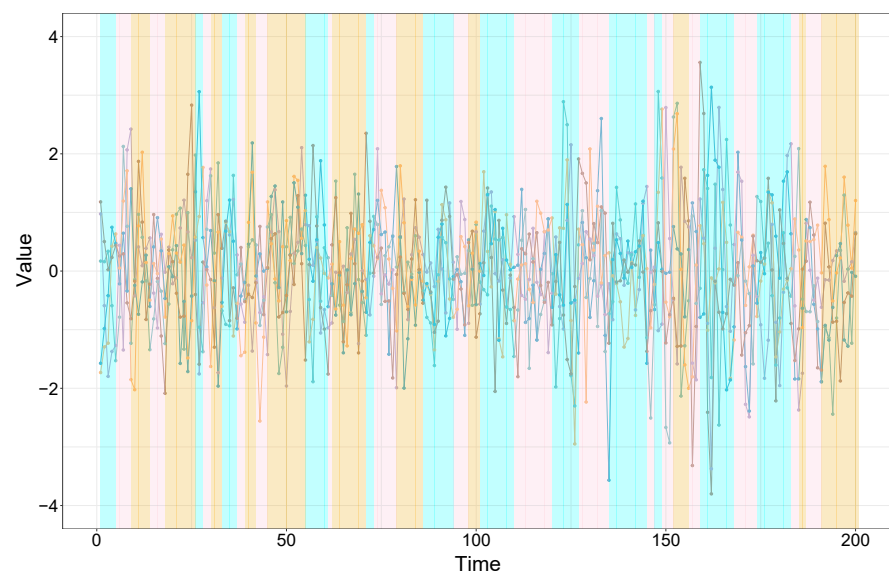


Figure 4. Simulated time series with dimension $d = 5$ and total time points $T = 1000$ (the last 200 time points are exhibited). The five dimensions are displayed in different colors. The color intervals indicate the state sequence.

We investigated the performance of HO-rSLDS by comparing with the competing models, while the estimation accuracy is evaluated by the normalized mean squared error (NMSE) between the true and estimated emission model parameters and the signal-to-noise ratio (SNR) of the inferred optimized switching state sequence ($\hat{z}_{1:T}$), given by

$$\text{NMSE} = \frac{1}{K} \sum_{k=1}^K \frac{\|\hat{C}_k - C_k\|_F}{\|C_k\|_F}, \quad (116)$$

$$\text{SNR} = \frac{\sum_{t=1}^T \mathbb{I}(z_t = \hat{z}_t)}{T}, \quad (117)$$

where $\|\cdot\|_F$ denotes Frobenius norm, \hat{C}_k and $\hat{z}_{1:T}$ are estimated values. For AR-HMM, the NMSE is evaluated based on the true and estimated VAR emission parameters. To obtain $\hat{z}_{1:T}$, we set $\hat{z}_t = \arg \max_k \gamma_t(k)$. In addition, we avoid the optimization about discrete state number (K) and the dependence order (p) in HO-rSLDS by directly setting them to the true value.

The results of estimated NMSE and SNR for the three models are summarized in Tables 1 and 2, respectively. It is observed that HO-rSLDS outperforms AR-HMM and SLDS with lower average NMSE and higher average SNR for all the autoregressive orders and state duration distribution combinations. In addition, when $p = 1$, for both geometric and nonparametric state duration, SLDS outperforms AR-HMM slightly, which may be because the synthetic data are generated by a simulated extensive SLDS structure. However, when $p > 1$ and there is long-term dependence involved in the latent variables or observation data, it is just the opposite, since AR-HMM with higher p can capture the temporal relationships in the observation data while SLDS cannot. Moreover, when the transition of the underlying discrete states is characterized by a nonparametric duration distribution, the estimation accuracy of AR-HMM and SLDS is reduced, while HO-rSLDS remains stable with good estimation accuracy. This is because state transitions in both AR-HMM and SLDS are characterized by a Markov assumption and restricted to geometric duration distribution. Thus, more weight is given to shorter consecutive time periods within a certain state, which means the regimes switch frequently, which may not be appropriate in cases without sufficient prior information about a geometric state duration. By comparison, HO-rSLDS exhibits robustness to the state duration distributions since it improves the state transition by leveraging the stick-breaking logistic regression. The current state z_t no longer depends on only the preceding state z_{t-1} , but also the latent variable x_{t-1} , and, furthermore, the previous observations.

A more exhaustive comparison is illustrated in Figure 5 where HO-rSLDS outperforms the competing models with better estimation accuracy and statistical significance in all the cases (t -test, ** $p < 0.01$).

Table 1. NMSE derived from AR-HMM, SLDS, and HO-rSLDS based on the synthetic time series with different state duration distribution and autoregressive order in the generative model. All the results are the average of the 100 experiments (mean \pm std).

NMSE		$p = 1$	$p = 2$	$p = 3$
Geometric	AR-HMM	0.25 \pm 0.06	0.33 \pm 0.07	0.37 \pm 0.09
	SLDS	0.23 \pm 0.05	0.39 \pm 0.09	0.44 \pm 0.11
	HO-rSLDS	0.16 \pm 0.04	0.17 \pm 0.05	0.20 \pm 0.06
Nonparametric	AR-HMM	0.32 \pm 0.07	0.41 \pm 0.09	0.48 \pm 0.11
	SLDS	0.31 \pm 0.09	0.44 \pm 0.11	0.49 \pm 0.13
	HO-rSLDS	0.18 \pm 0.06	0.20 \pm 0.05	0.21 \pm 0.06

Table 2. SNR derived from AR-HMM, SLDS, and HO-rSLDS based on the synthetic time series with different state duration distribution and autoregressive order in the generative model. All the results are the average of the 100 experiments (mean \pm std).

	SNR	$p = 1$	$p = 2$	$p = 3$
Geometric	AR-HMM	0.79 ± 0.08	0.76 ± 0.07	0.69 ± 0.08
	SLDS	0.85 ± 0.06	0.68 ± 0.09	0.64 ± 0.09
	HO-rSLDS	0.93 ± 0.04	0.92 ± 0.05	0.90 ± 0.05
Nonparametric	AR-HMM	0.73 ± 0.08	0.65 ± 0.07	0.59 ± 0.10
	SLDS	0.80 ± 0.06	0.60 ± 0.09	0.54 ± 0.11
	HO-rSLDS	0.91 ± 0.04	0.87 ± 0.05	0.86 ± 0.05

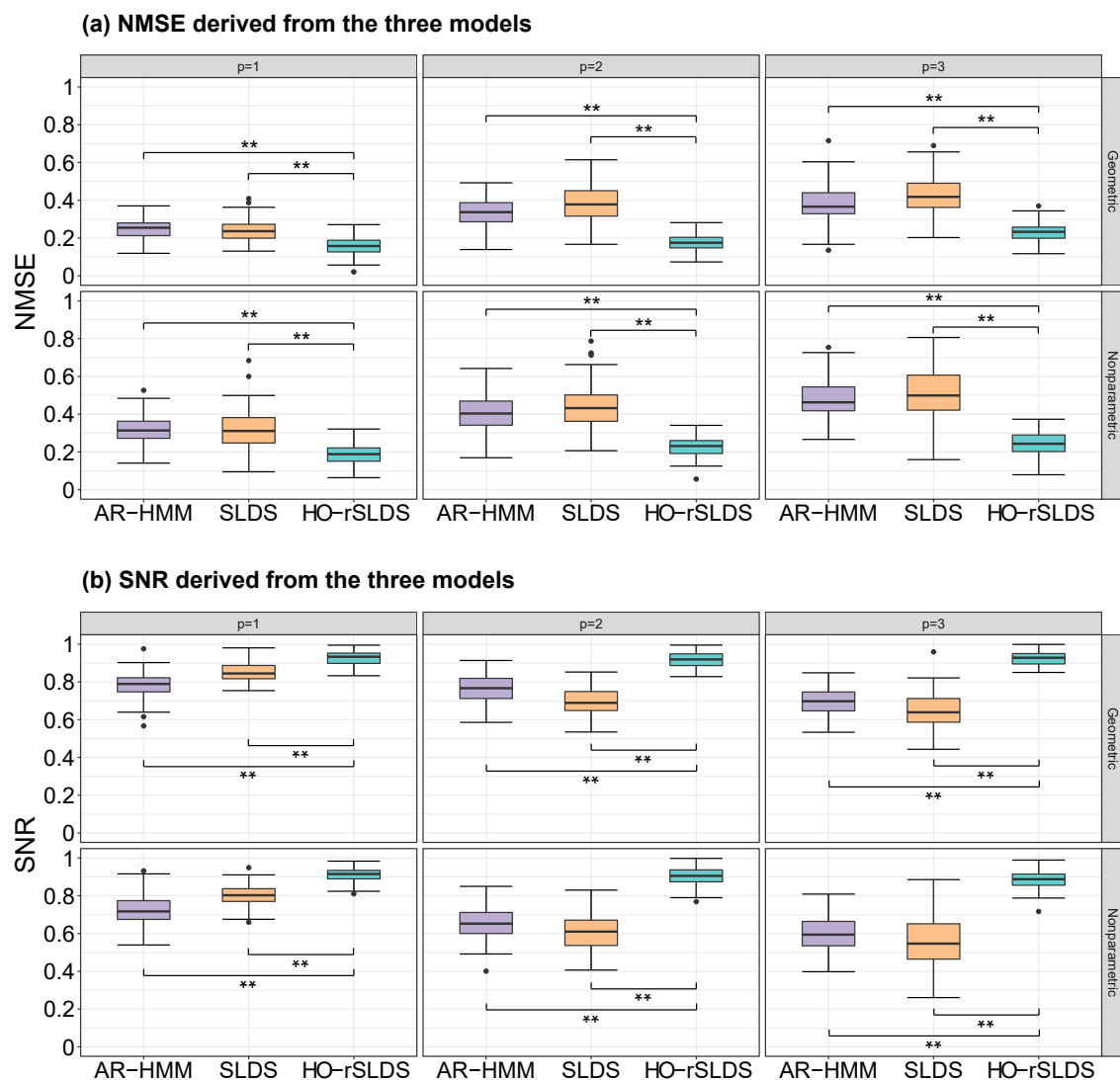


Figure 5. Box-plot of NMSE and SNR derived from AR-HMM, SLDS and HO-rSLDS based on the 100 numerical experiments. HO-rSLDS outperforms the competing models with significant lower NMSE and higher SNR on average for all the autoregressive order and state duration distribution combination. (*t*-test, ** $p < 0.01$).

5. Dynamic Functional Connectivity Analysis in fMRI Data

To show that HO-rSLDS can also work well with real-world data, we evaluate the performance of HO-rSLDS by comparing it with standard SLDS based on public functional magnetic resonance imaging (fMRI) data. The data are collected in the Human Connectome

Project (HCP) 1200 Parcellation Timeseries Netmats (HCP1200-TPN) dataset (Available at <https://www.humanconnectome.org>, (accessed on 6 March 2024)). Please see [28] for a detailed explanation of the entire acquisition protocol. In this study, we apply HO-rSLDS to data from 20 unrelated subjects with dimension $d = 15$ and times series length $T_0 = 4800$ for each subject included in the HCP1200-PTN release to perform dynamic functional connectivity (DFC) analysis, which can be used to analysis the dynamical temporal coherence among endogenous fluctuations in distributed brain regions [29,30]. In general, HO-rSLDS and SLDS assume that there are metastable states with characteristic connectivity pattern in the brain, while the connectivity pattern can be represented by the statistical correlations directly. Specifically, by Lemma 1, the functional connectivity pattern of state k in HO-rSLDS and SLDS can be obtained as

$$\text{DFC}_k = \frac{\sum_{t=1}^T \gamma_t(k) [\mathbb{E}C_k(\mathbb{E}x_t x_t^\top - \mathbb{E}x_t \mathbb{E}x_t^\top) \mathbb{E}C_k^\top + \mathbb{E}S_k]}{\sum_{t=1}^T \gamma_t(k)}. \quad (118)$$

In addition, there are two unknown parameters: the state number K and the dependence order p . We use the mixture minimum description length (MMDL) considering optimal code lengths for each state in a mixed model [31], given by

$$\text{MMDL}(K) := -\ln \mathcal{L}(K) + \frac{1}{2} \ln(T) K(K-1) + \sum_{k=1}^K \frac{1}{2} \ln(T \bar{\gamma}_k) d^2, \quad (119)$$

where $\mathcal{L}(K)$ is the variational lower bound evaluated during the model inference and $\bar{\gamma}_k := \sum_{t=1}^T \gamma_t(k)$. The concept of MMDL may be elucidated by the minimal description length principle in [32]. The state number K is determined by minimizing the MMDL on a specified list of K . In addition, we fit VAR model on the fMRI time series and the optimal dependence order p will be determined by AIC, BIC and HQ criteria [33], given by

$$\text{AIC} := \ln(\det(\hat{\Sigma})) + \frac{2pd^2}{T}, \quad (120)$$

$$\text{BIC} := \ln(\det(\hat{\Sigma})) + \frac{\ln(T)pd^2}{T}, \quad (121)$$

$$\text{HQ} := \ln(\det(\hat{\Sigma})) + \frac{2 \ln(\ln T)pd^2}{T}, \quad (122)$$

where $\hat{\Sigma}$ is the estimated covariance matrix of noise in VAR model. To avoid the model order too large to permit feasible computation, we choose the lag order on the condition when AIC/BIC/HQ show no further substantial decreases at higher orders [34]. As illustrated in Figure 6, we set $K = 4$ according to the minimal MMDL and $p = 4$ since the downward trend of AIC, BIC and HQ tends to be flat with higher order.

In Figure 7, we present the DFC patterns derived from the proposed HO-rSLDS and the standard SLDS. The 15 brain regions producing the fMRI time series are indexed by numbers 1 to 15. The states are ordered by their fractional occupancy (defined as the percentage of time allocated to that specific state); it is evident that the 4 states from SLDS have a more average fractional occupancy distribution. Moreover, it is observed that DFC patterns derived from SLDS are more similar than those from HO-rSLDS, where there is only a slight difference in DFC patterns between State4 and the other three states from SLDS. We use cosine similarity to measure the difference of the DFC patterns characterized by each state, given by

$$\cos(F_1, F_2) = \frac{\text{Tr}(F_1 F_2)}{\|F_1\|_F \|F_2\|_F}, \quad (123)$$

where F_1 and F_2 are square matrices. By this means, the averaged cosine similarity for SLDS is 0.9189 while that for HO-rSLDS is 0.6731. Although the real DFC patterns of

this fMRI data are unknown, compared to standard SLDS, HO-rSLDS can indeed dig out more information about the DFC patterns by the model superiority. For example, the State4 in HO-rSLDS, with the lowest fractional occupancy, can be considered as a strongly connected state where the absolute functional connectivity is significantly larger than that of the other three states. Moreover, compared with standard SLDS, there is significantly positive or negative functional connectivity in the four states derived from HO-rSLDS. In addition, there is similar functional connectivity between State1 and State2, which results in a higher cosine similarity than the averaged value for HO-rSLDS (0.7776). The similar functional connectivity is not exhibited obviously in State3 (such as the negative connectivity between region 8 and regions 11, 12, 13). However, these findings cannot be obtained by standard SLDS.

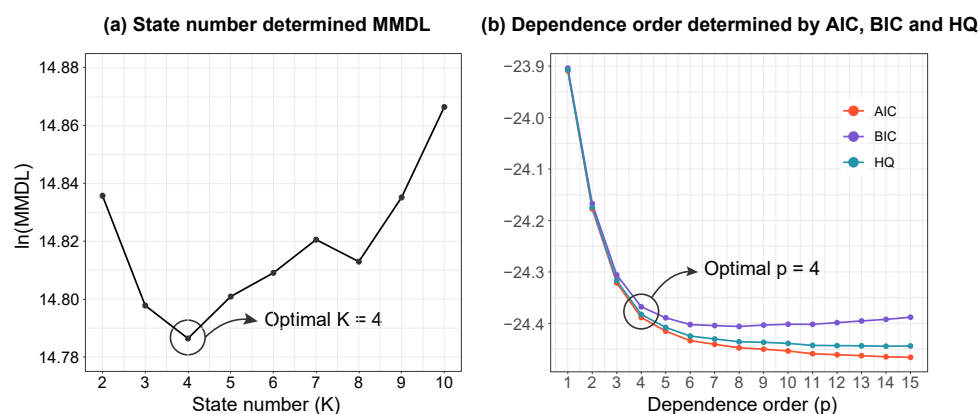
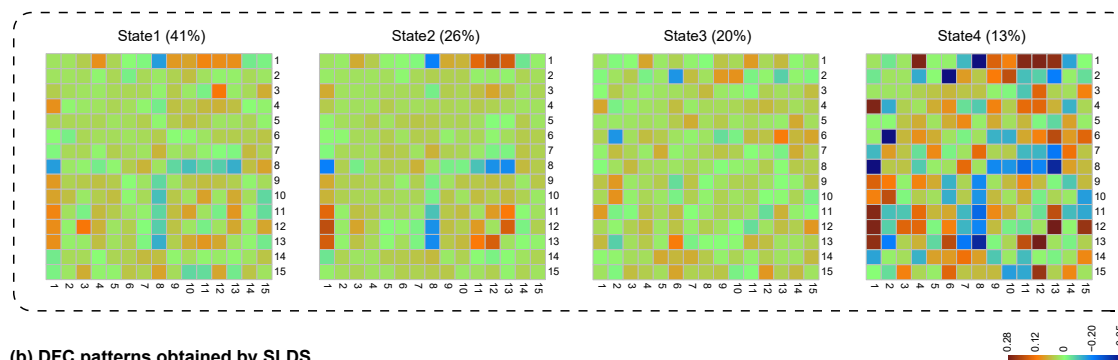


Figure 6. (a) State number (K) determined by MMDL. The best $K = 4$ is chosen with respect to the minimal MMDL. The y-axis has been logarithmically scaled to highlight the variation. (b) Dependence order p is determined by AIC, BIC, and HQ criteria. The ideal p is set to 4 since there is no further substantial decrease with higher order.

(a) DFC patterns obtained by LT-rSLDS



(b) DFC patterns obtained by SLDS

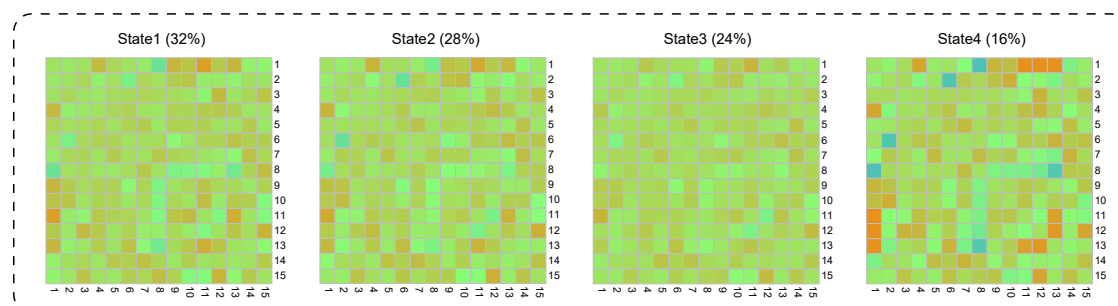


Figure 7. DFC patterns derived from HO-rSLDS (a) and standard SLDS (b). The 15 brain regions are indexed by the numbers 1 to 15, and the states are ordered by their fractional occupancy (number in the brackets). The color bar is shared by the two group heatmaps.

6. Conclusions

In this study, we have established a new method named HO-rSLDS as an extension of SLDS that improves the statistical modeling of dynamics from complex systems. First, HO-rSLDS addresses the problem of discovering a higher-order temporal dependence involved in the latent variables or observation data by allowing the current latent variable to be associated with the measurements of several previous time steps as in AR-HMM. Second, HO-rSLDS improves the switching state transition by utilizing stick-breaking logistic regression and Pólya-gamma augmentation, making the transition dependent on the latent variable and thus overcoming the limitations of the Markov assumption, leading to a geometric distribution of state durations. Moreover, the symmetric dependency between the switching states and the latent variable is recovered from the asymmetric relationships as in standard SLDS. We have presented the detailed inference algorithms of the HO-rSLDS for message passing and parameter learning. In numerical experiments, we concluded that HO-rSLDS outperforms the competing models with higher estimation accuracy and robustness. The utility and versatility of the developed HO-rSLDS have been demonstrated on real-world data as well. The application of fMRI data for DFC analyses indicates the superiority of the method in discovering abundant information about DFC patterns. A potential limitation of the proposed HO-rSLDS is that the state number and dependency order need to be predetermined in applications, while the hyperparameter optimization is always a non-trivial task. Future research will improve the proposed model by utilizing Bayesian nonparametric methods so that the complicated hyperparameter optimization can be efficiently avoided [35]. The proposed model improves the current methods for learning the switching linear dynamical modes, which will facilitate the identification and assessment of the dynamics of complex systems.

Author Contributions: Conceptualization, H.W. and J.C.; methodology, H.W.; software, H.W.; validation, H.W.; formal analysis, H.W. and J.C.; investigation, H.W.; resources, H.W. and J.C.; data curation, J.C.; writing—original draft preparation, H.W.; writing—review and editing, J.C.; visualization, H.W.; supervision, J.C.; project administration, J.C.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation (grant number 81671633), and the project was supported by the Open Fund of Hubei Longzhong Laboratory.

Data Availability Statement: The data that supported the findings of this study are included in this article.

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

Abbreviations

The following abbreviations are used in this manuscript:

SLDS	Switching linear dynamical systems
HO-SLDS	Switching linear dynamical systems with higher-order dependence
HO-rSLDS	Recurrent switching linear dynamical systems with higher-order dependence
HMM	Hidden Markov model
VAR	Vector autoregressive
AR-HMM	Autoregressive hidden Markov model
AR-HSMM	Autoregressive hidden semi-Markov model
NMSE	Normalized mean squared error
SNR	Signal-to-noise ratio
fMRI	Functional magnetic resonance imaging
DFC	Dynamic functional connectivity

References

1. Fox, E.; Sudderth, E.B.; Jordan, M.I.; Willsky, A.S. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Trans. Signal Process.* **2011**, *59*, 1569–1585. [\[CrossRef\]](#)
2. Pandarinath, C.; O'Shea, D.J.; Collins, J.; Jozefowicz, R.; Stavisky, S.D.; Kao, J.C.; Trautmann, E.M.; Kaufman, M.T.; Ryu, S.I.; Hochberg, L.R.; et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **2018**, *15*, 805–815. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Frigola, R.; Chen, Y.; Rasmussen, C.E. Variational Gaussian process state-space models. In Proceedings of the Annual Conference on Neural Information Processing Systems 2014, Montreal, QC, Canada, 8–13 December 2014; Volume 27.
4. Krishnan, R.; Shalit, U.; Sontag, D. Structured inference networks for nonlinear state space models. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
5. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 2020.
6. de Souza Baptista, R.; Bó, A.P.; Hayashibe, M. Automatic human movement assessment with switching linear dynamic system: Motion segmentation and motor performance. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2016**, *25*, 628–640. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Liu, H.; Wang, L. Human motion prediction for human-robot collaboration. *J. Manuf. Syst.* **2017**, *44*, 287–294. [\[CrossRef\]](#)
8. Alameda-Pineda, X.; Drouard, V.; Horaud, R.P. Variational inference and learning of piecewise linear dynamical systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 3753–3764. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Ma, T.; Srinivasan, S.; Lazarou, G.; Picone, J. Continuous speech recognition using linear dynamic models. *Int. J. Speech Technol.* **2014**, *17*, 11–16. [\[CrossRef\]](#)
10. Pagan, A.R.; Pesaran, M.H. Econometric analysis of structural systems with permanent and transitory shocks. *J. Econ. Dyn. Control* **2008**, *32*, 3376–3395. [\[CrossRef\]](#)
11. Haluszczynski, A.; Räth, C. Controlling nonlinear dynamical systems into arbitrary states using machine learning. *Sci. Rep.* **2021**, *11*, 12991. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Smith, J.F.; Pillai, A.; Chen, K.; Horwitz, B. Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. *Neuroimage* **2010**, *52*, 1027–1040. [\[CrossRef\]](#)
13. Wang, E.T.; Vannucci, M.; Haneef, Z.; Moss, R.; Rao, V.R.; Chiang, S. A Bayesian switching linear dynamical system for estimating seizure chronotypes. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2200822119. [\[CrossRef\]](#)
14. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **1989**, *77*, 257–286. [\[CrossRef\]](#)
15. Vidaurre, D. A new model for simultaneous dimensionality reduction and time-varying functional connectivity estimation. *PLoS Comput. Biol.* **2021**, *17*, e1008580. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Hamada, R.; Kubo, T.; Ikeda, K.; Zhang, Z.; Shibata, T.; Bando, T.; Hitomi, K.; Egawa, M. Modeling and prediction of driving behaviors using a nonparametric Bayesian method with AR models. *IEEE Trans. Intell. Veh.* **2016**, *1*, 131–138. [\[CrossRef\]](#)
17. Hout, J.W.; Frame, M.E.; Blaha, L.M. Unsupervised parsing of gaze data with a beta-process vector auto-regressive hidden Markov model. *Behav. Res. Methods* **2018**, *50*, 2074–2096. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Vidaurre, D.; Quinn, A.J.; Baker, A.P.; Dupret, D.; Tejero-Cantero, A.; Woolrich, M.W. Spectrally resolved fast transient brain states in electrophysiological data. *Neuroimage* **2016**, *126*, 81–95. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Glennie, R.; Adam, T.; Leos-Barajas, V.; Michelot, T.; Photopoulou, T.; McClintock, B.T. Hidden Markov models: Pitfalls and opportunities in ecology. *Methods Ecol. Evol.* **2023**, *14*, 43–56. [\[CrossRef\]](#)
20. Johnson, M.J.; Willsky, A.S. Bayesian Nonparametric Hidden Semi-Markov Models. *J. Mach. Learn. Res.* **2013**, *14*, 673–701.
21. Linderman, S.W.; Johnson, M.J.; Miller, A.C.; Adams, R.P.; Blei, D.M.; Paninski, L. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 914–922.
22. Polson, N.G.; Scott, J.G.; Windle, J. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Stat. Assoc.* **2013**, *108*, 1339–1349. [\[CrossRef\]](#)
23. Jacobs, W.R.; Baldacchino, T.; Dodd, T.; Anderson, S.R. Sparse Bayesian nonlinear system identification using variational inference. *IEEE Trans. Autom. Control* **2018**, *63*, 4172–4187. [\[CrossRef\]](#)
24. Yu, S.Z. Hidden semi-Markov models. *Artif. Intell.* **2010**, *174*, 215–243. [\[CrossRef\]](#)
25. Liu, Q.; Liu, W.; Dong, M.; Li, Z.; Zheng, Y. Residual useful life prognosis of equipment based on modified hidden semi-Markov model with a co-evolutional optimization method. *Comput. Ind. Eng.* **2023**, *182*, 109433. [\[CrossRef\]](#)
26. Särkkä, S.; Svensson, L. *Bayesian Filtering and Smoothing*; Cambridge University Press: Cambridge, UK, 2023; Volume 17.
27. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [\[CrossRef\]](#)
28. Glasser, M.F.; Sotiropoulos, S.N.; Wilson, J.A.; Coalson, T.S.; Fischl, B.; Andersson, J.L.; Xu, J.; Jbabdi, S.; Webster, M.; Polimeni, J.R.; et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **2013**, *80*, 105–124. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Hutchison, R.M.; Womelsdorf, T.; Allen, E.A.; Bandettini, P.A.; Calhoun, V.D.; Corbetta, M.; Della Penna, S.; Duyn, J.H.; Glover, G.H.; Gonzalez-Castillo, J.; et al. Dynamic functional connectivity: Promise, issues, and interpretations. *Neuroimage* **2013**, *80*, 360–378. [\[CrossRef\]](#) [\[PubMed\]](#)

30. Motlaghian, S.; Vahidi, V.; Baker, B.; Belger, A.; Bustillo, J.; Faghiri, A.; Ford, J.; Iraj, A.; Lim, K.; Mathalon, D.; et al. A method for estimating and characterizing explicitly nonlinear dynamic functional network connectivity in resting-state fMRI data. *J. Neurosci. Methods* **2023**, *389*, 109794. [[CrossRef](#)]
31. Zhang, G.; Cai, B.; Zhang, A.; Stephen, J.M.; Wilson, T.W.; Calhoun, V.D.; Wang, Y.P. Estimating dynamic functional brain connectivity with a sparse hidden Markov model. *IEEE Trans. Med. Imaging* **2019**, *39*, 488–498. [[CrossRef](#)] [[PubMed](#)]
32. Barron, A.; Rissanen, J.; Yu, B. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760. [[CrossRef](#)]
33. Gredenhoff, M.; Karlsson, S. Lag-length selection in VAR-models using equal and unequal lag-length procedures. *Comput. Stat.* **1999**, *14*, 171–187. [[CrossRef](#)]
34. Brovelli, A.; Ding, M.; Ledberg, A.; Chen, Y.; Nakamura, R.; Bressler, S.L. Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9849–9854. [[CrossRef](#)]
35. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Hierarchical Dirichlet Processes. *J. Am. Stat. Assoc.* **2006**, *101*, 1566–1581. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.