Tractable Inference and Observation Likelihood Evaluation in Latent Structure Influence Models

Sajjad Karimi, *Student Member, IEEE*, and Mohammad Bagher Shamsollahi, *Senior Member, IEEE*

***Abstract*—Latent Structure Influence Models (LSIMs) are a particular kind of Coupled Hidden Markov Models (CHMMs). Against CHMMs, LSIMs overcome the exponential growth of state-space parameters by considering the influence model for coupled Markov chains. Nevertheless, the exact inference in LSIMs requires exponential complexity. We propose a new recursive formulation to compute marginal forward and backward parameters by $\mathcal{O}(T(NC)^2)$ instead of $\mathcal{O}(TN^{2C})$ for C channels of N states apiece observing T data points. This formulation is derived systematically and carefully to increase the inference accuracy. Furthermore, a solution is presented for the evaluation problem of LSIMs based on the proposed marginal forward parameter. This solution is essential in statistical multi-channel time-series classification. The results show that the proposed algorithm is generally more accurate and reliable than other existing algorithms. Novelties in deriving the marginal backward parameter plays an important role in this superiority. The Hellinger distance is computed between the proposed and exact forward and one-slice parameters for various simulation scenarios. Distances are small enough, indicating that the proposed inference algorithm is sufficiently close to exact inference for various channels, hidden state numbers, and other parameters. Statistical multi-channel time-series classification is also considered for both proposed and exact algorithms. Classification results are almost similar, indicating that the proposed approximate inference is proper and acceptable in the classification task. Finally, the iEEG dataset's parameter learning indicates that the proposed inference algorithm leads to a higher log-likelihood than the existing algorithms.**

***Index Terms*—Approximate inference, Boyen-Koller algorithm, convex combination, coupled hidden Markov model, forward and backward parameters, influence model, latent structure influence model, squared euclidean distance.**

I. Introduction

MODELING complex dynamic systems consisting of multiple interacting processes is essential in various fields of science and engineering. Standard Hidden Markov models (HMMs) can be used to model multi-channel time-series of complex dynamic systems. Nevertheless, if several interacting channels (or processes) generate a multi-channel time-series, an HMM with a single hidden variable is ill-suited to it. Coupled hidden Markov models (CHMMs) are the extension of HMMs that contain multiple interacting channels [1], [2]. CHMMs have been studied in many applications such as audio-visual speech recognition (AVSR) [3], [4], dynamic functional connectivity (dFC) in fMRI [5], EEG and ECG classification [6]–[8], disease interactions [9], freeway traffic modelling [10], [11] and financial crisis detection [12].

CHMMs were proposed and investigated in [1] and [13] for continuous and discrete amplitude observations. In a CHMM, each channel has its Markov chain, associated with its observations (channel observation can be univariate or multivariate). Transition probabilities of the current state of each channel depend on all previous hidden states. In general, this structure implies that the state space parameters grow exponentially concerning the number of channels. A C-channel CHMM with N hidden states per channel can be transformed into an equivalent HMM (with $\mathcal{O}(N^C)$ hidden states) using the Cartesian product of all hidden states. So, the exact inference has a computation complexity $\mathcal{O}(TN^{2C})$, which makes it impractical for applications with a large number of channels.

Several approaches have been proposed in the literature to overcome the problem of state-space parameters. Brand assumed a simplification considering a factorization of transition matrix as follows

$$P(q_t^{\xi}|q_{t-1}^1,\ldots,q_{t-1}^C) \equiv \prod_{c=1}^{C} P(q_t^{\xi}|q_{t-1}^c), \tag{1}$$

where q_t^{ξ} denotes state of channel ξ at time t [1], [14]. So, state conditional probability in the left side of (1) is substituted by the product of marginal conditional probabilities. Even though, the Brand's assumption reduces transition probability parameter space but a normalizing value is necessary to hold equality on both sides of (1), which is also reported in [15].

The convex combination model [13] or the influence model [16] is the next approach well-defined as opposed to Brand's assumption. The influence model also prevents the exponential growth of transition probabilities parameters, and transition matrices are factorized as follows

$$P(q_t^{\xi}|q_{t-1}^1,\ldots,q_{t-1}^C) = \sum_{c=1}^{C} \theta^{c,\xi} P(q_t^{\xi}|q_{t-1}^c),$$

$$\theta^{c,\xi} \geq 0, \sum_{c=1}^{C} \theta^{c,\xi} = 1, \tag{2}$$

Manuscript received April 19, 2020; revised July 30, 2020 and September 1, 2020; accepted September 4, 2020. Date of publication September 21, 2020; date of current version October 15, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark A. Davenport. *(Corresponding author: Sajjad Karimi.)*

The authors are with Biomedical Signal and Image Processing Laboratory (BiSIPL), School of Electrical Engineering, Sharif University of Technology, Tehran 009821, Iran (e-mail: sajjadkarimi91@gmail.com; mbshams@sharif.ir).

Digital Object Identifier 10.1109/TSP.2020.3025522

where, $\theta^{c,\xi}$ is coupling weight from channel c to channel ξ, showing how much q_{t-1}^{c} influences distribution of q_t^{ξ}. Set of coupling weights can be viewed as a weighted-directed graph or network that firstly introduced in [16] as influence model. So, there are a variety of network measures that can describe this network [17]. In [18], convex combination was introduced for the modeling of higher-order Markov chain. This model was also used to stochastic language modeling [19]. After that, convex combination was used in mixed memory Markov models with discrete observation [13], and then it was extended in the context of continuous observation modeling of complex stochastic systems [15]. In [16], the influence model (an alternative name of convex combination) was employed to represent dynamical interaction over networks. This model was then used as the interaction model of Markov chains in CHMMs (known as LSIMs), which was widely applied in social computing in several studies [20]–[23].

Like CHMMs, joint forward and backward parameters can be computed in $\mathcal{O}(TN^{2C})$ for a C-channel LSIM with N hidden states [1], [15], [24]. This computational complexity grows exponentially concerning the number of channels, which is demanding and time-consuming. For example, if an LSIM model (with 10 hidden states per channel), is applied to 100 milliseconds (sampled at 1000 Hz) of a 32-channel EEG time-series, it needs 10^{66} computations for the exact algorithm, which can be reduced to 10^{7} using the proposed algorithm. Even though the influence model reduced transition parameters, but the computational complexity of exact inference still grows exponentially with the number of channels.

There are several approaches to cope with the computational complexity in CHMMs. In [25], it was shown that weak interactions could be omitted, and a complex system consisting of multiple channels can be decomposed into several independent smaller subsystems [26]. N-heads dynamic programming was also proposed to perform approximate inference in CHMMs with Brand's assumption [1]. A recent study suggested another approximate inference algorithm based on the Brand's assumption with $\mathcal{O}(T(NC)^2)$ computational complexity [8]. The algorithm computes the marginal forward and backward parameters recursively using two different simplifying assumptions, and it also improved the performance of Apnea Bradycardia detection.

There are also two approximate inference algorithms for LSIMs considering the influence model [27], [28]. The first algorithm uses a nonlinear mapping based on Structured Variational Inference (SVI), and the marginal forward and backward parameters are calculated recursively with computational complexity $\mathcal{O}(T(NC)^2)$ [27]. The latter algorithm was developed using mean-field approximation and variational inference [28]. This algorithm calculates the marginal one-slice parameter considering the Completely Factorized Variational Inference (CFVI) with computational complexity $\mathcal{O}(T(NC)^3)$. As shown in the results section, despite of more computation cost of the CFVI algorithm, its inference error is higher than the SVI algorithm.

CHMMs and LSIMs enrich the capability of HMMs in analyzing multi-channel datasets. These models can improve accuracy and performance in various applications, including modeling, segmentation, and classification tasks. Thus an efficient and tractable inference algorithm is necessary for CHMMs and LSIMs. The exponential growth of state-space parameters is a crucial weakness of CHMMs in datasets with numerous channels. We focused on developing an inference algorithm for LSIMs since existing approximate inference algorithms have multiple limitations. Brand's assumption needs normalizing values at the order of $\mathcal{O}(N^C)$ according to $q_{t-1}^{1}, \ldots, q_{t-1}^{C}$, and these values may interrupt simplification procedures of the mentioned frameworks. In contrast, the influence model does not need any normalization value, but there are also some points associated with SVI and CFVI algorithms. Marginal forward and backward parameters were not compared to exact ones to analyze the error of approximate inference. The marginal backward parameter error is sensitive to weighted out-degree and increases for channels with small weighted out-degree (precise and accurate analysis was omitted here due to space constraints). This sensitivity can interrupt the monotone convergence of log-likelihood values in the Expectation-Maximization (EM) algorithm. So, contributions of this work are summarized as follows.

- A new formulation is derived to compute recursively marginal forward and backward parameters (inference) in LSIMs. This formulation outperforms the accuracy of existing approximate inference algorithms due to a systematic and constructive derivation. The main advantage and improvement of the proposed formulation mostly come from the novelty in the derivation of marginal backward parameter.
- A fast and closed-form solution is presented to find optimum mixing weights of a mixture model for discrete conditional probabilities by minimizing squared Euclidean distance. This solution allows us to complete the recursion of the marginal backward parameter.
- Marginal forward, backward and one-slice parameters are formulated in recursive closed-form expressions. The complexity order of proposed formulations is $\mathcal{O}(T(NC)^2)$, and real datasets' applications also reveal the more reliable performance of these formulations over existing ones with the same complexity order.
- The evaluation problem of LSIMs is solved using the proposed marginal forward parameter, and this provides a fast multi-channel time-series classification.

Rest of this article is organized as follows. In the next section, the model and mathematical formulation are described. Then, procedures for the construction of simulated data are explained, and validation criteria are described. Following these methodological aspects, the results of the proposed inference algorithm are reported on simulated and real datasets.

II. Proposed LSIM Framework

In this section, symbols and variables are adequately defined with the same notations as [8]. We then define the marginal forward parameter and propose a recursive formulation to compute it efficiently, and marginal backward and one-slice parameters are appropriately defined. Achieving a recursion for the

marginal backward parameter is more challenging compared to the marginal forward parameter. At the end of this section, we present a formulation to solve the evaluation problem of LSIMs using the marginal forward parameter.

A. Notations

Assume there is an LSIM with C channels, and its observations are available for $t = 1, \ldots, T$. Let us denote $S^c = \{S_1^c, S_2^c, \ldots, S_{M(c)}^c\}$ to be state space of channel c in the LSIM. Let $q_t^c \in S^c$ and $o_t^c \in \mathbb{R}^{L(c)}$ be state and observation of channel c at time t, respectively and $L(c)$ is observation dimension of channel c. Initial state probabilities of each channel are denoted by $\pi_m^c = P(q_1^c = S_m^c)$ and $\pi = \{\pi_m^c | m = 1, \ldots, M(c), c = 1, \ldots, C\}$. Also let $a_{m,n}^{c,\xi} = P(q_t^\xi = S_n^\xi | q_{t-1}^c = S_m^c)$ and $\theta^{c,\xi}$ is coupling weight from channel c to channel ξ. The sets of all transition matrices and coupling weights are denoted by A and Θ respectively. The emission probabilities of the observation given its hidden state is written as $b_m^c(o_t^c) = f(o_t^c | q_t^c = S_m^c)$ where o_t^c may be either discrete or continuous. In this study, observations are assumed to be continuous amplitude, and emission probabilities $b_m^c(o_t^c)$ belong to Gaussian Mixture Model (GMM) families as follows

$$\begin{aligned} b_m^c(o_t^c) &= \sum_{k=1}^{D(c)} \omega_{m,k}^c \mathcal{N}(\mu_{m,k}^c, \Sigma_{m,k}^c) \\ &= \sum_{k=1}^{D(c)} \omega_{m,k}^c b_{m,k}^c(o_t^c), \end{aligned} \tag{3}$$

where $D(c)$ is the number of Gaussian in channel c and $\omega_m^c = \{\omega_{m,1}^c, \ldots, \omega_{m,D(c)}^c\}$, $\mu_m^c = \{\mu_{m,1}^c, \ldots, \mu_{m,D(c)}^c\}$ and $\Sigma_m^c = \{\Sigma_{m,1}^c, \ldots, \Sigma_{m,D(c)}^c\}$ are weights, means and covariance matrices of GMM in channel c at state m, respectively. Sets of all mixing weights, means and covariance matrices are also denoted by ω, μ and Σ. Thus, the LSIM is characterized by $\lambda = \{\pi, A, \Theta, \omega, \mu, \Sigma\}$.

Set of observations at time t is denoted by $o_t = \{o_t^1, o_t^2, \ldots, o_t^C\}$ and set of observations in interval $t_s : t_p$ is denoted by $o_{t_s:t_p} = \{o_{t_s}, o_{t_s+1}, \ldots, o_{t_p}\}$. Besides, let $S = S^1 \times \ldots \times S^C$ be joint state space of all channels, and let $q_t = \{q_t^1, \ldots, q_t^C\} \in S$ be the random variable describing the state in S at time t. There are also three simplifying definitions as $v_t^c(m) = \{q_t^c = S_m^c\}$, $v_t(n_1, \ldots, n_C) = \{q_t^1 = S_{n_1}^1, \ldots, q_t^C = S_{n_C}^C\}$ and $v_t(\mathbf{n}) = v_t(n_1, \ldots, n_C)$.

B. Forward Parameter

Forward and backward parameters play a central role in evaluation, inference, and learning in HMMs, CHMMs, and LSIMs. Following previous studies [2], [8], [29], marginal forward parameter is defined as

$$\alpha_{t|x}^\xi(m) = P(v_t^\xi(m) | o_{1:x}), \tag{4}$$

where, for $x = t-1$, t and $x = T$, the above quantity is termed as predicted, filtered and smoothed probability, respectively.

The standard prediction equation (also known as Chapman-Kolmogorov equation) is used to compute forward parameter recursively as

$$\begin{aligned} &\alpha_{t|t-1}^\xi(m) \\ &= \sum_{n_1=1}^{M(1)} \cdots \sum_{n_C=1}^{M(C)} P(v_t^\xi(m) | v_{t-1}(\mathbf{n}), o_{1:t-1}) P(v_{t-1}(\mathbf{n}) | o_{1:t-1}). \end{aligned} \tag{5}$$

In the first term of the right side of equation, $o_{1:t-1}$ can be omitted since given previous joint states, current state becomes independent from past observations. Using the influence model, it follows that

$$\alpha_{t|t-1}^\xi(m) = \sum_{n_1=1}^{M(1)} \cdots \sum_{n_C=1}^{M(C)} \sum_{c=1}^{C} \theta^{c,\xi} a_{m,n_c}^{c,\xi} P(v_{t-1}(\mathbf{n}) | o_{1:t-1}). \tag{6}$$

Now by changing summations, the following recursion is obtained

$$\alpha_{t|t-1}^\xi(m) = \sum_{c=1}^{C} \theta^{c,\xi} \sum_{n_c=1}^{M(c)} a_{m,n_c}^{c,\xi} \alpha_{t-1|t-1}^c(n_c). \tag{7}$$

Finally, forward parameter is recursively computed according to

$$\alpha_{t|t-1}^\xi(m) = \sum_{c=1}^{C} \theta^{c,\xi} \sum_{n_c=1}^{M(c)} a_{m,n_c}^{c,\xi} \alpha_{t-1|t-2}^c(n_c) \tilde{\mathbf{b}}_{n_c}^c(o_{t-1}^c), \tag{8}$$

where $\tilde{\mathbf{b}}_{n_c}^c(o_t^c)$ is defined as $\frac{\alpha_{t|t}^c(n_c)}{\alpha_{t|t-1}^c(n_c)}$. From Bayes' rule and independence of the observation of one channel from the other channels given its hidden state, it can be deduced that

$$P(v_t(\mathbf{n}) | o_{1:t}) \propto P(v_t(\mathbf{n}) | o_{1:t-1}) \prod_{c=1}^{C} b_{n_c}^c(o_t^c). \tag{9}$$

Boyen-Koller (BK) algorithm is used to simplify $P(v_t(\mathbf{n}) | o_{1:t-1})$ and decomposing joint probability in independent clusters [25], [30], [31]. We assume that q_t^c is independent from $q_t^{c'}$ given $o_{1:t-1}$ for any $c' \in \{1, \ldots, C\}$, and after marginalization, it holds that:

$$\begin{aligned} &P(v_t^c(n_c) | o_{1:t}) \propto P(v_t^c(n_c) | o_{1:t-1}) b_{n_c}^c(o_t^c) \\ &\times \underbrace{\sum_{n_1=1}^{M(1)} \cdots \sum_{n_C=1}^{M(C)}}_{\text{except c}} \prod_{c'=1}^{C} b_{n_{c'}}^{c'}(o_t^{c'}) P(v_t^{c'}(n_c) | o_{1:t-1}). \end{aligned} \tag{10}$$

Since normalizing constant is simply the sum over n_c of the right side of previous equation, the expectation is simplified, and it follows that

$$\alpha_{t|t}^c(n_c) = \frac{\alpha_{t|t-1}^c(n_c) b_{n_c}^c(o_t^c)}{\sum_{n_c=1}^{M(c)} \alpha_{t|t-1}^c(n_c) b_{n_c}^c(o_t^c)} = \frac{\alpha_{t|t-1}^c(n_c) b_{n_c}^c(o_t^c)}{f(o_t^c | o_{1:t-1})}. \tag{11}$$

So, $\tilde{\mathbf{b}}^c_{n_c}(o^c_t)$ formula is obtained

$$\tilde{\mathbf{b}}^c_{n_c}(o^c_t) = \frac{b^c_{n_c}(o^c_t)}{\sum_{n_c=1}^{M(c)} \alpha^c_{t|t-1}(n_c) b^c_{n_c}(o^c_t)}. \tag{12}$$

Decomposition of observation is also immediately resulted from BK algorithm and similar inspections as

$$f(o_t|o_{1:t-1}) = \prod_{c=1}^{C} f(o^c_t|o_{1:t-1}). \tag{13}$$

C. Backward Parameter

A favorite definition of the marginal one-slice parameter (smoothed probability) is the product of marginal forward and backward parameters. We define the marginal backward parameter based on the ratio of the one-slice parameter over the forward parameter as follows

$$\beta^\xi_t(m) = \frac{\alpha^\xi_{t|T}(m)}{\alpha^\xi_{t|t-1}(m)} = \tilde{\mathbf{b}}^\xi_m(o^\xi_t) \frac{f(o_{t+1:T}|v^\xi_t(m), o_{1:t})}{f(o_{t+1:T}|o_{1:t})}. \tag{14}$$

Backward recursion is derived by summing on the next hidden states of all channels, using Bayes' rule and properties of Markov chains to omit conditioning on $o_{t+1:T}$ given $v_{t+1}(n_1, \ldots, n_C)$ (see Appendix A) as follows

$$\beta^\xi_t(m) = \frac{1}{\alpha^\xi_{t|t-1}(m)} \sum_{n_1=1}^{M(1)} \cdots \sum_{n_C=1}^{M(C)} P(v^\xi_t(m)|v_{t+1}(\mathbf{n}), o_{1:t}) P(v_{t+1}(\mathbf{n})|o_{1:T}). \tag{15}$$

A mixture model approximation will be used to simplify $P(v^\xi_t(m)|v_{t+1}(\mathbf{n}), o_{1:t})$ as follows

$$P(v^\xi_t(m)|v_{t+1}(\mathbf{n}), o_{1:t}) \approx \sum_{w=1}^{C} \hat{\mathbf{d}}^{\xi,w}_t P(v^\xi_t(m)|v^w_{t+1}(n_w), o_{1:t}). \tag{16}$$

Assuming the set $\{\hat{\mathbf{d}}^{\xi,w}_t\}_{w=1}^C$ is available, and minimizing an appropriate statistical distance (such as Kullback-Leibler divergence) between both sides of (16) leads to this set of mixing weights. Substituting (16) in (15) and changing summations simplifies backward recursion by the following formulas

$$\begin{aligned}\beta^\xi_t(m) &= \sum_{w=1}^{C} \hat{\mathbf{d}}^{\xi,w}_t \sum_{n_w=1}^{M(w)} \frac{P(v^\xi_t(m), v^w_{t+1}(n_w)|o_{1:t})}{\alpha^\xi_{t|t-1}(m)\alpha^w_{t+1|t}(n_w)} \alpha^w_{t+1|T}(n_w) \\ &= \frac{\alpha^\xi_{t|t}(m)}{\alpha^\xi_{t|t-1}(m)} \sum_{w=1}^{C} \hat{\mathbf{d}}^{\xi,w}_t \\ &\quad \times \sum_{n_w=1}^{M(w)} P(v^w_{t+1}(n_w)|v^\xi_t(m), o_{1:t}) \beta^w_{t+1}(n_w).\end{aligned} \tag{17}$$

Afterward, the backward parameter is recursively computed according to the following relation

$$\beta^\xi_t(m) = \tilde{\mathbf{b}}^\xi_m(o^\xi_t) \sum_{w=1}^{C} \hat{\mathbf{d}}^{\xi,w}_t \sum_{n_w=1}^{M(w)} \rho^{\xi,w}_{t+1}(m, n_w) \beta^w_{t+1}(n_w), \tag{18}$$

where, $\rho^{\xi,w}_{t+1}(m, n_w)$ is defined as $P(v^w_{t+1}(n_w)|v^\xi_t(m), o_{1:t})$. The influence model, BK algorithm and Markov chains properties also simplifies computation of $\rho^{\xi,w}_{t+1}(m, n_w)$, and the following formula is achieved

$$\begin{aligned}&\rho^{\xi,w}_{t+1}(m, n_w) \\ &= \theta^{\xi,w} a^{\xi,w}_{m,n_w} + \sum_{\substack{c=1 \\ c\neq\xi}}^{C} \theta^{c,w} \sum_{m_c=1}^{M(c)} a^{c,w}_{m_c,n_w} \alpha^c_{t|t}(m_c) \\ &= \alpha^w_{t+1|t}(n_w) + \theta^{\xi,w} \left(a^{\xi,w}_{m,n_w} - \sum_{m_\xi=1}^{M(\xi)} a^{\xi,w}_{m_\xi,n_w} \alpha^\xi_{t|t}(m_\xi) \right).\end{aligned} \tag{19}$$

An attractive and intuitive interpretation of the above equation is that if $\theta^{\xi,w}$ is zero, $\rho^{\xi,w}_{t+1}(m, n_w)$ is equal to $\alpha^w_{t+1|t}(n_w)$ in (19) as expected.

To complete the backward recursion in (18), we must compute the set $\{\hat{\mathbf{d}}^{\xi,w}_t\}_{w=1}^C$ efficiently while minimizing a suitable distribution distance criterion. Various studies have used Kullback-Leibler divergence (KL-divergence) as an appropriate distance criterion [32]–[34]. So in this study, KL-divergence is chosen as a preliminary distance measure between probabilities on both sides of (16), and minimizing KL-divergence leads to the set $\{\hat{\mathbf{d}}^{\xi,w}_t\}_{w=1}^C$. Both sides of (16) include conditional probabilities, so the expected KL-divergence is considered instead of the KL-divergence [35]–[37]. Expected KL-divergence is a function of mixing weights, according to [36]

$$\begin{aligned}\mathbf{KL}(d^{\xi,1}_t, \ldots, d^{\xi,C}_t) &= \sum_{n_1=1}^{M(1)} \cdots \sum_{n_C=1}^{M(C)} P(v_{t+1}(\mathbf{n})|o_{1:t}) \\ &\times \left(\sum_{m=1}^{M(\xi)} P(v^\xi_t(m)|v_{t+1}(\mathbf{n}), o_{1:t}) \right. \\ &\left. \times \log \frac{P(v^\xi_t(m)|v_{t+1}(\mathbf{n}), o_{1:t})}{\sum_{w=1}^{C} d^{\xi,w}_t P(v^\xi_t(m)|v^w_{t+1}(n_w), o_{1:t})} \right).\end{aligned} \tag{20}$$

Unfortunately, $\mathbf{KL}(d^{\xi,1}_t, \ldots, d^{\xi,C}_t)$ and $\frac{\partial \mathbf{KL}(d^{\xi,1}_t, \ldots, d^{\xi,C}_t)}{\partial d^{\xi,c}_t}$ need $\mathcal{O}(N^{C+1})$ computations, which is time-consuming and demanding, and the minimization of $\mathbf{KL}(d^{\xi,1}_t, \ldots, d^{\xi,C}_t)$ also needs more computations, and these computations stop the proposed fast recursive formulation. Consequently, we changed the KL-divergence with alternative statistical distances to have a fast solution for $\{\hat{\mathbf{d}}^{\xi,w}_t\}_{w=1}^C$. Since KL-divergence belongs to Bregman divergences, we focused on these divergences. Bregman divergences include a large number of useful distances such as Squared Euclidean distance (SED), KL-divergence, squared Mahalanobis distance, Itakura Saito, and I-divergence [38].

We achieved a fast and closed-form solution by considering SED instead of KL-divergence. The procedures of finding this solution are described in detail on the following. Substituting KL-divergence by SED in (20), the new distance function is obtained

$$\mathbf{SED}(d_t^{\xi,1},\ldots,d_t^{\xi,C}) = \sum_{n_1=1}^{M(1)} \cdots \sum_{n_C=1}^{M(C)} P(v_{t+1}(\mathbf{n})|o_{1:t}) \times \sum_{m=1}^{M(\xi)} \left(\sum_{w=1}^{C} d_t^{\xi,w} P(v_t^{\xi}(m)|v_{t+1}^{w}(n_w), o_{1:t}) - P(v_t^{\xi}(m)|v_{t+1}(\mathbf{n}), o_{1:t}) \right)^2. \quad (21)$$

The computational complexity of $\mathbf{SED}(d_t^{\xi,1},\ldots,d_t^{\xi,C})$ is $\mathcal{O}(N^2C)$ due to the marginalization of joint probabilities. Fortunately, the computational complexity of $\frac{\partial \mathbf{SED}(d_t^{\xi,1},\ldots,d_t^{\xi,C})}{\partial d_t^{\xi,c}}$ is $\mathcal{O}(N^2)$ which is more straightforward than $\mathbf{SED}(d_t^{\xi,1},\ldots,d_t^{\xi,C})$. So optimum mixing weights are obtained from the below minimization problem

$$\hat{\mathbf{d}}_t^{\xi,1},\ldots,\hat{\mathbf{d}}_t^{\xi,C} = \underset{d_t^{\xi,1},\ldots,d_t^{\xi,C}}{\operatorname{argmin}} \mathbf{SED}(d_t^{\xi,1},\ldots,d_t^{\xi,C}). \quad (22)$$

Partial derivative of the above minimization problem is taken with respect to $d_t^{\xi,c}$ as follows

$$\frac{\partial \mathbf{SED}(d_t^{\xi,1},\ldots,d_t^{\xi,C})}{\partial d_t^{\xi,c}} = 2\sum_{\substack{w=1 \\ w\neq c}}^{C} h_t^{\xi} d_t^{\xi,w} + 2f_t^{\xi,c} d_t^{\xi,c} - 2f_t^{\xi,c}, \quad (23)$$

where, h_t^{ξ} and $f_t^{\xi,c}$ are defined as

$$h_t^{\xi} = \sum_{m=1}^{M(\xi)} (\tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi})\alpha_{t|t-1}^{\xi}(m))^2$$

$$f_t^{\xi,c} = \sum_{n_c=1}^{M(c)} \sum_{m=1}^{M(\xi)} \frac{(\rho_{t+1}^{\xi,c}(m,n_c)\tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi})\alpha_{t|t-1}^{\xi}(m))^2}{\alpha_{t+1|t}^{c}(n_c)}. \quad (24)$$

So, considering all partial derivatives below linear equation system is achieved for minimization of (22) in $\mathcal{O}(N^2\,C)$

$$\mathbf{H}_t^{\xi}\mathbf{d}_t^{\xi} = \mathbf{f}_t^{\xi}$$

$$\mathbf{H}_t^{\xi}(c,w) = \begin{cases} f_t^{\xi,c}, & \text{if } c = w \\ h_t^{\xi}, & \text{otherwise} \end{cases}$$

$$\mathbf{f}_t^{\xi}(c) = f_t^{\xi,c}. \quad (25)$$

Optimum mixing weights of the above linear equation system are achieved easily by $\hat{\mathbf{d}}_t^{\xi} = (\mathbf{H}_t^{\xi})^{-1}\mathbf{f}_t^{\xi}$ in $\mathcal{O}(C^3 + N^2\,C)$, and overall computation for all channels is $\mathcal{O}(C^4 + (NC)^2)$. Luckily, $\mathbf{H}_t^{\xi}$ has a particular structure (be constant at non-diagonal elements), and the Sherman-Morrison formula can be applied in the computation of $(\mathbf{H}_t^{\xi})^{-1}$. Consequently $\hat{\mathbf{d}}_t^{\xi}$ can be achieved for all channels in $\mathcal{O}((NC)^2)$ as follows

$$\hat{\mathbf{d}}_t^{\xi,c} = \frac{f_t^{\xi,c} - g_t^{\xi}}{f_t^{\xi,c} - h_t^{\xi}}$$

$$g_t^{\xi} = \frac{h_t^{\xi} \sum_{c=1}^{C} \frac{f_t^{\xi,c}}{f_t^{\xi,c} - h_t^{\xi}}}{1 + h_t^{\xi} \sum_{c=1}^{C} \frac{1}{f_t^{\xi,c} - h_t^{\xi}}}. \quad (26)$$

Here, the recursions of proposed forward, backward and one-slice parameters are completed, and pseudocode of the algorithm is summarized in Appendix B.

D. Evaluation Problem

Evaluation problem is defined as given an observed sequence $o_{1:T}$ and the model parameters λ, how do we compute the probability that the model λ produced the observed sequence [39]. This problem is almost solved using forward and backward parameters. Because of primary definitions of proposed forward and backward parameters are similar to [8], the evaluation problem's solution will also be achieved in the same way as [8].

Firstly, joint distribution $f(o_{1:T}|\lambda)$ is described according to conditional probabilities using the chain rule as follows

$$f(o_{1:T}|\lambda) = f(o_1|\lambda)\prod_{t=2}^{T} f(o_t|o_{1:t-1},\lambda). \quad (27)$$

Substituting (13) in (27), the following relation is obtained

$$f(o_{1:T}|\lambda) = \prod_{c=1}^{C} f(o_1^c|\lambda) \times \prod_{t=2}^{T}\prod_{c=1}^{C} f(o_t^c|o_{1:t-1},\lambda). \quad (28)$$

Consequently, $f(o_t^c|o_{1:t-1},\lambda)$ can be calculated using proposed forward parameter according to

$$f(o_t^c|o_{1:t-1},\lambda) = \sum_{n_c=1}^{M(c)} f(o_t^c, v_t^c(n_c)|o_{1:t-1},\lambda) = \sum_{n_c=1}^{M(c)} b_{n_c}^c(o_t^c)\alpha_{t|t-1}^c(n_c). \quad (29)$$

So, the evaluation problem can be approximately solved in computational complexity $\mathcal{O}(T(NC)^2)$ by the proposed algorithm. In the next, this approximate solution will be compared to the exact solution achieved from the equivalent Cartesian product of the LSIM, with a computational complexity of $\mathcal{O}(TN^{2C})$.

III. Numerical Experiments

In this section, we describe data simulation procedures and performance criteria to evaluate the proposed framework. We can compute exact forward, backward, and one-slice parameters using the equivalent Cartesian product of the LSIM with a computational complexity of $\mathcal{O}(TN^{2C})$ [15], [39]. Also, the evaluation problem can be solved exactly by the same computational complexity.

A. Data Simulation

To generate simulated time-series, firstly, model parameters $\lambda = \{\pi, A, \Theta, \omega, \mu, \Sigma\}$ must be appropriately initialized. In this study, random parameter initialization is applied to cover various scenarios as below.

We consider an LSIM with C channels where each channel has a random state number between 2 to 6. Initial state probabilities $\pi^c = \{\pi^c_1, \ldots, \pi^c_{M(c)}\}$ are drawn from uniform distribution $\mathcal{U}(0,1)$, and then normalized by dividing to the sum of them. In the same way, each row of transition matrices $a^{c,\xi}_{m,n}$ is also initialized and normalized. The observation dimension of each channel also initialized randomly between 1 to 5. Emission probabilities belong to Gaussian distribution. Distribution on $\mathcal{N}(m,1)$ initializes elements of mean vector $\mu^c_{m,k}$. For simplicity, covariance matrices of emission probabilities are assumed to be diagonal, and $\mathcal{U}(1,5)$ initializes their elements.

Coupling weights $\theta^{c,\xi}$ are also initialized similarly to transition matrices $a^{c,\xi}_{m,n}$. Regarding real datasets, structures of the coupling matrix are often similar to sparse matrices [21], [27]. So, we consider an additional step for coupling weights to cover structures more complexity. The coupling weights of randomly selected channels were multiplied by 0.01, and the coupling matrix was renormalized again to have a new coupling structure with channels having a negligible influence on the system.

B. Performance Criteria

Despite approximate or exact inference, marginal forward and one-slice parameters are discrete distributions. So, statistical distances are an appropriate criterion to measure deviance between approximate and exact inferences and determine the quality of the proposed algorithm. We select Hellinger distance because it is a normalized distance whose maximum value is one. It belongs to the f-divergence type of statistical distance [40] and is defined as (P is the exact distribution)

$$D_H(P\|Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{n=1}^{N}(\sqrt{P(n)} - \sqrt{Q(n)})^2}. \tag{30}$$

C. Multi-Channel Time-Series Classification

Applications of the evaluation problem are often in multi-channel time-series classification. So, a classification task is considered to compare results of both exact and proposed approximate solutions. Assuming there are two sets of parameters λ_1 and λ_2 initialized independently. Both models generate their multi-channel time-series with duration T, and $o^{\lambda_1}_{1:T}$ and $o^{\lambda_2}_{1:T}$ indicate them. Each multi-channel time-series is classified according to its log-likelihoods (ll_λ) condition on both models. The log-likelihood of observation is the logarithm of its probability, obtained by evaluation problem. Proposed approximate and exact solutions are used to compute the log-likelihood of observation, and if the proposed approximate solution in (28) is acceptable, then it is expected to have almost equal classification accuracy for both algorithms.

TABLE I
SPECIFICATION OF DERIVED IEEG FOR SELECTED SUBJECTS

Subjects	1	2	3	4	5	6	7	8	9	10
Channels	33	31	14	34	118	74	57	43	57	57
Trials	111	113	115	118	120	108	100	90	120	116

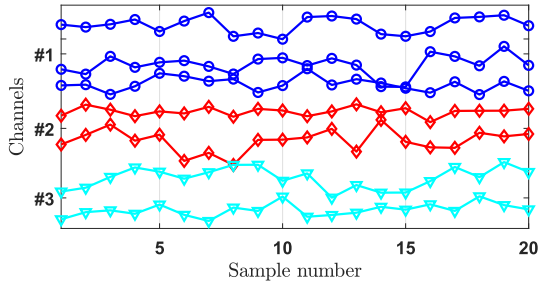


Fig. 1. A multi-channel time-series generated by an LSIM with three channels.

D. Intracranial EEG Modelling

Finally, LSIMs are applied to intracranial EEG (iEEG) using an existing learning algorithm [21], [27]. Learning algorithm formulas are based on the marginal forward, backward and one-slice parameters and reestimate LSIM parameters. iEEG parameters are learned by both proposed forward and backward parameters and the SVI algorithm. Then, the log-likelihood convergence path of both learning algorithms is considered as the final criterion to indicate the superiority of the proposed backward definition.

The selected dataset includes iEEG recordings of medial temporal, lateral frontal, and orbitofrontal regions in 10 human adults completing 120 trials of a visuospatial working memory task [41]. Participants were epileptic patients with channels in frontal and medial temporal lobes. Primary (filtered) and derived (fully preprocessed) iEEG data and analysis scripts are described in detail at [42] and are available online (http://dx.doi.org/10.6080/K0VX0DQD)

Herein, derived iEEG of all subjects are used, having a different number of channels and recording trials (see Table I). An FIR bandpass filter extracts the theta band (3-7 Hz) of derived iEEG (same as [42]), and resulted signals are downsampled to 20 Hz. LSIM parameters of resulted iEEG are learned using an existing learning algorithm for each subject [27].

IV. RESULTS AND DISCUSSION

In this section, we report Hellinger distances, classification, and modeling on simulated time-series and real iEEG using exact, proposed, and other existing algorithms.

In the first simulation scenario, the channels varied from 2 to 6, LSIM parameters were reinitialized 10 000 times to generate a duration of 20 samples. Fig. 1 shows a simulated multi-channel time-series from an LSIM.

The Hellinger distances were calculated between exact and proposed marginal forward (one-slice) parameters for all time points and channels. So, there are a sufficient number of Hellinger distances (more than 400 000), and the histogram of them seems valid to determine their distributions. As can be seen

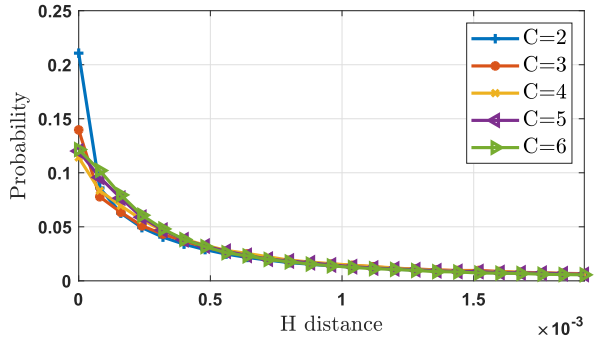


Fig. 2. Histogram of Hellinger distances between exact and proposed forward parameter.

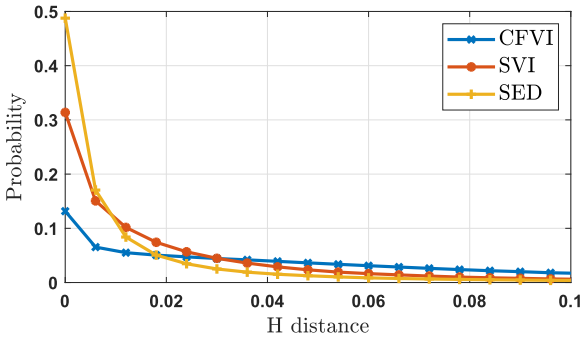


Fig. 3. Histogram of Hellinger distances between exact and approximate one-slice parameters for the six-channel simulated dataset.

TABLE II
AVERAGE OF THE ONE-SLICE PARAMETER HELLINGER DISTANCES FOR DIFFERENT ALGORITHMS

Algs \ C	2	3	4	5	6
SED	**0.021**	**0.024**	**0.024**	**0.023**	**0.022**
SVI	0.029	0.033	0.035	0.035	0.034
CFVI	0.101	0.102	0.099	0.095	0.091

in Fig. 2, Hellinger distances have small values, and the proposed marginal forward parameter seems acceptable. These results also indicate that the proposed marginal forward parameter error is not sensitive to the number of channels and hidden state cardinality.

Hellinger distances were also calculated between exact and proposed marginal one-slice parameters (SED algorithm) to evaluate the proposed marginal backward parameter. The proposed marginal one-slice is the same as the definitions of two existing SVI and CFVI algorithms. So, the Hellinger distances of these algorithms were also compared with the results of the SED algorithm. The marginal one-slice parameter has higher Hellinger distances than the marginal forward parameter due to the multiplication error of marginal forward and backward parameters. As shown in Fig. 3 (SED curve), Hellinger distance criteria increased about 10-times due to multiplication error. This figure also provides a comparison of SED with SVI and CFVI. The histogram of SED has higher values around zero, showing the better performance of SED against SVI and CFVI. Table II presents average values of Hellinger distances,

TABLE III
PERCENTAGE OF CLASSIFICATION ACCURACY FOR BOTH EXACT (EX) AND APPROXIMATE (AP) ALGORITHMS

	$T=2$		$T=3$		$T=4$		$T=5$	
C	EX	AP	EX	AP	EX	AP	EX	AP
2	88.7	88.7	92.6	92.7	94.8	94.7	96.3	96.3
3	93.4	93.4	96.5	96.4	97.8	97.8	98.5	98.4
4	96.2	96.3	98.2	98.2	99.0	98.9	99.4	99.4
5	97.6	97.6	99.0	98.9	99.6	99.6	99.8	99.8
6	98.5	98.7	99.5	99.6	99.8	99.8	99.9	99.9

TABLE IV
PERCENTAGE OF CLASSIFICATION ACCURACY OF CO-LABELING BETWEEN THE EXACT AND APPROXIMATE ALGORITHMS

C \ T	2	3	4	5
2	99.7	99.7	99.8	99.8
3	99.6	99.8	99.9	99.9
4	99.5	99.7	99.8	99.9
5	99.4	99.7	99.9	99.9
6	99.3	99.8	99.9	99.9

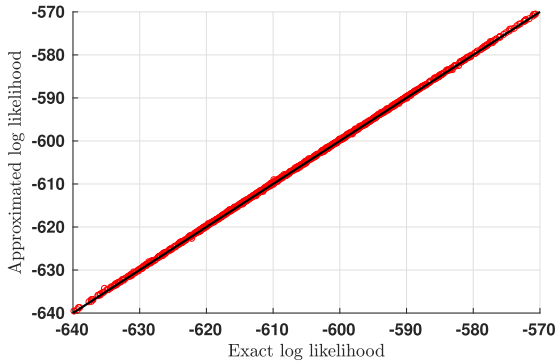


Fig. 4. Scatter plot of exact and approximate log-likelihoods ($C=3$).

confirming that SED has less average Hellinger distance for all number of channels.

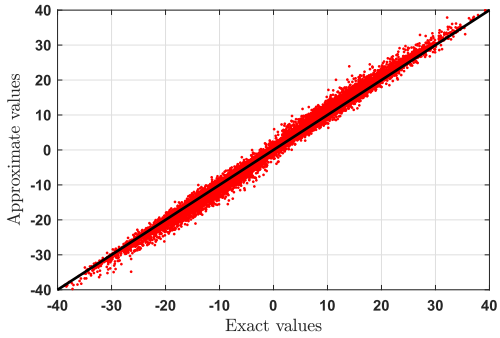
Additionally, exact and approximate log-likelihoods were calculated for each simulated multi-channel time-series to investigate the proposed solution's accuracy for the evaluation problem. So, there are 10 000 exact and approximate log-likelihoods for a specific C considering different structures. The scatter plot of exact and approximate log-likelihoods (Fig. 4) substantially matches the identity function (black line), which is also valid for other C, and correlation coefficients of exact and approximate log-likelihoods are very close to one in all cases. Thus, the proposed solution to the evaluation problem looks adequate.

A classification problem considers more details about the proposed solution, and classifying a multi-channel time-series is one of the most common areas of application for the evaluation problem. In a classification problem there are two LSIMs with λ_1 and λ_2 parameters, and a multi-channel time-series must be assigned to one of them. So, two log-likelihoods are computed conditioned on both LSIMs parameters λ_1 and λ_2, as noted by ll_{λ_1} and ll_{λ_2}. If $d_{\lambda_1,\lambda_2} = ll_{\lambda_1} - ll_{\lambda_2} > 0$, then multi-channel time-series is assigned to model λ_1 and vice versa. We computed d_{λ_1,λ_2} by exact and proposed solution for different values of C

TABLE V
LOG-LIKELIHOOD VALUES AT ITERATION OF 1000 ($d = ll_{SED} - ll_{SVI}$)

Subjects	1	2	3	4	5	6	7	8	9	10
ll_{SED}	-968037	-757545	-226634	-1091790	-4422301	-2545309	-1349206	-766910	-1462599	-2070715
ll_{SVI}	-969001	-758367	-233295	-1096344	-4423051	-2546818	-1350387	-774702	-1466722	-2066046
d	963	819	6662	4554	750	1509	1181	7792	4124	4669

(a) Correlation coefficient =0.997 ($C = 6$ and $T = 2$)



(b) Correlation coefficient =0.999 ($C = 6$ and $T = 5$)

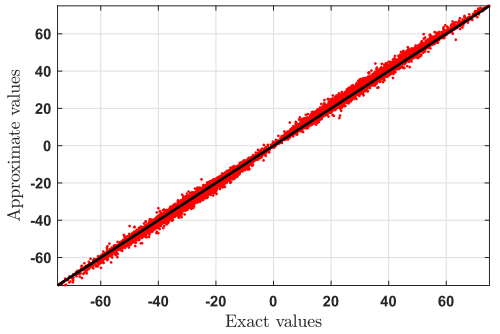


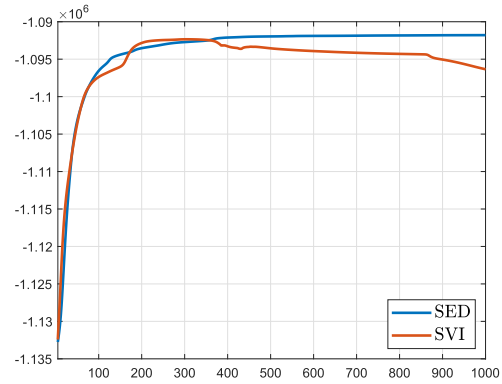
Fig. 5. Scatter plot between exact and approximate d_{λ_1,λ_2}.

($2 \leq C \leq 6$) and T ($2 \leq T \leq 5$), where the LSIM parameters were reinitialized 10 000 times for each C and T.

Table III presents classification accuracies for different values of C and T. As can be seen, both solutions have almost similar accuracies in different situations. So, the proposed solution is an acceptable approximation of the exact solution. Moreover, accuracy was improved by increasing C or T, which is equivalent to an increase in the given information.

Besides, the co-labeling of both solutions was considered to verify the proposed solution in more detail. Here, predicted labels of the exact solution were considered as new true-labels, and a new accuracy is calculated between these new true-labels and output labels of the proposed solution. This new accuracy shows the percentage of communal labels of both solutions. So, high accuracy values indicate that the proposed solution works the same as the exact solution. Table IV shows accuracies of co-labeling in different situations. As seen, increasing sequence duration (T) leads to better accuracy, whereas increasing channels, (C) decrease co-labeling accuracy. However, in general, the increasing effect of T wins decreasing effect of C, and the co-labeling accuracy is above 99.7 for $T \geq 5$. Fig. 5 depicts the scatter plots of exact and approximate d_{λ_1,λ_2} for $C = 6$ with duration $T = 2$ and $T = 5$. These scatter plots indicate that both exact and approximate d_{λ_1,λ_2} get closer to the identity function by increasing T, then correlation coefficients increase from 0.997 to 0.999.

(a) Log-likelihood convergence path at iterations 1 to 1000



(b) Log-likelihood convergence path at iterations 150 to 1000

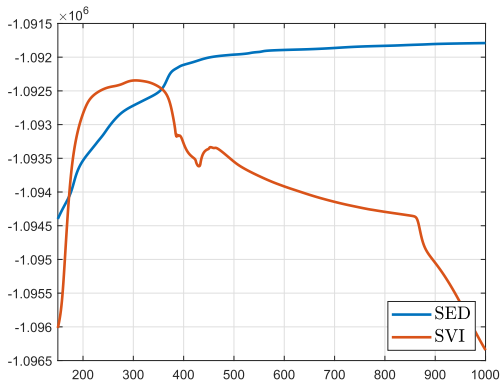


Fig. 6. Log-likelihood convergence path of SVI and SED-based learning algorithms for subject 4.

Finally, we applied the LSIMs learning algorithm to real iEEG data with different channels from 14 to 118 based on both SED and SVI algorithms. The learning algorithm performs the estimation of LSIM parameters until 1000 iterations for all subjects. Table V provides a comprehensive performance comparison of log-likelihood values at the last iteration. As can be seen, the SED-based algorithm had better log-likelihood for all subjects. Besides, the log-likelihood convergence path is plotted in Fig. 6 for subject 4, and subject 3 and subject 8 also have similar results. These plots show that the SED algorithm has monotonically increased of log-likelihood while the SVI algorithm has some decreasing intervals in its convergence path. Therefore, the SED algorithm not only converges to a model with higher likelihood but also conserves monotonically increasing compared to the SVI algorithm.

V. Conclusion

This study assumes the state space interaction of CHMMs as the influence model called as LSIMs. A new recursive formulation is proposed to compute marginal forward and backward parameters and the evaluation problem by $\mathcal{O}(T(NC)^2)$ instead of $\mathcal{O}(TN^{2C})$. This formulation is derived systematically and constructively step by step using the divide-and-conquer paradigm. The main improvement of the proposed approximate inference over existing algorithms comes from the contributions in the marginal backward parameter's derivation.

Hellinger distances indicate that the proposed parameters have smaller distances than existing algorithms of previous studies. Besides, the histogram of Hellinger distances expressed that the proposed parameters are very close to exact values. Results are valid concerning the number of channels and various model structures.

Furthermore, multi-channel time-series classification is performed by both proposed and exact solutions of the evaluation problem and achieved accuracies are almost equal under different conditions. Co-labeling is also investigated between proposed and exact solutions, and co-labeling accuracies indicate that increasing T causes the creation of more similar output labels between proposed and exact solutions. Modeling real iEEG data shows that the proposed approximate inference has monotonically increasing convergence, and it converges to a higher likelihood model compared to SVI.

LSIMs are new and original models that have an interpretable and linear wise structure in state-space. For example, coupling weights indicate relationships and connectivity between channels. However, this linear structure may degrade the power of LSIM compared to CHMMs, but it makes LSIMs practical in multi-channel datasets. We believe that LSIMs could emerge in various fields if its inference and learning problem are solved precisely.

The proposed inference algorithm is essential in solving the evaluation and learning problems. So, it is suggested to conduct future studies to focus on developing a novel framework to solve the LSIM learning problem based on proposed tractable forward and backward parameters. If the learning problem is solved accurately, and proof of convergence is presented like the HMM learning problem, this framework can be applied to various real datasets in different areas.

Appendix A
Conditional Independence of Future Observations Given Next Hidden States

An underlying property of Markov chains is that given present hidden state, present and future observations are independent of past. Considering this property, we prove that the probability of current hidden state of channel ξ is independent of future observations given next hidden states of all channels. Using Bayes' rule, this probability is rewritten as follows

$$P(v_t^{\xi}(m)|v_{t+1}(\mathbf{n}), o_{1:t}, o_{t+1:T}) = \frac{f(o_{t+1:T}|v_{t+1}(\mathbf{n}), o_{1:t}, v_t^{\xi}(m))}{f(o_{t+1:T}|v_{t+1}(\mathbf{n}), o_{1:t})} \times P(v_t^{\xi}(m)|v_{t+1}(\mathbf{n}), o_{1:t}). \quad (31)$$

Using Markov property, the numerator of (31) is simplified as follows

$$f(o_{t+1:T}|v_{t+1}(\mathbf{n}), o_{1:t}, v_t^{\xi}(m)) = f(o_{t+1:T}|v_{t+1}(\mathbf{n})). \quad (32)$$

In the same way, the denominator of (31) is also reduced to

$$f(o_{t+1:T}|v_{t+1}(\mathbf{n}), o_{1:t}) = f(o_{t+1:T}|v_{t+1}(\mathbf{n})). \quad (33)$$

Substituting (32) and (33) in (31), the proof is completed as follows

$$P(v_t^{\xi}(m)|v_{t+1}(\mathbf{n}), o_{1:t}, o_{t+1:T}) = P(v_t^{\xi}(m)|v_{t+1}(\mathbf{n}), o_{1:t}). \quad (34)$$

Appendix B
Proposed Forward-Backward Algorithm

The approximate forward and backward parameters are computed based on the following recursive procedures in Algorithm 1.

Algorithm 1: Forward-Backward Algorithm.

Require: $\lambda = \{\pi, A, \Theta, \omega, \mu, \Sigma\}$, $o_{1:T}$
• The forward recursion
$t \leftarrow 1$
for $\xi = 1 : C$ **do**
  for $m = 1 : M(\xi)$ **do**
    $\alpha_{t|t-1}^{\xi}(m) = \pi_m^{\xi}$
  end for
  for $m = 1 : M(\xi)$ **do**
    $\tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi}) = \frac{b_m^{\xi}(o_t^{\xi})}{\sum_{n_\xi=1}^{M(\xi)} b_{n_\xi}^{c}(o_t^{\xi})\alpha_{t|t-1}^{\xi}(n_\xi)}$
  end for
end for
for $t = 2 : T$ **do**

  for $\xi = 1 : C$ **do**
    for $m = 1 : M(\xi)$ **do**
      $\alpha_{t|t-1}^{\xi}(m) = \sum_{c=1}^{C} \theta^{c,\xi} \sum_{n_c=1}^{M(c)} a_{m,n_c}^{c,\xi} \alpha_{t-1|t-2}^{c}(n_c)\tilde{\mathbf{b}}_{n_c}^{c}(o_{t-1}^{c})$
    end for
    for $m = 1 : M(\xi)$ **do**
      $\tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi}) = \frac{b_m^{\xi}(o_t^{\xi})}{\sum_{n_\xi=1}^{M(\xi)} b_{n_\xi}^{\xi}(o_t^{\xi})\alpha_{t|t-1}^{\xi}(n_\xi)}$
    end for
  end for
  for $\xi, w = \{1 : C\} \times \{1 : C\}$ **do**
    for $m, n_w = \{1 : M(\xi)\} \times \{1 : M(w)\}$ **do**
      $\rho_t^{\xi,w}(m, n_w) = \alpha_{t|t-1}^{w}(n_w) + \theta^{\xi,w}(a_{m,n_w}^{\xi,w} - \sum_{m_\xi=1}^{M(\xi)} a_{m_\xi,n_w}^{\xi,w}\alpha_{t-1|t-2}^{\xi}(m_\xi)\tilde{\mathbf{b}}_m^{\xi}(o_{t-1}^{\xi}))$
    end for
    $f_t^{\xi,w} = \sum_{n_w=1}^{M(w)} \sum_{m=1}^{M(\xi)} \frac{(\rho_{t+1}^{\xi,w}(m,n_c)\tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi})\alpha_{t|t-1}^{\xi}(m))^2}{\alpha_{t+1|t}^{w}(n_w)}$
    $h_t^{\xi} = \sum_{m=1}^{M(\xi)} (\tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi})\alpha_{t|t-1}^{\xi}(m))^2$
  end for
end for
• The backward recursion
$t \leftarrow T$
for $\xi = 1 : C$ **do**
  for $m = 1 : N_\xi$ **do**
    $\beta_t^{\xi}(m) = \tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi})$
    $\alpha_{t|T}^{\xi}(m) = \alpha_{t|t-1}^{\xi}(m)\beta_t^{\xi}(m)$
  end for
end for

for $t = T-1:-1:1$ **do**
  for $\xi = 1:C$ **do**

$$\hat{\mathbf{d}}_t^{\xi,c} = \frac{f_t^{\xi,c} - g_t^{\xi}}{f_t^{\xi,c} - h_t^{\xi}},\ g_t^{\xi} = \frac{h_t^{\xi} \sum_{c=1}^{C} \frac{f_t^{\xi,c}}{f_t^{\xi,c} - h_t^{\xi}}}{1 + h_t^{\xi} \sum_{c=1}^{C} \frac{1}{f_t^{\xi,c} - h_t^{\xi}}}$$

    for $m = 1:M(\xi)$ **do**

$$\beta_t^{\xi}(m) = \tilde{\mathbf{b}}_m^{\xi}(o_t^{\xi}) \sum_{w=1}^{C} \hat{\mathbf{d}}_t^{\xi,w} \sum_{n_w=1}^{M(w)} \rho_{t+1}^{\xi,w}(m, n_w) \beta_{t+1}^{w}(n_w)$$

$$\alpha_{t|T}^{\xi}(m) = \alpha_{t|t-1}^{\xi}(m) \beta_t^{\xi}(m)$$

$$\beta_t^{\xi}(m) = \frac{\beta_t^{\xi}(m)}{\sum_{n_\xi=1}^{M(\xi)} \alpha_{t|T}^{\xi}(n_\xi)},\ \alpha_{t|T}^{\xi}(m) = \frac{\alpha_{t|T}^{\xi}(m)}{\sum_{n_\xi=1}^{M(\xi)} \alpha_{t|T}^{\xi}(n_\xi)}$$

    end for
  end for
end for
return $\alpha_{t|t-1}^{\xi}(m), \beta_t^{\xi}(m), \alpha_{t|T}^{\xi}(m)$

REFERENCES

[1] M. Brand, "Coupled hidden Markov models for modeling interacting processes," MIT Media Lab Perceptual Computing/Learning and Common Sense, Tech. Rep. 405, 1997.

[2] K. P. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. thesis, Univ. California, Berkeley, CA, USA, 2002.

[3] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 2, 2002, pp. II–2013.

[4] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 863–876, May 2015.

[5] T. Bolton and D. Van De Ville, "Sparse coupled hidden Markov models shed light on resting-state FMRI cross-network interactions," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, pp. 358–361, 2017.

[6] R. Zhao, G. Schalk, and Q. Ji, "Coupled hidden Markov model for electrocorticographic signal classification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 1858–1862,.

[7] S. Zhong and J. Ghosh, "HMMs and coupled HMMs for multi-channel EEG classification," in *Proc. Neural Netw., 2002. IJCNN'02. Proc. Int. Joint Conf.*, 2002, vol. 2, pp. 1154–1159.

[8] N. M. Ghahjaverestan *et al.*, "Coupled hidden Markov model-based method for apnea bradycardia detection," *IEEE J. Biomed. Health Inf.*, vol. 20, no. 2, pp. 527–538, Mar. 2016.

[9] C. Sherlock, T. Xifara, S. Telfer, and M. Begon, "A coupled hidden Markov model for disease interactions," *J. Roy. Stat. Soc.: Ser. C (Appl. Statist.)*, vol. 62, no. 4, pp. 609–627, 2013.

[10] J. Kwon and K. Murphy, "Modeling freeway traffic with coupled HMMs," Tech. Rep., Univ. California, Berkeley, 2000.

[11] Y. Qi and S. Ishak, "A hidden Markov model for short term prediction of traffic conditions on freeways," *Trans. Res. Part C: Emerg. Technol.*, vol. 43, no. 0968-090X, pp. 95–111, 2014.

[12] W. Cao, L. Cao, and Y. Song, "Coupled market behavior based financial crisis detection," in *Proc. Neural Netw. (IJCNN), Int. Joint Conf.*, 2013, pp. 1–8.

[13] L. K. Saul and M. I. Jordan, "Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Mach. Learn.*, vol. 37, no. 1, pp. 75–87, 1999.

[14] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Proc. Comput. Vis. Pattern Recognit., 1997. Proc., IEEE Comput. Soc. Conf.*, 1997, pp. 994–999.

[15] S. Zhong and J. Ghosh, "A new formulation of coupled hidden Markov models," Dept. Elect. Comput. Eng., Univ. Austin, Austin, TX, USA, 2001.

[16] C. Asavathiratham and G. C. Verghese, "The influence model: A tractable representation for the dynamics of networked Markov chains," Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2001.

[17] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[18] A. E. Raftery, "A model for high-order Markov chains," *J. Roy. Stat. Soc. Ser. B (Methodological)*, vol. 47, no. 3, pp. 528–539, 1985.

[19] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Comput. Speech Lang.*, vol. 8, no. 1, pp. 1–38, 1994.

[20] W. Dong *et al.*, "Influence modeling of complex stochastic processes," Master's thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2006.

[21] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proc. 9th Int. Conf. Multimodal Interfaces*, ACM, 2007 pp. 271–278.

[22] W. Dong, A. Mani, A. Pentland, B. Lepri, and F. Pianesi, "Modeling group discussion dynamics," *IEEE Trans. Auton. Mental Dev.*, to be published.

[23] W. Pan, W. Dong, M. Cebrian, T. Kim, J. H. Fowler, and A. S. Pentland, "Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems," *IEEE Signal Process. Mag.*, vol. 29, no. 2, pp. 77–86, Mar. 2012.

[24] I. Rezek, P. Sykacek, and S. J. Roberts, "Learning interaction dynamics with coupled hidden Markov models," *IEEE Proc.-Sci., Meas. Technol.*, vol. 147, no. 6, pp. 345–350, Nov. 2000.

[25] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 33–42.

[26] X. Boyen and D. Koller, "Approximate learning of dynamic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 396–402.

[27] W. Dong and A. Pentland, "Modeling influence between experts," in *Artif. Intell. Human Comput.*, vol. 4451, pp. 170–189, 2007.

[28] W. Dong *et al.*, "Modeling the structure of collective intelligence," Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2010.

[29] S.-Z. Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1947–1951, May 2006.

[30] K. Murphy and Y. Weiss, "The factored frontier algorithm for approximate inference in dbns," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 378–385.

[31] T. Heskes and O. Zoeter, "Expectation propagation for approximate inference in dynamic Bayesian networks," in *Proc. 18th Conf. Uncertainty Artif. Intell.*, 2002, pp. 216–223.

[32] T. T. Georgiou and A. Lindquist, "Kullback-Leibler approximation of spectral density functions," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 2910–2917, Nov. 2003.

[33] A. F. García-Fernández and B.-N. Vo, "Derivation of the phd and cphd filters based on direct Kullback–Leibler divergence minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5812–5820, Nov. 2015.

[34] J. E. Darling and K. J. DeMars, "Minimization of the Kullback–Leibler divergence for nonlinear estimation," *J. Guid., Control, Dyn.*, vol. 40, no. 7, pp. 1739–1748, 2017.

[35] D. V. Lindley, "On a measure of the information provided by an experiment," *Ann. Math. Statist.*, vol. 27, no. 4, pp. 986–1005, 1956.

[36] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Stat. Soc. Ser. B (Methodological)*, vol. 41, no. 2, pp. 113–147, 1979.

[37] D. Sun and J. O. Berger, "Reference priors with partial information," *Biometrika*, vol. 85, no. 1, pp. 55–71, 1998.

[38] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Oct. 2005.

[39] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[40] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Stat. Soc. Ser. B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[41] E. Johnson, "Intracranial EEG recordings of medial temporal, lateral frontal, and orbitofrontal regions in 10 human adults performing a visuospatial working memory task. crcns.org," 2018. [Online]. Available: https://dx.doi.org/10.6080/K0VX0DQD

[42] E. L. Johnson *et al.*, "Dynamic frontotemporal systems process space and time in working memory," *PLoS Bio.*, vol. 16, no. 3, 2018, Art. no. e2004274.