Clustering Hidden Markov Models with Variational Bayesian Hierarchical EM

Hui Lan, Ziquan Liu, Janet H. Hsiao, Dan Yu, Antoni B. Chan

***Abstract*—The hidden Markov model (HMM) is a broadly applied generative model for representing time series data, and clustering HMMs attracts increased interests from machine learning researchers. However, the number of clusters (K) and the number of hidden states (S) for cluster centers are still difficult to determine. In this paper, we propose a novel HMM-based clustering algorithm, the variational Bayesian hierarchical EM algorithm, which clusters HMMs through their densities and priors, and simultaneously learns posteriors for the novel HMM cluster centers that compactly represent the structure of each cluster. The numbers K and S are automatically determined in two ways. First, we place a prior on the pair (K, S) and approximate their posterior probabilities, from which the values with the maximum posterior are selected. Second, some clusters and states are pruned out implicitly when no data samples are assigned to them, thereby leading to automatic selection of the model complexity. Experiments on synthetic and real data demonstrate that our algorithm performs better than using model selection techniques with maximum likelihood estimation.**

***Index Terms*—Variational Bayesian, Hidden Markov mixture model, Clustering, Hierarchical EM**

I. Introduction

THE hidden Markov model (HMM) [1] is an effective method for statistically representing time series data, assuming that each observation in a sequence is generated conditioned on a discrete state of a hidden Markov chain, i.e., a hidden state sequence. HMM has been popularly applied in many areas that need to analyze time series data, such as speech recognition [2, 3], cognitive science [4, 5], and bioinformatics [6, 7]. Although neural networks (NN) [8, 9] and reinforcement learning [10] are also works in these areas, they typically require large datasets to prevent over-fitting and learn models that are difficult to interpret. In contrast, as a generative probabilistic model, HMMs work well on smaller datasets with Bayesian estimation preventing over-fitting, while also being interpretable models.

Clustering HMMs to explore the hidden cluster structure can be an effective method for discovering commonalities and differences among HMMs, and the cluster center serves as a representation of the HMMs in each cluster. In particular, recent works represent an individual's eye gaze pattern by estimating an HMM from their eye fixation sequences on a stimuli, and then clusters the individual HMMs into groups to discover common eye gaze strategies, represented by the HMM cluster centers. In the context of eye gaze, the hidden state of the HMM corresponds to a region-of-interest (ROI) on the stimuli, with the extent determined by the emission density, and the state transition matrix of the HMM contains the probabilities of viewing the next ROI after viewing the current ROI. This bottom-up clustering method has enabled interesting discoveries about the role of eye gaze in cognitive processes, including optimal strategies for face recognition [11–14], masking effects in visual search [15], and the association of eye gaze patterns to cognitive decline [5], emotion recognition [16, 17], scene perception [18], chronic pain [19, 20], and decision making [21].

In these previous works, the individual HMMs are estimated using a variational Bayesian method, which automatically determines the number of hidden states (i.e., number of ROIs) as well as other hyperparameters. However, in contrast, the clustering of HMMs is performed with a predefined number of clusters and states, which is manually set by the experimenters. Manual selection of the model hyperparameters could introduce experimenter bias into the data analysis, and thus a data-driven approach for automatically selecting the model hyperparameters is preferred. In this paper, we propose a Bayesian method for clustering HMMs that automatically estimates the number of clusters (i.e., number of gaze strategies) and number of states (i.e., number of ROIs).

Note that clustering HMMs is not the same as clustering time-series data with HMMs. The latter aims to form K groups from N observation sequences, with each group modeled by one HMM. For example, [22] proposed a Dirichlet process for learning an HMM mixture from music clips to represent a song, while [23] clusters sequences by forming a single HMM with a block-diagonal transition matrix and then training on all sequences with the Baum-Welch [24] algorithm. In contrast, clustering HMMs aims to form K groups of N HMMs, with each group represented by one HMM. Clustering HMMs is equivalent to building a large mixture model of HMMs, and then reducing it into a mixture model with fewer components and states, which can concisely represent the original HMMs. To this end, [25] proposes a variational hierarchical EM (VHEM) algorithm to cluster HMMs directly using their probability densities of the observation sequence, by minimizing the Kullback-Leibler divergence (KLD) [26] between the input HMMs and cluster center HMMs. In summary, clustering time series data with HMMs is a process mapping data to model, while clustering HMMs maps from model to model. Clustering HMMs is preferred in the above cognitive science works because it allows modeling and analysis of both individual

Hui Lan is with the School of Statistics and Data Science, Faculty of Science, Beijing University of Technology, Beijing, China.

Hui Lan, Ziquan Liu and Antoni Chan are with the Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China.

Janet H. Hsiao is with the Department of Psychology, The University of Hong Kong, Hong Kong SAR, China.

Dan Yu is with the Institute of Systems Science, Academy of Mathematics and System Science, Chinese Academy of Sciences, China.

differences and group similarities.

Our Contributions. In this paper, we propose a variational hierarchical EM algorithm that clusters HMMs within a Bayesian framework (VBHEM-H3M). VBHEM automatically determines the number of clusters and the number of hidden states, clusters HMMs directly, and estimates novel HMM cluster centers. Compared with VHEM, firstly, VBHEM automatically performs model selection (i.e., estimates the model hyperparameters, e.g., the number of clusters), while VHEM cannot. Automatic model selection is important for analyzing experiment data, since it removes the experimenter bias caused when manually selecting the model hyperparameters. Secondly, each input HMM of VBHEM is represented with a prior distribution over its parameters, which better incorporates the uncertainty of parameter estimation of the input HMMs as they are estimated from limited data samples. In contrast, this uncertainty information about the input HMMs is discarded by VHEM. In experiment, we demonstrate that using the input HMM uncertainty leads to better cluster performance; Thirdly, VBHEM computes a posterior distribution over the HMM cluster centers (c.f., a point-estimate obtained by VHEM), which gives a better characterization of the uncertainty in the estimated model. More detailed comparisons can be found in Sec. IV-D. Finally, we give complete derivations of Bayesian HEM algorithm for H3M with its prior distributions, which could be generalized to other mixture of mixture models.

The remainder of this paper is organized as follows. In Sec. II, we provide the related works. In Sec. III we review the necessary related knowledge. In Sec. IV we introduce a new objective function and derive the VBHEM algorithm. Sec. V presents experimental results obtained by applying VBHEM algorithm to synthetic data and real data. Finally, Sec. VI concludes this paper.

II. Related Work

Clustering HMMs. The existing approaches to cluster HMMs leverage different distances or similarity between two HMMs. [27] clustered HMMs by calculating a probability product kernel (PPK) similarity matrix between all HMMs, and then applying spectral clustering. [25] proposed the variational hierarchical EM (VHEM) algorithm to cluster HMMs directly using their probability densities of the observation sequence and estimate HMM cluster centers, via minimizing the KLD between input HMMs and cluster center HMMs. [28] modeled each gene sequence with an HMM and defined a distance matrix based on likelihood, and applied a hierarchical clustering algorithm to find the best clusters. [29] used a bagging method, where a large set of HMMs is computed from the data, the HMMs are grouped together based on KLD, and the cluster centers are found by averaging HMMs in the same group. For works [28, 29], data is used in the clustering process, while [25, 27] are only based on the input HMMs. [30] proposed a framework, Aggregated Wasserstein, for computing distances between two HMMs with state conditional distributions as Gaussians. However, clustering based on Wasserstein distance has not been studied for HMMs.

For the above HMM clustering methods, the numbers of clusters and states are set manually. Therefore, these methods need to resort to other model selection techniques, such as Akaike information criterion (AIC) [31], Bayesian information criterion (BIC) [32], Monte-Carlo cross-validation [33], and minimum description length (MDL) [34]. In our work, we propose a method that both clusters HMMs directly and automatically performs model selection via Bayesian formulation.

Bayesian Model Selection. The Bayesian view of model comparison involves the use of probabilities to represent uncertainty in the choice of model – the model structure is determined using the posterior distribution over the model structure, conditioned on the training data. Suppose we wish to compare a set of M models m_i where $i \in \{1, \dots, M\}$. We then need to evaluate the posterior distribution $p(m_i|Y) \propto p(m_i)p(Y|m_i)$, where the uncertainty is expressed through a model prior probability distribution $p(m_i)$, and the model evidence $p(Y|m_i)$ expresses the preference shown by the data Y for different models. Variational inference (VI) [35, 36] is widely used to approximate the model evidence for Bayesian models, while an alternative, but computationally expensive, strategy is Markov chain Monte Carlo (MCMC) sampling [37]. VI first posits a family of densities for the approximate posterior distribution, and then finds a member of that family that is closest to the target density, as measured by KLD.

The previous works using VI with Bayesian model selection focus on mixture of experts model [38], HMMs [39], and Dirichlet process mixture models [40]. In most cases, the model evidence $p(Y|m_i)$ is intractable, and thus the evidence lower bound (ELBO) is used as a model selection criterion. VI has been explored for mixture models [41, 42] and more generally [43]. [44] derived VI in the Bayesian framework for hidden Markov mixture models (H3M) [23], but [44] only considered learning H3M from data, not from the HMMs. [45] proposed a VB method for clustering Gaussians, i.e., learning a Gaussian mixture model from a set of input GMMs. However, a VB method for clustering HMMs, by learning an H3M from a set of input HMMs, has not been studied so far.

Other model selection methods. Other clustering methods also focus on automatic selection of the number of cluster centers, but are not proposed for HMMs. [46] proposed an approach based on the idea that cluster centers are characterized by a higher density than their neighbors, and by a relatively large distance from points with higher densities. However, this method needs to draw a rectangle to manually select the cluster centers. [47] proposed a cluster center fast determination (CCFD) algorithm, which overcomes this problem and realizes automatic selection of the cluster centers. CCFD has been applied to image segmentation [48] and hybrid data stream clustering [49].

III. Preliminaries

A. Hidden Markov (Mixture) Model

We first briefly review hidden Markov models (HMMs) and the hidden Markov mixture model (H3M) [23], and define the notation used in the derivation (see summary in Table I, II).

An H3M models a set of observation sequences as samples from a group of K hidden Markov models (HMMs), and is parameterized by $\mathcal{M} = \{\omega_i, \mathcal{M}_i\}_{i=1}^{K}$, where $\mathcal{M}_i$ is the i-th HMM and ω_i is the corresponding mixture component

TABLE I
NOTATIONS USED IN THE DERIVATION OF THE VBHEM-H3M ALGORITHM.

variables	*base model (b)*	*reduced model (r)*
number of HMM components	$K^{(b)}$	$K^{(r)}$
index for HMM components	$i \in \{1,2,...,K^{(b)}\} = [1, K^{(b)}]$	$j \in \{1,2,...,K^{(r)}\}$
number of HMM states	$S^{(b)}$	$S^{(r)}$
index for HMM state at time t	$\beta_t \in \{1,2,...,S^{(b)}\}$	$\rho_t \in \{1,2,...,S^{(r)}\}$
sequence length	τ	τ
HMM state sequence	$\boldsymbol{\beta} = \{\beta_t\}_{t=1}^{\tau}$, $\beta_t \in \{1,2,...,S^{(b)}\}$	$\boldsymbol{\rho} = \{\rho_t\}_{t=1}^{\tau}$, $\rho_t \in \{1,2,...,S^{(r)}\}$
models		
H3M	$\mathcal{M}^{(b)} = \{\boldsymbol{\omega}^{(b)}, \{\mathcal{M}_i^{(b)}\}_{i=1}^{K^{(b)}}\}$	$\mathcal{M}^{(r)} = \{\boldsymbol{\omega}^{(r)}, \{\mathcal{M}_j^{(r)}\}_{j=1}^{K^{(r)}}\}$
HMM component (of H3M)	$\mathcal{M}_i^{(b)} = \{\boldsymbol{\pi}^{(b),i}, \boldsymbol{A}^{(b),i}, \{\boldsymbol{\Theta}_\beta^{(b),i}\}_{\beta=1}^{S^{(b)}}\}$	$\mathcal{M}_j^{(r)} = \{\boldsymbol{\pi}^{(r),j}, \boldsymbol{A}^{(r),j}, \{\boldsymbol{\Theta}_\rho^{(r),j}\}_{\rho=1}^{S^{(r)}}\}$
Gaussian emission	$\boldsymbol{\Theta}_\beta^{(b),i} = \{\boldsymbol{\mu}_\beta^{(b),i}, \boldsymbol{\Lambda}_\beta^{(b),i}\}$	$\boldsymbol{\Theta}_\rho^{(r),j} = \{\boldsymbol{\mu}_\rho^{(r),j}, \boldsymbol{\Lambda}_\rho^{(r),j}\}$
latent variables		
assignment variable	$\boldsymbol{Z}'$	$\boldsymbol{Z}$
hidden state sequence	$\boldsymbol{X}'$	$\boldsymbol{X}$
H3M mixture weights	$\boldsymbol{\omega}^{(b)} = \{\omega_i^{(b)}\}$	$\boldsymbol{\omega}^{(r)} = \{\omega_j^{(r)}\}$
HMM initial state probability	$\boldsymbol{\pi}^{(b),i} = \{\pi_\beta^{(b),i}\}$	$\boldsymbol{\pi}^{(r),j} = \{\pi_\rho^{(r),j}\}$
HMM state transition matrix	$\boldsymbol{A}^{(b),i} = [a_{\beta,\beta'}^{(b),i}]$	$\boldsymbol{A}^{(r),j} = [a_{\rho,\rho'}^{(r),j}]$
Emission mean and covariance	$\{\boldsymbol{\mu}_\beta^{(b),i}, (\boldsymbol{\Lambda}_\beta^{(b),i})^{-1}\}$	$\{\boldsymbol{\mu}_\rho^{(r),j}, (\boldsymbol{\Lambda}_\rho^{(r),j})^{-1}\}$
prior distributions	*base model (b)*	*reduced model (r)*
hyperparameters	$\mathcal{P}^{(b)} = \{\boldsymbol{\alpha}_0^{(b)}, \{\boldsymbol{\eta}_0^{(b),i}, \boldsymbol{\epsilon}_{\beta,0}^{(b),i}, \boldsymbol{\gamma}_{\beta,0}^{(b),i}, \boldsymbol{m}_{\beta,0}^{(b),i}, \boldsymbol{W}_{\beta,0}^{(b),i}, \boldsymbol{\nu}_{\beta,0}^{(b),i}\}_{\beta=1,i=1}^{S^{(b)},K^{(b)}}\}$	$\mathcal{P}^{(r)} = \{\alpha_0^{(r)}, \eta_0^{(r)}, \epsilon_0^{(r)}, \gamma_0^{(r)}, \boldsymbol{m}_0^{(r)}, \boldsymbol{W}_0^{(r)}, \nu_0^{(r)}\}$
$p(\boldsymbol{\omega})$	$\mathrm{Dir}(\boldsymbol{\omega}^{(b)}\|\boldsymbol{\alpha}_0^{(b)})$	$\mathrm{Dir}(\boldsymbol{\omega}^{(r)}\|\alpha_0^{(r)})$
$p(\boldsymbol{\pi})$	$\mathrm{Dir}(\boldsymbol{\pi}^{(b),i}\|\boldsymbol{\eta}_0^{(b),i})$	$\mathrm{Dir}(\boldsymbol{\pi}^{(r),j}\|\eta_0^{(r)})$
$p(\boldsymbol{a})$, $\boldsymbol{a}$ is a row of $\boldsymbol{A}$	$\mathrm{Dir}(\boldsymbol{a}_\beta^{(b),i}\|\boldsymbol{\epsilon}_{\beta,0}^{(b),i})$	$\mathrm{Dir}(\boldsymbol{a}_\rho^{(r),j}\|\boldsymbol{\epsilon}_0^{(r)})$
$p(\boldsymbol{\mu}\|\boldsymbol{\Lambda})$	$\mathcal{N}(\boldsymbol{\mu}_\beta^{(b),i}\|\boldsymbol{m}_{\beta,0}^{(b),i}, (\gamma_{\beta,0}^{(b),i}\boldsymbol{\Lambda}_\beta^{(b),i})^{-1})$	$\mathcal{N}(\boldsymbol{\mu}_\rho^{(r),j}\|\boldsymbol{m}_0^{(r)}, (\gamma_0^{(r)}\boldsymbol{\Lambda}_\rho^{(r),j})^{-1})$
$p(\boldsymbol{\Lambda})$	$\mathcal{W}(\boldsymbol{\Lambda}_\beta^{(b),i}\|\boldsymbol{W}_{\beta,0}^{(b),i}, \nu_{\beta,0}^{(b),i})$	$\mathcal{W}(\boldsymbol{\Lambda}_\rho^{(r),j}\|\boldsymbol{W}_0^{(r)}, \nu_0^{(r)})$

TABLE II
NOTATIONS USED IN THE DERIVATION OF THE VBHEM-H3M ALGORITHM.

probability distributions	*notation*	*short-hand*
HMM state sequence (r)	$p(x = \boldsymbol{\rho}\|z_i = j, \mathcal{M}^{(r)})$	$p(\boldsymbol{\rho}\|\mathcal{M}_j^{(r)}) = \boldsymbol{\pi}_{\boldsymbol{\rho}}^{(r),j}$
HMM observation likelihood (r)	$p(\boldsymbol{y}^i\|z_i = j, \mathcal{M}^{(r)})$	$p(\boldsymbol{y}^i\|\mathcal{M}_j^{(r)})$
Gaussian emission likelihood (r)	$p(y_t^i\|x_t = \rho, \mathcal{M}_j^{(r)})$	$p(y_t^i\|\boldsymbol{\Theta}_{\rho_t}^{(r),j})$
HMM state sequence (b)	$p(x' = \boldsymbol{\beta}\|z'_n = i, \mathcal{M}^{(b)})$	$p(\boldsymbol{\beta}\|\mathcal{M}_i^{(b)}) = \boldsymbol{\pi}_{\boldsymbol{\beta}}^{(b),i}$
HMM observation likelihood (b)	$p(\boldsymbol{y}_n\|z'_n = i, \mathcal{M}^{(b)})$	$p(\boldsymbol{y}_n\|\mathcal{M}_i^{(b)})$
Gaussian emission likelihood (b)	$p(y_{nt}\|x'_{nt} = \beta, \mathcal{M}_i^{(b)})$	$p(y_{nt}\|\boldsymbol{\Theta}_{\beta_t}^{(b),i})$
expectations		
HMM observation sequence (b)	$\mathbb{E}_{\boldsymbol{y}\|z'=i,\mathcal{M}^{(b)}}[\cdot]$	$\mathbb{E}_{\mathcal{M}_i^{(b)}}[\cdot]$
expected log-likelihood	*lower bound*	*variational distribution*
$\mathbb{E}_{\boldsymbol{y}\|\mathcal{M}_i^{(b)}}[\log p(\boldsymbol{y}\|\mathcal{M}_j^{(r)})]$	$\mathcal{L}_{HMM}^{i,j}$	-
$\mathbb{E}_{\mathcal{M}_i^{(b)}}\mathbb{E}_{\boldsymbol{y}\|\mathcal{M}_i^{(b)}}[\log p(\boldsymbol{y}\|\mathcal{M}_j^{(r)})]$	$\mathcal{L}_{HMM}^{\bar{i},j}$	-
$\mathbb{E}_{\mathcal{M}_j^{(r)}}\mathbb{E}_{\mathcal{M}_i^{(b)}}\mathbb{E}_{\boldsymbol{y}\|\mathcal{M}_i^{(b)}}[\log p(\boldsymbol{y}\|\mathcal{M}_j^{(r)})]$	$\mathcal{L}_{HMM}^{\bar{i},\bar{j}}$	$q^{i,j}(\boldsymbol{\rho}\|\boldsymbol{\beta})$

weight. An observation sequence with length τ is denoted by $\boldsymbol{y} = (y_1, y_2, ..., y_\tau)$, and depends on a hidden state sequence $\boldsymbol{x} = (x_1, x_2, ..., x_\tau)$. The observation likelihood for $\boldsymbol{y} \sim \mathcal{M}$ is $p(\boldsymbol{y}|\mathcal{M}) = \sum_i \omega_i p(\boldsymbol{y}|\mathcal{M}_i)$, where the i-th HMM $\mathcal{M}_i$ with S states is specified by parameters $\mathcal{M}_i = \{\boldsymbol{\pi}^i, \boldsymbol{A}^i, \{\boldsymbol{\Theta}_\beta^i\}_{\beta=1}^S\}$. In detail, $\boldsymbol{\pi}^i = [\pi_1^i, ..., \pi_S^i]$ is the initial state probability, where $\pi_\beta^i = p(x_1 = \beta|\mathcal{M}_i)$. $\boldsymbol{A}^i = (a_{\beta,\beta'}^i)_{S\times S}$ is the state transition matrix, where $a_{\beta,\beta'}^i = p(x_{t+1} = \beta'|x_t = \beta, \mathcal{M}_i)$ is the transition probability from state β to β'. Thus the probability of a state sequence $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_\tau)$ will be $p(\boldsymbol{x} = \boldsymbol{\beta}|\mathcal{M}_i) = \pi_{\beta_1}^i \prod_{t=2}^\tau a_{\beta_{t-1},\beta_t}^i \triangleq \boldsymbol{\pi}_{\boldsymbol{\beta}}^i$. $\boldsymbol{\Theta}_\beta^i$ is the parameter set of emission density at state β. Here, we assume a Gaussian distribution, $p(y_t|x_t = \beta, \mathcal{M}_i) = \mathcal{N}(y_t|\boldsymbol{\mu}_\beta^i, (\boldsymbol{\Lambda}_\beta^i)^{-1}) \triangleq p(y_t|\boldsymbol{\Theta}_\beta^i)$, with mean $\boldsymbol{\mu}_\beta^i$ and precision matrix $\boldsymbol{\Lambda}_\beta^i$. Thus, the probability of an observation $\boldsymbol{y}$ generated by $\mathcal{M}_i$ will be $p(\boldsymbol{y}|\mathcal{M}_i) = \sum_{\boldsymbol{x}=\boldsymbol{\beta}} p(\boldsymbol{y}, \boldsymbol{x} = \boldsymbol{\beta}|\mathcal{M}_i) = \sum_{\boldsymbol{x}=\boldsymbol{\beta}} p(\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{\beta}, \mathcal{M}_i) p(\boldsymbol{x} = \boldsymbol{\beta}|\mathcal{M}_i)$, where $p(\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{\beta}, \mathcal{M}_i) = \prod_t^\tau p(y_t|\boldsymbol{\Theta}_{\beta_t}^i)$ and the summation is over all state sequences of length τ.

Considering the H3M with unknown number of components K and number of states S, and treating them as random variables, the observation likelihood will be $p(\boldsymbol{y}|\mathcal{M}) = \sum_{K,S} p(K,S) p(\boldsymbol{y}|\mathcal{M}, K, S)$, where $p(K,S)$ is a prior of pair (K,S) and the summation is over all candidate number of components $K \in [K_{min}, K_{max}]$ and states $S \in [S_{min}, S_{max}]$, where we use shorthand $[A,B] = \{A, \cdots, B\}$.

B. Variational Bayesian Inference

A central task in the application of probabilistic models is evaluating the posterior distribution $p(\boldsymbol{H}|\boldsymbol{Y})$ of the hidden (latent) variables $\boldsymbol{H}$ given the observed data $\boldsymbol{Y}$. In a fully Bayesian framework, any unknown model parameters are given prior distributions and are absorbed into the set of latent variables $\boldsymbol{H}$. When it is infeasible to evaluate the posterior distribution directly, e.g., it has a highly complex form, then variational inference can be used to approximate $p(\boldsymbol{H}|\boldsymbol{Y})$ with a *variational distribution* $q(\boldsymbol{H})$. Furthermore, to consider different model structures, the number of mixture components and hidden states (K,S) can be considered as latent variables with prior distributions. Hence, we introduce a variational distribution $q(\boldsymbol{H}, K, S)$ as an approximation of the true posterior distribution $p(\boldsymbol{H}, K, S|\boldsymbol{Y})$.

The VB framework for an H3M is formulated as follows. The *marginal log-likelihood* (i.e., model evidence) $\log p(\boldsymbol{Y})$ is decomposed into a lower-bound and Kullback-Leibler divergence (KLD) term [26] (see derivation in Appendix A),

$$\log p(\boldsymbol{Y}) = \mathcal{L}(q) + \mathrm{KL}(q||p)$$

where we define

$$\mathcal{L}(q) = \sum_{K,S} q(K,S) \left[\mathcal{L}_{(K,S)}(q) + \log \frac{p(K,S)}{q(K,S)} \right], \tag{1}$$

$$\mathcal{L}_{(K,S)}(q) = \int q(\boldsymbol{H}|K,S) \log \frac{p(\boldsymbol{Y}, \boldsymbol{H}|K,S)}{q(\boldsymbol{H}|K,S)} d\boldsymbol{H}, \tag{2}$$

$$\mathrm{KL}(q||p) = \sum_{K,S} \int q(\boldsymbol{H}, K, S) \log \frac{q(\boldsymbol{H}, K, S)}{p(\boldsymbol{H}, K, S|\boldsymbol{Y})} d\boldsymbol{H}.$$

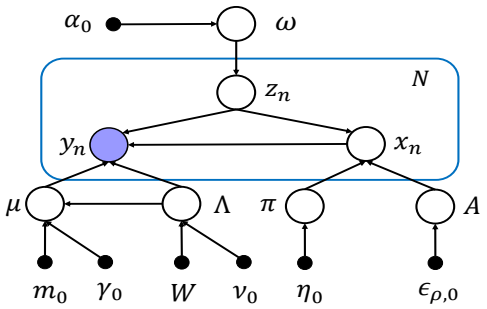


Fig. 1. Graphical model representing the Bayesian H3M. The plate denotes a set of N i.i.d. observations. $\boldsymbol{y}_n$ is one observation sequence, $\boldsymbol{x}_n$ is the state sequence that emits $\boldsymbol{y}_n$, and $\boldsymbol{z}_n$ indicates which HMM component $\boldsymbol{y}_n$ and $\boldsymbol{x}_n$ are assigned to. The variables outside the plate (e.g., $\boldsymbol{\mu}$) are HMM parameters $\boldsymbol{\omega} = \{\omega_j\}, \boldsymbol{\mu} = \{\mu_\rho^j\}, \boldsymbol{\Lambda} = \{\Lambda_\rho^j\}, \boldsymbol{\pi} = \{\pi_\rho^j\}, \boldsymbol{A} = \{\boldsymbol{A}_\rho^j\}$. $\{\alpha_0, \eta_0, \epsilon_{\rho,0}, \gamma_0, \boldsymbol{m}_0, \boldsymbol{W}_0, \nu_0\}$ are the hyperparameters.

Since $\mathrm{KL}(q||p)) \geq 0$, we have $\log p(\boldsymbol{Y}) \geq \mathcal{L}(q)$, which holds for any distribution $q(\boldsymbol{H}, K, S)$, and equality occurs when $q(\boldsymbol{H}, K, S) = p(\boldsymbol{H}, K, S|\boldsymbol{Y})$ (i.e. $\mathrm{KL}(q||p)) = 0$). Therefore, $\mathcal{L}(q)$ is a lower bound on $\log p(\boldsymbol{Y})$, and optimizing $\mathcal{L}(q)$ w.r.t $q(\boldsymbol{H}, K, S)$ will obtain an approximation of the true posterior distribution $p(\boldsymbol{H}, K, S|\boldsymbol{Y})$.

However, if we maximize $\mathcal{L}(q)$ w.r.t. $q(\boldsymbol{H}|K, S)$, the results for different pairs of (K, S) are coupled since they are conditioned on (K, S). Instead we first optimize each of the $q(\boldsymbol{H}|K, S)$ individually by optimizing $\mathcal{L}_{(K,S)}(q)$. Assuming that $q(\boldsymbol{H}|K, S) = \prod_l^L q_l(H_l|K, S)$ and $\{H_l\}_{l\in[L]}$ is a partition of $\boldsymbol{H}$, then the optimal solution $q_l^*(H_l|K, S)$ is [26]:

$$\log q_l^*(H_l|K, S) = \mathbb{E}_{l'\neq l}[\log p(\boldsymbol{Y}, \boldsymbol{H}|K, S)] + \text{const}, \quad (3)$$

$$\mathbb{E}_{l'\neq l}[\log p(\boldsymbol{Y}, \boldsymbol{H}|K, S] = \int (\log p(\boldsymbol{Y}, \boldsymbol{H}|K, S)) \prod_{l'\neq l} q_{l'}(H_{l'}|K, S) dH_{l'}.$$

Model Selection. For a set of candidate models, i.e. different pairs of (K, S). We can rewrite (1) as

$$\mathcal{L}(q) = \sum_{K,S} q(K, S) \log \frac{p(K, S) \exp\{\mathcal{L}_{(K,S)}(q)\}}{q(K, S)}. \quad (4)$$

Recognizing (4) as the negative KLD between $q(K, S)$ and the unnormalized distribution $p(K, S) \exp\{\mathcal{L}_{(K,S)}(q)\}$. The lower bound $\mathcal{L}$ will be maximized when the KLD is minimized when

$$q^*(K, S) \propto p(K, S) \exp\left\{\mathcal{L}_{(K,S)}(q)\right\}. \quad (5)$$

Thus, the optimal model structure is found by $(K^*, S^*) = \arg\max_{K,S} q^*(K, S)$.

IV. Variational Bayesian Hierarchical EM Algorithm for H3Ms

In this section, we derive a variational Bayesian (VB) hierarchical EM algorithm, which takes as input an H3M with priors over each parameter and outputs posteriors over the parameters of an equivalent H3M with fewer number of components (VBHEM-H3M). Formally, let $\mathcal{M}^{(b)} = \{\omega_i^{(b)}, \mathcal{M}_i^{(b)}\}_{i=1}^{K^{(b)}}$ represents a "base" H3M with $K^{(b)}$ components (HMMs) and $S^{(b)}$ states for each component. The input is the prior $p(\mathcal{M}^{(b)})$, consisting of priors over each parameter of $\mathcal{M}^{(b)}$, denoted by $p(\boldsymbol{\omega}^{(b)})$, $p(\boldsymbol{\mu}_\beta^{(b),i}|\boldsymbol{\Lambda}_\beta^{(b),i})$, $p(\boldsymbol{\Lambda}_\beta^{(b),i})$, $p(\boldsymbol{\pi}^{(b),i})$, and $p(\boldsymbol{a}_\beta^{(b),i})$, where $\boldsymbol{a}_\beta^{(b),i}$ is a row of $\boldsymbol{A}^{(b),i}$, $\beta \in [1, S^{(b)}]$. Our goal is to simplify the base model $\mathcal{M}^{(b)}$ to a "reduced" mixture model $\mathcal{M}^{(r)}$ and automatically determine the number of components $K^{(r)}$ and states $S^{(r)}$ in $\mathcal{M}^{(r)}$. Rather than learning a single model $\mathcal{M}^{(r)}$ as in VHEM, VBHEM-H3M estimates a posterior distribution over the reduced model's parameters and structures. The reduced model is denoted by $\mathcal{M}^{(r)} = \{\omega_j^{(r)}, \mathcal{M}_j^{(r)}\}_{j=1}^{K^{(r)}}$ with $K^{(r)}$ components and $S^{(r)}$ states, where $K^{(b)} > K^{(r)}$ and $S^{(b)} \geq S^{(r)}$. In the Bayesian framework, we also assume priors on all unknown parameters in $\mathcal{M}^{(r)}$, denoted by $p(K^{(r)}, S^{(r)})$, $p(\boldsymbol{\omega}^{(r)})$, $p(\boldsymbol{\mu}_\rho^{(r),j}|\boldsymbol{\Lambda}_\rho^{(r),j})$, $p(\boldsymbol{\Lambda}_\rho^{(r),j})$, $p(\boldsymbol{\pi}^{(r),j})$, and $p(\boldsymbol{a}_\rho^{(r),j})$, where $\boldsymbol{a}_\rho^{(r),j}$ is a row of $\boldsymbol{A}^{(r),j}$, $\rho \in [1, S^{(r)}]$. Note that we will always use superscripts (b) and (r) to distinguish the parameters for base and reduced model, i and j to index the mixture component in the base and reduced models, and β and ρ to index the hidden states in the base and reduced models, respectively. Table I and II summarize the notation used in the derivation, including the variable names, latent variables, model names, prior distributions and expectations and expected log-likelihood.

A. Framework

One possible solution to estimate $\mathcal{M}^{(r)}$ is to directly sample from $\mathcal{M}^{(b)}$ and then estimate $\mathcal{M}^{(r)}$ with any needed number of components and states by the EM algorithm. However, this would be inefficient when handling large-scale high-dimensional data. Also, the number of components and states must be set by hand, which introduces experimenter bias. Instead, we take our inspiration from VBmerge [45], which reduces a Gaussian mixture model (GMM) by directly clustering the Gaussian components in the Bayesian framework, and the number of components is simultaneously determined.

We define a set of N sequence samples $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N\}$, where $\boldsymbol{y}_n = (y_{n,1}, y_{n,2}, ..., y_{n,\tau})$ is a sequence, and $y_{n,t} \in \mathbb{R}^d$ is the observation at time t. The generative process of the data set $\boldsymbol{Y}$ is:

1) Sample a base model $\mathcal{M}^{(b)} \sim p(\mathcal{M}^{(b)})$;
2) Sample (i.i.d.) data sequences $\boldsymbol{y}_n \sim \mathcal{M}^{(b)}$, $n = [1, N]$.

Thus, the marginal likelihood over the data according to the base model prior $p(\mathcal{M}^{(b)})$ is

$$p'(\boldsymbol{Y}) = \int p(\boldsymbol{Y}|\mathcal{M}^{(b)}) p(\mathcal{M}^{(b)}) d\mathcal{M}^{(b)} \quad (6)$$

A similar generative process also exists for the reduced model, and thus the marginal likelihood of the data according to the reduced model prior $p(\mathcal{M}^{(r)})$ is

$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y}|\mathcal{M}^{(r)}) p(\mathcal{M}^{(r)}) d\mathcal{M}^{(r)}. \quad (7)$$

A typical VB method constructs a lower bound of the evidence $\log p(\boldsymbol{Y})$, under the model being learned, which in our case is the reduced model $\mathcal{M}^{(r)}$. However, in our scenario, we do not have direct access to the data $\boldsymbol{Y}$, but instead have access to the model $\mathcal{M}^{(b)}$ that generates the data. Thus, our starting point is the expected evidence, where the "evidence" $\boldsymbol{Y} \sim p'$ is generated from the input model $\mathcal{M}^{(b)}$, and evaluated according to the reduced model $\mathcal{M}^{(r)}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}\sim p'} \log p(\boldsymbol{Y}) \\
&= \int \Big[\int p(\boldsymbol{Y}|\mathcal{M}^{(b)}) p(\mathcal{M}^{(b)}) d\mathcal{M}^{(b)} \Big] \log p(\boldsymbol{Y}) d\boldsymbol{Y} \\
&= \mathbb{E}_{\mathcal{M}^{(b)}} \mathbb{E}_{\boldsymbol{Y}|\mathcal{M}^{(b)}} \log p(\boldsymbol{Y}), \quad (8)
\end{aligned}$$

where the exchange of the integral is guaranteed by Fubini's theorem. Substituting the lower bound in (1) of the marginal log-likelihood into (8), we have

$$\mathbb{E}_{\mathcal{M}^{(b)}}\mathbb{E}_{\boldsymbol{Y}|\mathcal{M}^{(b)}}\log p(\boldsymbol{Y}) \geq \sum_{K^{(r)}}\sum_{S^{(r)}} q(K^{(r)},S^{(r)})\cdot \left[\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q)] + \log\frac{p(K^{(r)},S^{(r)})}{q(K^{(r)},S^{(r)})}\right] = \bar{\mathcal{L}}(q), \quad (9)$$

where the inequality holds due to the non-negativity of probability density functions. Note that the expectation $\mathbb{E}_{\boldsymbol{Y}}[\cdot] = \mathbb{E}_{\mathcal{M}^{(b)}}\mathbb{E}_{\boldsymbol{Y}|\mathcal{M}^{(b)}}[\cdot]$ only influences $\mathcal{L}_{(K^{(r)},S^{(r)})}(q)$ on the RHS of (9). The set of hidden variables is

$$\boldsymbol{H} = \big\{\boldsymbol{Z}, \{\boldsymbol{\omega}^{(r),j}, \boldsymbol{\pi}^{(r),j}, \boldsymbol{A}^{(r),j}, \boldsymbol{\mu}^{(r),j}, \boldsymbol{\Lambda}^{(r),j}\}_{j=1}^{K^{(r)}}\big\}.$$

Looking at each hidden variable H_l in $\boldsymbol{H}$ and using (3), the optimal solution for $q_l^*(H_l|K^{(r)},S^{(r)})$ is

$$\log q_l^*(H_l|K^{(r)},S^{(r)}) \propto \mathbb{E}_{l'\neq l}[\mathbb{E}_{\boldsymbol{Y}}\log p(\boldsymbol{Y},\boldsymbol{H}|K^{(r)},S^{(r)})], \quad (10)$$

and using (5), the optimal $q^*(K^{(r)},S^{(r)})$ is

$$\log q^*(K^{(r)},S^{(r)}) \propto \log p(K^{(r)},S^{(r)}) + \mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q)]. \quad (11)$$

From (10-11), our algorithm contains two steps:

Step 1: For each candidate pair $(K^{(r)},S^{(r)})$, calculate each optimal solution $q_l^*(H_l|K^{(r)},S^{(r)})$.

Step 2: Find the optimal model structure through $(K^{(r),*},S^{(r),*}) = \arg\max_{(K^{(r)},S^{(r)})}\log q^*(K^{(r)},S^{(r)})$.

However, the expectation in (10) cannot be calculated in the closed-form. We will show how to approximate that in the following sections.

B. Priors

In this section, we introduce the conjugate prior distributions over the parameters of the H3M $\mathcal{M}$ with K components and S states (see Fig. 1),

$$\begin{aligned}
p(\mathcal{M}|K,S) &= p(\boldsymbol{\omega}|K,S)\prod_{i=1}^{K} p(\mathcal{M}_i|K,S)\\
&= p(\boldsymbol{\omega}|K,S)\prod_{i} p(\boldsymbol{\pi}^i|K,S)p(\boldsymbol{A}^i|K,S)p(\boldsymbol{\mu}^i,\boldsymbol{\Lambda}^i|K,S).
\end{aligned}$$

Here, we assume that (K,S) are fixed and do not explicitly condition on them to remove clutter. The priors for $\mathcal{M}^{(b)}$ are

$$\begin{aligned}
p(\boldsymbol{\omega}^{(b)}) &= \mathrm{Dir}(\boldsymbol{\omega}^{(b)}|\boldsymbol{\alpha}_0^{(b)}),\\
p(\boldsymbol{\pi}^{(b),i}) &= \mathrm{Dir}(\boldsymbol{\pi}^{(b),i}|\boldsymbol{\eta}_0^{(b),i}),\\
p(\boldsymbol{A}^{(b),i}) &= \prod_{\beta} p(\boldsymbol{a}_\beta^{(b),i}) = \prod_{\beta}\mathrm{Dir}(\boldsymbol{a}_\beta^{(b),i}|\boldsymbol{\epsilon}_{\beta,0}^{(b),i}),\\
p(\boldsymbol{\mu}^{(b),i},\boldsymbol{\Lambda}^{(b),i}) &= \prod_{\beta} p(\boldsymbol{\mu}_\beta^{(b),i},\boldsymbol{\Lambda}_\beta^{(b),i}),\\
p(\boldsymbol{\mu}_\beta^{(b),i},\boldsymbol{\Lambda}_\beta^{(b),i}) &= \mathcal{N}(\boldsymbol{\mu}_\beta^{(b),i}|\boldsymbol{m}_{\beta,0}^{(b),i},(\gamma_{\beta,0}^{(b),i}\boldsymbol{\Lambda}_\beta^{(b),i})^{-1})\\
&\quad\cdot\mathcal{W}(\boldsymbol{\Lambda}_\beta^{(b),i}|\boldsymbol{W}_{\beta,0}^{(b),i},\nu_{\beta,0}^{(b),i}),
\end{aligned}$$

where $\mathrm{Dir}(\cdot|\boldsymbol{\alpha})$ is a Dirichlet distribution with concentration vector $\boldsymbol{\alpha}$, $\mathcal{N}(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\mathcal{W}(\cdot|\boldsymbol{W},\nu)$ is a Wishart distribution with scale matrix $\boldsymbol{W}$ and degrees-of-freedom ν (see Appendix B for the details of each distribution). The hyperparameters of the base model are summarized as the set $\mathcal{P}^{(b)}=\{\boldsymbol{\alpha}_0^{(b)}, \{\boldsymbol{\eta}_0^{(b),i},\boldsymbol{\epsilon}_{\beta,0}^{(b),i},\boldsymbol{\gamma}_{\beta,0}^{(b),i},\boldsymbol{m}_{\beta,0}^{(b),i},\boldsymbol{W}_{\beta,0}^{(b),i},\boldsymbol{\nu}_{\beta,0}^{(b),i}\}_{\beta=1,i=1}^{S^{(b)},K^{(b)}}\}$.

Algorithm 1 Optimizing the Variational Distribution

Input: hyperparameter sets $\mathcal{P}^{(b)}$ and $\mathcal{P}^{(r)}$, and the number of virtual samples N, clusters $K^{(r)}$, and states $S^{(r)}$.
Output: variational distributions $q^*(\boldsymbol{Z})$, $q^*(\boldsymbol{\omega}^{(r)})$, $q^*(\boldsymbol{\pi}^{(r),j})$, $q^*(\boldsymbol{A}^{(r),j})$, $q^*(\boldsymbol{\mu}^{(r),j},\boldsymbol{\Lambda}^{(r),j})$, $j\in[1,K^{(r)}]$.

1: Pre-process base model using (16)-(21).
2: **repeat**
3: VBH E-step : compute responsibilities $\hat{z}_{i,j}$ using (22).
4: VBH M-step: update variational parameters $\boldsymbol{\alpha}^{(r)}$, $\boldsymbol{\eta}^{(r),j}$, $\boldsymbol{\epsilon}^{(r),j}$, $\boldsymbol{m}^{(r),j}$, $\boldsymbol{\lambda}^{(r),j}$, $\boldsymbol{\Lambda}^{(r),j}$ and $\boldsymbol{\nu}^{(r),j}$ for each j using (24)-(25) .
5: **until** convergence of $\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)]$.

The priors for reduced model $\mathcal{M}^{(r)}$ share the same probabilities form as $p(\mathcal{M}^{(b)})$ but with simpler hyperparameters. We set $\boldsymbol{\alpha}_0^{(r),j}\equiv\alpha_0^{(r)}$ (scalar), $\boldsymbol{\eta}_0^{(r),j}\equiv\eta_0^{(r)}$, $\boldsymbol{\epsilon}_{\rho,0}^{(r),j}\equiv\epsilon_0^{(r)}$, $\boldsymbol{m}_{\rho,0}^{(r),j}\equiv\boldsymbol{m}_0^{(r)}$, $\gamma_{\rho,0}^{(r),j}\equiv\gamma_0^{(r)}$, $\boldsymbol{W}_{\rho,0}^{(r),j}\equiv\boldsymbol{W}_0^{(r)}$, and $\nu_{\rho,0}^{(r),j}\equiv\nu_0^{(r)}$ which are common for all components j and states ρ in the reduced model. Similarly, we summary the hyperparameters as the set $\mathcal{P}^{(r)} = \{\alpha_0^{(r)},\eta_0^{(r)},\epsilon_0^{(r)},\gamma_0^{(r)},\boldsymbol{m}_0^{(r)},\boldsymbol{W}_0^{(r)},\nu_0^{(r)}\}$.

For the priors over the number of components $K^{(r)}$ and states $S^{(r)}$, we assume a Poisson distribution on $K^{(r)}$ and a uniform distribution on $S^{(r)}$,

$$p(K^{(r)}=K,S^{(r)}=S) = \frac{\lambda_0^K e^{-\lambda_0}}{K!}\frac{1}{S_{max}-S_{min}+1}.$$

This prior allows us to express a preference for different models, through the hyperparameter λ_0.

C. Optimizing the Variational Distribution

We next explain how to utilize the prior distributions over the parameters of base model $\mathcal{M}^{(b)}$ and optimize the variational distribution. We assume grouped observations $\boldsymbol{Y} = \{\boldsymbol{Y}_1,...,\boldsymbol{Y}_{K^{(b)}}\}$ as in [50]. The subset $\boldsymbol{Y}_i$ has size $N_i = N\omega_i^{(b)}$, and consists of all $\boldsymbol{y}_n$ that are generated by $\mathcal{M}_i^{(b)}$. Similarly the grouped assignments are $\boldsymbol{Z} = \{\boldsymbol{z}_1,...,\boldsymbol{z}_{K^{(b)}}\}$, and $\boldsymbol{z}_i$ is a 1-of-$K^{(r)}$ binary vector where each element is an indicator variable z_{ij}, with $z_{ij}=1$ if the observations $\boldsymbol{Y}_i$ are assigned to the j-th reduced model $\mathcal{M}_j^{(r)}$, and $z_{ij}=0$ otherwise.

Revisiting (10), the joint distribution of random variables $\boldsymbol{Y}$ and $\boldsymbol{H}$ condition on $K^{(r)}$ and $S^{(r)}$ is

$$\begin{aligned}
&\log p(\boldsymbol{Y},\boldsymbol{H}|K^{(r)},S^{(r)}) \quad (12)\\
&= \log p(\boldsymbol{Y}|\boldsymbol{Z},\{\mathcal{M}_j^{(r)}\},K^{(r)},S^{(r)}) + \log p(\boldsymbol{\omega}^{(r)}|K^{(r)})\\
&\quad + \log p(\boldsymbol{Z}|\boldsymbol{\omega}^{(r)},K^{(r)}) + \sum_j \log p(\mathcal{M}_j^{(r)}|K^{(r)},S^{(r)}).
\end{aligned}$$

In the following, we consider a fixed pair $(K^{(r)},S^{(r)})$, and do not explicitly write the dependence to reduce clutter. Note that taking the expectation $\mathbb{E}_{\boldsymbol{Y}}[\cdot] = \mathbb{E}_{\mathcal{M}^{(b)}}\mathbb{E}_{\boldsymbol{Y}|\mathcal{M}^{(b)}}[\cdot]$ w.r.t. (12) only affects the first term in the RHS of (12), i.e., (see Appendix C for the detailed derivation) ,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Y}}\log p(\boldsymbol{Y}|\boldsymbol{Z},\{\mathcal{M}_j^{(r)}\}) \quad (13)\\
&= \sum_{i,j} z_{ij}N\mathbb{E}_{\boldsymbol{\omega}^{(b)}}[\omega_i^{(b)}]\mathbb{E}_{\mathcal{M}_i^{(b)}}\mathbb{E}_{\boldsymbol{y}|\mathcal{M}_i^{(b)}}[\log p(\boldsymbol{y}|\mathcal{M}_j^{(r)})].
\end{aligned}$$

Then, we consider a variational distribution which factorizes between the latent variables and the parameters so that

$$q(\boldsymbol{Z},\ \boldsymbol{\omega}^{(r)}, \boldsymbol{\pi}^{(r)}, \boldsymbol{A}^{(r)}, \boldsymbol{\mu}^{(r)}, \boldsymbol{\Lambda}^{(r)}) = q(\boldsymbol{Z})q(\boldsymbol{\omega}^{(r)}, \boldsymbol{\pi}^{(r)}, \boldsymbol{A}^{(r)}, \boldsymbol{\mu}^{(r)}, \boldsymbol{\Lambda}^{(r)}).$$

The corresponding sequential update equations for these factors can be derived from (10). In particular, the functional form of the factors $q(\boldsymbol{Z})$ and $q(\boldsymbol{\omega}^{(r)}, \boldsymbol{\pi}^{(r)}, \boldsymbol{A}^{(r)}, \boldsymbol{\mu}^{(r)}, \boldsymbol{\Lambda}^{(r)})$ will be determined automatically by optimization of the variational distribution. The whole algorithm for optimizing the variational distribution is summarized in Alg. 1. We explain each step in the following sections.

1) Pre-processing the Input Prior over HMMs: In our algorithm, the input is the prior of base model, and in (13) the expectation w.r.t. the base model results in a new *equivalent* base model. In detail, for the expected log-likelihood, we have

$$\mathbb{E}_{\mathcal{M}_i^{(b)}}\mathbb{E}_{\boldsymbol{y}|\mathcal{M}_i^{(b)}}[\log p(\boldsymbol{y}|\mathcal{M}_j^{(r)})] \geq \mathcal{L}_{HMM}^{\bar{i},j},$$

where

$$\mathcal{L}_{HMM}^{\bar{i},j} = \sum_{\boldsymbol{\beta}} \mathbb{E}_{\{\boldsymbol{\pi}^{(b),i}, \boldsymbol{A}^{(b),i}\}}[\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(b),i}] \sum_{\boldsymbol{\rho}} q^{i,j}(\boldsymbol{\rho}|\boldsymbol{\beta}) \cdot \tag{14}$$

$$\left\{ \log \frac{\boldsymbol{\pi}_{\boldsymbol{\rho}}^{(r),j}}{q^{i,j}(\boldsymbol{\rho}|\boldsymbol{\beta})} + \sum_t \mathbb{E}_{\boldsymbol{\Theta}_{\beta_t}^{(b),i}} \mathbb{E}_{y|\boldsymbol{\Theta}_{\beta_t}^{(b),i}} \left[\log \mathcal{N}(y|\boldsymbol{\Theta}_{j,\rho_t}^{(r)}) \right] \right\}$$

where we introduce a variational distribution $q^{i,j}(\boldsymbol{\rho}|\boldsymbol{\beta})$ on the state sequence $\boldsymbol{\rho}$, which depends on a state sequence $\boldsymbol{\beta}$ from $\mathcal{M}_i^{(b)}$. $q^{i,j}(\boldsymbol{\rho}|\boldsymbol{\beta})$ represents the probability of the state sequence $\boldsymbol{\rho}$ in HMM $\mathcal{M}_j^{(r)}$, when $\mathcal{M}_j^{(r)}$ is used to explain the observation sequence in $\boldsymbol{Y}_i$ that evolved through state sequence $\boldsymbol{\beta}$. The summation over state sequences $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ in (14) can be efficiently calculated using a recursive algorithm from [51].

For all the expectations in (13) and (14),

$$\mathbb{E}_{\boldsymbol{\omega}^{(b)}}[\omega_i^{(b)}] = \tilde{\omega}_i^{(b)},$$

$$\mathbb{E}_{\{\boldsymbol{\pi}^{(b),i}, \boldsymbol{A}^{(b),i}\}}[\boldsymbol{\pi}_{\boldsymbol{\beta}}^{(b),i}] \geq \tilde{\pi}_{\beta_1}^{(b),i} \cdot \prod_{t=2}^{\tau} \tilde{a}_{\beta_{t-1}\beta_t}^{(b),i}, \tag{15}$$

$$\mathbb{E}_{\boldsymbol{\Theta}_{\beta_t}^{(b),i}} \mathbb{E}_{y|\boldsymbol{\Theta}_{\beta_t}^{(b),i}} \log \mathcal{N}(y|\boldsymbol{\Theta}_{\rho_t}^{(r),j}) = -\frac{1}{2}(\tilde{\boldsymbol{\mu}}_{\beta_t}^{(b),i} - \boldsymbol{\mu}_{\rho_t}^{(r),j})^T \boldsymbol{\Lambda}_{\rho_t}^{(r),j} (\tilde{\boldsymbol{\mu}}_{\beta_t}^{(b),i} - \boldsymbol{\mu}_{\rho_t}^{(r),j}) - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Lambda}_{\rho_t}^{(r),j} [\tilde{\boldsymbol{\Lambda}}_{\beta_t}^{(b),i}]^{-1}\right) + \frac{1}{2} \log |\boldsymbol{\Lambda}_{\rho_t}^{(r),j}| - \frac{1}{2} \log(2\pi)$$

and we have defined (see Appendix B for results),

$$\tilde{\omega}_i^{(b)} = \mathbb{E}_{\boldsymbol{\omega}^{(b)}}[\omega_i^{(b)}], \tag{16}$$

$$\tilde{\pi}_{\beta_1}^{(b),i} = \exp\left\{\mathbb{E}_{\boldsymbol{\pi}^{(b),i}}[\log \pi_{\beta_1}^{(b),i}]\right\}, \tag{17}$$

$$\tilde{a}_{\beta_{t-1}\beta_t}^{(b),i} = \exp\left\{\mathbb{E}_{\boldsymbol{a}_{\beta_{t-1}}^{(b),i}}[\log a_{\beta_{t-1}\beta_t}^{(b),i}]\right\}, \tag{18}$$

$$\tilde{\boldsymbol{\mu}}_{\beta_t}^{(b),i} = \mathbb{E}_{\{\boldsymbol{\mu}_{\beta_t}^{(b),i}\}}[\boldsymbol{\mu}_{\beta_t}^{(b),i}], \tag{19}$$

$$[\tilde{\boldsymbol{\Lambda}}_{\beta_t}^{(b),i}]^{-1} = c_{\beta_t}^{(b),i} \mathbb{E}_{\boldsymbol{\Lambda}_{\beta_t}^{(b),i}}[(\boldsymbol{\Lambda}_{\beta_t}^{(b),i})^{-1}], \tag{20}$$

$$c_{\beta_t}^{(b),i} = \frac{\gamma_{\beta_t}^{(b),i} + 1}{\gamma_{\beta_t}^{(b),i}}. \tag{21}$$

Thus, the pre-processed base model is an equivalent model with parameters as in (16-20).

Effect of Input Priors: We have stated that the main motivation for introducing an expectation $\mathbb{E}_{\boldsymbol{Y}}[\cdot]$ is to directly cluster the input HMMs without requiring the original data (through $\mathbb{E}_{\boldsymbol{Y}|\mathcal{M}^{(b)}}[\cdot]$) . A common way to estimate an input HMM is using VBHMM [39, 52], which computes an approximate posterior over the HMM parameters given the data. When clustering HMMs learned with VBHMM, we can take advantage of the computed posteriors (through $\mathbb{E}_{\mathcal{M}^{(b)}}[\cdot]$) by using the output posteriors of VBHMM as the priors $p(\mathcal{M}_i^{(b)})$ for the input HMMs for clustering. This should give more robust results than just using the point-estimate from MLE or MAP estimation, since the model uncertainty is accounted for.

Specifically, we compare VHEM using point-estimates as input and VBHEM using prior distributions as input. Consider VHEM using the expectation of the base model prior as the point-estimate for the input HMM, whereas VBHEM uses the prior distribution for each base model as input. Both VHEM and VBHEM use the same weight $\tilde{\omega}_i^{(b)}$ and mean $\tilde{\boldsymbol{\mu}}_{\beta_t}^{(b),i}$ for the input HMM. However, the initial probabilities and transition probabilities are different, but still order-preserving, and the effect of VBHEM is to "saturate" the initial state probability distribution, by increasing the probability of higher probability states and decreasing the probability of low probability states (see Appendix F). Thus, VBHEM has a tendency to keep the the states with high probability, especially when the sequence length τ is long. Finally, the covariance matrix $(\tilde{\boldsymbol{\Lambda}}_{\beta_t}^{(b),i})^{-1} \geq (\boldsymbol{\Lambda}_{\beta_t}^{(b),i})^{-1}$ because the coefficient $c_{\beta_t}^{(b),i} \geq 1$. Thus, the HMM with the updated emission probability in (19-21) incorporates the model uncertainty through generating more diverse data on each state. The benefit of modeling the input uncertainty is validated in the experiments in Sec V-A.

2) Variational E-Step: In the variational E-Step, we derive the optimal variational distribution of the assignment variables $\boldsymbol{Z}$ (i.e., calculate the responsibilities). The conditional distribution of $\boldsymbol{Z}$, given the mixing coefficients $\boldsymbol{\omega}^{(r)}$, is

$$p(\boldsymbol{Z}|\boldsymbol{\omega}^{(r)}) = \prod_i \prod_j \left[\omega_j^{(r)}\right]^{N_i z_{ij}}.$$

Making use of the result (10), the optimized $q(\boldsymbol{Z})$ is

$$\log q^*(\boldsymbol{Z}) \propto \mathbb{E}_{\boldsymbol{\omega}^{(r)}, \{\mathcal{M}_j^{(r)}\}} \mathbb{E}_{\boldsymbol{Y}} \log p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\omega}^{(r)}, \{\mathcal{M}_j^{(r)}\}).$$

After normalization, we have (see App. D for derivations):

$$q^*(\boldsymbol{Z}) = \prod_i \prod_j [\hat{z}_{ij}]^{z_{ij}},$$

$$\hat{z}_{ij} = \mathbb{E}[z_{ij}] = \frac{(\tilde{\omega}_j^{(r)})^{\tilde{N}_i} \exp\left(\tilde{N}_i \hat{\mathcal{L}}_{HMM}^{\bar{i},\bar{j}}\right)}{\sum_{j'} (\tilde{\omega}_{j'}^{(r)})^{\tilde{N}_i} \exp\left(\tilde{N}_i \hat{\mathcal{L}}_{HMM}^{\bar{i},\bar{j}'}\right)}, \tag{22}$$

where $\tilde{\omega}_j^{(r)} = \exp\left(\mathbb{E}_{\boldsymbol{\omega}^{(r)}}[\log \omega_j^{(r)}]\right)$ and $\tilde{N}_i = \tilde{\omega}_i^{(b)} N$. $\hat{\mathcal{L}}_{HMM}^{\bar{i},\bar{j}}$ is the optimized expected lower bound (w.r.t model j),

$$\hat{\mathcal{L}}_{HMM}^{\bar{i},\bar{j}} = \max_{q^{i,j}} \mathcal{L}_{HMM}^{\bar{i},\bar{j}}, \quad \mathcal{L}_{HMM}^{\bar{i},\bar{j}} = \mathbb{E}_{\mathcal{M}_j^{(r)}} \mathcal{L}_{HMM}^{\bar{i},j}.$$

In (22), the quantities $\hat{z}_{ij}$ are the responsibilities. Note that the optimal solution $q^*(\boldsymbol{Z})$ depends on moments evaluated w.r.t the distributions of other variables, and thus the variational update equations are coupled and must be solved iteratively.

3) Variational M-step: In the variational M-step, we update the variational parameters $\boldsymbol{\alpha}^{(r)}$, $\boldsymbol{\eta}^{(r),j}$, $\boldsymbol{\epsilon}^{(r),j}$, $\boldsymbol{m}^{(r),j}$, $\boldsymbol{\lambda}^{(r),j}$, $\boldsymbol{\Lambda}^{(r),j}$ and $\boldsymbol{\nu}^{(r),j}$ for each component (i.e., compute the optimized distributions of parameters $\boldsymbol{\omega}^{(r)}$, $\boldsymbol{\pi}^{(r),j}$, $\boldsymbol{A}^{(r),j}$,

Algorithm 2 VBHEM-H3M

Input: hyperparameter sets $\mathcal{P}^{(b)}$ and $\mathcal{P}^{(r)}$, the number of virtual samples N, candidates for number of clusters and states $K^{(r)} \in [K^{(r)}_{min}, K^{(r)}_{max}]$ and $S^{(r)} \in [S^{(r)}_{min}, S^{(r)}_{max}]$.
Output: reduced H3M $\mathcal{M} = \{\omega_j, \mathcal{M}_j\}_{j=1}^{K}$.

1: **for** each pair $(K^{(r)}, S^{(r)})$ **do**
2: **repeat**
3: Run Alg.1 and obtain $\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)]$.
4: Update hyperparameters $\alpha_0^{(r)}, \eta_0^{(r)}, \epsilon_{\rho,0}^{(r)}$, $\boldsymbol{m}_0^{(r)}$, $\boldsymbol{W}_0^{(r)}$, $\gamma_0^{(r)}, \nu_0^{(r)}$ using gradient ascent on $\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)]$.
5: **until** convergence
6: **end for**
7: Select the reduced H3M $\mathcal{M}^{(r)} = \{\omega_j, \mathcal{M}_j\}_{j=1}^{K^{(r)*}}$ with maximum $\log q^*(K^{(r)}, S^{(r)})$.
8: Prune out component j with low weight $\mathbb{E}[\omega_j^{(r)}]$, and state ρ with low probability $\mathbb{E}[\pi_\rho^{(r),j}] + \sum_{\rho'} \mathbb{E}[a_{\rho',\rho}^{(r),j}]$.

$\boldsymbol{\mu}^{(r),j}, \boldsymbol{\Lambda}^{(r),j}$). Revisiting (10), we have

$$\log q^*(\boldsymbol{\omega}^{(r)}, \boldsymbol{\pi}^{(r)}, \boldsymbol{A}^{(r)}, \boldsymbol{\mu}^{(r)}, \boldsymbol{\Lambda}^{(r)}) \propto \mathbb{E}_{\boldsymbol{Z}}\mathbb{E}_{\boldsymbol{Y}} \log p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\omega}^{(r)}, \{\mathcal{M}_j^{(r)}\}). \quad (23)$$

Decomposing the RHS of (23), the optimized variational posterior factorizes as

$$q^*(\boldsymbol{\omega}^{(r)}, \boldsymbol{\pi}^{(r)}, \boldsymbol{A}^{(r)}, \boldsymbol{\mu}^{(r)}, \boldsymbol{\Lambda}^{(r)}) = q^*(\boldsymbol{\omega}^{(r)}) \prod_j q^*(\boldsymbol{\pi}^{(r),j}) q^*(\boldsymbol{A}^{(r),j}) q^*(\boldsymbol{\mu}^{(r),j}, \boldsymbol{\Lambda}^{(r),j}),$$

and the form of each q distribution is automatically determined via the expectation in (23) (see Appendix D for the details),

$$q^*(\boldsymbol{\omega}^{(r)}) = \mathrm{Dir}(\boldsymbol{\omega}^{(r)}|\boldsymbol{\alpha}^{(r)}),$$
$$q^*(\boldsymbol{\pi}^{(r),j}) = \mathrm{Dir}(\boldsymbol{\pi}^{(r),j}|\boldsymbol{\eta}^{(r),j}),$$
$$q^*(\boldsymbol{a}_\rho^{(r),j}) = \mathrm{Dir}(\boldsymbol{a}_\rho^{(r),j}|\boldsymbol{\epsilon}_\rho^{(r),j}),$$

where parameter vectors $\boldsymbol{\alpha}^{(r)}$, $\boldsymbol{\eta}^{(r),j}$ and $\boldsymbol{\epsilon}_\rho^{(r),j}$ have elements

$$\begin{aligned} \alpha_j^{(r)} &= \alpha_0^{(r)} + N^j, & N^j &= \textstyle\sum_{i=1}^{K^{(b)}} \hat{z}_{ij}\tilde{N}_i \\ \eta_{\rho_1}^{(r),j} &= \eta_0^{(r)} + N_{\rho_1}^j, & N_{\rho_1}^j &= \textstyle\sum_{i=1}^{K^{(b)}} \hat{z}_{ij}\tilde{N}_i\hat{\nu}_1^{i,j}(\rho_1), \\ \epsilon_{\rho,\rho'}^{(r),j} &= \epsilon_{\rho,0}^{(r)} + N_{\rho,\rho'}^j, & N_{\rho,\rho'}^j &= \textstyle\sum_{i=1}^{K^{(b)}} \hat{z}_{ij}\tilde{N}_i\hat{\xi}^{i,j}(\rho,\rho'). \end{aligned} \quad (24)$$

Here $\hat{\nu}_1^{i,j}(\rho_1)$ and $\hat{\xi}^{i,j}(\rho,\rho')$ have the same form as in VHEM [25]. In each iteration, we update N^j (the number of samples assigned to j-th component $\mathcal{M}_j^{(r)}$), $N_{\rho_1}^j$ (the number of samples which have been assigned to $\mathcal{M}_j^{(r)}$ and have initial state ρ_1), and $N_{\rho,\rho'}^j$ (the number of samples which have been assigned to $\mathcal{M}_j^{(r)}$ and have transition from state ρ to ρ').

Consider the expectation of $\omega_j^{(r)}$ w.r.t. a Dirichlet distribution, $\mathbb{E}[\omega_j^{(r)}] = \frac{\alpha_0^{(r)}+N^j}{K^{(r)}\alpha_0^{(r)}+N}$. If a component for which $N^j \simeq 0$ and $\alpha_j^{(r)} \simeq \alpha_0^{(r)}$ and the prior is broad so that $\alpha_0^{(r)} \to 0$, then $\mathbb{E}[\omega_j^{(r)}] \to 0$ and j-th component plays no role in the model and will be pruned out automatically. Hyperparameters $\eta_{\rho_1}^{(r),j}$ and $\epsilon_{\rho,\rho'}^{(r),j}$ can be analyzed in a similar way, and the states with near 0 initial probability $\mathbb{E}[\pi_{\rho_1}^{(r),j}]$ and no transitions from other states $\sum_\rho \mathbb{E}[a_{\rho,\rho'}^{(r),j}]$ will have no role, and can be pruned.

Finally, using the product rule, the variational posterior distribution $q(\boldsymbol{\mu}_\rho^{(r),j}, \boldsymbol{\Lambda}_\rho^{(r),j})$ can be written as

$$q^*(\boldsymbol{\mu}_\rho^{(r),j}, \boldsymbol{\Lambda}_\rho^{(r),j}) = q(\boldsymbol{\mu}_\rho^{(r),j}|\boldsymbol{\Lambda}_\rho^{(r),j}) q(\boldsymbol{\Lambda}_\rho^{(r),j})$$
$$= \mathcal{N}(\boldsymbol{\mu}_\rho^{(r),j}|\boldsymbol{m}_\rho^{(r),j}, (\gamma_\rho^{(r),j}\boldsymbol{\Lambda}_\rho^{(r),j})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_\rho^{(r),j}|\boldsymbol{W}_\rho^{(r),j}, \nu_\rho^{(r),j}),$$

i.e., a Gaussian-Wishart distribution, where we have defined

$$\gamma_\rho^{(r),j} = \gamma_0^{(r)} + N_\rho^j, \quad \nu_\rho^{(r),j} = \nu_0^{(r)} + N_\rho^j + 1, \quad (25)$$
$$\boldsymbol{m}_\rho^{(r),j} = \frac{1}{\gamma_\rho^{(r),j}}(\gamma_0^{(r)}\boldsymbol{m}_0^{(r)} + N_\rho^j \bar{\boldsymbol{y}}_\rho^j),$$
$$(\boldsymbol{W}_\rho^{(r),j})^{-1} = (W_0^{(r)})^{-1} + N_\rho^j \boldsymbol{S}_\rho^j + N_\rho^j \boldsymbol{C}_\rho^j + \frac{\gamma_0^{(r)} N_\rho^j}{\gamma_0^{(r)} + N_\rho^j}(\bar{\boldsymbol{y}}_\rho^j - \boldsymbol{m}_0^{(r)})(\bar{\boldsymbol{y}}_\rho^j - \boldsymbol{m}_0^{(r)})^T.$$

The sufficient synthetic statistics in (25) are defined as:

$$N_\rho^j = \sum_i \hat{z}_{ij}\tilde{N}_i \sum_\beta \hat{\nu}^{i,j}(\rho,\beta),$$
$$\bar{\boldsymbol{y}}_\rho^j = \frac{1}{N_\rho^j}\sum_i \hat{z}_{ij}\tilde{N}_i \sum_\beta \hat{\nu}^{i,j}(\rho,\beta)\tilde{\boldsymbol{\mu}}_\beta^{(b),i},$$
$$\boldsymbol{S}_\rho^j = \frac{1}{N_\rho^j}\sum_i \hat{z}_{ij}\tilde{N}_i \sum_\beta \hat{\nu}^{i,j}(\rho,\beta)(\tilde{\boldsymbol{\mu}}_\beta^{(b),i} - \bar{\boldsymbol{y}}_\rho^j)(\tilde{\boldsymbol{\mu}}_\beta^{(b),i} - \bar{\boldsymbol{y}}_\rho^j)^T,$$
$$\boldsymbol{C}_\rho^j = \frac{1}{N_\rho^j}\sum_i \hat{z}_{ij}\tilde{N}_i \sum_\beta \hat{\nu}^{i,j}(\rho,\beta)(\tilde{\boldsymbol{\Lambda}}_\beta^{(b),i})^{-1}, \quad (26)$$

where $\hat{\nu}^{i,j}(\rho,\beta)$ has the same form as that in VHEM [25], and N_ρ^j is the expected number of samples that have been assigned to $\mathcal{M}_j^{(r)}$ with state ρ during the whole time.

From (25), as more samples are assigned to $\mathcal{M}_j^{(r)}$ with state ρ (i.e., N_ρ^j increases), the $\gamma_\rho^{(r),j}$ will increase and the covariance of posterior of $\boldsymbol{\mu}_\rho^{(r),j}$ will decrease; At the same time, the degree of freedom $\nu_\rho^{(r),j}$ will increase, which leads to increasing precision of the posterior of $\boldsymbol{\mu}_\rho^{(r),j}$. The update equation of $\boldsymbol{m}_\rho^{(r),j}$ is a mix between the prior and the soft sample mean $\bar{\boldsymbol{y}}_\rho^j$. Similarly, the update for $\boldsymbol{W}_\rho^{(r),j}$ is the mix between the prior and the soft sample covariance $\boldsymbol{S}_\rho^j$ and mean base covariance $\boldsymbol{C}_\rho^j$. Thus the optimization of the variational posterior involves cycling between two stages analogous to the E and M steps of EM algorithm (see Alg. 1).

D. Comparison with VHEM-H3M

We compare our VBHEM-H3M algorithm with the VHEM-H3M algorithm of [25] in this section. Firstly, our method VBHEM uses a Bayesian framework, compared to VHEM, which is not Bayesian. We provide priors for all the unknown parameters in the reduced model $\mathcal{M}^{(r)}$ and estimate an approximate posterior distribution of $\mathcal{M}^{(r)}$, rather than a point-estimate for each parameter as in VHEM.

Secondly, in addition to the prior on the reduced model $\mathcal{M}^{(r)}$, which is common for traditional Bayesian inference, we also assume priors on the input base model $\mathcal{M}^{(b)}$, while VHEM uses the point-estimate of the parameters of the input HMMs. When we assume a delta function prior for $\mathcal{M}_i^{(b)}$, then VBHEM is a Bayesian version of VHEM, where the point-estimate base models are inputs and Bayesian priors are placed on the reduced models. In this case, equality holds in (15) and

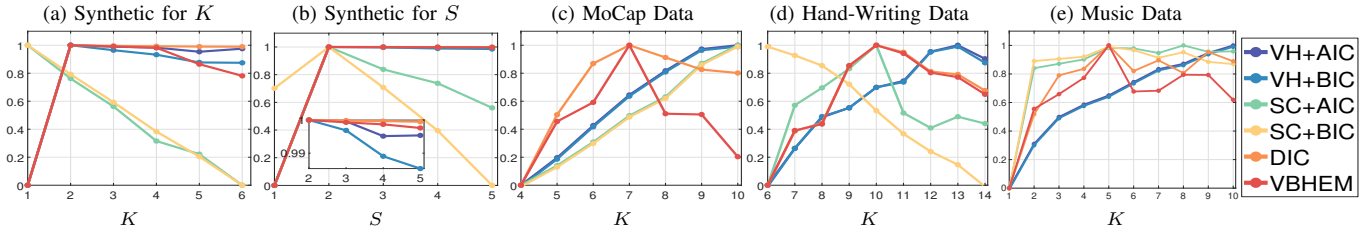


Fig. 2. Model selection results ($K^{(r)*}$, $S^{(r)*}$) from VH+AIC, VH+BIC, SC+AIC, SC+BIC, DIC and VBHEM on different datasets: (a), (b) Synthetic data, (c) Mocap data; (d) Hand-writing data; (e) Music data. The penalized log-likelihood scores are plotted for AIC, BIC, and DIC methods, and thus larger scores indicate better model fits (consistent with VBHEM). The scores for each method are normalized between [0,1] for better visualization. For VBHEM, the model selection curve for S in (b) uses (27) with $K^{(r)} = K^{(r)*}$, and the model selection curve for K is based on (28).

$c_\beta^{(b),i} = 1$, and thus the expectations in (17 -20) yield back the point-estimates.

Thirdly, a coefficient $c_\beta^{(b),i}$ is introduced in VBHEM in (20). It is worth emphasizing that $c_\beta^{(b),i} > 1$, which makes $\boldsymbol{C}_\rho^j$ in (26) "larger" (in the positive definite sense), i.e., variance $(\boldsymbol{\Lambda}_\rho^{(r),j})^{-1} \propto (\boldsymbol{W}_\rho^{(r),j})^{-1}$, increases, compared to VHEM where $c_\beta^{(b),i} = 1$. Note that this will help to mitigate the problem that variational inference generally underestimates the variance of the posterior density [35].

Fourthly, VBHEM enlarges (reduces) the responsibilities of the components with large (small) weight. The assignment variable $\hat{z}_{ij}$ in VHEM is

$$\hat{z}_{ij} = \frac{\omega_j^{(r)} \exp\left(N_i \mathcal{L}_{HMM}^{i,j}\right)}{\sum_{j'} \omega_{j'}^{(r)} \exp\left(N_i \mathcal{L}_{HMM}^{i,j'}\right)},$$

in contrast to (22) for VBHEM. For VBHEM, the power of $\tilde{\omega}_j^{(r)}$ in (22) increases the gap among the weights, e.g., the ratio between the maximum and minimum of $\tilde{\boldsymbol{\omega}}^{(r)}$, $b = \frac{\max_j \tilde{\omega}_j^{(r)}}{\min_j \tilde{\omega}_j^{(r)}} > 1$, is smaller than $b^{\tilde{N}_i}$, thus the probability of component with largest $\tilde{\omega}_j^{(r)}$ will increase, and vice versa for the smallest $\tilde{\omega}_j^{(r)}$.

Fifthly, as discussed in the next section, VBHEM can simultaneously perform model selection, while VHEM cannot.

E. Optimizing the Hyperparameters

Given a pair of $(K^{(r)}, S^{(r)})$, our model contains hyperparameters $\mathcal{P}^{(r)} = \{\alpha_0^{(r)}, \eta_0^{(r)}, \epsilon_{\rho,0}^{(r)}, \boldsymbol{m}_0^{(r)}, \boldsymbol{W}_0^{(r)}, \gamma_0^{(r)}, \nu_0^{(r)}\}$ (now assumed λ_0 is known and determined according to data). One approach for estimating the hyperparameters is to maximize the marginal log-likelihood of the data (i.e., empirical Bayes, type-II maximum likelihood), or a lower bound when the marginal log-likelihood is intractable. Applying this to our model, we maximize the expected lower bound $\bar{\mathcal{L}}(q)$, and proceed by firstly maximizing $\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)]$ under a fixed model structure $(K^{(r)}, S^{(r)})$ (see Appendix E for details of $\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)]$). For the continuous parameters (e.g., α_0), $\mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)]$ is maximized using gradient ascent. For the discrete parameters $K^{(r)}$ and $S^{(r)}$, we train on a range of possible $K^{(r)}$ and $S^{(r)}$, and select the pair that yields the highest $\bar{\mathcal{L}}(q)$. Recall from (11), we have

$$\log q^*(K^{(r)}, S^{(r)}) \propto \log p(K^{(r)}) + \log p(S^{(r)}|K^{(r)}) + \mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)].$$

Thus, the model selection can be done by:

1) For each candidate $K^{(r)}$, select the optimal

$$S^{(r)*}(K^{(r)}) = \arg\max_{S^{(r)}} \log p(S^{(r)}|K^{(r)}) + \mathbb{E}_{\boldsymbol{Y}}[\mathcal{L}_{(K^{(r)},S^{(r)})}(q^*)] \quad (27)$$

2) Select the optimal

$$K^{(r)*} = \arg\max_{K^{(r)}} \log q^*\left(K^{(r)}, S^{(r)*}(K^{(r)})\right). \quad (28)$$

The entire VBHEM algorithm is summarized in Alg. 2.

V. Experiments

We present experiment results on synthetic data and real data to demonstrate that our proposed VBHEM-H3M[1] can be effectively applied in several domains. The experiments on the synthetic data show the performance of VBHEM in various aspects, including estimating the parameters, automatically choosing the number of clusters K and the number of states S, sensitivity analysis of VBHEM to (K, S). The experiments on real data include a motion capture (MoCap) dataset[2], Eye Movement dataset [5], Hand-writing dataset[3] and Music dataset [53], showing that VBHEM correctly finds the number of clusters K. In the experiments, we remove components/states with weights lower than 10^{-3} (for all the cluster centers from all the compared methods). Finally, we compare our VBHEM-H3M using marginal likelihood (denoted as VBHEM), with VBHEM using DIC [54] (denoted as DIC), VHEM [25] with AIC [31] and BIC [32] (denoted as VH+AIC, and VH+BIC), PPK-SC [27] with AIC and BIC (denoted as SC+AIC, and SC+BIC) and CCFD algorithm [47] (see Appendix G-A for details of each compared method).

A. Synthetic Data

a) Experiment 1: We first consider a 2-dimensional case of a deceptively simple "toy" problem, which is also considered in [23, 55]. The ground truth is a 2D H3M with 2 components (HMMs), and each with 2 states. The transition matrices are

$$\boldsymbol{A}^{(1)} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}, \boldsymbol{A}^{(2)} = \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}. \quad (29)$$

The two HMMs share the same 2D Gaussian emission densities $\mathcal{N}(y|\boldsymbol{\mu}_s^{(i)}, \boldsymbol{\Sigma}_s^{(i)})$, $s{=}1,2, i{=}1,2$, where means $\boldsymbol{\mu}^{(i)}{=}[\boldsymbol{\mu}_1^{(i)}, \boldsymbol{\mu}_2^{(i)}]{=}\begin{bmatrix} 0 & 3 \\ 0 & 3 \end{bmatrix}$, and variances $\boldsymbol{\Sigma}_1^{(i)}{=}\boldsymbol{\Sigma}_2^{(i)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The weights and initial probabilities are all uniform.

The synthetic experiment proceeds as follows: (1) generate 20 sample sets from each HMM, where each sample set contains 25 sequences with length $\tau = 50$ [56]; (2) add noise $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, 0.1 \cdot \mathrm{I}_2)$ to each observation; (3) estimate the posteriors over HMM parameters for each noisy sample set via the *EMHMM toolbox* [4][4], resulting in 20 HMMs

[1] source code is released at https://doi.org/10.5281/zenodo.4468501
[2] http://mocap.cs.cmu.edu/
[3] https://archive.ics.uci.edu/ml/datasets/Character+Trajectories
[4] available at http://visal.cs.cityu.edu.hk/research/emhmm/

TABLE III
EXPERIMENT RESULTS FROM CLUSTERING SYNTHETIC DATA, AVERAGED OVER 1000 TRIALS. RI, ACC, OVER-EST, UNDER-EST ARE RAND-INDEX, ACCURACY, OVERESTIMATE AND UNDERESTIMATE, RESPECTIVELY. THE NUMBERS IN PARENTHESES ARE THE STANDARD DEVIATIONS.

	Ri ↑	Purity ↑	Acc % ↑		Over-est % ↓		Under-est % ↓	
			K	S	K	S	K	S
VHEM	0.827(.01)	0.832(.01)	65.9(.02)	20.0(.00)	0.6(.01)	60.0(.00)	33.5(.01)	20.0(.00)
VH+AIC	**1.000(.00)**	**1.000(.00)**	**100.0(.00)**	90.5(.29)	0.0(.00)	9.5(.29)	0.0(.00)	0.0(.00)
VH+BIC	**1.000(.00)**	**1.000(.00)**	**100.0(.00)**	98.0(.14)	0.0(.00)	2.0(.14)	0.0(.00)	0.0(.00)
SC+AIC	0.508(.10)	0.520(.10)	4.0(.20)	**100.0(.00)**	0.0(.00)	0.0(.00)	96.0(.20)	0.0(.00)
SC+BIC	0.487(.00)	0.500(.00)	0.0(.00)	**100.0(.00)**	0.0(.00)	0.0(.00)	100.0(.00)	0.0(.00)
DIC	0.999(.02)	**1.000(.00)**	99.5(.07)	99.5(.07)	0.5(.07)	0.5(.07)	0.0(.00)	0.0(.00)
CCFD	0.990(.10)	0.990(.10)	99.0(.10)	-	0.0(.00)	-	1.0(.10)	-
VBHEM (ours)	**1.000(.00)**	**1.000(.00)**	**100.0(.00)**	**100.0(.00)**	0.0(.00)	0.0(.00)	0.0(.00)	0.0(.00)
VBHEM ($\gamma_0^{(r)}, \nu_0^{(r)} \to \infty$)	0.487(.00)	0.500(.00)	0.0(.00)	0.0(.00)	0.0(.00)	0.0(.00)	100(.00)	100(.00)

TABLE IV
EXPERIMENT RESULTS FROM CLUSTERING SYNTHETIC DATA USING VBHEM FOR DIFFERENT PAIRS OF (K, S), AVERAGED OVER 1000 TRIALS.

	(K,S)	Ri ↑	Purity ↑	Acc % ↑	
				K	S
VHEM	(3,3)	0.909(.00)	0.899(.00)	79.5(.01)	17.4(.03)
	(3,5)	0.897(.01)	0.899(.00)	51.5(.20)	14.5(.02)
	(5,3)	0.846(.01)	0.784(.01)	27.0(.14)	13.2(.01)
VH+AIC	(3,3)	0.977(.07)	0.967(.10)	90.5(.29)	81.7(.37)
	(3,5)	0.964(.03)	**1.000(.00)**	30.2(.46)	41.3(.30)
	(5,3)	0.976(.02)	0.999(.00)	15.9(.37)	56.9(.41)
VH+BIC	(3,3)	0.977(.07)	0.967(.10)	90.5(.29)	81.7(.37)
	(3,5)	0.976(.02)	0.998(.01)	35.4(.48)	58.2(.31)
	(5,3)	0.970(.03)	0.983(.06)	15.9(.37)	62.6(.39)
SC+AIC	(3,3)	0.986(.02)	**1.000(.00)**	75.3(.43)	95.0(.22)
	(3,5)	0.995(.01)	**1.000(.00)**	84.9(.36)	60.0(.49)
	(5,3)	0.963(.02)	0.988(.02)	15.4(.36)	85.9(.35)
SC+BIC	(3,3)	0.986(.02)	**1.000(.00)**	75.3(.44)	95.0(.22)
	(3,5)	0.987(.03)	0.997(.01)	80.0(.40)	55.0(.50)
	(5,3)	0.964(.02)	0.982(.03)	25.4(.44)	90.5(.29)
DIC	(3,3)	0.999(.00)	**1.000(.00)**	95.0(.22)	98.8(.05)
	(3,5)	**1.000(.00)**	**1.000(.00)**	**100.0(.00)**	53.3(.37)
	(5,3)	0.910(.11)	0.800(.23)	70.9(.45)	32.2(.32)
CCFD	(3,3)	0.989(.05)	0.983(.07)	95.6(.21)	-
	(3,5)	0.955(.09)	0.933(.13)	80.2(.40)	-
	(5,3)	0.975(.04)	0.940(.10)	70.2(.46)	-
VBHEM (ours)	(3,3)	**1.000(.00)**	**1.000(.00)**	**100.0(.00)**	**100.0(.00)**
	(3,5)	**1.000(.00)**	**1.000(.00)**	**100.0(.00)**	**96.0(.16)**
	(5,3)	**0.996(.02)**	**0.990(.04)**	**95.6(.21)**	**92.3(.27)**

$(K^{(b)} = 2 \times 20, S^{(b)} = 2)$; (4) use these posteriors as the input HMMs, and run VBHEM with $K^{(r)} \in [1, 6]$ and $S^{(r)} \in [1, 5]$ to automatically determine the model structure. We run steps (1-4) 1000 times with VBHEM with different random initializations. We set $N = 100 \times K^{(b)}$, and $\lambda_0 = 1$. For comparison we also run steps (1-3) with VHEM, PPK-SC and CCFD for 1000 trials. Note that VHEM clusters HMMs for each $(K^{(r)}, S^{(r)})$ pair separately, and likewise for PPK-SC. For VHEM and PPK-SC, we compute AIC and BIC (denoted as VH+AIC, VH+BIC and SC+AIC, SC+BIC) to select the model structure among all $K^{(r)}$ and $S^{(r)}$ combinations with smallest AIC or BIC. We evaluate the methods using five criteria: the accuracy of selecting the correct K or S, the percentage of over-estimating K or S, the percentage of under-estimating K or S, the Rand-index (Ri) [57], which measures the correctness of the computed clustering against the ground-truth clustering, and the Purity [58], which measures the extent to which clusters contain a single class.

The average results over 1000 trials are summarized in Table III. Our VBHEM-H3M is the most consistent (100% accuracy) in selecting the correct number of components $K = 2$ and number of states $S = 2$, as compared to the other methods. Furthermore, VBHEM obtains perfect Rand-index (Ri) of 1 and perfect Purity of 1, and thus finds the correct clustering consistently. For VHEM, note that we also pruned out the components and states with no samples assigned. VHEM has 65.9% accuracy in selecting the true number of components K. VHEM overestimates the number of states S in 60.0% of the trials, i.e., the maximum-likelihood-based VHEM is likely to over-fit the emission model. The over-fitting problem is mitigated using Bayesian methods, i.e., VBHEM, or by adding complexity penalization terms, such as AIC and BIC. Using complexity terms, VH+AIC and VH+BIC, are slightly less accurate than VBHEM in estimating S.

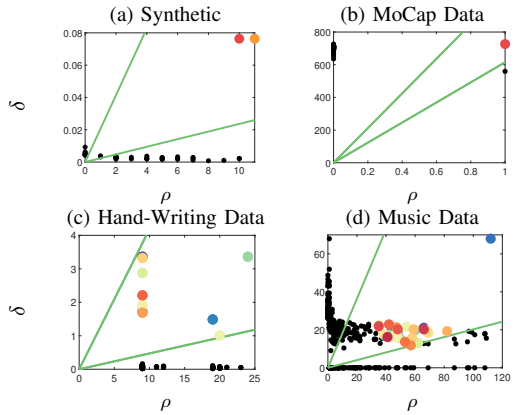


Fig. 3. The decision graph in CCFD algorithm for (a) Synthetic data ($K = 2$), (b) MoCap data set ($K = 1$), (c) Hand-writing data set ($K = 10$), (d) Music data set ($K = 32$).

Spectral clustering (SC+AIC, SC+BIC) performed worst in selecting K, resulting in lower Rand-index and Purity, but was perfect at selecting S. Since SC clusters the HMMs learned with different states numbers, it can perfectly select S possibly because the input HMMs under that S are learned well. Thus, SC-AIC and SC+BIC cannot obtain the true K are because the AIC penalty is too heavy for SC to select K, and the BIC penalty is even worse. DIC performs well at selecting the number of components and states, but slightly worse than VBHEM. Note that DIC takes advantage of the good posterior estimate of the reduced model from VBHEM. CCFD also works well when selecting K (99% accuracy), and shows good clustering performance with Rand-index 0.990 and Purity 0.990. The decision graph shows CCFD can successfully find two cluster centers (see Fig. 3a).

TABLE V
EXPERIMENT RESULTS FROM CLUSTERING MOCAP DATA SET, AVERAGED OVER 20 TRIALS.

	Ri ↑	Purity↑	Acc↑	Over-est↓	Under-est↓
			K %		
VHEM	0.823(.00)	0.534(.01)	14(.01)	43(.00)	43(.01)
VH+AIC	0.827(.01)	0.502(.04)	0(.00)	100(.00)	0(.00)
VH+BIC	0.827(.01)	0.502(.04)	0(.00)	100(.00)	0(.00)
VH/τ+AIC	0.827(.01)	0.502(.04)	0(.00)	100(.00)	0(.00)
VH/τ+BIC	0.827(.01)	0.502(.04)	0(.00)	100(.00)	0(.00)
SC+AIC	0.795(.01)	0.349(.03)	0(.00)	100(.00)	0(.00)
SC+BIC	0.794(.01)	0.348(.03)	0(.00)	100(.00)	0(.00)
SC/τ+AIC	0.794(.01)	0.347(.03)	0(.00)	100(.00)	0(.00)
SC/τ+BIC	0.794(.01)	0.347(.03)	0(.00)	100(.00)	0(.00)
DIC	0.993(.01)	0.991(.03)	40(.49)	55(.50)	5(.22)
DIC/τ	0.898(.04)	0.688(.08)	0(.00)	0(.00)	100(.00)
CCFD	0.129(.00)	0.143(.00)	0(.00)	0(.00)	100(.00)
VBHEM (ours)	**0.994(.01)**	**1.000(.00)**	**90(.31)**	10(.31)	0(.00)

To show the effectiveness of the prior distributions on the input, we also train VBHEM where the priors of the mean and precision matrices are collapsed into delta function priors by letting $\gamma_0^{(r)}, \nu_0^{(r)} \to \infty$, which is equivalent to using point-wise estimates of the mean and precision. The results are in the last row of Table III, and always underestimate K and S, i.e., the HMMs could not be separated. Thus, using prior distributions on the input HMMs can better handle the uncertainty and leads to better clustering results.

Fig. 2a shows the model selection criteria for varying $K^{(r)}$ and fixed $S^{(r)*}$, while Fig. 2b shows the criteria for fixed $K^{(r)*}$ and varying $S^{(r)}$. In Fig. 2a, VH+AIC, VH+BIC, DIC and VBHEM all successfully find the true number of components, while SC+AIC and SC+BIC underestimate the number of components. In Fig. 2b, the six methods all successfully select the true number of states. Finally, Appendix G-B shows an example result.

b) Experiment 2: We next test the robustness of VBHEM for different settings of true number of clusters and true number of states. We generate the synthetic data for different K and S by: (1) randomly generating K HMMs, each with S states; (2) use *EMHMM toolbox* [4] to learn 20 HMMs ($K^{(b)} = 20 \times K$) for each given true HMM. Here we test three pairs of $(K, S) \in \{(3,3), (3,5), (5,3)\}$. For each (K, S) pair, we run VBHEM with $K^{(r)} \in [1, 10]$ and $S^{(r)} \in [1, 10]$ for 1000 times with different random initializations. For comparison we also run VHEM, PPK-SC, and CCFD for 1000 trials. Note that the number of states can be different for each HMM component in the reduced H3M $\mathcal{M}^{(r)}$; e.g., for the selected $S^{(r)} = 5$, the final number of states can be less than 5, since some states may be pruned out in Step 8 of Alg. 2.

Table IV shows the average performance over 1000 trials, showing that our VBHEM-H3M method outperforms other methods. The Rand-index, Purity and Accuracy do not change significantly with different (K, S), and thus VBHEM is robust to changes in the number of clusters/states. DIC performs well for setting $(3,3)$, but the accuracy of S decrease sharply as the model complexity in S increases, and likewise for K. VHEM, VH+AIC and VH+BIC have similar Rand-index and Purity, and perform better when (K, S) are small, but are not robust when they increase. Likewise SC+AIC and SC+BIC are not robust when K and S are increased. CCFD also obtains good Rand-index and Purity, but the accuracy decreases as K increases, which shows that it is less robust than our method.

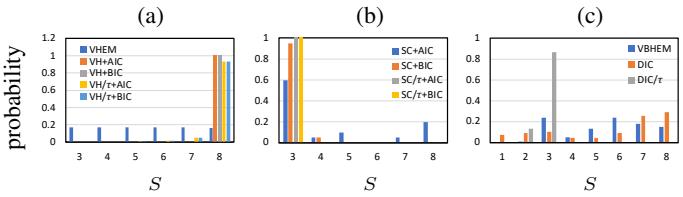
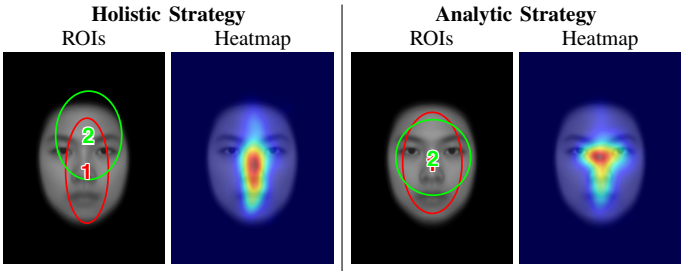


Fig. 4. Selection of number of states for different methods in MoCap experiment, including (a) VHEM, (b) SC, (c) VBHEM and DIC.

B. Motion Capture Data

This experiment uses Motion Capture data (MoCap), which are time series representing human locomotions and actions. We use 63 motion examples spanning 7 different classes (*sit*, *run*, *jump*, *yoga*, *swim*, *dance*, and *baseball*). Each example is a sequence of 123-dimensional vectors representing the (x, y, z)-coordinates of 41 body markers tracked spatially through time. For each example, we learn the HMM posteriors over parameters using *EMHMM toolbox*. This HMM summarizes the appearance (Gaussian emission) and dynamics (state prior and transition) of the particular motion sequence it represents. We then use these posteriors as input to our algorithm to find the true number of motion classes. For running VBHEM, we set $K^{(r)} \in [4, 10]$ and $S^{(r)} \in [3, 8]$, $N = 10K^{(b)}$, $\tau = 10$, and $\lambda_0 = 1$. This experiment is repeated 20 times with different random initialization, and the average results are reported in Table V.

Our VBHEM obtains the true number of motion classes with 90% accuracy and perfect Purity 1, and outperforms other methods. Fig. 2c plots the model selection curve, indicating that VBHEM has a peak at $K = 7$, leading to the correct choice of number of clusters. VHEM has 14% accuracy in selecting the correct $K = 7$, while overestimating and underestimating K equally 43% of the time, which is close to random chance. Fig. 2c shows the model selection criteria for the various methods versus K. VH+AIC, VH+BIC, SC+AIC and SC+BIC have a tendency to overestimate K, as their curves alway increase as K increases, resulting in selection of $K = 10$. This is because the log-likelihood approximation used by VHEM and PPK-SC increase as the model complexity increases, and the increase cannot be effectively penalized by the AIC or BIC terms. Moreover, it demonstrates that as the data becomes noisier (as in real-world data) and the dimension increases, our method performs better than BIC and AIC. DIC achieves an Accuracy of 40% in determining K, which performs better than other methods, but is still inferior to our VBHEM. Although DIC has a 55% probability of overestimating K, it obtains high Rand-index and Purity, which means DIC is still forming consistent groups of data. CCFD performs the worst in this experiment and always underestimates K. The main reason is that the cut-off d_c is too small; $d_c = d_{min} + (d_{max} - d_{min}) * p$, where d_{min} and d_{max} are respectively the minimum and maximum among all distances of two HMMs, and p is searched in $[1\%, 20\%]$. However, in this experiment d_{min} is about 560, d_{max} is about 730, and the most of distances are centered on 700 (see Fig. 3b). Thus d_c is still small even when $p = 20\%$ (about 600), and thus most of the densities are $\rho_i = 0$, which causes CCFD to fail to find the cluster centers.



transition matrix

	To Red	To Green
Priors	.84	.16
From Red	.99	.01
From Green	.05	.95

transition matrix

	To Red	To Green
Priors	.88	.12
From Red	.16	.84
From Green	.25	.75

Fig. 5. HMMs estimated from VBHEM clustering of eye gaze data: (left) holistic and (right) analytic eye gaze strategies.

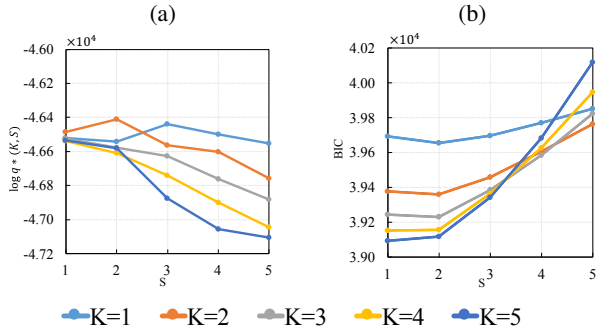


Fig. 6. Model selection curves for the experiments on eye movement data: (a) VBHEM and (b) BIC.

The analysis of the results show that VH+AIC, VH+BIC, SC+AIC, SC+BIC and DIC might have failed because the log-likelihood approximation term is too large compared to the penalty terms. To alleviate this problem, we modify these methods by dividing this term by the length of the sequence τ. The results are shown in Table V, denoted as VH/τ+BIC, VH/τ+AIC, SC/τ+BIC, SC/τ+AIC and DIC/τ. The clustering results do not change or even get worse, which shows this process does not help.

Finally, Fig. 4 presents the selected S for each method. VBHEM has a high probability to select $S = 3$ or $S = 6$ in the reduced H3M. VHEM is not effective at selecting S, and thus the percentages are mostly uniform. When using the model selection methods, VHEM tends to select the largest candidate S, while PPK-SC tends to select the smallest candidate S. DIC is more likely to select $S = 7$ or $S = 8$, while DIC/τ mostly selects 3 states. Among the methods with decent Rand-index, our method selects a more parsimonious model while also having high Rand-index (0.998). While there is no ground-truth value for the number of states, a more parsimonious model may be preferred since it more succinctly summarizes the data, and also is easier to interpret compared to a model with many states.

C. Eye Movement Data

In this experiment, we use the Eye Movement dataset from [5], which is a collection of eye fixation trajectories from 68 participants (34 older adults and 34 young adults) while performing a recognition experiment on face images. An eye movement data is a sequence of 2-dimensional vectors representing the (x, y)-coordinates of the location of an eye fixation on the face image over time. Previous work [59] models the regions of interest (ROIs) as GMMs and study the correlation between GMMs and cognitive abilities. [4, 5, 60] also consider the temporal dynamics, where each participant's eye movements are modeled with an HMM, including both person-specific ROIs and transitions among the ROIs. Individual HMMs are then clustered using VHEM into two groups [4, 5] or three groups [60] in order to discover common eye gaze strategies among the participants, but requires setting the number of group and number of states by hand.

TABLE VI
THE NUMBER OF PARTICIPANTS BELONGING TO HOLISTIC AND ANALYTIC PATTERN THOUGH VBHEM AND VHEM.

		Holistic	Analytic
VBHEM (Ours)	Total	39	29
	Old	21	13
	Young	18	16
VHEM [5]	Total	33	35
	Old	21	13
	Young	12	22

1) Clustering results: We use VBHEM to automatically choose the number of clusters and states to discover common patterns among individuals. We set $K^{(r)} \in [1, 5]$ and $S^{(r)} \in [1, 5]$, $N = 10K^{(b)}$, and $\lambda_0 = 1$ for running VBHEM and VH+BIC. VBHEM automatically selects an H3M with 2 components and 2 states, and the estimated group HMMs are shown in Fig. 5. Fig. 6a plots the model selection curve for VBHEM. The number of selected states is inversely proportional to the number of components. Given $K = 1$, then the best selection is $S = 3$; increasing K will decrease the best selection of S. The reason is that when using an H3M to model the given data, more components means less data for each component to model, and thus less hidden states are needed within each component. The pair $(K, S) = (2, 2)$ has the maximum $\log q^*(K, S)$ among all candidate model structures, and thus is the overall selected model structure. Compared with BIC (see Fig. 6b), $S = 2$ is selected when $K \in \{1, 2, 3\}$, and $S = 1$ is selected when $K \in \{4, 5\}$. The final selection is $(5, 1)$; BIC cannot effectively penalize the growth of the log-likelihood caused by increasing K, which leads to the selection of the largest K.

2) Results analysis: The HMMs clustering results are displayed in Fig. 5. The pattern in Fig. 5 (left) resembles a holistic pattern, a scan path typically started at the nose/mouth region, and then staying around the same region. In contrast, the pattern in Fig. 5 (right) resembles an analytic pattern, a scan path typically started around the face center, and then transitioned to the eye region. (i.e., more frequent fixation transitions between the eyes; [4]). The differences can also be seen in the corresponding fixation heatmap shown in Fig. 5. The two representative HMMs significantly differ from each other based on the KL divergence test [4]; using data from holistic HMMs, $t(38)$=7.10, p<0.001; using data from analytic HMMs, $t(28)$=6.10, p<.001.

Table VI shows the number of young and older participants belonging to holistic and analytic pattern, and compare with the VHEM results from [5]. There are 39 adults assigned to the holistic pattern and 29 adults assigned to the analytic pattern. Comparing with the results from [5] using VHEM, the

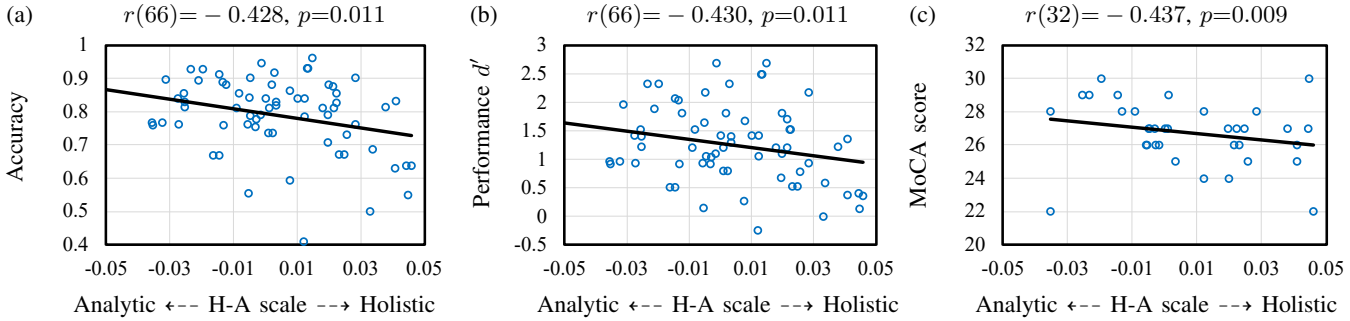


Fig. 7. Correlation analysis between H-A score and (a) recognize accuracy, (b) recognition performance d', (c) MoCA score.

main difference is that VBHEM assigns 6 additional young adults into the holistic group. This difference may be due to our analytic HMM focusing more on the regions around the two eyes than the analytic HMM in [5] (our transition matrix shows a larger probability to stay in the eye regions, while the transition matrix in [5] is more uniform). Thus, those participants that do not show significant focus on the eyes are assigned to holistic pattern by our algorithm.

As we may be interested in individual differences in eye movement patterns, we also quantify the degree to which a subject's eye-movement pattern resembles the representative holistic and analytic HMMs using the H-A scale [5]. For each subject, the H-A scale measures a normalized difference between the log-likelihoods of a subject's eye movement data being generated by the representative holistic and analytic HMM models,

$$\text{H-A Scale} = \frac{\log p(Y|\mathcal{M}_h) - \log p(Y|\mathcal{M}_a)}{|\log p(Y|\mathcal{M}_h) + \log p(Y|\mathcal{M}_a)|},$$

where Y is the fixation data for the individual, and $\mathcal{M}_h$ and $\mathcal{M}_a$ are the representative holistic and analytic HMMs learned by VBHEM. A positive H-A value indicates that the subject's gaze pattern is more similar to a holistic pattern, while a negative value indicates similarity to an analytic pattern. According to the H-A values, older adults tended to exhibit holistic patterns (M=0.0089), and younger adults tended to exhibit analytic patterns (M=-0.0029), and this difference was statistically significant according to a two-sample t-test, $t(66)$=2.139, p=0.036. The same comparison between older/young adults using the H-A scale built using VHEM (i.e. from [5]) yielded a marginal difference (older adults M=0.0032, young adults M=-0.010), $t(66)$=1.916, p=0.060. We also performed correlation analysis to see how eye gaze patterns (as quantified by H-A scale) are correlated with the subjects' behavioral data, as in [5]. The subjects' recognition accuracy was negatively correlated with H-A scale, $r(66)=-0.428$, p=.011. The lower the recognition accuracy, the more holistic the eye-gaze pattern (see Fig. 7a). In addition, the participants' recognition performance d' was also negatively correlated with H-A scale, $r(66)=-0.430$, p=.011 (see Fig.7b). Finally, the MoCA scores[5] for the older adults was negatively correlated with the H-A scale, $r(32)=-0.437$, p=.009. In other words, the lower the MoCA score (the more cognitive impairment), the more holistic the pattern (see Fig.7c). These results from

[5]Montreal Cognitive Assessment (MoCA) is a valid brief assessment tool for screening of people with mild cognitive impairment, and 22 points or more (out of 30) is a normal score

TABLE VII
EXPERIMENT RESULTS FOR CLUSTERING HANDWRITING TRAJECTORIES, AVERAGED OVER 10 TRIALS.

	Ri ↑	Purity↑	Acc↑	Over-est↓	Under-est↓
			K %		
VHEM	0.966(.00)	0.864(.00)	11(.00)	44(.00)	44(.00)
VH+AIC	0.980(.00)	**1.000(.00)**	0(.00)	100(.00)	0(.00)
VH+BIC	0.980(.00)	**1.000(.00)**	0(.00)	100(.00)	0(.00)
VH/τ+AIC	0.958(.01)	0.816(.06)	0(.00)	0(.00)	100(.00)
VH/τ+BIC	0.899(.01)	0.600(.00)	0(.00)	0(.00)	100(.00)
SC+AIC	0.930(.04)	0.751(.11)	10(.32)	0(.00)	90(.32)
SC+BIC	0.904(.02)	0.630(.07)	0(.00)	0(.00)	100(.00)
SC/τ+AIC	0.895(.02)	0.589(.03)	0(.00)	0(.00)	100(.00)
SC/τ+BIC	0.895(.02)	0.589(.03)	0(.00)	0(.00)	100(.00)
DIC	**0.985(.00)**	0.942(.02)	**70(.48)**	30(.48)	0(.00)
DIC/τ	0.964(.03)	0.810(.09)	20(.42)	0(.00)	80(.42)
CCFD	0.982(.00)	0.940(.05)	40(.52)	0(.00)	60(.52)
VBHEM (ours)	**0.985(.00)**	0.942(.02)	**70(.48)**	30(.48)	0(.00)

VBHEM are consistent with the previous study [4, 5] – two strategies of eye movements (holistic and analytic patterns) in face recognition tasks are discovered by clustering HMMs, and the H-A scales are negatively correlated with recognition performance and MoCA. However, here we use VBHEM to automatically determine the number of clusters and the number of states, whereas [5] set these values by hand.

D. On-Line Hand-Writing Data Set

In this experiment, we evaluate VBHEM for clustering characters from the Character Trajectories Data Set, which consists of 2858 examples for 20 characters from the same writer. Each example is the trajectory of one character that corresponds to a single pen-down segment. The data was captured using a WACOM tablet at 200Hz. and consists of (x, y)-coordinates and pen tip force, and the data has been numerically differentiated and Gaussian smoothed [61].

In consideration of the aim for VBHEM is to cluster data while automatically choosing the number of clusters. We simplify the experiment by only selecting 10 character types. For each character, we randomly select 25 examples to learn an HMM, and repeat 10 times, resulting in a base H3M with $K^{(b)} = 100$ components. We then perform clustering using VBHEM, VHEM, PPK-SC and CCFD. For VBHEM, we set $K^{(r)} \in [6, 14]$, $S^{(r)} = 6$, $N = 10K^{(b)}$, $\tau = 100$, and $\lambda_0 = 1$. The experiment is repeated 10 times with different initializations, and the average results are in Table VII.

VBHEM and DIC obtain the same best performance among the compared methods. Indeed the model selection curves for VBHEM and DIC (Fig. 2d) are coincident. Comparatively, our method of model selection is more straightforward than DIC, which requires further approximation on top of the VB framework.

TABLE VIII
EXPERIMENT RESULTS FROM CLUSTERING MUSIC GENRE DATASET, AVERAGED OVER 10 TRIALS.

	Ri ↑	Purity ↑	Acc ↑	Over-est ↓	Under-est ↓
			K %		
VHEM	0.641(.02)	0.435(.03)	10(.00)	50(.00)	40(.00)
VH+AIC	0.735(.03)	0.502(.05)	0(.00)	100(.00)	0(.00)
VH+BIC	0.735(.03)	0.502(.05)	0(.00)	100(.00)	0(.00)
VH/τ+AIC	0.827(.01)	0.502(.04)	0(.00)	100(.00)	0(.00)
VH/τ+BIC	0.827(.01)	0.502(.04)	0(.00)	100(.00)	0(.00)
SC+AIC	0.742(.01)	0.424(.04)	0(.00)	100(.00)	0(.00)
SC+BIC	0.679(.08)	0.384(.06)	10(.32)	50(.53)	40(.52)
SC/τ+AIC	0.432(.12)	0.260(.04)	0(.00)	0(.00)	100(.00)
SC/τ+BIC	0.470(.10)	0.276(.04)	0(.00)	0(.00)	100(.00)
DIC	0.757(.02)	0.554(.05)	20(.42)	60(.52)	20(.42)
DIC/τ	0.197(.00)	0.200(.00)	00(.00)	0(.00)	100(.00)
CCFD	0.427(.30)	0.385(.24)	0(.00)	40(.52)	60(.52)
VBHEM (ours)	**0.870(.03)**	**0.591(.07)**	**60(.52)**	30(.50)	10(.32)

VHEM by itself cannot perform model selection, and always uses the given number of components. VH+AIC and VH+BIC always overestimate the number of clusters (see Fig. 2d), which indicates that the penalties are too small compared with the log-likelihood. Although they overestimate the number of clusters, the Purity is 100%, indicating that each cluster still consists of only one character. All the PPK-SC methods underestimate the number of clusters, which leads to worse Rand-index and Purity. Moreover, normalizing VHEM and PPK-SC by the length of the sequences (VH/τ and SC/τ) always underestimates the number of clusters, which implies that the complexity penalties are too heavy. Finally, CCFD performs slightly worse than our method in terms of Rand-index and Purity. CCFD correctly estimates K in 40% of the trials, and underestimates K otherwise. For example, Fig. 3c plots the decision graph that successfully finds $K = 10$ cluster centers in one trial.

E. Music Data Set

In this experiment, we evaluate VBHEM-H3M for clustering songs from different music genres. As H3Ms allow to account for timbre (i.e., through the Gaussian emission process) as well as longer term temporal dynamics (i.e., through the HMM hidden state process), when modeling musical signals. Thus, clustering the represented H3Ms of songs is expected to reveal the true genre group.

The music dataset from [53] contains 10 genres, each with 100 songs stored at 22,050 Hz, 16-bit. The acoustic content of a song is represented as a time series of audio features, by computing the first 13 Mel frequency cepstral coefficients (MFCCs) [1] over half-overlapping windows of 46ms of audio signal, and augmented with first and second instantaneous derivatives. The song is then represented as a collection of audio fragments, which are sequences of 125 audio features (about 6 seconds of audio), using a dense sampling with 80% overlap. Each song is represented by 6 HMMs.

In this experiment, we use 5 genres (*hip-hop*, *classical*, *jazz*, *metal* and *country*), and randomly choose 10 songs ($K^{(b)} = 300$) as the inputs in each run. We repeat each method 10 times with different inputs, and the average results are shown in Table VIII. Our VBHEM-H3M outperforms other methods in terms of Rand-index, Purity and Accuracy. VHEM by itself always estimates the number of clusters to be the given number of clusters. Regardless of the information criteria used, VH always overestimates the number of clusters, and normalizing by the sequence length (VH/τ+AIC and VH/τ+BIC) only slightly increases the Rand-index. Although SC/τ+AIC selects the true number of cluster K in one trial, the 70% probabilities of underestimating K still leading to worse Rand-index and Purity than SC+AIC. Moreover, as the penalty increase, from SC+AIC to SC/τ+BIC, the performances get worse, which implies that the penalty is not appropriate. DIC has slightly lower Purity than our method and normalizing with the length of sequence also does not help. CCFD could not find the true number of genres K. In 4 trials, K was overestimated, e.g., the decision graph of one trial where CCFD selected $K = 32$ is in Fig. 3d. In the remaining 6 trials, CCFD failed to separate the music HMMs and formed only one cluster. Fig. 2e shows the model selection curve for VBHEM, VH+BIC, SC+BIC and DIC. VBHEM and DIC have peaks at $K = 5$, while VH+BIC and SC+BIC overestimate the number of clusters.

VI. CONCLUSIONS

We have derived the VBHEM algorithm for clustering HMMs, which automatically determines the number of clusters and the number of states. We show the efficacy of VBHEM on both synthetic dataset and real-world datasets, including motion capture, eye fixation sequences, character trajectories, and music. For the synthetic datasets considered, our VBHEM recovers the correct number of components/states in the H3M model, and finds good posterior estimates of the component HMMs. For the real datasets, we obtained results for clustering and model selection that are better or comparable to other methods. For future work, we now use the same value of $S^{(r)}$ for all the reduced HMMs, and we can consider using different values of $S_j^{(r)}$ for each component j, but this requires a more efficient search process over K and S to make it scalable.

ACKNOWLEDGMENT

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (GRF 17609117), and by a Strategic Research Grant from City University of Hong Kong (Project No. 7005218).

REFERENCES

[1] L. Rabiner, *Fundamentals of Speech Recognition*. PTR Prentice Hall, 1993.

[2] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[3] J.-J. Aucouturier and S. Mark, "Segmentation of musical signals using hidden Markov models," in *Proc. 110th Conv. Audio Eng. Soc.*, 2001.

[4] T. Chuk, A. B. Chan, and J. H. Hsiao, "Understanding eye movements in face recognition using hidden Markov models," *J. Vision*, vol. 14, no. 11, pp. 8–8, 2014.

[5] C. Y. Chan, A. B. Chan, T. M. Lee, and J. H. Hsiao, "Eye-movement patterns in face recognition are associated with cognitive decline in older adults," *Psychon. Bull. Rev.*, pp. 1–8, 2018.

[6] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational

biology: Applications to protein modeling," *J. Mol. Biol.*, vol. 235, no. 5, pp. 1501–1531, 1994.

[7] R. J. Boys and D. A. Henderson, "A Bayesian approach to DNA sequence segmentation," *Biometrics*, vol. 60, no. 3, pp. 573–581, 2004.

[8] Y. Tian and Z. Wang, "H∞ performance state estimation for static neural networks with time-varying delays via two improved inequalities," *IEEE Trans. Circuits Syst. II, Exp. Briefs,*, vol. 68, no. 1, pp. 321–325, 2020.

[9] Y. Tian and Z. Wang, "Extended dissipativity analysis for Markovian jump neural networks via double-integral-based delay-product-type Lyapunov functional," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020.

[10] Z. Zhang, D. Wang, and J. Gao, "Learning automata-based multiagent reinforcement learning for optimization of cooperative tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020.

[11] T. Chuk, K. Crookes, W. G. Hayward, A. B. Chan, and J. H. Hsiao, "Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures," *Cognition*, vol. 169, pp. 102–117, 2017.

[12] T. Chuk, A. B. Chan, and J. H. Hsiao, "Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling," *Vision Res.*, vol. 141, pp. 204–216, 2017.

[13] J. An and J. H. Hsiao, "Modulation of mood on eye movement pattern and performance in face recognition," *Emotion*, vol. 21(3), pp. 617–630, 2021.

[14] J. H. Hsiao, J. An, Y. Zheng, and A. B. Chan, "Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements," *Cognition*, vol. 211, 104616, 2021.

[15] J. H. Hsiao, A. B. Chan, J. An, S.-L. Yeh, and J. Li, "Understanding the collinear masking effect in visual search through eye tracking," *Psychon. Bull. Rev.*, 2021, (in press).

[16] J. Zhang, A. B. Chan, E. Y. Lau, and J. H. Hsiao, "Individuals with insomnia misrecognize angry faces as fearful faces due to missing the eyes: An eye-tracking study," *Sleep*, vol. 42, no. 2, p. zsy220, 2019.

[17] F. H. Chan, T. J. Barry, A. B. Chan, and J. H. Hsiao, "Understanding visual attention to face emotions in social anxiety using hidden Markov models," *Cogn. Emot.*, vol. 34(8), pp. 1704–1710, 2020.

[18] J. H. Hsiao, H. Lan, Y. Zheng, and A. B. Chan, "Eye movement analysis with hidden markov models (emhmm) with co-clustering," *Behavior Research Methods*, pp. 1–14, 2021.

[19] F. H. Chan, H. Suen, J. H. Hsiao, A. B. Chan, and T. J. Barry, "Interpretation biases and visual attention in the processing of ambiguous information in chronic pain," *Eur. J. Pain*, vol. 24, no. 7, pp. 1242–1256, 2020.

[20] F. H. Chan, T. Jackson, J. H. Hsiao, A. B. Chan, and T. J. Barry, "The interrelation between interpretation biases, threat expectancies and pain-related attentional processing," *Eur. J. Pain*, vol. 24(10), pp. 1956–1967, 2020.

[21] T. Chuk, A. B. Chan, S. Shimojo, and J. H. Hsiao, "Eye movement analysis with switching hidden Markov models," *Behav. Res. Methods*, vol. 52, pp. 1026–1043, 2020.

[22] Y. Qi, J. W. Paisley, and L. Carin, "Dirichlet process HMM mixture models with application to music analysis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–465.

[23] P. Smyth, "Clustering sequences with hidden Markov models," in *Proc. NeurIPS*, 1997, pp. 648–654.

[24] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41(1), pp. 164–171, 1970.

[25] E. Coviello, A. B. Chan, and G. R. G. Lanckriet, "Clustering hidden Markov models with variational HEM," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 697–747, Jan. 2014.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[27] T. Jebara, Y. Song, and K. Thadani, "Spectral clustering and embedding with hidden Markov models," in *Proc. ECML*, 2007, pp. 164–175.

[28] M. Soruri, J. Sadri, and S. H. Zahiri, "Gene clustering with hidden Markov model optimized by PSO algorithm," *Pattern Anal. Appl.*, vol. 21, no. 4, pp. 1121–1126, 2018.

[29] M. Kanaan, K. Benabdeslem, and H. Kheddouci, "A generative time series clustering framework based on an ensemble mixture of HMMs," in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2020, pp. 793–798.

[30] Y. Chen, J. Ye, and J. Li, "Aggregated Wasserstein distance and state registration for hidden Markov models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2133–2147, 2019.

[31] H. Akailte, "A new look at statistical model identification," *IEEE Trans.*, 1973.

[32] G. Schwarz *et al.*, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6(2), pp. 461–4, 1978.

[33] J. Shao, "Linear model selection by cross-validation," *J. Am. Stat. Assoc.*, vol. 88(422), pp. 486–94, 1993.

[34] J. Rissanen, "Hypothesis selection and testing by the MDL principle," *Comput. J.*, vol. 42, no. 4, pp. 260–269, 1999.

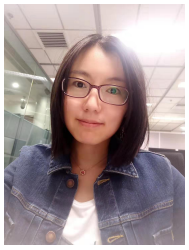
[35] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Am. Stat. Assoc.*, vol. 112(518), pp. 859–77, 2017.

[36] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 2008–2026, 2018.

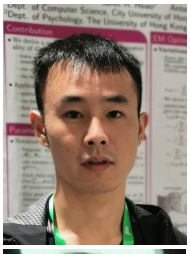
[37] A. E. Gelfand and A. F. Smith, "Sampling-based approaches to calculating marginal densities," *J. Am. Stat. Assoc.*, vol. 85, no. 410, pp. 398–409, 1990.

[38] S. R. Waterhouse, D. MacKay, and A. J. Robinson, "Bayesian methods for mixtures of experts," in *Proc. NeurIPS*, 1996, pp. 351–57.

[39] D. J. MacKay, "Ensemble learning for hidden Markov models," Citeseer, Tech. Rep., 1997.
[40] D. M. Blei, M. I. Jordan *et al.*, "Variational inference for Dirichlet process mixtures," *Bayesian Anal.*, vol. 1, no. 1, pp. 121–143, 2006.
[41] N. Ueda and Z. Ghahramani, "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Netw.*, vol. 15, no. 10, pp. 1223–1241, 2002.
[42] C. A. McGrory and D. Titterington, "Variational approximations in Bayesian model selection for finite mixture distributions," *Comput. Stat. Data Anal.*, vol. 51, no. 11, pp. 5352–5367, 2007.
[43] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West *et al.*, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian statistics*, vol. 7, no. 453-464, p. 210, 2003.
[44] M. Danielson, "Hidden Markov models with variational inference in marketing science," Ph.D. dissertation, University of St Andrews, 2021.
[45] P. Bruneau, M. Gelgon, and F. Picarougne, "Parsimonious reduction of Gaussian mixture models with a variational-Bayes approach," *Pattern Recognit.*, vol. 43, no. 3, pp. 850 – 858, 2010.
[46] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.
[47] J. Chen, X. Lin, H. Zheng, and X. Bao, "A novel cluster center fast determination clustering algorithm," *Appl. Soft. Comput.*, vol. 57, pp. 539–555, 2017.
[48] J. Chen, H. Zheng, X. Lin, Y. Wu, and M. Su, "A novel image segmentation method based on fast density clustering algorithm," *Eng. Appl. Artif. Intell.*, vol. 73, pp. 92–110, 2018.
[49] J. Chen, X. Lin, Q. Xuan, and Y. Xiang, "FGCH: a fast and grid based clustering algorithm for hybrid data stream," *Appl. Intell.*, vol. 49, no. 4, pp. 1228–1244, 2019.
[50] N. Vasconcelos and A. Lippman, "Learning mixture hierarchies," in *Proc. NeurIPS*, 1999, pp. 606–12.
[51] J. R. Hershey, P. A. Olsen, and S. J. Rennie, "Variational Kullback-Leibler divergence for hidden Markov models," in *Proc. IEEE Workshop on ASRU*, 2007, pp. 323–328.
[52] M. J. Beal *et al.*, *Variational Algorithms for Approximate Bayesian Inference*. University of London, 2003.
[53] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.
[54] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, vol. 64(4), pp. 583–639, 2002.
[55] J. Alon, S. Sclaroff, G. Kollios, and V. Pavlovic, "Discovering clusters in motion time-series data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2003.
[56] A. B. Chan and J. H. Hsiao, "EMHMM simulation study," 2018.
[57] L. Hubert and P. Arabie, "Comparing partitions," *J. Classif.*, vol. 2, no. 1, pp. 193–218, 1985.
[58] C. D. Manning, H. Schütze, and P. Raghavan, *Introduction to Information Retrieval*. Cambridge Univ. Press, 2008.
[59] Z. Liu, L. Yu, J. H. Hsiao, and A. B. Chan, "Parametric manifold learning of Gaussian mixture models," in *Proc. IJCAI*, 2019, pp. 3073–9.
[60] T. Chuk, K. Crookes, W. G. Hayward, A. B. Chan, and J. H. Hsiao, "Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures," *Cognition*, vol. 169, pp. 102–117, 2017.
[61] B. H. Williams, M. Toussaint, and A. J. Storkey, "Extracting motion primitives from natural handwriting data," in *Proc. ICANN*, 2006, pp. 634–643.



Hui Lan received the B.S. degree in information and computational science from Dalian University of Technology, Dalian, China, in 2015, and the Ph.D. degree in statistics from the University of Chinese Academy of Sciences, China, in 2021. She is currently a postdoctoral researcher in the School of Statistics and Data Science, Faculty of Science, Beijing University of Technology, China. Her research interests include machine learning, statistics analysis and optimization.



Ziquan Liu is currently a PhD student in the Department of Computer Science, City University of Hong Kong. He received the B.Eng. degree in information engineering and B.S. in mathematics from Beihang Univerisity, Beijing, in 2017. His research interests include low-dimensional representation, deep neural networks and Bayesian inference.

Janet H. Hsiao Dr. Janet Hsiao is an associate professor at the University of Hong Kong in the Department of Psychology. Before joining HKU, she was a postdoctoral researcher in the Temporal Dynamics of Learning Center (TDLC) at the University of California, San Diego (UC San Diego). She received the Ph.D. degree in Informatics (Cognitive Science) from the University of Edinburgh in 2006. She received the M.Sc. degree in Computing Science from Simon Fraser University in 2002, and the B.Sc. degree in Computer Science from National Taiwan University in 1999. Her research interests include cognitive science, computational modeling, and eye movement analysis.

Dan Yu received the B.S. in mathematics from University of Science and Technology of China, Hefei, China, in 1984, the M.S. degree in Probability and Statistics from Peking University, Beijing, China, in 1991, and the PhD degree from Academy of Mathematics and Systems Science Chinese Academy of Sciences, Beijing, China, in 1996. He is currently a Professor with the Institute of Systems Science, Academy of Mathmatics and systems Sciences Chinese Academy of Sciences. His research interests include industrial statistics and reliability analysis.

Antoni B. Chan received the B.S. and M.Eng. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2000 and 2001, and the Ph.D. degree in electrical and computer engineering from the University of California, San Diego (UCSD), San Diego, in 2008. He is currently an Professor in the Department of Computer Science, City University of Hong Kong. His research interests include computer vision, machine learning, pattern recognition, and music analysis.