Hierarchical Learning of Hidden Markov Models with Clustering Regularization

Hui Lan[1] **Antoni B. Chan**[1]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong

Abstract

Hierarchical learning of generative models is useful for representing and interpreting complex data. For instance, one application is to learn an HMM to represent an individual's eye fixations on a stimuli, and then cluster individuals' HMMs to discover common eye gaze strategies. However, learning the individual representation models from observations and clustering individual models to group models are often considered as two separate tasks.In this paper, we propose a novel tree structure variational Bayesian method to learn the individual model and group model simultaneously by treating the group models as the parents of individual models, so that the individual model is learned from observations and regularized by its parents, and conversely, the parent model will be optimized to best represent its children. Due to the regularization process, our method has advantages when the number of training samples decreases. Experimental results on the synthetic datasets demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

The hidden Markov model (HMM) [Rabiner, 1993] 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. HMMs have been popularly applied in many areas that need to analyze time series data, such as speech recognition [Juang and Rabiner, 1991, Aucouturier and Mark, 2001], cognitive science [Chan et al., 2018, Hsiao et al., 2021b], and music analysis [Qi et al., 2007]. In particular, recent works using HMMs to model eye fixation sequences has enabled interesting discoveries about the role of eye gaze in cognitive processes, including three processing states in information search tasks [Simola et al., 2008], optimal strategies for face recognition [Chuk et al., 2017b,a, An and Hsiao, 2021, Hsiao et al., 2021a], masking effects in visual search [Hsiao et al., 2021b], and the association of eye gaze patterns to cognitive decline [Chan et al., 2018], emotion recognition [Zhang et al., 2019, Chan et al., 2020a], chronic pain [Chan et al., 2020c,b], and decision making [Chuk et al., 2020].

The previous methods of hierarchical modelling of HMMs learn sequentially; first the individual models are learned from observations, then the group models are learned from the individual models, as shown in Fig. 1(a). Individual HMMs can be learned from observations using two typical methods: 1) the Baum-Welch (EM) algorithm [Baum et al., 1970], which computes the maximum likelihood parameter estimation of the HMM; 2) variational Bayesian EM (VBEM) algorithm [Beal et al., 2003], which computes the posterior distribution over each parameter of HMM through maximizing the evidence lower bound (ELBO). Learning the group models is equivalent to clustering individual HMMs, with each cluster center representing one group model. Coviello et al. [2014] proposed a variational hierarchical EM (VHEM) algorithm, which clusters HMMs directly using their probability densities of the observation sequence, and estimates HMM cluster centers. Lan et al. [2021] proposed a variational Bayesian hierarchical EM (VBHEM) algorithm, a Bayesian version of VHEM.

In the above, the individual models and group models are learned as separate tasks. When the data is sufficient, separately learning individual and group models is fine, such as experiments in [Coviello et al., 2014]. However, when the data is insufficient, the individual model may overfit, which affects the group model. For example, in face recognition [Chuk et al., 2014] or in scene perception Hsiao et al. [2021c], only one eye fixation sequence is tracked per stimulus. In this case, joint learning of individual and group models will help learning of individual models through pooling of common information in the group.

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1628–1638.

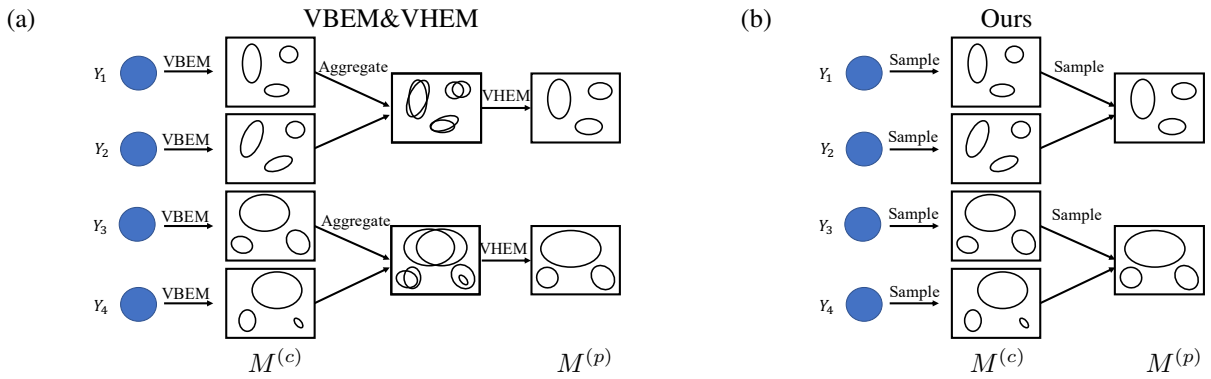

Figure 1: Illustration of the hierarchical modelling of HMMs, learning individual (child) models $M^{(c)}$ and group (parent) models $M^{(p)}$ through: (a) VBEM and VHEM (sequentially); (b) Our method (simultaneously). $\{Y_i\}_{i=1}^4$ represent 4 observation sequences. The HMMs are visualized by its 3 Gaussian emissions.

In this paper, we propose to estimate the individual and the group models simultaneously, as shown in Fig. 1(b), so that the group models can regularize the individual models. This is similar to the hierarchical generative process: 1) there are group models; 2) individual models are sampled from the group models; 3) observations are sampled from the individual models. This generative process is similar to topic model in document corpora, where documents are organized in a multi-level hierarchy [Kim et al., 2013]. Although here we focus on HMMs, the framework could be applied to other probabilistic models.

Our Contributions. In this paper, we propose a novel tree structure variational Bayesian method to learning the individual models and group models simultaneously, where the group models are parents and the individual models are children. The group models regularize the individual models thereby alleviating overfitting of individual models, and the group models are estimated from the individual models, thus iteratively affecting each other. In experiments, we obtain good clustering performance, and the individual and group models are close to the ground-truth models, even for small sample size. Furthermore, the clustering is inherent in our model and does not resort to other existing clustering methods. Lastly, our child-parent framework is a generic regularization method, and could be applied to other probabilistic models.

2 RELATED WORK

Hierarchical Models and Inference. In Fig. 1(b), the hierarchical HMMs structure allows us to use the group HMMs as the prior on the individual HMMs and then to learn the individual HMMs. Dirichlet process (DP) provides nonparametric prior for the number of mixture components and is widely considered in learning of HMMs. Teh et al. [2006] introduced the hierarchical Dirichlet process HMM (HDP-HMM) to learn an HMM, where each HMM state is represented by a mixture model, and the mixture models in the different groups share mixture components. Qi et al. [2007] utilized DP HMM mixture models (DP-H3M) to build an H3M for a song, where each HMM represents a song clip, i.e., the individual HMM is the same as one component of the group H3M. In contrast, for our method, the mixture models do not share parameters, and each individual model has its own distribution that is not the same as one of the group models. The nested HDP [Paisley et al., 2014] is a novel prior to perform word-specific path clustering on a shared tree, which also has shared parameters and has not been applied to HMMs.

The HDP-HMM uses Markov chain Monte Carlo (MCMC) [Hastings, 1970, Gelfand and Smith, 1990] for posterior inference. MCMC explores the parameter space relying on sampling. However, when the model is complex (such as the hierarchal HMM), performing Bayesian inference via MCMC can be exceedingly expensive. We resort to variational Bayesian methods for inference on the individual models and group models. Variational inference (VI) is an alternative to MCMC, which relies on optimization rather than sampling. For mixture models, VI may perform better than a more general MCMC technique (e.g., Hamiltonian Monte Carlo), even for small datasets [Kucukelbir et al., 2015]. Zhang et al. [2016] derived a VI for the HDP-HMM.

Regarding VI, the hierarchical structure of priors has been introduced to relax the mean-field assumption of variational distributions, e.g., hierarchical variational models (HVMs) [Ranganath et al., 2016] and Ladder-VAE [Sønderby et al., 2016]. HVM is a two-level model that first draws variational parameters from a prior and then draws latent variables from the corresponding likelihood. In this perspective, our model is a three-level model since we also have a hyper-prior over variational parameters, i.e., $p(M^p)$. Also, Bouchacourt et al. [2018] developed a multi-level variational autoencoder (ML-

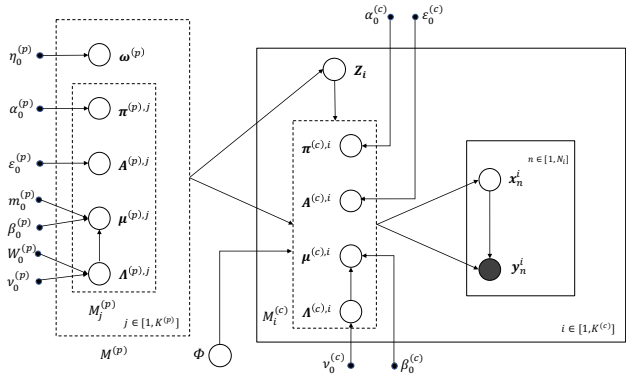

Figure 2: Graphical model representation of our model. The solid line plate denotes a set of i.i.d. samples. y_n^i is an observation sequence, and x_n^i is the state sequence that emits y_n^i. Z_i indicates the assignment of $M_i^{(c)}$ to an HMM component in $M^{(p)}$. The dashed line plate denotes an HMM: child HMM $M_i^{(c)}$ with parameters $\{\pi^{(c),i}, A^{(c),i}, \mu^{(c),i}, \Lambda^{(c),i}\}$ and parent $M_j^{(p)}$ with parameters $\{\pi^{(p),j}, A^{(p),j}, \mu^{(p),j}, \Lambda^{(p),j}\}$. The variables outside the plate, $\{\eta_0^{(p)}, \alpha_0^{(p)}, \epsilon_0^{(p)}, m_0^{(p)}, \beta_0^{(p)}, W_0^{(p)}, \nu_0^{(p)}, \alpha_0^{(c)}, \epsilon_0^{(c)}, \nu_0^{(c)}, \beta_0^{(c)}\}$ are hyperparameters, and $\Phi = \{\phi^{i,j}\}_{i=1,j=1}^{K^{(c)},K^{(p)}}$ are a set of state permutation matrices.

VAE) for learning a disentangled representation of a set of grouped observations. However, these works focus on the hierarchical structure of latent variables (or latent code), such as assignment variable, not on the model parameters, while our method focuses on the hierarchical structure of the model parameters.

Regularization via Clustering. In our model, the group models are actually cluster centers when clustering the individual models, and the clustering regularizes the learning of individuals. The idea of using clustering as regularization has been explored in other domains. Pang et al. [2014] simultaneously regularized the between- and within-class scatter matrices to learn regularized linear discriminant analysis. Price et al. [2021] proposed a penalized likelihood framework for estimating the C precision matrices with cluster regularization. Cluster-based regularization has also been of interest in semisupervised learning, active learning, transfer learning, and other areas of AI [Soares et al., 2012, Sellars et al., 2020, Hubert and Arabie, 1985, Zhao et al., 2019, Long et al., 2013]. Soares et al. [2012] proposed a robust algorithm, cluster-based regularization (ClusterReg) for semisupervised learning (SSL), which takes advantage of partitions resulting from a clustering algorithm, and uses such information to regularize prediction. The method is also extended to Ensemble Learning [Soares et al., 2017]. Sellars et al. [2020] used clustering based regularization to improve decision boundaries within a novel SSL framework called two-cycle learning. In contrast to these methods, our method is a Bayesian generative model, where the clustering is inherent in our model, and does not resort to other existing clustering methods.

Hidden Markov (Mixture) Model. We briefly review hidden Markov models (HMMs) and the hidden Markov mixture model (H3M) [Smyth, 1997], and define the notation used in the paper. An H3M models a set of observation sequences as samples from a group of K hidden Markov models (HMMs), and is parameterized by $M = \{\omega_i, M_i\}_{i=1}^K$, where M_i is the ith HMM and ω_i is the corresponding mixture component weight. An observation sequence with length τ is denoted by $y = (y_1, y_2, ..., y_\tau)$, and depends on a hidden state sequence $x = (x_1, x_2, ..., x_\tau)$. The observation likelihood for $y \sim M$ is $p(y|M) = \sum_i \omega_i p(y|M_i)$, where the ith HMM M_i with S states is specified by parameters $M_i = \{\pi^i, A^i, \{\Theta_k^i\}_{k=1}^S\}$. In detail, $\pi^i = [\pi_1^i, ..., \pi_S^i]$ is the initial state probability, where $\pi_k^i = p(x_1 = k|M_i)$. $A^i = [a_{k,k'}^i]_{S\times S}$ is the state transition matrix, where $a_{k,k'}^i = p(x_{t+1} = k'|x_t = k, M_i)$ is the transition probability from state k to k'. Θ_k^i is the parameter set of emission density at state k. Here, we assume the emissions are Gaussian distributions, $p(y_t|x_t = k, M_i) = \mathcal{N}(y_t|\mu_k^i, (\Lambda_k^i)^{-1})$, with mean μ_k^i and precision matrix Λ_k^i.

3 METHODOLOGY

In this section, we introduce our tree structure variational Bayesian method based on HMMs. Our hierarchical model consists of the following generative process: 1) a parent H3M is sampled from a prior, with each HMM component corresponding to a group; 2) a child HMM is sampled around the parent model, specifically, via a distribution formed using the parent H3M parameters; 3) observations are sampled from the child HMM. Here we use "parent" and "child" to refer to the group and individual models.

3.1 FRAMEWORK

Formally, consider a set of $K^{(c)}$ grouped samples $Y = \{Y_1, Y_2, \ldots, Y_{K^{(c)}}\}$, where $Y_i = \{y_1^i, y_2^i, \ldots, y_{N_i}^i\}$ are drawn from the ith child model $M_i^{(c)}$, and each sample $y_n^i = (y_{n,1}^i, y_{n,2}^i, \ldots, y_{n,\tau}^i)$ is a time series with length τ, and $y_{n,t}^i \in \mathbb{R}^d$. Each observation (or emission) $y_{n,t}^i$ at time t depends on the state of a discrete hidden variable $x_{n,t}^i$, and the sequence of hidden states $x_n^i = (x_{n,1}^i, x_{n,2}^i, \ldots, x_{n,\tau}^i)$ evolves as a first-order Markov chain. The hidden variable $x_{n,\tau}^i$ can take one of $S^{(c)}$ values, i.e., $x_{n,t}^i \in \{1, \ldots, S^{(c)}\}$. Each child model $M_i^{(c)}$ is an hidden Markov model (HMM) with $S^{(c)}$ states. The parent model $M^{(p)}$ is a hidden Markov mixture model (H3M) consisting of components $M_j^{(p)}$, $j \in \{1, \ldots, K^{(p)}\}$, $K^{(p)} \leq K^{(c)}$, and each $M_j^{(p)}$ has $S^{(p)}$ states, $S^{(p)} \leq S^{(c)}$. The child and parent are connected

through a child-parent distribution, from which the child model $M_i^{(c)}$ is sampled around the corresponding parent model $M_j^{(p)}$. Note that we will always use superscripts (p) and (c) to distinguish the parameters for parent and child model, i and j to index the components in the mixture model of parents and children, and k and l to index the hidden states in the parent and child model, respectively.

An illustration of the probabilistic graphical model is shown in Fig. 2. Given the hyperparameters related to the parent model (e.g., $\alpha_0^{(p)}$), a parent model is sampled from its prior, i.e., $M_j^{(p)} \sim p(M_j^{(p)})$, and weight $\omega^{(p)} \sim p(\omega^{(p)})$ (the hyperparameters are omitted to reduce clutter). Then, we introduce the assignment variable Z_i, which takes one of $K^{(p)}$ values to assign the child model $M_i^{(c)}$ to one of the components in parent model, and the state permutation matrix $\phi^{i,j}$, which has size $S^{(c)} \times S^{(p)}$, to match the states between two HMMs $M_i^{(c)}$ and $M_j^{(p)}$. We assume the priors $p(Z_i = j|\omega^{(p)}) = \omega_j^{(p)}$ and $p(\phi^{i,j}) = \prod_{k=1}^{S^{(c)}} \prod_{l=1}^{S^{(p)}} (1/S^{(p)})^{\phi_{k,l}^{i,j}}$, respectively. Given the assignment $Z_i = j$, the generative process of the data Y_i is:

1. Sample a child model $M_i^{(c)} \sim p(M_i^{(c)}|M_j^{(p)})$;
2. Sample (i.i.d.) sequences $y_n^i \sim M_i^{(c)}$, $n = 1, ..., N_i$.

Note that only the variables $\{y_n^i\}_{i=1,n=1}^{K^{(c)},N_i}$ are observed. All the parameters (white circles) are unknown and are treated as hidden variables. The child models are affected by observations and parent models together, and the parent models are affected by the child models. Also, one child in our model only has one parent, but one parent may have several children, as shown in Fig. 2.

3.2 VARIATIONAL INFERENCE

The very large parameter space resulting from the hierarchical HMM comes with a large computational burden. Thus, we resort to the variational Bayes, which is a faster alternative to MCMC methods, to approximate the posterior distributions of the hidden variables. First, we posit a family of approximate densities $q(\cdot) \in \mathcal{Q}$ over the hidden variables. Then, we find the member of that family that minimizes the Kullback-Leibler (KL) divergence to the exact posterior. Finally, we approximate the posterior with the optimized member of the family $q^*(\cdot)$ [Blei et al., 2017].

ELBO. Formally, our goal is to find the best candidate in the family $\mathcal{Q}$, i.e., the one closest in KL divergence to the exact posterior. Inference now amounts to solving the following optimization problem,

$$q^*(H) = \arg\min_{q(H)\in\mathcal{Q}} \mathrm{KL}\big(q(H)||p(H|Y)\big), \tag{1}$$

where $H = \{M^{(p)}, M^{(c)}, Z, \Phi, X\}$ and the assumed $\mathcal{Q}$ can be found in Sec. 3.4. Minimizing the KL divergence in (1) is equivalent to maximizing the evidence lower bound (ELBO) [Blei et al., 2017] as below (see Appendix A for details),

$$\begin{aligned}\log p(Y) &\geq \sum_i \mathbb{E}_{q(M^{(p)})}\big[\log p(Y_i|M^{(p)})\big] - \mathrm{KL}(q(M^{(p)})||p(M^{(p)})) \\ &\triangleq \mathcal{L}(M^{(p)}, M^{(c)}),\end{aligned}$$

where

$$\begin{aligned}\mathcal{L}(M^{(p)}, M^{(c)}) &= \sum_i \mathbb{E}_{q(M_i^{(c)})}\big[\log p(Y_i|M_i^{(c)})\big] \\ &+ \sum_i \sum_j \hat{z}_{ij} \mathbb{E}_{q(M_i^{(c)})} \mathbb{E}_{q(M_j^{(p)})}\big[\log p(M_i^{(c)}|M_j^{(p)})\big] \\ &+ \sum_i \sum_j \hat{z}_{ij} \mathbb{E}_{q(\omega^{(p)})}\big[\log \omega_j^{(p)}\big] + \mathbb{E}_{\omega^{(p)}} \log p(\omega^{(p)}) \\ &+ \sum_j \mathbb{E}_{q(M_j^{(p)})}\big[\log p(M_j^{(p)})\big] - \sum_i \sum_j \hat{z}_{ij} \log \hat{z}_{ij} \\ &- \sum_i \mathbb{E}_{q(M_i^{(c)})}\big[\log q(M_i^{(c)})\big] - \mathbb{E}_{q(\omega^{(p)})}\big[\log q(\omega^{(p)})\big] \\ &- \sum_j \mathbb{E}_{q(M_j^{(p)})}\big[\log q(M_j^{(p)})\big],\end{aligned} \tag{2}$$

and $\hat{z}_{i,j} = \mathbb{E}_{q(Z)}[z_{ij}]$. The best child model $M_*^{(c)}$ and the best parent model $M_*^{(p)}$ are obtained via

$$\big\{M_*^{(p)}, M_*^{(c)}\big\} = \arg\max_{M^{(p)}, M^{(c)}} \mathcal{L}(M^{(p)}, M^{(c)}).$$

It is important to note that although here we focus on HMMs, our framework could be used for any child and parent probability distributions. The whole algorithm is summarized in Alg. 1. We explain each step in the following subsections, in particular the prior distributions and the variational distributions for approximating the posterior.

3.3 PRIORS DISTRIBUTIONS

Priors on Child Models. The key element in our ELBO in (2) is to construct the child-parent model $p(M_i^{(c)}|M_j^{(p)})$. It can be considered as the prior on child model $M_i^{(c)}$ given assignment to parent model $M_j^{(p)}$, from which instances of $M_i^{(c)}$ are generated. Since $M_i^{(c)}$ and $M_j^{(p)}$ are both HMMs with their own states, a hidden state permutation matrix $\phi^{i,j}$ has been introduced to match their states, where $\phi_{k,l}^{i,j} = 1$ if the kth state in $M_i^{(c)}$ corresponds to the lth state in $M_j^{(p)}$, otherwise $\phi_{k,l}^{i,j} = 0$. We assume that each child state k is assigned to only one parent state l, i.e., $\sum_{l=1}^{S^{(p)}} \phi_{k,l}^{i,j} = 1$, $\forall k \in \{1, ..., S^{(c)}\}$, although multiple child states can be assigned to the same parent state.

With $\phi^{i,j}$, we obtain a lower bound on $\log p(M_i^{(c)}|M_j^{(p)})$,

$$\log p(M_i^{(c)}|M_j^{(p)}) = \log \int p(M_i^{(c)}, \phi^{i,j}|M_j^{(p)}) d\phi^{i,j}$$

$$\geq \mathbb{E}_{q(\phi^{i,j})}\left[\log p(M_i^{(c)}|\phi^{i,j}, M_j^{(p)})\right] - \mathrm{KL}(q(\phi^{i,j})||p(\phi^{i,j})),$$

where we introduce a variational posterior distribution $q(\phi^{i,j})$. In the generative process, we assume that the state permutation $\phi^{i,j}$ is applied to the parent model and then the child model parameters (initial probability, transition matrix, and emission densities) are sampled. By marginalizing over the state permutation matrix distribution $q(\phi^{i,j})$, we avoid the issue of multiple equivalent parameterizations of the hidden states.

Next, we construct the child-parent model, where the parent HMM parameters serve as the "mean" of the prior distributions of the child HMM parameters,

$$\begin{aligned}\log p(M_i^{(c)}|\phi^{i,j}, M_j^{(p)}) &= \log p(\pi^{(c),i}|\alpha_0^{(c)}, \phi^{i,j}, \pi^{(p),j}) \\ &+ \log p(\mu^{(c),i}, \Lambda^{(c),i}|\phi^{i,j}, \mu^{(p),j}, \beta_0^{(c)}, \Lambda^{(p),j}, \nu_0^{(c)}) \\ &+ \log p(A^{(c),i}|\epsilon_0^{(c)}, \phi^{i,j}, A^{(p),j}), \end{aligned} \quad (3)$$

where $\alpha_0^{(c)}$, $\epsilon_0^{(c)}$, $\beta_0^{(c)}$, and $\nu_0^{(c)}$ are scalar hyperparameters to control the regularization effect from the parent models. Eq. 3 is the key that connects the parent model $M_j^{(p)}$ and child model $M_i^{(c)}$. Specifically, we take the priors on the child model parameters to be their corresponding conjugate priors [Diaconis et al., 1979]:

1. Prior on ith child initial probabilities $\pi^{(c),i}$,
$$p(\pi^{(c),i}|\alpha_0^{(c)}, \phi^{i,j}, \pi^{(p),j}) = \mathrm{Dir}(\pi^{(c),i}|\tilde{\alpha}^{i,j}),$$
where $\tilde{\alpha}^{i,j}{=}\alpha_0^{(c)}\phi^{i,j}\pi^{(p),j}$ is a concentration hyperparameter, and $\phi^{i,j}\pi^{(p),j}$ is a state permutation of $\pi^{(p),j}$.
2. Prior on ith child transition matrix $A^{(c),i}$,
$$p(A^{(c),i}|\epsilon_0^{(c)}, \phi^{i,j}, A^{(p),j}) = \prod_k \mathrm{Dir}(a_k^{(c),i}|\tilde{\epsilon}_k^{i,j}),$$
where $a_k^{(c),i}$ is the kth row of $A^{(c),i}$, $\tilde{\epsilon}^{i,j} = \epsilon_0^{(c)}\phi^{i,j}A^{(p),j}(\phi^{i,j})^T$ is the concentration hyperparameter of the kth row of the permuted matrix.
3. Priors on ith child emission mean and precision,
$$\begin{aligned}&p(\mu^{(c),i}, \Lambda^{(c),i}|\phi^{i,j}, \mu^{(p),j}, \beta_0^{(c)}, \Lambda^{(p),j}, \nu_0^{(c)}) \\ &= \prod_k \prod_l \left[\mathcal{N}(\mu_k^{(c),i}|\mu_l^{(p),j}, (\beta_0^{(c)}\Lambda_k^{(c),i})^{-1})\right. \\ &\qquad \left. \cdot\, \mathcal{W}(\Lambda_k^{(c),i}|\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)})\right]^{\phi_{k,l}^{i,j}},\end{aligned}$$
where $\mathcal{W}(\cdot|W,\nu)$ is a Wishart distribution with scale matrix W and degrees-of-freedom ν. With this prior, we have $\mathbb{E}[\mu_k^{(c),i}] = \sum_l \phi_{k,l}^{i,j}\mu_l^{(p),j}$ and $\mathbb{E}[\Lambda_k^{(c),i}] = \sum_l \phi_{k,l}^{i,j}\Lambda_l^{(p),j}$.

Priors on Parent Models. Next, we consider the priors on parent models,

$$p(M^{(p)}) = p(\omega^{(p)}) \prod_{j=1}^{K^{(p)}} p(M_j^{(p)}),$$

$$p(M_j^{(p)}) = p(\pi^{(p),j}) \prod_{l=1}^{S^{(p)}} p(a_l^{(p),j})p(\mu_l^{(p),j}, \Lambda_l^{(p),j}).$$

Similar to the child models, we assume conjugate priors on parent models parameters to simplify the analysis,

$$\begin{aligned}&\omega^{(p)} \sim \mathrm{Dir}(\omega^{(p)}|\eta_0^{(p)}), \\ &\pi^{(p),j} \sim \mathrm{Dir}(\pi^{(p),j}|\alpha_0^{(p)}), \quad a_l^{(p),j} \sim \mathrm{Dir}(a_l^{(p),j}|\epsilon_0^{(p)}), \\ &\mu_l^{(p),j}|\Lambda_l^{(p),j} \sim \mathcal{N}(\mu_l^{(p),j}|m_0^{(p)}, (\beta_0^{(p)}\Lambda_l^{(p),j})^{-1}), \\ &\Lambda_l^{(p),j} \sim \mathcal{W}(\Lambda_l^{(p),j}|W_0^{(p)}, \nu_0^{(p)}).\end{aligned}$$

The hyperparameters $\eta_0^{(p)}$, $m_0^{(p)}$, $W_0^{(p)}$, $\alpha_0^{(p)}$, $\epsilon_0^{(p)}$, $\beta_0^{(p)}$, and $\nu_0^{(p)}$ are all scalars. Note that $p(M^{(p)})$ serves as both the prior on $M^{(p)}$ and the hyper-prior on $M^{(c)}$.

3.4 VARIATIONAL DISTRIBUTIONS

The ELBO in (2) is maximized via coordinating ascent w.r.t. the variational distribution over each hidden variable, i.e., iteratively optimizing each factor, $q(M^{(p)})$, $q(M^{(c)})$, $q(Z)$, $q(\Phi)$, and $q(X)$, while holding the others fixed, resulting in the approximate posterior distributions for each hidden variable. Specifically, our method has two alternating steps:

1. Given $q(M^{(p)})$ and $q(M^{(c)})$, update $q(X)$, $q(\Phi)$, and $q(Z)$ via maximizing (2).
2. Update $q(M^{(p)})$ and $q(M^{(c)})$ via maximizing (2).

Here we restrict the family of distributions $q(H)$ with a mean field assumption, i.e., the q distribution factorizes w.r.t. each parameter. In the optimization, many of the parameters have closed form updates (similar to Beal et al. [2003]), while others need numeric solvers. We derive the optimal variational distributions in the following.

3.4.1 Variational Distributions for Z, Φ, and X

We provide the optimal variational distributions for the assignment variables Z, the state permutation matrices Φ, and the hidden state sequences X. With the mean field assumption of variational distribution q, we are only interested in the functional dependence of the RHS in (2) on the variables Z, Φ, and X, respectively.

For the variational distribution $q(Z) = \prod_i \prod_j (z_{ij})^{\hat{z}_{ij}}$ and

$$\log z_{ij} \propto \mathbb{E}_{M_i^{(c)}}\mathbb{E}_{M_j^{(p)}} \log p(M_i^{(c)}|M_j^{(p)}) + \mathbb{E}_{\omega^{(p)}} \log \omega_j^{(p)}.$$

After normalizing, the optimal solution is

$$\hat{z}_{ij} = \frac{\tilde{\omega}_j^{(p)} \exp\left\{\mathbb{E}_{M_i^{(c)}}\mathbb{E}_{M_j^{(p)}} \log p(M_i^{(c)}|M_j^{(p)})\right\}}{\sum_l \tilde{\omega}_l^{(p)} \exp\left\{\mathbb{E}_{M_i^{(c)}}\mathbb{E}_{M_l^{(p)}} \log p(M_i^{(c)}|M_l^{(p)})\right\}}, \quad (4)$$

where $\tilde{\omega}_j^{(p)} = \mathbb{E}[\log \omega_j^{(p)}]$ and the expectation term is approximated by a lower bound (see Appendix B.1). $\hat{z}_{ij}$ is the responsibility for the parent model $M_j^{(p)}$ explaining the child model $M_i^{(c)}$ and the corresponding observations Y_i.

For the variational distribution $q(\Phi) = \prod_i \prod_j q(\phi^{i,j})$, $q(\phi^{i,j}) = \prod_k \prod_l (\phi_{k,l}^{i,j})^{\hat{\phi}_{k,l}^{i,j}}$, and thus $\mathbb{E}[\phi_{k,l}^{i,j}] = \hat{\phi}_{k,l}^{i,j}$. Moreover, we assume $\sum_k \phi_{k,l}^{i,j} \geq 1$, which means at least one state in $M_i^{(c)}$ is assigned to the lth state in $M_j^{(p)}$. There is no closed form solution for $\phi^{i,j}$ and we solve for the optimal $\hat{\phi}^{i,j}$ via the optimization problem

$$\max \ \mathcal{L}(\hat{\phi}^{i,j}) \quad s.t. \sum_l \hat{\phi}_{k,l}^{i,j} = 1, \quad \sum_k \hat{\phi}_{k,l}^{i,j} \geq 1, \quad (5)$$

where

$$\mathcal{L}(\hat{\phi}^{i,j}) \propto \mathbb{E}_{M_i^{(c)}} \mathbb{E}_{M_j^{(p)}} \log p(M_i^{(c)} | M_j^{(p)}),$$

which contains all terms in (2) involving $\hat{\phi}^{i,j}$ (see Appendix B.2). $\hat{\phi}_{k,l}^{i,j}$ provides the probabilities of the kth child state corresponding to the lth parent state.

For the variational distribution $q(X) = \prod_i \prod_n q(x_n^i)$, since each x_n^i is independent of the parent model, $q(x_n^i)$ can be solved using the traditional variational Bayesian EM for HMMs [Beal et al., 2003] given y_n^i. The responsibilities

$$r_{n,t,k}^i = \mathbb{E}[x_{n,t,k}^i], \ \ r_{n,t,k,k'}^i = \mathbb{E}[x_{n,t,k'}^i x_{n,t-1,k}^i], \quad (6)$$

$k, k' \in \{1, \ldots, S^{(c)}\}$, are solved using the forward-backward algorithm. $r_{n,t,k}^i$ is the responsibility for the kth Gaussian for observation $y_{n,t}^i$, and $r_{n,t,k,k'}^i$ is a transition responsibility.

3.4.2 Variational Distributions for Child Models

With the conjugate priors of $M^{(c)}$ assumed in Sec. 3.3, $q(M^{(c)})$ is determined automatically by optimization of the variational distributions and has the same form as the priors (similar to Beal et al. [2003]). For $q(M_i^{(c)})$,

$$\pi^{(c),i} \sim \mathrm{Dir}(\pi^{(c),i} | \alpha^{(c),i}), \quad a_k^{(c),i} \sim \mathrm{Dir}(a_k^{(c),i} | \epsilon_k^{(c),i}),$$
$$\mu_k^{(c),i} \sim \mathcal{N}(\mu_k^{(c),i} | m_k^{(c),i}, (\beta_k^{(c),i} \Lambda_k^{(c),i})^{-1}),$$
$$\Lambda_k^{(c),i} \sim \mathcal{W}(\Lambda_k^{(c),i} | W_k^{(c),i}, \nu_k^{(c),i}),$$

for $i \in \{1, \cdots, K^{(c)}\}$, $k \in \{1, \cdots, S^{(c)}\}$. The parameters $\alpha^{(c),i}$, $\epsilon_k^{(c),i}$, $m_k^{(c),i}$, $\beta_k^{(c),i}$, $W_k^{(c),i}$, $\nu_k^{(c),i}$ are all updated by closed form solutions that combine observations and priors.

Specifically, for initial and transition probabilities,

$$\alpha_k^{(c),i} = N_{1,k}^i + \alpha_0^{(c)} \sum_j \hat{z}_{ij} \sum_l \hat{\phi}_{k,l}^{i,j} \hat{\pi}_l^{(p),j}, \quad (7)$$

Algorithm 1 Co-learning $M^{(p)}$ and $M^{(c)}$

Input: data $Y = \{Y_1, Y_2, \ldots, Y_{K^{(c)}}\}$, $S^{(c)}$, $K^{(c)}$, $S^{(p)}$, $K^{(p)}$, and hyperparameters $\alpha_0^{(c)}$, $\epsilon_0^{(c)}$, $\beta_0^{(c)}$, $\nu_0^{(c)}$, $\eta_0^{(p)}$, $m_0^{(p)}$, $W_0^{(p)}$, $\alpha_0^{(p)}$, $\epsilon_0^{(p)}$, $\beta_0^{(p)}$, $\nu_0^{(p)}$.
Output: $M^{(p)}$, $M^{(c)}$ and Z.

```
Initialize M^(p) and M^(c).
repeat
    for i = 1 to K^(c) and j = 1 to K^(p) do
        Compute φ^{i,j} via (5).
        Compute z_ij via (4).
    end for
    for each Y_i, i = 1 to K^(c) do
        Compute responsibilities r^i_{n,t,k} and r^i_{n,t,k,k'} via (6);
        Update M_i^(c): update α_k^{(c),i}, ε_{k,k'}^{(c),i}, m_k^{(c),i}, β_k^{(c),i}, ν_k^{(c),i}, and W_k^{(c),i} via (7), (8), and (9), respectively, for k = 1, ..., S^(c).
    end for
    Update M^(p) via (11) and solving (12) and (13).
until L(M^(p), M^(c)) converges.
```

where on the RHS, $N_{1,k}^i = \sum_n r_{n,1,k}^i$ is the number of observations in Y_i with the kth state at $t = 1$, and the second term is the number of virtual samples provided by parents models with the kth state at $t = 1$ and $\hat{\pi}_l^{(p),j} = \mathbb{E}[\pi_l^{(p),j}]$. Similarly,

$$\epsilon_{k,k'}^{(c),i} = N_{k,k'}^i + \epsilon_0^{(c)} \sum_j \hat{z}_{ij} \sum_l \sum_{l'} \hat{\phi}_{k,l}^{i,j} \hat{a}_{l,l'}^{(p),l} \hat{\phi}_{k',l'}^{i,j}, \quad (8)$$

where $N_{k,k'}^i = \sum_n \sum_{t=2}^{\tau} r_{n,t,k,k'}^i$ is the number of observations which transition from kth state to k'th state in the sequences, and the second term is the number of virtual samples provided by parents models with the same transition and $\hat{a}_{l,l'}^{(p),l} = \mathbb{E}[a_{l,l'}^{(p),l}]$.

For the emission probability density, we have

$$m_k^{(c),i} = \tfrac{1}{\beta_k^{(c),i}} \Big[\beta_0^{(c)} \bar{m}_k^i + N_k^i \bar{y}_k^i \Big], \quad (9)$$
$$\beta_k^{(c),i} = N_k^i + \beta_0^{(c)}, \quad \nu_k^{(c),i} = N_k^i + \nu_0^{(c)},$$
$$W_k^{(c),i} = N_k^i S_k^i + \tfrac{N_k^i \beta_0^{(c)}}{\beta_k^{(c),i}} (\bar{y}_k^i - \bar{m}_k^i)(\bar{y}_k^i - \bar{m}_k^i)^T$$
$$+ \beta_0^{(c)} C_k^i + \sum_j \hat{z}_{ij} \sum_l \hat{\phi}_{k,l}^{i,j} \Big(\tfrac{\beta_0^{(c)}}{\beta_l^{(p),j}} + \nu_0^{(c)} \Big) \mathbb{E}\big[(\Lambda_l^{(p),j})^{-1} \big],$$

where

$$N_k^i = \sum_n \sum_t r_{n,t,k}^i, \quad \bar{y}_k^i = \frac{1}{N_k^i} \sum_n \sum_t r_{n,t,k}^i y_{n,t}^i,$$
$$S_k^i = \frac{1}{N_k^{(i)}} \sum_n \sum_t r_{n,t,k}^i (y_{n,t}^i - \bar{y}_k^i)(y_{n,t}^i - \bar{y}_k^i)^T,$$
$$\bar{m}_k^i = \sum_j \hat{z}_{ij} \sum_l \hat{\phi}_{k,l}^{i,j} m_l^{(p),j},$$

$$C_k^i = \sum_j \hat{z}_{ij} \sum_l \hat{\phi}_{k,l}^{i,j} (m_l^{(p),j} - \bar{m}_k^i)(m_l^{(p),j} - \bar{m}_k^i)^T,$$
$$\mathbb{E}\big[(\Lambda_l^{(p),j})^{-1}\big] = (W_l^{(p),j})^{-1}/(\nu_l^{(p),j} - d - 1). \quad (10)$$

The regularization effect of the parent model is seen in the update steps, where each update is a mix of observations (from data) and a regularization term (from parents). For example, in (9), $m_k^{(c),i}$ is updated by $\beta_0^{(c)}$ virtual samples $\bar{m}_k^i$ from parents and N_k^i observations $\bar{y}_k^i$ which are assigned to the ith child with kth state. $\beta_k^{(c),i}$ and $\nu_k^{(c),i}$ are updated by the number of observations assigned to the kth child state N_k^i plus the virtual samples size $\beta_0^{(c)}$ and $\nu_0^{(c)}$, respectively. For $W_k^{(c),i}$, the first line is the same with VBEM, and the second line shows the variance provided by the parent models. Note that, with the constraint $\sum_k \hat{\phi}_{k,l}^{i,j} \geq 1$, the regularization effect will not lose efficacy – even if N_k^i, $N_{k,k'}^i$ and $N_{1,k}^i$ are all zeros for the kth state in child $M_i^{(c)}$, that state will not degenerate. Comparing with VBEM, our method is equivalent to giving each parameter an exclusive prior, and the prior is provided by its parent models and averaged with weight $\hat{z}_{ij}\hat{\phi}_{k,l}^{i,j}$.

3.4.3 Variational Distributions for Parent Models

The functional form of the factors $q(M^{(p)})$ cannot be determined automatically by optimization of the variation distribution. Thus, we assume $q(M^{(p)})$ has the following form,

$$\omega^{(p)} \sim \mathrm{Dir}(\omega^{(p)}|\eta^{(p)}),$$
$$\pi^{(p),j} \sim \mathrm{Dir}(\pi^{(p),j}|\alpha^{(p),j}), \quad a_l^{(p),j} \sim \mathrm{Dir}(a_l^{(p),j}|\epsilon_l^{(p),j}),$$
$$\mu_l^{(p),j} \sim \mathcal{N}(\mu_l^{(p),j}|m_l^{(p),j}, (\beta_l^{(p),j}\Lambda_l^{(p),j})^{-1}),$$
$$\Lambda_l^{(p),j} \sim \mathcal{W}(\Lambda_l^{(p),j}|W_l^{(p),j}, \nu_l^{(p),j}).$$

The variational parameters are estimated by optimizing (2) w.r.t. each parameter (see Appendix C).

For variational parameters $\{\eta_j^{(p)}, \beta_l^{(p),j}, m_l^{(p),j}\}$, the updates are

$$\eta_j^{(p)} = \sum_i \hat{z}_{ij} + \eta_0^{(p)}, \quad \beta_l^{(p),j} = \beta_0^{(p)} + \beta_0^{(c)}\gamma_l^j, \quad (11)$$
$$m_l^{(p),j} = \Big[\beta_0^{(c)} B_l^j + \beta_0^{(p)}\nu_l^{(p),j} W_l^{(p),j}\Big]^{-1} \cdot (\beta_0^{(c)} m_l^j + \beta_0^{(p)}\nu_l^{(p),j} W_l^{(p),j} m_0^{(p)}),$$

and the aggregated statistics from child models are

$$B_l^j = \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{kl}^{i,j} \nu_k^{(c),i} W_k^{(c),i},$$
$$m_l^j = \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{kl}^{i,j} \nu_k^{(c),i} W_k^{(c),i} m_k^{(c),i},$$
$$\gamma_l^j = \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{kl}^{i,j} \frac{1}{d} \mathrm{Tr}\Big(\frac{(W_l^{(p),j})^{-1}\nu_k^{(c),i} W_k^{(c),i}}{\nu_l^{(p),j} - d - 1}\Big).$$

The child models conversely influence the parent models through: 1) the total precision B_l^j, which is the summation of precisions of the child models that are assigned to the lth state in jth parent model $M_j^{(p)}$; 2) the total modified mean m_l^j, which is the summation of means of the child models that are assigned to the lth state in $M_j^{(p)}$; (iii) and the total number of assignments γ_l^j, which are the number of child models assigned to the lth state in $M_j^{(p)}$, and $\frac{1}{d}\mathrm{Tr}\Big(\frac{(W_l^{(p),j})^{-1}\nu_k^{(c),i} W_k^{(c),i}}{\nu_l^{(p),j} - d - 1}\Big)$ is near to 1 if the precisions $\Lambda_l^{(p),j}$ and $\Lambda_k^{(c),i}$ are similar.

For $\alpha^{(p),j}$, $\epsilon_l^{(p),j}$, and $\nu_l^{(p),j}$, there are no closed-form solutions, and they are solved through numeric solvers. For $\alpha^{(p),j}$ and $\epsilon_l^{(p),j}$, we solve the optimization problems:

$$\begin{aligned} &\max\ \mathcal{L}(\alpha^{(p),j}), \quad \text{s.t. } \alpha^{(p),j} > 0, \\ &\max\ \mathcal{L}(\epsilon_l^{(p),j}), \quad \text{s.t. } \epsilon_l^{(p),j} > 0, \\ &\max\ \mathcal{L}(\nu_l^{(p),j}), \quad \text{s.t. } \nu_l^{(p),j} > d + 1, \end{aligned} \quad (12)$$

where $\mathcal{L}(\alpha^{(p),j})$, $\mathcal{L}(\epsilon_l^{(p),j})$, and $\mathcal{L}(\nu_l^{(p),j})$ contain all terms involving $\alpha^{(p),j}$, $\epsilon_l^{(p),j}$, and $\nu_l^{(p),j}$ in objective function (2), respectively. We restrict $\nu_l^{(p),j} > d + 1$, and thus the expectation in (10) will always exist.

Interestingly, for $W_l^{(p),j}$, setting the derivative of $\mathcal{L}(W_l^{(p),j})$ w.r.t. $W_l^{(p),j}$ to zero, we obtain a special case of an Algebraic Riccati Equation (see Appendix D),

$$-2cW_l^{(p),j} + W_l^{(p),j} R W_l^{(p),j} - Q = 0, \quad (13)$$

where

$$c = \frac{\nu_0^{(p)} - N_l^j \nu_0^{(c)}}{2},$$
$$R = \nu_l^{(p),j}\big[\beta_0^{(p)}(m_l^{(p),j} - m_0^{(p)})(m_l^{(p),j} - m_0^{(p)})^T + (W_0^{(p)})^{-1}\big],$$
$$Q = \frac{1}{\nu_l^{(p),j} - D - 1}\Big(\frac{\beta_0^{(c)}}{\beta_l^{(p),j}} + \nu_0^{(c)}\Big)\sum_i \hat{z}_{ij}\sum_k \hat{\phi}_{k,l}^{i,j}\nu_k^{(c),i} W_k^{(c),i}.$$

Note that R and Q are both symmetric positive definite matrices, since they are the weighted sums of symmetric positive definite matrices with positive coefficients. The following Lemma guarantees that $W_l^{(p),j}$ is a symmetric positive definite matrix.

Lemma 3.1 *The Algebraic Riccati Equation (ARE)*

$$-2cP + PRP - Q = 0$$

has a unique positive definite solution when R and Q are symmetric positive definite matrix.

The proof can be found in [Fazayeli and Banerjee, 2016]. There are several numerical methods to solve the ARE (see [Anderson and Moore, 2007]). In this paper, we use the Matlab ARE solver (icare) to find the solution of (13).

Table 1: Synthetic experiment results, averaged over 100 trials: (a) SSKL between parents $M^{(p)}$ and ground-truth; (b) clustering Rand-index; (c) SSKL between children $M^{(c)}$ and ground-truth. Standard deviations are in parentheses.

	Methods	$\tau = 100$	$\tau = 80$	$\tau = 50$	$\tau = 30$	$\tau = 10$
(a)	VBEM+VHEM	18.70 (61.2)	41.94 (103.4)	28.37 (95.6)	29.97 (59.8)	35.62 (12.3)
	EM+VHEM	387.49 (234.3)	284.65 (183.4)	208.83 (131.2)	247.81 (194.5)	158.68 (134.8)
	VBEM+VBHEM	69.64 (110.5)	61.43 (95.1)	58.67 (93.9)	142.12 (92.9)	89.68 (125.9)
	EM+VBHEM	220.59 (153.7)	259.98 (172.3)	99.12 (68.7)	312.63 (164.0)	233.65 (159.7)
	Ours	**0.63** (1.6)	**0.73** (1.8)	**1.10** (2.5)	**1.98** (4.5)	**3.43** (5.6)
(b)	VBEM+VHEM	**1.000** (.00)	**1.000** (.00)	**1.000** (.00)	**1.000** (.00)	**1.000** (.00)
	EM+VHEM	0.646 (.22)	0.837 (.22)	0.830 (.20)	0.781 (.19)	0.843 (.19)
	VBEM+VBHEM	0.990 (.07)	0.990 (.07)	0.990 (.07)	0.990 (.07)	0.990 (.07)
	EM+VBHEM	0.843 (.22)	0.780 (.24)	0.705 (.22)	0.660 (.18)	0.758 (.18)
	Ours	**1.000** (.00)	**1.000** (.00)	**1.000** (.00)	**1.000** (.00)	**1.000** (.00)
(c)	VBEM	0.75 (2.24)	0.85 (2.32)	1.44 (3.71)	2.83 (1.23)	6.70 (8.88)
	EM	32.20 (15.47)	31.77 (15.32)	26.01 (12.34)	25.72 (11.91)	21.43 (34.74)
	Ours	**0.31** (0.76)	**0.36** (0.93)	**0.54** (1.28)	**1.21** (1.18)	**2.23** (5.32)

4 EXPERIMENTS

In this section, we test our method on synthetic data and real data to show that our method can: (1) learn $M^{(c)}$ and $M^{(p)}$ simultaneously; (2) jointly regularize $M^{(c)}$ and $M^{(p)}$; and (3) give good clustering results.

4.1 SYNTHETIC DATA

We generate a synthetic dataset via: (1) randomly generating 2 ground-truth (GT) HMMs with 3 states in $\mathbb{R}^2$; (2) sample 50 sample sets from each GT HMM and each sample consists of 30 sequences with length $\tau = 100$ [Chan and Hsiao, 2018]; (3) add noise $e \sim \mathcal{N}(0_2, \mathrm{I}_2)$ to each observation. Thus, $K^{(p)} = 2$, $K^{(c)} = 100$, $S^{(c)} = 3$, and $S^{(p)} = 3$.

4.1.1 Experiment setup

We compare our method with two-stage hierarchical model learning methods: EM+VHEM, EM+VBHEM, VBEM+VHEM, and VBEM+VBHEM. The child HMMs are learned with standard EM or VBEM. The parent HMMs are learned via hierarchical clustering, VHEM [Coviello et al., 2014] or VBHEM [Lan et al., 2021]. We test on sequences with different lengths, $\tau \in \{100, 80, 50, 30, 10\}$, to show the effect on small sample size.

In order to quantify how close $M^{(p)}$ and $M^{(c)}$ are to the GT, we define a simple symmetric KL divergence (SSKL) between two HMMs, $M_i = \{\pi_i, A_i, \theta_i\}$, $i = 1, 2$, as the sum of SKL between each pair of HMM parameters (π_i, A_i, θ_i), which is minimized over all state permutations,

$$\mathrm{SSKL}(M_1||M_2) = \min_{SP} \big\{\mathrm{SKL}(\pi_1||\pi_2) + \mathrm{SKL}(A_1||A_2) + \mathrm{SKL}(\Theta_1||\Theta_2)\big\},$$

where SP is all possible state permutations. Note that here we compare the HMM parameters separately, rather than the whole HMM (as a time-series density), since we are interested in recovering and interpreting the GT parameters.

4.1.2 Experiment results

The SSKL between $M^{(p)}$ and GT are shown in Tab. 1(a). Our method has significantly lower SSKL than other methods on sequences of every length, indicating that our method well recovers the GT parent models. As the sequence length τ decreases, the SSKL will increase in most cases – our method has the least amount of increase and smallest standard deviation.

Fig. 3 shows five different GT HMMs for each test. Our method is closer to the GT compared with other methods, even though the emission distributions may have large overlaps. When $\tau = 10$, our method is the only one that is close to the GT, which is due to the good regularization from the parent model, which pool common information from the children. Some methods poorly estimate the emission densities (e.g., the green state in EV(a)) because of wrong clustering or badly learned individual HMMs.

The SSKL between $M^{(c)}$ and GT HMMs are shown in Tab. 1(c). Our method also has lowest SSKL and smallest standard deviation. Fig. 3 (bottom) shows the individual models learned by three methods. VBEM and EM methods tend to overfit, e.g., the HMMs for VB(b) and EM(e), while our method overcomes this problem and learns better individual HMMs.

Next, we compare the clustering performance using Rand-index [Hubert and Arabie, 1985] against the ground-truth clusters, shown in Tab. 1(b). Our method and VBEM+VHEM have perfect clustering results, while

Figure 3: Illustration of HMMs from synthetic data. The first row is the ground-truth (GT). The following rows show parent HMMs in $M^{(p)}$ from VBEM+VHEM (BV), EM+VHEM (EV), VBEM+VBHEM (BB), EM+VBHEM (EB), and our method, respectively. The last three rows show child HMMs in $M^{(c)}$ from VBEM (VB), EM, and our method.

Table 2: Averaged log-likelihood for different number of sample sets N. Standard deviation in parenthesis.

	$N = 10$	$N = 20$	$N = 50$	$N = 100$
VBEM	-8.76 (.024)	-8.51 (.014)	-8.47 (.007)	-8.43 (.005)
Ours	**-8.59** (.029)	**-8.47 (.012)**	**-8.44 (.006)**	**-8.42 (.005)**

VBEM+VBHEM is second best. Although VBEM+VHEM obtains perfect clustering results, our estimates of the parent model are much better than those of VBEM+VHEM (see Tab. 1a), which is due to our iterative updating between the child and parent models. Our child model estimates are also better than VBEM (Tab. 1(c)). The individual HMM from VBEM may have wrong emission densities (see Fig. 3(c), VB), leading to bad group HMMs (Fig. 3(c), VB+BV).

Finally, to demonstrate the regularization effect, we calculate the log-likelihood (LL) of held-out test data ($K = 2, S = 3, \tau = 10$) for models learned with VBEM and Ours using different numbers of sample sets. The results are averaged over 100 trials, and appear in Tab. 2. The LLs of our method are consistently higher than VBEM, especially for small N (for each N, paired t-test $p < .001$).

In short, the experiment results show that our method can learn $M^{(c)}$ and $M^{(p)}$ simultaneously, $M^{(c)}$ and $M^{(p)}$ regularize each other, and also obtain good clustering results.

4.2 EYE MOVEMENT DATA

Hsiao et al. [2021c] collect eye movement data from stimuli with different feature layouts to explore participant groups with consistent eye movement patterns. The data contains the eye movement sequences of 61 participants when they view 120 stimuli. With these data, we demonstrate that our method has good regularization effect.

Following [Hsiao et al., 2021c], for each stimulus, we have $K^{(c)} = 61$ and set $S^{(c)} = 3$, $S^{(p)} = 3$, and $K^{(p)} = 2$.

Table 3: Eye movement data: Average log-likelihood on held-out test data. Standard deviation in parenthesis.

	$\tau = 3$	$\tau = 5$	$\tau = 10$	$\tau = 15$
VBEM	-123.94 (42.83)	-47.51 (13.48)	-15.52 (2.41)	-14.09 (5.14)
Our	**-14.73** (2.49)	**-12.22** (0.94)	**-11.44** (0.47)	**-11.56** (0.67)

Each individual HMM models the eye fixation sequence of one participant viewing one stimulus. We train individual HMM using different length of fixations sequence, i.e., $\tau \in \{3, 5, 10, 15\}$ and use the remaining data as test data, and compare with VBEM. We calculate the log-likelihood of individual HMMs on test data to examine the performance of regularization effect.

Tab. 3 shows the log-likelihood on test data, averaged over participants and stimuli. Comparing with VBEM, our method has larger log-likelihood under all scenarios, especially when $\tau = 3$. Moreover, with the decrease of the length of training data, our method degrades less significantly than VBEM, which shows the individual HMM learned by our method has a good generalization ability. The illustration of individual HMMs is shown in the Appendix. In summary, our framework effectively learns HMMs with good performance.

5 CONCLUSION

In this paper, we propose a tree structure variational Bayesian method, which learn individual models and group models simultaneously. The group model regularizes the individual models, which shows advantages on small data sets. We test our method on synthetic data and real data, and demonstrate our method learns individual models and group models together and has good clustering results. In the future, we will extend our method to automatically determine the number of components and the number of states in individual models and group models. We will also apply our model to real applications in other fields.

Acknowledgements

We sincerely thank Dr. Janet H. Hsiao for collecting the eye movement data and helpful discussions. The work described in this paper was supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005218).

References

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

B. D. Anderson and J. B. Moore. *Optimal control: Linear quadratic methods*. Courier Corporation, 2007.

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

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.*, 41(1):164–171, 1970.

M. J. Beal et al. *Variational Algorithms for Approximate Bayesian Inference*. University of London, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *JASA*, 112(518): 859–877, 2017.

D. Bouchacourt, R. Tomioka, and S. Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference*, volume 32, 2018.

A. B. Chan and J. H. Hsiao. EMHMM simulation study, 2018.

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.*, pages 1–8, 2018.

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. *Cognition and Emotion*, 34(8):1704–1710, 2020a.

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*, 24(10):1956–1967, 2020b.

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*, 24(7):1242–1256, 2020c.

T. Chuk, A. B. Chan, and J. H. Hsiao. Understanding eye movements in face recognition using hidden Markov models. *J. Vision*, 14(11):8–8, 2014.

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.*, 141:204–216, 2017a.

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*, 169:102–117, 2017b.

T. Chuk, A. B. Chan, S. Shimojo, and J. H. Hsiao. Eye movement analysis with switching hidden Markov models. *Behavior Research Methods*, 52:1026–1043, 2020.

E. Coviello, A. B. Chan, and G. R. G. Lanckriet. Clustering hidden Markov models with variational HEM. *J. Mach. Learn. Res.*, 15(1):697–747, Jan. 2014. ISSN 1532-4435.

P. Diaconis, D. Ylvisaker, et al. Conjugate priors for exponential families. *Ann. Stat.*, 7(2):269–281, 1979.

F. Fazayeli and A. Banerjee. The matrix generalized inverse Gaussian distribution: Properties and applications. In *ECML PKDD*, pages 648–664. Springer, 2016.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *JASA*, 85(410):398–409, 1990.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

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*, 211, 104616, 2021a.

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.*, 2021b. (in press).

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*, 2021c.

L. Hubert and P. Arabie. Comparing partitions. *J. Classif.*, 2(1):193–218, 1985.

B. H. Juang and L. R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

D.-k. Kim, G. Voelker, and L. Saul. A variational approximation for topic modeling of hierarchical corpora. In *ICML*, pages 55–63. PMLR, 2013.

A. Kucukelbir, R. Ranganath, A. Gelman, and D. M. Blei. Automatic variational inference in Stan. *arXiv preprint arXiv:1506.03431*, 2015.

H. Lan, Z. Liu, J. H. Hsiao, D. Yu, and A. B. Chan. Clustering hidden Markov models with variational Bayesian hierarchical EM. *IEEE TNNLS*, 2021.

M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang. Transfer learning with graph co-regularization. *IEEE TKDE*, 26 (7):1805–1818, 2013.

J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan. Nested hierarchical Dirichlet processes. *IEEE TPAMI*, 37(2): 256–270, 2014.

Y. Pang, S. Wang, and Y. Yuan. Learning regularized lda by clustering. *IEEE TNNLS*, 25(12):2191–2201, 2014.

B. S. Price, A. J. Molstad, and B. Sherwood. Estimating multiple precision matrices with cluster fusion regularization. *J. Comput. Graph. Stat.*, pages 1–30, 2021.

Y. Qi, J. W. Paisley, and L. Carin. Dirichlet process HMM mixture models with application to music analysis. In *ICASSP*, volume 2, pages II–465. IEEE, 2007.

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

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *ICML*, pages 324–333. PMLR, 2016.

P. Sellars, A. Aviles-Rivero, and C. B. Schönlieb. Two cycle learning: Clustering based regularisation for deep semi-supervised classification. *arXiv preprint arXiv:2001.05317*, 2020.

J. Simola, J. Salojärvi, and I. Kojo. Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive systems research*, 9(4):237–251, 2008.

P. Smyth. Clustering sequences with hidden Markov models. In *Proc. NeurIPS*, pages 648–654, 1997.

R. G. Soares, H. Chen, and X. Yao. Semisupervised classification with cluster regularization. *IEEE TNNLS*, 23(11): 1779–1792, 2012.

R. G. Soares, H. Chen, and X. Yao. A cluster-based semisupervised ensemble for multiclass classification. *IEEE TETCI*, 1(6):408–420, 2017.

C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder variational autoencoders. *arXiv preprint arXiv:1602.02282*, 2016.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581, 2006.

A. Zhang, S. Gultekin, and J. Paisley. Stochastic variational inference for the HDP-HMM. In *AISTAT*, pages 800–808. PMLR, 2016.

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*, 42 (2):zsy220, 2019.

K. Zhao, J. Xu, and M.-M. Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the CVPR*, pages 1136–1144, 2019.