Hierarchical Learning of Hidden Markov Models with Clustering Regularization Supplementary material

Hui Lan[1,2,3] **Janet H. Hsiao**[4,5] **Antoni B. Chan**[1]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong
[2]Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Department of Psychology, University of Hong Kong, Pok Fu Lam, Hong Kong
[5]The State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Pok Fu Lam, Hong Kong

A DERIVATION OF ELBO

A central task in the application of probabilistic models is the evaluation of the posterior distribution of the latent variables given the observed (visible) data variables. For many models of practical interest, it will be infeasible to evaluate the posterior distribution since the dimensionality of the latent space is too high to work with directly or because the posterior distribution has a highly complex form for which expectations are not analytically tractable. This is the case in our approach. Therefore, we resort to the approximation techniques called variational inference or variational Bayes to estimate $M^{(c)}$ and $M^{(p)}$. Similar with Beal et al. [2003], we can decompose the log marginal probability $\log p(Y)$ as

$$\begin{aligned}\log p(Y) &= \log \int p(Y|M^{(p)})p(M^{(p)})dM^{(p)} \\ &\geq \sum_i \mathbb{E}_{q(M^{(p)})}\big[\log p(Y_i|M^{(p)})\big] - \mathrm{KL}(q(M^{(p)})||p(M^{(p)})),\end{aligned} \tag{1}$$

where $\mathrm{KL}(q(x)||p(x)) = \int q(x)\log\frac{q(x)}{p(x)}dx$ is the Kullback-Leibler divergence (KLD), which measures the similarity between distribution q and p. We treat the parents $M^{(p)}$ as a hidden variable with prior $p(M^{(p)})$, and $q(M^{(p)})$ is its variational posterior distribution. As $M^{(p)}$ is a mixture model, to simplify the log-likelihood $\log p(Y_i|M^{(p)})$, we introduce a hidden assignment variable Z_i, where $Z_i = j$ if Y_i and its associated child model $M_i^{(c)}$ are assigned to $M_j^{(p)}$. Next, we obtain a lower bound on the log-likelihood $\log p(Y_i|M^{(p)})$,

$$\begin{aligned}\log p(Y_i|M^{(p)}) &= \log \sum_j p(Y_i, Z_i = j|M^{(p)}) \\ &\geq \sum_j \hat{z}_{ij}\big[\log p(Y_i|M_j^{(p)}) + \log \omega_j^{(p)} - \log \hat{z}_{ij}\big],\end{aligned} \tag{2}$$

where $\omega_j^{(p)} = p(Z_i = j|\omega^{(p)})$ and we denote the variational distribution $z_{ij} = q(Z_i = j)$ and $\hat{z}_{ij} = \mathbb{E}[z_{ij}]$. Note that one child in our model only has one parent but one parent has several children, as shown in Fig. 2. 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.) data sequences $y_n^i \sim M_i^{(c)}, n = 1, ..., N_i$.

Thus, we define the likelihood $p(Y_i|M_j^{(p)})$ as the marginalization over the corresponding child model $M_i^{(c)}$, and obtain its lower bound,

$$\log p(Y_i|M_j^{(p)}) = \log \int p(Y_i, M_i^{(c)}|M_j^{(p)})dM_i^{(c)} \tag{3}$$

$$\geq \mathbb{E}_{q(M_i^{(c)})}\big[\log p(Y_i|M_i^{(c)}) + \log p(M_i^{(c)}|M_j^{(p)}) - \log q(M_i^{(c)})\big],$$

where $q(M_i^{(c)})$ is the variational posterior distribution of child $M_i^{(c)}$. Finally, combining (1), (2), and (3), we obtain the our objective function (lower bound),

$$\begin{aligned}\mathcal{L}(M^{(p)}, M^{(c)}) = &\sum_i \mathbb{E}_{q(M_i^{(c)})}\big[\log p(Y_i|M_i^{(c)})\big] + \sum_i\sum_j \hat{z}_{ij}\mathbb{E}_{q(M_i^{(c)})}\mathbb{E}_{q(M_j^{(p)})}\big[\log p(M_i^{(c)}|M_j^{(p)})\big] \\ &+ \sum_i\sum_j \hat{z}_{ij}\mathbb{E}_{q(\omega^{(p)})}\big[\log \omega_j^{(p)}\big] + \mathbb{E}_{\omega^{(p)}} \log p(\omega^{(p)}) + \sum_j \mathbb{E}_{q(M_j^{(p)})}\big[\log p(M_j^{(p)})\big] - \sum_i\sum_j \hat{z}_{ij}\log\hat{z}_{ij} \\ &- \sum_i \mathbb{E}_{q(M_i^{(c)})}\big[\log q(M_i^{(c)})\big] - \sum_j \mathbb{E}_{q(M_j^{(p)})}\big[\log q(M_j^{(p)})\big] - \mathbb{E}_{q(\omega^{(p)})}\big[\log q(\omega^{(p)})\big]. \qquad (4)\end{aligned}$$

where $\hat{z}_{i,j} = \mathbb{E}_{q(Z)}[z_{ij}]$. Since $\log p(Y) \geq \mathcal{L}(M^{(p)}, M^{(c)})$, we find the best child models $M_*^{(c)}$ and the best parent models $M_*^{(p)}$ through

$$\Big\{M_*^{(p)}, M_*^{(c)}\Big\} = \underset{M^{(p)}, M^{(c)}}{\arg\max}\ \mathcal{L}(M^{(p)}, M^{(c)}).$$

B VARIATIONAL DISTRIBUTION OF ϕ

In this section, we give the specific $\mathcal{L}(\hat{\phi}^{i,j})$. The term involved $\hat{\phi}^{i,j}$ in our objective function (4) is $\mathbb{E}_{q(M_i^{(c)})}\mathbb{E}_{q(M_j^{(p)})}\big[\log p(M_i^{(c)}|M_j^{(p)})\big]$, i.e.,

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

we derive $\mathbb{E}_{q(M_i^{(c)})}\mathbb{E}_{q(M_j^{(p)})}\big[\log p(M_i^{(c)}|M_j^{(p)})\big]$ first and then give $\mathcal{L}(\hat{\phi}^{i,j})$.

B.1 $\mathbb{E}_{q(M_i^{(c)})}\mathbb{E}_{q(M_j^{(p)})} \log p(M_i^{(c)}|M_j^{(p)})$

$$\begin{aligned}&\mathbb{E}_{q(M_i^{(c)})}\mathbb{E}_{q(M_j^{(p)})} \log p(M_i^{(c)}|M_j^{(p)}) \geq \mathbb{E}_{q(M_i^{(c)})}\mathbb{E}_{q(M_j^{(p)})}\mathbb{E}_{q(\phi^{i,j})}\Big[\log p(M_i^{(c)}|\phi^{i,j}, M_j^{(p)}) + \log p(\phi^{i,j}) - \log q(\phi^{i,j})\Big] \\ &= \mathbb{E}_{q(\pi^{(c),i})}\mathbb{E}_{q(\pi^{(p),j})}\mathbb{E}_{q(\phi^{i,j})}\big[\log p(\pi^{(c),i}|\alpha_0^{(c)}, \phi^{i,j}, \pi^{(p),j})\big] + \sum_k \mathbb{E}_{q(a_k^{(c),i})}\mathbb{E}_{q(A^{(p),j})}\mathbb{E}_{q(\phi^{i,j})}\big[\log p(a_k^{(c),i}|\epsilon_0^{(c)}, \phi^{i,j}, A^{(p),j})\big] \\ &\quad + \sum_k\sum_l \mathbb{E}_{q(\phi^{i,j})}[\phi_{k,l}^{i,j}]\mathbb{E}_{q(\mu_k^{(c),i}, \Lambda_k^{(c),i})}\mathbb{E}_{q(\mu_l^{(p),j}, \Lambda_l^{(p),j})}\Big[\log \mathcal{N}(\mu_k^{(c),i}|\mu_l^{(p),j}, (\beta_0^{(c)}\Lambda_k^{(c),i})^{-1})\Big] \\ &\quad + \sum_k\sum_l \mathbb{E}_{q(\phi^{i,j})}[\phi_{k,l}^{i,j}]\mathbb{E}_{q(\Lambda_k^{(c),i})}\mathbb{E}_{q(\Lambda_l^{(p),j})}\Big[\log \mathcal{W}(\Lambda_k^{(c),i}|\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)})\Big]. \qquad (5)\end{aligned}$$

Consider the first line in the RHS of (5), we simplify the normalization terms in two Dirichlet distributions. Firstly, $\phi_{k,l}^{i,j}$ is a binary variable and $\sum_l \phi_{k,l}^{i,j} = 1$. Thus,

$$\log\Gamma(\alpha_0^{(c)}\sum_l \phi_{k,l}^{i,j}\pi_l^{(p),j}) \leq \sum_l \phi_{k,l}^{i,j}\log\Gamma(\alpha_0^{(c)}\pi_l^{(p),j}), \qquad (6)$$

since $\log\Gamma$ is convex. 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 l-th state in $M_j^{(p)}$. Since $\log\Gamma(\rho\cdot a) \geq \rho\cdot\log\Gamma(a),\ \text{if}\ a \geq 1, \rho \geq 1$ [1], we have

$$\log\Gamma(\alpha_0^{(c)}\sum_k\sum_l \phi_{k,l}^{i,j}\pi_l^{(p),j}) \geq \sum_k\sum_l \phi_{k,l}^{i,j}\pi_l^{(p),j}\log\Gamma(\alpha_0^{(c)}), \qquad (7)$$

[1] Let $f(a) = \log\Gamma(\rho a) - \rho\log\Gamma(a)$, $f'(a) = \rho(\psi(\rho a) - \psi(a)) \geq 0$, thus $f(a)$ is a monotonic increasing function and $f(a) \geq f(1) = 0$.

and $\alpha_0^{(c)} \geq 1$. Combining (6) and (7), firstly we obtain a lower bound on the normalization term $\log C(\tilde{\alpha}^{i,j})$ for $\mathrm{Dir}(\pi^{(c),i})$, and secondly, we bring $\phi_{k,l}^{i,j}$ out of the log-gamma function. The normalization term $\log C(\tilde{\epsilon}_k^{i,j})$ for prior on $a_k^{(c),i}$ is similar.

Next, for each term in the RHS of (5), we have,

1.
$$
\begin{aligned}
&\mathbb{E}_{q(\pi^{(c),i})}\mathbb{E}_{q(\pi^{(p),j})}\mathbb{E}_{q(\phi^{i,j})}\big[\log p(\pi^{(c),i}|\alpha_0^{(c)}, \phi^{i,j}, \pi^{(p),j})\big] \\
&= \mathbb{E}_{q(\pi^{(c),i})}\mathbb{E}_{q(\pi^{(p),j})}\mathbb{E}_{q(\phi^{i,j})}\Big[\log C(\tilde{\alpha}^{(c)}) + \sum_k \big(\sum_l \alpha_0^{(c)}\phi_{k,l}^{i,j}\pi_l^{(p),j} - 1\big)\log \pi_k^{(c),i}\Big] \\
&\geq \sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\hat{\pi}_l^{(p),j}\log\Gamma(\alpha_0^{(c)}) - \sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\mathbb{E}\big[\log\Gamma(\pi_l^{(p),j}\alpha_0^{(c)})\big] + \sum_k \big(\alpha_0^{(c)}\sum_l \hat{\phi}_{k,l}^{i,j}\hat{\pi}_l^{(p),j} - 1\big)\log\tilde{\pi}_k^{(c),i}.
\end{aligned}
$$

2.
$$
\begin{aligned}
&\sum_k \mathbb{E}_{q(a_k^{(c),i})}\mathbb{E}_{q(A^{(p),j})}\mathbb{E}_{q(\phi^{i,j})}\big[\log p(a_k^{(c),i}|\epsilon_0^{(c)}, \phi^{i,j}, A^{(p),j})\big] \\
&\geq \sum_k\sum_{k'}\Big[\sum_l\sum_{l'}\hat{\phi}_{k,l}^{i,j}\hat{\phi}_{k',l'}^{i,j}\hat{a}_{l,l'}^{(p),j}\Big]\log\Gamma(\epsilon_0^{(c)}) - \sum_k\sum_{k'}\Big[\sum_l\sum_{l'}\hat{\phi}_{k,l}^{i,j}\hat{\phi}_{k',l'}^{i,j}\Big]\mathbb{E}\log\Gamma(a_{l,l'}^{(p),j}\epsilon_0^{(c)}) \\
&\quad + \sum_k\sum_{k'}\Big[\epsilon_0^{(c)}\sum_l\sum_{l'}\hat{\phi}_{k,l}^{i,j}\hat{\phi}_{k',l'}^{i,j}\hat{a}_{l,l'}^{(p),j} - 1\Big]\log\tilde{a}_{k,k'}^{(c),i}.
\end{aligned}
$$

3.
$$
\begin{aligned}
&\sum_k\sum_l \mathbb{E}_{q(\phi^{i,j})}\big[\phi_{k,l}^{i,j}\big]\mathbb{E}_{q(\mu_k^{(c),i},\Lambda_k^{(c),i})}\mathbb{E}_{q(\mu_l^{(p),j},\Lambda_l^{(p),j})}\Big[\log\mathcal{N}(\mu_k^{(c),i}|\mu_l^{(p),j}, (\beta_0^{(c)}\Lambda_k^{(c),i})^{-1})\Big] \\
&= \sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\mathbb{E}_{q(\mu_k^{(c),i},\Lambda_k^{(c),i})}\mathbb{E}_{q(\mu_l^{(p),j},\Lambda_l^{(p),j})}\Big[\frac{D}{2}\log\frac{\beta_0^{(c)}}{2\pi} + \frac{1}{2}\log|\Lambda_k^{(c),i}| - \frac{\beta_0^{(c)}}{2}(\mu_k^{(c),i} - \mu_l^{(p),j})^T\Lambda_k^{(c),i}(\mu_k^{(c),i} - \mu_l^{(p),j})\Big] \\
&= \sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\Bigg\{\frac{D}{2}\log\frac{\beta_0^{(c)}}{2\pi} + \frac{1}{2}\log\tilde{\Lambda}_k^{(c),i} - \frac{\beta_0^{(c)}}{2}\Big[\frac{D}{\beta_k^{(c),i}} + \nu_k^{(c),i}(m_k^{(c),i} - m_l^{(p),j})^T W_k^{(c),i}(m_k^{(c),i} - m_l^{(p),j}) \\
&\quad + \frac{\nu_k^{(c),i}}{\beta_l^{(p),j}(\nu_l^{(p),j} - D - 1)}\mathrm{Tr}((W_l^{(p),j})^{-1}W_k^{(c),i})\Big]\Bigg\} \\
&= S^{(c)}\frac{D}{2}\log\frac{\beta_0^{(c)}}{2\pi} + \frac{1}{2}\sum_k\big[\log\tilde{\Lambda}_k^{(c),i} - \beta_0^{(c)}\frac{D}{\beta_k^{(c),i}}\big] - \frac{\beta_0^{(c)}}{2}\sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\nu_k^{(c),i}(m_k^{(c),i} - m_l^{(p),j})^T W_k^{(c),i}(m_k^{(c),i} - m_l^{(p),j}) \\
&\quad - \frac{\beta_0^{(c)}}{2}\sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\frac{\nu_k^{(c),i}}{\beta_l^{(p),j}(\nu_l^{(p),j} - D - 1)}\mathrm{Tr}\big((W_l^{(p),j})^{-1}W_k^{(c),i}\big).
\end{aligned}
$$

4.
$$
\begin{aligned}
&\sum_k \mathbb{E}_{q(\Lambda_k^{(c),i})}\mathbb{E}_{q(\Lambda_l^{(p),j})}\sum_l \mathbb{E}_{q(\phi^{i,j})}\big[\phi_{k,l}^{i,j}\big]\Big[\log\mathcal{W}(\Lambda_k^{(c),i}|\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)})\Big] \\
&= \sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\mathbb{E}_{q(\Lambda_k^{(c),i})}\mathbb{E}_{q(\Lambda_l^{(p),j})}\Bigg\{\log B(\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)}) + \frac{\nu_0^{(c)} - D - 1}{2}\log|\Lambda_k^{(c),i}| - \frac{\nu_0^{(c)}}{2}\mathrm{Tr}((\Lambda_l^{(p),j})^{-1}\Lambda_k^{(c),i})\Bigg\} \\
&= \sum_k\sum_l \hat{\phi}_{k,l}^{i,j}\Bigg\{-\frac{\nu_0^{(c)}}{2}\log\tilde{\Lambda}_l^{(p),j} + \frac{\nu_0^{(c)}}{2}\log|\nu_0^{(c)}| - \Big[\frac{\nu_0^{(c)}D}{2}\log 2 + \frac{D(D-1)}{4}\log\pi \\
&\quad + \sum_{d=1}^{D}\log\Gamma(\frac{\nu_0^{(c)} + 1 - d}{2})\Big] + \frac{\nu_0^{(c)} - D - 1}{2}\log\tilde{\Lambda}_k^{(c),i} - \frac{\nu_0^{(c)}\nu_k^{(c),i}}{2(\nu_l^{(p),j} - D - 1)}\mathrm{Tr}((W_l^{(p),j})^{-1}W_k^{(c),i})\Bigg\}.
\end{aligned}
$$

And,

$$
\mathbb{E}_{q(\Lambda_k^{(c),i})}\mathbb{E}_{q(\Lambda_l^{(p),j})}\Big[\log\mathcal{W}(\Lambda_k^{(c),i}|\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)})\Big]
$$

$$
\begin{aligned}
&= \mathbb{E}_{q(\Lambda_k^{(c),i})}\mathbb{E}_{q(\Lambda_l^{(p),j})}\Big\{ \log B(\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)}) + \frac{\nu_0^{(c)} - D - 1}{2} \log |\Lambda_k^{(c),i}| - \frac{\nu_0^{(c)}}{2} \operatorname{Tr}((\Lambda_l^{(p),j})^{-1}\Lambda_k^{(c),i}) \Big\} \\
&= -\frac{\nu_0^{(c)}}{2} \log \tilde{\Lambda}_l^{(p),j} + \frac{\nu_0^{(c)}}{2} \log |\nu_0^{(c)}| - \Big[\frac{\nu_0^{(c)} D}{2} \log 2 + \frac{D(D-1)}{4} \log \pi + \sum_{d=1}^{D} \log \Gamma(\frac{\nu_0^{(c)} + 1 - d}{2})\Big] \\
&\quad + \frac{\nu_0^{(c)} - D - 1}{2} \log \tilde{\Lambda}_k^{(c),i} - \frac{\nu_0^{(c)} \nu_k^{(c),i}}{2(\nu_l^{(p),j} - D - 1)} \operatorname{Tr}((W_l^{(p),j})^{-1} W_k^{(c),i}).
\end{aligned}
$$

B.2 $\mathcal{L}(\hat{\phi}^{i,j})$

$$
\begin{aligned}
\mathcal{L}(\hat{\phi}^{i,j}) = \sum_l \sum_k \hat{\phi}_{k,l}^{i,j} \Big\{ & \hat{\pi}_l^{(p),j} \log \Gamma(\alpha_0^{(c)}) - \mathbb{E} \log \Gamma(\alpha_0^{(c)} \pi_l^{(p),j}) + \alpha_0^{(c)} \hat{\pi}_l^{(p),j} \log \tilde{\pi}_k^{(c),i} \\
& + \mathbb{E} \log \mathcal{N}(\mu_k^{(c),i} | \mu_l^{(p),j}, (\beta_0^{(c)} \Lambda_k^{(c),i})^{-1}) + \mathbb{E} \log \mathcal{W}(\Lambda_k^{(c),i} | \Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)}) - \log S^{(p)} - \log \hat{\phi}_{k,l}^{i,j} \Big\} \\
& + \sum_k \sum_{k'} \hat{\phi}_{k\cdot}^{i,j} \Big\{ A^{(p),j} \log \Gamma(\epsilon_0^{(c)}) + \mathbb{E} \log \Gamma(\epsilon_0^{(c)} A^{(p),j}) + \epsilon_0^{(c)} A^{(p),j} \log \tilde{a}_{k,k'}^{(p),j} \Big\} (\hat{\phi}_{k'\cdot}^{i,j})^T,
\end{aligned}
$$

where $\log \tilde{\pi}_k^{(c),i} = \mathbb{E}[\log \pi_k^{(c),i}]$ and

1. $\mathbb{E} \log \Gamma(\alpha_0^{(c)} \pi_l^{(p),j})$

Kim et al. [2013] propose an upper bound for $\mathbb{E} \log \Gamma(\alpha_0^{(c)} \pi_l^{(p),j})$, i.e.,

$$
\mathbb{E} \log \Gamma(\alpha_0^{(c)} \pi_l^{(p),j}) \leq \log \Gamma(\alpha_0^{(c)} \hat{\pi}_l^{(p),j}) + \alpha_0^{(c)} (1 - \hat{\pi}_l^{(p),j})/\hat{\alpha}^{(p),j} + (1 - \alpha_0^{(c)} \hat{\pi}_l^{(p),j})[\log \hat{\pi}_l^{(p),j} + \psi(\hat{\alpha}^{(p),j}) - \psi(\alpha_l^{(p),j})], \tag{8}
$$

where $\hat{\pi}_l^{(p),j} = \mathbb{E}[\pi_l^{(p),j}]$ and $\hat{\alpha}^{(p),j} = \sum_l \alpha_l^{(p),j}$.

2. $\mathbb{E} \log \mathcal{N}(\mu_k^{(c),i} | \mu_l^{(p),j}, (\beta_0^{(c)} \Lambda_k^{(c),i})^{-1})$

There is

$$
\begin{aligned}
& \mathbb{E}_{q(\mu_k^{(c),i}, \Lambda_k^{(c),i})} \mathbb{E}_{q(\mu_l^{(p),j}, \Lambda_l^{(p),j})} \Big[\log \mathcal{N}(\mu_k^{(c),i} | \mu_l^{(p),j}, (\beta_0^{(c)} \Lambda_k^{(c),i})^{-1}) \Big] \\
&= \mathbb{E}_{q(\mu_k^{(c),i}, \Lambda_k^{(c),i})} \mathbb{E}_{q(\mu_l^{(p),j}, \Lambda_l^{(p),j})} \Big[\frac{D}{2} \log \frac{\beta_0^{(c)}}{2\pi} + \frac{1}{2} \log |\Lambda_k^{(c),i}| - \frac{\beta_0^{(c)}}{2} (\mu_k^{(c),i} - \mu_l^{(p),j})^T \Lambda_k^{(c),i} (\mu_k^{(c),i} - \mu_l^{(p),j}) \Big] \\
&= \frac{D}{2} \log \frac{\beta_0^{(c)}}{2\pi} + \frac{1}{2} \log \tilde{\Lambda}_k^{(c),i} - \frac{\beta_0^{(c)}}{2} \Big[\frac{D}{\beta_k^{(c),i}} + \nu_k^{(c),i} (m_k^{(c),i} - m_l^{(p),j})^T W_k^{(c),i} (m_k^{(c),i} - m_l^{(p),j}) \\
&\quad + \frac{\nu_k^{(c),i}}{\beta_l^{(p),j} (\nu_l^{(p),j} - D - 1)} \operatorname{Tr}((W_l^{(p),j})^{-1} W_k^{(c),i}) \Big],
\end{aligned}
$$

where $\log \tilde{\Lambda}_k^{(c),i} = \mathbb{E}[\log \Lambda_k^{(c),i}]$.

3. $\mathbb{E} \log \mathcal{W}(\Lambda_k^{(c),i} | \Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)})$

There is

$$
\begin{aligned}
& \mathbb{E}_{q(\Lambda_k^{(c),i})} \mathbb{E}_{q(\Lambda_l^{(p),j})} \Big[\log \mathcal{W}(\Lambda_k^{(c),i} | \Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)}) \Big] \\
&= \mathbb{E}_{q(\Lambda_k^{(c),i})} \mathbb{E}_{q(\Lambda_l^{(p),j})} \Big\{ \log B(\Lambda_l^{(p),j}/\nu_0^{(c)}, \nu_0^{(c)}) + \frac{\nu_0^{(c)} - D - 1}{2} \log |\Lambda_k^{(c),i}| - \frac{\nu_0^{(c)}}{2} \operatorname{Tr}((\Lambda_l^{(p),j})^{-1} \Lambda_k^{(c),i}) \Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{\nu_0^{(c)}}{2}\log\tilde{\Lambda}_l^{(p),j} + \frac{\nu_0^{(c)}}{2}\log|\nu_0^{(c)}| - \Big[\frac{\nu_0^{(c)}D}{2}\log 2 + \frac{D(D-1)}{4}\log\pi + \sum_{d=1}^{D}\log\Gamma(\frac{\nu_0^{(c)}+1-d}{2})\Big] \\
&\quad + \frac{\nu_0^{(c)}-D-1}{2}\log\tilde{\Lambda}_k^{(c),i} - \frac{\nu_0^{(c)}\nu_k^{(c),i}}{2(\nu_l^{(p),j}-D-1)}\,\mathrm{Tr}((W_l^{(p),j})^{-1}W_k^{(c),i}).
\end{aligned}
$$

4. $\mathbb{E}\log\Gamma(\epsilon_0^{(c)}A^{(p),j})$ is similar to $\mathbb{E}\log\Gamma(\alpha_0^{(c)}\pi_l^{(p),j})$.

With the above 1-4, we solve $\hat{\phi}^{i,j}$ via an optimization problem with constraints, i.e.,

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

C VARIATIONAL DISTRIBUTIONS FOR PARENT MODELS

In this section, we derive the parameters for variational distributions for parent model, and we consider each parameter in the following.

C.1 INITIAL PROBABILITY

For initial probability $\pi^{(p),j}$, the terms involved its parameter $\alpha^{(p),j}$ in variational distribution is

$$
\begin{aligned}
\mathcal{L}(\alpha^{(p),j}) &\propto \mathbb{E}\log p(\pi^{(c),i}|\alpha_0^{(c)},\phi^{i,j},\pi^{(p),j}) + \mathbb{E}\log p(\pi^{(p),j}) - \mathbb{E}\log q(\pi^{(p),j}) \\
&\geq \log\Gamma(\alpha_0^{(c)})\sum_l N_l^j\hat{\pi}_l^{(p),j} - \sum_l N_l^j\mathbb{E}\big[\log\Gamma(\pi_l^{(p),j}\alpha_0^{(c)})\big] + \alpha_0^{(c)}\sum_l \pi_l^j\hat{\pi}_l^{(p),j} + \sum_l(\alpha_0^{(p)}-\alpha_l^{(p),j})\log\tilde{\pi}_l^{(p),j} \\
&\quad - \log\Gamma(\hat{\alpha}^{(p),j}) + \sum_l\log\Gamma(\alpha_l^{(p),j}),
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\pi}_l^{(p),j} &= \mathbb{E}[\pi_l^{(p),j}] = \frac{\alpha_l^{(p),j}}{\hat{\alpha}_l^{(p),j}}, \quad \hat{\alpha}_l^{(p),j} = \sum_l \alpha_l^{(p),j}, \\
\log(\pi_l^{(p),j}) &= \psi(\alpha_l^{(p),j}) - \psi(\hat{\alpha}_l^{(p),j}), \\
N_l^j &= \sum_i\sum_k \hat{z}_{ij}\hat{\phi}^{(i,j)}(l|k), \\
\pi_l^j &= \sum_i\sum_k \hat{z}_{ij}\hat{\phi}_{k,l}^{i,j}\log\tilde{\pi}_k^{(c),i},
\end{aligned}
$$

and combining the inequality in Equ (8), we obtain $\mathcal{L}(\alpha^{(p),j})$. Solving the optimization problem

$$
\max\ \mathcal{L}(\alpha^{(p),j}), \quad s.t.\ \alpha_l^{(p),j} \geq 1,
$$

to solve $\alpha^{(p),j}$.

C.2 TRANSITION MATRIX

For transition matrix $A^{(p),j}$, the terms involved its parameter $\epsilon^{(p),j}$ in variational distribution is

$$
\mathcal{L}(\epsilon^{(p),j}) \propto \sum_i \hat{z}_{ij}\mathbb{E}\log P(A^{(c),i}|\epsilon_0^{(c)},\phi^{i,j},A^{(p),j}) + \mathbb{E}\log p(A^{(p),j}) - \mathbb{E}\log q(A^{(p),j}) \geq \sum_l \mathcal{L}(\epsilon_l^{(p),j}),
$$

and

$$
\mathcal{L}(\epsilon_l^{(p),j}) = \sum_{l'}\Big[N_{l,l'}^{(j)}\hat{a}_{l,l'}^{(p),j}\log\Gamma(\epsilon_0^{(c)}) - N_{l,l'}^{(j)}\mathbb{E}\log\Gamma(\epsilon_0^{(c)}a_{l,l'}^{(p),j}) + \epsilon_0^{(c)}\beta_{l,l'}^{(j)}\hat{a}_{l,l'}^{(p),j}\Big] - \log\Gamma(\hat{\epsilon}_l^{(p),j})
$$

$$+ \sum_{l'} \log \Gamma(\epsilon_{l,l'}^{(p),j}) - \sum_{l'} (\epsilon_{l,l'}^{(p),j} - \epsilon_0^{(c)}) \log \tilde{a}_{l,l'}^{(p),j}, \tag{9}$$

where

$$\begin{aligned}
\hat{a}_{l,l'}^{(p),j} &= \mathbb{E}[a_{l,l'}^{(p),j}] = \frac{\epsilon_{l,l'}^{(p),j}}{\hat{\epsilon}_l^{(p),j}}, \quad \hat{\epsilon}_l^{(p),j} = \sum_{l'} \epsilon_{l,l'}^{(p),j}, \\
\log \tilde{a}_{l,l'}^{(p),j} &= \psi(\epsilon_{l,l'}^{(p),j}) - \psi(\hat{\epsilon}_l^{(p),j}), \\
N_{l,l'}^{(j)} &= \sum_i \hat{z}_{ij} \sum_k \sum_{k'} \hat{\phi}_{k,l}^{i,j} \hat{\phi}_{k',l'}^{i,j}, \\
\beta_{l,l'}^{(j)} &= \sum_i \hat{z}_{ij} \sum_k \sum_{k'} \hat{\phi}_{k,l}^{i,j} \hat{\phi}_{k',l'}^{i,j} \log \tilde{a}_{k,k'}^{(c),i},
\end{aligned}$$

and combining the inequality in Equ (8), we obtain $\mathcal{L}(\epsilon_l^{(p),j})$. Solving the optimization problem

$$\max \ \mathcal{L}(\epsilon_l^{(p),j}), \qquad \text{s.t. } \epsilon_l^{(p),j} > 0,$$

to compute $\epsilon_l^{(p),j}$.

C.3 GAUSSIAN EMISSION DENSITY

The terms in objective function (4) involve $m_l^{(p),j}$, $\beta_l^{(p),j}$, $W_l^{(p),j}$, and $\nu_l^{(p),j}$ are

$$\begin{aligned}
&\mathcal{L}(m_l^{(p),j}, \beta_l^{(p),j}, W_l^{(p),j}, \nu_l^{(p),j}) \\
&= \sum_i \hat{z}_{ij} \mathbb{E}_{q(M_i^{(c)})} \mathbb{E}_{q(M_j^{(p)})} \big[\log p(M_i^{(c)} | M_j^{(p)}) \big] + \mathbb{E}_{q(M_j^{(p)})} \big[\log p(M_j^{(p)}) \big] - \mathbb{E}_{q(M_j^{(p)})} \big[\log q(M_j^{(p)}) \big] \\
&= \mathcal{L}(m_l^{(p),j}) + \mathcal{L}(\beta_l^{(p),j}) + \mathcal{L}(W_l^{(p),j}) + \mathcal{L}(\nu_l^{(p),j})
\end{aligned}$$

1. $\mathcal{L}(m_l^{(p),j})$

$$\begin{aligned}
\mathcal{L}(m_l^{(p),j}) = \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{k,l}^{i,j} \big[&- \frac{\beta_0^{(c)}}{2} \nu_k^{(c),i} (m_k^{(c),i} - m_l^{(p),j})^T \cdot W_k^{(c),i} (m_k^{(c),i} - m_l^{(p),j}) \big] \\
&- \frac{\beta_0^{(p)}}{2} \nu_l^{(p),j} (m_l^{(p),j} - m_0^{(r)})^T W_l^{(p),j} (m_l^{(p),j} - m_0^{(r)}).
\end{aligned}$$

Take the derivation of $\mathcal{L}(m_l^{(p),j})$ w.r.t. $m_l^{(p),j}$ and set to zero, we can obtain a closed-form solution of $m_l^{(p),j}$.

$$\begin{aligned}
m_l^{(p),j} = &\Big[\beta_0^{(c)} \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{kl}^{i,j} \nu_k^{(c),i} W_k^{(c),i} + \beta_0^{(p)} \nu_l^{(p),j} W_l^{(p),j} \Big]^{-1} \\
&\cdot \Big(\beta_0^{(c)} \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{kl}^{i,j} \nu_k^{(c),i} W_k^{(c),i} m_k^{(c),i} + \beta_0^{(p)} \nu_l^{(p),j} W_l^{(p),j} m_0^{(p)} \Big).
\end{aligned}$$

2. $\mathcal{L}(\beta_l^{(p),j})$

$$\mathcal{L}(\beta_l^{(p),j}) = \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{k,l}^{i,j} \Big[- \frac{\beta_0^{(c)} \nu_k^{(c),i}}{2\beta_l^{(p),j} (\nu_l^{(p),j} - D - 1)} \cdot \mathrm{Tr}((W_l^{(p),j})^{-1} W_k^{(c),i}) \Big] - \frac{\beta_0^{(p)} D}{2\beta_l^{(p),j}} - \frac{D}{2} \log \beta_l^{(p),j}.$$

Take the derivation of $\mathcal{L}(\beta_l^{(p),j})$ w.r.t. $\beta_l^{(p),j}$ and set to zero, we can obtain a closed-form solution of $\beta_l^{(p),j}$.

$$\beta_l^{(p),j} = \beta_0^{(p)} + \beta_0^{(c)} \sum_i \hat{z}_{ij} \sum_k \hat{\phi}_{kl}^{i,j} \frac{1}{d} \,\mathrm{Tr}\Big(\tfrac{(W_l^{(p),j})^{-1} \nu_k^{(c),i} W_k^{(c),i}}{\nu_l^{(p),j} - d - 1} \Big). \tag{10}$$

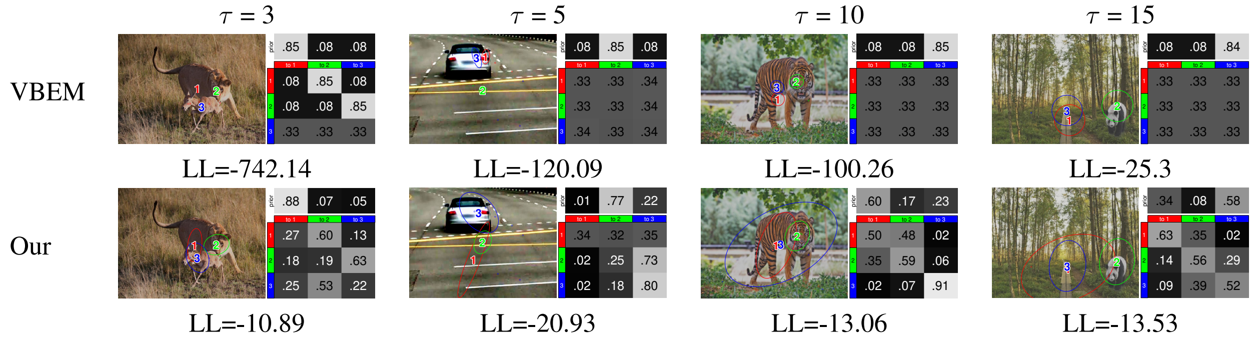

Figure 1: Illustration for individual HMMs learned by VBEM and our method, and LL means the log-likelihood on test data.

3. $\mathcal{L}(W_l^{(p),j})$

$$\begin{aligned}\mathcal{L}(W_l^{(p),j}) = &-\frac{1}{2}\operatorname{Tr}\left(\frac{(W_l^{(p),j})^{-1}}{(\nu_l^{(p),j}-D-1)}\sum_i z_{ij}\sum_k \hat{\phi}_{k,l}^{i,j}\cdot\left[\frac{\beta_0^{(c)}}{\beta_l^{(p),j}}+\nu_0^{(c)}\right]\nu_k^{(c),i}W_k^{(c),i}\right)+\frac{\nu_0^{(p)}-N_l^j\nu_0^{(c)}}{2}\log|W_l^{(p),j}|\\ &-\frac{1}{2}\nu_l^{(p),j}\operatorname{Tr}\left(W_l^{(p),j}\left[\beta_0^{(p)}(m_l^{(p),j}-m_0^{(p)})^T(m_l^{(p),j}-m_0^{(p)})+(W_0^{(p)})^{-1}\right]\right).\end{aligned}$$

Take the derivation of $\mathcal{L}(W_l^{(p),j})$ w.r.t. $W_l^{(p),j}$ and set to zero, we get an Algebraic Riccati Equation,

$$-2cW_l^{(p),j}+W_l^{(p),j}RW_l^{(p),j}-Q=0, \tag{11}$$

where

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

(11) with a unique positive definite solution. Thus, solve (11), we get $W_l^{(p),j}$. In this paper, we use the Matlab ARE solver (icare) to find the solution of (11).

4. $\mathcal{L}(\nu_l^{(p),j})$

$$\begin{aligned}\mathcal{L}(\nu_l^{(p),j}) = &\frac{1}{(\nu_l^{(p),j}-D-1)}\operatorname{Tr}\left((W_l^{(p),j})^{-1}\left[-\left(\frac{\beta_0^{(c)}}{2\beta_l^{(p),j}}+\frac{\nu_0^{(c)}}{2}\right)\sum_i z_{ij}\sum_k \hat{\phi}^{(i,j)}(l|k)\nu_k^{(c),i}W_k^{(c),i}\right]\right)\\ &-\left[\sum_{d=1}^{D}\psi\Big(\frac{\nu_l^{(p),j}+1-d}{2}\Big)+D\log 2+\log(|W_l^{(p),j}|)\right]\left[\frac{N_l^{(j)}\nu_0^{(c)}+\nu_l^{(p),j}-\nu_0^{(p)}}{2}\right]\\ &-\nu_l^{(p),j}\Bigg[\frac{1}{2}\operatorname{Tr}\left(W_l^{(p),j}\left[\beta_0^{(p)}(m_l^{(p),j}-m_0^{(p)})^T(m_l^{(p),j}-m_0^{(p)})+(W_0^{(p)})^{-1}\right]\right)\\ &-\frac{D}{2}\log 2-\frac{1}{2}\log|W_l^{(p),j}|-\frac{D}{2}\Bigg]+\sum_{d=1}^{D}\Gamma\Big(\frac{\nu_l^{(p),j}+1-d}{2}\Big).\end{aligned}$$

Solving the optimization problem

$$\max\ \mathcal{L}(\nu_l^{(p),j}),\quad s.t.\ \nu_l^{(p),j}\geq d+1,$$

to solve $\nu_l^{(p),j}$.

In this paper, we use the Global Optimization Toolbox (GlobalSearch combine fmincon solver) in Matlab to solver the optimization problems.

D ALGEBRAIC RICCATI EQUATION

The Algebraic Riccati Equation (ARE) has been used to study the matrix generalized inverse Gaussian distribution [Fazayeli and Banerjee, 2016], which is a distribution over symmetric positive semi-definite matrices. We consider the ARE when learning group model and here give a brief introduction. An ARE with respect to symmetric positive matrix $P \in \mathbb{R}^{d \times d}$ is

$$A^T P + PA + PRP + Q = 0, \tag{12}$$

where $A \in \mathbb{R}^{d \times d}$, and $Q, R \in \mathbb{S}_+^d$, where $\mathbb{S}_+^d$ denotes the space of symmetric $(d \times d)$ positive semi-definite matrix. The ARE (12) has a unique positive definite solution if and only if the associated Hamiltonian matrix $H = \begin{bmatrix} A & R \\ -Q & -A^T \end{bmatrix}$ has no imaginary eigenvalues [Boyd and Barratt, 1991].

References

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

Stephen P Boyd and Craig H Barratt. *Linear controller design: Limits of performance*. Citeseer, 1991.

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

Do-kyum Kim, Geoffrey Voelker, and Lawrence Saul. A variational approximation for topic modeling of hierarchical corpora. In *ICML*, pages 55–63. PMLR, 2013.