Derivation of the Variational HEM Algorithm for Hidden Markov Mixture Models

Emanuele Coviello
ECE Dept., UC San Diego
ecoviell@ucsd.edu

Antoni B. Chan
CS Dept., CityU of Hong Kong
abchan@cityu.edu.hk

Gert R.G. Lanckriet
ECE Dept., UC San Diego
gert@ece.ucsd.edu

Abstract

This is supplemental material for the NIPS paper "The variational hierarchical EM algorithm for clustering hidden Markov models" [1]. It contains the derivation of the VHEM-H3M algorithm, and the associated E-step and M-step computations.

1 Derivation of the VHEM algorithm for HMMs

We derive the VHEM algorithm for hidden Markov models. We present the problem formulation in Section 1.1, and derive the variational lower bound in Sections 1.2, the variational E-step in Section 1.3 and the M-step in Section 1.4

For the reader's convenience, we summarize the notation used in [1] in Table 1.

1.1 Formulation

Let $\mathcal{M}^{(b)}$ be a base hidden Markov mixture model with $K^{(b)}$ components. The goal of the VHEM algorithm is to find a reduced hidden Markov mixture model $\mathcal{M}^{(r)}$ with $K^{(r)} < K^{(b)}$ (i.e., fewer) components that represents $\mathcal{M}^{(b)}$ well. The likelihood of a random sequence $y_{1:\tau} \sim \mathcal{M}^{(b)}$ is given by

$$p(y_{1:\tau}|\mathcal{M}^{(b)}) = \sum_{i=1}^{K^{(b)}} \omega_i^{(b)} p(y_{1:\tau}|z^{(b)} = i, \mathcal{M}^{(b)}), \tag{1}$$

where $z^{(b)} \sim \text{multinomial}(\omega_1^{(b)}, \cdots \omega_{K^{(b)}}^{(b)})$ is the hidden variable that indexes the mixture components. $p(y_{1:\tau}|z = i, \mathcal{M}^{(b)})$ is the likelihood of $y_{1:\tau}$ under the ith mixture component, and $\omega_i^{(b)}$ is the mixture weight for the ith component. The likelihood of the random sequence $y_{1:\tau} \sim \mathcal{M}^{(r)}$ is

$$p(y_{1:\tau}|\mathcal{M}^{(r)}) = \sum_{j=1}^{K^{(r)}} \omega_j^{(r)} p(y_{1:\tau}|z^{(r)} = j, \mathcal{M}^{(r)}), \tag{2}$$

where $z^{(r)} \sim \text{multinomial}(\omega_1^{(r)}, \cdots, \omega_{K^{(r)}}^{(r)})$ is the hidden variable for indexing components in $\mathcal{M}^{(r)}$.

At a high level, the VHEM-H3M algorithm estimates the reduced H3M model $\mathcal{M}^{(r)}$ in (2) from *virtual* sequences distributed according to the base H3M model $\mathcal{M}^{(b)}$ in (1). From this estimation procedure, the VHEM algorithm provides:

1. a soft clustering of the original $K^{(b)}$ components into $K^{(r)}$ groups, where the cluster membership is encoded in assignment variables that represents the *responsibility* of each reduced mixture component over each base mixture component, i.e., $\hat{z}_{i,j} = P(z^{(r)} = j|z^{(b)} = i)$, for $i = 1, \ldots, K^{(b)}$ and $j = 1, \ldots, K^{(r)}$;

Table 1: Notation used in the derivation of the VHEM-H3M algorithm.

variables	*base model (b)*	*reduced model (r)*
index for HMM components	i	j
HMM states	β	ρ
HMM state sequence	$\beta_{1:\tau} = \{\beta_1, \cdots, \beta_\tau\}$	$\rho_{1:\tau} = \{\rho_1, \cdots, \rho_\tau\}$
index for component of GMM	m	ℓ
models		
H3M	$\mathcal{M}^{(b)}$	$\mathcal{M}^{(r)}$
HMM component	$\mathcal{M}_i^{(b)}$	$\mathcal{M}_j^{(r)}$
GMM emission	$\mathcal{M}_{i,\beta}^{(b)}$	$\mathcal{M}_{j,\rho}^{(r)}$
component of GMM	$\mathcal{M}_{i,\beta,m}^{(b)}$	$\mathcal{M}_{j,\rho,\ell}^{(r)}$
parameters		
H3M mixture weights	$\omega^{(b)} = \{\omega_i^{(b)}\}$	$\omega^{(r)} = \{\omega_j^{(r)}\}$
HMM initial state	$\pi^{(b),i} = \{\pi_\beta^{(b),i}\}$	$\pi^{(r),j} = \{\pi_\rho^{(r),j}\}$
HMM state transition matrix	$A^{(b),i} = [a_{\beta,\beta'}^{(b),i}]$	$A^{(r),j} = [a_{\rho,\rho'}^{(r),j}]$
GMM emission	$\{c_{\beta,m}^{(b),i}, \mu_{\beta,m}^{(b),i}, \Sigma_{\beta,m}^{(b),i}\}_{m=1}^M$	$\{c_{\rho,\ell}^{(r),j}, \mu_{\rho,\ell}^{(r),j}, \Sigma_{\rho,\ell}^{(r),j}\}_{\ell=1}^M$
probability distributions	*notation*	*short-hand*
HMM state sequence (b)	$p(x_{1:\tau} = \beta_{1:\tau} \vert z^{(b)} = i, \mathcal{M}^{(b)})$	$p(\beta_{1:\tau} \vert \mathcal{M}_i^{(b)}) = \pi_{\beta_{1:\tau}}^{(b),i}$
HMM state sequence (r)	$p(x_{1:\tau} = \rho_{1:\tau} \vert z^{(r)} = j, \mathcal{M}^{(r)})$	$p(\rho_{1:\tau} \vert \mathcal{M}_j^{(r)}) = \pi_{\rho_{1:\tau}}^{(r),j}$
HMM observation likelihood (r)	$p(y_{1:\tau} \vert z^{(r)} = j, \mathcal{M}^{(r)})$	$p(y_{1:\tau} \vert \mathcal{M}_j^{(r)})$
GMM emission likelihood (r)	$p(y_t \vert x_t = \rho, \mathcal{M}_j^{(r)})$	$p(y_t \vert \mathcal{M}_{j,\rho}^{(r)})$
Gaussian component likelihood (r)	$p(y_t \vert \zeta_t = \ell, x_t = \rho, \mathcal{M}_j^{(r)})$	$p(y_t \vert \mathcal{M}_{j,\rho,\ell}^{(r)})$
expectations		
HMM observation sequence	$\mathrm{E}_{y_{1:\tau} \vert z^{(b)}=i, \mathcal{M}^{(b)}}[\cdot]$	$\mathrm{E}_{\mathcal{M}_i^{(b)}}[\cdot]$
GMM emission	$\mathrm{E}_{y_t \vert x_t=\beta, \mathcal{M}_i^{(b)}}[\cdot]$	$\mathrm{E}_{\mathcal{M}_{i,\beta}^{(b)}}[\cdot]$
Gaussian component	$\mathrm{E}_{y_t \vert \zeta_t=m, x_t=\beta, \mathcal{M}_i^{(b)}}[\cdot]$	$\mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}[\cdot]$
expected log-likelihood	*lower bound*	*variational distribution*
$\mathrm{E}_{\mathcal{M}_i^{(b)}}[\log p(Y_i \vert \mathcal{M}^{(r)})]$	$\mathcal{L}_{H3M}^i$	$q_i(z_i = j) = z_{ij}$
$\mathrm{E}_{\mathcal{M}_i^{(b)}}[\log p(y_{1:\tau} \vert \mathcal{M}_j^{(r)})]$	$\mathcal{L}_{HMM}^{i,j}$	$q^{i,j}(\rho_{1:\tau} \vert \beta_{1:\tau}) = \phi_{\rho_{1:\tau} \vert \beta_{1:\tau}}^{i,j}$ $= \phi_1^{i,j}(\rho_1 \vert \beta_1) \prod_{t=2}^{\tau} \phi_t^{i,j}(\rho_t \vert \rho_{t-1}, \beta_t)$
$\mathrm{E}_{\mathcal{M}_{i,\beta}^{(b)}}[\log p(y \vert \mathcal{M}_{j,\rho}^{(r)})]$	$\mathcal{L}_{GMM}^{(i,\beta),(j,\rho)}$	$q_{\beta,\rho}^{i,j}(\zeta = \ell \vert m) = \eta_{\ell \vert m}^{(i,\beta),(j,\rho)}$

2. novel cluster centers represented by the individual mixture components of (2), i.e., $p(y_{1:\tau} \vert z^{(r)} = j, \mathcal{M}^{(r)})$ for $j = 1, \ldots, K^{(r)}$.

Finally, because we take the expectation over the virtual samples, the estimation is carried out in an efficient manner that requires only knowledge of the parameters of the base model without the need of generating actual virtual samples.

Notation. We will always use i and j to index the components of the base model, $\mathcal{M}^{(b)}$, and the reduced model, $\mathcal{M}^{(r)}$, respectively. To reduce clutter, we will also use the short-hand notation $\mathcal{M}_i^{(b)}$ and $\mathcal{M}_j^{(r)}$ to denote the ith component of $\mathcal{M}^{(b)}$ and the jth component of $\mathcal{M}^{(r)}$. Hidden states of the HMMs are denoted with β and ρ for the base model $\mathcal{M}_i^{(b)}$ and reduced model $\mathcal{M}_j^{(r)}$. The GMM emission models for each hidden state are denoted as $\mathcal{M}_{i,\beta}^{(b)}$ and $\mathcal{M}_{j,\rho}^{(r)}$. We will use m and ℓ for indexing the Gaussian mixture components of the emission models of the base and reduced models, respectively, The individual Gaussian components are denoted as $\mathcal{M}_{i,\beta,m}^{(b)}$ for the base model and $\mathcal{M}_{j,\rho,\ell}^{(r)}$ for the reduced model. Finally, we denote the parameters of i-th HMM component of the base mixture model as $\mathcal{M}_i^{(b)} = \{\pi^{(b),i}, A^{(b),i}, \{\{c_{\beta,m}^{(b),i}, \mu_{\beta,m}^{(b),i}, \Sigma_{\beta,m}^{(b),i}\}_{m=1}^M\}_{\beta=1}^S\}$, and for the j-th HMM in the reduced mixture as $\mathcal{M}_j^{(r)} = \{\pi^{(r),j}, A^{(r),j}, \{\{c_{\rho,\ell}^{(r),j}, \mu_{\rho,\ell}^{(r),j}, \Sigma_{\rho,\ell}^{(r),j}\}_{\ell=1}^M\}_{\rho=1}^S\}$.

When appearing in a probability distribution, the short-hand model notation (e.g., $\mathcal{M}_i^{(b)}$) always implies *conditioning* on the model being active. For example, we will use $p(y_{1:\tau}|\mathcal{M}_i^{(b)})$ as short-hand for $p(y_{1:\tau}|z^{(b)}=i,\mathcal{M}^{(b)})$, or $p(y_t|\mathcal{M}_{i,\beta}^{(b)})$ as short-hand for $p(y_t|x_t=\beta,z^{(b)}=i,\mathcal{M}^{(b)})$. Furthermore, we will use $\pi_{\beta_{1:\tau}}^{(b),i}$ as shorthand for the probability of the state sequence $\beta_{1:\tau}$ in the base HMM component $\mathcal{M}_i^{(b)}$, i.e., $p(\beta_{1:\tau}|\mathcal{M}_i^{(b)})$, and likewise for the reduced HMM component.

Expectations will also use the short-hand model notation to imply conditioning on the model. In addition, expectations are assumed to be taken with respect to the output variable ($y_{1:\tau}$ or y_t), unless otherwise specified. For example, we will use $\mathrm{E}_{\mathcal{M}_i^{(b)}}[\cdot]$ as short-hand for $\mathrm{E}_{y_{1:\tau}|\mathcal{M}^{(b)},z^{(b)}=i}[\cdot]$.

Table 1 summarizes the notation used in the derivation, including the variable names, model parameters, and short-hand notations for probability distributions and expectations. The bottom of Table 1 also summarizes the variational lower bound and variational distributions, which will be introduced subsequently.

1.2 Variational HEM algorithm

To learn the reduced model in (2), we consider a set of N virtual samples distributed accordingly to the base model $\mathcal{M}^{(b)}$ in (1), such that $N_i = N\omega_i^{(b)}$ samples are drawn from the ith component. We denote the set of N_i virtual samples for the ith component as $Y_i = \{y_{1:\tau}^{(i,m)}\}_{m=1}^{N_i}$, where $y_{1:\tau}^{(i,m)} \sim \mathcal{M}_i^{(b)}$, and the entire set of N samples as $Y = \{Y_i\}_{i=1}^{K^{(b)}}$. Note that, in this formulation, we are not considering virtual samples $\{x_{1:\tau}^{(i,m)}, y_{1:\tau}^{(i,m)}\}$ for each base component, according to its joint distribution $p(x_{1:\tau}, y_{1:\tau}|\mathcal{M}_i^{(b)})$. The reason is that the hidden state space of each base mixture component $\mathcal{M}_i^{(b)}$ may have a different representation (e.g., the numbering of the hidden states may be permuted between the components). This basis mismatch will cause problems when the parameters of $\mathcal{M}_j^{(r)}$ are computed from virtual samples of the hidden states of $\{\mathcal{M}_i^{(b)}\}_{i=1}^{K^{(b)}}$. Instead, we treat $X_i = \{x_{1:\tau}^{(i,m)}\}_{m=1}^{N_i}$ as "missing" information, and estimate them in the E-step. The log-likelihood of the virtual samples is

$$\log p(Y|\mathcal{M}^{(r)}) = \sum_{i=1}^{K^{(b)}} \log p(Y_i|\mathcal{M}^{(r)}) \tag{3}$$

where, in order to obtain a consistent clustering, we assume the entirety of samples Y_i is assigned to the same component of the reduced model [2].

The original formulation of HEM [2] maximizes (3) with respect to $\mathcal{M}^{(r)}$, and uses the law of large numbers to turn the virtual samples Y_i into an expectation over the base model components $\mathcal{M}_i^{(b)}$. In this paper, we will start with a different objective function to derive the VHEM algorithm. To estimate $\mathcal{M}^{(r)}$, we will maximize the average log-likelihood of all possible samples, weighted by their likelihood of being generated by $\mathcal{M}_i^{(b)}$, i.e., the *expected* log-likelihood of the virtual samples,

$$\mathcal{J}(\mathcal{M}^{(r)}) = \mathrm{E}_{\mathcal{M}^{(b)}}\left[\log p(Y|\mathcal{M}^{(r)})\right] = \sum_{i=1}^{K^{(b)}} \mathrm{E}_{\mathcal{M}_i^{(b)}}\left[\log p(Y_i|\mathcal{M}^{(r)})\right], \tag{4}$$

where the expectation is over the base model components $\mathcal{M}_i^{(b)}$. Maximizing (4) will eventually lead to the same estimate as maximizing (3), but allows us to strictly preserve the lower bound, which would otherwise be ruined when using the law-of-large numbers with (3).

A general approach to deal with maximum likelihood estimation in the presence of hidden variables (which is the case for H3Ms) is the EM algorithm [3]. Although in the traditional formulation the EM algorithm is presented as an alternation between an expectation step (E step) and a maximization step (M step), in this work we take a variational perspective [4, 5, 6], which views each step as a maximization step. The variational E-step first obtains a family of lower bounds to the log-likelihood (i.e., to equation 4), indexed by variational parameters, and then optimizes over the variational parameters to find the tightest bound. The corresponding M-step then maximizes the lower bound

(with the variational parameters fixed) with respect to the model parameters. One advantage of the variational formulation is that it readily allows for useful extensions to the EM algorithm, such as replacing a difficult inference in the E-step with a variational approximation. In practice, this is achieved by restricting the maximization in the variational E-step to a smaller domain for which the lower bound is tractable.

1.2.1 Lower bound to an expected log-likelihood

Before proceeding with the derivation of VHEM for H3Ms, we first need to derive a lower-bound to an expected log-likelihood term (e.g., (4)). In all generality, let $\{O, H\}$ be the observation and hidden variables of a probabilistic model, respectively, where $p(H)$ is the distribution of the hidden variables, $p(O|H)$ is the conditional likelihood of the observations, and $p(O) = \sum_H p(O|H)p(H)$ is the observation likelihood. We can define a *variational lower bound* to the observation log-likelihood [7, 8]:

$$\log p(O) \geq \log p(O) - D(q(H)||p(H|O)) \tag{5}$$

$$= \sum_H q(H) \log \frac{p(H)p(O|H)}{q(H)} \tag{6}$$

where $p(H|O)$ is the posterior distribution of H given observation O, and $D(p\|q) = \int p(y) \log \frac{p(y)}{q(y)} dy$ is the Kullback-Leibler (KL) divergence between two distributions, p and q. We introduce a variational distribution $q(H)$, which approximates the posterior distribution, where $\sum_H q(H) = 1$ and $q(H) \geq 0$. When the variational distribution equals the true posterior, $q(H) = P(H|O)$, then the KL divergence is zero, and hence the lower-bound reaches $\log p(O)$. When the true posterior is not possible to calculate, then typically q is restricted to some set of approximate posterior distributions $\mathcal{Q}$ that are tractable, and the best lower-bound is obtained by maximizing over $q \in \mathcal{Q}$,

$$\log p(O) \geq \max_{q \in \mathcal{Q}} \sum_H q(H) \log \frac{p(H)p(O|H)}{q(H)} \tag{7}$$

Using the lower bound in (7), we can now derive a lower bound to an expected log-likelihood expression. Let $\mathrm{E}_b[\cdot]$ be the expectation of O with respect to a distribution $p_b(O)$. Since $p_b(O)$ is non-negative, taking the expectation on both sides of (7) yields,

$$\mathrm{E}_b\left[\log p(O)\right] \geq \mathrm{E}_b\left[\max_{q \in \mathcal{Q}} \sum_H q(H) \log \frac{p(H)p(O|H)}{q(H)}\right] \tag{8}$$

$$\geq \max_{q \in \mathcal{Q}} \mathrm{E}_b\left[\sum_H q(H) \log \frac{p(H)p(O|H)}{q(H)}\right] \tag{9}$$

$$= \max_{q \in \mathcal{Q}} \sum_H q(H) \left\{\log \frac{p(H)}{q(H)} + \mathrm{E}_b\left[\log p(O|H)\right]\right\}, \tag{10}$$

where (9) follows from Jensen's inequality (i.e., $f(\mathrm{E}[x]) \leq \mathrm{E}[f(x)]$ when f is convex), and the convexity of the $\max$ function.

1.2.2 Variational lower bound

We now derive the lower bound of the expected log-likelihood cost function in (4). The derivation will proceed by successively applying the lower bound from (10) on each arising expected log-likelihood term, which results in a set of nested lower bounds. We first define the following three lower bounds:

$$\mathrm{E}_{\mathcal{M}_i^{(b)}}[\log p(Y_i|\mathcal{M}^{(r)})] \geq \mathcal{L}^i_{H3M}, \tag{11}$$

$$\mathrm{E}_{\mathcal{M}_i^{(b)}}[\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})] \geq \mathcal{L}^{i,j}_{HMM}, \tag{12}$$

$$\mathrm{E}_{\mathcal{M}_{i,\beta}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho}^{(r)})] \geq \mathcal{L}^{(i,\beta),(j,\rho)}_{GMM}. \tag{13}$$

The first lower bound, $\mathcal{L}^i_{H3M}$, is on the expected log-likelihood of an H3M $\mathcal{M}^{(r)}$ with respect to an HMM $\mathcal{M}^{(b)}_i$. The second lower bound, $\mathcal{L}^{i,j}_{HMM}$, is on the expected log-likelihood of an HMM $\mathcal{M}^{(r)}_j$, marginalized over observation sequences from a *different* HMM $\mathcal{M}^{(b)}_i$. Although the data log-likelihood $\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)$ can be computed exactly using the forward algorithm [9], calculating its expectation is not analytically tractable since $\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)$ is essentially a mixture model[1]. The third lower bound, $\mathcal{L}^{(i,\beta),(j,\rho)}_{GMM}$, is on the expected log-likelihood of a GMM emission density $\mathcal{M}^{(r)}_{j,\rho}$ with respect to another GMM $\mathcal{M}^{(b)}_{i,\beta}$. This lower bound does not depend on time, as we have assumed that the emission densities are time-invariant.

Looking at an individual term in (4), $p(Y_i|\mathcal{M}^{(r)})$ is the likelihood under a mixture of HMMs, as in (2), where the observation variable is Y_i and the hidden variable is z_i (the assignment of Y_i to a component $\mathcal{M}^{(r)}_j$). Hence, introducing the variational distribution $q_i(z_i)$ and applying (10), we have

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{M}^{(b)}_i}\left[\log p(Y_i|\mathcal{M}^{(r)})\right] \\
&\geq \max_{q_i} \sum_j q_i(z_i=j)\left\{\log \frac{p(z_i=j|\mathcal{M}^{(r)})}{q_i(z_i=j)} + \mathrm{E}_{\mathcal{M}^{(b)}_i}[\log p(Y_i|\mathcal{M}^{(r)}_j)]\right\} && (14)\\
&= \max_{q_i} \sum_j q_i(z_i=j)\left\{\log \frac{p(z_i=j|\mathcal{M}^{(r)})}{q_i(z_i=j)} + \mathrm{E}_{\mathcal{M}^{(b)}_i}[\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)^{N_i}]\right\} && (15)\\
&= \max_{q_i} \sum_j q_i(z_i=j)\left\{\log \frac{p(z_i=j|\mathcal{M}^{(r)})}{q_i(z_i=j)} + N_i\mathrm{E}_{\mathcal{M}^{(b)}_i}[\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)]\right\}, && (16)
\end{aligned}
$$

where in (15) we use the fact that Y_i is an i.i.d. set of N_i samples. In (16), $\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)$ is the observation log-likelihood of an HMM, which is a mixture distribution, and hence its expectation cannot be calculated directly. Instead we use ithe lower bound $\mathcal{L}^{i,j}_{HMM}$ defined in (12), yielding

$$
\mathrm{E}_{\mathcal{M}^{(b)}_i}\left[\log p(Y_i|\mathcal{M}^{(r)})\right] \geq \max_{q_i} \sum_j q_i(z_i=j)\left\{\log \frac{p(z_i=j|\mathcal{M}^{(r)})}{q_i(z_i=j)} + N_i\mathcal{L}^{i,j}_{HMM}\right\} \triangleq \mathcal{L}^i_{H3M}. \tag{17}
$$

Next, we calculate the lower bound $\mathcal{L}^{i,j}_{HMM}$. Starting with (12), we first rewrite the expectation $\mathrm{E}_{\mathcal{M}^{(b)}_i}$ to explicitly marginalize over the HMM state sequence $\beta_{1:\tau}$ from $\mathcal{M}^{(b)}_i$,

$$
\begin{aligned}
\mathrm{E}_{\mathcal{M}^{(b)}_i}[\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)] &= \mathrm{E}_{\beta_{1:\tau}|\mathcal{M}^{(b)}_i}\left[\mathrm{E}_{y_{1:\tau}|\beta_{1:\tau},\mathcal{M}^{(b)}_i}[\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)]\right] && (18)\\
&= \sum_{\beta_{1:\tau}} \pi^{(b),i}_{\beta_{1:\tau}} \mathrm{E}_{y_{1:\tau}|\beta_{1:\tau},\mathcal{M}^{(b)}_i}[\log p(y_{1:\tau}|\mathcal{M}^{(r)}_j)] && (19)
\end{aligned}
$$

For the HMM likelihood $p(y_{1:\tau}|\mathcal{M}^{(r)}_j)$, the observation variable is $y_{1:\tau}$ and the hidden variable is the state sequence $\rho_{1:\tau}$. We therefore introduce a variational distribution $q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})$ on the state sequence $\rho_{1:\tau}$, which depends on a particular sequence $\beta_{1:\tau}$ from $\mathcal{M}^{(b)}_i$. Applying (10) to (19), we

[1]For an observation sequence of length τ, an HMM with S states can be considered as a mixture model with $O(S^\tau)$ components.

have

$$\begin{aligned}
&\mathrm{E}_{\mathcal{M}_i^{(b)}}[\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})] \\
&\geq \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \max_{q^{i,j}} \sum_{\rho_{1:\tau}} q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \left\{ \log \frac{p(\rho_{1:\tau}|\mathcal{M}_j^{(r)})}{q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})} + \mathrm{E}_{y_{1:\tau}|\beta_{1:\tau},\mathcal{M}_i^{(b)}}[\log p(y_{1:\tau}|\rho_{1:\tau},\mathcal{M}_j^{(r)})] \right\} \qquad (20) \\
&= \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \max_{q^{i,j}} \sum_{\rho_{1:\tau}} q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \left\{ \log \frac{p(\rho_{1:\tau}|\mathcal{M}_j^{(r)})}{q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})} + \sum_t \mathrm{E}_{\mathcal{M}_{i,\beta_t}^{(b)}}[\log p(y_t|\mathcal{M}_{j,\rho_t}^{(r)})] \right\} \qquad (21) \\
&\geq \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \max_{q^{i,j}} \sum_{\rho_{1:\tau}} q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \left\{ \log \frac{p(\rho_{1:\tau}|\mathcal{M}_j^{(r)})}{q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})} + \sum_t \mathcal{L}_{GMM}^{(i,\beta_t),(j,\rho_t)} \right\} \triangleq \mathcal{L}_{HMM}^{i,j} \qquad (22)
\end{aligned}$$

where in (21) we use the conditional independence of the observation sequence given the state sequence, and in (22) we use the lower bound, defined in (13), on each expectation.

Finally, we derive the lower bound $\mathcal{L}_{GMM}^{(i,\beta),(j,\rho)}$ for (13). First we rewrite the expectation with respect to $\mathcal{M}_{i,\beta}^{(b)}$ to explicitly marginalize out the GMM hidden assignment variable ζ,

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}_{i,\beta}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho}^{(r)})] &= \mathrm{E}_{\zeta|\mathcal{M}_{i,\beta}^{(b)}}\left[\mathrm{E}_{\mathcal{M}_{i,\beta,\zeta}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho}^{(r)})]\right] \qquad (23) \\
&= \sum_{m=1}^{M} c_{\beta,m}^{(b),i} \mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}\left[\log p(y|\mathcal{M}_{j,\rho}^{(r)})\right] \qquad (24)
\end{aligned}$$

Note that $p(y|\mathcal{M}_{j,\rho}^{(r)})$ is a GMM emission distribution, and hence the observation variable is y, and the hidden variable is ζ. Therefore, we introduce the variational distribution $q_{\beta,\rho}^{i,j}(\zeta|m)$, which is conditioned on the observation y arising from the m-th component in $\mathcal{M}_{i,\beta}^{(b)}$, and apply (10),

$$\begin{aligned}
&\mathrm{E}_{\mathcal{M}_{i,\beta}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho}^{(r)})] \\
&\geq \sum_{m=1}^{M} c_{\beta,m}^{(b),i} \max_{q_{\beta,\rho}^{i,j}} \sum_{\zeta=1}^{M} q_{\beta,\rho}^{i,j}(\zeta|m) \left\{ \log \frac{p(\zeta|\mathcal{M}_{j,\rho}^{(r)})}{q_{\beta,\rho}^{i,j}(\zeta|m)} + \mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho,\zeta}^{(r)})] \right\} \triangleq \mathcal{L}_{GMM}^{(i,\beta),(j,\rho)}, \qquad (25)
\end{aligned}$$

where $\mathrm{E}_{\mathcal{M}_{i,\rho,m}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho,\ell}^{(r)})]$ is the expected log-likelihood of a Gaussian distribution $\mathcal{M}_{j,\rho,\ell}^{(r)}$ with respect to another Gaussian $\mathcal{M}_{i,\rho,m}^{(b)}$, which has a closed-form solution (see Section 1.3.1).

In summary, we have derived a variational lower bound of the expected log-likelihood of the virtual samples in (4),

$$\mathcal{J}(\mathcal{M}^{(r)}) = \mathrm{E}_{\mathcal{M}^{(b)}}\left[\log p(Y|\mathcal{M}^{(r)})\right] \geq \sum_{i=1}^{K^{(b)}} \mathcal{L}_{H3M}^{i}, \qquad (26)$$

which is composed of three nested lower bounds, corresponding to different model elements (the H3M, the component HMMs, and the emission GMMs),

$$\begin{aligned}
\mathcal{L}_{H3M}^{i} &= \max_{q_i} \sum_j q_i(z_i = j) \left\{ \log \frac{p(z_i = j|\mathcal{M}^{(r)})}{q_i(z_i = j)} + N_i \mathcal{L}_{HMM}^{i,j} \right\}, \qquad (27) \\
\mathcal{L}_{HMM}^{i,j} &= \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \max_{q^{i,j}} \sum_{\rho_{1:\tau}} q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \left\{ \log \frac{p(\rho_{1:\tau}|\mathcal{M}_j^{(r)})}{q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})} + \sum_t \mathcal{L}_{GMM}^{(i,\beta_t),(j,\rho_t)} \right\}, \qquad (28) \\
\mathcal{L}_{GMM}^{(i,\beta),(j,\rho)} &= \sum_{m=1}^{M} c_{\beta,m}^{(b),i} \max_{q_{\beta,\rho}^{i,j}} \sum_{\zeta=1}^{M} q_{\beta,\rho}^{i,j}(\zeta|m) \left\{ \log \frac{p(\zeta|\mathcal{M}_{j,\rho}^{(r)})}{q_{\beta,\rho}^{i,j}(\zeta|m)} + \mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho,\zeta}^{(r)})] \right\}, \qquad (29)
\end{aligned}$$

where $q_i(z_i)$, $q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})$, and $q^{i,j}_{\beta,\rho}(\zeta|m)$ are the corresponding variational distributions. Finally, the variational HEM algorithm for HMMs consists of two alternating steps:

- (variational E-step) given $\mathcal{M}^{(r)}$, calculate the variational distributions $q^{i,j}_{\beta,\rho}(\zeta|m)$, $q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})$, and $q_i(z_i)$ for the lower bounds in (29), (28), and (27);
- (M-step) update the model parameters via $\mathcal{M}^{(r)^*} = \text{argmax}_{\mathcal{M}^{(r)}} \sum_{i=1}^{K^{(b)}} \mathcal{L}^i_{H3M}$.

We next derive the E- and M-steps of the algorithm.

1.3 Variational E-step

The variational E-step consists of finding the variational distributions to maximize the lower bounds in (29), (28), and (27). In particular, given the nesting of the lower bounds, we proceed by first maximizing the GMM lower bound $\mathcal{L}^{(i,\beta),(j,\rho)}_{GMM}$ for each pair of emission GMMs in the base and reduced models. Next, the HMM lower bound $\mathcal{L}^{i,j}_{HMM}$ is maximized for each pair of HMMs in the base and reduced models, followed by maximizing the H3M lower bound $\mathcal{L}^i_{H3M}$ for each base HMM. Finally, a set of summary statistics are calculated, which will be used in the M-step.

1.3.1 Variational distributions

We first consider the forms of the three variational distributions, as well as the optimal parameters to maximize the corresponding lower bounds.

GMM: For the GMM lower bound $\mathcal{L}^{(i,\beta),(j,\rho)}_{GMM}$, we assume each variational distribution has the form

$$q^{i,j}_{\beta,\rho}(\zeta = l|m) = \eta^{(i,\beta),(j,\rho)}_{\ell|m} \tag{30}$$

where $\sum_{\ell=1}^{M} \eta^{(i,\beta),(j,\rho)}_{\ell|m} = 1$, and $\eta^{(i,\beta),(j,\rho)}_{\ell|m} \geq 0$, $\forall \ell$. Intuitively, $\boldsymbol{\eta}^{(i,\beta),(j,\rho)}$ is the responsibility matrix between each pair of Gaussian components in GMMs $\mathcal{M}^{(b)}_{i,\beta}$ and $\mathcal{M}^{(r)}_{j,\rho}$, where $\eta^{(i,\beta),(j,\rho)}_{\ell|m}$ represents the probability that an observation from component m of $\mathcal{M}^{(b)}_{i,\beta}$ corresponds to component ℓ of $\mathcal{M}^{(r)}_{j,\rho}$.

Substituting into (29), we have

$$\mathcal{L}^{(i,\beta),(j,\rho)}_{GMM} = \sum_{m=1}^{M} c^{(b),i}_{\beta,m} \max_{\eta^{(i,\beta),(j,\rho)}_{\ell|m}} \sum_{\ell=1}^{M} \eta^{(i,\beta),(j,\rho)}_{\ell|m} \left\{ \log \frac{c^{(r),j}_{\rho,\ell}}{\eta^{(i,\beta),(j,\rho)}_{\ell|m}} + \mathrm{E}_{\mathcal{M}^{(b)}_{i,\beta,m}}[\log p(y|\mathcal{M}^{(r)}_{j,\rho,\ell})] \right\}. \tag{31}$$

The maximizing variational parameters are obtained using Appendix C.2,

$$\hat{\eta}^{(i,\beta),(j,\rho)}_{\ell|m} = \frac{c^{(r),j}_{\rho,\ell} \exp\left\{ \mathrm{E}_{\mathcal{M}^{(b)}_{i,\beta,m}}[\log p(y|\mathcal{M}^{(r)}_{j,\rho,\ell})] \right\}}{\sum_{\ell'} c^{(r),j}_{\rho,\ell'} \exp\left\{ \mathrm{E}_{\mathcal{M}^{(b)}_{i,\beta,m}}[\log p(y|\mathcal{M}^{(r)}_{j,\rho,\ell'})] \right\}}, \tag{32}$$

where the expected log-likelihood of a Gaussian $\mathcal{M}^{(r)}_{j,\rho,\ell}$ with respect to another Gaussian $\mathcal{M}^{(b)}_{i,\rho,m}$ is computable in closed-form,

$$\begin{aligned} \mathrm{E}_{\mathcal{M}^{(b)}_{i,\beta,m}}[\log p(y|\mathcal{M}^{(r)}_{j,\rho,\ell})] = &-\frac{d}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma^{(r),j}_{\rho,\ell}\right| - \frac{1}{2}\mathrm{tr}\left((\Sigma^{(r),j}_{\rho,\ell})^{-1}\Sigma^{(b),i}_{\beta,m}\right) \\ &- \frac{1}{2}(\mu^{(r),j}_{\rho,\ell} - \mu^{(b),i}_{\beta,m})^T(\Sigma^{(r),j}_{\rho,\ell})^{-1}(\mu^{(r),j}_{\rho,\ell} - \mu^{(b),i}_{\beta,m}). \end{aligned} \tag{33}$$

HMM: For the HMM lower bound $\mathcal{L}^{i,j}_{HMM}$, we assume each variational distribution takes the form of a Markov chain,

$$q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) = \phi^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) = \phi^{i,j}_1(\rho_1|\beta_1)\prod_{t=2}^{\tau}\phi^{i,j}_t(\rho_t|\rho_{t-1},\beta_t), \tag{34}$$

where $\sum_{\rho_1=1}^{S} \phi_1^{i,j}(\rho_1|\beta_1) = 1$, and $\sum_{\rho_t=1}^{S} \phi_t^{i,j}(\rho_t|\rho_{t-1},\beta_t) = 1$. The variational distribution $q^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})$ represents the distribution of the state sequence $\rho_{1:\tau}$ in HMM $\mathcal{M}_j^{(r)}$, when $\mathcal{M}_j^{(r)}$ is used to explain the *observation* sequence generated by $\mathcal{M}_i^{(b)}$ that evolved through state sequence $\beta_{1:\tau}$.

Substituting into (28), we have

$$\mathcal{L}_{HMM}^{i,j} = \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \max_{\phi^{i,j}} \sum_{\rho_{1:\tau}} \phi^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \left\{ \log \frac{\pi_{\rho_{1:\tau}}^{(r),j}}{\phi^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})} + \sum_t \mathcal{L}_{GMM}^{(i,\beta_t),(j,\rho_t)} \right\}. \quad (35)$$

The maximization with respect to $\phi_t^{i,j}(\rho_t|\rho_{t-1},\beta_t)$ and $\phi_1^{i,j}(\rho_1|\beta_1)$ is carried out independently for each pair (i,j), and follow [10], which is further detailed in Appendix A. By separating terms and breaking up the summation $\beta_{1:\tau}$ and $\rho_{1:\tau}$, the optimal $\hat{\phi}_t^{i,j}(\rho_t|\rho_{t-1},\beta_t)$ and $\hat{\phi}_1^{i,j}(\rho_1|\beta_1)$ can be obtained using an efficient recursive iteration (similar to the forward algorithm).

H3M: For the H3M lower bound $\mathcal{L}_{H3M}^{i}$, we assume variational distributions of the form $q_i(z_i = j) = z_{ij}$, where $\sum_{j=1}^{K^{(r)}} z_{ij} = 1$, and $z_{ij} \geq 0$. Substituting into (27), we have

$$\mathcal{L}_{H3M}^{i} = \max_{z_{ij}} \sum_j z_{ij} \left\{ \log \frac{\omega_j^{(r)}}{z_{ij}} + N_i \mathcal{L}_{HMM}^{i,j} \right\}. \quad (36)$$

The maximizing variational parameters of (36) are obtained by using Appendix C.2,

$$\hat{z}_{ij} = \frac{\omega_j^{(r)} \exp(N_i \mathcal{L}_{HMM}^{i,j})}{\sum_{j'} \omega_{j'}^{(r)} \exp(N_i \mathcal{L}_{HMM}^{i,j'})}. \quad (37)$$

Note that in the standard HEM algorithm derived in [2, 11], the assignment probabilities z_{ij} are based on the expected log-likelihoods of the components, (e.g., $\mathrm{E}_{\mathcal{M}_i^{(b)}}[\log p(y_{1:\tau}|\mathcal{M}_j^{(r)})]$ for H3Ms). For the variational HEM algorithm, these expectations are now replaced with their lower bounds (in our case, $\mathcal{L}_{HMM}^{i,j}$).

1.3.2 Lower bound

Substituting the optimal variational distributions into (31), (35), and (36) gives the lower bounds,

$$\mathcal{L}_{H3M}^{i} = \sum_j \hat{z}_{ij} \left\{ \log \frac{\omega_j^{(r)}}{\hat{z}_{ij}} + N_i \mathcal{L}_{HMM}^{i,j} \right\}, \quad (38)$$

$$\mathcal{L}_{HMM}^{i,j} = \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho_{1:\tau}} \hat{\phi}^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \left\{ \log \frac{\pi_{\rho_{1:\tau}}^{(r),j}}{\hat{\phi}^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})} + \sum_t \mathcal{L}_{GMM}^{(i,\beta_t),(j,\rho_t)} \right\}, \quad (39)$$

$$\mathcal{L}_{GMM}^{(i,\beta),(j,\rho)} = \sum_{m=1}^{M} c_{\beta,m}^{(b),i} \sum_{\ell=1}^{M} \hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \left\{ \log \frac{c_{\rho,\ell}^{(r),j}}{\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)}} + \mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho,\ell}^{(r)})] \right\}. \quad (40)$$

The lower bound $\mathcal{L}_{HMM}^{i,j}$ requires summing over all sequences $\beta_{1:\tau}$ and $\rho_{1:\tau}$ This summation can be computed efficiently along with $\hat{\phi}_t^{i,j}(\rho_t|\rho_{t-1},\beta_t)$ and $\hat{\phi}_1^{i,j}(\rho_1|\beta_1)$ using a recursive algorithm from [10] and is described in Appendix A.

1.3.3 Summary Statistics

After calculating the optimal variational distributions, we calculate the following summary statistics, which are necessary for the M-step:

$$\nu_1^{i,j}(\rho_1, \beta_1) = \pi_{\beta_1}^{(b),i}\, \hat{\phi}_1^{i,j}(\rho_1|\beta_1) \tag{41}$$

$$\xi_t^{i,j}(\rho_{t-1}, \rho_t, \beta_t) = \left(\sum_{\beta_{t-1}=1}^{S} \nu_{t-1}^{i,j}(\rho_{t-1}, \beta_{t-1})\, a_{\beta_{t-1},\beta_t}^{(b),i} \right) \hat{\phi}_t^{i,j}(\rho_t|\rho_{t-1}, \beta_t) \text{ for } t = 2, \ldots, \tau \tag{42}$$

$$\nu_t^{i,j}(\rho_t, \beta_t) = \sum_{\rho_{t-1}=1}^{S} \xi_t^{i,j}(\rho_{t-1}, \rho_t, \beta_t) \text{ for } t = 2, \ldots, \tau \tag{43}$$

and the aggregate statistics

$$\hat{\nu}_1^{i,j}(\rho) = \sum_{\beta=1}^{S} \nu_1^{i,j}(\rho, \beta) \tag{44}$$

$$\hat{\nu}^{i,j}(\rho, \beta) = \sum_{t=1}^{\tau} \nu_t^{i,j}(\rho, \beta) \tag{45}$$

$$\hat{\xi}^{i,j}(\rho, \rho') = \sum_{t=2}^{\tau} \sum_{\beta=1}^{S} \xi_t^{i,j}(\rho, \rho', \beta). \tag{46}$$

The statistic $\hat{\nu}_1^{i,j}(\rho)$ is the expected number of times that the HMM $\mathcal{M}_j^{(r)}$ starts from state ρ, when modeling sequences generated by $\mathcal{M}_i^{(b)}$. The quantity $\hat{\nu}^{i,j}(\rho, \beta)$ is the expected number of times that the HMM $\mathcal{M}_j^{(r)}$ is in state ρ when the HMM $\mathcal{M}_i^{(b)}$ is in state β, when both modeling sequences generated by $\mathcal{M}_i^{(b)}$. Similarly, the quantity $\hat{\xi}^{i,j}(\rho, \rho')$ is the expected number of transitions from state ρ to state ρ' of the HMM $\mathcal{M}_j^{(r)}$, when modeling sequences generated by $\mathcal{M}_i^{(b)}$.

1.4 M-step

In the M-step, the lower bound in (26) is maximized with respect to the parameters $\mathcal{M}^{(r)}$,

$$\mathcal{M}^{(r)^*} = \underset{\mathcal{M}^{(r)}}{\operatorname{argmax}} \sum_{i=1}^{K^{(b)}} \mathcal{L}_{H3M}^{i}. \tag{47}$$

The derivation of the maximization is presented in Appendix B. Each mixture component of $\mathcal{M}^{(r)}$ is updated independently according to

$$\omega_j^{(r)^*} = \frac{\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}}{K^{(b)}}, \tag{48}$$

$$\pi_\rho^{(r),j^*} = \frac{\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}\omega_i^{(b)} \hat{\nu}_1^{i,j}(\rho)}{\sum_{\rho'=1}^{S}\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}\omega_i^{(b)} \hat{\nu}_1^{i,j}(\rho'))}, \quad a_{\rho,\rho'}^{(r),j^*} = \frac{\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}\omega_i^{(b)} \hat{\xi}^{i,j}(\rho, \rho')}{\sum_{\sigma=1}^{S}\sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}\omega_i^{(b)} \hat{\xi}^{i,j}(\rho, \sigma)}, \tag{49}$$

$$c_{\rho,\ell}^{(r),j^*} = \frac{\Omega_{j,\rho}\left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)}\right)}{\sum_{\ell'=1}^{M} \Omega_{j,\rho}\left(\hat{\eta}_{\ell'|m}^{(i,\beta),(j,\rho)}\right)}, \qquad \mu_{\rho,\ell}^{(r),j^*} = \frac{\Omega_{j,\rho}\left(\eta_{\ell|m}^{(i,\beta),(j,\rho)}\ \mu_{\beta,m}^{(b),i}\right)}{\Omega_{j,\rho}\left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)}\right)}, \tag{50}$$

$$\Sigma_{\rho,\ell}^{(r),j^*} = \frac{\Omega_{j,\rho}\left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)}\left[\Sigma_{\beta,m}^{(b),i} + (\mu_{\beta,m}^{(b),i} - \mu_{\rho,\ell}^{(r),j})(\mu_{\beta,m}^{(b),i} - \mu_{\rho,\ell}^{(r),j})^T\right]\right)}{\Omega_{j,\rho}\left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)}\right)}, \tag{51}$$

where $\Omega_{j,\rho}(\cdot)$ is the weighted sum operator over all base models, HMM states, and GMM components (ie., over all tuples (i, β, m)),

$$\Omega_{j,\rho}(f(i,\beta,m)) = \sum_{i=1}^{K^{(b)}} \hat{z}_{i,j}\omega_i^{(b)} \sum_{\beta=1}^{S} \hat{\nu}^{i,j}(\rho, \beta) \sum_{m=1}^{M} c_{\beta,m}^{(b),i}\, f(i,\beta,m). \tag{52}$$

Note that the covariance matrices of the reduced models (51) are never smaller (in magnitude) than the covariance matrices of the base models, due to the outer-product term. This regularization effect derives from the E-step, which averages all possible observations from the base model.

Appendix A. Derivation of the E-step

The maximization of (35) with respect to $\phi_t^{i,j}(\rho_t|\rho_{t-1},\beta_t)$ and $\phi_1^{i,j}(\rho_1|\beta_1)$ is carried out independently for each pair (i,j), and follow [10]. In particular it uses a backward recursion, starting with $\mathcal{L}_{\tau+1}^{i,j}(\beta_t,\rho_t)=0$, for $t=\tau,\ldots,2$,

$$\hat{\phi}_t^{i,j}(\rho_t|\rho_{t-1},\beta_t)=\frac{a_{\rho_{t-1},\rho_t}^{(r),j}\exp\left\{\mathcal{L}_{\text{GMM}}^{(i,\beta_t),(j,\rho_t)}+\mathcal{L}_{t+1}^{i,j}(\beta_t,\rho_t)\right\}}{\sum_\rho^S a_{\rho_{t-1},\rho}^{(r),j}\exp\left\{\mathcal{L}_{\text{GMM}}^{(i,\beta_t),(j,\rho)}+\mathcal{L}_{t+1}^{i,j}(\beta_t,\rho)\right\}} \tag{53}$$

$$\mathcal{L}_t^{i,j}(\beta_{t-1},\rho_{t-1})=\sum_{\beta=1}^S a_{\beta_{t-1},\beta}^{(b),i}\log\sum_{\rho=1}^S a_{\rho_{t-1},\rho}^{(r),j}\exp\left\{\mathcal{L}_{\text{GMM}}^{(i,\beta),(j,\rho)}+\mathcal{L}_{t+1}^{i,j}(\beta,\rho)\right\}, \tag{54}$$

and terminates with

$$\hat{\phi}_1^{i,j}(\rho_1|\beta_1)=\frac{\pi_{\rho_1}^{(r),j}\exp\left\{\mathcal{L}_{\text{GMM}}^{(i,\beta_1),(j,\rho_1)}+\mathcal{L}_2^{i,j}(\beta_1,\rho_1)\right\}}{\sum_\rho^S \pi_\rho^{(r),j}\exp\left\{\mathcal{L}_{\text{GMM}}^{(i,\beta_1),(j,\rho)}+\mathcal{L}_2^{i,j}(\beta_1,\rho)\right\}} \tag{55}$$

$$\mathcal{L}_{\text{HMM}}^{i,j}=\sum_{\beta=1}^S \pi_\beta^{(b),i}\log\sum_{\rho=1}^S \pi_\rho^{(r),j}\exp\left\{\mathcal{L}_{\text{GMM}}^{(i,\beta),(j,\rho)}+\mathcal{L}_2^{i,j}(\beta,\rho)\right\} \tag{56}$$

where (56) is the maxima of the terms in (35) in Section 1.3.1.

Appendix B. Derivation of the M-step

The M-steps involves maximizing the lower bound in (26) with respect to $\mathcal{M}^{(r)}$, while holding the variational distributions fixed,

$$\mathcal{M}^{(r)^*}=\underset{\mathcal{M}^{(r)}}{\text{argmax}}\sum_{i=1}^{K^{(b)}}\mathcal{L}_{H3M}^i. \tag{57}$$

Substituting (38) and (39) into the objective function of (57),

$$\mathcal{L}(\mathcal{M}^{(r)})=\sum_{i=1}^{K^{(b)}}\mathcal{L}_{H3M}^i \tag{58}$$

$$=\sum_{i,j}\hat{z}_{ij}\left\{\log\frac{\omega_j^{(r)}}{\hat{z}_{ij}}+N_i\sum_{\beta_{1:\tau}}\pi_{\beta_{1:\tau}}^{(b),i}\sum_{\rho_{1:\tau}}\hat{\phi}^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})\left[\log\frac{\pi_{\rho_{1:\tau}}^{(r),j}}{\hat{\phi}^{i,j}(\rho_{1:\tau}|\beta_{1:\tau})}+\sum_t\mathcal{L}_{GMM}^{(i,\beta_t),(j,\rho_t)}\right]\right\} \tag{59}$$

In the following, we detail the update rules for the parameters of the reduced model $\mathcal{M}^{(r)}$.

HMMs mixture weights

Collecting terms in (59) that only depend on the mixture weights $\{\omega_j^{(r)}\}_{j=1}^{K^{(r)}}$, we have

$$\widetilde{\mathcal{L}}(\{\omega_j^{(r)}\})=\sum_i\sum_j\hat{z}_{ij}\log\omega_j^{(r)}=\sum_j\left[\sum_i\hat{z}_{ij}\right]\log\omega_j^{(r)} \tag{60}$$

Given the constraints $\sum_{j=1}^{K^{(r)}}\omega_j^{(r)}=1$ and $\omega_j^{(r)}\geq 0$, (60) is maximized using the result in Appendix C.1, which yields the update in (48).

Initial state probabilities

The objective function in (59) factorizes for each HMM $\mathcal{M}_j^{(r)}$, and hence the parameters of each HMM are updated independently. For the j-th HMM, we collect terms in (59) that depend on the initial state probabilities $\{\pi_\rho^{(r),j}\}_{\rho=1}^S$,

$$\widetilde{\mathcal{L}}_j(\{\pi_\rho^{(r),j}\}) = \sum_i \hat{z}_{ij} N_i \sum_{\beta_1} \pi_{\beta_1}^{(b),i} \sum_{\rho_1} \hat{\phi}_1^{i,j}(\rho_1|\beta_1) \log \pi_{\rho_1}^{(r),j} \tag{61}$$

$$= \sum_{\rho_1} \sum_i \hat{z}_{ij} N_i \underbrace{\sum_{\beta_1} \pi_{\beta_1}^{(b),i} \hat{\phi}_1^{i,j}(\rho_1|\beta_1)}_{\hat{\nu}_1^{i,j}(\rho_1)} \log \pi_{\rho_1}^{(r),j} \tag{62}$$

$$= \sum_\rho \sum_i \hat{z}_{ij} N_i \hat{\nu}_1^{i,j}(\rho) \log \pi_\rho^{(r),j} \tag{63}$$

$$\propto \sum_\rho \left[\sum_i \hat{z}_{ij} \omega_i^{(b)} \hat{\nu}_1^{i,j}(\rho) \right] \log \pi_\rho^{(r),j}, \tag{64}$$

where in the (63) we have used the summary statistic defined in (41). Considering the constraints $\sum_{\rho=1}^S \pi_\rho^{(r),j} = 1$ and $\pi_\rho^{(r),j} \geq 0$, (64) is maximized using the result in Appendix C.1, giving the update formula in (49).

State transition probabilities

Similarly, for each HMM $\mathcal{M}_j^{(r)}$ and previous state ρ, we collect terms in (59) that depend on the transition probabilities $\{a_{\rho,\rho'}^{(r),j}\}_{\rho'=1}^S$,

$$\widetilde{\mathcal{L}}_{j,\rho}(\{a_{\rho,\rho'}^{(r),j}\}_{\rho'=1}^S) = \sum_i \hat{z}_{ij} N_i \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho_{1:\tau}} \hat{\phi}^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \log \pi_{\rho_{1:\tau}}^{(r),j} \tag{65}$$

$$\propto \sum_i \hat{z}_{ij} N_i \sum_{\beta_{1:\tau}} \left[\prod_{t=2}^\tau a_{\beta_{t-1},\beta_t}^{(b),i} \right] \sum_{\rho_{1:\tau}} \left[\prod_{t=2}^\tau \hat{\phi}_t^{i,j}(\rho_t|\rho_{t-1},\beta_t) \right] \left[\sum_{t=2}^\tau \log a_{\rho_{t-1},\rho_t}^{(r),j} \right] \tag{66}$$

$$= \sum_i \sum_{\rho'=1}^S \hat{z}_{ij} N_i \hat{\xi}^{i,j}(\rho,\rho') \log a_{\rho,\rho'}^{(r),j} \tag{67}$$

$$\propto \sum_{\rho'=1}^S \left[\sum_i \hat{z}_{ij} \omega_i^{(b)} \hat{\xi}^{i,j}(\rho,\rho') \right] \log a_{\rho,\rho'}^{(r),j}. \tag{68}$$

Considering the constraints $\sum_{\rho'=1}^S a_{\rho,\rho'}^{(r),j} = 1$ and $a_{\rho,\rho'}^{(r),j} \geq 0$, (68) is maximized using the result in Appendix C.1, giving the update in (49).

Emission probability density functions

The cost function (59) factors also for each GMM indexed by (j,ρ,ℓ). Collecting relevant terms in (59),

$$\widetilde{\mathcal{L}}(\mathcal{M}_{j,\rho,\ell}^{(r)}) = \sum_i \hat{z}_{ij} N_i \sum_{\beta_{1:\tau}} \pi_{\beta_{1:\tau}}^{(b),i} \sum_{\rho_{1:\tau}} \hat{\phi}^{i,j}(\rho_{1:\tau}|\beta_{1:\tau}) \sum_t \mathcal{L}_{GMM}^{(i,\beta_t),(j,\rho_t)} \tag{69}$$

$$= \sum_i \hat{z}_{ij} N_i \sum_{\beta=1}^S \hat{\nu}^{i,j}(\rho,\beta) \mathcal{L}_{GMM}^{(i,\beta),(j,\rho)} \tag{70}$$

$$\propto \sum_i \hat{z}_{ij} N_i \sum_{\beta=1}^S \hat{\nu}^{i,j}(\rho,\beta) \sum_{m=1}^M c_{\beta,m}^{(b),i} \; \hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \left[\log c_{\rho,\ell}^{(r),j} + \mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho,\ell}^{(r)})] \right] \tag{71}$$

$$= \Omega_{j,\rho} \left(\hat{\eta}_{\ell|m}^{(i,\beta),(j,\rho)} \left[\log c_{\rho,\ell}^{(r),j} + \mathrm{E}_{\mathcal{M}_{i,\beta,m}^{(b)}}[\log p(y|\mathcal{M}_{j,\rho,\ell}^{(r)})] \right] \right), \tag{72}$$

where in (72) we use the weighted-sum operator defined in (52), which is over all base model GMMs $\{\mathcal{M}^{(b)}_{i,\beta,m}\}$. The GMM mixture weights are subject to the constraints $\sum_{\ell=1}^{M} c^{(r),j}_{\rho,\ell} = 1, \forall j, \rho$. Taking the derivative with respect to each parameter and setting it to zero[2], gives the GMM update equations (50) and (51).

Appendix C. Useful optimization problems

Appendix C.1

The optimization problem

$$\begin{aligned} \max_{\alpha_\ell} \quad & \sum_{\ell=1}^{L} \beta_\ell \log \alpha_\ell \\ \text{s.t.} \quad & \sum_{\ell=1}^{L} \alpha_\ell = 1 \\ & \alpha_\ell \geq 0, \, \forall \ell \end{aligned} \tag{73}$$

is optimized by

$$\alpha^*_\ell = \frac{\beta_\ell}{\sum_{\ell'=1}^{L} \beta'_\ell}. \tag{74}$$

Appendix C.2

The optimization problem

$$\begin{aligned} \max_{\alpha_\ell} \quad & \sum_{\ell=1}^{L} \alpha_\ell \left(\beta_\ell - \log \alpha_\ell\right) \\ \text{s.t.} \quad & \sum_{\ell=1}^{L} \alpha_\ell = 1 \\ & \alpha_\ell \geq 0, \, \forall \ell \end{aligned} \tag{75}$$

is optimized by

$$\alpha^*_\ell = \frac{\exp \beta_\ell}{\sum_{\ell'=1}^{L} \exp \beta'_\ell}. \tag{76}$$

References

[1] Anonymous. The variational hierarchical EM algorithm for clustering hidden Markov models. *submitted to NIPS, paper ID 206*, 2012.

[2] N. Vasconcelos and A. Lippman. Learning mixture hierarchies. In *Advances in Neural Information Processing Systems*, 1998.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[4] R.M. Neal and G.E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *NATO ASI SERIES D BEHAVIOURAL AND SOCIAL SCIENCES*, 89:355–370, 1998.

[5] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[2] We also considered the constraints on the covariance matrices $\Sigma^{(r),j}_{\rho,\ell} \succ \mathbf{0}$.

[6] I. Csisz, G. Tusnády, et al. Information geometry and alternating minimization procedures. *Statistics and decisions*, 1984.

[7] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[8] Tommi S. Jaakkola. Tutorial on Variational Approximation Methods. In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

[9] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Upper Saddle River (NJ, USA), 1993.

[10] J.R. Hershey, P.A. Olsen, and S.J. Rennie. Variational Kullback-Leibler divergence for hidden Markov models. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 323–328. IEEE, 2008.

[11] A.B. Chan, E. Coviello, and G.R.G. Lanckriet. Clustering dynamic textures with the hierarchical em algorithm. In *Intl. Conference on Computer Vision and Pattern Recognition*, 2010.