Hierarchical Inference on Single-molecule Time Series, using VBEM and Emperical Bayes on HMM's

Contents

1 Model Properties

1.1 Data, latent states, parameters

$x = \{x_n\} = \{\{x_{n,t}\}\}$	Observation in trace $n \in \{1 \dots N\}$ at time $t \in \{1 \dots T_n\}$
$z = \{z_n\} = \{\{z_{n,t}\}\}$	State of molecule n at time t
$\theta = \{\theta_n\} = \{\pi_n, A_n, \mu_n, \lambda_n\}$	Parameters for trace n
$\pi_n = \{\pi_{n,k}\}$	Initial probabilities: Prob that trace starts in state k
$A_n = \{\{A_{n,k,l}\}\}$	Transition matrix: Prob of moving from state k to state l
$\mu_n = \{\mu_{n,k}\}$	FRET level for state k in trace n
$\lambda_n = \{\lambda_{n,k}\}$	FRET emissions precision (1/var) for state k in trace n
$u = \{\{u^\mu_k, u^\beta_k, u^W_k, u^\nu_k\}, \{u^A_k\}, \{u^\pi\}\}$	Hyperparameters for ensemble distribution (prior)

1.2 Evidence

$$
\begin{aligned}
p(x \mid u) &= \int d\theta \, p(x, \theta \mid u) \\
&= \int d\theta \, p(x \mid \theta) p(\theta \mid u) \\
&= \int d\theta \, \prod_n p(x_n \mid \theta_n) p(\theta_n \mid u) \\
&= \prod_n \int d\theta_n \, p(x_n \mid \theta_n) p(\theta_n \mid u)
\end{aligned} \tag{1}
$$

Likelihood

$$
\begin{aligned}
p(x \mid \theta) &= \prod_n p(x_n \mid \theta_n) \\
&= \prod_n \sum_{z_n} p(x_n, z_n \mid \theta_n) \\
&= \prod_n \sum_{z_n} p(x_n \mid z_n, \theta_n) p(z_n \mid \theta_n)
\end{aligned} \tag{2}
$$

1.3 Emissions model

$$
\begin{aligned}
p(x_n \mid z_n, \theta_n) &= \prod_t p(x_{n,t} \mid z_{n,t} \, , \, \theta_n) \\
&= \prod_t \prod_k p(x_{n,t} \mid \theta_{n,k})^{z_{n,t,k}}
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
p(x_{n,t} \mid \theta_{n,k}) &= N(x_{n,t} \,|\, \mu_{n,k}, \lambda_{n,k}) \\
&= (\lambda_{n,k}/2\pi)^{1/2} \exp[-\tfrac{1}{2}\Delta^2_{n,t,k}]
\end{aligned} \tag{4}
$$

$$
\Delta^2_{n,t,k} = \lambda_{n,k}(x_{n,t} - \mu_{n,k})^2 \tag{5}
$$

1.4 Transition probabilities (HMM)

$$
p(z_n \mid \theta_n) = \left[\prod_{t=2}^{T_n} p(z_{n,t} \mid z_{n,t-1}, \theta_n) \right] p(z_{n,1} \mid \theta_n) \tag{6}
$$

$$
p(z_{n,t} \mid z_{n,t-1} \, , \, \theta_n) = \prod_{k,l} (A_{n,k,l})^{z_{n,t-1,k} \, z_{n,t,l}} \tag{7}
$$

$$
p(z_1 \mid \theta_n) = \prod_k (\pi_{n,k})^{z_{n,1,k}} \tag{8}
$$

1.5 Ensemble Distributions (VBEM Priors)

$$
\begin{aligned}
p(\theta_n \mid u) &= p(\pi_n \mid u) p(A_n \mid u) p(\mu_n, \lambda_n \mid u) \\
&= p(\pi_n \mid u) \prod_k p(A_{n,k} \mid u) p(\mu_{n,k}, \lambda_{n,k} \mid u)
\end{aligned} \tag{9}
$$

$$\pi_n \sim \mathrm{Dir}(u^\pi) \tag{10}$$

$$A_{n,k} \sim \mathrm{Dir}(u_k^A) \tag{11}$$

$$\lambda_{n,k} \sim \mathrm{Wish}(u_k^W, u_k^\nu) \tag{12}$$

$$\mu_{n,k} \sim \mathrm{N}(u_k^\mu, u_k^\beta \lambda_{n,k}) \tag{13}$$

1.6 Algorithm Outline

Loop over iterations i until $\sum_n \mathcal{L}_n$ converges:

1. VBEM updates: obtain $q^{(i)}(\theta_n)$, $q^{(i)}(z_n)$, and $\mathcal{L}_n^{(i)}$ for each trace n, using $p^{(i-1)}(\theta_n)$ as the VBEM prior on the paramters.
2. Hierarchical updates: solve for

$$p^{(i)}(\theta \mid u) = \arg\max_{p(\theta)} \sum_n \mathcal{L}_n^{(i)}[q^{(i)}(z_n), q^{(i)}(\theta_n), p(\theta \mid u)]$$

2 Conjugate-Exponential Form

Given that all our likelihoods and prior are in the exponential family, the likelihood $p(x \mid \eta)$ and the prior $p(\eta \mid \nu, \chi)$ can be written in a common general form:

$$p(x \mid \eta) = f(x) g(\theta) \exp[\eta \cdot \upsilon(x)] \tag{14}$$

$$p(\eta \mid \nu, \chi) = h(\nu, \chi) g(\theta)^\nu \exp[\eta \cdot \chi] \tag{15}$$

where η represents the remapped parameters θ, and $\{\nu, \chi\}$ represent the remapped hyperparameters to this general form. The posterior $p(\eta \mid x, \nu, \chi)$ now takes the form:

$$p(\eta \mid x, \nu, \chi) \propto p(x \mid \eta) p(\eta \mid \nu, \chi) \tag{16}$$

$$= f(x) h(\nu, \chi) g^{\nu+1} \exp[\eta \cdot (\chi + \upsilon(x))] \tag{17}$$

which of course yields a distribution of the same form as the prior

$$p(\eta \mid x, \nu, \chi) = p(\eta \mid \tilde{\nu}, \tilde{\chi}) \tag{18}$$

with parameters

$$\tilde{\nu} = \nu + 1 \tag{19}$$

$$\tilde{\chi} = \chi + \upsilon(x) \tag{20}$$

In generalized exponential form, the hyperparameter ν can be interpreted as scale factor, that encodes the number of spreviously observed samples. The hyperparameter vector χ, in turn takes the role of the summed sufficient statistics $\upsilon(x)$ associated with each of the samples.

2.1 Normal-Gamma

$$p(x \mid \mu, \lambda) = N(x \mid \mu, \lambda) \tag{21}$$
$$p(\mu, \lambda \mid u^\mu, u^\beta, u^a, u^b) = N(\mu \mid u^\mu, \lambda u^\beta)\mathrm{Gamma}(\lambda \mid u^a, u^b) \tag{22}$$

$$\eta = \{\lambda, \lambda\mu\} \tag{23}$$
$$\nu = u^\beta = 2u^a - 1 \tag{24}$$
$$\chi = \{-\tfrac{1}{2}(u^\beta(u^\mu)^2 + 2u^b), u^\beta u^\mu\} \tag{25}$$

$$g(\eta) = (\eta_1/2\pi)^{1/2} \exp[-\eta_2^2/2\eta_1] \tag{26}$$
$$\upsilon(x) = \{-\tfrac{1}{2}x^2, x\} \tag{27}$$

$$f(x) = 1 \tag{28}$$
$$h(\nu, \chi) = (2\pi)^{(\nu-1)/2}\nu^{1/2}(-\chi_1 - \chi_2^2/\nu)^{(\nu+1)/2} \tag{29}$$

2.2 Normal-Wishart (1d)

$$p(x \mid \mu, \lambda) = N(x \mid \mu, \lambda) \tag{30}$$
$$p(\mu, \lambda \mid u^\mu, u^\beta, u^a, u^b) = N(\mu \mid u^\mu, \lambda u^\beta)\mathrm{Wish}(\lambda \mid u^W, u^\nu) \tag{31}$$

$$\eta = \{\lambda, \lambda\mu\} \tag{32}$$
$$\nu = u^\beta = u^\nu - 1 \tag{33}$$
$$\chi = \{-\tfrac{1}{2}(u^\beta(u^\mu)^2 + 1/u^W), u^\beta u^\mu\} \tag{34}$$

Functions $g(\eta)$, $u(x)$, $f(x)$ and $h(\nu, \chi)$ are the same as with a Normal-Gamma distribution.

2.3 Dirichlet

$$p(z \mid \pi) = \mathrm{Cat}(z \mid \pi) = \prod_k \pi_k^{z_k} \tag{35}$$
$$p(\pi \mid u^\pi) = \mathrm{Dir}(\pi \mid u^\pi) = \frac{\Gamma(\sum_k u_k^\pi)}{\prod_k \Gamma(u_k^\pi)} \prod_k \pi_k^{u_k^\pi - 1} \tag{36}$$

$$\eta = \{\ln \pi_k\} \tag{37}$$
$$\nu = 1 \tag{38}$$
$$\chi = \{u_k^\pi\} \tag{39}$$

$$g(\eta) = 1 \tag{40}$$
$$\upsilon(z) = \{z_k\} \tag{41}$$

$$
\begin{aligned}
f(x) &= 1 && (42)\\
h(\nu, \chi) &= \frac{\prod_k \Gamma(\chi_k + 1)}{\Gamma(\sum_k (\chi_k + 1))} && (43)
\end{aligned}
$$

3 Variational Bayes Expectation Maximization (VBEM)

Note: We will omit the n-subscript in this section, since VBEM is performed on one trace at a time.

When performing (structured) VBEM on a Hidden Markov Model, we introduce an approximating factorization for the posterior $p(z, \theta \mid x) \simeq q(z)q(\theta)$, that allows calculation of a lower bound on the log-evidence (using Jensen's inequality):

$$
\begin{aligned}
\ln p(x) &= \ln \left[\int d\theta \, \sum_z p(x, z, \theta) \right] \\
&= \ln \left[\int d\theta \, \sum_z q(z)q(\theta) \frac{p(x, z, \theta)}{q(z)q(\theta)} \right] \\
&\geq \int d\theta \, \sum_z q(z)q(\theta) \ln \left[\frac{p(x, z, \theta)}{q(z)q(\theta)} \right] \\
&= \mathcal{L}[q(z), q(\theta)]
\end{aligned} \tag{44}
$$

The lower bound $\mathcal{L}$ is tight if $q(z)q(\theta) = p(z, \theta \mid x)$:

$$
\begin{aligned}
\mathcal{L}[q(z), q(\theta)] &= \int d\theta \, \sum_z q(z)q(\theta) \ln \left[\frac{p(x, z, \theta)}{q(z)q(\theta)} \right] \\
&= \int d\theta \, \sum_z p(z, \theta \mid x) \ln \left[\frac{p(x, z, \theta)}{p(z, \theta \mid x)} \right] \\
&= \int d\theta \, \sum_z p(z, \theta \mid x) \ln \left[\frac{p(z, \theta \mid x)p(x)}{p(z, \theta \mid x)} \right] \\
&= \int d\theta \, \sum_z p(z, \theta \mid x) \ln p(x) \\
&= \ln p(x) \int d\theta \, \sum_z p(z, \theta \mid x) \\
&= \ln p(x)
\end{aligned} \tag{45}
$$

3.1 Updates

Loop until $\mathcal{L}$ converges. For i-th iteration:

1. E-step: keeping $q^{(i)}(\theta)$ fixed, solve for
$$q^{(i+1)}(z) = \arg\max_{q(z)} \mathcal{L}[q(z), q^{(i)}(\theta)]$$

2. M-step: keeping $q^{(i)}(z)$ fixed, solve for
$$q^{(i+1)}(\theta) = \arg\max_{q(\theta)} \mathcal{L}[q^{(i)}(z), q(\theta)]$$

3.2 E-step

To maximize $\mathcal{L}$ w.r.t. $q(z)$, we solve $\nabla_{q(z)}\mathcal{L} = 0$, introducing a Lagrange multiplier λ_z to ensure normalization:

$$\begin{aligned} 0 &= \nabla_{q(z)} \left[\mathcal{L}[q(z), q(\theta)] + \lambda_z \left(1 - \sum_{z'} q(z') \right) \right] \\ &= \left[\int d\theta \, q(\theta) \left(\ln p(x, z, \theta) - \ln q(z) - \ln q(\theta) - 1 \right) \right] - \lambda_z \end{aligned} \tag{46}$$

We can pull $\ln q(z)$ out of the integral, since it has no dependence on θ. This yields

$$\begin{aligned} \ln q(z) &= \left[\int d\theta \, q(\theta) \left(\ln p(x, z, \theta) - \ln q(\theta) - 1 \right) \right] - \lambda_z \\ &= E_{q(\theta)}[\ln p(x, z, \theta)] - E_{q(\theta)}[\ln q(\theta)] - (1 + \lambda_\theta) \\ &= E_{q(\theta)}[\ln p(x, z, \theta)] - \ln Z_{q(\theta)} \end{aligned} \tag{47}$$

here we have absorbed all terms without a z-dependence into a constant $\ln Z_{q(z)}$. The approximate posterior $q(z)$ is obtained by taking the exponent of the above equation

$$q(z) = \frac{1}{Z_{q(\theta)}} \exp \left[E_{q(\theta)}[\ln p(x, z, \theta)] \right] \tag{48}$$

where normalization of $q(z)$ implies

$$Z_{q(z)} = \sum_z \exp \left[E_{q(\theta)}[\ln p(x, z, \theta)] \right] \tag{49}$$

The expectation of $p(x, z, \theta)$ w.r.t. $q(\theta)$ expands to:

$$E_{q(\theta)}[\ln p(x, z, \theta)] = \int d\theta \, q(\theta) \left[\ln p(x \mid z, \theta) + \ln p(z \mid \theta) + \ln p(\theta \mid u) \right] \tag{50}$$

The z-dependent terms can be written as:

$$\begin{aligned} E_{q(\theta)}[\ln p(x \mid z, \theta)] &= \sum_t \sum_k z_{t,k} \int d\theta \, q(\theta) \left[\tfrac{1}{2} \ln (\lambda_k / 2\pi) - \tfrac{1}{2} \Delta^2 \right] \\ &= \sum_t z_t^\top \cdot E_{q(\theta)} [\tfrac{1}{2} \ln (\lambda_: / 2\pi) - \tfrac{1}{2} \Delta^2] \end{aligned} \tag{51}$$

and

$$\begin{aligned} E_{q(\theta)}[\ln p(z \mid \theta)] &= \sum_{t=2}^{T} \sum_{k,l} z_{t,l} z_{t-1,k} \int d\theta \, q(\theta) \ln A_{kl} \\ &\quad + \sum_k z_{1,k} \int d\theta \, q(\theta) \ln \pi_k \\ &= \sum_{t=2}^{T} z_{t-1}^\top \cdot E_{q(\theta)}[\ln A] \cdot z_t + z_t^\top \cdot E_{q(\theta)}[\ln \pi] \end{aligned} \tag{52}$$

Note that we do not need the expectation of the prior $E_{q(\theta)}[p(\theta)]$, since

$$
\begin{aligned}
q(z) =& \frac{\exp\left(E_{q(\theta)}[\ln p(x,z,\theta)]\right)}{\sum_z \exp\left(\ln E_{q(\theta)}[p(x,z,\theta)]\right)} \\
=& \frac{\exp\left(E_{q(\theta)}[\ln p(x,z \mid \theta)] + E_{q(\theta)}[\ln p(\theta)]\right)}{\sum_z \exp\left(E_{q(\theta)}[\ln p(x,z \mid \theta)] + E_{q(\theta)}[\ln p(\theta)]\right)} \\
=& \frac{\exp\left(E_{q(\theta)}[\ln p(x,z \mid \theta)]\right)}{\sum_z \exp\left(E_{q(\theta)}[\ln p(x,z \mid \theta)]\right)} \frac{\exp\left(E_{q(\theta)}[\ln p(\theta)]\right)}{\exp\left(E_{q(\theta)}[\ln p(\theta)]\right)} \\
=& \frac{\exp\left(E_{q(\theta)}[\ln p(x,z \mid \theta)]\right)}{\sum_z \exp\left(E_{q(\theta)}[\ln p(x,z \mid \theta)]\right)}
\end{aligned}
\tag{53}
$$

We see that the posterior $q(z)$ is parametrized by expectation under $q(\theta)$ of the squared Mahalanobis distance $E_{q(\theta)}[\Delta^2_{t,k}]$, and the logarithm of the parameters $E_{q(\theta)}[\ln \lambda]$, $E_{q(\theta)}[\ln A]$ and $E_{q(\theta)}[\ln \pi]$. This allows us to define

$$
\begin{aligned}
q(z) &= \frac{1}{\hat{Z}_{q(z)}} p^*(x,z) \\
p^*(x,z) &= \exp\left[E_{q(\theta)}[\ln p(x,z \mid \theta)]\right] \\
\hat{Z}_{q(z)} &= \sum_z p^*(x,z) = p^*(x) = Z_{q(z)}/E[p(\theta)]_{q(\theta)}
\end{aligned}
\tag{54}
$$

which decomposes into

$$p^*(x,z) = \left[\prod_t p^*(x_t \mid z_t)\right] p^*(z \mid \theta) \tag{55}$$

$$p^*(x_t \mid z_t = k) = (\lambda^*_k/2\pi)^{1/2} \exp\left[-\tfrac{1}{2}\Delta^{*2}_{t,k}\right] \tag{56}$$

$$p^*(z \mid \theta) = p(z \mid A^*, \pi^*) \tag{57}$$

by defining

$$\Delta^{*2} = E_{q(\theta)}[\Delta^2] \tag{58}$$

$$\ln \lambda^* = E_{q(\theta)}[\ln \lambda] \tag{59}$$

$$\ln A^* = E_{q(\theta)}[\ln A] \tag{60}$$

$$\ln \pi^* = E_{q(\theta)}[\ln \pi] \tag{61}$$

This result is a specific example of a general property of all exponential family models with conjugate likelihood/prior pairs: we can always find a set of point-estimates η^* such that (*reference Beal here*)

$$q(z) = \frac{1}{Z_{q(\eta)}} \exp[E_{q(\eta)}[\ln p(x,z,\eta)]] = \frac{1}{Z_{q(\eta)}} p(x,z,\eta^*) \tag{62}$$

In our specific case, this result implies that we could in principle compute some η^* for the natural parameters for the Normal-Wishart distribution $\eta = \{\lambda, \lambda\mu\}$, such that $p(x \mid \eta^*_k) = (\lambda^*_k/2\pi)^{1/2} \exp[-\frac{1}{2}\Delta^{*2}]$. However for the purposes of implementing the VBEM algorithm, this step is not required to calculate $q(z)$.

From the analytical forms of the priors, we can express the point estimates as (*TODO: verify the algebra here*):

$$\Delta^{*2} = (1/w_k^\beta) + w_k^\nu w_k^W (x - w_k^\mu)^2 \tag{63}$$

$$\ln \lambda^* = \psi(w_k^\nu) + \ln 2 w_k^W \tag{64}$$

$$\ln A_{k,l}^* = \psi\left(w_{k,l}^A\right) - \psi\left(\sum_l w_{k,l}^A\right) \tag{65}$$

$$\ln \pi_k^* = \psi\left(w_k^\pi\right) - \psi\left(\sum_k w_k^\pi\right) \tag{66}$$

In practice, we do not calculate $q(z)$ for all K^T possible paths through the state space (which would be numerically unfeasible). Rather, show in the next section that the updates for $q(\theta)$ only require knowledge of a set of point estimates of the state $z_{t,k}$ and transition correlation $z_{t-1,k} z_{t,l}$. We will show how to calculate these using the *forward-backward* algorithm at the end of the section.

3.3 M-step

In the m-step we maximize $\mathcal{L}$ w.r.t. $q(\theta)$. Again λ_θ is a Lagrange multiplier. We now take the functional derivative instead of a gradient, but the steps are essentially the same.

$$0 = \frac{\delta}{\delta q(\theta)} \left[\mathcal{L}[q(z), q(\theta)] + \lambda_\theta \left(1 - \int d\theta'\, q(\theta')\right)\right] \tag{67}$$

$$= \left[\sum_z q(z) \left(\ln p(x, z, \theta) - \ln q(z) - \ln q(\theta) - 1\right)\right] - \lambda_\theta \tag{68}$$

like in the E-step, this reduces to

$$\ln q(\theta) = \left[\sum_z q(z) \left(\ln p(x, z, \theta) - \ln q(z) - 1\right)\right] - \lambda_\theta \tag{69}$$

$$= E_{q(z)}[\ln p(x, z, \theta)] - E_{q(z)}[\ln q(z)] - (1 + \lambda_\theta) \tag{70}$$

$$= E_{q(z)}[\ln p(x, z, \theta)] - \ln Z_{q(\theta)} \tag{71}$$

with normalization constant $Z_{q(\theta)}$

$$\ln Z_{q(\theta)} = \int d\theta \sum_z q(z) \ln p(x, z, \theta) \tag{72}$$

The expectation of $\ln p(x, z, \theta)$ expands to:

$$\begin{aligned} E_{q(z)}[\ln p(x, z, \theta)] =& E_{q(z)}[\ln p(x \mid z, \theta) + E_{q(z)}[\ln p(z \mid \theta)] \\ &+ \ln p(\theta \mid u) \end{aligned} \tag{73}$$

where the z-dependent terms become:

$$E_{q(z)}[\ln p(x \mid z, \theta)] = \sum_t \sum_k E_{q(z)}[z_{t,k}] \left[\tfrac{1}{2} \ln \left(\lambda_k / 2\pi\right) - \tfrac{1}{2} \Delta^2_{t,k}\right] \tag{74}$$

$$\begin{aligned} E_{q(z)}[\ln p(z \mid \theta)] = &\sum_{t=2}^{T} \sum_{k,l} E_{q(z)}[z_{t,l} z_{t-1,k}] \ln A_{kl} \\ &+ \sum_k E_{q(z)}[z_{1,k}] \ln \pi_k \end{aligned} \tag{75}$$

the sufficient statistics for $q(z)$, which can be calculated with a forward backward algorithm (see below), are given by:

$$\gamma_{t,k} = E_{q(z)}[z_{t,k}] \tag{76}$$

$$\xi_{t,kl} = E_{q(z)}[z_{t-1,k} z_{t,l}] \tag{77}$$

and the expression for $q(\theta)$ can be rewritten as:

$$\begin{aligned} q(\theta) = &\frac{p(\theta|u)}{Z_{q(\theta)}} \prod_{t,k} \left(\left(\lambda_k / 2\pi\right)^{1/2} \exp\left[-\tfrac{1}{2} \Delta^2_{t,k}\right] \right)^{\gamma_{t,k}} \\ &\prod_{t=2,k,l} \left(A_{kl}\right)^{\xi_{t,kl}} \prod_k \left(\pi_k\right)^{\gamma_{1,k}} \end{aligned} \tag{78}$$

Again we see that we can write:

$$p^*(x, z, \theta) = \exp\left[\ln E_{q(z)}[p(x, z, \theta)]\right] \tag{79}$$

where the integral over $q(z)$ can be expressed through the substitutions

$$\begin{aligned} z^*_{t,k} &= \gamma_{t,k} \\ (z_{t-1,k} z_{t,l})^* &= \xi_{t,k,l} \end{aligned}$$

Note also that the following decomposition for $q(\theta)$ holds without further need for approximation:

$$q(\theta) = q(\mu, \lambda) q(A) q(\pi) \tag{80}$$

This in turn means we can write:

$$\begin{aligned} q(\mu, \lambda) = &p^*(x \mid z, \mu_n, \lambda_n) p(\mu, \lambda) \\ = &\prod_k \left[\prod_t p(x_t \mid \mu_k, \lambda_k)^{\gamma_{t,k}} \right] p(\mu_k, \lambda_k) \end{aligned} \tag{81}$$

$$q(A) = p^*(z_{2:T} \mid z_1, A) p(A) \tag{82}$$

$$q(\pi) = p^*(z_1 \mid \pi) \tag{83}$$

This means that the m-step reduces to calculation of a set of *variational* parameters w that determines $q(\theta|w)$ from the *hyperparameters* u that define $p(\theta|u)$ and the sufficient statistics for $p^*(x, z \mid \theta)$.

In order to calculated the updates for $q(\mu, \lambda \mid w)$ we will use the fact that the prior and likelihood are exponential family, so that they may be written as:

$$\begin{aligned}
q(\eta \mid \tilde{\nu}_n, \tilde{\chi}_n) =& h\left(\tilde{\nu_n}, \tilde{\chi_n}\right) g(\eta)^{\tilde{\nu}_n} \exp\left[\eta \cdot \tilde{\chi}_n\right] \\
=& Z_{q(\theta)}^{-1} f(x) g(\eta) \exp\left[\eta \cdot \upsilon(x_n)\right] \\
& h\left(\nu_n, \chi_n\right) g(\eta)^{\nu_n} \exp\left[\eta \cdot \chi_n\right] \\
=& \tilde{Z}_{q(\theta)}^{-1} g(\eta)^{\nu_n + 1} \exp\left[\eta \cdot (\chi_n + \upsilon(x_n))\right]
\end{aligned} \tag{84}$$

This allows us to rewrite equation (81) as:

$$q(\eta_k \mid \tilde{\nu}_k, \tilde{\chi}_k) = \tilde{Z}_{q(\theta)}^{-1} \left[\prod_t \left(g(\eta_k) \exp\left[\eta_k \cdot \upsilon(x_t)\right]\right)^{\gamma_{t,k}}\right] \tag{85}$$

$$g(\eta_k) \exp\left[\eta_k \cdot \chi_k\right] \tag{86}$$

which yields the updates

$$\tilde{\nu}_k = \nu_k + \sum_t \gamma_{t,k} \tag{87}$$

$$\tilde{\chi}_k = \chi_k + \sum_t \gamma_{t,k} \upsilon(x_t) \tag{88}$$

We can now substitute

$$\nu = u^\beta = u^\nu - 1 \tag{89}$$

$$\chi = \{-\tfrac{1}{2}(\nu(u^\mu)^2 + 1/u^W)\,,\ \nu u^\mu\} \tag{90}$$

$$\upsilon(x) = \{-\tfrac{1}{2}x^2, x\} \tag{91}$$

and define

$$N_k = \sum_t \gamma_{t,k} \tag{92}$$

$$\bar{X}_k = \sum_t \gamma_{t,k} x_t \tag{93}$$

$$\bar{X^2}_k = \sum_t \gamma_{t,k} x_t^2 \tag{94}$$

to obtain the following expressions for the variational parameters $q(\theta \mid w)$:

$$w_k^\mu = \tilde{\chi}_{k,2}/\tilde{\nu}_k = (u_k^\beta u_k^\mu + \bar{X}_k)/(u_k^\beta + N_k) \tag{95}$$

$$w_k^\beta = u_k^\beta + N_k \tag{96}$$

$$w_k^\nu = u_k^\nu + N_k \tag{97}$$

$$\begin{aligned}
(w_k^W)^{-1} =& -\tilde{\chi}_{k,2} - \tfrac{1}{2}\tilde{\chi}_{k,2}^2/\tilde{\nu}_k \\
=& \left((u_k^W)^{-1} + u_k^\beta (u_k^\mu)^2 + \bar{X^2}_k\right) \\
& - \frac{1}{2}\left((u_k^\beta u_k^\mu + \bar{X}_k)^2/(u_k^\beta + N_k)\right)
\end{aligned} \tag{98}$$

Finally, the updates for u^A and u^π can be obtained by substitution of the terms in equation (78) into equations (82) and (83):

$$w^A_{kl} = u^A_{kl} + \sum_{t=2}^{T} \xi_{t,kl} \tag{99}$$

$$w^\pi_k = u^\pi_k + \gamma_{1,k} \tag{100}$$

We now proceed to derive how γ and ξ can be calculated using the Forward-backward algorithm.

3.4 Forward-Backward Algorithm

The forward-backward algorithm is a method to calculate expectation values under the posterior $p(z|x,\theta)$, or in our case, the approximate posterior $q(z)$ of a Hidden Markov Model:

$$\gamma_{t,k} = E_{q(z)}[z_{t,k}] = p^*(x_1 \mid z_1)p^*(z_1) \tag{101}$$

$$\xi_{t,kl} = E_{q(z)}[z_{t-1,k} z_{t,l}] = p^*(z_{t-1} = k, z_{t-1} = l \mid x_{1:T}) \tag{102}$$

to do so we calculate two variables:

$$\alpha_{t,k} = p^*(x_{1:t}, z_t = k) \tag{103}$$

$$\beta_{t,k} = p^*(z_t = k \mid x_{t+1:T}) \tag{104}$$

such that

$$\gamma_{t,k} = p^*(z_t = k \mid x_{1:T}) = \frac{\alpha_{t,k}\beta_{t,k}}{p^*(x_{1:T})} \tag{105}$$

$$\xi_{t,k,l} = p^*(z_{t-1} = k, z_{t-1} = l \mid x_{1:T}) \tag{106}$$

$$= \frac{p^*(x_{1:T} \mid z_t, z_{t-1})p^*(z_t, z_{t-1})}{p^*(x_{1:T})} = \frac{\beta_{t,l} p^*(x_t \mid z_t = l) A_{kl} \alpha_{t-1,k}}{p^*(x_{1:T})} \tag{107}$$

and exploit the following recursive relationships:

$$\begin{aligned} \alpha_{t,k} &= p^*(x_{1:t}, z_t) \\ &= \sum_l p^*(x_t \mid z_t = k) p^*(z_t = k \mid z_{t-1} = l) p^*(x_{1:t-1}, z_{t-1} = l) \\ &= \sum_l p^*(x_t \mid z_t = k) A^*_{lk} \alpha_{t-1,l} \end{aligned} \tag{108}$$

$$\begin{aligned} \beta_{t,k} &= p^*(x_{t+1:T} \mid z_t) \\ &= \sum_l p^*(x_{t+2:T} \mid z_{t+1} = l) p^*(x_{t+1} \mid z_{t+1} = l) p^*(z_{t+1} = l \mid z_t = k) \\ &= \sum_l \beta_{t+1,l} p^*(x_{t+1} \mid z_{t+1} = l) A^*_{kl} \end{aligned} \tag{109}$$

We can now loop *forward* in time to recursively calculate α_t from α_{t-1} and backward in time to calculate β_t from β_{t+1}. The boundary conditions on these passes are:

$$\alpha_{1,k} = p^*(x_1, z_1) = p^*(x_1 \mid z_1)p^*(z_1) = \prod_k p^*(x_1 \mid z_1 = k)\pi_k^* \tag{110}$$

$$\beta_{T,k} = 1 \tag{111}$$

In practice, it proves more convenient to calculate a normalized version of $\hat{\alpha}$ and $\hat{\beta}$. To do so, we introduce a set of scaling factors c_t:

$$c_t = p^*(x_t \mid x_{1:t-1}) \tag{112}$$

such that normalized forward and backward variables can be defined as:

$$\begin{aligned}
\hat{\alpha}_{t,k} &= \frac{\alpha_{t,k}}{p^*(x_{1:t})} = \prod_{t'=1}^{t} \frac{1}{c_{t'}} \alpha_{t,k} \\
\hat{\beta}_{t,k} &= \frac{\beta_{t,k}}{p^*(x_{t+1:T} \mid x_{1:t})} = \prod_{t'=t+1}^{T} \frac{1}{c_{t'}} \beta_{t,k}
\end{aligned} \tag{113}$$

This choice of normalization implies:

$$\gamma_{t,k} = \frac{\alpha_{t,k}\beta_{t,k}}{p^*(x_{1:T})} = \frac{\alpha_{t,k}\beta_{t,k}}{p^*(x_{t+1:T} \mid x_{1:t})p^*(x_{1:t})} = \hat{\alpha}_{t,k}\hat{\beta}_{t,k} \tag{114}$$

$$\xi_{t,k,l} = \frac{\beta_{t,l}p^*(x_t \mid z_t = l)A_{kl}\alpha_{t-1,k}}{p^*(x_{1:T})} = \frac{c_t\hat{\beta}_{t,l}p^*(x_t \mid z_t = l)A_{kl}\hat{\alpha}_{t-1,k}}{p^*(x_{1:T})} \tag{115}$$

The following recursion relations hold for $\hat{\alpha}$ and $\hat{\beta}$:

$$c_t\hat{\alpha}_{t,k} = \sum_l p^*(x_t \mid z_t = k)A^*_{lk}\hat{\alpha}_{t-1,l} \tag{116}$$

$$c_{t+1}\beta_{t,k} = \sum_l \hat{\beta}_{t+1,l}p^*(x_{t+1} \mid z_{t+1} = l)A^*_{kl} \tag{117}$$

We can now solve for c_t from the recursion relation for $\hat{\alpha}$ using that $\sum_k \hat{\alpha}_{t,k} = 1$:

$$c_t = c_t \sum_k \hat{\alpha}_{t,k} = \sum_{k,l} p^*(x_t \mid z_t = k)A^*_{lk}\alpha_{t-1,l} \tag{118}$$

So the scale factors c_t are nothing but the normalization constant for $\hat{\alpha}_t$ and can therefore essentially be obtained for free during the forward pass. Note that these also give us an estimate for $p^*(x)$:

$$p^*(x) = p^*(x_{1:t}) = \prod_t c_t \tag{119}$$

which gives us the normalization constant for $q(z)$

$$\hat{Z}_{q(z)} = \ln p^*(x) = \sum_t \ln c_t \tag{120}$$

3.5 Calculation of the Evidence

The last thing that remains is to calculate the lower bound so we can check for convergence.

$$\mathcal{L}[q(z), q(\theta)] = \sum \int d\theta \sum_z q(z)q(\theta) \ln \left[\frac{p(x, z, \theta)}{q(z)q(\theta)} \right] \tag{121}$$

We can decompose the terms in this equation as:

$$\begin{aligned} \mathcal{L}[q(z), q(\theta)] = \sum & E_{q(z)q(\theta)} \left[\ln p(x, z \mid \theta) \right] \\ & + D_{KL}[q(\theta)|p(\theta)] - E_{q(z)} \left[\ln q(z) \right] \end{aligned} \tag{122}$$

Now note from equation (54) that $E_{q(z)} \left[\ln q(z) \right]$ can be written as:

$$E_{q(z)} \left[\ln q(z) \right] = E_{q(z)q(\theta)}[\ln p(x, z \mid \theta)] - \ln \hat{Z}_{q(z)} \tag{123}$$

So

$$\mathcal{L}[q(z), q(\theta)] = \sum \ln \hat{Z}_{q(z)} + D_{KL}[q(\theta) \,||\, p(\theta)] \tag{124}$$

The term $\ln \hat{Z}_{q(z)}$ is obtained from the forward backward algorithm. The Kullback-Leibler divergence between $q(\theta)$ and $p(\theta)$ decomposes into:

$$\begin{aligned} D_{KL}[q(\theta) \,||\, p(\theta)] = \sum_k & D_{KL}[q(\mu_k, \lambda_k) \,||\, p(\mu_k, \lambda_k)] \\ & + D_{KL}[q(A) \,||\, p(A)] + D_{KL}[q(\pi) \,||\, p(\pi)] \end{aligned} \tag{125}$$

The expression for the D_{KL} of a Gaussian-Wishart distribution is a bit painful, but can be obtained from Bishop equations (10.74) and (10.77).

$$D_{KL}[q(\mu_k, \lambda_k) \,||\, p(\mu_k, \lambda_k)] = E_{q(\mu_k,\lambda_k)}[p(\mu_k, \lambda_k)] - E_{q(\mu_k,\lambda_k)}[q(\mu_k, \lambda_k)] \tag{126}$$

which expands to

$$E_{q(\mu_k,\lambda_k)}[p(\mu_k, \lambda_k)] = \frac{1}{2} \left[\ln \left(\frac{u_k^\beta}{2\pi} \right) + \ln \lambda_k^* - \frac{u_k^\beta}{w_k^\beta} + u_k^\beta u_k^\nu w_k^W (w_k^\mu - u_k^\mu)^2 \right] \tag{127}$$

$$E_{q(\mu_k,\lambda_k)}[q(\mu_k, \lambda_k)] = \frac{1}{2} \left[\ln \left(\frac{u_k^\beta}{2\pi} \right) + \ln \lambda_k^* - 1 - H[\lambda_k] \right] \tag{128}$$

with

$$H[\lambda_k] = -\ln \left[(2/u_k^W)^{u_k^\nu/2} \Gamma(w_k^\nu/2) \right] - \frac{w_k^\nu - 2}{2} \ln w_k^\nu w_k^W + \frac{w^\nu}{2} \tag{129}$$

The KL divergences for A and π have simple closed-form expressions:

$$D_{KL}[q(A_k) \,||\, p(A_k)] = \sum_l [w_{k,l}^A - u_{k,l}^a][\psi(w_{k,l}^A) - \psi(u_{k,l}^A)] \tag{130}$$

$$D_{KL}[q(\pi) \,||\, p(\pi)] = \sum_l [w_l^\pi - u_l^\pi][\psi(w_l^\pi) - \psi(u_l^\pi)] \tag{131}$$

4 Hierarchical Updates (Empirical Bayes)

In the hierarchical step we maximize the summed lower bound log-evidence with respect to the ensemble distribution $p(\theta \,|\, u)$. This step can be understood as a type of Emperical Bayes method.

In the more general case of Empirical Bayes, one would introduce a prior $p(u)$, and run alternating variational updates to find (approximations) of the posteriors $p(\theta \mid x)$ and $p(u \mid x)$:

$$p(z, \theta \mid x) = \frac{p(x \mid z, \theta)}{p(x)} \int du \, p(z \mid \theta) p(\theta \mid u) p(u) \tag{132}$$

$$p(u \mid x) = \frac{p(u)}{p(x)} \sum_z \int d\theta \, p(x \mid z, \theta) p(z \mid \theta) p(\theta \mid u) \tag{133}$$

Of course, calculation of the hierarchical generalization of the evidence $p(x)$ would require an additional integral:

$$p(x) = \int du \, p(x \mid u) p(u) \tag{134}$$

One could now in principle attempt to construct a variational approach in terms of 3 distributions $q(z), q(\theta), q(u)$, that minimizes a lower bound on the log hierarchical evidence $\log p(x)$. However, this would be a lot of pain, for not so much gain.

A simpler approach is to construct an EM algorithm to obtain a point estimate for u. The quantity optimized is the summed lower bound log evidence over the ensemble:

$$\log p(x \mid u) \simeq \sum_n \mathcal{L}_n \tag{135}$$

The E-step amounts to running VBEM on every trace to construct:

$$q(\theta \mid w) = \prod_n q(\theta \mid w_n) \simeq p(\theta \mid x, u) \tag{136}$$

Whereas the M-step maximizes the summed lower bound w.r.t. u:

$$0 = \frac{\partial}{\partial u} \sum_n \mathcal{L}_n \tag{137}$$

$$= \frac{\partial}{\partial u} \sum_n \int d\theta_n \, q(\theta_n \mid w_n) \log p(\theta_n \mid u) \tag{138}$$

$$= \sum_n \int d\theta_n \, q(\theta_n \mid w_n) \frac{\partial_u p(\theta_n \mid u)}{p(\theta_n \mid u)} \tag{139}$$

Now note that $p(\theta)$ factorizes without need for further approximation

$$p(\theta \mid u) = p(\mu, \lambda \mid u^\mu, u^\beta, u^W, u^\nu) p(A \mid u^A) p(\pi \mid u^\pi) \tag{140}$$

so the updates for each factor can be computed separatedly.

4.1 Conjugate-Exponential Form

If we rewrite $p(\theta \,|\, u)$ to its conjugate exponential form $p(\eta \,|\, \nu, \chi)$, the expression in equation 139 takes the form:

$$0 = \sum_n \int d\eta_n \, q(\eta_n \mid \nu_n, \chi_n) \frac{\partial_{\nu,\chi} p(\eta_n \mid \nu, \chi)}{p(\eta_n \mid \nu, \chi)} \tag{141}$$

Here we adopt the convention where $\{\nu, \chi\}$ are taken to signify the hyperparameters of the ensemble distribution, whereas $\{\nu_n, \chi_n\}$ denotes the set of variational parameters for the approximate posterior of each trace.

The derivatives of $p(\eta \,|\, \nu, \chi)$ with respect to the hyperparameters are given by:

$$\frac{\partial p(\eta \mid \nu, \chi)}{\partial \nu} = \left[\frac{\partial_\nu h(\nu, \chi)}{h(\nu, \chi)} + \ln g(\eta) \right] p(\eta \mid \nu, \chi) \tag{142}$$

$$\nabla_\chi p(\eta \mid \nu, \chi) = \left[\frac{\nabla_\chi h(\nu, \chi)}{h(\nu, \chi)} + \eta \right] p(\eta \mid \nu, \chi) \tag{143}$$

If we now substitute these expressions in equation 141, we obtain the expressions:

$$0 = \frac{\partial}{\partial \nu} \sum_n \mathcal{L}_n = \sum_n E_{q(\eta_n)} \left[\frac{\partial_\nu h(\nu, \chi)}{h(\nu, \chi)} + \ln g(\eta) \right] = \tag{144}$$

Given that terms containing $h(\nu, \chi)$ have no dependence on η we can rewrite these equalities as:

$$E_{q(\eta)} \left[\ln g(\eta) \right] = \frac{1}{N} \sum_n E_{q(\eta_n)} \left[\ln g(\eta) \right] \tag{145}$$

$$= -\frac{\partial_\nu h(\nu, \chi)}{h(\nu, \chi)} \tag{146}$$

$$E_{q(\eta)} \left[\eta \right] = \frac{1}{N} \sum_n E_{q(\eta_n)} \left[\eta \right] \tag{147}$$

$$= -\frac{\nabla_\chi h(\nu, \chi)}{h(\nu, \chi)} \tag{148}$$

These equations implicitly specify the update conditions for the hyperparameters in terms of the averaged expectation values of η and $\ln g(\eta)$ under the approximate posterior for each trace.

The expectation values for $\ln g(\eta)$ and η can be computed by noting that the integral of a probability density function must always equal to 1, implying that it's derivatives w.r.t. ν and χ must be zero:

$$0 = \frac{\partial}{\partial \nu_n} \int d\eta_n \, q(\eta_n \mid \nu_n, \chi_n) = \frac{\partial_{\nu_n} h(\nu_n, \chi_n)}{h(\nu_n, \chi_n)} + E_{q(\theta_n)}[\ln g(\eta_n)] \tag{149}$$

$$0 = \nabla_{\chi_n} \int d\eta_n \, q(\eta_n \mid \nu_n, \chi_n) = \frac{\nabla_{\chi_n} h(\nu_n, \chi_n)}{h(\nu_n, \chi_n)} + E_{q(\theta_n)}[\eta_n] \tag{150}$$

So the logarithmic derivates of $h(\nu, \chi)$ in fact gives us the required expectation values, and the equations for the hyperparameter updates are in fact equivalent to the expressions:

$$E_{p(\eta)}\left[\ln g(\eta)\right] = E_{q(\eta)}\left[\ln g(\eta)\right] \tag{151}$$
$$E_{p(\eta)}\left[\eta\right] = E_{q(\eta)}\left[\eta\right] \tag{152}$$

or

$$\frac{\partial_\nu h(\nu,\chi)}{h(\nu,\chi)} = \frac{1}{N}\sum_n \frac{\partial_{\nu_n} h(\nu_n,\chi_n)}{h(\nu_n,\chi_n)} \tag{153}$$
$$\frac{\nabla_\chi h(\nu,\chi)}{h(\nu,\chi)} = \frac{1}{N}\sum_n \frac{\nabla_{\chi_n} h(\nu_n,\chi_n)}{h(\nu_n,\chi_n)} \tag{154}$$

4.2 Emission Distribution (Normal-Wishart)

For a 1-dimensional Normal-Wishart distribution the conjugate-exponential representation (section 2.2) takes the form:

$$\eta = \{\lambda, \lambda\mu\} \tag{155}$$
$$\nu = u^\beta = u^\nu - 1 \tag{156}$$
$$\chi = \{-\tfrac{1}{2}(u^\beta(u^\mu)^2 + 1/u^W), u^\beta u^\mu\} \tag{157}$$
$$g(\eta) = (\eta_1/2\pi)^{1/2}\exp[-\eta_2^2/2\eta_1] \tag{158}$$
$$h(\nu,\chi) = (2\pi)^{(\nu-1)/2}\nu^{1/2}(-\chi_1 - \chi_2^2/\nu)^{(\nu+1)/2} \tag{159}$$

The expressions for the expectation values of $\ln g$ and η become:

$$E_{q(\theta_n)}[\ln g] = -\frac{1}{2}[1/w_n^\beta + w_n^\nu w_n^W (w_n^\mu)^2 + \log(\pi/w_n^W) - \psi(w_n^\nu/2)] \tag{160}$$
$$E_{q(\theta_n)}[\lambda] = w_n^\nu w_n^W \tag{161}$$
$$E_{q(\theta_n)}[\lambda\mu] = w_n^\nu w_n^W w_n^\mu \tag{162}$$

We now obtain the following updates for u^μ and u^W:

$$u^\mu = E_{q(\theta)}[\lambda\mu]/E_{q(\theta)}[\lambda] \tag{163}$$
$$u^W = E_{q(\theta)}[\lambda]/u^\nu \tag{164}$$
$$\tag{165}$$

And an implicit expression for for $\nu = u^\beta = u^\nu - 1$ that must be solved numerically:

$$-\frac{1}{2}\left[\frac{1}{u^\nu - 1} + \frac{E_{q(\theta)}[\lambda\mu]^2}{E_{q(\theta)}[\lambda]} + \log\left(\frac{\pi u^\nu}{E_{q(\theta)}[\lambda]}\right) - \psi(u^\nu/2)\right] = \frac{1}{N}\sum_n E_{q(\theta_n)}[\ln g] \tag{166}$$

4.3 Inital State and Transition Probabilities (Dirichlet)

For a Dirichlet distribution the conjugate exponential forms (section 2.3) are given by:

$$\eta = \{\ln \pi_k\} \tag{167}$$
$$\chi = \{u_k^\pi\} \tag{168}$$
$$h(\chi) = \frac{\prod_k \Gamma(\chi_k + 1)}{\Gamma(\sum_k(\chi_k + 1))} \tag{169}$$

And the log expectation value of η is:

$$E_{q(\theta_n)}[\eta] = E_{q(\theta_n)}[\ln \pi] = \psi(w^{\pi}_{n,k}) - \psi(\sum_k w^{\pi}_{n,k}) \tag{170}$$

which again leads to a coupled set of implicit equations that must be solved numerically:

$$\psi(u^{\pi}_k) - \psi(\sum_k u^{\pi}_k) = \frac{1}{N}\sum_n \psi(w^{\pi}_{n,k}) - \psi(\sum_k w^{\pi}_{n,k}) \tag{171}$$

The updates for each row of the transition matrix are performed in the same manner

$$\psi(u^{A}_{kl}) - \psi(\sum_l u^{A}_{kl}) = \frac{1}{N}\sum_n \psi(w^{A}_{n,kl}) - \psi(\sum_l w^{A}_{n,kl}) \tag{172}$$

4.4 Mixtures of Priors

The empirical Bayes approach admits a straightforward generalization to inference over mixtures of ensemble distributions. Suppose that a latent state $y = 1 \ldots M$ encodes the membership of each trace with respect to a set of M sub-populations in the ensemble, which have different parameter distributions $p(\theta|u_m \mid)$. The evidence can now be expressed as a marginal over y:

$$\begin{aligned} p(x \mid u) &= \sum_y p(x \mid u, y) p(y) && (173)\\ &= \sum_m p(x \mid u_m) p(y = m) && (174)\\ &\geq \sum_{n,m} \exp(\mathcal{L}_{nm}) p(y = m) && (175) \end{aligned}$$

where $\mathcal{L}_{nm} \geq \log p(x_n \mid u_m)$ is the lower bound log evidence for trace n with respect to mixture component m.

An expectation maximization algorithm over this mixture can be now be constructed by introducing a variational posterior $q(y_n\!=\!m) = \omega_{nm}$ for each trace. The corresponding (approximate) E-step is now given by:

$$q^{(i+1)}(y_n\!=\!m) = \frac{\exp(\mathcal{L}_{nm}) p^{(i)}(y\!=\!m)}{\sum_l \exp(\mathcal{L}_{nl}) p^{(i)}(y\!=\!l)} = \omega^{(i+1)}_{nm} \tag{176}$$

And the M-step simply becomes:

$$p^{(i+1)}(y = m) = \frac{1}{N}\sum_n q^{(i+1)}(y_n = m) \tag{177}$$

The mixed version of the hierarchical algorithm now maximizes the lower bound

$$\begin{aligned} \log p(x \mid u) &\geq \sum_n \sum_{y_n} q(y_n) \log \left[\frac{p(x_n \mid u, y_n)}{q(y_n)}\right] && (178)\\ &\geq \sum_{n,m} \omega_{nm} \left[\mathcal{L}_{nm} - \log \omega_{nm}\right] && (179)\\ &= \mathcal{L} && (180) \end{aligned}$$

and the hierarchical update for the m-th subpopulation becomes equivalent solving of the equations

$$0 = \sum_{n,m} \omega_{nm} \frac{\partial \mathcal{L}_{nm}}{\partial u_m} \tag{181}$$

which produces a set of update equations analogous to the single-population case, where the expectation values with respect to the approximate posteriors are now weighted by ω:

$$E_{p(\theta \mid u_m)}[\ln g] = \frac{1}{\sum_n \omega_{nm}} \sum_n \omega_{nm} E_{q(\theta_n \mid w_{nm})}[\ln g] \tag{182}$$

$$E_{p(\theta \mid u_m)}[\eta] = \frac{1}{\sum_n \omega_{nm}} \sum_n \omega_{nm} E_{q(\theta_n \mid w_{nm})}[\eta] \tag{183}$$