# Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments

Jan-Willem van de Meent,[†] Jonathan E. Bronson,[‡] Chris H. Wiggins,[§] and Ruben L. Gonzalez Jr.[‡]

[†]Department of Statistics, [‡]Department of Chemistry, and [§]Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York

# Supplementary material

# Contents


**1 Generative Model for Coupled HMMs** **2**
1.1 Variable Definitions . . . 2
1.2 Evidence . . . 2
1.3 Likelihood . . . 2
1.4 Emissions model . . . 3
1.5 Transition probabilities (HMM) . . . 3
1.6 Ensemble Distributions (Priors) . . . 3
1.7 Evidence Lower Bound (ELBO) . . . 3
1.8 Algorithm Outline . . . 3

**2 Conjugate-Exponential Forms** **4**
2.1 Normal-Gamma . . . 4
2.2 Dirichlet . . . 5

**3 Variational Bayes Expectation Maximization (VBEM)** **5**
3.1 Updates . . . 6
3.2 E-step . . . 6
3.3 M-step . . . 8
3.4 Forward-Backward Algorithm . . . 10
3.5 Calculation of the Evidence . . . 11

**4 Empirical Bayes Updates** **12**
4.1 Conjugate-Exponential Form . . . 13
4.2 State Distributions (Dirichlet) . . . 13
4.3 Emission Distribution (Normal-Gamma) . . . 13
4.4 Inital State and Transition Probabilities (Dirichlet) . . . 14
4.5 Mixtures of Priors . . . 14

**5 Calculation of Derivative Kinetic Quantities** **15**
5.1 Kinetic Rates . . . 15
5.2 Life Time . . . 15
5.3 Free Energy . . . 15

# S1 Generative Model for Coupled HMMs

## S1.1 Variable Definitions

Observation in time series $n \in \{1 \dots N\}$ at time $t \in \{1 \dots T_n\}$

$$x = \{x_n\} = \{\{x_{n,t}\}\}$$

State of molecule $n$ at time $t$

$$z = \{z_n\} = \{\{z_{n,t}\}\}$$

Parameters for time series $n$

$$\theta = \{\theta_n\} = \{\pi_n, A_n, \mu_n, \lambda_n\}$$

Initial probabilities: Prob that time series $n$ starts in state $k$

$$\pi_n = \{\pi_{n,k}\}$$

Transition matrix: Prob of moving from state $k$ to state $l$

$$A_n = \{\{A_{n,kl}\}\}$$

$E_{\text{FRET}}$ observation mean for state $k$ in time series $n$

$$mu_n = \{\mu_{n,k}\}$$

$E_{\text{FRET}}$ emissions precision (1/var) for state $k$ in time series $n$

$$\lambda_n = \{\lambda_{n,k}\}$$

Hyperparameters for prior

$$\psi_0 = \{\{m_{0,k}, \beta_{0,k}, a_{0,k}, b_{0,k}\}, \{\alpha_{0,k}\}, \{\rho_0\}\}$$

Variational parameters for posterior of time series $n$

$$\psi_n = \{\{m_{n,k}, \beta_{n,k}, a_{n,k}, b_{n,k}\}, \{\alpha_{n,k}\}, \{\rho_n\}\}$$

## S1.2 Evidence

$$\begin{aligned} p(x \mid \psi_0) &= \int d\theta \, p(x, \theta \mid \psi_0) \\ &= \int d\theta \, p(x \mid \theta) p(\theta \mid \psi_0) \\ &= \int d\theta \prod_n p(x_n \mid \theta_n) p(\theta_n \mid \psi_0) \\ &= \prod_n \int d\theta_n \, p(x_n \mid \theta_n) p(\theta_n \mid \psi_0) \end{aligned} \tag{1}$$

## S1.3 Likelihood

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

### S1.4 Emissions model

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

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

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

### S1.5 Transition probabilities (HMM)

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

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

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

### S1.6 Ensemble Distributions (Priors)

$$
\begin{aligned}
p(\theta_n \mid \psi_0) &= p(\pi_n \mid \psi_0) p(A_n \mid \psi_0) p(\mu_n, \lambda_n \mid \psi_0) \\
&= p(\pi_n \mid \psi_0) \prod_k p(A_{n,k} \mid \psi_0) p(\mu_{n,k}, \lambda_{n,k} \mid \psi_0)
\end{aligned} \tag{9}
$$

$$
\pi_n \sim \mathrm{Dir}(\rho_0) \tag{10}
$$

$$
A_{n,k} \sim \mathrm{Dir}(\alpha_{0\,k}) \tag{11}
$$

$$
\lambda_{n,k} \sim \mathrm{Gamma}(a_{0\,k}, b_{0\,k}) \tag{12}
$$

$$
\mu_{n,k} \sim \mathrm{N}(m_{0\,k}, \beta_{0\,k}\lambda_{n,k}) \tag{13}
$$

### S1.7 Evidence Lower Bound (ELBO)

$$
L_n\big[q(z_n), q(\theta_n), \psi_0\big] = \int d\theta_n \sum_{z_n} q(z_n) q(\theta_n) \ln\left[\frac{p(x_n, z_n, \theta_n \mid \psi_0)}{q(z_n) q(\theta_n)}\right] \tag{14}
$$

### S1.8 Algorithm Outline

Loop over iterations $i$ until $\sum_n L_n$ converges:

1. VB updates: obtain $q^{(i)}(\theta_n)$, $q^{(i)}(z_n)$, and $L_n^{(i)}$ for each trace $n$, holding the prior $p^{(i-1)}(\theta_n \mid \psi_0)$ constant.

2. Empirical bayes updates: Holding $q(z_n)$ and $q(\theta_n)$ constant, solve for

$$
\psi_0 = \arg\max_{\psi_0} \sum_n L_n^{(i)}[q^{(i)}(z_n), q^{(i)}(\theta_n), \psi_0]
$$

As we will show, the variational posterior has the same analytical form as the prior $q(\theta_n) = p(\theta_n \mid \psi_n)$ and its updates correspond to calculating a set of variational parameters $\psi_n$. Calculation of $\psi_n$ only requires knowledge of two sets of expectation values $\gamma_{n,t,k} = E_{q(z_n)}[z_{n,t,k}]$ and $\xi_{n,t,k,l} = E_{q(z_n)}[z_{n,t+1,l} z_{n,t,k}]$, which can be calculated with a forward-backward algorithm where expectation values of $\exp(E_{q(\theta_n)}[\ln \theta_n])$ are substituted for the parameters. The empirical Bayes updates for $\psi_0$ can be calculated in terms of expectation values on $q(\theta_n)$.

## S2 Conjugate-Exponential Forms

In a model where the prior and likelihood are in the exponential family, it is possible to paramaterize these distributions as

$$p(x \mid \eta) = \exp[\eta \cdot \mathcal{T}(x) - A(\eta) + B(x)] \,, \tag{15}$$

$$p(\eta \mid \nu_0, \chi_0, \phi_0) = \exp\left[\eta \cdot \chi_0 - \nu_0 A(\eta) - A(\nu_0, \chi_0, \phi_0) + B(\eta, \phi_0)\right] . \tag{16}$$

Here $\eta$ represents the remapped parameters $\theta$, and $\{\nu_0, \chi_0, \phi_0\}$ represent the remapped hyperparameters $\psi_0$. The functions $A$ are sometimes known as log-normalizers, whereas the functions $B$ can be seen as log base measure. As with parameter distributions, where $p(\eta)$ is used to represent a different distribution than $p(x)$, we here employ the convention that the log normalizers $A(\eta)$ and $A(\nu_0, \chi_0, \phi_0)$, as well as the log base measures $B(x)$ and $B(\eta, \phi_0)$, take unique forms for each set of parameters.

Given this parameterization, the posterior $p(\eta \mid x, \nu_0, \chi_0, \phi_0)$ is now proportional to

$$p(\eta \mid x, \nu_0, \chi_0, \phi_0) \propto p(x \mid \eta) p(\eta \mid \nu_0, \chi_0, \phi_0) \tag{17}$$

$$= \exp[\eta \cdot (\chi_0 + \mathcal{T}(x)) - (\nu_0 + 1) A(\eta)] / Z(x, \nu_0, \chi_0, \phi_0) \tag{18}$$

In other words, the posterior has the same analytical form as prior

$$p(\eta \mid x, \nu_0, \chi_0, \phi_0) = p(\eta \mid \nu, \chi, \phi_0) \tag{19}$$

with an updated set of hyperparameters

$$\nu = \nu_0 + 1 \tag{20}$$

$$\chi = \chi_0 + \mathcal{T}(x) \tag{21}$$

We see that the hyperparameter $\nu$ can be interpreted as scale factor that tracks the number of previously observed samples. The hyperparameter vector $\chi$ in turn takes the role of the aggregate sufficient statistics $\mathcal{T}$ associated with these samples.

In any pair of conjugate distributions $\eta$, $\chi$ and $\mathcal{T}(x)$ must have the same dimensionality. This means that if $\eta$ has $D$ components, the hyperparameters $\{\nu, \chi\}$ have dimensionality $D$+1. In general a prior distribution need not have $D + 1$ parameters. For example, the Dirichlet distribution lacks $\nu_0$ and $\phi_0$ parameters. For a Normal-gamma prior $p(\mu, \lambda | m_0, \beta_0, a_0, b_0)$, 4 hyperparameters encode a distribution on 2 variables. In this case an extra hyperparameter $\phi_0$, which can be thought of as the difference in number of initial observations for the precision and mean, remains invariant in light of new data.

Our derivation of the EB estimation algorithm on coupled HMMs will assume that the prior and likelihood are conjugate exponential family. This means the approach derived here could be adapted to model any experiment where the measurement signal can be represented with an exponential family likelihood, though the corresponding updates for posterior parameters and hyperparameters would have to be re-derived.

### S2.1 Normal-Gamma

This Normal-Gamma distribution is a joint prior on the mean and precision of a Gaussian likelihood, where the aggregate statistics for the mean are equivalent to $\nu$ observations and the statistics for the precision are equivalent to $\nu + \phi$ observations.

$$p(x \mid \mu, \lambda) = N(x \mid \mu, \lambda) \tag{22}$$

$$p(\mu, \lambda \mid m, \beta, a, b) = \text{Norm}(\mu \mid m, \beta\lambda) \text{Gamma}(\lambda \mid a, b) \tag{23}$$

$$\eta = \{-\tfrac{1}{2}\lambda, \lambda\mu\} \tag{24}$$

$$\nu = \beta \tag{25}$$

$$\chi = \{2b + \beta m^2 \,, \beta m\} \tag{26}$$

$$\phi = 2a - \beta \tag{27}$$

$$
\begin{aligned}
\mathcal{T}(x) &= \{x^2, x\} && (28)\\
A(\eta) &= -\frac{1}{2}\left[\ln\left(-\eta_1 \,/\, \pi\right) + \eta_2^2 \,/\, (2\eta_1)\right] && (29)\\
B(\eta, \phi) &= -\frac{1}{2}(\phi + 1)\ln\left(-\eta_1 \,/\, \pi\right) && (30)\\
A(\nu, \chi, \phi) &= -\frac{1}{2}\Big[\ln(\nu) + (\nu + \phi - 2)\ln(2\pi) \\
&\qquad + (\nu + \phi)\ln\left[\tfrac{1}{2}(\chi_1 - \chi_2^2/\nu)\right] - 2\ln\Gamma\left[\tfrac{1}{2}(\nu + \phi)\right]\Big] && (31)\\
&= -\frac{1}{2}\ln(\beta) - (a-1)\ln(2\pi) - a\ln(b) + \ln\Gamma(a)
\end{aligned}
$$

*Note:* a Normal-gamma distribution is equivalent to a 1-dimensional Normal-Wishart

$$
p(\mu, \lambda \mid m, \beta, W, \nu) = \mathrm{Norm}(\mu \mid m, \beta\lambda)\mathrm{Wish}(\lambda \mid W, \nu) \tag{32}
$$

with parameters $\nu = 2a$ and $W = 1/(2b)$.

### S2.2 Dirichlet

$$
\begin{aligned}
p(z \mid \pi) &= \mathrm{Cat}(z \mid \pi) = \prod_k \pi_k^{z_k} && (33)\\
p(\pi \mid \rho) &= \mathrm{Dir}(\pi \mid \rho) = \frac{\Gamma(\sum_k \rho_k)}{\prod_k \Gamma(\rho_k)} \prod_k \pi_k^{\rho_k - 1} && (34)
\end{aligned}
$$

$$
\begin{aligned}
\eta &= \{\ln \pi\} && (35)\\
\chi &= \{\rho\} && (36)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{T}(z) &= \{z\} && (37)\\
A(\eta) &= \eta && (38)\\
B(\eta) &= -\eta && (39)\\
A(\chi) &= \log\Gamma(\textstyle\sum_k \chi_k) - \sum_k \log \chi_k && (40)
\end{aligned}
$$

## S3 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 $p(z, \theta \mid x, \psi_0) \simeq q(z)q(\theta)$, that allows calculation of a lower bound on the log-evidence (using Jensen's inequality):

$$
\begin{aligned}
\ln p(x \mid \psi_0) &= \ln\left[\int d\theta \sum_z p(x, z, \theta \mid \psi_0)\right] \\
&= \ln\left[\int d\theta \sum_z q(z)q(\theta)\frac{p(x, z, \theta \mid \psi_0)}{q(z)q(\theta)}\right] \\
&\geq \int d\theta \sum_z q(z)q(\theta)\ln\left[\frac{p(x, z, \theta \mid \psi_0)}{q(z)q(\theta)}\right] \\
&= L[q(z), q(\theta)]
\end{aligned} \tag{41}
$$

The lower bound $L$ is tight if $q(z)q(\theta) = p(z, \theta \mid x, \psi_0)$:

$$
\begin{aligned}
L[q(z), q(\theta)] &= \int d\theta \sum_z q(z)q(\theta) \ln\left[\frac{p(x, z, \theta \mid \psi_0)}{q(z)q(\theta)}\right] \\
&= \int d\theta \sum_z p(z, \theta \mid x, \psi_0) \ln\left[\frac{p(x, z, \theta \mid \psi_0)}{p(z, \theta \mid x, \psi_0)}\right] \\
&= \int d\theta \sum_z p(z, \theta \mid x, \psi_0) \ln\left[\frac{p(z, \theta \mid x, \psi_0) p(x \mid \psi_0)}{p(z, \theta \mid x, \psi_0)}\right] \\
&= \int d\theta \sum_z p(z, \theta \mid x, \psi_0) \ln p(x \mid \psi_0) \\
&= \ln p(x \mid \psi_0) \int d\theta \sum_z p(z, \theta \mid x, \psi_0) \\
&= \ln p(x \mid \psi_0)
\end{aligned}
\tag{42}
$$

In general we can we write the lower bound in terms of the evidence $p(x \mid \psi_0)$ and a Kullback-Leibler divergence

$$
L[q(z), q(\theta)] = \ln p(x \mid \psi_0) - D_{\mathrm{KL}}\big[q(z)q(\theta) \,\|\, p(z, \theta \mid x, \psi_0)\big], \tag{43}
$$

which is defined as

$$
D_{\mathrm{KL}}\big[q(z)q(\theta) \,\|\, p(z, \theta \mid x, \psi_0)\big] = \int d\theta \sum_z q(z)q(\theta) \ln\left[\frac{q(z)q(\theta)}{p(z, \theta \mid x, \psi_0)}\right]. \tag{44}
$$

The $D_{\mathrm{KL}}$ term is $\geq 0$ and is 0 only when $q(z)q(\theta) = p(z, \theta \mid x, \psi_0)$ and $L = \ln p(x \mid \psi_0)$. We can use $q(z)$ and $q(\theta)$ to approximate $p(z, \theta | x, \psi_0)$ by minimizing the Kullback-Leibler divergence, which is equivalent to maximizing the lower bound $L$.

## S3.1 Updates

Loop until $L$ converges. For $i$-th iteration:

1. E-step: keeping $q^{(i)}(\theta)$ fixed, solve for

$$
q^{(i+1)}(z) = \arg\max_{q(z)} L[q(z), q^{(i)}(\theta)]
$$

2. M-step: keeping $q^{(i)}(z)$ fixed, solve for

$$
q^{(i+1)}(\theta) = \arg\max_{q(\theta)} L[q^{(i)}(z), q(\theta)]
$$

## S3.2 E-step

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

$$
\begin{aligned}
0 &= \nabla_{q(z)} \left[ L[q(z), q(\theta)] + \lambda_z \left(1 - \sum_{z'} q(z')\right)\right] \\
&= \left[\int d\theta \, q(\theta) \left(\ln p(x, z, \theta \mid \psi_0) - \ln q(z) - \ln q(\theta) - 1\right)\right] - \lambda_z
\end{aligned}
\tag{45}
$$

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

$$
\begin{aligned}
\ln q(z) &= \left[\int d\theta \, q(\theta) \left[\ln p(x, z \mid \theta) + \ln p(\theta \mid \psi_0) - \ln q(\theta) - 1\right]\right] - \lambda_z \\
&= E_{q(\theta)}[\ln p(x, z \mid \theta)] - D_{\mathrm{KL}}[q(\theta) \,\|\, p(\theta \mid \psi_0)] - (1 + \lambda_\theta) \\
&= E_{q(\theta)}[\ln p(x, z \mid \theta)] - \ln Z[q(\theta)]
\end{aligned}
\tag{46}
$$

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

$$q(z) = \frac{1}{Z[q(\theta)]} \exp\left[E_{q(\theta)}[\ln p(x, z \mid \theta)]\right] \tag{47}$$

where normalization of $q(z)$ implies

$$Z[q(\theta)] = \sum_z \exp\left[E_{q(\theta)}[\ln p(x, z \mid \theta)]\right] \tag{48}$$

The $z$-dependent terms can be written as:

$$\begin{aligned} E_{q(\theta)}[\ln p(x \mid z, \theta)] &= \sum_t \sum_k z_{t,k} \int d\theta\, q(\theta) \left[\tfrac{1}{2} \ln(\lambda_k \,/\, 2\pi) - \tfrac{1}{2}\Delta^2\right] \\ &= \sum_t z_t^\top \cdot E_{q(\theta)}\left[\tfrac{1}{2} \ln(\lambda_: \,/\, 2\pi) - \tfrac{1}{2}\Delta^2\right] \end{aligned} \tag{49}$$

and

$$\begin{aligned} E_{q(\theta)}[\ln p(z \mid \theta)] &= \sum_{t=2}^{T} \sum_{k,l} z_{t,l} z_{t-1,k} \int d\theta\, q(\theta) \ln A_{kl} \\ &\quad + \sum_k z_{1,k} \int d\theta\, q(\theta) \ln \pi_k \\ &= \sum_{t=2}^{T} z_{t-1}^\top \cdot E_{q(\theta)}[\ln A] \cdot z_t + z_t^\top \cdot E_{q(\theta)}[\ln \pi] \end{aligned} \tag{50}$$

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

$$q(z) = \frac{1}{Z[q(\theta)]} p^*(x, z) \tag{51}$$

with

$$p^*(x, z) = \left[\prod_t p^*(x_t \mid z_t)\right] p^*(z \mid \theta) \tag{52}$$

$$p^*(x_t \mid z_t = k) = (\lambda^*_k / 2\pi)^{1/2} \exp\left[-\tfrac{1}{2}\Delta^{*2}_{t,k}\right] \tag{53}$$

$$p^*(z \mid \theta) = p(z \mid A^*, \pi^*) \tag{54}$$

where point estimates for the parameters are defined as

$$\Delta^{*2} = E_{q(\theta)}[\Delta^2] \tag{55}$$

$$\lambda^* = \exp(E_{q(\theta)}[\ln \lambda]) \tag{56}$$

$$A^* = \exp(E_{q(\theta)}[\ln A]) \tag{57}$$

$$\pi^* = \exp(E_{q(\theta)}[\ln \pi]) \tag{58}$$

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 $\eta^*$ such that

$$q(z) = \frac{1}{Z[q(\eta)]} \exp[E_{q(\eta)}[\ln p(x, z, \eta)]] = \frac{1}{Z[q(\eta)]} p(x, z, \eta^*) \tag{59}$$

In our specific case, this result implies that we could in principle compute some $\eta^*$ 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:

$$\Delta^{*2} = (1/\beta_k) + a_k(x - m_k)^2/b_k \tag{60}$$

$$\ln \lambda^* = \Psi(a_k) - \ln b_k \tag{61}$$

$$\ln A^*_{k,l} = \Psi\left(\alpha_{k,l}\right) - \Psi\left(\textstyle\sum_l \alpha_{k,l}\right) \tag{62}$$

$$\ln \pi^*_k = \Psi\left(\rho_k\right) - \Psi\left(\textstyle\sum_k \rho_k\right) \tag{63}$$

Here $\Psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function.

In practice, we do not calculate $q(z)$ for all $K^T$ possible paths through the state space (which would be numerically unfeasible). Rather, we show in the next section that the updates for $q(\theta)$ only require knowledge of expectation values $\gamma_{tk} = E_{q(z)}[z_{t,k}]$ and $\xi_{tkl} = E_{q(z)}[z_{t-1,k} z_{t,l}]$, which can be calculated with a standard *forward-backward* algorithm.

### S3.3 M-step

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

$$0 = \frac{\delta}{\delta q(\theta)} \left[ L[q(z), q(\theta)] + \lambda_\theta \left(1 - \int d\theta'\, q(\theta')\right)\right] \tag{64}$$

$$= \left[\sum_z q(z)\left(\ln p(x, z, \theta \mid \psi_0) - \ln q(z) - \ln q(\theta) - 1\right)\right] - \lambda_\theta \tag{65}$$

like in the E-step, this reduces to

$$\ln q(\theta) = \left[\sum_z q(z)\left(\ln p(x, z, \theta \mid \psi_0) - \ln q(z) - 1\right)\right] - \lambda_\theta \tag{66}$$

$$= E_{q(z)}[\ln p(x, z, \theta \mid \psi_0)] - E_{q(z)}[\ln q(z)] - (1 + \lambda_\theta) \tag{67}$$

$$= E_{q(z)}[\ln p(x, z, \theta \mid \psi_0)] - \ln Z[q(z)] \tag{68}$$

Again we have absorbed all terms without a $\theta$ dependence into a normalization constant $Z[q(z)]$, which in fact does not need to be calculated explicitly in order to derive our updates. The expectation term expands to:

$$\begin{aligned} E_{q(z)}[\ln p(x, z, \theta \mid \psi_0)] = & E_{q(z)}[\ln p(x \mid z, \theta) + E_{q(z)}[\ln p(z \mid \theta)] \\ & + \ln p(\theta \mid \psi_0) \end{aligned} \tag{69}$$

where the $z$-dependent terms become:

$$E_{q(z)}[\ln p(x \mid z, \theta)] = \sum_t \sum_k E_{q(z)}[z_{t,k}] \left[\tfrac{1}{2}\ln\left(\lambda_k / 2\pi\right) - \tfrac{1}{2}\Delta^2_{t,k}\right] \tag{70}$$

$$\begin{aligned} E_{q(z)}[\ln p(z \mid \theta)] = & \sum_{t=2}^{T} \sum_{k,l} E_{q(z)}[z_{t,l} z_{t-1,k}] \ln A_{kl} \\ & + \sum_k E_{q(z)}[z_{1,k}] \ln \pi_k \end{aligned} \tag{71}$$

The variational posterior $q(\theta)$ is therefore parameterized in terms of two sets of expected posterior statistics:

$$\gamma_{t,k} = E_{q(z)}[z_{t,k}] \tag{72}$$

$$\xi_{t,kl} = E_{q(z)}[z_{t-1,k} z_{t,l}] \tag{73}$$

The expression for $q(\theta)$ can now be rewritten as:

$$\begin{aligned} q(\theta) = & \frac{p(\theta|\psi_0)}{Z[q(z)]} \prod_{t,k} \left(\left(\lambda_k / 2\pi\right)^{1/2} \exp\left[-\tfrac{1}{2}\Delta^2_{t,k}\right]\right)^{\gamma_{t,k}} \\ & \prod_{t=2,k,l} \left(A_{kl}\right)^{\xi_{t,kl}} \prod_k \left(\pi_k\right)^{\gamma_{1,k}} \end{aligned} \tag{74}$$

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

$$q(\theta) = q(\mu, \lambda)q(A)q(\pi) \tag{75}$$

This in turn means we can write:

$$q(\mu, \lambda) \propto \prod_{k,t} p(x_t \mid \mu_k, \lambda_k)^{\gamma_{t,k}} p(\mu_k, \lambda_k \mid m_0, \beta_0, a_0, b_0) \tag{76}$$

$$q(A) \propto \prod_{t=2,k,l} (A_{kl})^{\xi_{t,kl}} p(A_k \mid \alpha_{0k}) \tag{77}$$

$$q(\pi) \propto \prod_{k} \pi_k^{\gamma_{1,k}} p(\pi \mid \rho_0) \tag{78}$$

Note that in each of these equations we now have a product of an exponential family likelihood with an exponential family prior, since the the normal likelihood is conjugate to a normal-gamma prior, and a multinomial distribution is conjugate to a Dirichlet prior. The variational posterior distribution is therefore in the same family as the prior, and the m-step updates reduce to calculating a set of posterior parameters $\psi$

For the distribution on $q(\mu, \lambda)$ the exponential form for these updates is simply

$$\nu_k = \nu_k + \sum_t \gamma_{t,k} \tag{79}$$

$$\chi_k = \chi_k + \sum_t \gamma_{t,k} \mathcal{T}(x_t) \tag{80}$$

and can now substitute

$$\nu = \beta_0 \tag{81}$$

$$\chi = \{2b + \beta m^2 \, , \, \beta m\} \tag{82}$$

$$\mathcal{T}(x) = \{x^2, x\} \tag{83}$$

and define

$$N_k = \sum_t \gamma_{t,k} \tag{84}$$

$$\langle X \rangle_k = \sum_t \gamma_{t,k} x_t \tag{85}$$

$$\langle X^2 \rangle_k = \sum_t \gamma_{t,k} x_t^2 \tag{86}$$

to obtain the following expressions for the variational parameters $\psi$

$$m_k = \chi_{k,2} / \nu_k = (\beta_{0k} m_{0k} + \langle X \rangle_k) / (\beta_{0k} + N_k) \tag{87}$$

$$\beta_k = \beta_{0k} + N_k \tag{88}$$

$$a_k = a_{0k} + \tfrac{1}{2} N_k \tag{89}$$

$$\begin{aligned} b_k &= \chi_{k,1} - \chi_{k,2}^2 \, / \, (2\nu_k) \\ &= b_{0k} + \frac{1}{2}\left[\beta_{0k} m_{0k}^2 + \langle X^2 \rangle_k - \frac{(\beta_{0k} m_{0k} + \langle X \rangle_k)^2}{\beta_{0k} + N_k}\right] \end{aligned} \tag{90}$$

Finally, the updates for $\alpha_0$ and $\rho_0$ can be obtained by substitution of the terms in equation (74) into equations (77) and (78):

$$\alpha_{n,kl} = \alpha_{0,kl} + \sum_{t=2}^{T} \xi_{n,t,kl} \tag{91}$$

$$\rho_k = \rho_{0k} + \gamma_{n,1,k} \tag{92}$$

We now proceed to derive how $\gamma$ and $\xi$ can be calculated using the Forward-backward algorithm.

## S3.4 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:

$$\gamma_{t,k} = E_{q(z)}[z_{t,k}] = p^*(x_1 \mid z_1)p^*(z_1) \tag{93}$$
$$\xi_{t,kl} = E_{q(z)}[z_{t-1,k}z_{t,l}] = p^*(z_{t-1} = k, z_{t-1} = l \mid x_{1:T}) \tag{94}$$

to do so we calculate two variables:

$$\alpha_{t,k} = p^*(x_{1:t}, z_t = k) \tag{95}$$
$$\beta_{t,k} = p^*(z_t = k \mid x_{t+1:T}) \tag{96}$$

such that

$$\gamma_{t,k} = p^*(z_t = k \mid x_{1:T}) = \frac{\alpha_{t,k}\beta_{t,k}}{p^*(x_{1:T})} \tag{97}$$
$$\xi_{t,k,l} = p^*(z_{t-1} = k, z_{t-1} = l \mid x_{1:T}) \tag{98}$$
$$= \frac{p^*(x_{1:T} \mid z_t, z_{t-1})p^*(z_t, z_{t-1})}{p^*(x_{1:T})} = \frac{\beta_{t,l}\, p^*(x_t \mid z_t = l)A_{kl}\alpha_{t-1,k}}{p^*(x_{1:T})} \tag{99}$$

and exploit the following recursive relationships:

$$\begin{aligned}\alpha_{t,k} &= p^*(x_{1:t}, z_t)\\ &= \sum_l p^*(x_t \mid z_t = k)p^*(z_t = k \mid z_{t-1} = l)p^*(x_{1:t-1}, z_{t-1} = l)\\ &= \sum_l p^*(x_t \mid z_t = k)A^*_{lk}\alpha_{t-1,l}\end{aligned} \tag{100}$$

$$\begin{aligned}\beta_{t,k} &= p^*(x_{t+1:T} \mid z_t)\\ &= \sum_l p^*(x_{t+2:T} \mid z_{t+1} = l)p^*(x_{t+1} \mid z_{t+1} = l)p^*(z_{t+1} = l \mid z_t = k)\\ &= \sum_l \beta_{t+1,l}\, p^*(x_{t+1} \mid z_{t+1} = l)A^*_{kl}\end{aligned} \tag{101}$$

We can now loop *forward* in time to recursively calculate $\alpha_t$ from $\alpha_{t-1}$ and backward in time to calculate $\beta_t$ from $\beta_{t+1}$. The boundary conditions on these passes are:

$$\alpha_{1,k} = p^*(x_1, z_1) = p^*(x_1 \mid z_1)p^*(z_1) = \prod_k p^*(x_1 \mid z_1 = k)\pi^*_k \tag{102}$$
$$\beta_{T,k} = 1 \tag{103}$$

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 $c_t$:

$$c_t = p^*(x_t \mid x_{1:t-1}) \tag{104}$$

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

$$\begin{aligned}\hat{\alpha}_{t,k} &= \frac{\alpha_{t,k}}{p^*(x_{1:t})} = \prod_{t'=1}^{t} \frac{1}{c_{t'}}\alpha_{t,k}\\ \hat{\beta}_{t,k} &= \frac{\beta_{t,k}}{p^*(x_{t+1:T} \mid x_{1:t})} = \prod_{t'=t+1}^{T} \frac{1}{c_{t'}}\beta_{t,k}\end{aligned} \tag{105}$$

This choice of normalization implies:

$$\gamma_{t,k} = \frac{\alpha_{t,k}\beta_{t,k}}{p^*(x_{1:T})} = \frac{\alpha_{t,k}\beta_{t,k}}{p^*(x_{t+1:T} \mid x_{1:t})p^*(x_{1:t})} = \hat{\alpha}_{t,k}\hat{\beta}_{t,k} \tag{106}$$
$$\xi_{t,k,l} = \frac{\beta_{t,l}\, p^*(x_t \mid z_t = l)A_{kl}\alpha_{t-1,k}}{p^*(x_{1:T})} = \frac{c_t\hat{\beta}_{t,l}\, p^*(x_t \mid z_t = l)A_{kl}\hat{\alpha}_{t-1,k}}{p^*(x_{1:T})} \tag{107}$$

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

$$c_t \hat{\alpha}_{t,k} = \sum_l p^*(x_t \mid z_t = k) A^*_{lk} \hat{\alpha}_{t-1,l} \tag{108}$$

$$c_{t+1} \beta_{t,k} = \sum_l \hat{\beta}_{t+1,l} p^*(x_{t+1} \mid z_{t+1} = l) A^*_{kl} \tag{109}$$

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

$$c_t = c_t \sum_k \hat{\alpha}_{t,k} = \sum_{k,l} p^*(x_t \mid z_t = k) A^*_{lk} \alpha_{t-1,l} \tag{110}$$

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 $p^*(x)$:

$$p^*(x) = p^*(x_{1:t}) = \prod_t c_t \tag{111}$$

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

$$\hat{Z}_{q(z)} = \ln p^*(x) = \sum_t \ln c_t \tag{112}$$

## S3.5 Calculation of the Evidence

The lower bound for the evidence

$$L[q(z), q(\theta)] = \sum \int d\theta \sum_z q(z) q(\theta) \ln \left[ \frac{p(x, z, \theta \mid \psi_0)}{q(z) q(\theta)} \right] \tag{113}$$

can be decomposed into the terms

$$\begin{aligned} L[q(z), q(\theta)] = \sum & E_{q(z)q(\theta)} \left[ \ln p(x, z \mid \theta) \right] \\ & - D_{KL}[q(\theta) \,||\, p(\theta \mid \psi_0)] - E_{q(z)} \left[ \ln q(z) \right] \end{aligned} \tag{114}$$

Now note from equation (51) that $E_{q(z)} \left[ \ln q(z) \right]$ can be written as:

$$E_{q(z)} \left[ \ln q(z) \right] = E_{q(z)q(\theta)} [\ln p(x, z \mid \theta)] - \ln Z[q(\theta)] \tag{115}$$

So

$$L[q(z), q(\theta)] = \ln Z[q(\theta)] - D_{KL}[q(\theta) \,||\, p(\theta \mid \psi_0)] \tag{116}$$

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

$$\begin{aligned} D_{KL}[q(\theta) \,||\, p((\mid \theta))] = & \sum_k D_{KL}[q(\mu_k, \lambda_k) \,||\, p(\mu_k, \lambda_k)] \\ & + D_{KL}[q(A) \,||\, p(A)] + D_{KL}[q(\pi) \,||\, p(\pi)] \end{aligned} \tag{117}$$

The Kullback-Leibler divergence of two exponential family distributions is

$$\begin{aligned} & D_{\mathrm{KL}} \left[ q(\eta \mid \nu, \chi, \phi_0) \,||\, p(\eta \mid \nu_0, \chi_0, \phi_0) \right] \\ & \quad = E_{q(\eta)} \left[ \eta \cdot (\chi - \chi_0) - A(\eta)(\nu - \nu_0) - A(\nu, \chi, \phi_0) + A(\nu_0, \chi_0, \phi_0) \right] \\ & \quad = E_{q(\eta)}[\eta] \cdot (\chi - \chi_0) - E_{q(\eta)}[A(\eta)](\nu - \nu_0) - A(\nu, \chi, \phi_0) + A(\nu_0, \chi_0, \phi_0) \end{aligned} \tag{118}$$

The two required expectation values can be obtained from the relationships

$$0 = \frac{\partial}{\partial \nu} \int d\eta \, q(\eta | \nu, \chi, \phi) \tag{119}$$

$$0 = \nabla_\chi \int d\eta \, q(\eta | \nu, \chi, \phi) \tag{120}$$

which yield

$$E_{q(\eta)}[A(\eta)] = -\nabla_\nu A(\nu, \chi, \phi_0) \tag{121}$$
$$E_{q(\eta)}[\eta] = \nabla_\chi A(\nu, \chi, \phi_0) \tag{122}$$

For a Normal-Gamma distribution we may now substitute the exponential forms

$$\begin{aligned}
\nu &= \beta \\
\chi &= \{2b + \beta m^2, \beta m\} \\
A(\nu, \chi, \phi) &= -\frac{1}{2}\Big[\ln(\nu) + (\nu + \phi - 2)\ln(2\pi) \\
&\qquad + (\nu + \phi)\ln\big[\tfrac{1}{2}(\chi_1 - \chi_2^2/\nu)\big] - 2\ln\Gamma\big[\tfrac{1}{2}(\nu + \phi)\big]\Big] \\
&= -\frac{1}{2}\ln(\beta) - (a-1)\ln(2\pi) - a\ln(b) + \ln\Gamma(a)
\end{aligned}$$

after which the expressions for expectation values are given by

$$E_{q(\eta)}[A(\eta)] = \frac{1}{2}\left[\frac{1}{\beta} + \frac{am^2}{b} + \ln(2\pi) - \Psi(a) + \ln(b)\right] \tag{123}$$
$$E_{q(\eta)}[\eta] = \Big\{-\frac{a}{2b}, \frac{am}{b}\Big\} \tag{124}$$

The KL divergences for $A$ and $\pi$ have simple closed-form expressions:

$$D_{KL}[q(A_k) \,\|\, p(A_k)] = \sum_l [\alpha_{k,l} - \alpha_{0\,k,l}][\psi_0(\alpha_{k,l}) - \psi_0(\alpha_{0\,k,l})] \tag{125}$$
$$D_{KL}[q(\pi) \,\|\, p(\pi)] = \sum_l [\rho_l - \rho_{0\,l}][\psi_0(\rho_l) - \psi_0(\rho_{0\,l})] \tag{126}$$

## S4 Empirical Bayes Updates

In (parametric) empirical Bayes estimation, we construct a generalized EM algorithm that obtains a point estimate $\psi_0$. The quantity optimized is the summed lower bound log evidence over the ensemble of time series:

$$\begin{aligned}
\ln p(x \mid \psi_0) &\geq \sum_n L_n & (127)\\
&= \sum_n E_{q(z_n)q(\theta_n)}\left[\ln\left(\frac{p(x_n, z_n, \theta_n \mid \psi_0)}{q(z_n)q(\theta_n)}\right)\right] & (128)\\
&= \sum_n \ln p(x_n \mid \psi_0) - D_{\mathrm{KL}}\big[q(z_n)q(\theta_n) \,\|\, p(z_n, \theta_n \mid x_n, \psi_0)\big] & (129)\\
&= \sum_n E_{q(z_n)q(\theta_n)}\big[\ln p(x_n \mid z_n, \theta_n)\big] - D_{\mathrm{KL}}\big[q(z_n)q(\theta_n) \,\|\, p(z_n, \theta_n \mid \psi_0)\big] & (130)
\end{aligned}$$

In the E-step the posterior $p(z_n, \theta_n \mid x_n, \psi_0)$ is approximated by maximizing the lower bound with respect to $q(z_n)$ and $q(\theta_n)$. In the M-step the prior $p(z_n, \theta_n \mid \psi_0)$ is used to approximate the variational posterior $q(z_n)q(\theta_n)$ by maximizing the lower bound with respect to $\psi_0$

$$\begin{aligned}
0 &= \nabla_{\psi_0} \sum_n L_n & (131)\\
&= \nabla_{\psi_0} \sum_n \int d\theta_n \, q(\theta_n \mid w_n) \ln p(\theta_n \mid \psi_0) & (132)\\
&= \sum_n \int d\theta_n \, q(\theta_n \mid w_n) \nabla_{\psi_0} \ln p(\theta_n \mid \psi_0) & (133)
\end{aligned}$$

From section 3.3 we note that $p(\theta)$ factorizes without need for further approximation

$$p(\theta \mid \psi_0) = p(\mu, \lambda \mid m_0, \beta_0, W_0, \nu_0)p(A \mid \alpha_0)p(\pi \mid \rho_0) \tag{134}$$

so the updates for $\{\mu, \lambda\}$, $A$, and $\pi$ can be computed separately.

## S4.1 Conjugate-Exponential Form

If we rewrite $p(\theta \mid \psi_0)$ to its conjugate exponential form $p(\eta \mid \nu_0, \chi_0, \phi_0)$, the expression in equation (133) becomes

$$0 = \sum_n E_{q(\eta_n)}\big[\nabla_{\nu_0,\chi_0,\phi_0}\left[\eta \cdot \chi_0 - A(\eta) \cdot \nu_0 + B(\eta, \phi_0) - A(\nu_0, \chi_0, \phi_0)\right]\big] \tag{135}$$

The empirical Bayes updates for the hyperparameters therefore reduce to finding solutions for 3 sets of equations

$$\nabla_{\nu_0} A(\nu_0, \chi_0, \phi_0) = -\frac{1}{N}\sum_n E_{q(\eta_n)}[A(\eta_n)] \tag{136}$$

$$\nabla_{\chi_0} A(\nu_0, \chi_0, \phi_0) = \frac{1}{N}\sum_n E_{q(\eta_n)}[\eta_n] \tag{137}$$

$$\nabla_{\phi_0} A(\nu_0, \chi_0, \phi_0) = \frac{1}{N}\sum_n E_{q(\eta_n)}[\nabla_{\phi_0} B(\eta_n, \phi_0)] \tag{138}$$

where each of the 3 expectation values can be calculated for a given $q(\theta_n \,|, \psi_n)$ in terms of the derivatives of the posterior log normalizer $A(\nu_n, \chi_n, \phi_0)$:

$$E_{q(\eta_n)}[A(\eta_n)] = -\nabla_{\nu_n} A(\nu_n, \chi_n, \phi_0) \tag{139}$$

$$E_{q(\eta_n)}[\eta_n] = \nabla_{\chi_n} A(\nu_0, \chi_0, \phi_0) \tag{140}$$

$$E_{q(\eta_n)}[\nabla_{\phi_0} B(\eta_n, \phi_0)] = \nabla_{\phi_0} A(\nu_n, \chi_n, \phi_0) \tag{141}$$

## S4.2 State Distributions (Dirichlet)

Empirical Bayes updates for a Dirichlet distribution simply match the log expectation values

$$E_{p(\theta_n)}[\log A_k] = \frac{1}{N} E_{q(\theta_n)}[\log A_k],$$
$$E_{p(\theta_n)}[\log \pi] = \frac{1}{N} E_{q(\theta_n)}[\log \pi].$$

These log expectation values can be expressed in terms of the digamma function $\Psi$

$$\Psi\big[\textstyle\sum_m \alpha_{0,km}\big] - \Psi\big[\alpha_{0,kl}\big]$$
$$= \frac{1}{N}\sum_n \Psi\big[\textstyle\sum_m \alpha_{n,km}\big] - \Psi\big[\alpha_{n,kl}\big].$$

While equations above have no closed-form solution, their stationary point can be found efficiently with a Newton iteration method [**?**].

## S4.3 Emission Distribution (Normal-Gamma)

For a 1-dimensional Normal-Gamma distribution substituation of the conjugate exponential forms (section 2.1) yields a set of update equations take the form

$$m_{0k} = \sum_n E_{q(\theta_n)}[\mu_{nk}\lambda_{nk}] / \sum_n E_{q(\theta_n)}[\lambda_{nk}], \tag{142}$$

$$1/\beta_{0k} = \frac{1}{N} E_{q(\theta_n)}[\mu_{nk}^2 \lambda_{nk}] - \frac{1}{N} E_{q(\theta_n)}[\lambda_{nk}\mu_{nk}]^2 / E_{q(\theta_n)}[\lambda_{nk}], \tag{143}$$

$$\Psi(a_{0k}) - \ln(a_{0k}) = \frac{1}{N} E_{q(\theta_n)}[\ln \lambda_{nk}] - \frac{1}{N} \ln E_{q(\theta_n)}[\lambda_{nk}], \tag{144}$$

$$b_k = \frac{N a_k}{E_{q(\theta_n)}[\lambda_{nk}]}, \tag{145}$$

As with the Dirichlet distribution, a Newton iteration method can be used to obtain $a_{0k}$. The prerequisite expectation values can be calculated from

$$E_{q(\theta_n)}[\lambda_{nk}] = a_{n,k}/b_{n,k}, \tag{146}$$

$$E_{q(\theta_n)}[\log \lambda_{nk}] = \psi(a_{n,k}) - \log(b_{n,k}), \tag{147}$$

$$E_{q(\theta_n)}[\mu_{nk}\lambda_{nk}] = m_{n,k} a_{n,k}/b_{n,k}, \tag{148}$$

$$E_{q(\theta_n)}[\mu_{nk}^2 \lambda_{nk}] = 1/\beta_{n,k} + m_{n,k}^2 a_{n,k}/b_{n,k}. \tag{149}$$

## S4.4 Inital State and Transition Probabilities (Dirichlet)

For a Dirichlet distribution the conjugate exponential forms (section 2.2) are given by:

$$\eta = \{\ln \pi_k\} \tag{150}$$

$$\chi = \{\rho_{0k}\} \tag{151}$$

$$h(\chi) = \frac{\prod_k \Gamma(\chi_k + 1)}{\Gamma(\sum_k (\chi_k + 1))} \tag{152}$$

And the log expectation value of $\eta$ is:

$$E_{q(\theta_n)}[\eta] = E_{q(\theta_n)}[\ln \pi] = \psi_0(\rho_{n,k}) - \psi_0(\sum_k \rho_{n,k}) \tag{153}$$

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

$$\psi_0(\rho_{0k}) - \psi_0(\sum_k \rho_{0k}) = \frac{1}{N} \sum_n \psi_0(\rho_{n,k}) - \psi_0(\sum_k \rho_{n,k}) \tag{154}$$

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

$$\psi_0(\alpha_{0kl}) - \psi_0(\sum_l \alpha_{0kl}) = \frac{1}{N} \sum_n \psi_0(\alpha_{n,kl}) - \psi_0(\sum_l \alpha_{n,kl}) \tag{155}$$

## S4.5 Mixtures of Priors

Empirical Bayes estimation can be extended to perform inference over unlabeled subpopulations by defining a mixture model $p(x_n, y_n \mid \psi_0, \zeta)$ on the evidence

$$p(x \mid \psi_0) = p(x \mid \psi_0, y) p(y \mid \zeta) \tag{156}$$

$$= \prod_{nm} \big(p(x \mid \psi_{0m}) \zeta_m\big)^{y_{nm}} \tag{157}$$

$$\geq \prod_{nm} \big(\exp(L_{nm}) \zeta_m\big)^{y_{nm}} \tag{158}$$

where $L_{nm} \geq \ln p(x_n \mid \psi_{0m})$ is the lower bound log evidence for trace $n$ with respect to mixture component $m$. An expectation maximization procedure can be constructed for this mixture model by introducing a variational posterior $q(z_n, \theta_n, y_n) = q(z_n \mid y_n) q(\theta_n \mid y_n) q(y_n)$ for each time series. The update equations for this EM procedure are

$$\frac{\delta L}{\delta q(y_n)} = 0 \qquad \frac{\delta L}{\delta q(z_n \mid y_n)} = 0 \qquad \frac{\delta L}{\delta q(\theta_n \mid y_n)} = 0\,, \qquad \frac{\partial L}{\partial \psi_{0m}} = 0 \qquad \frac{\partial L}{\partial \zeta} = 0\,. \tag{159}$$

The E-step of this EM procedure calculates a set of posterior responsibilities

$$\omega_{nm}^{(i+1)} = E_{q(y_n)}[y_{nm}] = \frac{\exp\big[L_{nm}^{(i)}\big] \zeta_m^{(i)}}{\sum_l \exp\big[L_{nl}^{(i)}\big] \zeta_m^{(i)}} \tag{160}$$

In the M-step we hold $q^{(i+1)}(y_n)$ fixed and maximize $L$ relative to $\psi_{0m}$ and $\zeta$. This amounts to performing EB analysis for subpopulation. In other words we first obtain VB estimates for $q(\theta_n \mid \psi_{nm})$ and then obtain a weighted update of for the hyperparameters

$$0 = \frac{\partial L^{(i+1)}}{\partial \psi_{0m}} = \sum_n \omega_{nm}^{(i+1)} \frac{\partial L_{nm}^{(i+1)}}{\partial \psi_{0m}}\,. \tag{161}$$

The updates in equations (136-138) now become

$$\nabla_{\nu_{0m}} A(\nu_{0m}, \chi_{0m}, \phi_{0m}) = -\sum_n \omega_{nm} E_{q(\eta_n)}[A(\eta_n)] \,/\, \sum_n \omega_{nm}\,, \tag{162}$$

$$\nabla_{\chi_{0m}} A(\nu_{0m}, \chi_{0m}, \phi_{0m}) = \sum_n \omega_{nm} E_{q(\eta_n)}[\eta_n]\,, \tag{163}$$

$$\nabla_{\phi_{0m}} A(\nu_{0m}, \chi_{0m}, \phi_{0m}) = \sum_n \omega_{nm} E_{q(\eta_n)}[\nabla_{\phi_{0m}} B(\eta_n, \phi_{0m})] / \sum_n \omega_{nm}\,. \tag{164}$$

Finally the mixture weights $\zeta^{(i+1)}$ are obtained from

$$\zeta_m^{(i+1)} = \frac{\sum_n \omega_{nm}^{(i+1)}}{\sum_{nm} \omega_{nm}^{(i+1)}}\,. \tag{165}$$

## S5 Calculation of Derivative Kinetic Quantities

### S5.1 Kinetic Rates

The kinetic rates $\kappa$ define a differential equation for the evolution of the probabality $\gamma_k(t)$ that a molecule is in state $k$ at time $t$

$$\frac{\partial \gamma_k(t)}{\partial t} = \sum_l \kappa_{lk} \gamma_l(t) \; . \tag{166}$$

The transition matrix $A$ and measurement time interval $\Delta t$ define a discretized version of this differential equation

$$\frac{1}{\Delta t}\big[\gamma_k(t+\Delta t) - \gamma_k(t)\big] = \frac{1}{\Delta t}\left[\sum_l A_{lk}\gamma_l(t) - \gamma_k(t)\right] = \sum_l \left[\frac{A_{lk} - \mathbb{I}}{\Delta t}\right]\gamma_l(t) \; . \tag{167}$$

In general, the transition matrix $A_{kl}$ can be expressed in terms of of $\kappa_{kl}$ and $\Delta t$ through the relationship

$$A = \exp[\kappa \Delta t] \; . \tag{168}$$

While any given $\kappa$ value uniquely determines $A$, the equation $\kappa = \ln[A]/\Delta t$ does not necessarily have a unique solution. However in the limit of small $\Delta t$ we may truncate the series expansion of the matrix exponent to first order

$$A = \mathbb{I} + (\kappa \Delta t) + \mathcal{O}[(\kappa \Delta t)^2] \; , \tag{169}$$

to obtain the same relationship

$$\kappa \simeq (A - \mathbb{I})/\Delta t \; . \tag{170}$$

### S5.2 Life Time

In order to obtain a distribution on the state life time $\tau_k$ we define

$$A_{kk} = \exp(-1/\tau_k) \; . \tag{171}$$

The marginal distribution on $A_{kk}$ is a Beta distribution

$$p(A_{kk} \,|\, \alpha) = \mathrm{Beta}(A_{kk} \,|\, a_k, b_k) \; , \tag{172}$$

$$= \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} (A_{kk})^{a_k - 1}(1 - A_{kk})^{b_k - 1} \; . \tag{173}$$

with

$$a_k = \alpha_{kk} \tag{174}$$

$$b_k = \Big(\sum_l \alpha_{kl}\Big) - \alpha_{kk} \; . \tag{175}$$

The probability density function for the life time is now given by

$$p(\tau_k \,|\, a_k, b_k) = \frac{\partial A_{kk}}{\partial \tau_k} p\big(A_{kk}(\tau_k) \,|\, a_k, b_k\big) \tag{176}$$

$$= \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \frac{1}{\tau_k^2} \Big(\exp[-1/\tau_k]\Big)^{a_k} \Big(1 - \exp[-1/\tau_k]\Big)^{b_k - 1} \tag{177}$$

### S5.3 Free Energy

In the limit $t \to \infty$, the markov chain for a set of probablities $\gamma_{kt}$ will converge to the stationary distribution $v_k$, which is given by the solution to the eigenvalue equation

$$v_k = \sum_l A_{lk} v_l \; . \tag{178}$$

In other words, the stationary distribution $v$ is the normalized eigenvector of $A^\top$ with eigenvalue 1. This quantity is related to the free energy $G_k$ of each state through

$$v_k \propto \exp[-G_k/k_B T] \,. \tag{179}$$

For a 2-state, system the eigen-vector of the transition matrix can be calculated trivially from the off-diagonal elements

$$A = \begin{bmatrix} (1-\delta) & \delta \\ \epsilon & (1-\epsilon) \end{bmatrix} \qquad u \propto \begin{bmatrix} \epsilon \\ \delta \end{bmatrix} \qquad G = k_B T \ln\left[\delta/\epsilon\right] \tag{180}$$

We approximate $G_k$ for each state by calculating a marginal

$$p(\delta_k, \epsilon_k \,|\, a_k, b_k, c_k, d_k) = \mathrm{Beta}(\delta_k \,|\, b_k, a_k)\mathrm{Beta}(\epsilon_k \,|\, c_k, d_k) \,. \tag{181}$$

with

$$a_k = \alpha_{kk} \tag{182}$$

$$b_k = \Big(\sum_l \alpha_{kl}\Big) - \alpha_{kk} \tag{183}$$

$$c_k = \Big(\sum_l \alpha_{lk}\Big) - \alpha_{kk} \tag{184}$$

$$d_k = \alpha_{kk} + \Big(\sum_{kl} \alpha_{kl}\Big) - b_k - c_k \tag{185}$$

In other words, for each state $k$ we collapse all states $l \neq k$ and calculate $G_k$ based on the resulting prior on a $2 \times 2$ transition matrix. We will now define $g_k = G_k/(k_B T)$ to calculate the marginal

$$p(g_k \,|\, a_k, b_k, c_k, d_k) = \int d\delta_k |J(\delta_k, g_k)| p(\delta_k, \exp[-g_k]\delta_k \,|\, a_k, b_k, c_k, d_k) \,, \tag{186}$$

where the Jacobian term is given by

$$|J(\delta_k, g_k)| = \delta_k \exp[-g_k] \tag{187}$$

The integral has no closed-form solution, but can be integrated numerically.