Biophysical Journal, Volume *116*

Supplemental Information

Bayesian-Estimated Hierarchical HMMs Enable Robust Analysis of Single-Molecule Kinetic Heterogeneity

Jason Hon and Ruben L. Gonzalez Jr.

SUPPLEMENTARY INFORMATION

Bayesian-estimated hierarchical HHMs enable robust analysis of single-molecule kinetic heterogeneity

Jason Hon[1] and Ruben L. Gonzalez Jr.[1,2]

[1] *Department of Chemistry, Columbia University, New York, New York 10027, USA*

[2] To whom correspondence may be addressed: Ruben L. Gonzalez Jr., Department of Chemistry, Columbia University, 3000 Broadway Av., MC 3126, New York, NY 10027, USA, Tel.: (212) 854-1096; FAX: (212) 932-1289; Email: rlg2118@columbia.edu

TABLE OF CONTENTS

S1 Generative Model for Hierarchical Hidden Markov Models (HHMMs)

S1.1 Overview

In this section, we will first define all the variables used to describe the algorithms for dynamic and static heterogeneity. Next, we show how these variables are organized to optimize the evidence – the probability that a given set of parameters, state occupancies, and observations give rise to the same dataset. To describe the evidence, we will begin with a formal definition of the evidence, then follow by defining the signal emissions model used herein, describe the prior distributions, and close with general outlines of the two algorithms.

S1.2 Variable Definitions

Variable	Definition
x_{nt}	Observations of a signal trajectory $n \in \{1, \dots, N\}$ at time $t \in \{1, \dots, T_n\}$.
z_{nt}^{d}	State of the biomolecule in signal trajectory $n \in \{1, \dots, N\}$ at time $t \in \{1, \dots, T_n\}$. The model for dynamic heterogeneity has $d \in \{1, \dots, D\}$ and the model for static heterogeneity has $d \in \{1,2\}$.
Ω_d	Size of the state space at level $d \in \{1, \dots, D\}$.
$\widetilde{\Omega}_d \equiv \frac{\Omega_d}{\Omega_{d-1}}$	Accessible state space at level $d \in \{1, \dots, D\}$.
θ	Collectively, the parameters for a population of signal trajectories.
ϕ	Collectively, the parameters for the distribution of signal emissions.

Variable	Definition
ϕ_i	Distribution of signal emissions for a given production state $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
μ_i	Mean of the normal distribution of signal emissions for a given production state $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
λ_i	Precision of the normal distribution of signal emissions for a given production state $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
m_i	Variational estimate for the mean, μ_i, of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
β_i	Variational estimate for the precision of the mean, μ_i, of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
a_i	Variational estimate for the scale of the gamma distribution of the precision, λ_i , of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
b_i	Variational estimate for the rate of the gamma distribution of the precision, λ_i , of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
$\pi_i^d(k)$	Initial state probabilities $d \in \{1, \dots, D\}, i \in \{1, \dots, \widetilde{\Omega}_d\}, k \in \{1, \dots, \Omega_{d-1}\}$.
$A_{ij}^d(k)$	Transition matrices $d \in \{1, \dots, D\}, i \in \{1, \dots, \widetilde{\Omega}_d\}, j \in \{1, \dots, \widetilde{\Omega}_d + 1\}, k \in \{1, \dots, \Omega_{d-1}\}$.
$A_{i,\widetilde{\Omega}_d+1}^d(k)$	Probability of transitioning between siblings of the parent of the ith node at level d.
$\rho_i^d(k)$	Variational estimate for the number of times a signal trajectory is first observed in state $d \in \{1, \dots, D\}, i \in \{1, \dots, \widetilde{\Omega}_d\}, k \in \{1, \dots, \Omega_{d-1}\}$.
$\alpha_{ij}^d(k)$	Variational estimate for the number of times a signal trajectory makes a transition between $i \in \{1, \dots, \widetilde{\Omega}_d\}$ and $j \in \{1, \dots, \widetilde{\Omega}_d + 1\}$ at level $d \in \{1, \dots, D\}$ positioned at the path $k \in \{1, \dots, \Omega_{d-1}\}$.
ψ_0	Collectively, hyperparameters for the prior distribution.
$m_{0,i}$	Prior estimate for the mean of the normal distribution of the mean μ_i of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
$\beta_{0,i}$	Prior estimate for the precision of the normal distribution of the mean μ_i of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.

Variable	Definition
$a_{0,i}$	Prior estimate for the scale of the gamma distribution of the precision, λ_i, of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
$b_{0,i}$	Prior estimate for the rate of the gamma distribution of the precision, λ_i, of the normal distribution of signal emissions, $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
$\rho^{d}_{0,i}(k)$	Prior estimate for the number of times a signal trajectory is first observed in state $d \in \{1, \dots, D\}, i \in \{1, \dots, \widetilde{\Omega}_d\}, k \in \{1, \dots, \Omega_{d-1}\}$.
$\alpha^{d}_{0,ij}(k)$	Prior estimate for the number of times a signal trajectory makes a transition between $i \in \{1, \dots, \widetilde{\Omega}_d\}$ and $j \in \{1, \dots, \widetilde{\Omega}_d + 1\}$ at level $d \in \{1, \dots, D\}$ positioned at the path $k \in \{1, \dots, \Omega_{d-1}\}$.
$L\left(q(\{z^{d}_{nt}\}), q(\theta)\right)$	Evidence.
γ^{i}_{nt}	Expected occupancy of the production state $i \in \{1, \dots, \widetilde{\Omega}_D\}$ of a biomolecule.
$\xi^{d}_{ntij}(k)$	Expected counts of the number of transitions between $i \in \{1, \dots, \widetilde{\Omega}_d\}$ and $j \in \{1, \dots, \widetilde{\Omega}_d + 1\}$ at level $d \in \{1, \dots, D\}$ positioned at the path $k \in \{1, \dots, \Omega_{d-1}\}$ in signal trajectory $n \in \{1, \dots, N\}$ at time $t \in \{2, \dots, T_n\}$.
g^{d}_{ni}	Expected counts of the number of times a signal trajectory begins in state $i \in \{1, \dots, \widetilde{\Omega}_d\}$ at level $d \in \{1, \dots, D\}$.
ζ_{ni}	Mixture coefficients in the static heterogeneity algorithm, probability that a signal trajectory $n \in \{1, \dots, N\}$ belongs to state $i \in \{1, \dots, \widetilde{\Omega}_D\}$.
$c_{nt}(k)$	Forward-backward scale variable.
$\hat{\alpha}^{d}_{nti}(k)$	Forward variable.
$\hat{\beta}^{d}_{nti}(k)$	Backward variable.
$\hat{\alpha}^{i}_{b^{d}_{n,t}}(k)$	Forward-upward variable.
$\hat{\alpha}^{i}_{e^{d}_{n,t}}(k)$	Forward-downward variable.
$\hat{\beta}^{i}_{b^{d}_{n,t}}(k)$	Backward-upward variable.
$\hat{\beta}^{i}_{e^{d}_{n,t}}(k)$	Backward-downward variable.
$par(z)$	The set of nodes in the state-space graph that point at z, or “parents”.

Variable	Definition
$par_k(z)$	The k^{th} super-parent of z.
$ch(z)$	The set of nodes in the state-space graph that z points at, or "children".
$sib(x)$	The set of nodes in the state-space graph that share nodes that point to x, or "siblings".
$\Gamma\,(z) = \int_0^\infty x^{z-1}e^{-x}dx$	Gamma function.
$\psi(z) = \frac{d\left[\ln\left(\Gamma\,(z)\right)\right]}{dz}$	Digamma function.

S1.3 Evidence

The evidence is the probability that the current set of observations was obtained from an experiment given any possible set of parameters and some amount of prior data. As in all Bayesian inference-based algorithms, we would like to use Bayesian inference to estimate the parameter distributions that will optimize the evidence:

$$p(\{x_{nt}\},\theta|\psi_0) = \int p(\{x_{nt}\}|\theta)p(\theta|\psi_0)\,d\theta.$$

Unfortunately, this calculation is analytically intractable in the case of the present model. As a consequence, we instead seek to estimate the parameter distributions that will maximize a lower bound of the evidence (1):

$$L\left(q\left(\{z_n^d\quad\}\right),q(\theta)\right) = \int d\theta \sum_n \sum_{z_{nt}^d} q(\theta)q(z_{nt}^d)\ln\frac{p(x_{nt},z_{nt}^d,\theta|\psi_0)}{q(z_{nt}^d)q(\theta)}.$$

This sum over z_{nt}^d runs over all of the possible values of all of the possible states of the signal trajectory of the biomolecule. This expression assumes that the joint probability can be factorized:

$$p\left(z_{nt}^d,\theta\middle|x_{nt},\psi_0\right) = q\left(z_{nt}^d\right)q(\theta),$$

an assumption that forms the basis of the variational approximation (1).

S1.4 Signal Emissions Model

The signal emissions model is the probability that an observation was obtained from an experiment given a particular set of parameters and a particular production state (*i.e.*, the state of the directly observed dimension) of the biomolecule. It is given by:

$$p(\{x_{nt}\}|z_{nt}^D,\theta) = \prod_{n=1}^{N}\prod_{t=1}^{T_n} p\left(x_{nt}\middle|\phi_{z_{nt}^D}\right).$$

Furthermore, we assume that $p\left(x_{nt}\middle|\phi_{z_{nt}^D}\right)$ follows a normal distribution:

$$p\left(x_{nt}\middle|\phi_{z_{nt}^D}\right) = \left(\frac{\lambda_{z_{nt}^D}}{\pi}\right)^{\frac{1}{2}} e^{-\frac{\lambda_{z_{nt}^D}}{2}\left(x_{nt}-\mu_{z_{nt}^D}\right)^2}.$$

Nonetheless, the signal emissions model may generally be modified to any appropriate distribution. Including the signal emission distributions, the likelihood function is given by:

$$L = \prod_{n=1}^{N}\left[\prod_{d=1}^{D}\pi_{d,z_{n1}^d}\prod_{t=1}^{T_n} p\left(x_{nt}\middle|\phi_{z_{nt}^D}\right)\right]$$

$$\left[\prod_{t=2}^{T_n-1}\prod_{d=1}^{D-1} A_{d,z_{nt}^d,exit}^{\delta_{z_{nt}^{d+1},z_{n,t+1}^{d+1}}} A_{d,z_{nt}^d,z_{n,t-1}^d}^{\delta_{z_{nt}^d,z_{n,t+1}^d}\left(1-\delta_{z_{nt}^{d+1},z_{n,t+1}^{d+1}}\right)} \pi_{d,z_{n,t+1}^d}\right]\left[\prod_{d=1}^{D} A_{d,z_{nT_n}^d,exit}\right].$$

S1.5 Prior Distributions

Prior information using the variational approximation and standard ("conjugate exponential") distributions allows us to write down the form of the prior distributions. These prior distributions have advantages and limitations that are discussed elsewhere (see reference 2). These are given by:

Variable	Prior Distribution
$\mu_i, i \in \{1, \dots, \widetilde{\Omega}_D\}$	$p(\mu_i\|\psi_0) = p(\mu_i\|m_{0,i},\beta_{0,i}) = \left(\frac{\beta_{0,i}}{2}\right)^{\frac{1}{2}} e^{-\frac{\beta_{0,i}}{2}(\mu_i-m_{0,i})^2}$
$\lambda_i, i \in \{1, \dots, \widetilde{\Omega}_D\}$	$p(\lambda_i\|\psi_0) = p(\lambda_i\|a_{0,i},b_{0,i}) = \frac{{b_{0,i}}^{a_{0,i}}}{\Gamma(a_{0,i})}{\lambda_i}^{a_{0,i}-1}e^{-b_{0,i}\lambda_i}$

Variable	Prior Distribution
$\{\pi_i^d(k)\}, i \in \{1, \dots, \widetilde{\Omega}_d\}$ $k \in \{1, \dots, \Omega_{d-1}\}$ $d \in \{1, \dots, D\}$	$p(\{\pi_i^d\}(k) \vert \psi_0) = p\left(\{\pi_i^d\}(k) \middle\vert \{\rho_{0,i}^d\}(k)\right) = \dfrac{\Gamma\left(\sum_{j=1}^{\widetilde{\Omega}_d} \rho_{0,j}^d(k)\right)}{\prod_{j=1}^{\widetilde{\Omega}_d} \Gamma\left(\rho_{0,j}^d(k)\right)} \prod_{j=1}^{\widetilde{\Omega}_d} \left(\pi_j^d(k)\right)^{\rho_{0,j}^d(k)-1}$
$A_{ij}^d(k), i \in \{1, \dots, \widetilde{\Omega}_d\}$ $j \in \{1, \dots, \widetilde{\Omega}_d + 1\}$ $k \in \{1, \dots, \Omega_{d-1}\}$ $d \in \{1, \dots, D\}$	$p(\{A_{ij}^d\}(k) \vert \psi_0, i) = p(\{A_{ij}^d\}(k) \vert \{\alpha_{0,ij}^d\}(k), i) = \dfrac{\Gamma\left(\sum_{j=1}^{\widetilde{\Omega}_d+1} \alpha_{0,ij}^d(k)\right)}{\prod_{j=1}^{\widetilde{\Omega}_d+1} \Gamma\left(\alpha_{0,ij}^d(k)\right)} \prod_{j=1}^{\widetilde{\Omega}_d+1} \left(A_{ji}^d(k)\right)^{\alpha_{0,ij}^d(k)-1}$

S1.6 Algorithm

The algorithm proceeds iteratively, first with an expectation step (E-step) and then a maximization step (M-step), until the change in the value of the evidence lower bound between consecutive iterations is negligible. Broadly, the E-step determines the expected state of the biomolecule in each signal trajectory at each time point and also calculates the value of the likelihood function (($p(x|\theta)$, generally; see Equation 2 in the main text of the article) The M-step takes the expected state occupancies of the biomolecule calculated in the E-step and uses these occupancies to re-estimate all of the parameters. We do not re-estimate the parameters of the prior distribution and, as such, it is important that the values chosen do not contribute more than the updates derived from observations in the M-step. By default, the algorithm utilizes naïve parameters for the prior distributions assuming observation of a dataset 100-fold smaller than that being analyzed.

S2 Variational Bayes Expectation Maximization (VBEM)

S2.1 The E-step

The E-step estimates the likelihood function while concurrently using the current estimates of the parameters, θ, to estimate the most likely state of each biomolecule n at time t,

z_{nt}^{d}. This is done in conceptually distinct ways for dynamic and for static heterogeneity. In the case of dynamic heterogeneity, wherein the biomolecules are described by signal trajectories that undergo stochastic, abrupt changes in the rates of transitions between observed states, we present an algorithm, termed the Forward-Backward Activation (FBA) algorithm and first described by Wakabayashi, *et al* (3), that generalizes the well-known forward-backward algorithm to account for transitions along indirectly observed dimensions. In the case of static heterogeneity, wherein the biomolecules are described by subpopulations of signal trajectories that are distinct in their rates of transitions between observed states, we present an algorithm, termed the Forward-Backward Mixture (FBM) algorithm, that, over a mixture of HMMs, determines the parameters that distinguish each subpopulation and describe the degree to which each signal trajectory belongs to each subpopulation. The goal of the E-step is to return summary statistics with which we can re-estimate all parameters in the M-step. The E-step is, in the cases of both dynamic and static heterogeneity, the step of the algorithm that consumes most of the computational time.

S2.1.1 Forward-Backward Activation (FBA) Algorithm – Dynamic Heterogeneity

We adapt the FBA algorithm initially described by Wakabayashi, *et al* (3). The FBA algorithm estimates quantities, termed the “forward-upward”, “forward-downward”, “backward-upward”, and “backward-downward” variables, that are used to calculate the state occupancies and inter-state transition counts. If each node of the tree has K children, then the algorithm is $O(NTK^{D})$ where N is the number of signal trajectories, T is the number of time points in each signal trajectory, D is the number of subpopulations of signal trajectories, and K is the number of production states. It is important to note that the FBA algorithm is linear in time, and is thus tractable for large datasets.

The forward-upward and forward-downward variables are calculated according to the following recursion, beginning at the top of the tree:

$$\hat{\alpha}^{i}_{b^{1}_{nt}}(1) = \sum_{j} \hat{\alpha}^{j}_{e^{1}_{nt-1}}(1) A^{1}_{ji}(1).$$

We define a scale factor that will eventually be used to calculate the likelihood function as well as keep the entirety of the calculation within computational precision:

$$\hat{\alpha}^{i}_{e^{D}_{nt}}(k) = \hat{\alpha}^{i}_{b^{D}_{nt}} p(x_{nt}|\phi_{i})$$

$$c_{nt} = \sum_{k \in \Omega_{D-1}} \sum_{i \in ch(k)} \hat{\alpha}^{i}_{e^{D}_{nt}}(k).$$

Using these two equations which generate the scale factor, we continue the recursion:

$$\hat{\boldsymbol{\alpha}}^{i}_{e^{D}_{nt}}(k) = \hat{\alpha}^{i}_{e^{D}_{nt}}(k) \prod_{t'=1}^{t} c^{-1}_{nt'}$$

$$\hat{\alpha}^{i}_{b^{d}_{nt}}(k) = \hat{\alpha}^{k}_{b^{d-1}_{nt}}\big(par(k)\big)\pi^{d}_{i}(k) + \sum_{j \in ch(k)} \hat{\alpha}^{j}_{e^{d}_{nt-1}}(k) A^{d}_{ji}(k), d \in \{2, \dots, D\}$$

$$\hat{\alpha}^{i}_{b^{D}_{nt}}(k) = \hat{\alpha}^{k}_{b^{D-1}_{nt}}\big(par(k)\big)\pi^{D}_{i}(k) + \sum_{j \in ch(k)} \hat{\boldsymbol{\alpha}}^{j}_{e^{D}_{nt-1}}(k) A^{D}_{ji}(k)$$

$$\hat{\alpha}^{i}_{e^{d}_{nt}}(k) = \sum_{j \in ch(i)} \hat{\alpha}^{j}_{e^{d+1}_{nt}}(i) A^{d+1}_{j,\widetilde{\Omega}_{d}+1}, d \in \{1, \dots, D-1\}.$$

The forward-upward variables have time-boundary conditions:

$$\hat{\alpha}^{i}_{b^{1}_{n1}}(k) = \pi^{1}_{i}$$

$$\hat{\alpha}^{i}_{b^{d}_{n1}}(k) = \hat{\alpha}^{k}_{b^{d-1}_{n1}}\big(par(k)\big)\pi^{d}_{i}, d \in \{2, \dots, D\}.$$

Similarly, the backward-upward and backward-downward variables are calculated according to the following recursion:

$$\hat{\beta}^{i}_{e^{1}_{nt}}(1) = \sum_{j} \hat{\beta}^{j}_{b^{1}_{nt+1}}(1) A^{1}_{ij}(1)$$

$$\hat{\beta}^{i}_{b^{D}_{nt}}(k) = \hat{\beta}^{i}_{e^{D}_{nt}} p(x_{nt}|\phi_{i}) \prod_{t'=1}^{t} c^{-1}_{nt}$$

$$\hat{\beta}^{i}_{e^{d}_{nt}}(k) = \hat{\beta}^{k}_{e^{d-1}_{nt}}\big(par(k)\big)A^{d}_{i,\widetilde{\Omega}_{d}+1}(k) + \sum_{j \in ch(k)} \hat{\beta}^{j}_{b^{d}_{nt+1}}(k)A^{d}_{ij}(k)\,, d \in \{2, \dots, D\}$$

$$\hat{\beta}^{i}_{b^{d}_{nt}}(k) = \textstyle\sum_{j \in ch(i)} \hat{\beta}^{j}_{b^{d+1}_{nt}}(i)\pi^{d+1}_{j}\,, d \in \{1, \dots, D-1\}.$$

It should be noted that we have introduced a scaling variable alongside the backward-downward variables. The backward-downward variables have time-boundary conditions:

$$\hat{\beta}^{i}_{e^{1}_{nT_n}}(k) = A^{1}_{i,\widetilde{\Omega}_{1}+1}$$

$$\hat{\beta}^{i}_{e^{d}_{n1}}(k) = \hat{\beta}^{k}_{b^{d-1}_{n1}}\big(par(k)\big)A^{d}_{i,\widetilde{\Omega}_{d}+1}, d \in \{2, \dots, D\}.$$

Finally, we need to prepare the variables needed for the M-step, as well as calculate the likelihood function. This is done by setting:

$$p(\{x_{nt}\}) = \prod_{n,t} c_{nt}$$

$$\gamma^{i}_{nt} = \sum_{k \in \Omega_{D-1}} \hat{\alpha}^{i}_{e^{D}_{nt}}(k)\hat{\beta}^{i}_{e^{D}_{nt}}(k)$$

$$g^{d}_{ni} = \hat{\alpha}^{i}_{b^{d}_{n1}}(k)\hat{\beta}^{i}_{b^{d}_{n1}}(k) + \sum_{t=1}^{T_n-1} \hat{\alpha}^{k}_{b^{d-1}_{n,t+1}}\big(par(k)\big)\pi^{d}_{i}\hat{\beta}^{i}_{b^{d}_{n,t+1}}(k)$$

$$\xi^{d}_{ntij}(k) = \sum_{t=1}^{T_n-1} \hat{\alpha}^{i}_{e^{d}_{n,t}}(k)A^{d}_{ij}\hat{\beta}^{j}_{b^{d}_{n,t+1}}(k)$$

$$\xi^{d}_{nti,\widetilde{\Omega}_{d}+1} = \hat{\alpha}^{i}_{e^{d}_{nT_n}}(k)\hat{\beta}^{i}_{b^{d}_{nT_n}}(k) + \sum_{t=1}^{T_n-1} \hat{\alpha}^{i}_{e^{d}_{n,t}}(k)A^{d}_{i,\widetilde{\Omega}_{d}+1}\hat{\beta}^{k}_{e^{d}_{n,t+1}}\big(par(k)\big).$$

S2.1.2 Forward-Backward Mixture (FBM) Algorithm – Static Heterogeneity

We adapt the E-step of the FBM algorithm from reference 1. This FBM algorithm estimates quantities, termed the “forward” and “backward” variables, that are used to acquire the mixture coefficients, state occupancies, and inter-state transition counts. The complexity of the algorithm is $O(NTK^2D)$ where N is the number of signal trajectories, T is the number of time points in each signal trajectory, D is the number of subpopulations of signal trajectories, and K is the number of production states. It is important to note that, like the FBA algorithm, the FBM algorithm is linear in time, and is thus tractable for large datasets.

The forward variable, $\hat{\alpha}_{nti}^{d}(k)$, and backward variable, $\hat{\beta}_{nti}^{d}(k)$, reduce in the d dimension because the subpopulations that do not interconvert do not have subsequent transitions along any indirectly observed dimensions:

$$\hat{\alpha}_{nti}^{d}(k) = \hat{\alpha}_{nti}(k) \equiv p(x_{nt}|\phi_i) \sum_{j} \hat{\alpha}_{n,t-1,j}(k) A_{ji}(k)$$

$$\hat{\beta}_{nti}^{d}(k) = \hat{\beta}_{nti}(k) \equiv \textstyle\sum_{j} \hat{\beta}_{n,t+1,j}(k) p\left(x_{n,t+1}|\phi_j\right) A_{ij}(k).$$

The boundary conditions are as follows:

$$\hat{\alpha}_{n1i}(k) = p(x_{nt}|\phi_i)\pi_i(k)$$

$$\hat{\beta}_{nT_n i}(k) = p\left(x_{nT_n}|\phi_i\right),$$

where it should be noted that we have removed an unnecessary index from $\pi_i^d(k)$. These variables are normalized to supply the likelihood function as well as a convenient scale and recursion, all to guarantee computational precision:

$$c_{n} \quad (k) \equiv \sum_{j} \hat{\alpha}_{ntj}(k) A_{ji}(k)$$

$$\widehat{\boldsymbol{\alpha}}_{nti}(k) \equiv \hat{\alpha}_{nti}(k) \prod_{t^{\prime}=1}^{t} c_{nt^{\prime}}^{-1}(k)$$

$$\widehat{\boldsymbol{\beta}}_{nti}(k) \equiv \hat{\beta}_{nti}(k) \textstyle\prod_{t^{\prime}=1}^{t} c_{nt^{\prime}}^{-1}(k).$$

To complete the algorithm, we use the above variables to calculate the variables of primary interest:

$$p(\{x_{nt}\}) = q\left(z_{nt}^{d}\right) = \prod_{n,k,t} c_{nt}(k)$$

$$\gamma_{nt}^{i} = \sum_{k} \widehat{\boldsymbol{\alpha}}_{nti}(k) \widehat{\boldsymbol{\beta}}_{nti}(k)$$

$$g_{ni} = \hat{\boldsymbol{\alpha}}_{n1i}(k)\hat{\boldsymbol{\beta}}_{n1i}(k)$$

$$\xi_{ntij}(k) = \frac{c_{nt}(k)p(x_{nt}|\phi_i)\hat{\boldsymbol{\alpha}}_{n,t-1,j}(k)\hat{\boldsymbol{\beta}}_{nti}(k)A_{ji}(k)}{\prod_t c_{nt}(k)}$$

$$\zeta_{ni} = \tfrac{\prod_t c_{nt}(k)}{\sum_k \prod_t c_{nt}(k)}.$$

At this point, all variables required for the M-step of the FBM algorithm have been prepared.

S2.2 M-step

Parameters are updated during the M-step. This is done simultaneously and, as such, all parameters on the right-hand side of the equations belong to the previous iteration and those on the left-hand side belong to the current iteration. Priors are not updated in this model, as we do not utilize the empirical Bayes' framework. The M-step is iterated with the E-step above until the evidence lower bound converges within a set tolerance. There are only trivial differences between the FBA and FBM algorithms in the M-step. This is because, at the M-step, the FBM algorithm is simply a limiting case of the FBA algorithm and, as such, the two algorithms share the same parameter distribution structures and differ only in the details of the model topology. Derivations of these equations may be found in reference 1 and are derived by maximizing the evidence lower bound.

$$\beta_i = \beta_{0,i} + \sum_{n,t} \gamma_{nt}^i \,, i \in \{1, \dots, \widetilde{\Omega}_D\}$$

$$a_i = a_{0,i} + \frac{1}{2}\sum_{n,t} \gamma_{nt}^i \,, i \in \{1, \dots, \widetilde{\Omega}_D\}$$

$$b_i = b_{0,i} + \frac{1}{2}\left(\beta_{0,i}m_{0,i}^2 + \sum_{n,t} x_{nt}^2\gamma_{nt}^i - \frac{\left(\beta_{o,i}m_{0,i} + \sum_{n,t} x_{nt}\gamma_{nt}^i\right)^2}{\beta_i}\right), i \in \{1, \dots, \widetilde{\Omega}_D\}$$

$$\lambda_i = \frac{a_i}{b_i}, i \in \{1, \dots, \widetilde{\Omega}_D\}$$

$$m_i = \frac{\lambda_i}{\beta_i}\left(\sum_{n,t} x_{nt}\gamma_{nt}^{i} + m_{0,i}\beta_{0,i}\right), i \in \left\{1, \dots, \widetilde{\Omega}_D\right\}$$

$$\mu_i = m_i, i \in \left\{1, \dots, \widetilde{\Omega}_D\right\}$$

$$\rho_i^d(k) = \ \rho_{0,i}^d(k) + \textstyle\sum_n g_{ni}^d,\ d \in \{1, \dots, D\}, i \in \left\{1, \dots, \widetilde{\Omega}_d\right\}, k \in \{1, \dots, \Omega_{d-1}\}$$

$$\pi_i^d(k) = e^{\psi(\rho_i^d) - \sum_i \psi(\rho_i^d)},\ d \in \{1, \dots, D\}, i \in \left\{1, \dots, \widetilde{\Omega}_d\right\}, k \in \{1, \dots, \Omega_{d-1}\}$$

$$\alpha_{ij}^d(k) = \alpha_{0,ij}^d(k) + \sum_{n,t} \xi_{ntij}^d(k), i \in \left\{1, \dots, \widetilde{\Omega}_d\right\}, j \in \left\{1, \dots, \widetilde{\Omega}_d + 1\right\},$$
$$k \in \{1, \dots, \Omega_{d-1}\}, d \in \{1, \dots, D\}$$

$$A_{ij}^d(k) = e^{\psi(\rho_i^d) - \sum_i \psi\left(A_{ij}^d(k)\right)}\ i \in \left\{1, \dots, \widetilde{\Omega}_d\right\}, j \in \left\{1, \dots, \widetilde{\Omega}_d + 1\right\},$$
$$k \in \{1, \dots, \Omega_{d-1}\}, d \in \{1, \dots, D\}.$$

S2.3 Calculation of Evidence Lower Bound

The evidence lower bound is given by:

$$L\left(q\left(\left\{z_{nt}^d\right\}\right), q(\theta)\right) = p(\{x_{nt}\}) - D_{KL}(\phi||\psi_0) - D_{KL}\left(\left\{\rho_i^d\right\}||\psi_0\right) - D_{KL}\left(\left\{\alpha_{ij}^d(k)\right\}||\psi_0\right),$$

where $p(\{x_{nt}\})$ is calculated in the E-step and the D_{KL} terms are given by:

$$D_{KL}(\phi||\psi_0) = \left[(a_i - 1)\psi(a_i) + \log\left(\frac{a_i}{b_i}\right) - a_i + \log\left(\frac{\Gamma\left(a_{0,i}\right)}{\Gamma\left(a_i\right)}\right) + a_{0,i}\log\left(b_{0,i}\right)\right.$$
$$\left. - \left(a_{0,i} - 1\right)(\psi(a_i) + \log(b_i)) + \frac{a_i b_i}{b_{0,i}}\right] + \left[\log\left(\frac{\beta_{0,i}}{\beta_i}\right) + \frac{\beta_i + \left(m_i - m_{0,i}\right)^2}{2\beta_{0,i}} - \frac{1}{2}\right]$$

$$D_{KL}\left(\{\rho_i^d\}||\psi_0\right) = \log\sum_{i=1}^{\widetilde{\Omega}_d}\Gamma\left(\rho_i^d\right) - \log\sum_{i=1}^{\widetilde{\Omega}_d}\Gamma\left(\rho_{0,i}^d\right) + \sum_{i=1}^{\widetilde{\Omega}_d}\log\Gamma\left(\rho_i^d\right) - \sum_{i=1}^{\widetilde{\Omega}_d}\log\Gamma\left(\rho_{0,i}^d\right)$$

$$+\sum_{i=1}^{\widetilde{\Omega}_d}\left(\rho_i^d - \rho_{0,i}^d\right)\left(\psi\left(\rho_i^d\right) - \psi\left(\sum_{i=1}^{\widetilde{\Omega}_d}\rho_i^d\right)\right)$$

$$D_{KL}\left(\{\alpha_{ij}^d(k)\}||\psi_0\right) = \sum_{j=1}^{\widetilde{\Omega}_d+1}\left(\log\sum_{i=1}^{\widetilde{\Omega}_d}\Gamma\left(\alpha_{ij}^d(k)\right) - \log\sum_{i=1}^{\widetilde{\Omega}_d}\Gamma\left(\alpha_{0,ij}^d(k)\right) + \sum_{i=1}^{\widetilde{\Omega}_d}\log\Gamma\left(\alpha_{ij}^d(k)\right) -\right.$$

$$\left.\sum_{i=1}^{\widetilde{\Omega}_d}\log\Gamma\left(\alpha_{0,ij}^d(k)\right) + \sum_{i=1}^{\widetilde{\Omega}_d}\left(\alpha_{ij}^d(k) - \alpha_{0,ij}^d(k)\right)\left(\psi\left(\alpha_{ij}^d(k)\right) - \psi\left(\sum_{i=1}^{\widetilde{\Omega}_d}\alpha_{ij}^d(k)\right)\right)\right).$$

S3 Calculation of Kinetic Rates

S3.1 Static Heterogeneity

Calculation of the kinetic rates follows:

$$k_{ij}^d \approx A_{ij}^d, i \neq j,$$

where the rate constants are in units of time-steps.

S3.2 Dynamic Heterogeneity

Calculation of the kinetic rates follows:

$$d^* \equiv \min\left(d|par_d(i) = par_d(j)\right)$$

$$k_{ij}^d \approx \left[\prod_{m=1}^{d-1}\sum_j A_{j,\widetilde{\Omega}_m+1}^m\left(ch_m(i)\right)\right]$$

$$\prod_{m=d}^{d^*} A_{par_{m-d}(i),\widetilde{\Omega}_m+1}^m\left(par_{m-d+1}(i)\right)\pi_{par_m(j)}^m\left[A_{par_{d^*}\left(par_{d^*-1}(i)\right),par_{d^*}\left(par_{d^*-1}(j)\right)}^{d^*}\left(par_{d^*+1}(i)\right)\right]$$

$$i \neq j,$$

where the rate constants are in units of time-steps (4).

References

1. Bishop, C. M. Pattern Recognition and Machine Learning. New York: Springer (2006).
2. Gutiérrez-Peña, E. & Muliere, P. Conjugate Priors Represent Strong Pre-Experimental Assumptions. *Scand. J. Stat.* **31**, 235-246 (2004).
3. Wakabayashi, K. & Miura, T. Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models, in *Advances in Neural Information Processing Systems 25*. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., editors. Curran Associates, Inc., (2012) 1493–1501.
4. Weiland, M., A. Smaill, and P. Nelson. 2005. Learning musical pitch structures with hierarchical hidden markov models. Proc. Journees Informatiques Music. .