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

ABSTRACT Many single-molecule experiments aim to characterize biomolecular processes in terms of kinetic models that specify the rates of transition between conformational states of the biomolecule. Estimation of these rates often requires analysis of a population of molecules, in which the conformational trajectory of each molecule is represented by a noisy, time-dependent signal trajectory. Although hidden Markov models (HMMs) may be used to infer the conformational trajectories of individual molecules, estimating a consensus kinetic model from the population of inferred conformational trajectories remains a statistically difficult task, as inferred parameters vary widely within a population. Here, we demonstrate how a recently developed empirical Bayesian method for HMMs can be extended to enable a more automated and statistically principled approach to two widely occurring tasks in the analysis of single-molecule fluorescence resonance energy transfer (smFRET) experiments: 1), the characterization of changes in rates across a series of experiments performed under variable conditions; and 2), the detection of degenerate states that exhibit the same FRET efficiency but differ in their rates of transition. We apply this newly developed methodology to two studies of the bacterial ribosome, each exemplary of one of these two analysis tasks. We conclude with a discussion of model-selection techniques for determination of the appropriate number of conformational states. The code used to perform this analysis and a basic graphical user interface front end are available as open source software.

INTRODUCTION

Owing to a host of technological innovations over the past two decades, single-molecule techniques are now reaching a level of maturity that makes it possible to perform detailed mechanistic investigations of some of the cell's most fundamental and complex biomolecular processes (1–5). A large class of such single-molecule experiments seeks to establish a kinetic model, defined in terms of a set of structural conformations of the molecule (hereafter referred to as states) and the rates of transition between these states. This kinetic model must be inferred from a set of experimental signal-versus-time trajectories that report on conformational transitions in tens, hundreds, or even thousands of signal trajectories. Unfortunately, however, the analysis of large populations of trajectories presents several challenges that currently impair our ability to accurately infer such kinetic models. Specifically, it remains difficult or impossible to 1), accurately determine the number of states that are present in each noisy signal trajectory; 2), rigorously infer a single kinetic model that is consistent with the entire population of signal trajectories; 3), directly compare kinetic models for populations of trajectories recorded under different experimental conditions; and 4), confidently detect degenerate states that exhibit the same signal output but that differ in their transition rates. Overcoming these challenges, therefore, promises to increase the ease, confidence, and accuracy with which kinetic models can be inferred from this class of single-molecule experiments.

The analysis of individual, noisy signal trajectories has been greatly facilitated by the use of hidden Markov models (HMMs) (6–8). In the biophysical community, these methods were introduced within the context of patch-clamp experiments on ion channels (9–11), and have since been applied within a variety of single-molecule experimental platforms, including optical trapping (12), magnetic tweezers (13), and single-molecule fluorescence resonance energy transfer (smFRET) experiments (14–19). In HMM approaches, a statistical model defines an expected distribution of measurement values in terms of a set of parameters, such as the centers and widths of Gaussian peaks representing the signal values associated with each conformational state, and the transition probabilities between states. Given this model, maximum likelihood (ML) techniques (14,18,20,21), such as those employed in the smFRET data analysis software packages HaMMy (14) and SMART (18), can determine the most likely set of parameters and conformational trajectory for each measured signal trajectory. A well-known deficiency of ML methods, however, is that the likelihood can always be improved by adding more states to the kinetic model, making it difficult to distinguish real conformational states from states that arise from overfitting the inherently noisy individual signal trajectories. Variational Bayesian (VB) techniques (15,16,19,22), such as those employed in the smFRET data analysis software package vbFRET (15,16), improve upon ML methods by introducing a prior distribution, which specifies the expected range of parameter values, allowing maximization of the evidence, a likelihood that is averaged over this prior distribution. Unlike the likelihood, the

Submitted July 24, 2013, and accepted for publication December 31, 2013.
*Correspondence: rlg2118@columbia.edu
Editor: David Rueda

0006-3495/14/03/1327/11 $2.00
http://dx.doi.org/10.1016/j.bpj.2013.12.055

evidence is more likely to peak when the signal trajectory is modeled with the optimal number of states. Thus, VB methods can be used to perform model selection, that is, to determine the number of states that yields the best average agreement between the data and the model (see Methods for further background).

Although maximization of the evidence has proven an effective model-selection strategy, it does not completely eliminate overfitting, and particularly underfitting, of the signal trajectories. For example, single-molecule FRET efficiency (E_{FRET}) trajectories that are particularly noisy (i.e., with a standard deviation in the E_{FRET} value of ~ 0.15) and/or include transitions that are fast relative to the rate of data acquisition (i.e., more than one transition every five time points) are particularly prone to underfitting (15). Moreover, existing ML and VB techniques have an important shortcoming that has significant theoretical and practical implications: they can only be used to model individual signal trajectories, or multiple signal trajectories (17) only if they are modeled with exactly the same parameters. For example, it is a common occurrence that the same state gives rise to a signal centered at $E_{\mathrm{FRET}} = 0.30$ in one trajectory and $E_{\mathrm{FRET}} = 0.35$ in another, due to variations in the the photophysical properties of the fluorophores, slight structural differences in the molecule, and offsetting errors in the measured fluorescence intensity. Although it might be trivial for an experimentalist to recognize that the $E_{\mathrm{FRET}} = 0.30$ and $E_{\mathrm{FRET}} = 0.35$ measurements are different manifestations of the same state, the ML and VB techniques described above cannot model this situation. From a theoretical perspective, it is unsatisfying that the existing algorithms cannot account for such a fundamental component of all real experiments that is obvious to the human eye. From a practical perspective, this shortcoming means that rather than simultaneously modeling a large population of signal trajectories to naturally infer a single kinetic model that is most consistent with the entire population, the experimentalist must instead individually model each trajectory and subsequently perform a significant amount of postprocessing to infer and validate the single, consensus kinetic model.

Recently, we have developed an empirical Bayesian (EB) technique (23,24) that improves upon VB methods by inferring the features of the prior distribution, which in VB methods must be specified by the experimentalist. In EB estimation, the variation in parameter values predicted by the prior distribution is matched to the variation in inferred parameter values over the population of trajectories, enabling a single, consensus kinetic model to be learned from the simultaneous analysis of a large population of signal trajectories (see the Methods section for a more detailed introduction). We have benchmarked this EB technique using computer-simulated data, demonstrating that, relative to both ML and VB methods, it exhibits a greater resistance to both over- and underfitting of signal trajectories, and we have provided a basic example showing that this EB technique can be used to analyze experimental E_{FRET} trajectories (25).

In this article, we use experimental smFRET data reporting on the mechanism of protein synthesis by the bacterial ribosome to demonstrate how our previously developed EB method (25) can be extended to perform two very frequently encountered smFRET data analysis tasks: 1), the comparison of the number of states, their occupancy, and associated transition rates, across experiments recorded for the same biomolecular system but under different experimental conditions (e.g., in the absence, presence, and/or varying concentrations of a particular buffer or biomolecular component), and 2), the detection of states that exhibit the same E_{FRET} value but have different transition rates. Currently, most experimentalists treat these problems by performing inference on the individual trajectories, deciding via a separate assessment (e.g., via a transition density plot (14) or similar (26) metric) how many states they believe are in the data and then binning the inference results in an ad hoc postprocessing step. This process is time-consuming, may be prone to user bias, and lacks metrics for assessing the accuracy of the outcomes. The two extensions of EB estimation presented here, in contrast, allow users to quickly perform analysis in a more automated, statistically rigorous, and reproducible manner, greatly reducing the potential for user bias.

Collectively, the results of these analyses highlight the considerable advantages of EB methods over ML and VB methods and demonstrate how the simultaneous analysis of large populations of signal trajectories using EB methods uniquely enables us to 1), automate identification of a common set of states across various experimental conditions; 2), detect small, but statistically significant, differences in a single state across different experimental conditions; 3), characterize the dependence of the thermodynamic and kinetic properties of states on experimental conditions; and 4), identify kinetically distinct subpopulations within a single experiment.

METHODS

Bayesian inference in coupled HMMs

Bayesian inference seeks to determine the probability of a set of unknown variables in light of a set of observed data. In the context of single-molecule studies, these unknown variables are a set of model parameters θ and a state sequence z_t, whereas the observations are a signal trajectory, x_t. A graphical model defines a statistical relationship between these variables that can commonly be factored into two terms

$$p(x, z, \theta|\psi_0) = p(x|z, \theta)\, p(z, \theta|\psi_0). \quad (1)$$

The two distributions $p(x|z, \theta)$ and $p(z, \theta|\psi_0)$, known as the likelihood and prior distribution, respectively, describe our assumptions about the model. The likelihood describes the measurement signal we expect to see given the state trajectory, z_t, of the molecule and a set of emission model

parameters that describe the distribution of measurement values associated with each state. The prior distribution encodes our expectations about the transition probabilities and emission model parameters. Based on these assumptions, the goal of Bayesian inference is now to reason about the so-called posterior probability of the state trajectory (z_t) and model parameters (θ) in light of a set of measurements (x_t). Bayes' rule states that this posterior probability $p(z,\theta|x,\psi)$ can be expressed as

$$p(z,\theta|x,\psi_0) = \frac{p(x|z,\theta)\,p(z,\theta|\psi_0)}{p(x|\psi_0)}. \quad (2)$$

The prior distribution for an HMM can be written as $p(z,\theta|\psi_0) = p(z|\theta)\,p(\theta|\psi_0)$, where the probability $p(z|\theta)$ depends on two model parameters. The first is a transition matrix, A_{kl}, that specifies the probability of entering state l from state k at any given time. The second is a set of probabilities π_k that specify the likelihood of starting in state k. The form of the likelihood $p(x|z,\theta)$ depends on the type of experimental technique considered. In the case of smFRET experiments, a common approach (14–16,18,25) is to model the signal for each state k as a Gaussian peak with center μ_k and width σ_k, or precision $\lambda_k = 1/\sigma_k^2$. The parameters that describe any given trajectory are therefore $\theta = \{\mu,\lambda,A,\pi\}$. The prior distribution $p(\theta|\psi_0)$ on the parameters can itself be defined in terms of a set of hyperparameters $\psi_0 = \{m_0,\beta_0,a_0,b_0,\alpha_0,\rho_0\}$ (see the Supporting Material).

The structure of the probabilistic relationships that define an HMM can be represented as a network, or more precisely as a directed acyclic graph (22,27). In this network, the nodes are individual variables and edges signify dependencies. Such a graphical model for a coupled HMM on N trajectories with K states is shown in Fig. 1. The dependency structure between variables in this model reflects three fundamental assumptions about the data: 1), at each time, there is a fixed probability of entering into a given state, which depends only on the current state, and has no memory of earlier parts of the state trajectory; 2), observations associated with a given state are independent and identically distributed; and 3), the parameters θ_n of each trajectory are coupled through a shared prior, $p(\theta_n|\psi_0)$, whose distribution reflects the variability of parameter values in an experiment.

The main difficulty in Bayesian inference is that the posterior $p(z,\theta|x,\psi_0)$ can typically not be calculated directly. This is because the normalizing term $p(x|\psi_0)$ in Eq. 2, known as the evidence, involves an intractable integral. In the EB approach used here, we approximate the evidence $p(x|\psi)$ with the same techniques as those employed in VB estimation: we use a pair of distributions $q(z)$ and $q(\theta|\psi)$ to approximate the posterior with a factorized form:

$$p(z,\theta|x,\psi_0) \simeq q(z)\,q(\theta|\psi). \quad (3)$$

Whereas ML methods obtain a point estimate for the optimal parameters θ, this approach yields a distribution $q(\theta|\psi)$ defined in terms of a set of posterior parameters ψ. The relationship between ψ and ψ_0 reflects an important principle of Bayesian statistics. The posterior parameters have the same form as the prior parameters, but define a more tightly peaked distribution that reflects our increased knowledge in light of the measurements. More precisely put, ψ can be calculated from a set of sufficient statistics, $\mathcal{T}$ (see section S2 in the Supporting Material). For an HMM, these statistics are given by

$$\gamma_{tk} = E_{q(z)}[z_{tk}],\ \xi_{kl} = \sum_t E_{q(z)}\left[z_{(t+1)l}z_{tk}\right], \quad (4)$$

$$\Gamma_k = \sum_t \gamma_{tk},\ X_k = \sum_t \gamma_{tk}x_t,\ U_k = \sum_t \gamma_{tk}x_t^2. \quad (5)$$

The statistics $\mathcal{T} = \{\gamma,\xi,\Gamma,X,U\}$ summarize the information contained in each trajectory in terms of the amount of time spent in each state, Γ_k, the number of transitions between states, ξ_{kl}, the mean X_k/Γ_k measurement value for each state, and its variance, $U_k/\Gamma_k - (X_k/\Gamma_k)^2$.

The posterior parameters can be calculated directly from the sufficient statistics and the prior parameters (for details, see section S3.3 of the Supporting Material). For example, the posterior for the transition probabilities $q(A|\alpha)$,

$$\alpha_{kl} = \xi_{kl} + \alpha_{0,kl}, \quad (6)$$

is simply the sum of the number of transitions ξ that we believe we have seen in the trajectory, and the equivalent number of transitions of the prior α_0.

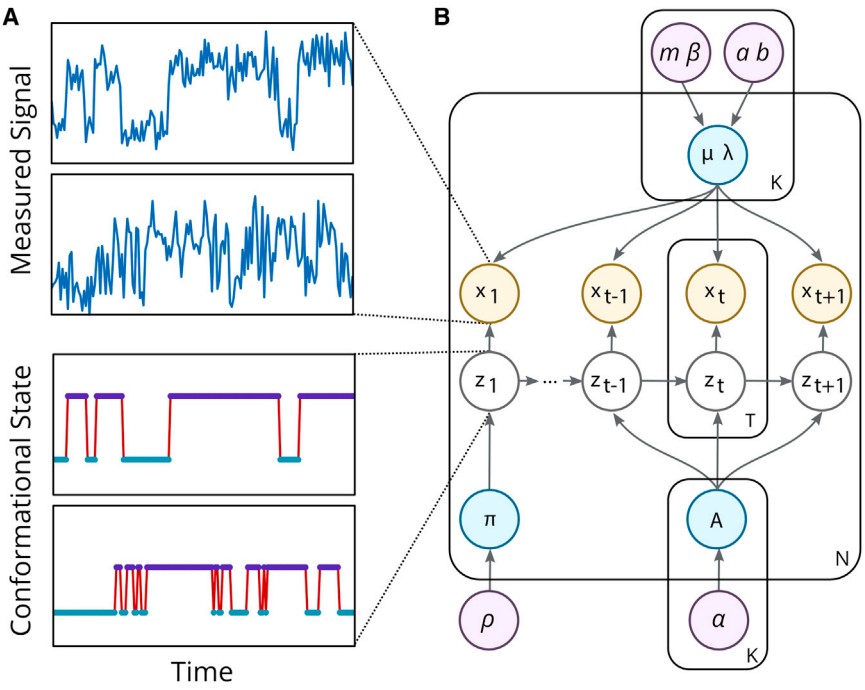

FIGURE 1 Graphical model for the coupled Bayesian HMM used in EB and VB methods. (*A*) smFRET signals and sequence of latent states for two trajectories in an experiment. (*B*) Graphical model showing an HMM for N trajectories with K states. The parameters $\theta_n = \{\mu_{n,k},\lambda_{n,k},A_{n,kl},\pi_{n,k}\}$ of each trajectory are distributed according to $p(\theta|\psi)$ with hyperparameters $\psi = \{m_k,\beta_k,a_k,b_k,\alpha_{kl},\rho_k\}$. ML methods use a non-Bayesian variant of this HMM, which omits the hyperparameters, ψ. To see this figure in color, go online.

In general, placing a prior on the parameters is equivalent to assuming that one has already seen a number of data points with statistics $\mathcal{T}_0$ before seeing the measurements, x_t. The number of equivalent observations associated with $\mathcal{T}_0$ determine how quickly the posterior will change in light of new observations.

EB estimation (23–25) extends VB estimation to perform simultaneous inference on populations of trajectories. To do so, we learn N approximate posterior distributions $q(\theta_n|\psi_n)$ for each trajectory x_n. The prior, $p(\theta|\psi_0)$, is subsequently chosen by way of a self-consistency requirement; the range of θ_n values predicted by the posterior distributions should match that of the prior. This is equivalent to choosing a set of prior parameters whose distribution is as close as possible to the average posterior (see section S4 of the Supporting Material).

In a mathematical sense, this estimation procedure approximates the log evidence $\log p(x|\psi_0)$ with a lower bound L,

$$L = \sum_n E_{q(z_n)q(\theta_n|\psi_n)}\left[\log \frac{p(x_n, z_n, \theta_n|\psi_0)}{q(z_n)q(\theta_n|\psi_n)}\right], \tag{7}$$

by iteratively finding solutions to the equations

$$\frac{\delta L}{\delta q(z_n)} = 0, \quad \frac{\delta L}{\delta q(\theta_n|\psi_n)} = 0, \quad \frac{\partial L}{\partial \psi_0} = 0. \tag{8}$$

A full derivation of each of these update steps in this algorithm can be found in sections S3 and S4 of the Supporting Material for this article.

In summary, the EB approach to kinetic analysis uses HMMs to calculate two sets of quantities. For each trajectory, we obtain a set of trajectory statistics, $\mathcal{T}_n$, which report on the occupancy, transitions and measurement values associated with each state. The second quantity is a set of prior parameters, $\psi_0 = \psi(\mathcal{T}_0)$, which represent the characteristics that all signal trajectories have in common. Finally, a set of posterior parameters, $\psi_n = \psi(\mathcal{T}_n + \mathcal{T}_0)$, encodes what we know about the parameters of individual trajectories in light of the measured signal. Note that the prior parameters ψ_0 can be equivalently defined in terms of a set of prior statistics, $\mathcal{T}_0$, whereas the posterior statistics are simply the sum of the prior statistics and the trajectory statistics.

We reiterate that EB estimation differs from VB estimation only in the fact that the hyperparameters ψ_0 are not chosen by the user and held fixed, but are set to the values that maximize the evidence as part of the inference procedure. This allows for more accurate inference, as knowledge of typical parameter values results in better estimates of $\mathcal{T}_n$. Moreover, since the learned EB prior is typically less broadly peaked than the postulated prior in VB methods, the effective number of observations for each posterior is larger, resulting in tighter confidence bounds on parameter estimates for individual trajectories (25). Indeed, past analysis of simulated data, for which the true state sequence is known, has shown that EB inference systematically outperforms VB and ML methods, in terms of both parameter estimation and model-selection tasks (25).

Analysis of labeled and unlabeled subpopulations of signal trajectories

In this section, we extend the EB method to perform commonly occurring advanced analysis tasks, which we illustrate in the next sections using two experimental smFRET studies that each investigate aspects of translation, the mechanism by which the bacterial ribosome synthesizes the protein that is encoded by a messenger RNA (mRNA) template (see Tinoco and Gonzalez (1) for a review). The goal of analysis in the first example is to coherently detect the set of states that can be sampled across experiments performed in the presence and absence of other biomolecular components, and subsequently separately estimate the transition rates for each experiment. In the second example, our goal is to extend the EB method to detect subpopulations of trajectories that sample the same two states, but to do so using different transition rates.

The common denominator in both these analysis tasks is that we seek to use measurements of large populations of trajectories to identify a common set of states and determine how transition rates differ for subpopulations of molecules within this aggregate data. In the case of the first set of experiments, we have labeled subpopulations consisting of sets of signal trajectories recorded under identical experimental conditions, and we simply wish to obtain per-experiment estimates of the transition rates based on a shared definition of states. In the case of the second study, each experiment contains two unlabeled subpopulations and the set of signal trajectories associated with each subpopulation must be inferred from the data.

To allow more straightforward analysis of labeled and unlabeled subpopulations, we will extend the EB estimation procedure in the following manner. Rather than estimate a single set of prior parameters, ψ_0, from the trajectory statistics, $\mathcal{T}_n$, we split our population in into M fractions with prior parameters ψ_{0m}. We introduce a new variable, y_{nm}, for the population membership of each signal trajectory. This variable is simply a binary indicator that is 1 if trajectory n is part of population m. For labeled populations, the values for y are known, and we can estimate distributions for individual populations from the restricted set of posterior distributions

$$p(\theta|\psi_{0m}) \simeq \sum_n y_{nm} q(\theta|\psi_n) / \sum_n y_{nm}. \tag{9}$$

In the case of unlabeled subpopulations, y must be inferred from the data. To do so, we generalize the EB approach to a mixture of distributions, $p(x_n|\psi_{0m})$, where we assume a discrete prior, $p(y|\phi)$, on the subpopulation membership. The evidence can now be expressed as a marginal over all possible y values,

$$p(x|\psi_0) = \sum_y p(x|y, \psi_0) p(y|\phi), \tag{10}$$

$$= \sum_n \sum_{y_n} \prod_m p(x|\psi_{0m})^{y_{nm}} \phi_m^{y_{nm}}. \tag{11}$$

An expectation maximization algorithm over this mixture can be constructed by introducing a variational posterior $q(y)$ and maximizing the lower bound,

$$L = E_{q(z|y)q(\theta|y)q(y)}[\log p(x, y, z, \theta|\psi_0)]. \tag{12}$$

We can subsequently estimate the statistic $\omega_{nm} = E_{q(y)}[y_{nm}]$ from the lower bounds, $L_{nm} \geq \log p(x_n|\psi_{0m})$

$$\omega_{nm} = \frac{\exp(L_{nm})\phi_m}{\sum_{m'} \exp(L_{nm'})\phi'_m}. \tag{13}$$

In the resulting EB procedure, the expectation values with respect to the approximate posteriors are now weighted by the population weights (see section S4.5 of the Supporting Material)

$$p(\theta|\psi_{0m}) \simeq \sum_n \omega_{nm} q(\theta|\psi_{nm}) / \sum_n \omega_{nm}. \tag{14}$$

Software implementation

All analysis algorithms are implemented in MATLAB, with essential inner components (i.e., the forward-backward and viterbi algorithms) written in C as MEX files. Our implementation uses multiple processors when available. We performed a simple benchmark in Matlab 2013a on a Macbook equipped with a four-core 2.3GHz Core i7 processor, using a computer-simulated data set with $N = 350$ trajectories of average length $T = 112$. Analysis with two to six states required 240 s using eight nodes and 600 s using a single

node. In comparison, our previously released vbFRET software (15) required 1500 s to analyze the same data set on the same machine.

A line-by-line derivation of the implemented EB estimation algorithm and its extensions can be found in the Supporting Material. A command-line version of the source code used in this publication, along with a GUI frontend for basic EB estimation tasks, is available at http://ebfret.github.io. This software supports a new single-molecule data format that has been designed in collaboration with the Herschlag group at Stanford to enable exchange of data and analysis results between research groups (M. Greenfeld, J.-W. van de Meent, D. S. Pavlichin, H. Mabuchi, C. H. Wiggins, R. L. Gonzalez Jr., and D. Herschlag, unpublished).

RESULTS

Labeled subpopulations: The role of IF3 conformational dynamics in regulating translation initiation

We begin by showing how the extended EB estimation procedure described by Eq. 9 can be used to characterize the dependence of conformational state occupancies, emission model parameters, and transition probabilities on experimental conditions. We do so by analyzing a set of previously published smFRET (29) experiments that investigate the role of initiation factor (IF) 3 in regulating the fidelity with which the bacterial ribosome initiates translation at the triplet-nucleotide start codon of the mRNA to be translated.

During bacterial translation initiation, the small, or 30S, ribosomal subunit, IF1, IF2, IF3, a specialized formylmethionyl initiator transfer RNA (fMet-tRNA^{fMet}), and the mRNA to be translated form a 30S initiation complex (30S IC) in which the triplet-nucleotide anticodon of fMet-tRNA^{fMet} is basepaired to the mRNA start codon within the peptdiyl-tRNA binding (P) site of the 30S subunit (30). Subsequent joining of the large, or 50S, ribosomal subunit to the 30S IC results in the formation of a translation-elongation-competent 70S initiation complex (70S IC). Because errors in fMet-tRNA^{fMet} or start-codon selection can result in mistranslation of the mRNA sequence, regulating the fidelity of initiation is crucial to protein synthesis and cellular fitness. Thus, the major role of IF1, IF2, and IF3 during translation initiation is to control the fidelity of this process by, among other mechanisms, coupling the 50S-subunit-joining step of the initiation process to the correct selection of fMet-tRNA^{fMet} and the start codon; the role of IF3 in this mechanism is to prevent 50S subunit joining until fMet-tRNA^{fMet} and the start codon have been correctly selected into the P site.

Here, we present analysis of smFRET experiments investigating the role that IF3 conformational dynamics plays in coupling correct fMet-tRNA^{fMet} and start codon selection to 50S subunit joining (29). IF3 is composed of two globular domains connected by a flexible linker. When these domains are labeled with FRET donor and acceptor fluorophores, the value of $E_{FRET} = I_A/(I_D + I_A)$, where I_A and I_D are the emission intensities of the acceptor and donor fluorophores, respectively, provides a noisy measure of the intramolecular distance between the two domains. Histograms of the observed E_{FRET} values (Fig. 2 *A*) show two dominant peaks, corresponding to a low-FRET extended conformational state, and a high-FRET compact conformational state of 30S IC-bound IF3, whose relative occupancies depend on the presence of the other IFs and fMet-tRNA^{fMet} on the

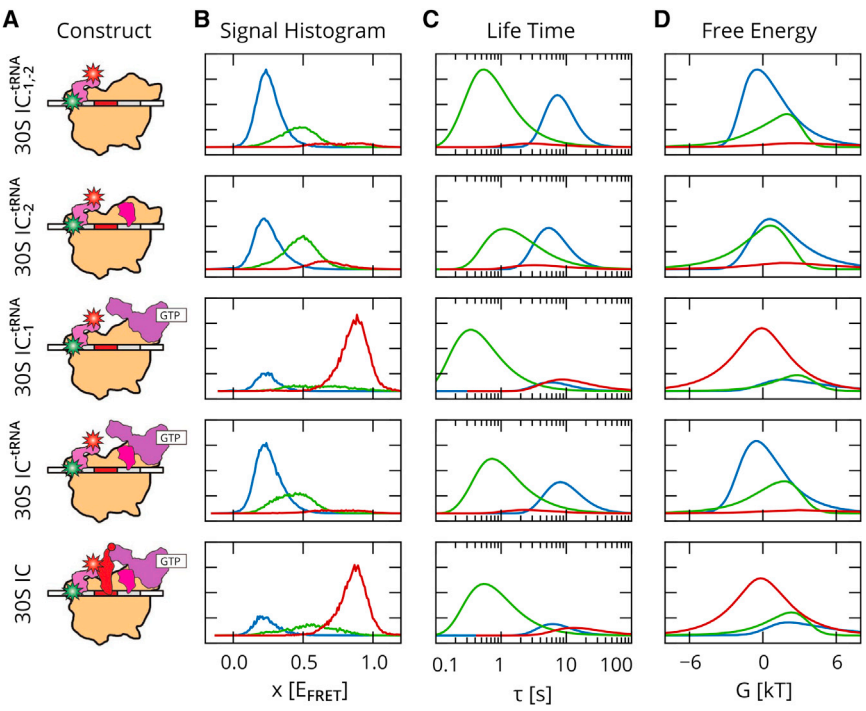

FIGURE 2 smFRET study of IF3 conformational dynamics on the 30S initiation complex of the bacterial ribosome. (*A*) Schematic illustrations of experimental contsructs 30S IC$^{-tRNA}_{-1,-2}$, 30S IC$^{-tRNA}_{-2}$, 30S IC$^{-tRNA}_{-1}$, 30S IC^{-tRNA}, and 30S IC^{fMet}. (*B*) Per-state observation histograms. (*C*) Lifetime distributions. (*D*) Free-energy distributions. States 1–3 are represented by blue, green, and red lines, respectively. To see this figure in color, go online.

IC. In addition to these two states, there appear to be one or more intermediate conformational states, which tend to be shorter-lived and have E_{FRET} values that are less well-defined.

Previous analysis was performed with the vbFRET software (15) that obtains VB estimates for each individual E_{FRET} trajectory. In this particular set of experiments, most trajectories are static (i.e., no conformational transitions are observed before the fluorophores photobleach). This makes it more difficult to distinguish between intermediate and extended or compact states, since within individual trajectories, there are few transitions that reveal the location of a state relative to others. For this reason, the resulting E_{FRET} means of states in each trajectory were assigned to three empirically chosen bins with intervals (0,0.3), (0.3,0.7), and (0.7,1.0), where all potential intermediate states were grouped into the middle interval. The compact state was found to be highly populated in a correctly assembled 30S IC, whereas the extended state is highly populated in incorrectly assembled or incomplete 30S ICs, which either lack IFs, lack fMet-tRNA^{fMet}, contain an incorrect elongator aminoacyl-tRNA, or contain an incorrect near-start codon (29).

In our analysis, we first performed EB inference on the aggregate data from five experiments that were recorded under different conditions: 30S IC$^{-tRNA}_{-1,-2}$ (lacking IF1, IF2, and tRNA), 30S IC$^{-tRNA}_{-2}$ (lacking IF2 and tRNA), 30S IC$^{-tRNA}_{-1}$ (lacking IF1 and tRNA), 30S IC^{-tRNA} (lacking tRNA), and 30S IC^{fMet} (a correctly assembled 30S IC). This aggregate dataset contained 4233 trajectories with $4.0 \cdot 10^5$ total data points. Three states were used to facilitate comparison with the previous results based on VB analysis. After inference, separate parameter distributions were estimated from the sufficient statistics of each individual experiment, as described in Eq. 9. The results of this analysis, which does not require that the user manually assign the E_{FRET} means of states in each trajectory to empirically chosen bins, are in excellent agreement with previous results based on explicitly defined bin intervals. Fig. 2 shows observation histograms for each state, as well as distributions of the lifetime and free energy of each state relative to the other states (see section S5 of the Supporting Material for the definitions of these quantities). The width of each distribution provides us with a confidence interval on each of the parameters. The fractional occupancies obtained for each experiment (Table 1) similarly show a close correspondence to the values obtained with the VB-based results.

TABLE 1 Relative occupancies of IF3 states obtained from VB and EB- analysis of labeled subpopulations

	VB + binning			EB		
Construct	ext.	int.	cpt.	ext.	int.	cpt.
30S IC$^{-tRNA}_{-1,-2}$	0.54	0.40	0.06	0.63	0.30	0.07
30S IC$^{-tRNA}_{-2}$	0.52	0.45	0.03	0.47	0.43	0.10
30S IC$^{-tRNA}_{-1}$	0.23	0.11	0.66	0.14	0.15	0.72
30S IC^{-tRNA}	0.56	0.42	0.02	0.60	0.34	0.06
30S IC^{fMet}	0.15	0.17	0.68	0.15	0.21	0.64

Relative occupancies of the extended (*ext.*), intermediate (*int.*), and compact (*cpt.*) conformations of IF3, obtained from binned analysis with vbFRET (29) and EB-based analysis of labeled subpopulations.

Unlabeled subpopulations: the influence of EF-G binding on the GS1-GS2 equilibrium

We now demonstrate that the extended EB estimation procedure described by Eq. 14 can be used to identify kinetically distinct subpopulations of states and estimate the transition rates for each subpopulation of states. As an example of this use case, we perform analysis of a set of smFRET experiments investigating the role of elongation factor (EF) G, a member of the GTPase family of translation factors, during translation elongation.

After the addition of each amino acid to the nascent polypeptide chain during translation elongation, EF-G binds to the ribosomal pretranlsocaiton (PRE) complex and hydrolyzes one molecule of GTP as it promotes the movement of the ribosome along the mRNA by precisely one triplet-nucleotide codon, a process termed translocation (Fig. 3 *A*). The overall process of translocation can be

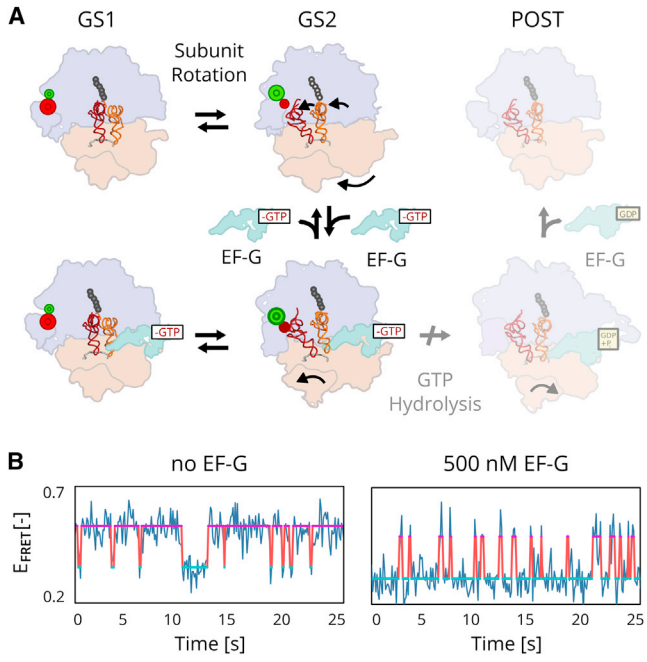

FIGURE 3 smFRET experiments (31) measuring the influence of EF-G on the GS1-GS2 equilibrium in the bacterial ribosome. (*A*) The kinetic pathway for translocation is believed to have three steps: a reversible rotation of the two subunits (*purple* and *orange*), followed by the binding of EF-G (*green*), which stabilizes the rotated GS2 state long enough for a GTP-driven transition to the posttranslocation (*POST*) complex, blocked here by substitution of GTP by a nonhydrolyzable analog. (*B*) smFRET signals reporting on the GS1-GS2 transition show a shift of the equilibrium from the GS1 state (*magenta line*) toward the GS2 state (*cyan line*) in the presence of EF-G. To see this figure in color, go online.

broken up into three smaller multistep processes. The first of these is a thermally driven, reversible transition between two global states (denoted as GS1 and GS2) of the PRE complex. The overall process of translocation can be broken up into three smaller multistep processes. This conformational transition is followed by binding of EF-G to the PRE complex, resulting in a transient stabilization of the GS2 state of the PRE complex that is long enough to enable the third step, a GTP hydrolysis-driven movement of the ribosome along the mRNA. The effect that binding of EF-G has on the dynamic equilibrium between the GS1 and GS2 states of the PRE complex can be studied using smFRET by labeling two ribosomal structural elements with a FRET donor-acceptor pair and substituting GTP with a nonhydrolyzable analog (GDPNP) that prevents GTP hydrolysis and the associated movement of the ribosome along its mRNA template.

Fig. 3 *B* shows two E_{FRET} trajectories that exhibit thermally driven, reversible transitions between GS1 and GS2. The first trajectory is from an experiment that was recorded in the absence of EF-G and shows a preference for the GS1 state. The second trajectory, from an experiment that was recorded in the presence of 500 nM EF-G and 1 mM GDPNP, shows a dramatic shift of the equilibrium toward the GS2 state. Qualitative comparison of these two trajectories suggests that EF-G destabilizes the GS1 state and stabilizes the GS2 state in the subpopulation of EF-G-bound PRE complexes. To quantify this difference in transition rates and characterize its dependence on EF-G concentration, we must obtain separate estimates for the distribution on transition rates for the EF-G-free and EF-G-bound subpopulations of PRE complexes in an experiment.

EB analysis of a series of experiments performed at increasing EF-G concentrations is shown in Fig. 4. As with the previous experiment we first analyze the aggregate data to identify two states. The aggregate data for seven different EF-G concentrations contained 2472 trajectories with 2.3×10^5 total data points. As can be seen in the observation histograms (Fig. 4 *A*), the occupancy of the GS2 state (*cyan line*) increases with the EF-G concentration. Conventional EB analysis with a single population (Fig. 4 *B*) naturally reveals a bimodal signature in the posterior (*solid lines*)

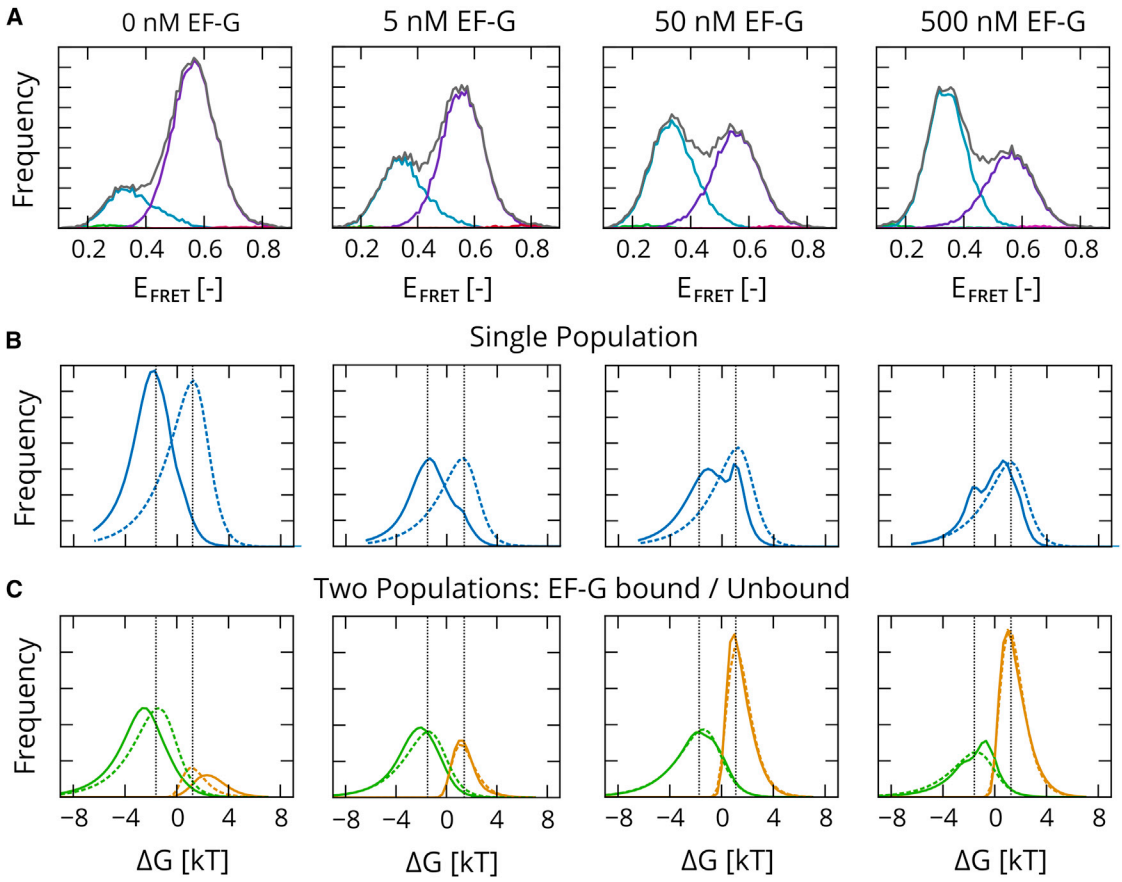

FIGURE 4 Analysis of GS1-GS2 equilibrium as a function of EF-G concentration. (*A*) Histogram of aggregate measurements, split by GS1 (*magenta line*) and GS2 (*cyan line*) states. (*B*) EB prior (*dashed line*) and mean posterior (*solid line*) on the free-energy difference $\Delta G = G_{GS1} - G_{GS2}$. A bimodal signature in the posterior is visible in experiments where EF-G is present. (*C*) Prior and posterior after unlabeled subpopulation analysis, showing an increasing occupancy of the bound fraction (*orange line*) relative to the nonbound fraction (*green line*) as a function of EF-G concentration. To see this figure in color, go online.

that hints at the existence of two (unlabeled) subpopulations. This signature is absent from the prior (*dashed lines*), since EB analysis assumes all transition probabilities are governed by the same prior distribution. Because a very limited number of transitions between GS1 and GS2 can be observed in any one signal trajectory before one of the fluorophores photobleaches, it is not possible to obtain a precise estimate of the transition rates for each individual PRE complex. As a result, the two peaks in Fig. 4 *B* have a very high degree of overlap, showing that it would be difficult to determine the population membership for each signal trajectory using any form of binning approach. This ambiguity of subpopulation membership is greatly reduced when using the subpopulation analysis technique described in the previous section (see also Section S4.5 of the Supporting Material), which produces two much-better-resolved peaks (Fig. 4 *C*). Table 2 lists the population fraction and free energy difference obtained from EB estimation with unlabeled subpopulations. As should be expected, the relative size of the EF-G-bound subpopulation increases as the concentration of EF-G increases.

Model selection

One of the stated advantages of the VB and EB methods is that they optimize a lower bound for the log evidence, a quantity that may be used to decide among analysis results with different numbers of states. Previous benchmarks using computer-simulated data have shown that EB estimation systematically outperforms VB and ML methods in model-selection tasks (25). Not only does EB estimation more accurately determine the number of states in individual trajectories, preventing both under- and overfitting, but the method can also determine the correct number of states starting from a larger number of candidate states, leaving superfluous states unpopulated.

In practice, experimental data differ from simulated data in that they are never in precise agreement with a given statistical model. In smFRET experiments, for example, we assume a Gaussian distribution of E_{FRET} values for each state. With one exception (17), all HMM approaches for analysis of (time-binned) smFRET data make this same assumption (14–16,18). In reality, however, the E_{FRET} value exhibits a sigmoidal dependence on the distance between the fluorophores, resulting in a distribution of E_{FRET} values that is skewed toward the middle of the spectrum and exhibits a subtle, but systematic, deviation from the idealized Gaussian shape assumed in the model. Distributions of E_{FRET} values further show heavy tails that likely arise from artifacts such as intermittent photoblinking of fluorophores (32), incorrect detection of the photobleaching transition, and errors in determining the background fluorescence intensity of individual trajectories.

TABLE 2 EF-G concentration dependence in unlabeled subpopulation analysis of GS1-GS2 equilibrium

EF-G	0 nM	5 nM	50 nM	500 nM	1000 nM
$\rho_{+\mathrm{EF-G}}$	0.13	0.30	0.56	0.65	0.67
$\Delta G_{+\mathrm{EF-G}}$	1.7	1.2	1.3	1.4	1.4
$\Delta G_{-\mathrm{EF-G}}$	−2.4	−1.7	−0.8	−0.4	−0.4

Fraction of EF-G bound complexes, $\rho_{+\mathrm{EF-G}}$, and the free energy difference between the GS1 and GS2 state, ΔG, for the bound and unbound subpopulation, as a function of EF-G concentration.

In general, systematic discrepancies and artifacts can cause a statistical algorithm to correct for the fact that observed measurement values are not precisely distributed according to the assumed model by populating additional states, as was found to be the case in our initial analysis of experimental data (25). In Fig. 5, we revisit this notion by examining results obtained by estimating models with 2–10 states on the same two data sets that were analyzed in the previous sections. As in previous work (25), we calculate an effective number of states, $K_{\mathrm{eff}} = \exp[-\sum_{k=1}^{K}\zeta_k \log \zeta_k]$, in terms of $\zeta_k = \sum_n \Gamma_{nk} / \sum_{nk}\Gamma_{nk}$, the fraction of time points assigned to each state. When performing analysis on simulated data, there is typically a range of solutions for different K that yield the same (correct) K_{eff} value and leave any additional states empty (25). Consistent with our previous study (25), the results in Fig. 5 *A* show that K_{eff} steadily increases with the number of candidate states, and it is not clear that there is an optimum K_{eff} value beyond which the lower bound, L, decreases. In other words, the fit of experimental data to the model can be improved by adding incremental low-occupancy states that capture outliers in the data, even when using model-selection criteria. This is undesirable behavior, as such outlier states are more likely to be indicative of measurement artifacts than of actual conformational states of interest. However, it is important to note that this behavior is different from the typical overfitting that is associated with ML estimation. ML methods obtain a better fit by assigning natural statistical variations to separate states, and will do so even for simulated data that is in perfect agreement with the hypothesized model. EB analysis generally obtains the correct result on simulated data but uncovers unnatural variations in experimental data that are real from a statistical point of view but do not contain useful information about actual conformational transitions.

Examples of these systematic discrepancies can be seen in Fig. 5 *B*, which shows the averaged posterior distribution on the state centers, μ_{nk}, and state dwell times, τ_{nk}, obtained by analyzing the aggregate data sets from the previous sections with increasing number of states. When plotted on a logarithmic scale, a Gaussian distribution will have a parabolic shape. The curves for μ_{nk} clearly show both asymmetries and aberrant tails that deviate from this idealized form. As a result, it is generally difficult to say whether too many states are used, since the curves obtained at higher K do show a closer agreement with the shape assumed in the model.

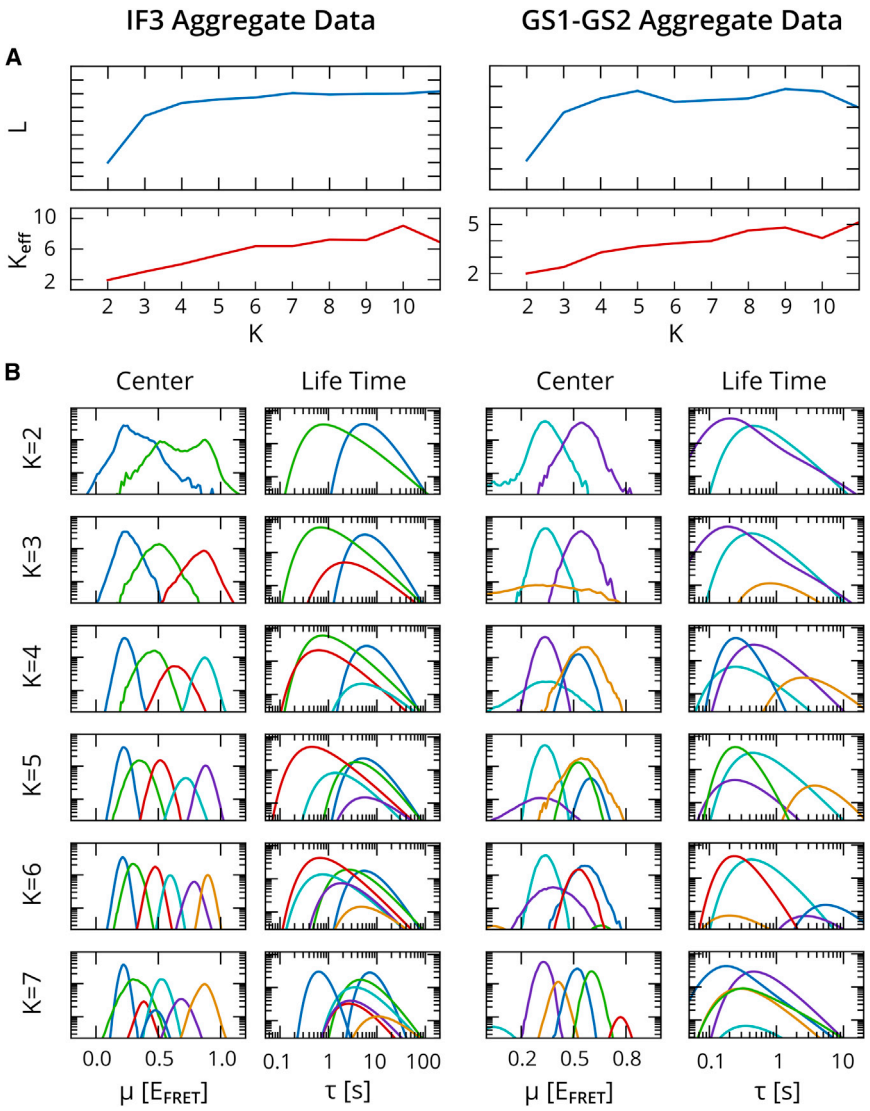

FIGURE 5 EB analysis of IF3 and GS1-GS2 aggregate data for an increasing number of states, K. (*A*) Evidence lower bound L and effective number of populated states, K_{eff}, as a function of K. (*B*) Averaged posterior on state centers, μ, and lifetimes, τ. To see this figure in color, go online.

For this reason, we suggest that users do not indiscriminately rely on the lower bound for model selection; thus, some prudent decision-making with regard to model selection may still be required on the part of the experimentalist. One rule of thumb is to treat states observed in <5% of the trajectories with some caution. Additional states may simply 1), capture artifacts, such as intermediate points between a transition (15); 2), split a single state into a short-lived and long-lived variant (which may mean that a subpopulation as described in Methods is necessary); or 3), isolate the non-Gaussian tails of actual states. Moreover, any decreases in the lower bound indicate that the method has converged to a local maximum rather than the globally optimal result, since adding an empty state to the previous result should result in the same, larger L value. In this case, the user may either opt to perform additional restarts with random initializations of ψ_0, to make it more likely that the global optimum is found for each number of candidate states, or accept the point where L begins to decrease as a bound on the number of states that can be confidently inferred, given computational limitations. As an example, the GS1/GS2 experiment shows a decrease in L at $K = 6$, whereas the lifetime plot for the blue state falls off the scale at $K = 5$, suggesting that $K = 4$ is the largest number of states that is credible. Also note that these four states form two pairs with similar E_{FRET} values but different lifetimes, which is consistent with our knowledge that this experiment in fact does contain kinetically distinct subpopulations. Finally, we note that the conformational trajectory can be inferred with more confidence when more transitions are observed, as it allows the inference procedure to more confidently situate one state relative to others. In cases such as the IF3/30S IC experiment, where the majority of trajectories do not exhibit transitions, analysis results could be improved by shuttering the excitation source to, optimally, obtain a state lifetime of the order of 10 time points.

In summary, although EB methods provide model-selection criteria that are superior to those employed in ML and

VB estimation (when applied to computer-simulated data), a methodological caveat in any statistical analysis is that model-selection criteria are only as accurate as the representation of the measurement data in the model. We emphasize that this limitation is by no means unique to EB analysis. ML and VB approaches typically use precisely the same Gaussian distribution of measurement values and suffer from the same defects. It is merely the case that these issues are obfuscated when signal trajectories are analyzed individually, since an individual signal trajectory rarely contains enough data points to make discrepancies between the data and the model apparent, and the experimentalist makes a judgment call as to how many conformational states they think are required as part of the data inference postprocessing. The advantage of the EB methodology is that analyzing all trajectories at once allows us to identify systematic deviations between data and model, allowing us to assess whether there is sufficient agreement between the data and the model for model-selection criteria to be effective.

DISCUSSION

Although HMMs have proven an immensely popular and effective tool for inferring states and transition rates from individual signal trajectories, combining results from the analysis of multiple trajectories has remained a difficult task. Typically, users manually specify a set of bin intervals, as was done in our previous, VB-based analysis of the IF3 data (29), that allow states identified in individual signal trajectories to be clustered according to their inferred parameter values. In contrast, the EB method uniquely enables simultaneous inference on multiple signal trajectories in a statistically robust manner that eliminates the need for user-defined bin intervals and is consequently less prone to user bias.

By exploiting the advantages of simultaneously analyzing multiple E_{FRET} trajectories using the EB method, we have developed estimation procedures that uniquely enable us to automate widely encountered tasks in the analysis of smFRET experiments. The first of these tasks is exemplified by our analysis of the IF3 experiments, which demonstrates how E_{FRET} trajectories from a large number of experiments recorded under different experimental conditions can first be simultaneously analyzed to identify a common set of states and then be subsequently reanalyzed to calculate a separate prior distribution for each experiment, allowing characterization of how the state occupancies and transition rates vary between experiments. The second task is exemplified by our analysis of the GS1/GS2 experiments, which demonstrates how the simultaneous analysis of an entire population of E_{FRET} trajectories can be used to automatically identify and characterize subpopulations of molecules occupying functionally and/or conformationally distinct states that exhibit similar E_{FRET} values but differ in the rates of transitions between states.

For each set of experiments, the results of the EB-based analysis are largely consistent with previous results based on VB methods. However, although the previous VB-based analysis required the use of experiment-specific postprocessing procedures that are time-consuming to implement, subject to user bias, and difficult to validate, our EB method can be used to obtain results rapidly and with little to no manual intervention by the user. Moreover, the EB approach optimizes a well-defined, statistical, model-selection criterion, the lower bound for the log evidence, which in principle can be used to compare and decide among different analyses of the same data.

Our EB-based analysis of smFRET data also demonstrates that comparing the prior and posterior distributions can often provide useful qualitative diagnostics that indicate whether a given model is appropriate for the data. In the case of the GS1/GS2 experiments, for example, we are able to calculate a posterior distribution on the free-energy difference between states that reveals a systematic mismatch between the single population of PRE complexes that is assumed in conventional EB analysis and the two subpopulations of PRE complexes that are actually present in the experiment (i.e., EF-G-free and EF-G-bound). This mismatch is resolved when we extend our EB method to identify the two subpopulations within the set of multiple E_{FRET} trajectories. In a similar way, combining results from multiple trajectories using our EB method allows us to see that the distribution of E_{FRET} values associated with a given conformational state often exhibits heavy tails and is skewed relative to the Gaussian distribution that is typically assumed in HMM analyses of smFRET data. Whereas discrepancies between the data and the statistical model will always exist, they are much more difficult to detect in individual trajectories (e.g., in ML- and VB-based HMM analyses of smFRET data). An important advantage of the EB method, therefore, is that it can tease out such discrepancies, which inform us as to how our assumptions about the data need to be adjusted in the next iteration of statistical model design.

We conclude by noting that the EB estimation framework is applicable to a wide range of single-molecule techniques. Although here we have analyzed smFRET experiments exclusively, our approach is by no means restricted to this platform. Adaptation of the EB algorithm presented here to the analysis of optical trapping and magnetic tweezers experimental data is possible with minimal modifications and we have recently collaborated to adapt the EB algorithm presented here to the analysis of tethered particle motion experiments (33).

SUPPORTING MATERIAL

A detailed model specification and derivation of update equations is available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00143-X.

The authors thank Margaret Elvekrog, Kevin Emmett, Jingyi Fei, Jason Hon, Daniel MacDougall, Jordan McKittrick, and the reviewers, for their comments on this manuscript. It is also our pleasure to acknowledge helpful discussions with Martin Lindén, Frank Wood, Matt Hoffman, and David Blei.

This work was supported by a National Science Foundation CAREER Award (MCB 0644262) and a National Institutes of Health (NIH) National Institute of General Medical Sciences grant (R01 GM084288) to R.L.G.; a NIH National Centers for Biomedical Computing grant (U54CA121852) to C.H.W.; and a Rubicon fellowship (680-50-1016) from the Netherlands Organization for Scientific Research (NWO) to J.W.M.

REFERENCES

1. Tinoco, Jr., I., and R. L. Gonzalez, Jr. 2011. Biological mechanisms, one molecule at a time. *Genes Dev.* 25:1205–1231.
2. Joo, C., H. Balci, …, T. Ha. 2008. Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.* 77:51–76.
3. Borgia, A., P. M. Williams, and J. Clarke. 2008. Single-molecule studies of protein folding. *Annu. Rev. Biochem.* 77:101–125.
4. Neuman, K. C., and A. Nagy. 2008. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods.* 5:491–505.
5. Cornish, P. V., and T. Ha. 2007. A survey of single-molecule techniques in chemical biology. *ACS Chem. Biol.* 2:53–61.
6. Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE.* 77:257–286.
7. Eddy, S. R. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* 6:361–365.
8. Bilmes, J. 1998. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of California Berkeley, ICSI-TR-97-021.
9. Chung, S. H., J. B. Moore, …, P. W. Gage. 1990. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 329:265–285.
10. Qin, F., A. Auerbach, and F. Sachs. 1997. Maximum likelihood estimation of aggregated Markov processes. *Proc. Biol. Sci.* 264:375–383.
11. Qin, F., A. Auerbach, and F. Sachs. 2000. A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* 79:1915–1927.
12. Smith, D. A., and R. M. Simmons. 2001. Models of motor-assisted transport of intracellular particles. *Biophys. J.* 80:45–68.
13. Kruithof, M., and J. van Noort. 2009. Hidden Markov analysis of nucleosome unwrapping under force. *Biophys. J.* 96:3708–3715.
14. McKinney, S. A., C. Joo, and T. Ha. 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* 91:1941–1951.
15. Bronson, J. E., J. Fei, …, C. H. Wiggins. 2009. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* 97:3196–3205.
16. Bronson, J. E., J. M. Hofman, …, C. H. Wiggins. 2010. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics.* 11 (*Suppl 8*):S2.
17. Liu, Y., J. Park, …, T. Ha. 2010. A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B.* 114:5386–5403.
18. Greenfeld, M., D. S. Pavlichin, …, D. Herschlag. 2012. Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data. *PLoS ONE.* 7:e30024.
19. Okamoto, K., and Y. Sako. 2012. Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.* 103:1315–1324.
20. Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* 39:1–38.
21. Baum, L., T. Petrie, …, N. Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41:164–171.
22. Jordan, M., Z. Ghahramani, and T. Jaakkola. 1999. An introduction to variational methods for graphical models. *Mach. Learn.* 233:183–233.
23. Berger, J. 1982. Bayesian robustness and the Stein effect. *J. Am. Stat. Assoc.* 77:358–368.
24. Kass, R., and D. Steffey. 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.* 84:717–726.
25. van de Meent, J.-W., J. E. Bronson, …, C. H. Wiggins. 2013. Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *Proc. Int. Conf. Mach. Learn.* 28:361–369.
26. Blanco, M., and N. G. Walter. 2010. Analysis of complex single-molecule FRET time trajectories. *Methods Enzymol.* 472:153–178.
27. Bishop, C. M. 2006. Pattern recognition and machine learning. Springer, New York.
28. Reference deleted in proof.
29. Elvekrog, M. M., and R. L. Gonzalez, Jr. 2013. Conformational selection of translation initiation factor 3 signals proper substrate selection. *Nat. Struct. Mol. Biol.* 20:628–633.
30. Laursen, B. S., H. P. Sørensen, …, H. U. Sperling-Petersen. 2005. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* 69:101–123.
31. Fei, J., J. E. Bronson, …, R. L. Gonzalez, Jr. 2009. Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *Proc. Natl. Acad. Sci. USA.* 106:15702–15707.
32. Blanchard, S. C., R. L. Gonzalez, …, J. D. Puglisi. 2004. tRNA selection and kinetic proofreading in translation. *Nat. Struct. Mol. Biol.* 11:1008–1014.
33. Johnson, S., J.-W. van de Meent, C. H. Wiggins, R. Phillips, and M. Lindén. Multiple Lac-mediated loops revealed by Bayesian statistics and tethered particle motion. Preprint, submitted February 4, 2014. arXiv. http://arxiv.org/abs/1402.0894v1.

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

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_{FRET} observation mean for state k in time series n

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

E_{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)

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

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

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

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

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

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 ψ_n. Calculation of ψ_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 ψ_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 η represents the remapped parameters θ, and $\{\nu_0, \chi_0, \phi_0\}$ represent the remapped hyperparameters ψ_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 ν can be interpreted as scale factor that tracks the number of previously observed samples. The hyperparameter vector χ in turn takes the role of the aggregate sufficient statistics $\mathcal{T}$ associated with these samples.

In any pair of conjugate distributions η, χ and $\mathcal{T}(x)$ must have the same dimensionality. This means that if η 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 ν_0 and ϕ_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 ϕ_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 ν observations and the statistics for the precision are equivalent to $\nu + \phi$ observations.

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

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

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

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

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

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

$$\mathcal{T}(x) = \{x^2, x\} \tag{28}$$

$$A(\eta) = -\frac{1}{2}\left[\ln\left(-\eta_1 / \pi\right) + \eta_2^2 / \left(2\eta_1\right)\right] \tag{29}$$

$$B(\eta, \phi) = -\frac{1}{2}(\phi + 1)\ln\left(-\eta_1 / \pi\right) \tag{30}$$

$$\begin{aligned} A(\nu, \chi, \phi) = -\frac{1}{2}\Big[& \ln(\nu) + (\nu + \phi - 2)\ln(2\pi) \\ & + (\nu + \phi)\ln\left[\tfrac{1}{2}(\chi_1 - \chi_2^2/\nu)\right] - 2\ln\Gamma\left[\tfrac{1}{2}(\nu + \phi)\right]\Big] \\ = -\frac{1}{2}\ln(\beta) - (a-1)\ln(2\pi) - a\ln(b) + \ln\Gamma(a) & \end{aligned} \tag{31}$$

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

$$p(z \mid \pi) = \mathrm{Cat}(z \mid \pi) = \prod_k \pi_k^{z_k} \tag{33}$$

$$p(\pi \mid \rho) = \mathrm{Dir}(\pi \mid \rho) = \frac{\Gamma(\sum_k \rho_k)}{\prod_k \Gamma(\rho_k)} \prod_k \pi_k^{\rho_k - 1} \tag{34}$$

$$\eta = \{\ln \pi\} \tag{35}$$

$$\chi = \{\rho\} \tag{36}$$

$$\mathcal{T}(z) = \{z\} \tag{37}$$

$$A(\eta) = \eta \tag{38}$$

$$B(\eta) = -\eta \tag{39}$$

$$A(\chi) = \log\Gamma(\textstyle\sum_k \chi_k) - \sum_k \log\chi_k \tag{40}$$

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_{KL} term is ≥ 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 λ_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 θ. 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:

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

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 η^* 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 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:

$$\Delta^{*2} = (1/\beta_k) + a_k(x - m_k)^2/b_k \tag{60}$$
$$\ln \lambda^* = \Psi(a_k) - \ln b_k \tag{61}$$
$$\ln A^*_{k,l} = \Psi(\alpha_{k,l}) - \Psi\left(\textstyle\sum_l \alpha_{k,l}\right) \tag{62}$$
$$\ln \pi^*_k = \Psi(\rho_k) - \Psi\left(\textstyle\sum_k \rho_k\right) \tag{63}$$

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 λ_θ is a Lagrange multiplier. We now take the functional derivative instead of a gradient, but the steps are essentially the same.

$$0 = \frac{\delta}{\delta q(\theta)}\left[L[q(z), q(\theta)] + \lambda_\theta\left(1 - \int d\theta'\, q(\theta')\right)\right] \tag{64}$$
$$= \left[\sum_z q(z)\left(\ln p(x, z, \theta \mid \psi_0) - \ln q(z) - \ln q(\theta) - 1\right)\right] - \lambda_\theta \tag{65}$$

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 θ 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:

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

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

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:

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

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 ψ

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 ψ

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

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

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

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

Finally, the updates for α_0 and ρ_0 can be obtained by substitution of the terms in equation (74) into equations (77) and (78):

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

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

We now proceed to derive how γ and ξ 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 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:

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

$$\beta_{T,k} = 1 \tag{103}$$

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

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

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

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

$$E_{q(\eta)}[\eta] = \left\{-\frac{a}{2b}, \frac{am}{b}\right\} \tag{124}$$

The KL divergences for A and π 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 ψ_0. The quantity optimized is the summed lower bound log evidence over the ensemble of time series:

$$\ln p(x \mid \psi_0) \geq \sum_n L_n \tag{127}$$

$$= \sum_n E_{q(z_n)q(\theta_n)}\left[\ln\left(\frac{p(x_n, z_n, \theta_n \mid \psi_0)}{q(z_n)q(\theta_n)}\right)\right] \tag{128}$$

$$= \sum_n \ln p(x_n \mid \psi_0) - D_{\mathrm{KL}}\big[q(z_n)q(\theta_n) \,||\, p(z_n, \theta_n \mid x_n, \psi_0)\big] \tag{129}$$

$$= \sum_n E_{q(z_n)q(\theta_n)}\big[\ln p(x_n \mid z_n, \theta_n)\big] - D_{\mathrm{KL}}\big[q(z_n)q(\theta_n) \,||\, p(z_n, \theta_n \mid \psi_0)\big] \tag{130}$$

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 ψ_0

$$0 = \nabla_{\psi_0} \sum_n L_n \tag{131}$$

$$= \nabla_{\psi_0} \sum_n \int d\theta_n \, q(\theta_n \mid w_n) \ln p(\theta_n \mid \psi_0) \tag{132}$$

$$= \sum_n \int d\theta_n \, q(\theta_n \mid w_n) \nabla_{\psi_0} \ln p(\theta_n \mid \psi_0) \tag{133}$$

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 π can be computed separatedly.

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 $q(\theta_n \mid, \psi_n)$ 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 Ψ

$$\Psi\Big[\textstyle\sum_m \alpha_{0,km}\Big] - \Psi\Big[\alpha_{0,kl}\Big]$$

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

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 η 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

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

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

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

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

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

In the M-step we hold $q^{(i+1)}(y_n)$ fixed and maximize L relative to ψ_{0m} and ζ. 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 κ define a differential equation for the evolution of the probability $\gamma_k(t)$ 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 Δt 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 κ_{kl} and Δt through the relationship

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

While any given κ value uniquely determines A, the equation $\kappa = \ln[A]/\Delta t$ does not necessarily have a unique solution. However in the limit of small Δt 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 τ_k we define

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

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

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

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

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 γ_{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

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

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 2×2 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.