

# vbdiff: variational Bayes analysis of simple Brownian mixtures

Alec Heckert

February 2024

`vbdiff` is a simple routine for inference on multi-state Brownian mixtures. It relies on a variational Bayesian framework to estimate a distribution over three factors: (1) the diffusion coefficients for each state, (2) the mixing coefficients, and (3) the origin state for each individual trajectory. Unlike more sophisticated frameworks like `vbSPT`, it does not consider state transitions. As a consequence, it is much faster than these more general models.

Here, we derive `vbdiff`, starting from basic one-state Brownian motion. Many of these results are standard but scattered over different texts; the goal of this document is provide a unified, accessible introduction to the variational Brownian mixture model. We also give the evidence lower bound (ELBO) for `vbdiff` in closed form. This can be used, for example, to choose the appropriate number of states when modeling a Brownian mixture [1].

## Contents

<b>1</b>	<b>Brownian motion</b>	<b>2</b>
<b>2</b>	<b>Bayesian inference on single-state Brownian motion</b>	<b>3</b>
<b>3</b>	<b>Brownian mixtures</b>	<b>4</b>
<b>4</b>	<b>Approximative inference on Brownian mixtures</b>	<b>6</b>
4.1	Choice of hyperparameters . . . . .	7
<b>5</b>	<b>Evidence lower bound (ELBO)</b>	<b>8</b>
<b>A</b>	<b>Appendix A: Derivation of approximate posterior for Brownian mixtures</b>	<b>9</b>

# 1 Brownian motion

A continuous-time stochastic process  $B_t \in \mathbb{R}$  ( $t \in \mathbb{R}$ ) is Brownian if

$$B_t - B_s \sim \mathcal{N}(0, 2D|t - s|) \quad \text{for any } t, s \quad (1)$$

$$\text{Cov}(B_{t_1} - B_{t_0}, B_{t_3} - B_{t_2}) = 0 \quad \text{if } t_0 \leq t_1 \leq t_2 \leq t_3 \quad (2)$$

where  $\mathcal{N}(0, 2D|t - s|)$  is a normal distribution with mean 0 and variance  $2D|t - s|$ .

Suppose we measure a Brownian trajectory in a single particle tracking (SPT) experiment. We sample the particle's position at  $m+1$  timepoints  $0, \Delta t, 2\Delta t, \dots, m\Delta t$  where  $\Delta t > 0$  is the frame interval (we'll neglect gaps in the trajectory for now). Let  $\Delta B_i = B_{i\Delta t} - B_{(i-1)\Delta t}$  be the  $i^{\text{th}}$  increment ("jump"). Then the joint distribution over these jumps is

$$\begin{aligned} p_{\Delta B_1, \dots, \Delta B_m}(\Delta b_1, \dots, \Delta b_m) &= \prod_{i=1}^m \frac{\exp(-\Delta b_i^2 / 4D\Delta t)}{\sqrt{4\pi D\Delta t}} \\ &= \frac{1}{(4\pi D\Delta t)^{m/2}} \exp\left(-\frac{1}{4D\Delta t} [\Delta b_1^2 + \dots + \Delta b_m^2]\right) \end{aligned}$$

This distribution does not depend on the individual jumps, only on the sum of their squares. It is straightforward to show that the sum of squares  $X = \Delta B_1^2 + \dots + \Delta B_m^2$  has a gamma distribution. First, the cdf for  $\Delta B_i^2$  is

$$\begin{aligned} F_{\Delta B^2}(x) &:= \Pr(\Delta B^2 \leq x) \\ &= \Pr(\Delta B \leq \sqrt{x}) - \Pr(\Delta B < -\sqrt{x}) \\ &= F_{\Delta B}(\sqrt{x}) - F_{\Delta B}(-\sqrt{x}) \end{aligned}$$

Differentiating both sides, we obtain the pdf

$$\begin{aligned} f_{\Delta B^2}(x) &= \frac{\partial F_{\Delta B^2}(x)}{\partial x} \\ &= \frac{1}{2\sqrt{x}} (f_{\Delta B}(\sqrt{x}) + f_{\Delta B}(-\sqrt{x})) \end{aligned}$$

Substituting the normal distribution from equation 1 as  $f_{\Delta B}$ , we have

$$f_{\Delta B^2}(x) = \frac{e^{-x/4D\Delta t}}{\sqrt{4\pi D\Delta t x}}$$

which we recognize as a gamma distribution with parameters  $\alpha = 1/2$  and  $\beta = 1/4D\Delta t$ :

$$\Delta B^2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{4D\Delta t}\right)$$

Now, the sum of  $m$  independent and identically distributed gamma random variables with the distribution  $\text{Gamma}(\alpha, \beta)$  has the distribution  $\text{Gamma}(m\alpha, \beta)$ . From this, we have

$$X := \Delta B_1^2 + \dots + \Delta B_m^2 \sim \text{Gamma}\left(\frac{m}{2}, \frac{1}{4D\Delta t}\right) \quad (3)$$

which corresponds to the pdf

$$f_X(x) = \frac{x^{\frac{m}{2}-1} e^{-x/4D\Delta t}}{\Gamma\left(\frac{m}{2}\right) (4D\Delta t)^{m/2}}, \quad x \geq 0 \quad (4)$$

Now, suppose that a particle's motion in  $d$  spatial dimensions can be modeled as independent Brownian motions in each dimension. Due to the Pythagorean theorem, the sum of its squared displacement is now  $X = X_{\text{dimension 1}} + X_{\text{dimension 2}} + \dots$ , where the terms on the right-hand side are the sums of its squared displacements along each spatial dimension. Insofar as these sums are iid, we have

$$X \sim \text{Gamma}\left(\frac{md}{2}, \frac{1}{4D\Delta t}\right) \quad (5)$$

Frequently, we will find it convenient to parametrize the distribution by the *scale* parameter  $\phi = 4D\Delta t$  instead of  $D$ , in which case the pdf becomes

$$f_X(x|\phi) = \frac{x^{\frac{md}{2}-1} e^{-x/\phi}}{\Gamma\left(\frac{md}{2}\right) \phi^{\frac{md}{2}}}, \quad x \geq 0 \quad (6)$$

Finally, at the time of writing we are only interested in 2D tracking ( $d = 2$ ), and so the likelihood simplifies to

$$\begin{aligned} X &\sim \text{Gamma}(m, \phi) \\ \Rightarrow f_X(x|\phi) &= \frac{x^{m-1} e^{-x/\phi}}{\Gamma(m) \phi^m}, \quad x \geq 0 \end{aligned} \quad (7)$$

where as before  $\phi = 4D\Delta t$ . For the sake of readability we use equation 7 throughout this text, with the understanding that generalizing to  $d \neq 2$  requires substituting  $m$  with  $md/2$ .

## 2 Bayesian inference on single-state Brownian motion

As a building block to mixture models, we first consider a simple Bayesian model for one-state Brownian motion. Suppose we have a 2D Brownian trajectory with  $m$  jumps and sum of squared displacements  $x$ . Our goal is to infer a distribution over the scale parameter (diffusion coefficient)  $\phi$  given the observed  $x$ . Bayes' theorem prescribes

$$f_{\phi|X}(\phi|x) \propto f_X(x|\phi) f_{\phi}(\phi)$$

where  $f_X(x|\phi)$  is given by equation 7. For the prior  $f_\phi(\phi)$ , we assume an inverse gamma distribution:

$$\begin{aligned}\phi &\sim \text{InvGamma}(\alpha_0, \beta_0) \\ \Rightarrow f_\phi(\phi) &= \frac{\beta_0^{\alpha_0} e^{-\beta_0/\phi}}{\Gamma(\alpha_0) \phi^{\alpha_0+1}}, \quad \phi \geq 0\end{aligned}$$

Substituting into Bayes' theorem and retaining only factors that depend directly on  $\phi$ ,

$$f_{\phi|X}(\phi|x) \propto \frac{e^{-(\beta_0+x)/\phi}}{\phi^{\alpha_0+m+1}}$$

After renormalization, we obtain

$$f_{\phi|X}(\phi|x) = \frac{(\beta_0 + x)^{\alpha_0+m-1} e^{-(\beta_0+x)/\phi}}{\Gamma(\alpha_0 + m) \phi^{\alpha_0+m+1}}$$

which we recognize as another inverse gamma distribution

$$\phi \mid X \sim \text{InvGamma}(\alpha_0 + m, \beta_0 + x)$$

So - due to the conjugacy of the Brownian likelihood ?? and the inverse gamma prior - inference boils down to simply adding shape and scale parameters. We can obtain a point estimate of the diffusion coefficient by taking the mean of the posterior distribution:

$$\mathbb{E}[D|x] = \frac{1}{4\Delta t} \mathbb{E}[\phi|x] = \frac{\beta_0 + x}{4\Delta t (\alpha_0 + m - 1)}$$

Notice that the  $-1$  in the denominator is analogous to the Bessel correction for the sample variance, as the diffusion coefficient is essentially a variance.

### 3 Brownian mixtures

Most samples of interest in SPT contain a mixture of particles in different states characterized by distinct diffusion coefficients. Suppose we have a dataset with  $N$  trajectories, each of which originates from one of  $K$  distinct states characterized by scale parameters  $\phi_1, \dots, \phi_K$ . There are multiple unknowns here: we don't know the scale parameters for each state, the relative abundances ("occupations") of each state, nor which state each trajectory comes from. To represent this uncertainty, we construct the following mixture model:

$$\boldsymbol{\tau} \sim \text{Dirichlet}(\alpha_0, \dots, \alpha_0), \quad \boldsymbol{\tau} \in C^{K-1} \quad (8)$$

$$\phi_j \sim \text{InvGamma}(\alpha_0, \beta_0), \quad j \in \{1, 2, \dots, K\} \quad (9)$$

$$\mathbf{Z}_i \mid \boldsymbol{\tau} \sim \boldsymbol{\tau}, \quad i \in \{1, 2, \dots, N\} \quad (10)$$

$$X_i \mid Z_{ij} = 1, \phi_j \sim \text{Gamma}(m_i, \phi_j), \quad i \in \{1, 2, \dots, N\} \quad (11)$$

Breaking this down:

- $\boldsymbol{\tau}$  is an unknown distribution over  $K$  states (in other words, it belongs to the simplex  $C^{K-1}$ ).
- Each state  $j$  is characterized by a scale parameter  $\phi_j$ .
- The priors over each  $\phi_j$  are independent and given by  $\text{InvGamma}(\alpha_0, \beta_0)$ .
- To each observation  $i$ , we associate a one-hot vector  $\mathbf{Z}_i \in \{0, 1\}^K$  so that  $Z_{ik} = 1$  if observation  $i$  originates from state  $j$  and  $Z_{ik} = 0$  otherwise. Given knowledge of  $\boldsymbol{\tau}$ , the event  $Z_{ij} = 1$  has probability  $\tau_j$ .
- $m_i$  is the number of jumps in trajectory  $i$ .
- $X_i$  is the sum of squared displacements for trajectory  $i$ .

We will use  $\mathbf{X} = (X_1, \dots, X_N)$  and  $\mathbf{Z} = (Z_1, \dots, Z_N)$  to represent the sequence of all  $X_i$  and  $\mathbf{Z}_i$ . Likewise, we'll use  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)$  to represent the sequence of scale parameters for all  $K$  states. Then the total probability function for this model factors as

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\phi}) p(\mathbf{Z}|\boldsymbol{\tau}) p(\boldsymbol{\tau}) p(\boldsymbol{\phi})$$

For later parts, it's useful to write out each of these terms explicitly:

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\phi}) = \prod_{i=1}^N \prod_{j=1}^K \left[ \frac{x_i^{m_i-1} e^{-x_i/\phi_j}}{\Gamma(m_i) \phi_j^{m_i}} \right]^{Z_{ij}} \quad (12)$$

$$p(\mathbf{Z}|\boldsymbol{\tau}) = \prod_{i=1}^N \prod_{j=1}^K \tau_j^{Z_{ij}} \quad (13)$$

$$p(\boldsymbol{\tau}) = \frac{1}{B(\alpha_0, \dots, \alpha_0)} \prod_{j=1}^K \tau_j^{\alpha_0-1} \quad (14)$$

$$p(\boldsymbol{\phi}) = \prod_{j=1}^K \frac{\beta_0^{\alpha_0} e^{-\beta_0/\phi_j}}{\Gamma(\alpha_0) \phi_j^{\alpha_0+1}} \quad (15)$$

(Here,  $B(\alpha_0, \dots, \alpha_0)$  is the multivariate beta function.)

Of all of these parameters, the only one we actually observe is  $\mathbf{X}$ . Our goal is to infer a posterior distribution over the remaining variables given an observed set of trajectories.

## 4 Approximative inference on Brownian mixtures

Given an observed set of trajectories  $\mathbf{X} = (X_1, \dots, X_N)$ , we want to infer the probability of the unobserved (latent) variables  $\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}$  in the mixture 8. In other words, we want to infer the posterior distribution  $p(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} | \mathbf{X})$ . Solving this problem exactly is intractable. There are two fallbacks: (a) approximate the distribution with a finite sequence of samples using MCMC methods or (b) construct a tractable approximation to the posterior. In `vbdiff`, we take the latter approach. We follow a classic variational Bayesian approach and paraphrase certain steps; for more detail see Bishop's book [2].

We will approximate the posterior with a tractable distribution  $q$ :

$$q(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}) \approx p(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} | \mathbf{X})$$

We will use the evidence lower bound (ELBO) as the objective function for  $q$ :

$$\mathcal{L}[q] := \mathbb{E}_{\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} \sim q} \left[ \log \left( \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})}{q(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})} \right) \right] \quad (16)$$

By factoring  $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}) = p(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} | \mathbf{X}) p(\mathbf{X})$  in equation 16, it straightforward to show that maximizing  $\mathcal{L}[q]$  corresponds to minimizing the Kullback-Leibler divergence between the true and approximative posteriors, and that  $\mathcal{L}[q]$  forms a lower bound on the marginal log likelihood  $\log p(\mathbf{X})$ :

$$\begin{aligned} \operatorname{argmax}_q \mathcal{L}[q] &= \operatorname{argmin}_q \text{KL}(q(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}) || p(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} | \mathbf{X})) \\ \mathcal{L}[q] &\leq \log p(\mathbf{X}) \end{aligned}$$

This last point is crucial, as it will form the basis for model selection considered later on. For now, notice that  $p(\mathbf{X})$  can be seen as the average probability of the observed data over all possible model parameters  $\mathbf{X}, \boldsymbol{\tau}, \boldsymbol{\phi}$  with some  $K$ .

What form will we choose for  $q$ ? As it turns out, we only need to make a single constraint to find a tractable  $q$ : assume that  $q$  factors into separate distributions over  $\mathbf{Z}$  and  $\boldsymbol{\tau}, \boldsymbol{\phi}$ :

$$q(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}) = q(\mathbf{Z}) q(\boldsymbol{\tau}, \boldsymbol{\phi})$$

This factorability criterion - a kind of mean field approximation - is sufficient to completely specify the form of the optimal  $q$ . It can be shown [2] that  $\operatorname{argmax}_q \mathcal{L}[q]$  satisfies

$$\begin{aligned} \log q(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\tau}, \boldsymbol{\phi} \sim q(\boldsymbol{\tau}, \boldsymbol{\phi})} [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})] \\ \log q(\boldsymbol{\tau}, \boldsymbol{\phi}) &= \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})] \end{aligned} \quad (17)$$

We solve these equations in Appendix A. Skipping to the punchline, the optimal approximation  $q$  is given by

$$\begin{aligned}
q(\mathbf{Z}, \boldsymbol{\tau}, \phi) &= q(\mathbf{Z}) q(\boldsymbol{\tau}) q(\phi) \\
q(\mathbf{Z}) &= \prod_{i=1}^N \prod_{j=1}^K r_{ij}^{Z_{ij}} \\
q(\boldsymbol{\tau}) &= \text{Dirichlet}(\boldsymbol{\tau} | \alpha_0 + \alpha_1, \dots, \alpha_0 + \alpha_K) \\
q(\phi) &= \prod_{j=1}^K \text{InvGamma}(\phi_j | \alpha_0 + \alpha_j, \beta_0 + \beta_j)
\end{aligned} \tag{18}$$

where

$$\begin{aligned}
r_{ij} &= \frac{\exp(\mathbb{E}_q[\log \tau_j] - x_i \mathbb{E}_q[\phi_j^{-1}] - m_i \mathbb{E}_q[\log \phi_j])}{\sum_{k=1}^K \exp(\mathbb{E}_q[\log \tau_k] - x_i \mathbb{E}_q[\phi_k^{-1}] - m_i \mathbb{E}_q[\log \phi_k])} \\
\alpha_j &= \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] m_i \\
\beta_j &= \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] x_i
\end{aligned} \tag{19}$$

Inference consists of cyclically updating  $r_{ij}$ ,  $\alpha_j$ , and  $\beta_j$  until convergence. This can be done by substituting the following expectations at each update:

$$\begin{aligned}
\mathbb{E}_q[Z_{ij}] &= r_{ij} \\
\mathbb{E}_q[\log \tau_j] &= \psi(\alpha_0 + \alpha_j) - \psi(K\alpha_0 + \alpha_1 + \dots + \alpha_K) \\
\mathbb{E}_q[\phi_j^{-1}] &= \frac{\alpha_0 + \alpha_j}{\beta_0 + \beta_j} \\
\mathbb{E}_q[\log \phi_j] &= \log(\beta_0 + \beta_j) - \psi(\alpha_0 + \alpha_j)
\end{aligned}$$

#### 4.1 Choice of hyperparameters

There are three hyperparameters in the mixture model 8:

- $\alpha_0$ , the pseudocounts accorded to each state in the prior;
- $\beta_0$ , the prior scale parameter for each state;
- $K$ , the number of states.

The last ( $K$ ) is treated in the next section 5. The number of pseudocounts is often set to  $\alpha_0 = 2$  in `vbdiff`, to avoid division by zero when evaluating prior means. For  $\beta_0$ , suppose we have some prior guess for the diffusion coefficient  $D_0$ . Since  $\phi = 4D\Delta t$  and the prior mean value of  $\phi$  is  $\beta_0/(\alpha_0 - 1)$ , a sensible choice is to set  $\beta_0 = 4\Delta t(\alpha_0 - 1)D_0$ . This corresponds to the average sum of squared displacements for a trajectory with  $\alpha_0$  jumps and diffusion coefficient  $D_0$ , sampled with frame interval  $\Delta t$ .

## 5 Evidence lower bound (ELBO)

A central hyperparameter in the mixture model 8 is the number of states  $K$ . How do we choose  $K$  in a principled way?

Besides providing an objective function for variational Bayesian inference, the evidence lower bound 16 can also be used for *model selection*. Because  $\mathcal{L}[q] \leq \log p(\mathbf{X})$ , the ELBO provides an approximation to the probability of the observed data  $\mathbf{X}$  averaged over all model parameters  $\mathbf{X}, \boldsymbol{\tau}, \boldsymbol{\phi}$  corresponding to a particular  $K$ . If  $K$  is too small, then there may not exist model parameters that adequately describe the data. If  $K$  is too high, then even if model parameters exist that adequately describe the data, most do not. In short, we should take the  $K$  with the highest  $\mathcal{L}[q]$ .

Expanding the ELBO equation 16, we have

$$\mathcal{L}[q] = \mathbb{E}_q [\log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})] \quad (\text{term A}) \quad (20)$$

$$+ \mathbb{E}_q [\log p(\mathbf{Z}|\boldsymbol{\tau})] \quad (\text{term B}) \quad (21)$$

$$+ \mathbb{E}_q [\log p(\boldsymbol{\tau})] \quad (\text{term C}) \quad (22)$$

$$+ \mathbb{E}_q [\log p(\boldsymbol{\phi})] \quad (\text{term D}) \quad (23)$$

$$- \mathbb{E}_q [\log q(\mathbf{Z})] \quad (\text{term E}) \quad (24)$$

$$- \mathbb{E}_q [\log q(\boldsymbol{\tau})] \quad (\text{term F}) \quad (25)$$

$$- \mathbb{E}_q [\log q(\boldsymbol{\phi})] \quad (\text{term G}) \quad (26)$$



Separately, these terms are:

$$\begin{aligned}
\mathbb{E}_q [\log p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})] &= \sum_{i=1}^N \sum_{j=1}^K \mathbb{E}_q [Z_{ij}] \left( (m_i - 1) \log x_i - \log \Gamma(m_i) \right. \\
&\quad \left. - m_i \mathbb{E}_q [\log \phi_j] - x_i \mathbb{E}_q [\phi_j^{-1}] \right) \\
\mathbb{E}_q [\log p(\mathbf{Z}|\boldsymbol{\tau})] &= \sum_{i=1}^N \sum_{j=1}^K \mathbb{E}_q [Z_{ij}] \mathbb{E}_z [\log \tau_j] \\
\mathbb{E}_q [\log p(\boldsymbol{\tau})] &= -\log B(\alpha_0, \dots, \alpha_0) + \sum_{j=1}^K (\alpha_0 - 1) \mathbb{E}_q [\log \tau_j] \\
\mathbb{E}_q [\log p(\boldsymbol{\phi})] &= \sum_{j=1}^K (\alpha_0 \log \beta_0 - \log \Gamma(\alpha_0) - \beta_0 \phi_j^{-1} - (\alpha_0 + 1) \log \phi_j) \\
\mathbb{E}_q [\log q(\mathbf{Z})] &= \sum_{i=1}^N \sum_{j=1}^K \mathbb{E}_q [Z_{ij}] \log r_{ij} \\
\mathbb{E}_q [\log q(\boldsymbol{\tau})] &= -\log B(\alpha_0 + \alpha_1, \dots, \alpha_0 + \alpha_K) + \sum_{j=1}^K (\alpha_0 + \alpha_j - 1) \mathbb{E}_q [\log \tau_j] \\
\mathbb{E}_q [\log q(\boldsymbol{\phi})] &= \sum_{j=1}^K \left( (\alpha_0 + \alpha_j) \log(\beta_0 + \beta_j) - \log \Gamma(\alpha_0 + \alpha_j) \right. \\
&\quad \left. - (\beta_0 + \beta_j) \mathbb{E}_q [\phi_j^{-1}] - (\alpha_0 + \alpha_j + 1) \mathbb{E}_q [\log \phi_j] \right)
\end{aligned}$$

where, as before,  $B(x_1, \dots, x_K)$  is the multivariate beta function and the inner expectations are given by

$$\begin{aligned}
\mathbb{E}_q [Z_{ij}] &= r_{ij} \\
\mathbb{E}_q [\phi_j^{-1}] &= \frac{\alpha_0 + \alpha_j}{\beta_0 + \beta_j} \\
\mathbb{E}_q [\log \phi_j] &= \log(\beta_0 + \beta_j) - \psi(\alpha_0 + \alpha_j) \\
\mathbb{E}_q [\log \tau_j] &= \psi(\alpha_0 + \alpha_j) - \psi(K\alpha_0 + \alpha_1 + \dots + \alpha_K)
\end{aligned}$$

where  $\psi$  is the digamma function.

## A Appendix A: Derivation of approximate posterior for Brownian mixtures

To identify the exact form of the approximative posterior  $q(\mathbf{X}, \boldsymbol{\tau}, \boldsymbol{\phi})$  for Brownian mixtures, we will solve the system of equations 17 given the model specified by 12, 13, 14, and 15.

First we consider the second equation in 17. Writing this out,

$$\begin{aligned}
\log q(\boldsymbol{\tau}, \boldsymbol{\phi}) &= \mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})] \\
&= \sum_{j=1}^K \left( \alpha_0 - 1 + \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] m_i \right) \log \tau_j \\
&\quad - \sum_{j=1}^K \left( \beta_0 + \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] x_i \right) \phi_j^{-1} \\
&\quad - \sum_{j=1}^K \left( \alpha_0 + 1 + \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] m_i \right) \log \phi_j \\
&\quad + \text{constant}
\end{aligned} \tag{27}$$

where the constant collects all terms that do not directly depend on  $\boldsymbol{\tau}$  or  $\boldsymbol{\phi}$ . We have made one arbitrary modification in equation 27, weighting the statistical weight of trajectory  $i$  toward state  $j$  by its number of jumps  $m_i$  rather than the same for all tracks. This accounts for a bias unmodeled by the naive mixture 8 - fast states tend to produce many short trajectories while slow states tend to produce a few long trajectories. Weighting statistical weight by the number of jumps in each trajectory removes some of this bias; the remaining bias can be addressed via defocalization corrections.

Examining the form of equation 27, we see that it is a product of a Dirichlet distribution over  $\boldsymbol{\tau}$  and independent inverse gamma distributions over each scale parameter  $\phi_j$ :

$$\begin{aligned}
q(\boldsymbol{\tau}, \boldsymbol{\phi}) &= q(\boldsymbol{\tau}) \prod_{j=1}^K q(\phi_j) \\
q(\boldsymbol{\tau}) &= \text{Dirichlet}(\boldsymbol{\tau} \mid \alpha_0 + \alpha_1, \dots, \alpha_0 + \alpha_K) \\
q(\phi_j) &= \text{InvGamma}(\alpha_0 + \alpha_j, \beta_0 + \beta_j) \\
\alpha_j &= \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] m_i \\
\beta_j &= \sum_{i=1}^N \mathbb{E}_q[Z_{ij}] x_i
\end{aligned} \tag{28}$$

We still need to solve  $q(\mathbf{Z})$ . Expanding the first equation in 17, we have

$$\begin{aligned}
\log q(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\tau}, \boldsymbol{\phi} \sim q(\boldsymbol{\tau}, \boldsymbol{\phi})} [\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\phi})] \\
&= \sum_{i=1}^N \sum_{j=1}^K Z_{ij} \left( \mathbb{E}_q[\log \tau_j] + (m_i - 1) \log x_i - x_i \mathbb{E}_q[\phi_j^{-1}] \right. \\
&\quad \left. - \log \Gamma(m_i) - m_i \mathbb{E}_q[\log \phi_j] \right) + \text{constant}
\end{aligned}$$

where, as before, we've aggregated terms that do not depend on  $\mathbf{Z}$  into the constant. Normalizing over all states for each trajectory  $i$ , we have

$$q(\mathbf{Z}) = \prod_{i=1}^N \prod_{j=1}^K r_{ij}^{Z_{ij}} \quad (29)$$

$$r_{ij} = \frac{\exp(\mathbb{E}_q[\log \tau_j] - x_i \mathbb{E}_q[\phi_j^{-1}] - m_i \mathbb{E}_q[\log \phi_j])}{\sum_{k=1}^K \exp(\mathbb{E}_q[\log \tau_k] - x_i \mathbb{E}_q[\phi_k^{-1}] - m_i \mathbb{E}_q[\log \phi_k])}$$

Together, equations 28 and 29 specify the approximative posterior. The expectation-maximization routine cyclically updates the expectations in these equations, obtaining a progressively better estimate of the posterior. Convergence is guaranteed because the ELBO is convex with respect to each factor in  $q$  [3]. Since we know the forms of  $q(\mathbf{Z})$ ,  $q(\boldsymbol{\tau})$ , and  $q(\boldsymbol{\phi})$ , the updates for each expectation can be evaluated in closed form:

$$\begin{aligned} \mathbb{E}_q[Z_{ij}] &= r_{ij} \\ \mathbb{E}_q[\tau_j] &= \psi(\alpha_0 + \alpha_j) - \psi(K\alpha_0 + \alpha_1 + \dots + \alpha_K) \\ \mathbb{E}_q[\phi_j^{-1}] &= \frac{\alpha_0 + \alpha_j}{\beta_0 + \beta_j} \\ \mathbb{E}_q[\log \phi_j] &= \log(\beta_0 + \beta_j) - \psi(\alpha_0 + \alpha_j) \end{aligned}$$

where  $\alpha_j, \beta_j$  are defined as in equation 28 and  $\psi(\tau) = \frac{\partial}{\partial \tau} \log \Gamma(\tau)$  is the digamma function.

## References

- [1] Corduneanu A. & Bishop C. M. Variational Bayesian model selection for mixture distributions. *Artificial Intelligence and Statistics*, 27-34 (2001).
- [2] Bishop C. M. *Pattern Recognition and Machine Learning*. Springer (2006).
- [3] Boyd S. & Vandenberghe L. *Convex Optimization*. Cambridge University Press (2004).