vbdiff: variational Bayes analysis of 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, and was intended for inference on large-scale single particle tracking (SPT) datasets.

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].

This document was written rather quickly and paraphrases some sections of the author's thesis. If you find errors, I would be grateful (and will buy you a coffee) - let me know at alecheckert@gmail.com.

Contents

1 Brownian motion

This section reviews Brownian mixture models, starting from simple one-state Brownian motion. It defines several terms relevant for the next section on the `vbdiff` algorithm. The central result is the mixture model represented by relations 9 - 12.

1.1 Single state Brownian motion

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

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

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

where $\mathcal{N}(0, 2D|t-s|)$ is a normal distribution with mean 0 and variance $2D|t-s|$. $D \ge 0$ is a parameter called the *diffusion coefficient*, parametrizing the scale of the motion.

Suppose we measure a Brownian trajectory in a single particle tracking (SPT) experiment. We sample the particle's position at $m+1$ timepoints $0, \Delta t, 2\Delta t, ..., m\Delta t$ 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^{th} increment ("jump"). Then the joint distribution over these jumps is

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

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 $X = \Delta B_1^2 + ... + \Delta B_m^2$ has a gamma distribution. First, the cdf for ΔB_i^2 is

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

Differentiating both sides, we obtain the pdf

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

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

$$f_{\Delta B^2}(x) = \frac{e^{-x/4D\Delta t}}{\sqrt{4\pi D\Delta t x}}, \quad x \geq 0$$

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 Gamma (α, β) has the distribution Gamma $(m\alpha, \beta)$. From this, we have

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

which corresponds to the pdf

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

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 $X = X_{\text{dimension 1}} + X_{\text{dimension 2}} + ...$, 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

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

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

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

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

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

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 $md/2$.

1.2 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) ϕ 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:

$$\begin{aligned} \phi &\sim \text{InvGamma}\left(\alpha_0, \beta_0\right) \\ \Rightarrow f_\phi\left(\phi\right) &= \frac{\beta_0^{\alpha_0} e^{-\beta_0/\phi}}{\Gamma\left(\alpha_0\right) \phi^{\alpha_0+1}}, \quad \phi \geq 0 \end{aligned} \tag{8}$$

Substituting into Bayes' theorem and retaining only factors that depend directly on ϕ,

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

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

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

So - because the Brownian likelihood 7 is conjugate to the inverse gamma prior 8 - 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:

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

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 over position.

1.3 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 $\phi_1, ..., \phi_K$. 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:

$$\boldsymbol{\tau} \sim \text{Dirichlet}\left(\alpha_0, ..., \alpha_0\right), \qquad \boldsymbol{\tau} \in C^{K-1} \tag{9}$$

$$\phi_j \sim \text{InvGamma}\left(\alpha_0, \beta_0\right), \qquad j \in \{1, 2, ..., K\} \tag{10}$$

$$\boldsymbol{Z}_i \mid \boldsymbol{\tau} \sim \boldsymbol{\tau}, \qquad i \in \{1, 2, ..., N\} \tag{11}$$

$$X_i \mid Z_{ij} = 1, \phi_j \sim \text{Gamma}\left(m_i, \phi_j\right), \qquad i \in \{1, 2, ..., N\} \tag{12}$$

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 ϕ_j.
- The priors over each ϕ_j are independent and given by InvGamma(α_0, β_0).
- To each observation i, we associate a one-hot vector $\boldsymbol{Z}_i \in \{0, 1\}^K$ 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 τ_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 $\boldsymbol{X} = (X_1, ..., X_N)$ and $\boldsymbol{Z} = (Z_1, ..., Z_N)$ to represent the sequence of all X_i and $\boldsymbol{Z}_i$. Likewise, we'll use $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$ to represent the sequence of scale parameters for all K states. Then the total probability function for this model factors as

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

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

$$p\left(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\phi}\right) = \prod_{i=1}^{N} \prod_{j=1}^{K} \left[\frac{x_i^{m_i - 1} e^{-x_i / \phi_j}}{\Gamma(m_i) \phi_j^{m_i}} \right]^{Z_{ij}} \tag{13}$$

$$p\left(\boldsymbol{Z} | \boldsymbol{\tau}\right) = \prod_{i=1}^{N} \prod_{j=1}^{K} \tau_j^{Z_{ij}} \tag{14}$$

$$p\left(\boldsymbol{\tau}\right) = \frac{1}{B(\alpha_0, ..., \alpha_0)} \prod_{j=1}^{K} \tau_j^{\alpha_0 - 1} \tag{15}$$

$$p\left(\boldsymbol{\phi}\right) = \prod_{j=1}^{K} \frac{\beta_0^{\alpha_0} e^{-\beta_0 / \phi_j}}{\Gamma(\alpha_0) \phi_j^{\alpha_0 + 1}} \tag{16}$$

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

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

2 Variational Bayes inference on Brownian mixtures

2.1 vbdiff

Given an observed set of trajectories $\boldsymbol{X} = (X_1, ..., X_N)$, we want to infer the probability of the unobserved (latent) variables $\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}$ in the mixture 9 - 12. In other words, we want to infer the posterior distribution $p\left(\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} | \boldsymbol{X}\right)$.

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:

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

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

$$\mathcal{L}\left[q\right] := \mathbb{E}_{\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} \sim q}\left[\log\left(\frac{p\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}\right)}{q\left(\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}\right)}\right)\right] \tag{17}$$

By factoring $p\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}\right) = p\left(\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi} | \boldsymbol{X}\right) p\left(\boldsymbol{X}\right)$ in equation 17, 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 $\log p(\boldsymbol{X})$:

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

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

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

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 $\underset{q}{\operatorname{argmax}}\, \mathcal{L}[q]$ satisfies

$$\begin{aligned} \log q\left(\boldsymbol{Z}\right) &= \mathbb{E}_{\boldsymbol{\tau},\boldsymbol{\phi}\sim q(\boldsymbol{\tau},\boldsymbol{\phi})}\left[\log p\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\tau},\boldsymbol{\phi}\right)\right] \\ \log q\left(\boldsymbol{\tau},\boldsymbol{\phi}\right) &= \mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left[\log p\left(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\tau},\boldsymbol{\phi}\right)\right] \end{aligned} \tag{18}$$

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

$$\begin{aligned} q\left(\boldsymbol{Z},\boldsymbol{\tau},\boldsymbol{\phi}\right) &= q\left(\boldsymbol{Z}\right) q\left(\boldsymbol{\tau}\right) q\left(\boldsymbol{\phi}\right) \\ q\left(\boldsymbol{Z}\right) &= \prod_{i=1}^{N}\prod_{j=1}^{K} r_{ij}^{Z_{ij}} \\ q\left(\boldsymbol{\tau}\right) &= \text{Dirichlet}\left(\boldsymbol{\tau}|\alpha_0+\alpha_1,...,\alpha_0+\alpha_K\right) \\ q\left(\boldsymbol{\phi}\right) &= \prod_{j=1}^{K} \text{InvGamma}\left(\phi_j|\alpha_0+\alpha_j,\beta_0+\beta_j\right) \end{aligned} \tag{19}$$

where

$$\begin{aligned} r_{ij} &= \frac{\exp\left(\mathbb{E}_q\left[\log\tau_j\right] - x_i\mathbb{E}_q\left[\phi_j^{-1}\right] - m_i\mathbb{E}_q\left[\log\phi_j\right]\right)}{\sum_{k=1}^{K}\exp\left(\mathbb{E}_q\left[\log\tau_k\right] - x_i\mathbb{E}_q\left[\phi_k^{-1}\right] - m_i\mathbb{E}_q\left[\log\phi_k\right]\right)} \\ \alpha_j &= \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right]m_i \\ \beta_j &= \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right]x_i \end{aligned} \tag{20}$$

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

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

where $\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\log\Gamma(\alpha)$ is the digamma function.

2.2 Choice of hyperparameters

There are three hyperparameters in the mixture model 9:

- α_0, the pseudocounts accorded to each state in the prior;
- β_0, the prior scale parameter for each state;
- K, the number of states.

The last (K) is treated in the next section 2.3. The number of pseudocounts is often set to $\alpha_0 = 2$ in `vbdiff`, to avoid division by zero when evaluating prior means. For β_0, suppose we have some prior guess for the diffusion coefficient D_0. Since $\phi = 4D\Delta t$ and the prior mean value of ϕ 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 α_0 jumps and diffusion coefficient D_0, sampled with frame interval Δt.

2.3 Evidence lower bound (ELBO)

A central hyperparameter in the mixture model 9 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 17 can also be used for *model selection*. Because $\mathcal{L}[q] \leq \log p(\boldsymbol{X})$, the ELBO provides an approximation to the probability of the observed data $\boldsymbol{X}$ averaged over all model parameters $\boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}$ 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 $\mathcal{L}[q]$. This means we have both model parameters that adequately describe the data and do not have vast tracts of parameter space that do not describe the data.

Expanding the ELBO equation 17, we have

$$\begin{aligned}
\mathcal{L}[q] = \mathbb{E}_q\left[\log p\left(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\tau},\boldsymbol{\phi}\right)\right] &\quad \text{(term A)} &(21)\\
+ \mathbb{E}_q\left[\log p\left(\boldsymbol{Z}|\boldsymbol{\tau}\right)\right] &\quad \text{(term B)} &(22)\\
+ \mathbb{E}_q\left[\log p\left(\boldsymbol{\tau}\right)\right] &\quad \text{(term C)} &(23)\\
+ \mathbb{E}_q\left[\log p\left(\boldsymbol{\phi}\right)\right] &\quad \text{(term D)} &(24)\\
- \mathbb{E}_q\left[\log q\left(\boldsymbol{Z}\right)\right] &\quad \text{(term E)} &(25)\\
- \mathbb{E}_q\left[\log q\left(\boldsymbol{\tau}\right)\right] &\quad \text{(term F)} &(26)\\
- \mathbb{E}_q\left[\log q\left(\boldsymbol{\phi}\right)\right] &\quad \text{(term G)} &(27)
\end{aligned}$$

Separately, these terms are:

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

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

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

where ψ is the digamma function.

2.4 Experimental biases

`vbdiff` also incorporates two experimental biases common in SPT experiments: localization error and defocalization.

2.4.1 Localization error

In real SPT experiments, the position of a particle isn't known exactly, but is always associated with some measurement error (known as *localization error*). We

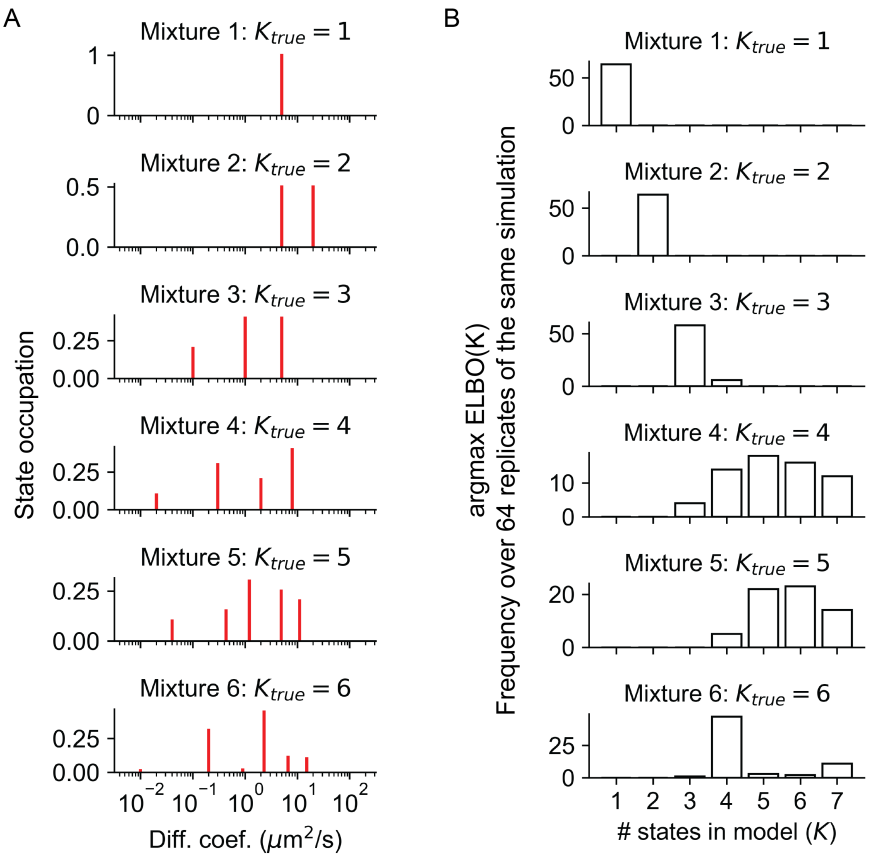

Figure 1: Using the ELBO to determine the number of states K in a Brownian mixture. We simulated Brownian mixtures with different numbers of states (K_{true}), then attempted to recover the correct K using vbdiff. To do this, we fit each simulated mixture to models with $K = 1$, 2, 3, 4, 5, 6, or 7 and took the K with the highest ELBO. Brownian mixture simulations were performed with the strobesim package, which simulates SPT-like conditions including 20 nm localization error, 700 nm depth of field, bleaching rate 10 Hz, frame rate 200 Hz, and approximately 7000 observed trajectories per simulation. Mean trajectory length is $\sim 3 - 4$ frames under these conditions (depending on the simulation). (A) Each of the six Brownian mixtures simulated. Red bars indicate ground truth states and the height of the bars indicates each state's *occupation* - the fraction of particles originating from that state. (B) Histogram of recovered K over 64 replicates of each simulated mixture. Notice that uncertainty in K rises as K_{true} rises.

model a Brownian motion with localization error as the process $\tilde{B}_t = B_t + \sigma_{\text{loc}} W_t$ where B_t is a Brownian motion, W_t is a Gaussian white noise process, and σ^2_{loc} is the localization error variance.

If the particle's position is measured at regular timepoints $0, \Delta t, 2\Delta t, ...$ so that the i^{th} increment is $\Delta\tilde{B}_i = \tilde{B}_{i\Delta t} - \tilde{B}_{(i-1)\Delta t}$, then the increments have the covariance

$$\text{Cov}\left(\Delta\tilde{B}_i, \Delta\tilde{B}_j\right) = \begin{cases} 2D\Delta t + 2\sigma^2_{\text{loc}} & \text{if } i = j \\ -\sigma^2_{\text{loc}} & \text{if } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

Models that incorporate the full covariance function, including the negative covariance between sequential jumps, include state arrays. State arrays also allow each trajectory to have a distinct localization error. In `vbdiff` we take a simpler approach by neglecting the off-diagonal covariances and assuming that localization error variance is a known constant, identical for all trajectories. Then the same arguments from Section 1 generate the likelihood function

$$X_i | D_j, \sigma^2_{\text{loc}} \sim \text{Gamma}\left(\frac{m_i d}{2}, \frac{1}{4(D_j\Delta t + \sigma^2_{\text{loc}})}\right) \tag{28}$$

where X is the sum of squared displacements in trajectory i, D_j is the diffusion coefficient for the j^{th} state, m_i is the number of trajectories in trajectory i, and d is the spatial dimension (usually 2). The core `vbdiff` algorithm remains exactly the same, except we define $\phi_j = 4(D\Delta t + \sigma^2_{\text{loc}})$ as the parameter for each state.

2.4.2 Defocalization

In real SPT experiments, particles are only observed when their positions coincide with the imaging system's depth of field. As a result, the jumps of fast particles are more likely to land outside the depth of field than the jumps of slow particles, creating a state bias toward slow particles.

In `vbdiff`, we address this bias in the following way. Suppose, as before, that the "observed" state occupations are represented by $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_K)$. Let $\eta_j(\phi_j)$ be the probability to observe a jump from state j given its mobility parameter ϕ_j and the imaging system's depth of field. Then the "true" state occupations are some vector $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_K)$ such that $\sum_{j=1}^K \mu_j = 1$ and

$$\tau_j = \frac{\eta_j(\phi_j)\mu_j}{\sum_{k=1}^K \eta_k(\phi_k)\mu_k} \tag{29}$$

(Note: We could also incorporate photobleaching into $\eta_j(\phi_j)$. But since we usually assume photobleaching affects all states equally, it factors out from the

numerator and denominator in equation 29 and we're left with only the defocalization part.)

We can include this in the stochastic model by modifying equation 14 so that it incorporates both the probability to sample a particle in state j (μ_j) as well as the probability to observe a jump from that particle ($\eta_j(\phi_j)$):

$$\log p\left(\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\phi}\right) \propto \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\left(\log \eta_j(\phi_j) + \log \mu_j\right)$$

To make things simpler, we will assume that η_j is constant at each iteration of `vbdiff`. For instance, we can evaluate η_j at the current posterior mean value of ϕ_j. Then

$$\log p\left(\boldsymbol{Z}|\boldsymbol{\mu}\right) \propto \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\left(\log \eta_j + \log \mu_j\right)$$

To understand how to incorporate this into the `vbdiff` algorithm, it is useful to examine the max likelihood solution to $\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}}{\operatorname{argmax}}\, p\left(\boldsymbol{Z}|\boldsymbol{\mu}\right)$. We seek the $\boldsymbol{\mu}$ that maximizes the Lagrangian

$$L(\boldsymbol{\mu},\lambda) = \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\left(\log \eta_j + \log \mu_j\right) - \lambda\sum_{j=1}^{K}\eta_j\mu_j$$

where λ is a Lagrange multiplier corresponding to the normalization constraint $\sum_{j=1}^{K}\eta_j\mu_j = 1$. Taking the first derivative with respect to μ_j and λ, setting to zero, and solving the resulting system of equations gives

$$\hat{\mu}_j = \frac{\eta_j^{-1}\tau_j}{\sum\limits_{k=1}^{K}\eta_k^{-1}\tau_k}$$

where we have used $\tau_j = \frac{1}{N}\sum\limits_{i=1}^{N} Z_{ij}$ to paraphrase the result. This simple correction suggests that in the update scheme for `vbdiff` we should adjust the occupations for state j by a factor proportional to η_j^{-1}:

$$q\left(\boldsymbol{\tau}\right) = \text{Dirichlet}\left(\alpha_0 + \overline{\eta}_1^{-1}\alpha_1, ..., \alpha_0 + \overline{\eta}_K^{-1}\alpha_K\right)$$

where $\overline{\eta}_j \propto \eta_j$ such that the overall statistical weight of the trajectories (number of jumps in the dataset) is unchanged.

In `vbdiff`, we evaluate $\eta_j = \eta_j(\phi_j)$ at each iteration using a numerical propagator approach (see `emdiff/defoc.py`).

A Appendix A: Derivation of vbdiff approximate posterior

To identify the exact form of the approximative posterior $q(\boldsymbol{X}, \boldsymbol{\tau}, \boldsymbol{\phi})$ for Brownian mixtures, we will solve the system of equations 18 given the model specified by 13, 14, 15, and 16.

First we consider the second equation in 18. Writing this out,

$$\begin{aligned}
\log q\left(\boldsymbol{\tau}, \boldsymbol{\phi}\right) &= \mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}\left[\log p\left(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\tau}, \boldsymbol{\phi}\right)\right] \\
&= \sum_{j=1}^{K}\left(\alpha_0 - 1 + \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right] m_i\right)\log\tau_j \\
&\quad - \sum_{j=1}^{K}\left(\beta_0 + \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right] x_i\right)\phi_j^{-1} \\
&\quad - \sum_{j=1}^{K}\left(\alpha_0 + 1 + \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right] m_i\right)\log\phi_j \\
&\quad + \text{constant}
\end{aligned} \tag{30}$$

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 30, 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 9 - 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 30, we see that it is a product of a Dirichlet distribution over $\boldsymbol{\tau}$ and independent inverse gamma distributions over each scale parameter ϕ_j:

$$\begin{aligned}
q\left(\boldsymbol{\tau}, \boldsymbol{\phi}\right) &= q\left(\boldsymbol{\tau}\right)\prod_{j=1}^{K} q\left(\phi_j\right) \\
q\left(\boldsymbol{\tau}\right) &= \text{Dirichlet}\left(\boldsymbol{\tau} \mid \alpha_0 + \alpha_1, ..., \alpha_0 + \alpha_K\right) \\
q\left(\phi_j\right) &= \text{InvGamma}\left(\alpha_0 + \alpha_j, \beta_0 + \beta_j\right) \\
\alpha_j &= \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right] m_i \\
\beta_j &= \sum_{i=1}^{N}\mathbb{E}_q\left[Z_{ij}\right] x_i
\end{aligned} \tag{31}$$

We still need to solve $q(\boldsymbol{Z})$. Expanding the first equation in 18, we have

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

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

$$\begin{aligned}q\left(\boldsymbol{Z}\right) &= \prod_{i=1}^{N}\prod_{j=1}^{K} r_{ij}^{Z_{ij}} \\ r_{ij} &= \frac{\exp\left(\mathbb{E}_q\left[\log\tau_j\right] - x_i\mathbb{E}_q\left[\phi_j^{-1}\right] - m_i\mathbb{E}_q\left[\log\phi_j\right]\right)}{\sum_{k=1}^{K}\exp\left(\mathbb{E}_q\left[\log\tau_k\right] - x_i\mathbb{E}_q\left[\phi_k^{-1}\right] - m_i\mathbb{E}_q\left[\log\phi_k\right]\right)}\end{aligned} \tag{32}$$

Together, equations 31 and 32 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 $q(\boldsymbol{Z})$, $q(\boldsymbol{\tau})$, and $q(\boldsymbol{\phi})$, the updates for each expectation can be evaluated in closed form:

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

where α_j, β_j are defined as in equation 31 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).