Biophysical *Journal*
Article

Variational Algorithms for Analyzing Noisy Multistate Diffusion Trajectories

Martin Lindén[1,*] and Johan Elf[1,*]
[1]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

ABSTRACT Single-particle tracking offers a noninvasive high-resolution probe of biomolecular reactions inside living cells. However, efficient data analysis methods that correctly account for various noise sources are needed to realize the full quantitative potential of the method. We report algorithms for hidden Markov-based analysis of single-particle tracking data, which incorporate most sources of experimental noise, including heterogeneous localization errors and missing positions. Compared to previous implementations, the algorithms offer significant speedups, support for a wider range of inference methods, and a simple user interface. This will enable more advanced and exploratory quantitative analysis of single-particle tracking data.

INTRODUCTION

Experimental techniques to track the conformational and binding states of single biomolecules can offer unique mechanistic insights into life at the molecular level but increasingly rely on statistical computing to extract quantitative and reproducible results. A simple example is super-resolved single-particle tracking (SPT) (1), in which changes in diffusion constant or between different modes of motion offer a noninvasive probe of binding and unbinding reactions in living cells (2–4).

Detecting and following single fluorophores can be challenging, and statistical methods to optimize the spot detection (5) and assembling of molecular trajectories in the presence of uncertain spot detections (6) are active research areas. Next, accurate quantitative analysis of trajectory data requires a faithful account of localization noise, which come in the form of localization errors and motion blur, sometimes referred to as "static" and "dynamic" errors, respectively (7,8). In particular, live cell imaging often lead to heterogeneous and asymmetric localization errors, for example due to photobleaching, variability between and across cells, out-of-focus motion, or the dependence of localization errors on the diffusion constant (9,10). Several emerging techniques for three-dimensional localization also give different precision in the axial and lateral directions (11).

A fundamental unknown in many live cell SPT studies is the number of underlying molecular states, e.g., binding states, which may differ in diffusion constant. Counting diffusive states in SPT data presents a statistical model selection problem that has so far only been solved with simplified noise models (2), which may be inappropriate in many live cell applications (10,12).

Here, we extend our previous hidden Markov model (HMM) analysis (10) by deriving and implementing variational algorithms that increase computational speed by more than an order of magnitude, allow statistical model selection using Bayesian or information-theoretic methods, and can be generalized to a wider class of localization error models. The methods are available in a user-friendly open source software suite.

METHODS

Variational diffusive HMM

The starting point for our analysis is a standard model for camera-based SPT that includes a combination of averaging (motion blur) and localization errors, in which the detected positions x_t are related to the underlying particle trajectory $y(t)$ through

$$x_t = \int_0^{\Delta t} f(t')y(t+t')dt' + \sqrt{v_t}\xi_t^{(x)}. \qquad (1)$$

Here, v_t is the localization error (variance) in frame t, and $\xi_t^{(\cdot)}$ are independent unit normal random variables. The shutter function $f(t)$ describes how the image acquisition is distributed throughout the frame, e.g., $f(t) = 1/\Delta t$ for continuous exposure and acquisition (8).

Submitted March 6, 2018, and accepted for publication May 23, 2018.
*Correspondence: martin.linden@icm.uu.se or johan.elf@icm.uu.se
Editor: Antoine van Oijen.
https://doi.org/10.1016/j.bpj.2018.05.027

We model the particle motion $y(t)$ as free diffusion, with a time-dependent diffusion constant governed by a hidden Markov process s_t with N discrete states.

For a fast variational algorithm, we seek a model in the exponential family of probability distributions (13), which yield variational algorithms of a particularly simple form that are often analytically tractable (14). This is achieved by modeling the two terms in Eq. 1 separately, i.e., keep both the true hidden path y_t and the true exposure-averaged positions z_t (the integral in Eq. 1) as explicit variables. In discrete time, this leads to the following model:

$$y_{t+1} = y_t + \sqrt{2D_{s_t}\Delta t}\xi_t^{(y)}, \tag{2}$$

$$z_t = (1-\tau)y_t + \tau y_{t+1} + \sqrt{\beta 2D_{s_t}\Delta t}\xi_t^{(z)}, \tag{3}$$

$$x_t = z_t + \sqrt{v_t}\xi_t^{(x)}, \tag{4}$$

where τ and β are blur coefficients that depend on the shutter function (10). With $f(t) = 1/\Delta t$ for continuous illumination, we get $\tau = 1/2$ and $\beta = 1/12$ (for details, see Supporting Materials and Methods, Section S1). Position coordinates in two- or three-dimensional trajectories are treated independently, which means that we neglect possible correlations between localization errors in different directions. As detailed in Supporting Materials and Methods, this model allows variational algorithms for both maximal likelihood estimation and variational Bayes inference (VB) (13,15). Missing positions due to, e.g., fluorophore blinking, are handled by formally setting $v_t = \infty$, which eliminates contributions from Eq. 4 for those points.

Our focus in this work is the case in which the localization variances v_t are input data estimated from the localization of single spots (10). However, one could also treat v_t as model parameters, for example as a single average error ($v_t = v$), dependent on the hidden state ($v_t = v_{s_t}$), and/or varying with coordinate dimension. These modified models remain in the exponential family and thus allow similarly efficient variational algorithms that differ only in details compared to our main case.

Simulated Trajectories

For the model selection experiments in Fig. 1 and Fig. S1, we used synthetic trajectories simulated using the analysis model, Eqs. 2, 3, and 4. We simulated a three-state model with the following parameters: diffusion constants $D_1 = 0.1\,\mu\text{m}^2\text{s}^{-1}$, $D_2 = 6\,\mu\text{m}^2\text{s}^{-1}$, and $D_3 = 3\,\mu\text{m}^2\text{s}^{-1}$. The kinetic model is an irreversible cycle $D_1 \rightarrow D_2 \rightarrow D_3 \rightarrow D_1 \rightarrow \ldots$ with exponentially distributed waiting times with average 100 ms (see Fig. 1 *b*). The positions are simulated according to Eqs. 2, 3, and 4, with time step $\Delta t = 5$ ms, and motion blur corresponding to an exposure time of $t_E = 1.5$ ms ($\tau = 0.15$, $\beta = 0.0775$). Trajectories were confined to $\left|z_t^{(z)}\right| < 500\,\text{nm}$ using a trajectory-wise method of images, i.e., reflecting trajectory parts outside this interval back in again ($z_t^{(z)}$ is the z component of z_t).

For the static localization errors v_t, we use a simple model of spot widening due to defocus Δz (9),

$$\sigma(\Delta z, D)^2 = \sigma_0^2\left(1 + \left(\frac{\Delta z}{L_z}\right)^2\right) + \frac{a^2}{12} + \frac{1}{3}Dt_E, \tag{5}$$

with minimal spot width $\sigma_0 = 100\,\text{nm}$, $L_z = 240\,\text{nm}$ (approximating $\lambda = 638$ nm, $NA = 1.4$, in water), and $a = 80\,\text{nm}$. The a^2 term approximates the effect of finite pixel size (16) and the Dt_E term spot-widening due to motion blur (9). We then compute v_t for use in Eq. 4 from the approximate Cramer-Rao lower bound (16):

$$v_t = 2\frac{\sigma\left(z_t^{(z)}, D_{s_t}\right)^2}{N_{phot.}}\left(\frac{16}{9} + \frac{8\pi b^2 \sigma\left(z_t^{(z)}, D\right)^2}{N_{phot.}a^2}\right), \tag{6}$$

with $N_{phot.} = 200$ photons per spot. This gave $14\,\text{nm} < \sqrt{v_t} < 41\text{nm}$. Fig. 1 *b* shows the curve for $D = 6\,\mu\text{m}^2\text{s}^{-1}$. We analyzed the x and y components of x_t and chose trajectory lengths to be exponentially distributed with mean length 25Δt, but discarded trajectories with length below 5Δt. For the statistical model selection study (Fig. 1; Fig. S1), we sampled data sets of various sizes (50 data sets with up to 32,000 steps, 24 data sets with 60,000 steps) from a large data set of several hundred thousand positions such that all model selection techniques used the same set of trajectories.

Simulated microscopy

Simulated video-microscopy images for transfer-RNA (tRNA) tracking was generated using the SMeagol simulation software (12) with the spatial reaction-diffusion model illustrated in Fig. 4. We simulated uniform exposure during 1.5 of the 5 ms sampling time. Camera noise was generated using a high-gain approximation of electron-multiplying charge-coupled device noise (16) with offset 200, gain 77, and Gaussian readout noise with standard deviation 20. We used 80 nm pixels and a uniform background fluorescence that decayed from two to one photon/pixel with a decay rate of 2 s^{-1}. For the optics, we used a Gibson-Lanni point-spread function (PSF) model (17) generated by PSFgenerator (18) with $\lambda = 680\,\text{nm}$ and $NA = 1.49$. This is a spherically symmetric PSF suitable for isotropic emitters or fluorophores with high rotational mobility. Fluorescent spot intensity was set to give on average of 200 photons per frame, and the average bleaching time was chosen to 20 frames. Using custom MATLAB (The MathWorks, Natick, MA) scripts, we simulated 200-frame movies with ~30 cells spread evenly across a 512 × 130 pixel field of view, with a few active fluorescent spots per cell. An example of one such cell is shown in Fig. 4 *c*.

Spot detection and localization

We use the fast radial symmetry transform (19) for spot detection and estimated spot positions and localization uncertainty using a symmetric Gaussian spot model and maximal aposteriori estimates on 9-by-9 regions of interests, as described by Lindén et al. (10). Spots with $\sqrt{v_t} > 80\,\text{nm}$ were discarded from the analysis.

RESULTS

Model selection

The number of diffusive states is often a biological unknown of great interest, but because different numbers of diffusive states correspond to statistical models with different numbers of parameters, the counting of states is a nontrivial problem of statistical model selection.

Bayesian reasoning, including model selection, is an extension of formal logic to uncertain statements that yields unique and consistent results (20). Assuming equal prior preference for a set of candidate models with uncertain parameters and unobserved degrees of freedom (latent variables), the Bayesian approach uses marginalization to select the model with the largest (log) evidence

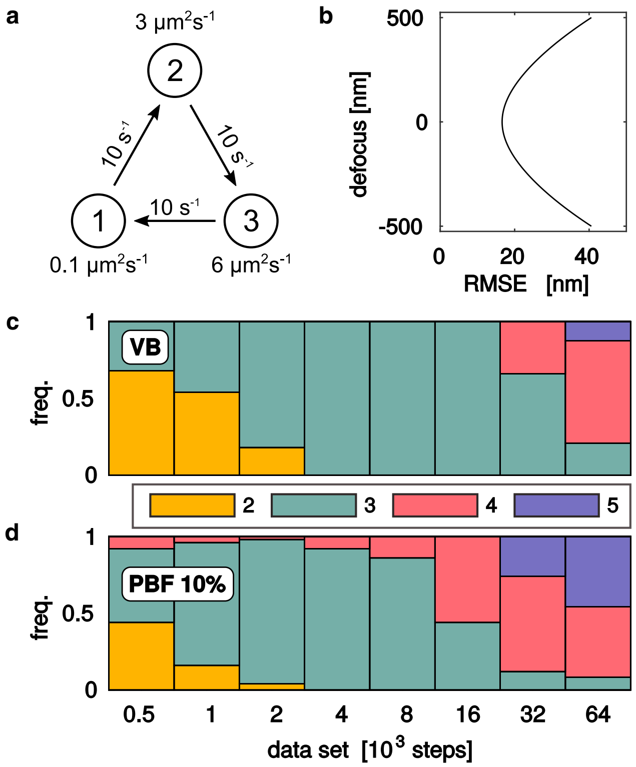

FIGURE 1 Statistical model selection. We generated a range of synthetic data sets with (*a*) three diffusive states and (*b*) defocus-dependent localization root mean-square errors, and estimated the number of states using (*c*) variational maximal evidence (VB), and (*d*) cross-validation using variational pseudo-Bayes factors using 10% of the data in the validation set (PBF 10%). For details, see Methods. To see this figure in color, go online.

(13,15), which in compact notation can be written as follows:

$$\ln p(x|M) = \ln \int dn\, d\theta\, p(x, n|\theta, M) p_0(\theta|M). \qquad (7)$$

Here, x denotes the observed data, M denotes the model, $n = \{s_t, y_t, z_t\}$ the latent variables (summed or integrated out as appropriate), and θ_M denotes the unknown parameters with prior distribution $p_0(\bullet|M)$ for the different models M.

In our case and many others, the evidence in Eq. 7 is analytically intractable, in which case a variational Bayes (VB) approximation can be an attractive approach (2,13–15,21–24). VB yields a lower bound $\ln L \leq \ln p(x|M)$ usable for approximate Bayesian model selection as well as approximate posterior distributions (variational distributions) of parameters and hidden states (13–15). Moreover, because it involves direct optimization of the lower bound $\ln L$, VB algorithms have an intrinsic parsimony that depopulates superfluous states and can be utilized for efficient greedy model search algorithms (2,24).

However, Bayesian inference may be statistically inefficient. In particular, the common practice of using uninformative priors to minimize bias in parameter estimates means that the prior likelihood for any particular parameter value is low. This in turn can lead to overly steep penalties against models with many parameters, a phenomenon known as Lindley's paradox, which means that an unnecessarily large amount of data is needed to resolve some feature of interest (25–27). An alternative non-Bayesian approach that avoids this difficulty is to rank competing models by their estimated predictive performance (26,27). Next, we explore a predictive approach to model selection for SPT analysis.

The most well-known predictive performance measure is Akaike's information criterion (AIC) (28), but this is only asymptotically valid for large data sets. For small data sets, one could instead use cross-validation, in which the data set is divided into two parts: one for estimating model parameters ("training") and one for estimating predictive performance ("validation"). In practice, the performance is estimated from averaging over several such divisions. We implemented a variant of cross-validation with a Bayesian flavor, pseudo-Bayes factors (PBF) (27), which include prior distributions but do not suffer from Lindley's paradox and are easy to compute with our variational algorithm (see Supporting Materials and Methods, Section S2.5).

Fig. 1 compares VB and PBF model selection on synthetic test data with three diffusive states and parameters that resemble in vivo SPT experiments in bacteria (2,4) (see Methods). Broadly, one expects predictive model selection to avoid Lindley's paradox and penalize complex models less severely than Bayesian methods as the amount of data increases. However, this comes at the expense of consistency, i.e., there is no guaranteed convergence to the correct model size (28,29). These expectations are qualitatively borne out in Fig. 1, where the Bayesian VB criterion is more prone than PBF to select too few states for small data sets, but less prone to select too many states for large data sets.

There is no general rule for selecting training and validation subsets. HMMs also suffer from the additional complication that individual observations are correlated because of the hidden state dynamics, which complicates cross-validation if the data is a single trajectory (30). Here, we focus on SPT experiments that produce a large number of trajectories (2,4) that can be used as atoms for constructing training and validation sets. Our simulated trajectories have an average length of 30, and Fig. 1 *b* uses randomly sampled validation sets containing ~10% of the data. Some other choices, including AIC and the Bayesian vbSPT (VB for SPT) code (2), are explored in Fig. S1, but do not perform better.

Speedup

In addition to more flexibility in modeling and inference methods, the algorithms presented here are also considerably faster compared to our previous implementation (10). This is mainly because variational algorithms based on Eqs. 2, 3 and 4 are analytically tractable and hence avoid a costly numerical optimization step. However, we have also found a more efficient algorithm for partial matrix inversion (31). Fig. 2 shows the time per iteration for a three-state model on data sets of different sizes for the algorithm presented here, that of (10), and vbSPT (2). Compared to the former, we see speedups of one to two orders of magnitude for experimentally relevant data set sizes of $10^4 - 10^5$ positions as well as better scaling. However, vbSPT is faster still, which is expected because it is based on a much simpler model and thus has less to do during each iterative update.

Finding the global optimum

Variational learning of a model and its parameters, diffusion constants, and transition rates involves finding the overall best fit to the data, but VB and other expectation-maximization-type algorithms only converge toward local optima. An additional global search is needed.

The simplest approach is to converge multiple models from different starting points. To speed things up, we use the built-in parsimony of the VB algorithms to start from complex many-state models and then systematically search for simpler ones by removing un- or low-populated states (2). However, the extra complexity of our model compared to standard HMMs (2) makes this approach more challenging to apply.

One attractive feature of our model is the ability to handle long trajectories with missing positions. However, when fitting high-dimensional models to data with missing positions, groups of superfluous states sometimes converge toward identical parameters and finite occupancy associated with the missing positions. Because this is clearly unphysical, we choose to remove such state clusters before commencing normal model pruning.

Another challenge is related to the presence of two types of latent variables for the discrete diffusive states (s_t) and uncertainty in true particle positions (y_t, z_t), respectively. Although hard to quantify, it seems reasonable to expect more latent variables to yield a more complex search landscape, with more local optima for the search to get trapped in compared to ordinary HMMs in which the particle positions are not latent variables (2). More concretely, the variational treatment uses a threefold factorization ansatz (parameters, hidden states, and hidden particle trajectories), and to initialize the local optimization iterations, two of three factors need to be initialized.

We use randomly selected parameter values and explore different strategies to initialize either hidden states or hidden trajectories: uniform hidden state occupancy, hidden trajectories modeled directly on observed data (with no uncertainty or correlations between y_t and z_t), and hidden trajectory models generated by a running average in which a pure diffusion model is fit to small windows of various lengths. In our testing, different methods perform best on different types of data, meaning that a wide range of initialization methods are needed to maximize the chance of finding the global optimum.

Fig. 3 shows an analysis of a data set from simulated images (see Methods) using 50 independent initializations of model parameters with 15 states and 10 different initializations of latent variables with each parameter set. Fig. 3 *a* shows the lower bounds of models originating from a single initial parameter set, with each line corresponding to models generated by the reductive search starting from one latent variable initialization. We see that the initialization with the largest number of nonspurious states does not lead to the best overall model, and that the search lines sometimes cross, meaning that relative ranking among the different

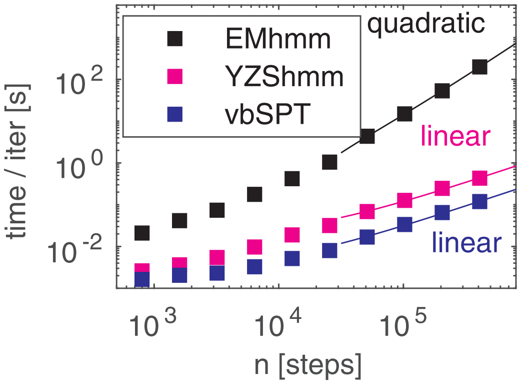

FIGURE 2 Speed of our variational algorithm (YZShmm) compared to that of Ref. (10) (EMhmm) and vbSPT (2). Time per iteration versus number of steps in the data, which has three diffusive states, was measured on a dual 6-core Intel Xeon 2.4GHz computer running MATLAB R2017a. Linear and quadratic scaling laws are guides to the eye. To see this figure in color, go online.

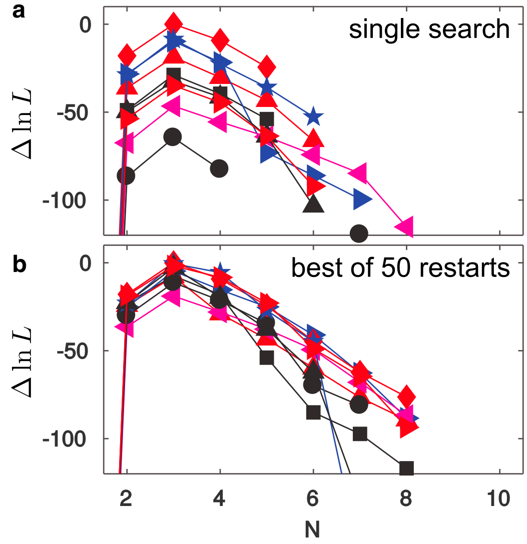

FIGURE 3 Model search with different initialization strategies. Each color/marker combination shows the relative lower bound $\Delta \ln L$ from the best model of each size for different initialization strategies. (*a*) Model search from a single parameter initialization is shown. (*b*) The best models from 50 independent initializations are shown. To see this figure in color, go online.

reduction searches can change as states are removed. Looking at the best models from 50 independent parameter initializations (Fig. 3 *b*), we again see search lines crossing and note that the two highest-ranked model sizes originate from different initialization methods.

Application: simulated tRNA tracking

Quantitative live cell SPT is complex, and errors may arise during measurement, spot detection, localization, trajectory building, and trajectory analysis. Comprehensive tests of the whole analysis chain are needed to validate quantitative interpretations of the experiments under particular conditions. To evaluate the capabilities of our trajectory analysis only, we seek test data with known ground truth and sufficient realism to be experimentally relevant. We use simulated video microscopy (12) to produce realistic test data and run spot detection and localization using our standard methods (see Methods) but use our knowledge of the simulated ground truth to produce trajectories free from false positives and linking errors that may lead to bad performance that does not reflect the intrinsic quality of the trajectory analysis. We allow at most three consecutive missing positions before starting a new trajectory.

As a test problem, we consider tracking tRNA molecules in *Escherichia coli* cells (4,32) (see Fig. 4), which presents several interesting difficulties. At least three discernible diffusive states may be expected: a ribosome-bound state (B, slow diffusion), an unbound state (U, fast diffusion), and a ternary complex (TC, intermediate diffusion). The ribosome-bound state further displays spatial structure in the form of nucleoid exclusion (32) as well as nonexponential waiting times (33) because tRNA goes through several reaction steps before dissociating from the ribosome (34). We constructed a simplified kinetic and spatial model incorporating these features and generated synthetic fluorescent microscopy data with a 200 Hz frame rate (12) (see Methods). We simulated a range of rate constants corresponding to a total bound state mean dwell time of $\tau_B = 2/k_u$ between 50 and 400 ms, whereas the steady-state occupancy is kept fixed at 20/30/50 (B/U/TC).

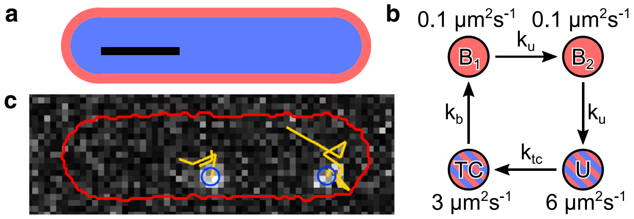

FIGURE 4 Model for simulated tracking of fluorescent tRNA molecules. (*a*) Shows the cross section of the simulation geometry, which consists of concentric cylinders with spherical end-caps, represents the nucleoid (*blue*) floating in the cytoplasm (*red*). Scale bars, (*black*) 1 μm. (*b*) Shows the simulated kinetic model of the tRNA cycle. Two states B_1, B_2 with low diffusion coefficients represent ribosome-bound states and are excluded from the nucleoid, whereas the unbound (U) and ternary complex (TC) states are free to roam the whole cell. (*c*) Shows a simulated frame with two fluorophores in a single cell, with cell outline (*red*) and particle tracks (*yellow*) added. Pixel size, 80 nm. To see this figure in color, go online.

Not all estimated parameters have direct simulated counterparts. For example, nucleoid exclusion means that the $TC \rightarrow B$ reaction cannot take place in the nucleoid region, lowering the effective value of k_b. Nucleoid exclusion also distorts the state occupancy of the detected spot population because defocused spots are more difficult to detect, and with our simulated focus in the cell midplane, the bound states are relatively enriched in defocused regions (4). Fig. 5 shows comparisons in which these complications are minimal, which means that we ignore overall occupancy and transitions out of the TC state.

The true model contains three diffusion constants but four kinetic states. Starting with the number of states (Fig. 5 *a*), we see mostly three states and note that the VB and PBF model selection agree half the time and that the PBF favors more states in cases of disagreement. Plotting the diffusion constants from VB-selected models (Fig. 5 *b*), we see that it finds diffusion constants close to the true values, although for the fastest kinetics ($\tau_B = 50$ ms), the high-D states are biased toward each other, probably because the faster dynamics produces more short events that make it more difficult for the HMM to distinguish the two fast states correctly. There is also a general downward bias in the highest diffusion constant, most likely a confinement artifact (see Fig. S2).

Regarding the kinetics, a detailed look at the four-state model for the fastest kinetics (Fig. 5 *c*) shows a striking resemblance to the true rate model in Fig. 5 *d*. The unidirectional cycle is clearly visible in the estimated parameters, and the transition probabilities corresponding to k_u and k_{tc} closely resemble the underlying ground truth. The mean dwell times of this model are comparable to the average trajectory length of ~0.12 s. For models with slower kinetics, the bound state dwell time is well captured (Fig. 5 *e*, *black*), but otherwise, the kinetics is not as well reproduced. Only one bound state is identified, the unbound dwell times (Fig. 5 *f*) are not close to the true values, and the transition matrices (data not shown) do not resemble the cyclic pattern of the underlying model. However, even the more limited ability to measure the mean dwell time of a slow or immobile state when that dwell time exceeds the average trajectory length could be of biological interest, for example to study the interactions of small molecules such as tRNA or proteins interacting with larger structures such as ribosomes or DNA (4).

DISCUSSION

Together with methods to extract both positions and position uncertainty from images of single spots (10), the variational algorithm we present here makes it possible to significantly decrease analysis artifacts associated with variable localization quality due to, for example, out-of-focus motion,

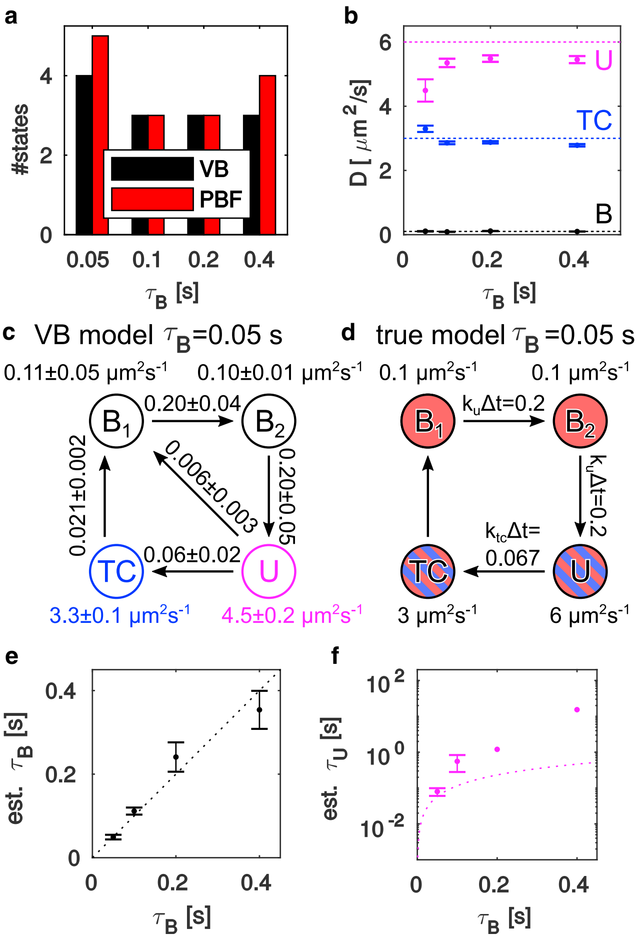

FIGURE 5 Analysis of simulated microscopy data. The different ground truth models are denoted by their total bound state dwell times: $\tau_B = 0.05$, 0.1, 0.2, and 0.4s, respectively. (*a*) Shows the number of states selected by the VB and PBF criteria. (*b*) Shows diffusion coefficients for the VB-selected models. Dashed colored lines indicate the true diffusion constants of the U, TC, and $B_{1,2}$ states. For the 0.05 s model, two states near 0.1 $\mu m^2 s^{-1}$ are found. (*c*) VB-estimated parameters and (*d*) ground truth parameters of the four-state 0.05 s model are shown with transition probabilities per time step below $10^{-8} \Delta t^{-1}$ suppressed. States are named and colored according to the obvious similarity with the true scheme in (*c*). (*e*) Bound and (*f*) unbound state mean dwell times, computed from the transition probability matrix. For the 0.05s model, we added the dwell times of the two B states. Dashed lines indicate the true mean dwell times. Error bars indicate bootstrap SEs. In (*f*), missing error bars indicate where not all bootstrap replicas gave finite mean dwell times. All data sets contain ~16,000 steps. To see this figure in color, go online.

gradual bleaching or stage drift, or fast fluorophore blinking. We are curious to see how these tools will help researchers make more nuanced interpretations of their in vivo SPT data.

Compared to our previous implementation (10), the algorithms presented here are significantly faster and support maximal likelihood as well as VB inference. This makes exploratory analysis of large data sets practical and allows a more comprehensive statistical analysis. We compared model selection by the purely Bayesian VB approach (14,15) to two methods based on predictive performance, the AIC (28), and a variational implementation of cross-validation using PBFs (27). Although no method avoided overfitting completely in our challenging synthetic data set, VB overfitted the least, and we recommend that for applications.

However, in light of the theoretical arguments against a purely Bayesian approach for some model selection problems (25,26), we think the non-Bayesian methods merit further study. For example, corrections to AIC have been derived for Markov switching regression models (35) and might be generalized to our class of HMMs as well. It is also possible that the PBFs would perform better when evaluated with Monte Carlo methods (27) than with the variational approximations we used here.

There are many interesting directions to further optimize and expand these types of analysis and algorithms.

The complex statistical model used here makes it possible to tackle complex data but computationally difficult to find the globally best model. We use a brute force approach with randomly initialized greedy search for easy parallelization. This is computationally costly, and the total analysis time with our code can be 10–100 times slower than a simplified analysis with vbSPT (2) on the same trajectory set. More sophisticated global optimization schemes in which the different search processes communicate may be more efficient, for example by avoiding redundant efforts when multiple initializations converge to the same model.

Another interesting direction for further development may be to incorporate other types of heterogeneity, such as variability in the underlying diffusion constants or other model parameters (23) or explicit models of spatial structure (36).

Third, more complex motion or kinetic models could be used. Our diffusive HMM may be extended in several useful ways within the exponential family of models that enable efficient variational algorithms (14); localization errors could be treated as model parameters rather than external observations, possibly depending on the chemical state or coordinate dimension (see Supporting Materials and Methods, Section S5). There are also combinations of directed motion and confinement in harmonic potentials that still lead to Gaussian motion models (3,37). Introducing explicit termination rates could correct bias that arises from correlations between chemical states and trajectory termination (38), for example when fast-diffusing molecules move out of focus faster than slow-diffusing ones (39).

Software

Our algorithms are freely available as open source MATLAB code from https://github.com/bmelinden/vbSPTu. The vbSPTu software suite includes a GUI to run a simple standard analysis, support for scripting large analysis tasks, and low-level tools for creating customized analysis.

SUPPORTING MATERIAL

Supporting Materials and Methods and two figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30665-9.

AUTHOR CONTRIBUTIONS

M.L. and J.E. planned the research and wrote the article. M.L. designed and implemented the analysis methods and generated and analyzed the data.

ACKNOWLEDGMENTS

We are grateful to Ivan Volkov, Magnus Johansson, Elias Amselem, David Fange, and members of the Elf laboratory for sharing their insight on imaging and SPT and to Irmeli Barkefors for constructive comments on drafts of the article.

This project was funded by grants to J.E. from the Knut and Alice Wallenberg Foundation and the European Research Council (ERC-2013-CoG 616047 SMILE).

REFERENCES

1. Manley, S., J. M. Gillette, …, J. Lippincott-Schwartz. 2008. High-density mapping of single-molecule trajectories with photoactivated localization microscopy. *Nat. Methods.* 5:155–157.
2. Persson, F., M. Lindén, …, J. Elf. 2013. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods.* 10:265–269.
3. Monnier, N., Z. Barry, …, M. Bathe. 2015. Inferring transient particle transport dynamics in live cells. *Nat. Methods.* 12:838–840.
4. Volkov, I. L., M. Lindén, …, M. Johansson. 2018. tRNA tracking for direct measurements of protein synthesis kinetics in live cells. *Nat. Chem. Biol.* 14:618–626.
5. Smith, C. S., S. Stallinga, …, D. Grunwald. 2015. Probability-based particle detection that enables threshold-free and robust in vivo single-molecule tracking. *Mol. Biol. Cell.* 26:4057–4062.
6. Chenouard, N., I. Bloch, and J. C. Olivo-Marin. 2013. Multiple hypothesis tracking for cluttered biological image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* 35:2736–3750.
7. Savin, T., and P. S. Doyle. 2005. Static and dynamic errors in particle tracking microrheology. *Biophys. J.* 88:623–638.
8. Berglund, A. J. 2010. Statistics of camera-based single-particle tracking. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 82:011917.
9. Deschout, H., K. Neyts, and K. Braeckmans. 2012. The influence of movement on the localization precision of sub-resolution particles in fluorescence microscopy. *J. Biophotonics.* 5:97–109.
10. Lindén, M., V. Ćurić, …, J. Elf. 2017. Pointwise error estimates in localization microscopy. *Nat. Commun.* 8:15115.
11. Rieger, B., and S. Stallinga. 2014. The lateral and axial localization uncertainty in super-resolution light microscopy. *ChemPhysChem.* 15:664–670.
12. Lindén, M., V. Ćurić, …, J. Elf. 2016. Simulated single molecule microscopy with SMeagol. *Bioinformatics.* 32:2394–2395.
13. Bishop, C. 2006. Pattern Recognition and Machine Learning. Springer, New York.
14. Beal, M. 2003. Variational algorithms for approximate Bayesian inference. PhD thesis. University of Cambridge.
15. MacKay, D. 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, United Kingdom.
16. Mortensen, K. I., L. S. Churchman, …, H. Flyvbjerg. 2010. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nat. Methods.* 7:377–381.
17. Gibson, S. F., and F. Lanni. 1992. Experimental test of an analytical model of aberration in an oil-immersion objective lens used in three-dimensional light microscopy. *J. Opt. Soc. Am. A.* 9:154–166.
18. Kirshner, H., F. Aguet, …, M. Unser. 2013. 3-D PSF fitting for fluorescence microscopy: implementation and localization application. *J. Microsc.* 249:13–25.
19. Loy, G., and A. Zelinsky. 2003. Fast radial symmetry for detecting points of interest. *IEEE Trans. Pattern Anal. Mach. Intell.* 25:959–973.
20. Cox, R. T. 1946. Probability, frequency and reasonable expectation. *Am. J. Phys.* 14:1–13.
21. 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.
22. 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.
23. van de Meent, J. W., J. E. Bronson, …, R. L. Gonzalez, Jr. 2014. Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophys. J.* 106:1327–1337.
24. Johnson, S., J. W. van de Meent, …, M. Lindén. 2014. Multiple LacI-mediated loops revealed by Bayesian statistics and tethered particle motion. *Nucleic Acids Res.* 42:10265–10277.
25. Cousins, R. D. 2017. The Jeffreys-Lindley paradox and discovery criteria in high energy physics. *Synthese.* 194:395–432.
26. LaMont, C. H., and P. A. Wiggins. 2016. The Lindley paradox: The loss of resolution in Bayesian inference. *arXiv*, arXiv:1610.09433 https://arxiv.org/abs/1610.09433.
27. Gelfand, A. E., and D. K. Dey. 1994. Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Stat. Soc. B Met.* 56:501–514.
28. Burnham, K. P., and D. R. Anderson. 2013. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.
29. Shao, J. 1993. Linear model selection by cross-validation. *J. Am. Stat. Assoc.* 88:486–494.
30. Celeux, G., and J.-B. Durand. 2008. Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Stat.* 23:541–564.
31. Meurant, G. 1992. A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. *SIAM J. Matrix Anal. Appl.* 13:707–728.
32. Plochowietz, A., I. Farrell, …, A. N. Kapanidis. 2017. *In vivo* single-RNA tracking shows that most tRNA diffuses freely in live bacteria. *Nucleic Acids Res.* 45:926–937.
33. Kienker, P. 1989. Equivalence of aggregated Markov models of ion-channel gating. *Proc. R. Soc. Lond. B Biol. Sci.* 236:269–309.
34. Steitz, T. A. 2008. A structural understanding of the dynamic ribosome machine. *Nat. Rev. Mol. Cell Biol.* 9:242–253.
35. Smith, A., P. A. Naik, and C.-L. Tsai. 2006. Markov-switching model selection using Kullback-Leibler divergence. *J. Econom.* 134:553–577.
36. El Beheiry, M., S. Türkcan, …, J. B. Masson. 2016. A primer on the Bayesian approach to high-density single-molecule trajectories analysis. *Biophys. J.* 110:1209–1215.
37. Calderon, C. P. 2016. Motion blur filtering: a statistical approach for extracting confinement forces and diffusivity from a single blurred trajectory. *Phys. Rev. E.* 93:053303.
38. Kolomeisky, A. B., and M. E. Fisher. 2000. Periodic sequential kinetic models with jumping, branching and deaths. *Physica A.* 279:1–20.
39. Kues, T., and U. Kubitscheck. 2002. Single molecule motion perpendicular to the focal plane of a microscope: application to splicing factor dynamics within the cell nucleus. *Single Mol.* 3:218–224.

Biophysical Journal, Volume *115*

Supplemental Information

Variational Algorithms for Analyzing Noisy Multistate Diffusion Trajectories

Martin Lindén and Johan Elf

Variational algorithms for analyzing noisy multi-state diffusion trajectories – supplementary material

Martin Lindén* and Johan Elf†
Department of Cell and Molecular Biology, Uppsala University, Sweden.
(Dated: May 25, 2018)

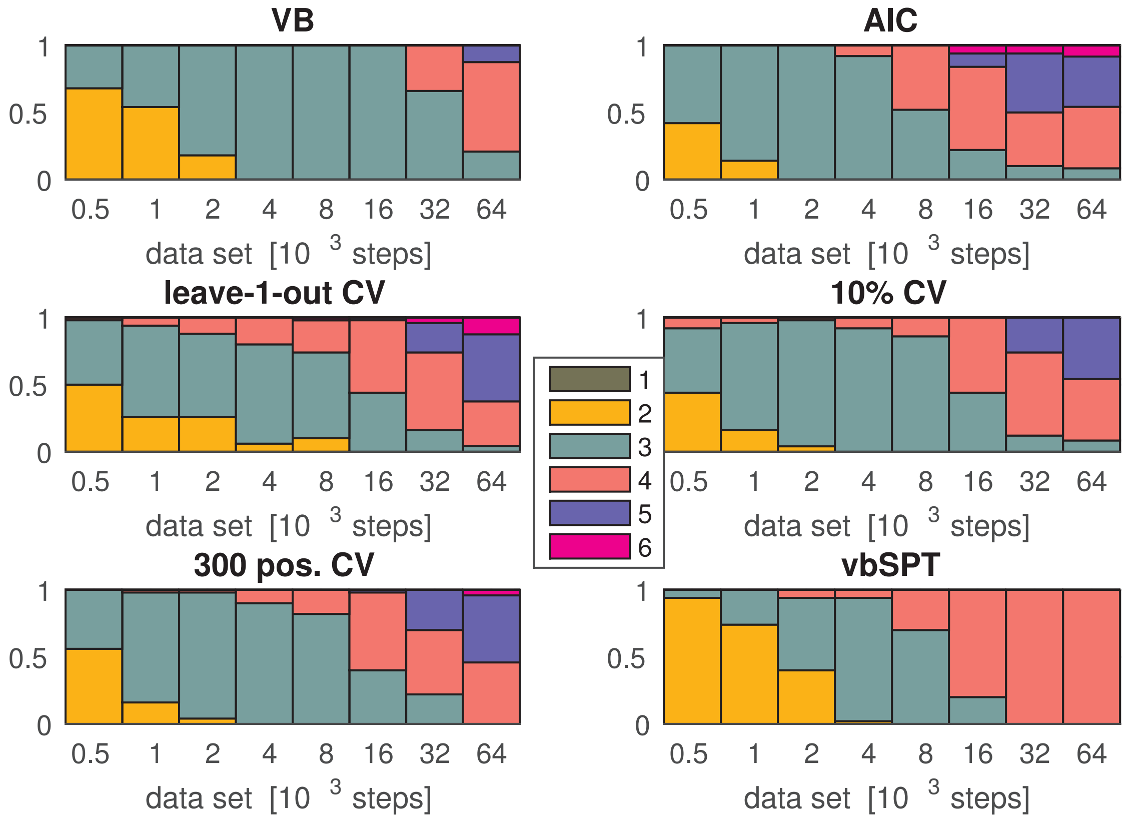

Figure S1: Performance of different model selection criteria on the simulated 3-state data sets used in Fig. 1. The graphs show the fraction of data sets classified to different number of states, where 3 states is the true answer. Model selection criteria: Criteria: Variational Bayes ("VB"), Akaike's information criterion ("AIC"), pseudo-Bayes factor (PBF) cross-validation on single trajectories ("leave-1-out CV"), PBF cross-validation on randomly selected trajectories corresponding to 10% of the total data set ("10% CV"), PBF cross-validation on randomly selected trajectories with about 300 positions in total ("300 pos. CV"), variational Bayes with vbSPT [1], which uses a model without localization errors or motion blur ("vbSPT").

* martin.linden@icm.uu.se
† johan.elf@icm.uu.se

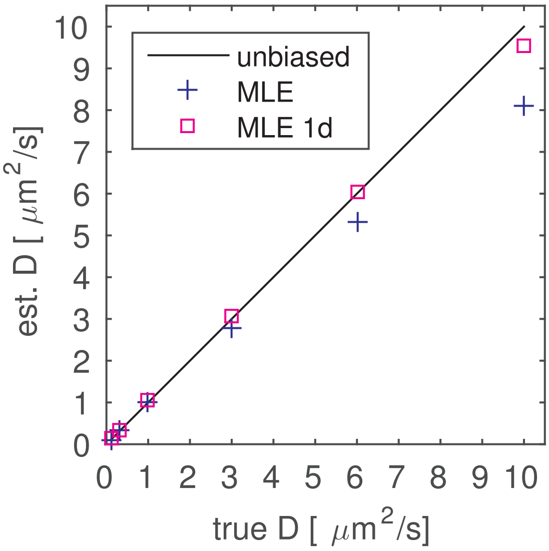

Figure S2: Diffusion constant estimates in an *E coli* geometry. Maximum likelihood estimates of diffusion constant, from simulated movies with a single diffusive state and otherwise as for Fig. 5. We estimated based on both x, y-coordinates (+) and position along the long cell axis ("MLE 1d", square). The downward bias in the x, y-analysis for high diffusion constants is comparable to the one seen for slow kinetics in Fig. 5b, which also used x, y-coordinates, but using the long axis (1d) coordinate only reduces the bias. This rules out state-switching or bias from prior parameters (which are absent here). However, the fact that the 1d analysis is less biased argues strongly for a confinement artifact, since motion perpendicular to the cell long axis is more severely confined. Bootstrap standard error (not shown) are smaller than the symbols.

In the next sections, we go through the trajectory model and derivation of the variational inference algorithms in detail. We first consider the case where both positions and localization uncertainties are estimated from the data. Variations of the algorithms where localization uncertainties are instead model parameters are considered in Sec. S5. This text is an excerpt from the vbSPTu software documentation, available at `https://github.com/bmelinden/`.

CONTENTS

S1. MODEL WITH ESTIMATED UNCERTAINTIES

Here, we simply state a diffusive hidden Markov model which assumes free diffusion in dim dimensions and generalizes the Berglund model of camera-based tracking [2] to free diffusion in dim dimensions, with diffusion constant goverend by a distrete state Markov process, and arbitrary uncorrelated localization errors. This model was previously considered in Ref. [3], and we refer to the supporting information of that work for details of the derivation. Here, we closely follow that formulation with some crucial differences in mathematical formulation (but not in physical interpretation) as noted below, which lead to much more efficient inference algorithms. Later, we also consider some variations of the model.

S1.1. Data

We assume that we measure a set of positions x_{tm} with associated uncertainties (variances of a Gaussian distribution) v_{tm}, where $t = 1, 2, \ldots, T$ describes time points in a trajectory, and $m = 1, 2, \ldots, \mathrm{dim}$ is the coordinate dimension. We allow for missing data points, which we keep track of with the indicator variable

$$o_t = \begin{cases} 1 & \text{, if there is data at time } t, \\ 0 & \text{, if not.} \end{cases} \tag{S1}$$

As we will see, one can in practise dispense with o_t almost everywhere if missing data points are assigned the value $v_{tm} \to \infty$, so that $1/v_{tm} = 0$. Data in several dimensions are assumed independent, which among other things mean that we neglect correlations between the localization errors in different coordinates. The application we have in mind is mostly data sets consisting of many independent trajectories, but for notational simplicity we do the math for a single trajectory only, except when otherwise noted.

S1.2. Equations of motion

a. Hidden states $s_t \in \{1, 2, \ldots, N\}$, $t = 1, 2, \ldots, T$:

$$p(s_1) = \pi_{s_1}, \quad p(s_{t+1}|s_t) = A_{s_t, s_{t+1}}. \tag{S2}$$

We follow recent versions of vbSPT [1] and parameterize A_{ij} in a form that makes it easier to formulate priors using physical insights about dwell times, and set

$$A_{ij} = \delta_{ij}(1 - a_i) + (1 - \delta_{ij}) a_i B_{ij} = (1 - a_i)^{\delta_{ij}} a_i^{(1-\delta_{ij})} B_{ij}^{(1-\delta_{ij})}, \tag{S3}$$

with constraints

$$0 \le a_j \le 1, \quad B_{ii} = 0, \quad \sum_{j \ne i} B_{ij} = 1. \tag{S4}$$

This means that $a_i = p(s_{t+1} \ne i | s_t = i)$ is the probability to exit state i, and $B_{ij} = p(s_{t+1} = j | s_t = i, i \ne j)$ is a matrix of jump probabilities, conditional on a jump actually occurring. While, non-standard for HMMs, we believe that this allows for formulating prior distributions that are easier to interpret, and thus can be formulated with greater confidence. However, in Sec. S4.2 we show that a certain class of prior choices lead to algorithms that are mathematically equivalent to the more conventional formulation where A is explicitly kept as a model parameter.

b. True diffusive path

$$y_{t+1,m} = y_{tm} + \sqrt{\lambda_{s_t}} \varepsilon_{tm}, \quad \varepsilon_{tm} \in \mathrm{N(0,1)} \text{ iid.}, \tag{S5}$$

where $\lambda_{s_t} = 2 D_{s_t} \Delta t$ is the diffusive step-length variance.

c. Measured and motion-averaged positions We model two sources of position noise, motion blur and localization errors, by the equation

$$x_{tm} = \underbrace{\int_0^{\Delta t} f(t') y_m(t + t') dt'}_{\equiv z_{tm}} + \sqrt{v_{tm}} \xi_{tm}^{(x)}, \tag{S6}$$

where $y_m(t)$ is the continuous true diffusion path.

The first term z_{tm} represent the motion-averaged position, and can be rewritten in terms of the discrete true positions y_{tm} [3], as

$$z_{tm} = (1-\tau)y_{tm} + \tau y_{t+1,m} + \sqrt{\beta\lambda_{s_t}}\zeta_{tm}. \tag{S7}$$

Here, τ and $\beta = \tau(1-\tau) - R$ are blur coefficients introduced in Ref. [3], and R is the original blur coefficient of Berglund [2]:

$$\tau = \frac{1}{\Delta t}\int_0^{\Delta t} f(t)t\,dt, \quad F(t) = \int_0^t f(t')dt', \quad R = \int_0^{\Delta t} F(t)(1-F(t))dt, \tag{S8}$$

where $f(t)$ is the shutter function [2], which describes the distribution of emitted light in each exposure. For example, continuous exposure during a time t_E in the beginning of each frame ($0 < t_E \leq \Delta t$) leads to [3]

$$f(t) = \begin{cases} \frac{1}{t_E}, & 0 \leq t \leq t_E, \\ 0, & t_E < t < \Delta t. \end{cases} \Rightarrow \tau = \frac{1}{2}\frac{t_E}{\Delta t}, \quad R = \frac{1}{6}\frac{t_E}{\Delta t}, \quad \beta = \frac{1}{4}\frac{t_E}{\Delta t}\Big(\frac{4}{3} - \frac{t_E}{\Delta t}\Big). \tag{S9}$$

The second term in Eq. (S6) represents a localization error, and thus we can write the measured position as

$$x_{tm} = z_{tm} + \sqrt{v_{tm}}\xi_{tm}, \tag{S10}$$

where v_t is the variance of the localization uncertainty at time t.

This formulation differs from our previous treatment [3] in that we do not integrate out z_{tm} at this point, although it can be done analytically. Instead, we keep both y_{tm} and z_{tm} as hidden path variables. While this doubles the number of Gaussian nuisance variables, it keeps the model in the exponential family, which leads to analytically tractable variational algorithms that allow a fully Bayesian treatment, large computational speed-ups, and treatment of several variant models, compared to our previous work.

d. Model parameters

$$\theta = \{\lambda, \pi, B, a\}. \tag{S11}$$

S1.3. Likelihood

Putting it all together, the complete data likelihood can be written

$$p(x,z,y,s|\pi,a,B,\lambda) = p(x|z)p(z|y,s,\lambda)p(y|s,\lambda)p(s|\pi,a,B), \tag{S12}$$

with

$$\ln p(s|\pi,a,B) = \sum_{i=1}^{N} \delta_{i,s_1}\ln\pi_i \tag{S13}$$

$$+\sum_{t=1}^{T-1}\sum_{k,j=1}^{N} \delta_{k,s_t}\delta_{j,s_{t+1}}\Big[\delta_{kj}\ln(1-a_k) + (1-\delta_{kj})\ln a_k + (1-\delta_{kj})\ln B_{kj}\Big], \tag{S14}$$

$$\ln p(y|s,\lambda) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{N}\sum_{m=1}^{\dim} \delta_{ks_t}\Big(\ln(2\pi\lambda_k) + \lambda_k^{-1}(y_{t+1,m} - y_{tm})^2\Big), \tag{S15}$$

$$\ln p(z|y,s,\lambda) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{N}\sum_{m=1}^{\dim} \delta_{ks_t}\Big(\ln(2\pi\beta\lambda_k) + (\beta\lambda_k)^{-1}(z_{tm} - (1-\tau)y_{tm} - \tau y_{t+1,m})^2\Big), \tag{S16}$$

$$\ln p(x|z) = -\frac{1}{2}\sum_{t=1}^{T} o_t \sum_{m=1}^{\dim}\Big(\ln(2\pi v_{tm}) + v_{tm}^{-1}(x_{tm} - z_{tm})^2\Big). \tag{S17}$$

S2. INFERENCE, MODEL SELECTION, AND VARIATIONAL APPROXIMATIONS

S2.1. Maximum evidence and Variational Bayes inference

In a pure Bayesian treatment [4], we add prior distributions $p(\theta|M) \equiv p_0(\theta)$ on the parameter values (conditional on model structure, M), and compute the evidence $p(x|M)$ for different models by integrating out both parameters and unobserved (nuisance) variables in the complete data likelihood,

$$p(x|M) = \int dz\, dy\, d\theta \sum_s p(x, z, y, s|\theta) p_0(\theta). \tag{S18}$$

This marginalization is intractable for our model, but we can make a mean-field approximation [4], and write

$$\ln p(x|M) \geq \ln L = \int dz\, dy\, d\theta \sum_s q(s)q(y,z)q(\theta) \ln \frac{p(x,z,y,s|\theta)p_0(\theta)}{q(s)q(y,z)q(\theta)}, \tag{S19}$$

where the inequality follows from Jensen's inequality, and $q(s)$, $q(y,z)$, $q(\theta)$ are arbitrary variational distributions, that we will optimize to maximize the lower bound $\ln L$, which is the mean-field approximation of $\ln L$. Using functional differentiation, and enforcing a normalization constraints of the variational distributions while optimizing $\ln L$, we get the variational equations

$$\ln q(\theta) = -\ln Z_\theta + \ln p_0(\theta) + \langle \ln p(s|\theta) \rangle_{q(s)} + \langle \ln p(y|s,\theta) \rangle_{q(s)q(y,z)} + \langle \ln p(z|s,y,\theta) \rangle_{q(s)q(y,z)}\,, \tag{S20}$$

$$\ln q(s) = -\ln Z_s + \langle \ln p(s|\theta) \rangle_{q(\theta)} + \langle \ln p(y|s,\theta) \rangle_{q(\theta)q(y,z)} + \langle \ln p(z|y,s,\theta) \rangle_{q(\theta),q(y,z)}\,, \tag{S21}$$

$$\ln q(y,z) = -\ln Z_{yz} + \langle \ln p(y|s,\theta) \rangle_{q(s)q(\theta)} + \langle \ln p(z|y,s,\theta) \rangle_{q(s)q(\theta)}\,, + \ln p(x|z), \tag{S22}$$

which we solve iteratively. From the likelihood terms one can see that a suitably factorized prior will lead to factorization of the entire variational parameter distributions. We will assume such a prior structure, and furthermore choose the functional forms of conjugate priors [4, 5] for computational tractability. We will also sometimes drop subscripts from the average brackets $\langle \cdot \rangle$ in unambiguous cases.

To actually compute the lower bound, we substitute the results of the variational updates back into $\ln L$. The result is especially convenient just after updating $q(s)$, in which case a lot of terms cancel, and we end up with

$$\ln L = \ln Z_s + \left\langle \ln \frac{p(x|z)}{q(y,z)} \right\rangle_{q(y,z)} + \left\langle \ln \frac{p_0(\theta)}{q(\theta)} \right\rangle_{q(\theta)}. \tag{S23}$$

The model selection criterion is then to prefer the model with the highest likelihood (or lower bound), or interpret $e^{\ln L}$ or e^F as proportional to the posterior probability that the model is true. The variational distributions can also be used for (approximate) inference about the parameters and hidden states and paths.

S2.2. Maximum aposteriori estimates (MAP)

Instead of integrating over the parameters, maximum aposteriori inference simply seek parameter values that maximize the likelihood,

$$\ln L = \max_\theta \ln \int dz\, dy \sum_s p(x,z,y,s|\theta) p_0(\theta). \tag{S24}$$

To derive approximate maximum likelihood inference, we make a variational ansatz with only $q(s)$ and $q(y,z)$, and write

$$\ln L = \max_\theta \int dz\, dy \sum_s q(s)q(y,z) \ln \frac{p(x,z,y,s|\theta)p_0(\theta)}{q(s)q(y,z)}, \tag{S25}$$

and maximize w.r.t. $q(s)$, $q(y,z)$, and θ, to get

$$\theta_{MAP} = \text{argmax}_\theta \Big[\ln p_0(\theta) + \langle \ln p(s|\theta) \rangle_{q(s)} + \langle \ln p(y|s,\theta) \rangle_{q(s)q(y,z)} + \langle \ln p(z|s,y,\theta) \rangle_{q(s)q(y,z)} \Big], \tag{S26}$$

$$\ln q(s) = -\ln Z_s + \ln p(s|\theta) + \langle \ln p(y|s,\theta) \rangle_{q(y,z)} + \langle \ln p(z|y,s,\theta) \rangle_{q(y,z)}\,, \tag{S27}$$

$$\ln q(y,z) = -\ln Z_{yz} + \langle \ln p(y|s,\theta) \rangle_{q(s)q(\theta)} + \langle \ln p(z|y,s,\theta) \rangle_{q(s)q(\theta)}\,, + \ln p(x|z). \tag{S28}$$

After updating $q(s)$, the lower bound is given by

$$\ln L = \ln Z_s + \left\langle \ln \frac{p(x|z)}{q(y,z)} \right\rangle_{q(y,z)} + \ln p_0(\theta_{MAP}). \tag{S29}$$

S2.3. Maximum likelihood estimates (MLE)

The Maximum likelihood estimate (MLE) is like MAP, but with priors removed. In practice, we will use MLE rather than MAP, but note that MAP inference might offer a way to numerically stabilize MLE in a principled way.

S2.4. Cross-validation with point-estimates

An alternative to using the Bayesian maximum evidence criterion for model selection, is to estimate the predictive performance of a model. This means that we imagine ensembles of training and validation data sets X_T and X_V, and seek to maximize

$$\ln P = \langle \ln p(X_V|X_T, M) \rangle_{X_V, X_T}, \tag{S30}$$

where the expectation is to be computed with respect to the true distribution of training and validation data sets. Training and validation data is assumed to be identically distributed (except for possibly being of different size), and their true distribution are not necessarily known, or part of the set of candidate models. The conditional dependence in Eq. (S30) should be interpreted as learning the model (integrating/maximizing parameters, and integrating out nuisance variables y, z, s) based on the training data.

In practice, we do not have an infinite amount of validation and training data, and instead divide out existing data sets into K different partitions $\{(x_V^{(1)}, x_T^{(1)}), (x_V^{(2)}, x_T^{(2)}), \dots, (x_V^{(K)}, x_T^{(K)})\}$, and approximate the predictive performance by a an average,

$$\langle \ln p(X_V|X_T, M) \rangle_{X_V, X_T} \approx \frac{1}{K} \sum_{j=1}^{K} \ln p(x_V^{(j)}|x_T^{(j)}, M) \tag{S31}$$

In practice, for each partition we learn (using MLE or MAP) a set of parameters $\theta_T^{(j)} = \theta(x_T^{(j)})$, and use

$$\ln p(x_V^{(j)}|x_T^{(j)}, M) = \ln p(x_V^{(j)}|\theta^{(j)}, M) = \ln \int dy_V \, dz_V \sum_{s_V} p(x_V^{(j)}, z_V^{(j)}, y_V^{(j)}, s_V^{(j)}|\theta_T^{(j)}, M) \tag{S32}$$

and do the marginalizations of y_V, z_V, s_V using the same variational approximation as the inference, but exclude explicit contributions from the prior to the predictive likelihood.

There is some freedom in choosing the size of the training and validation data sets, and to be able to compare different choices, some normalization may be in order. The question is further complicated by the fact that the statistically independent atoms of single particle tracking is (to good approximation) single trajectories, not single coordinate observations, and single trajectories vary in length.

LaMont and Wiggins [6] suggested normalizing data sets of independent observations to the size of the full data set. In our setting, this probably means rescaling the predictive performance of each validation set to the total data set size,

$$\hat{H}_{CV}(M) = \frac{1}{K} \sum_{j=1}^{K} \frac{N_V^{(j)} + N_T^{(j)}}{N_V^{(j)}} \ln p(x_V^{(j)}|\theta_T^{(j)}, M) = \frac{N}{K} \sum_{j=1}^{K} \frac{1}{N_V^{(j)}} \ln p(x_V^{(j)}|\theta_T^{(j)}, M), \tag{S33}$$

where $N_{\dots}$ means the number of coordinates (including missing positions) in the training/validation data sets, and $N = N_T + N_V$ if we always partition the data so that each data point is used exactly once. Equivalently, we could rescale to compute the predictive performance per observed position.

S2.5. Pseudo-Bayes factors

A Bayesian version of cross-validation is to include marginalization over parameters as well in the predictive performance [7]. In particular, we use the parameter posterior from the training set as a prior in evaluating the performance on the validation set. For brevity, we use the more compact notation

$$S_T = (z_T, y_T, s_T), \qquad S_V = (z_V, y_V, s_V), \qquad S = (z, y, s), \tag{S34}$$

$$\int dS_T = \int dy_T\, dz_T \sum_{s_T}, \qquad \int dS_V = \int dy_V\, dz_V \sum_{s_V}, \qquad \int dS = \int dy\, dz \sum_{s}, \tag{S35}$$

and the pseudo-Bayes factor for a single training-validation partition is then given by

$$\begin{aligned}\ln P_{PBF}(x_V^{(j)}, x_T^{(j)}) &= \ln \frac{\int dS_V^{(j)} dS_T^{(j)} d\theta p(x_V^{(j)}, S_V^{(j)}|\theta) p(x_T^{(j)}, S_T^{(j)}|\theta) p_0(\theta)}{\int dS_T^{(j)} d\theta' p(x_T^{(j)}, S_T^{(j)}|\theta') p_0(\theta')}\\ &= \ln \int dS_V^{(j)} dS_T^{(j)} d\theta p(x_V^{(j)}, S_V^{(j)}|\theta) p(x_T^{(j)}, S_T^{(j)}|\theta) p_0(\theta) - \ln \int dS_T^{(j)} d\theta p(x_T^{(j)}, S_T^{(j)}|\theta) p_0(\theta)\\ &= \ln p(x_V^{(j)}, x_T^{(j)}|M) - \ln p(x_T^{(j)}|M) = \ln p(x|M) - \ln p(x_T^{(j)}|M) \approx \ln L[x] - \ln L[x_T^{(j)}].\end{aligned} \tag{S36}$$

that is, the difference log evidence between the total and training data. On the other hand, one could also approximate the training posterior by the variational parameter distributions from the training set,

$$\frac{\int dS_T^{(j)} p(x_T^{(j)}, S_T^{(j)}|\theta) p_0(\theta)}{\int dS_T^{(j)} d\theta' p(x_T^{(j)}, S_T^{(j)}|\theta) p_0(\theta)} \approx q(\theta; x_T^{(j)}) \equiv q_T^{(j)}(\theta), \tag{S37}$$

which means that this approximate posterior is used as the prior for evaluating the validation set,

$$\ln P_{PBF}(x_V^{(j)}, x_T^{(j)}) \approx \ln L[x_V^{(j)}; q_T^{(j)}(\theta)]. \tag{S38}$$

It is not quite clear which of these approximations is preferred theoretically. The lower bound difference of Eq. (S36) seems more systematic in that it only approximates integrals with no reference to additional probabilistic interpretations. On the other hand, the variational posterior, Eq. (S37), represents the approximate the Bayesian predictive distribution available for practical use [4, 5], and hence is arguably a reasonable choice for assessing predictive performance. Numerically, using the variational parameter posterior avoids potential cancellation problems inherent in computing small differences between large numbers. Since the validation data sets will in general be smaller than the training sets, the extra computational cost of converging both training and validation sets (as opposed to only the training set, since the full data set is already converged) should be negligible. As for cross-validation, we compute an average over many partitions,

$$\hat{H}_{PBF}(M) \approx \frac{N}{K} \sum_{j=1}^{K} \frac{1}{N_V} \ln P_{PBF}(x_V^{(j)}, x_T^{(j)}) \tag{S39}$$

and also normalize by validation set size in case of varying trajectory lengths. In order to generate the predictive/AIC limit asymptotically, the validation sets should be chosen to be small [6, 7]. On the other hand, small validation data sets means higher statistical errors in $\hat{H}_{PBF}$.

S3. VARIATIONAL BAYES ALGORITHM

S3.1. Inital state and transition probabilities

The variational equations for the parameters governing the hidden states are

$$\ln q(\pi_m) = -\ln Z_\pi + \ln p_0(\pi_m) + \langle \delta_{m,s_1} \rangle \ln \pi_m, \tag{S40}$$

$$\ln q(a_k) = -\ln Z_a + \ln p_0(a_k) + \ln(1 - a_k) \sum_{t=1}^{T-1} \left\langle \delta_{k,s_t} \delta_{k,s_{t+1}} \right\rangle + \ln a_k \sum_{t=1}^{T-1} \left(1 - \left\langle \delta_{k,s_t} \delta_{k,s_{t+1}} \right\rangle \right), \tag{S41}$$

$$\ln q(B_{kj}) = -\ln Z_{B,k} + \ln B_{kj} \sum_{t=1}^{T-1} \left\langle \delta_{k,s_t} \delta_{j,s_{t+1}} \right\rangle, \quad k \neq j. \tag{S42}$$

Except for the summation bounds on t, this is the same as in vbSPT [1, software documentation], and all relevant statistics are given in the count matrix

$$\hat{w}_{ij} = \sum_{t=1}^{T-1} \left\langle \delta_{k,s_t} \delta_{j,s_{t+1}} \right\rangle \tag{S43}$$

and expected occupancy $\langle \delta_{i,s_t} \rangle$. Using conjugate priors, π and the rows (minus diagonal elements) of B get Dirichlet distributions, while each a_k is beta distributed (the 2-component Dirichlet).

$$q(\pi) = \text{Dir}(\pi | w^{(\pi)}), \quad w_j^{(\pi)} = \tilde{w}_j^{(\pi)} + \langle \delta_{j,s_1} \rangle \,, \tag{S44}$$

$$q(B) = \prod_{j=1}^{N} \text{Dir}(B_{j,:} | w_{j,:}^{(B)}), \quad w_{jk}^{(B)} = \tilde{w}_{jk}^{(B)} + \hat{w}_{ij}, \; (k \neq j), \tag{S45}$$

$$\begin{aligned} q(a) = \prod_{j=1}^{N} \beta(a_j | w_{j1}^{(a)}, w_{j2}^{(a)}), \quad w_{j1}^{(a)} &= \tilde{w}_{j1}^{(a)} + \sum_{t=1}^{T-1} \left\langle \delta_{j,s_t} (1 - \delta_{j,s_{t+1}}) \right\rangle = \tilde{w}_{j1}^{(a)} + \sum_{k \neq j} \hat{w}_{jk}, \\ w_{j2}^{(a)} &= \tilde{w}_{j2}^{(a)} + \sum_{t=1}^{T-1} \left\langle \delta_{j,s_t} \delta_{j,s_{t+1}} \right\rangle \qquad = \tilde{w}_{j2}^{(a)} + \hat{w}_{jj}, \end{aligned} \tag{S46}$$

with where $\tilde{w}_j^{(\pi)}$, $\tilde{w}_{jk}^{(B)}$, and $\tilde{w}_{jk}^{(a)}$ are pseudo-counts in the prior distributions. The total number of pseudo-counts (for each distribution) is called the prior strength and denoted $w_0^{(\cdot)} = \sum_k w_k^{(\cdot)}$.

The Dirichlet density function, in this case for a vector $\mathbf{x}$, is

$$\text{Dir}(\mathbf{x}|\mathbf{w}) = \frac{1}{B(\mathbf{w})} = \prod_j x_j^{(w_j - 1)}, \quad B(\mathbf{w}) = \frac{\prod_j \Gamma(w_j)}{\Gamma(w_0)}, \quad w_0 = \sum_k w_k \tag{S47}$$

with the constraints $0 \leq x_j \leq 1$ and $\sum_j x_j = 1$. The beta distribution is the special case of two components (x and $1 - x$),

$$\beta(x|u, v) = \frac{\Gamma(u + v)}{\Gamma(u)\Gamma(v)} x^{u-1} (1 - x)^{v-1}. \tag{S48}$$

The following average and mode values will be needed:

$$\langle \ln \pi_i \rangle = \psi(w_i^{(\pi)}) - \psi(w_0^{(\pi)}), \qquad w_0^{(\pi)} = \sum_{i=1}^{N} w_i^{(\pi)}, \tag{S49}$$

$$\langle \ln a_j \rangle = \psi(w_{j1}^{(a)}) - \psi(w_{j0}^{(a)}), \qquad w_{j0}^{(a)} = w_{j1}^{(a)} + w_{j2}^{(a)}, \tag{S50}$$

$$\langle \ln(1 - a_j) \rangle = \psi(w_{j2}^{(a)}) - \psi(w_{j0}^{(a)}), \tag{S51}$$

$$\langle \ln B_{jk} \rangle = \psi(w_{jk}^{(B)}) - \psi(w_{j0}^{(B)}), \qquad w_{j0}^{(B)} = \sum_{k=1, k \neq j}^{N} w_{jk}^{(B)}, \tag{S52}$$

Some additional variational mode*, $\langle$mean$\rangle$ and variances are

$$\begin{aligned}
\pi_i^* =& \frac{w_i^{(\pi)}-1}{w_0^{(\pi)}-N}, & \langle\pi_i\rangle =& \frac{w_i^{(\pi)}}{w_0^{(\pi)}}, & \mathrm{Var}[\pi_i] =& \frac{w_i^{(\pi)}(w_0^{(\pi)}-w_i^{(\pi)})}{(w_0^{(\pi)})^2(w_0^{(\pi)}+1)}, && \text{(S53)}\\
a_i^* =& \frac{w_{i1}^{(a)}-1}{w_{i0}^{(a)}-2}, & \langle a_i\rangle =& \frac{w_{i1}^{(a)}}{w_{i0}^{(a)}}, & \mathrm{Var}[a_j] =& \frac{w_{j1}^{(a)}w_{j2}^{(a)}}{(w_{j0}^{(a)})^2(1+w_{j1}^{(a)})}, && \text{(S54)}\\
(1-a_i)^* =& \frac{w_{i2}^{(a)}-1}{w_{i0}^{(a)}-2}, & \langle 1-a_i\rangle =& \frac{w_{i2}^{(a)}}{w_{i0}^{(a)}}, & \mathrm{Var}[1-a_j] =& \mathrm{Var}[a_j] && \text{(S55)}\\
B_{jk}^* =& \frac{w_{jk}^{(B)}-1}{w_{j0}^{(B)}-N+1} & \langle B_{jk}\rangle =& \frac{w_{jk}^{(B)}}{w_{j0}^{(B)}}, & \mathrm{Var}[B_{jk}] =& \frac{w_{jk}^{(B)}(w_{j0}^{(B)}-w_{jk}^{(B)})}{(w_{j0}^{(B)})^2(1+w_{j0}^{(B)})}, && \text{(S56)}\\
A_{jj}^* =& ???, & \langle A_{jj}\rangle =& \langle 1-a_j\rangle\,, & & && \text{(S57)}\\
A_{jk}^* =& ???, & \langle A_{jk}\rangle =& \langle a_i\rangle\,\langle B_{jk}\rangle\,. & & && \text{(S58)}
\end{aligned}$$

The variational distributions of a, B induces a joint distribution on A, which can be written (for row i), as

$$q(A_{i,:})dA_{i,:} = q_a(a_i(A_{i,:}))q_B(B_{i,:}(A_{i,:}))\left|\frac{\partial(a_i, B_{i,:})}{\partial A_{i,:}}\right| dA_{i,:}. \tag{S59}$$

This is difficult to do analytically, and we leave the posterior mode of the transition matrix unknown.

S3.2. Dwell times

Mean dwell times (in units of Δt) is $\tau_j = a_j^{-1}$. This gives the variational density function

$$q(\tau_j) = q(a_j(\tau_j))\left|\frac{da_j}{d\tau_j}\right| = \frac{\Gamma(w_{j1}^{(\mathbf{a})})\Gamma(w_{j2}^{(\mathbf{a})})}{\Gamma(w_{j0}^{(\mathbf{a})})}\tau_j^{-w_{j0}^{(\mathbf{a})}}(\tau_j-1)^{w_{j2}^{(\mathbf{a})}-1}, \quad \tau_j \geq 1, \tag{S60}$$

which means that

$$\begin{aligned}
\langle\tau_j\rangle =& \left\langle a_j^{-1}\right\rangle = \frac{w_{j0}^{(\mathbf{a})}}{w_{j1}^{(\mathbf{a})}} = \frac{1}{\langle a_j\rangle}, && \text{(S61)}\\
\tau_j^* =& \frac{w_{j0}^{(\mathbf{a})}}{1+w_{j1}^{(\mathbf{a})}}, && \text{(S62)}\\
\left\langle\tau_j^2\right\rangle =& \left\langle a_j^{-2}\right\rangle_{q(\mathbf{a})} = \langle\tau_j\rangle\,\frac{w_{j0}^{(\mathbf{a})}-1}{w_{j1}^{(\mathbf{a})}-1} = \langle\tau_j\rangle^2\,\frac{w_{j0}^{(\mathbf{a})}-1}{w_{j0}^{(\mathbf{a})}-\langle\tau_j\rangle}, && \text{(S63)}\\
\mathrm{Var}(\tau_j) =& \left\langle\tau_j^2\right\rangle - \langle\tau_j\rangle^2 = \frac{\langle\tau_j\rangle^2\left(\langle\tau_j\rangle-1\right)}{w_{j0}^{(\mathbf{a})}-\langle\tau_j\rangle}. && \text{(S64)}
\end{aligned}$$

or

$$\begin{aligned}
w_{j1}^{(\mathbf{a})} =& \frac{w_{j0}^{(\mathbf{a})}}{\langle\tau_j\rangle} = 1 + \frac{\langle\tau_j\rangle\left(\langle\tau_j\rangle-1\right)}{\mathrm{Var}(\tau_j)},\\
w_{j2}^{(\mathbf{a})} =& w_{j0}^{(\mathbf{a})}\frac{\langle\tau_j\rangle-1}{\langle\tau_j\rangle} = (\langle\tau_j\rangle-1)w_{j1}^{(\mathbf{a})}. && \text{(S65)}
\end{aligned}$$

S3.3. Step variance

$$\begin{aligned}\ln q(\lambda_k) &= -\ln Z_\lambda + \ln p_0(\lambda_k)\\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\dim}\langle\delta_{ks_t}\rangle\left(2\ln\lambda_k+\lambda_k^{-1}\Big[\left\langle(y_{t+1,m}-y_{tm})^2\right\rangle+\beta^{-1}\left\langle(z_{tm}-(1-\tau)y_{tm}-\tau y_{t+1,m})^2\right\rangle\Big]\right).\end{aligned} \tag{S66}$$

If the prior is inverse gamma, then so is $q(\lambda_k)$:

$$q(\lambda_k)=\frac{c_k^{n_k}}{\Gamma(n_k)}\lambda_k^{-n_k-1}e^{-c_k/\lambda_k}. \tag{S67}$$

with

$$n_k=\tilde{n}_k+\hat{n}_k,\quad c_k=\tilde{c}_k+\hat{c}_k. \tag{S68}$$

Here, $\tilde{n}_k$, $\tilde{c}_k$ are prior parameters, and the data-dependent terms are given by

$$\hat{n}_k=\dim\times\sum_{t=1}^{T}\langle\delta_{ks_t}\rangle\,, \tag{S69}$$

and

$$\begin{aligned}\hat{c}_k=\frac{1}{2}\sum_{t=1}^{T}\langle\delta_{ks_t}\rangle\sum_{m=1}^{\dim}\Bigg\{&\Big(\langle y_{t+1,m}\rangle-\langle y_{tm}\rangle\Big)^2+\beta^{-1}\Big(\langle z_{tm}\rangle-(1-\tau)\langle y_{tm}\rangle-\tau\langle y_{t+1,m}\rangle\Big)^2\\ &+\big(1+\beta^{-1}(1-\tau)^2\big)\Sigma_{y_{tm},y_{tm}}+\big(1+\beta^{-1}\tau^2\big)\Sigma_{y_{t+1,m},y_{t+1,m}}+\beta^{-1}\Sigma_{z_{tm},z_{tm}}\\ &\underbrace{-2\big(1-\beta^{-1}\tau(1-\tau)\big)}_{=2R/\beta}\Sigma_{y_{tm},y_{t+1,m}}-2\beta^{-1}(1-\tau)\Sigma_{z_{tm},y_{tm}}-2\beta^{-1}\tau\Sigma_{z_{tm},y_{t+1,m}}\Bigg\},\end{aligned} \tag{S70}$$

where Σ are joint covariance matrices of $q(y_{:,m},z_{:,m})$. We need the following averages (dropping the subscript):

$$\langle\lambda\rangle=\frac{c}{n-1}, \tag{S71}$$

$$\mathrm{std}[\lambda]=\frac{c}{(n-1)\sqrt{n-2}}, \tag{S72}$$

$$\left\langle\lambda^{-1}\right\rangle=\frac{n}{c}, \tag{S73}$$

$$\langle\ln\lambda\rangle=\ln c-\psi(n), \tag{S74}$$

$$\lambda^*=\frac{c}{n+1}, \tag{S75}$$

Here, ψ is the digamma function, and the asterix * denotes the mode (most likely value). Since $\lambda=2D\Delta t$, the variational distribution for the diffusion constant is

$$q(D_k)=\frac{(c_k/2\Delta t)^{n_k}}{\Gamma(n_k)}D_k^{-n_k-1}e^{-(c_k/2\Delta t)/D_k}, \tag{S76}$$

i.e., D is also inverse gamma, with

$$c_k^{(D)}=c_k^{(\lambda)}/2\Delta t,\quad n_k^{(D)}=n_k^{(\lambda)}. \tag{S77}$$

S3.4. Hidden states

$$\ln q(s) = -\ln Z_s + \sum_{i=1}^{N} \delta_{i,s_1} \left\langle \ln \pi_i \right\rangle \tag{S78}$$

$$+\sum_{t=1}^{T-1} \sum_{k,j=1}^{N} \delta_{k,s_t}\delta_{j,s_{t+1}} \Big[\delta_{kj} \left\langle \ln(1-a_k) \right\rangle + (1-\delta_{kj}) \left\langle \ln a_k \right\rangle + (1-\delta_{kj}) \left\langle \ln B_{kj} \right\rangle \Big] \tag{S79}$$

$$-\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{m=1}^{\dim} \delta_{is_t}\Big[\left\langle \ln(\lambda_i) \right\rangle + \left\langle \lambda_i^{-1} \right\rangle \left\langle (y_{tm} - y_{t+1,m})^2 \right\rangle \Big] \tag{S80}$$

$$-\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{m=1}^{\dim} \delta_{is_t}\Big[\left\langle \ln(\lambda_i) \right\rangle + \beta^{-1} \left\langle \lambda_i^{-1} \right\rangle \left\langle (z_{tm} - (1-\tau)y_{tm} - \tau y_{t+1,m})^2 \right\rangle \Big] \tag{S81}$$

$$= -\ln Z_s + \sum_{t=1}^{T-1} \sum_{k,j=1}^{N} \delta_{k,s_t}\delta_{j,s_{t+1}} \ln Q_{jk} + \sum_{t=1}^{T}\sum_{i=1}^{N} \delta_{is_t} \ln H_{ti}, \tag{S82}$$

with

$$\ln Q_{kj} = \delta_{kj} \left\langle \ln(1-a_k) \right\rangle + (1-\delta_{kj}) \Big[\left\langle \ln a_k \right\rangle + \left\langle \ln B_{kj} \right\rangle \Big], \tag{S83}$$

$$\ln H_{ti} = \delta_{1t} \left\langle \ln \pi_i \right\rangle - \dim \times \left\langle \ln(\lambda_i) \right\rangle \tag{S84}$$

$$-\frac{1}{2}\left\langle \lambda_i^{-1} \right\rangle \sum_{m=1}^{\dim} \Bigg\{ \Big(\left\langle y_{t+1,m} \right\rangle - \left\langle y_{tm} \right\rangle \Big)^2 + \beta^{-1}\Big(\left\langle z_{tm} \right\rangle - (1-\tau) \left\langle y_{tm} \right\rangle - \tau \left\langle y_{t+1,m} \right\rangle \Big)^2 \tag{S85}$$

$$+\Big(1 + \frac{(1-\tau)^2}{\beta}\Big)\Sigma_{y_{tm}y_{tm}} + \Big(1 + \frac{\tau^2}{\beta}\Big)\Sigma_{y_{t+1,m}y_{t+1,m}} + \frac{1}{\beta}\Sigma_{z_{tm}z_{tm}} \tag{S86}$$

$$\underbrace{-2\Big(1 - \frac{\tau(1-\tau)}{\beta}\Big)}_{=2R/\beta}\Sigma_{y_{tm}y_{t+1,m}} - 2\frac{1-\tau}{\beta}\Sigma_{y_{tm}z_{tm}} - \frac{2\tau}{\beta}\Sigma_{y_{t+1,m}z_{tm}} \Bigg\}. \tag{S87}$$

Averages $\left\langle \delta_{j,s_t} \right\rangle$, $\sum_{t=1}^{T-1} \left\langle \delta_{k,s_t}\delta_{j,s_{t+1}} \right\rangle$, and the normalization constant $\ln Z_s$ are computed with the standard forward-backward algorithm.

S3.5. Hidden trajectories

$$\ln q(y,z) = const. - \frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\dim} o_t \frac{(x_{tm} - z_{tm})^2}{v_{tm}} - \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{m=1}^{\dim} \left\langle \delta_{is_t} \right\rangle \left\langle \lambda_i^{-1} \right\rangle (y_{t+1,m} - y_{tm})^2$$

$$- \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{m=1}^{\dim} \left\langle \delta_{is_t} \right\rangle \left\langle (\beta\lambda_i)^{-1} \right\rangle (z_{tm} - (1-\tau)y_{tm} - \tau y_{t+1,m})^2 \tag{S88}$$

This is a product of dim multivariate normals. We introduce an effective step variance defined by

$$\frac{1}{\alpha_t} = \sum_{i=1}^{N} \left\langle \delta_{is_t} \right\rangle \left\langle \lambda_i^{-1} \right\rangle, \tag{S89}$$

and expand:

$$\begin{aligned}\ln q(y,z) &= const. - \frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\dim}\Bigg\{\frac{o_t}{v_{tm}}x_{tm}^2 + \frac{o_t}{v_{tm}}z_{tm}^2 - 2\frac{o_t}{v_{tm}}x_{tm}z_{tm} + \frac{y_{tm}^2}{\alpha_t} + \frac{y_{t+1,m}^2}{\alpha_t} - 2\frac{y_{tm}y_{t+1,m}}{\alpha_t} \\ &\quad + \frac{z_{tm}^2}{\beta\alpha_t} + \frac{(1-\tau)^2}{\beta\alpha_t}y_{tm}^2 + \frac{\tau^2}{\beta\alpha_t}y_{t+1,m}^2 - 2\frac{1-\tau}{\beta\alpha_t}y_{tm}z_{tm} - 2\frac{\tau}{\beta\alpha_t}y_{t+1,m}z_{tm} + 2\frac{\tau(1-\tau)}{\beta\alpha_t}y_{tm}y_{t+1,m}\Bigg\} \\ &= const. - \frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\dim}\Bigg\{\Big(1+\frac{(1-\tau)^2}{\beta}\Big)\frac{y_{tm}^2}{\alpha_t} + \Big(1+\frac{\tau^2}{\beta}\Big)\frac{y_{t+1,m}^2}{\alpha_t} + 2\Big(\underbrace{\frac{\tau(1-\tau)}{\beta} - 1}_{=R/\beta}\Big)\frac{y_{tm}y_{t+1,m}}{\alpha_t} \\ &\quad - 2\frac{1-\tau}{\beta\alpha_t}y_{tm}z_{tm} - 2\frac{\tau}{\beta\alpha_t}y_{t+1,m}z_{tm} + \Big(\frac{o_t}{v_{tm}} + \frac{1}{\beta\alpha_t}\Big)z_{tm}^2 - 2\frac{o_t}{v_{tm}}x_{tm}z_{tm} + \frac{o_t}{v_{tm}}x_{tm}^2\Bigg\}.\end{aligned} \tag{S90}$$

Written in matrix notation, with

$$z_m = \big[z_{1m}, z_{2m}, \dots, z_{Tm}\big]^\dagger, \quad y_m = \big[y_{1m}, y_{2m}, \dots, y_{Tm}, y_{T+1,m}\big]^\dagger, \tag{S91}$$

we have

$$\ln q(y,z) = const. - \frac{1}{2}\sum_{m=1}^{\dim}\Bigg\{\Big[y_m^\dagger,\ z_m^\dagger\Big]\begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{yz}^\dagger & A_{zzm}\end{bmatrix}\begin{bmatrix} y_m \\ z_m\end{bmatrix} - 2\Big[0,\ x_m^\dagger V_m\Big]\begin{bmatrix} y_m \\ z_m\end{bmatrix}\Bigg\}, \tag{S92}$$

with $A_{yy} \in \mathbf{R}^{(T+1)\times(T+1)}$, $A_{yz} \in \mathbf{R}^{(T+1)\times T}$, and $A_{zz} \in \mathbf{R}^{T\times T}$, given by

$$A_{yy} = \begin{bmatrix} \frac{1}{\alpha_1}\Big(1+\frac{(1-\tau)^2}{\beta}\Big) & \frac{R}{\beta\alpha_1} & 0 & \cdots & \\ \frac{R}{\beta\alpha_1} & \frac{1}{\alpha_2}\Big(1+\frac{(1-\tau)^2}{\beta}\Big) + \frac{1}{\alpha_1}\Big(1+\frac{\tau^2}{\beta}\Big) & \ddots & & \\ 0 & \ddots & \ddots & & \\ \vdots & & & & \\ & & & \frac{1}{\alpha_T}\Big(1+\frac{(1-\tau)^2}{\beta}\Big) + \frac{1}{\alpha_{T-1}}\Big(1+\frac{\tau^2}{\beta}\Big) & \frac{R}{\beta\alpha_T} \\ & & & \frac{R}{\beta\alpha_T} & \frac{1}{\alpha_T}\Big(1+\frac{\tau^2}{\beta}\Big) \end{bmatrix}, \tag{S93}$$

and

$$\begin{aligned} A_{yz} &= \frac{1}{\beta}\begin{bmatrix} \frac{1-\tau}{\alpha_1} & 0 & 0 & \cdots & \\ \frac{\tau}{\alpha 1} & \frac{1-\tau}{\alpha_2} & 0 & & \\ 0 & \frac{\tau}{\alpha 2} & \frac{1-\tau}{\alpha_3} & 0 & \\ 0 & 0 & \frac{\tau}{\alpha 3} & & \\ \vdots & & & \ddots & \\ & & & & \frac{1-\tau}{\alpha_T} \\ & & & & \frac{\tau}{\alpha_T} \end{bmatrix}, & \text{(S94)} \\ A_{zzm} &= \operatorname{diag}\Bigg(\Big[\frac{o_1}{v_{1m}} + \frac{1}{\beta\alpha_1}, \frac{o_2}{v_{2m}} + \frac{1}{\beta\alpha_2}, \dots, \frac{o_T}{v_{Tm}} + \frac{1}{\beta\alpha_T}\Big]\Bigg), & \text{(S95)} \\ V_m &= \operatorname{diag}\Bigg(\Big[\frac{o_1}{v_{1m}}, \frac{o_2}{v_{2m}}, \dots, \frac{o_T}{v_{Tm}}\Big]\Bigg). & \text{(S96)} \end{aligned}$$

Finally, we rewrite $\ln q(y,z)$ in the canonical form for multivariate normal distributions,

$$\begin{aligned} \ln q(y,z) &= -\ln Z_{yz} - \frac{1}{2}\sum_{m=1}^{\dim}\Big[y_m^\dagger - \langle y_m\rangle^\dagger\ ,\ z_m^\dagger - \langle z_m\rangle^\dagger\Big]\begin{bmatrix} \Sigma_{yym} & \Sigma_{yzm} \\ \Sigma_{yzm}^\dagger & \Sigma_{zzm}\end{bmatrix}^{-1}\begin{bmatrix} y_m - \langle y_m\rangle \\ z_m - \langle z_m\rangle\end{bmatrix} & \text{(S97)} \\ &= -\ln Z_{yz} - \frac{1}{2}\sum_{m=1}^{\dim}\Big[y_m^\dagger - \langle y_m\rangle^\dagger\ ,\ z_m^\dagger - \langle z_m\rangle^\dagger\Big]\begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{yz}^\dagger & A_{zzm}\end{bmatrix}\begin{bmatrix} y_m - \langle y_m\rangle \\ z_m - \langle z_m\rangle\end{bmatrix}, & \text{(S98)} \end{aligned}$$

where the covariance matrix is given by

$$\begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{zy} & A_{zzm} \end{bmatrix} \begin{bmatrix} \Sigma_{yym} & \Sigma_{yzm} \\ \Sigma_{zym} & \Sigma_{zzm} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{S99}$$

Expanding Eq. (S98), we get

$$\ln q(yz) = const. - \frac{1}{2}\sum_{m=1}^{\dim} \Bigg\{ \Big[y_m^\dagger,\ z_m^\dagger \Big] \begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{yz}^\dagger & A_{zzm} \end{bmatrix} \begin{bmatrix} y_m \\ z_m \end{bmatrix} - 2\Big[\langle y_m \rangle^\dagger, \langle z_m \rangle^\dagger \Big] \begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{yz}^\dagger & A_{zzm} \end{bmatrix} \begin{bmatrix} y_m \\ z_m \end{bmatrix}, \tag{S100}$$

and comparing the linear term with that of Eq. (S92), we see that the expectation values are given by

$$\begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{zy} & A_{zzm} \end{bmatrix} \begin{bmatrix} \langle y_m \rangle \\ \langle z_m \rangle \end{bmatrix} = \begin{bmatrix} 0 \\ V_m x_m \end{bmatrix}. \tag{S101}$$

The A-matrix blocks have some nice sparsity and symmetry properties: diagonal (A_{zzm}), symmetric tri-diagonal (A_{yy}), or asymmetric and non-square bi-diagonal (A_{yz}). We also note that A_{zz} is guaranteed to be invertible, since it is diagonal and $\beta\lambda_t > 0$ for all t. Furthermore, since $q(y, z)$ is multivariate normal, the marginal distributions for y and z are too. This means that both Σ_{yym} and Σ_{zzm} are invertible, since otherwise the marginals would not be properly defined.

Algorithmically, we can avoid full matrix inversions, since only a special subset of covariances are needed, namely

$$\mathrm{Var}(y_{tm}) = \Sigma_{y_{tm}y_{tm}}, \quad \Sigma_{y_{tm}y_{t+1,m}}, \quad \Sigma_{y_{tm}z_{tm}}, \quad \mathrm{Var}(z_{tm}) = \Sigma_{z_{tm}z_{tm}}, \quad \Sigma_{y_{t+1m}z_{tm}}. \tag{S102}$$

We also need the determinant of the full covariance matrix for the lower bound, but as we show below, this can be reduced to computing $|A_{zzm}|$ (easy, since A_{zzm} is diagonal), and $|\Sigma_{yym}^{-1}|$ (also fairly easy, since this matrix is symmetric and tridiagonal).

In all, this suggests that further analytical calculations would be valuable.

a. Covariance matrices Manipulations of the inversion equation (S99) leads to

$$\begin{aligned}
\Sigma_{yym}^{-1} &= \underbrace{(A_{yy} - A_{yz}A_{zzm}^{-1}A_{zy})}_{\text{symmetric tridiagonal}}, && \text{(S103)}\\
\Sigma_{yzm} &= (A_{yy} - A_{yz}A_{zzm}^{-1}A_{zy})^{-1}A_{yz}A_{zzm}^{-1} = \Sigma_{yym}A_{yz}A_{zzm}^{-1}, && \text{(S104)}\\
\Sigma_{zym} &= \Sigma_{yzm}^\dagger, && \text{(S105)}\\
\Sigma_{zzm} &= A_{zzm}^{-1}(I + A_{zy}\Sigma_{yzm}) \qquad = A_{zzm}^{-1} + A_{zzm}^{-1}A_{zy}\Sigma_{yym}A_{yz}A_{zzm}^{-1}. && \text{(S106)}
\end{aligned}$$

To figure out which elements are needed, we write out some matrixx elements explicitly:

$$\begin{aligned}
\Sigma_{y_{tm},y_{tm}} &= \big(\Sigma_{yym}\big)_{t,t} \quad , && \text{(S107)}\\
\Sigma_{y_{tm},y_{t+1,m}} &= \big(\Sigma_{yym}\big)_{t,t+1}, && \text{(S108)}\\
\Sigma_{y_{tm},z_{tm}} &= \big(\Sigma_{yzm}\big)_{t,t} \quad =\ldots=\Big(\quad (1-\tau)\frac{\Sigma_{y_{tm},y_{tm}}}{\beta\alpha_t} + \tau\frac{\Sigma_{y_{tm},y_{t+1,m}}}{\beta\alpha_t}\Big)\big(A_{zzm}^{-1}\big)_{t,t} \quad , && \text{(S109)}\\
\Sigma_{y_{t+1,m},z_{tm}} &= \big(\Sigma_{yzm}\big)_{t+1,t} =\ldots=\Big(\quad (1-\tau)\frac{\Sigma_{y_{tm},y_{t+1,m}}}{\beta\alpha_t} + \tau\frac{\Sigma_{y_{t+1,m},y_{t+1,m}}}{\beta\alpha_t}\Big)\big(A_{zzm}^{-1}\big)_{t,t}, && \text{(S110)}\\
\Sigma_{z_{tm},z_{tm}} &= \big(\Sigma_{zzm}\big)_{t,t} \quad =\ldots=\Big(1 + (1-\tau)\frac{\Sigma_{y_{tm},z_{tm}}}{\beta\alpha_t} + \tau\frac{\Sigma_{y_{t+1,m},z_{tm}}}{\beta\alpha_t}\Big)\big(A_{zzm}^{-1}\big)_{t,t} \quad . && \text{(S111)}
\end{aligned}$$

From this, we see that we only need the diagonal and first off-diagonal of Σ_{yym}.

It is indeed possible to invert symmetric positive definite tridiagonal matrices from the main diagonal and outwards, for example as described in Ref. [8], which means that this partial inversion can be done in linear time. We use recursion relations from Ref. [8], rewritten so as to minimize the risk of numerical over- or underflow. These relations use a Cholesky factorization as an intermediate step, which is useful for solving triangular systems of equations. They also yield the determinant $|\Sigma_{yym}^{-1}|$.

b. Mean values To compute the mean values $\langle z_m\rangle$, $\langle y_m\rangle$ efficiently, we manipulate Eq. (S101) to get

$$\underbrace{\big(A_{yy} - A_{yz}A_{zzm}^{-1}A_{yz}^{\dagger}\big)}_{=\,\Sigma_{yym}^{-1},\text{ symmetric tridiagonal}}\quad \langle y_m\rangle = A_{yz}A_{zzm}^{-1}V_m x_m, \tag{S112}$$

$$\langle z_m\rangle = A_{zzm}^{-1}\big(V_m x_m + A_{yz}^{\dagger}\langle y_m\rangle\big). \tag{S113}$$

Thus, computing mean values requires inverting a diagonal matrix (A_{zzm}) and solving a symmetric tri-diagonal linear system of equations. As mentioned above, one first step in this solution would be a cholesky factorization, which we get as a by-product of partially inverting Σ_{yym}. Σ_{yym}^{-1} is a symmetric tri-diagonal matrix with elements

$$\Sigma_{yym}^{-1} = A_{yy} - A_{yz}A_{zzm}^{-1}A_{yz}^{\dagger} = \begin{bmatrix} a_1 & c_1 & 0 & \cdots & & \\ c_1 & a_2+b_1 & c_2 & & & \\ 0 & c_2 & a_3+b_2 & \ddots & & \\ \vdots & & \ddots & \ddots & & \\ & & & & a_T+b_{T-1} & c_T \\ & & & & c_T & b_T \end{bmatrix}, \tag{S114}$$

with

$$a_{tm} = \frac{1}{\alpha_t}\Big(1 + \frac{(1-\tau)^2}{\beta}\Big) - \frac{(1-\tau)^2}{\beta^2\alpha_t^2}\Big(\frac{o_t}{v_{tm}} + \frac{1}{\beta\alpha_t}\Big)^{-1}, \tag{S115}$$

$$b_{tm} = \frac{1}{\alpha_t}\Big(1 + \frac{\tau^2}{\beta}\Big) - \frac{\tau^2}{\beta^2\alpha_t^2}\Big(\frac{o_t}{v_{tm}} + \frac{1}{\beta\alpha_t}\Big)^{-1}, \tag{S116}$$

$$c_{tm} = \frac{1}{\alpha_t}\frac{R}{\beta} - \frac{\tau(1-\tau)}{\beta^2\alpha_t^2}\Big(\frac{o_t}{v_{tm}} + \frac{1}{\beta\alpha_t}\Big)^{-1}. \tag{S117}$$

The RHS of the $\langle y_m\rangle$ system is given by

$$A_{yz}A_{zzm}^{-1}V_m x_m = \frac{(1-\tau)}{\beta}\begin{bmatrix} \frac{1}{\alpha_1}\Big(\frac{o_1}{v_{1m}} + \frac{1}{\beta\alpha_1}\Big)^{-1}\frac{x_{1m}}{v_{1m}} \\ \frac{1}{\alpha_2}\Big(\frac{o_2}{v_{2m}} + \frac{1}{\beta\alpha_2}\Big)^{-1}\frac{x_{2m}}{v_{2m}} \\ \vdots \\ \frac{1}{\alpha_T}\Big(\frac{o_T}{v_{Tm}} + \frac{1}{\beta\alpha_T}\Big)^{-1}\frac{x_{Tm}}{v_{Tm}} \\ 0 \end{bmatrix} + \frac{\tau}{\beta}\begin{bmatrix} 0 \\ \frac{1}{\alpha_1}\Big(\frac{o_1}{v_{1m}} + \frac{1}{\beta\alpha_1}\Big)^{-1}\frac{x_{1m}}{v_{1m}} \\ \frac{1}{\alpha_2}\Big(\frac{o_2}{v_{2m}} + \frac{1}{\beta\alpha_2}\Big)^{-1}\frac{x_{2m}}{v_{2m}} \\ \vdots \\ \frac{1}{\alpha_T}\Big(\frac{o_T}{v_{Tm}} + \frac{1}{\beta\alpha_T}\Big)^{-1}\frac{x_{Tm}}{v_{Tm}} \end{bmatrix}, \tag{S118}$$

and the $\langle y\rangle$-dependent part of $\langle z\rangle$ is given by

$$A_{zzm}^{-1}A_{yz}^{\dagger}\langle y_m\rangle = \Big[\dots, \Big(\frac{o_t}{v_{tm}} + \frac{1}{\beta\alpha_t}\Big)^{-1}\frac{1}{\beta\alpha_t}\Big((1-\tau)\langle y_{tm}\rangle + \tau\langle y_{tm}\rangle\Big), \dots\Big]^{\dagger} \in R^{T\times 1}. \tag{S119}$$

S3.6. The lower bound

We recall the expression for the lower bound just after updating $q(s)$,

$$F = \ln Z_s + \Big\langle \ln\frac{p(x|z)}{q(y,z)}\Big\rangle_{q(y,z)} + \Big\langle \ln\frac{p_0(\theta)}{q(\theta)}\Big\rangle_{q(\theta)}. \tag{S23}$$

Here, $\ln Z_s$ is the normalization constant in Eq. (S81), the contribution from measurement errors is

$$\begin{aligned}\langle \ln p(x|z)\rangle_{q(y,z)} &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\text{dim}} o_t\Big(\ln(2\pi v_{tm}) + v_{tm}^{-1}\big\langle (x_{tm} - z_{tm})^2\big\rangle\Big) \\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\text{dim}} o_t\Big(\ln(2\pi v_{tm}) + \frac{(x_{tm} - \langle z_{tm}\rangle)^2 + \Sigma_{z_{tm},z_{tm}}}{v_{tm}}\Big),\end{aligned} \tag{S120}$$

and the variational terms from $q(y,z)$ are given, since $q(y_m,z_m)$ are multivariate normal distributions of dimension $2T+1$, by

$$\begin{aligned}\langle \ln q(y,z)\rangle_{q(y,z)} &= -\frac{2T+1}{2}(1+\ln 2\pi)\times \dim + \frac{1}{2}\sum_{m=1}^{\dim}\ln\left|\Sigma_m^{-1}\right| \\ &= -\frac{2T+1}{2}(1+\ln 2\pi)\times \dim + \frac{1}{2}\sum_{m=1}^{\dim}\ln\begin{vmatrix} A_{yy} & -A_{yz} \\ -A_{zy} & A_{zzm}\end{vmatrix}. \end{aligned} \tag{S121}$$

The determinants can be simplified in various ways. Since A_{zzm} is diagonal and $A_{zy}=A_{yz}^{\dagger}$, we can use a block LU decomposition and write

$$\begin{bmatrix} A_{yy} & -A_{yz} \\ -A_{zy} & A_{zzm}\end{bmatrix} = \begin{bmatrix} I & -A_{yz}A_{zzm}^{-1} \\ 0 & I\end{bmatrix}\begin{bmatrix} A_{yy}-A_{yz}A_{zzm}^{-1}A_{zy} & 0 \\ 0 & A_{zzm}\end{bmatrix}\begin{bmatrix} I & 0 \\ -A_{zzm}^{-1}A_{zy} & I\end{bmatrix}, \tag{S122}$$

which means that

$$\left|\Sigma^{-1}\right| = |A_{zzm}|\left|A_{yy}-A_{yz}A_{zzm}^{-1}A_{zy}\right| = |A_{zzm}|\left|\Sigma_{yym}^{-1}\right|. \tag{S123}$$

Here, A_{zzm} is diagonal positive definite, so that determinant is simply the product of the diagonal elements. The second factor is a tridiagonal symmetric matrix which should be positive definite, since it is the inverse of a (symmetric, positive definite) covariance matrix.

Finally, parameter contributions are given by the negative Kullback-Leibler divergence from the variational distribution to the prior,

$$\begin{aligned}\langle \ln p_0(\theta)\rangle_{q(\theta)} - \langle \ln q(\theta)\rangle_{q(\theta)} &= -\left\langle \ln\frac{q(\theta)}{p_0(\theta)}\right\rangle_{q(\theta)} \\ &= -\left\langle \ln\frac{q(\pi)}{p_0(\pi)}\right\rangle_{q(\pi)} - \left\langle \ln\frac{q(a)}{p_0(a)}\right\rangle_{q(a)} - \left\langle \ln\frac{q(B)}{p_0(B)}\right\rangle_{q(B)} - \left\langle \ln\frac{q(\lambda)}{p_0(\lambda)}\right\rangle_{q(\lambda)} \end{aligned} \tag{S124}$$

Here, the contributions from π,a,B are just as in vbSPT. The step length variance is not the same variable as used in vbSPT, but it's KL divergence term turns out to be the same:

$$\left\langle \ln\frac{q(\lambda_j)}{p_0(\lambda_j)}\right\rangle_{q(\lambda)} = \ldots = \tilde{n}_j\ln\frac{c_j}{\tilde{c}_j} - n_j\Big(1-\frac{\tilde{c}_j}{c_j}\Big) - \ln\frac{\Gamma(n_j)}{\Gamma(\tilde{n}_j)} + (n_j-\tilde{n}_j)\psi(n_j) \tag{S125}$$

S4. MAXIMUM LIKELIHOOD AND MAXIMUM APOSTERIORI INFERENCE

S4.1. Parameter updates and log likelihood

It might also be useful to perform maximum likelihood inference in the model parameters, to get unbiased estimates and an impression about the influence of the priors. As we saw in Secs. S2.2-S2.3, to derive approximate maixmum likelihood (MLE) or maximum aposteriori (MAP) estimates, we simply drop the variational ansatz for the model parameters from the above variational treatment, and replace the optimization w.r.t. $\ln q(\theta)$ by optimizing the parameter values. Skipping a lot of details, the parameter updates for a single trajectory are given by the classical update formulae, which we give below for without (MLE) and with (MAP) the conjugate priors used above:

$$\begin{aligned}
&(MLE) && (MAP) && \text{(S126)}\\
\pi_k^* &= \frac{\langle \delta_{k,s_1}\rangle}{\sum_m \langle \delta_{m,s_1}\rangle}, & \pi_k^* &= \frac{w_k^{(\pi)}-1}{w_0^{(\pi)}-N}, && \text{(S127)}\\
a_k^* &= \frac{\sum_{t=1}^{T-1}\big(1-\langle \delta_{k,s_t}\delta_{k,s_{t+1}}\rangle\big)}{\sum_{j=1}^{N}\langle \delta_{k,s_t}\delta_{j,s_{t+1}}\rangle} = \frac{\sum_{j\neq k}\hat{w}_{kj}}{\sum_j \hat{w}_{kj}}, & a_k^* &= \frac{w_{k1}^{(a)}-1}{w_{k0}^{(a)}-2} = \frac{\tilde{w}_{k1}^{(a)}-1+\sum_{j\neq k}\hat{w}_{kj}}{\tilde{w}_{k0}^{(a)}-2+\sum_j \hat{w}_{kj}}, && \text{(S128)}\\
B_{kj}^* &= \frac{\sum_{t=1}^{T-1}\langle \delta_{k,s_t}\delta_{j,s_{t+1}}\rangle}{\sum_{t=1}^{T-1}\big(1-\langle \delta_{k,s_t}\delta_{k,s_{t+1}}\rangle\big)} = \frac{\hat{w}_{kl}}{\sum_{j\neq k}\hat{w}_{kl}}, & B_{kj}^* &= \frac{w_{kj}^{(B)}-1}{\sum_{j\neq k}(w_{kj}^{(B)}-1)} = \frac{\tilde{w}_{kj}^{(B)}-1+\hat{w}_{kj}}{\sum_{j\neq k}(\tilde{w}_{kj}^{(B)}-1+\hat{w}_{kj})}, && \text{(S129)}\\
A_{kj}^* &= \frac{\sum_{t=1}^{T-1}\langle \delta_{k,s_t}\delta_{j,s_{t+1}}\rangle}{\sum_j\sum_{t=1}^{T-1}\langle \delta_{k,s_t}\delta_{j,s_{t+1}}\rangle}, & A_{kj}^* &= \begin{cases}(1-a_k^*), & j=k\\ a_k^* B_{kj}^*, & j\neq k,\end{cases} && \text{(S130)}\\
\lambda_k^* &= \frac{\hat{c}_k}{\hat{n}_k}, & \lambda_k^* &= \frac{c_k}{n_k+1} = \frac{\tilde{c}_k+\hat{c}_k}{\tilde{n}_k+1+\hat{n}_k}, && \text{(S131)}
\end{aligned}$$

where $\hat{c}_k, \hat{n}_k$ are the data dependent terms of Eqs. (S69) and (S70). The MAP notation is the same as used for the variational parameter updates, and inclues the same conjugate priors. The iterations for hidden states and path can now be carried out by using delta functions for the variational parameter distributions, $q(\theta) = \delta(\theta-\theta^*)$. The lower bound on the likelihood can also be computed with the results above, except that the prior terms are absent (MLE) or just the log prior evaluated at the MAP parameter values. After the s-update, the lower bound is given by

$$\ln L = \underbrace{\ln Z_s}_{\text{forward-backward}} + \underbrace{\langle \ln p(x|z)\rangle_{y,z}}_{\text{eq. (S120)}} - \underbrace{\langle \ln q(y,z)\rangle_{y,z}}_{\text{eq. (S121)}} + \underbrace{\ln p_0(\theta)}_{\text{MAP only}}\ . \tag{S132}$$

S4.2. Transition matrix parameterization

It is perhaps more common to parameterize A directly, in which case conjugate priors lead to Dirichlet-distributed matrix rows, with parameters

$$w_{ij}^{(A)} = \tilde{w}_{ij}^{(A)} + \hat{w}_{ij}, \Rightarrow A_{ij}^{MLE} = \frac{\hat{w}_{ij}}{\sum_j \hat{w}_{ij}}, \quad A_{ij}^{MAP} = \frac{w_{ij}^{(A)}-1}{\sum_j (w_{ij}^{(A)}-1)} = \frac{\tilde{w}_{ij}^{(A)}-1+\hat{w}_{ij}}{\sum_j(w_{ij}^{(A)}-1+\hat{w}_{ij})} \tag{S133}$$

Not surprising, the MLE estimate is equivalent to the a_i,B_{ij} parameterization. However, for the MAP estimates to be consistent, we require

$$A_{kk}^{MAP} = \frac{\tilde{w}_{kk}^{(A)}-1+\hat{w}_{kk}}{\sum_j(w_{kj}^{(A)}-1+\hat{w}_{kj})} = \frac{\tilde{w}_{k2}^{(a)}-1+\hat{w}_{kk}}{\tilde{w}_{k1}^{(a)}+\tilde{w}_{k2}^{(a)}-2+\sum_j \hat{w}_{kj}} = 1-a_k^*, \tag{S134}$$

$$A_{kj}^{MAP} = \frac{\tilde{w}_{kj}^{(A)}-1+\hat{w}_{kj}}{\sum_j(w_{kj}^{(A)}-1+\hat{w}_{kj})} = \frac{\tilde{w}_{k1}^{(a)}-1+\sum_{j\neq k}\hat{w}_{kj}}{\underbrace{\sum_{j\neq k}(\tilde{w}_{kj}^{(B)}-1)+\sum_{j\neq k}\hat{w}_{kj}}_{(\dagger)}} \times \frac{\tilde{w}_{kj}^{(B)}-1+\hat{w}_{kj}}{\tilde{w}_{k1}^{(a)}+\tilde{w}_{k2}^{(a)}-2+\sum_j \hat{w}_{kj}} = a_k^* B_{kj}^*, \tag{S135}$$

where the second equation applies only for $k \neq j$. The simplest solution seems to be to set $(\dagger) = 1$, and then equate nominators and denominators separately, since we seek a solution independent of $\hat{w}_{kj}$. This leads to a unique solution

$$\tilde{w}_{k2}^{(a)} = \tilde{w}_{kk}^{(A)}, \qquad \tilde{w}_{kj}^{(B)} = \tilde{w}_{kj}^{(A)}, \qquad \tilde{w}_{k1}^{(a)} = 1 + \sum_{j\neq k}(\tilde{w}_{kj}^{(B)}-1). \tag{S136}$$

and thus we see that not all conjugate priors on a_k, B_{kj} are equivalent to A_{ij} priors in this sense. Of special interest is the flat prior $\tilde{w}_{ij}^{(A)} = 1$, which corresponds to

$$\tilde{w}_{k1}^{(a)} = 1, \quad \tilde{w}_{k2}^{(a)} = 1, \quad \tilde{w}_{kj}^{(B)} = 1, \tag{S137}$$

that leads to $A^{MAP} = A^{MLE}$ (in both parameterizations).

S5. LEARNING THE LOCALIZATION UNCERTAINTY

In some cases, point-wise uncertainty estimates might not available, in which case one can try to infer localization errors from the trajectories instead. It is obviously not a good idea to try to infer point-wise localization uncertainties (an underdetermined problem), but it might work to model an average localization uncertainty, or an average but state-dependent localization uncertainty. Here, we explore those possibilities.

S5.1. Average localization uncertainty

We start with a single uniform localization error. Compared to the point-wise uncertainty model, this case differs in the model of measured positions, which are instead given by

$$x_{tm} = z_{tm} + \sqrt{v}\xi_{tm}, \tag{S138}$$

corresponding to the likelihood term

$$\ln p(x|z,v) = -\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{\mathrm{dim}} o_t\Big(\ln(2\pi v) + v^{-1}(x_{tm}-z_{tm})^2\Big), \tag{S139}$$

where v is a single model parameter to be learned. For a maximum likelihood algorithm (MLE), the parameter update is then given by maximizing

$$\langle \ln p(x|z)\rangle_{q(y,z)} = const. - \hat{n}^{(v)}\ln v - \frac{\hat{c}^{(v)}}{v} \Rightarrow v^* = \frac{\hat{c}^{(v)}}{\hat{n}^{(v)}}, \tag{S140}$$

$$\hat{n}^{(v)} = \frac{\mathrm{dim}}{2}\sum_{t=1}^{T} o_t, \tag{S141}$$

$$\hat{c}^{(v)} = \frac{1}{2}\sum_{t=1}^{T} o_t \sum_{m=1}^{\mathrm{dim}} \left\langle (x_{tm}-z_{tm})^2\right\rangle_{q(y,z)} = \frac{1}{2}\sum_{t=1}^{T} o_t \sum_{m=1}^{\mathrm{dim}} \Big((x_{tm}-\langle z_{tm}\rangle)^2 + \Sigma_{z_t,z_t,m}\Big). \tag{S142}$$

The MLE parameter is further used to substitute $v_{mt}^{-1} \to 1/v^*$ in A_{zzm} and V_m in the $q(y,z)$ update.

The lower bound is also affected through the modified term $\langle \ln p(x|z)\rangle_{q(y,z)}$, and now becomes

$$\langle \ln p(x|z,v)\rangle_{q(y,z)} = -\hat{n}^{(v)}(\ln 2\pi + \ln v) - \frac{\hat{c}^{(v)}}{v}. \tag{S143}$$

In a variational Bayes (VB) algorithm , the conjugate prior is inverse gamma, which leads to

$$\ln q(v) = -\ln Z_v - (n^{(v)}+1)\ln v - \frac{c^{(v)}}{v}, \quad n^{(v)} = \tilde{n}^{(v)} + \hat{n}^{(v)}, \quad c^{(v)} = \tilde{c}^{(v)} + \hat{c}^{(v)}, \tag{S144}$$

where ˜ indicates prior parameters, and

$$\langle v\rangle_{q(v)} = \frac{c^{(v)}}{n^{(v)}-1}, \quad v^*|_{q(v)} = \frac{c^{(v)}}{n^{(v)}+1}, \quad \left\langle v^{-1}\right\rangle_{q(v)} = \frac{n^{(v)}}{c^{(v)}}, \quad \langle \ln v\rangle_{q(v)} = \ln c^{(v)} - \psi(n^{(v)}). \tag{S145}$$

The average $\left\langle v^{-1}\right\rangle_{q(v)}$ is used to substitute v_{mt}^{-1} in A_{zzm} and V_m in the $q(y,z)$ update, and $\langle \ln v\rangle_{q(v)}$ is used in the lower bound term

$$\langle \ln p(x|z,v)\rangle_{q(y,z)q(\theta)} = -\hat{n}^{(v)}(\ln 2\pi + \langle \ln v\rangle_{q(\theta)}) - \hat{c}^{(v)}\left\langle v^{-1}\right\rangle_{q(\theta)}. \tag{S146}$$

For prior specification, the RMS error, given by $r = \sqrt{v}$, is a more intuitive quantity. It's distribution is

$$f(r) = \frac{2c^n}{\Gamma(n)} r^{-(2n+1)} e^{-c/r^2}, \tag{S147}$$

and by making use of the asymptotic series expansion

$$\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)} = \sqrt{m}\Big(1 - \frac{1}{8m} + \frac{1}{128m^2} + O(m^{-3})\Big) \tag{S148}$$

from Mathworld[9], we get

$$\langle r\rangle =\sqrt{c}\frac{\Gamma(n-\frac{1}{2})}{\Gamma(n)}\approx\sqrt{\frac{c}{n-1}}\Big(1-\frac{1}{8(n-1)}+\dots\Big), \tag{S149}$$

$$r^* =\sqrt{\frac{c}{n+\frac{1}{2}}},\quad \langle r^2\rangle=\langle v\rangle=\frac{c}{n-1}. \tag{S150}$$

Further,

$$\mathrm{Var}[r]=\frac{c}{n-1}\bigg(1-(n-1)\Big(\frac{\Gamma(n-\frac{1}{2})}{\Gamma(n)}\Big)^2\bigg)\approx\frac{c}{4(n-1)^2}\Big(1-\frac{3}{32(n-1)}+\dots\Big), \tag{S151}$$

which with leads to

$$\mathrm{std}[r]=\frac{\sqrt{c}}{2(n-1)}\Big(1-\frac{3}{64(n-1)}+\dots\Big). \tag{S152}$$

The first order approximations

$$\langle r\rangle\approx\sqrt{\frac{c}{n-1}}=\sqrt{\langle v\rangle},\quad \mathrm{std}[r]\approx\frac{\sqrt{c}}{2(n-1)},\quad \frac{\mathrm{std}[r]}{\langle r\rangle}\approx\frac{1}{2\sqrt{n-1}}, \tag{S153}$$

are better than about 10% for $n>2$, or $\mathrm{std}[r]/r^*\leq 0.79$, or $\mathrm{std}[r]/\langle r\rangle<0.54$. It seems likely that one would be able to make an informed guess about the average localization error with smaller uncertainty than that.

If not, since the number of precision parameters does not change with model dimension, it is possible to use a Jeffreys prior $\sim 1/v$, corresponding to $\tilde{n}^{(v)}=c^{(v)}=0$. Note that this makes $\langle v^{-1}\rangle_{q(v)}=\hat{c}_{(v)}/\hat{n}_{(v)}$, which coincides with the MLE update $1/v^*$.

S5.2. State-wise localization uncertainty

One could also imagine modeling distinct localization uncertainties for different states. For example, this could be caused by motion blur, which contributes a diffusion-dependent terms to the localization uncertainty, or if different states may have different spatial distributions. This case is a little more complicated, since it adds additional interactions to the model. Here, the measurement model is modified to

$$x_{tm}=z_{tm}+\sqrt{v_{s_t}}\xi_{tm}, \tag{S154}$$

corresponding to the likelihood term

$$\ln p(x|z,s,v)=-\frac{1}{2}\sum_{k=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{\dim}o_t\delta_{k,s_t}\Big(\ln(2\pi v_k)+v_k^{-1}(x_{tm}-z_{tm})^2\Big), \tag{S155}$$

where there are now N precision parameters v_k. This term will also influence the hidden state distribution.

a. Maximum likelihood First, the parameter updates are given by maximizing

$$\langle\ln p(x|z,s,v)\rangle_{q(y,z)q(s)}=const.-\sum_{k=1}^{N}\Big[\hat{n}_k^{(v)}\ln v_k-\frac{\hat{c}^{(v)}}{v_k}\Big]\Rightarrow v_k^*=\frac{\hat{c}_k^{(v)}}{\hat{n}_k^{(v)}}, \tag{S156}$$

$$\hat{n}_k^{(v)}=\frac{\dim}{2}\sum_{t=1}^{T}o_t\langle\delta_{ks_t}\rangle, \tag{S157}$$

$$\hat{c}_k^{(v)}=\frac{1}{2}\sum_{t=1}^{T}o_t\langle\delta_{ks_t}\rangle\sum_{m=1}^{\dim}\big\langle(x_{tm}-z_{tm})^2\big\rangle_{q(y,z)} \tag{S158}$$

$$=\frac{1}{2}\sum_{t=1}^{T}o_t\langle\delta_{ks_t}\rangle\sum_{m=1}^{\dim}\Big((x_{tm}-\langle z_{tm}\rangle)^2+\Sigma_{z_t,z_t,m}\Big). \tag{S159}$$

Second, there is an additional term in the hidden state distribution,

$$\ln H_{tk} = \ldots - \frac{o_t}{2}\left[\dim \ln v_k^* + \frac{1}{v_k^*}\sum_{m=1}^{\dim}\Big((x_{tm} - \langle z_{tm}\rangle)^2 + \Sigma_{z_t,z_t,m}\Big)\right]. \tag{S160}$$

Third, the localization precision contribution to the trajectory distributions become

$$\ln q(y,z) = -\frac{1}{2}\sum_{k=1}^{N}\sum_{t=1}^{T} o_t \frac{\langle \delta_{k,s_t}\rangle}{v_k^*}\sum_{m=1}^{\dim}(x_{tm} - z_{tm})^2 + \ldots, \tag{S161}$$

which means that we can introduce an effective time-dependent localization uncertainty given by

$$\tilde{v}_t = \left(\sum_{k=1}^{N}\frac{\langle \delta_{k,s_t}\rangle}{v_k^*}\right)^{-1}, \tag{S162}$$

which is substituted for v_{tm} in A_{zzm} and V_m for the $q(y,z)$ update.

Finally, the fact that $\langle \ln p(x|z,s,v)\rangle_{q(y,z)q(s)}$ contributes to $\ln q(s)$ means that needs not be explicitly accounted for in the lower bound expression, since it is already included in $\ln Z_s$. The lower bound (after an s-update) then simplifies to

$$\ln L = \ln Z_s - \langle \ln q(y,z)\rangle_{q(y,z)}\,. \tag{S163}$$

b. Variational Bayes For the precision parameters, this is analogous to the 1-parameter case, with inverse gamma priors on v_k leading to inverse gamma distributions for $q(v_k)$,

$$\ln q(v_k) = -\ln Z_{v_k} - n_k^{(v)}\ln v_k - \frac{c^{(v)}}{v_k}, \quad n_k^{(v)} = \tilde{n}_k^{(v)} + \hat{n}_k^{(v)}, \quad c_k^{(v)} = \tilde{c}_k^{(v)} + \hat{n}_k^{(v)}, \tag{S164}$$

and expectation values as in Eq. (S145).

For the hidden states, the additional contribution to H is

$$\ln H_{tk} = \ldots - \frac{o_t}{2}\left[\dim \langle \ln v_k\rangle + \left\langle v_k^{-1}\right\rangle \sum_{m=1}^{\dim}\Big((x_{tm} - \langle z_{tm}\rangle)^2 + \Sigma_{z_t,z_t,m}\Big)\right]. \tag{S165}$$

For the trajectory distribution, it is again convenient to introduce an effective time-dependent localization precision

$$\tilde{v}_t = \left(\sum_{k=1}^{N}\langle \delta_{k,s_t}\rangle \left\langle v_k^{-1}\right\rangle\right)^{-1} = \left(\sum_{k=1}^{N}\langle \delta_{k,s_t}\rangle \frac{n_k^{(v)}}{c_k^{(v)}}\right)^{-1}, \tag{S166}$$

and substitute in A_{zzm} and V_m.

Again, $\ln p(z|x,s,\theta)$ makes no explicit contribution to the lower bound, since it is already included in $\ln Z_s$.

[1] Fredrik Persson, Martin Lindén, Cecilia Unoson, and Johan Elf. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Meth.*, 10(3):265–269, 2013. doi:10.1038/nmeth.2367.

[2] Andrew J. Berglund. Statistics of camera-based single-particle tracking. *Phys. Rev. E*, 82(1):011917, 2010. doi: 10.1103/PhysRevE.82.011917.

[3] Martin Lindén, Vladimir Ćurić, Elias Amselem, and Johan Elf. Pointwise error estimates in localization microscopy. *Nat Commun*, 8:15115, 2017. doi:10.1038/ncomms15115.

[4] David MacKay. *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.

[5] Matthew Beal. *Variational algorithms for approximate Bayesian inference.* PhD thesis, University of Cambridge, UK, 2003. URL `http://www.cse.buffalo.edu/faculty/mbeal/thesis/`.

[6] Colin H. LaMont and Paul A. Wiggins. The Lindley paradox: The loss of resolution in Bayesian inference. *arXiv:1610.09433 [math, stat]*, 2016. arXiv: 1610.09433.

[7] A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Stat. Soc. B Met.*, 56 (3):501–514, 1994. doi:10.2307/2346123.

[8] G. Meurant. A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. *SIAM. J. Matrix Anal. & Appl.*, 13(3):707–728, 1992. doi:10.1137/0613045.

[9] Note1. `http://mathworld.wolfram.com/GammaFunction.html`.