Bayesian inference of scaled versus fractional Brownian motion

To cite this article: Samudrajit Thapa et al 2022 J. Phys. A: Math. Theor. 55 194003

View the article online for updates and enhancements.
Bayesian inference of scaled versus fractional Brownian motion

Samudrajit Thapa\textsuperscript{1,2}*, Seongyu Park\textsuperscript{3}*, Yeongjin Kim\textsuperscript{3}, Jae-Hyung Jeon\textsuperscript{3}*, Ralf Metzler\textsuperscript{4} and Michael A Lomholt\textsuperscript{5}*

\begin{itemize}
\item \textsuperscript{1} Sackler Center for Computational Molecular and Materials Science, Tel Aviv University, Tel Aviv 6997801, Israel
\item \textsuperscript{2} School of Mechanical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel
\item \textsuperscript{3} Department of Physics, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea
\item \textsuperscript{4} Institute of Physics and Astronomy, University of Potsdam, D-14476 Potsdam-Golm, Germany
\item \textsuperscript{5} PhyLife, Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark
\end{itemize}

E-mail: mlomholt@sdu.dk

Received 30 December 2021, revised 8 March 2022
Accepted for publication 24 March 2022
Published 12 April 2022

Abstract

We present a Bayesian inference scheme for scaled Brownian motion, and investigate its performance on synthetic data for parameter estimation and model selection in a combined inference with fractional Brownian motion. We include the possibility of measurement noise in both models. We find that for trajectories of a few hundred time points the procedure is able to resolve well the true model and parameters. Using the prior of the synthetic data generation process also for the inference, the approach is optimal based on decision theory. We include a comparison with inference using a prior different from the data generating one.

Keywords: Bayesian inference, scaled Brownian motion, single particle tracking

(Some figures may appear in colour only in the online journal)

1. Introduction

With Robert Brown’s observation of the random motion of micron-sized granules contained in pollen grains [1], Albert Einstein’s explanation of the physical origin of Brownian motion [2], and Jean Perrin’s [3] and Ivar Nordlund’s [4] subsequent quantitative measurements, the field
of single particle tracking was born. More recently, with the advent of fluorescence microscopy, it has become possible to track single particles inside highly complex environments such as biological cells. This has revealed motion distinct from pure Brownian, and a wealth of anomalous diffusion models have been put forward to quantify this behaviour [5, 6]. The question then arises, how does one determine which of these models best fit the observed random motion?

Mathematical models of randomly fluctuating data—such as the position time series measured in single particle tracking experiments—are comprehensibly described by writing down the probability of the complete data given the model: \( P(\text{data}|\text{model}) \). But when we do inference we would like to make deductions in the opposite direction. In Bayesian inference such deductions are quantified by allowing probability to be interpreted as belief in different models (including the parameter values necessary to completely specify the model), and then the beliefs are updated when data is revealed through the use of Bayes’ formula, which up to a normalisation constant says that \( P(\text{model}|\text{data}) \propto P(\text{data}|\text{model})P(\text{model}) \) [7–10]. The main controversies around Bayesian inference originates from the need to assign probabilities \( P(\text{model}) \) that represent prior beliefs in the model in the absence of any data. For this assignment, there is no universally accepted procedure.

The objections leading to this controversy, however, are not valid when the data are generated by procedures where probability distributions for all random assignments are known. In this case \( P(\text{model}) \) is a known probability distribution and Bayes’ formula has the status of a mathematical theorem with satisfied prerequisites. The problem of inferring a parameter value in a way that optimises the average value of a certain score function thus becomes a well-defined problem in decision theory with an optimal solution [9, 11].

In the AnDi challenge [12] artificial trajectories are generated from five disclosed models of anomalous diffusion with announced limits on the model parameters. The precise parameter distributions, however, are not disclosed. This introduces some psychological guesswork into the competition, since arbitrary choices have to be made, for instance, for the Bayesian parameter priors. Similarly, for machine learning methods choices also have to be made for the distributions of parameters with which the training sets are generated. However, the AnDi challenge is very close to having disclosed the data generating procedure completely. Thus one would expect that an approach based on Bayesian inference and decision theory in principle should be able to come very close to an optimal solution.

Although an approach based on Bayesian inference and decision theory is optimal in principle, carrying out the necessary computations to evaluate the solution can be a difficult problem, in particular when the model contains hidden variables that needs to be integrated out. Thus the AnDi challenge team that many of the present authors participated in (BIT) did not succeed in implementing effective computational solutions for all five models in time for the challenge.

In this article we present our Bayesian inference approach for one of the AnDi challenge models: scaled Brownian motion (SBM)\(^6\). SBM is a simple Markovian model of anomalous diffusion, where Brownian motion is modified by allowing the diffusion coefficient to depend on time [17]. This approach with time dependent diffusion coefficient is a convenient way to generalise for instance formulas for photobleaching recovery data [18]. We combine our implementation for SBM with the implementation of Bayesian inference for fractional Brownian motion (FBM) presented in [14] to demonstrate model selection and inference of the anomalous diffusion exponent for these two models. FBM is a modification of Brownian motion, where correlations are allowed for the motion at different times [19]. Figure 1 shows examples for subdiffusive and superdiffusive two-dimensional trajectories of SBM and FBM. The

\(^6\)We refer interested readers to [13–16] for details of our implementation of Bayesian inference for other models of diffusion.
trajectories of these processes can look qualitatively different because while SBM is a process with non-stationary increments, FBM has stationary increments. Subdiffusive SBM diffuses progressively slowly with time, whereas superdiffusive SBM diffuses progressively fast. However, the trajectories of FBM and SBM can look similar, as is shown in figure 1 by selecting trajectories that look alike from a few realisations.

The article is organised with our inference methods being presented in section 2, before we present results showing how effective they are for parameter estimation and model selection in section 3 and conclude in section 4.

2. Methods

In this section we briefly review Bayesian inference before presenting the likelihood functions of the models in subsections 2.2–2.5. Finally, we state our computational methods.

2.1. Bayesian inference

If we have a time series \( \Delta x_N \) of \( N \) steps generated from a known model \( M_i \) with parameters \( \theta \) taken from a known probability distribution \( \pi_i(\theta) = P(\theta|M_i) \), Bayes’ theorem tells us that the probability distribution (called posterior) for the unknown parameter \( \theta \) is

\[
P(\theta|\Delta x_N, M_i) = \frac{\mathcal{L}_i(\theta)\pi_i(\theta)}{E_i},
\]

where \( \mathcal{L}_i(\theta) = P(\Delta x_N|\theta, M_i) \) is called the likelihood function and the normalisation constant (called evidence) is \( E_i = P(\Delta x_N|M_i) = \int d\theta \mathcal{L}_i(\theta)\pi(\theta) \). If, additionally, the model \( M_i \) is randomly chosen from a set of \( M \) models: \( M_i, i = 1, \ldots, M \), with probabilities \( \pi(M_i) \), then we can again use Bayes’ theorem to find the probability \( P(M_i|\Delta x_N) \) that the model was \( M_i \) given an artificially generated trajectory \( \Delta x_N \),

\[
P(M_i|\Delta x_N) = \frac{E_i\pi(M_i)}{\sum_{i=1}^{M} E_i\pi(M_i)},
\]

3
If the priors $\pi_i(\theta)$ and $\pi(M_i)$ are not known, for instance, if the trajectory is experimental data from a system that is not well known, then Bayesian inference requires that we interpret the priors as quantifying our lack of knowledge about the parameters and the true model. For instance, if we have no basis for assuming that one model is more probable than another, we should choose $\pi(M_i) = 1/M$. For a continuous parameter with unlimited range the choice of a uniform prior is generally not possible. In this case one then has to shape the prior non-uniformly according to judgement about realistic values of the parameter.

2.2. Scaled Brownian motion

Scaled Brownian motion (SBM) is a popular stochastic process that has been used to model anomalous diffusion observed in experiments where the diffusion coefficient seems to be time-dependent [20–25]. It is particularly convenient to describe experiments with fluorescence recovery after photobleaching [18] in which the diffusion coefficient appears to depend on time as a power law: $D(t) = \alpha K_\alpha (t + t_0)^{\alpha - 1}$. Here $K_\alpha$ is the anomalous diffusion coefficient, $\alpha$ the anomalous diffusion exponent, and $t_0$ the ageing time of the system prior to $t = 0$ (all taken to be real and positive numbers). SBM in one dimension can be described mathematically as having the mean squared displacement

$$\langle [x(t) - x(0)]^2 \rangle = 2K_\alpha [(t + t_0)^\alpha - t_0^\alpha]$$

and zero mean independent Gaussian increments [17]. The corresponding probability density $P(x, t)$ for the position $x$ satisfies a diffusion equation with explicit time dependence

$$\frac{\partial P}{\partial t} = D(t)\frac{\partial^2 P}{\partial x^2}$$

and the increments are not stationary in general since

$$\langle [x(t + \Delta t) - x(t)]^2 \rangle = 2K_\alpha [(t + \Delta t + t_0)^\alpha - (t + t_0)^\alpha]$$

depends on the time $t$ unless we are in the case of pure Brownian motion with $\alpha = 1$. SBM exhibits a number of interesting properties, such as ageing and non-ergodic behaviour. Regarding ageing, this can be seen from equation (5), since the statistics of the increments depend on the time $t + t_0$ since initiation of the system, which is the characteristic feature of ageing. For a further discussion of these properties we refer the reader to [26–29].

Since the increments are independent and Gaussian, we can obtain the probability density for a set of increments $\Delta x_i = (\Delta x_1, \ldots, \Delta x_N)$ where $\Delta x_i = x_i - x_{i-1}$ and $x_i = x(t_i)$ for equidistant times $t_i = i\Delta t$. The probability density is a product of the densities for each increment,

$$P(\Delta x_i | \theta, M_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \sigma_i^2}} \exp \left( -\frac{\Delta x_i^2}{2\sigma_i^2} \right),$$

where we have labelled the SBM model $M_1$, $\theta = (\alpha, K_\alpha, t_0)$ and $\sigma_i^2 = 2K_\alpha [\Delta t^{\alpha} [(i + t_0/\Delta t)^\alpha - (i - 1 + t_0/\Delta t)^\alpha]]$.

2.3. Scaled Brownian motion with measurement noise

The experimental observation of single particle motion is complicated by noise in the measurement procedure [30, 31]. Here we model measurement noise as independent Gaussian
displacements $\eta_i$ of the underlying true positions $x_i^{\text{true}}$ resulting in a measured position at time $t_i$ which is
\[ x_i = x_i^{\text{true}} + \eta_i, \]
where $x_i^{\text{true}}$ follows the statistics of SBM as presented in the previous subsection and $\eta_i$ are independent with zero mean and variance $\langle \eta_i^2 \rangle = \sigma_{\text{err}}^2$. For this situation, we can find a formula for the probability of the observed trajectory
\[ P(\Delta x_N|\theta, M_1) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi \sigma_i^2}} \exp \left( -\frac{1}{2\sigma_i^2} (\Delta x_i - \Delta \hat{x}_i)^2 \right), \]
where we have recursively for $i \geq 1$
\[ \Delta \hat{x}_{i+1} = E(\Delta x_{i+1}|\Delta x_i) = -\frac{\sigma_{\text{err}}^2}{\sigma_i^2} (\Delta x_i - \Delta \hat{x}_i), \]
\[ \sigma_{i+1}^2 = \text{Var}(\Delta x_{i+1}|\Delta x_i) = \sigma_i^2 + \sigma_{\text{err}}^2 \left( 2 - \frac{\sigma_{\text{err}}^2}{\sigma_i^2} \right), \]
with base case $\sigma_1^2 = \sigma_{\text{err}}^2 + 2 \sigma_{\text{err}}^2$ and $\Delta \hat{x}_1 = 0$. Here $E(a|b)$ means the expected value of $a$ given $b$ and $\text{Var}(a|b)$ means variance of $a$ given $b$. The derivation of the above formulas is a straightforward generalisation of the derivation in [13], where the $\sigma_i$ are assumed to be identical. This is the case since according to the derivation in [13] the joint probability distribution for $x_{i+1}$ and $\eta_{i+1}$ at time $t_{i+1}$ only depends on the corresponding distribution at time $t_i$ and what happens during the time interval to $t_{i+1}$.

Note from equation (9) that the measurement noise introduces anti-correlations between steps: if a step was further to the right than expected, then the next step will tend to be to the left. This is similar to subdiffusive FBM (i.e., with $\alpha < 1$), but there the correlations are long-range.

2.4. Fractional Brownian motion

Fractional Brownian motion (FBM) [19, 32] is a generalisation of Brownian motion to include correlated increments that can successfully model anomalous diffusion in numerous experiments, particularly in visco-elastic systems [33–42]. FBM with measurement noise is a stationary Gaussian process whose increments are correlated with covariance $\langle \Delta x_i \Delta x_j \rangle = \gamma(i - j)$ where
\[ \gamma(n) = \begin{cases} 2K_n \Delta t^n + 2 \sigma_{\text{err}}^2, & n = 0 \\ K_n \Delta t^n (2^n - 2) - \sigma_{\text{err}}^2, & |n| = 1 \\ K_n \Delta t^n (|n + 1^n| + |n - 1^n| - 2 |n|), & |n| \geq 2 \end{cases} \]
with $K_n$ being the diffusion coefficient [14]. Note that for $\alpha = 1$, this becomes identical to SBM with $\alpha = 1$ and arbitrary $t_0$. The probability for a trajectory (the likelihood function) can be written as
\[ P(\Delta x_N|\theta, M_2) = \frac{1}{(2\pi)^{N/2} |\Gamma_N|^{1/2}} \exp \left( -\frac{1}{2} \Delta x_N^\top \Gamma_N^{-1} \Delta x_N \right), \]
where $\Gamma_N$ is the covariance matrix with components $\Gamma_{i,j} = \gamma(i - j)$ and we have labelled FBM as $M_2$. FBM does not display ageing, and it is ergodic, albeit with slow convergence of
the time-averaged mean-squared displacement to its ensemble averaged limit [43–45]. Despite these differences with SBM it obeys the same time dependent diffusion equation, equation (4), and has mean square displacement as in equation (3), but with no effect of ageing, i.e., \( t_0 = 0 \). To calculate the likelihood function for FBM numerically, we employ the Durbin-Levinson algorithm as done previously for Bayesian inference with FBM in [14].

2.5. Multiple dimensions

The likelihood functions for SBM and FBM described above straightforwardly generalises to multiple dimensions, by assuming independent steps in each dimension. The likelihood function then becomes a product of the likelihood functions in each dimension with shared \( \alpha, K_\alpha \) and possibly \( t_0 \). For instance in two dimensions we will have

\[
L_i(\theta) = P(\Delta x_N | \theta, M_i)P(\Delta y_N | \theta, M_i),
\]

where \( \Delta y_N \) contains the coordinates along the second dimension. Keeping in mind the large number of single particle tracking experiments in two dimensions [46, 47], we will mainly consider two dimensions in the following. The results are similar in other dimensions and we present results from one-dimensional trajectories in the appendix A.

2.6. Computational methods

To evaluate the model evidences \( E_i \) and sample from the posteriors \( P(\theta | \Delta x_N, M_i) \) we use the nested sampling algorithm of skilling [48, 49]. Our implementation is an update of the one presented in [14], where the random walk Monte Carlo steps have been replaced by the discontinuous Hamiltonian Monte Carlo method presented in [50]. We have uploaded our implementation to GitHub: https://github.com/Samudrajit11/SBM.

3. Results

Here we specify our priors and parameters for the synthetic data generation before giving our results for the parameter estimation in subsection 3.3 and model selection in 3.4. Finally, we test our implementation on continuous time random walk (CTRW) trajectories, which illustrates the use of Bayesian inference on data not generated by a model included in the inference.

3.1. Priors

We choose a uniform prior on \( \alpha \) restricted to \( 0 < \alpha < 2 \) for both SBM and FBM. We parameterise the diffusion coefficient \( K_\alpha \) of FBM as \( K_\alpha = \sigma^2/(2\Delta t' \alpha) \) with \( \log_{10} \sigma \) having a standard normal distribution, where \( \sigma \) is the step-deviation. In the case of SBM, the step-deviation \( \sigma_1 \) (and therefore diffusion coefficient) changes with time. We choose the standard deviation of the first step \( \sigma_1 \) as a parameter with \( \log_{10} \sigma_1 \) having a standard normal distribution. For the noise strength \( \sigma_{mn} \) we choose a uniform prior in the range \( 0 < \sigma_{mn} < 1 \) for the analysis of all simulated data sets except for those generated with \( \sigma_{mn} = 10 \). We choose a uniform prior on \( \sigma_{mn} \) in the range \( 0 < \sigma_{mn} < 10 \) for the latter case, such that the true value of the noise-strength lies within the prior range. For the analysis with a wrong prior on \( \alpha \) (see below) we use a linear prior \( p(\alpha) = \frac{\alpha}{\alpha} \) with \( 0 < \alpha < 2 \). Finally, the models have prior \( \pi(M_i) = 1/2 \).

3.2. Details of simulated data sets

In order to quantify the performance of our code, we simulated SBM and FBM trajectories to constitute data sets that were used in the analysis described in the subsequent sections. All
Figure 2. MAE vs $\alpha_{GT}$ plot with the estimated MAE obtained from the analysis of $N = 100$ two-dimensional SBM trajectories for each $\alpha_{GT}$. Each sub-plot shows the results from the analysis of SBM trajectories generated with different noise-strengths $\sigma_{mn}$. The error-bars are the standard error on the mean.

The SBM trajectories were generated with $t_0 = 0$, $\Delta t = 1$ and $\log_{10}\sigma_1$ drawn from a standard normal distribution. The FBM trajectories were also generated with $\Delta t = 1$ and $\log_{10}\sigma$ drawn from a standard normal distribution. The other simulation parameters such as $\alpha_{GT}$, $N$ and $\sigma_{mn}$ are specified in the subsequent sections for each data set used in the analysis. We remark here that a choice of $\sigma_{mn} = 0$ means no measurement noise, while for $\sigma_{mn} = 1$ the measurement noise will typically be comparable in strength to $\sigma_1$, i.e., the diffusive noise of the first single step. The codes for the simulations of both FBM and SBM trajectories are home-written and presented at: https://github.com/Samudrajit11/SBM. The FBM generation algorithm is the same that was used in [14].
3.3. Quantification of parameter estimation

To quantify the parameter estimation results we use the mean absolute error (MAE), which for a parameter $\theta$ is defined as [12]

$$\text{MAE} = \frac{1}{\tilde{N}} \sum_{j=1}^{\tilde{N}} |\theta_{j,p} - \theta_{j,\text{GT}}|,$$  \hspace{1cm} (14)

where the subscript $j$ denotes the trajectory number, $\theta_{j,p}$ is the inferred parameter value, $\theta_{j,\text{GT}}$ is the true value (ground truth) and $\tilde{N}$ is the total number of trajectories. In what follows we choose $\tilde{N} = 100$ trajectories for each MAE estimate, except when we compare the results in the case of different priors on $\alpha$. For the latter case we choose $\tilde{N} = 1000$ for each MAE estimate. This is because, to highlight the effect of the choice of prior distribution, we need a sufficiently large number of trajectories to achieve a significant difference between the two priors.

In order to compute the MAE in equation (14) we consider the median of the posterior distribution (equation (1)) weighted with the corresponding model probability (equation (2)) as the inferred value. Therefore the inferred $\alpha$ from a particular trajectory $j$ is obtained as

$$\alpha_{j,p} = \tilde{\alpha}_{j,1} P(M_1 | (\Delta x_N)_j) + \tilde{\alpha}_{j,2} P(M_2 | (\Delta x_N)_j),$$  \hspace{1cm} (15)

where $\tilde{\alpha}_{j,i}$ is the median of the posterior distribution of $\alpha$ for model $M_i$ and trajectory $j$, whose steps are collected in $(\Delta x_N)_j$. We have chosen the median of the posterior as estimator of $\alpha$, since this choice is the optimal one based on decision theory, i.e., on average it minimises the MAE [9, 11].

Figure 2 shows the results of MAE on the estimations of $\alpha$ as a function of $\alpha_{\text{GT}}$ used to generate the two-dimensional SBM trajectories. As expected, we see the estimations get better with the length $N$ of the trajectories. The analysis of noise-free trajectories ($\sigma_{mn} = 0$) shows that the estimation is better for high and low values of $\alpha_{\text{GT}}$, whereas it gets relatively worse close to $\alpha_{\text{GT}} = 1$. A comparison of these results with those from the analysis of noisy trajectories shows interesting differences. While in general the analysis of noisy trajectories gives worse estimates of $\alpha$ as compared to the estimates from noise-free trajectories, the noise effects the estimates of very low and very high values of $\alpha$ much more than the values close to $\alpha_{\text{GT}} = 1$.

We attribute part of the explanation of this to the enlarged possibility that a predicted $\alpha$ is far away from the ground truth when the ground truth is extreme. For instance, if the measurement noise strength is completely obscuring the actual position, we would expect the posterior on $\alpha$ to become equal to the uniform prior. Averaging over this uniform distribution with $0 < \alpha < 2$ we obtain $\langle \text{MAE} \rangle = (\alpha_{\text{GT}} - 1)^2 / 2 + 1/2$, i.e., larger values in the cases where $\alpha_{\text{GT}}$ is extreme. Figure A1 shows similar results from the analysis of one-dimensional SBM trajectories.

On looking at the corresponding results from the analysis of two-dimensional FBM trajectories in figure 3, we find that for the cases of noise-free ($\sigma_{mn} = 0$) and relatively less noisy ($\sigma_{mn} = 0.1$) trajectories, the MAE as a function of the $\alpha_{\text{GT}}$ shows an interesting asymmetry. It decreases monotonically with increasing $\alpha_{\text{GT}}$. This can be understood from realising that the measurement noise is anti-persistent in nature, i.e., it induces anti-correlation between steps (see equation (9)), and because subdiffusive FBM itself is anti-persistent, it seems reasonable that it becomes increasingly difficult to estimate the anomalous diffusion exponent with increasing anti-persistence, i.e., for lower values of $\alpha_{\text{GT}}$. On increasing the noise-strength (the cases of $\sigma_{mn} = 1$ and $\sigma_{mn} = 10$), we recover the systematic increase in the values of MAE farther away from $\alpha_{\text{GT}} = 1$ in both directions like for SBM. Figure A2 shows similar results from the analysis of one-dimensional FBM trajectories.
Figure 3. MAE vs $\alpha_{GT}$ plot with the estimated MAE obtained from the analysis of $\tilde{N} = 100$ two-dimensional FBM trajectories for each $\alpha_{GT}$. Each sub-plot shows the results from the analysis of FBM trajectories generated with different noise-strengths $\sigma_{mn}$. The error-bars are the standard error on the mean.

Taking cognisance of the fact that parameter estimation can be performed with or without conditioning on particular models, in figure 4 we compare our estimation of $\alpha$ using equation (15)—where we considered the posterior median weighted with the model probabilities—with the inference using the posterior median of the true model. The figure on the left (right) shows results from the analysis of $\tilde{N} = 100$ SBM (FBM) trajectories of length $N = 200$ at each $\alpha_{GT}$. For both SBM and FBM trajectories, and for both $\sigma_{mn} = 0.1$ and $\sigma_{mn} = 10$, we find that using the posterior of the true model does not improve the MAE significantly.

In figure 5 we show what effect choosing a wrong prior on $\alpha$ has on its estimation. The results labelled ‘wrong prior’ correspond to a linear prior on $\alpha$ whereas those labelled ‘correct prior’ correspond to a uniform prior (see section 3.1). The figure on the left shows the MAE vs $N$ plot from the analysis of $\tilde{N} = 1000$ noise-free SBM trajectories for each $N$. Each trajectory was generated with a $\alpha_{GT}$ drawn from a uniform distribution in the range $0 < \alpha_{GT} < 2$. The figure on the right shows the corresponding results from the analysis of $\tilde{N} = 1000$ noise-free
Figure 4. Comparison of the MAE vs $\alpha_{GT}$ plot between the estimates from the weighted average (labelled ‘weighted’) and the estimates from the true model without considering model inference (labelled ‘true’). The error-bars are the standard error on the mean. The labels are the same for the figure on the right as shown in the figure on the left.

Figure 5. MAE vs $N$ plot with the estimated MAE obtained from the analysis of $\tilde{N} = 1000$ two-dimensional SBM (left) and FBM (right) trajectories. The results labelled as ‘wrong prior’ were inferred using a linear prior on $\alpha$. The error-bars are the standard error on the mean.

FBM trajectories for each $N$. In both cases we observe that choosing the correct prior results in a better estimate of $\alpha$ at low $N$. The estimations get better with longer trajectories and converges for the two choices of prior. Since the increments of FBM are correlated there are effectively fewer independent data points and it makes sense that the convergence for the two choices of prior happens more slowly with increasing $N$ in that case.
3.4. Quantification of model selection

To quantify the model comparison results we use the $F_1$ metric, which is defined as [12]

$$F_1 = \frac{\text{True positive}}{N_{\text{total}}},$$

(16)

where true positive denotes the total number of trajectories assigned to the correct model after model comparison and $N_{\text{total}}$ denotes the total number of trajectories considered. We assign each trajectory $\Delta x_N$ to the model $M_i$ that has the highest value of the posterior model probability $P(M_i | \Delta x_N)$ as defined in equation (2). Except when we compare the effect of different choices of priors, we consider two models (FBM and SBM) with $N = 100$ trajectories each, and therefore with $N_{\text{total}} = 200$. Just as in the case of parameter estimation, to study the effect of different choices of priors on $\alpha$ we consider the $F_1$ score from FBM and SBM trajectories but with $N = 1000$ trajectories each, and therefore with $N_{\text{total}} = 2000$. Note that $F_1$ is always
Figure 7. Plot of the confusion matrix highlighting the asymmetry between subdiffusive and superdiffusive cases during model inference. For the subdiffusive case ($\alpha_{GT} = 0.4$) shown in the upper panel, with increasing noise-strength model inference gets asymmetrically worse, with more SBM trajectories incorrectly inferred as FBM than FBM trajectories inferred as SBM. The lower panel shows that for the superdiffusive case ($\alpha_{GT} = 1.6$), model inference gets more symmetrically worse with increasing noise-strength.

in the range $0 \leq F_1 \leq 1$ with a high value of $F_1$ corresponding to better model prediction. The error on the estimate of $F_1$ is obtained as the standard error on the mean of a Bernoulli random variable. This is estimated as $\sqrt{p(1-p)/N_{total}}$ where $p$ is the fraction of trajectories for which the model prediction is correct.

Figure 6 shows the $F_1$ vs $\alpha_{GT}$ plot for different lengths of the trajectories, $N$ and each sub-plot shows results from the analysis of trajectories generated with a specific noise-strength $\sigma_{mn}$. For all noise-strengths, at $\alpha_{GT} = 1$, $F_1 \approx 0.5$ which is expected because at this value of $\alpha_{GT}$ both FBM and SBM are identical to Brownian motion and therefore should be indistinguishable. In addition to the expected result that the $F_1$ score gets better with increasing length of the trajectories, and worse with increasing strength of the measurement noise, we see some interesting features. The analysis of noise-free trajectories ($\sigma_{mn} = 0$) gives $F_1$ scores more or less symmetric around $\alpha_{GT} = 1$, with a score getting systematically better with $\alpha_{GT}$ farther away from $\alpha_{GT} = 1$ in both directions. The addition of measurement noise breaks this symmetry, with the $F_1$ score getting profoundly lower with increasing noise-strength for subdiffusive trajectories as compared to superdiffusive trajectories. This can be understood in terms of the measurement noise being anti-persistent in nature. With increasing noise-strength the SBM trajectories develop anti-persistent characteristics (see equation (9)) making them difficult to distinguish from subdiffusive FBM trajectories which are inherently anti-persistent. This should lead a higher number of SBM trajectories incorrectly predicted as being FBM trajectories in comparison to the number of FBM trajectories falsely detected as SBM trajectories. Moreover, this effect should occur in the subdiffusive case but not in the superdiffusive case where noise-free FBM is persistent, not anti-persistent. This is exactly what we see in figure 7, which shows the comparison of model inference via the confusion matrix plot. The upper panel shows the results from subdiffusive ($\alpha_{GT} = 0.4$) SBM and FBM trajectories while the lower panel shows the results from superdiffusive ($\alpha_{GT} = 1.6$) SBM and FBM trajectories.

The trajectories were generated with $N = 200$ and fixed $\sigma_{mn}$ (as specified on the title of each sub-plot) for both cases. The model prediction is very good for both cases, and for both FBM and SBM trajectories in the case of small noise-strengths ($\sigma_{mn} = 0$ and $\sigma_{mn} = 0.1$). However,
Figure 8. $F_1$ score vs $N$ plot from the analysis of two-dimensional noise-free SBM and FBM trajectories. The $F_1$ score at each $N$ are from the analysis of $\tilde{N} = 1000$ SBM and $\tilde{N} = 1000$ FBM trajectories generated with that particular $N$ and a $\alpha_{GT}$ chosen from a uniform distribution in the range $0 < \alpha < 2$. The results labelled as ‘wrong prior’ were inferred using a linear prior on $\alpha$. The error-bars are the standard error on the mean for a Bernoulli random variable.

for large noise-strengths ($\sigma_{mn} = 1$ and $\sigma_{mn} = 10$) we clearly see an asymmetry in the model prediction between subdiffusive FBM and SBM trajectories. With increasing noise-strength, indeed a higher number of SBM trajectories are predicted as FBM whereas the model prediction for FBM trajectories remain reasonably well. This asymmetry in the model prediction results is not seen in the superdiffusive case where, with increasing noise-strength, the incorrect attribution of FBM trajectories to SBM and vice-versa increases almost symmetrically, as expected.

Figure 8 shows what effect using a wrong prior on $\alpha$ has on model prediction. We consider the same data sets used in figure 5 where we discussed the effect of a wrong prior on parameter estimation. We observe a systematic decrease in $F_1$ score on using a wrong prior on $\alpha$ for low $N$. However the score itself is very close to the one obtained using the correct prior, and the differences between the two choices of prior vanishes with more data.

3.5. CTRW trajectories

The results from Bayesian inference obviously depend on the list of models considered in the analysis. We highlight this in figure 9 where we present results from the analysis of simulated two-dimensional CTRW trajectories of length $N = 200$, with an added measurement noise of strength $\sigma_{mn} = 0.1$. CTRW has been used to model numerous experiments exhibiting anomalous diffusion [51–54]. The chosen CTRW model is a renewal process with Gaussian jump-lengths with an asymptotic power law waiting time between the jumps [55, 56]. The exact probability density function of the waiting time is a one-sided $\alpha$-stable distribution [55, 56], where $0 \leq \alpha \leq 1$ is also the anomalous diffusion exponent [55, 56]. The CTRW trajectories
Figure 9. Left: MAE vs $\alpha_{GT}$ plot from two-dimensional CTRW trajectories of $N = 200$ points with added measurement noise of strength $\sigma_{mn} = 0.1$ (labelled 'CTRW'). $\tilde{N} = 100$ trajectories were generated for each $\alpha_{GT}$. The estimates from FBM and SBM trajectories (labelled 'FBM' and 'SBM' respectively)—with $\sigma_{mn} = 0.1$ and $N = 200$ points from figures 2 and 3—are presented for comparison. All the MAE estimates are done using equations (14) and (15) with two models, FBM and SBM. Right: model inference using FBM and SBM models on two-dimensional CTRW trajectories generated with added measurement noise of strength $\sigma_{mn} = 0.1$. $\tilde{N} = 100$ trajectories were generated for each $\alpha_{GT}$.

We implemented Bayesian inference for SBM, and tested the procedure in combination with FBM on synthetic data. The results obtained in these tests are an upper limit on how well any inference method can perform, since the procedure is singled out as optimal by Bayes’ theorem.

4. Conclusion

We implemented Bayesian inference for SBM, and tested the procedure in combination with FBM on synthetic data. The results obtained in these tests are an upper limit on how well any inference method can perform, since the procedure is singled out as optimal by Bayes’ theorem.

7 We choose $\Delta x = 0.01$ and $\Delta t = 1$ in the simulations. See reference [57] for details.
and decision theory. In this connection we note that the issue of falsely inferring many SBM trajectories as FBM (see our figure 7) also applies for the top machine learning method in the AnDi competition (see figure 3(e) in [12]). For the computational efficiency of the Bayesian approach it is important that the hidden variables (e.g., measurement noise) can be integrated out analytically, which can be achieved for SBM and FBM. For models where this analytic integration is too difficult, one can instead attempt to integrate the hidden variables numerically or with Monte Carlo simulations. However, if this is too expensive computationally, one might resort to machine learning methods as an alternative for both parameter and model inference [12]. But note that while a Bayesian analysis automatically provides uncertainty estimates for the inferred parameters, this is not the case for neural networks [59]. This is because the result of a Bayesian analysis is a posterior distribution for the parameters, unlike point estimates from a neural network.

Bayesian inference is sometimes criticised for being subjective, because a prior has to be chosen, which specifies the beliefs in the models and parameters prior to any data being revealed. The problem diminishes, as more data is obtained, something we also tested here by comparisons with results from using a wrongly skewed prior on the $\alpha$-parameter. We would like to point out, that machine learning methods suffer from the same issue of having to choose a prior distribution, since distributions of models and parameters have to be chosen for the data that is used for training these methods.

For real experimental data any simple mathematical model will not be an exact match of the underlying process. Thus the methods discussed in this article cannot stand alone, but has to be supplemented by model checking [10] (also called goodness-of-fit tests) to check how well the inferred model describes the observed data, and possibly, if too big discrepancies are found, improved models have to be designed and the inference process reiterated with these models.

**Acknowledgments**

S Thapa acknowledges support in the form of a Sackler postdoctoral fellowship and funding from the Pikovsky-Valazzi matching scholarship, Tel Aviv University. J-H Jeon acknowledges support from the National Research Foundation (NRF) of Korea (No. 2020R1A2C4002490). R Metzler acknowledges funding from the German Science Foundation (DFG, Grant No. ME 1525/12-1) and the Foundation for Polish Science (Fundacja na rzecz Nauki Polskiej, FNR) within an Alexander von Humboldt Honorary Polish Research Scholarship.

**Data availability statement**

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/Samudrajit11/SBM.

**Appendix A. Additional figures (from the analysis of one-dimensional trajectories)**

See figures A1–A3.
Figure A1. MAE vs $\alpha_{GT}$ plot with the estimated MAE obtained from the analysis of $\tilde{N} = 100$ one-dimensional SBM trajectories for each $\alpha_{GT}$. Each sub-plot shows the results from the analysis of SBM trajectories generated with different noise-strengths $\sigma_{mn}$. The error-bars are the standard error on the mean. This figure is equivalent to figure 2, except that it is for one-dimensional trajectories instead of two-dimensional ones.
Figure A2. MAE vs $\alpha_{GT}$ plot with the estimated MAE obtained from the analysis of $\tilde{N} = 100$ one-dimensional FBM trajectories for each $\alpha_{GT}$. Each sub-plot shows the results from the analysis of FBM trajectories generated with different noise-strengths $\sigma_m$. The error-bars are the standard error on the mean. This figure is equivalent to figure 3, except that it is for one-dimensional trajectories instead of two-dimensional ones.
Figure A3. $F_1$ score vs $\alpha_{GT}$ plot from the analysis of one-dimensional FBM and SBM trajectories. The $F_1$ scores at each $\alpha_{GT}$ are from the analysis of $\tilde{N} = 100$ SBM and $\tilde{N} = 100$ FBM trajectories generated with that particular $\alpha_{GT}$. The error-bars are the standard error on the mean for a Bernoulli random variable. This figure is equivalent to figure 6, except that it is for one-dimensional trajectories instead of two-dimensional ones.

ORCID iDs

Samudrajit Thapa https://orcid.org/0000-0002-4340-5900
Seongyu Park https://orcid.org/0000-0002-6478-660X
Jae-Hyung Jeon https://orcid.org/0000-0003-0658-9672
Ralf Metzler https://orcid.org/0000-0002-6013-7020
Michael A Lomholt https://orcid.org/0000-0002-3698-0958
References

[1] Brown R 1828 A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies *Phil. Mag.* 4 161


[38] Skilling J 2006 Nested sampling for general Bayesian computation Bayesian Anal. 1 833–59
[58] Thiel F and Sokolov I M 2014 Scaled Brownian motion as a mean-field model for continuous-time random walks Phys. Rev. E 89 012115