

Bayesian inference with information content model check for Langevin equations

Krog, Jens F. C. ; Lomholt, Michael Andersen

Published in:
Physical Review E

DOI:
[10.1103/PhysRevE.96.062106](https://doi.org/10.1103/PhysRevE.96.062106)

Publication date:
2017

Document version
Accepted manuscript

Citation for pulished version (APA):
Krog, J. F. C., & Lomholt, M. A. (2017). Bayesian inference with information content model check for Langevin equations. *Physical Review E*, 96(6), [062106]. <https://doi.org/10.1103/PhysRevE.96.062106>

Terms of use

This work is brought to you by the University of Southern Denmark through the SDU Research Portal. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:

- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

Bayesian inference with information content model check for Langevin equations

Jens Krog and Michael A. Lomholt

*MEMPHYS - Center for Biomembrane Physics, Department of Physics, Chemistry,
and Pharmacy, University of Southern Denmark, 5230 Odense M, Denmark*

(Dated: November 7, 2017)

The Bayesian data analysis framework has been proven to be a systematic and effective method of parameter inference and model selection for stochastic processes. In this work we introduce an information content model check which may serve as a *goodness-of-fit*, like the χ^2 procedure, to complement conventional Bayesian analysis. We demonstrate this extended Bayesian framework on a system of Langevin equations, where coordinate dependent mobilities and measurement noise hinder the normal mean squared displacement approach.

PACS numbers: 02.50.-r, 05.10.Gg, 05.40.-a

I. INTRODUCTION

Stochastic processes are commonly used for modeling of single particle tracking (SPT) experiments in biophysics. The motion of the particle is assumed to be partly diffusive, but the possibility of effects such as active transport and drift due to flow or some effective potential often cannot be ruled out a priori. Standard tools to distinguish these effects are the estimation of mean square displacement (MSD) or correlations [1, 2]. However, if the environment is heterogeneous then the situation easily becomes too complex for estimators such as the MSD to resolve the mechanisms of the underlying dynamics.

An alternative approach is offered by Bayesian analysis [3–5], which gives a systematic method for comparing different models of the investigated particle dynamics. The approach simultaneously provides a way of selecting the most probable model among different candidates and provides the most probable associated parameter values together with their uncertainties. A Bayesian approach to analysing SPT trajectories has previously been applied to model selection among hidden Markov models [6–9], diffusion in varying potentials [10–12], and parameter estimation for optical traps [13], to name just a few examples.

While the Bayesian approach is excellent at choosing the best among specified candidate models and their parameters it does not provide a measure of how well the specified models actually describe the data, i.e., it lacks *goodness-of-fit* measures like the χ^2 procedure of least-squares fitting.

In this article we demonstrate how Bayesian inference can be applied efficiently on SPT data with models that takes the form of a Langevin equation using the nested sampling algorithm introduced by Skilling [14]. The Langevin equation form allows for position or time dependent forces as well as mobilities. Additionally we supplement the Bayesian approach by coupling it with a goodness-of-fit test that checks whether typical trajectories generated from the models using typical inferred parameters reproduce the information content of the observed trajectories at different time scales. The approach is implemented in Matlab, and the code is available at

GitHub [15].

The article is organised as follows. In Section II we introduce the general Langevin equation on which we apply the Bayesian inference and information content model check as explained in Section III. In Section IV we demonstrate how the approach works on synthetic data with a position dependent mobility and we finally conclude in Section V.

II. GENERIC STOCHASTIC DIFFERENTIAL EQUATIONS

We will assume that a theoretical analysis of the system at hand has resulted in a stochastic differential equation for an observable x on the form of a Langevin equation

$$\frac{dx}{dt} = a(x, t) + b(x, t)\xi(t). \quad (1)$$

Both the deterministic term $a(x, t)$ and the coefficient $b(x, t)$ of the stochastic term should be known functions except for a limited number of unknown parameters. The stochastic nature is supplied through the noise function $\xi(t)$. In the following we will assume that $\xi(t)$ is a *white noise* idealization of the physical noise, such that

$$\langle \xi(t)\xi(t') \rangle = \delta(t - t'), \quad (2)$$

with an Ito interpretation when $b(x, t)$ is not a constant [16]. The characteristics of the process under investigation will then be determined from the other two functions, a and b in Eq. (1). These functions may take on many forms, depending on the system and observable of interest.

An example of a system that can be brought on such a form is a particle under the influence of a force f and with a position dependent mobility $\mu(x)$. The probability distribution $p(x, t)$ for the particle position x at time t obeys a Fokker-Planck equation

$$\partial_t p(x, t) = -\partial_x [\mu(x) f p(x, t) - k_B T \mu(x) \partial_x p(x, t)], \quad (3)$$

where $k_B T$ is Boltzmann's constant times temperature. This equation generalises the usual high friction limit of

Kramers' equation [17] to the situation with position dependent mobility. The order of the second partial derivative ∂_x relative to $\mu(x)$ is fixed by the condition that the Boltzmann distribution should be a stationary solution when the force f is conservative. We have not indicated a dependence of position or time in the force f , since in later sections we will take it to be constant. The Fokker-Planck equation can be converted to the form of the Langevin equation (1) giving for the Ito interpretation of the noise [16]

$$\frac{dx}{dt} = \mu(x)f + k_B T \frac{\partial}{\partial x} \mu(x) + \sqrt{2k_B T \mu(x)} \xi(t), \quad (4)$$

where the coordinate dependent $\mu(x)$ appears in the deterministic as well as the stochastic term. We only study the one dimensional case here, but generalizations of this equation to higher dimensions are straightforward.

If the noise is assumed to be negligible, then Eq. (4) can be used to obtain deterministic equations of motion for x , which can be fitted to obtain the correct form of the deterministic terms. On the other hand, if the mobility and the force are independent of x , then mean-squared-displacement techniques may be employed to unveil characteristics of the stochastic dynamics. The Bayesian inference approach, however, relies on none of these assumptions, and can be used to infer the functional form of $\mu(x)$, and thus the best underlying model, in any case where a Langevin equation on the form (4) can be found.

III. BAYESIAN PARAMETER INFERENCE AND MODEL TESTING

In the following we will refer to a specific functional form of the underlying Langevin equation (1) as a *model*, M . Labeling the different models via a subscript i , model selection through the knowledge of some experimentally observed data \mathbf{x} then follows the logic of Bayes' theorem

$$P(M_i|\mathbf{x}) = \frac{P(\mathbf{x}|M_i)P(M_i)}{P(\mathbf{x})}. \quad (5)$$

Here, the left hand side is called the *posterior* probability for the model M_i given the data, while $P(\mathbf{x}|M_i)$ is the *likelihood* of M_i . $P(M_i)$ is the *prior* probability of the model, and may be set equal for all models to achieve impartial judgement. The denominator in (5) is independent of the model, and may be seen as a normalization constant.

To compare the posterior probabilities of a set of models, one must then simply calculate the likelihoods $P(\mathbf{x}|M_i)$. As each model M_i contains a number of parameters θ , this is found by marginalizing over the unknown parameters,

$$P(\mathbf{x}|M_i) = \int P(\mathbf{x}|\theta, M_i)P(\theta|M_i)d\theta, \quad (6)$$

where $P(\mathbf{x}|\theta, M_i)$ and $P(\theta|M_i)$ are the likelihood function and prior probability, respectively of the parameters for the model M_i . Considering each possible set

of parameters θ as representing a unique model, $Z_i = P(\mathbf{x}|M_i)$ is often referred to as the evidence for that class of models, and used to compare the relative merit of these classes.

Assuming that the positions x_0, x_1, \dots, x_N constituting \mathbf{x} are sampled at closely spaced times t_0, t_1, \dots, t_N such that the force f and mobility μ vary very little during a time interval we can integrate Eq. (4) over each time interval to obtain the likelihood function

$$P(\mathbf{x}|\theta, M_i) = \prod_{j=1}^N \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left[-\frac{(x_j - \bar{x}_j)^2}{2\sigma_j^2} \right]. \quad (7)$$

with the following means and variances for each new position

$$\begin{aligned} \bar{x}_j &\equiv \langle x(t_j) \rangle \\ &= x_{j-1} + [\mu(x_{j-1})f + k_B T \partial_x \mu(x_{j-1})] (t_j - t_{j-1}), \end{aligned} \quad (8)$$

$$\begin{aligned} \sigma_j^2 &\equiv \langle (x_j - \bar{x}_j)^2 \rangle \\ &= 2k_B T \mu(x_{j-1}) (t_j - t_{j-1}). \end{aligned} \quad (9)$$

By defining the priors $P(\theta|M_i)$ the integrand of Eq. (6) is known, and a numerical integration may be performed to obtain the evidence Z_i and posterior probability

$$P(\theta|\mathbf{x}, M_i) = \frac{P(\mathbf{x}|\theta, M_i)P(\theta|M_i)}{Z_i} \quad (10)$$

for the parameters of each model. The evidences, depending on the given likelihoods $P(\mathbf{x}|\theta, M_i)$ and on the user defined priors $P(\theta|M_i)$, constitute a degree of plausibility of the model M_i . An evidence ratio $Z_1/Z_2 = 10$ may be interpreted, assuming impartial priors, as a 91% chance of M_1 being the better model of the two. As long as careful attention is given to the process of prior assignment, this estimate will yield reliable conclusions. One thing to be aware of, though, is situations where one model is an expansion of another simpler model by the addition of a parameter θ that can deviate from the fixed value in the simpler model θ_{norm} . If the prior for θ is very broad in the sense that very little is known about the true value of θ , then the Bayesian analysis may yield the highest evidence for the model with $\theta = \theta_{\text{norm}}$, even though the true value is different from θ_{norm} . This reflects the fact that for very broad priors a small deviation will be evaluated as unimportant within the Bayesian framework, until a sufficient amount of data is accumulated. Finding relatively small effects requires relatively large amounts of data in any analysis. The dependence of model selection on the broadness of a prior is, however, absent in frequentistic approaches. The discrepancy between the two approaches is called Lindley's paradox and has been discussed extensively. See for instance [18] for a discussion in the context of high energy physics.

A. Measurement noise effects

If the subject of investigation is a process where an error is connected to the data collection process itself, it may be necessary to distinguish between this measurement noise and the noise native to the underlying stochastic dynamics.

The presence of measurement noise η shifts the true values \mathbf{x}^{true} , such that we observe

$$x_j = x_j^{\text{true}} + \eta_j, \quad (11)$$

where we assume a zero mean Gaussian measurement noise with $\langle \eta_i \eta_j \rangle = \delta_{i,j} \sigma_{\text{mn}}^2$. This is appropriate for situations where the coordinate is determined by averaging an intensity profile at each measurement.

If the inherent noise of the system does not depend on the coordinate, we can adjust the mean \bar{x}_j and variance σ_j^2 of the true coordinates to incorporate the measurement noise through an iteratively calculated mean \tilde{x}_j and variance $\tilde{\sigma}_j^2$ [19]. In Appendix A we derive (alternatively, see for instance [19]) that for constant mobility μ and force f

$$\tilde{x}_j = \bar{x}_j - \frac{\sigma_{\text{mn}}^2}{\tilde{\sigma}_{j-1}^2} (x_{j-1} - \tilde{x}_{j-1}) \quad (12)$$

$$\tilde{\sigma}_j^2 = \sigma_j^2 + \sigma_{\text{mn}}^2 \left(2 - \frac{\sigma_{\text{mn}}^2}{\tilde{\sigma}_{j-1}^2} \right), \quad (13)$$

with the starting values $\tilde{x}_1 = \bar{x}_1$ and $\tilde{\sigma}_1^2 = \sigma_1^2 + 2\sigma_{\text{mn}}^2$. The likelihood is then computed by replacing \bar{x}_j and σ_j in Eq. (7) with \tilde{x}_j and $\tilde{\sigma}_j$.

If μ and f are not independent of position x , then the evaluation of the the mean and variance given in Eqs. (8) and (9) is in principle not possible, since μ and f depend on the true positions. However, if μ and f varies only slowly on the length scale σ_{mn} associated with the measurement noise, then we can, to a good approximation, use the observed positions when evaluating Eqs. (8) and (9). In the following we will assume this weak dependence of μ and f on position and evaluate them in this way.

We remark that we model the noise as independent between measurements of the position. However, there is a number of effects that can lead to correlations in the measurement noise. For instance, a cameras finite exposure time gives rise to scalded motion blur [20]. Including such effects in the modelling would of course be beneficial if they are expected to be substantial.

B. Information and model testing

So far we have only discussed how the Bayesian setup generates relative degrees of plausibility of different models, but not touched upon the problem of determining whether or not the set of models available actually can match the data well. A way to test the latter is to compare properties of the data with replicated data from the

models. The nested sampling framework offers a readily accessible property: the *Shannon information* [21]. For a model M we can quantify the information that some observation \mathbf{x} gives us about the parameters of the model as

$$\mathcal{H} = \int P(\boldsymbol{\theta}|\mathbf{x}, M) \ln \frac{P(\boldsymbol{\theta}|\mathbf{x}, M)}{P(\boldsymbol{\theta}|M)} d\boldsymbol{\theta}. \quad (14)$$

Using Eq. (10) the corresponding expression takes the form [22]:

$$\mathcal{H} = \frac{1}{Z(M)} \int P(\mathbf{x}|\boldsymbol{\theta}, M) P(\boldsymbol{\theta}|M) \ln \left(\frac{P(\mathbf{x}|\boldsymbol{\theta}, M)}{Z(M)} \right) d\boldsymbol{\theta}. \quad (15)$$

This can also be read as the posterior average of minus the information content in the observation, with the factor $Z(M)$ included to make the argument of the logarithm unitless.

In order to evaluate the merit of a model, we use the fact that the information content should be similar to typical values for replicated observations generated from the model with parameters taken from the posterior. We thus have a measure to be used for comparison. By drawing a sample of parameters $\boldsymbol{\theta}^*$ from the posterior distribution

$$P(\boldsymbol{\theta}^*|M, \mathbf{x}) = \frac{P(\boldsymbol{\theta}^*, \mathbf{x}|M)}{P(\mathbf{x}|M)} = \frac{P(\mathbf{x}|\boldsymbol{\theta}^*, M)}{P(\mathbf{x}|M)} P(\boldsymbol{\theta}^*|M). \quad (16)$$

and using these to generate a replicated trajectory \mathbf{x}^* we can calculate

$$h(\mathbf{x}^*, \boldsymbol{\theta}^*) = \ln \left(\frac{P(\mathbf{x}^*|\boldsymbol{\theta}^*, M)}{Z(M)} \right), \quad (17)$$

where $Z(M) = P(\mathbf{x}|M)$. Generating several pairs of $\boldsymbol{\theta}^*$ and \mathbf{x}^* we obtain a cluster of h -values. If the models describe the original data \mathbf{x} well with respect to information content, then \mathcal{H} should be somewhere within this cluster of h -values for replicated data. We demonstrate this with examples in Section IV.

As an alternative way of representing the check of whether the information content of the original data is typical with respect to the posterior of the Bayesian inference we calculate, by averaging over many draws of $\boldsymbol{\theta}^*$ and corresponding \mathbf{x}^* , the posterior predictive p value [23]

$$p = \langle \Theta[h(\mathbf{x}^*, \boldsymbol{\theta}^*) - h(\mathbf{x}, \boldsymbol{\theta}^*)] \rangle, \quad (18)$$

where $\Theta(x)$ is the Heaviside function. This p value is the probability of replicated data being more extreme with respect to information content than the observation \mathbf{x} . If the original data x is well described by the model, then the p value should not be extreme, i.e., improbably close to zero or one. In particular, if the model contains the true underlying model as a special case and the posterior probability for the parameters is narrowly distributed around their true values, then the above p value should be uniformly distributed between zero and one.

However, since the posterior probability distribution is obtained based on the same observational data that is used for the calculation of the p value, then the distribution of p values will tend to concentrate more towards a value of one half [23]. In any case we conclude that the model describes the data well with respect to information content if the p value is not extremely close to zero or one. The information measure regarded here is convenient because it is readily available from the calculations used for the Bayesian analysis. For a more involved procedure addressing a uniformly distributed measure see [24]. Note that this analysis does not require the comparison between different models, but yields a check of the absolute quality of a model in contrast to the relative one yielded by the evidence ratios. If an observed data set is not typical for the model with respect to information content, then the model is likely to miss a key aspect of the underlying process, even though this model has yielded the highest evidence of the ones tested.

Since the information content of the data can be quite closely related to the evidence of the model, then the model check as described above might not be a very stringent test. We suspect that this is the reason why the authors have not been able to find this particular model check described elsewhere. To improve on the stringency of the model check, we additionally rescale the data by only considering every n 'th data point in the series and in turn generate replicated tracks with correspondingly longer but fewer time steps. The information content of a reduced track \mathbf{x}_n^* for parameters θ^* is then calculated as

$$h(\mathbf{x}_n^*, \theta^*) = \ln \left(\frac{P(\mathbf{x}_n^* | \theta^*, M)}{Z_n(M)} \right), \quad (19)$$

where $Z_n(M) = \int P(\mathbf{x}_n | \theta, M) P(\theta | M) d\theta$ with \mathbf{x}_n being the reduced original data. The p value is then found as before, and the posterior average becomes

$$\mathcal{H}_n = \frac{1}{Z(M)} \int P(\mathbf{x} | \theta, M) P(\theta | M) \ln \left(\frac{P(\mathbf{x}_n | \theta, M)}{Z_n(M)} \right) d\theta. \quad (20)$$

Obtaining these quantities for a number of different n provides a more stringent test that models will fail if their probability distribution do not scale in a proper way with step size compared with the distribution ingrained in the experimentally observed steps. We demonstrate this later, for instance for the case of overlooked measurement noise displayed in Fig. 7. To make the above reasoning more precise we have detailed it for a case of diffusive Gaussian models in Appendix B.

We have framed the discussion in this subsection in the language of information theory, since our inspiration for choosing this particular model check has come from the concept of ‘‘typical set’’ [3]: in essence the model check tests whether the observed data belongs to the typical set of the inferred model. Another way to phrase this is, that the check examines whether the observed data belongs to the very many possible trajectories of the model that does not have an unlikely high or small value for

their probability density, i.e., the model fails the check if the observed trajectory is an extreme case of high or low probability density relative to typical trajectories produced by the model.

We remark that there is a number of approaches to model selection involving information criteria, which rank models based on their predictive accuracy as measured by information content of the data [23, 25]. Sometimes in these approaches the models are optimized on part of the data and then cross-validated on another part. This is different from the approach in this article, where the full data-set is used for model selection and parameter fitting, and then also used for model checking.

IV. IMPLEMENTATION AND EXAMPLES

To demonstrate the Bayesian inference method and the information content model check, we will examine artificial data samples generated by the Langevin equation on the form shown in Eq. (4), with the mobility,

$$\mu(x) = D_0 |x|^\alpha, \quad (21)$$

similar to what has been theoretically investigated in [26]. D_0 has dimension of $\text{length}^{1-\alpha} \text{time}^{-1} \text{force}^{-1}$, and the force f as well as the exponent α will be assumed to be either zero or a nonzero constant. By letting the force be constant while the mobility varies we essentially analyse a situation complementary to the setup investigated in [11], where the force was assumed to be a polynomial in x . We stress that any set of functions $\mu(x), f(x)$ could have been chosen, as long as $\mu(x)$ was differentiable. By choosing values for D_0, α, f , and σ_{mn}^2 as well as a starting point x_0 , we use (12) and (13) to generate the steps iteratively. To specify the assumptions of the models we consider we use the subscripts ‘‘pull’’ and ‘‘free’’ to signify whether or not the model assumes a nonzero force f , while ‘‘mn’’ and ‘‘clean’’ specifies whether or not the model takes into account measurement noise. Throughout all examples we set $k_B T = 1$.

To calculate the respective evidences $Z_i = P(\mathbf{x} | M_i)$, as given in Eq. (6), we specify the use of a Jeffreys prior for $D_0 \in [10^{-4}; 10^2]$, such that

$$P(D_0 | M_i) = \frac{1}{D_0 \ln 10^6} \Theta \left(3 - \left| \log_{10} \left(\frac{D_0}{10^{-1}} \right) \right| \right), \quad (22)$$

allocating the same amount of probability within each decade, while assigning the uniform priors

$$P(\alpha | M_i) = \frac{1}{4} \Theta(2 - |\alpha|), \quad (23)$$

$$P(f | M_{\text{pull}}) = \frac{1}{2} \Theta(1 - |f|), \quad (24)$$

$$P(\sigma_{\text{mn}}^2 | M_{\text{mn}}) = \frac{1}{100} \Theta(100 - \sigma_{\text{mn}}^2). \quad (25)$$

We choose to use the Jeffreys prior for parameters where the correct order of magnitude might not be known beforehand, and uniform priors for parameters that may vanish or be negative, or where the order of magnitude

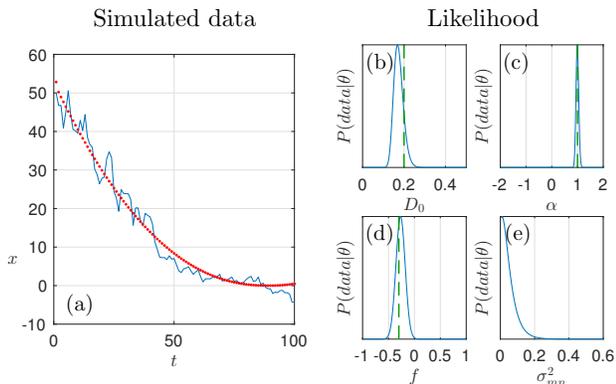


FIG. 1. (a) Sample track along with a deterministic fit. (b) - (e) Likelihoods for variations of single parameters (right) around the maximum likelihood point for the best candidate model, which in this case is $M_{\text{pull, clean}}$. The dashed lines represent the true values.

may be estimated from the experimental background information. If the amount of data is not too small then the specific choice of prior should have negligible influence on the inference results for the parameter values as long as the prior is nonzero in the peak of the likelihood. As discussed earlier, then the specific choice of prior can influence the result for model selection (Lindley's paradox). Thus care must be taken to make the prior reflect the a priori expectations for the parameters in the specific experiment. For a suggestion of a non-informative approach towards this using reference priors see [27].

The integral in Eq. (6) is evaluated numerically with the nested sampling algorithm [28], in which stochastic sampling of the parameter space is performed via a Markov chain Monte Carlo (MCMC) technique. From an initial uniform distribution of sample points within parameter space, the sample with the lowest likelihood is progressively removed and replaced by a new sample within the remaining volume of parameter space with higher likelihood. The MCMC technique generates this new sample and ensure that it is independent of the remaining samples and as such is once again uniformly distributed. Through such sampling, inferred parameter values are automatically included through the estimated posterior probability landscape. The inferred posterior distribution can then be used to generate the samples needed for the information content estimates in Eq. (17).

A. Noise free data

In the left side of Fig. 1 an artificial sample track is displayed that we will apply the Bayesian inference method on. For now we consider data with no measurement noise, so we focus on two models, $M_{\text{free, clean}}$ and $M_{\text{pull, clean}}$, where both D_0 and α are free parameters, while $\sigma_{mn}^2 = 0$, but where only the latter model allows a nonzero force.

Integrating over all of parameter space, we find

$$\ln Z_{\text{pull, clean}} - \ln Z_{\text{free, clean}} = 2.42 \pm 0.15. \quad (26)$$

Thus $Z_{\text{pull, clean}}/Z_{\text{free, clean}} \approx 11.3$, which strongly favors the model with a nonzero force. We obtain uncertainty for the ratio of the evidences via the inferred information [14]. For $M_{\text{pull, clean}}$ we obtain the inferred values shown in Table I.

Parameter	Inferred value	True value	Fitted value
D_0	0.174 ± 0.039	0.2	N/A
α	1.001 ± 0.091	1.0	0.553 ± 0.050
f	-0.277 ± 0.097	-0.3	N/A
$D_0 \times f$	-0.048 ± 0.022	-0.06	-0.148 ± 0.012

TABLE I. Bayesian-inferred, true and naively fitted parameters for the data in Fig. 1.

As expected the parameter values are inferred with reasonable uncertainty. The plots in the right side of Fig. 1 show the development of the likelihood function as the parameters are varied one at a time, showing a slightly skewed distribution for D_0 as it is restricted to be positive. This illustrates how Bayesian inference describes the entire parameter space in terms of the associated likelihood of the data. In addition, by examining the likelihood as a function of the noise parameter we see how the best model correctly is the one without noise.

For comparison, we also estimate the parameters in this example via simple fitting, by removing all (physical) noise terms, i.e. second derivative terms in the Fokker-Planck equation (3) from which we then recover a deterministic differential equation¹:

$$\frac{dx_{det}}{dt} = \mu(x)f = D_0 f |x_{det}|^\alpha, \quad (27)$$

where $x_{det}(t)$ is uniquely determined by the initial condition. Solving this equation then leads to an expression for $x_{det}(t)$ which is fitted to the data in Fig. 1 via the native `fit` function in Matlab, showing the best fit in red dots. Note that only the product of D_0 and f appears in the deterministic expression (27), such that neither may be fitted individually. As is evident from Table I, this naive fitting procedure fails to estimate the parameters correctly although the parameter space has been decreased. This is not surprising, as the simple fitting procedure considers the noise to be independently and identically distributed. In reality, the positions are highly correlated in the fluctuating time series considered here. As the deterministic approach does not include any knowledge about the fluctuations of the measured quantities, regular χ^2 testing cannot be employed as a model

¹ As the noise term is coordinate dependent, the mean squared displacement method of estimating noise cannot be employed.

check. The ability to deal with probabilities for fluctuating data rather than just an averaged curve enables the Bayesian approach to estimate the noise and hence the mobility μ independently of the force².

Taking the data from Fig. 1 and the corresponding inference, we calculate the information for changed timescales by considering only the data points x_{1+in} for integer i . We draw parameters to generate new tracks via Eq. (16) with $M_{\text{pull, clean}}$. The generation of one hundred tracks for each scaling and the associated measures of information $h(\mathbf{x}_n, \boldsymbol{\theta}^*)$ are shown in Fig. 2 along with the mean of the information for the model as given in Eq. (20). The value of $Z_n(M)$ is estimated with prior samples from the nested sampling run on the non-scaled original data. Any error in this estimate will only cause a shift in the collective information content estimates, leaving all comparisons and p values unaffected.

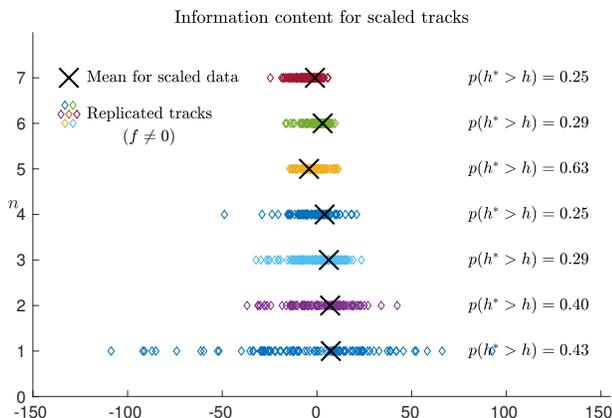


FIG. 2. The inferred mean information \mathcal{H}_n for the data set in Fig. 1 assuming the model $M_{\text{pull, clean}}$ along with the information $h^* = h(\mathbf{x}_n^*, \boldsymbol{\theta}^*)$ for replicated tracks from the same model and corresponding p values.

Evidently, the information measures for the replicated tracks are distributed around the information of the original data such that we conclude that the original data set is within the “typical set” for the model with respect to information content.

On the other hand, if only the model with no force (M_{free}) is used to perform the parameter inference and replicated track generation, the picture in Fig. 3 appears. When the force is set to zero, the replicated data for the model deviate considerably from the original data in terms of the information. Note that this is the case even though the original analysis has located the best possible

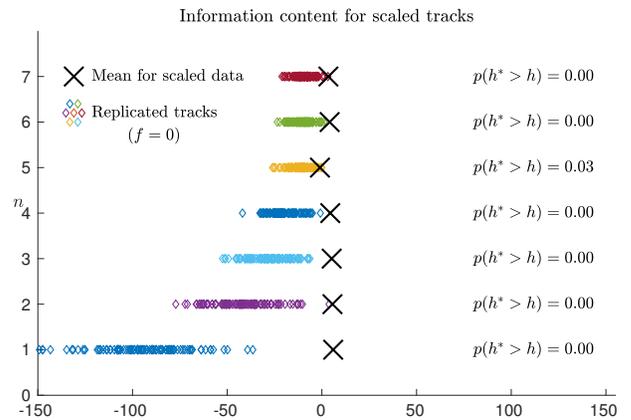


FIG. 3. Information measures as in Fig. 2 for the data set in Fig. 1 assuming the model M_{free} .

available parameters for the model. This warns us that a critical aspect of the process has been overlooked, namely the inclusion of the force.

Since the force parameter adds a drift to the system, the lack of such a parameter might have been easily detectable from visual inspection of the data alone. Conversely, a significant measurement noise might not be detected by simple inspection. If this effect is neglected in the analysis however, the comparison of the information of the original data versus that of replicated data will reveal the flaw in the model.

B. Noisy data

The generation of noisy data is simply done by generating clean data as above and then adding noise with the appropriate variance afterwards. In Fig. 4 an artificial data set contaminated with random noise is shown alongside slices of the likelihood function around the inferred maximum likelihood parameters.

By reusing the priors from the previous example to compare the calculated evidences for the models with and without measurement noise ($M_{\text{free, clean}}, M_{\text{free, mn}}$), we find

$$\ln Z_{\text{free, mn}} - \ln Z_{\text{free, clean}} = 10.62 \pm 0.14, \quad (28)$$

such that $Z_{\text{free, mn}}/Z_{\text{free, clean}} \sim 4.1 \times 10^4$, firmly establishing superiority of the noise model. The inferred parameters for the model are shown in Table II.

Parameter	Inferred value	True value
D_0	0.21 ± 0.22	0.2
α	0.94 ± 0.49	1.0
σ_{mn}^2	2.49 ± 0.61	2.0

TABLE II. Inferred and true parameters for the data in Fig. 4

² The shortcomings for the fitting procedure may of course be remedied by more sophisticated methods, such as a stepwise fitting procedure with changing variance as the one employed in [29], which, however, requires many more data points. Such techniques take a step towards the Bayesian approach by analysing the maximum likelihood parameters, while not exploring their distribution.

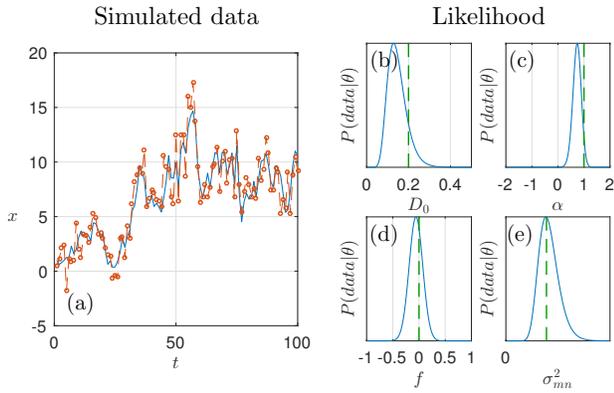


FIG. 4. (a) A clean sample track in solid blue along with the same track with measurement noise pollution in orange circles. (b) - (e) Likelihoods for variations of single parameters around the maximum likelihood point for the best candidate model ($M_{\text{free,mn}}$). The true values are marked by the green dashes. Note that the likelihood curves are one dimensional slices of a four dimensional function.

The inferred diffusion coefficient has a large uncertainty which is expected in the presence of large measurement noise. At this level, however, the true value is still well within the error. Note that the likelihood function for D_0 is shown at a specific (the maximum likelihood) point for the remaining parameters, which is why it does not correspond at a one-to-one basis to the estimate given in Table II. We display the likelihood evolution under variation of the force parameter in order to visualize the likelihood landscape around this point, even though the models with zero force yield higher evidences. As expected, the force is inferred to be close to zero. Evidently the Bayesian method can detect and correctly estimate measurement noise in this example.

As discussed at the end of Subsection III A, an important caveat is that the coordinate dependent mobility conflicts with the assumptions for the noise inference. If the noise is too large, then the mobility at the true position may be poorly estimated by the mobility at the measured position, reducing reliability of the analysis. On the contrary, if the measurement noise is too small to influence the data significantly, the noise will likely not be discovered in smaller data sets.

To demonstrate possible inaccuracies of the first kind, we infect the data in Fig. 1 with a very large measurement noise ($\sigma_{\text{mn}}^2 = 40$ in native units) and display the inference results in Fig. 5. The large noise hinders the inference of the diffusion constant, while still correctly enables the model comparison (with $Z_{\text{pull,mn}}/Z_{\text{pull, clean}} \sim 10^{18}$) and inference of the noise parameter.

From our experience, the largest inherent inaccuracy we encounter is the statistics itself, i.e., the amount of data points. In addition, if the coordinate does not change by more than an order or magnitude, the exponent α is hard to determine, and therefore all other parameters are inaccurately estimated. This error can be

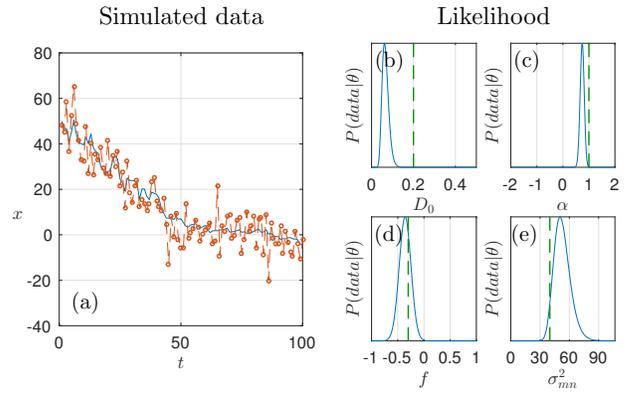


FIG. 5. (a) The data from Fig. 1(a) is reproduced in solid blue while the orange circles represent data with a large added measurement noise ($\sigma_{\text{mn}}^2 = 40$). (b) - (e) Likelihoods for variations of single parameters around the maximum likelihood point for the best candidate model ($M_{\text{pull,mn}}$) are shown, displaying inaccuracy caused by the large measurement noise.

amended, if the system is predicted to have a particular value for α .

Once again, we compare the inferred information to that of replicated tracks generated from parameters drawn via Eq. (16) for the model with noise. The results are shown in Fig. 6. The large measurement error

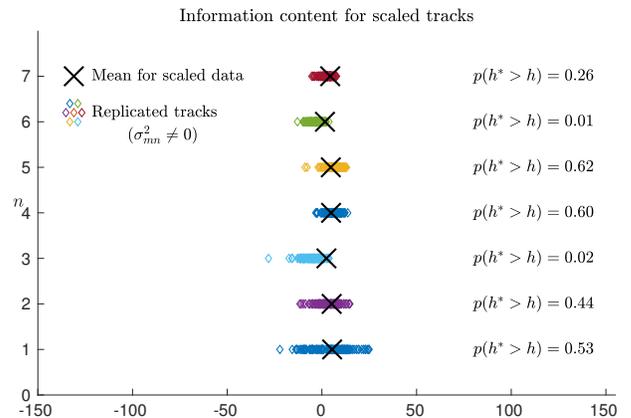


FIG. 6. Information measures as in Fig. 2 for the data set in Fig. 5 assuming the model $M_{\text{pull,mn}}$.

effects dominate this data set, and the analysis deems the data set to be typical for the model, providing a check of whether critical aspects have been overlooked.

In contrast, carrying out the analysis without a measurement error, the information comparison yields the results shown in Fig. 7. In this case, where the measurement error has been overlooked, the replicated data yield different amounts of information than the actual data. Note once again, that this is despite the fact that the Bayesian inference has found the most likely values for the parameters to use for the model. Even though the measurement error in this case would neither pro-

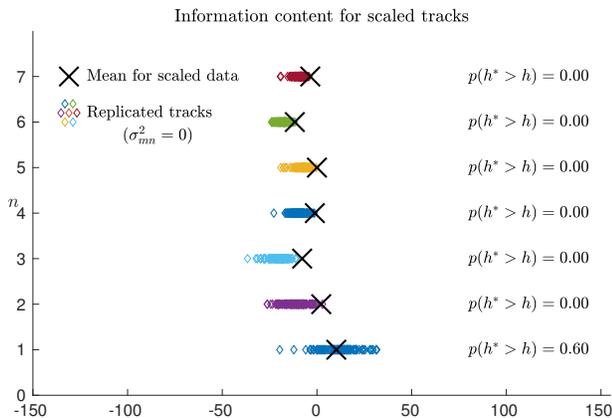


FIG. 7. Information measures as in Fig. 2 for the data set in Fig. 5 assuming the model $M_{\text{pull,clean}}$.

duce a mean drift or a relatively substantial contribution to the mean squared displacement at large times, it is revealed by the information comparison and determined by the Bayesian inference. Note also that the information content model check for the unscaled data does not reveal that a substantial measurement error has been overlooked. This exemplifies that the undersampling of data can be crucial for the information content model check as discussed further in Appendix B.

V. CONCLUSIONS

We have demonstrated that Bayesian inference offers a straightforward method of analysis to systems where forces and/or mobilities are coordinate dependent while also taking into account the stochastic nature of the dynamics. This was found to be contrasted with a simple analysis where deterministic equations of motion were fitted to a trajectory yielding parameter values off the underlying true values. While also being able to compare the relative merit of several models, the Bayesian analysis provides estimates of the parameters for each model via their posterior probability distribution that are consistent with the underlying true values.

One limitation of a strict Bayesian approach is that it does not necessarily reveal serious shortcomings of the candidate models, since it only supplies their relative merits. Here, we found that a combination with a frequentistic approach turns out to be useful. The information content model check combined with undersampling was shown to be able to detect such shortcomings of models even in the case where the model was the best “fit”. This includes “obvious” parameters such as an extra force as well as measurement error parameters which are harder to discover by visual inspection. The test will not capture all kinds of shortcomings though. Thus it could be beneficial to supplement it with further model checks. In particular ones that match key properties of the specific system at hand.

We expect the Bayesian inference coupled with the information content model check approach presented in this article to have many applications within the field of single particle tracking.

ACKNOWLEDGEMENTS

The authors acknowledge support from the Danish Council for Independent Research - Natural Sciences (FNU), grant number 4002-00428B. We thank Tobias Ambjörnsson and Christoffer Lagerholm for stimulating discussions.

Appendix A: Measurement noise inference

The conclusions of this section have been drawn in [19]. We rederive the results here in order to align the notation. We assume that the actual coordinate x_i^{true} is polluted by noise η_i to produce the observed coordinates

$$\begin{aligned} x_i &= x_i^{\text{true}} + \eta_i \\ &= x_{i-1}^{\text{true}} + \Delta x_i^{\text{true}} + \eta_i \\ &= x_{i-1} - \eta_{i-1} + \Delta x_i^{\text{true}} + \eta_i. \end{aligned} \quad (\text{A1})$$

where the η_i are independent and Gaussianly distributed with zero mean and variance σ_{mn}^2 and $\Delta x_i^{\text{true}} = x_i^{\text{true}} - x_{i-1}^{\text{true}}$.

For the displacement Δx_i^{true} we assume that the average $E(\Delta x_i^{\text{true}}) = \Delta \bar{x}$ and variance $\text{Var}(\Delta x_i^{\text{true}}) = \sigma_d^2$ are independent of the coordinates. Thus after the first step their conditional covariances are

$$\begin{aligned} \text{Cov}(x_1, x_1 | x_0) &= \sigma_d^2 + 2\sigma_{\text{mn}}^2 \\ \text{Cov}(x_1, \eta_1 | x_0) &= \sigma_{\text{mn}}^2 \\ \text{Cov}(\eta_1, \eta_1 | x_0) &= \sigma_{\text{mn}}^2, \end{aligned} \quad (\text{A2})$$

and

$$P\left(\begin{bmatrix} \eta_1 \\ x_1 \end{bmatrix} | x_0\right) \sim N\left(\begin{bmatrix} 0 \\ x_0 + \Delta \bar{x} \end{bmatrix}, \begin{bmatrix} \sigma_{\text{mn}}^2 & \sigma_{\text{mn}}^2 \\ \sigma_{\text{mn}}^2 & \sigma_d^2 + 2\sigma_{\text{mn}}^2 \end{bmatrix}\right). \quad (\text{A3})$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ represents a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Gamma}$. Now assume that we know \tilde{x}_i and $\tilde{\sigma}_i^2$ such that

$$P\left(\begin{bmatrix} \eta_i \\ x_i \end{bmatrix} | \mathbf{x}_{i-1}\right) \sim N\left(\begin{bmatrix} 0 \\ \tilde{x}_i \end{bmatrix}, \begin{bmatrix} \sigma_{\text{mn}}^2 & \sigma_{\text{mn}}^2 \\ \sigma_{\text{mn}}^2 & \tilde{\sigma}_i^2 \end{bmatrix}\right). \quad (\text{A4})$$

where \mathbf{x}_{i-1} is a vector containing as elements x_0, x_1, \dots, x_{i-1} . To iterate we then need to find \tilde{x}_{i+1} and $\tilde{\sigma}_{i+1}^2$. We can infer $P(\eta_i | \mathbf{x}_i)$ via

$$P(\eta_i | \mathbf{x}_i) = \frac{P(\eta_i, x_i | \mathbf{x}_{i-1})}{P(x_i | \mathbf{x}_{i-1})}. \quad (\text{A5})$$

where

$$P(x_i | \mathbf{x}_{i-1}) = \frac{1}{\sqrt{2\pi\tilde{\sigma}_i^2}} \exp\left(-\frac{1}{2\tilde{\sigma}_i^2}(x_i - \tilde{x}_i)^2\right). \quad (\text{A6})$$

Combining Eqs. (A4), (A5) and (A6), one obtains the Gaussian probability density function $P(\eta_i|\mathbf{x}_i)$, from which we read off the conditional mean and variance for η_i

$$E(\eta_i|\mathbf{x}_i) = \frac{\sigma_{\text{mn}}^2}{\tilde{\sigma}_i^2}(x_i - \tilde{x}_i) \quad (\text{A7})$$

$$\text{Var}(\eta_i|\mathbf{x}_i) = \sigma_{\text{mn}}^2 \left(1 - \frac{\sigma_{\text{mn}}^2}{\tilde{\sigma}_i^2}\right) \quad (\text{A8})$$

to be used with Eq. (A1) to obtain

$$\tilde{x}_{j+1} = E(x_{j+1}|\mathbf{x}_j) = x_j + \Delta\bar{x} - \frac{\sigma_{\text{mn}}^2}{\tilde{\sigma}_j^2}(x_j - \tilde{x}_j), \quad (\text{A9})$$

$$\tilde{\sigma}_{j+1}^2 = \text{Var}(x_{j+1}|\mathbf{x}_j) = \sigma_d^2 + \sigma_{\text{mn}}^2 \left(2 - \frac{\sigma_{\text{mn}}^2}{\tilde{\sigma}_j^2}\right), \quad (\text{A10})$$

with the initial values given in Eq. (A3).

Appendix B: Information for scaled tracks

In this appendix we will demonstrate that diffusive Gaussian models will pass the information content model check if no scaling of steps is performed. To do this we begin by looking at the logarithm of a Gaussian likelihood

$$\begin{aligned} h(\mathbf{x}^*, \boldsymbol{\theta}) &= \ln P(\mathbf{x}^*|\boldsymbol{\theta}, M) - \ln Z(M) \\ &= -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln \det \boldsymbol{\Gamma} - \ln Z(M) \\ &\quad - \frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}). \end{aligned} \quad (\text{B1})$$

Here N is the size of the replicated observations \mathbf{x}^* , $\boldsymbol{\Gamma}$ is the model covariance, $\boldsymbol{\mu}$ the mean and T denotes transposition. We would like to compare the typical $h(\mathbf{x}^*, \boldsymbol{\theta})$ with the corresponding values for the real observations $h(\mathbf{x}, \boldsymbol{\theta})$. We will do this by comparing their averages over the random draws of \mathbf{x}^* and $\boldsymbol{\theta}$, i.e., comparing

$$\mathcal{H}^* = \int h(\mathbf{x}^*, \boldsymbol{\theta}) P(\mathbf{x}^*|\boldsymbol{\theta}, M) P(\boldsymbol{\theta}|\mathbf{x}, M) d\mathbf{x}^* d\boldsymbol{\theta} \quad (\text{B2})$$

with

$$\begin{aligned} \mathcal{H} &= \int h(\mathbf{x}, \boldsymbol{\theta}) P(\mathbf{x}^*|\boldsymbol{\theta}, M) P(\boldsymbol{\theta}|\mathbf{x}, M) d\mathbf{x}^* d\boldsymbol{\theta} \\ &= \int h(\mathbf{x}, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{x}, M) d\boldsymbol{\theta}. \end{aligned} \quad (\text{B3})$$

Since a difference between \mathcal{H}^* and \mathcal{H} can only arise from the last term in Eq. (B1) we will define

$$\tilde{h}(\mathbf{x}^*, \boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}) \quad (\text{B4})$$

and compare the corresponding $\tilde{\mathcal{H}}^*$ and $\tilde{\mathcal{H}}$ defined similarly to Eqs. (B2) and (B3).

$\tilde{\mathcal{H}}^*$ is straightforwardly evaluated. If we average $\tilde{h}(\mathbf{x}^*, \boldsymbol{\theta})$ over \mathbf{x}^* , indicating this average over the distribution $P(\mathbf{x}^*|\boldsymbol{\theta}, M)$ by $\langle \dots \rangle_{\mathbf{x}^*}$, we get

$$\begin{aligned} \langle (\mathbf{x}^* - \boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}) \rangle_{\mathbf{x}^*} &= \\ \text{Tr} [\boldsymbol{\Gamma}^{-1} \langle (\mathbf{x}^* - \boldsymbol{\mu})(\mathbf{x}^* - \boldsymbol{\mu})^T \rangle_{\mathbf{x}^*}] &= \text{Tr} [\boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}] = N, \end{aligned} \quad (\text{B5})$$

since $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ is the mean and covariance of the components of \mathbf{x}^* . Thus we have that $\tilde{\mathcal{H}}^* = -N/2$.

To evaluate $\tilde{\mathcal{H}}$ we need to specify some of the parameter dependence in the model. Here we will assume that all the elements of the covariance matrix scale in the same way with a single parameter, e.g., a diffusion constant. Thus we can write $\boldsymbol{\Gamma} = \lambda \boldsymbol{\Gamma}_1$ where λ represents the scaling with the parameter and $\boldsymbol{\Gamma}_1$ is independent of λ . Furthermore, we will assume a Jeffreys prior for λ , i.e., that over a broad range of λ we have $P(\boldsymbol{\theta}|M) = \lambda^{-1} f(\hat{\boldsymbol{\theta}})$, where f is some function of the remaining parameters $\boldsymbol{\theta}$. The range over which this is true is assumed to be much broader than the peak of the posterior for λ . With these assumptions we have $\boldsymbol{\Gamma}^{-1} = \lambda^{-1} \boldsymbol{\Gamma}_1^{-1}$ and we can rewrite $\tilde{\mathcal{H}}$ conveniently as

$$\begin{aligned} \tilde{\mathcal{H}} &= \int \tilde{h}(\mathbf{x}, \boldsymbol{\theta}) \frac{P(\mathbf{x}|\boldsymbol{\theta}, M) P(\boldsymbol{\theta}|M)}{Z(M)} d\boldsymbol{\theta} \\ &= Z(M)^{-1} \int \lambda^{-1} f(\hat{\boldsymbol{\theta}}) [(2\pi)^N \lambda^N \det \boldsymbol{\Gamma}_1]^{-1/2} \\ &\quad \times (-\lambda) \frac{d}{d\lambda} \exp \left[-\frac{1}{2\lambda} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Gamma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] d\lambda d\hat{\boldsymbol{\theta}} \end{aligned} \quad (\text{B6})$$

Doing a partial integration, using the assumption of broadness of the prior λ -range relative to the posterior to discard the boundary terms, we get

$$\begin{aligned} \tilde{\mathcal{H}} &= Z(M)^{-1} \int \left(-\frac{N}{2} \right) \lambda^{-1} f(\hat{\boldsymbol{\theta}}) [(2\pi)^N \lambda^N \det \boldsymbol{\Gamma}_1]^{-1/2} \\ &\quad \times \exp \left[-\frac{1}{2\lambda} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Gamma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] d\lambda d\hat{\boldsymbol{\theta}} \end{aligned} \quad (\text{B7})$$

If we take the factor $-N/2$ out of the integral, then the remaining integral becomes exactly the evidence $Z(M)$. Thus we have that $\tilde{\mathcal{H}} = -N/2$ just like for $\tilde{\mathcal{H}}^*$.

We have thus found that on average the replicated observations \mathbf{x}^* will have the same information content as the real observations for this class of Gaussian models. We therefore extend the test by scaling the time steps, i.e., omitting data points. To see why this can make a diffusive Gaussian model fail, we can look at deterministic motion with constant velocity, i.e., steps given by $\Delta x_i = v \Delta t$, where v is the velocity and Δt the duration of the steps. If we apply the above test with Brownian motion without drift as the model then we would get

$$\tilde{h}(\mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^N \frac{\Delta x_i^2}{2D \Delta t} = -\frac{1}{2} N \frac{v^2 \Delta t}{2D} \quad (\text{B8})$$

where D is the diffusion constant of the Brownian diffusion model. As the derivation above shows, then the Bayesian analysis will provide a posterior for D that is peaked around $D = v^2 \Delta t / 2$, such that $\tilde{h}(\mathbf{x}, \boldsymbol{\theta})$ has typical values around $-N/2$. However, if one now only keeps every n 'th data point, then one will have a step duration of $\Delta t' = n \Delta t$ with a total of $N' = N/n$ steps. This will give values of $\tilde{h}(\mathbf{x}_n, \boldsymbol{\theta})$ for the remaining data points \mathbf{x}_n that are unchanged in this case, since $N' \Delta t' = N \Delta t$.

But the values of $\tilde{h}(\mathbf{x}_n^*, \boldsymbol{\theta})$ will on average be reduced by n since $\tilde{H}_n^* = -N'/2$ for the reduced replicated data sets \mathbf{x}_n^* (this follows from a calculation similar to Eq. (B5)). Thus the predicted scaling of the step deviation does not match the scaling inherent in the observations, and therefore Brownian motion without drift will fail the model check when the scaling of the data is included in the check.

-
- [1] H. Qian, M. P. Sheetz, and E. L. Elson, *Biophys. J.* **60**, 910 (1991).
- [2] Y. Meroz and I. M. Sokolov, *Phys. Rep.* **573**, 1 (2015).
- [3] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003).
- [4] E. T. Jaynes and G. L. Bretthorst, *Probability Theory - The Logic of Science* (Cambridge University Press, 2003).
- [5] D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, 2006).
- [6] R. Das, C. W. Cairo, and D. Coombs, *PLoS Comput. Biol.* **5**, e1000556 (2009).
- [7] F. Persson, M. Lindén, C. Unoson, and J. Elf, *Nat. Meth.* **10**, 265 (2013).
- [8] N. Monnier, Z. Barry, H. Y. Park, K.-C. Su, Z. Katz, B. P. English, A. Dey, K. Pan, I. M. Cheeseman, R. H. Singer, and M. Bathe, *Nat. Meth.* **12**, 838 (2015).
- [9] I. Sgouralis and S. Press, *Biophysical Journal* **112**, 2021 (2017).
- [10] J.-B. Masson, D. Casanova, S. Türkcan, G. Voisinne, M. R. Popoff, M. Vergassola, and A. Alexandrou, *Phys. Rev. Lett.* **102**, 048103 (2009).
- [11] D. L. Ensign and V. S. Pande, *Phys. Rev. E* **82**, 016705 (2010).
- [12] S. Türkcan, A. Alexandrou, and J.-B. Masson, *Biophysical Journal* **102**, 2288 (2012).
- [13] S. Bera, S. Paul, R. Singh, D. Ghosh, A. Kundu, A. Banerjee, and R. Adhikari, *Sci. Rep.* **7**, 41638 (2017).
- [14] J. Skilling, *AIP Conf. Proc.* **735**, 395 (2004).
- [15] <https://github.com/krogjens/sde>.
- [16] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, Berlin, 2004).
- [17] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 2nd ed. (Elsevier, Amsterdam, 1997).
- [18] R. D. Cousins, *Synthese* **194**, 395 (2017).
- [19] C. E. Rasmussen and C. K. I. Williams, *Adaptive Computation and Machine Learning* (MIT Press, 2006).
- [20] A. J. Berglund, *Phys. Rev. E* **82**, 011917 (2010).
- [21] C. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).
- [22] E. T. Jaynes, "Delaware seminar in the foundations of physics," (Springer, Berlin, Heidelberg, 1967) Chap. Foundations of Probability Theory and Statistical Mechanics, pp. 77–101.
- [23] A. Gelman *et al.*, *Bayesian Data Analysis* (CRC Press, Boca Raton, 2013).
- [24] Y. Hong and H. Li, *Rev. Financ. Stud.* **18**, 37 (2005).
- [25] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer, 2002).
- [26] A. G. Cherstvy, A. V. Chechkin, and R. Metzler, *New Journal of Physics* **15**, 083039 (2013).
- [27] J. O. Berger and J. M. Bernardo, *J. Am. Stat. Assoc.* **84**, 200 (1989).
- [28] J. Skilling, *Bayesian Analysis* **1**, 833 (2006).
- [29] C. Calderon, M. Thompson, J. Casolari, R. Paffenroth, and W. Moerner, *J. Phys. Chem. B* **117**, 15701 (2013).