ESTIMATION OF THE DIFFUSION COEFFICIENT WITH A GENERATIVE MODEL

ALAN ZHOU

Abstract. Estimation of the diffusion coefficient and the uncertainty on its measurements is often the first step in any particle tracking analysis. The most common method of doing so is by a mean-squared displacement (MSD) analysis whereby the diffusion coefficient is calculated from fitting a curve to the measurement of MSD over various lag times. However, methods of accurately evaluating MSD are ad hoc and not standardized. In this paper we present a number of alternative methods for estimating the diffusion coefficient in pure diffusion by using a generative model for the data. With simulated data we compare the efficacy of maximum likelihood estimation, Markov chain Monte Carlo (MCMC), and Bayesian marginalization with one of the most common MSD analysis methods, Crocker Grier Weeks (CGW) analysis. We find that methods involving a generative model offer several advantages to CGW.

1. Introduction

Estimation of the diffusion coefficient and uncertainty on its measurements is essential in any colloidal tracking experiment. With an accurate constraint on the diffusion coefficient, one can discover properties of the experimental setup. As an example, from the Stokes-Einstein equation which describes the diffusion of spherical particles through a liquid with low Reynolds number [1]:

$$D = \frac{k_b T}{6\pi\mu a} \tag{1.1}$$

where D is the diffusion coefficient, k_b is Boltzmann's constant, T is the temperature of the medium, μ is the dynamic viscosity of the medium, and a is the radius of a spherical particle. Thus knowledge of D, as well as knowledge of two out of the three of T, μ, and a can directly lead to an accurate estimation of the third parameter.

1.1. Mean Square Displacement.

For a pure diffusion process, the mean squared displacement, $< r^2 >$ is characterized by

$$< r^2 >= n_{dim} D\tau \tag{1.2}$$

Date: May 5, 2016.

Key words and phrases. Diffusion Coefficient, Bayesian Analysis.

where n_{dim} is the number of dimensions and τ is the lag time between measurements.

While accurate measurement of the MSD can provide an unbiased and efficient estimator for the value of D in pure diffusion[2], it is worth noting two key problems. To understand these problems let us first assume[1] that we have a list of particle positions in one dimension $x_1, x_2, ..., x_n$. Then for displacements of 1 time step, we have $N = n - 1$ displacements $\Delta x_{2,1}, \Delta x_{3,2}, ..., \Delta x_{n,n-1}$ where $\Delta x_{i,i-k} = x_i - x_{i-k}$. For displacements of 2 time steps, we could compute N-2 displacements $\Delta x_{3,1}, \Delta x_{4,2}, ..., \Delta x_{n,n-2}$. However, in this case adjacent displacements overlap (e.g., $\Delta x_{3,1}$ and $\Delta x_{4,2}$) and so are not independent of one another. In that case there exist at most n/N independent displacements, where n is the number of time steps. Thus for larger values of τ (longer lag times) there exist fewer independent points from which to calculate the MSD.

The second, more concerning problem is that for displacements of greater than 1 time step, any displacement measurement is correlated to any smaller displacements contained within it. That is, for any displacement $\Delta x_{i,i-k} = x_i - x_{i-k}$, contained within are, for example, the displacements $\Delta x_{i,i-k/2}$ and $\Delta x_{i-k/2,i-k}$. Many MSD methods attempt to resolve the first problem, but the second problem is more difficult to deal with. A model that works with the original data set (the position and time coordinates of particles rather than MSD) will circumvent these problems. Such a method involves writing down a generative model for the data set.

1.1.1. *CGW analysis.* One of the most common methods of performing MSD analysis, CGW analysis involves fitting doing a weighted fit of a line to the mean squared displacement, $< r^2 >$. The uncertainty on the MSD for a given lag time is

$$\sigma_{MSD} = MSD * \sqrt{\frac{2}{N_{ind} - 1}} \tag{1.3}$$

where N_{ind} is the number of independent displacements. In order to resolve the fact that N_{ind} is small for high lag times, Crocker, Grier, and Weeks offer an adjustment for N_{ind}:

$$N_{ind,CGW} = 2\frac{N - n_{\text{steps}}}{n_{\text{steps}}} \tag{1.4}$$

where N is the number of evenly-spaced timesteps in the data and n_{steps} is the number of timesteps for a given lag time. This factor of 2 is very ad hoc, and while it seems reasonable, there does not seem to exist a mathematical justification for it [4].

[1]This argument closely follows that made by Jerome Fung in his P.h.D. dissertation[3]

2. The Generative Model

In order to analyze particle positional and time data, a generative model must first be written down, that is, a model that accurately describes the motion of diffusing particles. A generative model not only allows us to simulate data, it also allows us to derive the likelihood function, from which we will do our analyses, namely maximum likelihood estimation (MLE), Markov chain Monte Carlo (MCMC), and marginalization.

2.1. **Generative Model and the Likelihood Function.** Because the diffusion process follows the normal distribution, writing down a generative model is straightforward. For any given particle at a given position, the PDF for the distance a particle will travel by Brownian motion from its current spot is given by

$$p(\Delta r|\theta_i) = \frac{1}{\sqrt{2\pi\beta}} \exp \frac{-(\Delta r)^2}{2\beta^2} \tag{2.1}$$

where $\beta = \sqrt{6D\tau}$ and $D = \frac{k_b T}{6\pi\mu a}$ and θ_i are the parameters in the model, namely T, μ, and a.

If we assume that the error term in the measurement of a position is Gaussian, we can see that the probability for a given displacement, Δr_i, occurring, is the convolution of two Gaussians[5]:

$$p(\Delta r_i|M, \theta_i, I) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_i^2 + \beta^2}} \exp \frac{-(\Delta r_i - m(x_i|\theta_i))^2}{2(\sigma_i^2 + \beta^2)} \tag{2.2}$$

where y_i is the ith measured value, $m(x_i|\theta_i)$ is the value predicted by the model[2], M, and σ_i is the uncertainty on the displacement measurement given by

$$\sigma_i = \sqrt{\sigma_{m,i}^2 + \sigma_{m,i-1}^2} \tag{2.3}$$

where $\sigma_{m,j}^2$ is the measurement error on the jth positional measurement.

If we extend this model to n measurements, and thus N = n-1 displacements (for time steps of 1), we return the probability of the entire data set, $\mathscr{D}$, the product of N such Gaussians. This is our likelihood function, $\mathscr{L}(\mathscr{D}) = (\mathscr{D}|M, \theta_i, I)$:

$$\mathscr{L}(\mathscr{D}) = p(\Delta r_1, \Delta r_2, ..., \Delta r_N|M, \theta_i, I) = \Pi_{i=1}^N p(\Delta r_i|M, \theta_i, I) \tag{2.4}$$

$$= (2\pi)^{-N/2}\{\Pi_{i=1}^n(\beta^2 + \sigma_i^2)^{-1/2}\} \exp\{\sum_{i=1}^N(-\frac{-(\Delta r_i)^2}{2(\beta^2 + \sigma_i^2)})\} \tag{2.5}$$

where $\Delta r_i = (x_i - x_{i-1})$

[2]Of course, $m(x_i|\theta_i) = 0$. That is, the most probable case is 0 movement.

As a function of the the coordinate positions of each particle in the ith frame x_i, y_i, z_i, the likelihood function is:

$$p(\mathscr{D}|M,\theta_i,I) = \tag{2.6}$$
$$(2\pi)^{-N/2}\{\Pi_{i=1}^{n}(\beta^2+\sigma_i^2)^{-1/2}\}\exp\{\sum_{i=1}^{n}(-\frac{-(x_i-x_{i-1})^2+(y_i-y_{i-1})^2+(z_i-z_{i-1})^2}{2(\beta^2+\sigma_i^2)})\}$$

3. Experiment

For our experiment we simulated a 3D random walk for a single particle beginning at the origin for 100,000 steps and parameters of $\tau = 1s$, $\sigma = 10^{-6}m$, $D = 10^{-10}m^2/s = 100\mu m^2/s$.

3.1. **Maximum Likelihood Estimation (MLE).** To find the 68% confidence interval on our MLE result we took the interval

$$\log\mathscr{L} = \log\hat{\mathscr{L}} - 0.5 \tag{3.1}$$

where $\log\hat{\mathscr{L}}$ is the maximum value of the log of the likelihood function.

For the interval we took $\log\mathscr{L}$ over we used values of D between 10^{-11} and 10^{-9} over a logspace of 10000 points.

3.2. **MCMC.** For our Markov chain Monte Carlo analysis, we used an affine invariant ensemble sampler[6] with 16 walkers and 500 steps. We used a JeffreysPrior with a lower bound of $10^{-12}m^2/s$ and an upper bound of $10^{-10}m^2/s$.

3.3. **Marginalization.** In general, for a single parameter, there is little need to do MCMC as marginalization involves evaluating only a single integral. However, in this case, we were unable to perform marginalization on larger data sets (>1000 data points), due to computational limits. This is discussed in the Appendix.

3.4. **CGW Anaylsis.** For our CGW analysis we calculated the MSD for every lag time from $n_{\text{steps}} = 1$ to $n_{\text{steps}} = 20000$ and did a weighted fit of a line of the form $y = 6D\tau$ to the line, where the weights on each point are given by equation 1.3.

4. Results

For MCMC we obtained 68% credibility interval of $D = 100.04^{+.435}_{-.426}\mu m^2/s$
For MLE we obtained a 68% confidence interval of $D = 100.02^{+.0415}_{-.414}\mu m^2/s$
For CGW we obtained a 68% confidence interval of $D = 99.53^{.0581}_{-.0581}\mu m^2/s$

5. Conclusion

As we can see from the estimated values and credibility and confidence intervals on MCMC and MLE, respectively, analyses of the diffusion coefficient that involve a generative model can offer very accurate measurements with credibility/confidence intervals that are on the same order of magnitude or smaller than traditional MSD analysis.

It is worth noting that in practice, particle trajectories are seldom only subject to pure diffusion motion. Indeed one of the great advantages of MSD analysis is the wide availability of closed-form analytical solutions for the dependence of the MSD on lag time. As examples, equations relating MSD and lag time are readily available for anamalous diffusion, confined diffusion, and flow or directed motion[2].

However, by writing down a generative model and likelihood function for each of these kinds of particle motion, we speculate that analyses, especially of the Bayesian kind, offer the same kinds of advantages over traditional MSD analyses as we have demonstrated here. Namely, that ad hoc procedures can be avoided, as well as removing the problem of correlation between different lag time measurements. Additionally, the advantage of MSD analysis decreases for these different systems; the MSD as an estimator can become biased for particle motion more complex than pure diffusion[2].

Appendix A. Marginalization

In general, marginalization involves an integral of the likelihood function over the prior range. If we take a look at our likelihood function, equation 2.6, we can immediately see that a problem emerges with the factor of

$$\Pi_{i=1}^{n} \frac{1}{\sqrt{(\beta^2 + \sigma_i^2)}}$$

Here, for large n, we run into computational limits and the likelihood function becomes very large for $\beta^2 + \sigma_i^2 < 1$ and very small for $\beta^2 + \sigma_i^2 > 1$. This results in overflow (the computer evaluates $\mathscr{L}(\mathscr{D}) = \infty$) or underflow ($\mathscr{L}(\mathscr{D}) = 0$), respectively. However, since we are only interested in the PDF of the posterior distribution for D, which is eventually normalized, we can find a constant c such that

$$0 < \Pi_{i=1}^{n} \frac{c}{\sqrt{(\beta^2 + \sigma_i^2)}} < \infty$$

Note that this c must be inside the product or else the the product evaluates infinity before the constant can be applied. For small values of $n < 1000$, it is relatively easy to guess a value of c by hand, but for $n > 1000$, guessing by hand becomes virtually impossible and even automating guesses becomes difficult. It is very likely that there exists a computational trick for evaulating this product, but we are not aware of such a trick.

References

[1] Christina Cruickshank Miller. The stokes-einstein law for diffusion in solution. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 106(740):724–749, 1924.

[2] Nilah Monnier. *Bayesian Inference Approaches for Particle Trajectory Analysis in Cell Biology*. PhD thesis, Harvard University, 2013.

[3] Jerome Fung. *Measuring the 3D Dynamics of Multiple Colloidal Particles with Digital Holographic Microscopy*. PhD thesis, Harvard University, 2013.

[4] Eric R. Weeks John C. Crocker. Microrheology tools for idl.

[5] Phil Gregory. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematic Support*. Cambridge University Press, New York, USA, 2010.

[6] Dustin Lang Jonathan Goodman Daniel Foreman-Mackey, David W. Gogg. emcee: The mcmc hammer. *arXiv*, 2013.

E-mail address: alanzhou@college.harvard.edu