Bayesian parameter estimation for stochastic models of biological cell migration

P. Dieterich[a] and R. Preuss[b]

[a]Institut für Physiologie,
Medizinische Fakultät Carl Gustav Carus der TU Dresden,
01307 Dresden, Germany,

[b]Max Planck Institute for Plasma Physics,
EURATOM Association,
Boltzmannstr. 2, 85748 Garching, Germany

Abstract

Cell migration plays an essential role under many physiological and patho-physiological conditions. It is of major importance during embryonic development and wound healing. In contrast, it also generates negative effects during inflammation processes, the transmigration of tumors or the formation of metastases. Thus, a reliable quantification and characterization of cell paths could give insight into the dynamics of these processes. Typically stochastic models are applied where parameters are extracted by fitting models to the so-called mean square displacement of the observed cell group. We show that this approach has several disadvantages and problems. Therefore, we propose a simple procedure directly relying on the positions of the cell's trajectory and the covariance matrix of the positions. It is shown that the covariance is identical with the spatial aging correlation function for the supposed linear Gaussian models of Brownian motion with drift and fractional Brownian motion. The technique is applied and illustrated with simulated data showing a reliable parameter estimation from single cell paths.

1 Introduction

Active motion is one of the primary characteristics of living systems. The observation of individual paths of whole cells or of sub-cellular components as channels in living cells has revealed a variety of complex properties beyond a simple diffusive behavior (see e.g. [1]). Spatial heterogeneities of the environment or strong temporal correlations are possible origins of anomalous sub- and super-diffusion behavior of these objects.

Generalized Langevin equations including memory kernels [5], fractional Fokker-Planck equations [7] or so-called continuous time random walk (CTRW) processes with power-law waiting times or step lengths are applied to model these observations.

However, the estimation of parameters for these stochastic processes applied to biological systems is still under debate: Typically, the mean squared displacement (*msd*) describing the temporal development of the mean squared distance of a group of objects from a common starting point is applied to fit experimental data and theoretical modeling results. However, this approach suffers from several disadvantages and problems: Frequently, only a small number of experimental paths is available causing strong uncertainties into the calculations of the *msd*. In addition, the application of ensemble- (EA) or time-averaging (TA) can lead to different results, known as (weak-)ergodicity breaking [3]. Averages are defined as:

$$msd_{EA}(t) = \frac{1}{N}\sum_{k=1}^{N}[x_k(t) - x_k(0)]^2, \tag{1}$$

$$msd_{TA}(t) = \frac{1}{T-t}\int_{0}^{T-t} dt'[x(t+t') - x(t')]^2, \tag{2}$$

for positions of the trajectories $\{x_k\}$ with $k = 1...N$ (N as the number of trajectories), and T the temporal length of the path. Problems are directly seen even in the simplest case of simulated random walk trajectories as illustrated in Fig. 1. Especially, time-averaged *msd* values display strong scatterings. These would lead to an artificial heterogeneity of diffusion coefficients if one would try to fit individual *msd* data to the theoretical random walk *msd* of $2\ D\ t$ where D denotes the diffusion coefficient and t the time. In addition, the application of (arbitrary) cutoffs – as often done in literature – is doubtful and reliable estimates of errors especially for the time-averaged quantities are missing. Finally, even simple ensemble averaging causes correlations that have to be included into a consistent data analysis.

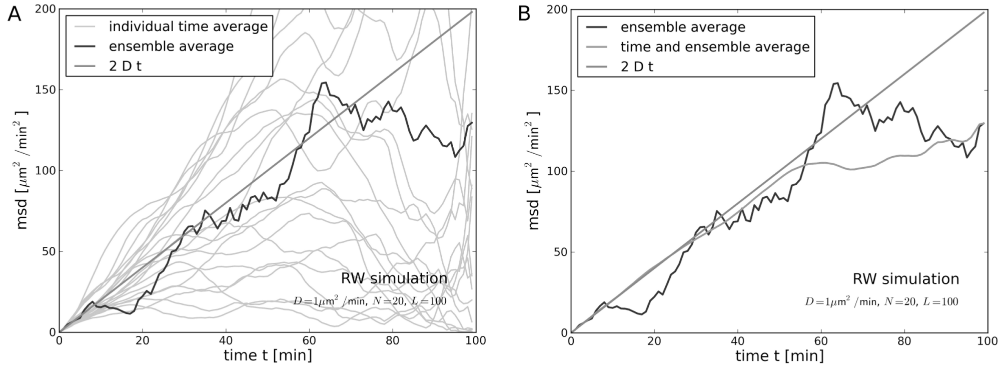

Figure 1: The *msd* was calculated for simulated random walk trajectories according to Eq. 2 ($D = 1\mu m^2/min$, $\Delta t = 1min$, $N = 20$, $L = 100$). A. Curves show the scattering of individual time-averaged *msd* values compared to the theoretical behavior 2 D t of the ensemble averaged *msd*. B. In addition, the figure shows that the combined time- and ensemble-averaged *msd* also deviates from the theoretical behavior for larger times.

2 Technique

2.1 Brownian random motion

As the application of fits to the *msd* suffers from several problems, it is instructive to focus directly on the original data positions obtained from simulations or experiments. The random walk model with drift provides a simple starting point. Positions x_i at time $t_i = i\ \Delta t$ are given by

$$x_{i+1} = x_i + \sqrt{2D\ \Delta t}\ \xi_{i+1} + v_d\ \Delta t \quad , \tag{3}$$

where ξ_i denote uncorrelated Gaussian random variables with zero mean and unit variance, D is the diffusion coefficient, v_d denotes the drift velocity and Δt the sampling time interval. Expanding the discrete form of Eq. 3 for all positions $i = 0...L$ leads to a compact matrix notation of the process:

$$\begin{pmatrix} x_1 - x_0 \\ x_2 - x_0 \\ x_3 - x_0 \\ \vdots \\ x_L - x_0 \end{pmatrix} = \sqrt{2D\Delta t} \underbrace{\begin{pmatrix} 1 & & & & \\ 1 & 1 & & 0 & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix}}_{=:\ \mathbf{T}} \cdot \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \vdots \\ \xi_L \end{pmatrix} + v_d \begin{pmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_L \end{pmatrix} + \sigma_x \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_L \end{pmatrix} \quad . \tag{4}$$

The matrix term on the right side describes the stochastic process and clearly indicates the correlation of points along a trajectory. The second term includes the linear drift which may also have a more complicated time-dependent form.

The last term proportional to σ_x was added to describe spatial uncertainties of the measurement of positions or to include additional fluctuations not included in the model such as biological noise [1]. Again, η_i represent uncorrelated Gaussian variables with zero mean and unit variance.

To further proceed, it is necessary to calculate expectation values of positions. These expectation values are evaluated over all realizations of ξ and η

$$\langle ... \rangle = \int_{-\infty}^{\infty} \prod_{i=1}^{L} d\xi_i \; d\eta_i \;\; ... \;\; p(\xi_i) \; p(\eta_i) \quad , \tag{5}$$

where the variables ξ_i and η_i are distributed according to

$$p(\xi_i) = \frac{1}{(2\pi)^{L/2}} \; \exp\left\{ -\frac{1}{2} \sum_{i=1}^{L} \xi_i^2 \right\} , \quad p(\eta_i) = \frac{1}{(2\pi)^{L/2}} \; \exp\left\{ -\frac{1}{2} \sum_{i=1}^{L} \eta_i^2 \right\} \quad . \tag{6}$$

In the following we set the initial point to $x_0 = 0$ without loss of generality. Calculation of averages based on Eqs. 4 – 6 delivers the mean position $\langle x(t_i) \rangle$ at time t_i

$$\langle x_i \rangle = v_d \; \Delta t \; i \tag{7}$$

and the covariance of positions

$$\langle [x_i - v_d \; t_i] \; [x_j - v_d \; t_j] \rangle = 2 \; D \; \Delta t \; \min(i,j) + \delta_{i,j} \; \sigma_x^2 \quad . \tag{8}$$

Thereby $\mathbf{T} \; \mathbf{T}^T = \min(i,j)$ with $\mathbf{T}$ as defined in Eq. 4 has been used. It is noteworthy that the first term of this result agrees with the well known aging correlation function $2D \; \min(t_i, t_j)$ of the random walk process. As mean and variances of x_i are given, a maximum entropy treatment (see e.g. [2] pages 450f) directly delivers the likelihood for the data $\{x_i\}$:

$$p(\{x_i\} | v_d, D, I) = \frac{1}{\sqrt{(2\pi)^L \det \mathcal{C}}} \; \exp\left\{ -\frac{1}{2} \sum_{i,j}^{L} (x_i - v_d t_i) \; [\mathcal{C}^{-1}]_{i,j} \; (x_j - v_d t_j) \right\} \tag{9}$$

with $\mathcal{C}_{i,j} = 2 \; D \; \Delta t \; \min(i,j) + \delta_{i,j} \; \sigma_x^2$ for the random walk process. This agreement should hold in general for linear Gaussian processes. In addition, Eq. 9 shows that the calculation of the *msd* is not necessary to perform a parameter estimation for these linear Gaussian processes. Whereas v_d enters in the numerator of the likelihood, the diffusion coefficient and the position uncertainty enters the denominator within the covariance structure. With $\sigma_x = 0$ the inverse covariance can be calculated analytically showing a simple tri-diagonal behavior connecting neighboring data positions. Finally, this approach is applicable for single trajectories allowing biologists to find differences even between individual cells.

2.2 Fractional Brownian motion

Analogously, dynamics of fractional Brownian motion [6, 10] can be formulated as a discrete matrix equation. In contrast to Brownian motion, the process is driven by correlated Gaussian random variables ζ_i with the following property:

$$\langle \zeta_i \zeta_j \rangle = \frac{\sigma^2}{2} \left\{ |j-i-1|^{2H} - 2|j-i|^{2H} + |j-i+1|^{2H} \right\} \quad , \tag{10}$$

with variance σ^2 and the so-called Hurst coefficient H ($0 < H < 1$). Expectation values have now to be calculated with respect to the correlations defined in Eq. 10. After some algebra of the discrete formalism with the correlated random variables (setting the drift term equal to zero) one obtains the following covariance

$$\langle x(t_i) x(t_j) \rangle = \frac{\sigma^2}{2} \left\{ |t_i|^{2H} + |t_j|^{2H} - |t_i - t_j|^{2H} \right\} \quad , \tag{11}$$

which is identical with the aging correlation function of fractional Brownian motion. Depending on the value of the Hurst coefficient H, the *msd* of these processes

$$\langle x(t_i)^2 \rangle = \sigma^2 \, t_i^{2H} \tag{12}$$

shows sub-diffusive or super-diffusive behavior for $H < 1/2$ and $H > 1/2$, respectively. In the following, simulations of paths from fractional Brownian motion will be used to test our formalism with systems including temporal power-law correlations.

3 Application

At first, we simulate paths of Brownian random motion according to Eq. 3 to test the performance of the proposed procedure. As the number of parameters in our models is small, we have performed all integrals by simple lattice integration. We assumed flat priors for all parameters under consideration. Mean and variances of parameters were calculated from the resulting posterior. Fig. 2 shows the dependency of the estimated parameters $\langle D \rangle$ and $\langle v_d \rangle$

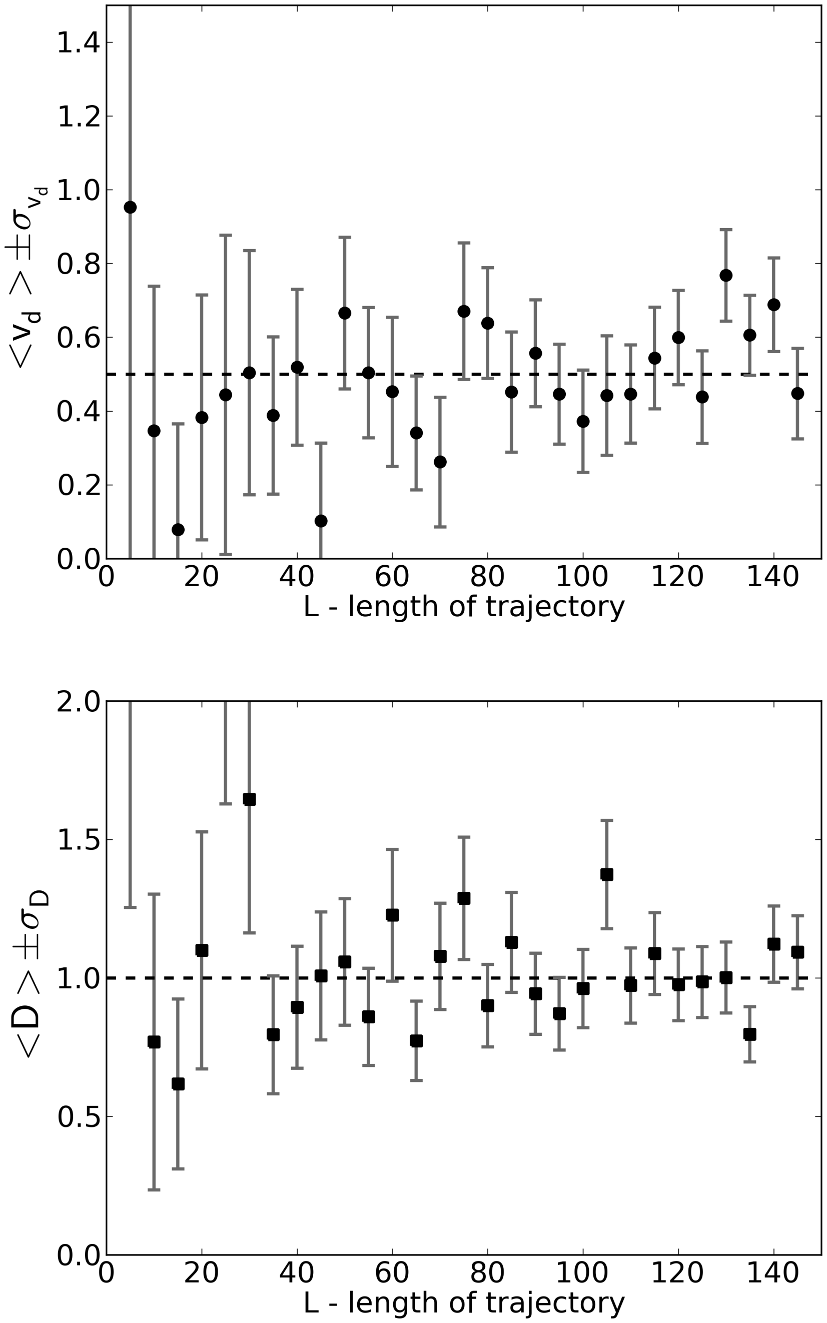

Figure 2: Parameters and uncertainties as a function of the length L of the trajectory. Simulations were performed for a random walk model with diffusion coefficient $D = 1$ and drift velocity $v_d = 0.5$.

including their uncertainties as a function of the length of the trajectory. Uncertainties are reduced with increasing path length L. The estimated parameters are in agreement with the parameters used in simulation indicated as dashed lines. This figure shows that the analysis leads to reasonable parameter estimates for a single trajectory. Even shorter trajectories with $L \sim 30$ deliver parameters within the range of the applied simulation parameter.

Fig. 3 shows the variability of estimating $\langle D \rangle$ and $\langle v_d \rangle$ for 20 independently simulated random walk trajectories. Most estimated parameter values are in agreement with the simulation parameter within one standard deviation. Using a product of the likelihoods as given in Eq. 9 of independent paths would allow to calculate the expectation values for the diffusion and drift coefficient of the observed group.

In a second step, we apply the formalism to fractional Brownian motion. Simulations of trajectories were performed with Hosking's algorithm [4]. As a result Fig. 4 shows the posterior $p(H|\{d_i\})$ as a function of the Hurst coefficient H for simulated trajectories with a different number of data points L. For both cases of sub- and super-diffusion the posterior localizes around the theoretical values of $H = 0.3$ and $H = 0.7$ with an increasing number of data points. This indicates that the proposed formalism can also be applied to processes with power-law correlations.

4 Conclusions and outlook

We have proposed and applied a simple procedure to extract parameters of stochastic processes given a single experimental or simulated path of this process. The procedure relies on including the so-called spatial aging correlation function as the covariance matrix within the likelihood function. The technique performs well if the covariance matrix is known for the assumed stochastic process and seems to be valid in general for linear Gaussian processes. In addition, it avoids problems of *msd* application and directly uses all simulated or experimentally measured data points. The technique is explicitly applied to experimental results of rolling leukocytes in the contribution of Moskopp et al. [8] within these proceedings.

If the covariance matrix is not known theoretically one can try to extract it directly from the data. This poses several problems and can even lead to doubtful parameter estimations. Preuss and Dieterich [9] discuss these issues within these proceedings and apply a renormalization of the covariance matrix and its inverse by a Bayesian approach including the uncertainties of the data covariance matrix. As shown by simulations of Brownian and fractional Brownian motion this approach rescues parameter estimation where

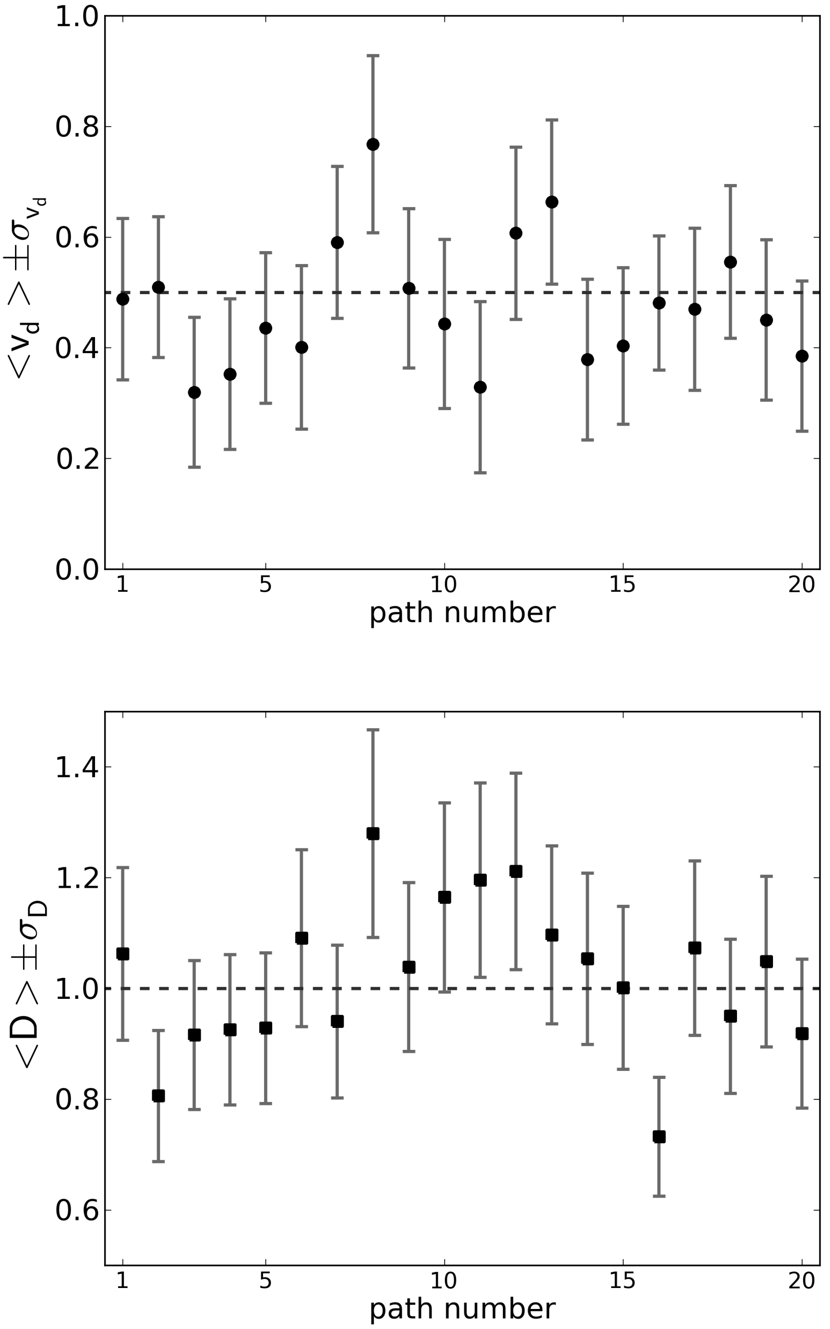

Figure 3: Parameters and uncertainties for 20 independently simulated paths. Simulation was performed with diffusion coefficient $D = 1$ and drift velocity $v_d = 0.5$. Trajectories consist of $L = 100$ data points.

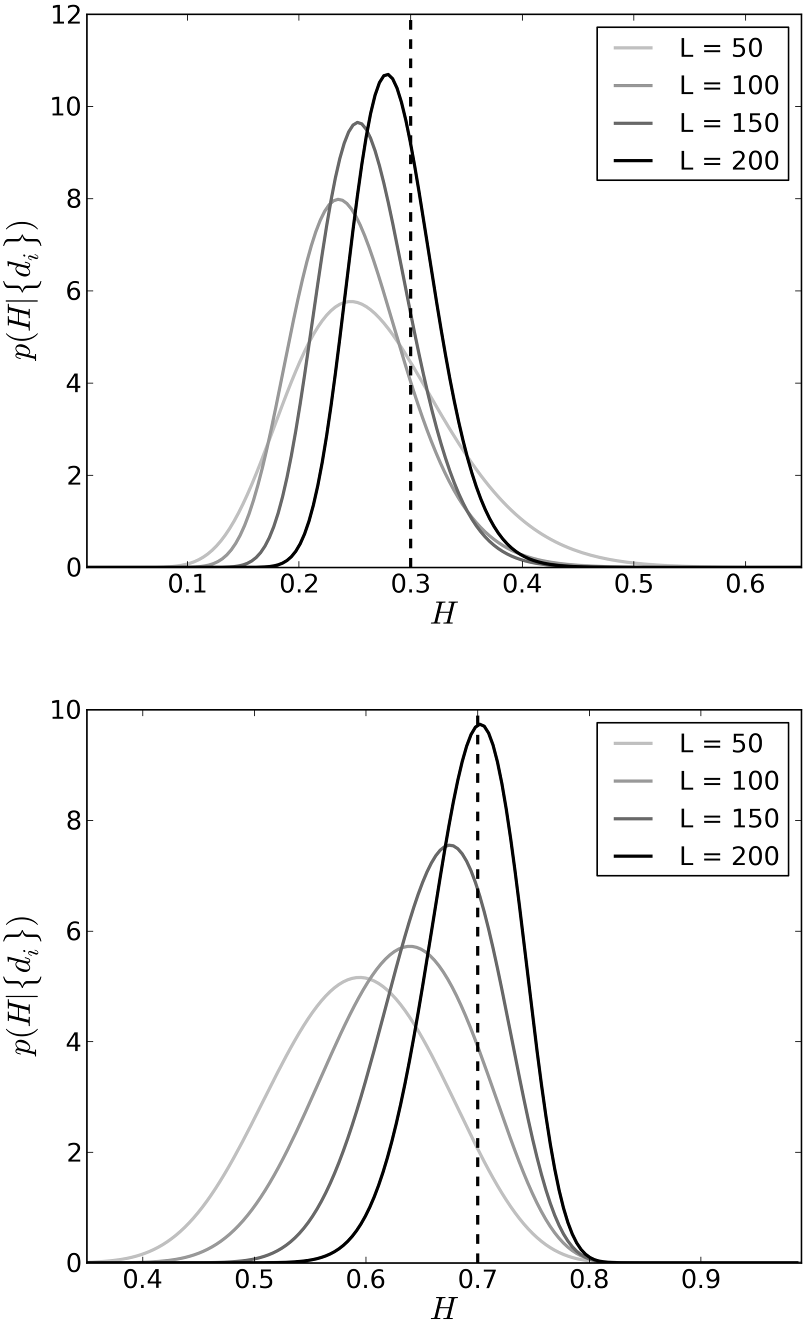

Figure 4: Posteriors of fractional Brownian motion: The simulation of paths was performed for sub-diffusion with $H = 0.3$ (left side) and for super-diffusion with $H = 0.7$ (right side). Each posterior corresponds to a simulated trajectory with L data points. The posterior converges towards the theoretical values of H indicated by dashed vertical lines with an increasing number of data points L.

the naive usage of raw covariance would lead to wrong results.

The current approach is limited to linear Gaussian processes. However, one could also try to apply it to more complicated models as the fractional Langevin [5] or the non-Gaussian fractional diffusion equation [7] where aging correlation functions are known. Finally Bayesian model selection could help to select the best aging correlation function out of several candidates. This is of major importance for experimental analyses where the underlying aging correlation typically is unknown in advance.

In summary, the suggested technique offers a reliable way to perform Bayesian parameter estimation for stochastic models directly from single trajectories. This is of special interest to biologists and enables to distinguish between individual cell tracks.

References

[1] Peter Dieterich, Rainer Klages, Roland Preuss, and Albrecht Schwab. Anomalous dynamics of cell migration. *Proc Natl Acad Sci USA*, 105(2):459–463, Jan 2008.

[2] P. C. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support.* Cambridge University Press, 2005.

[3] Y. He, S. Burov, R. Metzler, and E. Barkai. Random time-scale invariant diffusion and transport coefficients. *Phys Rev Lett*, 101(5):058101, Aug 2008.

[4] J. R. M. Hosking. Modeling persistence in hydrological time series using fractional differencing. *Water Resour. Res.*, 20(12):1898–1908, January 1984.

[5] E. Lutz. Fractional Langevin equation. *Physical Review E*, 6405(5):051106, November 2001.

[6] Benoit B. Mandelbrot and John W. Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM Rev.*, 10(4):422–437, 1968.

[7] R. Metzler and J. Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1):1–77, December 2000.

[8] M. L. Moskopp, R. Preuss, A. Deussen, T. Chavakis, and P. Dieterich. Bayesian data analysis of the dynamics of rolling leukocytes. In *Proceedings of the 32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Garching*, 2012.

[9] R. Preuss and P. Dieterich. Employment of the covariance matrix in parameter estimation for stochastic processes in cell biology. In *Proceedings of the 32nd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Garching*, 2012.

[10] Gary M. Raymond and James B. Bassingthwaighte. Deriving dispersional and scaled windowed variance analyses using the correlation function of discrete fractional gaussian noise. *Physica A: Statistical and Theoretical Physics*, 265(1-2):85–96, March 1999.