Solving missing data in tracked image data: Ornstein-Uhlenbeck process and data augmentation
Supervisor: Burroughs (Mathematics). Second supervisor: McAinsh.

Tracking of objects in image sequences is ubiquitous in biology, from single molecules to clusters that appear as spots under the limitations of light microscopy. Such tracks contain valuable information on the mechanisms of movement of those objects, and the fitting of mathematical models to that data is an active area. Such model fitting has attracted sophisticated statistical techniques to ensure the interpretations are reliable and dealing with problems of taking into account measurement errors and missing data points. Markov chain Monte Carlo (MCMC) techniques are probably the most powerful and reliable.

This mini-project tackles the problem of missing data points in chromosome tracking during cell division. Chromosomes are observed to oscillate during *metaphase* of the cell division cycle - this is the phase when the paired duplicated chromosomes are in a waiting pattern whilst correct attachments are checked, prior to separation of the duplicated chromosmes to the daughter cells, ensuring each daughter cell gets one and only one copy of each chromosome. This checking system is often corrupted in cancerous cells, leading to genome instability. *Crucially this oscillation provides significant information on how chromosomes are moved around the cell, achieved by fitting a dynamic model to the tracking data* [1,2].

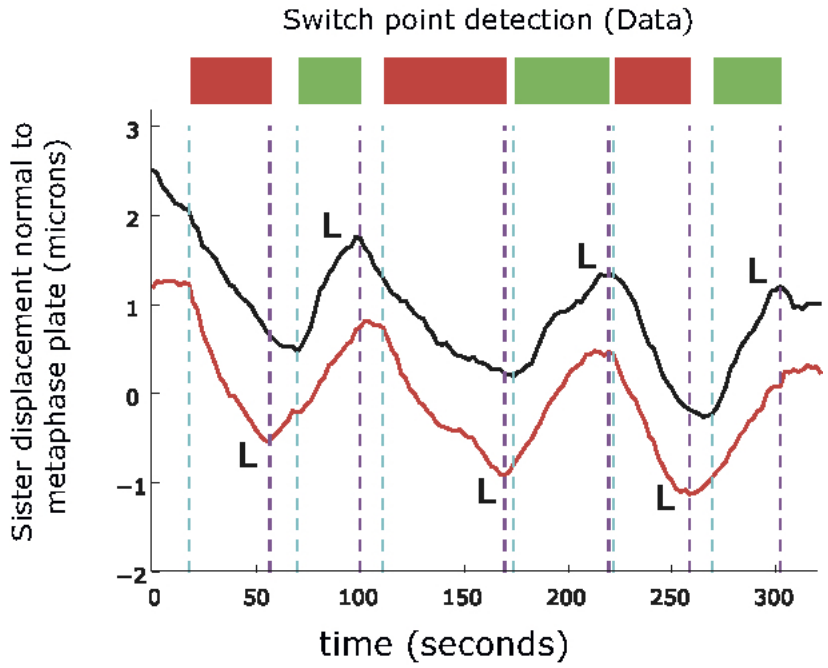

Figure 1: Tracked oscillatory trajectory of a chromosome pair (black, red) with a complete set of time points annotated with switching points (vertical lines) as determined from an MCMC switching point inference algorithm. Sisters moving towards the same left (red), right (green) pole is ahown in the top bar.

The model. The data consists of paired tracks $X^1(t), X^2(t)$ (paired chromosomes, hereafter called particles) undergoing an approximate saw-tooth oscillation, see Fig. 1. To catch the saw-tooth behaviour there are two hidden variables corresponding to the direction of motion of the two chromosomes, $\sigma^k(t)$, values in $\{+,-\}$ for moving to the right or left respectively. The system biophysics (eg the particles are connected by a spring) gives the following dynamics (a coupled pair

of linear stochastic differential equations),

$$\begin{pmatrix} dX^1 \\ dX^2_t \end{pmatrix} = dt \begin{pmatrix} -v_{\sigma^1} + \kappa L \\ v_{\sigma^2} - \kappa L \end{pmatrix} + dt\kappa \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X^1_t \\ X^2_t \end{pmatrix} - dt\alpha \begin{pmatrix} X^1_t \\ X^2_t \end{pmatrix} + \sqrt{(}2D) \begin{pmatrix} dB^1_t \\ dB^2_t \end{pmatrix} \tag{1}$$

where $\alpha, \kappa, L, v_{\pm}, D$ are parameters.

Fitting the model. As the particles are observed every 2s, we integrate the sDE, using a Euler approximation giving a Gaussian model for the differences $X^k_{t+\Delta t} - X^k_t$, time step $\Delta t = 2$s. We then used MCMC techniques to fit this model to data, [1,2]. However this model requires complete observation data, $X_{k\Delta t}, k = 1...N$ which is restrictive. In fact we lose about 1/2 our tracked data, the majority in fact lacking one missing observation in the track.

Solving the missing data problem. There are a number of ways to deal with missing data in model fitting applications. Since our algorithm is Bayesian, the easiest idea is to model the missing data itself as hidden (called *data augmentation*). The algorithm then has an extra layer where the missing data values are also updated (using a so called *Brownian bridge*). Unfortunately such algorithms can be computationally expensive depending on how many data points are missing. The alternative method is to have an algorithm that can utilize arbitrary time steps, i.e. the algorithm utilizes the conditional probability $\pi(X_{t+s}|X_t, \theta)$. This is of course only practical if there is an analytic form. Since our sDE is linear, in fact it is an *Ornstein-Uhlenbeck* process, it can be solved in the variables $X_1 - X_2$, $X_1 + X_2$. The question is whether this solution is tractable enough to use in an MCMC algorithm, and able to give a efficient method to solve the missing data problem.

The project. In this miniproject you will develop an MCMC algorithm to fit the sDE to experimental data that is robust to missing data, using data augmentation and/or the OU process. You will test the algorithm(s) on simulated data. On simulated and experimental data you will test the effect of missing data points on the estimated parameters by using complete trajectories and removing data points. Finally, if time permits you will fit the model to experimental data with missing data.

Data availability. We have extensive tracked data sets (100s of cells, 1000s of trajectories generated in the McAinsh lab in Warwick Medical School). Data from a variety of platforms is available (spinning disc, light sheet).

Desirable skills An understanding of Markov processes and MCMC. Knowledge of stochastic differential equations would be an advantage. Programming in MatLab.

Opportunities for a PhD. This mini-project can lead to a PhD with Burroughs and McAinsh, with 3i as external partner. 3i are willing to part fund the PhD. The PhD project could focus on developing models and associated fitting algorithms for chromosome movements throughout the full course of mitosis, previously an impossible task until the emergence of new lattice light sheet microscopes that can provide long tracks (40 min) of sufficient spatio-temporal resolution for modelling.

References:

[1] Super-resolution kinetochore tracking reveals the mechanisms of human sister kinetochore directional switching. Nigel J. Burroughs, Edward F. Harry and Andrew D. McAinsh. eLife 2015;10.7554/eLife.09500.

[2] Inferring the Forces Controlling Metaphase Kinetochore Oscillations by Reverse Engineering System Dynamics. Jonathan W. Armond, Edward F. Harry, Andrew D. McAinsh, Nigel J. Burroughs, PLOS Comp. Biol. 2015, DOI: 10.1371/journal.pcbi.1004607.

[3] SOFTWARE: KiT: A MATLAB package for kinetochore tracking. Jonathan W. Armond, Elina Vladimirou, Andrew D. McAinsh and Nigel J. Burroughs.Bioinformatics, 2016.