From Statistical Physics to Data-Driven Modelling with Applications in Quantitative Biology : Tutorial 1 S.C., R.M., F.Z.

Bayesian inference and single-particle tracking

Corrections

Solution

Data Analysis

The trajectory in the (x, y) plane given in the data for $M = 1000$ is plotted in figure **??** (left). It has the characteristics of a random-walk: the space is not regularly filled, but the trajectory densely explore one region before "jumping" to another region.

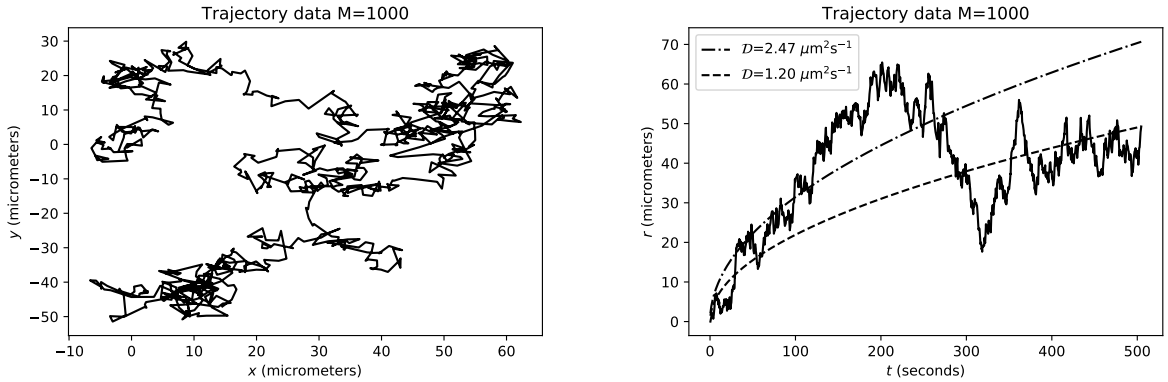


Figure 1: Left: trajectory of the particle. Right: displacement from the origin as a function of time.

The displacement $r = \sqrt{x^2 + y^2}$ as a function of time is plotted in figure **??** (right). On average, it grows as the square root of the time, but on a single trajectory we observe large fluctuations. The random walk in two dimensions is described by the relation:

$$\langle r^2(t) \rangle = 4\mathcal{D}t \ , \tag{1}$$

where $\mathcal{D}$ is the diffusion coefficient whose physical dimensions are $[\mathcal{D}] = l^2 t^{-1}$. Here lengths are given in μm and times in s. A first estimate of $\mathcal{D}$ from the data can be obtained by just considering the largest time and estimating

$$\mathcal{D}_0 = \frac{r^2(t_{max})}{4t_{max}} \tag{2}$$

giving $\mathcal{D}_0 = 1.20 \mu m^2 \ s^{-1}$ for the data set with $M = 1000$. Another estimate of $\mathcal{D}$ can be obtained as the average of the square displacement from one data point to the next one divided by the time interval. We define the differences between two successive positions and between two successive recording times

$$\delta x_i = x_{i+1} - x_i \ , \qquad \delta y_i = y_{i+1} - y_i \ , \qquad \delta t_i = t_{i+1} - t_i \ . \tag{3}$$

Note that $i = 1, \ldots, M-1$. The square displacement in a time step is $\delta r_i^2 = \delta x_i^2 + \delta y_i^2$ and the estimate of $\mathcal{D}$ is

$$\mathcal{D}_1 = \frac{1}{4(M-1)} \sum_{i=1}^{M-1} \frac{\delta r_i^2}{\delta t_i} \ , \tag{4}$$

giving $\mathcal{D}_1 = 2.47\mu m^2\ s^{-1}$ for the same data set. These estimates are compared with the trajectory in figure **??** (right).

Posterior distribution

Due to diffusion δx_i and δy_i are Gaussian random variables with variances $2\mathcal{D}\,\delta t_i$. We have

$$p(\delta x_i|\mathcal{D},\delta t_i) = \frac{1}{\sqrt{4\pi\mathcal{D}\,\delta t_i}} e^{-\frac{\delta x_i^2}{4\,\mathcal{D}\,\delta t_i}}\ , \tag{5}$$

and $p(\delta y_i|\mathcal{D},\delta t_i)$ has the same form. The probability of a time series of increments $\{\delta x_i,\delta y_i\}_{i=1,\ldots,M-1}$, given $\mathcal{D}$ is therefore:

$$\begin{aligned} P(\{\delta x_i,\delta y_i\}|\mathcal{D},\{\delta t_i\}) &= \prod_{i=1}^{M-1} \frac{1}{4\pi\mathcal{D}\,\delta t_i} e^{-\frac{\delta x_i^2}{4\,\mathcal{D}\,\delta t_i}-\frac{\delta y_i^2}{4\,\mathcal{D}\,\delta t_i}} \\ &= C e^{-B/\mathcal{D}}\,\mathcal{D}^{-(M-1)}\ , \end{aligned} \tag{6}$$

where $C = \prod_{i=1}^{M-1}\frac{1}{4\pi\delta t_i}$ and $B = \sum_{i=1}^{M-1}\frac{\delta r_i^2}{4\,\delta t_i}$. Note that to infer $\mathcal{D}$ we do not need the absolute values of (x_i,y_i), but only their increments on each time interval.

According to Bayes' Theorem,

$$P(\mathcal{D}|\{\delta x_i,\delta y_i,\delta t_i\}) = \frac{P(\{\delta x_i,\delta y_i\}|\mathcal{D},\{\delta t_i\})P(\mathcal{D})}{\int_0^\infty \mathrm{d}\mathcal{D}\ P(\{x_i,y_i\}|\mathcal{D},\{\delta t_i\})P(\mathcal{D})}\ . \tag{7}$$

We consider an improper uniform prior $P(\mathcal{D})$ =const. This can be thought as a uniform prior in $[\mathcal{D}_{\min},\mathcal{D}_{\max}]$, in the limit $\mathcal{D}_{\min}\to 0$ and $\mathcal{D}_{\max}\to\infty$. Thanks to the likelihood, the posterior remains normalisable in this limit.

Note that, introducing

$$\mathcal{D}^* = \frac{B}{M-1} = \frac{1}{4(M-1)}\sum_{i=1}^{M-1}\frac{\delta r_i^2}{\delta t_i} = \mathcal{D}_1\ , \tag{8}$$

we can write the posterior as

$$\begin{aligned} P(\mathcal{D}|M,\mathcal{D}^*) &= \frac{e^{-(M-1)\mathcal{D}^*/\mathcal{D}}\,\mathcal{D}^{-(M-1)}}{\int_0^\infty \mathrm{d}\mathcal{D} e^{-(M-1)\mathcal{D}^*/\mathcal{D}}\,\mathcal{D}^{-(M-1)}} \\ &= \frac{e^{-(M-1)\mathcal{D}^*/\mathcal{D}}\ \mathcal{D}^{-(M-1)}\ [(M-1)\mathcal{D}^*]^{(M-2)}}{(M-3)!}\ , \end{aligned} \tag{9}$$

where the denominator is easily computed by changing variable to $u=\mathcal{D}^*/\mathcal{D}$ and recognising that $\int_0^\infty \mathrm{d}t\, e^{-t}\,t^{M-3} \equiv \Gamma(M-2)\equiv(M-3)!$, where Γ indicates the Gamma function. Note that, as in Laplace problem,

$$P(\mathcal{D}|M,\mathcal{D}^*) \propto e^{(M-1)f_{\mathcal{D}^*}(\mathcal{D})}\ , \qquad f_{\mathcal{D}^*}(\mathcal{D}) = -\frac{\mathcal{D}^*}{\mathcal{D}} - \log\mathcal{D}\ . \tag{10}$$

The most likely value of $\mathcal{D}$ is precisely $\mathcal{D}^*$, which is the maximum of $f_{\mathcal{D}^*}(\mathcal{D})$ and of the posterior, and coincides with the previous estimate $\mathcal{D}_1$.

The average value of $\mathcal{D}$ can also be computed by the same change of variables,

$$\langle\mathcal{D}\rangle = \frac{(M-1)}{(M-3)}\mathcal{D}^*\ , \tag{11}$$

and converges to $\mathcal{D}^*$ for $M\to\infty$. The variance of $\mathcal{D}$ is

$$\sigma_{\mathcal{D}}^2 = \frac{(M-1)^2}{(M-3)^2\,(M-4)}(\mathcal{D}^*)^2\ , \tag{12}$$

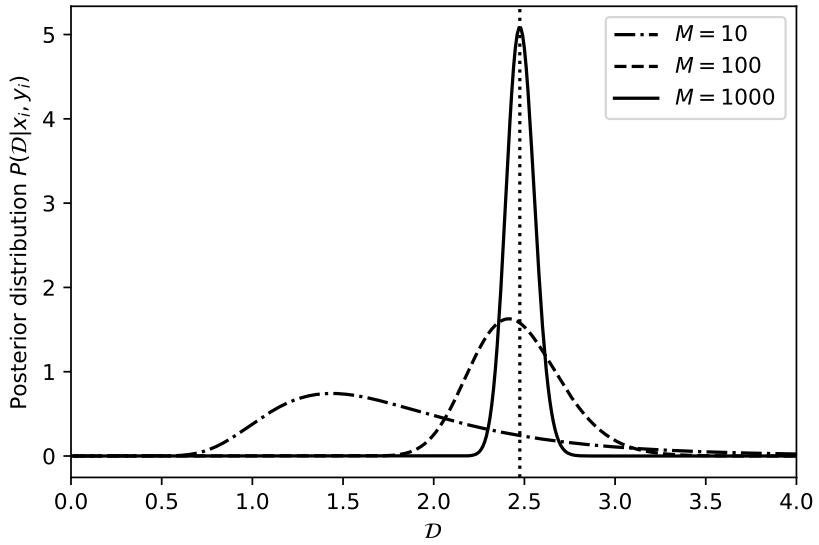
and it decreases proportionally to $1/M$ for large M.

Numerical analysis of the data

The trajectories given in the data files give the following results:

Name-file	M	$\mathcal{D}$	$\mathcal{D}^*$	$\langle\mathcal{D}\rangle$	$\sigma_{\mathcal{D}}$
dataN10d2.5.dat	10	2.5	1.43	1.84	0.75
dataN100d2.5.dat	100	2.5	2.41	2.46	0.25
dataN1000d2.5.dat	1000	2.5	2.47	2.48	0.08

An example of the posterior distribution is given in the following figure:



Note that for large value of M it is not possible to calculate directly the $(M-3)!$ in the posterior distribution, Eq. (**??**). It is better to use the Stirling's formula:

$$(M-3)! \approx \sqrt{2\pi}\,(M-3)^{M-3+1/2}\,e^{-(M-3)} \tag{13}$$

One therefore obtains

$$P(\mathcal{D}|\{x_i,y_i\}) \approx \frac{e^{-B/\mathcal{D}}\,\mathcal{D}^{-(M-1)}\sqrt{M-3}}{\sqrt{2\pi}\,e}\left(\frac{B\,e}{M-3}\right)^{(M-2)} . \tag{14}$$

We see that this is a very good approximation for $M = 10$ and we can use it for $M = 100$ and $M = 1000$.

Diffusion constant and characteristic size of the diffusing object

The order of magnitude of the diffusion constant can be obtained by the Einstein-Stokes relation: $\mathcal{D} = \frac{k_B T}{6\pi\eta\ell}$, where ℓ is the radius of the object (here considered as spherical), and η is the viscosity of the medium. Considering the viscosity of the water $\eta = 10^{-3}$ Pa s and $k_B T = 4 \times 10^{-21}$ J, one obtains the following orders of magnitude:

object	ℓ (nm)	$\mathcal{D}$ (μm^2 s^{-1})
small protein (lysozime) (100 residues)	1	200
large protein (1000 residues)	10	20
influenza viruses	100	2
small bacteria (e-coli)	2000	0.1

Therefore the data could correspond to an influenza virus diffusing in water.

The paper by Brune & Kim (1992) reports the following values: for a small protein (lysozyme) $\mathcal{D} = 10^{-6}$ cm^{2}s^{-1}, and for a tobacco virus $\mathcal{D} = 4\ 10^{-8}$cm^{2}s^{-1}, in agreement with the above orders of magnitude. In the paper by Robson et al. the diffusion coefficient of proteins complexes inside bacteria, and with widths approximately equal to $300 - 400$ nm, are estimated to be equal to $\mathcal{D} = 10^{-2}\mu$m^{2}s^{-1}. Differences with the order of magnitude given above are due to the fact that the diffusion is confined and the medium is the interior of the cell, with larger viscosity than water.