Statistical analysis of super-resolution single particle trajectories

Matteo Dora

matteo.dora@polito.it

5th February 2018

1 Introduction

A dataset of single particle trajectories sampled at 125 ms was analyzed. The goal was to recover a description of the physical and dynamical properties of the particles (e.g. diffusion, drift field) by means of statistical analysis.

In particular I was interested in studying two features of the drift field: attractors and channels. Attractors (or wells) are regions of the space where the drift field vanishes and the particles can escape only by diffusion. Channels instead are paths along which particles are funneled together and pushed at high speed from one region to another.

The dataset consisted in $\sim$350 000 datapoints belonging to $\sim$20 000 trajectories. The spatial distribution of the datapoints is visible in figure 1. The supporting code is available in [4].

2 Physical model

2.1 Effective dynamical model

At the microscopic level the motion of the molecules can be described by the Langevin equation, which in the case of biological processes we are interested in can be considered in its large friction limit (Smoluchowski's equation)

$$\dot{\boldsymbol{x}} = \frac{\boldsymbol{F}(\boldsymbol{x})}{\gamma} + \sqrt{2D}\dot{\boldsymbol{w}} \tag{1}$$

where $\boldsymbol{F}(\boldsymbol{x})$ is the drift force exerted on the particle at position $\boldsymbol{x}$, γ is the friction coefficient, D is the diffusion coefficient and $\boldsymbol{w}(t)$ is a two-dimensional Wiener process. For this microscopic model we expect the diffusion to be due to thermal agitation and we can consider it isotropic.

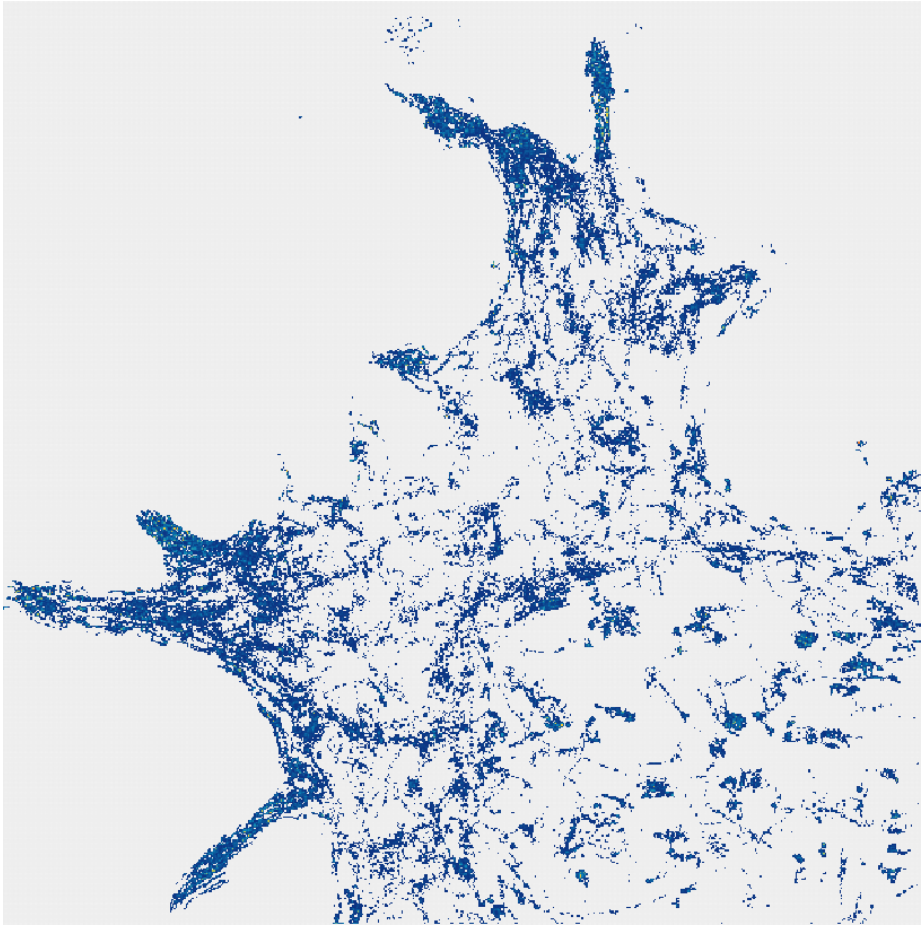


Figure 1: Dataset points at a resolution of 100 nm

However, the empirical data have a finite resolution in space and time due to obvious limits of the acquisition device. It is thus not possible to recover the microscopic model, since we miss information about the local behaviour both in space (e.g. the presence of microscopic obstacles) and time (e.g. thermal fluctuations are much faster than the measuring timescale). Yet we can build a coarse-grained model transforming eq. 1 into the effective stochastic equation [1] [2]

$$\dot{\boldsymbol{x}} = \boldsymbol{a}(\boldsymbol{x}) + \sqrt{2}\boldsymbol{B}\dot{\boldsymbol{w}} \tag{2}$$

where $\boldsymbol{a}(\boldsymbol{x})$ is the effective drift field and $\boldsymbol{D} \equiv \boldsymbol{B}^T\boldsymbol{B}$ is the effective diffusion tensor. Note that in this coarse-grained model the diffusion coefficient cannot be considered isotropic because it takes into account the microscopic local features (e.g. obstacles).

2.2 Reconstructing the drift and diffusion coefficients

It is possible to reconstruct the effective drift and diffusion tensor of the coarse-grained model of eq. 2 from the trajectories increments [2] [3] with

$$a^i(\boldsymbol{x}) = \lim_{\Delta t \to 0} \frac{\mathbb{E}\left[x^i(t+\Delta t) - x^i(t) \mid \boldsymbol{x}(t) = \boldsymbol{x}\right]}{\Delta t} \tag{3}$$

$$2D^{ij}(\boldsymbol{x}) = \lim_{\Delta t \to 0} \frac{\mathbb{E}\left[\left(x^i(t+\Delta t) - x^i(t)\right)\left(x^j(t+\Delta t) - x^j(t)\right) \mid \boldsymbol{x}(t) = \boldsymbol{x}\right]}{\Delta t} \tag{4}$$

where the expected value is taken on the trajectories that pass through $\boldsymbol{x}$ at time t.

We can translate eq. 3 and 4 into statistical estimators by grouping the trajectories in a small neighbourhood of $\boldsymbol{x}$ on which we can perform the average. For convenience I chose to divide the space using a grid of identical square bins $S_l(\boldsymbol{x})$ characterized by side l and center $\boldsymbol{x}$.

Defining $\Delta \boldsymbol{x}_k(t_m) \equiv \boldsymbol{x}_k(t_m+\Delta t) - \boldsymbol{x}_k(t_m)$, i.e the m-th step of the trajectory k, the estimators can then be expressed as

$$\hat{a}^i(\boldsymbol{x}) = \frac{1}{N} \sum_{\boldsymbol{x}_k(t_m) \in S_l(\boldsymbol{x})} \frac{\Delta x_k^i(t_m)}{\Delta t} \tag{5}$$

$$\hat{D}^{ij} = \frac{1}{2}\frac{1}{N} \sum_{\boldsymbol{x}_k(t_m) \in S_l(\boldsymbol{x})} \frac{\Delta x_k^i(t_m) \Delta x_k^j(t_m)}{\Delta t} \tag{6}$$

where N is the number of steps falling into the square $S_l(\boldsymbol{x})$.

2.3 Error of the estimators

To determine the error of the estimators I considered the discretized version of eq. 2 (see appendix of [1])

$$\Delta \boldsymbol{x}(t_m) = \boldsymbol{a}(\boldsymbol{x})\Delta t + \sqrt{2\Delta t}\boldsymbol{B}\boldsymbol{\eta}_m \tag{7}$$

where $\boldsymbol{\eta}_m$ is a two-dimensional white Gaussian noise with unit variance.

The estimator error for the drift in the bin $S_l(\boldsymbol{x})$ containing N steps is then

$$\boldsymbol{e_a}(\boldsymbol{x}) = \hat{\boldsymbol{a}}(\boldsymbol{x}) - \boldsymbol{a}(\boldsymbol{x}) = \frac{1}{N}\sqrt{\frac{2}{\Delta t}}\boldsymbol{B}\sum_{k=1}^{N}\boldsymbol{\eta}_k \tag{8}$$

that is a normally distributed random variable with covariance matrix $\frac{2\boldsymbol{D}}{N\Delta t}$.

A similar result holds for the diffusion estimator. Considering the case of

isotropic diffusion for simplicity

$$\hat{D}(\boldsymbol{x}) = D(\boldsymbol{x})\frac{1}{N}\sum_{k=1}^{N}(\eta_k)^2 + O(\sqrt{\Delta t}) \tag{9}$$

which is a random variable with mean D and variance inversely proportional to N.

In both cases the standard error is inversely proportional to $\sqrt{N}$, so we can control it by choosing a proper partitioning and by discarding the bins that contain a number of datapoints lower than a given threshold. In the following, I considered a threshold value of 100 datapoints per bin.

3 Attractors in the drift field

An interesting feature of the drift field is represented by attractors, so the first purpose in the analysis of the data is to locate them and identify their properties.

We can give a description of the attractors using the methods described in [1]. First of all, we imagine that the effective drift field $\boldsymbol{a}(\boldsymbol{x})$ is generated from a potential $U(\boldsymbol{x})$

$$\boldsymbol{a}(\boldsymbol{x}) = -\nabla U(\boldsymbol{x}). \tag{10}$$

In this setting a point attractor (x_0, y_0) is a local minima of the potential. If we assume the well to be circular, we can describe the potential to the lowest order in $(x - x_0), (y - y_0)$ in a small neighbourhood of the attractor by

$$U(x, y) = U_0 + \frac{W}{r^2}(x - x_0)^2 + \frac{W}{r^2}(y - y_0)^2 + \text{higher order terms} \tag{11}$$

where the weight W represents the potential difference at a distance r from (x_0, y_0), which can be used as an indicator of the strength of the potential well.

We can fit eq. 11 with respect to the empirical drift $\boldsymbol{a}(x, y)$ using the least squares method. The sum of squared residuals is

$$Res = \sum_{k=1}^{N} \|\boldsymbol{a}(x_k, y_k) + \nabla U(x_k, y_k)\|^2 \tag{12}$$

Considering r fixed we can determine the value of W that minimizes the error. Eq. 12 turns out to be easy to handle analytically, leading to

$$W = -\frac{r^2}{2}\frac{\sum_{k=1}^{N} a^x(x_k, y_k)x_k + a^y(x_k, y_k)y_k}{\sum_{k=1}^{N} x_k^2 + y_k^2} \tag{13}$$

3.1 Locating the attractors

In the previous section I have given a characterisation of the attractors, but we have yet to locate them in the space. To do that one could scan systematically over the data to find the regions matching the features defined above, as described in [1].

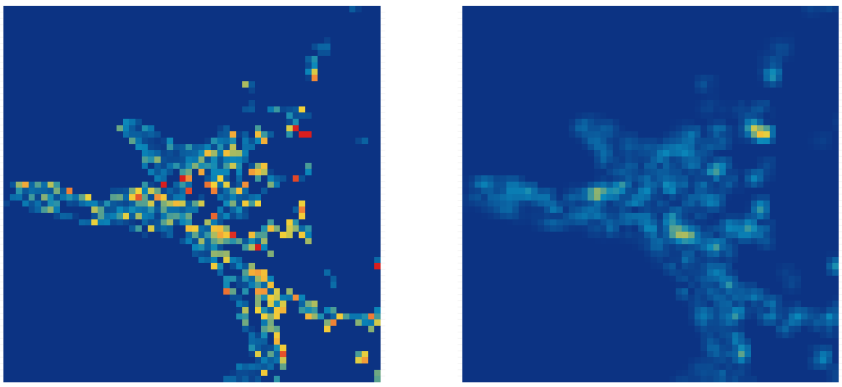
Here instead I tried a different approach by simulating the motion of particles in the effective drift field. Starting from a uniform distribution of the particles, I identify as attractors the bins showing the highest particle density after a sufficiently long simulation time.

To simulate the drift field we have to do some preliminary work. First of all, the trajectories do not cover the whole space, so there are regions where we are unable to estimate the drift. Moreover, we filtered out the bins that do not reach a minimum threshold of samples, adding more holes in the domain of the drift field.

We can identify two main kinds of artifacts which would hinder the simulation: bins such that all their 8 neighbours have an undefined field (isolated bins) and bins with undefined field surrounded by well defined bins (holes). Running the numerical simulation in this discontinuous field would not be effective, since the particles would get trapped in the isolated bins or holes.

To account for this, I manipulated the empirical field by filtering out the isolated bins and then smoothening the field to eliminate the holes. For the smoothening part I used an approximated Gaussian kernel defined as

$$K = \frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{14}$$



(a) Estimated field (b) Smoothened field

Figure 2: The modulus of the drift field before (a) and after (b) the application of the Gaussian filter (subregion of the total domain).

The simulated trajectories are obtained by forward Euler's scheme

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \tilde{\boldsymbol{a}}(\boldsymbol{x_n})\Delta t \tag{15}$$

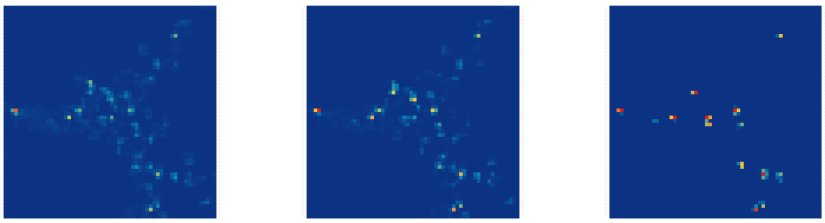
where $\tilde{\boldsymbol{a}}$ is the filtered and smoothed drift field.

We can choose a quite large simulation timestep as the dynamics is deterministic. The coarse grained field has a spatial resolution of l corresponding to the side of the square bins, so we are interested in having, on average, a step size close to the order of l. I chose

$$\Delta t = \frac{1}{10}\frac{l}{\langle|\tilde{\boldsymbol{a}}|\rangle} \tag{16}$$

where $\langle\cdot\rangle$ denotes the mean over the whole spatial domain of $\tilde{\boldsymbol{a}}$.

The results of the simulation are shown in figure 3. An example of the attractors localized is visible in figure 4.



(a) t = 5 (b) t = 10 (c) t = 500

Figure 3: The density of particles evolution during the simulation in a subregion of the total domain. After 500 steps the particles appear trapped into the attractors.

4 Channels

4.1 Channels as potential valleys

Characterizing the channels is a more challenging task. In fact, if attractors are features well localized in space, channels on the contrary are structures which spread on a moderately large region of space. It becomes difficult to relate them to local features of the field as we did with attractors.

Because of the non-locality, analyzing directly the drift field seems complicated. Instead, I assumed again the field to be generated taking the gradient of a scalar potential U (eq. 10). Putting aside the drift field $\boldsymbol{a}$ for one moment, we can try to find a description of channels with respect to the potential U.

From this viewpoint, channels are valleys in the surface defined by the potential function, i.e. they are one-dimension local minima of U. Features of this type for a two-dimensional scalar function are well known in computer

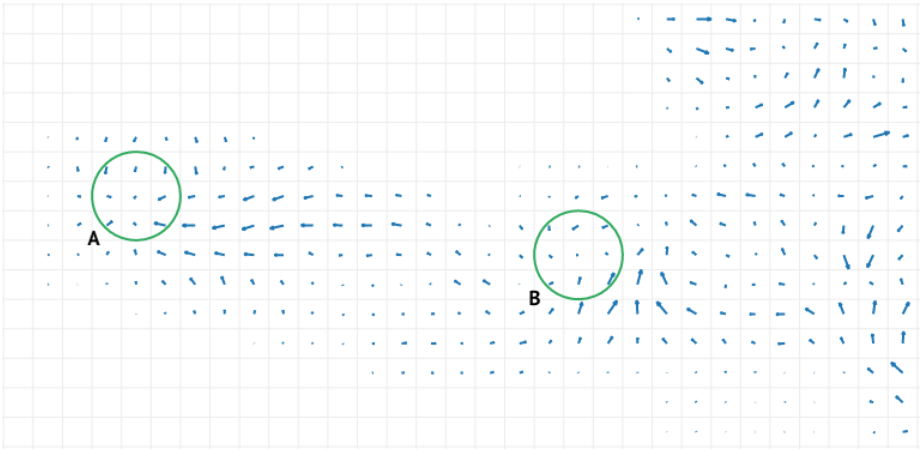


Figure 4: Two examples of attractors localized via the numerical simulation. Their weights are $W_A = 7,03 \times 10^{-5}$ and $W_B = 2,06 \times 10^{-5}$.

vision, where ridge (and valley) detection is used to identify the main axis of long objects. Reconstructing the potential function would allow to translate the problem from channel detection to ridge detection, thus inheriting from computer vision an already developed toolset for studying these structures.

4.2 Building the potential

The empirical field $\boldsymbol{a}$ is defined on the square lattice of side l formed by the centres of each bin S_l. I indicate with $\boldsymbol{a}_{i,j}$ the estimated value of the drift field at row i and column j of the lattice. I similarly define U on a square lattice formed by the middle points between the bin centres. In this way, eq. 10 can be represented by the discrete partial derivatives

$$a^x_{i,j} = -\frac{U_{i+1,j} - U_{i,j}}{l} \tag{17}$$

$$a^y_{i,j} = -\frac{U_{i,j+1} - U_{i,j}}{l} \tag{18}$$

Fixing an initial condition as $U_{k,m} = c$, it is possible to recover the values of U on the whole lattice iteratively using

$$U_{i+1,j} = -a^x_{i,j} l + U_{i,j} \tag{19}$$

$$U_{i,j+1} = -a^y_{i,j} l + U_{i,j} \tag{20}$$

For this method to work effectively we need a cluster of contiguous sites where the field $\boldsymbol{a}$ is defined and one initial condition for U is fixed. This is not the case for the rough estimated field, since it is full of holes and discontinuities that would prevent us from building a smooth potential. Note also that for isolated bins we have no possibility of reconstructing the potential. To overcome

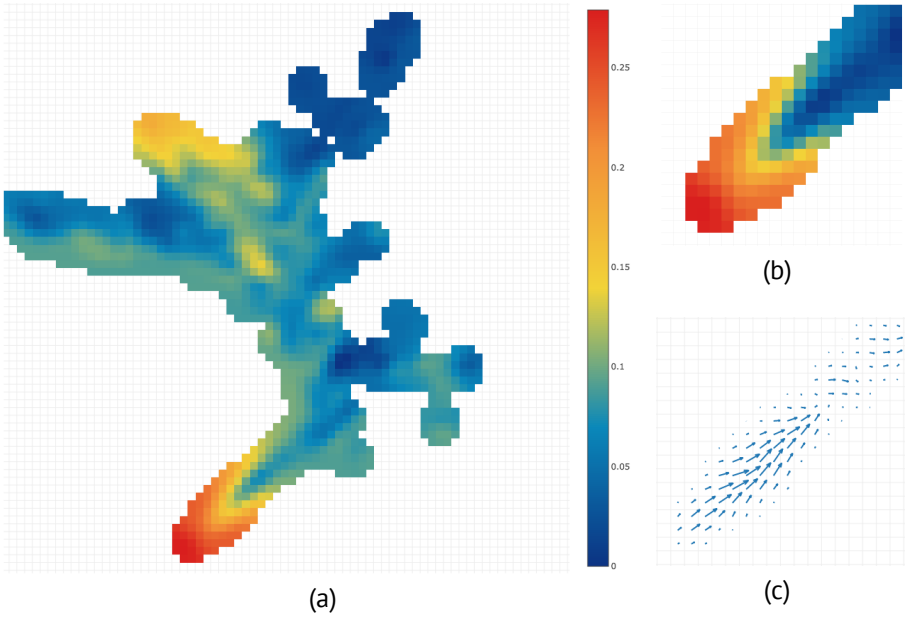


Figure 5: The reconstructed potential for a cluster of contiguous bins. A steep valley is easily identifiable in the bottom left tail of (a). The detail this region is shown in (b) and the corresponding field vectors in (c).

this problems I used the same approach described in section 3.1 to smoothen the empirical field by first eliminating isolated bins and then convolving with the kernel defined in eq. 14. After applying the convolution we obtain a drift field with large clusters of connected bins on which it is possible to reconstruct the potential.

Since the drift field is not perfectly conservative, integrating in one direction or another may produce different potential values for the same lattice site. To minimize these differences I built the potential exploring the adjacent sites in a breadth first order and taking the mean when multiple choices of the integration direction were available (detailed implementation can be found in the supporting code [4]).

The potential reconstructed with this method and the location of a valley are visibile in figure 5. A 3D view of the same potential is visible in fig. **??**.

Lastly, the knowledge of the potential offers a new way to locate attractors, as they are local minima of U. We can compare results of this method with those of the numerical simulation used in section 3.1. The finding is coherent: a comparison of the attractors of figure 4 is visible in figure 6.

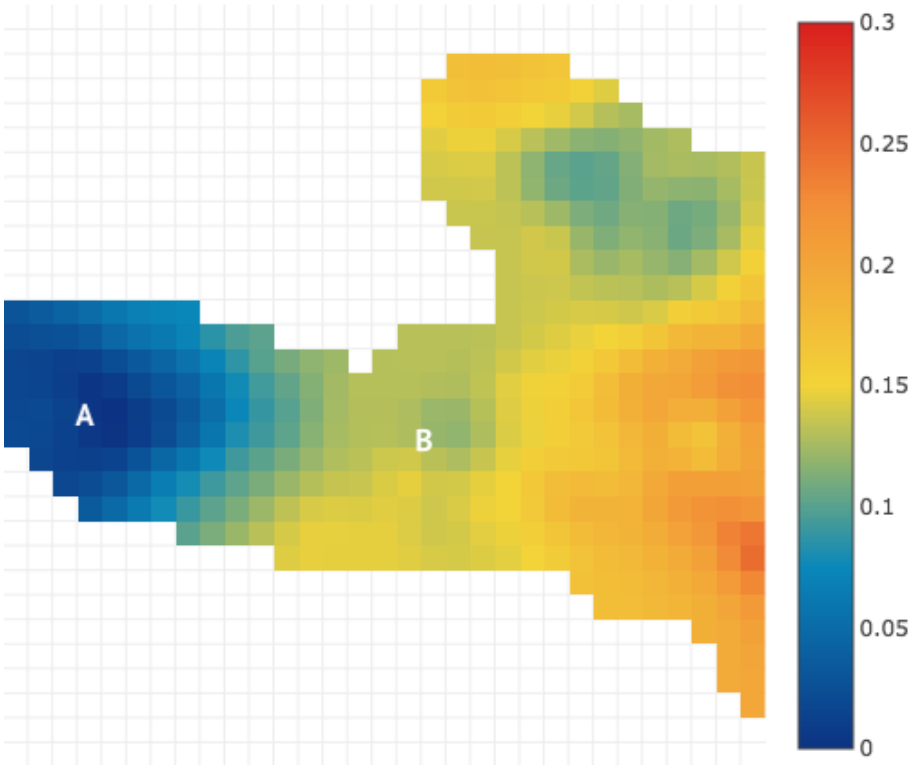


Figure 6: Localization of the two potential wells already shown in figure 4 using the potential. I recall their weights $W_A = 7,03 \times 10^{-5}$ and $W_B = 2,06 \times 10^{-5}$.

References

[1] Hozé N, Nair D, Hosy E, Sieben C, Manley S, et al. 2012. *Heterogeneity of receptor trafficking and molecular interactions revealed by superresolution analysis of live cell imaging.* PNAS 109:17052–57

[2] Hozé N, Holcman D. 2014. *Residence times of receptors in dendritic spines analyzed by stochastic simulations in empirical domains.* Biophys. J. 107:3008–17

[3] Schuss, Z. 2010. *Theory and applications of stochastic processes: an analytical approach.* Springer, New York.

[4] Supporting code: https://github.com/mattbit/spt