**Journal of Statistical Physics manuscript No.**
(will be inserted by the editor)

**Daan Crommelin**

# Estimation of space-dependent diffusions and potential landscapes from non-equilibrium data

Revised version, 16 July 2012. Accepted for publication.

**Abstract** In scientific topics ranging from protein folding to the thermohaline ocean circulation, it is useful to model the effective macroscopic dynamics of complex systems as noise-driven motion in a potential landscape. In this paper we consider the estimation of such models from a collection of short non-equilibrium trajectories between two points in phase-space. We generalize a recently introduced spectral methodology for the estimation of diffusion processes from timeseries, so that it can be used for non-equilibrium data. This methodology makes use of the spectral properties (leading eigenvalue-eigenfunction pairs) of the Fokker-Planck operator associated with the diffusion process. It is well suited to infer stochastic differential equations that give effective, coarse-grained descriptions of multiscale systems. The generalization to the non-equilibrium situation is illustrated with numerical examples in which potentials and diffusion coefficients are estimated from ensembles of short trajectories.

**Keywords** parameter estimation · diffusion process · non-equilibrium data · stochastic differential equation · subsampling

## 1 Introduction

The description of complex processes as noise-driven motion in a potential landscape has been an appealing concept in various areas of science, such as biophysics, chemistry and climate science [11,12,1,7,6]. In such a description, the effective macroscopic dynamics of the system of interest is modeled as a diffusion process in a reduced phase-space of one or two key variables (e.g. reaction coordinates). The drift of the process is given by the gradient of a potential. Finding the correct drift and diffusion coefficients of the effective stochastic process

D.T. Crommelin
CWI Amsterdam, Science Park 123, 1098 XG, Amsterdam, Netherlands
E-mail: Daan.Crommelin@cwi.nl

is challenging. In many practical cases, an analytical derivation of the effective drift and diffusion coefficients is not possible, so that they must be estimated from simulation or observation data.

Estimation of drifts and diffusions can be a difficult task, for several reasons. The dynamics in the reduced phase-space is a projection of the full dynamics, so that it will typically not be a Markov process on short timescales, only on long timescales. For the estimation, these short timescales must be avoided in order to avoid biased estimates [9,4], by using data sampled at time intervals that are sufficiently long. Estimation from such "low-frequency data" is challenging, because various popular estimators are only valid in the limit of small sampling intervals. Their results are affected by time discretization errors in the case of non-infinitesimal sampling intervals. Furthermore, the diffusion coefficient is often allowed to be coordinate-dependent, rather than restricted to be constant in space. This makes the estimation more complicated. Finally, the available data can consist of a collection of short non-equilibrium trajectories, rather than one long equilibrium trajectory.

In this paper, we will generalize a recently introduced methodology for the estimation of diffusion processes [2,4], so that it can be used for non-equilibrium data. This methodology makes use of the spectral properties (leading eigenvalue-eigenfunction pairs) of the Fokker-Planck operator associated with the diffusion process. It was shown to be suitable for estimation from low-frequency data, because it makes no time discretization errors [4]. By generalizing to the non-equilibrium situation, it can be used for estimating potentials and space-dependent diffusion coefficients from ensembles of short trajectories.

We consider 1-dimensional diffusion processes on a domain $\Omega \subseteq \mathbb{R}$, described by the stochastic differential equation (SDE)

$$dX_t = B(X_t)dt + \sqrt{2D(X_t)}\,dW_t\,. \tag{1}$$

Here, $B(x)$ is the drift function, $D(x)$ the (possibly space-dependent) diffusion coefficient, and $W_t$ a standard Wiener process. The SDE (1), and all other SDEs in this paper, are Ito SDEs. Typically, the drift is determined by the gradient of a potential, $B(x) = -\partial_x V(x)$, as in the case of overdamped Langevin dynamics.

In [2] and [4], a spectral procedure was introduced to estimate $B$ and $D$ from a sampling of the process in equilibrium. Thus, the focus was on estimating the drift and diffusion functions from a single, long equilibrium trajectory of the process. Here we are interested in estimation of $B$ and $D$ from non-equilibrium data, consisting of an ensemble of short trajectories. The probability distribution of the data in this ensemble can be very different from the invariant probability distribution associated with (1) (in fact, it may be the case that the observed process (1) has no invariant measure at all). The prototype example we consider is the case where the process starts at an initial state $x_i$ and the trajectory stops when the process reaches the final state $x_f$. In this paper we generalize the procedure presented in [4] so that it can be used to estimate $B$ and $D$ from a collection of such trajectories.

The key element is to view the ensemble of trajectories as a series of observations of a process with re-injection, as was recently proposed in [13]. Because the trajectories stop at $x_f$, there is an absorbing boundary at $x_f$. Concatenating these trajectories is equivalent to the situation where the process is instantaneously reinjected at $x_i$ whenever it hits $x_f$. We start with considering a fixed initial point $x_i$.

Later on we generalize to the situation where $x_i$ is not the same for each trajectory in the ensemble, but randomly sampled from a given distribution.

In section 2 we give a summary of the estimation procedure introduced in [2,4]. How this procedure can be adapted to deal with a reinjection process is discussed in section 3. Random initial points $x_i$ are considered in section 4. In section 5, we investigate estimation in a multiscale setting, where it is necessary to use data with long sampling intervals to be able to obtain unbiased estimates of the coarse-grained process. We finish with a conclusion in section 6.

## 2 Spectral estimation of diffusion processes: a summary

In this section we give a brief summary of the estimation methodology presented in [2,4]. We focus on 1-dimensional diffusion processes, but note that the methodology can also be used for multivariate diffusions (see [2,4] for more details and numerical examples).

We denote by $L$ the diffusion operator associated with (1):

$$L = B(x)\partial_x + D(x)\partial_{xx}\,, \tag{2}$$

and by $L^*$ its adjoint in $L_2(\Omega, dx)$:

$$L^* r = -\partial_x(Br) + \partial_{xx}(Dr) \tag{3}$$

for suitable functions $r(x)$. $L^*$ is also known as the Fokker-Planck operator. In absence of special boundaries, the probability density of the process (1), denoted $\rho(x,t)$, evolves according to the Fokker-Planck equation

$$\partial_t \rho = L^* \rho\,. \tag{4}$$

The formal solution of this equation is $\rho(x,t+\tau) = (P_\tau^* \rho)(x,t)$ with $P_\tau^* = \exp(\tau L^*)$.

Estimating $B(x)$ and $D(x)$ from timeseries of $X_t$ is challenging because the data almost always have a finite sampling interval. A timeseries of $X_t$ with sampling interval $\tau$ gives direct information about $P_\tau^*$, but not about $B$ and $D$. With only few exceptions, the time-$\tau$ transition probabilities embodied by $P_\tau^*$ are unknown functions of $B$ and $D$, causing great difficulties for estimation. Approximations such as $P_\tau^* \approx 1 + \tau L^*$ and $D(x) \approx (2\tau)^{-1}\mathbb{E}(\,(X_{t+\tau} - X_t)^2 \,|\, X_t = x)$ are frequently used, but these are only valid for small $\tau$ (formally, valid in the limit $\tau \to 0$). However, one cannot always use small $\tau$, for example because the available data have large $\tau$, or because of the need to avoid non-Markov effects at short timescales, as explained earlier (see also section 5 of [4] for a detailed discussion of this issue).

The semigroup structure $P_\tau^* = \exp(\tau L^*)$ implies that if $(\psi(x), \Lambda)$ is an eigenfunction-eigenvalue pair of $P_\tau^*$, then $(\psi(x), \lambda)$ with

$$\lambda = \tau^{-1} \log \Lambda \tag{5}$$

is an eigenpair of $L^*$. This relation is valid regardless of the value of $\tau$, and provides a way to estimate (the coefficients of) $L^*$ without making time discretization errors (errors due to finite $\tau$). The methodology in [2,4] is based on this property, and consists of two steps. First we estimate the leading eigenpairs $(\psi_k, \Lambda_k)$ of $P_\tau^*$. These give us, after application of (5), the leading eigenpairs $(\psi_k, \lambda_k)$ of $L^*$. The

pairs are ordered by decreasing eigenvalue, $1 = \Lambda_1 > |\Lambda_2| \geq |\Lambda_3| \geq ...$, implying that $\psi_1$ is the invariant probability density: $P_\tau^* \psi_1 = \psi_1$ and $L^* \psi_1 = 0$, cf. (4). In the second step, we reconstruct the functions $B$ and $D$ by minimizing the residuals $L^* \psi_k - \lambda_k \psi_k$ under variation of $B$ and $D$.

In [4], several possibilities to estimate the eigenpairs and minimize the residuals are discussed. For the estimation of eigenpairs, we focus here on what is named the "binning method" in [4]. For this method, the state space $\Omega$ is discretized into bins $\Omega_i$, $i = 1,..,M$, and $P_\tau^*$ is approximated by the set of probabilities to jump between bins over a time interval $\tau$. An alternative method is a Galerkin approximation in which the domain of the operator $P_\tau$ is projected onto a finite basis of smooth functions, see [4]. The binning method can be viewed as a discontinuous Galerkin method, where the discretization of $P_\tau^*$ is a $M \times M$ stochastic matrix whose elements are the transition probabilities $p_{ij} = \text{Prob}(X_{t+\tau} \in \Omega_j | X_t \in \Omega_i)$. The maximum likelihood estimator $\hat{P}$ for this matrix is easily calculated, by counting transitions between bins as observed in the data, and normalizing afterwards. The left eigenvectors and eigenvalues of $\hat{P}$ are the estimates of the (spatially discretized) eigenpairs of $P_\tau^*$. Using (5) we obtain estimates $(\hat{\psi}_k, \hat{\lambda}_k)$ of the eigenpairs of $L^*$. Overall, the binning method is straightforward to use, although some care has to be taken not to choose $M$ too small if $\tau$ is (very) small (see [4] for a discussion of this).

From $(\hat{\psi}_k, \hat{\lambda}_k)$, we infer $B$ and $D$ by minimizing the residuals $L^* \hat{\psi}_k - \hat{\lambda}_k \hat{\psi}_k$. It is possible to minimize the norms of the residuals, as was proposed in [2], but in order to do so, one must provide the first and second derivatives of $\hat{\psi}_k$ to calculate $L^* \hat{\psi}_k$. These derivatives are a major source of error, because sampling errors on $\hat{\psi}_k$ are strongly amplified by differentiation. Therefore a modified procedure was introduced in [4] that allows to avoid differentiation of $\hat{\psi}_k$. The residuals are integrated against test functions and then minimized. The test functions are smooth, $\sigma_j(x) \in C^2(\Omega)$, $j = 1,..,N_\sigma$, with known derivatives. We use the adjoint property to rewrite $\langle L^* \psi_k, \sigma_j \rangle$ as $\langle \psi_k, L\sigma_j \rangle$ (where $\langle .,. \rangle$ denotes the $L^2(\Omega, dx)$ inner product), so that we can use the (exact, error-free) derivatives of $\sigma_j$ rather than the (estimated) derivatives of $\hat{\psi}_k$.

Let $\theta$ be the set of unknown parameters in $B$ and $D$, so $L = L(\theta)$. The spectral estimator for $\theta$ is

$$\hat{\theta} = \arg\min_\theta \sum_{k=1}^{K} \sum_{j=1}^{N_\sigma} \alpha_{kj} |\langle \hat{\psi}_k, L(\theta)\sigma_j \rangle - \langle \hat{\lambda}_k \hat{\psi}_k, \sigma_j \rangle|^2 . \tag{6}$$

The $\alpha_{kj}$ are non-negative weights; we set $\alpha_{kj} = 1$ for all $k, j$ in this paper. The inner products $\langle \hat{\psi}_k, f \rangle$, where $f$ is $L(\theta)\sigma_j$ or $\sigma_j$, can be evaluated either by using numerical integration (quadrature) or by casting them as expectations:

$$\langle \psi_k, f \rangle = \int_\Omega dx\, \psi_1(x) \xi_k(x) f(x) = \mathbb{E}\, \xi_k(X_t) f(X_t) , \tag{7}$$

where $\xi_k$ is defined such that $\psi_k = \psi_1 \xi_k$ (recall that $\psi_1$ is the invariant density of the process).

A necessary condition for a unique minimum in (6) is $\dim(\theta) \leq K N_\sigma$. If $\theta \mapsto L(\theta)$ is a linear map, i.e. $L(c\,\theta) = c\,L(\theta)$ for any scalar $c$, an additional condition is $K \geq 2$. We note that the minimization problem is convex quadratic if $L$ is linear in $\theta$, making numerical solution of (6) straightforward.

## 3 Fokker-Planck equation for a reinjection process

The estimation procedure summarized in the previous section was developped with the equilibrium situation in mind, where the available data consists of a single long trajectory of the process (1) in equilibrium. If the data consists of an ensemble of non-equilibrium trajectories, each starting at $x_i$ and terminating at $x_f$, the procedure must be adapted. As was already explained in the introduction, a sequence of trajectories from $x_i$ to $x_f$ can be regarded as a single trajectory of the process (1) with an absorbing boundary at $x_f$ and instantaneous reinjection at $x_i$ [13]. The Fokker-Planck equation for such a reinjection process is

$$\partial_t \rho(x,t) = (L^*\rho)(x,t) + \delta(x-x_i)D(x_f)(\partial_x\rho)(x_f,t) \qquad x > x_f\,, \tag{8a}$$
$$\rho(x_f,t) = 0\ \forall t\,, \tag{8b}$$

where we have assumed $x_i > x_f$. The FP operator $L^*$ was defined in (3). The domain of the reinjection process is $\Omega = [x_f, \infty)$.

Observations from a long, single trajectory of (1) without absorbtion/reinjection would approach the equilibrium PDF $\rho_{\text{eq}}(x)$ defined by $L^*\rho_{\text{eq}} = 0$ (assuming there exists such equilibrium PDF). By contrast, observations from the concatenation of short trajectories from $x_i$ to $x_f$ approach the non-equilibrium stationary PDF $\rho_{\text{neq}}(x)$ defined by

$$(L^*\rho_{\text{neq}})(x) = -\delta(x-x_i)D(x_f)(\partial_x\rho_{\text{neq}})(x_f) \quad x > x_f\,, \tag{9a}$$
$$\rho_{\text{neq}}(x) = 0 \quad x \le x_f \tag{9b}$$

The probability flux associated with $\rho_{\text{neq}}$ is defined by $J_{\text{neq}} = B\rho_{\text{neq}} - \partial_x(D\rho_{\text{neq}})$. Because of the absorbtion/reinjection mechanism, the outflow of $J_{\text{neq}}$ at $x_f$ equals the inflow at $x_i$, hence the source term at the right hand side of (9a).

We define the modified Fokker-Planck operator $\tilde{L}^*$ as

$$(\tilde{L}^*\rho)(x,t) = (L^*\rho)(x,t) + \delta(x-x_i)D(x_f)(\partial_x\rho)(x_f,t)\,, \tag{10}$$

so the Fokker-Planck equation for the reinjection process is $\partial_t\rho = \tilde{L}^*\rho$. The corresponding finite-time transition operator is $\tilde{P}^*_\tau = \exp(\tau\tilde{L}^*)$. This operator, and its leading eigenpairs, can be estimated from the non-equilibrium data by binning, in the same way as explained in the previous section. The absorption/reinjection is treated as a jump from the bin containing $x_f$ to the bin containing $x_i$.

Because of the source term in $\tilde{L}^*$, a smart choice of test functions $\sigma_j$ is required to be able to infer $B$ and $D$ from the eigenpairs. Assume $(\tilde{\psi}_k, \tilde{\lambda}_k)$ is an eigenpair of $\tilde{L}^*$. The eigenfunction satisfies $\tilde{\psi}_k(\infty) = \tilde{\psi}'_k(\infty) = \tilde{\psi}_k(x_f) = 0$. It is easy to show that

$$\begin{aligned}\langle \tilde{L}^*\tilde{\psi}_k, \sigma_j\rangle &= \int_{x_f}^{\infty} dx\,(\tilde{L}^*\tilde{\psi}_k)(x)\sigma_j(x) \\ &= \langle \tilde{\psi}_k, L\sigma_j\rangle + D(x_f)\tilde{\psi}'_k(x_f)\big(\sigma_j(x_i) - \sigma_j(x_f)\big)\end{aligned} \tag{11}$$

with $L$ as defined in (2). By choosing $\sigma_j$ such that $\sigma_j(x_i) = \sigma_j(x_f)$ we have $\langle \tilde{L}^*\tilde{\psi}_k, \sigma_j\rangle = \langle \tilde{\psi}_k, L\sigma_j\rangle$. This implies that we can estimate the parameters of $L$ from the eigenpairs of $\tilde{L}^*$, using the methodology from [4]. Note that we do not

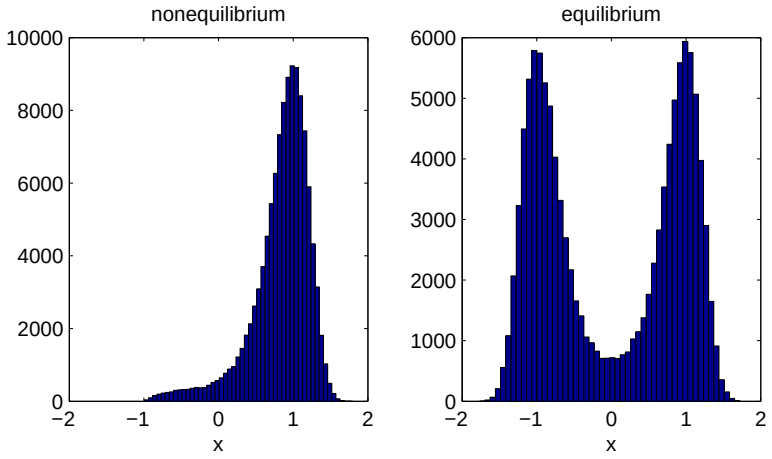


**Fig. 1** Distribution of data for the system with double-well potential and constant diffusion considered in example 3.1. The left panel shows a histogram of the data that results if the absorbtion/reinjection mechanism is active (i.e., non-equilibrium data). The right panel contains a histogram of the data that results if the absorption/reinjection mechanism is absent (i.e., equilibrium data).

need to estimate $\tilde{\psi}_k'$, $\tilde{\psi}_k''$, nor do we need to estimate $D(x_f)\rho_{\mathrm{neq}}'(x_f)$ by other means (as is required in [13]).

Summarizing: Let $(\hat{\tilde{\psi}}_k, \hat{\tilde{\lambda}}_k)$ be estimates of the leading eigenpairs of the modified Fokker-Planck operator $\tilde{L}^*$ in (10), obtained via the eigenpairs of $\tilde{P}_\tau^*$. The spectral estimator for the parameters $\theta$ of $L$ is

$$\hat{\theta} = \arg\min_{\theta} \sum_{k=1}^{K} \sum_{j=1}^{N_\sigma} |\langle \hat{\tilde{\psi}}_k, L\sigma_j \rangle - \langle \hat{\tilde{\lambda}}_k \hat{\tilde{\psi}}_k, \sigma_j \rangle|^2 , \tag{12}$$

provided the test functions satisfy

$$\sigma_j(x_i) = \sigma_j(x_f) . \tag{13}$$

### 3.1 Example: constant $D$

In this example, the drift in (1) is $B(x) = -V'(x)$ with double-well potential $V(x) = (1-x^2)^2$. The diffusion is constant (i.e., additive noise) with $D = 0.5$. The absorbing boundary is at $x_f = -1$, reinjection is at $x_i = 1$. Thus, trajectories start at the bottom of one potential well and end at the bottom of the other well. For this system, we generate data with constant sampling interval $\tau$ by numerically integrating the SDE (1), using the Euler-Maruyama scheme with time step $10^{-4}$. The absorption/reinjection mechanism is implemented by letting $X_t$ jump to $X_t + (x_i - x_f)$ whenever $X_t \le x_f$ during the integration.

Figure 1 shows the distributions of the data that are generated if the absorbtion/reinjection mechanism is present (left panel) and if it is absent (right panel).

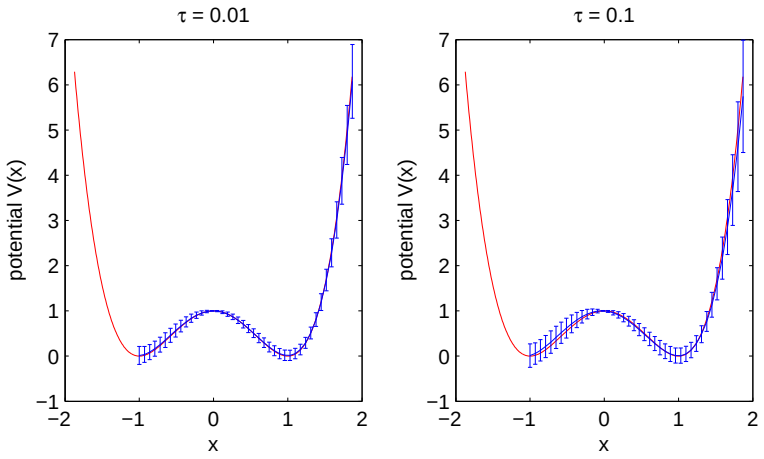


**Fig. 2** (Color online) Results for example 3.1 with double-well potential (cubic drift) and constant diffusion. The parameters of drift and diffusion are estimated from 100 different nonequilibrium datasets, each consisting of 100 trajectories starting at $x_i = 1$ and terminating at $x_f = -1$. Shown in blue are the mean of the estimated potentials and the corresponding errorbars (indicating the standard deviation). The curve in red is the true potential. The timeseries were sampled with interval $\tau = 0.01$ (left panel) and $\tau = 0.1$ (right panel). Results for the estimated diffusion coefficient are given in table 1.

The latter case is the equilibrium situation, added here for comparison. Its distribution differs strongly from the nonequilibrium situation that is the central topic of this study.

We fit a drift function of the form $B = b_1 + b_2 x + b_3 x^2 + b_4 x^3$ and constant diffusion $D$. This drift corresponds to a potential $V = c_0 - b_1 x - \frac{1}{2} b_2 x^2 - \frac{1}{3} b_3 x^3 - \frac{1}{4} b_4 x^4$, where $c_0$ is an irrelevant overall constant that we will set to $c_0 = 1$. The parameters $b_i$ and $D$ are estimated from timeseries that each contain 100 consecutive trajectories from $x_i$ to $x_f$ (i.e., nonequilibrium data). For the spectral estimation procedure, we use $M = 200$ bins and $K = 3$ eigenpairs. The test functions are linear combinations of the functions $x^2, x^3 - x, x^4$. They are obtained by Gram-Schmidt orthonormalization with respect to the observed distribution of $X_t$. Thus, if the data consists of $X_0, X_\tau, \ldots, X_{N\tau}$, the test functions are such that $(N+1)^{-1} \sum_{n=0}^{N} \sigma_i(X_{n\tau}) \sigma_j(X_{n\tau}) = \delta_{ij}$. They satisfy the requirement $\sigma_j(x_i) = \sigma_j(x_f)\ \forall j$ by construction.

The estimation is repeated with 100 different timeseries, using sampling intervals $\tau = 0.01$ as well as $\tau = 0.1$. Figure 2 shows the mean of the 100 estimated potentials, together with the true potential. The standard deviations (std) of the estimates are indicated by the errorbars of width 2 std . The mean of the estimates agrees very well with the true potential. For the larger value of $\tau$, the errorbars are larger, but the mean remains unbiased.

The mean and standard deviation of the estimated diffusion parameter $D$ are given in table 1. For comparison, we also include results for $D$ estimated using the

**Table 1** Estimates of the diffusion coefficient $D$ in example 3.1. Shown are the means and standard deviations of the estimates obtained with the spectral procedure and with the quadratic variation (QV) estimator (14). The QV estimates have lower variance than the spectral estimates, but are substantially biased for the longer sampling intervals ($\tau = 0.1$).

| | true | spectral, $\tau = 0.01$ | spectral, $\tau = 0.1$ | QV, $\tau = 0.01$ | QV, $\tau = 0.1$ |
|---|---|---|---|---|---|
| mean | 0.5 | 0.487 | 0.465 | 0.485 | 0.385 |
| std | | 0.059 | 0.092 | 0.002 | 0.006 |

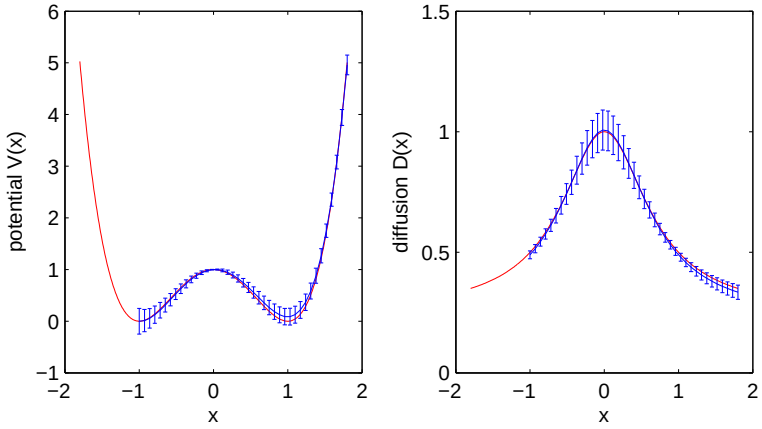


**Fig. 3** (Color online) Results for example 3.2 with double-well potential (cubic drift) and space-dependent diffusion. The mean and errorbars of the estimated potentials are shown in blue in the left panel; those of the estimated diffusions are in the right panel. The curves in red are the true potential and diffusion, respectively.

quadratic variation (QV) of the path:

$$\hat{D}_{\mathrm{qv}} = \frac{1}{2N\tau} \sum_{n=0}^{N} (X_{(n+1)\tau} - X_{n\tau})^2 \tag{14}$$

For the QV estimates, the jumps from $x_f$ to $x_i$ are omitted from the calculation (recall that these jumps are not part of the physical trajectories; they are the consequence of concatenating the individual trajectories from $x_i$ to $x_f$, therefore they must be left out of the QV estimate). As can be seen in the table, the QV estimates have lower variance than the spectral estimates, but are substantially biased for the longer sampling intervals ($\tau = 0.1$). This is because the QV estimator is consistent (unbiased) in the limit $\tau \to 0$; away from this limit it is affected by time discretization errors. As was discussed in section 2, the spectral procedure is not affected by these errors.

### 3.2 Example: space dependent $D$

In this example, the diffusion coefficient is no longer constant, but space dependent: $D(x) = \frac{1}{4}(1 + \frac{3}{1+2x^2})$. The potential is the same as in the previous example,

$V(x) = (1-x^2)^2$. We fit the drift $B(x) = b_1 + b_2 x + b_3 x^2 + b_4 x^3$ (as before) and diffusion $D(x) = d_1 + d_2/(1+2x^2)$. The sampling interval used here is $\tau = 0.01$. For numerical integration of the SDE we use the Milstein scheme with time step $10^{-4}$. All other details, such as number of trajectories and choice of test functions, are the same as in the previous example.

In figure 3 we have plotted the mean estimated potential and mean estimated diffusion, as well as their errorbars. They are in good agreement with the true potential and diffusion (both also shown).

## 4 Random $x_i$

If the starting point $x_i$ is not the same for each short trajectory, but instead drawn randomly from a distribution $\rho_0(x)$, the modified Fokker-Planck operator in (10) must be generalized to

$$(\tilde{L}^*\rho)(x,t) = (L^*\rho)(x,t) + \rho_0(x) D(x_f)(\partial_x \rho)(x_f,t)\,, \tag{15}$$

see [13]. As a result we have

$$\langle \tilde{L}^* \psi_k, \sigma_j \rangle = \langle \psi_k, L\sigma_j \rangle + D(x_f)\psi_k'(x_f)\big(\langle \rho_0, \sigma_j \rangle - \sigma_j(x_f)\big)\,. \tag{16}$$

Thus, we have to generalize the condition $\sigma_j(x_i) = \sigma_j(x_f)$ to $\langle \rho_0, \sigma_j \rangle = \sigma_j(x_f)$, in order to eliminate all boundary terms. Constructing such $\sigma_j$ is easy: if $\tilde{\sigma}_j$ is an arbitrary function, then $\sigma_j = \tilde{\sigma}_j + \alpha x$ with $\alpha = (\tilde{\sigma}_j(x_f) - \langle \rho_0, \tilde{\sigma}_j \rangle)/(\langle \rho_0, x \rangle - x_f)$ will satisfy this generalized condition.

### 4.1 Example: Gaussian distribution for $x_i$

We use the same model as in example 3.2 (i.e., space-dependent diffusion), but instead of keeping $x_i$ fixed at 1, we draw the $x_i$ from a Gaussian distribution with mean 1 and standard deviation 0.25. We keep $x_f$ fixed at -1. Results from the same kind of numerical experiment as in example 3.2 (100 timeseries with each 100 trajectories $x_i \to x_f$) are shown in figure 4. For numerical integration we use again the Milstein scheme with time step $10^{-4}$. The test functions are constructed as described just above, starting from $\tilde{\sigma}_j \in \{x^2, x^3 - x, x^4\}$. The results are comparable to the case with constant $x_i$ (figure 3), except that the errors on the estimated diffusion are somewhat higher here.

## 5 Multiscale systems, non-Markov data and low sampling frequency

The spectral estimation procedure from [2,4], generalized to non-equilibrium data in this paper, is suitable for situations where one wishes to model the coarse-grained dynamics of an observed multiscale system with a diffusion process. In such situations, the aim is typically to find effective models for the observed dynamics for selected (slow) variables. The effective model should be consistent with the coarse-grained (long timescale) features of these variables, but it can be inconsistent with the dynamics on short timescales. If the effective model is

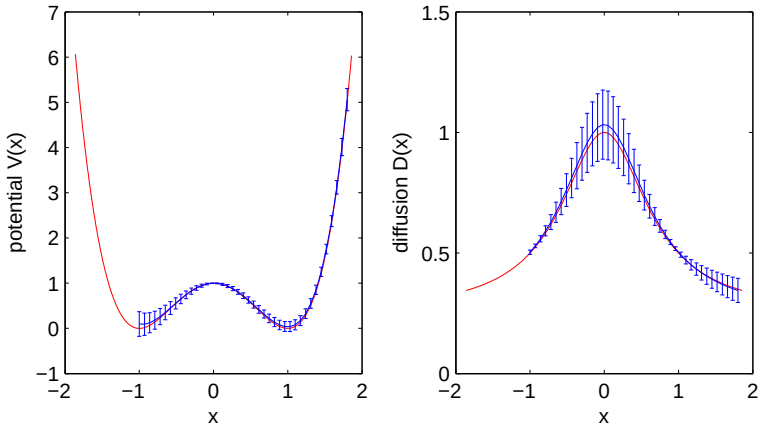


**Fig. 4** (Color online) Results for example 4.1, with random initial points $x_i$ for each trajectory. As in example 3.2, the potential has a double well (cubic drift) and the diffusion is space-dependent. The mean and errorbars of the estimated potentials are shown in blue in the left panel; those of the estimated diffusions are in the right panel. The curves in red are the true potential and diffusion, respectively.

inferred from observations, these short timescales should be avoided, implying that data with a low sampling frequency must be used for estimation. If $\tau$ is too short, the estimated drifts and diffusions can be strongly biased [9,8,4]. As already discussed in section 2, the spectral estimation procedure has no inherent time discretization error, unlike some other estimators (e.g. the QV estimator in (14)). Clearly, this is an advantage whenever data with long sampling intervals is used for estimation.

The question whether one can infer a correct coarse-grained model from observations of a multiscale system, can be systematically investigated in the context of multiscale diffusion processes, as was done in e.g. [9,8,4]. Under some mild assumptions, it can be shown rigorously that the slow dynamics of these multiscale processes converge to an effective (averaged or homogenized) diffusion process in the limit of large scale separation [10]. In some cases, one can derive analytical expressions for the effective drift and diffusion coefficients. These analytical results can be compared with results from estimation, for assessment of estimation procedures in a multiscale setting.

A detailed analysis of the spectral estimation procedure in the context of multiscale processes is given in section 5 of [4]. We will summarize a few key results of this analysis here. Starting point is the multiscale diffusion process $(X_t, Y_t) \in \Omega_x \times \Omega_y \subset \mathbb{R}^n \times \mathbb{R}^m$ with SDEs

$$dX_t = \left(\frac{1}{\varepsilon}F_1(X_t,Y_t) + F_0(X_t,Y_t)\right)dt + \alpha(X_t,Y_t)dW_t^x \tag{17a}$$

$$dY_t = \frac{1}{\varepsilon^2}G(X_t,Y_t)dt + \frac{1}{\varepsilon}\beta(X_t,Y_t)dW_t^y \tag{17b}$$

where $\varepsilon$ is a small parameter, and $W_t^x$ and $W_t^y$ are independent Wiener processes of dimension $n$ and $m$, respectively. In the limit $\varepsilon \to 0$, the slow variable $X_t$ converges

in law to the solution $\bar{X}_t$ of the effective (homogenized) SDE

$$d\bar{X}_t = \bar{F}(\bar{X}_t)dt + \bar{\alpha}(\bar{X}_t)dW_t^x, \tag{18}$$

provided the following assumptions hold: (i) if $X_t$ is fixed at $x$, the fast variable $Y_t$ is ergodic with unique invariant measure $\mu_x(y)$, and (ii) the centering condition

$$\int_{\Omega_y} \mu_x(dy)F_1(x,y) = 0 \tag{19}$$

is satisfied for all $x \in \Omega_x$. We will also assume that $\mu_x(y)$ admits a density $\rho_x(y)$, i.e. $\mu_x(dy) = \rho_x(y)dy$.

The Fokker-Planck operator $L^{h*}$ of the homogenized system (18) has leading eigenpairs that are asymptotically close to the leading eigenpairs of the Fokker-Planck operator $L^*$ of the full multiscale system (17). More precisely:

$$L^*\psi_k = \lambda_k\psi_k, \qquad L^{h*}u_k = \lambda_k^h u_k \tag{20a}$$
$$\psi_k(x,y) = u_k(x)\rho_x(y) + O(\varepsilon) \tag{20b}$$
$$\lambda_k = \lambda_k^h + O(\varepsilon) \tag{20c}$$

A similar results holds for the diffusion operators (or backward Fokker-Planck operators) $L$ and $L^h$, see [4].

If we can only observe the slow variable $X_t$ of (17), but not the fast variable $Y_t$, it is still possible to estimate the leading eigenpairs of the multiscale operators $L$ and $L^*$. However, the sampling interval should be large enough in this case. With only $X_t$ observed, we effectively observe the projected operator $\Pi P_\tau$ rather than $P_\tau$, where $P_\tau = \exp(\tau L)$, as in section 2, and $\Pi$ is the projection operator defined as $(\Pi h)(x) = \int \rho_x(y)h(x,y)dy$. If $\tau \gg \varepsilon^2$, the leading eigenpairs of $\Pi P_\tau$ and $P_\tau$ (and their adjoints) are again $O(\varepsilon)$ close. However, for the estimation we need the leading eigenvalues of $L$ rather than those of $P_\tau$ (or its projected counterpart $\Pi P_\tau$). Let $\Lambda_k^\Pi$ and $\Lambda_k$ be leading eigenvalues of $\Pi P_\tau$ and $P_\tau$, respectively, and define $\lambda_k^\Pi = \tau^{-1}\log\Lambda_k^\Pi$, $\lambda_k = \tau^{-1}\log\Lambda_k$, see (5). As mentioned, $\Lambda_k^\Pi - \Lambda_k = O(\varepsilon)$ if $\tau \gg \varepsilon^2$. Under the stricter requirement $\tau = \varepsilon^q$ with $0 \le q < 1$ we have $\lambda_k^\Pi - \lambda_k = O(\varepsilon^{1-q})$, and thus $\lambda_k^\Pi \to \lambda_k$ as $\varepsilon \to 0$.

Summarizing: if we observe the slow variable $X_t$ but not the fast variable $Y_t$ of (17), we can estimate the leading eigenpairs $(u_k^\Pi, \Lambda_k^\Pi)$ of $(\Pi P_\tau)^*$. Applying (5) gives the eigenpairs $(u_k^\Pi, \lambda_k^\Pi)$. If $\tau \gg \varepsilon$, these eigenpairs are asymptotically close the the leading eigenpairs $(u_k, \lambda^h)$ of the Fokker-Planck operator $L^{h*}$ of the homogenized system (18). Thus, we have $(u_k^\Pi, \lambda_k^\Pi) \to (u_k, \lambda_k^h)$ as $\varepsilon \to 0$, provided $\tau \gg \varepsilon$. This implies that with the spectral estimation procedure, we can infer the correct coarse-grained process (18) from the eigenpairs $(u_k^\Pi, \lambda_k^\Pi)$. This will be demonstrated in the following numerical examples.

### 5.1 Example: multiplicative red noise and Stratonovich corrections

We revisit the example with space-dependent diffusion and fixed $x_i$ (section 3.2), but replace the white noise by red noise. Thus, we have

$$dX_t = -V'(X_t)dt + \frac{1}{\varepsilon}\sqrt{2D(X_t)}\,Y_t dt\,, \tag{21a}$$

$$dY_t = -\frac{1}{\varepsilon^2}Y_t dt + \frac{1}{\varepsilon}dW_t\,, \tag{21b}$$

with $\varepsilon \ll 1$. As can be seen, the fast variable ("red noise") $Y_t$ is an Ornstein Uhlenbeck process. The dynamics of $X_t$ on $O(1)$ timescales can be well approximated by the homogenized SDE

$$dX_t = F(X_t)\,dt + \sqrt{2D(X_t)}\,dW_t \tag{22}$$

with

$$F(x) = -V'(x) + \tfrac{1}{2}\sqrt{2D(x)}\sqrt{2D(x)}' = -V'(x) + \tfrac{1}{2}D'(x)\,. \tag{23}$$

The term $D'(x)/2$ is the well-known Stratonovich correction, the contribution to the drift that arises if one interpretes an SDE in the Stratonovich sense and passes over to the corresponding Ito form. We refer to [10] for a detailed introduction of SDE homogenization and related techniques.

As in example 3.2, $V(x) = (1-x^2)^2$ and the space-dependent diffusion is $D(x) = \frac{1}{4}(1+\frac{3}{1+2x^2})$. Furthermore, we set $\varepsilon = 0.01$. We generate data with the multiscale system (21), using the Euler-Maruyama integration scheme with time step $10^{-5}$. The variable $X_t$ is subject to absorption at $x_f = -1$ immediately followed by reinjection at $x_i = 1$. There is no absorption/reinjection for $Y_t$. Using only observations of $X_t$ with sampling interval $\tau = 0.01$, we fit the SDE (22) with drift $B(x) = b_1 + b_2 x + b_3 x^2 + b_4 x^3 + b_5 x/(1+2x^2)^2$ and diffusion $D(x) = d_1 + d_2/(1+2x^2)$. Details of the estimation are identical to example 3.2.

In figure 5 we show the resulting mean and standard deviations (errorbars) of the estimated potentials and diffusions. The potential and diffusion functions predicted by homogenization theory are plotted as well (these correspond to the parameter values $(b_1,b_2,b_3,b_4,b_5,d_1,d_2) = (0,4,0,-4,-1.5,0.25,0.75)$).

### 5.2 Example: subsampling and biased estimates

Inferring the correct coarse-grained diffusion process from data of the underlying multiscale process is nontrivial, as was analysed in detail in [9,8,4]. When estimating the homogenized process (22) from observations of the slow variable $X_t$ in (21), one obtains strongly biased results if the sampling interval $\tau$ is too short, due to non-Markov effects at short timescales. In such a case, subsampling (skipping datapoints in order to increase $\tau$) is needed to avoid bias. However, if one uses an estimator that suffers from time discretization errors, the results deteriorate with growing $\tau$. The QV estimator (14) is an example. By contrast, the spectral procedure central to this paper is not affected by time discretization errors (however, if $\tau$ is very large, one needs very long timeseries in order to overcome sampling error).

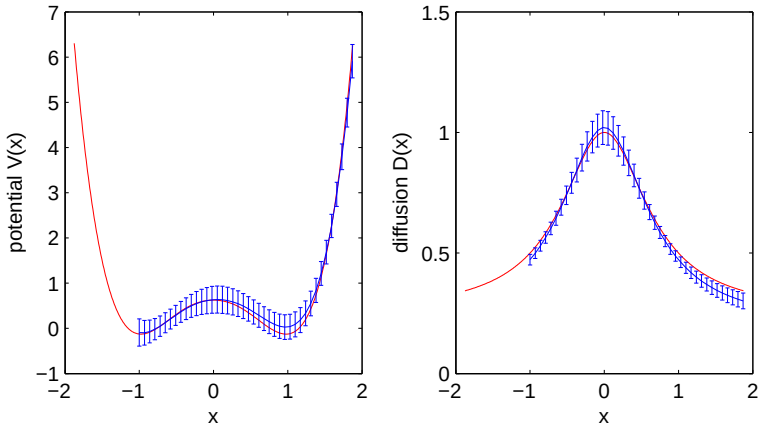


**Fig. 5** (Color online) Results for the multiscale example 5.1. The mean and errorbars of the estimated potentials are shown in blue in the left panel; those of the estimated diffusions are in the right panel. The curves in red are the potential and diffusion predicted by homogenization theory.

It allows one to do estimation at values of $\tau$ large enough to avoid the non-Markov effects, without being affected by time discretization errors.

To demonstrate this issue, we return to the example with constant $D$ (section 3.1), and replace the white noise by red noise, similar to the previous example. Thus, we consider the multiscale system (21) with $D = 0.5$ and $V(x) = (x^2-1)^2$. As before, the variable $X_t$ is subject to absorption at $x_f = -1$ immediately followed by reinjection at $x_i = 1$. There is no absorption/reinjection for $Y_t$. We set $\varepsilon = 0.01$. Using observations of $X_t$, we fit the SDE (22) with drift $B(x) = b_1 + b_2 x + b_3 x^2 + b_4 x^3$ and diffusion $D$. Details of the estimation are identical to example 3.1. We also estimate $D$ with the QV estimator (14).

For the estimation, we use timeseries consisting of 300 trajectories from $x_i$ to $x_f$. This is more than in previous examples, thereby enabling us to focus better on the $\tau$-dependent bias because of the smaller sampling error. The trajectories are generated with the Euler-Maruyama integration scheme with time step $10^{-5}$. They are sampled at intervals that are integer multiples of 0.0005, i.e. $\tau = h\,0.0005$ with $1 \le h \le 400$. Furthermore, the estimation is repeated using 10 different timeseries (each with 300 $x_i \to x_f$ trajectories). The mean of the estimates of $D$ is shown in figure 6, for all values of $\tau$. Both the estimates obtained with the spectral procedure and those from the QV estimator are visibly affected by non-Markov effects at the smallest values of $\tau$. In the range $0.005 < \tau < 0.1$, the mean of the spectral estimates is 0.48, only slightly below the correct value 0.5. The standard deviation in this range is about 0.015. For $\tau > 0.1$, sampling error becomes substantial (eventually, for very large values of $\tau$, the estimates tend to decrease again). By contrast, the QV estimates reach a maximum near the correct value of $D$ at $\tau \approx 0.005$ and decreases again for larger $\tau$, due to the time discretization error inherent to the QV estimator (14). The QV estimates have much smaller variance (about 0.0025) than the spectral estimates, but they are significantly biased for nearly all values of $\tau$.

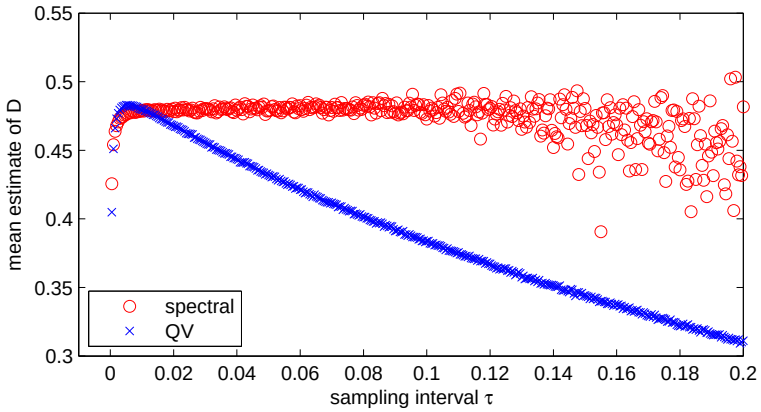


**Fig. 6** (Color online) Results for the multiscale example 5.2 with constant $D$. The spectral estimation procedure is affected by non-Markov effects at small values of $\tau$, but gives good results if larger $\tau$ are used. For $\tau > 0.1$, sampling errors become substantial. The quadratic variation (QV) estimator (14) is affected both by non-Markov effects at small $\tau$ and by time discretization errors at larger values of $\tau$. The correct diffusion coefficient is $D = 0.5$.

Rather similar results, using equilibrium data, can be found in the last example of [4]. A detailed analysis of the small $\tau$ limit is included there as well.

## 6 Conclusion

In this paper, we considered the estimation of diffusion processes from collections of short trajectories of the process. Each trajectory starts in the initial point $x_i$ and ends in the final point $x_f$. The distribution of the data (observations) of the trajectories can be far from the equilibrium distribution of the diffusion process. These non-equilibrium datasets can result from e.g. laboratory experiments or numerical experiments where a system is brought in a certain state $x_i$, after which it is released and observed until it reaches $x_f$.

We generalized a spectral estimation approach, introduced recently [2,4], so that it can be used for inferring diffusion processes from such non-equilibrium data. This was made possible by viewing the collection of trajectories from $x_i$ to $x_f$ as a single trajectory of a process with absorption at $x_f$, immediately followed by reinjection at $x_i$, as proposed in [13]. In sections 3 and 4 the estimation of potential functions and diffusion coefficients from non-equilibrium data with the spectral method was discussed and demonstrated with numerical examples, showing good results.

Because the spectral estimation procedure has no inherent time discretization error, it is a suitable method for situations where an effective, coarse-grained diffusion process must be estimated from data of a multiscale system. In these situations, data with long sampling intervals $\tau$ must be used in order to avoid biased results, posing problems for estimation methods that are only valid in the limit $\tau \to 0$. In section 5 we showed the favorable properties of the spectral method in

this respect, using examples with non-equilibrium data sampled from a multiscale system.

The discussion in this paper was limited to cases where $\tau$ is constant throughout the dataset. However, we expect that estimation from data with non-constant sampling intervals is well possible, following the approach proposed in [3]. Although the context in [3] was the estimation of Markov jump processes by the spectral method, we anticipate that the treatment of data with non-constant $\tau$ presented there will carry over to diffusion processes.

Furthermore, the focus throughout the paper was on inference of 1-dimensional processes. Clearly, generalization to processes with $\dim \geq 2$ will be useful for various practical applications. For equilibrium data, several 2-dimensional numerical examples were already treated in [2] and [4], with positive results. For the case of non-equilibrium data as considered here, generalization to higher dimensions is more complicated, because it involves generalization of the stopping point $x_f$ to a hypersurface. We leave this for future study.

Finally, we note that the spectral procedure as presented here is a method for parametric estimation. The potentials and diffusion functions are expanded in a finite number of basis functions, requiring estimation of the expansion coefficients. We intend to investigate the extension of our procedure to nonparametric estimation in future work.

**Acknowledgements** The author thanks Eric Vanden-Eijnden for helpful discussions. This research was supported by the Netherlands Organization for Scientific Research (NWO) through the research cluster Nonlinear Dynamics of Natural Systems.

## References

1. Best, R.B. and Hummer, G.: Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. USA*, **107**, 1088-1093 (2010).
2. Crommelin, D.T. and Vanden-Eijnden, E.: Reconstruction of diffusions using spectral data from timeseries. *Comm. Math. Sci.*, **4**, 651-668 (2006).
3. Crommelin, D.T. and Vanden-Eijnden, E.: Data-based inference of generators for Markov jump processes using convex optimization. *Multiscale Model. Simul.*, **7**, 1751-1778 (2009).
4. Crommelin, D.T. and Vanden-Eijnden, E.: Diffusion estimation from multiscale data by operator eigenpairs. *Multiscale Model. Simul.*, **9**, 1588-1623 (2011).
5. Hummer, G.: Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, **7**, 34 (2005).
6. Kwasniok, F. and Lohmann, G.: Deriving dynamical models from paleoclimatic records: Application to glacial millennial-scale climate variability. *Phys. Rev. E*, **80**, 066104 (2009).
7. Lucarini, V., Faranda, D. and Willeit, M.: Bistable systems with Stochastic Noise: Virtues and Limits of effective one-dimensional Langevin equations. *Nonlin. Processes Geophys.*, **19**, 9-22 (2012).
8. Papavasiliou, A., Pavliotis, G.A. and Stuart, A.M.: Maximum likelihood drift estimation for multiscale diffusions. *Stochastic Process. Appl.*, **119**, 3173-3210 (2009).
9. Pavliotis, G.A. and Stuart, A.M.: Parameter estimation for multiscale diffusions. *J. Statist. Phys.*, **127**, 741-781 (2007).
10. Pavliotis, G.A. and Stuart, A.M.: *Multiscale methods. Avergaging and homogenization*. New York: Springer (2008).
11. Socci, N.D., Onuchic, J.N. and Wolynes, P.G.: Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.*, **104**, 5860 (1996).
12. Yang, S., Onuchic, J.N. and Levine, H.: Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, **125**, 054910 (2006).

13. Zhang, Q., Brujić, J. and Vanden-Eijnden, E.: Reconstructing free energy profiles from nonequilibrium relaxation trajectories. *J. Statist. Phys.*, **144**, 344-366 (2011).