RECONSTRUCTION OF DIFFUSIONS USING SPECTRAL DATA FROM TIMESERIES *

DAAN CROMMELIN † AND ERIC VANDEN-EIJNDEN ‡

Abstract. A numerical technique for the reconstruction of diffusion processes (diffusions, in short) from data is presented. The drift and diffusion coefficients of the generator of the diffusion are found by minimizing an object function which measures the difference between the eigenspectrum of the operator and a reference eigenspectrum. The reference spectrum can be obtained, in discretized form, from timeseries through the construction of a discrete-time Markov chain. Discretization of the Fokker-Planck operator turns minimization of the object function into a quadratic programming problem on a convex domain, for which well-established solution methods exist. The technique is a generalization of a reconstruction procedure for continuous-time Markov chain generators, recently developed by the authors. The technique also allows to derive the coefficient in the homogenized diffusion for the slow variables in system with multiple timescales.

Key words. diffusions; stochastic differential equations; parameter estimation.

AMS subject classifications: 60J60; 60H10; 62G05; 62M10

1. Introduction

The reconstruction of diffusions (that is, stochastic differential equations) from data is a topic of practical importance in many fields, including molecular dynamics, atmosphere-ocean sciences, and econophysics. To tackle this problem people [1, 2, 3, 4, 5, 6] have employed a method based on the statistical definitions of drift and diffusion: if $b(x) \in \mathbb{R}^n$ and $a(x) \in \mathbb{R}^n \times \mathbb{R}^n$ are, respectively the drift vector and the diffusion tensor of a diffusion with sample path $X_t \in \Omega \subseteq \mathbb{R}^n$, then

$$b(x) = \lim_{\Delta t \to 0} \mathbb{E}_x(X_{\Delta t} - x), \qquad a(x) = \lim_{\Delta t \to 0} \mathbb{E}_x(X_{\Delta t} - x) \otimes (X_{\Delta t} - x) \tag{1.1}$$

where $\mathbb{E}_x$ denotes expectation conditional on $X_0 = x$. The advantage of these definitions is that they offer a very simple and direct way to determine $b(x)$ and $a(x)$ locally from the time-series. However, in practical applications, the use of these definitions may be problematic, for several reasons: the temporal resolution of the available time-series can be rather coarse (i.e., far from the $\Delta t \to 0$ limit); the implicit assumption that the data has an underlying diffusion is often not justified in the limit $\Delta t \to 0$; the conditional expectation on $X_0 = x$ may be difficult to enforce since the process X_t must be binned and the binning may be coarse; and, finally, the definitions above are very rigid and do not allow to include some *a priori* information on $b(x)$ and $a(x)$ one may have.

Other approaches to reconstruction of drift and diffusion include methods based on maximum likelihood estimators [7, 8, 9]. These are well suited for the parametric reconstruction of the drift and the diffusion (i.e. when their functional form is known up to some parameters that need to be determined) from time-series that are sampled continuously (in contrast to being sampled at discrete lags). However, these approaches are not so well-suited for non-parametric estimation. Problems also arise when the temporal resolution of the available time-series is coarse or the data has no exact underlying diffusion.

*April 28, 2006, submitted to Comm. Math. Sci.

†Courant Institute, New York University (crommelin@cims.nyu.edu)

‡Courant Institute, New York University and Department of Mathematics, University of California, Berkeley (eve2@cims.nyu.edu)

In this paper we propose an alternative method of reconstructing the drift vector and diffusion tensor from data, one that makes use of information about the eigenspectrum of the generator of the diffusion. Recall that the generator L of the diffusion with drift vector $b(x)$ and diffusion tensor $a(x)$ is the elliptic operator

$$L = b(x) \cdot \nabla + \tfrac{1}{2} a(x) : \nabla\nabla \tag{1.2}$$

A reconstruction procedure can be considered successful if it results in a generator L whose eigenspectrum sufficiently resembles a reference eigenspectrum. The main contribution of the present paper is to show how to perform this reconstruction via the minimization of a quadratic, convex object function. The reference spectrum used in this object function can be obtained in different ways from the data in a preliminary step which is independent from the reconstruction itself: for instance (and this is the approach that we will follow here) one can construct a time-discrete Markov chain from the given timeseries after suitable discretization (or binning) of the state-space and calculate the spectrum of the resulting transition probability matrix P. Since P can be calculated at many different timelags Δt, we can avoid taking the limit $\Delta t \to 0$. The (spatially discrete) spectrum of P can then be used to estimate the spectrum of the generator L for instance by interpolation, as explained below.

The use of spectral information for the identification of diffusion processes has been employed before, for example by [10]. The more general issue of reconstructing coefficient functions of differential equations from spectral data is a well-known problem in the field of inverse problems, see for example [11, 12]. New in our study is the proposal of a numerical scheme for the approximation of drift and diffusion coefficients from spectral data, thereby going beyond purely analytical methods. Moreover, the scheme has several desirable properties: it is very versatile as it allows for non-parametric estimation but any *a priori* information on $b(x)$ and $a(x)$ can be incorporated straightforwardly; it can handle data that does not have an exact underlying diffusion process (e.g. being non-Markov on short time intervals); it is not limited to scalar diffusions, but can also deal with higher-dimensional diffusions; and, finally, the central element in the algorithm is the minimization of a quadratic, convex object function, a problem usually referred to as quadratic programmming for which efficient numerical methods exist.

In [13], a numerical scheme is proposed to find optimal continuous-time Markov chain generators using spectral information obtained from timeseries. The scheme discussed in this paper is a generalization of the algorithm in [13] to diffusions. The possibility of diffusion reconstruction was mentioned briefly in [13]; the present study focusses entirely on diffusion processes. Some subtle, but important, differences with the procedure in [13] are a) the Markov chain generator constraints are replaced by the constraint of positive semi-definite diffusion matrices, and b) spatial discretization of the problem is necessary for numerical implementation of the scheme for diffusions, but is not an issue for the Markov chain generator reconstruction. It is far from obvious that an approach to reconstruction that was shown to work well for continuous-time Markov chains [13] will also work for diffusion reconstruction. In this paper we will show that the approach gives good results for diffusions as well.

In section 2 the object function that is central to our approach is presented. In section 3 we discuss practical implementation issues, in particular how to estimate the reference spectrum from the data and then how to carry on the reconstruction step by minimization of the object function. Numerical examples in 1 and 2 dimensions are presented in section 4. In section 5 we discuss a generalization to systems with

multiple timescales, in which the data is generated by a diffusion with fast and slow components but what is really sought are the effective drift and diffusion coefficients appearing in the homogenized equation for the slow component alone. We show that our method can be straightforwardly generalized to handle this case in a very suitable way. Finally some conclusions are given in section 6.

2. Variational formulation of the reconstruction problem

Consider a diffusion with generator L over the state-space $\Omega \subseteq \mathbb{R}^n$. Assume that the process is ergodic and, for simplicity, assume also that the spectrum of L is discrete. Denote by $\{\psi_k(x), \phi_k(x), \lambda_k\}_{k\in\mathbb{N}}$ the set of (left and right) eigenfunctions and eigenvalues,

$$L^*\psi_k = \lambda_k \psi_k, \qquad L\phi_k = \lambda_k \phi_k, \tag{2.1}$$

where L^* denotes the adjoint of L in $L_2(\Omega, dx)$. It is always possible to properly normalize the eigenfunctions (here $\bar{\psi}_k$ is the complex conjugate of ψ_k)

$$\int_\Omega \bar{\psi}_k(x)\phi_l(x)dx = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

and order them according to decreasing Re λ_k (notice that λ_k, ψ_k, ϕ_k can be complex). The first mode, $k = 1$, contains the equilibrium probability density $m(x)$ of the process: $\psi_1(x) = m(x)$, $\phi_1(x) = 1$, $\lambda_1 = 0$; all subsequent eigenvalues have Re $\lambda_k < 0$ by ergodicity.

From (1.2), the operator L is determined by $b(x)$ and $a(x)$, i.e. one can write $L = L(b,a)$. The question that we address next is the following: how should one choose $b(x)$ and $a(x)$ so that the corresponding operator $L(b,a)$ has an eigenspectrum $\{\psi_k(x), \phi_k(x), \lambda_k\}$ that resembles a given reference spectrum $\{\tilde{\psi}_k(x), \tilde{\phi}_k(x), \tilde{\lambda}_k\}$ as closely as possible? Since there is no guarantee that there exist b and a such that $L(b,a)$ has *exactly* the set $\{\tilde{\psi}_k(x), \tilde{\phi}_k(x), \tilde{\lambda}_k\}$ as its spectrum, an approximation procedure is needed. Our approach is to minimize an object function that measures how close $L(b,a)$ is to having $\{\tilde{\psi}_k(x), \tilde{\phi}_k(x), \tilde{\lambda}_k\}$ as its spectrum.

We propose the following object function:

$$E = \sum_{k=1}^{K} \left(\alpha_k \|L^*\tilde{\psi}_k - \tilde{\lambda}_k \tilde{\psi}_k\|^2 + \beta_k \|L\tilde{\phi}_k - \tilde{\lambda}_k \tilde{\phi}_k\|^2 + \gamma_k |\langle \tilde{\psi}_k, L\tilde{\phi}_k\rangle - \tilde{\lambda}_k|^2 \right) \tag{2.3}$$

where $\|\cdot\|^2$ denotes the L_2 norm on Ω, $\|f\|^2 = \int_\Omega |f(x)|^2 dx$, and $\langle\cdot,\cdot\rangle$ is the associated inner product, $\langle f(x), g(x)\rangle = \int_\Omega \bar{f}(x) g(x) dx$. In (2.3), $K \in \mathbb{N}$ is a parameter determined by the highest mode one is able to estimate reliably. The parameters $\alpha_k, \beta_k, \gamma_k \in (0,\infty)$ are weights which permits to put emphasis a certain modes which we may find more important than others like e.g. the equilibrium probability density function, $m(x) = \psi_1(x)$.

It is possible to do the reconstruction using only the generator and not its adjoint (set $\alpha_k = \gamma_k = 0$ for all k) or only the adjoint and not L itself (set $\beta_k = \gamma_k = 0$ for all k), provided one supplies enough reference eigenfunctions. For example, for reconstruction of a 2-dimensional diffusion one needs in principle a set of 5 reference eigenfunctions, which can be $\psi_1, \psi_2, \psi_3, \phi_2, \phi_3$ but also $\psi_1, \ldots, \psi_5$. Since estimates of the leading eigenmodes from data are the most reliable, the former set is preferable over the latter. Therefore both the operator and its adjoint are used in (2.3).

The object function (2.3) is to be minimized under variation of L, i.e. under variation of $b(x)$ and $a(x)$, under the following additional constraint that $a(x)$ is a non-negative definite tensor: for any $v\in\mathbb{R}^n$, we must have

$$a(x):vv\geq 0. \tag{2.4}$$

Thus (2.3) is positive semi-definite and convex (at least if K is large enough), as well as quadratic in b and a, and it must be minimized on a convex domain consistent with (2.4). This is an advantage since it guarantees that the minimum of (2.3) is unique and it can be found by well-established solution methods (see section 3). In principle, instead of $L_2(\Omega,dx)$ other norms can be used in (2.3), like e.g. $L_2(\Omega,m(x)dx)$; as long as they are quadratic norms they will result in object functions that are still quadratic in b and a. Similarly, if some *a priori* information is known about $b(x)$ and $a(x)$, for instance say, $b(x)=b_1f_1(x)+\cdots+b_Nf_N(x)$ and $a(x)=a_1g_1(x)+\cdots+a_Ng_N(x)$ where $f_1,\ldots,g_N$ are known functions, (2.3) reduces to a quadratic object function in $(b_1,\ldots,a_N)$. Many other variations are possible, and lead to numerical problems which can be tackled by straightforward generalization of the method used in section 3, but here we will stick to (2.3) with $L_2(\Omega,dx)$ and non-parametric estimation of $b(x)$ and $a(x)$.

3. Numerical implementation

To minimize (2.3) in practice, one is faced with two preliminary tasks. First, we must represent the functions $b(x)$ and $a(x)$ by a finite number of quantities to be minimized over, which, in effect, amounts to discretizing (2.3). Second, we must estimate the eigenfunctions $\tilde\psi_k$ and $\tilde\phi_k$ and the eigenvalues $\tilde\lambda_k$ for $k=1,\ldots,K$. Even though the estimation of $\tilde\psi_k$, $\tilde\phi_k$ and $\tilde\lambda_k$ must obviously be consistent with the discretization of (2.3), it is important to realize that these two tasks are separated ones.

Let us consider the discretization of (2.3) first, assuming that $\tilde\psi_k$, $\tilde\phi_k$ and $\tilde\lambda_k$ for $k=1,\ldots,K$ are known. In essence, this amounts to representing $b(x)$ and $a(x)$ on some appropriate basis of functions, then truncating the series at some order. This can be done using the Fourier representation, wavelets of various sorts, etc. or, as will be done here, by representation on a predefined grid (notwithstanding this representation, we maintain that the reconstruction remains nonparametric, since we do not impose a specific functional form such as expansion in a few Fourier modes). For simplicity, let us consider the one dimensional case when $\Omega\in[0,1]$. Given $N\in\mathbb{N}$, let $\Delta x=1/N$, and define

$$b_i=b(i\Delta x),\qquad a_i=a(i\Delta x),\qquad i=1,\ldots N \tag{3.1}$$

Within this representation of $b(x)$ and $a(x)$ it is then natural to discretize $\psi_k(x)$ and $\phi_k(x)$ accordingly, and therefore approximate (2.3) by

$$\begin{aligned}\tilde E=\sum_{k=1}^{K}\Big(&\alpha_k\sum_{i=1}^{N}\Big|-D_i(b\tilde\psi_k)+\tfrac12 D_i^2(a\tilde\psi_k)-\tilde\lambda_k\tilde\psi_{k,i}\Big|^2\\&+\beta_k\sum_i\Big|b_iD_i\tilde\phi_k+\tfrac12 a_iD_i^2\tilde\phi_k-\tilde\lambda_k\tilde\phi_{k,i}\Big|^2\\&+\gamma_k\Big|\sum_i\tilde\psi_{k,i}\Big(b_iD_i\tilde\phi_k+\tfrac12 a_iD_i^2\tilde\phi_k\Big)-\tilde\lambda_k\Big|^2\Big).\end{aligned} \tag{3.2}$$

where $\tilde\phi_{k,i}=\tilde\phi_k(i\Delta x)$, $\tilde\psi_{k,i}=\tilde\psi_k(i\Delta x)$, and D_i and D_i^2 denote respectively the first

and second finite difference operators:

$$D_i f = \frac{f_{i+1} - f_{i-1}}{2\Delta x}, \qquad D_i^2 f = \frac{f_{i+1} + f_{i-1} - 2f_i}{\Delta x^2}, \tag{3.3}$$

Assuming periodic boundary conditions the index $i = N+1$ is to be identified with $i = 1$ and $i = 0$ with $i = N$. (3.2) is to be minimized subject to the following N inequality constraints (corresponding to (2.4)):

$$a_i \geq 0 \qquad \forall i = 1, \ldots, N. \tag{3.4}$$

Notice that (3.2) can be rewritten as

$$E = \langle v, Hv \rangle + \langle v, F \rangle + E_0 \tag{3.5}$$

where E_0 is a constant, H is a positive definite symmetric matrix and v is the $2N$-dimensional vector containing drift and diffusion: $v = (b_1, \ldots, b_N, a_1, \ldots, a_N)$. This formulation shows explicitly that minimization of the object function (3.2) is a $2N$-dimensional quadratic programming problem with the constraint in (3.4).

The generalization of (3.1), (3.2), (3.3) and (3.4) to situations in higher dimension possibly with different boundary conditions is straightforward so we shall omit writing it down explicitly. The only point worth mentioning is the equivalent of the constraint (3.4). In 2 dimensions, if $a_{i,j} = a(i\Delta x, j\Delta y)$, it becomes

$$\det a_{i,j} \geq 0 \qquad \text{trace } a_{i,j} \geq 0, \qquad \forall i, j = 1, \ldots, N, \tag{3.6}$$

and similarly in higher dimension. Constraints like (3.6) complicate matters slightly because they are nonlinear (although they still define a convex domain). They can be simplified if one assumes that $a(x)$ is diagonal, in which case it simply reduces to the requirement that each diagonal entry of $a_{i,j}$ be non-negative.

Once we have decided to use (3.2), the next task is to estimate $\tilde{\phi}_{k,i} = \tilde{\phi}_k(i\Delta x)$, $\tilde{\psi}_{k,i} = \tilde{\psi}_k(i\Delta x)$ and $\tilde{\lambda}_k$ for $k = 1, \ldots, K$. Here too, several strategies are possible and we will focus on the simplest for illustration. The idea is to use the strategy developed in [13] in the context of reconstruction of generators of continuous-time Markov chain on a discrete state space. Suppose that we bin Ω into $M \in \mathbb{N}$ non-overlapping bins, B_m, $m = 1, \ldots, M$, such that $\cup_m B_m = \Omega$. For instance, in the example above where $\Omega = [0,1]$, we may take $B_m = [(m-1)/M, m/M]$. Notice that the number of bins may be different from the number of discretization points in (3.1), i.e. we may have $M \neq N$ (and typically, $M < N$ because efficient sampling is costly if the bins are small; as will become clear shortly, we should however take $M > K$). Once the bins have been defined, the sampling path of the process can be discretized accordingly, by taking

$$Z_t = \text{index } m \text{ of the bin such that } X_t \in B_m. \tag{3.7}$$

Associated with with the discrete variable Z_t, we can then define an $M \times M$ transition probability matrix P as

$$P_{mm'} = \frac{\sum_{j=1}^{N_t} \mathbf{1}(Z_{j\Delta t} = m) \mathbf{1}(Z_{(j+1)\Delta t} = m')}{\sum_{j=1}^{N_t} \mathbf{1}(Z_{j\Delta t} = m)} \tag{3.8}$$

where $\mathbf{1}(z = m) = 1$ if $z = m$ and 0 otherwise, $T = N_t \Delta t$ is the length of the observed time-series, and Δt is the discrete lag at which this time-series in sampled. (3.8)

exactly is the estimator for $e^{\Delta t L}$ that is used in [13]. Of course, in [13], the generator L itself is an $M \times M$ matrix, whereas here L is an operator. Nevertheless, the first K eigenvectors $\hat{\phi}_{k,m}$ and $\hat{\psi}_{k,m}$ and eigenvalues Λ_k, $k=1,\ldots,K$, of P can be used to estimate the ones needed in (3.2). From Λ_k, we obtain

$$\tilde{\lambda}_k = \Delta t^{-1} \log \Lambda_k \tag{3.9}$$

From $\hat{\phi}_{k,m}$ and $\hat{\psi}_{k,m}$, we take

$$\tilde{\phi}_k(x_m) = \hat{\phi}_{k,m}, \qquad \tilde{\psi}_k(x_m) = \hat{\psi}_{k,m} \tag{3.10}$$

where $x_m \in B_m$, $m=1,\ldots,M$ are reprentative points in each bins. From (3.10), $\tilde{\phi}_{k,i}$, $\tilde{\psi}_{k,i}$ can then be obtained by suitable interpolation. In the example below, we actually take a slightly more sophisticated approach to interpolate $\tilde{\phi}_{k,i}$ and $\tilde{\psi}_{k,i}$ from $\hat{\phi}_{k,m}$ and $\hat{\psi}_{k,m}$ to eliminate spurious oscillations in the $\tilde{\phi}_{k,i}$ and $\tilde{\psi}_{k,i}$ caused by sampling errors.

Once $\tilde{\phi}_{k,i}$, $\tilde{\psi}_{k,i}$ and $\tilde{\lambda}_k$ have been determined, the minimization of (3.2) subject to (3.4) can be done straightforwardly using a standard quadratic programming package. Here, we simply used the internal quadratic programming routine from Matlab.

To conclude this section, we discuss briefly the possible sources of error in the algorithm we propose. The most important is the estimation of the reference eigenmodes from the timeseries. Clearly, the quality of the estimates depends on the amount of available data. In [13], error estimates are discussed for the eigenmodes of a Markov chain constructed from data. However, other methods to estimate the reference eigenmodes from data are possible as well (see for example [10] for ideas on this). Another source is the discretization of the object function, for which various strategies can be chosen (finite differences on a uniform grid, as used here, is just one of those strategies). We intend to explore some of the alternative strategies to both eigenmode estimation and discretization in future work. A third possible source of error is the minimization algorithm. However, since the minimization problem that results from our approach falls within the category of quadratic programming (with a convex object function, on a convex domain), for which well-established solution methods exist, we consider this to be the least important source of error. Finally, it must be noted that in situations with distinctly non-Markov data, it may be simply impossible to find a diffusion process that does justice to most aspects of the data. This is then not a shortcoming of the reconstruction approach, but an inherent limitation to any attempt to fit a diffusion process to data.

4. Numerical examples

4.1. 1-dimensional

As a 1-dimensional example, we consider a diffusion process on the domain $[-\pi,\pi]$ with periodic boundary conditions with drift and diffusion

$$b(x) = 1+\cos(x) \qquad a(x) = 1+\tfrac{1}{2}\sin(x) \tag{4.1}$$

Using a timeseries generated by numerical integration of this diffusion, we reconstructed the drift and diffusion coefficients, following the algorithm presented in this paper. The integration was carried out with an Euler scheme with timestep 10^{-4}, resulting in a timeseries of 10^6 datapoints with a time-interval $h=0.1$ between consecutive points.

For the reconstruction we used the object function in the form (3.2), with weights $\alpha_k = |\tilde{\lambda}_k \tilde{\psi}_k|^{-2}$, $\beta_k = |\tilde{\lambda}_k \tilde{\phi}_k|^{-2}$ and $\gamma_k = |\tilde{\lambda}_k|^{-2}$ (i.e. all eigenmodes had equal relative

weight). Also, $\alpha_1 = \alpha_2$ and $\beta_1 = \gamma_1 = 0$, since the errors on ϕ_1 and λ_1 are zero by construction. Only the leading three eigenmodes were used in the object function (i.e, $K = 3$). In principle, two eigenmodes ($K = 2$) is enough for the reconstruction of a 1-dimensional drift and diffusion; however, since in this example the $k = 2$ mode is complex, we also included its complex conjugate ($k = 3$).

The reference eigenmodes needed in the object function were obtained from the timeseries according to the simple strategy described in the previous section: we constructed the $M \times M$ stochastic matrix P using (3.8), at timelag $\Delta t = h$, and calculated its spectrum $\{\hat{\psi}_k, \hat{\phi}_k, \Lambda_k\}$. The number of bins M and the number of discretization points N were taken to be the same: $M = N = 60$.

The reference eigenvalues $\tilde{\lambda}_k$ were calculated according to (3.9). For the reference eigenvectors, rather than using the unfiltered $\hat{\psi}_k$ and $\hat{\phi}_k$ (whose small-scale errors, due to finite sample size, unnecessarily distort the outcome), we Fourier-filtered the $\hat{\psi}_k$ and $\hat{\phi}_k$ by discarding the Fourier modes with wavenumbers higher than 6. The thus obtained vectors were used as reference vectors $\tilde{\psi}_k$ and $\tilde{\phi}_k$.

Figure 4.1 shows the resulting, reconstructed drift and diffusion, together with the actual ones in (4.1). Both drift and diffusion are well recovered. Also shown in figure 4.2 are the leading eigenvectors as obtained from the data $(\hat{\psi}_1, \hat{\psi}_2)$, their filtered versions $(\tilde{\psi}_1, \tilde{\psi}_2)$ used as reference modes, and the eigenvectors of the reconstructed diffusion process (ψ_1, ψ_2) (calculated by discretizing the Fokker-Planck operator using the reconstructed b and a). Note that the third eigenvector in this example, ψ_3, is simply the complex conjugate of ψ_2. As can be seen, the eigenvectors of the reconstructed process are indistinguishable by eye from the reference vectors $\tilde{\psi}_k$ (which are again hardly distinguishable from the unfiltered eigenvectors). The match between the other observed, filtered and reproduced eigenvectors $(\hat{\phi}_k, \tilde{\phi}_k, \phi_k)$ is equally good (not shown). Finally, the leading eigenvalues $\lambda_{2,3}$ resemble the reference values $\tilde{\lambda}_{2,3}$ very closely: $\lambda_{2,3} = -0.6508 \pm 0.9086i$ for the reconstructed process, whereas the reference values are $\tilde{\lambda}_{2,3} = -0.6512 \pm 0.9086i$.

It must be stressed that the (small) errors in the reconstructed drift and diffusion are almost exclusively due to errors in the estimates of the eigenmodes. The match between the reference eigenmodes and the eigenmodes of the reconstructed diffusion process is extremely close. Obviously, the technique described in this paper is mostly of interest in situations where the drift and diffusion are not known *a priori*. In that case, the main criterion for the succes of the reconstruction procedure is the similarity between the measured eigenmodes and the eigenmodes of the reconstructed diffusion process.

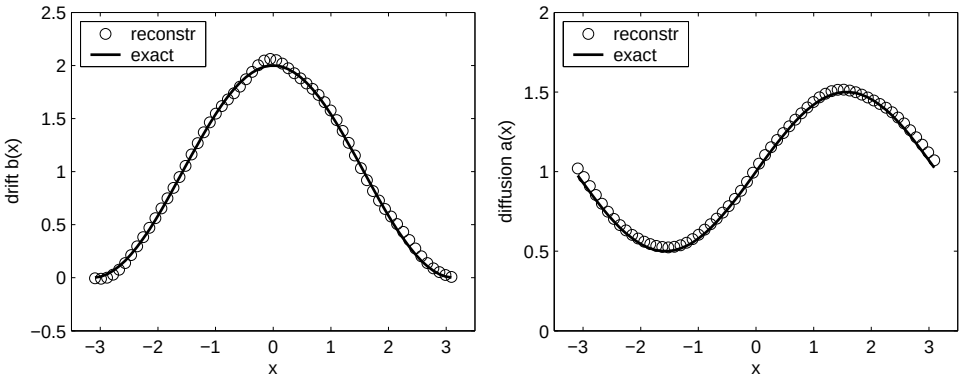

FIG. 4.1. *Reconstructed drift* $b(x) = 1 + \cos(x)$ *and diffusion* $a(x) = 1 + \frac{1}{2}\sin(x)$ *from data.*

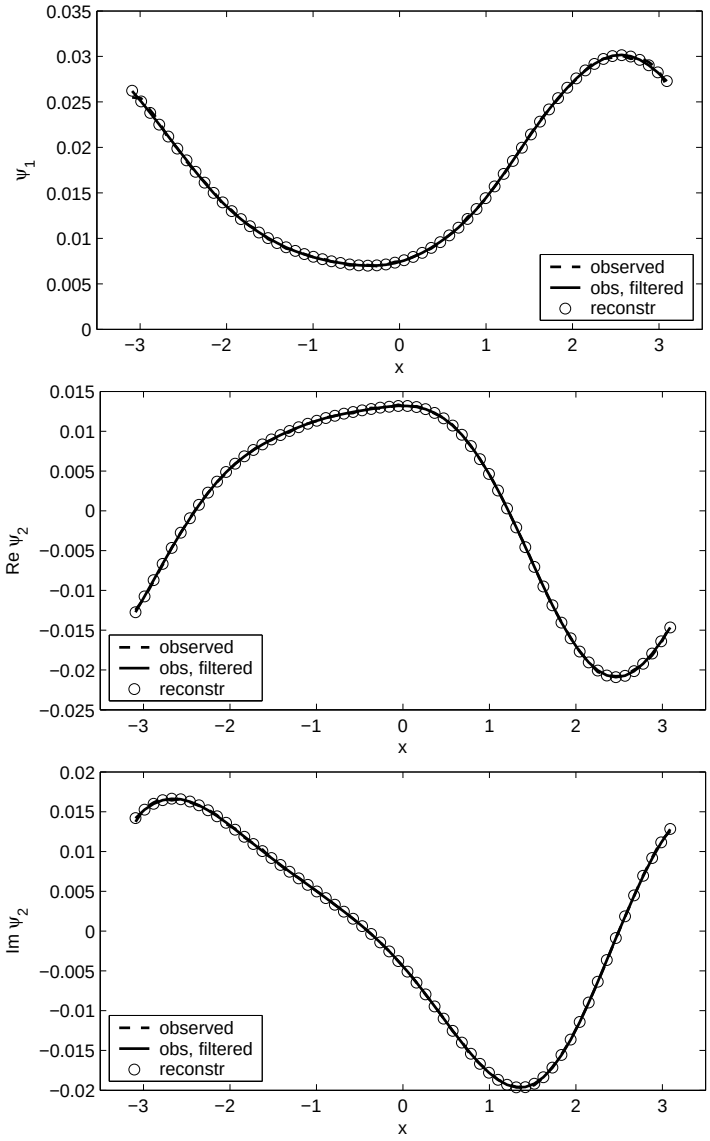

FIG. 4.2. *Invariant distribution (ψ_1) and real and imaginary part of next leading eigenmode (ψ_2) for the diffusion process characterized by the reconstructed drift and diffusion shown in figure 4.1. Also shown are ψ_1 and ψ_2 (both raw and Fourier-filtered) as obtained from the data.*

4.2. 2-dimensional

As a 2-dimensional example we consider the system with drift and diffusion

$$\begin{aligned} b^x(x,y) &= 1 \\ b^y(x,y) &= \frac{3+\cos y}{2} \\ a^{xx}(x,y) &= \left(\frac{3+\sin x}{4}\right)^2 \\ a^{yy}(x,y) &= 1 \end{aligned} \tag{4.2}$$

defined on the domain $(x,y) \in [0,2\pi] \times [0,2\pi]$ with periodic boundary conditions. As in the 1-dimensional example, we generated a timeseries of 10^6 datapoints with time interval $h = 0.1$ between points. The weights in the object function were set the same as in the 1-d example. We used the leading 5 eigenmodes ($K = 5$), obtained by constructing a Markov chain using 35 bins in both x- and y-direction (i.e. 1225 bins in total), and diagonalising the stochastic matrix P. The eigenvectors were Fourier-filtered (only wavenumbers up to 3 retained).

The resulting reconstructed drift and diffusion are shown in figures 4.3–4.6. Figures 4.3 and 4.4 show $b^y(x,y)$ and $a^{xx}(x,y)$, both the reconstructed and the exact fields. The other elements, b^x and a^{yy}, are not shown since they are (nearly) constant. The (absolute) errors in all fields $(b^x, b^y, a^{xx}, a^{yy})$ are shown in figures 4.5 and 4.6, and can be seen to be roughly an order of magnitude smaller than the values of the fields themselves. Finally, in figure 4.7 the invariant distribution (ψ_1) is shown, both the reference distribution $(\tilde{\psi}_1)$ which was obtained from the data and subsequently Fourier-filtered, and the invariant distribution of the diffusion process with the reconstructed drift and diffusion coefficients. The match is very good. The next leading eigenvectors of the reconstructed process match equally well with the reference vectors (not shown); the same holds for the eigenvalues (reference: $0, -0.3074 \pm 1.0224i, -0.5697 \pm 1.4692i$, reconstructed: $0, -0.3065 \pm 1.0219i, -0.5695 \pm 1.4690i$).

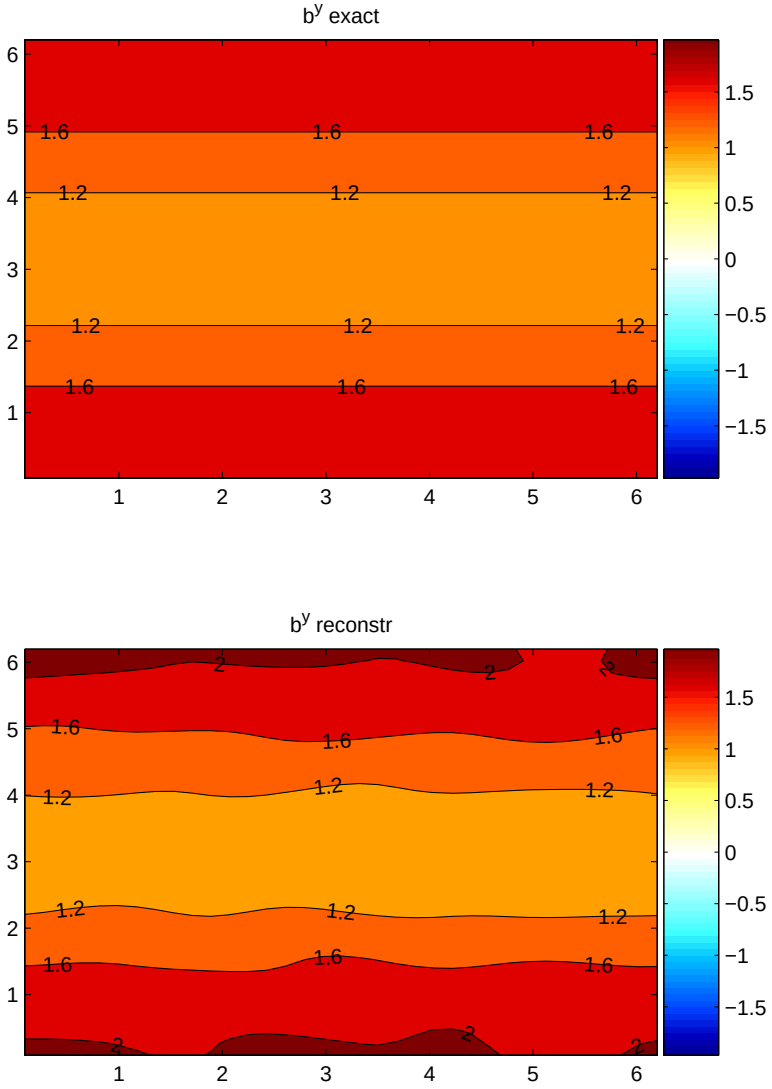

FIG. 4.3. *Reconstructed y-component of the drift,* $b^y(x,y)=\frac{3+\cos y}{2}$*, from data.*

5. Generalization to systems with multiple time-scales

Consider a diffusion with two time-scales such as

$$\begin{cases} \dot{x} = b_x(x,y) + \sigma_x(x,y)\dot{W}_x \\ \dot{y} = \dfrac{1}{\varepsilon} b_y(x,y) + \dfrac{1}{\sqrt{\varepsilon}} \sigma_y(x,y)\dot{W}_y \end{cases} \tag{5.1}$$

where $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$, W_x and W_y are independent Brownian motions and $\varepsilon \ll 1$ is a parameter measuring the time-scale ratio between the fast variables y and the slow variables x. It is well know that if the fast process y is ergodic for each fixed value of x with respect to the equilibrium distribution $\mu_x(y)$, then in the limit as $\varepsilon \to 0$, the

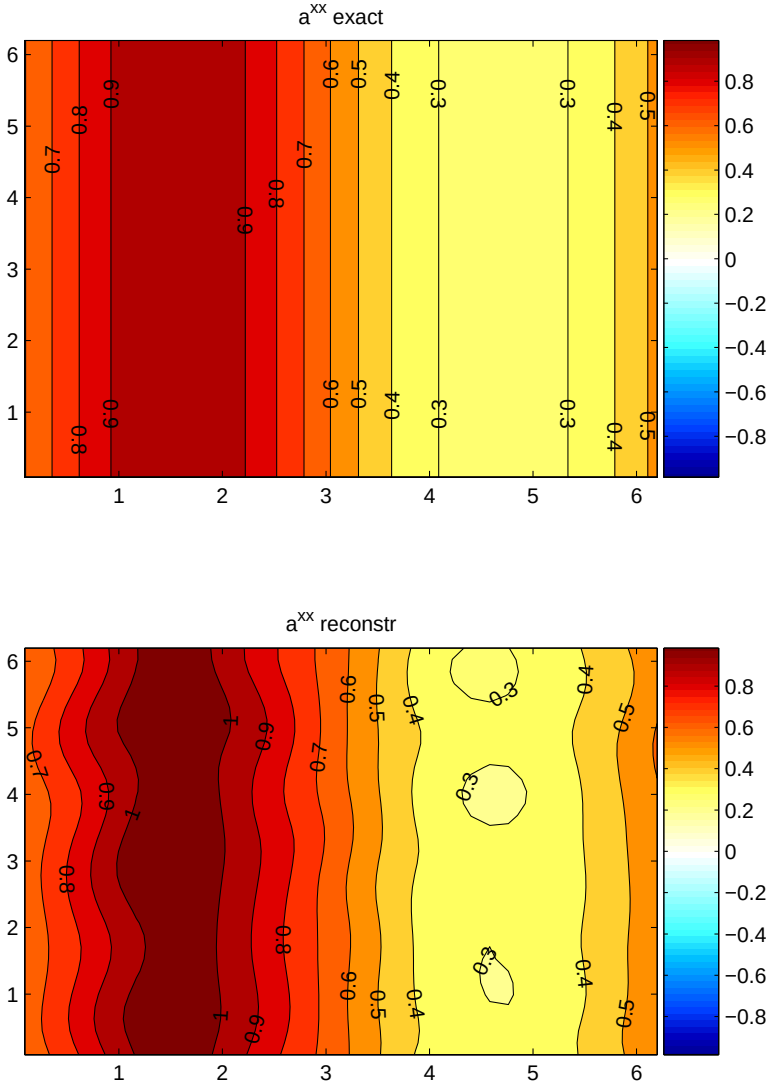

FIG. 4.4. *Reconstructed xx-component of the diffusion, $a^{xx}(x,y)=\left(\frac{3+\sin x}{4}\right)^2$, from data.*

statistics of the slow process $x(t)$ can be approximated by the solution of the limiting diffusion [14, 15, 16, 17, 18] (see also [19] for an overview of related problems)

$$\dot{\bar{x}} = \bar{b}(\bar{x}) + \bar{\sigma}(\bar{x})\dot{W}_x, \tag{5.2}$$

where

$$\bar{b}(x) = \int_{\mathbb{R}^m} b_x(x,y)d\mu_x(y) \tag{5.3}$$

and $\bar{\sigma}(x)$ is the square-root of the diffusion tensor

$$\bar{a}(x) = \int_{\mathbb{R}^m} \sigma_x(x,y)\sigma_x^T(x,y)d\mu_x(y) \tag{5.4}$$

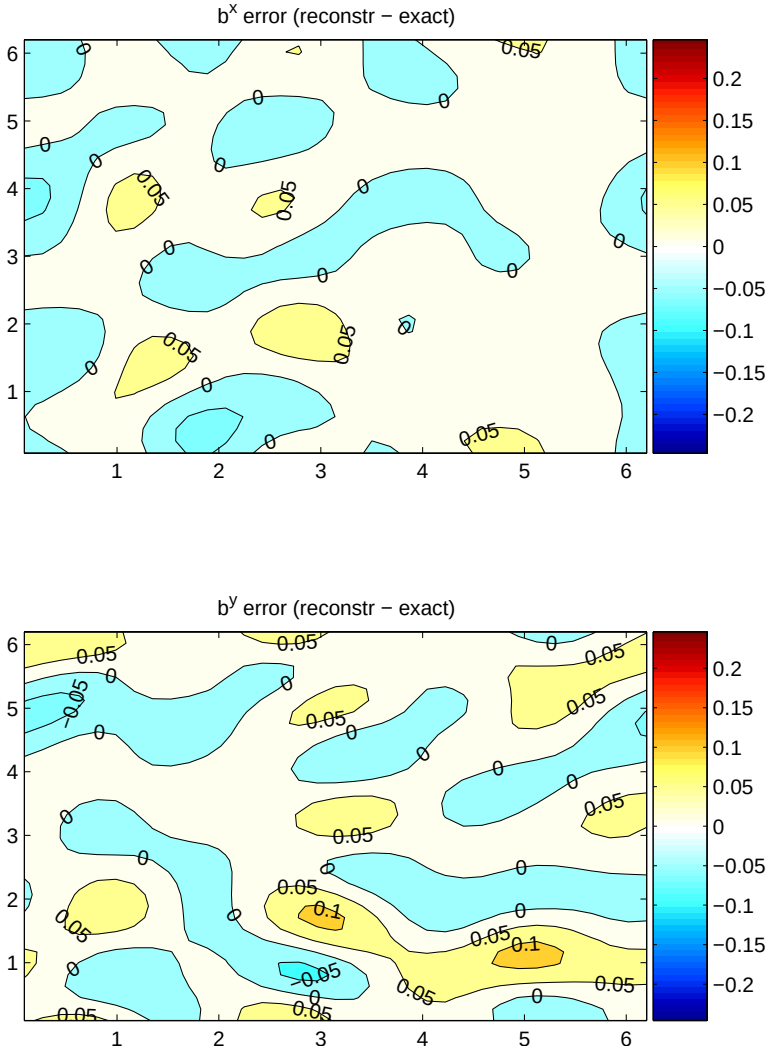

FIG. 4.5. *Errors in the drift vector b reconstructed from data.*

Suppose that we observe the time-series of the original $x(t)$ but what we are really interested in are the effective drift (5.3) and diffusion (5.4) of the homogenized process $\bar{x}(t)$. Our method is very-well suited to handle this case. Indeed, it can be shown that the spectrum of the original process solution of (5.1) can be split into two groups: on group which contains eigenfunctions with eigenvalues that are $O(\varepsilon^{-1})$ and another which contains eigenfunctions with eigenvalues that are $O(1)$. In addition, the eigenfunctions in the second group are approximately independent of y and each of them is close to an eigenfunction of the homogenized process $\bar{x}(t)$, i.e.

$$\phi_k(x,y) = \bar{\phi}_k(x) + O(\varepsilon) \tag{5.5}$$

and a similar result holds for the right eigenfunctions $\psi_k(x,y)$ (for the readers convenience (5.5) is established by formal asymptotics in the Appendix to this paper).

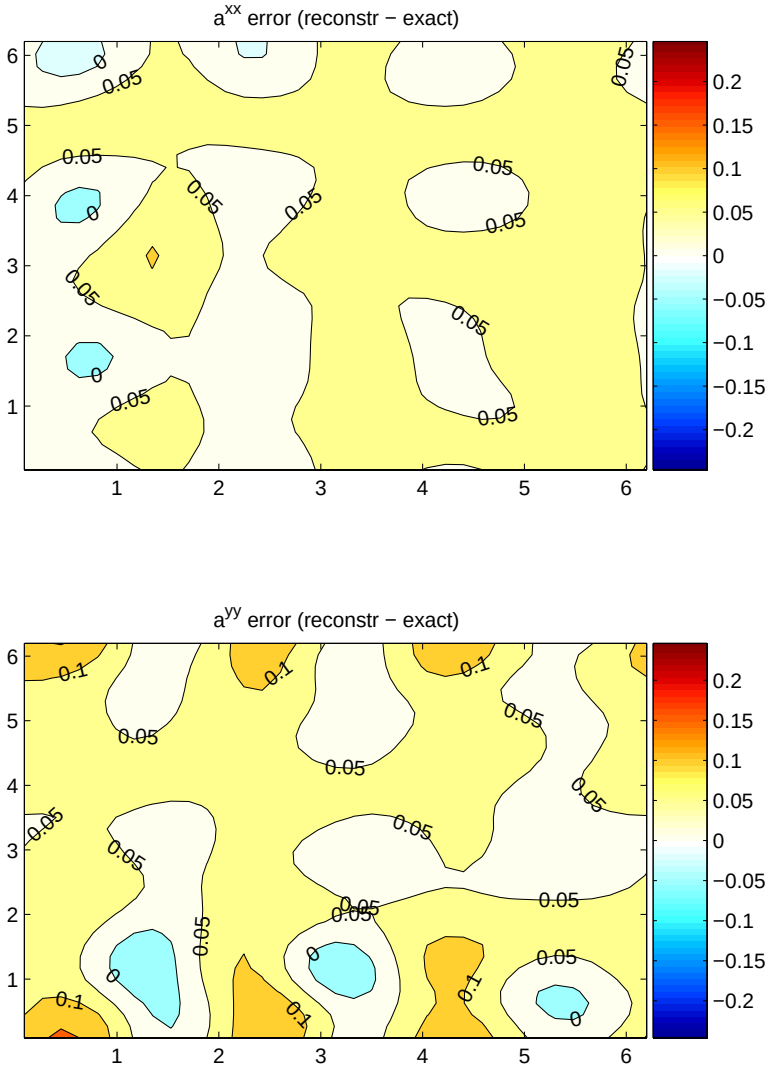

FIG. 4.6. *Errors in the diffusion matrix a reconstructed from data.*

From these results, it follows that if one compute the spectrum of the time-series associated with $x(t)$ by the method explained in section 3, keep only those eigenfunction with $O(1)$ eigenvalues, and use those in the reconstruction procedure, what it will give are an approximation of the effective drift (5.3) and diffusion (5.4) of the homogenized process. Note that this approximation will $O(\varepsilon)$ accurate (since ε is small but finite in the time-series for $x(t)$ generated from (5.1)), but there will be no time-discretization error, i.e. the sampling can be done at an arbitrary lag Δt. This is a big advantage on methods based on the direct reconstruction using the formulas in (1.1), since with those some subsampling must be used in the present situation to capture the homogenized coefficients (i.e. one must have $\Delta t \gg \varepsilon$) but this introduces an additional $O(\Delta t)$ error.

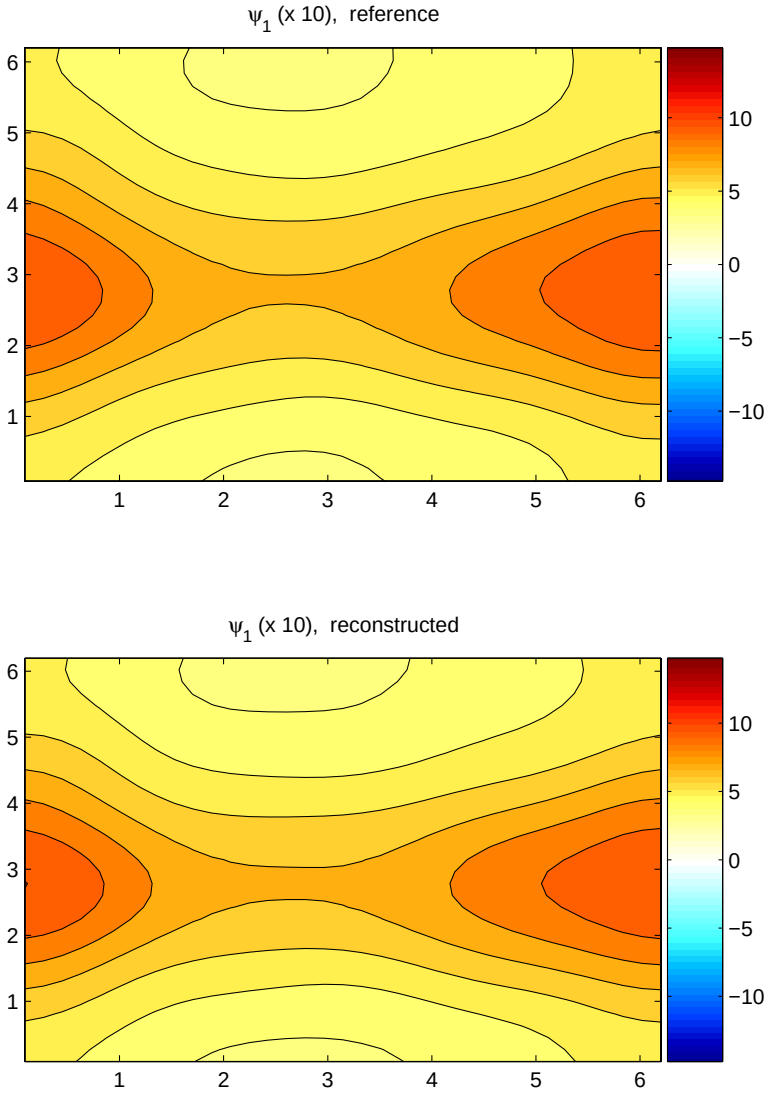

FIG. 4.7. *Invariant distribution ψ_1 of the 2-dimensional diffusion process, both the reference and the reconstructed distribution.*

As an illustration, we consider the two-timescale system

$$\begin{cases} \dot{x} = \sin y + \sqrt{1+\frac{1}{2}\sin y}\,\dot{W}_x \\ \dot{y} = \dfrac{1}{\varepsilon}(y - \sin x) + \dfrac{1}{\sqrt{\varepsilon}}\dot{W}_y \end{cases} \tag{5.6}$$

where the slow variable x takes values in $[0, 2\pi]$ with periodic boundary conditions and the fast variable y in $\mathbb{R}$.

When x is fixed, y is an Ornstein-Uhlenbeck process with mean $\sin x$ and variance $\frac{1}{2}$. Therefore the effective drift and diffusion coefficients in the homogenized equation

for $\bar{x}$ are

$$\bar{b}(x) = \int_{\mathbb{R}} \sin y \frac{e^{-(y-\sin x)^2}}{\sqrt{\pi}} dy = \mathrm{e}^{-1/4} \sin(\sin x) \tag{5.7}$$

and

$$\bar{a}(x) = \int_{\mathbb{R}} \left(1 + \tfrac{1}{2}\sin y\right) \frac{e^{-(y-\sin x)^2}}{\sqrt{\pi}} dy = 1 + \tfrac{1}{2}\mathrm{e}^{-1/4} \sin(\sin x) \tag{5.8}$$

A timeseries of 10^6 points (time interval 0.1) was generated by integrating the system (5.6) with $\varepsilon = 10^{-3}$. Using only the data for x, we reconstructed the drift and diffusion in the exact same way as in the 1-dimensional example in section 4 (including equal settings for the number of bins, Fourier filtering and object function weights). The resulting b and a, together with the exact homogenized coefficients $\bar{b}$ and $\bar{a}$ from (5.7) and (5.8), are shown in figure 5.1. The reconstructed and the exact coefficients match very well. The invariant distributions for x, both for the reconstructed process and as observed from the timeseries (fourier-filtered and unfiltered), are shown in figure 5.2, and are indistinguishable by eye.

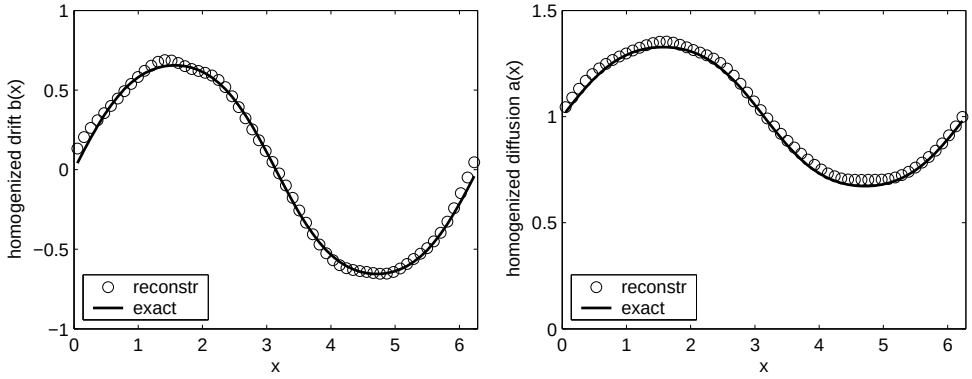

FIG. 5.1. *Homogenized drift and diffusion for the slow variable x of the two timescale system (5.6), obtained using the reconstruction procedure. Also shown are the exact homogenized coefficients given in (5.7) and (5.8).*

6. Conclusion

We have proposed a new algorithm for the reconstruction of drifts and diffusions from timeseries. Our approach is based on the use of (leading) eigenmodes of the diffusion process that is to be reconstructed. Estimates of these eigenmodes can be obtained easily by constructing a Markov chain from the timeseries. The algorithm centers on the minimization of the convex, quadratic, positive semi-definite object function (2.3). The minimization can be carried out numerically using well-established techniques for quadratic programming problems. The result of the minimization are drift and diffusion coefficients which are such that the corresponding diffusion generator L has an eigenspectrum that resembles the reference eigenspectrum (estimated from the available data) as closely as possible.

The validity of our method has been illustrated with concrete examples of the reconstruction of 1-dimensional and 2-dimensional diffusion processes, which were

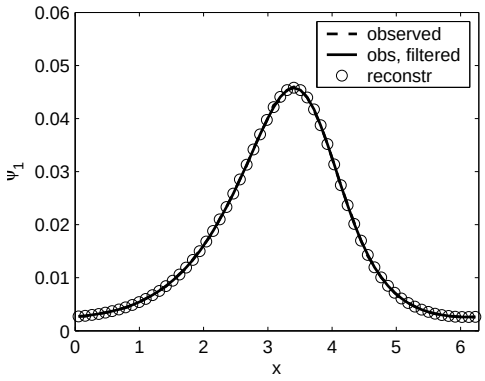

FIG. 5.2. *Invariant distribution ψ_1 of the slow variable x. The observed distribution was obtained from a timeseries generated using (5.6); the reconstructed distribution is the ψ_1 associated with the reconstructed drift and diffusion shown in figure 5.1.*

presented in section 4, and show that the algorithm is numerically feasible and gives good results. As shown in section 5 the algorithm can also be straightforwardly generalized to situations with multiple time-scales in which one is interested in constructing the effective drift and diffusion coefficients of the homogenized equation for the slow component of the process.

Our approach is very flexible and many generalizations/modifications are possible: for instance we have focussed in this study on non-parametric estimation of drift and diffusion, but the algorithm can be easily adapted to allow for parametric estimation (as briefly explained in sections 2 and 3). We have shown numerical examples in 1 and 2 dimensions; for reconstruction of processes in dim>2, the representation of eigenfunctions, drift and diffusion on a uniform grid will become quite costly. Parametric estimation will then be a useful alternative, which can reduce the size of the minimization problem significantly. In general, for reconstruction of a d-dimensional diffusion process, one needs to reconstruct the d-dim. vector $b(x)$ and the $d\times d$ matrix $a(x)$, giving $(d^2+3d)/2$ functions to reconstruct (if $a(x)$ is assumed to be symmetric). If each of them is spatially represented on n gridpoints or with n basis functions (e.g. polynomials, fourier modes, etc.), one gets a minimization problem of dimension $n(d^2+3d)/2$. Since the number of points on a uniform grid scales as $n=m^d$ with m points along each coordinate axis, representation using a low number of basis functions will be particularly attractive in higher dimensions. We shall focus on these issues in future work.

Acknowledgments. We thank Andy Majda for many useful discussions and comments. This work was sponsored in part by NSF through Grants DMS01-01439, DMS02-09959, DMS02-39625 and DMS-0222133.

Appendix A. Spectral analysis of (5.1).

Let L be the generator associated with the diffusion process in (5.1). We can decompose L as follows:

$$L = \frac{1}{\varepsilon} L_1 + L_0 \tag{A1}$$

with

$$\begin{cases} L_1 = b_y \dfrac{\partial}{\partial y} + \dfrac{1}{2}\sigma_y \sigma_y^T \dfrac{\partial^2}{\partial y^2}, \\ L_0 = b_x \dfrac{\partial}{\partial x} + \dfrac{1}{2}\sigma_x \sigma_x^T \dfrac{\partial^2}{\partial x^2}. \end{cases} \tag{A2}$$

In principle, the coefficient functions b_x, b_y, σ_x, σ_y all depend on both x and y. For the asymptotic analysis of the eigenvalue problem

$$\left(\frac{1}{\varepsilon} L_1 + L_0\right) \phi_k(x,y) = \lambda_k \phi_k(x,y) \tag{A3}$$

we expand ϕ_k as $\phi_k = \phi_k^{(0)} + \varepsilon \phi_k^{(1)} + \ldots$, and consider a Laurent expansion for λ_k:

$$\lambda_k = \frac{1}{\varepsilon} \lambda_k^{(-1)} + \lambda_k^{(0)} + \varepsilon \lambda_k^{(1)} + \ldots. \tag{A4}$$

At leading order, $\mathcal{O}(\varepsilon^{-1})$, we find the equation

$$L_1 \phi_k^{(0)} = \lambda_k^{(-1)} \phi_k^{(0)}, \tag{A5}$$

i.e. another eigenvalue problem. The solutions to this problem with $\lambda_k^{(-1)} \neq 0$ form the first group of solutions to (A3): eigenfunctions with eigenvalues that are $O(\varepsilon^{-1})$. The second group of solutions to (A5) has $\lambda_k^{(-1)} = 0$ (i.e. they are solutions to (A3) with $O(1)$ eigenvalues). Considering the equations at $O(\varepsilon^{-1})$ as well as $O(1)$ for this second group, we find

$$\begin{cases} L_1 \phi_k^{(0)} = 0, \\ L_1 \phi_k^{(1)} = \lambda_k^{(0)} \phi_k^{(0)} - L_0 \phi_k^{(0)}. \end{cases} \tag{A6}$$

The first equation implies that $\phi_k^{(0)}$ lies in the null-space of L_1, i.e. $P\phi_k^{(0)} = \phi_k^{(0)}$ where P denotes the expectation with respect to $\rho_x(y)$, the equilibrium distribution for y with x fixed (for the process (5.1)): given any suitable function $f(x,y)$,

$$Pf(x,y) = \int_{\mathbb{R}^m} \rho_x(y) f(x,y) dy. \tag{A7}$$

Equivalently, $\rho_x(y)$ is the solution to $L_1^* \rho_x(y) = 0$ with L_1^* the adjoint of L_1 in $L_0(\mathbb{R}^m, dy)$ and $P\phi_k^{(0)} = \phi_k^{(0)}$ implies that $\phi_k^{(0)}$ may depend of x but is independent of y. Solvability of the second equation in (A6) requires

$$PL_0 P \phi_k^{(0)} = \lambda_k^{(0)} \phi_k^{(0)}. \tag{A8}$$

The operator PL_0P is the generator of the homogenized process (5.2). Moreover, from (A8) follows (5.5): each eigenmode of the homogenized process approximates to $O(1)$ in ε an eigenmode of the original process with an $O(1)$ eigenvalue.

REFERENCES

[1] S. Siegert, R. Friedrich, J. Peinke, Analysis of data sets of stochastic systems, Phys. Lett. A 243 (1998) 275–280.
[2] R. Friedrich, S. Siegert, J. Peinke, S. Lück, M. Siefert, M. Lindemann, J. Raethjen, G. Deuschl, G. Pfister, Extracting model equations from experimental data, Phys. Lett. A 271 (2000) 217–222.
[3] J. Gradišek, S. Siegert, R. Friedrich, I. Grabec, Analysis of time series from stochastic processes, Phys. Rev. E 62 (2000) 3146–3155.
[4] P. Sura, Stochastic analysis of Southern and Pacific sea surface winds, J. Atmos. Sci. 60 (2003) 654–666.
[5] J. Berner, Detection and stochastic modeling of nonlinear signatures in the geopotential height field of an atmospheric general circulation model, Ph.D. thesis, Universität Bonn (2003).
[6] J. Berner, Linking nonlinearity and non-Gaussianity of planetary wave behavior by the Fokker-Planck equation, J. Atmos. Sci. 62 (2005) 2098–2117.
[7] M. Kessler, M. Sørensen, Estimating equations based on eigenfunctions for a discretely observed diffusion process, Bernoulli 5 (1999) 299–314.
[8] Y. Aït-Sahalia, Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach, Econometrica 70 (2002) 223–262.
[9] L. Kelly, E. Platen, M. Sørensen, Estimation for discretely observed diffusions using transform functions, J. Appl. Prob. 41A (2004) 99–118.
[10] L. Hansen, J. Scheinkman, N. Touzi, Spectral methods for identifying scalar diffusions, J. Econometrics 86 (1998) 1–32.
[11] I. Gelfand, B. Levitan, On the determination of a differential equation from its spectral function, Am. Math. Soc. Transl. 1 (1955) 253–304.
[12] J. McLaughlin, Analytical methods for recovering coefficients in differential equations from spectral data, SIAM Rev. 28 (1986) 53–72.
[13] D. Crommelin, E. Vanden-Eijnden, Fitting timeseries by continuous-time Markov chains: A quadratic programming approach, J. Comp. Phys. , submitted.
[14] R. Z. Khasminskii, Principle of averaging for parabolic and elliptic differential equations and for Markov processes with small diffusion, Theory Probab. Appl. 8 (1963) 1–21.
[15] R. Z. Khasminskii, A limit theorem for the solutions of differential equations with random right-hand sides, Theory Probab. Appl. 11 (1966) 390–406.
[16] T. Kurtz, A limit theorem for perturbed operator semigroups with applications to random evolutions, J. Functional Anal. 12 (1973) 55–67.
[17] G. Papanicolaou, Some probabilistic problems and methods in singular perturbations, Rocky Mountain J. Math. 6 (1976) 653–674.
[18] G. C. Papanicolaou, Introduction to the asymptotic analysis of stochastic equations, in: R. DiPrima (Ed.), Modern modeling of continuum phenomena, Vol. 16 of Lectures in Applied Mathematics, American Mathematical Society, 1977, pp. 109–147.
[19] D. Givon, R. Kupferman, A. Stuart, Extracting macroscopic dynamics: model problems and algorithms, Nonlinearity 17 (2004) R55–127.