DIFFUSION ESTIMATION FROM MULTISCALE DATA BY OPERATOR EIGENPAIRS[*]

DAAN CROMMELIN[†] AND ERIC VANDEN-EIJNDEN[‡]

Abstract. In this paper we present a new procedure for the estimation of diffusion processes from discretely sampled data. It is based on the close relation between eigenpairs of the diffusion operator $\mathscr{L}$ and those of the conditional expectation operator P_t, a relation stemming from the semigroup structure $P_t = \exp(t\mathscr{L})$ for $t \geq 0$. It allows for estimation without making time discretization errors, an aspect that is particularly advantageous in case of data with low sampling frequency. After estimating eigenpairs of $\mathscr{L}$ via eigenpairs of P_t, we infer the drift and diffusion functions that determine $\mathscr{L}$ by fitting $\mathscr{L}$ to the estimated eigenpairs using a convex optimization procedure. We present numerical examples where we apply the procedure to one- and two-dimensional diffusions, reversible as well as nonreversible.

In the second part of the paper we consider estimation of coarse-grained (homogenized) diffusion processes from multiscale data. We show that eigenpairs of the homogenized diffusion operator are asymptotically close to eigenpairs of the underlying multiscale diffusion operator. This implies that we can infer the correct homogenized process from data of the multiscale process, using the estimation procedure discussed in the first part of the paper. This is illustrated with numerical examples.

Key words. parameter estimation, diffusion process, stochastic differential equation, generator, discrete sampling, multiscale analysis, homogenization, subsampling

AMS subject classifications. 62M05, 60J60, 60J35, 60H10, 47A75, 62F12, 60H30, 35B27, 34E13, 47D07

1. Introduction.

Estimation of stochastic models from timeseries is an important tool in scientific disciplines ranging from econometrics [23, 2, 4] to chemistry [17, 25, 41, 9] and atmosphere-ocean science [35, 7, 40]. A widely used class of such models are diffusion processes, described by stochastic differential equations (SDEs):

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t \tag{1.1}$$

where $X_t \in \Omega \subseteq \mathbb{R}^d$ and W_t is a d-dimensional Wiener process.

Inferring the drift $b(x)$ and the diffusion $a(x) = \sigma(x)\sigma(x)^T$ from timeseries data is a challenging task, facing two major practical issues. The first is that of *discrete-time data*. In applications, the available timeseries data is nearly always discrete in time, whereas a diffusion is a continuous-time process. With only few exceptions, the finite-time transition densities of a diffusion process are unknown functions of $b(x)$ and $a(x)$. This causes great difficulties for estimation, in particular in case of low-frequency data (i.e., data with long sampling intervals).

Reflecting this difficulty, and the variety of approaches proposed to overcome it, the literature on diffusion estimation from discrete-time data is extensive. It includes likelihood-based estimation as well as Bayesian methods, in which transition densities are approximated with simulations [34, 19, 20, 36, 13, 8] or with closed-form expansions [1, 3]. Alternative approaches include the use of estimating functions [10, 27, 11] and spectral methods [23, 22, 14]. An overview of different approaches can be found in [38]; the difficulties of estimation from low-frequency data are highlighted in [22].

The second major difficulty is that of *model misspecification*, occurring when the data is not consistent with the chosen model class (in this case, the class of diffusion

[*] Revised version, 31 August 2011. To appear in SIAM Multiscale Model. Simul.

[†] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands (`Daan.Crommelin@cwi.nl`).

[‡] Courant Institute of Mathematical Sciences, New York, USA (`eve2@cims.nyu.edu`).

processes). If the data differs significantly from a diffusion process, estimation of a "best-fit" diffusion process can be a delicate task. A notable example arises if one observes a process with multiple scales in space and/or time and one wishes to model the coarse-grained dynamics of this process with a diffusion process. In this case the chosen model should be consistent with the coarse-grained features of the data, but not necessarily with its "fine-grained" (small-scale) features. Because of the inconsistency of model and data at small scales, care has to be taken when inferring a coarse-grained model from multiscale data, as was shown for example in [32].

In this paper we present a methodology for estimation that allows to tackle both issues. In summary, the methodology consists of two steps. First we estimate eigenfunctions and eigenvalues of the operator $P_t = \exp(t\mathscr{L})$, where the generator $\mathscr{L}$ is the diffusion operator associated with (1.1),

$$\mathscr{L} = \sum_{i=1}^{d} b_i(x)\,\frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2}{\partial x_i\,\partial x_j}\,. \tag{1.2}$$

Eigenpairs of $\mathscr{L}$ follow directly from estimated eigenpairs of P_t. In the next step, we solve the inverse problem of inferring the coefficients (b and a) of $\mathscr{L}$ from its eigenpairs. This is done by casting the inverse problem as a convex minimization problem.

The use of eigenpairs solves the difficulty of discrete-time data, because the relation between P_t and $\mathscr{L}$ is exact for any $t \geq 0$. Furthermore, the proposed methodology gives a handle on model misspecification, because of the formulation as a minimization problem and the possibility to infer b and a from a small number of eigenpairs. This allows to use the eigenpairs that best represent the coarse-grained features of the observed process. In section 5 we analyze, for a broad class of multiscale diffusions, how the correct coarse-grained process can be inferred from data of a multiscale process with the methodology proposed here.

Estimation procedures that use estimates of eigenpairs to infer $\mathscr{L}$ were proposed in [23, 27, 22, 14]. They all exploit the close relation between the spectrum of $\mathscr{L}$ and that of the conditional expectation operator P_t of the process X_t. This operator is defined by

$$\big(P_t\, f\big)(x) = \mathbb{E}(f(X_t)\,|\,X_0 = x) \tag{1.3}$$

for suitable functions $f(x)$ and $t \geq 0$. $\mathscr{L}$ is the generator associated with P_t:

$$\mathscr{L}\, f(x) = \lim_{t\downarrow 0}\frac{(P_t f)(x) - f(x)}{t} \tag{1.4}$$

For a diffusion process, $\mathscr{L}$ is the diffusion operator (1.2). As mentioned before, P_t and $\mathscr{L}$ are related via

$$P_t = \exp(t\mathscr{L})\,, \tag{1.5}$$

and similarly for the adjoints in $L^2(\Omega, dx)$ of $\mathscr{L}$ and P_t, denoted $\mathscr{L}^*$ and P_t^*. As a consequence,

$$P_t\phi = \Lambda\phi \quad \text{implies} \quad \mathscr{L}\phi = \lambda\phi \tag{1.6a}$$

$$P_t^*\psi = \bar{\Lambda}\psi \quad \text{implies} \quad \mathscr{L}^*\psi = \bar{\lambda}\psi \tag{1.6b}$$

with

$$\lambda = \frac{1}{t} \log \Lambda \tag{1.7}$$

The procedures in [23, 27, 22] require either explicit expressions of $b(x)$ and $a(x)$ in terms of the eigenfunctions and eigenvalues, or a priori knowledge of the eigenfunctions. In [14] it was proposed to estimate eigentriplets (ϕ, ψ, λ) and minimize the residuals $\mathscr{L}\phi - \lambda\phi$ and $\mathscr{L}^*\psi - \bar{\lambda}\psi$ under variation of $b(x)$ and $a(x)$. For this procedure it is not necessary to know the eigenfunctions a priori, nor to have explicit expressions of $b(x)$ and $a(x)$ in terms of (ϕ, ψ, λ) available. Also, sampling error or model misspecification can cause problems (e.g., $a(x)$ may become negative for some x) if explicit expressions are used. Such problems can be avoided by minimizing residuals under appropriate constraints (such as $a(x) \geq 0$).

In this paper we expand and modify the approach from [14] in several ways. In section 3 we put the estimation of eigentriplets (ϕ, ψ, λ) in the framework of Galerkin methods. This leads to two alternative ways to estimate eigentriplets, depending on whether the Galerkin basis functions are smooth or discontinuous (piecewise constant). In the latter case, P_t^* is effectively approximated by the transition probability matrix of a finite-state Markov chain, the method also used in [14].

Next, in section 4 we present a modification of the minimization procedure proposed in [14]. Rather than minimizing the residuals $\mathscr{L}\phi - \lambda\phi$ or $\mathscr{L}^*\psi - \bar{\lambda}\psi$ themselves, we integrate them against suitable test functions and minimize the integrals. This modified procedure has a natural connection with the Galerkin method for estimating eigentriplets. It also allows for estimation of $\mathscr{L}$ without requiring estimates of the derivatives of the eigenfunctions, thereby circumventing a major source of error. The proposed procedure is suitable for estimation of reversible as well as nonreversible diffusion processes.

In section 5 we investigate the eigenspectrum of diffusion operators with a multiscale character. We consider multiscale diffusions whose slow dynamics can effectively be described by an homogenized diffusion process. We show that the leading eigentriplets of the multiscale diffusion operator and those of the homogenized operator are, in essence, the same at leading order in ϵ, where $\epsilon \ll 1$ is a measure for the scale separation in the multiscale process. This makes inference procedures that use the eigenspectrum attractive for estimation of a coarse-grained process from multiscale data. Included in section 5 is a discussion of partially observed diffusions and subsampling.

The paper finishes with a conclusion and discussion in section 6. Numerical examples will be presented throughout the paper.

2. Mathematical preliminaries.

In this section, we summarize some properties of the diffusion operator and its eigenvalues and eigenfunctions. We also fix some conventions, definitions and notations that will be used in the paper.

We define $\Omega \subseteq \mathbb{R}^d$ to be the domain of the process X_t. Throughout the paper, we assume that the process has an invariant measure (denoted μ) that admits a density ρ, i.e. $\mu(dx) = \rho(x)dx$, $x \in \Omega$. We also assume that the process is ergodic and that ρ is unique. Furthermore, b and a do not depend explicitly on time, so the process is time-homogeneous.

We use the notation $\langle ., . \rangle_\omega$ for the $L^2(\Omega, \omega dx)$ inner product with some weight function $\omega(x)$. A process X_t is said to be *reversible* if its associated $\mathscr{L}$ is selfadjoint with respect to the $L^2(\Omega, \mu)$ inner product (note that $\mathscr{L}^*$ is defined as the adjoint in $L^2(\Omega, dx)$ rather than in $L^2(\Omega, \mu)$, therefore reversibility does not imply $\mathscr{L} = \mathscr{L}^*$).

We consider the Sobolev space $H^2(\Omega, \mu)$ as the domain of $\mathscr{L}$. For the domain of P_t, denoted $\mathscr{F}$, taking $\mathscr{F} = \mathrm{dom}(\mathscr{L})$ seems the most natural choice. However, it will be convenient to consider a larger space, $\mathscr{F} = L^2(\Omega, \mu)$. This allows us to use functions that approximate the eigenfunctions of $\mathscr{L}$ but do not approximate their derivatives.

The eigentriplets of P_t and $\mathscr{L}$ are ordered by decreasing $|\Lambda|$. Thus, $1 = \Lambda_1 > |\Lambda_2| \geq |\Lambda_3| \geq |\Lambda_4| \geq ...$ (where the strict inequality $|\Lambda_2| < 1$ follows from the assumption of ergodicity). We assume that the discrete spectrum of P_t is non-empty and that its essential spectrum is bounded by a radius smaller than some appropriate $|\Lambda_k|$. The eigenfunctions are normalized so that they form a bi-orthonormal set: $\langle \psi_k, \phi_l \rangle_1 = \delta_{kl}$. The ordering by decreasing $|\Lambda_k|$ implies that $\psi_1 = \rho$, $\phi_1 = 1$ and $0 = \lambda_1 > \mathrm{Re}\,\lambda_2 \geq \mathrm{Re}\,\lambda_3 \geq \mathrm{Re}\,\lambda_4 \geq ....$ Finally, we will make use of the functions ξ_k, defined such that

$$\psi_k = \rho\, \xi_k\,. \tag{2.1}$$

If X_t is a reversible process, $\xi_k = \phi_k$.

Finally, an overbar denotes complex conjugation, and the Hermitian transpose of a matrix A is denoted A^*. Also, we will occasionally use the abbreviated notation $\mathscr{L} = b \cdot \nabla + \frac{1}{2} a : \nabla\nabla$ for the diffusion operator (1.2).

3. Statistical inference of operator eigenpairs. By the relations (1.6a) and (1.6b), estimates of eigenpairs of $\mathscr{L}$ and $\mathscr{L}^*$ can be obtained by estimating eigenpairs of P_t and P_t^*. The relation (1.7) is nontrivial in case of complex eigenvalues, because of the non-uniqueness of the logarithm. This subtlety is discussed in detail in [15]; here, we use the principal branch of the logarithm in case of complex eigenvalues. For reversible diffusion processes, all eigenvalues are real and the relation (1.7) is unambiguous.

In this section we discuss estimation of eigenpairs of P_t and P_t^* using Galerkin methods to discretize $\mathscr{F}$. For simplicity we assume that the data has a constant sampling interval $t = \tau$, i.e. we have data $X_0, X_\tau, X_{2\tau}, ..., X_{N\tau}$ from which we want to infer eigenpairs of P_τ and P_τ^*. In [15], estimation from data with nonconstant (e.g., random) sampling intervals is discussed. Although the context there was generator estimation for Markov jump processes, the spectral estimation procedure in [15] is similar to what is proposed here, and many of the ideas carry over to diffusion estimation.

3.1. Galerkin method. In the Galerkin method for estimating eigenpairs, the domain $\mathscr{F}$ of P_τ is approximated by its projection into a finite-dimensional subspace $\mathscr{F}_M$. Correspondingly, P_τ is approximated by a matrix-valued operator mapping this subspace to itself. We refer to [6] (and refences therein) for a discussion of Galerkin approximations for eigenvalue problems involving linear operators. If the operator is self-adjoint (as in the case of a reversible diffusion process), the Galerkin method is also known as the Rayleigh-Ritz method. In [16, 22], the Galerkin method to estimate eigenpairs is referred to as the sieve method.

3.1.1. Galerkin approximation for eigenpairs of P_τ. The Galerkin method starts from a weak formulation of the eigenvalue problem for P_τ. Let the set of independent functions $f_i : \Omega \to \mathbb{R}$, $i = 1, ..., M$, be a basis for $\mathscr{F}_M \subset \mathscr{F}$. We want to find pairs (Λ_k^g, ϕ_k^g) with $\Lambda_k^g \in \mathbb{C}$, $|\Lambda_k^g| \leq 1$ and $\phi_k^g \in \mathscr{F}_M \setminus \{0\}$ such that

$$\langle P_\tau \phi_k^g, f_i \rangle_\rho = \langle \Lambda_k^g \phi_k^g, f_i \rangle_\rho \qquad \text{for all } \; i = 1, ..., M \tag{3.1}$$

We expand ϕ_k^g, the Galerkin approximation of ϕ_k, on the basis $f_1, ..., f_M$,

$$\phi_k^g(x) = \sum_{i=1}^{M} v_{ki} f_i(x)\,, \tag{3.2}$$

and define V as the matrix of expansion coefficients v_{ki} ($\in \mathbb{C}$). Furthermore, we define the matrices R and T with elements

$$R_{ij} = \langle f_i, f_j \rangle_\rho\,, \tag{3.3a}$$
$$T_{ij} = \langle P_\tau f_i, f_j \rangle_\rho\,. \tag{3.3b}$$

Because all f_i are real functions, R and T are real matrices. Also, R is symmetric. In matrix notation, it can be seen that the weak formulation (3.1) of the eigenvalue problem for P_τ is the generalized eigenvalue problem

$$V\,T = D_\Lambda\,V\,R \tag{3.4}$$

where D_Λ is the diagonal matrix

$$D_\Lambda = \mathrm{diag}(\Lambda_1^g, ..., \Lambda_M^g)\,. \tag{3.5}$$

The adjoint problem can be treated similarly, resulting in

$$T\,W^* = R\,W^*\,D_\Lambda\,, \tag{3.6}$$

where W is the matrix of expansion coefficients for the ξ_k^g, cf. (2.1),

$$\xi_k^g(x) = \sum_{i=1}^{M} w_{ki} f_i(x)\,, \tag{3.7}$$

and we have used the identity

$$\langle P_\tau^* \psi_k^g, f_i \rangle_1 = \langle \xi_k^g, P_\tau f_i \rangle_\rho \tag{3.8}$$

Thus, the operator eigenvalue problem $P_\tau \phi_k = \Lambda_k \phi_k$ and the adjoint problem $P_\tau^* \psi_k = \bar{\Lambda}_k \psi_k$ are converted into the generalized matrix eigenvalue problems (3.4) and (3.6).

We will assume that (T, R) form a regular matrix pair, implying that they can both be diagonalized with the same pair of matrices. Bi-orthonormality of the eigenfunctions translates into

$$V\,R\,W^* = \mathbf{I}\,, \tag{3.9}$$

where $\mathbf{I}$ is the unit matrix. Combining (3.9) with either (3.4) or (3.6) gives

$$V\,T\,W^* = D_\Lambda\,. \tag{3.10}$$

3.1.2. Estimators for the Galerkin method. The inner products that define R and T in (3.3) can be written as expectations with respect to the law of X_t:

$$\langle f_i, f_j \rangle_\rho = \mathbb{E} f_i(X_t) f_j(X_t)\,, \tag{3.11a}$$
$$\langle P_\tau f_i, f_j \rangle_\rho = \mathbb{E} f_i(X_{t+\tau}) f_j(X_t)\,. \tag{3.11b}$$

Because we have assumed ergodicity of the process X_t, we can estimate the matrix elements of R and T from the timeseries, using for example the estimators

$$\hat{R}_{ij} = \frac{1}{N} \sum_{n=0}^{N-1} f_i(X_{n\tau}) f_j(X_{n\tau}) \tag{3.12a}$$

$$\hat{T}_{ij} = \frac{1}{N} \sum_{n=0}^{N-1} f_i(X_{(n+1)\tau}) f_j(X_{n\tau}) \tag{3.12b}$$

In [22], the estimators

$$\hat{R}'_{ij} = \frac{1}{N} \Big(\frac{1}{2} f_i(X_0) f_j(X_0) + \frac{1}{2} f_i(X_{N\tau}) f_j(X_{N\tau}) + \sum_{n=1}^{N-1} f_i(X_{n\tau}) f_j(X_{n\tau}) \Big) \tag{3.13a}$$

$$\hat{T}'_{ij} = \frac{1}{2N} \sum_{n=0}^{N-1} \Big(f_i(X_{(n+1)\tau}) f_j(X_{n\tau}) + f_i(X_{n\tau}) f_j(X_{(n+1)\tau}) \Big) \tag{3.13b}$$

are proposed. The validity of $\hat{T}'$ as an estimator of T is limited to reversible processes, where $\langle P_\tau f_i, f_j \rangle_\rho = \langle f_i, P_\tau f_j \rangle_\rho$ and thus $T^T = T$.

We solve the eigenproblems (3.4) and (3.6) by substituting $\hat{T}, \hat{R}$ for T, R, resulting in the estimates $\hat{V}$, $\hat{W}$ and $\hat{D}_\Lambda$:

$$\hat{V} \hat{T} = \hat{D}_\Lambda \hat{V} \hat{R} , \qquad \hat{T} \hat{W}^* = \hat{R} \hat{W}^* \hat{D}_\Lambda . \tag{3.14}$$

The estimated (eigen)functions $\hat{\phi}^g_k$ and $\hat{\xi}^g_k$ are obtained by using the elements of $\hat{V}$ and $\hat{W}$ in the expansions (3.2) and (3.7). Note that this procedure does not give estimates of the ψ^g_k. To obtain those, one first has to estimate the invariant density ρ. However, estimates of ψ^g_k are not needed in the inference procedure discussed in section 4. The estimates $\hat{\xi}^g_k$ and $\hat{\phi}^g_k$ suffice.

3.1.3. Discontinuous Galerkin method: binning. A particular version of the Galerkin method occurs if the basis functions are chosen to be indicator functions on subdomains (”bins”) Ω_i of Ω. It is also used in Ulam's method for approximating invariant measures of mappings, see e.g. [26, 21, 18]. One discretizes Ω by covering it with a non-overlapping finite set Ω_i, $i \in S = \{1, ..., M\}$:

$$\sum_{i=1}^{M} \Omega_i = \Omega , \qquad \Omega_i \cap \Omega_j = \emptyset \quad \text{if} \quad i \neq j . \tag{3.15}$$

As mentioned, the basis functions are indicator functions on the subdomains:

$$f_i(x) = \mathbf{1}_{\Omega_i}(x) \tag{3.16}$$

Hence, they are discontinuous. With this choice for f_i, R in (3.3a) becomes a diagonal matrix,

$$R_{ij} = \delta_{ij} \rho_i \quad \text{with} \quad \rho_i = \int_{\Omega_i} \rho(x) dx , \tag{3.17}$$

so that (3.4) and (3.6) are reduced from generalized to regular eigenvalue problems.

With (3.16), the estimators $\hat{R}$ and $\hat{T}$ in (3.12) become

$$\hat{R}_{ij} = \delta_{ij}\hat{\rho}_j \quad \text{with} \quad \hat{\rho}_j = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{1}_{\Omega_j}(X_{n\tau})\,, \tag{3.18a}$$

$$\hat{T}_{ij} = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{1}_{\Omega_j}(X_{n\tau})\mathbf{1}_{\Omega_i}(X_{(n+1)\tau})\,. \tag{3.18b}$$

Calculating the spectrum from these estimators is equivalent to calculating the spectrum of the maximum-likelihood estimator (MLE) $\hat{P}$ for transitions on S. The elements of $\hat{P}$ are estimators for the conditional probabilities $p_{ij} = \mathbb{P}(X_{t+\tau} \in \Omega_j | X_t \in \Omega_i)$. They are given by

$$\hat{p}_{ij} = \begin{cases} \dfrac{k_{ij}^{(N)}}{\sum_j k_{ij}^{(N)}} & \text{if} \quad \sum_j k_{ij}^{(N)} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

where $K^{(N)}$ is the frequency matrix with elements

$$k_{ij}^{(N)} = \sum_{n=0}^{N-1} \mathbf{1}_{\Omega_i}(X_{n\tau})\mathbf{1}_{\Omega_j}(X_{(n+1)\tau})\,. \tag{3.20}$$

Comparing (3.18), (3.19) and (3.20), we see that

$$\hat{R}\,\hat{P} = \hat{T}^T\,. \tag{3.21}$$

Let us assume that $\hat{P}$ admits the spectral decomposition

$$\hat{P} = \hat{U}\hat{D}_\Lambda\hat{U}^{-1}\,. \tag{3.22}$$

This identity is equivalent to the generalized eigenvalue problems in (3.14) if we identify

$$\hat{U} = \hat{V}^T\,, \qquad \hat{U}^{-1} = (\hat{R}\,\hat{W}^*)^T\,. \tag{3.23}$$

3.1.4. Galerkin representation of $\mathscr{L}$ and its spectrum. If it is assumed that the basis functions $f_i(x)$ are all twice differentiable, so that $f_i \in \mathrm{dom}(\mathscr{L})$, the Galerkin method can also be applied to the eigenvalue problems $\mathscr{L}\phi_k = \lambda_k\phi_k$ and $\mathscr{L}^*\psi_k = \bar{\lambda}_k\psi_k$. Because of the relations (1.6a) and (1.6b) this results in

$$V\,Q = D_\lambda\,V\,R \tag{3.24a}$$

$$Q\,W^* = R\,W^*\,D_\lambda \tag{3.24b}$$

where Q is the matrix with elements

$$Q_{ij} = \langle \mathscr{L}f_i, f_j\rangle_\rho \tag{3.25}$$

and D_λ is the matrix

$$D_\lambda = \mathrm{diag}(\lambda_1^g, ..., \lambda_M^g) \tag{3.26}$$

with $\lambda_k^g = \tau^{-1}\log \Lambda_k^g$, cf. (1.7). Using (3.9) we find

$$V\,Q\,W^* = D_\lambda \tag{3.27}$$

The last identity suggests to infer $\mathscr{L}$ by minimizing the residual matrix $V\,Q\,W^* - D_\lambda$. This will be discussed in section 4.

If the f_i are not smooth, as in the binning method, the resulting eigenfunction approximations are not in dom($\mathscr{L}$). This poses no problem, because the procedure presented in section 4 allows us to infer $\mathscr{L}$ without letting $\mathscr{L}$ (or its adjoint) act on the estimated eigenfunctions. Thus, although $\mathscr{L}$ is a differential operator, it is not necessary to estimate the derivatives of ϕ_k, ξ_k or ψ_k in order to infer $\mathscr{L}$.

To conclude, we point out once more that for diffusion processes, eigenfunctions of P_τ or P_τ^* are also eigenfunctions of $\mathscr{L}$ or $\mathscr{L}^*$, see (1.6a) and (1.6b). Thus, the matrix estimates $\hat{V}$ and $\hat{W}$, obtained with the Galerkin method, determine the (eigen)function estimates $\hat{\phi}_k^g$ and $\hat{\xi}_k^g$ associated with $\mathscr{L}$ and $\mathscr{L}^*$. The eigenvalues of P_τ and $\mathscr{L}$ are related through (1.7), so that the diagonal matrix with estimates of the eigenvalues of $\mathscr{L}$ is

$$\hat{D}_\lambda = \mathrm{diag}(\hat{\lambda}_1^g, ..., \hat{\lambda}_M^g) = \tau^{-1}\mathrm{diag}(\log \hat{\Lambda}_1^g, ..., \log \hat{\Lambda}_M^g) \tag{3.28}$$

3.2. Sampling and discretization errors. There are two sources of error for the estimated triplets $(\hat{\phi}_k^g, \hat{\xi}_k^g, \hat{\lambda}_k^g)$: finite sample size N and finite discretization level M. The former results in sampling error (the difference $\hat{\phi}_k^g - \phi_k^g$), the latter in discretization error (the difference $\phi_k^g - \phi_k$). For the total error we have

$$\|\hat{\phi}_k^g - \phi_k\| \le \|\hat{\phi}_k^g - \phi_k^g\| + \|\phi_k^g - \phi_k\|\,, \tag{3.29}$$

and similarly for $\|\hat{\xi}_k^g - \xi_k\|$ and $|\hat{\lambda}_k^g - \lambda_k|$. We will not analyse convergence in detail here, but we have some remarks about it. It is reasonable to expect, under mild conditions (e.g., $\max_j \mathrm{Vol}(\Omega_j) \to 0$ as $M \to \infty$ in the case of binning, or more generally $\mathscr{F}_M \to \mathscr{F}$ as $M \to \infty$), that the sampling and discretization errors vanish as $N, M \to \infty$ (and τ remains fixed):

$$\|\hat{\phi}_k^g - \phi_k^g\| \to 0 \quad \text{as} \quad N \to \infty\,, \tag{3.30a}$$

$$\|\phi_k^g - \phi_k\| \to 0 \quad \text{as} \quad M \to \infty\,, \tag{3.30b}$$

so that

$$\lim_{M\to\infty}\lim_{N\to\infty} \|\hat{\phi}_k^g - \phi_k\| \to 0\,. \tag{3.31}$$

Note that even though the eigenvalues are scalars, they are affected by discretization error, in the sense that in general, $|\lambda_k^g - \lambda_k| > 0$ if $M < \infty$.

By ergodicity, the estimators $\hat{P}$, $\hat{R}$ and $\hat{T}$ converge to P, R, T as $N \to \infty$. The convergence of the eigenvalues and eigenvectors of $\hat{P}$ to those of P as $\hat{P} \to P$ was analyzed in detail in [15]. For other Galerkin approximations than the binning method, the analysis is more complicated because it involves a generalized eigenvalue problem instead of a regular one. In [22], rigorous results are given for the case of a reversible scalar diffusion on a bounded domain. The asymptotics of the discretization errors as $M \to \infty$ is treated in many texts on Galerkin methods, see e.g. [6] and references therein. In [37, 24], the approximation of transfer operators (such as P_t^*) and their spectra by discretization of Ω is investigated extensively. The literature on

Ulam's method also contains convergence results relevant in this context [26, 21, 18]. We leave further analysis for a future study.

For the particular case of the binning method, the error due to finite M and hence to finite bin volumes is tightly connected to the sampling interval of the data: the smaller τ, the smaller the bins must be to avoid bias in the estimated eigenvalues. This will be demonstrated in the next section. It can be particularly problematic in case of multivariate processes: M will increase very rapidly by decreasing bin volumes, easily leading to an intractable number of bins and/or severe undersampling.

Finally, we note that besides finite sample size and finite discretization level, model misspecification may also be a source of error. The observations may have been generated by a process that is not a diffusion. Alternatively, it may be the case that one observes a true diffusion, but only part of it, or that the data is contaminated by observation error. In section 5 we consider a generic situation of model misspecification, by analyzing estimation of a coarse-grained (homogenized) diffusion process from data of a multiscale diffusion, and quantifying the model errors involved.

3.3. Numerical example: OU process. As an illustration of the issues discussed in this section, we present a numerical example. From discretely sampled timeseries of the Ornstein-Uhlenbeck (OU) process, we estimate the leading eigenfunctions $\phi_k(x)$ and eigenvalues λ_k. Because the OU process is one of the rare cases for which the spectrum of the diffusion operator is known exactly, we can assess the estimation errors on the spectrum.

The SDE for the OU process is

$$dX_t = -X_t\,dt + dW_t \tag{3.32}$$

with $X_t \in \mathbb{R}$. As usual, W_t is a Wiener process. The associated diffusion operator is

$$\mathscr{L} = -x\frac{\partial}{\partial x} + \frac{1}{2}\frac{\partial^2}{\partial x^2} \tag{3.33}$$

As is well known, the OU process is reversible and its invariant density is $\rho(x) = \pi^{-1/2}\exp(-x^2)$. The eigenvalues of $\mathscr{L}$ are $0, -1, -2, -3, ...$; the eigenfunctions are the Hermite polynomials, $\phi_1 = 1$, $\phi_2 = 2x$, $\phi_3 = 4x^2 - 2$, $\phi_4 = 8x^3 - 12x$, etc.

We generate a timeseries of $N = 10^4$ datapoints with sampling interval $\tau = 0.1$ by numerically integrating the SDE (3.32). From this timeseries we estimate the spectrum of $\mathscr{L}$. For the Galerkin method with smooth basis functions we use $f_i(x) = x^i$ with $i = 0, ..., M = 10$. For the binning method we use 100 equally sized bins ($M = 100$), with the first (last) located such that the minimum (maximum) of the timeseries falls in it. The matrices R and T are estimated using (3.13a), (3.13b).

In figure 1 we show the estimate $\hat{\phi}_2$ as well as the exact $\phi_2 = 2x$ (for comparison). To highlight the region where ρ is not small, we plot $\rho\,\hat{\phi}_2$ and $\rho\,\phi_2$ rather than $\hat{\phi}_2$ and ϕ_2. As is clear from the figure, both methods reproduce the correct ϕ_2 quite well.

The estimated eigenvalues show significant bias due to finite bin size when the number of bins is small. This bias is investigated in table 1. We estimate the leading eigenvalue λ_2 of $\mathscr{L}$ from 100 different sample paths of the OU process, using both methods. The table shows the means and standard deviations of $\hat{\lambda}_2^g$; the exact value is $\lambda_2 = -1$. The calculations are repeated for varying discretization levels ($M = 100, 20, 10$ for binning, $M = 10, 7, 4$ for smooth Galerkin) and varying sampling intervals ($\tau = 0.01, 0.1, 1$). Most striking is the strong bias of the binning method if both τ and M are small. The bias largely disappears if either τ or M increases. The

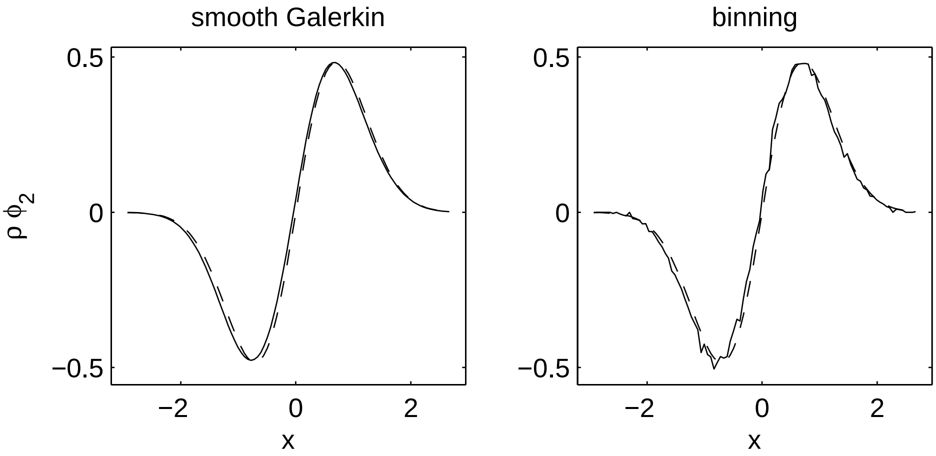

FIG. 1. *Estimated and exact eigenfunctions of the diffusion operator for the Ornstein-Uhlenbeck process. The eigenfunctions are multiplied with the (exact) invariant density $\rho = \pi^{-1/2}\exp(-x^2)$ of the OU process, in order to highlight the region where ρ is not small. Dashed curves are for $\rho\,\phi_2$ with ϕ_2 the exact eigenfunction, $\phi_2 = 2x$. Solid curves are for $\rho\,\hat{\phi}_2$, where $\hat{\phi}_2$ is the estimate of ϕ_2. In the left panel, $\hat{\phi}_2$ was obtained using the Galerkin method with smooth basis functions $x^0, x^1, ..., x^{10}$. In the right panel, the binning method with 100 bins was used.*

TABLE 1

Estimates of the eigenvalue λ_2 for the diffusion operator of the Ornstein-Uhlenbeck process. From 100 different sample paths of the process, each with $N = 10^4$ data points and sampling interval τ, λ_2 was estimated using both the binning and the smooth Galerkin method. M is the discretization level (number of bins, or number of smooth basis functions, see text). Shown are the mean and standard deviation (std) of the 100 estimates $\hat{\lambda}_2^g$, for varying τ and M. The true value is $\lambda_2 = -1$. As can be seen, the binning method is strongly biased if both the sampling interval and the number of bins are small.

data		binning			smooth Galerkin		
N	τ	M	mean $\hat{\lambda}_2^g$	std $\hat{\lambda}_2^g$	M	mean $\hat{\lambda}_2^g$	std $\hat{\lambda}_2^g$
10^4	0.01	100	-1.02	0.16	10	-1.01	0.15
10^4	0.01	20	-1.79	0.23	7	-1.02	0.16
10^4	0.01	10	-3.35	0.42	4	-1.02	0.13
10^4	0.1	10	-1.43	0.10	10	-0.99	0.06
10^4	1	10	-1.05	0.02	10	-0.99	0.03

estimates using smooth Galerkin basis functions are not noticeably affected by small sampling intervals.

In this numerical example, the smooth Galerkin method is superior to the binning method. However, it must be kept in mind that here, it was easy to choose a suitable basis of functions $f_i(x)$ for smooth Galerkin. Because the eigenfunctions are Hermite polynomials in case of an OU process, by taking $f_i(x) = x^i$ we selected just the right subspace $\mathscr{F}_M$ to approximate the leading eigenfunctions. For many other processes, a good choice of smooth basis functions will be more difficult.

4. Inference of $\mathscr{L}$ from eigenpairs. The previous section was devoted the problem of estimating eigenpairs of $\mathscr{L}$ and $\mathscr{L}^*$ from observations of X_t. In this section we discuss how $\mathscr{L}$ can be inferred from these eigenpairs.

Given the leading eigentriplets $(\phi_k, \psi_k, \lambda_k)$, $k \leq K$, we want to identify $b(x)$ and $a(x)$ such that

$$\mathscr{L}(b,a)\phi_k = \lambda_k \phi_k \tag{4.1a}$$

$$\mathscr{L}^*(b,a)\psi_k = \bar{\lambda}_k \psi_k \tag{4.1b}$$

for all $k \leq K$, under the constraint that a be positive semi-definite everywhere. Other constraints on b or a may apply in specific situations (depending on e.g. application, geometry, boundary conditions); we do not specify these further. We summarize the constraints by requiring $b \in \Theta_b$, $a \in \Theta_a$.

This inverse problem can be approached in several ways. In [23, 22], (4.1) is solved exactly for univariate diffusions, resulting in explicit expressions for b and a in terms of λ_2, ψ_1, ϕ_2 and its derivatives $\partial_x \phi_2$, $\partial_{xx}\phi_2$. However, this procedure requires estimates of $\partial_x \phi_2$, $\partial_{xx}\phi_2$, introducing a major source of error (in [22] this differentiation is interpreted as an ill-posed operation). Furthermore, (4.1) may have no solution $(b,a) \in (\Theta_b, \Theta_a)$ at all, due to e.g. sampling error or model misspecification. Finally, it will be difficult to generalize this approach to multivariate processes.

An alternative to solving (4.1) exactly is to minimize $\|\mathscr{L}(b,a)\phi_k - \lambda_k \phi_k\|^2$ and/or $\|\mathscr{L}^*(b,a)\psi_k - \bar{\lambda}_k \psi_k\|^2$, summed over $k \leq K$. This approach was proposed in [14], where the binning method was used for estimation of eigenpairs, and b and a were discretized on the same set of bins as the eigenfunctions. The procedure in [14] is nonparametric, but requires estimates of eigenfunction derivatives. With enough data and small bins, so that these derivatives can be calculated reliably by finite differences, this is a feasible strategy, as was demonstrated in [14]. Notwithstanding, in this section we present a modification of the procedure from [14]. This modified procedure is much more robust against sampling error, because it allows to avoid eigenfunction differentiation. It also makes a natural connection with the Galerkin representation of $\mathscr{L}$.

4.1. A new objective function. Let $\sigma_i(x)$, $i = 1, ..., N_\sigma$ be a collection of test functions, $\sigma_n \in \mathrm{dom}(\mathscr{L})$. Instead of minimizing the residuals $\mathscr{L}^*\psi_k - \bar{\lambda}_k \psi_k$ directly, we can integrate them against the σ_i and minimize the (squared) integrals. Using the adjoint property as well as (2.1) gives $\langle \mathscr{L}^*\psi_k - \bar{\lambda}_k \psi_k, \sigma_i \rangle_1 = \langle \xi_k, \mathscr{L}\sigma_i \rangle_\rho - \bar{\lambda}_k \langle \xi_k, \sigma_i \rangle_\rho$. Hence, given the estimates $(\hat{\bar{\lambda}}_k, \hat{\xi}_k)$, $k \leq K$, we propose to estimate b, a by minimization of the objective function

$$E(b,a) = \sum_{k=1}^{K} \sum_{i=1}^{N_\sigma} \alpha_{ki} \left| \langle \hat{\xi}_k, \mathscr{L}(b,a)\sigma_i \rangle_\rho - \hat{\bar{\lambda}}_k \langle \hat{\xi}_k, \sigma_i \rangle_\rho \right|^2 , \tag{4.2}$$

with nonnegative constant weights α_{ki}. We discuss three different ways to use (4.2).

4.1.1. Smooth Galerkin. If the eigenpairs are estimated with smooth Galerkin basis functions, it is natural to use the estimated eigenfunctions as test functions:

$$\sigma_i = \hat{\phi}_i^g \tag{4.3}$$

We take the weights $\alpha_{ki} = c_k c_i$ with $c_k \in [0, \infty)$. The objective function (4.2) now reads

$$E^g(b,a) = \|\hat{V}\hat{Q}\hat{W}^* - \hat{D}_\lambda\|_c^2 \tag{4.4}$$

where $\|\cdot\|_c^2$ denotes a weighted Frobenius norm: given any square matrix A with entries a_{ij},

$$\|A\|_c^2 = \sum_{i,j} c_i c_j |a_{ij}|^2 \tag{4.5}$$

The matrices $\hat{V}$, $\hat{W}$ and $\hat{D}_\lambda$ were defined in section 3.1. The elements of the matrix $\hat{Q} = \hat{Q}(b,a)$ are estimates of $\langle \mathscr{L}(b,a) f_i, f_j \rangle_\rho$, see (3.25). Note that ρ is the true invariant density of the process X_t, not to be confused with the invariant density associated with $\mathscr{L}(b,a)$. Similar to the matrices T and R, see (3.11), the elements of $Q(b,a)$ can be cast as expectations, therefore they can be estimated with

$$\hat{Q}_{ij}(b,a) = \frac{1}{N+1} \sum_{n=0}^{N} f_j(X_{n\tau})(\mathscr{L}(b,a) f_i)(X_{n\tau})\,. \tag{4.6}$$

We remark that the identity (3.27) already suggested an objective function that minimizes $\hat{V}\hat{Q}\hat{W}^* - \hat{D}_\lambda$, as in (4.4). Furthermore, (4.4) is almost identical to the objective function used in [15] for the inference of generators for Markov jump processes from discrete samplings. There, Q itself is the generator, whereas here Q is a matrix that represents the action of the generator $\mathscr{L}(b,a)$ on the subspace $\mathscr{F}_M$. Following [15], we propose to relate the weights to the eigenvalues:

$$c_k = |\hat{\Lambda}_k^g|^\delta \tag{4.7}$$

with some $\delta \geq 0$.

4.1.2. Binning. If the eigenpairs are estimated with the binning method, we choose smooth test functions and write the inner product in (4.2) as the expectation

$$\mathbb{E}\left[\, \hat{\xi}_k^g(X_t)(\mathscr{L}\sigma_i)(X_t) - \hat{\bar{\lambda}}_k^g \hat{\xi}_k^g(X_t)\sigma_i(X_t)\right]. \tag{4.8}$$

We denote by $\hat{\Sigma}_{ki}$ the estimator of this expectation:

$$\hat{\Sigma}_{ki} = \frac{1}{N+1} \sum_{n=0}^{N} \left[\, \hat{\xi}_k^g(X_{n\tau})(\mathscr{L}\sigma_i)(X_{n\tau}) - \hat{\bar{\lambda}}_k^g \hat{\xi}_k^g(X_{n\tau})\sigma_i(X_{n\tau})\right] \tag{4.9}$$

Then (4.2) can be written as

$$E^b(b,a) = \sum_{k,i} \alpha_{ki} \left|\hat{\Sigma}_{ki}\right|^2 \tag{4.10}$$

We emphasize that in (4.10), no eigenfunction derivatives are used. The functions $\hat{\xi}_k^g$ are obtained from $\hat{\xi}_k^g = \hat{\psi}_k^g / \hat{\psi}_1^g$, cf. (2.1). Clearly, this is ill-defined at points where $\hat{\psi}_1^g = 0$. However, $\hat{\xi}_k^g$ is only evaluated at the observed datapoints $X_{n\tau}$, where $\hat{\psi}_1^g > 0$.

4.1.3. Mixed. One can mix the previous approaches, by estimating eigenpairs with the smooth Galerkin method and using test functions that are not eigenfunctions. If we pick the Galerkin basis functions as test functions, $\sigma_i = f_i$, we obtain the objective function

$$E^m(b,a) = \sum_{k,i} \alpha_{ki} \left| \left(\hat{W}\hat{Q}^T - \hat{D}_{\bar{\lambda}}\hat{W}\hat{R}\right)_{ki} \right|^2 \tag{4.11}$$

where we have used that $\hat{Q}$ is real and thus $\hat{Q}^* = \hat{Q}^T$, as well as $\hat{D}_\lambda^* = \hat{D}_{\bar{\lambda}}$. In (4.11), as in (4.10), $\mathscr{L}(b,a)$ acts on the test functions and not on the estimated eigenfunctions.

4.2. Inference by minimization. The true drift and diffusion functions b_*, a_* associated with the observed process X_t are estimated by minimization:

$$(\hat{b}, \hat{a}) = \underset{b \in \Theta_b,\, a \in \Theta_a}{\arg\min} \; E(b, a) \tag{4.12}$$

where E is the objective function (4.2). We focus on the situation where b and a are each expanded on a basis of linearly independent functions (e.g. polynomials):

$$b(x) = \sum_{j=1}^{N_b} b_j g_j(x)\,, \qquad a(x) = \sum_{j=1}^{N_a} a_j h_j(x)\,. \tag{4.13}$$

The expansion coefficients are denoted by θ,

$$\theta = (b_1, ..., b_{N_b}, a_1, ..., a_{N_a}) \in \Theta\,, \tag{4.14}$$

where $\Theta = \{\theta \,|\, b \in \Theta_b, a \in \Theta_a\}$. In appendix A, we give the expressions for the objective functions E^g, E^b and E^m that follow from (4.13).

In section 4.4 we present examples with low-order expansions ($N_b + N_a \leq 9$). Whether the procedure can be successfully extended to the nonparametric case (limit of infinite expansions) remains to be investigated. The condition $KN_\sigma \geq N_b + N_a$ (discussed below) may be an obstacle for this. Alternative ideas to solving (4.1) nonparametrically for b and a may be found in literature on inverse problems for elliptic systems, e.g. [5, 28, 30]. A much studied problem there is to estimate a from $\partial_x(a\,\partial_x u) + f = 0$, where f is given and u is observed (possibly with observation errors, see e.g. [28]). Although there are differences with the problem considered here, the ideas may have value for finding a procedure to solve (4.1), alternative to what is proposed here and in [14]. We leave this for future study.

With (4.13), $\mathscr{L}$ is linear in θ and the objective function is of the form $E = |A\theta - \gamma|^2$, where A is a $(KN_\sigma) \times (N_b + N_a)$ matrix. Thus, E is convex quadratic and we are dealing with a least squares problem. E is strictly convex if null(A)=0, or equivalently $\langle \mathscr{L}^*(\theta)\hat{\psi}_k, \sigma_i \rangle_1 = 0 \;\forall\; k \leq K, i \leq N_\sigma$ iff $\theta = 0$. Two necessary conditions for this are (i) $KN_\sigma \geq N_b + N_a$, and (ii) $K > 1$ (to see this, note that $\hat{\psi}_1$ is a probability density, so there can be $\theta \neq 0$ such that $\mathscr{L}^*(\theta)\hat{\psi}_1 = 0$).

If E is strictly convex and, additionally, Θ is convex, (4.12) has a unique solution, i.e. there is a single global minimum of E and no other local minimum [12]. The existence of a unique minimum is computationally advantageous. If E is convex quadratic and Θ is convex and determined by linear constraints, (4.12) is a quadratic program (QP) and can be solved using well-established, efficient numerical methods (see e.g. [29]). Notwithstanding, in the numerical examples in this paper we estimate θ with the unconstrained minimum of the objective function.

4.3. Procedure summary. The entire procedure can be summarized as follows. Starting from a timeseries $X_0, X_\tau, ..., X_{N\tau}$ one has to make several steps. For the smooth Galerkin method:

1s. Choose the functions $g_j(x)$ and $h_j(x)$ for the expansions (4.13) of $b(x)$ and $a(x)$, and determine the parameter domain Θ.

2s. Choose the smooth Galerkin basis functions $f_i(x)$. Calculate the estimators $\hat{R}$ and $\hat{T}$ (3.12).

3s. Solve the generalized eigenvalue problems (3.14), resulting in $\hat{V}$, $\hat{W}$ and $\hat{D}_\Lambda$. Calculate $\hat{D}_\lambda$ from $\hat{D}_\Lambda$ (3.28).

4s. Fix the weights α_{ki}. Minimize E^g (4.4), (A.1) or E^m (4.11), (A.6) under variation of $\theta \in \Theta$.

If the binning method is used, one can calculate the MLE $\hat{P}$ and its decomposition (3.22), resulting in $\hat{U}$, $\hat{U}^{-1}$ and $\hat{D}_\Lambda$. As was shown in section 4.1.2, this is equivalent to calculating $\hat{R}$, $\hat{T}$ and solving (3.14). Hence, for the binning method the procedure is:

1b. Choose the functions $g_j(x)$ and $h_j(x)$ for the expansions (4.13) of $b(x)$ and $a(x)$, and determine the parameter domain Θ.
2b. Choose the subdomains Ω_i, thereby determining the Galerkin basis functions $f_i(x) = \mathbf{1}_{\Omega_i}(x)$.
3b. Calculate the MLE $\hat{P}$ and its decomposition (3.22), resulting in $\hat{U}$, $\hat{U}^{-1}$ and $\hat{D}_\Lambda$. Calculate $\hat{D}_\lambda$ from $\hat{D}_\Lambda$ (3.28).
4b. Construct $\hat{\xi}^g_k(x) = \sum_i \mathbf{1}_{\Omega_i}(x)\,(\hat{U}^{-1})_{ki}/(\hat{U}^{-1})_{1i}$.
5b. Choose the test functions σ_i and the weights α_{ki}.
6b. Minimize E^b (4.10), (A.4) under variation of $\theta \in \Theta$.

4.4. Numerical examples. In this section we present several examples, where we estimate parameters from sample paths of various processes, observed at discrete points in time. We numerically investigate consistency, bias and variance of the estimated parameters, we compare the performance of the different objective functions (E^g, E^b, E^m) and their dependence on choices of Galerkin basis functions, number of bins, test functions and weights.

In each example, the true set of parameters will be denoted θ_* and the estimated set $\hat{\theta}$. For a single parameter θ_i, the bias is the difference between the expectation of $\hat{\theta}_i$ and θ_{i*}. We approximate the expectation by calculating the mean of an ensemble of estimates, i.e. $\text{bias}(\hat{\theta}_i) = \text{mean}(\hat{\theta}_i) - \theta_{i*}$. From the same ensemble we calculate the variance $\text{var}(\hat{\theta}_i)$. As measures for the "collective" bias and variance of all elements in $\hat{\theta}$ we use

$$\text{bias}(\hat{\theta}) = \Big(\sum_i \big[\text{bias}(\hat{\theta}_i)\big]^2\Big)^{1/2}, \qquad \text{var}(\hat{\theta}) = \sum_i \text{var}(\hat{\theta}_i) \tag{4.15}$$

Thus, $\text{bias}(\hat{\theta})$ is defined as the L^2 distance in parameter space between the expectation of $\hat{\theta}$ and θ_*; $\text{var}(\hat{\theta})$ is simply the sum of the variances of the individual parameter estimates.

4.4.1. Diffusion on T^1. The first example concerns the process with SDE

$$dX_t = (1 + 0.5\sin(X_t))dt + \sqrt{1 + 0.3\cos(X_t)}\,dW_t \tag{4.16}$$

where $X_t \in [0, 2\pi]$ with periodic boundary conditions. This process has nonconstant diffusion (multiplicative noise) and is nonreversible (indicated by the presence of complex pairs of eigenvalues, e.g. $\lambda_{2,3} \approx -0.56 \pm i\,0.96$). We fit the diffusion operator with drift $b(x) = b_1 + b_2\cos(x) + b_3\sin(x)$ and diffusion $a(x) = a_1 + a_2\cos(x) + a_3\sin(x)$ to timeseries generated by the SDE (4.16). As is clear, the true set of parameters is $\theta_* = (b_{1*}, ..., a_{3*}) = (1, 0, 0.5, 1, 0.3, 0)$.

We generate timeseries by numerically integrating the SDE (4.16) using the Euler scheme with time step 0.001. Their length N varies from 10^3 to 10^6; all timeseries have sampling interval $\tau = 0.1$. For the smooth Galerkin method we use Fourier basis functions, $\cos(ix)$ and $\sin(ix)$ with $i = 0, 1, ..., N_f$. The test functions used in E^b are also Fourier functions. For each value of N we infer the parameters from 100 different numerically generated paths of X_t.

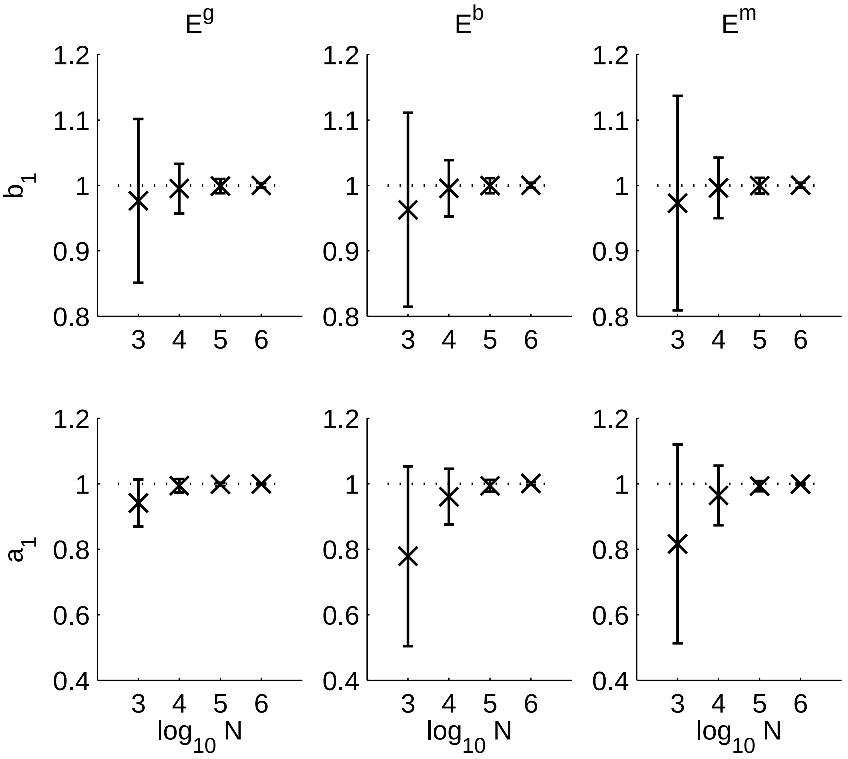

FIG. 2. *Example on T^1. The parameters of the process (4.16) are estimated using the three different objective functions E^g (4.4), E^b (4.10) and E^m (4.11). For precise choices of Galerkin basis functions, test functions etc., see text. Shown are the means and standard deviations of the parameters b_1 and a_1 inferred from 100 different sample paths, each N datapoints long. The dotted lines indicate the value of the true parameters b_{1*}, a_{1*}.*

First we compare the three different objective functions, E^g, E^b, and E^m. With E^g we set $N_f = 5$ for the basis functions (i.e., $M = 11$) and $\delta = \log(0.5)/\log|\hat{\Lambda}_2|$ for the weights c_k (4.7), so that $c_2 = 0.5$. For E^b we use $M = 200$ bins and test functions $\sigma_i = \cos(x), \sin(x), \cos(2x), \sin(2x)$. The weights are set to $\alpha_{ki} = 1$ for all $k = 1, 2, 3$ and $i = 1, ..., 4$, and zero otherwise. The settings for E^m are consistent with those used for E^g and E^b (i.e., $N_f = 5$ and $\alpha_{ki} = 1$ for $k = 1, 2, 3$ and $i = 2, ..., 5$). In figure 2 we show the means and standard deviations of the 100 estimates for b_1 and a_1 for increasing values for N, as representative examples of individual parameters. In figure 3 we plot bias($\hat{\theta}$) and var($\hat{\theta}$), as defined in (4.15). All three objective functions show convergence of the estimates, i.e. $\hat{\theta} \to \theta_*$ (or bias($\hat{\theta}$) $\to 0$) as N grows. The decrease of var($\hat{\theta}$) is nearly proportional to N^{-1}. The Galerkin objective function E^g shows the best performance, in particular for the diffusion parameters a_j.

A natural question to ask is how the accuracy of the estimation is affected by the choices for the Galerkin basis functions, test functions and weights. The left panel of figure 4 shows bias($\hat{\theta}$) versus N for E^g with $N_f = 2, 5$ and 8 (weights are such that $c_2 = 0.5$). With $N_f = 2$ the bias only marginally decreases beyond $N = 10^4$, a sign

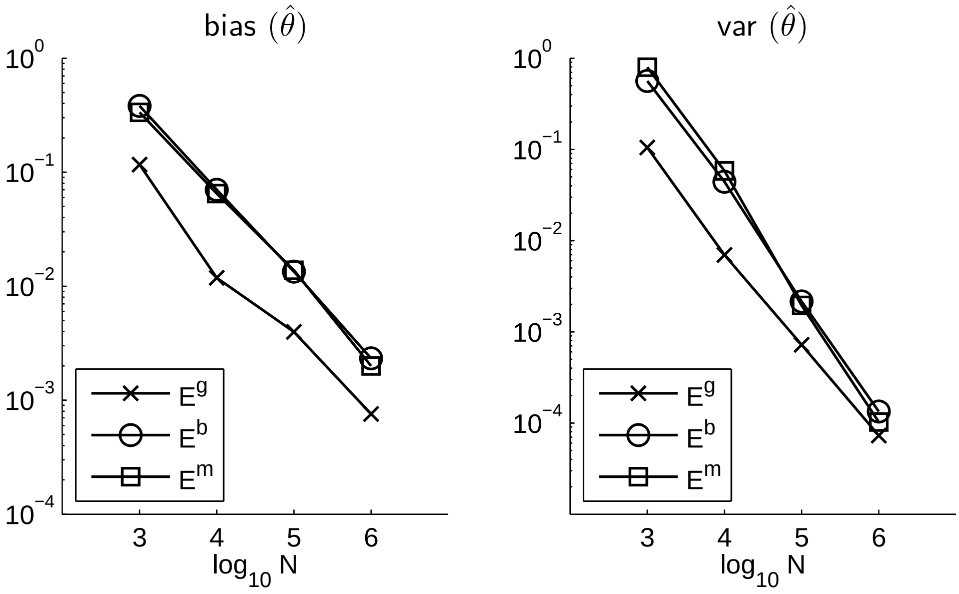

FIG. 3. *Example on T^1. Results are from the same calculations as as those in figure 2. Shown are the bias and variance (for all parameters $b_1, ..., a_3$), bias$(\hat{\theta})$ and var$(\hat{\theta})$, as defined in (4.15).*

that the convergence is halted by discretization error.

In the right panel of figure 4 we use E^b with varying numbers of bins, $M = 20, 70, 200$. As before, the test functions are $\cos(jx)$ and $\sin(jx)$ with $j = 1, 2$), and the weights are set at $\alpha_{ki} = 1$ for $k = 1, 2, 3$ and $i = 1, ..., 4$. With $M = 20$ the bias no longer decreases beyond $N = 10^4$, with $M = 70$ the decrease halts beyond $N = 10^5$. This is due to discretization error (finite bin volumes). With short timeseries ($N = 10^3$), $M = 20$ gives only marginally better results than $M = 200$ and $M = 70$, so there is no reason not to choose a large number of bins.

As for the weights, the higher δ, the steeper the weights c_k (4.7) decrease with increasing k, hence the more weight is put on the leading estimated eigenfunctions in the objective function E^g. With $c_2 = 1$ (and hence $c_k = 1$ for all k) there is too much weight on the non-leading eigenpairs; with $c_2 = 0.1$ there is too much emphasis on $k = 1$. The intermediate value $c_2 = 0.5$ (i.e. $\delta = \log(0.5)/\log|\hat{\Lambda}_2|$) gives the best results (figure not shown). Finally, increasing the number of test functions from $N_\sigma = 4$ to $N_\sigma = 6, 8$ in E^b was found to degrade the performance (results not shown).

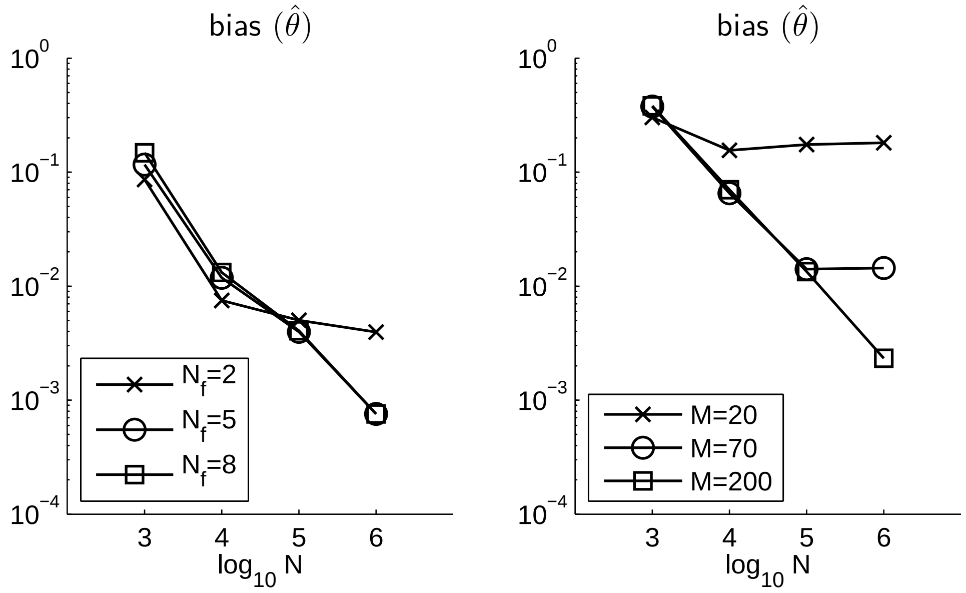

FIG. 4. *Example on T^1. Left: results using E^g with different values for N_f, the highest wavenumber of the Fourier Galerkin basis functions. Right: results from E^b with different values for M, the number of bins. Both panels show bias$(\hat{\theta})$, as defined in (4.15).*

4.4.2. Double-well potential in $\mathbb{R}^1$. In this example we consider a diffusion process on $\mathbb{R}^1$ with SDE

$$dX_t = -V'(X_t)dt + \sqrt{1+X_t^2}\,dW_t \quad \text{with} \quad V(x) = (1-x^2)^2\,. \tag{4.17}$$

This process is reversible and is driven by multiplicative noise. It has a bimodal invariant density, $\rho \propto (1+x^2)^7 \exp(-4\,x^2)$, due to the double-well structure of the potential V. The process is metastable: it switches between the two wells on a relatively long timescale, as indicated by the separation between the second and third eigenvalue: $\lambda_2 \approx -0.5$, $\lambda_3 \approx -4.7$, $\lambda_4 \approx -9.1$.

We fit the diffusion operator with $b(x) = b_1 + b_2 x + b_3 x^2 + b_4 x^3$ and $a(x) = a_1 + a_2 x + a_3 x^2$. The vector of true parameters is $\theta_* = (b_{1*}, ..., a_{3*}) = (0, 4, 0, -4, 1, 0, 1)$. Sample paths are generated with sampling interval $\tau = 0.1$ using the Milstein scheme with time step 0.0001. Both for the Galerkin basis functions and for the test functions we take monomials: $f_i = x^i$ with $i = 0, 1, .., N_f$ and $\sigma_i = x^i$ with $i = 1, ..., N_\sigma$.

In figures 5 and 6 we compare E^g, E^b and E^m. For E^g we take $N_f = 15$ and δ such that $c_2 = 0.5$. For E^b we use 200 bins, $K = 2$ and $N_\sigma = 4$. For E^m we take $N_f = 15$, $K = 2$ and $N_\sigma = 4$. Different from the previous example, E^b and E^m perform significantly better than E^g (compare figures 3 and 6). We hypothesize that

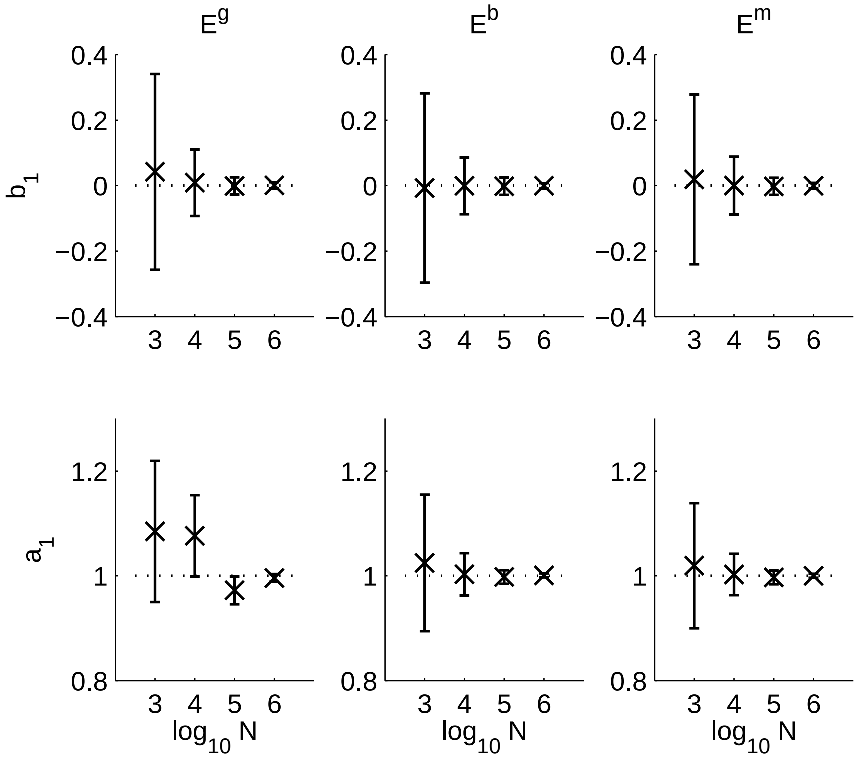

FIG. 5. *Example with double-well potential on $\mathbb{R}^1$. The parameters of the process (4.17) are estimated using the three different objective functions E^g (4.4), E^b (4.10) and E^m (4.11). For precise choices of Galerkin basis functions, test functions etc., see text. Shown are the means and standard deviations of the parameters b_1 and a_1 inferred from 100 different sample paths, each N datapoints long. The dotted lines indicate the value of the true parameters b_{1*}, a_{1*}.*

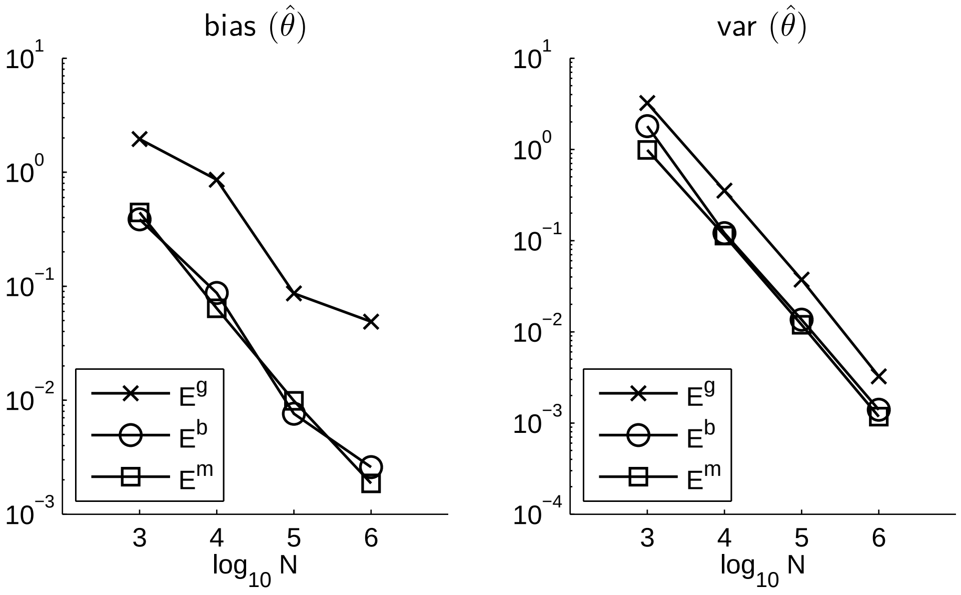

FIG. 6. *Example with double-well potential on $\mathbb{R}^1$. Results are from the same calculations as those in figure 5. Shown are the bias and variance (for all parameters $b_1, ..., a_3$), bias($\hat{\theta}$) and var($\hat{\theta}$), as defined in (4.15).*

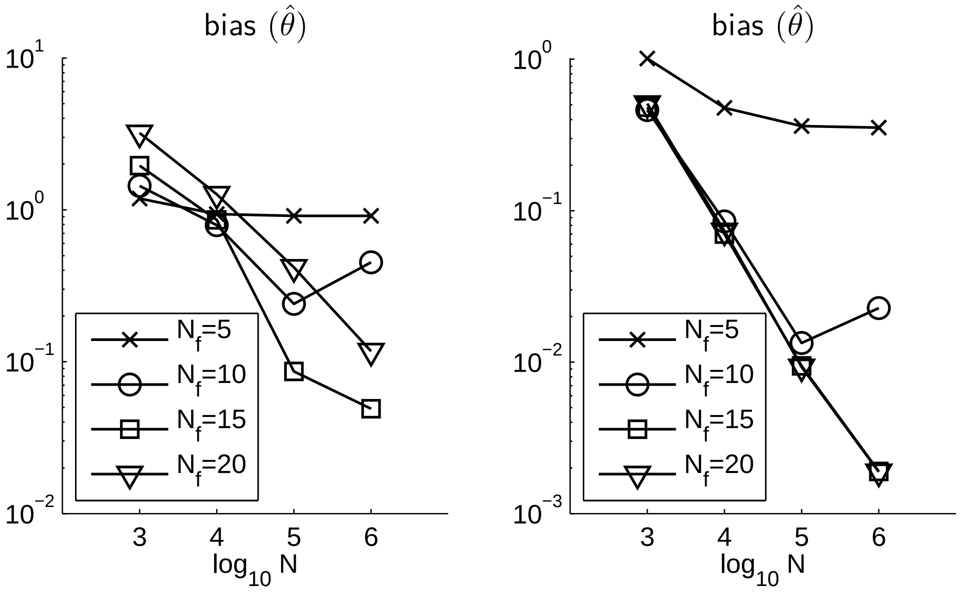

FIG. 7. *Example with double-well potential on $\mathbb{R}^1$. Left: results using E^g with increasing number of smooth Galerkin basis functions (N_f). Right: results using E^m with increasing N_f. Both panels show bias($\hat{\theta}$), as defined in (4.15). Note the different vertical scalings in both panels.*

this is due to the high degree ($N_f = 15$) of the polynomials used to represent the eigenfunctions in case of E^g and E^m. With E^m, this is mitigated because only up to quartic test functions are used ($N_\sigma = 4$).

In figure 7 we show results obtained with E^g and E^m using different numbers of Galerkin basis functions ($N_f = 5, 10, 15, 20$). If N_f is too low, the eigenfunctions are not well enough represented for further bias reduction beyond $N = 10^4$. With N_f too high, the results using E^g are affected by the high polynomial degree of the test functions. By contrast, E^m performs well with $N_f = 15, 20$.

	b_1	b_2	b_3	b_4	b_5	a_1	a_2	a_3	a_4
true	2	0.5	-0.3	1	0.5	2	1	1	0.2
E^g mean	2.00	0.50	-0.30	1.00	0.50	1.99	1.004	0.995	0.20
E^g std	0.02	0.02	0.02	0.01	0.02	0.01	0.007	0.009	0.02
E^b mean	1.7	0.42	-0.3	1.00	0.50	1.6	1.027	0.8	0.2
E^b std	0.3	0.07	0.1	0.01	0.01	0.3	0.005	0.1	0.2

TABLE 2
Results for T^2 example.

4.4.3. Diffusion on T^2. With this example we demonstrate that the inference approach discussed in this paper is capable of handling a challenging case: we consider a nonreversible, multivariate (2-dimensional) process with both cross-diffusion ($a_{12} = a_{21} \neq 0$) and multiplicative noise ($\nabla a_{11} \neq 0$). The SDE of the process is

$$dX_t = (2 + 0.5\cos(X_t) - 0.3\cos(Y_t))\,dt + dV_t + \sqrt{1 + 0.2\cos(Y_t)}\,dW_t\,, \tag{4.18a}$$
$$dY_t = (1 + 0.5\cos(Y_t))\,dt + dV_t\,. \tag{4.18b}$$

V_t and W_t are independent Wiener processes. The domain is doubly periodic, $(X_t, Y_t) \in [0, 2\pi] \times [0, 2\pi]$. The diffusion matrix associated with this process is

$$a(x, y) = \begin{pmatrix} 2 + 0.2\cos(y) & 1 \\ 1 & 1 \end{pmatrix} \tag{4.19}$$

The leading eigenvalues form complex pairs, indicating that the process is nonreversible: $\lambda_{2,3} \approx -0.56 \pm i0.96$.

We fit a diffusion process with drift

$$b(x, y) = b_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + b_2 \begin{pmatrix} \cos(x) \\ 0 \end{pmatrix} + b_3 \begin{pmatrix} \cos(y) \\ 0 \end{pmatrix} + b_4 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + b_5 \begin{pmatrix} 0 \\ \cos(y) \end{pmatrix} \tag{4.20}$$

and diffusion

$$a(x, y) = a_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + a_3 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + a_4 \begin{pmatrix} \cos(y) & 0 \\ 0 & 0 \end{pmatrix} \tag{4.21}$$

The true parameter values are $(b_{1*}, ..., a_{4*}) = (2, 0.5, -0.3, 1, 0.5, 2, 1, 1, 0.2)$.

The parameters are estimated from 100 different sample paths of the process (X_t, Y_t), each sample path being 10^5 datapoints long with sampling interval $\tau = 0.1$. We use E^g and E^b. For E^g we use Galerkin basis functions $\cos(mx)\cos(ny)$, $\cos(mx)\sin(ny)$, $\sin(mx)\cos(ny)$ and $\sin(mx)\sin(ny)$, with $m, n = 0, 1, 2$. For the weights we take δ such that $c_2 = 0.5$. For E^b we use 50×50 bins, $K = 3$ and test functions $\cos(x)$, $\sin(x)$, $\cos(y)$, $\sin(y)$, $\cos(x)\cos(y)$, $\sin(x)\cos(y)$, $\cos(x)\sin(y)$, $\sin(x)\sin(y)$.

In table 2 we summarize the results by showing the means and standard deviations of the estimated parameters. The binning-based approach has difficulties estimating the parameters that appear in the SDE for X_t; the other parameters (b_4, b_5, a_2) are estimated rather well. We hypothesize that this is due to the cross-diffusion term and the multiplicative noise term in the SDE for X_t. The approach using E^g gives good results, showing no significant bias and fairly small errors.

5. Inference of multiscale diffusions. We consider the diffusion process $(X_t, Y_t) \in \Omega_x \times \Omega_y \subset \mathbb{R}^n \times \mathbb{R}^m$ with SDEs

$$dX_t = \left(\frac{1}{\epsilon}F_1(X_t, Y_t) + F_0(X_t, Y_t)\right) dt + \alpha(X_t, Y_t) dW_t^x \tag{5.1a}$$

$$dY_t = \frac{1}{\epsilon^2}G(X_t, Y_t)dt + \frac{1}{\epsilon}\beta(X_t, Y_t)dW_t^y \tag{5.1b}$$

where W_t^x and W_t^y are independent Wiener processes of dimension n and m, respectively. It is assumed that (i) if X_t is fixed at x, Y_t is ergodic with unique invariant measure $\mu_x(y)$ and (ii) the centering condition

$$\int_{\Omega_y} \mu_x(dy) F_1(x, y) = 0 \quad \forall\ x \in \Omega_x \tag{5.2}$$

is satisfied. For systems of this type, it is known that in the limit $\epsilon \to 0$, X_t converges in law to the solution $\bar{X}_t$ of the effective (homogenized) SDE

$$d\bar{X}_t = \bar{F}(\bar{X}_t)dt + \bar{\alpha}(\bar{X}_t)dW_t^x \tag{5.3}$$

Explicit expressions for the homogenized drift and diffusion $\bar{F}$ and $\bar{\alpha}$ can be found in e.g. [33]. In what follows, it is assumed that the conditional invariant measure for Y_t admits a density $\rho_x(y)$, i.e.

$$\mu_x(dy) = \rho_x(y)dy\,. \tag{5.4}$$

The central question in this section is whether it is possible to estimate the homogenized process (5.3) from data of the multiscale process (5.1). As was discussed and analyzed in [32, 31], estimates can be strongly biased if the sampling interval of the multiscale data is too short (see also [2, 42] for related results, albeit in a different framework). However, different estimation procedures can lead to different perspectives on this question. Here, we put it in the perspective of the estimation-by-eigenpairs procedure discussed in section 4. In section 5.1 we will analyse the eigenfunctions and eigenvalues of the diffusion operator (and its adjoint) associated with (5.1) and relate them to the spectrum of homogenized diffusion operator associated with (5.3). The relation between these spectra is of importance for estimation, as will be discussed in section 5.2. In section 5.3 we consider estimation of (5.3) from partial observations of (5.1) (only X_t observed, not Y_t) and show how partial observation may necessitate subsampling. We present numerical examples in sections 5.4 and 5.5.

5.1. Asymptotics of the diffusion operator and its adjoint. The diffusion operator corresponding to (5.1) is

$$\mathscr{L} = \mathscr{L}_0 + \frac{1}{\epsilon}\mathscr{L}_1 + \frac{1}{\epsilon^2}\mathscr{L}_2 \tag{5.5}$$

with

$$\mathscr{L}_0 = F_0 \cdot \nabla_x + \tfrac{1}{2}(\alpha\alpha^T) : \nabla_x \nabla_x \tag{5.6a}$$

$$\mathscr{L}_1 = F_1 \cdot \nabla_x \tag{5.6b}$$

$$\mathscr{L}_2 = G \cdot \nabla_y + \tfrac{1}{2}(\beta\beta^T) : \nabla_y \nabla_y \tag{5.6c}$$

It is known (e.g. [33]) that the diffusion operator of the homogenized system with SDE (5.3) is

$$\mathscr{L}^H = \Pi\,(\mathscr{L}_0 - \mathscr{L}_1\mathscr{L}_2^{-1}\mathscr{L}_1)\Pi\,, \tag{5.7}$$

where the projection operator Π is defined as

$$(\Pi\, h)(x) = \int_{\Omega_y} dy\,\rho_x(y)h(x,y)\,. \tag{5.8}$$

The operator $\mathscr{L}_2$ is not invertible in general, but condition (5.2) guarantees that $\mathscr{L}_1\Pi\, h$ is orthogonal to the nullspace of the adjoint of $\mathscr{L}_2$, for arbitrary functions $h(x,y)$. By the Fredholm alternative, the equation $\mathscr{L}_2 H = \mathscr{L}_1\Pi\, h$ has a solution H, loosely written as $H = \mathscr{L}_2^{-1}\mathscr{L}_1\Pi\, h$.

The eigenpair (ϕ_k, λ_k), solving

$$\mathscr{L}\phi_k = \lambda_k\phi_k\ , \tag{5.9}$$

can be approximated using the expansions

$$\phi_k = \phi_k^{(0)} + \epsilon\phi_k^{(1)} + \epsilon^2\phi_k^{(2)} + \dots \tag{5.10a}$$

$$\lambda_k = \frac{1}{\epsilon^2}\lambda_k^{(-2)} + \frac{1}{\epsilon}\lambda_k^{(-1)} + \lambda_k^{(0)} + \dots \tag{5.10b}$$

By equating terms of equal power in ϵ, it can be shown (see appendix B) that the leading eigenpairs of $\mathscr{L}$ satisfy

$$\lambda_k^{(-2)} = 0\,, \qquad \lambda_k^{(-1)} = 0\,, \qquad \Pi\,(\mathscr{L}_0 - \mathscr{L}_1\mathscr{L}_2^{-1}\mathscr{L}_1)\phi_k^{(0)} = \lambda_k^{(0)}\phi_k^{(0)} \tag{5.11}$$

where $\phi_k^{(0)}$ depends on x only (i.e., $\Pi\,\phi_k^{(0)} = \phi_k^{(0)}$). This implies that the leading eigenvalues and eigenfunctions of the diffusion operator $\mathscr{L}$ of the full system (5.1) on the one hand and those of the diffusion operator $\mathscr{L}^H$ of the homogenized system (5.3) on the other hand, are the same at leading order:

$$\mathscr{L}\phi_k = \lambda_k\phi_k, \qquad \mathscr{L}^H\phi_k^{(0)} = \lambda_k^{(0)}\phi_k^{(0)} \tag{5.12a}$$

$$\phi_k(x,y) = \phi_k^{(0)}(x) + O(\epsilon) \tag{5.12b}$$

$$\lambda_k = \lambda_k^{(0)} + O(\epsilon) \tag{5.12c}$$

A similar result holds for the eigenpairs of the adjoint operator $\mathscr{L}^*$, see appendix B. The leading eigenfunctions $\psi_k(x,y)$ of $\mathscr{L}^*$ associated with the full system (5.1) have leading order terms $\psi_k^{(0)}(x,y)$ that can be written as $\psi_k^{(0)}(x,y) = u_k(x)\rho_x(y)$. The functions $u_k(x)$ are eigenfunctions of the adjoint $\mathscr{L}^{H*}$ of the diffusion operator of the homogenized system (5.3):

$$\mathscr{L}^*\psi_k = \bar{\lambda}_k\psi_k, \qquad \mathscr{L}^{H*}u_k = \bar{\lambda}_k^{(0)}u_k \tag{5.13a}$$

$$\psi_k(x,y) = u_k(x)\rho_x(y) + O(\epsilon) \tag{5.13b}$$

$$\lambda_k = \lambda_k^{(0)} + O(\epsilon) \tag{5.13c}$$

5.2. Implications for statistical inference. The results (5.12) and (5.13) have important implications for statistical inference of the homogenized diffusion process (5.3) from data of the multiscale process (5.1). From a timeseries for the slow variable(s) X_t of the full system (5.1), can we infer the correct homogenized process (5.3)? In [32] it was shown that for certain types of estimators (quadratic variation of the path for estimating the diffusion, path-space likelihood with respect to a pure diffusion for estimating the drift), one has to be careful about choosing the sampling interval τ. If τ is too short, these estimators will result in biased estimates for the homogenized process, due to finite ϵ. In such cases, subsampling of the data is necessary. However, for longer sampling intervals, the finite difference approximation underlying the estimators in [32] becomes inaccurate. Given the time scale separation ϵ in (5.1), it was found in [32] that the sampling interval should be between $O(\epsilon)$ and $O(1)$. This range may be too narrow so that the estimates suffer from either the error due to finite ϵ or the error due to finite τ (or both).

With the spectral approach discussed in this paper, the situation is different. First of all, there is no finite difference approximation involved that deteriorates with growing τ. The relations (1.6a) and (1.6b) are exact, so that we can avoid approximations whose errors only disappear in the limit $\tau \to 0$. Furthermore, if one can estimate the leading eigenpairs of $\mathscr{L}$ and/or $\mathscr{L}^*$ correctly, one can infer the correct homogenized process, due to the close relations (5.12) and (5.13) between the leading eigenpairs of $\mathscr{L}, \mathscr{L}^*$ and $\mathscr{L}^H, \mathscr{L}^{H*}$. As will be discussed below, estimates of eigenpairs of $\mathscr{L}, \mathscr{L}^*$ can be affected by too small τ if the multiscale process (X_t, Y_t) is only partially observed. However, if τ grows this error (or bias) vanishes, without trading it for finite τ errors.

5.3. Partially observed diffusions. If one observes only the slow variable(s) X_t of the full multiscale system (5.1) and not the fast ones Y_t, clearly it is not possible to estimate eigenfunctions of $\mathscr{L}$ or $\mathscr{L}^*$ that are dependent on both x and y. However, the leading eigenfunctions can be constructed, to leading order in ϵ, from X_t data only, because of their structures as given in (5.12) and (5.13).

Having only X_t data available, one can only use Galerkin basis functions that depend on x but not on y. With $f_i = f_i(x)$ for all i, the inner products in (3.3b) become

$$\langle P_\tau f_i, f_j \rangle_\rho = \int_{\Omega_x} dx\, \tilde{\rho}(x) f_j(x) (\Pi\, P_\tau f_i)(x) = \langle \Pi\, P_\tau f_i, f_j \rangle_{\tilde{\rho}} \tag{5.14}$$

where $\tilde{\rho}(x)$ is the marginal invariant density for X_t:

$$\tilde{\rho}(x) = \int_{\Omega_y} \rho(x,y) \quad \Leftrightarrow \quad \rho(x,y) = \tilde{\rho}(x)\rho_x(y)\,, \tag{5.15}$$

and Π is the projection operator defined in (5.8). Thus, one effectively observes the operator $\Pi\, P_\tau$ instead of P_τ. However, we show below that under appropriate conditions (notably, $\tau \gg \epsilon^2$), the operator $\Pi\, P_\tau$ has eigenpairs that are $O(\epsilon)$ close to the leading eigenpairs of P_τ. As a consequence, the leading eigenpairs of $\mathscr{L}$ can be inferred, to $O(\epsilon)$ accuracy, without observing Y_t. Together with (5.12) this implies we can infer the leading eigenpairs of $\mathscr{L}^H$ to $O(\epsilon)$ accuracy from X_t data only.

THEOREM 1. *Let (Λ_k, ϕ_k) be a leading eigenpair of P_τ (i.e. $k \in K_0$, see appendix*

B), and let the following conditions hold:

$$\begin{aligned} &\textit{(i)} \quad \tau \gg \epsilon^2 && (5.16a) \\ &\textit{(ii)} \quad |\Lambda_k - \Lambda_{k'}| \gg \epsilon \ \textit{ for all } \ k' \neq k && (5.16b) \end{aligned}$$

Then

$$\exists \ (\Lambda_k^g, \phi_k^g) \ \ \textit{such that} \begin{cases} \textit{(a)} \ \Pi \, P_\tau \phi_k^g = \Lambda_k^g \phi_k^g \\ \textit{(b)} \ \phi_k - \phi_k^g = O(\epsilon) \\ \textit{(c)} \ \Lambda_k - \Lambda_k^g = O(\epsilon) \end{cases} \tag{5.17}$$

Thus, under conditions (i) and (ii), the operator $\Pi \, P_\tau$ has an eigenpair (Λ_k^g, ϕ_k^g) that is $O(\epsilon)$ close to the eigenpair (Λ_k, ϕ_k), $k \in K_0$, of P_τ. Condition (ii) ensures that the eigenvalue Λ_k has multiplicity 1 and is well separated from all other eigenvalues. To prove theorem 1, we will use that for $k \in K_0$,

$$\begin{aligned} \phi_k - \Pi \, \phi_k &= O(\epsilon) \, , && (5.18a) \\ \psi_k - \rho_x \Pi^* \psi_k &= O(\epsilon) \ , && (5.18b) \end{aligned}$$

resulting from (5.12) and (5.13). The projection operator Π^* is defined in (B.15). We will also need the following lemma:

LEMMA 2. *If $\tau \gg \epsilon^2$, then $P_\tau h - P_\tau \Pi \, h = O(\epsilon)$ for any $h(x,y) \in \mathscr{F}$.*

Proof of lemma 2. We split $h - \Pi \, h = h^0 + h^\perp$, where h^0 lies in the subspace spanned by the ϕ_k with $k \in K_0$ and $h^\perp$ lies in the subspace spanned by all other eigenfunctions. Because of (5.18) we have $\langle \psi_k, h - \Pi \, h \rangle_1 = O(\epsilon)$ if $k \in K_0$, and therefore $h^0 = O(\epsilon)$. The spectral radius of P_τ being 1, this implies $P_\tau h^0 = O(\epsilon)$. Furthermore, $\|P_\tau h^\perp\| \leq |\Lambda_l| \, \|h^\perp\|$ where $\Lambda_l = \exp(\tau \lambda_l)$ is the largest eigenvalue with $l \notin K_0$. Because $\lambda_l = O(\epsilon^{-2})$ if $l \notin K_0$, setting $\tau = \epsilon^q$ with $q < 2$ gives $|\Lambda_l| = \exp(-c\epsilon^{q-2})$ with $c > 0$ a real constant of order 1 in ϵ. Thus, as $\epsilon \to 0$, $|\Lambda_l|$ approaches zero at a rate that is exponential in ϵ if $q < 2$. □

Proof of theorem 1. Because $P_\tau \phi_k = \Lambda_k \phi_k$, we have $(P_\tau \Pi + P_\tau(1 - \Pi))\phi_k = \Lambda_k \phi_k$. By lemma 2, $\|P_\tau(1 - \Pi)\|$ is $O(\epsilon)$ if condition (i) is satisfied. Using (ii), it then follows from operator perturbation theory that $P_\tau \Pi$ has an eigenpair $(\Lambda_k^g, \tilde{\phi}_k^g)$ that is $O(\epsilon)$ close to the eigenpair (Λ_k, ϕ_k) of $P_\tau \Pi + P_\tau(1 - \Pi)$. Thus, $\Lambda_k^g - \Lambda_k = O(\epsilon)$ and $\tilde{\phi}_k^g - \phi_k = O(\epsilon)$. Furthermore, because $P_\tau \, \Pi \, \tilde{\phi}_k^g = \Lambda_k^g \, \tilde{\phi}_k^g$ we have $\Pi \, P_\tau \, \Pi \, \tilde{\phi}_k^g = \Lambda_k^g \, \Pi \, \tilde{\phi}_k^g$, therefore (Λ_k^g, ϕ_k^g) with $\phi_k^g = \Pi \, \tilde{\phi}_k^g$ is an eigenpair of $\Pi \, P_\tau$. Finally, ϕ_k, $\tilde{\phi}_k^g$ and $\Pi \, \phi_k$ are all $O(\epsilon)$ close to each other, see also (5.18), so that $\phi_k - \phi_k^g = O(\epsilon)$. □

For the adjoint operator P_τ^* a similar result holds, as is shown below. We note that for the adjoint of $\Pi \, P_\tau$ we have $\langle v, \Pi \, P_\tau h \rangle_1 = \langle \Pi^* P_\tau^* \rho_x v, h \rangle_1$ for appropriate functions $v(x)$ and $h(x)$. Hence, $(\Pi \, P_\tau)^* v = \Pi^* P_\tau^* \rho_x v$.

THEOREM 3. *Let $(\bar{\Lambda}_k, \psi_k)$ be a leading eigenpair of P_τ^* $(k \in K_0)$, and let condi-*

tions (i) and (ii) from theorem 1 hold. Then

$$\exists\ (\bar\Lambda^g_k, u^g_k) \ \ \textit{such that} \begin{cases} \textit{(a)}\ \ (\Pi\, P_\tau)^* u^g_k = \bar\Lambda^g_k u^g_k \\ \textit{(b)}\ \ u_k - u^g_k = O(\epsilon) \\ \textit{(c)}\ \ \Lambda_k - \Lambda^g_k = O(\epsilon) \end{cases} \tag{5.19}$$

where u_k is defined in (5.13).

Thus, under conditions (i) and (ii), the adjoint operator $(\Pi\, P_\tau)^*$ has an eigenpair $(\bar\Lambda^g_k, u^g_k)$ that is $O(\epsilon)$ close to the Π^* projection of the eigenpair $(\bar\Lambda_k, \psi_k)$ of P^*_τ (in the sense that u^g_k is $O(\epsilon)$ close to $\Pi^*\psi_k$).

Proof of theorem 3. Consider two functions $h \in \mathrm{dom}(P_\tau)$, $r \in \mathrm{dom}(P^*_\tau)$. We have $\langle r, P_\tau(1-\Pi)h\rangle_1 = \langle (1-\rho_x\Pi^*)P^*_\tau r, h\rangle_1$. Thus, $(1-\rho_x\Pi^*)P^*_\tau$ is $O(\epsilon)$ if condition (i) is satisfied, by lemma 2 and the fact that h and r are arbitrary. We rewrite $P^*_\tau\psi_k = (\rho_x\Pi^*P^*_\tau + (1-\rho_x\Pi^*)P^*_\tau)\psi_k = \bar\Lambda_k\psi_k$. Because, as noted, $(1-\rho_x\Pi^*)P^*_\tau$ is a small perturbation (under (i)), $\rho_x\Pi^*P^*_\tau$ has an eigenpair $(\bar\Lambda^g_k, \psi^g_k)$ that is $O(\epsilon)$ close to $(\bar\Lambda_k, \psi_k)$. Furthermore, we note that if $(\Pi\, P_\tau)^* u^g_k = \bar\Lambda^g_k u^g_k$ then $\rho_x\Pi^*P^*_\tau\rho_x u^g_k = \bar\Lambda^g_k\rho_x u^g_k$, i.e. $(\bar\Lambda^g_k, \rho_x u^g_k)$ is an eigenpair of $\rho_x\Pi^*P^*_\tau$. Hence, we identify $\psi^g_k = \rho_x u^g_k$. □

Theorems 1 and 3 show that if Λ_k, $k \in K_0$, is well separated from other eigenvalues and $\tau \gg \epsilon^2$, then P_τ and $\Pi\, P_\tau$ (and their adjoints) have eigenpairs that are $O(\epsilon)$ close. However, an additional constraint on τ is needed to ensure that $\lambda_k = \tau^{-1}\log\Lambda_k$ and $\lambda^g_k = \tau^{-1}\log\Lambda^g_k$ are also close to each other. To see this, we write $\Lambda^g_k = \Lambda_k + \epsilon\,\delta\Lambda_k$, and note that Λ^g_k and Λ_k are $O(1)$ unless $\tau \gg 1$. By substituting $\tau = \epsilon^q$ and using Taylor expansion for the logarithm, we arrive at

$$\lambda^g_k = \lambda_k + \epsilon^{1-q}\frac{\delta\Lambda_k}{\Lambda_k} + O(\epsilon^{2-q})\,. \tag{5.20}$$

Thus, $\lambda^g_k \to \lambda_k$ as $\epsilon \to 0$, provided $0 \le q < 1$. Put differently:

$$\lambda^g_k = \lambda_k + O(\epsilon) \quad \text{if } \tau = O(1) \text{ in } \epsilon \tag{5.21}$$

If τ is too small, the eigenpairs of P_τ and $\Pi\, P_\tau$ are no longer $O(\epsilon)$ close to each other. To see what happens if τ is very small, consider the expansion $P_\tau = \exp(\tau\mathscr{L}) = 1 + \tau\mathscr{L} + \frac{1}{2}\tau^2\mathscr{L}^2 + ...$ Substitution of (5.5) and $\tau = \epsilon^q$, $q > 3$, gives

$$\langle P_\tau f_i, f_j\rangle_\rho = \langle f_i, f_j\rangle_\rho + \epsilon^q\langle \mathscr{L}_0 f_i, f_j\rangle_\rho + O(\epsilon^{2q-3})\,. \tag{5.22}$$

To see this, note that $\mathscr{L}_2 f_i = 0$ because f_i depends only on x, and $\langle\mathscr{L}_1 f_i, f_j\rangle_\rho = 0$ due to the centering condition (5.2). Furthermore, $\langle\mathscr{L}_0 f_i, f_j\rangle_\rho = \langle\Pi\,\mathscr{L}_0 f_i, f_j\rangle_{\bar\rho}$ if $f_i = f_i(x)$ and $f_j = f_j(x)$. Thus, if $\tau \ll \epsilon^3$ the solutions to the generalized eigenvalue problem (3.4) are approximations of the eigenpairs of $\exp(\tau\,\Pi\,\mathscr{L}_0)$ rather than $\exp(\tau\mathscr{L})$. We remark that $\Pi\,\mathscr{L}_0$ corresponds to the diffusion operator that would result from averaging (rather than homogenizing) the multiscale system (5.1). We refer to [33] for more details about averaging.

The analysis in this section makes clear that partial observation may necessitate subsampling. If only X_t, generated by (5.1), is observed and not Y_t, a τ that is too small results in eigenpair estimates that are not close to the eigenpairs of $\mathscr{L}^H$ and

$\mathscr{L}^{H*}$. The estimation procedure discussed in section 4 will then give biased estimates of the homogenized drift and diffusion. In such case, subsampling (skipping data points in order to increase τ) is needed to arrive at correct estimates. Our analysis points out that the sampling interval should scale as $\tau = \epsilon^q$ with $0 \leq q < 1$ for estimating the correct homogenized diffusion process from X_t data of (5.1).

5.4. Numerical example: noise-driven motion in a multiscale potential. We consider a system studied previously in [32, 31], consisting of noise-driven motion in a potential with two spatial scales. The SDE of this system reads

$$dX_t = -V'(X_t)dt + \sqrt{2\sigma}\, dW_t \tag{5.23}$$

where, as usual, W_t is a Wiener process and $V'(x) = dV/dx$. The diffusion coefficient σ is constant (additive noise). The potential consists of two parts:

$$V(x) = \alpha V_0(x) + p(\tfrac{x}{\epsilon}) \qquad \text{with} \quad V_0(x) = \tfrac{1}{2}x^2 \quad \text{and} \quad p(y) = \cos(y) \tag{5.24}$$

Strictly speaking, $\mathscr{L}$ associated with (5.23) is slightly different from (5.6) because W_t^x and W_t^y are correlated, see [32, 31]. However, as this does not change the asymptotic results in section 5.1, we will not discuss this further.

The effective SDE of the homogenized system is

$$dX_t = -V_h'(X_t)dt + \sqrt{2\Sigma}\, dW_t \tag{5.25}$$

with

$$V_h(x) = K_h \alpha V_0(x) \qquad \text{and} \qquad \Sigma = K_h \sigma \tag{5.26}$$

The constant K_h is determined by the small-scale part of the potential $p(y)$ and its period L:

$$K_h = \frac{L^2}{Z\hat{Z}}, \qquad Z = \int_0^L dy \mathrm{e}^{-p(y)/\sigma}, \quad \hat{Z} = \int_0^L dy \mathrm{e}^{p(y)/\sigma} \tag{5.27}$$

With $p(y) = \cos(y)$ and thus $L = 2\pi$ we find $K_h = 0.1924$. Following [32], we set $\epsilon = 0.1$, $\alpha = 1$ and $\sigma = 1/2$. The homogenized SDE then reads

$$dX_t = -K_h X_t dt + \sqrt{K_h}\, dW_t \tag{5.28}$$

We fit drift and diffusion functions

$$b(x) = b_1 x\,, \qquad a(x) = a_1 \tag{5.29}$$

to X_t data generated by the multiscale process (5.23). We use 100 sample paths of the process, each with a total time length $T = 4 \times 10^4$, obtained by numerical integration with time step 10^{-4}. We sample them at various intervals. The parameters b_1, a_1 are estimated using the objective functions E^g, E^b and E^m.

Because the homogenized process (5.28) is an Ornstein-Uhlenbeck process, the leading eigenfunctions of $\mathscr{L}^H$ are Hermite polynomials and the subspace $\mathscr{F}_M = \operatorname{span}\{1, x, x^2\}$ captures the leading 3 eigenfunctions of $\mathscr{L}^H$. Hence, we use Galerkin basis functions $f_1 = 1, f_2 = x, f_3 = x^2$ for E^g and E^m. Furthermore, for E^b we use 200 bins, $K = 2$ and test functions $\sigma_1 = x, \sigma_2 = x^2$. For E^m we use the same test functions. The weights in E^g are such that $c_2 = 0.5$. In figure 8 we plot the

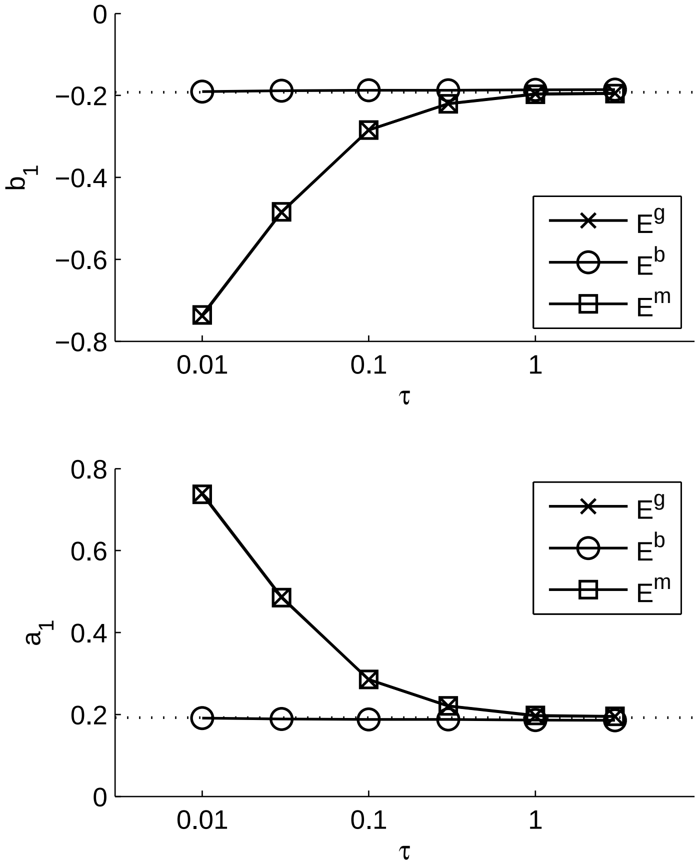

FIG. 8. *Example with multiscale potential. The drift and diffusion functions (5.29) are fitted to X_t data of the multiscale process (5.23) with sampling interval τ. The parameters b_1 and a_1 are estimated using E^b (200 bins), E^g and E^m ($\mathscr{F}_M = \{1, x, x^2\}$ for both). The dotted lines indicate the values predicted by homogenization theory.*

mean values of the estimated parameters, for different values of the sampling interval ($\tau = 0.01, 0.03, 0.1, 0.3, 1, 3$). The standard deviations of the estimates are not shown; they are small (ranging from 0.001 to 0.01).

It can be seen that the estimates obtained with E^b are consistent with the values predicted by homogenization theory ($-b_1 = a_1 = K_h$), for all values of τ. The estimates from E^g and E^m are only consistent with homogenization theory if τ is large enough. These results are in agreement with the discussion in section 5.3. Because of the large number of bins used in the binning method, we resolve the small-scale features of the eigenfunctions, induced by the small-scale part of the potential. Thus, with the binning method we estimate the (approximate) eigenpairs of P_τ itself. With the smooth Galerkin method with $\mathscr{F}_M = \text{span}\{1, x, x^2\}$, these small-scale features are not resolved at all. We effectively observe only the large-scale part of the process, therefore we get estimates of eigenpairs of ΠP_τ instead of P_τ. As was discussed in section 5.3, the eigenpairs of P_τ and ΠP_τ can differ significantly in case of small τ. This affects the results from both E^g and E^m. In the limit $\tau \to 0$, the estimates from E^g and E^m approach $b_1 = -1$, $a_1 = 1$, consistent with ΠP_τ approximating

$\exp(\tau\,\Pi\,\mathscr{L}_0)$ if $\tau \ll \epsilon^3$ (as discussed in section 5.3). The quadratic variation estimator used in [32] also overestimates a_1 for small τ (and approaches 1 as $\tau \to 0$).

5.5. Numerical example: One OU process driving another. In this example we consider the case where a (slow) variable X_t is forced by a fast stochastic variable Y_t:

$$dX_t = F_0(X_t)dt + \frac{\gamma}{\epsilon}Y_t dt \tag{5.30a}$$

$$dY_t = -\frac{\beta}{\epsilon^2}Y_t dt + \frac{\sigma}{\epsilon}\,dW_t\,, \tag{5.30b}$$

where γ, β, σ are all real $O(1)$ constants ($\beta > 0$) and $\epsilon \ll 1$. As can be seen, Y_t is an OU process. Comparing to equation (5.1) we see that we have $F_1(X_t, Y_t) = \gamma Y_t$ and $F_0(X_t, Y_t) = F_0(X_t)$. Because Y_t has mean zero, condition (5.2) is satisfied. Hence, there is an homogenized equation for X_t [33], reading

$$d\bar{X}_t = F_0(\bar{X}_t)dt + s\,dW_t \tag{5.31}$$

with $F_0(x)$ as in (5.30a) and the constant s given by

$$s^2 = 2\gamma^2 \int_0^\infty d\tau\, \mathbb{E}\, Y_t^o Y_{t+\tau}^o \tag{5.32}$$

Here, Y_t^o is the solution to (5.30b) with $\epsilon = 1$. The expectation in (5.32) is with respect to the law of Y_t^o, so s^2 is proportional to the integrated autocorrelation function of Y_t^o. Because Y_t^o is an OU process we can calculate this function exactly:

$$\mathbb{E}\, Y_t^o Y_{t+\tau}^o = \frac{\sigma^2}{2\beta}\exp(-\beta\tau) \qquad \text{and} \qquad s^2 = \frac{\gamma^2\sigma^2}{\beta^2} \tag{5.33}$$

In what follows, we set $\epsilon = 0.1$ and $\beta = \sigma = \gamma = 1$, so that $s = 1$. For F_0 we choose a simple form, $F_0(x) = -x$. Thus, heuristically speaking, X_t is an OU process forced by ”red noise” (the fast OU process Y_t) instead of the usual ”white noise” (the Wiener process). In the homogenized equation, the red noise gets replaced by white noise of an appropriate amplitude (determined by s). Because of the form of (5.30a), Stratonovich corrections do not play a role in going from (5.30) to (5.31).

Similar to the previous example, we fit drift and diffusion functions

$$b(x) = b_1 x\,, \qquad a(x) = a_1 \tag{5.34}$$

to timeseries of X_t generated by the multiscale process (5.30). We use 100 different sample paths of the process, each of total time length $T = 10^5$, obtained by numerical integration with time step 10^{-4}. We vary the sampling interval τ from 0.01 (yielding 10^7 data points) to 3 (3.3×10^4 data points).

The parameters b_1 and a_1 are estimated using E^g, E^b and E^m. Settings for these objective functions (Galerkin basis functions, etc.) are the same as in the previous example. For comparison, we also estimate a_1 from the quadratic variation of the path,

$$a_1^{\mathrm{qv}} = \frac{1}{h\tau\, N_h}\sum_{i=1}^{N_h}(X_{ih\tau} - X_{(i-1)h\tau})^2 \tag{5.35}$$

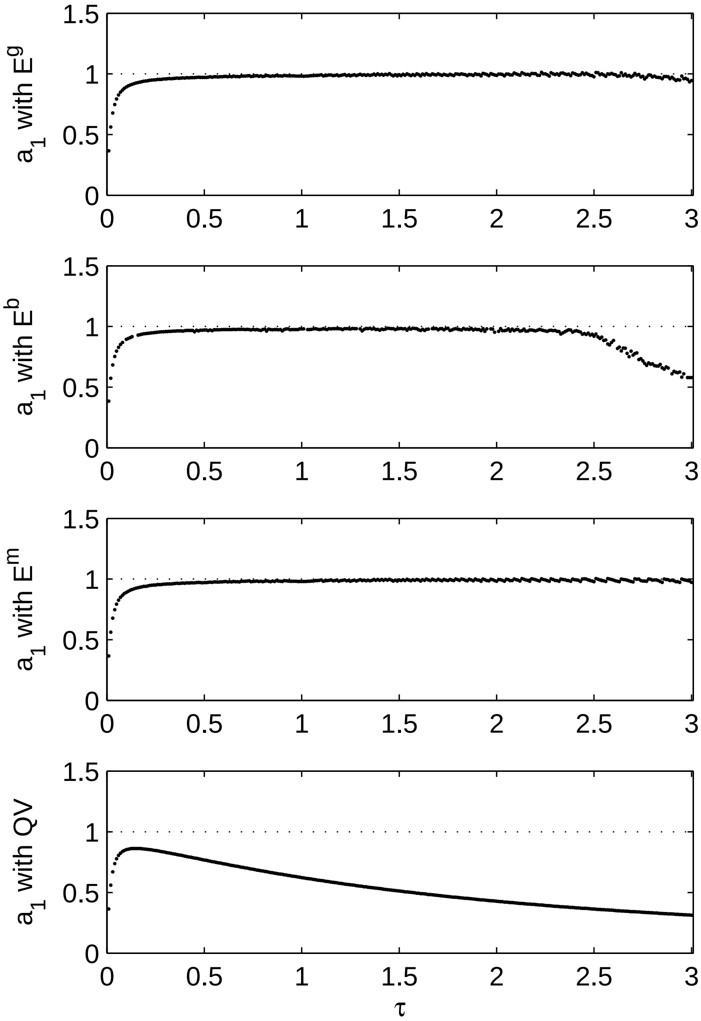

Fig. 9. *Example with a fast OU process driving a slow one. The drift and diffusion functions (5.34) are fitted to X_t data of the multiscale process (5.30) with sampling interval τ. Shown are the mean values of the a_1 estimates obtained from 100 different sample paths of total time length $T = 10^5$. From top to bottom: results using E^g, E^b, E^m and quadratic variation (QV), see (5.35). Homogenization theory predicts $a_1 = 1$, indicated by the dotted lines.*

with $N_h = \lfloor N/h \rfloor$. As discussed before, homogenization theory predicts $b_1 = -1$ and $a_1 = 1$.

In figure 9 the mean values for the estimates of a_1 are plotted. Away from the small τ limit, the estimates obtained with E^g, E^b, E^m are consistent with homogenization theory. For very long sampling intervals, these estimates are affected by sampling error, visible in the decrease of the E^b estimates as $\tau > 2.5$ (with E^g and E^m this occurs at even longer τ, beyond the range of the figure). The quadratic variation underestimates a_1 at every τ; it peaks at $a_1^{\mathrm{qv}} \approx 0.85$ around $\tau = 0.15$. This underestimation is not due to sampling error. There is no plot of b_1 estimates; for all τ we find that b_1 is near $-a_1$, deviations are largest for large τ.

All estimates tend to zero in the limit of small τ, due to the fact that the process (5.30) is only partially observed. As is discussed in section 5.3, for very small τ the operator $\Pi\, P_\tau$ approaches $\exp(\tau\, \Pi\, \mathscr{L}_0)$. The multiscale process (5.30) in this example is a hypoelliptic diffusion, i.e. $\alpha = 0$ in (5.1) and (5.6). Therefore $\Pi\, \mathscr{L}_0 = (\Pi\, F_0) \cdot \nabla_x$ (in fact, in this example $\Pi\, F_0 = F_0$ because $F_0 = F_0(x)$). As a consequence, the matrix Q^x defined as

$$Q^x_{ij} = \langle \Pi\, \mathscr{L}_0 f_i, f_j \rangle_{\bar{\rho}} = \langle \mathscr{L}_0 f_i, f_j \rangle_\rho \tag{5.36}$$

can be shown to be antisymmetric: because $\mathscr{L}^* \rho = 0$ we have

$$0 = \langle \mathscr{L} f_i f_j, \rho \rangle_1 = \langle f_i \mathscr{L}_0 f_j + f_j \mathscr{L}_0 f_i, \rho \rangle_1 \,, \tag{5.37}$$

where we use that $\mathscr{L}_2 f_i = 0$ because $f_i = f_i(x)$, and $\langle \mathscr{L}_1 f_i, f_j \rangle_\rho = 0$ because of (5.2). Q^x being antisymmetric, its eigenvalues must have zero real part.

As mentioned, $\mathscr{L}^* \rho = 0$ so that $\langle \mathscr{L} g, \rho \rangle_1 = 0$ for any function $g \in \mathrm{dom}(\mathscr{L})$. Assume that for all $i, j \leq M$ there exist functions $g_{ij}(x) \in \mathrm{dom}(\mathscr{L})$ such that $f_i \nabla_x f_j = \nabla_x g_{ij}$. Then $f_i(\mathscr{L}_0 + \epsilon^{-1} \mathscr{L}_1) f_j = (\mathscr{L}_0 + \epsilon^{-1} \mathscr{L}_1) g_{ij} = \mathscr{L} g_{ij}$ and the elements of Q^x satisfy

$$Q^x_{ij} = \langle f_j \mathscr{L}_0 f_i, \rho \rangle_1 = \langle \mathscr{L} g_{ij}, \rho \rangle_1 = 0 \tag{5.38}$$

Hence, $Q^x = 0$ and therefore all its eigenvalues are zero. If x is one-dimensional, as in the current example, the functions $g_{ij}(x)$ always exist: they are the antiderivatives of $f_i \nabla_x f_j$.

If the eigenvalues of Q^x are zero, it means that in the limit of small τ, we are fitting a diffusion process to eigenpairs whose eigenvalues approach zero. As a consequence, the fitted drift and diffusion approach zero too. It explains why $b_1 \to 0$ and $a_1 \to 0$ as $\tau \to 0$, see figure 9.

6. Conclusion.

In this paper we considered estimation of diffusion processes from discrete-time data. The paper consists of two parts. In the first part (sections 3 and 4) we presented a new estimation method, applicable for a broad class of diffusion processes (scalar as well as multivariate, reversible and nonreversible, with nonlinear drifts and/or multiplicative noises). In the second part of the paper (section 5) we discussed estimation of coarse-grained (homogenized) diffusion processes from multiscale data and we investigated the performance of the method presented in sections 3 and 4 in this context.

The estimation method presented in sections 3 and 4 relies on the close relation between eigenpairs of the diffusion operator $\mathscr{L}$ and those of the conditional expectation operator P_t, see (1.6a). This relation is a consequence of the semigroup structure $P_t = \exp(t\mathscr{L})$ for $t \geq 0$. A similar relation holds for the adjoint operators, see (1.6b). Hence, eigenpairs of $\mathscr{L}$ and $\mathscr{L}^*$ can be inferred by estimating eigenpairs of P_t and P_t^*: the eigenfunctions are identical, and the eigenvalues are related as in (1.7).

In section 3, we showed how to estimate eigenpairs of P_t and P_t^* by means of the Galerkin method. Both smooth and discontinuous approaches were discussed. The next step, inferring the drift $b(x)$ and diffusion $a(x)$ that determine $\mathscr{L}$ from eigenpairs of $\mathscr{L}$ and/or $\mathscr{L}^*$, was considered in section 4. We presented a new method to infer b and a from eigenpairs, in which residuals $(\mathscr{L}^* - \hat{\hat{\lambda}}_k)\hat{\psi}_k$ are minimized via minimization of an objective function. We integrate the residuals against smooth test functions and build an objective function from the squared integrals. This allows us to infer b and a without estimating eigenfunction derivatives, thereby avoiding a major source of error. If b and a are linear in their parameters, as in (4.13), the objective function is convex quadratic and has a unique minimum. The total computational cost of estimating eigenpairs and inferring b and a is small (e.g., in the 2-dimensional example in section 4.4.3, estimating the parameters from $N = 10^5$ datapoints takes us around 5 seconds) .

In several numerical examples, the performance of the newly presented method was investigated, demonstrating the overall feasibility of this method and its good results in some highly nontrivial examples. One of the examples (section 4.4.2) involved a (mildly) metastable system, where the process switches between the wells of a double-well potential. The long-timescale dynamics of a metastable system (hopping between metastable states) is captured by the leading eigenmodes [37, 24], so that our spectral procedure, in which these eigenmodes play a central role, is in a good position to estimate such a system. Indeed, the spectral method was well capable of estimating the system in section 4.4.2. It is important that the available timeseries data contain enough switches between metastable states, in order to obtain good estimates of the leading eigenvalues. We note that the leading eigenvalues can be estimated correctly even if the sampling intervals of the data are too long to obtain good estimates of the non-leading eigenvalues (representing the fast dynamics of the metastable system). This was analysed in detail in [15].

In this paper we focussed on estimation from data with constant sampling intervals. However, our estimation procedure can be generalized to deal with data with nonconstant sampling intervals. In [15], it was shown how the spectral estimation procedure can be used to estimate generators for Markov jump processes from data with nonconstant τ. We expect that a similar generalization to nonconstant τ can be formulated for diffusion estimation.

We note that there are some limitations to the estimation procedure as presented here. It is assumed throughout that the diffusion process to be estimated has an invariant measure. Thus, for processes such as pure diffusion on $\mathbb{R}^1$ ($dX_t = \sigma dW_t$) or geometric Brownian motion, which have no invariant measure, the procedure does not apply. However, we are currently investigating how to adapt the spectral procedure to deal with such processes; we will report on this work elsewhere. A second limitation is that the estimation procedure cannot be used to estimate parameters in equations for unobserved variables. An example is the Heston model for option prices (or other models for stochastic volatility). One can observe the option price but not the volatility, so our procedure does not enable estimation of parameters that appear in the volatility SDE.

Because the relations (1.6a), (1.6b) and (1.7) are exact, the use of eigenpairs makes it possible to estimate b and a from discrete-time samplings without making time discretization errors. This makes the eigenpair approach particularly attractive in case of data with long sampling intervals (low-frequency data). It is also advantageous when fitting a coarse-grained diffusion process to multiscale data. As was shown in [32], a too short sampling interval can lead to biased estimates for the coarse-grained process. An estimation method that allows to use longer sampling intervals without introducing time discretization errors is obviously attractive in this situation.

Estimation of homogenized diffusion processes from data of multiscale diffusions was investigated in detail in section 5. We showed that the leading eigenpairs of the homogenized diffusion operator $\mathscr{L}^H$ and those of the underlying multiscale operator $\mathscr{L}$ are the same at leading order in ϵ (where ϵ measures the scale separation, $\epsilon \ll 1$), see section 5.1. Moreover, those eigenpairs can be estimated from data of the slow variables alone, provided the sampling interval is long enough ($\tau = \epsilon^q$ with $0 \leq q < 1$), as was discussed in section 5.3. The necessity to subsample (or more precisely, the necessity to avoid very short sampling intervals) is a consequence of partial observation (only the slow variables are observed, not the fast ones). The analysis of sections 5.1-5.3 was illustrated with two numerical examples in sections 5.4 and 5.5. Both showed that the estimation method presented in this paper is well suited to infer correct homogenized diffusion processes from multiscale data.

Acknowledgments. Comments from anonymous referees helped to improve the presentation of the paper and are gratefully acknowledged. Research by D. Crommelin was supported by the Netherlands Organization for Scientific Research (NWO) through the research cluster Nonlinear Dynamics of Natural Systems. Research by E. Vanden-Eijnden was supported by NSF grants DMS02-39625 and DMS07-08140, and by ONR grant N00014-04-1-0565.

Appendix A. Expressions for E^g, E^b and E^m.

We assume b and a have expansions as in (4.13). Then (4.4) becomes

$$E^g(b_1,...,b_{N_b},a_1,...,a_{N_a}) = \Big\| \sum_j b_j \hat{B}^g_j + \sum_j a_j \hat{A}^g_j - \hat{D}_\lambda \Big\|^2_c \tag{A.1}$$

with

$$\hat{B}^g_j = \hat{V}\hat{B}'_j\hat{W}^* \,, \qquad \hat{A}^g_j = \hat{V}\hat{A}'_j\hat{W}^* \,. \tag{A.2}$$

The elements of the matrices $\hat{B}'_j$ and $\hat{A}'_j$ are

$$\hat{B}'_{jmn} = \frac{1}{N+1}\sum_{i=0}^{N} \Big[f_n(X_{i\tau})\big(g_j\cdot(\nabla f_m)\big)(X_{i\tau}) \Big] \tag{A.3a}$$

$$\hat{A}'_{jmn} = \frac{1}{N+1}\sum_{i=0}^{N} \Big[f_n(X_{i\tau})\big(\tfrac{1}{2}h_j : (\nabla\nabla f_m)\big)(X_{i\tau}) \Big] \tag{A.3b}$$

For the binning approach, the expansions (4.13) lead to

$$E^b(b_1,...,b_{N_b},a_1,...,a_{N_a}) = \sum_{k,n}\alpha_{kn}\Big| \big(\sum_j b_j \hat{B}^b_j + \sum_j a_j \hat{A}^b_j - \hat{C}^b \big)_{kn} \Big|^2 \tag{A.4}$$

with

$$\hat{B}^b_{jkn} = \frac{1}{N+1}\sum_{i=0}^{N} (\hat{\xi}_k g_j \cdot \nabla\sigma_n)(X_{i\tau}) \tag{A.5a}$$

$$\hat{A}^b_{jkn} = \frac{1}{N+1}\sum_{i=0}^{N} (\tfrac{1}{2}\hat{\xi}_k h_j : \nabla\nabla\sigma_n)(X_{i\tau}) \tag{A.5b}$$

$$\hat{C}^b_{kn} = \hat{\bar{\lambda}}_k \frac{1}{N+1}\sum_{i=0}^{N} (\hat{\xi}_k \sigma_n)(X_{i\tau}) \tag{A.5c}$$

For the "mixed" approach, finally, (4.13) results in

$$E^m(b_1,...,b_{N_b},a_1,...,a_{N_a}) = \sum_{k,n}\alpha_{kn}\Big| \big(\sum_j b_j \hat{B}^m_j + \sum_j a_j \hat{A}^m_j - \hat{C}^m \big)_{kn} \Big|^2 \tag{A.6}$$

with

$$\hat{B}^m_j = (\hat{B}'_j\hat{W}^*)^* \,, \qquad \hat{A}^m_j = (\hat{A}'_j\hat{W}^*)^* \,, \qquad \hat{C}^m = \hat{D}_{\bar{\lambda}}\hat{W}\hat{R} \tag{A.7}$$

and B'_j and A'_j as in (A.3).

We note that the matrices $\hat{B}^g_j$, $\hat{A}^g_j$ and $\hat{D}_\lambda$ need to be evaluated only once, at the beginning of the minimization of E^g (and not at every step of the minimization algorithm). The same holds for $\hat{B}^b_j$, $\hat{A}^b_j$, $\hat{C}^b$ in the minimization of E^b and for $\hat{B}^m_j$, $\hat{A}^m_j$, $\hat{C}^m$ in the minimization of E^m.

Appendix B. Asymptotics of $\mathscr{L}$ and $\mathscr{L}^*$.

B.1. Eigenpairs of $\mathscr{L}$. We consider the asymptotics of the multiscale diffusion operator (5.5), (5.6). Let $\mathscr{L}_2^*$ be the adjoint of $\mathscr{L}_2$ in $L_2(\Omega_y, dy)$. The assumption that Y_t is ergodic with unique invariant measure if X_t is fixed implies that the null spaces of both $\mathscr{L}_2$ and $\mathscr{L}_2^*$ are one-dimensional. Assuming that μ_x has a density, $d\mu_x(y) = \rho_x(y)dy$, we have

$$\mathscr{L}_2 1(y) = 0, \qquad \mathscr{L}_2^* \rho_x(y) = 0 \tag{B.1}$$

where $1(y) = 1 \,\forall\, y \in \Omega_y$.

Substituting the expansions (5.10) in (5.5), (5.6), we obtain for the eigenpairs (ϕ_k, λ_k) of $\mathscr{L}$ a sequence of problems:

$$\begin{aligned}
O(\epsilon^{-2}) &\qquad \mathscr{L}_2\phi_k^{(0)} = \lambda_k^{(-2)}\phi_k^{(0)} && \text{(B.2a)}\\
O(\epsilon^{-1}) &\qquad \mathscr{L}_1\phi_k^{(0)} + \mathscr{L}_2\phi_k^{(1)} = \lambda_k^{(-1)}\phi_k^{(0)} + \lambda_k^{(-2)}\phi_k^{(1)} && \text{(B.2b)}\\
O(1) &\qquad \mathscr{L}_2\phi_k^{(2)} + \mathscr{L}_1\phi_k^{(1)} + \mathscr{L}_0\phi_k^{(0)} = \lambda_k^{(-2)}\phi_k^{(0)} + \lambda_k^{(-1)}\phi_k^{(1)} + \lambda_k^{(0)}\phi_k^{(0)} && \text{(B.2c)}\\
&\ \vdots
\end{aligned}$$

The equation at leading order is itself an eigenvalue equation. Solutions with nonzero $\lambda_k^{(-2)}$ give the leading order terms for eigenpairs (ϕ_k, λ_k) with eigenvalues of order $O(\epsilon^{-2})$. More interesting to us are solutions of (B.2a) with $\lambda_k^{(-2)} = 0$. The corresponding $\phi_k^{(0)}$ lies in the null space of $\mathscr{L}_2$, hence it can be a function of x but must be constant in y.

We define K_0 to be the set of all indices k for which $\lambda_k^{(-2)} = 0$ and hence $\mathscr{L}_2\phi_k^{(0)} = 0$:

$$K_0 := \{k \mid \lambda_k^{(-2)} = 0\} \tag{B.3}$$

Furthermore, for all $k \in K_0$ we have $\Pi\,\phi_k^{(0)} = \phi_k^{(0)}$, with Π as defined in (5.8). Now we consider (B.2b) for $k \in K_0$:

$$\mathscr{L}_2\phi_k^{(1)} = \lambda_k^{(-1)}\phi_k^{(0)} - \mathscr{L}_1\phi_k^{(0)} \qquad (k \in K_0) \tag{B.4}$$

The solvability condition for this equation is

$$\Pi\,(\mathscr{L}_1\phi_k^{(0)} - \lambda_k^{(-1)}\phi_k^{(0)}) = 0 \tag{B.5}$$

Because $\phi_k^{(0)} = \phi_k^{(0)}(x)$ and because of the assumption (5.2), we have

$$\Pi\,\mathscr{L}_1\phi_k^{(0)} = 0 \tag{B.6}$$

and thus

$$\lambda_k^{(-1)} = 0 \qquad (k \in K_0) \tag{B.7}$$

Equation (B.2b) with $\lambda_k^{(-2)} = \lambda_k^{(-1)} = 0$ gives $\mathscr{L}_2\phi_k^{(1)} = -\mathscr{L}_1\phi_k^{(0)}$. Since $\Pi\,\mathscr{L}_1\phi_k^{(0)} = 0$, we may write

$$\phi_k^{(1)} = -\mathscr{L}_2^{-1}\mathscr{L}_1\phi_k^{(0)} \tag{B.8}$$

Finally, we go to equation (B.2c) and substitute $\lambda_k^{(-2)} = \lambda_k^{(-1)} = 0$ as well as (B.8):

$$\mathscr{L}_2\phi_k^{(2)} = -\mathscr{L}_0\phi_k^{(0)} + \mathscr{L}_1\mathscr{L}_2^{-1}\mathscr{L}_1\phi_k^{(0)} + \lambda_k^{(0)}\phi_k^{(0)} \tag{B.9}$$

The solvability condition for this equation gives

$$\mathscr{L}^H\phi_k^{(0)} = \lambda_k^{(0)}\phi_k^{(0)} \tag{B.10}$$

where $\mathscr{L}^H$ is the diffusion operator (5.7).

B.2. Eigenpairs of $\mathscr{L}^*$. Let $\psi_k(x,y)$ be an eigenfunction of the adjoint $\mathscr{L}^*$ of the multiscale diffusion operator (5.5), (5.6). Then

$$\mathscr{L}^*\psi_k = \left(\mathscr{L}_0^* + \frac{1}{\epsilon}\mathscr{L}_1^* + \frac{1}{\epsilon^2}\mathscr{L}_2^*\right)\psi_k = \bar\lambda_k\psi_k \tag{B.11}$$

with

$$\begin{aligned}
\mathscr{L}_0^*\psi_k &= \nabla_x\cdot(F_0\psi_k) + \tfrac{1}{2}\nabla_x\nabla_x : (\alpha\alpha^T\psi_k) && \text{(B.12a)}\\
\mathscr{L}_1^*\psi_k &= \nabla_x\cdot(F_1\psi_k) && \text{(B.12b)}\\
\mathscr{L}_2^*\psi_k &= \nabla_y\cdot(G\psi_k) + \tfrac{1}{2}\nabla_y\nabla_y : (\beta\beta^T\psi_k) && \text{(B.12c)}
\end{aligned}$$

Expanding $\psi_k = \psi_k^{(0)} + \epsilon\psi_k^{(1)} + \epsilon^2\psi_k^{(2)} + ...$ and λ_k as in (5.10), we obtain the sequence

$$\begin{aligned}
O(\epsilon^{-2}) &\quad \mathscr{L}_2^*\psi_k^{(0)} = \bar\lambda_k^{(-2)}\psi_k^{(0)} && \text{(B.13a)}\\
O(\epsilon^{-1}) &\quad \mathscr{L}_1^*\psi_k^{(0)} + \mathscr{L}_2^*\psi_k^{(1)} = \bar\lambda_k^{(-1)}\psi_k^{(0)} + \bar\lambda_k^{(-2)}\psi_k^{(1)} && \text{(B.13b)}\\
O(1) &\quad \mathscr{L}_2^*\psi_k^{(2)} + \mathscr{L}_1^*\psi_k^{(1)} + \mathscr{L}_0^*\psi_k^{(0)} = \bar\lambda_k^{(-2)}\psi_k^{(0)} + \bar\lambda_k^{(-1)}\psi_k^{(1)} + \bar\lambda_k^{(0)}\psi_k^{(0)} && \text{(B.13c)}\\
&\vdots
\end{aligned}$$

For all $k \in K_0$ we have $\bar\lambda_k^{(-2)} = 0$ and

$$\psi_k^{(0)}(x,y) = u_k(x)\rho_x(y) \qquad (k \in K_0) \tag{B.14}$$

Similar to the definition of Π (5.8), we define Π^* as

$$\Pi^*\, h(x,y) = \int_{\Omega_y} dy\, h(x,y) \tag{B.15}$$

For $k \in K_0$, the solvability condition for (B.13b) gives:

$$\Pi^*\,(\mathscr{L}_1^*\psi_k^{(0)} - \bar\lambda_k^{(-1)}\psi_k^{(0)}) = 0 \tag{B.16}$$

Assumption (5.2) and (B.14) imply

$$\Pi^*\,\mathscr{L}_1^*\psi_k^{(0)} = 0 \tag{B.17}$$

and we find, as before, that $\bar\lambda_k^{(-1)} = 0$ if $\bar\lambda_k^{(-2)} = 0$. Furthermore, we have

$$\psi_k^{(1)} = -(\mathscr{L}_2^*)^{-1}\mathscr{L}_1^*\psi_k^{(0)} \tag{B.18}$$

Finally, the solvability condition for (B.13c) gives us, for $k \in K_0$:

$$\Pi^* \, (\mathscr{L}_0^* - \mathscr{L}_1^* (\mathscr{L}_2^*)^{-1} \mathscr{L}_1^*) \psi_k^{(0)} = \bar{\lambda}_k^{(0)} \Pi^* \, \psi_k^{(0)} \tag{B.19}$$

Recalling (B.14) and the fact that $\Pi^* \, \psi_k^{(0)} = u_k$ we get the following eigenequation for $u_k(x)$:

$$\mathscr{L}^{H*} u_k = \bar{\lambda}_k^{(0)} u_k \tag{B.20}$$

where $\mathscr{L}^{H*}$ is defined as

$$\mathscr{L}^{H*} \cdot = \Pi^* \, (\mathscr{L}_0^* - \mathscr{L}_1^* (\mathscr{L}_2^*)^{-1} \mathscr{L}_1^*) \rho_x \cdot \tag{B.21}$$

It can be shown that $\mathscr{L}^{H*}$ is the adjoint of $\mathscr{L}^H$ (5.7) in $L_2(\Omega_x, dx)$. Also, the operator $(\mathscr{L}_0 - \mathscr{L}_1 \mathscr{L}_2^{-1} \mathscr{L}_1)^*$ is the adjoint of $\mathscr{L}_0 - \mathscr{L}_1 \mathscr{L}_2^{-1} \mathscr{L}_1$ in $L_2(\Omega_x \times \Omega_y, dx\, dy)$, and is equal to $\mathscr{L}_0^* - \mathscr{L}_1^* (\mathscr{L}_2^*)^{-1} \mathscr{L}_1^*$.

REFERENCES

[1] Aït-Sahalia, Y. (2002) Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, **70**, 223-262.

[2] Aït-Sahalia, Y., Mykland, P.A. and Zhang, L. (2005) How often to sample a continuous-time process in the presence of market microstructure noise. *Rev. Financ. Studies*, **18**, 351-416.

[3] Aït-Sahalia, Y. (2008) Closed-form likelihood expansions for multivariate diffusions. *Ann. Statist.*, **36**, 906-937.

[4] Aït-Sahalia, Y. and Hansen, L.P. (ed.) (2009) *Handbook of Financial Econometrics*. Amsterdam: Elsevier.

[5] Banks, H.T. and Kunisch, K. (1989) *Estimation techniques for distributed parameter systems*. Boston: Birkhäuser.

[6] Beattie, C. (2000) Galerkin eigenvector approximations. *Math. Comp.*, **69**, 1409-1434.

[7] Berner, J. (2005) Linking Nonlinearity and non-Gaussianity of Planetary Wave Behavior by the Fokker-Planck Equation. *J. Atmos. Sci.*, **62**, 2098-2117.

[8] Beskos, A., Papaspiliopoulos, O., Roberts, G.O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Statist. Soc. B*, **68**, 333-382.

[9] Best, R.B. and Hummer, G. (2010) Coordinate-dependent diffusion in protein folding. *Proc. Natl. Acad. Sci. USA*, **107**, 1088-1093.

[10] Bibby, B.M. and Sørensen, M. (1995) Martingale estimating functions for discretely observed diffusion processes. *Bernoulli*, **1**, 17-39.

[11] Bibby, B.M., Jacobsen, M. and Sørensen, M. (2009) Estimating functions for discretely sampled diffusion-type models, in *Handbook of Financial Econometrics*, Eds. Y. Aït-Sahalia and L. P. Hansen. Amsterdam: Elsevier.

[12] Bonnans, J.F., Gilbert, J.C., Lemaréchal, C. and Sagastizábal, C.A. (2003) *Numerical optimization. Theoretical and practical aspects*. New York: Springer.

[13] Brandt, M.W. and Santa-Clara, P. (2002) Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. *J. Financial Economics*, **63**, 161-210.

[14] Crommelin, D.T. and Vanden-Eijnden, E. (2006) Reconstruction of diffusions using spectral data from timeseries. *Comm. Math. Sci.*, **4**, 651-668.

[15] Crommelin, D.T. and Vanden-Eijnden, E. (2009) Data-based inference of generators for Markov jump processes using convex optimization. *SIAM Multiscale Model. Simul.*, **7**, 1751-1778.

[16] Darolles, S. and Gouriéroux, C. (2001) Truncated dynamics and estimation of diffusion equations. *J. Econometrics*, **102**, 1-22.

[17] Deuflhard, P. and Schütte, Ch. (2004) Molecular Conformation Dynamics and Computational Drug Design, in *Applied Mathematics Entering the 21st Century*, J. M. Hill and R. Moore (ed.), SIAM.

[18] Ding, J., Li, T.Y. and Zhou, A. (2002) Finite approximations of Markov operators. *J. Comput. Appl. Math.*, **147**, 137-152.

[19] Elarian, O.S., Chib, S. and Shephard, N. (2001) Likelihood inference for discretely observed non-linear diffusions. *Econometrica*, **69**, 959-993.

[20] Eraker, B. (2001) MCMC analysis of diffusion models with application to finance. *J. Business and Econom. Statist.*, **19**, 177-191.

[21] Froyland, G. (1998) Approximating physical invariant measures of mixing dynamical systems in higher dimensions. *Nonlinear Anal. TMA*, **32**, 831-860.

[22] Gobet, E., Hoffmann, M. and Reiß, M. (2004) Nonparametric estimation of scalar diffusions based on low-frequency data. *Ann. Statist.*, **32**, 2223-2253.

[23] Hansen, L.P., Scheinkman, J.A. and Touzi, N. (1998) Spectral methods for identifying scalar diffusions. *J. Econometrics*, **86**, 1-32.

[24] Huisinga, W. (2001) *Metastability of Markovian systems. A transfer operator based approach to molecular dynamics.*. PhD thesis, Free University Berlin.

[25] Hummer, G. (2005) Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, **7**, 34.

[26] Hunt, F.Y. and Miller, W.M. (1992) On the approximation of invariant measures. *J. Statist. Phys.*, **66**, 535-548.

[27] Kessler, M. and Sørensen, M. (1999) Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**, 299-314.

[28] Luce, R. and Perez, S. (1999) Parameter identification for an elliptic partial differential equation with distributed noisy data. *Inverse Problems*, **15**, 291-307.

[29] Nocedal, J. and Wright, S. J. (2006) *Numerical optimization* (2nd edition). New York: Springer.

[30] Nolen, J. and Papanicolaou, G. (2009) Fine scale uncertainty in parameter estimation for elliptic equations. *Inverse Problems*, **25**, 115021.

[31] Papavasiliou, A., Pavliotis, G.A. and Stuart, A.M. (2009) Maximum likelihood drift estimation for multiscale diffusions. *Stochastic Process. Appl.*, **119**, 3173-3210.

[32] Pavliotis, G.A. and Stuart, A.M. (2007) Parameter estimation for multiscale diffusions. *J. Statist. Phys.*, **127**, 741-781.

[33] Pavliotis, G.A. and Stuart, A.M. (2008) *Multiscale methods. Avergaging and homogenization.* New York: Springer.

[34] Pedersen, A.R. (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, **22**, 55-71.

[35] Penland, C. and Magorian, T. (1993) Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *J. Clim.*, **6**, 1067-1076.

[36] Roberts, G.O. and Stramer, O. (2001) On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, **88**, 603-621.

[37] Schütte, Ch. (1998) *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules.* Habilitation thesis, Free University Berlin.

[38] Sørensen, H. (2004) Parametric inference for diffusion processes observed at discrete points in time: a survey. *Int. Statist. Rev.*, **72**, 337-354.

[39] G. W. Stewart and J. Sun, *Matrix perturbation theory*, San Diego: Academic Press, San Diego, 1990.

[40] Sura, P., Newman, M. and Alexander, M.A. (2006) Daily to decadal sea surface temperature variability driven by state-dependent stochastic heat fluxes, *J. Phys. Oceanogr.* **36**, 1940-1958.

[41] Yang, S., Onuchic, J.N. and Levine, H. (2006) Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, **125**, 054910.

[42] Zhang, L., Mykland, P.A. and Aït-Sahalia, Y. (2005) A tale of two time scales: determining integrated volatility with noisy high-frequency data. *J. Amer. Stat. Assoc.*, **100**, 1394-1411.