SIAM J. Applied Dynamical Systems
Vol. 15, No. 3, pp. 1327–1351

Data-Driven Reduction for a Class of Multiscale Fast-Slow Stochastic Dynamical Systems*

Carmeline J. Dsilva†, Ronen Talmon‡, C. William Gear†, Ronald R. Coifman§, and Ioannis G. Kevrekidis¶

Abstract. Multi-time-scale stochastic dynamical systems are ubiquitous in science and engineering, and the reduction of such systems and their models to only their slow components is often essential for scientific computation and further analysis. Rather than being available in the form of an explicit analytical model, often such systems can only be observed as a data set which embodies dynamics on several time scales. We focus on applying and adapting data-mining and manifold learning techniques to detect the slow components in a class of such multiscale data. Traditional data-mining methods are based on metrics (and thus, geometries) which are not informed of the multiscale nature of the underlying system dynamics; such methods cannot successfully recover the slow variables. Here, we present an approach which utilizes both the local geometry and the *local noise dynamics* within the data set through a metric which is both insensitive to the fast variables and more general than simple statistical averaging. Our analysis of the approach provides conditions for successfully recovering the underlying slow variables, as well as an empirical protocol guiding the selection of the method parameters. Interestingly, the recovered underlying variables are *gauge invariant*—they are insensitive to the measuring instrument/observation function.

Key words. multiscale dynamical systems, Mahalanobis distance, diffusion maps

AMS subject classifications. 37M10, 62-07

DOI. 10.1137/151004896

1. Introduction.

We often encounter (whether in model simulations or in experiments) complex, multiscale dynamical systems that can be *effectively simplified*, or *reduced*; that is, one can, at least in principle, write predictive equations in terms of a smaller set of state variables. This situation typically arises in systems containing several disparate time scales; if we are interested in studying the long-term dynamics, we may not want/need to resolve

*Received by the editors January 21, 2015; accepted for publication (in revised form) by J. Guckenheimer March 8, 2016; published electronically July 7, 2016.
http://www.siam.org/journals/siads/15-3/100489.html

†Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544 (cdsilva@princeton.edu, wgear@princeton.edu). The work of the first author was supported by the U.S. Department of Energy Computational Science Graduate Fellowship (CSGF), grant DE-FG02-97ER25308, and by the National Science Foundation Graduate Research Fellowship, grant DGE 1148900.

‡Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, 3200003, Israel (ronen@ef.technion.ac.il). The work of the second author was supported by the European Union's Seventh Framework Programme (FP7) under Marie Curie grant 630657, by the Horev Fellowship, and by the National Science Foundation, award 1309858.

§Department of Mathematics, Yale University, New Haven, CT 06520 (coifman@math.yale.edu). The work of the fourth author was supported by the National Science Foundation, award 1309858.

¶Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 (yannis@arnold.princeton.edu). The work of the fifth author was supported by the National Science Foundation and the Air Force Office of Scientific Research.

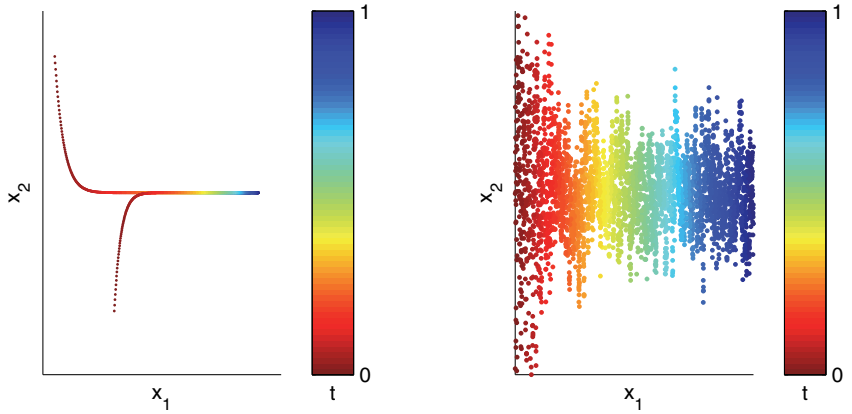


Figure 1. (a) *Schematic of a two-dimensional, two-time-scale ($\tau_1 = 50$ and $\tau_2 = 2$) ordinary differential equation system where the value of x_2 becomes slaved to the value of x_1. In such an example, traditional data-mining algorithms are sufficient to recover the slow variable.* (b) *Schematic of a two-scale, two-dimensional stochastic dynamical system where the statistics of x_2 become slaved to x_1. In such an example, traditional data-mining algorithms will not recover the slow variable if the variance in the fast variable is too large.*

the fast time scales. The term *closure* is used in this context (one can avoid a full dynamic description of both fast and slow state variable dynamics by modeling the effect of the fast variables on the dynamics of the slow variables): model equations *close* at the level of the slow variables only.

Figure 1 illustrates two distinct realizations of this problem: Figure 1(a) arises in a prototypical *deterministic* singularly perturbed case of two coupled ordinary differential equations (ODEs), where the *value* of the fast variable is quickly slaved to the *value* of the slow one. The two-dimensional dynamics are quickly attracted and then slowly evolve on a reduced-dimensional "slow manifold." Figure 1(b) arises in a prototypical *stochastic* Markovian model of two coupled stochastic differential equations (SDEs): a slow SDE and a fast SDE. Here one may argue that it is the *statistics* of the fast variable that become quickly slaved to the *value* of the slow variable. Resolving the dynamics of all state variables—fast as well as slow—at all relevant scales can be challenging, especially when an accurate model of the system is not known and we have only observation data.

In this paper we focus on finding a description that is visually evident in the caricatures of Figure 1: the reduction in dimensionality (from two to one) is clear in the data of Figure 1(a), and observing the data points immediately suggests the possibility of a reduced, one-dimensional description. Yet, inspection of the data in Figure 1(b) (without additional dynamical information) *does not* suggest that a reduced *effective* description may be possible. The coloring of the data by time in Figure 1(b) does, however, suggest that the actual instantaneous x_2 value does not contribute to the macroscale evolution, and that, therefore, a (slow) description might be formulated knowing only the x_1 values. As the formal mathematics for reduction of this class of systems suggests, it is the *average* (the "statistics") of the fast dy-

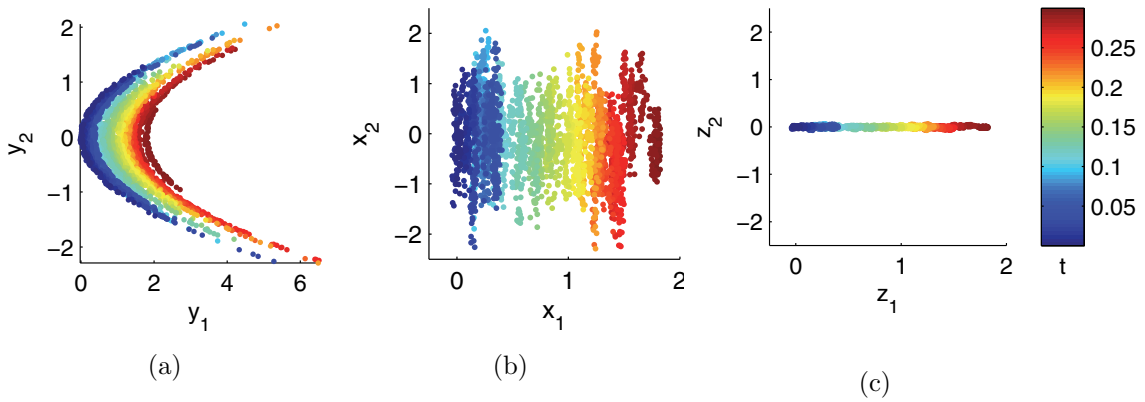


Figure 2. (a) *Example of observed data* $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t))$. (b) *Data from* (a)*, transformed to remove the effect of the function* $\mathbf{f}$. (c) *Data from* (b)*, rescaled so that the fast direction is collapsed. The Mahalanobis distance between the* $\mathbf{y}$ *data approximates the Euclidean distance between the* $\mathbf{z}$ *data without explicitly constructing the above transformations.*

namics that affects the slow dynamics. Indeed, Figure 1(b) was plotted with prior knowledge of the relevant fast and slow variables. Moreover, in this simple case, "coloring with time" suffices to uncover the relevant (slow and fast) variables. In the more general case, however, we are not able to observe the "pure slow" and the "pure fast" variables (namely, the x_1 and x_2 variables of Figure 1(b)); our observations (or our computations) are, in general, *nonlinear functions* of the "pure" state variables (see the schematic in Figure 2(a)). In this paper we show how, for a class of multiscale stochastic dynamical systems that are reducible (in the sense we will designate in what follows), it is possible (by including some *local noise* dynamic information) to devise data-mining protocols that recover the effective (lower) problem dimension (and the identification of the "slow" variables). More specifically, we present, for this class of systems, a method that can successfully both decouple the fast variables from the slow variables (see the schematic in Figure 2(b)) and, at the same time, attenuate ("collapse") the range of the fast variable variation (see the schematic in Figure 2(c)), thus revealing that the system is effectively lower-dimensional.

Our introduction so far has been in terms of only data and schematic caricatures, and, deliberately, no equations have been given. When explicit models of dynamical systems (whether deterministic or stochastic) are available, there exists a broad and deep arsenal of techniques for establishing conditions under which a multi-time-scale dynamical system is effectively reducible. See, for example, Mori–Zwanzig theory [34, 52], singular perturbations theory/slow manifolds for fast-slow systems of ODEs in, e.g., [24, 37], Khasminskii theory for systems of SDEs [28, 6, 35], as well as many others [4, 8, 22]. This approach also includes techniques that (in some cases) construct/approximate the effective reduced model (the effective equation's right-hand side, deterministic or stochastic): given, for example, a system of ODEs with an explicit functional form, one can make numerical approximations, such as the quasi-steady state approximation [40] or the partial equilibrium approximation [17] to reduce the system dimensionality. There has been some recent analytical work on extending and generalizing

such ideas to more complex systems of equations, e.g., [1, 5, 11, 12, 21, 36, 44]. We remark that for well-studied systems, one often has some a priori knowledge of the appropriate observables (such as phase field variables) with which to formulate the reduced dynamics [7, 50]. Yet, such observables may not be immediately obvious upon inspection (to the researcher) for new complex systems.

The particular class of "effectively reducible" systems we target here is (or can be usefully approximated by) the class of systems of (nonlinear) SDEs with additive Brownian noise [33, 32]. More specifically, we refer to systems for which the averaging conditions analyzed in Khasminskii [28, 6] are assumed to hold, so that a meaningful (approximate) effective reduced SDE with white additive noise (or a meaningful ODE) exists. Our particular method relies crucially on an additional property that further narrows our target system class: we need to assume that the (unavailable) reduced effective SDE has constant, decoupled additive noise terms (or that it can be smoothly and invertibly transformed to such an SDE system). This latter feature has also been described as "reducible diffusions" in the literature [1]; note, however, that in this context "reduction" refers to the simplicity of uncoupling the noises, and not to the reduction in the number of variables.

Our work starts with the assumption that the observations/measurements come from such a reducible multi-time-scale stochastic system (reducible both in terms of the number of slow variables and in terms of noise "simplicity"). We have no equations and do not aspire to derive them; we have no way of *proving* that the unavailable equations are indeed reducible, nor do we aspire here to *test* this assumption based on data. Yet some of the main mathematical constructions involved in the relevant proofs (such as the fast sampling of a quasi-equilibrium measure by the fast stochastic variables, whichever they happen to be) do play a role in our data-mining computations. Moving averages and subsampling, for example, have also often been used in simple cases as appropriate functions of variables in which to formulate slow lower-dimensional models [36]. Data-based hypothesis testing and the design of experiments/computations to collect additional observations for that purpose are vital and are only mentioned here; we plan to address these issues in the future.

We employ established manifold learning techniques, in particular, diffusion maps, to process observations from our target class of multiscale stochastic dynamical systems. The main enabling extension is the introduction of an appropriate metric between observation points, informed by the local noise dynamics. At the heart of most manifold learning methods lies a notion of similarity between data points, usually through a distance metric [2, 9, 10, 38, 48]. These pairwise distances are then integrated into a *global* parametrization of the data, typically through the solution of an eigenproblem. Standard "off-the-shelf" manifold learning techniques which utilize the Euclidean distance are not appropriate for analyzing our type of data, since this metric does not account for the disparate time scales. Research efforts have addressed the construction of more informative distance metrics, which are less sensitive to noise and can better recover the true underlying structure in the data by suppressing unimportant sources of variability [3, 19, 39, 41, 51]. The Mahalanobis distance is one such metric. It has been shown that the Mahalanobis distance can remove the effect of *observing* the underlying system variables through a complex, nonlinear function [14, 42, 45]. Here, we will build on the knowledge of how to remove the effects of such nonlinear observation functions, so as to usefully "unscramble" the slow and fast stochastic variables, paving the

way to model reduction. Our approach will build a parametrization of the data which is consistent with the underlying slow variables.

Nonlinearity in the context of manifold learning does not refer to whether the right-hand side of the model equations (in either the original or the transformed variables) may be linear or nonlinear; both are equally acceptable in our case. In our context, it is the nonlinearity of the reduced manifold in the original embedded space that is important for parsimonious reduction (as opposed to, say, linear data reduction to hyperplanes through principal component analysis). It is remarkable that yet a third nonlinearity (the nonlinearity of the observation transformation) almost miraculously disappears when the Mahalanobis metric is incorporated into the diffusion map harmonic analysis machinery.

This paper is organized as follows. In section 2, we present the notation used in this paper. In section 3, we introduce a modified version of the Mahalanobis distance and discuss its properties as well as its specific role in model reduction. In section 4, diffusion maps, the particular manifold learning method used in this paper, are described. In section 5, we present detailed analysis for our method and provide conditions under which it will successfully recover the slow variables. Furthermore, based on this analysis, we present data-driven protocols to appropriately tune the parameters of the method. Finally, in section 6 we show experimental results on illustrative examples.

We note that our presentation and discussion addresses two-time-scale stochastic systems; it is straightforward (even if nontrivial) to "telescope" the approach to similar systems with multi-time-scale separations [18]. Indeed, if there are n_1 "slow" variables, followed by n_2 "faster" variables (after a "$1/\epsilon_1$ gap") as well as n_3 "even faster" variables (after another "$1/\epsilon_2$ gap"), we can observe short bursts of the data over appropriate "very short" time scales (corresponding with the second gap), so as to furnish us with an (n_1+n_2)-dimensional effective model. Then, by observing the results over appropriately short (but not as short as before) time scales, associated with the first gap, the procedure will produce a final n_1-dimensional model.

One final note before starting: in this paper we are going to assume that "sufficient data" are always available, either in advance or because we can always generate them as necessary. This is natural when we want to reduce multiscale simulation data and have available the "fine scale" code that produces them.

2. Notation.

In our setting, we observe samples $\mathbf{y}(t_1), \ldots, \mathbf{y}(t_N)$ from a stochastic dynamical system $\mathbf{y}(t)$ collected at times $t_1, \ldots, t_N$. For our class of "doubly reducible" systems, we assume that there exists a nonlinear coordinate transform $\mathbf{f}$ such that $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t))$, where $\mathbf{x}(t)$ is the following autonomous two-time-scale system of SDEs:

$$
\begin{aligned}
dx_i(t) &= a_i(\mathbf{x}(t))dt + dW_i(t), && 1 \le i \le m, \\
dx_i(t) &= \frac{a_i(\mathbf{x}(t))}{\epsilon}dt + \frac{1}{\sqrt{\epsilon}}dW_i(t), && m+1 \le i \le n,
\end{aligned}
\tag{2.1}
$$

where $a_i(\mathbf{x})$ are (in general nonlinear) functions that satisfy the conditions of SDE reducibility in [28], $W_i(t)$ are independent standard Brownian motions, $\mathbf{x}(t) = \begin{bmatrix} x_1(t) & \cdots & x_n(t) \end{bmatrix}^T \in \mathbb{R}^n$, and $\epsilon \ll 1$. The exposition here is only for (underlying) systems with additive Brownian noise (or systems that can be effectively modeled by them). The parameter ϵ induces a separation

of time scales such that $\mathcal{O}(1)$ changes in $x_1, \ldots, x_m$ occur over times scales $\mathcal{O}(1)$, and $\mathcal{O}(1)$ changes in $x_{m+1}, \ldots, x_n$ occur over time scales $\mathcal{O}(\epsilon)$.

Our goal is to extract a parametrization of the observed data $\mathbf{y}(t)$ that is consistent with the underlying *slow* variables $x_1, \ldots, x_m$. In order to factor out the effects of the nonlinear observation function $\mathbf{f}$,[1] we further assume that $\mathbf{g} = \mathbf{f}^{-1}$ is well defined on the image of $\mathbf{f}$, and both $\mathbf{f}$ and $\mathbf{g}$ are continuously differentiable to fourth order. For the examples presented here, we have $\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^n$; however, we note that in general, the function $\mathbf{f} : \mathbb{R}^n \mapsto \mathbb{R}^d$. We have assumed an m-dimensional slow manifold embedded in $\mathbb{R}^n$. Clearly, we need at least $d = m$ measurement functions to accomplish this; Whitney embedding theorem guarantees that this can be done with $d = 2m$ "generic" measurement functions, but we may well need less. Even if we have $d < m$ measurements, we know that in specific cases (in the spirit of time-delay phase space reconstructions in dynamical systems) we can compensate for the dimensionality gap.

In terms of data requirements, we need to have observations spanning time intervals ($\mathcal{O}(1)$) much larger than ϵ, to ensure that fast variables have had sufficient time to sample their "quasi-stationary" measure, while the slow variables have also had sufficient time to appreciably evolve. For the examples presented, the data $\mathbf{y}(t_1), \ldots, \mathbf{y}(t_N)$ come from a single trajectory sampled at a uniform time interval dt, i.e., $t_i = i \cdot dt$, such that $Ndt \gg \epsilon$; some effects of the sampling time interval will be quantified below.

3. Local data-driven metric. In order to recover the slow variables from data, we will utilize a local metric that collapses the fast directions. Typically, such a metric averages out the fast variables. However, simple averages are inadequate to describe data which are observed through a complicated nonlinear function. Instead, we propose to use the Mahalanobis distance, which measures distances normalized by the respective variances in each local principal direction of the noise. Using this metric, we still retain information about both the fast and slow directions and can more clearly observe complex dynamic behavior within the data set.

If two points $\mathbf{x}_1$ and $\mathbf{x}_2$ are drawn from an n-dimensional Gaussian distribution with covariance $\mathbf{C}_x$, the Mahalanobis distance between the points is defined as [30]

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_M = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{C}_x^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}. \tag{3.1}$$

Now, our data do not come from a given, stationary equilibrium distribution such as a Gaussian distribution, but rather they arise from observations of the stochastic dynamical system (2.1). In the spirit of the Mahalanobis distance (3.1), we will also define a "Mahalanobis distance" between our sampled points, in which the inverse covariance $\mathbf{C}_x^{-1}$ will be replaced by the average of the inverses of the two covariances *of the distributions of the noise* at the two state space points in question, and will be obtained by "freezing" time and sampling the two local noise distributions. It is important to explicitly notice that this is not the meaning one normally ascribes to the word "covariance" in the context of a time series. For a dynamical system like (2.1), and as we will discuss in some more detail below, if the time series sampling window is short enough and the sampling time step fine enough, the "usual" sample

[1] Note that it is equally informative to call it a "measurement" function.

covariance could be used to estimate our "local noise covariance"; yet, we reiterate, to avoid confusion, that the covariance in our Mahalanobis distances is the local noise covariance. We estimate it, as we discuss below, by many parallel short trajectories starting *at the same point*; conceptually, it could also be estimated from a single, short enough, finely enough sampled interval of the time series.

In (2.1) the noise covariance does not depend on $\mathbf{x}$; i.e., $\mathbf{C}_x^{-1} = \mathrm{diag}(e_1, \ldots, e_n)$ is a constant matrix, where

$$\begin{aligned} e_i &= 1, \qquad 1 \le i \le m, \\ e_i &= \epsilon, \quad m+1 \le i \le n. \end{aligned} \tag{3.2}$$

As a result, the Mahalanobis distance (3.1) between any two samples of the system $\mathbf{x}_1$ and $\mathbf{x}_2$ (sampled at some t_1 and t_2, respectively) is given by

$$\|\mathbf{x}_2 - \mathbf{x}_1\|_M^2 = \sum_{i=1}^{n} e_i \left((x_2)_i - (x_1)_i\right)^2, \tag{3.3}$$

where $(x)_i$ denotes the ith coordinate of $\mathbf{x}$. Note that the metric in (3.3) is less sensitive to variations in the fast variables: it can be rewritten as

$$\|\mathbf{x}_2 - \mathbf{x}_1\|_M^2 = \|\mathbf{z}_2 - \mathbf{z}_1\|_2^2, \tag{3.4}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ can be viewed as samples from a stochastic process $\mathbf{z}(t)$ of the same dimension as $\mathbf{x}(t)$, rescaled so that each variable has unit diffusivity; i.e., the relation between their coordinates is given by

$$z_i = \sqrt{e_i} x_i. \tag{3.5}$$

The relation described in (3.4) can be interpreted from two standpoints. From a data analysis standpoint, the Mahalanobis distance provides a *data-driven, constructive* way to normalize the sampled data from the system into a space where the diffusion is a unit diffusion (as guaranteed for reducible systems by [1]). From a dynamical system point of view, (3.5) implies that in this normalized space, the fast variables ($m+1 \le i \le n$) are collapsed by a factor of $\sqrt{\epsilon}$, whereas the slow variables ($1 \le i \le m$) remain unchanged, thereby achieving slow-fast reduction in a purely data-driven manner.

Further intuition regarding the effect of the Mahalanobis distance can be drawn from the following description. Consider a short burst of simulations of (2.1) over time scale $\delta t = o(\epsilon)$. The collection of these bursts yields a cloud of samples which is broadly distributed ($\mathcal{O}(1)$) in the last $n-m$ fast directions but narrowly distributed ($\mathcal{O}(\epsilon)$) in the first m slow ones. In the Mahalanobis distance, this cloud of samples can be "summarized" by the covariance of the local noise, and the inverse of this covariance matrix collapses the fast directions, while leaving the slow directions unchanged.

As previously mentioned, in what we propose to accomplish we will not have access to the original variables $\mathbf{x}(t)$ from the underlying original SDE system (2.1); instead, we will have nonlinear observations $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t))$ that couple the (originally uncoupled) fast and slow

noise directions. The traditional Mahalanobis distance (3.1) is defined for a fixed covariance, whereas in the system $\mathbf{y}(t)$ the covariance of the noise possibly changes as a function of position due to nonlinearities in the observation function $\mathbf{f}$ and in the drift $\mathbf{a}(\mathbf{x})$. Consequently, we use the following *modified definition* of a Mahalanobis-inspired distance between two samples $\mathbf{y}_1$ and $\mathbf{y}_2$ (sampled at some t_1 and t_2):

$$\|\mathbf{y}_2 - \mathbf{y}_1\|_M^2 = \frac{1}{2}(\mathbf{y}_2 - \mathbf{y}_1)^T \left(\mathbf{C}^\dagger(\mathbf{y}_1) + \mathbf{C}^\dagger(\mathbf{y}_2)\right)(\mathbf{y}_2 - \mathbf{y}_1), \tag{3.6}$$

where $\mathbf{C}(\mathbf{y}_j)$ is the covariance of the noise of the observed stochastic process *at the point* $\mathbf{y}_j$, and $\dagger$ denotes the Moore–Penrose pseudoinverse (since d may exceed n). In what follows, by abuse of terminology, we will sometimes refer to the distance in (3.6) simply as the "Mahalanobis distance."

To motivate this definition of the Mahalanobis distance, we first consider the simple linear case where $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, with $\mathbf{A} \in \mathbb{R}^{d\times n}$. The covariance of the observed stochastic process $\mathbf{f}(\mathbf{x})$ is given by $\mathbf{C} = \mathbf{A}\mathbf{C}_x\mathbf{A}^T$. Let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ be the singular value decomposition (SVD) of $\mathbf{A}$, where $\mathbf{U} \in \mathbb{R}^{d\times n}$, $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$, and $\mathbf{V} \in \mathbb{R}^{n\times n}$. The pseudoinverse of the covariance matrix is $\mathbf{C}^\dagger = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{C}_x^{-1}\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^T$. Consequently, the Mahalanobis distance (3.6) is reduced to

$$\begin{aligned}
\|\mathbf{y}_2 - \mathbf{y}_1\|_M^2 &= (\mathbf{y}_2 - \mathbf{y}_1)^T\mathbf{C}^\dagger(\mathbf{y}_2 - \mathbf{y}_1) \\
&= (\mathbf{x}_2 - \mathbf{x}_1)^T\mathbf{A}^T\mathbf{C}_x^{-1}\mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1) \\
&= (\mathbf{x}_2 - \mathbf{x}_1)^T\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{C}_x^{-1}\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T(\mathbf{x}_2 - \mathbf{x}_1) \\
&= (\mathbf{x}_2 - \mathbf{x}_1)^T\mathbf{C}_x^{-1}(\mathbf{x}_2 - \mathbf{x}_1) \\
&= \|\mathbf{x}_2 - \mathbf{x}_1\|_M^2 = \|\mathbf{z}_2 - \mathbf{z}_1\|_2^2.
\end{aligned} \tag{3.7}$$

Hence, our modified Mahalanobis distance between samples of $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t))$ (3.6) allows us to obtain the Euclidean distance *between samples of the rescaled variables* $\mathbf{z}$, where the fast coordinates are collapsed. This would have also been the result achieved by the traditional Mahalanobis distance (3.4), applied directly to the unavailable samples of $\mathbf{x}(t)$.

For general nonlinear observation functions $\mathbf{f}$, it was shown in [42] via Taylor expansion that the Mahalanobis distance (3.6) between the samples from the accessible system $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t))$ approximates the Euclidean distance between samples of the rescaled variables $\mathbf{z}(t)$,

$$\|\mathbf{y}_2 - \mathbf{y}_1\|_M^2 = \|\mathbf{z}_2 - \mathbf{z}_1\|_2^2 + \mathcal{O}(\|\mathbf{y}_2 - \mathbf{y}_1\|_2^4), \tag{3.8}$$

provided that $\mathbf{f}$ is bi-Lipschitz.

The Mahalanobis distance incorporates information about the dynamics and relevant time scales, so that using traditional data-mining techniques with this metric will allow us to detect the slow variables in our data [43]. In particular, because we integrate these distances into a manifold learning algorithm *which effectively takes only local distances into account*, we will recover a parametrization of the data which is consistent with the underlying system variables $\mathbf{x}(t)$, even when the data are "obscured" by the nonlinear observation function $\mathbf{f}$. But clearly from (3.8), this approximation will be accurate only when $\|\mathbf{y}_2 - \mathbf{y}_1\|_2^4 \ll \|\mathbf{z}_2 - \mathbf{z}_1\|_2^2$. This implies that our computations (and the resulting embeddings) are approximately *gauge*

invariant, i.e., they are (approximately) insensitive to the coordinate system (the measurement instrument, the observation function) used [49]. Following [42], in section 5.1 we provide a more detailed analysis of the error incurred by this approximation by exploiting the assumption that both $\mathbf{f}$ and $\mathbf{f}^{-1}$ are differentiable to fourth order.

Another source of error, which has not been addressed thus far in the exposition, stems from the fact that we do not know the (possibly varying) covariance matrix of the noise in the observed system $\mathbf{y}(t)$. Thus, in order to approximate the Mahalanobis distance (3.6) directly from data, the covariance matrix at each sample must be estimated. The error incurred due to this estimation is the focus of section 5.2.

4. Diffusion maps for global parametrization.

In this section we describe the construction of a *global* parametrization of the data that embodies only their slow variable components, solely from the pairwise distances in (3.6). To accomplish this, we will use diffusion maps [9, 10], a kernel-based manifold learning technique. Diffusion maps are typically applied to high-dimensional data that happen to lie on low-dimensional, possibly nonlinear manifolds; for sufficiently densely sampled data, the technique will yield a parsimonious parametrization of this low-dimensional manifold. The parametrization is obtained through solving an eigenvector problem for the Laplace (or Laplace–Beltrami) operator *on the data*; the data should then be thought of as a (generally, irregular) discretization mesh on a manifold. The diffusion map problem is, in effect (and can be asymptotically proven to be [9]), an approximation of the Laplacian eigenproblem on this sampled manifold. Thus, a few leading eigenfunctions of the Laplace (or the Laplace–Beltrami) operator on a manifold can be used to parametrize it (see, for example, [23]).

In our case, however, the data are not just embedded in a high-dimensional ambient space; they actually *are* high-dimensional, since both the fast and the slow directions are extensively sampled by the dynamics. Using our particular Mahalanobis-inspired metric, we will in effect *dramatically collapse* the fast directions, making the data *appear* lower-dimensional (i.e., making them appear to lie on the subspace of the slow variables). This metric thus allows the local noise dynamics to inform the data-mining process, so that an intrinsic geometry of only the slow manifold is recovered.

Given N samples $\mathbf{y}_1, \ldots, \mathbf{y}_N$ of the stochastic system $\mathbf{y}(t)$, we first construct the kernel matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$, where

$$W_{ij} = \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{\sigma_{kernel}^2}\right), \quad i, j = 1, \ldots, N. \tag{4.1}$$

Here, $\|\cdot\|$ denotes the appropriate norm (in our case, the Mahalanobis distance (3.6)), and σ_{kernel} is the kernel scale and denotes a characteristic distance within the data set. Note that σ_{kernel} induces a notion of locality: if $\|\mathbf{y}_i - \mathbf{y}_j\| \gg \sigma_{kernel}$, then W_{ij} is negligible. Therefore, we only need our metric to be informative within a ball of radius $c\sigma_{kernel}$, where c is a constant of $\mathcal{O}(1)$. We then construct the diagonal matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$, with

$$D_{ii} = \sum_{j=1}^{N} W_{ij}, \quad i = 1, \ldots, N. \tag{4.2}$$

We compute the eigenvalues $\lambda_0, \ldots, \lambda_{N-1}$ and eigenvectors $\phi_0, \ldots, \phi_{N-1}$ of the matrix $\mathbf{A} = \mathbf{D}^{-1}\mathbf{W}$ and order them such that $1 = \lambda_0 \geq |\lambda_1| \geq \cdots \geq |\lambda_{N-1}|$. Note that $\phi_0 = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T$ is the trivial eigenvector associated with $\lambda_0 = 1$; the next few eigenvectors provide embedding coordinates for the data, so that $(\phi_j)_i$, the ith entry of ϕ_j, provides the jth embedding coordinate for $\mathbf{y}_i$ (modulo higher harmonics which characterize the same direction in the data [16, 13]).

5. Estimation of the Mahalanobis distance. Errors in the estimation of the Mahalanobis distance from data arise from three sources. One source of error, addressed in section 5.1, stems from the approximation of the function $\mathbf{f}$ locally as a linear function by truncating the Taylor expansion of $\mathbf{f}$ at first order. We control this error due to the truncation of the Taylor expansion by adjusting σ_{kernel} in the Gaussian kernel in (4.1); the higher-order terms in this expansion will be small for samples which are close enough, so, adjusting σ_{kernel} will allow us to only use—in our diffusion map computational scheme—Mahalanobis distances which are sufficiently accurate approximations of the Euclidean distances between the true (unavailable, original) variables.

As already noted above, our Mahalanobis distance (3.6) cannot be computed directly since we do not know the covariance matrices. Instead, we need to *estimate* the covariances from data samples. Let $\mathbf{y}(t_0)$ denote the sample, at time t_0, of the system $\mathbf{y}(t)$. In this paper, we estimate the local covariance of the noise at $\mathbf{y}(t)$, i.e., $\mathbf{C}(\mathbf{y}(t_0))$, empirically from a set of values $\mathbf{y}(t_1), \ldots, \mathbf{y}(t_q)$ drawn from the local distribution of the noise at $\mathbf{y}(t_0)$, by running q parallel, independent simulations for a sufficiently short duration of time $\delta t = o(\epsilon)$, each simulation starting from $\mathbf{y}(t_0)$; we refer to this as a *simulation burst*. We note that one can also consider a single time series of length $q\delta t = o(\epsilon)$ starting from $\mathbf{y}(t_0)$, and then estimate the covariance from *the increments* $\Delta\mathbf{y}(t_i) = \mathbf{y}(t_i) - \mathbf{y}(t_{i-1})$. Although—for simplicity and convenience—we will present analysis and results only for the first type of estimation, the second case is more often encountered in practice. An additional source of error, discussed in section 5.2, will arise from disregarding the drift in the simulation bursts and assuming that the computation of the increments $\Delta\mathbf{y}(t_1), \ldots, \Delta\mathbf{y}(t_q)$ involves drawing from the same local distribution of the noise at $\mathbf{y}(t_0)$. In the estimate we make, we control the error incurred by circumventing the variation of the drift through adjusting the time scale of our simulation bursts δt. The third source of error comes from finite sampling effects in the covariance estimation.

In this paper, we address and discuss the first two sources of error, whereas the finite sampling effects are the subject of future work. We present both analytical results for the error bounds, as well as an empirical methodology to set the parameters σ_{kernel} and δt for our method so as to accurately recover the slow variables. Figure 3 schematically illustrates how choosing the sizes δt (or $q\delta t$ if the alternate method is used) and the parameter σ_{kernel} affects the first two sources of error.

5.1. Error due to the nonlinearity of the observation function $\mathbf{f}$. We define the error incurred by using the Mahalanobis distance in $\mathbf{y}$ to approximate the distance in the "true" (uncoupled and rescaled) variables $\mathbf{z}$ (3.5) as

$$E_M(\mathbf{y}_1, \mathbf{y}_2) = \|\mathbf{z}_2 - \mathbf{z}_1\|_2^2 - \|\mathbf{y}_2 - \mathbf{y}_1\|_M^2. \tag{5.1}$$

By Taylor expansion of $\mathbf{g}(y) = \mathbf{f}^{-1}(y)$ around $\mathbf{y}_1$ and $\mathbf{y}_2$ and averaging the two expansions,

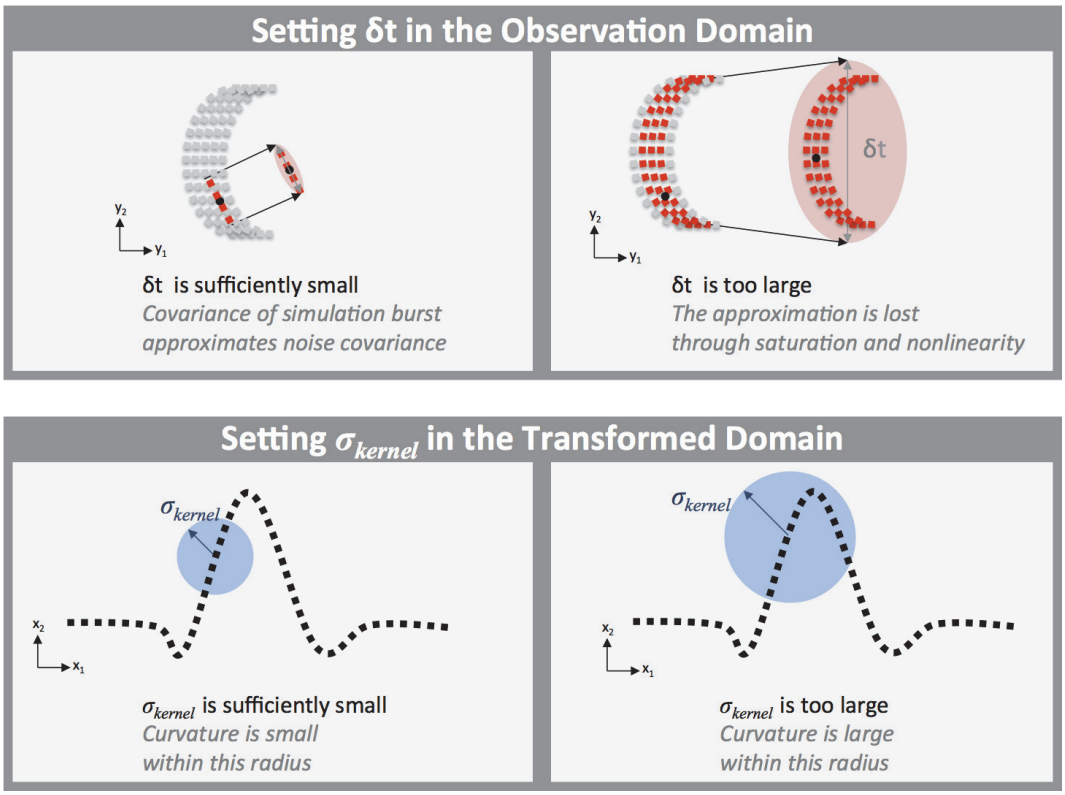


Figure 3. *Illustration of how to choose δt and σ_{kernel} appropriately. Curvature effects and other nonlinearities should be negligible within a time window δt and within a ball of radius σ_{kernel}.*

we obtain

(5.2)

$$
\begin{aligned}
E_M(\mathbf{y}_1,\mathbf{y}_2) = &\frac{1}{2}\sum_{i=1}^{n}\sum_{jkl=1}^{d}\left(g_{i,(j)}(\mathbf{y}_1)g_{i,(k,l)}(\mathbf{y}_1) - g_{i,(j)}(\mathbf{y}_2)g_{i,(k,l)}(\mathbf{y}_2)\right)\\
&\qquad\times((y_2)_j-(y_1)_j)((y_2)_k-(y_1)_k)((y_2)_l-(y_1)_l)\\
&+\frac{1}{8}\sum_{i=1}^{n}\sum_{jklm=1}^{d}\left(g_{i,(j,k)}(\mathbf{y}_1)g_{i,(l,m)}(\mathbf{y}_1) + g_{i,(j,k)}(\mathbf{y}_2)g_{i,(l,m)}(\mathbf{y}_2)\right)\\
&\qquad\times((y_2)_j-(y_1)_j)((y_2)_k-(y_1)_k)((y_2)_l-(y_1)_l)((y_2)_m-(y_1)_m)\\
&+\frac{1}{6}\sum_{i=1}^{n}\sum_{jklm=1}^{d}\left(g_{i,(j)}(\mathbf{y}_1)g_{i,(k,l,m)}(\mathbf{y}_1) + g_{i,(j)}(\mathbf{y}_2)g_{i,(k,l,m)}(\mathbf{y}_2)\right)\\
&\qquad\times((y_2)_j-(y_1)_j)((y_2)_k-(y_1)_k)((y_2)_l-(y_1)_l)((y_2)_m-(y_1)_m)\\
&+\mathcal{O}\left(\|\mathbf{y}_2-\mathbf{y}_1\|_2^6\right),
\end{aligned}
$$

where

$$
\begin{aligned}
g_{i,(j)} &= \sqrt{e_i}\frac{\partial g_i}{\partial y_j}, \\
g_{i,(j,k)} &= \sqrt{e_i}\frac{\partial^2 g_i}{\partial y_j \partial y_k}, \\
g_{i,(j,k,l)} &= \sqrt{e_i}\frac{\partial^3 g_i}{\partial y_j \partial y_k \partial y_l}.
\end{aligned} \tag{5.3}
$$

Singer and Coifman [42] showed that the error incurred by using the Mahalanobis distance to approximate the L_2-distance between samples of $\mathbf{z}(t)$ is $\mathcal{O}(\|\mathbf{y}_1 - \mathbf{y}_2\|_2^4)$ (see the accompanying supplementary materials file 100489_01.pdf [local/web 92.1KB] for details). We now see from (5.2) that the error is an explicit function of the second- and higher-order derivatives of $\mathbf{g} = \mathbf{f}^{-1}$ and the L_2-distance between samples of the observed system, i.e., $\|\mathbf{y}_2 - \mathbf{y}_1\|_2$. Note that this component of the overall error does not depend on the dynamics of the underlying stochastic process (as, in this subsection, we assume the covariance matrices of the noise of the system at each point are known), but is only a function of the observation function $\mathbf{f}$ and the distance between the samples, $\|\mathbf{y}_2 - \mathbf{y}_1\|_2$. The parameter σ_{kernel} in the diffusion map calculation determines how much E_M contributes to the overall analysis. From (4.1), distances which are much greater than σ_{kernel} are negligible in the diffusion map computation because of the exponential kernel. Therefore, we want to choose σ_{kernel}^2 on the order of $\|\mathbf{y}_2 - \mathbf{y}_1\|_M^2$ in a regime where $|E_M(\mathbf{y}_1, \mathbf{y}_2)| \ll \|\mathbf{y}_2 - \mathbf{y}_1\|_M^2$. This is illustrated in Figure 3, where we want to choose σ_{kernel} small enough so that the curvature and other nonlinear effects (captured in the error term E_M) are negligible. This will ensure that the errors in the Mahalanobis distance approximation do not significantly affect our overall analysis.

On first inspection, it would appear that our analysis indicates that σ_{kernel} should be chosen arbitrarily small. However, to obtain a meaningful parametrization of the data set, there must be a nonnegligible number of data samples within a ball of radius σ_{kernel} around each sample. Therefore, the sampling density on the underlying manifold provides a lower bound for σ_{kernel}; what is a feasible (or a realistic) sampling density for a particular case is clearly problem dependent.

5.2. Error in the local noise covariance estimation due to nonconstant drift. To compute the Mahalanobis distance in (3.6), we require $\mathbf{C}$, the covariance of the noise of the observed stochastic process $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t))$. We use simulation bursts to locally explore the dynamics in order to estimate this local noise covariance at a point $\mathbf{y}(t)$ from data [47, 46]. To emphasize the fact that time-dependent simulations are used to implement this local covariance estimation, in this section we use a slightly modified notation and explicitly note the time dependence of the samples involved. We write the elements of the estimated covariance matrix $\widehat{\mathbf{C}}(\mathbf{y}(t), \delta t)$ as

$$
\widehat{C}_{ij}(\mathbf{y}(t), \delta t) = \frac{1}{\delta t}\left(\mathbb{E}\left[y_i(t+\delta t) y_j(t+\delta t) \mid \mathbf{y}(t)\right] - \mathbb{E}\left[y_i(t+\delta t) \mid \mathbf{y}(t)\right]\mathbb{E}\left[y_j(t+\delta t) \mid \mathbf{y}(t)\right]\right), \tag{5.4}
$$

where $\delta t = o(\epsilon)$ is the length of the simulation burst with initial condition $\mathbf{y}(t)$ and $\mathbb{E}[\cdot]$ denotes the expected value.

Due to the drift in the stochastic process and the (perhaps nonlinear) observation function $\mathbf{f}$, we incur some error by approximating the covariance at a point $\mathbf{y}(t)$ using simulations of length $\delta t > 0$. Define the error in this approximation as

$$\mathbf{E}_C(\mathbf{y}(t), \delta t) = \widehat{\mathbf{C}}(\mathbf{y}(t), \delta t) - \mathbf{C}(\mathbf{y}(t)). \tag{5.5}$$

By Itô–Taylor expansion of $\mathbf{f}$ and $\mathbf{x}(t)$ [29], the entries in the error matrix are given by

(5.6)

$$\begin{aligned} &E_{C,ij}(\mathbf{x}(t), \delta t) \\ &\quad = \frac{1}{\delta t} \sum_{k=1}^{n} f_{i,(k)}(\mathbf{x}(t)) \mathbb{E}\left[\int_t^{t+\delta t} \left(\int_{s_2}^{t+\delta t} f_{j,(k,0)}(\mathbf{x}(s_1)) ds_1 + \int_t^{s_2} f_{j,(0,k)}(\mathbf{x}(s_1)) ds_1 \right) ds_2 \right] \\ &\quad + \frac{1}{\delta t} \sum_{k=1}^{n} f_{j,(k)}(\mathbf{x}(t)) \mathbb{E}\left[\int_t^{t+\delta t} \left(\int_{s_2}^{t+\delta t} f_{i,(k,0)}(\mathbf{x}(s_1)) ds_1 + \int_t^{s_2} f_{i,(0,k)}(\mathbf{x}(s_1)) ds_1 \right) ds_2 \right] \\ &\quad + \frac{1}{\delta t} \sum_{k,l=1}^{n} \mathbb{E}\left[\int_t^{t+\delta t} \left(\int_t^{s_2} f_{i,(k,l)}(\mathbf{x}(s_1)) dW_{s_1,k} \right) \left(\int_t^{s_2} f_{j,(k,l)}(\mathbf{x}(s_1)) dW_{s_1,k} \right) ds_2 \right] \\ &\quad + \mathcal{O}(\delta t^{3/2}), \end{aligned}$$

where

(5.7)

$$\begin{aligned} f_{i,(k)} &= \frac{1}{\sqrt{e_k}} \frac{\partial f_i}{\partial x_k}, \\ f_{i,(k,l)} &= \frac{1}{\sqrt{e_k e_l}} \frac{\partial^2 f_i}{\partial x_k \partial x_l}, \\ f_{i,(k,0)} &= \frac{1}{\sqrt{e_k}} \sum_{l=1}^{n} \left(\frac{\partial}{\partial x_k} \left(\frac{a_l(\mathbf{x})}{e_l} \frac{\partial f_i}{\partial x_l} \right) + \frac{1}{2e_l} \frac{\partial^3 f_i}{\partial x_k \partial x_l^2} \right), \\ f_{i,(0,k)} &= \frac{1}{\sqrt{e_k}} \sum_{l=1}^{n} \left(\frac{a_l(\mathbf{x})}{e_l} \frac{\partial^2 f_i}{\partial x_k \partial x_l} + \frac{1}{2e_l} \frac{\partial^3 f_i}{\partial x_k \partial^2 x_l} \right). \end{aligned}$$

From (5.6), the error in the covariance is $\mathcal{O}(\delta t)$ (as the integrals over s_i are each $\mathcal{O}(\delta t)$, and the dW integrals are each $\mathcal{O}(\sqrt{\delta t})$) and a function of the derivatives of the observation function $\mathbf{f}$ and the drift $\mathbf{a}$. To obtain accurate covariance estimation, we set δt such that $\|\mathbf{E}_C\| \ll \|\mathbf{C}\|$ (as illustrated in the schematic of Figure 3). Note that in practice, we compute $\widehat{\mathbf{C}}(\mathbf{y}(t))$ by running many parallel simulations of length δt starting from the state space point $\mathbf{y}(t)$, and then we use the sample average to approximate the expected values in (5.4). We therefore incur additional error due to finite sampling; as already stated, this error is ignored for the purposes of our analysis here, and quantifying it is the subject of future work.

Our analysis reveals that the errors decrease with decreasing δt; at first inspection, one would want to set δt arbitrarily small to obtain the highest accuracy possible. However, often in practice, one cannot implement an arbitrarily refined sampling rate, and a smaller δt will then result in fewer samples with which to approximate the local noise covariance.[2] Therefore,

[2]The work of Pavliotis and Stuart [36] about the need for subsampling in the estimation of multiscale diffusions may arise as yet another consideration in providing a lower bound for σ_{kernel}; we will not discuss this additional constraint any further here.

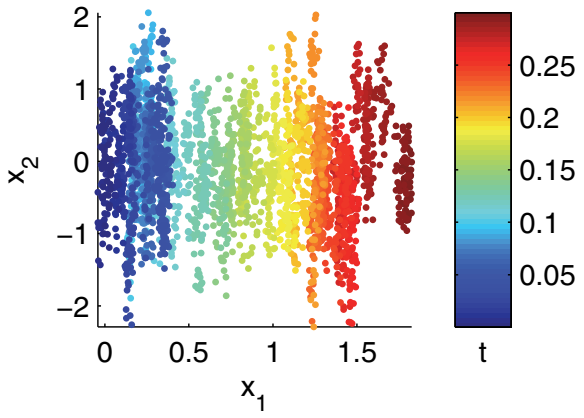


Figure 4. *Data simulated from* (6.1) *($\epsilon = 10^{-3}$) for* 3000 *time steps with $dt = 10^{-4}$. The data are colored by time.*

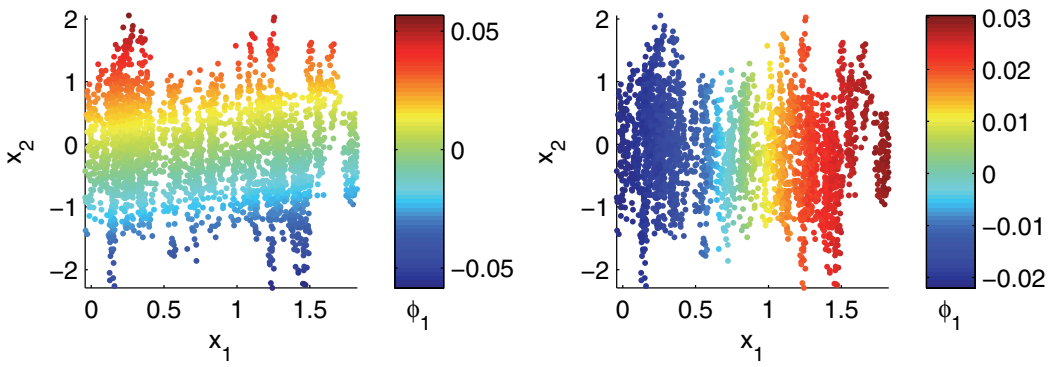
when also accounting for these finite sampling errors, one should take δt as long as possible while still maintaining negligible errors from the observation function $\mathbf{f}$ and the drift $\mathbf{a}$.

6. Illustrative examples. For illustrative purposes, we consider the two-dimensional SDE

$$\begin{aligned} dx_1(t) &= 3dt + dW_1(t), \\ dx_2(t) &= -\frac{x_2(t)}{\epsilon}dt + \frac{1}{\sqrt{\epsilon}}dW_2(t), \end{aligned} \tag{6.1}$$

with $\epsilon = 10^{-3}$, as a specific example of (2.1). x_1 is the slow variable, and x_2 is a fast noise whose equilibrium measure is bounded and $\mathcal{O}(1)$. Figure 4 shows data simulated from this SDE colored by time. We would like to recover a parametrization of this data which is one-to-one with the slow variable x_1.

The reader's first reaction is that this is too simple an example (the drifts are too linear, too constant, too uncoupled, etc.), but it is selected for clarity of representation of what we perceive as the important issues; what we discuss will also be valid for "more generic" examples with drifts that are nonlinearly state-dependent. We offer one additional (small, but nontrivial) remark: this sense of simplicity, which one gets from inspection of the formula, would not be perceived if one were provided only with measurement data from this stochastic process, and our goal here is to develop a method that is data-driven, without knowledge of the underlying formulas. Consequently, we also do not attempt here to test (based on the data) whether the dynamical system that gave rise to the data is indeed "reducible" in both the sense of (a) being effectively describable by a lower-dimensional SDE, and (b) the existence of a variable transformation that renders the noise constant and diagonal. In this work we *assume* both of these "reducibilities" and proceed to effectively reduce *the data*. Testing these reducibilities, even with explicit formulas for the SDE, is case-dependent nontrivial research; formulating data-based tests for this hypothesis is clearly even more difficult and is—as far as we know—largely uncharted research territory.



(a) (b)

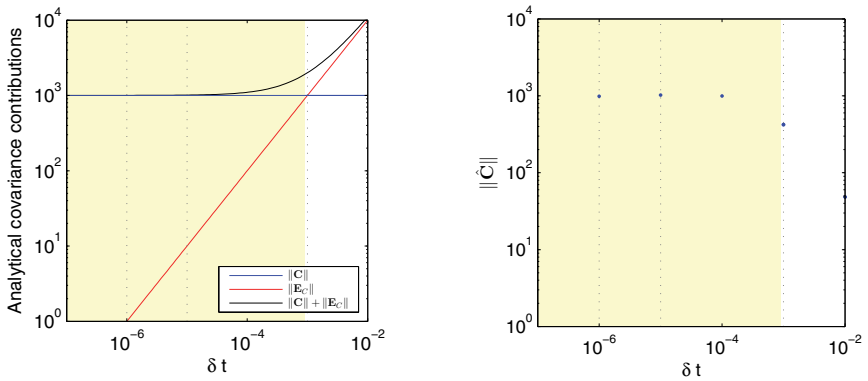
Figure 5. *Comparison of using the Euclidean distance and the Mahalanobis distance in multiscale data mining.* (a) *The data from Figure 4, colored by the first diffusion map coordinate when using the Euclidean distance in the kernel in (4.1). Note that we do* not *recover the slow variable.* (b) *The data from Figure 4, colored by the first diffusion map coordinate when using the Mahalanobis distance in the kernel in (4.1). The good correspondence between this coordinate and the slow variable is visually obvious.*

6.1. Linear observation function. In the first example, our observation function $\mathbf{f}$ will be the identity function,

$$\begin{aligned} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} &= \mathbf{f}(\mathbf{x}(t)) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \\ \mathbf{g}(\mathbf{y}(t)) &= \mathbf{f}^{-1}(\mathbf{y}(t)) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}, \end{aligned} \tag{6.2}$$

where the fast and slow noise terms remain uncoupled. In this case, there is no error incurred due to the observation function $\mathbf{f}$ ($E_M = 0$), as the second- and higher-order derivatives of $\mathbf{g}$ are identically 0.

6.1.1. Importance of using the Mahalanobis distance. We want to demonstrate the utility of using the Mahalanobis distance compared to using the typical Euclidean distance. We compute the diffusion map embedding for the data in Figure 4, using both the standard Euclidean distance and the Mahalanobis distance for the computation of the kernel in (4.1). The data, colored by ϕ_1 using the two different metrics, are shown in Figure 5. When using the standard Euclidean distance, which does not account for the underlying dynamics, the first nontrivial diffusion map coordinate visibly recovers the fast variable x_2; it would appear that data analysis suggests that the fast mode is the dominant factor (Figure 5(a)). In contrast, the slow variable is recovered as the dominant factor when using the Mahalanobis distance; indeed, the coloring in Figure 5(b) (where the data are colored by the first diffusion map coordinate) is consistent with the coloring in Figure 4 (where the data are colored by time).



(a) (b)

Figure 6. *Errors in covariance estimation for the linear example.* (a) *The analytical contributions to the covariance for the example in section 6.1 as a function of* δt. (b) *The average estimated covariance* $\|\widehat{\mathbf{C}}\|$ *as a function of* δt. *The average is computed over* 10 *data points and using* 50 *sample points to estimate each covariance. The shaded yellow region indicates the range of* δt *over which the errors in the estimated covariance are less than the norm of the covariance.*

6.1.2. Errors in covariance estimation. For the example in (6.2), the analytical covariance matrix is

$$\mathbf{C}(\mathbf{x}(t)) = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\epsilon} \end{bmatrix}. \tag{6.3}$$

From (5.6), we find the error matrix

$$\mathbf{E}_C(\mathbf{x}(t), \delta t) = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{\delta t}{\epsilon^2} \end{bmatrix} + \mathcal{O}(\delta t^{3/2}). \tag{6.4}$$

Therefore, $\|\mathbf{C}\| = \mathcal{O}(\frac{1}{\epsilon})$ and $\|\mathbf{E}_C\| = \mathcal{O}(\frac{\delta t}{\epsilon^2})$. These terms are shown in Figure 6(a) as a function of δt. We want to choose $\delta t = o(\epsilon)$ such that $\|\mathbf{E}_C\| \ll \|\mathbf{C}\|$ (the yellow shaded region in Figure 6 indicates where $\|\mathbf{E}_C\| < \|\mathbf{C}\|$), so that the errors in the estimated covariance are small with respect to the covariance itself.

When we do not analytically know the functions $\mathbf{f}$ or $\mathbf{g}$, or the value of ϵ, we can detect such a regime empirically by estimating the covariance for several values of δt. This provides an estimate of $\widehat{\mathbf{C}} = \mathbf{C} + \mathbf{E}_C$ as a function of δt. From Figure 6(a), we expect a "knee" in the plot of $\|\widehat{\mathbf{C}}\|$ versus δt when $\|\mathbf{E}_C\|$ becomes larger than $\|\mathbf{C}\|$. Figure 6(b) shows the empirical $\|\widehat{\mathbf{C}}\|$ as a function of δt for the data in Figure 4, and the knee in this curve is consistent with the intersection in Figure 6(a).

6.1.3. Recovery of the fast variable. Note that, for the example in (6.2), $\mathbf{E}_C$ is a constant diagonal matrix. Therefore, taking δt too large will not lead to nonlinear effects or mixing of the fast and slow variables. Rather, changing δt will only affect the perceived ratio of the fast and slow time scales (the fast "direction" will soon saturate).

To see this behavior in our diffusion map results, we must first discuss the interpretation of the diffusion map eigenspectrum. The diffusion map eigenvectors provide embedding coordinates for the data, and the corresponding eigenvalues provide a measure of the importance of each coordinate. However, some eigenvectors can be harmonics of previous eigenvectors [20, 13]; for example, for a data set parametrized by a variable x, both $\cos x$ and $\cos 2x$ will appear as diffusion map eigenvectors (see [16] for a more detailed discussion). These harmonics do not capture any new direction within the data set, but do appear as additional eigenvector/eigenvalue pairs. Therefore, for the two-dimensional data considered here, the fast variable will not necessarily appear as the second (nontrivial) eigenvector. As the time-scale separation increases, the relative importance of the slow and fast directions will also increase. This implies that the eigenvalue corresponding to the eigenvector which parametrizes the fast direction will decrease, and the number of harmonics of the slow mode which appear before the fast mode will increase.

Figure 7 shows results for three different values of δt (the corresponding values are indicated by the dashed lines in Figure 6). When the time scale of the simulation burst used to estimate the local covariance (indicated by the red clouds in the top row of figures) is sufficiently shorter than that of the "saturation" time for the fast variable, the estimated local covariance is accurate and the (variability in the) fast direction is attenuated significantly relative to the slow variable. This implies that the diffusion map coordinate parametrizing the fast variable will first arise (and thus, the fast variable will be "recovered") *noticeably* far down in the ordered diffusion map eigenvectors.

The left and middle columns of Figure 7 show that, for this example, when the simulation burst is shorter than the fast saturation time, the fast variable is recovered as ϕ_{10}. However, if the time scale of the burst is *longer* than the saturation time of the fast variable, the estimated covariance changes: the variance in the slow direction continues to grow, while the variance in the fast direction is fixed. This implies that the apparent time-scale separation is smaller, the attenuation of the fast variable is less pronounced relative to the slow variable, and the fast variable is recovered in an earlier eigenvector (in our ordering of the spectrum). The right column of Figure 7 shows that, when the burst is now longer than the equilibration time, the fast variable appears earlier in the eigenvalue spectrum and is recovered as ϕ_6.

We only discuss these "long" bursts in relation to the recovery (should that be desirable) of the fast variable. We have been focusing here on discovering the slow variables; if the burst time is long enough for the fast variable to appear "early," we are clearly running for too long, and the local noise covariance estimation is already flawed.

6.2. Nonlinear observation function. In the second example, our data from section 6.1 will be warped into "half-moon" shapes via the function

$$\begin{aligned} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} &= \mathbf{f}(\mathbf{x}(t)) = \begin{bmatrix} x_1(t) + x_2^2(t) \\ x_2(t) \end{bmatrix}, \\ \mathbf{g}(\mathbf{y}(t)) &= \mathbf{f}^{-1}(\mathbf{y}(t)) = \begin{bmatrix} y_1(t) - y_2^2(t) \\ y_2(t) \end{bmatrix}. \end{aligned} \tag{6.5}$$

Figure 8 shows the data from Figure 4 transformed by the function $\mathbf{f}$ in (6.5) and colored by time. This is a difficult class of problem in practice, as none of the observed/measured

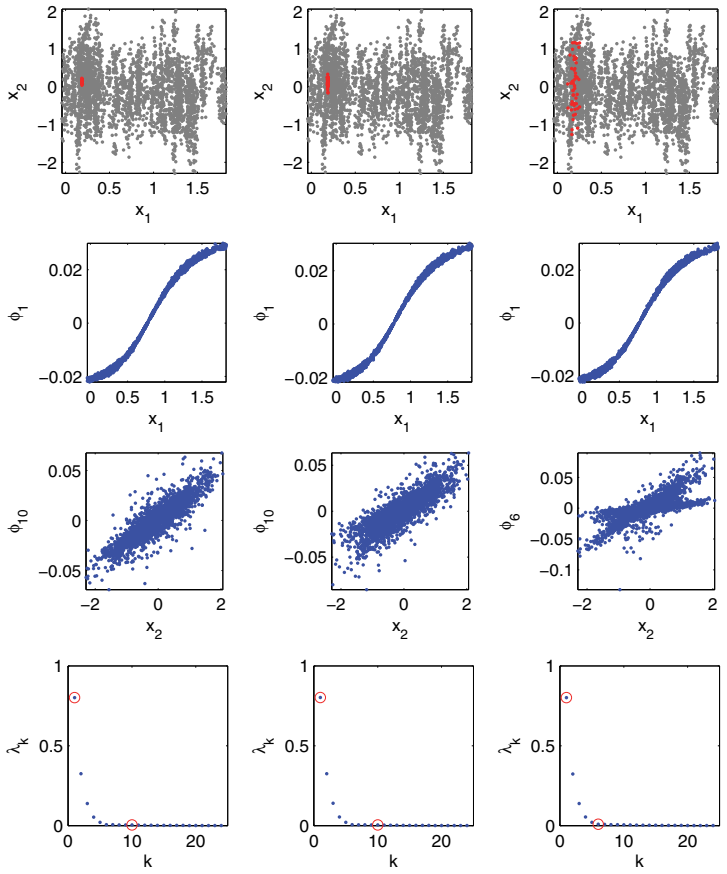


Figure 7. *Relationship between changing δt and recovery of the variables. From left to right, the columns correspond to $\delta t = 10^{-6}, 10^{-5}, 10^{-3}$. Row* 1: *Data (gray) and representative burst (red) used to estimate the local covariance. Row* 2: *Correlation between the first diffusion map coordinate and the slow variable x_1. Row* 3: *Correlation between the relevant diffusion map coordinate and the fast variable x_2. Note that for $\delta t = 10^{-6}$ and $\delta t = 10^{-5}$, x_2 is correlated with ϕ_{10}. When $\delta t = 10^{-3}$, x_2 is correlated with ϕ_6. Row* 4: *Diffusion map eigenvalue spectra. The eigenvalues corresponding to the coordinates for the slow and fast modes are indicated by red circles. Note that when δt is too large, the* apparent *(as opposed to actual) time-scale separation decreases and the coordinate corresponding to the fast variable appears earlier in the spectrum.*

variables is purely slow, and the *observed* system appears, at first inspection, to possess no separation of time scales. For this example, the analytical covariance and inverse covariance are

$$\mathbf{C}(\mathbf{x}(t)) = \frac{1}{\epsilon}\begin{bmatrix} \epsilon + 4x_2^2(t) & 2x_2(t) \\ 2x_2(t) & 1 \end{bmatrix}, \tag{6.6}$$
$$\mathbf{C}^{\dagger}(\mathbf{x}(t)) = \begin{bmatrix} 1 & -2x_2(t) \\ -2x_2(t) & \epsilon + 4x_2^2(t) \end{bmatrix}.$$

The fast and slow variables are now coupled through the function $\mathbf{f}$, and the Euclidean distance is not informative about the fast *or* about the slow variables. We need to use the

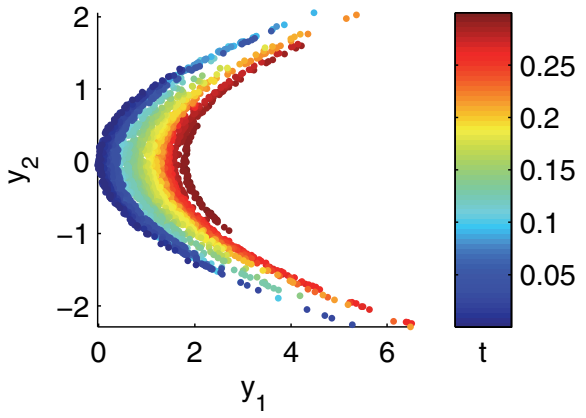


Figure 8. *The data from Figure 4, transformed by* $\mathbf{f}$ *in* (6.5).

Mahalanobis distance to obtain a parametrization that is consistent with the underlying fast-slow dynamics.

6.2.1. Errors in Mahalanobis distance. We can bound the Mahalanobis distance by the eigenvalues of $\mathbf{C}^\dagger$:

$$\lambda_{C^\dagger,1}\|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_2^2 \le \|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2 \le \lambda_{C^\dagger,2}\|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_2^2, \tag{6.7}$$

where $\lambda_{C^\dagger,1} \le \lambda_{C^\dagger,2}$ are the two eigenvalues of $\mathbf{C}^\dagger$. Therefore, for the example in (6.5), we have

$$E_M(\mathbf{y}(t_1),\mathbf{y}(t_2)) = -(y_2(t_2)-y_2(t_1))^4. \tag{6.8}$$

Figure 9(a) shows $\|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$ and $|E_M|$ as a function of $\|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_2$. The Mahalanobis distance is an accurate approximation to the true distance $\|\mathbf{z}(t_2)-\mathbf{z}(t_1)\|_2$ when $|E_M| \ll \|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$ (the shaded yellow region in the plot indicates where $|E_M| < \|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$). We want to choose σ_{kernel}^2 in a regime where $|E_M(\mathbf{y}(t_1),\mathbf{y}(t_2))| \ll \|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$, so that the distances we utilize in the diffusion map calculation are accurate. We can find such a regime empirically by plotting $\|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$ as a function of $\|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_2$, and assessing when the relationship deviates from quadratic. This is shown in Figure 9(b), and the deviation from quadratic behavior is consistent with the intersection of the analytical expressions plotted in Figure 9(a).

Figure 10 (rows 1 and 2) shows the data from Figure 8, colored by ϕ_1 for two different values of σ_{kernel}. The corresponding values of σ_{kernel}^2 are indicated by the dashed lines. When σ_{kernel}^2 corresponds to a region where $|E_M| \ll \|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$, ϕ_1 is well correlated with the slow variable. However, when σ_{kernel}^2 corresponds to a region where $|E_M| \gg \|\mathbf{y}(t_2)-\mathbf{y}(t_1)\|_M^2$, the slow variable is no longer recovered.

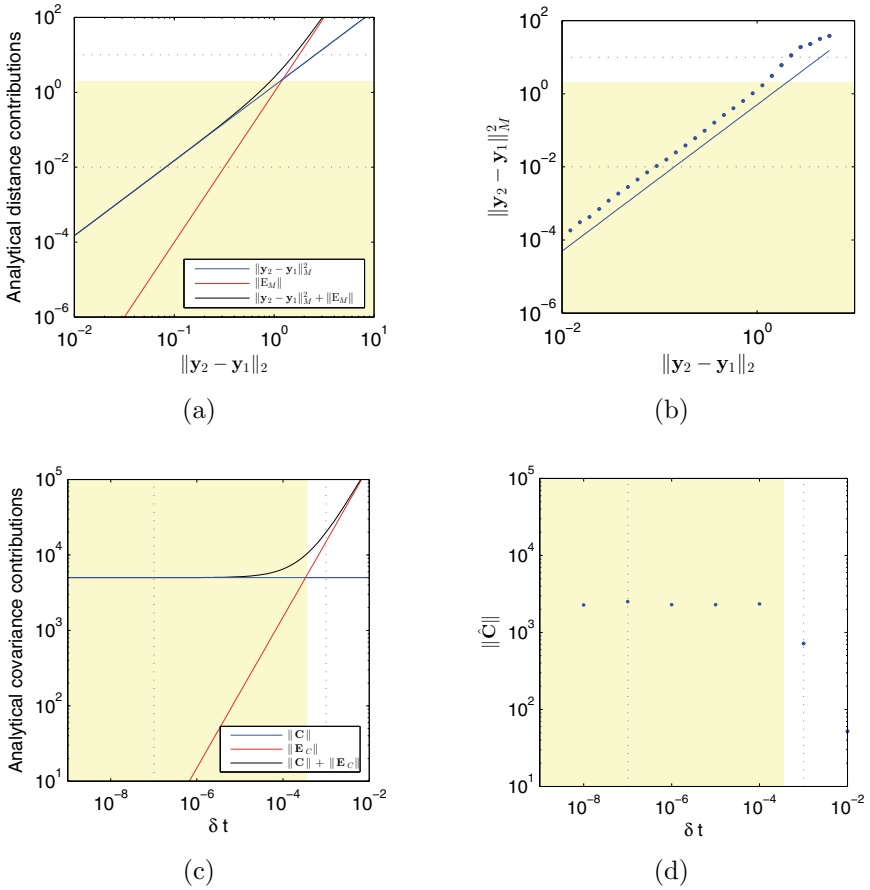


Figure 9. *Errors in the Mahalanobis distance and the covariance estimation for the nonlinear example in (6.5). (a) The analytical expressions for the contributions to the distance approximation as a function of* $\|\mathbf{y}_2-\mathbf{y}_1\|_2$. *(b) The average estimated Mahalanobis distance* $\|\mathbf{y}_2-\mathbf{y}_1\|_M^2$ *as a function of the distance* $\|\mathbf{y}_2-\mathbf{y}_1\|_2$. *The line* $\|\mathbf{y}_2-\mathbf{y}_1\|_M^2=\|\mathbf{y}_2-\mathbf{y}_1\|_2$ *is shown for reference. The yellow region indicates the range in which* σ_{kernel} *should be chosen. (c) The analytical expressions for the contributions to the covariance as a function of* δt. *(d) The average estimated covariance* $\|\widehat{\mathbf{C}}\|$ *as a function of* δt. *The yellow region indicates the range of* δt *over which the errors in the estimated covariance are small relative to the norm of the covariance.*

6.2.2. Errors in covariance estimation. From (5.6), we find that, for the example in (6.5),

(6.9)
$$E_{C,11}(\mathbf{x}(t),\delta t)=\frac{2\delta t}{\epsilon^2}-\frac{8x_2(t)}{\epsilon^2\delta t}\mathbb{E}\left[\int_t^{t+\delta t}\left(\int_{s_2}^{t+\delta t}2x_2(s_1)ds_1+\int_t^{s_2}x_2(s_1)ds_1\right)ds_2\right]+\mathcal{O}(\delta t^{3/2}),$$
$$E_{C,12}(\mathbf{x}(t),\delta t)=E_{C,21}(\mathbf{x}(t),\delta t)$$
$$=-\frac{x_2(t)\delta t}{\epsilon^2}-\frac{2}{\epsilon^2\delta t}\mathbb{E}\left[\int_t^{t+\delta t}\left(\int_{s_2}^{t+\delta t}2x_2(s_1)ds_1+\int_t^{s_2}x_2(s_1)ds_1\right)ds_2\right]+\mathcal{O}(\delta t^{3/2}),$$
$$E_{C,22}(\mathbf{x}(t),\delta t)=-\frac{\delta t}{\epsilon^2}+\mathcal{O}(\delta t^{3/2}).$$

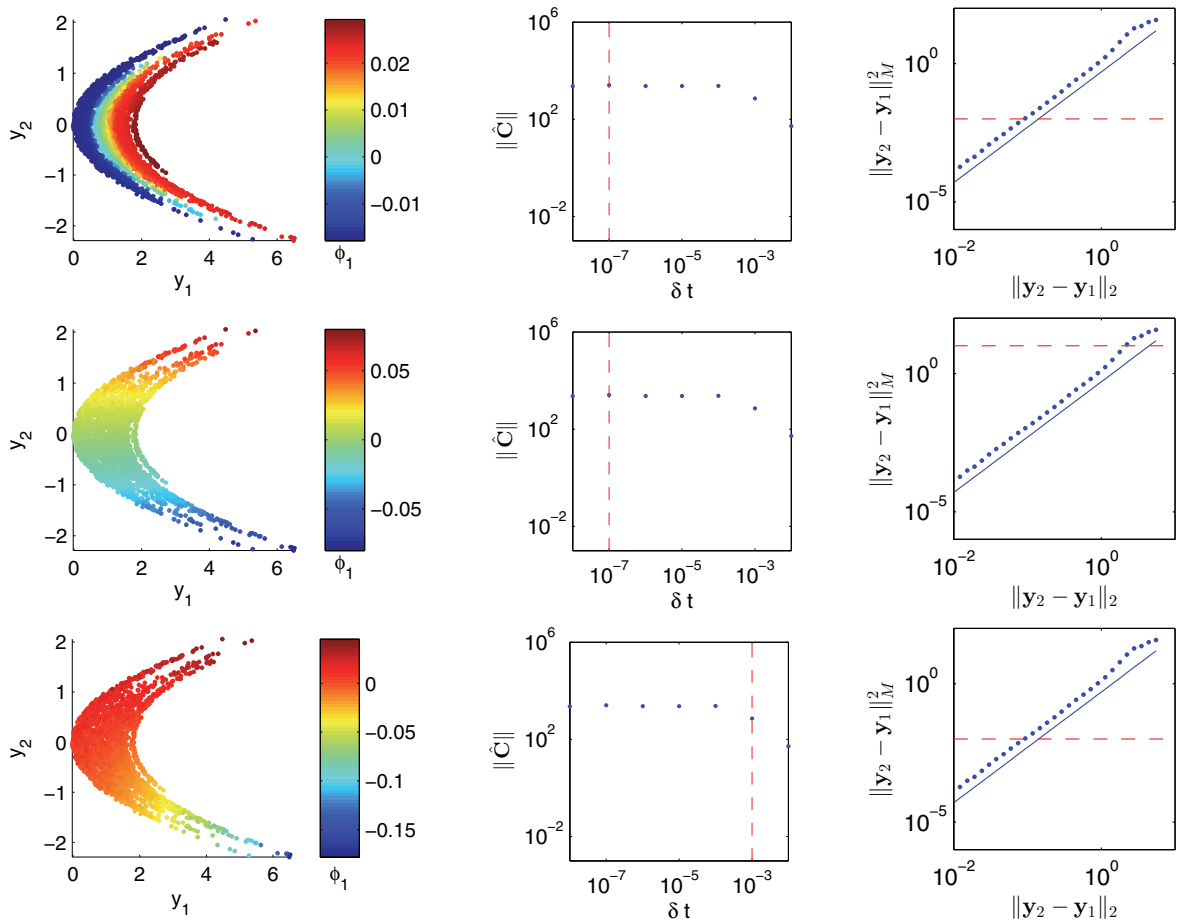


Figure 10. *Data from Figure 8, colored by the first diffusion map variable ϕ_1 using the Mahalanobis distance for three different parameter settings. The relevant values of δt and σ_{kernel} are indicated by the red dashed lines on the corresponding plots. Row 1: $\delta t = 10^{-7}$ and $\sigma^2_{kernel} = 10^{-2}$. Note that the parametrization is one-to-one with the slow variable. Row 2: $\delta t = 10^{-7}$ and $\sigma^2_{kernel} = 10^{1}$. We do not recover the slow variable because σ_{kernel} is too large. Row 3: $\delta t = 10^{-3}$ and $\sigma^2_{kernel} = 10^{-2}$. We do not recover the slow variable because δt is too large.*

The error in the covariance is $\mathcal{O}(\frac{\delta t}{\epsilon^2})$. As expected, the error grows with increasing δt. We can also see the explicit dependence of the covariance error (for fixed δt) on the time scale separation ϵ; larger time scale separation results in a larger covariance error, as a more refined simulation burst is required to estimate the covariance of the fast directions. $\|\mathbf{C}\|$ and $\|\mathbf{E}_C\|$ are plotted as a function of δt in Figure 9(c); the shaded yellow portion denotes the region where $\|\mathbf{E}_C\| < \|\mathbf{C}\|$. As in the previous example, we can empirically find where $\|\mathbf{E}_C\| \ll \|\mathbf{C}\|$ by plotting $\|\widehat{\mathbf{C}}\|$ as a function of δt and looking for a knee in the plot. These results are shown in Figure 9(d).

Figure 10 (rows 1 and 3) shows the data from Figure 8, colored by ϕ_1 for two different values of δt. The corresponding values of δt are indicated by the dashed lines in Figures 9(c) and 9(d). When δt corresponds to a region where $\|\mathbf{E}_C\| \ll \|\mathbf{C}\|$, the slow variable is recovered by the

first diffusion map coordinate. However, when δt corresponds to a region where $\|\mathbf{E}_C\| \gg \|\mathbf{C}\|$, the slow variable is no longer recovered.

7. Conclusions. We have presented a methodology for computing a parametrization of a data set, assumed to arise from observations of a class of multiscale (fast-slow) stochastic dynamical systems, which "respects" the slow variables in the underlying dynamical system. The approach utilizes diffusion maps, a kernel-based manifold learning technique, with a modified Mahalanobis pairwise distance. We showed that this Mahalanobis distance attenuates ("collapses") the fast directions within a data set, allowing for successful recovery of the slow variables. Furthermore, we showed how to estimate the covariances (required for this Mahalanobis distance computation) directly from data. A key point in our approach is that the embedding coordinates we compute are not only insensitive to the fast variables, but are also (approximately) *invariant to nonlinear observation functions*. Therefore, the approach can also be used for *data fusion*: data collected from the same system via different observation functions can be combined and merged into a single "intrinsic" (or "canonical") coordinate system.

In the examples presented, the initial data came from a single trajectory of a dynamical system, and the local covariance at each point in the trajectory was estimated using brief simulation bursts. However, the initial data need not be collected from a single trajectory, and other sampling schemes could be employed. Brief time series were required here to estimate the local noise covariances, but given a simulator, one could reinitialize brief simulation bursts which are sufficiently short and refined from each sample point if necessary.

In our examples, we controlled the time scale of sampling; we could therefore set the time scale over which to estimate the noise covariance and could set the simulation time step to be arbitrarily small. However, in some settings, such as previously collected historical data, it is not uncommon to have a fixed sampling rate and be unable to reinitialize simulations at will. In such cases, and for the particular data, we might not be able to find an appropriate kernel scale given the fixed δt such that we can accurately recover the slow variables. For these cases, the data cannot be processed as given; there is a possibility of remedying the situation by constructing useful intermediate observers, such as histograms, Fourier coefficients, or scattering transform coefficients [31, 46, 47]. Such intermediates are more complex statistical functions than simple averages and may be able to capture additional structure within the data. They also reduce the effects of noise and permit a larger time step. However, constructing such intermediates often requires additional a priori knowledge about the system dynamics and noise structure, and it was not pursued here.

As repeatedly stated, in our analysis, we have ignored finite sampling effects in our estimation. In reality, both the number of samples used to estimate the covariances as well as the density of sampled points on the manifold affect the recovered parametrization and provide additional constraints on δt and σ_{kernel}. Future work involves extending our analysis to the finite sample case and providing guidelines for the amount of data required to apply our approach.

The method presented here provides a bridge between traditional data mining and a class of multi-time-scale dynamical systems. With this interface established, one can envisage using such data-driven methodologies to extract reduced models (either explicitly or implicitly via

an equation-free framework [15, 25, 26, 27]) which also respect the underlying slow dynamics and geometry of the data. Such reduced models hold the promise of accelerated analysis and reduced simulation of dynamical systems whose effective dynamics are obscure upon simple inspection.

REFERENCES

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

[2] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396.

[3] T. Berry, J. R. Cressman, Z. Gregurić-Ferenček, and T. Sauer, *Time-scale separation from diffusion-mapped delay coordinates*, SIAM J. Appl. Dyn. Syst., 12 (2013), pp. 618–649, doi:10.1137/12088183X.

[4] J. J. Brey, R. Zwanzig, and J. R. Dorfman, *Nonlinear transport equations in statistical mechanics*, Phys. A, 109 (1981), pp. 425–444.

[5] C. P. Calderon, *Fitting effective diffusion models to data associated with a "glassy" potential: Estimation, classical inference procedures, and some heuristics*, Multiscale Model. Simul., 6 (2007), pp. 656–687, doi:10.1137/050643647.

[6] S. Cerrai, *A Khasminskii type averaging principle for stochastic reaction-diffusion equations*, Ann. Appl. Probab., 19 (2009), pp. 899–948.

[7] L.-Q. Chen, *Phase-field models for microstructure evolution*, Annu. Rev. Mater. Res., 32 (2002), pp. 113–140.

[8] A. J. Chorin, O. H. Hald, and R. Kupferman, *Optimal prediction and the Mori–Zwanzig representation of irreversible processes*, Proc. Natl. Acad. Sci. USA, 97 (2000), pp. 2968–2973.

[9] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.

[10] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci., 102 (2005), pp. 7426–7431.

[11] M.-N. Contou-Carrere, V. Sotiropoulos, Y. N. Kaznessis, and P. Daoutidis, *Model reduction of multi-scale chemical Langevin equations*, Systems Control Lett., 60 (2011), pp. 75–86.

[12] G. Q. Dong, L. Jakobowski, M. A. J. Iafolla, and D. R. McMillen, *Simplification of stochastic chemical reaction models with fast and slow dynamics*, J. Biol. Phys., 33 (2007), pp. 67–95.

[13] C. J. Dsilva, R. Talmon, R. R. Coifman, and I. G. Kevrekidis, *Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study*, Appl. Comput. Harmon. Anal., (2015), doi:10.1016/j.acha.2015.06.008.

[14] C. J. Dsilva, R. Talmon, N. Rabin, R. R. Coifman, and I. G. Kevrekidis, *Nonlinear intrinsic variables and state reconstruction in multiscale simulations*, J. Chem. Phys., 139 (2013), 184109.

[15] R. Erban, I. G. Kevrekidis, D. Adalsteinsson, and T. C. Elston, *Gene regulatory networks: A coarse-grained, equation-free approach to multiscale computation*, J. Chem. Phys., 124 (2006), 084106.

[16] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, *Systematic determination of order parameters for chain dynamics using diffusion maps*, Proc. Natl. Acad. Sci. USA, 107 (2010), pp. 13597–13602.

[17] P. M. Gallagher, A. L. Athayde, and C. F. Ivory, *The combined flux technique for diffusion–reaction problems in partial equilibrium: Application to the facilitated transport of carbon dioxide in aqueous bicarbonate solutions*, Chem. Eng. Sci., 41 (1986), pp. 567–578.

[18] C. W. Gear and I. G. Kevrekidis, *Telescopic projective methods for parabolic differential equations*, J. Comput. Phys., 187 (2003), pp. 95–109.

[19] S. Gepshtein and Y. Keller, *Image completion by diffusion maps and spectral relaxation*, IEEE Trans. Image Process., 22 (2013), pp. 2983–2994.

[20] S. Gerber, T. Tasdizen, and R. Whitaker, *Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps*, in Proceedings of the 24th International Conference on Machine Learning, ACM, New York, 2007, pp. 281–288.

[21] D. Givon, R. Kupferman, and A. Stuart, *Extracting macroscopic dynamics: Model problems and algorithms*, Nonlinearity, 17 (2004), pp. R55–R127.
[22] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, *Mori–Zwanzig formalism as a practical computational tool*, Faraday Discuss., 144 (2010), pp. 301–322.
[23] P. W. Jones, M. Maggioni, and R. Schul, *Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels*, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 1803–1808.
[24] J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics*, App. Math. Sci. 34, Springer-Verlag, New York, Berlin, 1981.
[25] I. G. Kevrekidis, C. W. Gear, and G. Hummer, *Equation-free: The computer-aided analysis of complex multiscale systems*, AIChE J., 50 (2004), pp. 1346–1355.
[26] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, and C. Theodoropoulos, *Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis*, Commun. Math. Sci., 1 (2003), pp. 715–762.
[27] I. G. Kevrekidis and G. Samaey, *Equation-free multiscale computation: Algorithms and applications*, Annu. Rev. Phys. Chem., 60 (2009), pp. 321–344.
[28] R. Z. Khasminskii, *On the principle of averaging the Itô's stochastic differential equations*, Kybernetika (Prague), 4 (1968), pp. 260–279.
[29] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. (N. Y.) 23, Springer-Verlag, Berlin, 1992.
[30] P. C. Mahalanobis, *On the generalized distance in statistics*, Proc. Natl. Inst. Sci. (Calcutta), 2 (1936), pp. 49–55.
[31] S. Mallat, *Group invariant scattering*, Comm. Pure Appl. Math., 65 (2012), pp. 1331–1398.
[32] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov, *Convergence of numerical time-averaging and stationary measures via Poisson equations*, SIAM J. Numer. Anal., 48 (2010), pp. 552–577, doi:10.1137/090770527.
[33] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed., Cambridge University Press, Cambridge, UK, 2009.
[34] H. Mori, *Transport, collective motion, and Brownian motion*, Progr. Theoret. Phys., 33 (1965), pp. 423–455.
[35] E. Pardoux and A. Yu. Veretennikov, *On the Poisson equation and diffusion approximation.* I, Ann. Probab., 29 (2001), pp. 1061–1085.
[36] G. A. Pavliotis and A. M. Stuart, *Parameter estimation for multiscale diffusions*, J. Statist. Phys., 127 (2007), pp. 741–781.
[37] A. J. Roberts, *Normal form transforms separate slow and fast modes in stochastic dynamical systems*, Phys. A, 387 (2008), pp. 12–38.
[38] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290 (2000), pp. 2323–2326.
[39] Y. Rubner, C. Tomasi, and L. J. Guibas, *The earth mover's distance as a metric for image retrieval*, Int. J. Comput. Vis., 40 (2000), pp. 99–121.
[40] L. A. Segel and M. Slemrod, *The quasi-steady-state assumption: A case study in perturbation*, SIAM Rev., 31 (1989), pp. 446–477, doi:10.1137/1031091.
[41] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, *Fisher vector faces in the wild*, in Proceedings of the British Machine Vision Conference, Vol. 1, Bristol, UK, 2013.
[42] A. Singer and R. R. Coifman, *Non-linear independent component analysis with diffusion maps*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 226–239.
[43] A. Singer, R. Erban, I. G. Kevrekidis, and R. R. Coifman, *Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps*, Proc. Natl. Acad. Sci., 106 (2009), pp. 16090–16095.
[44] V. Sotiropoulos, M.-N. Contou-Carrere, P. Daoutidis, and Y. N. Kaznessis, *Model reduction of multiscale chemical Langevin equations: A numerical case study*, IEEE/ACM Trans. Comput. Biol. Bioinform., 6 (2009), pp. 470–482.
[45] R. Talmon and R. R. Coifman, *Empirical intrinsic geometry for nonlinear modeling and time series filtering*, Proc. Natl. Acad. Sci., 110 (2013), pp. 12535–12540.
[46] R. Talmon and R. R. Coifman, *Intrinsic modeling of stochastic dynamical systems using empirical geometry*, Appl. Comput. Harmon. Anal., 35 (2014), pp. 138–160.

[47] R. Talmon, S. Mallat, H. Zaveri, and R. R. Coifman, *Manifold learning for latent variable inference in dynamical systems*, IEEE Trans. Signal Process., 63 (2015), pp. 3843–3856.
[48] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science, 290 (2000), pp. 2319–2323.
[49] A. B. Tumpach, *Gauge invariance of degenerate Riemannian metrics*, Notices Amer. Math. Soc., 63 (2016), pp. 342–350.
[50] A. A. Wheeler, W. J. Boettinger, and G. B. McFadden, *Phase-field model for isothermal phase transitions in binary alloys*, Phys. Rev. A, 45 (1992), pp. 7424–7439.
[51] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, *Distance metric learning with application to clustering with side-information*, in Advances in Neural Information Processing Systems 15 (NIPS 2002), MIT Press, Cambridge, MA, 2003, pp. 505–512.
[52] R. Zwanzig, *Memory effects in irreversible thermodynamics*, Phys. Rev., 124 (1961), p. 983–992.