Learning Stochastic Dynamical Systems via Bridge Sampling

Anonymous Author(s)
Affiliation
Address
email

Abstract

We develop algorithms to automate discovery of stochastic dynamical system models from noisy, vector-valued time series. By discovery, we mean learning both a nonlinear drift vector field and a diagonal diffusion matrix for an Itô stochastic differential equation in $\mathbb{R}^d$. We parameterize the vector field using tensor products of Hermite polynomials, enabling the model to capture highly nonlinear and/or coupled dynamics. We solve the resulting estimation problem using expectation maximization (EM). This involves two steps. We augment the data via diffusion bridge sampling, with the goal of producing time series observed at a higher frequency than the original data. With this augmented data, the resulting expected log likelihood maximization problem reduces to a least squares problem. Through experiments on systems with dimensions one through eight, we show that this EM approach enables accurate estimation for multiple time series with possibly irregular observation times. We study how the EM method performs as a function of the noise level in the data, the volume of data, and the amount of data augmentation performed.

1 Introduction

Traditional mathematical modeling in the sciences and engineering often has as its goal the development of equations of motion that describe observed phenomena. Classically, these equations of motion usually took the form of deterministic systems of ordinary or partial differential equations (ODE or PDE, respectively). Especially in systems of contemporary interest in biology and finance where intrinsic noise must be modeled, we find stochastic differential equations (SDE) used instead of deterministic ones. Still, these models are often built from first principles, after which the model's predictions (obtained, for instance, by numerical simulation) are compared against observed data.

Recent years have seen a surge of interest in using data to automate discovery of ODE, PDE, and SDE models. These machine learning approaches complement traditional modeling efforts, using available data to constrain the space of plausible models, and shortening the feedback loop linking model development to prediction and comparison to real observations. We posit two additional reasons to develop algorithms to learn SDE models. First, SDE models—including the models considered here—have the capacity to model highly nonlinear, coupled stochastic systems, including systems whose equilibria are non-Gaussian and/or multimodal. Second, SDE models often allow for interpretability. Especially if the terms on the right-hand side of the SDE are expressed in terms of commonly used functions (such as polynomials), we can obtain a qualitative understanding of how the system's variables influence, regulate, and/or mediate one other.

In this paper, we develop an algorithm to learn SDE models from high-dimensional time series. To our knowledge, this is the most general expectation maximization (EM) approach to learning an SDE with multidimensional drift vector field and diagonal diffusion matrix. Prior EM approaches were restricted to one-dimensional SDE [8], or used a Gaussian process approximation, linear drift approximation, and approximate maximization [22]. To develop our method, we use diffusion bridge sampling as in [12, 29], which focused on Bayesian nonparametric methods for SDE in $\mathbb{R}^1$. After augmenting the data using bridge sampling, we are left with a least-squares problem, generalizing the work of [6] from the ODE to the SDE context.

In the literature, variational Bayesian methods are the only other SDE learning methods that have been tested on high-dimensional problems [31]. These methods use approximations consisting of linear SDE with time-varying coefficients [1], kernel density estimates [2], or Gaussian processes [3]. In contrast, we parameterize the drift vector field using tensor products of Hermite polynomials; as mentioned above, the resulting SDE has much higher capacity than linear and/or Gaussian process models.

Many other techniques explored in the statistical literature focus on scalar SDE [4, 13, 14, 30].

As mentioned, differential equation discovery problems have attracted considerable recent interest. A variety of methods have been developed to learn ODE [6, 7, 17, 23, 25, 26, 28] as well as PDE [18, 19, 21, 24]. Unlike many of these works, we do not focus on model selection and/or regularization; if needed, our methods can be combined with model selection procedures developed in the ODE context [10, 11].

- results 1D, 2D, 3D damped duffing, 3D lorenz
- plots error of theta vs noise, error vs amount of data (number of data points) parametric curves for noise levels, brownian bridge plots for illustration, ...

2 Problem Setup

Let W_t denote Brownian motion in $\mathbb{R}^d$—informally, an increment dW_t of this process has a multivariate normal distribution with zero mean vector and covariance matrix $I dt$. Let X_t denote an $\mathbb{R}^d$-valued stochastic process that evolves according to the Itô SDE

$$dX_t = f(X_t)dt + \Gamma dW_t. \tag{1}$$

For rigorous definitions of Brownian motion and SDE, see [5, 32]. The nonlinear vector field $f\Omega \subset \mathbb{R}^d \to \mathbb{R}^d$ is the *drift* function, and the $d \times d$ matrix Γ is the *diffusion* matrix. To reduce the number of model parameters, we assume $\Gamma = \operatorname{diag} \gamma$.

Our goal is to develop an algorithm that accurately estimates the functional form of f and the vector γ from time series data.

Parameterization. We parameterize f using Hermite polynomials. The n-th Hermite polynomial takes the form

$$H_n(x) = (\sqrt{2\pi} n!)^{-1/2}(-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2} \tag{2}$$

Let $\langle f, g \rangle_w = \int_{\mathbb{R}} f(x)g(x) \exp(-x^2/2)\, dx$ denote a weighted L^2 inner product. Then, $\langle H_i, H_j \rangle_w = \delta_{ij}$, i.e., the Hermite polynomials are orthonormal with respect to the weighted inner product. In fact, with respect to this inner product, the Hermite polynomials form an orthonormal basis of $L^2_w(\mathbb{R}) = \{f \ \langle f, f \rangle_w < \infty\}$.

Now let $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}^d_+$ denote a multi-index. We use the notation $|\alpha| = \sum_j \alpha_j$ and $x^\alpha = \prod_j (x_j)^{\alpha_j}$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. For $x \in \mathbb{R}^d$ and a multi-index α, we also define

$$H_\alpha(x) = \prod_{j=1}^{d} H_{\alpha_j}(x_j). \tag{3}$$

We write $f(x) = (f_1(x), \ldots f_d(x))$ and then parameterize each component

$$f_j(x) = \sum_{m=0}^{M} \sum_{|\alpha|=m} \beta^j_\alpha H_\alpha(x). \tag{4}$$

We see that the maximum degree of $H_\alpha(x)$ is $|\alpha|$. Hence we think of the double sum in (4) as first summing over degrees and then summing over all terms with a fixed maximum degree. We say maximum degree because, for instance, $H_2(z) = (z^2 - 1)/(\sqrt{2\pi}2)^{1/2}$ contains both degree 2 and degree 0 terms.

There are $\binom{m+d-1}{d-1}$ possibilities for a d-dimensional multi-index α such that $|\alpha| = m$. Summing this from $m = 0$ to M, there are $\widetilde{M} = \binom{M+d}{d}$ total multi-indices in the double sum in (4). Let (i) denote the i-th multi-index according to some ordering. Then we can write

$$f_j(x) = \sum_{i=1}^{\widetilde{M}} \beta^j_{(i)} H_{(i)}(x). \tag{5}$$

Essentially, we parameterize f using tensor products of Hermite polynomials.

Data. We consider our data $\mathbf{x} = \{x_j\}_{j=0}^L$ to be direct observations of X_t at discrete points in time $\mathbf{t} = \{t_j\}_{t=0}^L$. Note that these time points do not need to be equispaced. In the derivation that follows, we will consider the data $(\mathbf{t}, \mathbf{x})$ to be one time series. Later, we indicate how our methods generalize naturally to multiple time series, i.e., repeated observations of the same system.

To achieve our estimation goal, we apply expectation maximization (EM). We regard $\mathbf{x}$ as the incomplete data. Let $\Delta t = \max_j (t_j - t_{j-1})$ be the maximum interobservation spacing. We think of the missing data $\mathbf{z}$ as data collected at a time scale $h \ll \Delta t$ fine enough such that the transition density of (1) is approximately Gaussian. To see how this works, let $\mathcal{N}(\mu, \Sigma)$ denote a multivariate normal with mean vector μ and covariance matrix Σ. Now discretize (1) in time via the Euler-Maruyama method with time step $h > 0$; the result is

$$\widetilde{X}_{n+1} = \widetilde{X}_n + f(\widetilde{X}_n)h + h^{1/2}\Gamma Z_{n+1}, \tag{6}$$

where $Z_{n+1} \sim \mathcal{N}(0, I)$ is a standard multivariate normal, independent of X_n. This implies that

$$(\widetilde{X}_{n+1} | \widetilde{X}_n = v) \sim \mathcal{N}(v + f(v)h, h\Gamma^2). \tag{7}$$

As h decreases, $\widetilde{X}_{n+1} | \widetilde{X}_n = v$—a Gaussian approximation—will converge to the true transition density $X_{(n+1)h} | X_{nh} = v$, where X_t refers to the solution of (1).

Diffusion Bridge. To augment or complete the data, we employ diffusion bridge sampling, using a Markov chain Monte Carlo (MCMC) method that goes back to [16, 20]. Let us describe our version here. We suppose our current estimate of $\boldsymbol{\theta} = (\beta, \gamma)$ is given. Define the diffusion bridge process to be (1) conditioned on both the initial value x_i at time t_i, and the final value x_{i+1} at time t_{i+1}. The goal is to generate sample paths of this diffusion bridge. By a sample path, we mean $F - 1$ *new* samples $\{z_{i,j}\}_{j=1}^{F-1}$ at times $t_i + jh$ with $h = (t_{i+1} - t_i)/F$.

To generate such a path, we start by drawing a sample from a Brownian bridge with the same diffusion as (1). That is, we sample from the SDE

$$d\widehat{X}_t = \Gamma dW_t \tag{8}$$

conditioned on $\widehat{X}_{t_i} = x_i$ and $\widehat{X}_{t_{i+1}} = x_{i+1}$. This Brownian bridge can be described explicitly

$$\widehat{X}_t = \Gamma(W_t - W_{t_i}) + x_i - \frac{t - t_i}{t_{i+1} - t_i}(\Gamma(W_{t_{i+1}} - W_{t_i}) + x_i - x_{i+1}) \tag{9}$$

Here $W_0 = 0$ (almost surely), and $W_t - W_s \sim \mathcal{N}(0, (t-s)I)$ for $t > s \geq 0$.

Let $\mathbb{P}$ denote the law of the diffusion bridge process, and let $\mathbb{Q}$ denote the law of the Brownian bridge (9). Using Girsanov's theorem [15], we can show that

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = C \exp\left(\int_{t_i}^{t_{i+1}} f(\widehat{X}_s)^T \Gamma^{-2} \, d\widehat{X}_s - \frac{1}{2} \int_{t_i}^{t_{i+1}} f(\widehat{X}_s)^T \Gamma^{-2} f(\widehat{X}_s) \, ds \right), \tag{10}$$

where the constant C depends only on x_i and x_{i+1}. The left-hand side is a Radon-Nikodym derivative, equivalent to a density or likelihood; the ratio of two such likelihoods is the accept/reject ratio in the Metropolis algorithm [27].

Putting the above pieces together yields the following Metropolis algorithm to generate diffusion bridge sample paths. Fix $F \geq 2$ and $i \in \{0, \ldots, L-1\}$. Assume we have stored the previous Metropolis step, i.e., a path $\mathbf{z}^{(\ell)} = \{z_{i,j}^{(\ell)}\}_{j=1}^{F-1}$.

1. Use (9) to generate samples of $\widehat{X}_t$ at times $t_i + jh$, for $j = 1, 2, \ldots, F-1$ and $h = (t_{i+1} - t_i)/F$. This is the proposal $\mathbf{z}^* = \{z_{i,j}^*\}_{j=1}^{F-1}$.
2. Numerically approximate the integrals in (10) to compute the likelihood of the proposal. Specifically, we compute
$$\begin{aligned} p(\mathbf{z}^*)/C = \sum_{j=0}^{F-1} f(z_{i,j}^*)^T \Gamma^{-2} (z_{i,j+1}^* - z_{i,j}^*) \\ - \frac{h}{4} \sum_{j=0}^{F-1} \left[f(z_{i,j}^*)^T \Gamma^{-2} f(z_{i,j}^*) + f(z_{i,j+1}^*)^T \Gamma^{-2} f(z_{i,j+1}^*) \right] \end{aligned}$$
We have discretized the stochastic $d\widehat{X}_s$ integral using Itô's definition, and we have discretized the ordinary ds integral using the trapezoidal rule.
3. Accept the proposal with probability $p(\mathbf{z}^*)/p(\mathbf{z}^{(\ell)})$—note the factors of C cancel. If the proposal is accepted, then set $\mathbf{z}^{(\ell+1)} = \mathbf{z}^*$. Else set $\mathbf{z}^{(\ell+1)} = \mathbf{z}^{(\ell)}$.

We initialize this algorithm with a Brownian bridge path, run for 10 burn-in steps, and then use subsequent steps as the diffusion bridge samples we seek.

Expectation Maximization (EM). Let us now give details to justify the intuition expressed above, that employing the diffusion bridge to augment the data on a fine scale will enable estimation. Let $\mathbf{z}^{(r)} = \{z_{i,j}^{(r)}\}_{j=1}^{F-1}$ be the r-th diffusion bridge sample path. We interleave this sampled data together with the observed data $\mathbf{x}$ to create the completed time series
$$\mathbf{y}^{(r)} = \{y_j^{(r)}\}_{j=1}^{N},$$
where $N = LF + 1$. By interleaving, we mean that $y_{1+iF}^{(r)} = x_i$ for $i = 0, 1, \ldots, L$, and that $y_{1+j+iF}^{(r)} = z_{i,j}$ for $j = 1, 2, \ldots, F-1$ and $i = 0, 1, \ldots, L-1$. With this notation, we can more easily express the EM algorithm. Let us assume that we currently have access to $\boldsymbol{\theta}^{(k)}$, our estimate of the parameters after k iterations. If $k = 0$, we set $\boldsymbol{\theta}^{(0)}$ equal to an initial guess. Then we follow two steps

1. For the expectation step, we first generate an ensemble of R diffusion bridge sample paths. Interleaving as above, this yields R completed time series $\mathbf{y}^{(r)}$ for $r = 1, \ldots, R$. In what follows, we will use an average over this ensemble to approximate the expected value. Let h_j denote the elapsed time between observations y_j and y_{j+1}. Using the completed data, the temporal discretization (6) of the SDE, the Markov property, and property (7), we have
$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \mathbb{E}_{\mathbf{z}|\mathbf{x},\boldsymbol{\theta}^{(k)}} [\log p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})] \qquad (11) \\ &\approx \frac{1}{R} \sum_{r=1}^{R} \log p(\mathbf{y}^{(r)} \mid \boldsymbol{\theta}) \\ &= \frac{1}{R} \sum_{r=1}^{R} \sum_{n=1}^{N-1} \log p(y_{n+1}^{(r)} \mid y_n^{(r)}, \boldsymbol{\theta}) \\ &= -\frac{1}{R} \sum_{r=1}^{R} \sum_{n=1}^{N-1} \left[\sum_{j=1}^{d} \frac{1}{2} \log(2\pi h_n \gamma_j^2) \right. \\ &\qquad \left. + \frac{1}{2h_n} \left\| \Gamma^{-1} \Big(y_{n+1}^{(r)} - y_n^{(r)} - h_n \sum_{\ell=1}^{\widetilde{M}} \beta_{(\ell)} H_{(\ell)} \big(y_n^{(r)} \big) \Big) \right\|_2^2 \right]. \qquad (12) \end{aligned}$$

2. For the M step, we maximize in stages

$$\beta^{(k+1)} = \arg\max_{\beta} Q((\beta, \gamma^{(k)}), \boldsymbol{\theta}^{(k)})$$
$$\gamma^{(k+1)} = \arg\max_{\gamma} Q((\beta^{(k+1)}, \gamma), \boldsymbol{\theta}^{(k)})$$

The maximization over β is a least squares problem. The solution is given by forming the matrix

$$\mathcal{M}_{k,\ell} = \frac{1}{R} \sum_{r=1}^{R} \sum_{j=1}^{N} h_j \phi_k^T(y_{j-1}^{(r)}) \Gamma^{-2} \phi_\ell^T(y_{j-1}^{(r)}) \tag{13}$$

and the vector

$$\rho_k = \frac{1}{R} \sum_{r=1}^{R} \sum_{j=1}^{N} \phi_k^T(y_{j-1}^{(r)}) \Gamma^{-2} (y_j^{(r)} - y_{j-1}^{(r)}). \tag{14}$$

We then solve the system $\mathcal{M}\beta = \rho$ for β. Now that we have β, we maximize over γ. The solution can be obtained in closed form

$$\gamma_i^2 = \frac{1}{RNh} \sum_{r=1}^{R} \sum_{j=1}^{N} ((y_j^{(r)} - y_{j-1}^{(r)} - h \sum_{\ell=1}^{M} \beta_\ell \phi_\ell(y_{j-1}^{(r)})) \cdot e_i)^2 \tag{15}$$

where e_i is the i^{th} canonical basis vector in $\mathbb{R}^d$.

We iterate the above two steps until $\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| / \|\boldsymbol{\theta}^{(k)}\| < \delta$ for some tolerance $\delta > 0$.

When the data consists of multiple time series $\{\mathbf{t}^{(i)}, \mathbf{x}^{(i)}\}_{i=1}^{S}$, everything scales accordingly. For instance, we create an ensemble of R diffusion bridge samples for each of the S time series. If we index the resulting completed time series appropriately, we simply replace R by RS in (13), (14), and (15) and keep everything else the same.

There are three sources of error in the above algorithm. The first relates to replacing the expectation by a sample average; the induced error should, by the law of large numbers, decrease as $R^{-1/2}$. The second stems from the approximate nature of the computed diffusion bridge samples—as indicated above, we use numerical integration to approximate the Girsanov likelihood. The third source of error is in using the Gaussian transition density to approximate the true transition density of the SDE. Both the second and third sources of error vanish in the $F \to \infty$ limit [9].

3 Experiments

We present a series of increasingly higher-dimensional experiments with synthetic data. To generate this data, we start with a known stochastic dynamical system of the form (1). Using Euler-Maruyama time stepping starting from a randomly chosen initial condition, we march forward in time from $t = 0$ to a final time $t = T$. Here T is problem-specific; for the one-dimensional example, $T = 10$.

In all examples, we use a fine internal time step to generate the data, but we *save* the data at a much coarser time scale. For instance, in the one-dimensional example, we step forward internally at a time step of $h = 0.0001$, but we save the data at increments of 0.01 units of time, essentially discarding 99% of the simulated trajectory. We use a fine internal time step to reduce, to the extent possible, numerical error in the simulated data. We save the data on a coarse time scale to test the data augmentation method proposed in this paper. In all examples, we choose initial conditions so that simulated trajectories remain bounded.

To study how the EM method performs as a function of noise strength, data volume, and data augmentation, we perform four sets of experiments. When we run EM, we randomly generate the initial guess $\beta^{(0)} \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.5)$. We set the EM tolerance parameter $\delta = 0.01$. The only regularization we include is to threshold β—values less than 0.01 in absolute value are reset to zero. Finally, in the MCMC diffusion bridge sampler, we use 10 burn-in steps and then create an ensemble of size $R = 100$.

To quantify error, we use the Frobenius norm of the difference between estimated $\widetilde{\beta}$ and true β matrices

$$\varepsilon = \sqrt{\sum_i \|\beta_{(i)} - \widetilde{\beta}_{(i)}\|^2} \tag{16}$$

The $\widetilde{\beta}$ coefficients are the Hermite coefficients of the estimated drift vector field f. For each example system, we compute the true Hermite coefficients β by multiplying the true ordinary polynomial coefficients by a change-of-basis matrix that is easily computed.

We test the method using stochastic systems in dimensions $d = 1, 2, 3$. For the one-dimensional system, we use

$$dX_t = (-1 + X_t + X_t^2)dt + \gamma dW_t.$$

In two dimensions, we use a stochastic Duffing oscillator with no damping or driving:

$$\begin{aligned} dX_{1,t} &= X_{2,t}dt + \gamma_1 dW_{1,t} \\ dX_{2,t} &= (-X_{1,t} - X_1^3(t))dt + \gamma_2 dW_{2,t} \end{aligned}$$

For the three-dimensional case, we consider two different systems. The first is a stochastic, damped, driven Duffing oscillator:

$$\begin{aligned} dX_{1,t} &= X_{2,t}dt + \gamma_1 dW_{1,t} \\ dX_{2,t} &= (X_{1,t} - X_{1,t}^3 - 0.3X_{2,t} + 0.5\cos(X_{3,t}))dt + \gamma_2 dW_{2,t} \\ dX_{3,t} &= 1.2dt + \gamma_3 dW_{3,t} \end{aligned}$$

The second is the stochastic Lorenz system

$$\begin{aligned} dX_{1,t} &= 10(X_{2,t} - X_{1,t})dt + \gamma_1 dW_{1,t} \\ dX_{2,t} &= (X_{1,t}(28 - X_{3,t}))dt + \gamma_2 dW_{2,t} \\ dX_{3,t} &= (X_{1,t}X_{2,t} - (8/3)X_{3,t})dt + \gamma_3 dW_{3,t} \end{aligned}$$

In what follows, we refer to these systems as the $1d$, $2d$, Duffing, and Lorenz systems.

Experiment 1: Varying Number of Time Series. Here we vary data volume by stepping the number S of time series from $S = 1$ to $S = 10$. Each time series has length $L + 1 = 101$.

Experiment 2: Varying Length of Time Series. Here we vary data volume by stepping the length $L + 1$ of the time series from $L + 1 = 11$ to $L + 1 = 101$, keeping the number of time series fixed at $S = 10$.

Experiment 3: Varying Noise Strength. Here we vary the noise strength γ, stepping from 0.5 to 0.0001 while keeping other parameters constant. Specifically, we take $S = 10$ time series each of length $L + 1 = 101$.

Experiment 4: Varying Data Augmentation. Here we vary the number F of interleaved diffusion bridge samples from $F = 1$ to $F = 9$. Note that for $F = 1$, no diffusion bridge is created; the likelihood is computed by applying the Gaussian transition density directly to the observed data.

References

[1] C. Archambeau, M. Opper, Y. Shen, D. Cornford, and J. S. Shawe-taylor. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, pages 17–24, 2008.

[2] P. Batz, A. Ruttor, and M. Opper. Variational estimation of the drift for stochastic differential equations from the empirical density. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(8):083404, Aug. 2016. ISSN 1742-5468. doi: 10.1088/1742-5468/2016/08/083404. URL `http://stacks.iop.org/1742-5468/2016/i=8/a=083404?key=crossref.4719755181cc98b942ea066bb4d58264`.

[3] P. Batz, A. Ruttor, and M. Opper. Approximate Bayes learning of stochastic differential equations. *arXiv preprint arXiv:1702.05390*, 2017. URL `https://arxiv.org/abs/1702.05390`.

[4] H. S. Bhat and R. W. M. A. Madushani. Nonparametric Adjoint-Based Inference for Stochastic Differential Equations. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 798–807, Oct. 2016. doi: 10.1109/DSAA.2016.69.

[5] R. N. Bhattacharya and E. C. Waymire. *Stochastic Processes with Applications*. SIAM, Aug. 2009. ISBN 978-0-89871-689-4. Google-Books-ID: 89dZjIid6XcC.

[6] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113 (15):3932–3937, Apr. 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517384113. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1517384113`.

[7] S. Chen, A. Shojaie, and D. M. Witten. Network Reconstruction From High-Dimensional Ordinary Differential Equations. *Journal of the American Statistical Association*, 112(520): 1697–1707, Oct. 2017. ISSN 0162-1459. doi: 10.1080/01621459.2016.1229197. URL `https://doi.org/10.1080/01621459.2016.1229197`.

[8] Z. Ghahramani and S. T. Roweis. Learning nonlinear dynamical systems using an EM algorithm. *Advances in Neural Information Processing Systems (NIPS)*, pages 431–437, 1999. URL `https://papers.nips.cc/paper/1594-learning-nonlinear-dynamical-systems-using-an-em-algorithm.pdf`.

[9] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Science & Business Media, June 2011. ISBN 978-3-540-54062-5. Google-Books-ID: BCvtssom1CMC.

[10] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):52–63, June 2016. doi: 10.1109/TMBMC.2016.2633265.

[11] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proc. R. Soc. A*, 473(2204): 20170009, Aug. 2017. ISSN 1364-5021, 1471-2946. doi: 10.1098/rspa.2017.0009. URL `http://rspa.royalsocietypublishing.org/content/473/2204/20170009`.

[12] F. v. d. Meulen, M. Schauer, and J. v. Waaij. Adaptive nonparametric drift estimation for diffusion processes using Faber–Schauder expansions. *Statistical Inference for Stochastic Processes*, pages 1–26, June 2017. ISSN 1387-0874, 1572-9311. doi: 10.1007/s11203-017-9163-7. URL `https://link.springer.com/article/10.1007/s11203-017-9163-7`.

[13] H.-G. Müller, F. Yao, and others. Empirical dynamics for longitudinal data. *The Annals of Statistics*, 38 (6):3458–3486, 2010. URL `http://projecteuclid.org/euclid.aos/1291126964`.

[14] J. Nicolau. NONPARAMETRIC ESTIMATION OF SECOND-ORDER STOCHASTIC DIFFERENTIAL EQUATIONS. *Econometric Theory*, 23(05):880, Oct.

2007. ISSN 0266-4666, 1469-4360. doi: 10.1017/S0266466607070375. URL http://www.journals.cambridge.org/abstract_S0266466607070375.

[15] O. Papaspiliopoulos and G. Roberts. Importance sampling techniques for estimation of diffusion models. *Statistical methods for stochastic differential equations*, 124:311–340, 2012.

[16] O. Papaspiliopoulos, G. O. Roberts, and O. Stramer. Data Augmentation for Diffusions. *Journal of Computational and Graphical Statistics*, 22(3):665–688, July 2013. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2013.783484. URL http://www.tandfonline.com/doi/abs/10.1080/10618600.2013.783484.

[17] M. Quade, M. Abel, J. N. Kutz, and S. L. Brunton. Sparse Identification of Nonlinear Dynamics for Rapid Model Recovery. *arXiv:1803.00894 [nlin, physics:physics]*, Mar. 2018. URL http://arxiv.org/abs/1803.00894. arXiv: 1803.00894.

[18] M. Raissi and G. E. Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357: 125–141, Mar. 2018. ISSN 0021-9991. doi: 10.1016/j.jcp.2017.11.039. URL http://www.sciencedirect.com/science/article/pii/S0021999117309014.

[19] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348: 683–693, Nov. 2017. ISSN 0021-9991. doi: 10.1016/j.jcp.2017.07.050. URL http://www.sciencedirect.com/science/article/pii/S0021999117305582.

[20] G. O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika*, 88(3):603–621, Oct. 2001. ISSN 0006-3444. doi: 10.1093/biomet/88.3.603. URL https://academic.oup.com/biomet/article/88/3/603/340094.

[21] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, Apr. 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1602614. URL http://advances.sciencemag.org/content/3/4/e1602614.

[22] A. Ruttor, P. Batz, and M. Opper. Approximate Gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems*, pages 2040–2048, 2013. URL http://papers.nips.cc/paper/4967-approximate-gaussian-process-inference-for-the-drift-function-in-stoc

[23] H. Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 473 (2197):20160446, Jan. 2017. ISSN 1364-5021, 1471-2946. doi: 10.1098/rspa.2016.0446. URL http://rspa.royalsocietypublishing.org/lookup/doi/10.1098/rspa.2016.0446.

[24] H. Schaeffer, R. Caflisch, C. D. Hauck, and S. Osher. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences*, 110(17):6634–6639, Apr. 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1302752110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1302752110.

[25] H. Schaeffer, G. Tran, and R. Ward. Extracting Sparse High-Dimensional Dynamics from Limited Data. *arXiv:1707.08528 [math]*, July 2017. URL http://arxiv.org/abs/1707.08528. arXiv: 1707.08528.

[26] T. B. Schön, A. Svensson, L. Murray, and F. Lindsten. Probabilistic learning of nonlinear dynamical systems using sequential Monte Carlo. *arXiv preprint arXiv:1703.02419*, 2017. URL https://arxiv.org/abs/1703.02419.

[27] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, May 2010. ISSN 0962-4929, 1474-0508. doi: 10.1017/S0962492910000061. URL http://www.journals.cambridge.org/abstract_S0962492910000061.

[28] G. Tran and R. Ward. Exact Recovery of Chaotic Systems from Highly Corrupted Data. *Multiscale Modeling & Simulation*, 15(3):1108–1129, Jan. 2017. ISSN 1540-3459. doi: 10.1137/16M1086637. URL https://epubs.siam.org/doi/abs/10.1137/16M1086637.

[29] F. van der Meulen, M. Schauer, and H. van Zanten. Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *Computational Statistics & Data Analysis*, 71:615–632, Mar. 2014. ISSN 0167-9473. doi: 10.1016/j.csda.2013.03.002. URL http://www.sciencedirect.com/science/article/pii/S016794731300090X.

[30] N. Verzelen, W. Tao, H.-G. Müller, and others. Inferring stochastic dynamics from functional data. *Biometrika*, 99(3):533–550, 2012. URL http://nicolas.verzelen.free.fr/pdf/2012-Ver-Biometrika.pdf.

[31] M. D. Vrettas, M. Opper, and D. Cornford. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. *Physical Review E*, 91(1):012148, Jan. 2015. doi: 10.1103/PhysRevE.91.012148. URL https://link.aps.org/doi/10.1103/PhysRevE.91.012148.

[32] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer-Verlag, Berlin Heidelberg, 6 edition, 2003. ISBN 978-3-540-04758-2. URL //www.springer.com/us/book/9783540047582.