Journal of Machine Learning Research 24 (2023) 1-39 Submitted 5/22; Revised 3/23; Published 4/23

Infinite-dimensional optimization and Bayesian nonparametric learning of stochastic differential equations

Arnab Ganguly AGANGULY@LSU.EDU
Department of Mathematics
Louisiana State University
Baton Rouge, LA 70820, USA

Riten Mitra RITENDRANATH.MITRA@LOUISVILLE.EDU
Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY 40202, USA

Jinpu Zhou ZJINPU1@LSU.EDU
Department of Mathematics
Louisiana State University
Baton Rouge, LA 70820, USA

Editor: Ryan Adams

Abstract

[1] The paper has two major themes. The first part of the paper establishes certain general results for infinite-dimensional optimization problems on Hilbert spaces. These results cover the classical representer theorem and many of its variants as special cases and offer a wider scope of applications. The second part of the paper then develops a systematic approach for learning the drift function of a stochastic differential equation by integrating the results of the first part with Bayesian hierarchical framework. Importantly, our Bayesian approach incorporates low-cost sparse learning through proper use of shrinkage priors while allowing proper quantification of uncertainty through posterior distributions. Several examples at the end illustrate the accuracy of our learning scheme.

Keywords: Reproducing kernel Hilbert spaces (RKHS), infinite-dimensional optimization, representer theorem, nonparametric learning, stochastic differential equations, diffusion processes, Bayesian methods

1. Introduction

The temporal dynamics of a variety of systems arising from systems biology, environmental science, engineering, physics, medicine can be captured by stochastic differential equations (SDEs) driven by appropriate drift and noise functions (c.f (3.1)). SDEs are also central to modern financial mathematics where they are used to model short term interest rates, asset and options pricing, their volatility. Understanding behaviors of these systems requires not just building mathematical models but integrating it with the available data. For instance, advanced technologies like single-cell imaging can attest to the stochasticity of cellular processes (Friedman et al. (2010); Coulon et al. (2013)). While this molecular noise is a rich source of information about the process dynamics, utilizing this source in a systematic manner requires building stochastic temporal models that

1. Authors contributed equally.

are calibrated according to the available data. Building such data-driven models characterizing the inner-workings of these systems is instrumental for advancement of quantitative biology and other quantitative disciplines.

There is a substantial volume of research on both theoretical and computational aspects of parametric SDE models and its statistical inference, a very limited list of references for which is Elerian et al. (2001); Roberts and Stramer (2001); Kutoyants (2004); Golightly and Wilkinson (2005); Bishwal (2008); Golightly and Wilkinson (2008); Archambeau and Opper (2011); Cseke et al. (2013); Beskos et al. (2006); Fearnhead et al. (2008); Sutter et al. (2016); Whitaker et al. (2017). Bishwal (2002, 2018, 2022b) analyzed asymptotic properties of statistical estimators for SDEs driven by fractional Brownian Motion and certain stochastic PDE models. Parameter estimation for stochastic volatility models are studied in Bishwal (2022a) Specifically, for these models the driving functions of the SDE are assumed to be known barring a finite-dimensional parameter θ, which then needs to be estimated from the available data. In reality, for a large class of physical systems functional or parametric forms of the underlying SDEs are not precisely known. However, to get a workable mathematical model a heavy set of assumptions is usually imposed on the system which in many cases is not practical — the resulting model might be too simplistic and might only work in certain ideal situations. For example, in biochemical systems, under a set of assumptions including spatial homogeneity, the intensity function of each reaction driving the stochastic dynamics is assumed to be of the form of a known polynomial function multiplied by the corresponding reaction rate constant (unknown parameter). Model calibration then requires estimation of these reaction rates from the given data (e.g. see Golightly and Wilkinson (2006); Boys et al. (2008); Golightly and Wilkinson (2011); Koeppl et al. (2012)). However, most cellular reactions do not occur in spatially homogeneous environments. Moreover there are often many unknown factors (e.g, undiscovered reactions or species) affecting the reaction rates – assuming that they are constants despite these can lead to simplistic models which might not be able to explain observed behavior of these systems satisfactorily. This highlights the importance of developing truly data-driven models, where some of the key driving functions (like b, σ) for a complex dynamical system of the form (3.1) are learnt entirely from the given data.

This, however, is an infinite-dimensional learning problem! Compared to the parametric case, very little is available in the literature for these nonparametric stochastic models. Most of the research in the area of machine learning and 'traditional' nonparametric statistics focus on regression or classification analysis involving i.i.d data points, which are comparatively much easier to work with. For stochastic dynamical systems that we are interested in, there exist some histogram based approaches using bins of size ϵ around each location x and computing appropriate local means in those bins (Friedrich et al. (2011)). Further refinements include replacing the bins with means of k-nearest neighbor (Hegger and Stock (2009)) and use of traditional Nadaraya-Watson type estimates (Lamouroux and Lehnertz (2009)). These methods unfortunately only work for a limited number of toy systems and require high number of data-points around each x. Some approaches involving Gaussian Process have also been used (see Ruttor et al.; Yildiz et al. (2018)), but they often rely on adhoc approximation including linearization which might not be desirable.

The present paper along with related future projects aims to develop a systematic Bayesian framework for addressing these types of complex problems. The data for these problems can come in a wide array of formats — ranging from a single path observed at high frequency to noisy partial observations observed at sparse times. This article is the first in the series of ongoing and

planned papers (Ganguly et al. (a,b)) that aims to develop learning schemes for these different data settings. This article specifically focuses on learning of the drift function of SDEs in the case of high frequency data by which we mean that it is of the form of a single discrete path $\{X(t_i) : i = 1, 2 \ldots, m\}$ where the gap $t_i - t_{i-1}$ between two successive observation times t_i and t_{i-1} is very small. Our first step toward estimating the driving functions of the SDE is to consider the problem of minimization of the negative log-likelihood subject to a penalty function over an appropriate function space. Reproducing kernel Hilbert spaces (RKHS) are most suitable function spaces for these kinds of infinite-dimensional optimization problems because of the well-known representer theorem which often converts a class of such problems into finite-dimensional ones. However the limitation of the representer theorem is that it requires the loss functional, $L(h)$, to depend on the input function h only through its values, $h(x_i)$, at a finite number of data points $\{x_i\}$ which makes it or its known variants inapplicable in many important cases.

This issue is addressed in the first part of the paper (Section 2), which studies infinite-dimensional optimization problems in a broader framework and proves certain general results (see Theorem 6 and its corollaries), special cases of which give the representer theorem on RKHS. Results of Section 2 should be of independent interest and are expected to find wider applications. The full generality of Theorem 6 is crucial in our upcoming papers involving more general stochastic models; in the current paper, only a slightly generalized version of the representer theorem is needed and it gives a representation of the minimizer of the penalized negative log-likelihood in an RKHS as a finite-sum with respect to the basis-functions, $\boldsymbol{\kappa}(\cdot, X(t_i))$, where $\boldsymbol{\kappa}$ is the associated kernel of the RKHS. We next develop a Bayesian hierarchical framework for estimating the coefficients of this finite-sum representation by putting appropriate prior distributions on them. The primary advantage of the Bayesian approach over point-optimization methods (like gradient descent) is the proper quantification of uncertainty through the posterior distributions of the estimators. Now the number of terms in this finite-sum expansion increases proportionately with the number of data points. It is therefore imperative that sparse learning is incorporated to reduce the complexity of the estimators. In our Bayesian paradigm, this is induced through proper shrinkage priors, and in this paper we employ a multivariate t-prior and an extension of Horseshoe like priors for this purpose. The interplay of shrinkage priors and the SDE dynamics is interesting to note. Shrinkage priors are effective in case of positive recurrence which forces the SDE to revisit the relevant parts of the state space numerous times over a finite time horizon. This implies that not all of the basis functions $\boldsymbol{\kappa}(\cdot, X(t_i))$ are needed in the finite-sum expansion of the estimator of the drift function; only a limited selection is enough for accuracy, and proper shrinkage priors help to identify this selection. The use of shrinkage priors in the context of SDEs is novel and to the best of our knowledge has not been studied before.

The layout of the article is as follows. Section 2 studies optimization problem in the setting of a general Hilbert space. Section 3 introduces the SDE model and formulates the Bayesian framework with shrinkage priors for learning the drift function. The learning algorithms are also presented. Numerical examples are discussed in Section 4. Finally, some concluding remarks can be found in Section 5.

Notation: $\mathbb{R}^{m\times n}$ denotes the space of $m \times n$ real matrices. $\text{vec}_{m\times n} : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn}$ will denote the vectorization function for $m \times n$ matrices. For a matrix $A \in \mathbb{R}^{m\times n}$, $A(i, *)$ and $A(*, j)$ respectively denote the i-th row and j-th column of A. For two Hilbert (or Banach) spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, $L(\mathcal{H}_1, \mathcal{H}_2)$ denotes the space of linear bounded operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. $\mathcal{H}_1 \oplus_e \mathcal{H}_2$ will denote

the external direct sum of $\mathcal{H}_1$ and $\mathcal{H}_2$. $\mathcal{N}_d(\mu, \Sigma)$ will refer to the d-dimensional Normal distribution with mean μ and covariance matrix Σ, and for notational convenience $\mathcal{N}_d(\cdot|\mu, \Sigma)$ will denote the corresponding density function. Similar convention will be followed for other named distributions:

- $\mathrm{t}_d(\nu, \mu, V)$: d-dimensional t-distribution with degrees of freedom ν, mean μ and scale matrix V; $\mathrm{t}_d(\cdot|\nu, \mu, V)$: corresponding density function (c.f (3.7)).
- $\mathcal{G}(a, b),\ \mathcal{IG}(a, b)$: Gamma and Inverse Gamma distributions with parameters a and b; $\mathcal{G}(\cdot|a, b),\ \mathcal{IG}(\cdot|a, b)$: corresponding density functions.
- $\mathcal{W}_d(\nu, V),\ \mathcal{IW}_d(\nu, V)$: d-dimensional Wishart and Inverse-Wishart distributions with degrees of freedom $\nu > d - 1$ and scale matrix V; $\mathcal{W}_d(\cdot|\nu, V)$ and $\mathcal{IW}_d(\cdot|\nu, V)$: the corresponding density functions.
- $\mathcal{F}(\nu_1, \nu_2, c)$: $\mathcal{F}$-distribution with degrees of freedom ν_1, ν_2 and scaling parameter c; $\mathcal{F}(\cdot|\nu_1, \nu_2, c)$: the corresponding density function (c.f (3.8)).

2. Infinite-dimensional optimization

The main goal of the section is to characterize the solutions of a broad class of minimization problems which in particular will generalize the classical representer theorem in a significant way.

2.1 Classical representer theorem and its limitations

The representer theorem is a seminal result in learning theory which converts a class of infinite-dimensional optimization problems on an RKHS to a tractable finite-dimensional one. It was first derived in Kimeldorf and Wahba (1971) for quadratic loss and penalty functions in the setting of Chebyshev splines and was later extended to more general RKHS framework in Wahba (1990). Extensions to more general loss and penalty functions have been done in Cox and O'Sullivan (1990); Schölkopf et al. (2001) (also see Scholkopf and Smola (2001)). We first recall its statement.

Let $\mathcal{H}$ be a Hilbert space, $\mathbb{U}$ an aribitrary space, and $\kappa : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ a symmetric kernel satisfying (a) for each $u \in \mathbb{U}$, $\kappa(u, \cdot) \in \mathcal{H}$, and (b) for every $h \in \mathcal{H}$ and $u \in \mathbb{U}$, $\langle h, \kappa(u, \cdot)\rangle = h(u)$. Property (b) refers to the reproducing property of the kernel κ, and such a kernel κ is called a *reproducing kernel* and the associated Hilbert space is called an ***RKHS***, which is also unique for a given reproducing kernel κ.

What makes RKHS particularly useful in learning theory is the fact that one can construct an RKHS $\mathcal{H}_\kappa$ starting from a positive definite kernel κ. This is possible due to Moore-Aronszajn theorem which says that the space $\mathcal{H}_\kappa \equiv \overline{\mathrm{Span}}\{\kappa(\cdot, u) : u \in \mathbb{U}\}$ is the RKHS corresponding to the kernel κ. Here the overbar denotes closure of a set, and the closure is taken with the norm, $\|\cdot\|_\kappa$ defined by

$$\|h\|_\kappa = \sum_{i,j=1}^{l} c_i c_j \kappa(u_i, u_j), \quad h = \sum_{j=1}^{l} \kappa(\cdot, u_j) c_j,\ c_j \in \mathbb{R}.$$

The following version of the representer theorem is adopted from Schölkopf et al. (2001).

Theorem 1 *Let $L : (\mathbb{U} \times \mathbb{R}^2)^m \to [-\infty, \infty]$ be any function, and $J : [0, \infty) \to [0, \infty)$ strictly increasing, and $\kappa : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$ a reproducing kernel. Let $\mathcal{H}_\kappa$ be the RKHS of functions $h : \mathbb{U} \to \mathbb{R}$*

corresponding to κ. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \in \mathbb{U} \times \mathbb{R}$ be fixed. Then any h^* minimizing the objective function

$$L(h) \equiv L\left((x_1, y_1, h(x_1)), (x_2, y_2, h(x_2)), \ldots, (x_m, y_m, h(x_m))\right) + J(\|h\|), \quad h \in \mathcal{H}_\kappa$$

admits a representation of the form $h^*(u) = \sum_{k=1}^m c_k^* \kappa(x_k, u)$ with $c_k^* \in \mathbb{R}$.

The representer theorem is a potent tool in learning theory as it provides a computable expression of the minimizer of a class of loss functionals as a finite sum expansion with respect to the concrete basis functions involving the underlying kernel. One of the post popular applications is estimation of the function h of a regression model, $y = h(x) + \varepsilon$, from the data points $\{(x_i, y_i)\}$. But as already noted in the Introduction one constraint in the representer theorem is the requirement that the loss functional L depends on its argument function h only through $h(x_i),\ i = 1, 2, \ldots, m$ — its values at the data points $\{x_i\}$. This inhibits its direct applications to models where the dependence on the loss functional on h is more intricate. We illustrate this with a couple of examples.

Example 1 (Linear functional regression) Consider the model

$$y = \mathcal{L}_x h + \varepsilon, \tag{2.1}$$

where for each x, $\mathcal{L}_x$ is a linear functional acting on h, and ε captures the noise of the system. Thus here the function h is observed (with errors) through a family of linear functionals. For example, consider the regression model, $Y = h(Z) + \varepsilon$, where Z is not directly observed. Instead for a third random variable X, the conditional distribution of $Z|X = x$, $\gamma(\cdot|x)$, is known (or at least can be well approximated). Integrating the effect of Z, the conditional model of Y given X is of the form (2.1), where for a given x, $\mathcal{L}_x h = \int h(u)\gamma(du|x)$.

Given data points $\{(x_i, y_i) : i = 1, 2, \ldots, m\}$, the natural approach to learn h is again through the minimization problem of the form

$$\min_{h \in \mathcal{H}_\kappa} L\left((x_1, y_1, \mathcal{L}_{x_1} h), (x_2, y_2, \mathcal{L}_{x_2} h), \ldots, (x_m, y_m, \mathcal{L}_{x_m} h)\right) + J(\|h\|). \tag{2.2}$$

It is clear that the classical representer theorem cannot be applied here directly as the loss function does not depend on h through the values $h(x_i)$, but it does so through certain integrals of h.

Example 2 (Fredholm integral equation of first kind) Consider the Fredholm equation of the first kind: $g(x) = \int_{\mathcal{E}} R(x, u)h(u)du$, where $\mathcal{E} \subset \mathbb{R}^d$. Here given (possibly noisy) values, y_i, of g at finitely many points x_i, the goal is to learn the best possible function h. The data generating models is thus of the form

$$y_i = \int_{\mathcal{E}} R(x_i, y)h(y)dy + \epsilon_i, \quad i = 1, 2, \ldots, m.$$

where ϵ_i captures the noise in the observations. As before, learning of h requires one to consider the minimization problem of the form

$$\min_{h \in \mathcal{H}_\kappa} L\left((x_1, y_1, Rh(x_1)), (x_2, y_2, Rh(x_2)), \ldots, (x_m, y_m, Rh(x_m))\right) + J(\|h\|), \tag{2.3}$$

where, by a slight abuse of notation, R also denotes the operator / integral transform corresponding to the kernel $R(\cdot, \cdot)$; that is, $Rh(x) = \int_{\mathcal{E}} R(x, u)h(u)du$. Note again that the classical representer theorem is not applicable as L depends on h not through its values $h(x_i)$ but through the above integrals.

2.2 A broad class of optimization problems in Hilbert space

Let $F : \mathcal{H} \times [0, \infty) \to \mathbb{R}$. We are interested in the minimization problem

$$\min_{h \in \mathcal{H}} F\left(h, \langle Qh, h\rangle^{1/2}\right), \tag{2.4}$$

where $Q \in L(\mathcal{H}, \mathcal{H})$ is a self-adjoint, positive semidefinite (p.s.d.) continuous linear operator. Notice that this class of minimization problems is equal to the class of problems of the type $\min_{h \in \mathcal{H}} F(h, \|Rh\|)$, where $R \in L(\mathcal{H}, \mathcal{H})$. It is useful to note here that by the Hellinger-Toeplitz theorem (or simply by the closed graph theorem) if $Q : \mathcal{H} \to \mathcal{H}$ is a self-adjoint linear operator with $\mathrm{Dom}(Q) = \mathcal{H}$, then Q has to be continuous, that is, $Q \in L(\mathcal{H}, \mathcal{H})$.

Recall that $Q \in L(\mathcal{H}, \mathcal{H})$ is *positive* or *positive semi-definite* (p.s.d.) if $\langle Qh, h\rangle \geqslant 0$ for any $h \neq 0$, *positive definite* (p.d.) if the previous inequality is strict for all $h \neq 0$, and *uniformly positive definite* (uniformly p.d.) if there exists a $\lambda > 0$ such that $\langle Qh, h\rangle \geqslant \lambda \|h\|^2$ for all $h \in \mathcal{H}$. If H_0 is a subset of $\mathcal{H}$, then the restriction of Q to H_0, $Q\big|_{H_0}$, is p.s.d. (p.d.) if $\langle Qh, h\rangle \geqslant 0$ (> 0) for any $0 \neq h \in H_0$, and *uniformly p.d.* if for some $\lambda > 0$, $\langle Qh, h\rangle \geqslant \lambda \|h\|^2$ for all $h \in H_0$. Lemma 1 in the Appendix is useful in characterizing uniformly p.d. operators.

For $Q \in L(\mathcal{H}, \mathcal{H})$, define

$$\mathcal{N}_Q = \{h \in \mathcal{H} : \langle Qh, h\rangle = 0\}. \tag{2.5}$$

Clearly, if Q is self-adjoint and p.s.d., $\mathcal{N}_Q$ is a closed subspace of $\mathcal{H}$, and a p.s.d. operator Q is p.d. if and only if $\mathcal{N}_Q = \{0\}$. Further note that if $\mathcal{M}$ is a subspace of $\mathcal{H}$, then $Q\big|_{\mathcal{M}}$ is p.d. if $\mathcal{N}_Q \cap \mathcal{M} = \{0\}$. When Q is p.s.d., $h \to \langle h, Qh\rangle^{1/2}$ defines a seminorm; it is a proper norm when Q is p.d., in which case we write $\|h\|_Q \equiv \langle h, Qh\rangle^{1/2}$. $\|\cdot\|_Q$ is equivalent to the original $\|\cdot\|$ norm if and only if Q is uniformly p.d.

By a solution to the problem (2.4) we will mean a *(global) minimizer* $h^* \in \mathcal{H}$ such that

$$F\left(h^*, \langle Qh^*, h^*\rangle^{1/2}\right) = \inf_{h \in \mathcal{H}} F\left(h, \langle Qh, h\rangle^{1/2}\right) \stackrel{def}{=} F^*.$$

In contrast, an element $h_0 \in \mathcal{H}$ is a *local minimizer* of the problem (2.4) if there exists a $r > 0$, such that $F\left(h_0, \langle Qh_0, h_0\rangle^{1/2}\right) = \inf_{h \in B(h_0, r)} F\left(h, \langle Qh, h\rangle^{1/2}\right)$. Here $B(h_0, r)$ is the open ball in $\mathcal{H}$ with center at h_0 and radius r.

Lower semicontinuity (l.s.c.) plays an important role in the solution of a minimization problem. Since there are different notions of l.s.c. in a Hilbert space, we first recall their definitions.

Definition 2 *(i) A function $G : \mathcal{H} \to [-\infty, \infty]$ is said to be strongly lower-semicontinuous (l.s.c.) or l.s.c. in the norm topology if $\liminf_{n\to\infty} G(h_n) \geqslant G(h)$, whenever $h_n \to h$ (in $\mathcal{H}$-norm); or equivalently, the sublevel sets $\{h : G(h) \leqslant a\}$ are closed in the norm topology of $\mathcal{H}$.*

(ii) A function $G : \mathcal{H} \to [-\infty, \infty]$ is said to be weakly sequentially l.s.c. if $\liminf_{n\to\infty} G(h_n) \geqslant G(h)$, whenever $h_n \stackrel{w}{\to} h$, or equivalently, the sublevel sets $\{h : G(h) \leqslant a\}$ are weakly sequentially closed.

(iii) A function $G : \mathcal{H} \to [-\infty, \infty]$ is said to be weakly l.s.c. if the sublevel sets $\{h : G(h) \leqslant a\}$ are closed in the weak topology on $\mathcal{H}$.

(iv) If $H_0 \subset \mathcal{H}$, the restriction of G to H_0, $G\big|_{H_0}$, is weakly sequentially l.s.c. if $\liminf_{n\to\infty} G(h_n) \geqslant G(h)$ *whenever* $\{h, h_n, n \geqslant 1\} \subset H_0$ *and* $h_n \overset{w}{\to} h$ *in* H_0 *in the sense for any* $g \in H_0$, $\langle h_n, g\rangle \to \langle h, g\rangle$. *Strong l.s.c. of* $G\big|_{H_0}$ *is defined similarly.*

All notions of l.s.c. are equivalent when $\mathcal{H}$ is finite-dimensional, but that is obviously not the case when $\mathcal{H}$ is infinite-dimensional. For infinite-dimensional Hilbert spaces, it should be noted that the notion of weakly sequentially l.s.c. is not equivalent to that of weakly l.s.c. (since the weak topology on $\mathcal{H}$ is not metrizable). In fact, we have the following hierarchy:

$$G \text{ is weakly l.s.c. } \Rightarrow G \text{ is weakly sequentially l.s.c. } \Rightarrow G \text{ is strongly l.s.c.}$$

This is immediate because a subset $C \subset \mathcal{H}$ is weakly closed $\Rightarrow C$ is weakly sequentially closed $\Rightarrow C$ is closed in the norm topology. Thus the assumption of strong l.s.c. on a function G is a weaker assumption than that of weak l.s.c. of G. However, under the additional assumption of quasiconvexity, all notions of l.s.c. are equivalent (see Remark 4-(iii) below).

Definition 3 *A function* $G : \mathcal{H} \to [-\infty, \infty]$ *is* quasiconvex *if for any* $\delta \in [0, 1]$ *and* $h, h' \in \mathcal{H}$,

$$G(\delta h + (1-\delta)h') \leqslant \max\{G(h), G(h')\}, \tag{2.6}$$

or equivalently, the sublevel sets $\{h : G(h) \leqslant a\}$ *are convex. It will be called* almost quasiconvex, *if* (2.6) *holds for* $0 < \delta < 1$ *when* $G(h) \neq G(h')$.

G *is* strictly quasiconvex *if the inequality in* (2.6) *is strict for* $0 < \delta < 1$ *and* $h \neq h'$. *It will be called* almost strictly quasiconvex *if* (2.6) *holds with strict inequality for* $G(h) \neq G(h')$ *and* $0 < \delta < 1$.

Note that for almost quasiconvex or almost strictly quasiconvex functions no stipulations are made if $G(h) = G(h')$.

Remark 4

(i) The definition of strict quasiconvexity is not uniform in the literature. Slight variants of the definition given above have been used in the literature. In particular, Greenberg and Pierskalla (1971) used strict quasiconvexity for functions which we call here almost strictly quasiconvex.

(ii) A strictly quasiconvex function is of course quasiconvex, and an almost strictly quasiconvex function is almost quasiconvex. But an almost strictly quasiconvex function need not be quasiconvex. The standard example given in Greenberg and Pierskalla (1971) is $G : \mathbb{R} \to \mathbb{R}$ defined by $G(x) = 1_{\{0\}}(x)$. It's clear G is almost strictly quasiconvex, but the sublevel set $\{x : G(x) \leqslant 0\} = \mathbb{R} - \{0\}$, which is not convex; hence G is not quasi-convex.

(iii) A convex function is of course both quasiconvex and almost strictly quasiconvex, and a strictly convex function is strictly quasiconvex. The equivalence of strong and weak l.s.c. of a function $G : \mathcal{H} \to [-\infty, \infty]$ under the assumption of quasiconvexity is simply a consequence of Mazur's lemma which, in particular, states that a convex subset $C \subset \mathcal{H}$ is closed in the norm topology iff it is closed in the weak topology.

Lemma 5 *Let $G : \mathcal{H} \to [-\infty, \infty]$ be weakly sequentially l.s.c., and $\limsup_{\|h\| \to \infty} G(h) = \infty$. Then there exists a global minimizer $h^* \in \mathcal{H}$ such that $G(h^*) = \min_{h \in \mathcal{H}} G(h) = \inf_{h \in \mathcal{H}} G(h)$.*

Theorem 6 *Let $\mathcal{H}$ be a Hilbert space, and $F : \mathcal{H} \times [0, \infty) \to [-\infty, \infty]$. Consider the minimization problem* (2.4) *where $Q \in L(\mathcal{H}, \mathcal{H})$ is self-adjoint and p.s.d.. Let $\mathcal{M}$ be a closed subspace of $\mathcal{H}$, and the following conditions hold: (a) $F(h, u) \geqslant F(\mathcal{P}_{\mathcal{M}} h, u)$, $h \in \mathcal{H}, u \in [0, \infty)$, where $\mathcal{P}_{\mathcal{M}} : \mathcal{H} \to \mathcal{M}$ is the (orthogonal) projection operator onto the subspace $\mathcal{M}$, (b) for each fixed $h \in \mathcal{H}$, the mapping $u \in \mathbb{R} \longrightarrow F(h, u)$ is non-decreasing, and (c) $Q\mathcal{M} \subset \mathcal{M}$.*

(i) Then $\inf_{h \in \mathcal{H}} F\left(h, \langle Qh, h\rangle^{1/2}\right) = \inf_{h \in \mathcal{M}} F\left(h, \langle Qh, h\rangle^{1/2}\right)$. If $h^* \in \mathcal{H}$ is a global minimizer of $F\left(h, \langle Qh, h\rangle^{1/2}\right)$, then so is $\mathcal{P}_{\mathcal{M}} h^*$; in other words existence of a minimizer also guarantees existence of a minimizer lying in $\mathcal{M}$. If in addition for each $h \in \mathcal{H}$, the mapping $u \to F(h, u)$ is strictly increasing and $\mathcal{N}_Q \subset \mathcal{M}$ (or equivalently, $\mathcal{N}_Q \cap \mathcal{M}^\perp = \{0\}$), then any (global) minimizer h^* of the minimization problem (when it exists) lies in $\mathcal{M}$.

(ii) If for each fixed $u \in [0, \infty)$, the mapping $h \in \mathcal{H} \longrightarrow F(h, u)$ is almost quasiconvex, and for each $h \in \mathcal{H}$, the mapping $u \to F(h, u)$ is strictly increasing and $\mathcal{N}_Q \subset \mathcal{M}$, then any local minimizer h^0 of (2.4) *(when it exists) lies in $\mathcal{M}$.*

(iii) If F is almost strictly quasiconvex (in particular, convex), then any local minimizer h^0 is also a global minimizer. If F is strictly quasiconvex, then the global minimizer of F, when it exists, is unique.

Remark 7 If Q is self-adjoint and $\mathcal{M}$ is a closed subspace then $Q\mathcal{M} \subset \mathcal{M}$ (condition (c) in Theorem 6) is equivalent to $Q\mathcal{M}^\perp \subset \mathcal{M}^\perp$ which in turn is equivalent to commutativity of Q and $\mathcal{P}_{\mathcal{M}}$. The first equivalence is easy to see. It is also immediate that if Q and $\mathcal{P}_{\mathcal{M}}$ commute, then $Q\mathcal{M} \subset \mathcal{M}$. To see the other direction of the second equivalence, we have for any $h \in \mathcal{H}$

$$Q\mathcal{P}_{\mathcal{M}} h + Q(I - \mathcal{P}_{\mathcal{M}})h = Qh = \mathcal{P}_{\mathcal{M}} Qh + (I - \mathcal{P}_{\mathcal{M}})Qh.$$

Now $Q\mathcal{M} \subset \mathcal{M}$ and $Q\mathcal{M}^\perp \subset \mathcal{M}^\perp$ imply that $Q\mathcal{P}_\mathcal{M} h \in \mathcal{M}$ and $Q(I - \mathcal{P}_\mathcal{M})h \in \mathcal{M}^\perp$, and, of course, by the definition of $\mathcal{P}_\mathcal{M}$, $\mathcal{P}_\mathcal{M} Qh \in \mathcal{M}$ and $(I - \mathcal{P}_\mathcal{M})Qh \in \mathcal{M}^\perp$. Since $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^\perp$, we must have $\mathcal{P}_\mathcal{M} Qh = Q\mathcal{P}_\mathcal{M} h$.

Lemma 5 and Theorem 6, whose proofs are given in the Appendix, together give conditions for existence of the solution (minimizer) to the problem (2.4) and for the solution to lie in a designated subspace $\mathcal{M}$.

Connection to optimization problems in learning: In many applications, particularly those arising in the context of learning problems, F is of the form $F(h, u) = F_0(h) + J(u)$, where F_0 can be viewed as a loss functional and an associated penalty function on the size of h is defined through J. A typical choice of J and the operator Q are $J(u) = u^2$, $Q = I$, which defines the popular square-norm penalty function, $\|h\|^2$. The following corollary is essentially a restatement of Theorem 6 in this case. Importantly, Theorem 6 or Corollary 8 below allows use of seminorms $\langle h, Qh\rangle^{1/2}$, which are different from the original $\mathcal{H}$-norm, inside the penalty function J. Since Q does not need to be uniformly p.d. or even p.d., they are not necessarily equivalent to the $\mathcal{H}$-norm. The desired subspace, $\mathcal{M}$, in these problems is often finite-dimensional leading to a tractable finite-sum expansion form of the minimizer.

Corollary 8 *Suppose F is of the form $F(h, u) = F_0(h) + J(u)$, where $J : [0, \infty) \to [0, \infty)$ is strictly increasing. Consider the minimization problem* (2.4)*. Assume that the linear operator $Q : \mathcal{H} \to \mathcal{H}$ of* (2.4) *is self-adjoint and p.s.d., $\mathcal{N}_Q \cup Q\mathcal{M} \subset \mathcal{M}$, where $\mathcal{M}$ is a closed subspace of $\mathcal{H}$, and $F_0(h) \geqslant F_0(\mathcal{P}_\mathcal{M} h)$ for all $h \in \mathcal{H}$. Then the set of global minimizers,*

$$M_0 \stackrel{def}{=} \left\{ h^* \in \mathcal{H} : F\left(h^*, \langle Qh^*, h^*\rangle^{1/2}\right) = F^* = \inf_{h\in\mathcal{H}} F\left(h, \langle Qh, h\rangle^{1/2}\right) \right\} \subset \mathcal{M}. \tag{2.7}$$

Suppose in addition $F_0\big|_\mathcal{M} : \mathcal{M} \to \mathcal{H}$ is weakly l.s.c., J is l.s.c. and either (a) $F_0\big|_\mathcal{M}$ is bounded below, J is coercive (that is, $\limsup_{u\to\infty} J(u) = \infty$), and $Q\big|_\mathcal{M}$ is uniformly p.d. (in particular, Q is p.d. because of the eariler assumption $\mathcal{N}_Q \subset \mathcal{M}$) , or (b) $\limsup_{h\in\mathcal{M},\ \|h\|\to\infty} F_0(h) = \infty$. Then $M_0 \neq \emptyset$.

In many applications it is desirable to consider minimization problems where penalty is imposed on the size of only a part of the function h. Below we demonstrate that Corollary 8 covers such cases. In machine-learning, such minimization problems arise when partial structure of the unknown function h to be learned is known, and the so-called semiparametric representer theorem (which is a special case of Corollary 8 or Corollary 9 below) is a useful result covering a subset of such instances.

For two Hilbert spaces $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$, recall that the external direct sum $\mathcal{H}_1 \oplus_e \mathcal{H}_2$ is the space $\mathcal{H}_1 \times \mathcal{H}_2$ equipped with the inner product

$$\langle (h_1, h_2), (h_1', h_2')\rangle_e = \langle h_1, h_1'\rangle_1 + \langle h_2, h_2'\rangle_2.$$

Corollary 9 *Let $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_1)$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_2)$ be two Hilbert spaces and $\mathcal{H} = \mathcal{H}_1 \oplus_e \mathcal{H}_2$. Suppose F is of the form $F(h, u) = F_0(h) + J(u)$, where $J : [0, \infty) \to [0, \infty)$ is strictly increasing. Consider the minimization problem*

$$\min_{h=(h_1,h_2)\in\mathcal{H}} F_0(h) + J\left(h_1, \langle Q_1 h_1, h_1\rangle^{1/2}\right) \tag{2.8}$$

where $Q_1 \in L(\mathcal{H}_1, \mathcal{H}_2)$ is self-adjoint and p.s.d.. Let $\mathcal{M}_1$ be a closed subspace of $\mathcal{H}_1$, and assume that $\mathcal{N}_{Q_1} \cup Q_1\mathcal{M}_1 \subset \mathcal{M}_1$, $F_0(h) = F_0(h_1, h_2) \geqslant F_0(P_{\mathcal{M}_1}h_1, h_2)$ for all $h = (h_1, h_2) \in \mathcal{H}$. Then the set of global minimizers, $M_0 \subset \mathcal{M}_1 \oplus_e \mathcal{H}_2$

Suppose in addition $F_0\big|_{\mathcal{M}_1 \oplus_e \mathcal{H}_2} : \mathcal{M}_1 \oplus_e \mathcal{H}_2 \to \mathcal{H}$ is weakly sequentially l.s.c., $\limsup\limits_{\substack{\|h\|\to\infty \\ h\in\mathcal{M}_1\oplus\mathcal{H}_2}} F_0(h) = \infty$, and J is l.s.c. Then $M_0 \neq \emptyset$.

Proof Define $\mathcal{M} = \mathcal{M}_1 \oplus_e \mathcal{H}_2$, $Q : \mathcal{H} \to \mathcal{H}$ by $Qh = Q(h_1, h_2) = (Q_1h_1, 0)$ and notice that $\mathcal{N}_Q = \mathcal{N}_{Q_1} \oplus \mathcal{H}_2$. The assertion now follows from Corollary 8. ■

Remark 10 If $\mathcal{M}$ is a finite-dimensional subspace of $\mathcal{H}$, which, as mentioned, is an important case in practice, and $F(h, u) = F_0(h) + J(u)$, then strong l.s.c. of $F_0\big|_{\mathcal{M}}$, which is easier to check, is equivalent to weak l.s.c. (and hence weak sequential l.s.c.) of $F_0\big|_{\mathcal{M}}$. No additional assumption of quasiconvexity of $F_0\big|_{\mathcal{M}}$ is needed. Furthermore, in this case $Q\big|_{\mathcal{M}}$ is p.d. iff it is uniformly p.d. Thus the conditions of Corollary 8 are easier to check.

Revisiting the representer theorem

We now show that the representer theorem along with most of its extensions is a special case of Theorem 6. Importantly, we present the generalized semiparametric version of it for vector-valued functions and also include conditions for existence of the minimizer. Representer theorem for vector-valued functions is comparatively less studied, but some notable versions have been proved in Micchelli and Pontil (2005) (also see Alvarez et al. (2012) for a review of results on learning vector-valued functions). We chose the range of the functions to be finite-dimensional vector space only for ease of presentation, but the same proof (with the appropriate changes) holds if the range of the functions is infinite-dimensional.

The definition of RKHS of vector-valued functions is very similar to that of the scalar-valued functions with the primary difference being that the associated kernel $\boldsymbol{\kappa}$ is now matrix-valued.

Definition 11 *Let $\mathbb{U}$ be an arbitrary space. A symmetric function $\boldsymbol{\kappa} : \mathbb{U} \times \mathbb{U} \to \mathbb{R}^{n\times n}$ is a* reproducing kernel *if for any $u, u' \in \mathbb{U}$, $\boldsymbol{\kappa}(u, u')$ is a $n \times n$ p.s.d. matrix.*

The RKHS associated with a reproducing kernel $\boldsymbol{\kappa}$ is a Hilbert space $\mathcal{H}_{\boldsymbol{\kappa}}$ of functions $h : \mathbb{U} \to \mathbb{R}^n$, such that for every fixed $u \in \mathbb{U}$ and a (column) vector $c \in \mathbb{R}^n$, (i) the mapping $u' \to \boldsymbol{\kappa}(u', u)c$ is an element of $\mathcal{H}_{\boldsymbol{\kappa}}$, and (ii) $\langle h, \boldsymbol{\kappa}(\cdot, u)c\rangle = h(u)^T c$.

Property (ii) is the reproducing property of the kernel $\boldsymbol{\kappa}$ in the vector framework. Like the scalar case, given a reproducing matrix-valued kernel $\boldsymbol{\kappa}$, an extension of Moore-Aronszajn theorem gives the the associated RKHS, $\mathcal{H}_{\boldsymbol{\kappa}} = \overline{\text{Span}}\{\boldsymbol{\kappa}(\cdot, u) : u \in \mathbb{U}\}$, where the closure is taken with respect to the norm, $\|\cdot\|_{\boldsymbol{\kappa}}$, now appropriately modified as

$$\|h\|_{\boldsymbol{\kappa}} = \sum_{i,j=1}^{l} c_i^T \boldsymbol{\kappa}(u_i, u_j)c_j, \quad h = \sum_{j=1}^{l} \boldsymbol{\kappa}(\cdot, u_j)c_j,\ c_j \in \mathbb{R}^n.$$

Corollary 12 (Semiparametric Representer Theorem) *Let $L : \mathbb{R}^{nm} \to [-\infty, \infty]$ be any function, and $J : [0, \infty) \to [0, \infty)$ nondecreasing, and $\boldsymbol{\kappa} : \mathbb{U} \times \mathbb{U} \to \mathbb{R}^{n\times n}$ a reproducing kernel. Let*

$\mathcal{H}_{\boldsymbol{\kappa}}$ be the RKHS of functions $h : \mathbb{U} \to \mathbb{R}^n$ corresponding to a symmetric positive definite kernel $\boldsymbol{\kappa}$. Let $x_1, x_2, \ldots, x_m \in \mathbb{U}$ be fixed. Let $\mathcal{G} = \text{span}\{\mathfrak{g}_1, \mathfrak{g}_2, \ldots, \mathfrak{g}_r\}$, where $\mathfrak{g}_1, \mathfrak{g}_2, \ldots, \mathfrak{g}_r$ are linearly independent functions mapping $\mathbb{U} \to \mathbb{R}^n$. Consider the objective function

$$L(h(x_1) + g(x_1), h(x_2) + g(x_2), \ldots, h(x_m) + g(x_m)) + J(\|h\|), \quad \bar{h} = (h, g) \in \mathcal{H}_{\boldsymbol{\kappa}} \oplus_e \mathcal{G}$$

Then the following hold.

(a) If a minimizer to the above objective function exists, then there also exists a minimizer $\bar{h}^*$ of the form

$$\bar{h}^*(u) = \left(\sum_{k=1}^{m} \boldsymbol{\kappa}(x_k, u) c_k^*, \sum_{i=1}^{r} \mathfrak{g}_i(u) \alpha_i^* \right) \tag{2.9}$$

for some constants $c_i^* \in \mathbb{R}^n$ and $\alpha_k^* \in \mathbb{R}$. If J is also strictly increasing then any minimizer (when it exists) is of the form (2.9)*.*

(b) If L and J are l.s.c. and L is coercive (that is, $\limsup_{\|z\| \to \infty} L(z) = \infty$), then there exists a minimizer h^* of the form (2.9)*.*

Notice that a Hilbertian structure can be put on $\mathcal{G}$ with the inner product

$$\langle g, g' \rangle_{\mathcal{G}} \stackrel{def}{=} \sum_{i,j=1}^{r} \alpha_i \alpha'_j, \quad g = \sum_{i=1}^{r} \alpha_i \mathfrak{g}_i, \quad g' = \sum_{i=1}^{r} \alpha'_i \mathfrak{g}_i,$$

Define the finite-dimensional subspace

$$\mathcal{M} = \left\{ \sum_{i=1}^{m} \boldsymbol{\kappa}(x_i, \cdot) c_i : c_i \in \mathbb{R}^n, i = 1, 2, \ldots, m \right\},$$

and observe that by the reproducing property for any $h \in \mathcal{H}_\kappa$ and $v \in \mathbb{R}^n$,

$$h(x_i)^T v = \langle h, \boldsymbol{\kappa}(x_i, \cdot) v \rangle = \langle \mathcal{P}_{\mathcal{M}} h, \boldsymbol{\kappa}(x_i, \cdot) v \rangle + \langle (I - \mathcal{P}_{\mathcal{M}}) h, \boldsymbol{\kappa}(x_i, \cdot) v \rangle = ((\mathcal{P}_{\mathcal{M}} h)(x_i))^T v.$$

The second term after the second equality is 0 because $\boldsymbol{\kappa}(x_i, \cdot) v \in \mathcal{M}$ and $(I - \mathcal{P}_{\mathcal{M}}) h \in \mathcal{M}^\perp$. Since the above equality is true for any $v \in \mathbb{R}^n$, it follows that $h(x_i) = \mathcal{P}_{\mathcal{M}} h(x_i)$. Consequently, $F_0(h, g) \stackrel{def}{=} L(h(x_1) + g(x_1), h(x_2) + g(x_2), \ldots, h(x_m) + g(x_m)) = F_0(\mathcal{P}_{\mathcal{M}} h, g)$. Moreover, it is easy to see that for each x_i, $\limsup_{\|h\|_{\boldsymbol{\kappa}} \to \infty,\, h \in \mathcal{M}} h(x_i) = \infty$ and $\limsup_{\|g\|_{\mathcal{G}} \to \infty} g(x_i) = \infty$, which, because of the hypothesis on L, in turn implies that $\limsup_{\substack{\|(h,g)\| \to \infty \\ (h,g) \in \mathcal{M} \oplus_e \mathcal{G}}} F_0(h, g) = \infty$. It follows that Corollary 12 is a restatement of Corollary 9 in this particular case.

Remark 13

(i) It is obvious that Corollary 12 covers minimization the objective function of the form

$$\tilde{L}\left((x_1, y_1, (h + g)(x_1)), (x_2, y_2, (h + g)(x_2)), \ldots, (x_m, y_m, (h + g)(x_m))\right) + J(\|h\|)$$

where the points $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^{d'}$, $i = 1, 2, \ldots, m$ are fixed. Indeed, in this case one simply defines the function $L : \mathbb{R}^m \to \mathbb{R}$ in Corollary 12 as

$$L(u_1, u_2, \ldots, u_m) = \tilde{L}\left((x_1, y_1, u_1), (x_2, y_2, u_2), \ldots, (x_m, y_m, u_m)\right).$$

(ii) Absence of the semiparametric part as encoded by the space $\mathcal{G}$ leads to the usual representer theorem. By Corollary 8 in this case, coercivity of J ($\limsup_{\|u\|\to\infty} J(u) = \infty$) with lower boundedness of L instead of coercivity of L also guarantees the existence of a minimizer in part (b). Also as evident from Theorem 6, seminorms of the form $\langle \cdot, Q\cdot\rangle^{1/2}$, which are different from the RKHS norm, can be used inside J.

(iii) Although informally, one can say that the minimizer in (2.9) is of the form $\bar{h}^*(u) = \sum_{k=1}^m \boldsymbol{\kappa}(x_k, u)c_k^* + \sum_{i=1}^r \mathfrak{g}_i(u)\alpha_i^*$, strictly speaking, such a representation is not correct, and mathematically it should be represented as a pair as in (2.9). This is because $\mathcal{H}_{\boldsymbol{\kappa}} \cap \mathcal{G}$ might not be $\{0\}$, in which case the mapping $(h, g) \in \mathcal{H}_{\boldsymbol{\kappa}} \oplus_e \mathcal{G} \to h + g \in \mathcal{H}_{\boldsymbol{\kappa}} + \mathcal{G}$ is not injective. In other words, the function $f = h + g$ might have different representations in $\mathcal{H} + \mathcal{G}$, and consequently the mapping $h + g \to L(h(x_1) + g(x_1), h(x_2) + g(x_2), \ldots, h(x_m) + g(x_m)) + J(\|h\|)$ is not a well-defined function!

A common choice of matrix-valued reproducing kernel is the class of separable kernels of the form $(\boldsymbol{\kappa}(u, u'))_{i,j} = k(u, u')\rho(i, j)$, where k and ρ are scalar kernels on $\mathbb{R}^d \times \mathbb{R}^d$ and $\{1, 2, \ldots, d\} \times \{1, 2, \ldots, d\}$, respectively. This is of course same as the class of kernels having the representation $\boldsymbol{\kappa}(u, u') = k(u, u')B$ with B being an $n \times n$ p.s.d. matrix. Note for most learning problems one can assume without loss of generality that $B = I_n$, as B can be "absorbed" in the coefficients c_i of the finite expansion of the form (2.9) by redefining c_i as Bc_i. More general class of matrix-kernels consists of $\boldsymbol{\kappa}$ of the form $\boldsymbol{\kappa}(u, u') = \sum_{r=1}^R k_r(u, u')B_r$. For a given set of data-points $\{x_1, x_2, \ldots, x_m\}$, the associated $nm \times nm$-dimensional Gram matrix $\boldsymbol{\mathcal{K}}$, which is important for determination of the coefficients of the finite expansion, is given by $\boldsymbol{\mathcal{K}} = \sum_{r=1}^R \mathcal{K}_r \otimes B_r$. Here $\mathcal{K}_r = ((k_r(x_i, x_j)))_{m\times m}$ is the usual Gram matrix corresponding to the scalar kernel k_r.

We now show how Corollary 8 (or more generally Theorem 6) can be employed to obtain a computable representation of the solution to the minimization problems mentioned in Examples 1 and Examples 2 — two scenarios where the usual representer theorem cannot be used.

Revisiting Example 1 - Linear functional regression

To obtain a solution to the problem (2.2) define the finite-dimensional vector space

$$\mathcal{M} = \text{span}\left\{f_i : f_i(u) = \mathcal{L}_{x_i}\boldsymbol{\kappa}(u, \cdot), i = 1, 2, \ldots, m\right\}.$$

Here, however, we first need to check that $\mathcal{M}$ is indeed a subspace of $\mathcal{H}_{\boldsymbol{\kappa}}$ (as it is not obvious). Nevertheless, it is easy as we first note that by the Riesz representation theorem there is $g_i \in \mathcal{H}_{\boldsymbol{\kappa}}$, such that $\mathcal{L}_{x_i}h = \langle h, g_i\rangle$ for any $h \in \mathcal{H}_{\boldsymbol{\kappa}}$. Consequently,

$$f_i(u) = \mathcal{L}_{x_i}\boldsymbol{\kappa}(u, \cdot) = \langle \boldsymbol{\kappa}(u, \cdot), g_i\rangle = g_i(u),$$

where the last equality is because of the reproducing property. That is $f_i = g_i \in \mathcal{H}_{\boldsymbol{\kappa}}$; hence $\mathcal{M} \subset \mathcal{H}_{\boldsymbol{\kappa}}$ and $g_i \in \mathcal{M}$. As before writing $h \in \mathcal{H}_{\boldsymbol{\kappa}}$ as $h = \mathcal{P}_{\mathcal{M}}h + (I - \mathcal{P}_{\mathcal{M}})h$, we see that

$$\mathcal{L}_{x_i}h = \mathcal{L}_{x_i}\mathcal{P}_{\mathcal{M}}h + \mathcal{L}_{x_i}(I - \mathcal{P}_{\mathcal{M}})h = \mathcal{L}_{x_i}\mathcal{P}_{\mathcal{M}}h + \langle (I - \mathcal{P}_{\mathcal{M}})h, g_i\rangle = \mathcal{L}_{x_i}\mathcal{P}_{\mathcal{M}}h.$$

The last equality is because $(I - \mathcal{P}_{\mathcal{M}})h \in \mathcal{M}^\perp$, and we showed that $g_i \in \mathcal{M}$. Consequently, $F_0(h) \stackrel{def}{=} \tilde{L}\left((x_1, y_1, \mathcal{L}_{x_1}h), (x_2, y_2, \mathcal{L}_{x_2}h), \ldots, (x_m, y_m, \mathcal{L}_{x_m}h)\right) = F_0 \circ \mathcal{P}_{\mathcal{M}}(h)$, and hence by

Corollary 8 (also see Remark 10) a minimizer $h^* \in \mathcal{M}$; in other words h^* is of the form

$$h^*(u) = \sum_{i=1}^{m} \mathcal{L}_{x_i} \boldsymbol{\kappa}(u, \cdot) c_i.$$

Revisiting Example 2 - Fredholm integral equation of first kind

The following result characterizes the solution to the problem (2.3) in Example 2.

Corollary 14 *Consider the framework of Example 2. Let $\mathcal{E} \subset \mathbb{R}^d$ be compact, and let $R : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ be continuous. Let $\mathcal{H}_{\boldsymbol{\kappa}}$ be the RKHS corresponding to a reproducing kernel $\boldsymbol{\kappa}$. Assume that $\boldsymbol{\kappa} : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ is continuous. Let $L : \mathbb{R}^m \to [-\infty, \infty]$ be any function, and $J : [0, \infty) \to [0, \infty)$ nondecreasing. For fixed $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$ consider the objective function*

$$L(Rh(x_1), Rh(x_2), \ldots, Rh(x_m)) + J(\|h\|) \stackrel{def}{=} F(h, \|h\|), \quad h \in \mathcal{H}_{\boldsymbol{\kappa}}.$$

Then the following hold.

(a) If a minimizer to the above objective function exists, then there also exists a minimizer h^* of the form

$$h^*(u) = \sum_{i=1}^{m} R\boldsymbol{\kappa}(u, \cdot)(x_i) c_i = \sum_{i=1}^{m} c_i \int_{\mathcal{E}} \boldsymbol{\kappa}(u, z) R(x_i, z) dz \tag{2.10}$$

for some constants $c_i \in \mathbb{R}$. If J is also strictly increasing then any minimizer (when it exists) is of the form (2.9).

(b) Suppose L and J are l.s.c. and either (a) L is coercive or (b) J is coercive and L bounded below (for example, non-negative). Then there exists a minimizer h^* of the form (2.10).

Proof It's easy to see that the continuity of the mapping $\boldsymbol{\kappa} : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ gives continuity of the mapping $z \in \mathcal{E} \longrightarrow \boldsymbol{\kappa}(z, \cdot) \in \mathcal{H}_R$. Since $\mathcal{E}$ is assumed to be compact, the latter mapping is Bochner measurable, and thus so is the mapping $z \in \mathcal{E} \longrightarrow \boldsymbol{\kappa}(z, \cdot) R(x_i, z) \in \mathcal{H}_R$ for each $i = 1, 2, \ldots, m$. Moreover, the mapping $z \in \mathcal{E} \longrightarrow \|\boldsymbol{\kappa}(z, \cdot) R(x_i, z)\| = \boldsymbol{\kappa}(z, z) |R(x_i, z)| \in \mathbb{R}$ is obviously integrable (as it is continuous, and $\mathcal{E}$ is compact); hence the mapping $z \to \boldsymbol{\kappa}(z, \cdot) R(x_i, z)$ is Bochner integrable (e.g. see Yosida (1995)). Thus the functions f_i defined by the following Bochner integral:

$$f_i \stackrel{def}{=} \int_{\mathcal{E}} \boldsymbol{\kappa}(\cdot, z) R(x_i, z) dz$$

are elements of $\mathcal{H}_{\boldsymbol{\kappa}}$. Since the evaluation functionals are continuous on an RKHS, obviously, $f_i(u) = \int_{\mathcal{E}} \boldsymbol{\kappa}(u, z) R(x_i, z) dz = R\boldsymbol{\kappa}(u, \cdot)(x_i)$, where the integral in the middle is a regular Riemann integral.

Now define the finite dimensional subspace $\mathcal{M} \subset \mathcal{H}_{\boldsymbol{\kappa}}$ by

$$\mathcal{M} = \text{span}\{f_i : i = 1, 2, \ldots, m\}.$$

Writing $h \in \mathcal{H}_{\boldsymbol{\kappa}}$ as $h = \mathcal{P}_{\mathcal{M}} h + (I - \mathcal{P}_{\mathcal{M}})h$, we see that

$$\begin{aligned} Rh(x_i) &= (R\mathcal{P}_{\mathcal{M}} h)(x_i) + (R(I - \mathcal{P}_{\mathcal{M}})h)(x_i) = (R\mathcal{P}_{\mathcal{M}} h)(x_i) + \int_{\mathcal{E}} R(x_i, z)(I - \mathcal{P}_{\mathcal{M}})h(z)\, dz \\ &= (R\mathcal{P}_{\mathcal{M}} h)(x_i) + \int_{\mathcal{E}} R(x_i, z)\langle (I - \mathcal{P}_{\mathcal{M}})h, \boldsymbol{\kappa}(z, \cdot)\rangle\, dz \\ &= (R\mathcal{P}_{\mathcal{M}} h)(x_i) + \left\langle (I - \mathcal{P}_{\mathcal{M}})h, \int_{\mathcal{E}} R(x_i, z)\boldsymbol{\kappa}(z, \cdot) dz \right\rangle = (R\mathcal{P}_{\mathcal{M}} h)(x_i) + \langle (I - \mathcal{P}_{\mathcal{M}})h, f_i\rangle \\ &= (R\mathcal{P}_{\mathcal{M}} h)(x_i), \end{aligned}$$

where the fourth equality is by the property of Bochner integrals (and the fact that the mapping $g \to \langle (I - \mathcal{P}_{\mathcal{M}})h, g\rangle$ is a continuous linear functional). Consequently, $F_0(h) \stackrel{def}{=} L(Rh(x_1), Rh(x_2), \ldots, Rh(x_m)) = F_0 \circ \mathcal{P}_{\mathcal{M}}(h)$, and hence the conclusion of Corollary 14 is just a restatement of Corollary 8. ■

Relaxations of some of the assumptions including compactness of $\mathcal{E}$ in Corollary 14 are easily possible.

3. Framework of stochastic differential equations

We consider the d-dimensional SDE of the form

$$X(t) = x_0 + \int_0^t b(X(s))ds + \int_0^t \sigma(X(s))dW(s), \quad x_0 \in \mathbb{R}^d, \tag{3.1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and W is a d-dimensional Brownian motion. We assume that the functions b and σ are such that the above SDE admits a unique strong solution. This, for example, holds when b and σ are locally Lipschitz and $\sigma\sigma^T$ is non singular. The functional forms of b and σ are unknown, and our objective is to learn the SDE, that is, the associated driving functions from high-frequency data $\boldsymbol{X}_{t_1:t_m} \stackrel{def}{=} (X(t_1), X(t_2), \ldots, X(t_m))$, where $\Delta = t_i - t_{i-1} \ll 1$.

Our approach to this problem is to first consider an optimization problem in an appropriate RKHS. Assume that for each $t \geqslant 0$, the distribution of $X(t)$ given $X(0) = x_0$ admits a density $p_t(\cdot|x_0)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. This, for example, exists when for each x, $\sigma\sigma^T(x)$ is positive definite (see Rogers and Williams (2000)). The function $p_t(x|x_0)$ satisfies the Kolmogorov forward PDE (Fokker-Plank equation) $\partial_t p_t(x|x_0) = (\mathcal{L})^* p_t(x|x_0),\ p_0(\cdot|x_0) = \delta_{x_0}$ in weak sense. Here $(\mathcal{L})^*$ is the adjoint of the generator $\mathcal{L}$ defined by

$$\mathcal{L}f(x) = \sum_{i=1}^d b_i(x)\partial_i f(x) + \frac{1}{2}\sum_{1\leqslant i,j\leqslant d} (\sigma\sigma^T)_{ij}(x)\partial_{ij} f(x), \quad f \in C^2(\mathbb{R}^d, \mathbb{R}).$$

By time-homogeneity, the transition density of $X(t + s)$ given $X(t) = x$ is of course given by $p_s(\cdot|x)$. Therefore the likelihood of the data as a function of b and the inverse covariance matrix $A = (\sigma\sigma^T)^{-1}$, which is the joint density of $\boldsymbol{X}_{t_1:t_m}$, is given by

$$L(b, A|\boldsymbol{X}_{t_1:t_m}) = \prod_{i=1}^m p_\Delta(X(t_i)|X(t_{i-1})), \quad t_0 = 0,\ X(0) = x_0. \tag{3.2}$$

The natural loss function here is the negative log likelihood, $-\ln L$, and the functions b and $A = (\sigma\sigma^T)^{-1}$ are learned through minimizing it over an RKHS, subject to a penalty term. Now the transition densities $p_s(\cdot|\cdot)$ are usually not available in closed form, and in practice, we often work with a discretized version of the SDE (3.1). In this paper we will consider the Euler-Maruyama approximation of (3.1) given by

$$X(t_i) = X(t_{i-1}) + b(X(t_{i-1})\Delta + \sigma(X(t_{i-1}))(W(t_i) - W(t_{i-1})), \quad \Delta = t_i - t_{i-1} \ll 1 \tag{3.3}$$

which has a weak-error of order 1, same as the Milstein-scheme (Graham and Talay (2013)). The advantage of Euler-Maruyama (EM) approximation, is that the transition density of the discretized chain (3.3), which can be thought of as an approximation to that of the original process X, is simply given by

$$p_\Delta^{EM}(x'|x) = \mathcal{N}_d(x'|x + b(x)\Delta, \sigma\sigma^T(x)\Delta).$$

Consequently, the likelihood function L in (3.2) will be approximated by L^{EM}, the likelihood function of the EM chain (3.3), which is defined in a way similar to (3.2) with the approximate transition densities $p_\Delta^{EM}(X(t_i)|X(t_{i-1}))$ replacing the exact $p_\Delta(X(t_i)|X(t_{i-1}))$. Discretized chains corresponding to Milstein-scheme or higher-order approximations like Runge-Kutta type schemes do not have such simple closed forms of transition densities and are comparatively difficult to work with for development of learning algorithms.

Since our objective is to learn vector-valued functions, the corresponding minimization problem needs to be cast in RKHS corresponding to matrix-valued kernels (see Definition 11). Let $\boldsymbol{\kappa}_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $\boldsymbol{\kappa}_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d^2\times d^2}$ be reproducing kernels with associated RKHS $\mathcal{H}_0$ and $\mathcal{H}_1$. Let $\mathcal{H} = \mathcal{H}_0 \oplus_e \mathcal{H}_1$, and $J : [0,\infty) \to [0,\infty]$ a strictly increasing function. Then Corollary 12 gives the following result.

Theorem 15 *Consider the following minimization problem*

$$\min_{(b,A)\in\mathcal{H}} -\ln L^{EM}(b, A|X(t_1), X(t_2), \ldots, X(t_m)) + J(\|(b, vec_{d\times d}(A))\|).$$

where $A = (\sigma\sigma^T)^{-1}$. Then there exists a solution to the above minimization problem and every minimizer (b^*, A^*) is of the form

$$b^*(\cdot) = \sum_{i=1}^m \boldsymbol{\kappa}_0(\cdot, X(t_i))\beta_i^*, \quad vec_{d\times d}(A)(\cdot)) = \sum_{i=1}^m \boldsymbol{\kappa}_1(\cdot, X(t_i))\alpha_i^* \qquad \beta_i^* \in \mathbb{R}^d, \quad \alpha_i^* \in \mathbb{R}^{d^2} \tag{3.4}$$

Here $\text{vec}_{d\times d}(M)$ is vectorization of a $d \times d$ matrix M. The next part of the paper focuses on estimating the weight coefficients in the summations in (3.4).

Computational aspects

The computational part of the paper focuses only on the nonparametric learning of the *drift coefficient* b from high-frequency data. More specifically, we consider Itô diffusion with unknown drift function b but whose diffusion coefficient has the parametric form $\sigma(x) = \sigma_0(x)\varsigma$, with a *known*

function $\sigma_0 : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and an *unknown* $d \times d$ parameter matrix ς. The transition density of the discretized chain (3.3) in this case is given by

$$p_\Delta^{EM}(x'|x) = \mathcal{N}_d(x'|x + b(x)\Delta, \sigma_0(x)\varsigma\varsigma^T\sigma_0^T(x)\Delta), \tag{3.5}$$

The assumption of parametric form of the diffusion coefficient was necessary to develop Gibbs algorithms for learning b which is the focus of this part of the paper. Gibbs algorithm are easy to implement and often the preferred choice in high-dimension compared to a traditional random-walk Metropolis Hastings which requires a suitable proposal distribution. The case where both b and σ are unknown functions will involve significantly different techniques and in particular will require different MCMC algorithms and will be pursued in a future work. We however do note that the framework in this paper covers the important class of SDEs with constant diffusion coefficients.

Estimating the minimizer: Now there are two approaches to estimate the minimizer b^*, or equivalently, $\boldsymbol{\beta}^* \equiv (\beta_1^*, \beta_2^*, \ldots, \beta_m^*)$. The first obvious way is to solve the optimization problem either by an optimization algorithm (e.g. stochastic gradient descent) or in closed form when it is possible (e.g. in the case, the penalty function $J(u) = \|u\|^2$). This gives a point-estimate of $\boldsymbol{\beta}^*$, a main drawback of which, as already pointed out by Tipping (2001) in the regression case, is the absence of a reliable measure of uncertainty. Any ad-hoc post processing of the estimate to get some quantification of the uncertainty is artificial due to lack of probabilistic framework and often leads to unreliable results.

A natural remedy to the above problem is a Bayesian approach, which is the focus of this paper. This entails assigning a prior distribution $p_{prior}(\cdot)$ on the weight vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$, and estimating the posterior distribution, $p_{post}(\boldsymbol{\beta}|\boldsymbol{X}_{t_1:t_m})$. Justifying the finite expansion,

$$b(\cdot) = \sum_{i=1}^{m} \boldsymbol{\kappa}_0(\cdot, X(t_i))\beta_i \tag{3.6}$$

as an "ideal form" of the drift function b by Theorem 15, the posterior distribution, $p_{post}(\boldsymbol{\beta}|\boldsymbol{X}_{t_1:t_m})$, efficiently captures the *uncertainty* in our estimator in the m-dimensional parameter space. A common way to quantify and visualize uncertainty is to compute and plot $(1-\alpha)\%$-Bayesain credible bands around b using the posterior distribution, $p_{post}(\boldsymbol{\beta}|\boldsymbol{X}_{t_1:t_m})$. Since the closed form expressions of the posterior distributions are almost never available, empirical methods need to be used for such calculations.

The connection between the optimization problem and the Bayesian approach, as has been described numerous times in the literature in other contexts (e.g. see MacKay (1992)), is the observation that the negative of the cost function in Theorem 15 (seen as a function of $\boldsymbol{\beta}$) is the log posterior-density of $\boldsymbol{\beta}$ under the prior $p_{prior}(\boldsymbol{\beta}) \propto \exp\{-J(\boldsymbol{\beta})\}$, where by a slight abuse of notation, we denote $J(\boldsymbol{\beta}) = J(\boldsymbol{\beta}^T\boldsymbol{\mathcal{K}}_0\boldsymbol{\beta}) = J(\|b\|^2)$ with $b(\cdot) = \sum_{i=1}^m \boldsymbol{\kappa}_0(\cdot, X(t_i))\beta_i$. Here, $\boldsymbol{\mathcal{K}}_0 = ((\boldsymbol{\kappa}_0(X(t_i), X(t_j))))$ is the Gram matrix associated with the kernel $\boldsymbol{\kappa}_0$. Thus $\boldsymbol{\beta}^*$, the solution of the penalized optimization problem, is interpreted as a-posterior mode (MAP) of the posterior distribution of $\boldsymbol{\beta}$. Importantly, this observation shows that Bayesian approach allows one to use a much larger class of priors on $\boldsymbol{\beta}$ than the class of penalty functions to achieve desired objectives like sparsity; in particular, one can now use priors which do not have closed form expressions.

3.1 Sparsity and Shrinkage priors

Since for the SDE model, the RKHS framework requires that the number of terms in the finite expansion of b equals the number of data-points, m, getting a sparse estimate of $\{\beta_i : i = 1, 2, \ldots, m\}$ is necessary. This would not only lead to reduction in complexity but will protect us from an over-parametrized model. But it is important to understand why a sparse solution is expected in this case. Note that shrinkage priors in the context of SDEs hold an appeal that is interestingly different from that in usual regression setups. Here our "predictors" come in the form of correlated data. An efficient algorithm should not ideally place non-zero weights on all data-points that are very close to each other. Data points clustered together in a small region of the data space, will not provide information individually over and above what could be provided by few representative points of the cluster. Such clusters can be typically formed by slow movement of SDE resulting in two successive data-points, $X(t_i)$ and $X(t_{i+1})$, differing only by a little margin. It could also be formed by multiple visits of the SDE trajectory to the same regions of the data space due to positive recurrence or ergodicity of the system. In other words, the presence of both $\boldsymbol{\kappa}_0(\cdot, X(t_i))$ and $\boldsymbol{\kappa}_0(\cdot, X(t_j)$ is unnecessary in the finite-expansion of b when $X(t_i)$ and $X(t_j)$ are nearly identical, and only a subset of $\{\boldsymbol{\kappa}_0(\cdot, X(t_i))\}$ is relevant for learning b. In fact, this shows why we expect the methodology of the paper to work for SDEs which are positive recurrent (ergodic). It guarantees that we have enough data points to learn about the relevant weights β_i, which might not be true for other types of SDEs.

In the optimization framework, sparsity can be induced by different cost functions J in the minimization problem

$$\min_{\boldsymbol{\beta}} \left[-\ln L^{EM}(b, A | X(t_1), X(t_2), \ldots, X(t_m)) + J(\boldsymbol{\beta}) \right]$$

with $b(\cdot) = \sum_{i=1}^{m} \boldsymbol{\kappa}_0(\cdot, X(t_i))\beta_i$. While l^2 cost function often does not result in noticeable sparsity, other choices of J, for example, the lasso penalty of Tibshirani (1996) results in certain β_i's becoming zero. Within the Bayesian framework, popular choices of shrinkage prior $p_{prior}(\boldsymbol{\beta})$ lie in the normal scale-mixture family which in particular include t-prior (Tipping (2001)), double-exponential (Park and Casella (2008)) and Horseshoe priors (Carvalho et al. (2009, 2010)). A survey of some of the popular shrinkage priors used for penalized regression problems can be found in van Erp et al. (2019) (also see the references therein). The MAP estimate corresponding to double-exponential prior of course is the same as the lasso estimate, but the posterior mode often lacks nice theoretical properties and is also unsuitable from Bayesian perspective. In fact, a Bayesian approach which touts model averaging does not expect model-averaged weights to be exactly zero! It is more reasonable to consider a weaker-form of sparsity which aims to decrease $\|\boldsymbol{\beta}\|$ for some suitable norm — resulting in shrinkage rather than selection of the weights.

In the Bayesian framework, an established method inducing shrinkage is by choosing appropriate heavy-tailed distributions with sharp peak at 0 as shrinkage priors. While the sharp peak results in shrinkage of most of the coefficients, the heaviness of the tail allows truly relevant weights to shift away from 0. The use of shrinkage priors is a first, to our knowledge, in the context of SDE models.

In this paper we use two types of priors on $\{\beta_i\}$ to induce sparsity - t-distributions and the Horseshoe distribution. Since the weights β_i are vector valued, it should be noted that multidimensional versions of the above prior distributions need to be used. While multidimensional t-distribution is standard in the literature, such is not the case for Horseshoe. We describe a natural and easy-to-

implement adaptation of the classical Horseshoe to d-dimension later in the section.

t-prior: To induce sparsity we assume multivariate $\mathrm{t}_d(\cdot|\nu, 0, U)$ prior with ν degrees of freedom on each β_i. Recall that the multivariate $\mathrm{t}_d(\cdot|\nu, \mu, U)$ density function is given by

$$\mathrm{t}_d(x|\nu,\mu,U) = \frac{\Gamma\left((\nu+d)/2\right)}{\Gamma(\nu/2)\det(U)^{1/2}(\nu\pi)^{d/2}}\left[1+\frac{1}{\nu}(x-\mu)^T U^{-1}(x-\mu)\right]^{-\frac{\nu+d}{2}}, \quad x\in\mathbb{R}^d. \tag{3.7}$$

Now $\mathrm{t}_d(\cdot|\nu, \mu, U)$ can be written as a normal scale mixture with covariance matrix mixed with inverse Wishart distribution; more specifically,

$$\mathrm{t}_d(x|\nu,\mu,U) = \int_{\mathbb{R}^{d\times d}} \mathcal{N}_d(x|\mu,\lambda^2)\mathcal{IW}_d(\Lambda|\nu+d-1,U)\,d\Lambda.$$

This facilitates Gibbs sampling of the posterior by the standard technique of augmentation of the parameter space. To complete the Bayesian framework, we also need to assume prior on the starting data point $X(t_1)$. We assume $X(t_1)|x_0, \{\beta_i\}, \varsigma\varsigma^t \sim \mathcal{N}_d(\cdot|x_0 + b(x_0)\Delta, \sigma_0(x)\varsigma(\sigma_0(x)\varsigma)^T\Delta)$ with hyperparameter x_0. Notice that this prior is consistent with the dynamics of X (c.f (3.3) and (3.5)) and can be interpreted as follows: designating $t_1 - \Delta$ as the starting time, $t = 0$, we assume that $X(0) = x_0$, where we choose x_0 to be close to the first observation $X(t_1)$. Ideally, we should assign a proper prior to $X(0)$, for example, a uniform prior on a small ball around $X(t_1)$, but simply fixing $X(0) = x_0$ near the data-point $X(t_1)$ (or equivalently, assigning Dirac δ_{x_0} prior to $X(0)$) does not affect the performance of the algorithms. Our Bayesian hierarchical framework is described below.

Bayesian hierarchical framework I: t-prior

- $X(t_1 - \Delta) \equiv X(0) \sim \delta_{x_0}$.
- $\{\boldsymbol{X}_{t_1:t_m} = (X(t_1), X(t_2), \ldots, X(t_m)))\Big|\boldsymbol{\beta}, \varsigma, x_0\}$ governed by the transition probabilities (3.5), which is the result of Euler-Maruyama approximation, (3.3).
- Mean-zero Gaussian prior on the parameter $\boldsymbol{\beta}$: for $i = 1, 2, 3, \ldots, m$, $\beta_i \overset{iid}{\sim} \mathcal{N}_d(0, \Lambda_i)$, where each η_i is a $d \times d$ positive definite matrix.
- Inverse Wishart prior on the hyperparameter Λ_i: $\Lambda_i \sim \mathcal{IW}_d(\nu + d - 1, U)$ for $i = 1, 2, \ldots, m$.
- Inverse Wishart prior on the parameter $\varsigma\varsigma^T$: $\varsigma\varsigma^T \sim \mathcal{IW}_d(n, V)$.

The Λ_i controls the strength of the coefficients β_i and therefore the relevance of the data-point $X(t_i)$. For one-dimensional SDEs, this is of course equivalent to putting an inverse-gamma prior on the variance of the zero-mean normal distributions of β_i.

For multidimensional SDEs, an alternate simpler t-like prior can also be assigned to β_i by setting $\Lambda_i = \lambda_i^2 I_d$ with 1-dimensional inverse gamma prior on the scalar λ_i^2. The main advantage of the

simpler prior is that it requires much less number of hyperparameters than the multi-dimensional t_d-prior resulting in potential savings in computational complexity.

With the above priors, the conditional distribution of each of the parameters given the rest have closed forms and can be deduced from Lemma 2. This results in the following Gibb's algorithm for (approximately) generating $\boldsymbol{\beta}, \{\Lambda_i\}$ and $\varsigma\varsigma^T$ from the posterior distribution $p_{post}(\boldsymbol{\beta}, \{\Lambda_i\}, \varsigma\varsigma^T|\boldsymbol{X}_{t_1:t_m})$.

Algorithm 1: Gibb's algorithm for high frequency data.

Input: The data $\boldsymbol{X}_{t_1:t_m} = (X(t_1), X(t_2), \ldots, X(t_m))$, x_0 discretization step Δ, number of iterations L.

Output: $\boldsymbol{\beta}, \varsigma\varsigma^T, \{\Lambda_i\}$ from the posterior density.

while $l < L$ **do**
- Generate $\boldsymbol{\beta}|\boldsymbol{X}_{t_1:t_m}, \varsigma\varsigma^T, \{\Lambda_i\} \sim \mathcal{N}_d(\cdot|\mu, \boldsymbol{C})$ where μ and $\boldsymbol{C}$ are defined by (A.3). Calculate the drift function b by (3.6).
- Generate $\varsigma\varsigma^T|\boldsymbol{X}_{t_1:t_m}, \boldsymbol{\beta}, \{\Lambda_i\} \sim \mathcal{IW}_d(n+m, V_{post})$, where V_{post} is defined by (A.4).
- Generate $\Lambda_i|\boldsymbol{X}_{t_1:t_m}, \boldsymbol{\beta}, \varsigma\varsigma^T \sim \mathcal{IW}_d(\nu+d, U^{-1}+\beta_i\beta_i^T), \quad i = 1, 2, \ldots, m$ independently.
- $l = l + 1$.

end

Horseshoe type prior: We next employ a global-local class of priors from the normal scale-mixture family which has potentially better shrinkage characteristics than the t-prior (Polson and Scott (2011)). In our context of d-dimensional β_j, this is described by

$$\beta_j|\Lambda_j, \Xi \sim \mathcal{N}(0, \Lambda_j\Xi), \quad \Lambda_j \sim p_{prior}(\Lambda_j), \quad \Xi \sim p_{prior}(\Xi).$$

Ξ, which denotes the global variance component, is akin to the regularization parameter in the penalized optimization problem and its purpose is to attempt to shrink all the weights $\{\beta_j\}$. This requires Ξ to be small in some appropriate sense. The local variance component $\{\Lambda_j\}$ should be such that it can relax the shrinkage effect for those coefficients whose magnitude is large. Now it is easy to see from Lemma 2 that

$$\mathbb{E}\left[\boldsymbol{\beta}|\boldsymbol{X}_{t_1:t_m}, \{\Lambda_j\}, \Xi\right] = (I - S)\hat{\boldsymbol{\beta}}_{MLE},$$

where the shrinkage factor, $S = (I + \varsigma^{-2}\eta\mathcal{K}_0^T\mathcal{K}_0)^{-1}$ with $\eta = diag(\Lambda_1, \Lambda_2, \ldots, \Lambda_m) \otimes \Xi$, and $\hat{\boldsymbol{\beta}}_{MLE} = \Delta(\mathcal{K}_0^T\mathcal{K}_0)^{-1}\mathcal{K}_0\vartheta$ is the standard MLE estimate of $\boldsymbol{\beta}$ based on the likelihood, $L^{EM}(b|\boldsymbol{X}_{t_1:t_m})$. Here we assumed for simplicity that the diffusion coefficient, $\sigma(x) \equiv \varsigma I$, $\varsigma \in \mathbb{R}$. This points to the necessary characteristics of the prior distributions of the hyperparameters, Ξ and Λ_j: (a) $p_{prior}(\Xi)$ should have a sharp peak at 0, and (b) $p_{prior}(\Lambda_j)$ should have heavy tails.

Instead of choosing $d \times d$- dimensional probability distributions as priors for Λ_j and Ξ we set $\Lambda_j = \lambda_j^2 I_{d\times d}$ and $\Xi = \tau^2 I_{d\times d}$ with one-dimensional priors on λ_j^2 and τ^2 satisfying the above criteria. These choices of priors require a much smaller number of hyperparameters, leading to potentially significant savings in computational complexity while allowing an easier-to-implement extension of 1-D global-local priors for multidimensional parameters.

If $\tau^2 = 1$, then an inverse gamma-prior on λ_j^2 leads to (a multidimensional version of) t-prior on the β_j. Although the inverse-gamma is popular as a choice of mixing distribution for the variance

components (like λ_j^2 and τ^2) of normal-scale mixture family of priors, it can be informative in certain cases leading to non-robust estimation of the β_j. Moreover, for an inverse-gamma distribution, $p_{prior}(\tau^2) \to 0$ as $\tau^2 \to 0$. Since $p(\tau^2|\boldsymbol{X}_{t_1:t_m}) \propto p(\boldsymbol{X}_{t_1:t_m}|\tau^2)p_{prior}(\tau^2)$, $p(\tau^2|\boldsymbol{X}_{t_1:t_m}) \overset{\tau^2 \to 0}{\to} 0$ forcing the posterior distribution of τ to biased away from 0. In other words, with an inverse-gamma prior on τ^2, the posterior probability of τ^2 to be near 0 is likely to be small. Thus low probability is assigned to that part of the parameter space where benefits of shrinkage is desired! This has already been pointed out for regression problems by Gelman (2006) (also see Polson and Scott (2012)) and is also true for data from dynamical systems that are of interest in this paper. These issues with the inverse-gamma prior can be mitigated by averaging its scale parameter with another appropriate distribution, e.g. a gamma distribution (see Pérez et al. (2017)). This leads to a scaled $\mathcal{F}$-distribution. The density $\mathcal{F}(\cdot|\nu_1, \nu_2, c)$ of $\mathcal{F}$-distribution (or Beta-prime $(2\nu_1, 2\nu_2)$ distribution with scaling parameter c) with degrees of freedom ν_1, ν_2 and scaling parameter c is given by

$$\begin{aligned} \mathcal{F}(z|\nu_1, \nu_2, c) &= \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_2}{2}\right)\Gamma\left(\frac{\nu_1}{2}\right)c^{\nu_1/2}} z^{\frac{\nu_1}{2}-1}(1+z/c)^{-\frac{\nu_1+\nu_2}{2}} \\ &= \int \mathcal{IG}(z|\nu_2/2, \theta)\, \mathcal{G}(\theta|\nu_1/2, c^{-1})\, d\theta. \end{aligned} \tag{3.8}$$

Elementary formal calculations show that

$$\mathcal{F}(z|\nu_1, \nu_2, c) \overset{z \to 0}{\approx} z^{\nu_1/2-1}, \quad \mathcal{F}(z|\nu_1, \nu_2, c) \overset{z \to \infty}{\approx} z^{-\nu_2/2-1},$$

which in turn show that the first degree of freedom, ν_1, controls the behavior of $\mathcal{F}$-distribution around zero, while the behavior in tails is controlled by the second degree of freedom, ν_2. Choosing a smaller value of $\nu_1 < 2$ will result in a pole at 0, and smaller values of ν_2 will lead to heavier tails. Formal calculations also indicate that

$$\begin{aligned} p_{prior}(\beta_j = 0|\lambda_j^2) &= \int \mathcal{N}(\beta_j = 0|\lambda_j^2\tau^2 I)\mathcal{F}(\tau^2|\nu_1, \nu_2, c)d\tau^2 \\ &\propto \int (\tau^2)^{(\nu_1-d)/2-1}(1+\tau^2/c)^{-(\nu_1+\nu_2)/2}d\tau^2 = \infty \end{aligned}$$

if $\nu_1 \leqslant d$, as the last integral then is proportional to integral of an improper $\mathcal{F}$-density. $\mathcal{F}(\nu_1 = \nu_2 = 1, c = 1)$-prior on λ_j^2 and τ^2 (or equivalently, Half-Cauchy(0,1)-prior on λ_j and τ) leads to the Horseshoe prior (or more precisely, a multidimensional version of it) on β_j. However, in the case of correlated temporal data from dynamical systems, these default choices of $\nu_1 = \nu_2 = 1,\ c = 1$, can result in τ^2 to be near-zero value shrinking all the weights β_j substantially. It might be necessary to adjust the degrees of freedom parameters to counter such strong shrinking force - for example, by using a $\mathcal{F}$-prior on λ_j^2 having heavier tails (that is, lower value of second degree of freedom, ν_2) than $\mathcal{F}(\nu_1 = \nu_2 = 1, c = 1)$ to recover the relevant weights.

The Bayesian hierarchical framework with the above choices is summarized below.

Bayesian hierarchical framework II: Horseshoe-type priors

- $X(t_1 - \Delta) \equiv X(0) \sim \delta_{x_0}$.
- $\{\boldsymbol{X}_{t_1:t_m} \Big| \boldsymbol{\beta}, \alpha\}$ governed by the transition probabilities (3.5), which is the result of Euler-Maruyama approximation, (3.3).
- Independent mean-zero Gaussian priors on the parameters β_i: $\beta_i \sim \mathcal{N}_d(\cdot|0, \lambda_i^2\tau^2 I_d)$, where $\lambda_i^2, \tau^2 \in [0, \infty)$.
- Inverse Gamma priors on the hyperparameters λ_i^2 and τ^2: $\lambda_i^2 \sim \mathcal{IG}(\cdot|\alpha_i, \theta_i)$ for $i = 1, 2, \ldots, m$ and $\tau^2 \sim \mathcal{IG}(\cdot|\alpha^0, \theta^0)$.
- Gamma priors on the hyperparameters θ^0, θ_i: $\theta_i \sim \mathcal{G}(\cdot|\mathfrak{a}, \mathfrak{b})$, for $i = 1, 2, \ldots, m$, and $\theta^0 \sim \mathcal{G}(\mathfrak{a}^0, \mathfrak{b}^0)$.
- Inverse Wishart prior on the hyperparameters $\varsigma\varsigma^T$: $\varsigma\varsigma^T \sim \mathcal{IW}_d(\cdot|n, V)$.

Equation 3.8 leads to easy sampling of the parameters from the posterior distribution via Gibbs sampling. This is summarized in the algorithm below, and the computational details are given in Lemma 2 in the Appendix. The following notations are convenient for descriptions of Algorithm 2 and Lemma 2.

Notation: Let $\mathcal{F}$ denote the σ-field generated by $\boldsymbol{X}_{t_1:t_m} = (X(t_1), X(t_2), \ldots, X(t_m)))$, and the parameters $\boldsymbol{\beta}, \varsigma\varsigma^T, \{\lambda_i^2 : i = 1, 2 \ldots, m\}, \tau^2, \{\theta_i : i = 1, 2, \ldots, m\}, \theta^0$ (viewed as random variables on the same probability space.). Let $\mathcal{F}_{-\boldsymbol{\beta}}$ be the σ-field generated by the above random elements except $\boldsymbol{\beta}$, $\mathcal{F}_{-\{\lambda_i^2\}}$ the σ-field generated by the above random elements except $\{\lambda_i^2 : i = 1, 2, \ldots, m\}$. The σ-fields $\mathcal{F}_{-\varsigma\varsigma^T}, \mathcal{F}_{-\{\theta_i\}}, \mathcal{F}_{-\theta^0,\{\theta_i\}}$, etc are defined similarly.

Algorithm 2: Gibb's algorithm for high frequency data with Horseshoe prior.

Input: The data $\boldsymbol{X}_{t_1:t_m}$, x_0, discretization step Δ, number of iterations L.
Output: $\boldsymbol{\beta}, \lambda_j^2, \tau^2, \varsigma\varsigma^T$ from the posterior density.

while $l < L$ **do**
- Generate $\boldsymbol{\beta}|\mathcal{F}_{-\boldsymbol{\beta}} \sim \mathcal{N}_d(\cdot|\mu, \boldsymbol{C})$ where μ and $\boldsymbol{C}$ are defined by (A.3). Calculate the drift function b by (3.6).
- Generate $\varsigma\varsigma^T|\mathcal{F}_{-\varsigma\varsigma^T} \sim \mathcal{IW}_d(n + m, V_{post})$, where V_{post} is defined by (A.4).
- Generate $\lambda_k^2|\mathcal{F}_{-\{\lambda_i^2\}} \sim \mathcal{IG}\left(\cdot\Big|(d + 2\alpha_k)/2,\ \frac{1}{2}\beta_k^T\beta_k/\tau^2 + \theta_k\right), \quad k = 1, 2, \ldots, m$ independently.
- Generate $\tau^2|\mathcal{F}_{-\tau^2} \sim \mathcal{IG}\left(\cdot\Big|(md + 2\alpha^0)/2,\ \theta^0 + \frac{1}{2}\sum_{k=1}^m \beta_k^T\beta_k/\lambda_k^2\right)$.
- Generate $\{\theta_k\}$ and θ^0 as $\theta_k|\mathcal{F}_{-\theta^0,\{\theta_i\}} \sim \mathcal{G}\left(\cdot|\alpha_k + \mathfrak{a}, \mathfrak{b} + 1/\lambda_k^2\right)$, and $\theta^0|\mathcal{F}_{-\theta^0,\{\theta_i\}} \sim \mathcal{G}\left(\cdot|\alpha^0 + \mathfrak{a}^0, \mathfrak{b}^0 + 1/\tau^2\right)$.
- $l = l + 1$

end

An alternate option would have been to impose independent one-dimensional Horseshoe prior on each component $\beta_{jl}, l = 1, 2, \ldots, d;\ j = 1, 2, \ldots, m$. A version of this prior has previously been used by one of the authors for a multi-outcome regression model in Kundu et al. (2021). There the local shrinkage effects, while varying among individual predictor values, were shared across multiple dimensions of the same predictor, and the global component varied across different dimensions. While these types of priors may be more natural for the multi-outcome regression models of Kundu et al. (2021) to allow more intra-dimensional variability, their use in the context of multidimensional dynamical systems lacks strong justification. Rather the significantly higher number of additional hyperparameters that these priors require will lead to substantial increase in the complexity and run-time of the resulting Gibb's algorithm.

4. Simulation Results

We next demonstrate the effectiveness of our algorithm for four SDE models. The SDEs considered are ergodic with a unique stationary distribution. As mentioned earlier, this is exactly the class of models where we expect our algorithms to work best. Ergodicity will ensure that the SDE will visit the relevant states multiple times. This will lead to a sufficient number of data points corresponding to each such states over a finite-time interval which in turn will result in more accurate learning of the drift function b.

From a discrete path from each of the SDE models, we use our algorithms to generate samples of $\boldsymbol{\beta}$ from the posterior distribution. The (posterior) mean of these $\boldsymbol{\beta}$-samples gives the estimated function $\hat{b}$ via equation (3.6), which is plotted against the true b. We next used multiple samples from the posterior distribution to calculate empirical 95% Bayesian credible bands around b. The plots of these credible bands are given in the pictures below. The corresponding mean square error (MSE) is also reported.

While closeness between $\hat{b}$ and the true b demonstrates the effectiveness of our learning algorithms, a further validation of the algorithm comes from matching the equilibrium (or the stationary) distribution of the estimated SDE with that of the true one. This shows that the behavior of the estimated SDE matches with that of the true SDE at future times — further beyond the time-range of the observed data. This is important as it demonstrates the predictive power of the learned SDE model and shows that the closeness between the true and the estimated drift functions, b and $\hat{b}$, is indeed due to the accuracy of the algorithms and not due to overfitting. The latter despite giving good fit within the time-range of the data would often result in markedly different behaviors of the paths of the corresponding SDEs at unobserved future times. The closeness between the two stationary distributions is assessed through the Kolmogorov metric, $\sup_x |F_{st}(x) - \hat{F}_{st}(x)|$, where F_{st} and $\hat{F}_{st}$ respectively denote the cumulative distribution functions (CDFs) of the stationary distributions of the true and the estimated SDEs. Specifically, the former refers to the SDE driven by the true drift function b and the diffusion parameter ς^2 while the latter corresponds to the SDE driven by their estimated versions $\hat{b}$ and $\hat{\varsigma}^2$.

We used Gaussian kernels for our simulation studies. Specifically, for the 1-D models, we used the kernel $\kappa_0(x, y) = \exp(-(x - y)^2/2)$ and for the multidimensional Michaelis-Menten kinetics in Model 3, we used $\boldsymbol{\kappa}_0 = \kappa_0 I_3$.

For comparison purposes, we also calculated and plotted the estimator given by histogram, k-nearest neighbors and kernel-based methods. Their expressions are given in the Appendix. The

pictures below clearly demonstrate the accuracy of our algorithms over these existing methods.

Model 1: Double-well potential SDE
Our first model is an overdamped Langevin SDE representing the motion of a particle in a double-well potential given by $u(x) = x^4 - 2x^2$. The trajectory of the particle depends on two factors: a (deterministic) driving force $b(x) = -u'(x) = 4(x - x^3)$, and random perturbations modeled by an additive Brownian noise. The potential has two wells (minimum energy states) located at ± 1, and the driving random noise occasionally makes the particle transition from one minima to the other. The dynamics of the particle is thus highly non-linear and the corresponding SDE given by

$$dX(t) = 4X(t)(1 - X^2(t))dt + \varsigma dW(t).$$

Such SDEs are also important in mathematical finance. The two wells lead to a bimodal stationary distribution whose density is given by

$$\pi_{st}(x) \propto \exp\left(\frac{2x^2 - x^4}{2\varsigma^2}\right).$$

Our data points come from the above SDE with $\varsigma = 1$, and we use Algorithm 1 and Algorithm 2 to estimate the entire drift function b, and the diffusion parameter ς. For this we use a scaled $\mathrm{t}(\cdot|\nu = 2, c = 1, \mu = 0)$-prior on the weights β_k (that is, $\beta_k \sim \mathcal{N}(\cdot|0, \lambda_k^2), \lambda_k^2 \sim \mathcal{IG}(1, 2)$) in Algorithm 1 (with inverse-gamma replacing inverse-Wishart), and we use the parameters $\alpha_i = \alpha^0 = \mathfrak{a} = \mathfrak{a}^0 = 1/2, \mathfrak{b} = \mathfrak{b}^0 = 1$ (that is, classical HS prior) for Algorithm 2. For both the algorithms we use $\mathcal{IG}(1, 2)$-prior on the diffusion-parameter ς^2.

Figure 1 gives a visual representation of the performances of the algorithms: (**a**) plots the real drift function b and the estimated $\hat{b}$ in three cases - with no-shrinkage, shrinkage with t and HS priors on the weights; (**b**) and (**c**) plot 95% credible intervals corresponding to t and HS priors, (**d**) plots a histogram of the weights β_k, which shows the effect of shrinkage priors; (**e**) compares the stationary distributions of the SDE with estimated drift function $\hat{b}$ in three cases (no-shrinkage, t and HS shrinkage priors) with the true stationary distribution of the double-well potential SDE; (**f**) shows the corresponding P-P plots.

Figure 1-(**d**) is noteworthy as it shows that both t and HS priors were successful in giving sparse solutions for the weights, β_i, with HS prior producing significantly higher degree of sparsity compared to t-prior as evidenced from much sharper peak of the histogram near 0. At the same time other figures show that both the shrinkage priors lead to almost identical $\hat{b}$ matching the accuracy of the estimate without shrinkage. The MSE and the Kolmogorov metric values in all the cases are in the range 0.27-0.29 and 0.7 - 0.8, respectively.

Better accuracy is expected with more data, which can be either because of higher frequency of observations (that is, lower value of Δ) or more observations over longer time range $[0, T]$.

T	40	40	40	80	60	20
Δ	0.025	0.05	0.1	0.05	0.05	0.05
t-prior	0.3035	0.3966	0.7234	0.2818	0.2971	0.5128
HS-prior	0.3258	0.4193	0.9442	0.2890	0.3362	0.7106

Table 1: MSE of $\hat{b}$ for t and HS priors.

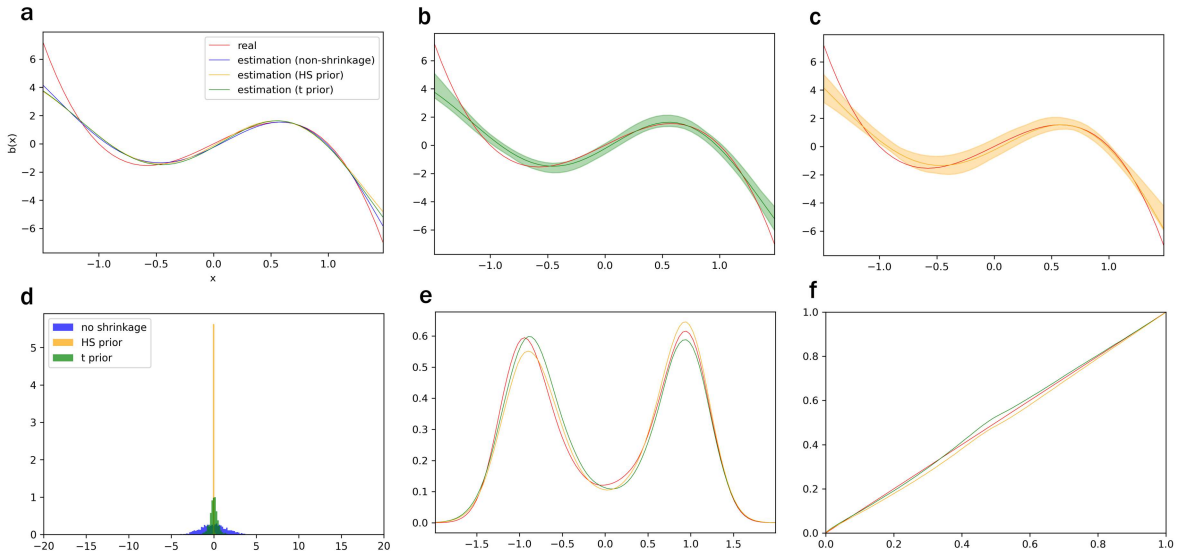

Figure 1: Double-well potential SDE. **a**: comparison of estimated function $\hat{b}$ with true b. **b, c**: estimated function $\hat{b}$ with the shaded areas showing the 95% confidence regions using t and HS priors. **d**: histogram of β_i's. **e**: comparison of the stationary distributions of the SDE driven by estimated $\hat{b}$ and true b. **f**: PP-plots of the stationary distributions of the estimated SDE against that of the original SDE.

T	40	40	40	80	60	20
Δ	0.025	0.05	0.1	0.05	0.05	0.05
t-prior	0.1494	0.2102	0.2047	0.0707	0.0627	0.2702
HS-prior	0.1325	0.2047	0.2486	0.0817	0.0913	0.2869

Table 2: Kolmogorov metric between the CDFs of the stationary distributions of the estimated and true SDEs.

This is corroborated by Table 1 and Table 2, which list the values of MSE and the Kolmogorov metric in two cases - (i) fixed observation-time range $[0, T]$, but increasing Δ, and (ii) fixed observation frequency Δ but increasing time range $[0, T]$.

We finally plot the estimators based on existing methods. The smoothing effect of kernel is visible in the traditional kernel-based estimator. It is clear that these methods cannot match the accuracy of the estimators given by Algorithms 1 and 2.

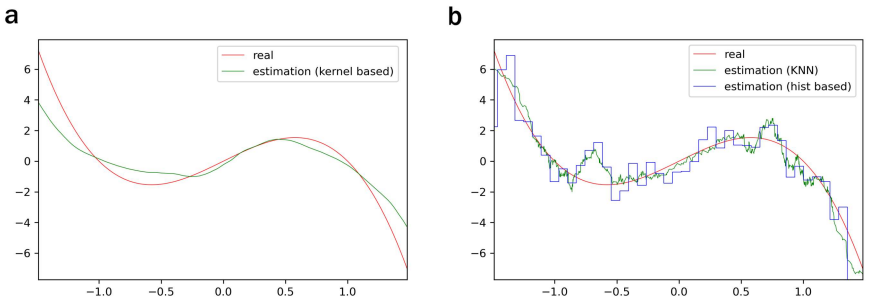

Figure 2: Comparison with other methods **a**: comparison of the estimated function $\hat{b}$ using kernel-based method with true b. **b**: comparison of the estimated function using histogram-based and k-nearest-neighbor (kNN)-based (with $k = 50$) methods.

Model 2: Variant of Double-well potential SDE

Our second model is a variant of the above double-well potential SDE with a multiplicative noise structure. The specific equation is given by

$$dX(t) = X(t)(1 - X^2(t))dt + \varsigma\sqrt{1 + X(t)^2}dW(t).$$

The multiplicative noise adds to the complexity of the already complex nonlinear dynamics of the original double-well process. The stationary density of the SDE is given by

$$\pi_{st}(x) \propto \varsigma^{-2}(1 + x^2)^{2\varsigma^{-2}-1} \exp\left(-x^2/\varsigma^2\right). \tag{4.1}$$

The stationary distribution is bimodal if $\varsigma < 1$, but it becomes unimodal if $\varsigma \geqslant 1$ with sharper peak with increasing ς. We consider two cases, $\varsigma = 1$ and $\varsigma = 0.5$.

Case: $\varsigma = 1$**:** We first consider (discrete) observations from (4.1) with true $\varsigma = 1$, and use Algorithm 1 and Algorithm 2 to estimate the drift function b and the diffusion parameter ς. For Algorithm 1, we use the same t-distribution as the last example. For Algorithm 2, classical HS prior was shrinking all the weights β_k to near 0, and it was necessary to use heavier-tailed distribution on the local variance component λ_k^2 (than $\mathcal{F}(\nu_1 = 1, \nu_2 = 1, c = 1)$-distribution) to counter the strong global shrinkage effect of τ^2. We use $\mathcal{F}(\nu_1 = 1, \nu_2 = 0.3, c = 1)$-distribution on λ_k^2 and the usual $\mathcal{F}(\nu_1 = 1, \nu_2 = 1, c = 1)$-distribution on τ^2, that is, the following values of hyperparameters: $\alpha_i = 0.5, \alpha^0 = \mathfrak{a} = \mathfrak{a}^0 = 1/2, \mathfrak{b} = \mathfrak{b}^0 = 1$. As before, we use $\mathcal{IG}(1, 2)$-prior on the diffusion-parameter ς^2 in both the algorithms. The efficacy of the algorithms is demonstrated in Figure 3. The MSE and the Kolmogorov metric values for cases corresponding to no-shrinkage, t-prior and the above HS-type prior are comparable and are again in the range 0.24-0.27 and about 0.07, respectively. The values of the estimate, $\hat{\varsigma^2}$, given by Algorithm 1 and Algorithm 2 are 0.998 and 0.974, respectively. Again, the global-local setup of a HS-type prior (Algorithm 2) was able to produce significantly higher shrinkage while achieving comparable level of accuracy.

As before, we list in Table 3 and Table 4 the values of MSE and the Kolmogorov metric in two cases - (i) fixed observation-time range $[0, T]$, but increasing Δ, and (ii) fixed observation frequency Δ but increasing time range $[0, T]$. As expected, better accuracy is obtained with more observations, with increasing time range $[0, T]$ of observations being more important than a fixed one with higher frequency of observations (that is smaller Δ). This is natural as data over longer time range reveals more about the behavior of the underlying SDE.

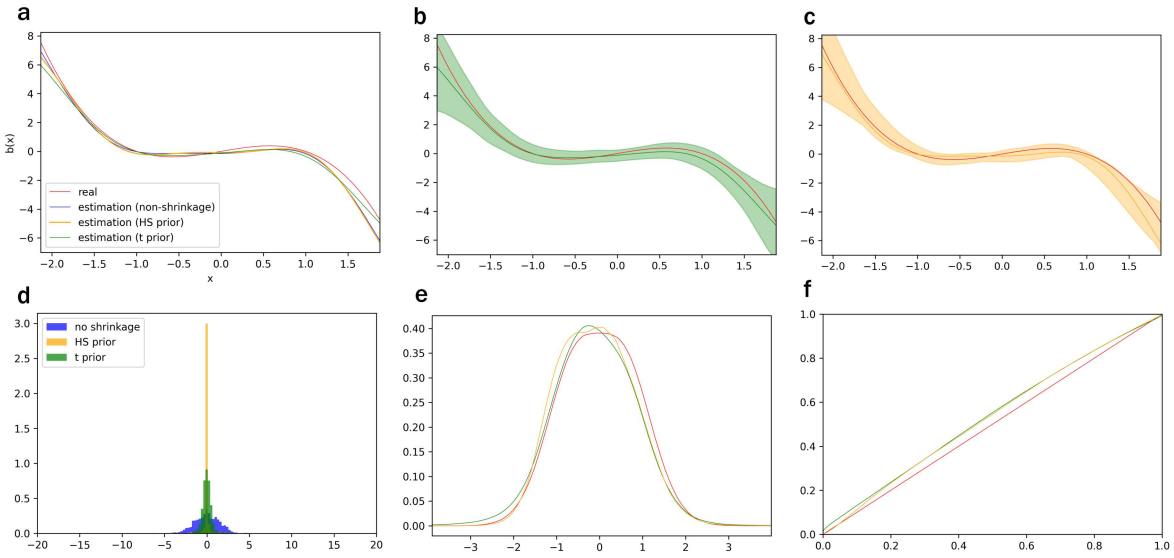

Figure 3: Variant of double-well model ($\varsigma = 1$). Descriptions of **a**, **b**, **c**, **d**, **e**, **f** are similar to that in Figure 1.

T	40	40	40	80	60	20
Δ	0.025	0.05	0.1	0.05	0.05	0.05
t-prior	0.3609	0.4512	0.5632	0.2702	0.3443	0.8265
HS-type-prior	0.3838	0.3972	0.7868	0.2404	0.3319	0.8858

Table 3: MSE with different priors.

T	40	40	40	80	60	20
Δ	0.025	0.05	0.1	0.05	0.05	0.05
t-prior	0.107	0.1091	0.1291	0.071	0.1225	0.1826
HS-prior	0.1187	0.107	0.1259	0.077	0.1268	0.1708

Table 4: Kolmogorov metric between the CDFs with different prior.

Case: $\varsigma = 0.5$**:** We also consider data points from (4.1) with $\varsigma = 0.5$ over the interval $[0, 40]$ (with $\Delta = 0.05$). As mentioned, the true stationary distribution in this case is distinctly bimodal. Bimodality and multiplicative noise make estimation of the drift function b particularly a challenging task. Figure 4 compares the estimated and the true b, and the corresponding stationary distributions. The hyperparameter values used in Algorithm 1 and Algorithm 2 are the same as in the previous case.

The estimated $\hat{b}$ for both the priors match closely with true b (on a large part of the x-axis), and the estimator $\hat{\varsigma^2} = 0.247$ and 0.258, which is almost same as the true $\varsigma^2 = 0.25$. But here Algorithm 1 with t-prior on the weights gives a much more accurate result than Algorithm 2 with

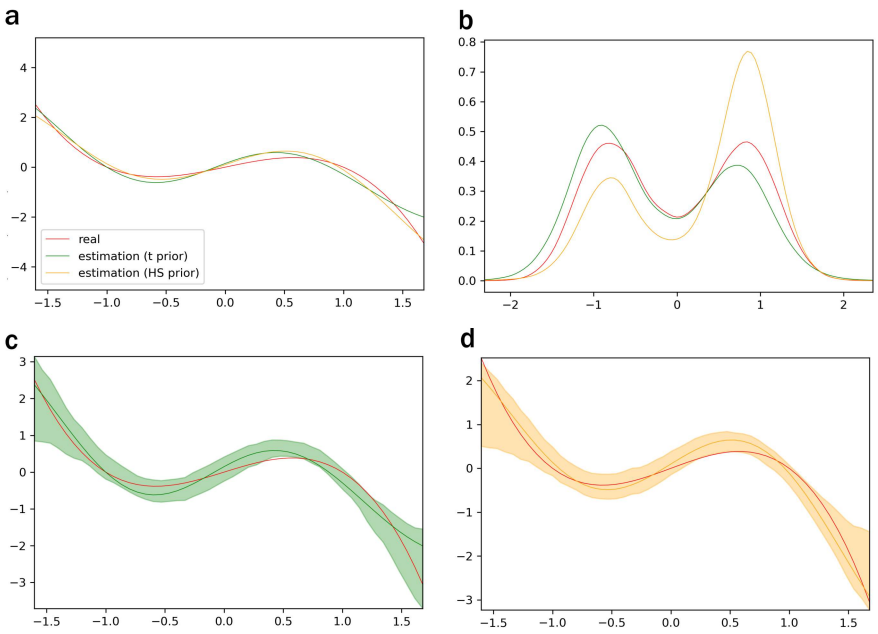

Figure 4: Variant of double-well model ($\varsigma = 0.5$). **a**: comparison of estimated function $\hat{b}$ with true b. **b**: comparison of the stationary distributions of the SDE driven by estimated $\hat{b}$ and true b. **c, d**: estimated function $\hat{b}$ with the shaded areas showing the 95% confidence regions.

the HS-type prior. This is clear from the plots of the different stationary distributions, where HS prior respectively underestimates and overestimates the modes at -1 and 1. The MSE values for estimated $\hat{b}$ corresponding to t and HS priors are respectively 0.051 and 0.04, which are comparable. But the Kolmogorov metric between the CDFs of the true stationary distribution (c.f (4.1)) and the stationary distribution with $\hat{b}$ as the drift in the case of t and HS priors is respectively 0.05 and 0.17 showing the edge that the Algorithm 1 had in this case.

Finally, Figure 5 below shows the estimated b given by the existing methods.

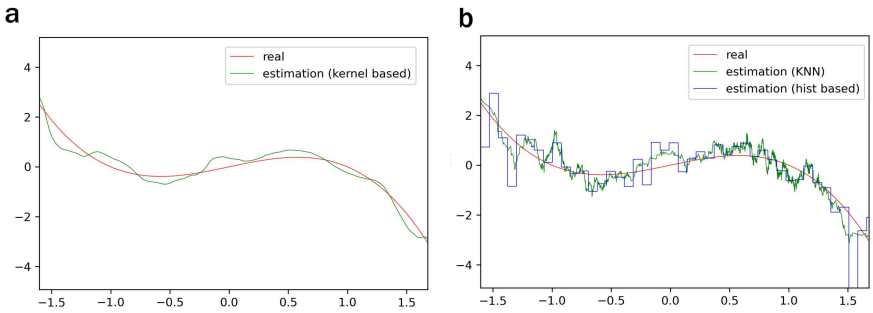

Figure 5: Comparison with other methods **a**: comparison of estimated function $\hat{b}$ using kernel-based method with true b. **b**: comparison of estimated function using histogram-based and k-nearest-neighbor (kNN)-based (with $k = 50$) methods.

Model 3: Michaelis-Menten Kinetics

The Michaelis-Menten is a well-known model in enzymatic kinetics describing the enzymatic substrate conversion process (Michaelis and Menten (2013); Srinivasan (2021)). The reaction system is given by

$$E + S \underset{k_2}{\overset{k_1}{\rightleftharpoons}} ES, \quad ES \underset{k_{m2}}{\overset{k_{m1}}{\rightleftharpoons}} E + P.$$

The full state of the system at time t is given by $X(t) = (X_E(t), X_S(t), X_{ES}(t), X_P)$. The system satisfies the conservation law: $X_E(t) + X_{ES}(t) = X_E(0) + X_{ES}(0)$. This gives a reduced 3-dimensional state which will still be denoted by $X(t) = (X_E(t), X_S(t), X_P)$. The differential equation describing the dynamics is governed by the drift function

$$b(x) = (-k_1 x_E x_S - k_{m2} x_E x_P + (k_{m1} + k_2) x_{ES}, -k_1 x_E x_S + k_{m1} x_{ES}, k_2 x_{ES} - k_{m2} x_E x_P).$$

Given a set of discrete observations from a stochastic version of this differential equation driven by additive Brownian noise $\varsigma_{3\times 3} B$, with $\varsigma = 0.1I$ over time-range $[0, 40]$ generated by taking $\Delta = 0.04$ and the conservation constant, $X_E(0) + X_{ES}(0) = 2$, we use Algorithm 1 and Algorithm 2 to estimate the entire drift function b and the (constant) diffusion matrix ς. For Algorithm 1, we use the hyperparameter values, $\nu = 5$, $U = 8I$, and $\mathcal{IW}_3(1 + \text{dim}, V = 2I_{3\times 3})$-prior (where, dimension, dim$= 3$) on $\varsigma\varsigma^T$. For Algorithm 1 we use the (multidimensional version of) classical HS prior, and the same inverse-wishart prior on $\varsigma\varsigma^T$. The MSE values in both cases came out to be about 0.004 (specifically, 0.00414 for HS and 0.00431 for t). Figure 6 - **a**, **b** and **c**, respectively, plots the first, second and third component of both the estimator $\hat{b}$ and the true b when HS-prior is used with z-coordinate fixed at 1.073.

The estimated diffusion matrix (via Algorithm 2) is given by

$$\hat{\varsigma\varsigma}^T = \begin{bmatrix} 0.01210109 & 0.0001914 & -0.00020334 \\ 0.0001914 & 0.01170018 & 0.00014556 \\ -0.00020334 & 0.00014556 & 0.01122171 \end{bmatrix}$$

which is very close to the true $\varsigma\varsigma^T = 0.01I$. The corresponding numbers for Algorithm 1 are very similar.

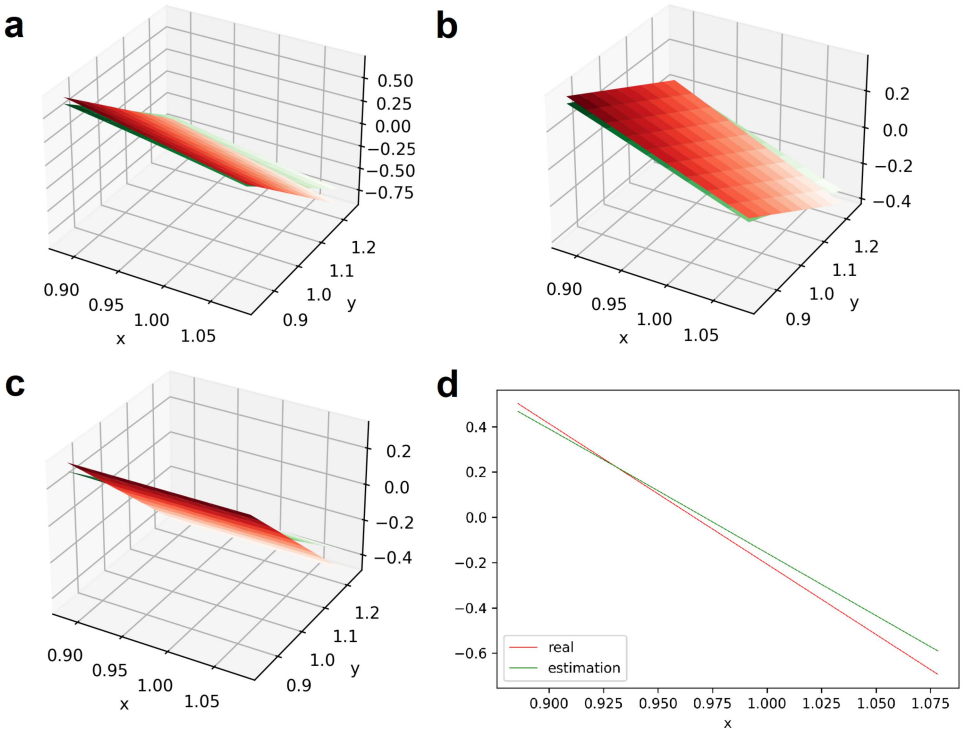

Figure 6: Michaelis-Menten Kinetics model with HS-prior. **a**, **b**, **c** show the plots of first, second and third component of the functions $\hat{b}$ and b with z fixed at 1.073. **d** shows a two-dimensional slice of **a** at $y = 1.060$.

5. Discussion

The paper presents a novel theoretical and computational paradigm for stochastic dynamic models which, on account of its generalizability, can potentially find its way to several interesting applications. We study two areas — (a) a class of infinite-dimensional optimization problems, which is broader than what the classical Representer Theorem covers, (b) Bayesian approach to nonparametric inference of stochastic dynamical systems. To our knowledge, this is the first instance of the merging of Bayesian methods, RKHS theory and stochastic differential equations into a single unified platform. The use of the resulting algorithms on data from well-known SDEs amply demonstrates their ability to learn the true drift functions to a high degree of accuracy. Specifically, their reliable prediction of long term dynamics beyond the range of data points is a strong testament to this fact. The accuracy measures obtained under the M-M kinetics model lend strong credence to the relevance of this approach for multivariate settings.

The hierarchical structure of the Bayesian framework makes the resulting inference scheme computationally scalable while opening the door to several model extensions. For instance, a semi-parametric model variant could be easily implemented in instances where the stochastic dynamics is known only partially. It is also of interest to study the effectiveness of other types of shrinkage priors in this context. These extensions would also broadly benefit from the convenience of Gibbs sampling schemes similar to the ones showcased in this article. The 'divide and conquer'

approach intrinsically encoded in such schemes would typically allow for multiple computational conveniences, like parallel computation as and when required.

Several ongoing works are focusing on more general models including sparse and noisy datasets, dynamical systems with jumps and multiscale stochastic systems, each of which has its own unique challenges. For example, for SDE models with noisy data the expansion in Theorem 15 does not directly hold as the actual trajectory of the underlying SDE is never observed. The generality of the optimization results in the first part of the paper will play a key role in these cases.

Acknowledgments

Research of A. Ganguly is supported in part by NSF DMS - 1855788 and Louisiana Board of Regents through the Board of Regents Support Fund (contract number: LEQSF(2016-19)-RD-A-04). Research of J. Zhou is supported in part by NSF DMS - 1855788.

Appendix A. Proofs and auxiliary results

The lemma below characterizes uniformly p.d. operators. A proof is given for completeness.

Lemma 1 *Let $Q \in L(\mathcal{H}, \mathcal{H})$ be a self-adjoint, p.d. operator. Then the following are equivalent:*

(i) Q is uniformly p.d. *(ii)* $\mathrm{Range}(Q)$ *is closed.* *(iii) Q is surjective.*

Proof (i) $\Rightarrow$ (ii): Suppose that $\{Qh_n\} \subset \mathrm{Range}(Q)$ such that $Qh_n \to g$ as $n \to \infty$. We need to show that $g = Qh$ for some $h \in \mathcal{H}$. Notice that in particular $\{Qh_n\}$ is Cauchy. Since Q is uniformly p.d., there exists $\lambda > 0$ such that $\langle Qh, h\rangle \geqslant \lambda \|h\|^2$. This implies $\{h_n\}$ is also a Cauchy sequence, and hence by completeness of $\mathcal{H}$ there exists h such that $h_n \to h$. By continuity of Q, we then have $Qh_n \to Qh$, and therefore, $Qh = g$.

(ii) $\Rightarrow$ (iii): Suppose $\mathrm{Range}(Q)^\perp \neq \{0\}$. Let $0 \neq y \in \mathrm{Range}(Q)^\perp$. Now $Qy \in \mathrm{Range}(Q)$; hence $\langle Qy, y\rangle = 0$. But since Q is p.d. this means that $y = 0$; in other words, $\mathrm{Range}(Q)^\perp = \{0\}$. Since $\mathrm{Range}(Q)$ is closed by the hypothesis, we get from $\mathcal{H} = \mathrm{Range}(Q) \oplus \mathrm{Range}(Q)^\perp$ that $\mathrm{Range}(Q) = \mathcal{H}$.

(iii) $\Rightarrow$ (i): Observe that since Q is self-adjoint and p.d., $\langle h', h\rangle_Q \stackrel{def}{=} \langle Qh', h\rangle$ defines a valid inner product. Furthermore, since Q is surjective, Q^{-1} is a bounded linear operator, that is, $Q^{-1} \in L(\mathcal{H}, \mathcal{H})$. Now by Cauchy-Schwartz inequality, $|\langle h', h\rangle_Q| \leqslant \|h'\|_Q \|h\|_Q$. Taking $h' = Q^{-1}h$, we get $\|h\| \leqslant \|Q^{-1}\|^{1/2} \|h\|_Q$ which establishes (i). ■

Proof [Lemma 5] Define $G^* \stackrel{def}{=} \inf_{h\in\mathcal{H}} G(h)$ and observe that if $G \equiv \infty$, the assertion is trivially true as then $G^* = \infty$, and any $h \in \mathcal{H}$ solves the minimization problem. So we assume that $G(h) < \infty$ for some $h \in \mathcal{H}$. Then $G^* < \infty$ (G^* still could be $-\infty$), and there exists a sequence $\{h_n\}$ such that $G(h_n) \to G^*$, as $n \to \infty$. Notice that this implies the sequence $\{\|h_n\|\}$ is bounded. Indeed, if this is not true then $\limsup_{n\to\infty} \|h_n\| = \infty$. But the hypothesis on G then implies that $G^* = \limsup_{n\to\infty} G(h_n) = \infty$, which contradicts the fact that $G^* < \infty$. Consequently, by Banach-Alaoglu (and Eberlein-Smulian theorem) there exists an $h^* \in \mathcal{H}$ and a subsequence $\{n_k\}$

such that $h_{n_k} \xrightarrow{w} h^*$. By the weak sequential l.s.c. of G we conclude

$$G^* = \lim_{k\to\infty} G(h_{n_k}) \geqslant G(h^*) \geqslant \inf_{h\in\mathcal{H}} G(h) = G^*.$$

This proves that the infimum of G is attained at h^*. ■

Proof [Theorem 6] (i) Fix $h \in \mathcal{H}$. Write $h = \mathcal{P}_{\mathcal{M}}h + (I - \mathcal{P}_{\mathcal{M}})h$. Next notice that since Q is self-adjoint,

$$\begin{aligned}\langle Qh, h\rangle &= \langle Q\mathcal{P}_{\mathcal{M}}h + Q(I-\mathcal{P}_{\mathcal{M}})h, \mathcal{P}_{\mathcal{M}}h + (I-\mathcal{P}_{\mathcal{M}})h\rangle \\ &= \langle Q\mathcal{P}_{\mathcal{M}}h, \mathcal{P}_{\mathcal{M}}h\rangle + 2\langle Q(I-\mathcal{P}_{\mathcal{M}})h, \mathcal{P}_{\mathcal{M}}h\rangle + \langle Q(I-\mathcal{P}_{\mathcal{M}})h, (I-\mathcal{P}_{\mathcal{M}})h\rangle \\ &= \langle Q\mathcal{P}_{\mathcal{M}}h, \mathcal{P}_{\mathcal{M}}h\rangle + \langle Q(I-\mathcal{P}_{\mathcal{M}})h, (I-\mathcal{P}_{\mathcal{M}})h\rangle\end{aligned}$$

because $\langle Q(I-\mathcal{P}_{\mathcal{M}})h, \mathcal{P}_{\mathcal{M}}h\rangle = \langle (I-\mathcal{P}_{\mathcal{M}})h, Q\mathcal{P}_{\mathcal{M}}h\rangle = 0$, as $(I-\mathcal{P}_{\mathcal{M}})h \in \mathcal{M}^\perp$ and $Q\mathcal{P}_{\mathcal{M}}h \in \mathcal{M}$ (because of the assumption (c), $Q\mathcal{M} \subset \mathcal{M}$). Since Q is p.s.d., it follows that $\langle Qh, h\rangle \geqslant \langle Q\mathcal{P}_{\mathcal{M}}h, \mathcal{P}_{\mathcal{M}}h\rangle$ with equality only when $(I-\mathcal{P}_{\mathcal{M}})h \in \mathcal{N}_Q$.

Since $F(h,\cdot) \geqslant F(\mathcal{P}_{\mathcal{M}}h,\cdot)$ and $F(h,\cdot)$ is non-decreasing by assumptions (a) and (b), we have

$$F\left(h, \langle Qh,h\rangle^{1/2}\right) \geqslant F\left(\mathcal{P}_{\mathcal{M}}h, \langle Qh,h\rangle^{1/2}\right) \geqslant F\left(\mathcal{P}_{\mathcal{M}}h, \langle Q\mathcal{P}_{\mathcal{M}}h, \mathcal{P}_{\mathcal{M}}h\rangle^{1/2}\right). \tag{A.1}$$

This proves both the first and the second assertions of (i). If $F(h,\cdot)$ is strictly increasing, then the second inequality in (A.1) is strict when $(I-\mathcal{P}_{\mathcal{M}})h \notin \mathcal{N}_Q$. Now $(I-\mathcal{P}_{\mathcal{M}})h \in \mathcal{M}^\perp$. Therefore, if $\mathcal{N}_Q \subset \mathcal{M}$, or equivalently, $\mathcal{N}_Q \cap \mathcal{M}^\perp = \{0\}$, then the second inequality in (A.1) is strict if and only if $h \neq \mathcal{P}_{\mathcal{M}}h$. Consequently, if $h^* \in \mathcal{H}$, is a global minimizer of (2.4), we must have $h^* = \mathcal{P}_{\mathcal{M}}h^*$, or equivalently, $h^* \in \mathcal{M}$. This proves the last part of (i).

(ii) We prove the statement by contradiction. Let h_0 be a local minimizer. Then there exists a $r > 0$ such that $F\left(h_0, \langle Qh_0,h_0\rangle^{1/2}\right) \leqslant F\left(h, \langle Qh,h\rangle^{1/2}\right)$ for all $h \in B(h_0, r)$. Suppose that $h_0 \notin \mathcal{M}$. Then $h_0 \neq \mathcal{P}_{\mathcal{M}}h_0$. Consequently, by the previous proof $\langle Qh_0, h_0\rangle > \langle Q\mathcal{P}_{\mathcal{M}}h_0, \mathcal{P}_{\mathcal{M}}h_0\rangle$. For $0 \leqslant \delta \leqslant 1$, define $h_\delta = \delta\mathcal{P}_{\mathcal{M}}h_0 + (1-\delta)h_0$. Note that by convexity of the mapping $h \to \langle Qh,h\rangle^{1/2}$, for any $0 < \delta < 1$,

$$\langle Qh_\delta, h_\delta\rangle^{1/2} \leqslant \delta\langle Q\mathcal{P}_{\mathcal{M}}h_0, \mathcal{P}_{\mathcal{M}}h_0\rangle^{1/2} + (1-\delta)\langle Qh_0,h_0\rangle^{1/2} < \langle Qh_0,h_0\rangle^{1/2}.$$

Thus for any $0 < \delta < 1$ by almost quasiconvexity of $F(\cdot, u)$ (c.f. Definition 3),

$$\begin{aligned}F\left(h_\delta, \langle Qh_\delta,h_\delta\rangle^{1/2}\right) < F\left(h_\delta, \langle Qh_0,h_0\rangle^{1/2}\right) &\leqslant F\left(h_0, \langle Qh_0,h_0\rangle^{1/2}\right) \vee F\left(\mathcal{P}_{\mathcal{M}}h_0, \langle Qh_0,h_0\rangle^{1/2}\right) \\ &= F\left(h_0, \langle Qh_0,h_0\rangle^{1/2}\right).\end{aligned} \tag{A.2}$$

The last equality is because $F\left(h_0, \langle Qh_0,h_0\rangle^{1/2}\right) \geqslant F\left(\mathcal{P}_{\mathcal{M}}h_0, \langle Qh_0,h_0\rangle^{1/2}\right)$ due to the assumption on F. Now notice that $\|h_\delta - h_0\| = \delta\|(I-\mathcal{P}_{\mathcal{M}})h_0\| < r$ for sufficiently small δ, and hence $F\left(h_0, \langle Qh_0,h_0\rangle^{1/2}\right) \leqslant F\left(h_\delta, \langle Qh_\delta,h_\delta\rangle^{1/2}\right)$ for sufficiently small δ. But that is a contradiction to (A.2).

(iii) is essentially a standard result in convex optimization. ■

Proof [**Corollary 8**] (2.7) follows from Theorem 6-(i). The fact that M_0 is nonempty (existence of minimizer) is a direct consequence of Lemma 5 applied in the setting of Hilbert subspace $\mathcal{M}$ (recall that $\mathcal{M}$ is closed). To see this we start by noting that the mapping $h \in \mathcal{M} \to \langle Qh, h\rangle^{1/2}$ is weakly l.s.c. This is because the sublevel sets $\left\{h \in \mathcal{M} : \langle Qh, h\rangle^{1/2} \leqslant a\right\} = \left\{h \in \mathcal{M} : \langle Qh, h\rangle \leqslant a^2\right\}$ are weakly closed since they are strongly closed (as the mapping $h \in \mathcal{M} \to \langle Qh, h\rangle$ is strongly continuous) and convex (due to convexity of $h \in \mathcal{M} \to \langle Qh, h\rangle$). Since $J : [0, \infty) \to [0, \infty)$ is l.s.c. and increasing, the (composition) mapping $h \in \mathcal{M} \to J\left(\langle Qh, h\rangle^{1/2}\right)$ is also weakly l.s.c. Hence, because of the hypothesis that F_0 is weakly sequentially l.s.c., the mapping $h \in \mathcal{M} \to F\left(h, \langle Qh, h\rangle^{1/2}\right)$ is weakly sequentially l.s.c.

Now clearly (b) implies that $\limsup_{h\in\mathcal{M},\ \|h\|\to\infty} F\left(h, \langle Qh, h\rangle^{1/2}\right) = \infty$. If (a) holds instead of (b), then we just need to observe that $\limsup_{h\in\mathcal{M},\ \|h\|\to\infty} J(\langle Qh, h\rangle^{1/2}) = \infty$. This follows as for some constant $\lambda > 0$, $\langle Qh, h\rangle \geqslant \lambda\|h\|^2$ for all $h \in \mathcal{M}$ (as $Q\big|_{\mathcal{M}}$ is uniformly p.d.) and $\limsup_{u\to\infty} J(u) = \infty$. Since F_0 is bounded below, $\limsup_{h\in\mathcal{M},\ \|h\|\to\infty} F(h, \langle Qh, h\rangle^{1/2}) = \infty$. In either case, the assertion follows from Lemma 5. ■

Recall the notations described before Algorithm 2.

Lemma 2 *Suppose that the joint distribution of $\boldsymbol{X}_{t_1:t_m} = (X(t_1), X(t_2), \ldots, X(t_m)))$ given the parameters $\boldsymbol{\beta}$ and $\varsigma\varsigma^T$ is described by the transition probabilities* (3.5)*. Assume that*

- $\beta_i|\lambda_i^2, \tau^2 \sim \mathcal{N}_d(\cdot|0, \lambda_i^2\tau^2 I)$,
- $\lambda_i^2|\theta_i \sim \mathcal{IG}(\cdot|\alpha_i, \theta_i)$ *for* $i = 1, 2, \ldots, m$, $\tau^2|\theta^0 \sim \mathcal{IG}(\cdot|\alpha^0, \theta^0)$*;*
- $\theta^0, \theta_i, i = 1, 2, , \ldots, m$ *are independent, and for each* $i = 1, 2, \ldots, m$, $\theta_i \sim \mathcal{G}(\cdot|\mathfrak{a}, \mathfrak{b})$ *and* $\theta^0 \sim \mathcal{G}(\cdot|\mathfrak{a}^0, \mathfrak{b}^0)$*;*
- $\varsigma\varsigma^T$*:* $\varsigma\varsigma^T \sim \mathcal{IW}_d(\cdot|n, V)$.

Then

(i) $\boldsymbol{\beta}|\mathcal{F}_{-\boldsymbol{\beta}} \sim N(\cdot|\mu, \boldsymbol{C})$ *where*

$$\begin{aligned} \boldsymbol{C}^{-1} &= \Delta\boldsymbol{\mathcal{K}}_0^T\boldsymbol{D}\boldsymbol{\mathcal{K}}_0 + \eta^{-1}, \quad \mu = \boldsymbol{C}\boldsymbol{\mathcal{K}}_0^T\boldsymbol{D}\boldsymbol{\vartheta} \\ \boldsymbol{D}_{dm\times dm} &= diag\left((\sigma\sigma^T(X(t_1)))^{-1}, (\sigma\sigma^T(X(t_2)))^{-1}, \ldots, (\sigma\sigma^T(X(t_m)))^{-1}\right) \\ \eta_{dm\times dm} &= diag\left(\lambda_1^2\tau^2, \lambda_2^2\tau^2, \ldots, \lambda_m^2\tau^2\right) \otimes I_d \\ \boldsymbol{\vartheta}_{dm\times 1} &= vec_{d\times m}\left(X(t_1) - x_0, X(t_2) - X(t_1), \ldots, X(t_m) - X(t_{m-1})\right), \end{aligned} \tag{A.3}$$

and $\boldsymbol{\mathcal{K}}_0 = ((\boldsymbol{\kappa}_0(X(t_i), X(t_j))))$ *is the Gram matrix associated with* $\boldsymbol{\kappa}_0$.

(ii) $\varsigma\varsigma^T|\mathcal{F}_{-\varsigma\varsigma^T} \sim \mathcal{IW}_d(n + m, V_{post})$*, where*

$$V_{post} = \Delta^{-1}\sum_{k=0}^{m-1}\left(\sigma_0(X(t_k))\right)^{-1}\left(\vartheta_{k+1} - b(X(t_k)\Delta)(\vartheta_{k+1} - b(X(t_k)\Delta)^T\left(\sigma_0^T(X(t_k))\right)^{-1} + V. \tag{A.4}$$

(iii) Conditioned on $\mathcal{F}_{-\{\lambda_i^2\}}$, $\lambda_k^2,\ k=1,2,\ldots,m$ *are independent, and*

$$\lambda_k^2|\mathcal{F}_{-\{\lambda_i^2\}} \ \sim\ \mathcal{IG}\left(\cdot\Big|(d+2\alpha_k)/2,\ \frac{1}{2}\beta_k^T\beta_k/\tau^2+\theta_k\right).$$

(iv) $\tau^2|\mathcal{F}_{-\tau^2} \ \sim\ \mathcal{IG}\left(\cdot\Big|(md+2\alpha^0)/2,\ \theta^0+\frac{1}{2}\sum_{k=1}^m\beta_k^T\beta_k/\lambda_k^2\right)$.

(v) Conditioned on $\mathcal{F}_{-\theta^0,\{\theta_i\}}$, $\theta^0,\theta_k,\ k=1,2,\ldots,m$ *are independent*

$$\theta_k|\mathcal{F}_{-\theta^0,\{\theta_i\}} \ \sim\ \mathcal{G}\left(\cdot|\alpha_k+\mathfrak{a},\mathfrak{b}+1/\lambda_k^2\right),\quad \theta|\mathcal{G}_{-\theta^0,\{\theta_i\}} \ \sim\ \mathcal{G}\left(\cdot|\alpha^0+\mathfrak{a}^0,\mathfrak{b}^0+1/\tau^2\right).$$

Proof By a slight abuse of notation, we use f as a generic symbol for various conditional densities below. Notice that

$$f\left(\boldsymbol{\beta}|\mathcal{F}_{-\boldsymbol{\beta}}\right)\propto\exp\left\{-\frac{1}{2}\left(\mathcal{E}_0+\sum_{k=1}^m\beta_k^T\beta_k/(\lambda_k^2\tau^2)\right)\right\},$$

where

$$\begin{aligned}\mathcal{E}_0&=\Delta^{-1}\sum_{k=0}^{m-1}\left[\left(\vartheta_{k+1}-\Delta b(X(t_k))\right)^T\left(\sigma\sigma^T(X(t_k))\right)^{-1}\left(\vartheta_{k+1}-\Delta b(X(t_k))\right)\right]\\&=\sum_{k=0}^{m-1}\Delta^{-1}\vartheta_{k+1}^T\left(\sigma\sigma^T(X(t_k))\right)^{-1}\vartheta_{k+1}+\Delta b^T(X(t_k)\left(\sigma\sigma^T(X(t_k))\right)^{-1}b(X(t_k))\\&\quad-2\vartheta_{k+1}^T\left(\sigma\sigma^T(X(t_k))\right)^{-1}b(X(t_k)).\end{aligned}$$

Here $\vartheta_{k+1}=X(t_{k+1})-X(t_k)$, and recall that $b(x)=\sum_{k=1}^m\boldsymbol{\kappa}_0(x,X(t_k))\beta_k$. Now

$$b(X(t_k))=\sum_{j=1}^m\boldsymbol{\kappa}_0(X(t_k),X(t_j))\beta_j=\boldsymbol{\mathcal{K}}_0(X(t_k),*)\boldsymbol{\beta},$$

and hence

$$\begin{aligned}\sum_{k=0}^{m-1}\vartheta_{k+1}^T\left(\sigma\sigma^T(X(t_k))\right)^{-1}b(X(t_k))&=\boldsymbol{\vartheta}^T\boldsymbol{D}\boldsymbol{\mathcal{K}}_{\mathbf{0}}\boldsymbol{\beta}\\\sum_{k=0}^{m-1}b^T(X(t_k)\left(\sigma\sigma^T(X(t_k))\right)^{-1}b(X(t_k))&=\boldsymbol{\beta}^T\boldsymbol{\mathcal{K}}_{\mathbf{0}}^T\boldsymbol{D}\boldsymbol{\mathcal{K}}_{\mathbf{0}}\boldsymbol{\beta}.\end{aligned}$$

Since $\sum_{k=1}^m\beta_k^T\beta_k/(\lambda_k^2\tau^2)=\boldsymbol{\beta}^T\boldsymbol{\eta}^{-1}\boldsymbol{\beta}$, it follows that

$$f\left(\boldsymbol{\beta}|\mathcal{F}_{-\boldsymbol{\beta}}\right)=N(\boldsymbol{\beta}|\mu,\boldsymbol{C}),$$

where

$$\boldsymbol{C}^{-1}=\Delta\boldsymbol{\mathcal{K}}_{\mathbf{0}}^T\boldsymbol{D}\boldsymbol{\mathcal{K}}_{\mathbf{0}}+\boldsymbol{\eta}^{-1},\quad \mu=\boldsymbol{C}\boldsymbol{\mathcal{K}}_{\mathbf{0}}\boldsymbol{D}\boldsymbol{\vartheta}$$

with $\boldsymbol{D}$ as in (A.3). Next note that

$$f\left(\varsigma\varsigma^T|\mathcal{F}_{-\varsigma\varsigma^T}\right) \propto \det\left((\varsigma\varsigma^T)\right)^{-m/2}\exp\left\{-\frac{1}{2}\mathcal{E}_0\right\}\det\left((\varsigma\varsigma^T)\right)^{\frac{-(n+d+1)}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(V(\varsigma\varsigma^T)^{-1}\right)\right\}$$
$$= \det\left((\varsigma\varsigma^T)\right)^{\frac{-(n+m+d+1)}{2}}\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big[\Big(\Delta^{-1}\sum_{k=0}^{m-1}\sigma_0^{-1}(X(t_k))(\vartheta_{k+1}-\Delta b(X(t_k)))$$
$$\times(\vartheta_{k+1}-\Delta b(X(t_k)))^T\left(\sigma_0^T(X(t_k))\right)^{-1}+V\Big)(\varsigma\varsigma^T)^{-1}\Big]\Big\},$$

which proves the assertion. Next note that

$$f\left(\lambda_1^2,\lambda_2^2,\ldots,\lambda_m^2|\mathcal{F}_{-\{\lambda_i^2\}}\right) \propto (\tau^2)^{-md/2}\prod_{k=1}^m(\lambda_k^2)^{-d/2}\exp\left\{-\frac{1}{2}(\lambda_k^2\tau^2)^{-1}\beta_k^T\beta_k\}\right\}$$
$$\times\prod_{k=1}^m(\lambda_k^2)^{-(\alpha_k+1)}\exp\left\{-\frac{\theta_k}{\lambda_k^2}\right\}$$
$$\propto \prod_{k=1}^m(\lambda_k^2)^{-(d+2\alpha_k)/2-1}\exp\left\{-\left(\frac{1}{2}\beta_k^T\beta_k/\tau^2+\theta_k\right)/\lambda_k^2\right\},$$

which proves the assertion. Similarly,

$$f\left(\tau^2|\mathcal{F}_{-\tau^2}\right) = (\tau^2)^{-md/2}\prod_{k=1}^m(\lambda_k^2)^{-d/2}\exp\left\{-\frac{1}{2}(\lambda_k^2\tau^2)^{-1}\beta_k^T\beta_k\}\right\}\times(\tau^2)^{-(\alpha^0+1)}\exp\left\{-\frac{\theta^0}{\tau^2}\right\}$$
$$\propto (\tau^2)^{-(md+2\alpha^0)/2-1}\exp\left\{-\left(\theta^0+\frac{1}{2}\sum_{k=1}^m\beta_k^T\beta_k/\lambda_k^2\right)\Big/\tau^2\right\},$$

and (iv) follows. Finally notice that

$$f\left(\theta,\theta_1,\theta_2,\ldots,\theta_m|\mathcal{F}_{-\theta,\{\theta_i\}}\right) \propto \prod_{k=1}^m\theta_k^{\alpha_k}(\lambda_k^2)^{-(\alpha_k+1)}\exp\left\{-\frac{\theta_k}{\lambda_k^2}\right\}\times\theta^{\alpha^0}(\tau^2)^{-(\alpha^0+1)}\exp\left\{-\frac{\theta^0}{\tau^2}\right\}$$
$$\times\prod_{k=1}^m\theta_k^{\mathfrak{a}-1}\exp\left\{-\mathfrak{b}\theta_k\right\}\times(\theta^0)^{\mathfrak{a}^0-1}\exp\left\{-\mathfrak{b}^0\theta^0\right\}$$
$$\propto \prod_{k=1}^m\theta_k^{\alpha_k+\mathfrak{a}-1}\exp\left\{-(\mathfrak{b}+1/\lambda_k^2)\theta_k\right\}\times(\theta^0)^{\alpha^0+\mathfrak{a}^0-1}\exp\left\{-(\mathfrak{b}^0+1/\tau^2)\theta\right\}$$

which proves (v).

■

Appendix B. Some existing estimators

Histogram and nearest-neighbor-based estimators: First, the range $[\min(\boldsymbol{X}_{t_1:t_m}),\max(\boldsymbol{X}_{t_1:t_m})]$ into n bins $\{B_i\equiv[\min(\boldsymbol{X}_{t_1:t_m})+(i-1)l,\min(\boldsymbol{X}_{t_1:t_m})+il) : i=1,2,\ldots,n\}$ with equal length

$l = (\max(\boldsymbol{X}_{t_1:t_m}) - \min(\boldsymbol{X}_{t_1:t_m}))/n$. The histogram-based regression estimator $\hat{b}^{hist}(x)$ is then defined as

$$b^{hist}(x) = \frac{1}{\Delta} \sum_{i=1}^{n} \left(1_{B_i}(x) \frac{\sum_{j=1}^{m-1} 1_{B_i}(X(t_j))(X(t_{j+1}) - X(t_j))}{\sum_{j=1}^{m-1} 1_{B_i}(X(t_j))} \right).$$

In other words the weight that is assigned to each x lying in the bin B_i is the average of all $\frac{1}{\Delta}(X(t_{j+1}) - X(t_j))$ such that $X(t_j))$ falls into B_i (Tuckey (1961); Lamouroux and Lehnertz (2009)). More sophisticated approaches rely on replacing the mean of the bins with the mean of the k-nearest neighbors (Hegger and Stock (2009)).

Traditional kernel-based estimator: This method uses traditional Nadaraya-Watson type estimates to circumvent assigning a simple average to each value of x (Lamouroux and Lehnertz (2009)). Specifically, the kernel-based regression estimator $\hat{b}^{ker}$ is given by

$$\hat{b}^{ker}(x) = \frac{1}{\Delta(m-1)} \sum_{j=1}^{m-1} w_{h,j}(x)(X(t_{j+1}) - X(t_j)),$$

where $w_{h,j}(x)$ is the Nadaraya–Watson-estimator,

$$w_{h,j}(x) = \frac{K_h(x - X(t_j))}{(m-1)^{-1} \sum_{j=1}^{m-1} K_h(x - X(t_j))},$$

and h is the bandwidth of the kernel. (Lamouroux and Lehnertz (2009)) suggested the Epanechnikov kernel, $K_h(x) = \max(0, \frac{3}{4\sqrt{5}} h^{-1}(1 - (\frac{(xh^{-1})^2}{5}))$, which is what we used in Figures 2 and 5.

References

Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Cédric Archambeau and Manfred Opper. Approximate inference for continuous-time Markov processes. In *Bayesian time series models*, pages 125–140. Cambridge Univ. Press, Cambridge, 2011.

Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O. Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):333–382, 2006. ISSN 1369-7412. With discussions and a reply by the authors.

J. P. N. Bishwal. The Bernstein-von Mises theorem and spectral asymptotics of Bayes estimators for parabolic SPDEs. *J. Aust. Math. Soc.*, 72(2):287–298, 2002. ISSN 1446-7887.

Jaya P. N. Bishwal. *Parameter estimation in stochastic differential equations*, volume 1923 of *Lecture Notes in Mathematics*. Springer, Berlin, 2008. ISBN 978-3-540-74447-4.

Jaya P. N. Bishwal. Bernstein–von Mises theorem and small noise asymptotics of Bayes estimators for parabolic stochastic partial differential equations. *Theory Stoch. Process.*, 23(1):6–17, 2018. ISSN 0321-3900.

Jaya P. N. Bishwal. *Parameter estimation in stochastic volatility models*. Springer, Cham, 2022a. ISBN 978-3-031-03860-0; 978-3-031-03861-7.

Jaya PN Bishwal. Mle evolution equation for fractional diffusions and Berry-Esseen inequality of stochastic gradient descent algorithm for american option. *European Journal of Statistics*, 2: 13–13, 2022b.

R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.*, 18(2):125–135, 2008. ISSN 0960-3174.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, 2009.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010. ISSN 0006-3444.

Antoine Coulon, Carson C Chow, Robert H Singer, and Daniel R Larson. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nature Reviews Genetics*, 14(8):572–584, 2013.

Dennis D. Cox and Finbarr O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.*, 18(4):1676–1695, 1990. ISSN 0090-5364.

Botond Cseke, Manfred Opper, and Guido Sanguinetti. Approximate inference in latent gaussian-markov models from continuous time observations. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 971–979. Curran Associates, Inc., 2013.

Ola Elerian, Siddhartha Chib, and Neil Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993, 2001. ISSN 0012-9682.

Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O. Roberts. Particle filters for partially observed diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(4):755–777, 2008. ISSN 1369-7412.

Nir Friedman, Long Cai, and X. Sunney Xie. Stochasticity in gene expression as observed by single-molecule experiments in live cells. *Israel Journal of Chemistry*, 49:333–342, 2010. doi: 10.1560/IJC.49.3-4.333.

R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar. Approaching complexity by stochastic methods: From biological systems to turbulence. *Physics Reports*, 506:87–162, 2011.

Arnab Ganguly, Riten Mitra, and Jinpu Zhou. Nonparametric learning of stochastic dynamical systems for sparse and noisy data. *In preparation*, a.

Arnab Ganguly, Riten Mitra, and Jinpu Zhou. Model reduction and learning of multiscale stochastic systems. *In preparation*, b.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1(3):515–533, 2006. ISSN 1936-0975.

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005. ISSN 0006-341X.

A. Golightly and D. J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Comput. Statist. Data Anal.*, 52(3):1674–1693, 2008. ISSN 0167-9473.

Andrew Golightly and Darren J. Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network models. *J. Comput. Biol.*, 13(3):838–851, 2006. ISSN 1066-5277. doi: 10.1089/cmb.2006.13.838. URL https://doi-org.libezp.lib.lsu.edu/10.1089/cmb.2006.13.838.

Andrew Golightly and Darren J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle markov chain monte carlo. *Interface Focus*, 1(6):807–820, 2011.

Carl Graham and Denis Talay. *Stochastic simulation and Monte Carlo methods*, volume 68 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2013. ISBN 978-3-642-39362-4; 978-3-642-39363-1. Mathematical foundations of stochastic simulation.

Harvey J Greenberg and William P Pierskalla. A review of quasi-convex functions. *Operations research*, 19(7):1553–1570, 1971.

Rainer Hegger and Gerhard Stock. Multidimensional langevin modeling of biomolecular dynamics. *The Journal of Chemical Physics*, 130(3):034106, 2009.

George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971. ISSN 0022-247X.

Heinz Koeppl, Christoph Zechner, Arnab Ganguly, Serge Pelet, and Matthias Peter. Accounting for extrinsic variability in the estimation of stochastic rate constants. *Internat. J. Robust Nonlinear Control*, 22(10):1103–1119, 2012. ISSN 1049-8923.

Debamita Kundu, Riten Mitra, and Jeremy T Gaskins. Bayesian variable selection for multioutcome models through shared shrinkage. *Scandinavian Journal of Statistics*, 48(1):295–320, 2021.

Yury A. Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Series in Statistics. Springer-Verlag London, Ltd., London, 2004. ISBN 1-85233-759-1.

David Lamouroux and Klaus Lehnertz. Kernel-based regression of drift and diffusion coefficients of stochastic processes. *Physics Letters A*, 373(39):3507–3512, 2009. ISSN 0375-9601.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Comput.*, 17(1):177–204, 2005. ISSN 0899-7667.

L Michaelis and MML Menten. The kinetics of invertin action: translated by trc boyde. *FEBS Lett*, 587:2712–2720, 2013.

Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

María-Eglée Pérez, Luis Raúl Pericchi, and Isabel Cristina Ramírez. The scaled Beta2 distribution as a robust prior for scales. *Bayesian Anal.*, 12(3):615–637, 2017. ISSN 1936-0975.

Nicholas G. Polson and James G. Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian statistics 9*, pages 501–538. Oxford Univ. Press, Oxford, 2011. With discussions by Bertrand Clark, C. Severinski, Merlise A. Clyde, Robert L. Wolpert, Jim e. Griffin, Philiip J. Brown, Chris Hans, Luis R. Pericchi, Christian P. Robert and Julyan Arbel.

Nicholas G. Polson and James G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Anal.*, 7(4):887–902, 2012. ISSN 1936-0975.

G. O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88(3):603–621, 2001. ISSN 0006-3444.

L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. ISBN 0-521-77593-0. Itô calculus, Reprint of the second (1994) edition.

Andreas Ruttor, Philipp Batz, and Manfred Opper. Approximate gaussian process inference for the drift function in stochastic differential equations. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Computational learning theory (Amsterdam, 2001)*, volume 2111 of *Lecture Notes in Comput. Sci.*, pages 416–426. Springer, Berlin, 2001.

Bharath Srinivasan. A guide to the michaelis–menten equation: steady state and beyond. *The FEBS Journal*, 2021.

Tobias Sutter, Arnab Ganguly, and Heinz Koeppl. A variational approach to path estimation and parameter inference of hidden diffusion processes. *J. Mach. Learn. Res.*, 17:Paper No. 190, 37, 2016. ISSN 1532-4435.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1(3):211–244, 2001. ISSN 1532-4435.

J.W. Tuckey. Curves as parameters, and touch estimation. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability.*, page 681–694. University of California Press., Berkeley, 1961.

Sara van Erp, Daniel L. Oberski, and Joris Mulder. Shrinkage priors for Bayesian penalized regression. *J. Math. Psych.*, 89:31–50, 2019. ISSN 0022-2496.

Grace Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990. ISBN 0-89871-244-0.

Gavin A. Whitaker, Andrew Golightly, Richard J. Boys, and Chris Sherlock. Bayesian inference for diffusion-driven mixed-effects models. *Bayesian Anal.*, 12(2):435–463, 2017. ISSN 1936-0975. doi: 10.1214/16-BA1009. URL https://doi-org.libezp.lib.lsu.edu/10.1214/16-BA1009.

Cagatay Yildiz, Markus Heinonen, Jukka Intosalmi, Henrik Mannerstrom, and Harri Lahdesmaki. Learning stochastic differential equations with gaussian processes without gradient matching. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2018.

Kôsaku Yosida. *Functional analysis*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. ISBN 3-540-58654-7. Reprint of the sixth (1980) edition.