Biometrika (2012), *pp.* 1–21 doi: 10.1093/biomet/ass034

Printed in Great Britain

Nonparametric estimation of diffusions: a differential equations approach

BY OMIROS PAPASPILIOPOULOS

Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
omiros.papaspiliopoulos@upf.edu

YVO POKERN

Department of Statistics, University College London, Gower Street, London WC1E 6BT, U.K.
yvo@stats.ucl.ac.uk

GARETH O. ROBERTS

Department of Statistics, The University of Warwick, Coventry CV4 7AL, U.K.
gareth.o.roberts@warwick.ac.uk

AND ANDREW M. STUART

Mathematics Institute, The University of Warwick, Coventry CV4 7AL, U.K.
a.m.stuart@warwick.ac.uk

SUMMARY

We consider estimation of scalar functions that determine the dynamics of diffusion processes. It has been recently shown that nonparametric maximum likelihood estimation is ill-posed in this context. We adopt a probabilistic approach to regularize the problem by the adoption of a prior distribution for the unknown functional. A Gaussian prior measure is chosen in the function space by specifying its precision operator as an appropriate differential operator. We establish that a Bayesian–Gaussian conjugate analysis for the drift of one-dimensional nonlinear diffusions is feasible using high-frequency data, by expressing the loglikelihood as a quadratic function of the drift, with sufficient statistics given by the local time process and the end points of the observed path. Computationally efficient posterior inference is carried out using a finite element method. We embed this technology in partially observed situations and adopt a data augmentation approach whereby we iteratively generate missing data paths and draws from the unknown functional. Our methodology is applied to estimate the drift of models used in molecular dynamics and financial econometrics using high- and low-frequency observations. We discuss extensions to other partially observed schemes and connections to other types of nonparametric inference.

Some key words: Finite element method; Gaussian measure; Inverse problem; Local time; Markov chain Monte Carlo; Markov process.

1. INTRODUCTION

Stochastic differential equations provide a rich framework for time series analysis and they are now used as statistical models throughout science. A typical specification is

$$\mathrm{d}V_s = \xi(V_s)\,\mathrm{d}s + \sigma(V_s)\,\mathrm{d}B_s \quad (s \in [0, T]), \tag{1}$$

where B is a standard Brownian motion. The weakly unique solution of (1), known as a diffusion process, is a strong Markov process. Whereas the mathematical theory underpinning stochastic differential equations is rich and developed, their likelihood-based estimation from data began only relatively recently. The estimation methodology has benefited from significant advances in the understanding of the low-frequency dynamics of diffusion processes, such as the analytic approximations in Aït-Sahalia (2002), and novel Monte Carlo data augmentation methods in Roberts & Stramer (2001), Durham & Gallant (2002), Beskos et al. (2006), and Golightly & Wilkinson (2008). These articles deal with parametric inference, where ξ and σ in (1) are specified parametrically, for partially observed diffusions, in the sense that (1) is observed only at discrete time-points and there might be latent components of V.

This article develops statistical and computational methodology for probabilistic nonparametric inference for partially observed diffusions. We first address a simpler problem: the nonparametric inference of α from a fully observed one-dimensional diffusion path X whose dynamics is given by

$$\mathrm{d}X_s = \alpha(X_s)\,\mathrm{d}s + \mathrm{d}B_s \quad (s \in [0, T]). \tag{2}$$

The assumption of continuous-time data practically means that the frequency of observation can be arbitrarily high, see § 6·1 for such an application. Even in this simple set-up, nonparametric maximum likelihood is ill-posed. The details are given in § 2, but the following is a brief description of the problem. The loglikelihood can be expressed as a quadratic function of the drift, with sufficient statistics given by the so-called local time process and the endpoints, X_0 and X_T. However, this quadratic representation is valid only in a weak sense, where α is sufficiently smooth. Statistically, this means that although the maximization problem is ill-posed, a Bayesian analysis that imposes enough smoothness on α using a prior distribution is well defined. Indeed, we show that a conjugate Bayesian analysis based on a Gaussian process prior for α is feasible.

Using differential operators, we construct Gaussian Markov priors, which lead to a mathematically tractable and computationally efficient posterior inference. The priors can be understood as the limit of Gaussian–Markov random fields (Rue & Held, 2005) and are close in spirit to the more recent work in Lindgren et al. (2011). We provide a finite element method for numerical calculation of the posterior moments, and for simulation from the posterior. Whilst any given finite element approximation may be viewed as parametric, we emphasize that the entire family of finite element approximations at different levels of resolution provides a framework for the approximation of the fully nonparametric posterior distribution to any desired degree of accuracy. This is one of the primary motivations for our approach.

The relationship between Gaussian process priors, differential operators and splines forms the basis for the regularized least squares approach to nonparametric estimation and its link to Bayesian statistics, as described by Wahba (1990). Our approach to drift estimation is a natural generalization of this approach but, because of the very different likelihood, the resulting posterior inference is considerably more complex than for nonparametric regression. Additionally, we note that estimation of α in (2) constitutes a qualitatively different version of the classical white noise model, described for example in Zhao (2000) and Wasserman (2006, Ch. 7). In § 7·1, we discuss the alternative nonparametric inference in the white noise model that involves expanding α in a given basis and estimating the coefficients.

We extend our methodology to inference on unknown drift functionals in partially observed models following a data augmentation approach. We concentrate on the estimation of ξ in (1) for discretely observed diffusions, assuming a parametric model for σ. We extend the existing data augmentation and Markov chain Monte Carlo algorithms for parametric diffusion models to this semiparametric framework. We apply our methods to previously analysed datasets in molecular dynamics and interest rates, where we demonstrate the efficiency of the proposed algorithms and the success of the model in uncovering the diffusion dynamics.

The probabilistic approach that we undertake, when coupled with data augmentation, allows nonparametric Bayesian estimation of the drift of latent diffusions that are involved in complex hierarchical models involving other stochastic processes. Our methods also extend to semiparametric modelling, where the drift of (2) is of the form $f(x) + g(x)\alpha(x)$ for known f and g and unknown α.

2. The likelihood function

We consider the estimation of α in (2). We assume that the derivative α' exists, and that α satisfies regularity conditions such that the following claims hold. We require that (2) admits a unique weak solution X on $[0, T]$. Let $\mathbb{P}_\alpha$ be the law of X on the space of real-valued continuous paths on $[0, T]$ and $\mathbb{W}$ the corresponding Wiener measure. We assume that $\mathbb{P}_\alpha$ is absolutely continuous with respect to $\mathbb{W}$ with Radon–Nikodym derivative $\mathrm{d}\mathbb{P}_\alpha/\mathrm{d}W = \exp\{-I(\alpha)\}$; a weak condition is that the diffusion does not explode (Elworthy, 1982, Theorem 11A). Our prior distributions on α will ensure non-explosion. Then, the negative log-density I between the two measures is

$$I(\alpha) = \frac{1}{2}\int_0^{\mathrm{T}} |\alpha(X_s)|^2 \,\mathrm{d}s - \int_0^{\mathrm{T}} \alpha(X_s)\,\mathrm{d}X_s, \tag{3}$$

which gives the negative loglikelihood for α in the context of diffusion processes. When α is specified in terms of a finite-dimensional parameter vector θ, (3) can be minimized to yield the maximum likelihood estimator for θ, see for example Prakasa Rao (1999). In a nonparametric framework, one might be tempted to minimize this functional over α; this turns out to be an ill-posed minimization problem, as we discuss below.

We will express the right-hand side of (3) as a quadratic functional of α. Let $A(x) = \int^x \alpha(u)\,\mathrm{d}u$ be an antiderivative of α. Then, applying Itô's formula to A, we get that

$$\mathrm{d}A(X_s) = \alpha(X_s)\,\mathrm{d}X_s + \tfrac{1}{2}\alpha'(X_s)\,\mathrm{d}s$$

and rewrite (3) as a Riemann integral:

$$I(\alpha) = \frac{1}{2}\int_0^{\mathrm{T}} \{|\alpha(X_s)|^2 + \alpha'(X_s)\}\,\mathrm{d}s - A(X_T) + A(X_0).$$

A key point of the development is the change from time to space integration. This is achieved by the introduction of the so-called local time process, and it yields a generalization of the change of variables formula. It is known that for any Borel measurable and locally integrable function f on $\mathbb{R}$ we have for each t, $\mathbb{P}_\alpha$-almost surely:

$$\int_0^t f(X_s)\,\mathrm{d}s = \int_{-\infty}^{\infty} L_t(u) f(u)\,\mathrm{d}u$$

where $L_t(x)$ is known as the local time process defined as

$$L_t(x) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_0^t 1\{X_s \in (u-\epsilon, u+\epsilon)\}\, \mathrm{d}s,$$

where the limit is both almost surely and in L^2, see for example Chung & Williams (1990, Corollary 7.4) and Kutoyants (2004, § 1.1.3). Note that $L_T(x) = 0$ for all $x < X_*(T)$ and $x > X^*(T)$, where $X_*(T) = \min\{X_s : s \in [0, T]\}$ and $X^*(T) = \max\{X_s : s \in [0, T]\}$. It is known that $L_t(x)$ is continuous in (t, x) but it is not differentiable; in particular, $L_T(\cdot)$ has the same regularity as Brownian motion, see Chung & Williams (1990, Ch. 7). Additionally, we define

$$\chi(u) = \begin{cases} 1, & X_0 < u < X_T, \\ -1, & X_T < u < X_0, \\ 0, & \text{otherwise}, \end{cases}$$

and we obtain

$$I(\alpha) = \frac{1}{2} \int_{-\infty}^{\infty} \{|\alpha(u)|^2 L_T(u) - 2\chi(u)\alpha(u) + \alpha'(u) L_T(u)\}\, \mathrm{d}u$$

where we emphasize that the integrand is zero for $u < X_*(T)$ and $u > X^*(T)$.

What makes the drift estimation problem nonstandard is the regularity of L_T. Consider the simplified functional

$$\tilde{I}(\alpha) = \frac{1}{2} \int_{-\infty}^{\infty} \{|\alpha(u)|^2 w(u) + \alpha'(u) w(u)\}\, \mathrm{d}u$$

and note that if w is differentiable and has compact support, we can use integration by parts to write $\tilde{I}(\alpha)$ as a quadratic function of α, which is uniquely minimized by $\alpha = w'/2w$. Pokern et al. (2009) study the minimization problem when w is a Brownian bridge, which is only Hölder continuous with exponent $1/2$, and thus has the same regularity as Brownian motion and the local time process of a diffusion. They show that in that case $\tilde{I}$ is unbounded from below.

In this paper we introduce a prior distribution supported only on sufficiently regular drift functions that can incorporate further knowledge or constraints about the unknown functional. The family of prior distributions we consider is motivated by the following formal calculation. Formal is understood as systematic but without a rigorous justification, a terminology which is standard in various areas of mathematics. We manipulate $I(\alpha)$ further, pretending that $L_T(\cdot)$ is differentiable. Using integration by parts, and the fact that L_T has compact support, we obtain that $I(\alpha)$ can be rewritten as

$$\frac{1}{2} \int_{-\infty}^{\infty} [|\alpha(u)|^2 L_T(u) - 2\alpha(u)\{\chi(u) + L_T'(u)/2\}]\, \mathrm{d}u \tag{4}$$

which is quadratic in α. This calculation suggests that the family of Gaussian process priors is conjugate to this likelihood function. It can be taken a step further by completing the square to identify the posterior mean m_1 and precision, i.e., inverse covariance, $\mathcal{Q}_1$ in terms of the prior mean m_0 and prior precision $\mathcal{Q}_0$. We obtain the formulae

$$(\mathcal{Q}_0 + L_T) m_1 = \mathcal{Q}_0 m_0 + \chi + \frac{1}{2} L_T', \quad \mathcal{Q}_1 = \mathcal{Q}_0 + L_T. \tag{5}$$

If $\mathcal{Q}_0$ is a differential operator, then this formulation of the Bayesian inverse problem leads to a powerful computational approach within which it is possible to perform nonparametric inference with precise control over the level of error arising from finite representation of nonparametric estimators, using ideas from numerical analysis. Furthermore, the formulae (5) that characterize the posterior measure in terms of its precision operator can be justified using the theory of weak solutions of differential equations together with properties of Gaussian measures; see Theorem 1.

3. Gaussian measures on function spaces via differential operators

3·1. *Approaches in the literature*

When working with infinite-dimensional spaces, e.g., unknown regression functions or spatial fields, it is common to specify a Gaussian distribution by means of its covariance operator or the corresponding covariance function. This is typically done in geostatistics (Diggle & Ribeiro, 2007) or in machine learning (Bishop, 2006, Ch. 6). On the other hand, it is standard to specify a finite-dimensional Gaussian distribution via its precision matrix, e.g., for stochastic processes on a graph. This approach is based on the key result that the elements of the precision matrix of a multivariate Gaussian distribution relate to the conditional correlation of the corresponding pair of variables given the rest. A convenient assumption from a modelling and computational point of view is that of a Markov dependence. The Markov property implies conditional independence, which translates into sparse precision matrices. The connection between the sparsity of the precision matrix and conditional independence is the key idea behind Gaussian graphical models. Computationally efficient inference is possible using sparse linear algebra methods, e.g., the Kalman filter; see for example Rue & Held (2005, Chs 2 and 3). A third approach, which is dominant in Bayesian nonparametric regression, is to express the unknown function in terms of an orthonormal basis and assign a Gaussian distribution on the coefficients in the expansion.

Our approach lies in the intersection of the first two paradigms, but it also has links with the third. We specify Gaussian measures on function spaces, but work directly with the precision operator. This is specified as a differential operator, which yields the continuous state-space analogue of the Markov property. The main motivation is to obtain a tractable and computable solution to (5) in a setting in which it is possible to rigorously establish the validity of the proposed prior-posterior update. The following subsections provide the necessary background and intuition to motivate this choice and draw connections to more familiar results for finite-dimensional Gaussian measures.

3·2. *Gaussian measures*

This section provides the required background on Gaussian measures in Hilbert space; details can be found for example in Da Prato & Zabczyk (1992). A random variable α on a separable Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ is said to be Gaussian if the law of $\langle \phi, \alpha \rangle$ is Gaussian for all $\phi \in \mathcal{H}$. Gaussian random variables are determined by their mean, $m_0 = E(\alpha) \in \mathcal{H}$, and their covariance operator $\mathcal{C} : \mathcal{H} \to \mathcal{H}$, such that

$$\langle \phi, \mathcal{C}\psi \rangle = E(\langle \phi, \alpha - m_0 \rangle \langle \alpha - m_0, \psi \rangle) \quad (\phi, \psi \in \mathcal{H}).$$

The variable is called nondegenerate if $\langle \phi, \mathcal{C}\phi \rangle > 0$ for all $\phi \in \mathcal{H} \backslash \{0\}$. Then $\mathcal{C}$ is strictly positive, self-adjoint, and trace class, and we can define $\mathcal{Q}$ to be its inverse which, because $\mathcal{C}$ is compact, will be densely defined on $\mathcal{H}$. We will refer to $\mathcal{Q}$ as the precision operator.

More structure is afforded when $\mathcal{H} = L^2([q,r],\mathbb{R}^d)$, in which case we can identify α with a random function/stochastic process $\{\alpha(u) : u \in [q,r]\}$. This will be the setting in this article with $d = 1$. Then, specializing the notation to $d = 1$, the covariance operator has a kernel $C : [q,r]^2 \to \mathbb{R}$, such that

$$(\mathcal{C}\phi)(u) = \int_q^r C(u,v)\phi(v)\,\mathrm{d}v,$$

where $C(u,v) = E[\{\alpha(u) - \alpha_0(u)\}\{\alpha(v) - \alpha_0(v)\}]$ is the covariance function defined for $\alpha_0(u) = E\{\alpha(u)\}$.

3·3. *Connection with differential equations*

In this article we consider precision operators $\mathcal{Q}$ that are real-valued linear differential operators on the interval $u \in [q,r]$. In this case $C(u,v)$ is the Green's function of $\mathcal{Q}$, i.e., for each fixed v the solution to

$$\mathcal{Q}C(u,v) = \delta(u-v), \quad u \in (q,r), \tag{6}$$

subject to the boundary conditions at $u = q$ and $u = r$, where δ denotes the Dirac delta function. Throughout this article the differential operator $\mathcal{Q}$ will have highest order term of the form $(-\eta)^k \mathrm{d}^{2k}/\mathrm{d}u^{2k}$ for some real $\eta > 0$ and integer $k > 0$. Hence, the domain of $\mathcal{Q}$ will be taken to be the Sobolev space H^{2k}, which consists of functions possessing $2k$ square integrable weak derivatives (Lieb & Loss, 2001, Ch. 7) intersected with spaces that impose the boundary conditions. Throughout we will work with weak solutions to differential equations, see § 4·1 for further discussion on this notion.

Equation (6) is solved by letting $\mathcal{Q}C(u,v) = 0$ for $u \neq v$, imposing the boundary conditions at $u = q$ and $u = r$, imposing continuity of the first $2k-2$ derivatives of $C(u,v)$ with respect to u at $u = v$, and imposing a jump of $(-\eta)^{-k}$ in the $(2k-1)$st derivative of $C(u,v)$ with respect to u, as u increases through v. In order to connect this perspective on Gaussian measures with the more standard Gaussian process viewpoint, we study some familiar examples.

Example 1. Consider standard Brownian motion on $[0,1]$. This has covariance function $C(u,v) = u \wedge v$, with $\wedge$ denoting the minimum, whereby it follows that $-\mathrm{d}^2/\mathrm{d}u^2$ with boundary conditions $c(0) = 0$ and $c'(1) = 0$ admits C as its Green's function, and hence is the precision operator of the Wiener measure. Similarly, for standard Brownian bridge on $[0,1]$, we get $C(u,v) = u \wedge v - uv$, which is the Green's function for the same differential operator with the different boundary conditions, $c(0) = c(1) = 0$.

Example 2. Consider the stationary Ornstein–Uhlenbeck process

$$\mathrm{d}\alpha_u = -\lambda_0^{1/2}\alpha_u\,\mathrm{d}u + \eta^{-1/2}\,\mathrm{d}B_u, \quad \alpha_0 \sim N(0, 2^{-1}\eta^{-1}\lambda_0^{-1/2}).$$

This process has covariance function $C(u,v) = (2\eta\lambda_0^{1/2})^{-1}\exp(-\lambda_0^{1/2}|u-v|)$, for $\lambda_0 \geqslant 0$, $\eta > 0$, which is the Green's function of $-\eta\,\mathrm{d}^2/\mathrm{d}u^2 + \eta\lambda_0$ with boundary conditions $c'(0) = \lambda_0^{1/2}c(0)$ and $c'(1) = -\lambda_0^{1/2}c(1)$.

The above examples give rise to second-order differential operators and it is the case that whenever the Gaussian process arises from a conditioned stochastic differential equation with invertible diffusion matrix the precision operator is a second-order differential operator. This is demonstrated in Hairer et al. (2005) where various types of conditioning are discussed. On

the other hand, when the diffusion matrix is degenerate, as arises for example when considering integrated Brownian motion, higher order differential operators can arise.

Example 3. Consider an integrated Ornstein–Uhlenbeck process:

$$\mathrm{d}\alpha_u = \beta_u\,\mathrm{d}t, \quad m\,\mathrm{d}\beta_u = -\beta_u\,\mathrm{d}t + \mathrm{d}B_u, \quad \alpha_0 = 0,\ \beta_0 \sim N\{0, (2m)^{-1}\}, \quad (m > 0).$$

This process conditioned on $\alpha_1 = 0$ has a Gaussian law with precision operator $-m\,\mathrm{d}^4/\mathrm{d}u^4 + \mathrm{d}^2/\mathrm{d}u^2$, subject to the boundary conditions $c(0) = c(1) = 0$, $mc''(0) = c'(0)$, and $mc''(0) = -c'(0)$. The proof of this is more involved than the previous examples, and may be found as Lemma 17 in Hairer et al. (2011).

Given any conditioned diffusion process, there is a prescription for calculating the precision operator. This is to adopt the physicists' convention that Brownian motion on $[q, r]$ has Lebesgue density proportional to $\exp\{-(1/2)\int_q^r |B'(u)|^2\,\mathrm{d}u\}$ and express $B'(u)$ in terms of the process $\alpha(u)$. Adding further conditioning and then writing the resulting density as the exponential of a quadratic form enables a formal identification of $\mathcal{Q}$. The result may then be rigorously verified by means of the Green's function approach; see Hairer et al. (2011) for details. However, our view is that, for the inverse problems arising in this paper, the natural way to specify Gaussian priors is directly through the precision operator. The link to conditioned stochastic processes is insightful, but not necessary in order to justify and implement statistical inference. We note, however, that the postulation of a precision operator given by a differential operator is a continuous-time analogue of the conditional Markov property for discrete random fields and is hence a natural choice of prior for nonparametric inference.

3·4. *Prior specification*

The Gaussian prior needs to comprise four key elements: a mean function that encodes any prior knowledge about the shape of the drift function to be inferred; a scale parameter determining the size of the variance about this mean; a specification of the almost sure smoothness of functions drawn from this prior; and a computationally efficient prior-posterior update, through equations (5), with controllable accuracy. These four elements can be achieved by working with a Gaussian prior $\mu_0 = N(m_0, \mathcal{C}_0)$ in which the covariance is specified via a precision operator

$$\mathcal{Q}_0 = \eta\left\{(-1)^k\frac{\mathrm{d}^{2k}}{\mathrm{d}u^{2k}} + \lambda_0\right\} \quad (\eta > 0,\ \lambda_0 \geqslant 0,\ k \in \mathbb{N}); \tag{7}$$

for simplicity we write $\lambda = \eta\lambda_0$ in what follows.

The domain of $\mathcal{Q}_0$ is a subset of the Sobolev space H^{2k}, i.e., the space of functions on (q, r) with $2k$ square integrable weak derivatives specified by different boundary conditions. The squared Sobolev norm, $\|\cdot\|^2_{H^{2k}}$, is simply the sum of the squares of the L^2 norms of the weak derivatives (Evans, 1998, Ch. 5). Periodic boundary conditions, according to which the value of the function and its $2k$ derivatives agree on endpoints q and r, are convenient from a mathematical perspective since they simplify considerably the proofs of § 4. Such conditions are standard in the theoretical analysis of partial differential equations, but they are also frequently assumed in nonparametric statistics, as for example in Zhao (2000). They are also the appropriate choice in certain applications, as for example the molecular dynamics application of § 6·1. We denote the Sobolev space with periodic boundary conditions by H^{2k}_{per}. On the other hand, in different applications, other boundary conditions are more appropriate; see § § 3·5 and 4·3.

The prior μ_0 satisfies the four criteria above: the mean m_0 encodes known properties of the shape of the drift function to be inferred; the parameters $\eta > 0$ and $\lambda_0 \geqslant 0$ set a scale for the prior variance about this mean function; k can be used to control regularity of draws from the prior as shown in Proposition 1; and efficient computations with controllable accuracy can be carried out, as will be demonstrated in § 4·2. In addition to controlling the scale of the prior variance, the two parameters η, λ_0 also control correlation lengths in the prior. Example 2, for $k = 1$, shows that λ_0 controls the speed of mean reversion to the prior mean, whereas the expected squared distance from the prior mean at any point in the domain is given by the stationary variance of the process, $2^{-1}\eta^{-1}\lambda_0^{-1/2}$. Similar properties arise for other values of k: when one keeps $\eta\lambda_0^{1-1/(2k)}$ constant while increasing λ_0, draws from the prior will revert to the mean more quickly, thus decreasing the correlation length, whilst keeping the total expected L^2 norm of the deviation from the mean constant. This follows from the following fact, which is based on the Karhunen–Loève expansion of $\mathcal{Q}_0$, see Proposition 1 and § 7·1. Under the prior, the total expected L^2 norm of the deviation from the prior mean is given by

$$\frac{1}{\eta}\sum_{j=1}^{\infty}\frac{1}{j^{2k}+\lambda_0} \approx \frac{1}{\eta\lambda_0^{1-1/2k}}\int_0^{\infty}\frac{1}{y^{2k}+1}\mathrm{d}y. \qquad \text{Q5}$$

PROPOSITION 1. *Assume that $m_0 \in H^k_{\mathrm{per}}$. Then the prior measure μ_0 is equivalent to the centred Gaussian $N(0, \mathcal{C}_0)$ and draws from μ_0 almost surely lie in the space H^s_{per} for any $s < k - 1/2$.*

Proof. The first statement is simply the Cameron–Martin theorem, which appears as Proposition 2.24 in Da Prato & Zabczyk (1992). The second statement follows directly from the asymptotic growth of the nth eigenvalue of $\mathcal{Q}_0$, which is proportional to n^{2k}, and use of the Karhunen–Loève expansion, as in Stuart (2010, Lemma 6.27). □

3·5. *Nonperiodic boundary conditions*

We have specified periodic boundary conditions for simplicity of exposition and will remain in this setting for the statement of the main theorems that underpin our approach. However, other choices of local boundary conditions specified at the endpoints $u = q$ and $u = r$ do not generally change the theory or practical implementation. The examples in § 3·3 show a variety of different boundary conditions, which typically lead to Dirichlet, Neumann, or various mixed boundary conditions, all of which can be handled by modifications of the periodic setting.

We describe one particular choice that we will deploy in the interest rate application in § 6·2. We will study a prior that corresponds to an integrated Brownian motion, with variance $\eta^{-1/2}$, and subject to the conditioning that the process has a Gaussian distribution at the endpoint with zero mean with variance σ^2. This gives rise to the prior precision operator (7) with $k = 2$, $\lambda_0 = 0$, and boundary conditions

$$c''(q) = 0, \quad \eta c'''(q) = -\sigma^{-2}c(q), \quad c''(r) = 0, \quad \eta c'''(r) = \sigma^{-2}c(r).$$

Bayesian inference using the family of Gaussian measures on $L^2[q, r]$ defined in (7) can be related to a penalized likelihood approach in which the unknown drift is penalized according to its Sobolev norm. This connection is established by using the physicists' convention already mentioned, according to which we can formally write a prior Lebesgue density $-2\log p_0(\alpha) = \int_q^r \alpha\mathcal{Q}_0\alpha\,\mathrm{d}u$ which we combine with (4) to obtain a penalized maximum likelihood estimator of the drift. For the type of differential precision operators we consider the penalty becomes the

square of the L^2 norm of the kth derivative of the drift. Such connections are established, for example, in Wahba (1990).

4. Gaussian posterior inference

4·1. *Derivation of the posterior distribution*

We consider the situation where we observe a diffusion path $\{X(s) : s \in [0, T]\}$ that is contained within a bounded interval $[q, r]$, and we aim to recover the drift α in (2). A Gaussian prior μ_0 on L^2_{per} has been chosen for the unknown drift in terms of a precision operator (7) with domain H^{2k}_{per}. In this setting, equations (5) become

$$\mathcal{Q}_1 m_1 = \left\{ \eta(-1)^k \frac{\mathrm{d}^{2k}}{\mathrm{d}u^{2k}} + \lambda \right\} m_0 + \chi + \frac{1}{2} L'_T, \tag{8}$$

$$= \eta(-1)^k \frac{\mathrm{d}^{2k}}{\mathrm{d}u^{2k}} + \lambda + L_T, \quad \mathcal{Q}_1, H^{2k}_{\text{per}} \tag{9}$$

A natural formulation of (8) is through the weak form as this provides a framework for its analysis and approximation. This is constructed as follows. Let $V = H^k_{\text{per}}$ and define the bilinear form $a : V \times V \to \mathbb{R}$

$$a(x, y) = \int_q^r \left\{ \eta \frac{\mathrm{d}^k x}{\mathrm{d}u^k}(u) \frac{\mathrm{d}^k y}{\mathrm{d}u^k}(u) + \lambda x(u) y(u) + L_T(u) x(u) y(u) \right\} \mathrm{d}u$$

and the linear form $r : V \to \mathbb{R}$ by

$$r(y) = \int_q^r \left\{ \frac{\mathrm{d}^k m_0}{\mathrm{d}u^k}(u) \frac{\mathrm{d}^k y}{\mathrm{d}u^k}(u) + \lambda m_0(u) y(u) - \frac{1}{2} L_T(u) y'(u) + \chi(u; X) y(u) \right\} \mathrm{d}u.$$

Weak solutions of (8) are functions $x \in V$ that satisfy $a(x, y) = r(y)$ for all $y \in V$, see Evans (1998, § 6.1.2). Using this weak form as the basis for analysis of (8), the formal calculations leading to the expression for the posterior mean can be justified, as summarized in the following theorems.

Theorem 1. *Consider a prior Gaussian measure $\mu_0 = N(m_0, \mathcal{C}_0)$ with mean $m_0 \in H^\tau_{\text{per}}$ with $\tau \geqslant k$ and precision operator $\mathcal{Q}_0$ specified as in (7) with domain H^{2k}_{per}. Then the posterior measure for $\alpha \in L^2([q, r])$, given a sample path X, is a Gaussian measure $\mu_1 = N(m_1, \mathcal{C}_1)$. The mean m_1 is the unique weak solution of (8) where the precision operator given by (9) has domain H^{2k}_{per}; the mean itself is an element of H^s_{per} for $s = \min\{\tau, 2k - 1/2 - \epsilon\}$ for any $\epsilon > 0$. Furthermore, μ_1 and μ_0 are equivalent.*

Theorem 2. *Consider the prior μ_0 and posterior μ_1 under the conditions of Theorem 1. Assume that the sample path $\{X(t)\}_{t \in [0,T]}$ is generated by (2) with drift α in H^k_{per}. Then, for any $\epsilon > 0$, $T^{1-1/2k-\epsilon} \|m_1 - \alpha\|^2_{L^2} \to 0$ in probability.*

The proofs of these theorems are technical and require novel theory. Theorems 1 and 2 may be found together with their proofs as Theorems 3.2 and 5.2 respectively, in an as yet unpublished paper by Y. Pokern, A. M. Stuart and J. H. van Zanten.

4·2. *Finite element method for posterior inference*

The weak formulation of § 4·1 naturally lends itself to a Galerkin approximation. We introduce a finite-dimensional subspace $V^h \subset V$ and seek to solve the following problem for $x^h \in V^h$: $a(x^h, y) = r(y)$, for all $y \in V^h$. We will employ finite element methods in which the finite-dimensional space V^h is spanned by functions with local support. This will lead to linear systems where the matrices to be inverted are banded, with the exception of top-right and bottom-left blocks enforcing the periodicity. The sparse structure of these matrices reflects two main characteristics of the underlying inference problem and one of the types of approximation employed: the operator $\mathcal{Q}_0$ is local due to the conditionally Markov structure of the prior; the operator $\mathcal{Q}_1$ is also local since the information from the data, summarized through L_T, also sees only local pointwise information; and the resulting matrices in the approximation are sparse since finite element bases, which have local support, are tuned to the local structure of the operator being inverted. It is primarily for this reason that we favour the use of finite elements in our numerical implementation of the Galerkin approximation, rather than spectral methods, which employ globally defined basis elements for V^h, see § 7·1.

A key aspect of the Galerkin approximation, and finite element methods in particular, is that they allow the development of error estimates controlling the approximation of an infinite-dimensional object. This theory allows us to tune the accuracy with which we represent the fully nonparametric posterior mean, which is described by equations (8) and (9), and the samples from the Gaussian posterior distribution. To illustrate the power of this theory, we explain in some detail the finite element theory covering the case $k = 2$.

The interval $[q, r]$ is decomposed into N intervals (u_j, u_{j+1}) and we denote the mesh size by $h = \max_j (u_{j+1} - u_j)$. We aim to approximate the posterior mean in the subspace $V^h = H^{2,h}_{\text{per}}$ by functions of the form

$$x^h(u) = \sum_{j=0}^{N-1} \sum_{i=0}^{3} 1_{[u_j, u_{j+1})}(u) x^h_{j,i} \Phi_i \left(\frac{u - u_j}{u_{j+1} - u_j} \right). \tag{10}$$

Here the Φ_i are finite element basis functions, which are third-order polynomials defined on $[0, 1]$ such that $\Phi_i(0) = \delta_{i,0}$, $\Phi_i'(0) = \delta_{i,1}$, $\Phi_i(1) = \delta_{i,2}$, $\Phi_i'(1) = \delta_{i,3}$, where $\delta_{\cdot,\cdot}$ is the Kronecker delta, i.e., $\delta_{i,k} = 1$ if $i = k$ and 0 otherwise. These are the Hermite basis functions displayed in Fig. 1. Also, we impose the conditions $x^h_{j,i+2} = x^h_{(j+1)|N,i}$ for $i \in \{0, 1\}$ and $j \in \{0, 1, \ldots, N - 1\}$, where $j \mid N$ denotes the modulus of j under division by N. These ensure that $x^h(u)$ defined by (10) is continuous with a continuous derivative across element boundaries, leading to so-called conforming finite elements. Substitution into the weak form, $a(x^h, y) = r(y)$, leads to a linear system of equations for the finite set of real numbers defining the function x^h: the coefficients in the expansion (10). This system of linear equations has a banded structure and may be inverted in $\mathcal{O}(N)$ operations. Furthermore, Céa's lemma, which we now state, establishes that the finite element error is bounded by a constant times the best possible approximation in the chosen finite-dimensional space (Braess, 1997, § II.4).

LEMMA 1. *There exists a constant C, independent of h, such that*

$$\|x - x^h\|_{H^2_{\text{per}}} \leqslant C \inf_{y^h \in H^{2,h}_{\text{per}}} \|x - y^h\|_{H^2_{\text{per}}}.$$

The power of this lemma is that it shows that, up to constants of proportionality, the error incurred by the finite element method can be found simply by looking at the error incurred

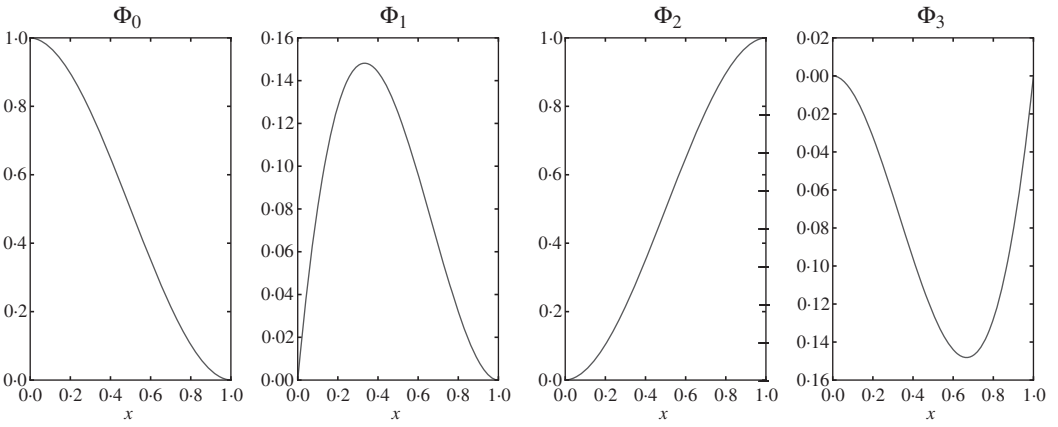

Fig. 1. Finite element basis functions $\Phi_0, \ldots, \Phi_3$.

through interpolation of the true solution in the finite element basis V^h. This is because the infimum on the right-hand side appearing in Lemma 1 is bounded by making this particular choice of y^h. To bound the interpolation error, one can analyse the error element by element, and use a scaling argument and the Bramble–Hilbert lemma (Reddy, 1984, p. 277) to obtain the following bound on the finite element error.

THEOREM 3. *There exists a constant C depending only on h, q, r, and k such that*

$$\|x - x^h\|_{H^2_{\mathrm{per}}} \leqslant Ch\|x\|_{H^3_{\mathrm{per}}}.$$

Theorem 1 states that the solution m_1 of (8) does indeed have regularity $x \in H^3_{\mathrm{per}}$, provided the prior mean is smooth enough, so that the approximation result in Theorem 3 gives quantitative bounds on the approximation of the fully nonparametric posterior. Thus, at increasing computational cost, it is possible to approximate the fully nonparametric solution to (8) to any desired precision. This error is in the H^2 norm and faster rates of convergence are obtained in weaker norms, such as L^2. For more details on the analysis underlying these finite element results, see the Supplementary Material.

The preceding analysis assumes that the local time L_T is known exactly, and that integrals of the local time against the finite element basis function can be computed exactly. In practice L_T is not known exactly, and integrals against it must be approximated. This is done by replacing local time by a piecewise constant function, corresponding to the use of a histogram approximation, see the Supplementary Material. In computational practice we aim to balance the error from approximation of the local time with that arising from the finite element approximation. More sophisticated approximation of the local time is also possible, for example, by using kernel density estimates. For low-frequency observations, where the missing paths can be filled in to an arbitrary frequency, the histogram approximation is natural.

4·3. *Nonperiodic boundary conditions*

It is straightforward to generalize away from the periodic setting without affecting the high-level structure of what has already been presented. However, the assumption of periodicity does have idiosyncracies and it is instructive to consider a nonperiodic case in some detail; we revisit the example of § 3·5, which is also used in the interest rate application of § 6·2. For simplicity we also assume that the prior mean is the zero function: $m_0 = 0$. The posterior mean is then given

by the solution to $\mathcal{Q}_1 m_1 = \chi + L_T'/2$ with the boundary conditions

$$m_1''(q) = 0, \quad \eta m_1'''(q) = -\sigma^{-2} m_1(q), \quad m_1''(r) = 0, \quad \eta m_1'''(r) = \sigma^{-2} m_1(r),$$

where $\mathcal{Q}_1 = \eta(-1)^k \mathrm{d}^{2k}/\mathrm{d}u^{2k} + \lambda + L_T$. The primary changes that are needed to incorporate these nonperiodic boundary conditions into the weak formulation of the problem for the posterior mean are as follows. We now seek $x \in H^k[q, r]$ such that

$$a(x, y) + a_{\mathrm{bdy}}(x, y) = r(y) + r_{\mathrm{bdy}}(y), \quad y \in H^k[q, r],$$

where to enforce our chosen boundary conditions we take

$$a_{\mathrm{bdy}}(x, y) = \frac{1}{\sigma^2} x(q) y(q) + \frac{1}{\sigma^2} x(r) y(r), \quad r_{\mathrm{bdy}}(y) = L_T(r) y(r) - L_T(q) y(q).$$

With these modifications the theory for the existence and regularity of the posterior mean, as well as the practice and theory of the finite element method, proceed as in the periodic case.

5. NONPARAMETRIC INFERENCE FOR DISCRETELY OBSERVED DIFFUSIONS

5·1. *Modelling*

In this section, we study drift and diffusion inference from (1) in the case of low-frequency data. We assume that we are given $n + 1$ discrete-time observations $\{V_j\}$, where $V_j = V_{t_j}$, $t_0 = 0 < t_1 < \cdots < t_n = T$, from the diffusion process V in (1). Unlike in the estimation framework of § 4, here we do not assume that the data are available at arbitrarily high frequency. Solely for simplicity we treat V_0 as fixed by design and model the rest of the observations conditionally on it.

Our approach consists of modelling parametrically the diffusion coefficient, $\sigma(v) = \sigma(v; \theta)$, and semiparametrically the drift $\xi(v)$. Inference for this model can be partially collapsed to inference for (2) by noting that if

$$\eta(v; \theta) = \int_0^v \frac{1}{\sigma(u; \theta)} \mathrm{d}u \tag{11}$$

and η^{-1} is the inverse transformation, then by direct application of Itô's formula the process $X = \eta(V; \theta)$ solves a stochastic differential equation of the form (2) where

$$\alpha(x) = \frac{\xi\{\eta^{-1}(x; \theta)\}}{\sigma\{\eta^{-1}(x; \theta); \theta\}} - \frac{1}{2}\sigma'\{\eta^{-1}(x; \theta); \theta\}. \tag{12}$$

Hence, we will treat (θ, α) as the unknown parameters that are assigned independent prior distributions: α is assigned a Gaussian prior measure as described in § 3·4 and θ a prior density $p_0(\theta)$. The drift of (1) is obtained by inverting (12):

$$\xi(v; \theta, \alpha) = \alpha\{\eta(v; \theta)\}\sigma(v; \theta) + \tfrac{1}{2}\sigma'(v; \theta)\sigma(v; \theta).$$

Section 7·2 discusses interesting alternatives for semiparametric modelling, especially with a view to the type of application considered in § 6·2.

The transformation of (1) to (2) via (11) is central to many Monte Carlo methods for diffusions and it considerably reduces the Monte Carlo variance; see for example Roberts & Stramer (2001), Durham & Gallant (2002), and Beskos et al. (2006).

5·2. *Markov chain Monte Carlo on path space*

The primary aim is inference for (θ, α) conditionally on the observed data $\{V_j\}$ by means of the posterior distribution. However, as in parametric models for diffusions, the posterior distribution is intractable due to the unavailability of continuous-time observations. The Monte Carlo approach to this problem is to resort to data augmentation methods. The Gibbs sampler with Metropolis–Hastings steps that we propose is a direct extension of that in Roberts & Stramer (2001) to the case where α is infinite dimensional. Below, we describe a theoretical algorithm describing this Gibbs loop for the sampling of infinite-dimensional missing paths and parameters. The Appendix contains expressions for the required conditional densities and the Supplementary Material includes details.

We would like to augment the parameter space with the latent process $X = \{X_s : s \in [0, T]\}$, since inference for α given X follows directly from § 4. Nevertheless, θ and X are deterministically linked via the data constraints $V_j = \eta^{-1}(X_{t_j}; \theta)$, so a Gibbs sampler that iteratively samples α, θ, and X from their conditional distributions would be reducible and not converge. The solution to this problem, which is common in inference for stochastic processes based on incomplete data, is given by noncentred reparameterizations (Papaspiliopoulos et al., 2007) that are aimed at removing strong prior dependence between parameters and auxiliary data. In this context, we take $Z = g(X)$, where

$$Z_s = X_s - \frac{t_j - s}{\Delta t_{j-1}} X_{t_{j-1}} - \frac{s - t_{j-1}}{\Delta t_{j-1}} X_{t_j} \quad (t_{j-1} \leqslant s \leqslant t_j;\ j = 1, \ldots, n), \tag{13}$$

which has the effect that $Z_{t_j} = 0$ for all j. Note that X can be reconstructed from Z, θ and $\{V_j\}$, by first obtaining $X_j = X_{t_j} = \eta(V_j; \theta)$ and then the interpolating paths by inverting (13); let $X = h(Z; \theta, \{V_j\})$ denote this transformation.

We can sample from the augmented posterior distribution $\mathbb{P}(\theta, \alpha, Z \mid \{V_j\})$ by the following algorithm, which after initialization iteratively simulates each of the three components according to its conditional distribution:

Step 1. Simulate θ from $\mathbb{P}(\theta \mid Z, \alpha, \{V_j\})$; set $X = h(Z; \theta, \{V_j\})$.

Step 2. Simulate α from $\mathbb{P}(\alpha \mid Z, \theta, \{V_j\}) = \mathbb{P}(\alpha \mid X)$.

Step 3. Simulate X from $\mathbb{P}(X \mid \alpha, \theta, \{V_j\}) = \mathbb{P}(X \mid \alpha, \{X_j\})$; set $Z = g(X)$.

This type of implementation, which involves alternating between the X and Z variables, is typical of Markov chain Monte Carlo algorithms based on noncentred parameterizations (Papaspiliopoulos et al., 2007). The distribution in Step 2 is the infinite-dimensional Gaussian law that we identified, and demonstrated how to approximate, in § 4. The density in Step 1 is given in the Appendix and typically it will not be possible to sample from it directly. Instead, a local Metropolis–Hastings step is used. A considerable reduction in complexity is achieved in Step 3, noting that, due to the Markov property, the diffusion bridges $X^{(j)} = \{X_s : s \in [t_{j-1}, t_j]\}$ are conditionally independent given the endpoints, $\{X_j\}$, and α. Each bridge is sampled using an independence Metropolis–Hastings step with Brownian bridge proposals, see Roberts & Stramer (2001), Papaspiliopoulos & Roberts (2012), the Appendix and Supplementary Material for details.

5·3. *Nonperiodic boundary conditions*

For nonperiodic boundary conditions, we fix a compact interval of interest, $[q, r] \subset \mathbb{R}$, and we perform inference for the drift typically by adopting the prior model with nonperiodic boundary conditions outlined in § 3·5. This leads to the posterior distribution, its weak formulation, and finite element approximation on that interval, as outlined in § 4·3. It may happen during imputation of segments of the diffusion, i.e., in step 3 of the algorithm in § 5·2, that $X_s \notin [q, r]$ for some $s \in [0, T]$. In this case, it becomes necessary to extend the drift function α beyond $[q, r]$. We extend α by a constant such that $\alpha(x) = \alpha(q)$ for all $x \leqslant q$ and $\alpha(x) = \alpha(r)$ for all $x \geqslant r$. Extension by constants means one only has to keep track of how much time the path X spends in $(-\infty, q]$ and $[r, \infty)$, respectively, to keep the procedure consistent.

6. NUMERICAL ILLUSTRATIONS

6·1. *Molecular dynamics*

In computational chemistry, molecular dynamics is often simulated using thermostatted Hamiltonian dynamical systems. The Cartesian positions of $m \in \mathbb{N}$ atoms, $Q_t \in \mathbb{R}^{3m}$ evolve according to Newtonian mechanics in a force field $F(Q_t)$, which is typically fitted empirically to match the observed behaviour of the molecules under study, and possibly subject to damping/driving to thermostat the system. This gives a dynamical system of dimension $6m$. For an accessible overview, see Schlick (2002) and, for the particular force field that we use here, see Brooks et al. (1983). We consider a single butane molecule, which is built around the positions of four carbon atoms, and is subjected to a Langevin thermostat. We study the dihedral angle $X_t = \omega(Q_t)$ subtended by the planes spanned by the first three carbon atoms and the last three carbon atoms; see Fig. 2(c). The available data cover a total time of $T = 4\,\text{ns} = 4 \times 10^{-9}\,\text{s}$ in time steps of $\Delta t = 1\,\text{fs} = 10^{-15}\,\text{s}$; see Fig. 2. The path $\omega(Q_t)$ is more regular than the fitted process at very short time scales, so we subsample at a time scale where the apparent diffusivity does not depend too sensitively on the subsampling chosen and change to a new time unit $1\,\text{u} = 3{\cdot}549\,\text{ps} = 3{\cdot}549 \times 10^{-12}\,\text{s}$ such that the apparent diffusivity is $1\,\text{rad}^2/\text{u}$, see the Supplementary Material for details of choice of time scale.

The unknown drift naturally lives on $[0, 2\pi]$ and given that the data are available at very high frequency we estimate it nonparametrically using the periodic methods of § 4. We adopt a Gaussian prior in the form (7) with $k = 2$ and hyperparameters η and λ fixed at $\eta = 0{\cdot}02\,\text{u}^2\,\text{rad} \approx 0{\cdot}2519\,\text{ps}^2\,\text{rad}$ and $\lambda = 0\,\text{ps/rad}^3$, and take $N = 50$ elements for the approximation. The resulting one standard deviation posterior credible region is displayed in Fig. 2(d). As can be seen, the data are quite informative in this example, and the posterior variance around the mean is quite small. However, the variance is larger away from the centre of the interval $[0, 2\pi]$; this is to be expected since, as the histogram of the data shows, there is more information at the centre of the interval.

6·2. *Interest rates*

The second example deals with nonconstant diffusivity and low-frequency data and the model is estimated using the methods of § 5. In particular, we analyse the well-known Eurodollar dataset, which consists of 5505 daily Eurodollar rates between 1973 and 1995 shown in Fig. 3, and has been analysed among others by Aït-Sahalia (1996), Roberts & Stramer (2001), and Beskos et al. (2006).

An off-the-shelf parametric model for this dataset is the Cox–Ingersoll–Ross model, which takes the form (1) with $\sigma(v; \theta) = \theta v^{1/2}$ and $\xi(v; a, b) = a + bv$. On the other hand, the analysis

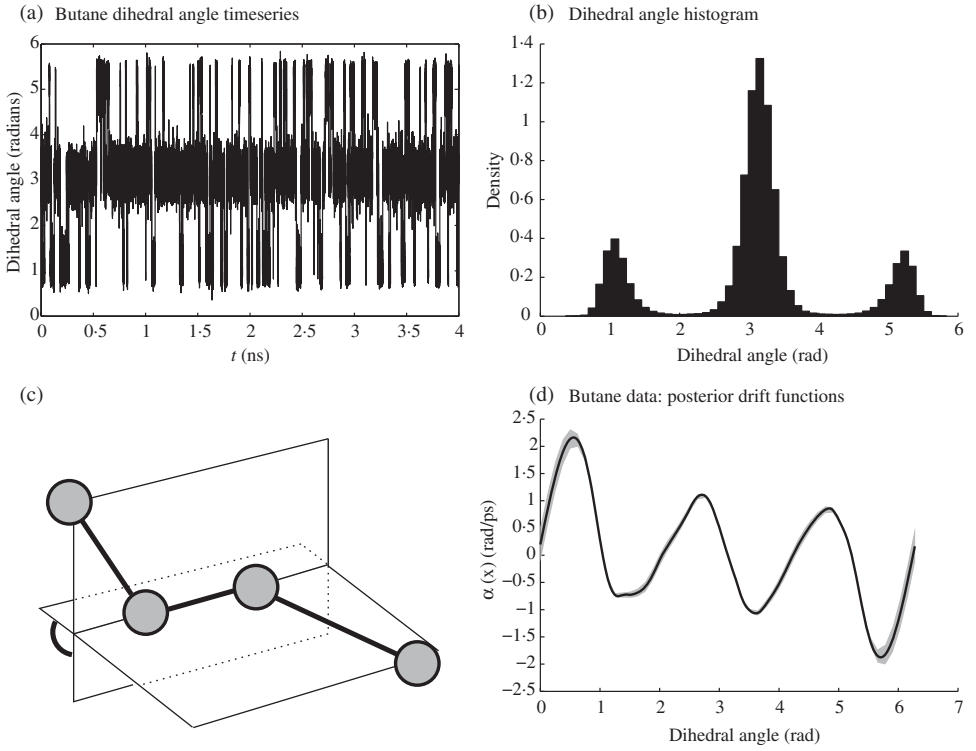

Fig. 2. Molecular dynamics example. (a) Time series of butane dihedral angle. (b) Histogram of the dihedral angle time series. (c) The dihedral angle of the molecule. (d) Posterior mean drift (solid black line) and one standard deviation posterior credible region (shaded in grey).

of Aït-Sahalia (1996) suggests a stronger restoring effect near zero than can be fitted using a linear drift and the need for a more flexible drift function. Hence, we also consider a semi-parametric model with the same diffusion coefficient as the Cox–Ingersoll–Ross model but a nonparametric drift; see also § 7·2.

The parametric model is estimated using the Roberts & Stramer (2001) algorithm, albeit with a slightly different parameterization. After the transformation (11), the drift becomes $\alpha(x;\theta) = [\{(2a/\theta^2) - 1/2\}]x^{-1} + bx/2$. We choose a time unit $1\,\mathrm{u} = 50{\cdot}7\mathrm{d} \approx 4{\cdot}38 \times 10^6\,\mathrm{s}$ such that the mean-field maximum likelihood estimate of θ is 1 $(\%/\mathrm{u})^{1/2}$ with details given in the Supplementary Material. We choose independent Gaussian priors for $\gamma_1 = 2a/\theta^2 - 1/2$ and $\gamma_2 = b/2$ with means 0 and $0\,\mathrm{u}^{-1}$ and variances 500 and $500\,\mathrm{u}^{-2}$, respectively, and an inverse Gamma prior for θ with parameters $\{2, 1(\%/\mathrm{u})^{1/2}\}$. While this prior gives positive probability to drifts that render the process transient, this is of no concern as the data are informative enough to rule out these parameter combinations.

For the nonparametric model, we employ the same prior for the diffusivity and impose a Gaussian prior on the drift $\alpha(\cdot)$ in (12). The Gaussian prior is taken to have the form (7) with prior mean $m_0 \equiv 0\,\mathrm{u}^{-1/2}$ and $k = 2$ with hyperparameters η and λ fixed at $\eta = 0{\cdot}5\,\mathrm{u}^{5/2}$ and $\lambda = 0\,\mathrm{u}^{-1/2}$. We use the nonperiodic boundary conditions discussed in § 4·3 with mean-zero on both sides and variance $\sigma^2 = 100\,\mathrm{u}^{-1}$.

We use $N = 100$ Hermite finite elements where the basis functions are piecewise third-order polynomials setting boundaries $q = (2\min_i V_i)^{1/2}$ and $r = 2(2\max_i V_i)^{1/2}$. We run 2500 iterations of the deterministic scan Gibbs sampler where the first 10 iterations have been discarded as burn-in; trace plots and histograms are given in Fig. 3. The Markov chain mixes well and the

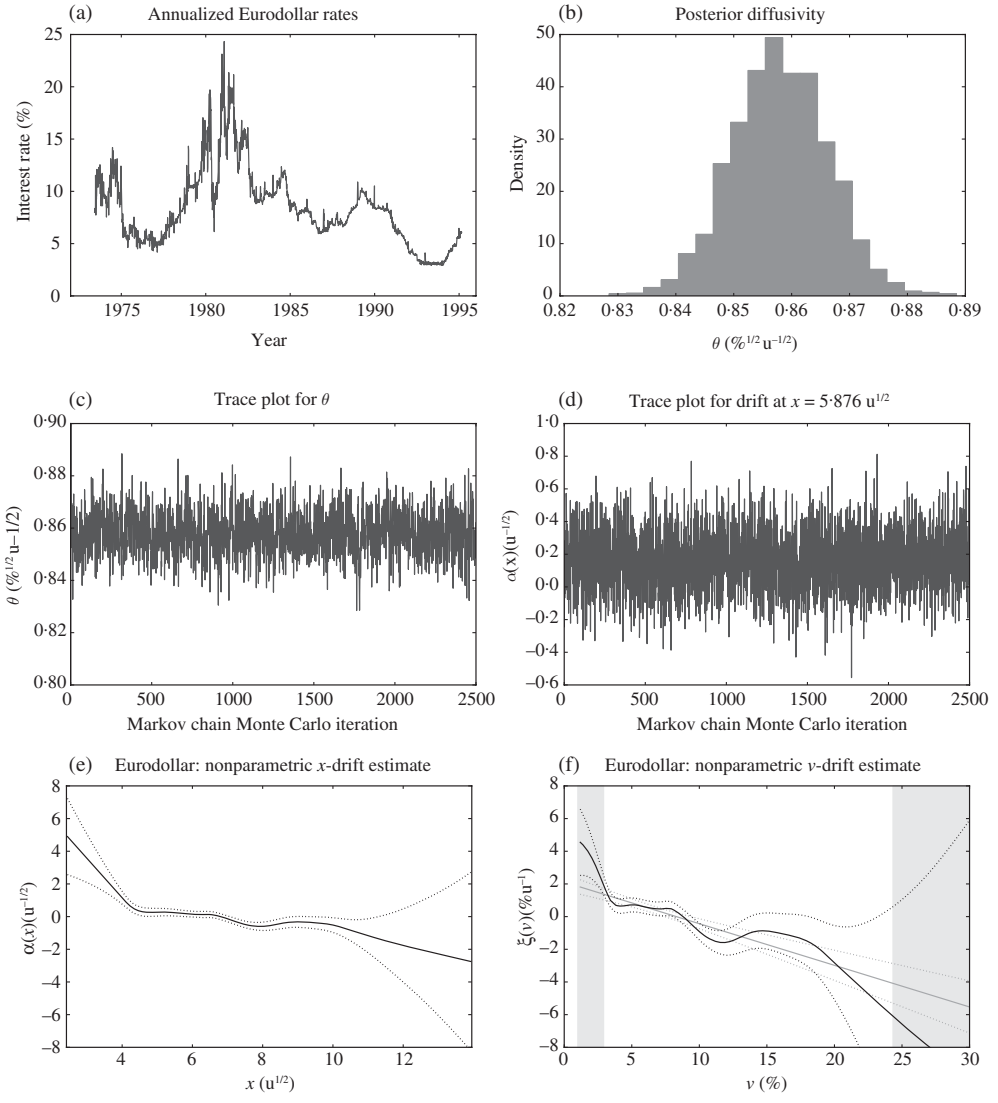

Fig. 3. Interest Rate Example. (a) Time series of annualized Eurodollar rates. (b) Histogram of draws from the posterior distribution of θ. (c) Trace plot for θ. (d) Trace plot for $\alpha(5{\cdot}876\,\mathrm{u}^{1/2})$. (e) Posterior mean of $\alpha(x)$ and one posterior standard deviation credible region by dotted lines. (f) Posterior mean of $\xi(v)$ for the parametric and semiparametric models with one posterior standard deviation credible regions superimposed by dotted lines; shaded areas indicate the region unsupported by direct observation but resolved by the finite element representation.

posterior drift contracts in the region where observations are available. Furthermore, towards low interest rates, the posterior drift $\xi(v)$ can credibly be extrapolated to rule out the simple linear drift model, thus confirming the observation in Aït-Sahalia (1996).

A diffusion might not be an appropriate model for this dataset for various reasons. One apparent feature is the presence of sharp changes in the rate. For example, Beskos et al. (2006) subsampled the dataset every 10th observation to obtain a better fit. A more systematic way to deal with this issue is to introduce the observation error. Doing so is a natural generalization that can be handled within the present framework.

7. Discussion

7·1. *The white noise model and the spectral method*

Recall the so-called white noise model, which is given by (2) but where α is a function only of the independent variable s, and the Brownian noise is scaled by $n^{-1/2}$:

$$\mathrm{d}X_s = \alpha_s\,\mathrm{d}s + n^{-1/2}\,\mathrm{d}B_s, \quad s \in [0,1]. \tag{14}$$

Asymptotic arguments are in terms of the no-noise limit $n \to \infty$ and this is closely related to a large T limit in (2); see Zhao (2000), Wasserman (2006, Ch. 7) and references therein.

For simplicity, we relabel the distribution-valued processes $\mathrm{d}X_s/\mathrm{d}s$ and $\mathrm{d}B_s/\mathrm{d}s$ by y_s and η_s and write (14) as

$$y = \alpha + n^{-1/2}\eta.$$

This equation defines the data likelihood, and its dependence on α. The white noise η is a mean zero process with covariance the identity I (Da Prato & Zabczyk, 1992). A formal argument, similar to that given in § 2, suggests that a Gaussian $N(m_0, \mathcal{C}_0)$ prior on α, with precision $\mathcal{Q}_0 = \mathcal{C}_0^{-1}$, leads to a Gaussian posterior with mean m_1 and precision $\mathcal{Q}_1$ given by

$$(\mathcal{Q}_0 + nI)m_1 = \mathcal{Q}_0 m_0 + ny, \quad \mathcal{Q}_1 = \mathcal{Q}_0 + nI.$$

Making these expressions rigorous requires care, but can be achieved using arguments similar to those in the proof of Theorem 1, when $\mathcal{Q}_0$ is defined as in (7) with domain H^{2k}_{per}. However, for the purposes of our discussion, this level of rigour is not needed.

We note the structural similarities with equations (5) that arise in our inverse problem. However, there are significant differences, the understanding of which can provide insight into the details of the approach we adopt in this paper. For the white noise model, the posterior precision $\mathcal{Q}_1$ and prior precision $\mathcal{Q}_0$ are diagonalizable in the Fourier basis. Working in this basis gives rise to an infinite set of independent scalar Bayesian linear Gaussian estimation problems of the form

$$y_i = \alpha_i + n^{-1/2}\eta_i$$

where the y_i, α_i, η_i are the expansion coefficients of y, α, η, respectively, in the orthonormal Fourier basis $\{\varphi_j\}$. By main properties of the white noise process, it follows that the coefficients η_i are independent standard Gaussian variates. Additionally, the Gaussian prior measure on α with precision operator (7) implies that a priori the coefficients α_i are independent zero-mean Gaussian random variables with standard deviations λ_j decaying like j^{-k}. This property can be checked using the Karhunen–Loève representation of draws from a Gaussian measure (Da Prato & Zabczyk, 1992, Equation 2.30), since the λ_j^2 are determined by the eigenvalue problem $\mathcal{C}_0\varphi_j = \lambda_j^2\varphi_j$; hence, they are the inverse of the eigenvalues of the differential operator (7), subject to periodic boundary conditions.

In contrast, our problem gives rise to a posterior precision $\mathcal{Q}_1$ that is not diagonalizable in the Fourier basis that diagonalizes the prior precision $\mathcal{Q}_0$. Statistically this means that if the prior is specified through independent Gaussian variates, the posterior will involve correlations: it is not possible to decouple into an infinite set of scalar estimation problems. The best we can do is to choose a basis in which the solution is banded and this is exactly what our finite element approach achieves. Furthermore, this banded structure is an explicit manifestation of the conditional independence structure in the prior and posterior distributions.

We could instead use a Fourier basis for the finite dimensional approximation of the posterior distribution identified in § 4; in the context of a numerical solution of differential equations, this

would correspond to the spectral method (Boyd, 2001). However, such a method would lead to full matrices and the conditional independence structure would not be explicit. Furthermore, the usual motivation for using Fourier-based methods is their exponential rate of convergence for smooth functions. However, as we now demonstrate, neither the white noise problem nor our diffusion estimation problem has smooth solutions. In both the posterior draws will lie in the space H^s_{per} for any $s < k - 1/2$ and, provided the prior mean is in H^τ for some $\tau \geqslant k$, the posterior mean will lie in H^s_{per} with $s = \min\{\tau, 2k - 1/2 - \epsilon\}$. This gap between the regularity of the posterior mean and that of draws from the prior is to be expected in conjugate Gaussian–Bayesian analyses because the Cameron–Martin space has measure zero. If the posterior is absolutely continuous with respect to the prior, which is often how Bayes' theorem is formulated in the infinite-dimensional setting, then the posterior mean must lie in the Cameron–Martin space of the posterior, and this will be the same as the Cameron–Martin space of the prior.

Finally, note also that methods based on polynomial chaos show some promise for the solution of inverse problems. In particular, they can exhibit rates of convergence that afford the possibility of improving on the computational complexity of Markov chain Monte Carlo methods; see Marzouk et al. (2007), for example.

7·2. *Semiparametric modelling*

Our approach can be generalized to a semiparametric framework where the drift of (2) is in the form $f(x) + g(x)\alpha(x)$, for parametrically specified f and g. The calculations to identify and approximate the posterior distribution can be carried through as in § 4, because the likelihood remains quadratic in α. The investigation of the strength of mean reversion of interest rates in § 6·2 motivates such an extension. In that context, it is interesting to assume that $\xi(v; a, b, \beta) = a + bv + \beta(v)$, where β is to be estimated nonparametrically from the data. After the transformation (11), the drift becomes

$$\left(\frac{2a}{\theta^2} - \frac{1}{2}\right) x^{-1} + \frac{b}{2}x + \frac{2}{\theta^2 x}\alpha(x),$$

where $\alpha(x) = \beta\{(\theta x/2)^2\}$ is to be estimated nonparametrically.

7·3. *Latent diffusion models*

The full potential of the probabilistic approach to function estimation, as opposed to other types of penalizations, is realized when considering more complex observation schemes than the discrete-time sampling considered in § 5. The key feature of our methodology is conditional conjugacy: given a complete diffusion trajectory and further hyperparameters, computationally efficient Gaussian inference for the drift is feasible. Thus, our approach can be implemented in a variety of other contexts. It directly covers the case where the diffusion is observed with error, either via conditionally independent noisy observations or via a discretely observed continuous-time process whose drift depends on the diffusion; these are versions of the so-called nonlinear filtering problem, see for example Del Moral & Miclo (2000) and Fearnhead et al. (2010). It can deal with the case that (1) drives the stochastic intensity of a Poisson process whose arrivals are observed. This arises for example in single molecule experiments (Kou et al., 2005), where a flexible model for the drift allows the identification of metastable states for the molecule; for details and further references in the context of Förster resonance energy transfer experiments, see Wu & Noé (2011).

7·4. *Multi-dimensional extensions*

The approach we introduce in this article can be extended for estimating the drift of multi-dimensional diffusions. The prior distributions of § 3 can be defined on higher dimensional spaces, whereas the likelihood can be expressed as a space integral using the occupation measure of the diffusion, which admits the local time process as its Lebesgue density in the one-dimensional case. Similarly, the finite elements implementation is numerically efficient for dimensions up to 3. On the other hand, the roughness of the occupation measure increases with dimension, so stronger smoothness conditions are required to estimate a drift of given regularity.

7·5. *Some related literature*

It is worth pointing out recent work on methods and theory for nonparametric estimation in diffusions. For frequentist nonparametric inference with low-frequency data see for example Comte et al. (2007) and Bandi & Phillips (2003) and references therein. There is also a growing body of theoretical literature concerning the rate of posterior contraction in Bayesian nonparametric drift estimation using Gaussian process priors, e.g., van Zanten & van der Vaart (2008) and van der Meulen & van Zanten (2012).

Acknowledgement

Papaspiliopoulos acknowledges financial support by the Spanish government through a Ramon y Cajal fellowship and a research grant. Stuart is grateful to the Engineering and Physical Sciences Research Council, U.K., and the European Research Council for financial support. Roberts would like to thank the Centre for Research in Statistical Methodology. The authors would like to thank Judith Rousseau for helpful discussions.

Supplementary material

Supplementary material available at *Biometrika* online describes the molecular dynamics models and data, the numerical solution of the differential equations in terms of a finite element representation and a weak solution, and the approximation of the local time process from data.

Appendix

Target distribution of the Markov chain Monte Carlo algorithm and approximations

Let $p_t(x, y; \alpha)$ denote the transition density of (2), $q_t(x, y)$ the transition density of the Brownian motion, $\mathbb{L}$ the Lebesgue measure, and $\mathbb{W}^{(t,0,0)}$ the Brownian bridge measure, i.e., the law of Brownian motion on $(0, t)$ conditioned to take the value 0 at the two endpoints. For η defined in (11), $\mathrm{d}\eta/\mathrm{d}v = 1/\sigma(v; \theta)$.

According to the notation of § 5·2, let $Z^{(j)} = \{Z_s : s \in [t_{j-1}, t_j]\}$ for $j = 1, \ldots, n$, $X = h(Z; \theta, \{V_j\})$ be the transformed path, $X_j = \eta(V_j; \theta)$ be the transformed endpoints, and $X^{(j)}$ be the transformed bridges. Let $I(\alpha, X, t, v)$, for $t < v$, be defined as in (3) where we explicitly denote the dependence on X. Finally, we assume that $\theta \in \mathbb{R}^p$, with Lebesgue density $p_0(\theta)$. Recall that the initial data point V_0 is treated as fixed for simplicity.

We have the following decomposition of the joint law of parameters, missing and observed data:

$$\mathbb{P}(\alpha, \theta, Z, \{V_j\}) = \mathbb{P}(\alpha)\mathbb{P}(\theta)\mathbb{P}(\{V_j\} \mid \alpha, \theta)\mathbb{P}(Z \mid \alpha, \theta, \{V_j\}).$$

Due to the Markov property,

$$\mathbb{P}(Z \mid \alpha, \theta, \{V_j\}) = \bigotimes_{j=1}^{N} \mathbb{P}(Z^{(j)} \mid \alpha, \theta, V_{j-1}, V_j),$$

$$\frac{\mathrm{d}\mathbb{P}(\{V_j\}\mid\alpha,\theta)}{\mathrm{d}\mathbb{L}^N}=\prod_{j=1}^{N}p_{\Delta t_{j-1}}(X_{j-1},X_j;\alpha)\frac{1}{\sigma(V_j;\theta)},$$

where in the second equality a change of variables is used. From Papaspiliopoulos & Roberts (2012) we can derive that

$$\frac{\mathrm{d}\mathbb{P}(Z^{(j)}\mid\alpha,\theta,V_{j-1},V_j)}{\mathrm{d}\mathbb{W}^{(\Delta t_{j-1},0,0)}}=\frac{q_{\Delta t_{j-1}}(X_{j-1},X_j)}{p_{\Delta t_{j-1}}(X_{j-1},X_j;\alpha)}\exp\{-2I(\alpha,X,t_{j-1},t_j)\},$$

so the density of the posterior measure $\mathbb{P}(\alpha,\theta,Z\mid\{V_j\})$ with respect to $\mu_0\otimes\mathbb{L}^p\bigotimes_{j=1}^{N}\mathbb{W}^{(\Delta t_{j-1},0,0)}$ is proportional to

$$\prod_{j=1}^{N}\frac{q_{\Delta t_{j-1}}(X_{j-1},X_j)}{\sigma(V_j;\theta)}\exp\left\{-2\sum_{j=1}^{N}I(\alpha,X,t_{j-1},t_j)\right\},$$

where μ_0 is the prior Gaussian measure for α defined in § 3·4. From this expression, it directly follows that $\mathbb{P}(\alpha\mid\theta,Z,\{V_j\})=\mathbb{P}(\alpha\mid X)$ where the latter is described in Theorem 1. Finally, it follows that the $X^{(j)}$ are conditionally independent given $\alpha,\theta,\{V_j\}$, with density proportional to $\exp\{-2I(\alpha,X^{(j)},t_{j-1},t_j)\}$, with respect to $\mathbb{W}^{(\Delta t_{j-1},0,0)}$. Typically, the conditional density of θ is sampled using a local Metropolis–Hastings step and the $X^{(j)}$ are sampled using an independence Metropolis–Hastings sampler with Brownian bridge proposals.

Implementation of the algorithm will typically require a finite-dimensional approximation of Z and discretization of the integrals involved in each $I(\alpha,X,t_{j-1},t_j)$. Hence, we simulate a skeleton of X at equally spaced times in each interval $[t_{j-1},t_j]$, and use a corresponding Riemann approximation to the integrals. In order to sample a new drift function, we approximate the local time implied by the imputed data points by computing a histogram, counting the number of imputed points falling in each interval defined in the finite element method. It is important to make the number of imputed points large enough such that the histogram is a fair representation of the true local time.

REFERENCES

AÏT-SAHALIA, Y. (1996). Testing continuous-time models of the spot interest rate. *Rev. Finan. Studies* **9**, 385–426.

AÏT-SAHALIA, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation. *Econometrica* **70**, 223–62.

BANDI, F. M. & PHILLIPS, P. C. B. (2003). Fully nonparametric estimation of scalar diffusion models. *Econometrica* **71**, 241–83.

BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G. O. & FEARNHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Statist. Soc.* B **68**, 333–82.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

BOYD, J. P. (2001). *Chebyshev and Fourier Spectral Methods: Second Revised Edition*, 2nd ed. Mineola: Dover Publications.

BRAESS, D. (1997). *Finite Elemente, Schnelle Löser und Anwendungen in der Elastizitätstheorie*. Berlin: Springer.

BROOKS, B., BRUCCOLERI, R. E., OLAFSON, B., STATES, D., SWAMINATHAN, S. & KARPLUS, M. (1983). CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.

CHUNG, K. & WILLIAMS, R. (1990). *Introduction to Stochastic Integration*. Boston: Birkhäuser.

COMTE, F., GENON-CATALOT, V. & ROZENHOLC, Y. (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13**, 514–43.

DA PRATO, G. & ZABCZYK, J. (1992). *Stochastic Equations in Infinite Dimensions*. Cambridge: Cambridge University Press.

DEL MORAL, P. & MICLO, L. (2000). *Branching and Interacting Particle Systems. Approximations of Feymann–Kac Formulae with Applications to Non-linear Filtering*, vol. 1729. Berlin: Springer.

DIGGLE, P. J. & RIBEIRO, P. J. (2007). *Model-Based Geostatics*. Berlin: Springer.

DURHAM, G. B. & GALLANT, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *J. Bus. Econ. Statist.* **20**, 297–338. With comments and a reply by the authors.

ELWORTHY, K. D. (1982). *Stochastic Differential Equations on Manifolds*. Cambridge: Cambridge University Press.

EVANS, L. (1998). *Partial Differential Equations*. Providence: American Mathematical Society.

Fearnhead, P., Papaspiliopoulos, O., Roberts, G. O. & Stuart, A. M. (2010). Random weight particle filtering of continuous time processes. *J. R. Statist. Soc.* B **72**, 497–513.
Golightly, A. & Wilkinson, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Comp. Statist. Data Anal.* **52**, 1674–93.
Hairer, M., Stuart, A. M., Voss, J. & Wiberg, P. (2005). Analysis of SPDEs arising in path sampling part I: the Gaussian case. *Comm. Math. Sci.* **3**, 587–603.
Hairer, M., Stuart, A. M. & Voss, J. (2011). Sampling conditioned hypoelliptic diffusions. *Ann. Appl. Prob.* **21**, 669–98.
Kou, S. C., Xie, X. S. & Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *Appl. Statist.* **54**, 469–506.
Kutoyants, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. Springer Series in Statistics. London: Springer.
Lieb, E. H. & Loss, M. (2001). *Analysis*, 2nd ed. Graduate Studies in Mathematics 14. Providence, RI: American Mathematical Society.
Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Statist. Soc.* B **73**, 423–498. With discussions and a reply by the authors.
Marzouk, Y. M., Najm, H. N. & Rahn, L. A. (2007). Stochastic spectral methods for efficient Bayesian solution of inverse problems. *J. Comput. Phys.* **224**, 560–86.
Papaspiliopoulos, O. & Roberts, G. O. (2012). Importance sampling techniques for estimation of diffusion models. In *Statistical Methods for Stochastic Differential Equations*, pp. 311–37. Monographs on Statistics and Applied Probability. Boca Raton: Chapman and Hall.
Papaspiliopoulos, O., Roberts, G. O. & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statist. Sci.* **22**, 59–73.
Pokern, Y., Stuart, A. M. & Vanden-Eijnden, E. (2009). Remarks on drift estimation for diffusion processes. *Multiscale Model. Simul.* **8**, 69–95.
Prakasa Rao, B. L. S. (1999). *Statistical Inference for Diffusion Type Processes*. Kendall's Library of Statistics 8. London: Edward Arnold.
Reddy, J. (1984). *An Introduction to the Finite Element Method*. New York: McGraw-Hill.
Roberts, G. O. & Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika* **88**, 603–21.
Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman & Hall.
Schlick, T. (2002). *Molecular Modeling and Simulation, an Interdisciplinary Guide*. New York: Springer.
Stuart, A. M. (2010). Inverse problems: a Bayesian perspective. *Acta Numer.* **19**, 451–559.
van der Meulen, F. & van Zanten, H. (2012). Consistent nonparametric Bayesian inference for discretely observed diffusions. *Bernoulli*, to appear.
van Zanten, H. & van der Vaart, A. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36**, 1435–63.
Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.
Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.
Wu, H. & Noé, F. (2011). A Bayesian framework for modeling multidimensional diffusion processes with nonlinear drift based on nonlinear and incomplete observations. *Phys. Rev. E* **83**, 036705.
Zhao, L. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28**, 532–52.

[*Received June* 2011. *Revised March* 2012]