Nonparametric estimation of diffusions: a differential equations approach

By Omiros Papaspiliopoulos
Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain

Yvo Pokern
Department of Statistics, UCL, UK

Gareth O. Roberts
Department of Statistics, Warwick University, UK

Andrew M. Stuart
Mathematics Institute, Warwick University, UK

Summary
We consider estimation of scalar functions which determine the dynamics of diffusion processes. It has been recently shown that nonparametric maximum likelihood 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 specified in the function space by means of its precision operator, which is defined as an appropriate differential operator. We establish that a Bayesian Gaussian conjugate analysis for the drift of one-dimensional non-linear diffusions is feasible given high-frequency data. This is achieved by expressing the log-likelihood as a quadratic function of the drift, with sufficient statistics given by the so-called 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 non-parametric inference.

Some key words: Markov processes, local time, finite element method, Gaussian measures, Markov chain Monte Carlo, inverse problems

1. Introduction
Stochastic differential equations (SDEs) provide a rich framework for time series analysis and they are now used as a statistical model throughout science. A typical (albeit not fully general) specification of an SDE is

\[ dV_s = \xi(V_s) \, ds + \sigma(V_s) \, dB_s, \quad s \in [0, T], \]  

(1)
where \( B \) is a standard Brownian motion. The weakly unique solution of (1) (when it exists, see for example Theorem 5.2.1 of Øksendal (2000)) is known as a diffusion process, and it is a strong Markov process. Whereas the mathematical theory underpinning SDEs is rich and developed, fitting SDEs to observed data using likelihood-based methods has taken off only relatively recently. This has been facilitated by significant advances in the understanding of the low frequency dynamics of SDEs. Such advances include analytic approximations, see for example Aït-Sahalia (2002), and novel Monte Carlo data augmentation methodology, as for example in: Roberts & Stramer (2001); Durham & Gallant (2002); Beskos et al. (2006); Golightly & Wilkinson (2008). These articles address the problem of parametric inference for partially observed diffusions. The latter means that (1) is observed only at discrete time points and there might be components of \( V \) which are latent. The former refers to \( \xi \) and \( \sigma \) in (1) being specified via functions known up to certain parameters which have to be estimated from the data. Statistical analysis will also typically involve inference for the unobserved paths of \( V \).

This article develops statistical and computational methodology for probabilistic nonparametric inference for partially observed diffusions. We first address a simpler problem, which turns out to be focal in our approach: the nonparametric inference of \( \alpha \) from a fully observed one-dimensional diffusion path \( X \) whose dynamics are given by the model

\[
dX_s = \alpha(X_s)ds + dB_s, \quad s \in [0, T].
\]

The assumption of continuous-time data practically means that the frequency of observation can be arbitrarily high, see Section 6.1 for such an application. It is well known that even in this simple setup, nonparametric maximum likelihood is ill-posed, see for example Pokern et al. (2009b) and Section 2. We express the log-likelihood as a quadratic function of the drift, with sufficient statistics given by the so-called local time process and the end points, \( X_0 \) and \( X_T \). However, this quadratic representation is only valid in a weak sense, where \( \alpha \) is sufficiently smooth. Statistically, this means that although the maximization problem is ill-posed, a Bayesian analysis which imposes enough smoothness on \( \alpha \) using a prior distribution is well defined. Indeed, we show that a conjugate Bayesian analysis based on a Gaussian process prior for \( \alpha \) is feasible. In more general terms, estimation of \( \alpha \) in (2) constitutes a qualitatively different version of the classical white noise model, see for example Zhao (2000) and Chapter 7 of Wasserman (2006); we discuss this link further in Section 7.1.

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 the more familiar Gaussian Markov random fields (GMRF), which are treated in detail in 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 rational framework for the approximation of the fully nonparametric posterior distribution to any desired degree of accuracy. This is one of the primary motivations for the approach we adopt here.

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 in the seminal book by Wahba (1990). Our approach to drift estimation is a natural generalization of this approach but, be-
cause of the very different likelihood, the resulting posterior inference is considerably more complex than for nonparametric regression. Complementary to this connection, we investigate inference for the white noise model using the paradigm introduced in this article, and contrast our approach with the more rehearsed approach of inferring the coefficients in a series expansion.

We extend our methodology to inference of unknown drift functionals in partially observed models following a data augmentation approach. We concentrate on the estimation of $\xi$ in (1) for discretely observed diffusions, assuming a parametric model for $\sigma$. We extend the existing data augmentation and Markov chain Monte Carlo algorithms for parametric diffusion models to this semi-parametric framework. We apply our methods to previously analyzed 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 which we undertake, when coupled with data augmentation, allows nonparametric Bayesian estimation of the drift of latent diffusions which are involved in complex hierarchical models involving other stochastic processes. Additionally, our methods extend to semi-parametric modelling, where the drift of (2) is of the form $f(x) + g(x)\alpha(x)$ for known functions $f$ and $g$ and unknown function $\alpha$; this setting, whilst being a straightforward generalization of what we describe here, may be useful in a variety of applications.

The structure of the paper is as follows. Section 2 expresses the log-likelihood as a quadratic function of $\alpha$ in (2) using the Girsanov theorem, integration by parts and the local time process. Section 3 introduces Gaussian Markov priors using differential operators. Section 4 contains the theory which validates our approach, and describes how to implement conjugate Gaussian inference using a finite element method. Section 5 designs path transformations and a data augmentation algorithm for nonparametric inference for discretely observed diffusions. Section 6 presents various numerical illustrations. Section 7 recasts the white noise model in the nonparametric framework we introduce here, it discusses semi-parametric models, application of our methodology to models with latent diffusions and multidimensional extensions. An Appendix contains results for the proposed Markov chain Monte Carlo scheme.

2. The likelihood function

We consider the estimation of $\alpha$ in (2). We assume that $\alpha'$ exists, and $\alpha$ satisfies certain regularity conditions such that the following claims hold. We require that (2) admits a unique weak solution on $[0, T]$ which we denote by $X$. Let $\mathbb{P}_\alpha$ be the distribution 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 $\mathbb{W}$; a weak condition is that the diffusion does not explode, for this see Theorem 11A in Elworthy (1982); our priors will satisfy non-explosivity with probability 1. Then, the negative log-density between the two measures is given

$$I[\alpha] = \frac{1}{2} \int_0^T |\alpha(X_s)|^2ds - \int_0^T \alpha(X_s)dX_s. \quad (3)$$

Equation (3) gives the negative log-likelihood for $\alpha$ in the context of diffusion processes. When $\alpha$ is specified in terms of a finite-dimensional parameter vector $\theta$, (3) can be minimized to yield the maximum likelihood estimator for $\theta$, see for example Prakasa Rao.
In a nonparametric framework one might be tempted to minimize this functional over \( \alpha \); 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 \( \alpha \). Firstly, let 
\[
A(x) = \int x \alpha(u) \, du
\]
be an antiderivative of \( \alpha \). Then, applying Itô's formula on \( A \) we get that
\[
dA(X_s) = \alpha(X_s) dX_s + \frac{1}{2} \alpha'(X_s) ds
\]
and re-write (3) as a Riemann integral:
\[
I[\alpha] = \frac{1}{2} \int_0^T (|\alpha(X_s)|^2 + \alpha'(X_s)) ds - A(X_T) + A(X_0).
\]

A key point of the development is the re-expression of the Riemann integral with respect to time as a space-integral. This is achieved by the introduction of the so-called local-time process, and it yields a generalization of the change of variables formula. In particular, it is known (see for example Corollary 7.4 of Chung & Williams (1990) and Section 1.1.3 of Kutoyants (2004)) that for any Borel measurable and locally integrable function \( f \) on \( \mathbb{R} \) we have for each \( t, \)
\[
\mathbb{P}_{\alpha} - a.s:
\int_0^t f(X_s) ds = \int_{-\infty}^{\infty} L_t(u) f(u) du
\]
where for every \( (t, x) \)
\[
L_t(x) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_0^t 1[|X_s| \in (u-\epsilon, u+\epsilon)] ds
\]
and the limit is both almost surely and in \( L^2 \). \( L_t(x) \) is known as the local time process. Note \( L_T(x) = 0 \) 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. Additionally, we define
\[
\chi(u) = \begin{cases} 
1 & \text{if } X_0 < u < X_T \\
-1 & \text{if } X_T < u < X_0 \\
0 & \text{otherwise}
\end{cases}
\]
and we obtain
\[
I[\alpha] = \frac{1}{2} \int_{-\infty}^{\infty} (|\alpha(u)|^2 L_T(u) - 2\chi(u)\alpha(u) + \alpha'(u)L_T(u)) du
\]
where we emphasize that the integrand is 0 outside \((X_*(T), X^*(T))\).

What makes the drift estimation problem non-standard is the regularity of \( L_T \). Consider the simplified functional
\[
\tilde{I}[\alpha] = \frac{1}{2} \int_{-\infty}^{\infty} (|\alpha(u)|^2 w(u) + \alpha'(u)w(u)) du
\]
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 \( \alpha \), which is uniquely minimized by \( \alpha = (\log w)' \). On the other hand, Pokern et al. (2009b) show that when \( w \) is a Brownian bridge, which is only Hölder continuous with exponent 1/2, \( \tilde{I} \) is unbounded from below.
The relevance of this finding to our problem, is that the local time process has the same fine scale properties as a Brownian bridge and it is these fine scale properties which lead to unboundedness from below.

In this paper we adopt a Bayesian approach whereby a prior distribution is introduced which is supported on sufficiently regular drift functions only. Additionally, the prior distribution can incorporate further knowledge (or constraints) about the unknown functional. The family of prior distributions we consider is motivated by the following formal\(^1\) calculation. 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 re-written as

\[
\frac{1}{2} \int_{-\infty}^{\infty} \left( |\alpha(u)|^2 L_T(u) - 2\alpha(u)(\chi(u) + L_T'(u)/2) \right) du
\]

which is a quadratic in \(\alpha\). 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 (inverse covariance) \(Q_1\) in terms of the prior mean \(m_0\) and prior precision \(Q_0\). We obtain the formulae

\[
(Q_0 + L_T)m_1 = Q_0m_0 + \chi + \frac{1}{2} L_T'
\]

\[
Q_1 = Q_0 + L_T.
\]

If \(Q_0\) 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 non-parametric inference with precise control over the level of error arising from finite representation of non-parametric estimators, using ideas from numerical analysis. Furthermore, the formulae (7),(8) which 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 in this paper.

3. Gaussian measures on function spaces via differential operators

We will elicit prior (mixed-)Gaussian measures for the drift function in (2) by specifying the prior mean function and the prior precision (inverse covariance) operator, which in this paper will always be a differential operator, including boundary conditions through its domain of definition. When working with infinite-dimensional spaces, e.g unknown regression functions or spatial fields, a common approach to specifying a Gaussian distribution is via its covariance operator or the corresponding covariance function. This is for example the norm in geostatistics (see for example Stein, 1999; Diggle & Ribeiro, 2007) or in machine learning (see for example Chapter 6 of Bishop, 2006). On the other hand, the specification of a Gaussian distribution via its precision matrix is standard when dealing with Gaussian processes on a lattice, e.g discrete-time stochastic processes or processes on a graph. This approach is based on the key though elementary result that the elements of the precision matrix of a multivariate Gaussian distribution relate to the conditional (also called partial) 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

\(^1\)“formal” is understood as systematic, but without a rigorous justification, a terminology which is standard in various areas of mathematics
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 Chapters 2 and 3 of Rue & Held (2005). 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. This relates to the white noise model which was mentioned in the Introduction and which is revisited in Section 7.1.

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 precisely to obtain a tractable and computable solution to (7) and (8), built on a setting in which it is possible to rigorously establish the validity of the proposed prior-posterior update encapsulated in these equations. 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. The presentation is structured as follows: subsection 3.1 recalls the basic theory of Gaussian measures on Hilbert space; subsection 3.2 connects this theory to precision operators of differential form, via the link between Green’s and covariance functions; subsection 3.3 describes the prior precision operators that are used throughout this paper, in the case of periodic boundary conditions; subsection 3.4 discusses generalization to the non-periodic setting.

### 3.1. 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 $\alpha$ on a separable Hilbert space $\mathcal{H}$ 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, $m_0 = \mathbb{E}[\alpha] \in \mathcal{H}$, and their covariance operator $C : \mathcal{H} \rightarrow \mathcal{H}$, such that

$$
\langle \phi, C \psi \rangle = \mathbb{E}[\langle \phi, \alpha - m_0 \rangle \langle \alpha - m_0, \psi \rangle] \quad \forall \phi, \psi \in \mathcal{H}.
$$

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

More structure is afforded when $\mathcal{H} = L^2([q, r], \mathbb{R}^d)$, in which case we can identify $\alpha$ 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 $C : [q, r]^2 \rightarrow \mathbb{R}$, such that

$$
(C\phi)(u) = \int_q^r C(u, v)\phi(v) dv,
$$

where $C$ is the covariance function

$$
C(u, v) = \mathbb{E}[\{\alpha(u) - \alpha_0(u)\}\{\alpha(v) - \alpha_0(v)\}]
$$

with $\alpha_0(u) = \mathbb{E}[\alpha(u)]$.  

3.2. Connection with differential equations

In this article we will be interested in situations where the precision operator $Q$ is a real-valued linear differential operator on the interval $u \in [q, r]$. In this case $C(u, v)$ is the Green’s function of $Q$, i.e., for each fixed $v$ the solution to

$$QC(u, v) = \delta(u - v), \quad u \in (q, r)$$

subject to the boundary conditions at $u = q$ and $u = r$, where $\delta$ denotes the Dirac delta function. Throughout this article the differential operator $Q$ will have highest order term of the form $(-1)^k \eta \frac{d^{2k}}{du^{2k}}$ for some real $\eta > 0$ and integer $k > 0$. Hence, the domain of $Q$ will be taken to be the Sobolev space $H^{2k}$, which consists of functions possessing $2k$ square integrable weak derivatives (see Chapter 7 of Lieb & Loss, 2001) intersected with spaces which impose the boundary conditions. Throughout we will work with weak solutions to differential equations (see Section 4.1 for further discussion on this notion).

The equation (9) is solved by letting $QC(u, v) = 0$ 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 $(-1)^k \eta^{-1}$ 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.

Consider standard Brownian motion on $[0, 1]$. This has covariance function $C(u, v) = u \wedge v$, whereby it follows that $-\frac{d^2}{du^2}$ with boundary conditions $c(0) = 0$ and $c'(1) = 0$ admits $C$ as its Green’s function, and is hence the precision operator of Brownian motion. Similarly, for standard Brownian bridge on $[0, 1]$, we get $C(u, v) = u \wedge v - uv$, which may be seen to be the Green’s function for the same differential operator but with different boundary conditions $c(0) = c(1) = 0$. Now consider the stationary Ornstein-Uhlenbeck (OU) process

$$d\alpha_u = -\alpha_u du + \sqrt{2} dB_u, \quad \alpha_0 \sim N(0, 1).$$

This process has covariance function $C(u, v) = \exp(-|u - v|)$, which can be seen to be the Green’s function of $-\frac{1}{2} \frac{d^2}{du^2} + \frac{1}{2}I$ with boundary conditions $c'(0) = c(0)$ and $c'(1) = -c(1)$.

The above examples give rise to second order differential operators and it is the case that, whenever the Gaussian process arises from a conditioned SDE with invertible diffusion matrix, then the precision operator is a second order differential operator. This is demonstrated in Hairer et al. (2005) where various conditionings are discussed including fixing the value of the process at the endpoints (bridge constraints) and conditioning upon another linear process whose drift is a linear function of the given SDE (the Kalman-Bucy filter). On the other hand, when the diffusion matrix is degenerate, as arises for example when considering integrated Brownian motion, then higher order differential operators can arise – see Hairer et al. (2011). To illustrate this latter point we consider the following diffusion process, corresponding to an integrated OU process:

$$d\alpha_u = -\alpha_u dt, \quad \alpha_0 = 0$$

$$md\beta_u = -\beta_u dt + dB_u, \quad \beta_0 \sim N(0, \frac{1}{2m}).$$

If we condition this on the end point constraint $\alpha_1 = 0$ then the corresponding precision operator is $-m \frac{d^4}{du^4} + \frac{d^2}{du^2}$ 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 \(\exp(-\int_q^r |B'(u)|^2du)\) 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 \(Q\). The result may then be rigorously verified by means of the Green’s function approach. However it is important to appreciate our standpoint which 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 by no means 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.

3·3. Prior specification

Our standpoint is that the Gaussian prior needs to comprise four key elements: (i) a mean function which encodes any prior knowledge about the shape of the drift function to be inferred; (ii) a scale parameter determining the size of the variance about this mean; (iii) a specification of the almost sure smoothness of functions drawn from this prior (and since we will work in a set-up in which the posterior is absolutely continuous with respect to the prior, this almost sure smoothness will be inherited by the posterior); (iv) a computationally efficient prior-posterior update, through the equations (7)-(8), with controllable accuracy. These four elements can be achieved by working with a Gaussian prior \(\mu_0 = N(m_0,C_0)\) in which the covariance is specified via a precision operator \(Q_0\) with the form

\[
Q_0 = \eta \left( (-1)^k \frac{d^{2k}}{du^{2k}} + \lambda_0 \right),
\]

for some real \(\eta > 0, \lambda_0 > 0\) and strictly positive integer \(k\); for simplicity we denote by \(\lambda\) the product \(\eta\lambda_0\) in what follows.

The domain of \(Q_0\) 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. Periodic boundary conditions, according to which the value of the function and its \(2k\) derivatives agree on endpoints \(q\) and \(r\), are very convenient from a mathematical perspective since they simplify considerably the proofs of Section 4. Such conditions are very standard in the theoretical analysis of partial differential equations, but they are also frequently assumed in nonparametric statistics, as for example in Zhao (2000). Additionally, they are the appropriate choice in certain applications, as for example the molecular dynamics application of Section 6·1 where \(X\) is an angle. We denote the Sobolev space with periodic boundary conditions as \(H^{2k}_{per}\). On the other hand, in different applications other boundary conditions are more appropriate and we discuss such alternatives in Sections 3·4 and 4·3.

The prior \(\mu_0\) satisfies the four criteria which we required above: (i) the mean \(m_0\) encodes known properties of the shape of the drift function to be inferred; (ii) the parameter \(\eta > 0\) sets a scale for the prior variance about this mean function; (iii) the following lemma shows that \(k\) can be used to control regularity of draws from the prior; (iv) the ability to compute efficiently and with controllable accuracy will be demonstrated in the remainder of the paper.
Lemma 1. Assume that $m_0 \in H^k_{\text{per}}$. Then the prior $\mu_0$ is equivalent (in the sense of measures) to the centred Gaussian $N(0, C_0)$ and draws from $\mu_0$ are almost surely in the space $H^s_{\text{per}}$ for any $s < k - \frac{1}{2}$.

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 $n^{th}$ eigenvalue of $Q_0$, which is proportional to $n^{2k}$, and use of the Karhunen-Loeve expansion, as in Lemma 6.27 of Stuart (2010). □

3.4. 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 which underpin our approach. However, other choices of local boundary conditions specified at the end points $u = q$ and $u = r$ do not change the theory or practical implementation in any significant way. The examples in subsection 3.2 show a variety of different boundary conditions, which typically lead to Dirichlet, Neumann or various mixed boundary conditions (such as Robin), all of which can be handled by modifications of the periodic setting that is our focus.

We describe one particular choice that we will deploy in the interest rate application in Section 6.2. We will study a prior which corresponds to an integrated Brownian motion, with variance $\eta^{-\frac{1}{2}}$, and subject to the conditioning that the process is distributed as a mean zero Gaussian with variance $\sigma^2$ at both end points. This gives rise to the prior precision operator (10) with $k = 2, \lambda_0 = 0$ and boundary conditions

\begin{align*}
  c''(q) &= 0, \quad \eta c'''(q) = -\sigma^{-2} c(q), \\
  c''(r) &= 0, \quad \eta c'''(r) = \sigma^{-2} c(r).
\end{align*}

Note that Bayesian inference using the family of Gaussian measures on $L^2[q, r]$ defined in (10) can be related to a penalized likelihood approach where the unknown drift is penalized according to its Sobolev norm. This connection is established by using the physicists’ convention that has already been mentioned, according to which we can formally write a prior Lebesgue density $-2 \log p_0(\alpha) = \int_q^r \alpha Q_0 \alpha \, du$ which we combine with (6) 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 $k$-th derivative of the drift. These type of connections are established, for example, in Wahba (1990).

4. Gaussian posterior inference

We consider the situation where we observe a diffusion path $(X(s), s \in [0, T])$ which is contained within a bounded interval $[q, r]$, and we aim to recover the drift $\alpha$ in (2). In subsection 4.1 we summarize the theoretical results concerning the prior-posterior update, under the prior $\mu_0$ defined in subsection 3.3 and under periodic boundary conditions for convenience. Subsection 4.2 concerns the application of the finite element method to the prior-posterior update equations. In subsection 4.3 we make some remarks concerning the situation where the sample path does not remain in a compact set, as required, for example in large time asymptotic studies in the non-periodic setting.
4.1. Derivation of the posterior distribution

The setting in this Section is as follows. We assume that we have observed a diffusion path \((X(s), s \in [0, T])\) from \((2)\) which is contained within \([q, r]\). A Gaussian prior \(\mu_0\) on \(L_{\text{per}}^2\) has been chosen for the unknown drift, as described in the previous section; in particular the precision operator is given by \((10)\) with domain \(H_{\text{per}}^{2k}\). Because the likelihood \((6)\) is quadratic we expect that the posterior \(\mu_1\) is a Gaussian \(N(m_1, C_1)\) with mean and covariance determined by the equations \((7), (8)\). For our particular prior these equations are simply

\[
\bigg(\eta(-1)^k \frac{d^{2k}}{d u^{2k}} + \lambda + L_T\bigg) m_1 = \bigg(\eta(-1)^k \frac{d^{2k}}{d u^{2k}} + \lambda\bigg) m_0 + \chi + \frac{1}{2} L_T',
\]

\[
Q_1 = \left(\eta(-1)^k \frac{d^{2k}}{d u^{2k}} + \lambda\right) + L_T, 
\]

with the domain of \(Q_1\) being \(H_{\text{per}}^{2k}\).

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

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

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

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

Weak solutions \((13)\) are functions \(x \in V\) which satisfy

\[
a(x, y) = r(y) \quad \forall y \in V.
\]

Using this weak form as the basis for analysis of the equation for the posterior mean, the formal calculations leading to the expression for the posterior mean can indeed be justified, as summarized in the following theorems. Theorem 1 may be found as Theorems 4.2 and Theorem 5.2 in Pokern et al. (2011); note also that Theorems 2 and 4 in Pokern et al. (2009a) contain related results, but that work does not identify the Gaussian with mean and covariance given by \((13), (14)\) as the desired posterior measure.

**Theorem 1.** Consider a prior Gaussian measure \(\mu_0 = N(m_0, C_0)\) with mean \(m_0 \in H_{\text{per}}^k\) with \(\tau \geq k\) and precision operator \(Q_0\) specified as in \((10)\) with domain \(H_{\text{per}}^{2k}\). Then the posterior measure for \(\alpha \in L^2([q, r])\) is a Gaussian \(\mu_1 = N(m_1, C_1)\) with mean \(m_1\) being the unique weak solution of \((13)\) and the precision operator \((14)\) with domain \(H_{\text{per}}^{2k}\) and \(m_1 \in H_{\text{per}}^s\) for \(s = \min\{\tau, 2k - \frac{1}{2} - \epsilon\}\) for any \(\epsilon > 0\). Furthermore \(\mu_1\) is equivalent to \(\mu_0\), in the sense of measures.

Furthermore, we have a form of posterior consistency, in the limit of large data. The following is proved as Theorem 6.2 in Pokern et al. (2011).

**Theorem 2.** Consider the prior \(\mu_0\) and posterior \(\mu_1\) under the conditions of Theorem 1. Assume that the sample path \(\{X(t)\}_{t \in [0, T]}\) is generated by \((2)\) with drift \(\alpha\) in \(H_{\text{per}}^k\). Then, for any \(\epsilon > 0\) \(T^{1-1/2k-\epsilon} \|m_1 - \alpha\|^2 \to 0\) in probability.
4.2. Finite element method for posterior inference

A natural starting point for approximation of the posterior mean is the weak formulation (15). In particular it is natural to seek a Galerkin approximation. We introduce a finite dimensional subspace $V^h \subset V$ and seek to solve the following problem for $x^h \in V^h$:

$$a(x^h, y) = r(y) \quad \forall y \in V^h.$$  \hfill (16)

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 type of approximation employed: (i) the operator $Q_0$ is local due to the conditionally Markov structure of the prior manifest as a precision which is a differential operator and hence sees only local pointwise information; (ii) the operator $Q_1$ is also local since the information from the data, summarized through $L_T$, also sees only local pointwise information; and (iii) 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 method, rather than spectral methods, which employ globally defined basis elements for $V^h$ (see Section 7.1 for further discussion).

A key aspect of the Galerkin method, 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 non-parametric posterior mean (and samples from the Gaussian distribution about it) which is described by equations (13), (14). 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 by $h = \max_j (u_{j+1} - u_j)$ the mesh size. We aim to approximate our posterior mean in the subspace $V^h = H_{\text{per}}^2$ by functions of the form

$$x^h(u) = \sum_{j=0}^{N-1} \sum_{i=0}^{3} 1_{[u_j, u_{j+1})}(u)x^h_{j,i}\Phi_i\left(\frac{u - u_j}{u_{j+1} - u_j}\right)$$  \hfill (17)

Here the $\Phi_i$ 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 $\delta_{i,k}$ is the Kronecker delta, i.e. $\delta_{i,k} = 1$ if $i = k$ and 0 otherwise. These are the Hermite basis functions displayed in Figure 1. Also, we impose the conditions $x^h_{j,i+2} = x^h_{(j+1)|N,4}$ for $i \in \{0, 1\}$ and $j \in \{0, 1, \ldots, N - 1\}$, where $j|N$ denotes the modulus of $j$ under division by $N$. These ensure that $x^h(u)$ defined by (17) is continuous with continuous derivative across element boundaries, leading to so-called conforming finite elements. Substitution into the weak form (16) leads to a linear system of equations for the finite set of real numbers defining the function $x^h$: the coefficients in the expansion (17). This system of linear equations has a banded structure and may be inverted in $O(N)$ 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; for the proof we refer to Braess (1997), section II.4.
LEMMA 2. There exists a constant $C$, independent of $h$, such that

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

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 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 2 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 (see e.g. p. 277 of Reddy (1984)) 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

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

Note that Theorem 1 states that the solution $m_1$ of (13) does indeed have regularity $x \in H^3_{\text{per}}$, 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. The theorem demonstrates that, at increasing computational cost, it is possible to approximate the fully nonparametric solution to (13) to any desired precision. We note also that this error is in the $H^2$ norm and that faster rates of convergence are obtained in weaker norms, such as $L^2$. For details of the analysis underlying these finite element results, see Pokern et al. (2009a).

The preceding analysis is based on an assumption 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. An error bound is given in Pokern et al. (2009a) which quantifies the error incurred in so-doing, by exploiting the known Hölder continuity of the local time. 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. However in this paper we use piecewise linear (histogram) approximation. For low frequency observations, where we use an augmentation scheme with the number of imputed data points between observations a free parameter, this is natural.

4.3. Nonperiodic boundary conditions

As mentioned when discussing the prior, it is straightforward to generalize away from the periodic setting without affecting the high-level structure of what we have presented in detail in the periodic case. However the assumption of periodicity does have idiosyncrasies and it is instructive to consider a nonperiodic case in some detail; we revisit the example of Section 3·4 which is also used in interest rate application of Section 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

\[
\left( \eta(-1)^k \frac{d^{2k}}{du^{2k}} + \lambda + L_T \right) m_1 = \chi + \frac{1}{2} L_T',
\]

with the boundary conditions

\[
m''_1(q) = 0, \quad \eta m'''_1(q) = -\sigma^{-2} m_1(q),
\]

\[
m''_1(r) = 0, \quad \eta m'''_1(r) = \sigma^{-2} m_1(r).
\]

Abstractly we have \( Q_1 m_1 = \chi + \frac{1}{2} L_T' \) and \( Q_1 \) is the posterior precision. We now detail the primary changes that need to be made to incorporate these nonperiodic boundary conditions into the weak formulation of the problem for the posterior mean, and hence the finite element method. The bilinear and linear forms \( a \) and \( r \) change to reflect the new boundary conditions. We now seek \( x \in H^k[q,r] \) such that

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

where \( r_{bdy} \) is a continuous linear form and \( a + a_{bdy} \) is a coercive, symmetric, continuous, positive quadratic form on \( H^k[q,r] \). To enforce our chosen boundary conditions we take

\[
a_{bdy}(x, y) = \frac{1}{\sigma^2} x(q)y(q) + \frac{1}{\sigma^2} x(r)y(r), \quad r_{bdy}(y) = L_T(r)y(r) - L_T(q)y(q). \tag{18}
\]

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

5. Nonparametric inference for discretely observed diffusions

5.1. Modelling

In this section, we build on the material from the previous section, concerned with nonparametric drift inference for (2), to 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 the estimation framework of Section 4, here we do not make the assumption 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 semi-parametrically the drift \( \xi(v) \). Inference for this model can be partially collapsed to inference for (2) by noting that if

\[
\eta(v; \theta) = \int_0^v \frac{1}{\sigma(u; \theta)} du
\]

and \( \eta^{-1} \) is the inverse transformation, then by direct application of Itô’s formula the process \( X := \eta(V; \theta) \) solves an SDE of the form (2) where

\[
\alpha(x) = \frac{\xi(\eta^{-1}(x; \theta))}{\sigma(\eta^{-1}(x; \theta); \theta)} - \frac{1}{2} \sigma'(\eta^{-1}(x; \theta); \theta) \cdot \frac{1}{2} \sigma'(x; \theta) \sigma(x; \theta).
\]

Hence, we will treat \((\theta, \alpha)\) as the unknown parameters which are assigned independent prior distributions: \( \alpha \) is assigned a Gaussian prior measure as described in Section 3·3 and \( \theta \) a prior density \( p_0(\theta) \). The drift of (1) is obtained by inverting (20):

\[
\xi(v; \theta, \alpha) = \alpha(\eta(v; \theta)) \sigma(v; \theta) + \frac{1}{2} \sigma'(v; \theta) \sigma(v; \theta).
\]

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

The transformation of (1) to (2) via (19) is central to many Monte Carlo methods for diffusions and it leads to a considerable reduction of the Monte Carlo variance; see for example the discussion in Roberts & Stramer (2001); Durham & Gallant (2002); Beskos et al. (2006).

5·2. Markov chain Monte Carlo on path space

The primary aim is inference for \((\theta, \alpha)\) 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 which we propose is a direct extension of that in Roberts & Stramer (2001) to the case where \( \alpha \) is infinite-dimensional. Below, we describe a theoretical algorithm describing this Gibbs loop for the sampling of infinite-dimensional missing paths and parameters. Details involved in the implementation of the Gibbs loop can be found in the original article Roberts & Stramer (2001) and in our Appendix. This Appendix contains expressions for the conditional densities which are involved in the implementation of the algorithm and it describes its practical implementation in terms of finite-dimensional approximations.

We would like to augment the parameter space with the latent process \( X = (X_s, s \in [0, T]) \), since inference for \( \alpha \) given \( X \) follows directly from Section 4. Nevertheless, \( \theta \) and \( X \) are deterministically linked via the data constraints \( V_j = \eta^{-1}(X_{t_j}; \theta) \), thus a Gibbs sampler which iteratively samples \( \alpha, \theta \) 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 the so-called non-centered reparameterisations (Papaspiliopoulos et al., 2007) which are aimed at removing strong prior dependence between parameters and auxiliary data. In this context, we take \( Z = g(X) \), where

\[
Z_s := X_s - \frac{t_j - s}{\Delta t_{j-1}} X_{t_{j-1}} - \frac{s - t_{j-1}}{\Delta t_{j-1}} X_{t_j}, \quad t_{j-1} \leq s \leq t_j, j = 1, \ldots, N
\]
which has the effect that $Z_{t_j} = 0$ for all $j$. Note that $X$ can be reconstructed from $Z$, $\theta$ and \{V_j\}, by first obtaining $X_{t_j} = \eta(V_j; \theta)$ and then the interpolating paths via inverting the above transformation; let $X = h(Z; \theta; \{V_j\})$ denote this transformation and $X_j := X_{t_j}$.

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

1. Simulate $\theta$ from $p(\theta \mid Z, \alpha, \{V_j\})$; set $X = h(Z; \theta; \{V_j\})$.
2. Simulate $\alpha$ from $P(\alpha \mid Z, \theta, \{V_j\}) = P(\alpha \mid X)$.
3. Simulate $X$ from $P(X \mid \alpha, \theta, \{V_j\}) = P(X \mid \alpha, \{X_j\})$; 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 non-centered parameterisations, see Papaspiliopoulos et al. (2007) for details. The distribution in Step 2 is the infinite dimensional Gaussian which we identified, and demonstrated how to approximate, in Section 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 by 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 $\{X_j\}$ and $\alpha$. Each bridge is sampled using an independence Metropolis-Hastings step using Brownian bridge proposals, see Roberts & Stramer (2001); Papaspiliopoulos & Roberts (2011) and the Appendix 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 subsection 3.4. This leads to the posterior model, its weak formulation and finite element approximation on that interval, as outlined in subsection 4.3. It may happen during imputation of segments of the diffusion, i.e. in step 3 of the algorithm in the preceding subsection, that $X_s \notin [q, r]$ for some $s \in [0, T]$. In this case, it becomes necessary to extend the drift function $\alpha$ beyond $[q, r]$. We extend $\alpha$ by a constant such that $\alpha(x) = \alpha(q)$ for all $x \leq q$ and $\alpha(x) = \alpha(r)$ for all $x \geq r$. 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

The methodology developed in Sections 4 and 5 is illustrated using two examples which we now outline. The first example arises from study of data generated by a computational molecular dynamics model of the Butane molecule. This model comprises a thermostatted Hamiltonian mechanical system for the positions and momenta of the atoms. The atomic positions are used to define a dihedral angle which encapsulates different molecular configurations via a scalar quantity. A time series for this dihedral angle is shown in the top left panel of Figure 2. Our objective is to find a diffusion model to explain this time series. The unknown drift naturally lives on $[0, 2\pi]$ in this example and we estimate the drift nonparametrically using the methods of Section 4, since the data are available at very high frequency, using the periodic formulation of the problem.

The second example deals with non-constant diffusivity and low frequency data and the model is estimated using the methods of Section 5. In particular, we analyze the well-
known Eurodollar dataset, which consists of 5505 daily Eurodollar rates between 1973 and 1995, which are shown in Figure 3, and has been analyzed among others by Aït-Sahalia (1996); Roberts & Stramer (2001); Beskos et al. (2006). An off-the-self model for this data set is the Cox-Ingersoll-Ross (CIR) model, which takes the form (1) with \( \sigma(v; \theta) = \theta \sqrt{v} \) and linear drift \( \xi(v; a, b) = a + bv \). On the other hand, the analysis 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. We estimate semi-parametrically the drift, and find clear evidence for nonlinearity for small values of the rate, here using a non-periodic formulation of the problem.

### 6.1. Molecular dynamics

In computational chemistry, molecular dynamics are often simulated using (possibly 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 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 the book by Schlick (2002) and for the particular force field that we use here, see Brooks et al. (1983). We consider the simple case of 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 the lower left panel in Figure 2 for a sketch. A time series and histogram from a long time series for this angle is given in the same figure. The data available cover a total time of \( T = 4\text{ns} = 4 \cdot 10^{-9}\text{s} \) in time steps of \( \Delta t = 1\text{fs} = 10^{-15}\text{s} \). 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 timescale such as to reach apparent diffusivity one.

We model this time series by the diffusion process (2) on the circle \([0, 2\pi]\) and seek to estimate the drift \( \alpha \) using \( N = 50 \) finite elements with a Gaussian prior in the form (10) with \( k = 2 \) and hyperparameters \( \eta \) and \( \lambda \) fixed at \( \eta = 0.02 \) and \( \lambda = 0 \). The resulting one standard deviation posterior credible region is displayed in the bottom right panel of Figure 2. As can be seen the data is quite informative in this example, and the posterior variance around the mean is quite small. Note, however, that this 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

We analyze the Eurodollar dataset, which is shown in Figure 3. We deploy two models: the parametric CIR diffusion process, which is of the form (1) with \( \sigma(v; \theta) = \theta \sqrt{v} \) and linear drift \( \xi(v; a, b) = a + bv \), and a nonparametric diffusion model which has the same diffusion coefficient as the CIR, but in which the drift is estimated nonparametrically as described in Section 5; see Section 7.2 for discussion of a different approach.

The parametric CIR is estimated using the Roberts & Stramer (2001) algorithm, albeit with a slightly different parameterisation. After the transformation (19) the drift becomes

\[
\alpha(x; \theta) = \left( \frac{2a}{\theta^2} - \frac{1}{2} \right) x^{-1} + \frac{b}{2} x.
\]
Fig. 2. Molecular Dynamics Example: Top Left: Time series of Butane Dihedral Angle, Top Right: Associated Histogram, Bottom Left: Sketch of Dihedral Angle, Bottom Right: Posterior Mean Drift and one Sigma Credible Region

We choose independent Gaussian priors for $\gamma_1 = \frac{2\theta}{\pi} - \frac{1}{2}$ and $\gamma_2 = \frac{1}{2}$ with mean zero and variance 500. Additionally, we use an inverse Gamma prior for the diffusivity parameter $\theta$ with parameters $(2, 1)$. While this prior gives positive probability to drifts which render the process transient, this is of no concern as the data are informative enough to essentially rule these parameter combinations out.

For the nonparametric model, we employ the same prior for the diffusivity and impose a Gaussian prior on the drift $\alpha(\cdot)$ in the equation (20). The Gaussian prior is taken to have the form (10) with prior mean $m_0 = 0$ and $k = 2$ with hyperparameters $\eta$ and $\lambda$ fixed at $\eta = 0.5$ and $\lambda = 0$. However, in this example we do not use periodic boundary conditions. Rather we specify Gaussian boundary conditions as discussed in subsection 4.3 with mean zero on both sides and variance $\sigma^2 = 100$.

We use $N = 100$ Hermite finite elements (where the basis functions are piecewise third order polynomials) setting boundaries $q = \sqrt{2} \min\ V_i$ and $r = 2\sqrt{2} \max\ V_i$. 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 Figure 3. It is evident that the MCMC chain mixes well and that the posterior drift has contracted 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) of a stronger than linear restoring effect for low interest rates.

It should be mentioned that an SDE might not be an appropriate model for this dataset for various reasons. One apparent feature against this model is the presence of sharp changes in the rate. For example, Beskos et al. (2006) subsampled the dataset every 10-th observation to obtain a process which is better approximated by an SDE. A more systematic way to deal with this issue is to introduce observation error. Doing so is a natural generalization which can be handled within the framework we introduce in this article.

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 $\alpha$ is only a function of the independent variable $s$, and the Brownian noise is scaled by $1/\sqrt{n}$:

$$dX_s = \alpha_s ds + \frac{1}{\sqrt{n}} dB_s, \quad s \in [0, T]. \quad (23)$$

Asymptotic arguments are in terms of the no-noise limit $n \to \infty$ and this is closely related to the large $T$ limit that we consider. An excellent reference to the white noise problem is Zhao (2000); see also Chapter 7 of Wasserman (2006) and references therein.

For simplicity we relabel the distribution-valued processes $dX_s/ds$ and $dB_s/ds$ by $y_s$ and $\eta_s$ and write (23) as

$$y = \alpha + \frac{1}{\sqrt{n}} \eta. \quad (24)$$

This equation defines the data likelihood, and its dependence on $\alpha$. Note that the white noise $\eta$ is a mean zero process with covariance the identity $I$ (Da Prato & Zabczyk, 1992). A formal argument, similar to that given in Section 2, suggests that a Gaussian $N(m_0, C_0)$ prior on $\alpha$, with precision $Q_0 = C_0^{-1}$, leads to a Gaussian posterior with mean $m_1$ and precision $Q_1$ given by

$$(Q_0 + nI) m_1 = Q_0 m_0 + ny \quad (25)$$

$$Q_1 = Q_0 + nI. \quad (26)$$

Making these expressions rigorous requires care, but can be achieved by means of analysis similar to that in Pokern et al. (2011), in the case where $Q_0$ is defined as in (10) with domain $H^{2k}_{per}$. However for the purposes of our discussion, this level of rigour is not needed. We note immediately the structural similarities with equations (7), (8) which arise in our inverse problem.

There are, however, 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 $Q_1$ and prior precision $Q_0$ are simultaneously diagonalizable in the same basis: the Fourier basis. Working in this basis gives rise to an infinite set of
Fig. 3. Eurodollar Data Set: Top Left: Time series of Annualized Interest Rates, Top Right: Posterior Diffusivity $\theta$, Middle Row: Trace plots for Diffusivity and Drift, Bottom Left: Posterior Drift $\alpha(x)$ (one sigma posterior credible region given by dotted lines). Bottom Right: Posterior Drift $\xi(v)$ (one sigma posterior credible region given by dotted lines), parametric posterior estimate superimposed, shaded areas indicate the region unsupported by direct observation but resolved by the finite element representation.
independent scalar Bayesian linear Gaussian estimation problems of the form

$$y_i = \alpha_i + \frac{1}{\sqrt{n}} \eta_i$$  \hspace{1cm} (27)$$

where the $y_i$ (resp. $\alpha_i$ and $\eta_i$) are the expansion coefficients of $y$ (resp. $\alpha$ and $\eta$) in the orthonormal Fourier basis $\{\varphi_j\}$. Thus the $\eta_i$ are i.i.d. $N(0,1)$ and the prior on the $\alpha_i$ are independent Gaussian with standard deviations $\lambda_j$ decaying like $j^{-k}$ when the prior precision (10) is used. Indeed, use of the Karhunen-Loeve representation of draws from a Gaussian measure (see Da Prato & Zabczyk (1992), equation (2.30)) the $\lambda^2_j$ are determined by the eigenvalue problem $C_0 \varphi_j = \lambda^2_j \varphi_j$ and are hence the inverse of the eigenvalues of the differential operator (10), subject to periodic boundary conditions.

In contrast, our problem gives rise to a posterior precision $Q_1$ which is not diagonalizable in the (Fourier) basis which diagonalizes the prior precision $Q_0$. Statistically this means that if the prior is specified through independent Gaussians, 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 Section 4; in the context of numerical solution of differential equations this would correspond to the spectral method. However, such a method would lead to full (not banded) 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 have smooth solutions. Indeed there is a clear similarity between the two problems in this regard. For both our problem and the white noise problem the drift of (2) is of the form $f(x) + g(x)\alpha(x)$, where $f$ and $g$ are known functions, possibly up to some unknown vector of parameters. The calculations can be carried through as in Section 4 to identify and approximate the posterior distribution. This is due to the simple fact that the likelihood is still quadratic in $\alpha$. There are several reasons why such an extension is appealing, and the following is one such example. Recall the CIR model of Section 6-2. Then, if we believe that the linear drift might not be appropriate to explain the data, it is interesting to ask that $\xi(v; a, b, \beta) = a + bv + \beta(v)$, where $\beta$ is to be estimated
nonparametrically from the data. After the transformation (19), 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 type of penalizations is realized when considering more complex observation schemes than the discrete-time sampling considered in Section 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 observation: these are versions of the so-called non-linear filtering problem, see for example Del Moral & Miclo (2000); 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, see for example Kou et al. (2005), where a flexible model for the drift is important since it allows the identification of metastable states for the molecule; for details and further references in the context of FRET experiments see Noé & Wu (2010).

7.4. Multidimensional extensions

The approach we introduce in this article can in principle be extended for estimating the drift of multi-dimensional diffusions. The prior distributions of Section 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 1-d case). Similarly, the finite elements implementation is numerically efficient for dimensions up to 3. On the other hand, the roughness of the occupation measure (appropriately defined) increases with dimension, hence stronger smoothness conditions are required to estimate a drift of given regularity.

7.5. Some related literature

Over and above the connections to related approaches we have already made in the article, it is worth pointing out recent work on methods and theory for nonparametric estimation in diffusions. For classical non-parametric 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, employing Gaussian process priors van Zanten & van der Vaart (2008): see Panzar & van Zanten (2009) and van der Meulen & van Zanten (2009).

Acknowledgements

Papaspiliopoulos would like to acknowledge financial support by the Spanish government through a “Ramon y Cajal” fellowship and grant MTM2009-09063. Stuart is grateful to EPSRC(UK) and ERC for financial support. Roberts would like to thank
CRiSM, particularly through the CRiSM Research Centre. The authors would like to thank Judith Rousseau for helpful discussions.

References


Due to the Markov property and observed data:

\[ \text{data point } V \]

\[ X \]

denote the dependence on \( t < v \), be defined as in (3) but with the integrals over \( \cdot \) defined in Section 5 of (2000). Bayesian aspects of some nonparametric problems. Zhao, L. (2006).


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 Wiener process, i.e the density of a Gaussian centered at \( x \) with variance \( t \), \( \mathbb{L} \) be the Lebesgue measure and \( \mathbb{B}^{(t,0,0)} \) the Brownian bridge measure, i.e the law of Brownian motion on \((0, t)\) conditioned to take the value 0 at the two endpoints. Note that for \( \eta \) defined in (19), \( d\eta/dv = 1/\sigma(v; \theta) \).

Let \( Z^{(j)} = (Z_s, s \in [t_{j-1}, t_j]) \) for \( j = 1, \ldots, N \). In the following \( X = h(Z; \theta, \{V_j\}) \) as defined in Section 5.2, \( X_j = \eta(V_j; \theta) \), and \( X^{(j)} \) defined accordingly. Let \( I[\alpha, X, t, v], \) for \( t < v \), be defined as in (3) but with the integrals over \((t, v)\), and 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 distribution of parameters, missing and observed data:

\[ \mathbb{P}(\alpha, \theta, Z, \{V_j\}) = \mathbb{P}(\alpha)\mathbb{P}(\theta)\mathbb{P}(\{V_j\} | \alpha, \theta)\mathbb{P}(Z | \alpha, \theta, \{V_j\}) \]

Due to the Markov property

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

\[ \frac{d\mathbb{P}(\{V_j\} | \alpha, \theta)}{d\mathbb{L}^N} = \prod_{j=1}^N p_{\Delta t_{j-1}}(X_{j-1}, X_j; \alpha) \frac{1}{\sigma(V_j; \theta)} \]
where in the second equality a change of variables is used. From Papaspiliopoulos & Roberts (2011) we can derive that
\[
\frac{dP(Z^{(j)} \mid \alpha, \theta, V_{j-1}, V_j)}{dW^{(\Delta t_{j-1}, 0, 0)}} = \frac{q_{\Delta t_{j-1}}(X_{j-1}, X_j)}{p_{\Delta t_{j-1}}(X_{j-1}, X_j; \alpha)} \exp\{-2I[\alpha, X, t_{j-1}, t_j]\}
\]
hence, the density of the posterior measure \(P(\alpha, \theta, Z \mid \{V_j\})\) with respect to \(\mu_0 \otimes \mathbb{P} \otimes \mathbb{R} \otimes \mathbb{N}\) 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 \(\mu_0\) is the prior Gaussian measure for \(\alpha\) defined in Section 3.3. From this expression it directly follows that \(P(\alpha \mid \theta, Z, \{V_j\}) = P(\alpha \mid X)\) where the latter is described in Theorem 1. Finally, it follows that the \(X^{(j)}\)'s, conditionally on \(\alpha, \theta, \{V_j\}\), are independent with density with respect to \(W^{(\Delta t_{j-1}, 0, 0)}\) which is proportional to
\[
\exp\{-2I[\alpha, X^{(j)}, t_{j-1}, t_j]\} .
\]
The conditional density of \(\theta\) is typically sampled using a local Metropolis-Hastings step, where a new value \(\theta^*\) is proposed according to some kernel \(q(\theta, \theta^*)\). The \(X^{(j)}\)'s are typically sampled using an independence Metropolis-Hastings sampler using 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 we 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 simply computing a histogram, counting the number of imputed points falling in each interval defined in the finite element method. It is vital to make the number of imputed points large enough such that the histogram is a fair representation of the true local time.