Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Reversible jump MCMC for nonparametric drift estimation for diffusion processes

Frank van der Meulen [a], Moritz Schauer [a,*,1], Harry van Zanten [b]

[a] Delft Institute for Applied Mathematics, Delft University of Technology, The Netherlands
[b] Korteweg–de Vries Institute for Mathematics, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 22 June 2012
Received in revised form 3 March 2013
Accepted 4 March 2013
Available online xxxx

Keywords:
Reversible jump Markov chain Monte Carlo
Discretely observed diffusion process
Data augmentation
Nonparametric Bayesian inference
Multiplicative scaling parameter
Series prior

ABSTRACT

In the context of nonparametric Bayesian estimation a Markov chain Monte Carlo algorithm is devised and implemented to sample from the posterior distribution of the drift function of a continuously or discretely observed one-dimensional diffusion. The drift is modeled by a scaled linear combination of basis functions with a Gaussian prior on the coefficients. The scaling parameter is equipped with a partially conjugate prior. The number of basis functions in the drift is equipped with a prior distribution as well. For continuous data, a reversible jump Markov chain algorithm enables the exploration of the posterior over models of varying dimension. Subsequently, it is explained how data-augmentation can be used to extend the algorithm to deal with diffusions observed discretely in time. Some examples illustrate that the method can give satisfactory results. In these examples a comparison is made with another existing method as well.

1. Introduction

Suppose we observe a diffusion process X, given as the solution of the stochastic differential equation (SDE)

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = x_0, \tag{1}$$

with initial state x_0 and unknown drift function b. The aim is to estimate the drift b when a sample path of the diffusion is observed continuously up till a time $T > 0$ or at discrete times $0, \Delta, 2\Delta, \dots, n\Delta$, for some $\Delta > 0$ and $n \in \mathbb{N}$.

Diffusion models are widely employed in a variety of scientific fields, including physics, economics and biology. Developing methodology for fitting SDEs to observed data has therefore become an important problem. In this paper we restrict the exposition to the case that the drift function is 1-periodic and the diffusion function is identically equal to 1. This is motivated by applications in which the data consists of recordings of angles; cf. e.g. Pokern (2007), Hindriks (2011) or Papaspiliopoulos et al. (2012). The methods we propose can however be adapted to work in more general setups, such as ergodic diffusions with non-unit diffusion coefficients. In the continuous observations case, a diffusion with periodic drift could alternatively be viewed as diffusion on the circle. Given only discrete observations on the circle, the information about how many turns around the circle the process has made between the observations is lost however and the total number of windings is unknown. For a lean exposition we concentrate therefore on diffusions with periodic drift on $\mathbb{R}$. In the discrete observations setting the true circle case could be treated by introducing a latent variable that keeps track of the winding number.

[*] Correspondence to: TU Delft, Mekelweg 4, 2628 CD Delft, The Netherlands. Tel.: +31 15 2782546.
E-mail address: m.r.schauer@tudelft.nl (M. Schauer).

[1] Supported in part by the Netherlands Organization for Scientific Research (NWO) via the STAR Mathematics Cluster.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.03.002

In this paper we propose a new approach to making nonparametric Bayesian inference for the model (1). A Bayesian method can be attractive since it does not only yield an estimator for the unknown drift function, but also gives a quantification of the associated uncertainty through the spread of the posterior distribution, visualized for instance by pointwise credible intervals. Until now the development of Bayesian methods for diffusions has largely focused on parametric models. In such models it is assumed that the drift is known up to a finite-dimensional parameter and the problem reduces to making inference about that parameter. See for instance the papers Eraker (2001), Roberts and Stramer (2001), Beskos, Papaspiliopoulos, Roberts et al. (2006), to mention but a few. When no obvious parametric specification of the drift function is available it is sensible to explore nonparametric estimation methods, in order to reduce the risk of model misspecification or to validate certain parametric specifications. The literature on nonparametric Bayesian methods for SDEs is however still very limited at the present time. The only paper which proposes a practical method we are aware of is Papaspiliopoulos et al. (2012). The theoretical, asymptotic behavior of the procedure of Papaspiliopoulos et al. (2012) is studied in the recent paper Pokern et al. (2013). Other papers dealing with asymptotics in this framework include Panzar and van Zanten (2009) and Van der Meulen and van Zanten (2013), but these do not propose practical computational methods.

The approach we develop in this paper extends or modifies that of Papaspiliopoulos et al. (2012) in a number of directions and employs different numerical methods. Papaspiliopoulos et al. (2012) consider a Gaussian prior distribution on the periodic drift function b. This prior is defined as a Gaussian distribution on $L^2[0, 1]$ with densely defined inverse covariance operator (precision operator)

$$\eta((-\Delta)^p + \kappa I), \tag{2}$$

where Δ is the one-dimensional Laplacian (with periodic boundaries conditions), I is the identity operator and $\eta, \kappa > 0$ and $p \in \mathbb{N}$ are fixed hyper parameters. It is asserted in Papaspiliopoulos et al. (2012) and proved in Pokern et al. (2013) that if the diffusion is observed continuously, then for this prior the posterior mean can be characterized as the weak solution of a certain differential equation involving the local time of the diffusion. Moreover, the posterior precision operator can be explicitly expressed as a differential operator as well. Posterior computations can then be done using numerical methods for differential equations.

To explain our alternative approach we note, as in Pokern et al. (2013), that the prior just defined can be described equivalently in terms of series expansions. Define the basis functions $\psi_k \in L^2[0, 1]$ by setting $\psi_1 \equiv 1$, and for $k \in \mathbb{N}$ $\psi_{2k}(x) = \sqrt{2}\sin(2k\pi x)$ and $\psi_{2k+1}(x) = \sqrt{2}\cos(2k\pi x)$. Then the prior is the law of the random function

$$x \mapsto \sum_{l=1}^{\infty} \sqrt{\lambda_l} Z_l \psi_l(x),$$

where the Z_l are independent, standard normal variables and for $l \geq 2$

$$\lambda_l = \left(\eta\left(4\pi^2\left\lceil\frac{l}{2}\right\rceil^2\right)^p + \eta\kappa\right)^{-1}. \tag{3}$$

This characterization shows in particular that the hyper parameter p describes the regularity of the prior through the decay of the Fourier coefficients and $1/\eta$ is a multiplicative scaling parameter. The priors we consider in this paper are also defined via series expansions. However, we make a number of substantial changes.

Firstly, we allow for different types of basis functions. Different basis functions instead of the Fourier-type functions may be computationally attractive. The posterior computations involve the inversion of certain large matrices and choosing basis functions with local support typically makes these matrices sparse. In the general exposition we keep the basis functions completely general but in the simulation results we will consider wavelet-type Faber–Schauder functions in addition to the Fourier basis. A second difference is that we truncate the infinite series at a level that we endow with a prior as well. In this manner we can achieve considerable computational gains if the data driven truncation point is relatively small, so that only low-dimensional models are used and hence only relatively small matrices have to be inverted. A last important change is that we do not set the multiplicative hyper parameter at a fixed value, but instead endow it with a prior and let the data determine the appropriate value.

We will present simulation results in Section 4 which illustrate that our approach indeed has several advantages. Although the truncation of the series at a data driven point involves incorporating reversible jump MCMC steps in our computational algorithm, we will show that it can indeed lead to a considerably faster procedure compared to truncating at some fixed high level. The introduction of a prior on the multiplicative hyper parameter reduces the risk of misspecifying the scale of the drift. We will show in Section 4.2 that using a fixed scaling parameter can seriously deteriorate the quality of the inference, whereas our hierarchical procedure with a prior on that parameter is able to adapt to the true scale of the drift. A last advantage that we will illustrate numerically is that by introducing both a prior on the scale and on the truncation level we can achieve some degree of adaptation to smoothness as well.

Computationally we use a combination of methods that are well established in other statistical settings. Within models in which the truncation point of the series is fixed we use Gibbs sampling based on standard inverse gamma-normal computations. We combine this with reversible jump MCMC to move between different models. For these moves, we can use an auxiliary Markov chain to propose a model, and subsequently draw coefficients from their posterior distribution within that model. Such a scheme has been proposed for example in Godsill (2001) for estimation in autoregressive time-series models. In case of discrete observations we also incorporate a data augmentation step using a Metropolis–Hastings sampler to generate diffusion bridges. Our numerical examples illustrate that using our algorithm it is computationally feasible to carry out nonparametric Bayesian inference for low-frequency diffusion data using a non-Gaussian hierarchical prior which is more flexible than previous methods.

A brief outline of the article is as follows: In Section 2 we give a concise prior specification. In the section thereafter, we present the reversible jump algorithm to draw from the posterior for continuous-time data. Data-augmentation is discussed in Section 3.3. In Section 4 we give some examples to illustrate our method. We end with a section on numerical details.

2. Prior distribution

2.1. General prior specification

To define our prior on the periodic drift function b we write a truncated series expansion for b and put prior weights on the truncation point and on the coefficients in the expansion. We employ general 1-periodic, continuous basis functions $\psi_l, l \in \mathbb{N}$. In the concrete examples ahead we will consider in particular Fourier and Faber–Schauder functions. We fix an increasing sequence of natural numbers $m_j, j \in \mathbb{N}$, to group the basis functions into *levels*. The functions $\psi_1, \ldots, \psi_{m_1}$ constitute level 1, the functions $\psi_{m_1+1}, \ldots, \psi_{m_2}$ correspond to level 2, etc. In this manner we can accommodate both families of basis functions with a single index (e.g. the Fourier basis) and doubly indexed families (e.g. wavelet-type bases) in our framework. Functions that are linear combinations of the first m_j basis functions $\psi_1, \ldots, \psi_{m_j}$ are said to belong to *model j*. Model j encompasses levels 1 up till j.

To define the prior on b we first put a prior on the model index j, given by certain prior weights $p(j), j \in \mathbb{N}$. By construction, a function in model j can be expanded as $\sum_{l=1}^{m_j} \theta_l^j \psi_l$ for a certain vector of coefficients $\theta^j \in \mathbb{R}^{m_j}$. Given j, we endow this vector with a prior by postulating that the coefficients θ_l^j are given by an inverse gamma scaling constant times independent, centered Gaussians with decreasing variances $\xi_l^2, l \in \mathbb{N}$. The choice of the constants ξ_l^2 is discussed in Sections 2.2.1 and 2.2.2.

Concretely, to define the prior we fix model probabilities $p(j), j \in \mathbb{N}$, decreasing variances ξ_l^2, positive constants $a, b > 0$ and set $\Xi^j = \operatorname{diag}(\xi_1^2, \ldots, \xi_{m_j}^2)$. Then the hierarchical prior Π on the drift function b is defined as follows:

$$\boxed{\begin{aligned} j &\sim p(j), \\ s^2 &\sim \mathrm{IG}(a, b), \\ \theta^j \mid j, s^2 &\sim N_{m_j}(0, s^2 \Xi^j), \\ b \mid j, s^2, \theta^j &\sim \sum_{l=1}^{m_j} \theta_l^j \psi_l. \end{aligned}}$$

2.2. Specific basis functions

Our general setup is chosen such that we can incorporate bases indexed by a single number and doubly indexed (wavelet-type) bases. For the purpose of illustration, one example for each case is given below. First, a Fourier basis expansion, which emphasizes the (spectral) properties of the drift in frequency domain and second, a (Faber–) Schauder system which features basis elements with local support.

2.2.1. Fourier basis

In this case we set $m_j = 2j - 1$ and the basis functions are defined as

$$\psi_1 \equiv 1, \qquad \psi_{2k}(x) = \sqrt{2}\sin(2k\pi x), \qquad \psi_{2k+1}(x) = \sqrt{2}\cos(2k\pi x), \quad k \in \mathbb{N}.$$

These functions form an orthonormal basis of $L^2[0, 1]$ and the decay of the Fourier coefficients of a function is related to its regularity. More precisely, if $f = \sum_{l \geq 1} \theta_l \psi_l$ and $\sum_{l \geq 1} \theta_l^2 l^{2\beta} < \infty$ for $\beta > 0$, then f has Sobolev regularly β, i.e. it has square integrable weak derivatives up to the order β. By setting $\xi_l^2 \sim l^{-1-2\beta}$ for $\beta > 0$, we obtain a prior which has a version with α-Hölder continuous sample paths for all $\alpha < \beta$. A possible choice for the model probabilities is to take them geometric, i.e. $p(j) \sim \exp(-Cm_j)$ for some $C > 0$.

Priors of this type are quite common in other statistical settings. See for instance Zhao (2000) and Shen and Wasserman (2001), who considered priors of this type in the context of the white noise model and nonparametric regression. The prior can be viewed as an extension of the one of Papaspiliopoulos et al. (2012) discussed in the introduction. The latter uses the same basis functions and decreasing variances with $\beta = p - 1/2$. It does not put a prior on the model index j however (it basically takes $j = \infty$) and uses a fixed scaling parameter whereas we put a prior on s. In Section 4 we argue that our approach has a number of advantages.

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

2.2.2. Schauder functions

The Schauder basis functions are a location and scale family based on the "hat" function $\Lambda(x) = (2x)1_{\left[0,\frac{1}{2}\right)}(x) + 2(x-1)1_{\left[\frac{1}{2},1\right]}(x)$. With $m_j = 2^{j-1}$, the Schauder system is given by $\psi_1 \equiv 1$ and for $l \geq 2$ $\psi_l(x) = \Lambda_l(x \bmod 1)$, where

$$\Lambda_{2^{j-1}+k}(x) = \Lambda(2^{j-1}x - k + 1), \quad j \geq 1,\ k = 1, \dots, 2^{j-1}.$$

These functions have compact supports. A Schauder expansion thus emphasizes local properties of the sample paths. For $\beta \in (0, 1)$, a function f with Faber–Schauder expansion $f = \sum_{l\geq 1} c_l\psi_l = c_1\psi_1 + \sum_{j\geq 1}\sum_{k=1}^{2^{j-1}} c_{2^{j-1}+k}\psi_{2^{j-1}+k}$ has Hölder regularity of order β if and only if $|c_l| \leq \text{const.} \times l^{-\beta}$ for every l (see for instance Kashin and Saakyan, 1989). It follows that if in our setup we take $\xi_{2^{j-1}+k} = 2^{-\beta j}$ for $j \geq 1$ and $k = 1, \dots, 2^{j-1}$, then we obtain a prior with regularity β. A natural choice for $p(j)$ is again $p(j) \sim \exp(-Cm_j)$.

The Schauder system is well known in the context of constructions of Brownian motion, see for instance Rogers and Williams (2000). The Brownian motion case corresponds to prior regularity $\beta = 1/2$.

3. The Markov chain Monte Carlo sampler

3.1. Posterior within a fixed model

When continuous observations $x^T = (x_t : t \in [0, T])$ from the diffusion model (1) are available, then we have an explicit expression for the likelihood $p(x^T \mid b)$. Indeed, by Girsanov's formula we almost surely have

$$p(x^T \mid b) = \exp\left(\int_0^T b(x_t)\,\mathrm{d}x_t - \frac{1}{2}\int_0^T b^2(x_t)\,\mathrm{d}t\right). \tag{4}$$

Cf. e.g. Liptser and Shiryaev (2001). Note in particular that the log-likelihood is quadratic in b.

Due to the special choices in the construction of our hierarchical prior, the quadratic structure implies that within a fixed model j, we can do partly explicit posterior computations. More precisely, we can derive the posterior distributions of the scaling constant s^2 and the vector of coefficients θ^j conditional on all the other parameters. The continuous observations enter the expressions through the vector $\mu^j \in \mathbb{R}^{m_j}$ and the $m_j \times m_j$ matrix Σ^j defined by

$$\mu_l^j = \int_0^T \psi_l(x_t)\,\mathrm{d}x_t, \quad l = 1, \dots, m_j, \tag{5}$$

and

$$\Sigma_{l,l'}^j = \int_0^T \psi_l(x_t)\psi_{l'}(x_t)\,\mathrm{d}t, \quad l, l' = 1, \dots, m_j. \tag{6}$$

Lemma 1. *We have*

$$\theta^j \mid s^2, j, x^T \sim N_{m_j}((W^j)^{-1}\mu^j, (W^j)^{-1}),$$
$$s^2 \mid \theta^j, j, x^T \sim \mathrm{IG}(a + (1/2)m_j, b + (1/2)(\theta^j)^T(\Xi^j)^{-1}\theta^j),$$

where $W^j = \Sigma^j + (s^2\Xi^j)^{-1}$.

Proof. The computations are straightforward. We note that by Girsanov's formula (4) and the definitions of μ^j and Σ^j we have

$$p(x^T \mid j, \theta^j, s^2) = e^{(\theta^j)^T\mu^j - \frac{1}{2}(\theta^j)^T\Sigma^j\theta^j} \tag{7}$$

and by construction of the prior,

$$p(\theta^j \mid j, s^2) \propto (s^2)^{-\frac{m_j}{2}} e^{-\frac{1}{2}(\theta^j)^T(s^2\Xi^j)^{-1}\theta^j}, \qquad p(s^2) \propto (s^2)^{-a-1}e^{-\frac{b}{s^2}}.$$

It follows that

$$p(\theta^j \mid s^2, j, x^T) \propto p(x^T \mid j, \theta^j, s^2)p(\theta^j \mid j, s^2) \propto e^{(\theta^j)^T\mu^j - \frac{1}{2}(\theta^j)^T W^j\theta^j},$$

which proves the first assertion of the lemma. Next we write

$$\begin{aligned} p(s^2 \mid \theta^j, j, x^T) &\propto p(x^T \mid j, \theta^j, s^2)p(\theta^j \mid j, s^2)p(s^2) \\ &\propto (s^2)^{-m_j/2-a-1}\exp\left(-\frac{-b - (1/2)(\theta^j)^T(\Xi^j)^{-1}\theta^j}{s^2}\right), \end{aligned}$$

which yields the second assertion. □

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

The lemma shows that Gibbs sampling can be used to sample (approximately) from the continuous observations posteriors of s^2 and θ^j within a fixed model j. In the next subsections we explain how to combine this with reversible jump steps between different models and with data augmentation through the generation of diffusion bridges in the case of discrete-time observations.

3.2. *Reversible jumps between models*

In this subsection we still assume that we have continuous data $x^T = (x_t : t \in [0, T])$ at our disposal. We complement the within-model computations given by Lemma 1 with (a basic version of) reversible jump MCMC (cf. Green, 1995) to explore different models. We will construct a Markov chain which has the full posterior $p(j, \theta^j, s^2 \mid x^T)$ as invariant distribution and hence can be used to generate approximate draws from the posterior distribution of the drift function b.

We use an auxiliary Markov chain on $\mathbb{N}$ with transition probabilities $q(j' \mid j), j, j' \in \mathbb{N}$. As the notation suggests, we denote by $p(j \mid s^2, x^T)$ the conditional (posterior) probability of model j given the parameter s^2 and the data x^T. Recall that $p(j)$ is the prior probability of model j. Now we define the quantities

$$B(j' \mid j) = \frac{p(x^T \mid j', s^2)}{p(x^T \mid j, s^2)}, \tag{8}$$

$$R(j' \mid j) = \frac{p(j')q(j \mid j')}{p(j)q(j' \mid j)}. \tag{9}$$

Note that $B(j' \mid j)$ is the Bayes factor of model j' relative to model j, for a fixed scale s^2. To simplify the notation, the dependence of this quantity on s^2 and x^T is suppressed.

The overall structure of the algorithm that we propose is that of a componentwise Metropolis–Hastings (MH) sampler. The scale parameter s^2 is taken as component I and the pair (j, θ^j) as component II. Starting with some initial value $(j_0, \theta^{j_0}, s_0^2)$, alternately moves from (j, θ^j, s^2) to $(j, \theta^j, (s')^2)$ and moves from (j, θ^j, s^2) to $(j', \theta^{j'}, s^2)$ are performed, where in each case the other component remains unchanged.

Updating the first component is done with a simple Gibbs move, that is a new value for s^2 is sampled from its posterior distribution described by Lemma 1, given the current value of the remaining parameters.

Move I. Update the scale. Current state: (j, θ^j, s^2).

- Sample $(s')^2 \sim \mathrm{IG}(a + (1/2)m_j, b + (1/2)(\theta^j)^T(\Xi^j)^{-1}\theta^j)$.
- Update the state to $(j, \theta^j, (s')^2)$.

The second component (j, θ^j) has varying dimension and a reversible jump move is employed to ensure detailed balance (e.g. Green, 2003; Brooks et al., 2011). To perform a transdimensional move, first a new model j' is chosen and a sample from the posterior for θ^j given by Lemma 1 is drawn.

Move II. Transdimensional move. Current state: (j, θ^j, s^2).

- Select a new model j' with probability $q(j' \mid j)$.
- Sample $\theta^{j'} \sim N_{m_j'}((W^{j'})^{-1}\mu^{j'}, (W^{j'})^{-1})$.
- Compute $r = B(j' \mid j)R(j' \mid j)$.
- With probability $\min\{1, r\}$ update the state to $(j', \theta^{j'}, s^2)$, else leave the state unchanged.

All together we have now constructed a Markov chain $Z_0, Z_1, Z_2, \ldots$ on the transdimensional space $E = \bigcup_{j\in\mathbb{N}} \{j\} \times \mathbb{R}^{m_j} \times (0, \infty)$, whose evolution is described by Algorithm 1.

Algorithm 1 Continuous observations algorithm.

```
*Initialization*. Set $Z_0 = (j_0, \theta^{j_0}, s_0^2)$.
*Transition*. Given the current state $Z_j = (j, \theta^j, s^2)$:
• Sample $(s')^2 \sim \mathrm{IG}(a + (1/2)m_j, b + (1/2)(\theta^j)^T(\Xi^j)^{-1}\theta^j)$.
• Sample $j' \sim q(j' \mid j)$,
• Sample $\theta^{j'} \sim N_{m_j'}((W^{j'})^{-1}\mu^{j'}, (W^{j'})^{-1})$.
• With probability $\min\{1, B(j' \mid j)R(j' \mid j)\}$, set
  $Z_{j+1} = (j', \theta^{j'}, (s')^2)$, else set $Z_{j+1} = (j, \theta^j, (s')^2)$.
```

Note that $r = (q(j \mid j')/q(j \mid j'))(p(j' \mid s^2, x^T)/p(j' \mid s^2, x^T))$, so effectively we perform Metropolis–Hastings for updating j. As a consequence, the vector of coefficients $\theta^{j'}$ needs to be drawn only in case the proposed j' is accepted.

The following lemma asserts that the constructed chain indeed has the desired stationary distribution.

Lemma 2. *The Markov chain $Z_0, Z_1, \ldots$ has the posterior $p(j, \theta^j, s^2 \mid x^T)$ as invariant distribution.*

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

Proof. By Lemma 1 we have that in Move I the chain moves from state (j, θ^j, s^2) to (j, θ^j, s'^2) with probability $p(s'^2 \mid j, \theta^j, x^T)$. Conditioning shows that we have detailed balance for this move, that is,

$$p(j, \theta^j, s^2 \mid x^T)p(s'^2 \mid j, \theta^j, x^T) = p(j, \theta^j, s'^2 \mid x^T)p(s^2 \mid j, \theta^j, x^T). \tag{10}$$

In view of Lemma 1 again, the probability that the chain moves from state (j, θ^j, s^2) to $(j', \theta^{j'}, s^2)$ in Move II equals, by construction,

$$p((j, \theta^j, s^2) \to (j', \theta^{j'}, s^2)) = \min\left\{1, \frac{q(j \mid j')}{q(j' \mid j)} \frac{p(j' \mid s^2, x^T)}{p(j \mid s^2, x^T)}\right\} q(j' \mid j)p(\theta^{j'} \mid j', s^2, x^T).$$

Now suppose first that the minimum is less than 1. Then using

$$p(j, \theta^j, s^2 \mid x^T) = p(\theta^j \mid j, s^2, x^T)p(j \mid s^2, x^T)p(s^2 \mid x^T)$$

and

$$p(j', \theta^{j'}, s^2 \mid x^T) = p(\theta^{j'} \mid j', s^2, x^T)p(j' \mid s^2, x^T)p(s^2 \mid x^T)$$

it is easily verified that we have the detailed balance relation

$$p(j, \theta^j, s^2 \mid x^T)p((j, \theta^j, s^2) \to (j', \theta^{j'}, s^2)) = p(j', \theta^{j'}, s^2 \mid x^T)p((j', \theta^{j'}, s^2) \to (j, \theta^j, s^2))$$

for Move II. The case that the minimum is greater than 1 can be dealt with similarly.

We conclude that we have detailed balance for both components of our MH sampler. Since our algorithm is a variable-at-a-time Metropolis–Hastings algorithm, this implies the statement of the lemma (see for example Section 1.12.5 of Brooks et al., 2011). □

3.3. *Data augmentation for discrete data*

So far we have been dealing with continuously observed diffusion. Obviously, the phrase "continuous data" should be interpreted properly. In practice it means that the frequency at which the diffusion is observed is so high that the error that is incurred by approximating the quantities (5) and (6) by their empirical counterparts, is negligible. If we only have low-frequency, discrete-time observations at our disposal, these approximation errors can typically not be ignored however and can introduce undesired biases. In this subsection we explain how our algorithm can be extended to accommodate this situation as well.

We assume now that we only have partial observations $x_0, x_\Delta, \ldots, x_{n\Delta}$ of our diffusion process, for some $\Delta > 0$ and $n \in \mathbb{N}$. We set $T = n\Delta$. The discrete observations constitute a Markov chain, but it is well known that the transition densities of discretely observed diffusions and hence the likelihood are not available in closed form in general. This complicates a Bayesian analysis. An approach that has been proven to be very fruitful, in particular in the context of parametric estimation for discretely observed diffusions, is to view the continuous diffusion segments between the observations as missing data and to treat them as latent (function-valued) variables. Since the continuous-data likelihood is known (cf. the preceding subsection), data augmentation methods (see Tanner and Wong, 1987) can be used to circumvent the unavailability of the likelihood in this manner.

As discussed in Van der Meulen and van Zanten (2013) and shown in a practical setting by Papaspiliopoulos et al. (2012), the data augmentation approach is not limited to the parametric setting and can be used in the present nonparametric problem as well. Practically it involves appending an extra step to the algorithm presented in the preceding subsection, corresponding to the simulation of the appropriate diffusion bridges. If we denote again the continuous observations by $x^T = (x_t : t \in [0, T])$ and the discrete-time observations by $x_\Delta, \ldots, x_{n\Delta}$, then using the same notation as above we essentially want to sample from the conditional distribution

$$p(x^T \mid j, \theta^j, s^2, x_\Delta, \ldots, x_{n\Delta}). \tag{11}$$

Exact simulation methods have been proposed in the literature to accomplish this, e.g. Beskos, Papaspiliopoulos, Roberts et al. (2006) and Beskos, Papaspiliopoulos, Roberts, Fearnhead et al. (2006). For our purposes exact simulation is not strictly necessary however and it is more convenient to add a Metropolis–Hastings step corresponding to a Markov chain that has the diffusion bridge law given by (11) as stationary distribution.

Underlying the MH sampler for diffusion bridges is the fact that by Girsanov's theorem, the conditional distribution of the continuous segment $X^{(k)} = (X_t : t \in [(k-1)\Delta, k\Delta])$ given that $X_{(k-1)\Delta} = x_{(k-1)\Delta}$ and $X_{k\Delta} = x_{k\Delta}$, is equivalent to the distribution of a Brownian bridge that goes from $x_{(k-1)\Delta}$ at time $(k-1)\Delta$ to $x_{k\Delta}$ at time $k\Delta$. The corresponding Radon–Nikodym derivative is proportional to

$$L_k(X^{(k)} \mid b) = \exp\left(\int_{(k-1)\Delta}^{k\Delta} b(X_t)\, \mathrm{d}X_t - \frac{1}{2}\int_{(k-1)\Delta}^{k\Delta} b^2(X_t)\, \mathrm{d}t\right). \tag{12}$$

We also note that due to the Markov property of the diffusion, the different diffusion bridges $X^{(1)}, \dots, X^{(n)}$, can be dealt with independently.

Concretely, the missing segments $x^{(k)} = (x_t : t \in ((k-1)\Delta, k\Delta))$, $k = 1, \dots, n$ can be added as latent variables to the Markov chain constructed in the preceding subsection, and the following move has to be added to Moves I and II introduced above. It is a standard Metropolis–Hastings step for the conditional law (11), with independent Brownian bridge proposals. For more details on this type of MH samplers for diffusion bridges we refer to Roberts and Stramer (2001).

Move III. Updating the diffusion bridges. Current state: $(j, \theta^j, s^2, x^{(1)}, \dots, x^{(n)})$:

- For $k = 1, \dots, n$, sample a Brownian bridge $w^{(k)}$ from $((k-1)\Delta, x_{(k-1)\Delta})$ to $(k\Delta, x_{k\Delta})$.
- For $k = 1, \dots, n$, compute $r_k = L_k(w^{(k)} \mid b)/L_k(x^{(k)} \mid b)$, for $b = \sum_{l \le m_j} \theta_l^j \psi_l$.
- Independently, for $k = 1, \dots, n$, with probability $\min\{1, r_k\}$ update $x^{(k)}$ to $w^{(k)}$, else retain $x^{(k)}$.

Of course the segments $x^{(1)}, \dots, x^{(n)}$ can always be concatenated to yield a continuous function on $[0, T]$. In this sense Move III can be viewed as a step that generates new, artificial continuous data given the discrete-time data. It is convenient to consider this whole continuous path on $[0, T]$ as the latent variable. When the new move is combined with the ones defined earlier a Markov chain $\widetilde{Z}_0, \widetilde{Z}_1, \widetilde{Z}_2, \dots$ is obtained on the space $\widetilde{E} = \bigcup_{j \in \mathbb{N}} \{j\} \times \mathbb{R}^{m_j} \times (0, \infty) \times C[0, T]$. Its evolution is described by Algorithm 2.

Algorithm 2 Discrete observations algorithm.

```
Initialization. Set $\tilde{Z}_0 = (j_0, \theta^{j_0}, s_0^2, x_0^T)$, where $x_0^T$ is for instance
obtained by linearly interpolating the observed data points.
Transition. Given the current state $\tilde{Z}_j = (j, \theta^j, s^2, x^T)$, construct $\tilde{Z}_{j+1}$ as follows:
• Sample $(s')^2 \sim \mathrm{IG}(a + (1/2)m_j, b + (1/2)(\theta^j)^T (\Xi^j)^{-1}\theta^j)$, update $s^2$ to $(s')^2$.
• Sample $j' \sim q(j' \mid j)$ and $\theta^{j'} \sim N_{m_j'}((W^{j'})^{-1}\mu^{j'}, (W^{j'})^{-1})$.
• With probability $\min\{1, B(j' \mid j)R(j' \mid j)\}$, update $(j, \theta^j)$ to $(j', \theta^{j'})$, else retain $(j, \theta^j)$.
• For $k = 1, \dots, n$, sample a Brownian bridge $w^{(k)}$ from $((k-1)\Delta, x_{(k-1)\Delta})$
  to $(k\Delta, x_{k\Delta})$ and compute $r_k = L_k(w^{(k)} \mid b)/L_k(x^{(k)} \mid b)$, for $b = \sum_{l \le m_j} \theta_l^j \psi_l$.
• Independently, with probability $\min\{1, r_k\}$, update $x^{(k)}$ to $w^{(k)}$, else retain $x^{(k)}$.
```

It follows from the fact that Move III is a MH step for the conditional law (11) and Lemma 2 that the new chain has the correct stationary distribution again.

4. Simulation results

The implementation of the algorithms presented in the preceding section involves the computation of several quantities, including the Bayes factors $B(j' \mid j)$ and sampling from the posterior distribution of θ^j given j and s^2. In Section 5 we explain in some detail how these issues can be tackled efficiently. In the present section we first investigate the performance of our method on simulated data.

For the drift function, we first choose the function $b(x) = 12(a(x) + 0.05)$ where

$$a(x) = \begin{cases} \dfrac{2}{7} - x - \dfrac{2}{7}(1 - 3x)\sqrt{|1 - 3x|} & x \in [0, 2/3) \\ -\dfrac{2}{7} + \dfrac{2}{7}x & x \in [2/3, 1]. \end{cases} \tag{13}$$

This function is Hölder-continuous of order 1.5 on $[0, 1]$. A plot of b and its derivative is shown in Fig. 1. Clearly, the derivative is not differentiable in 0, 1/3 and 2/3.

We simulated a diffusion on the time interval $[0, 200]$ using the Euler discretization scheme with time discretization step equal to 10^{-5}. Next, we retained all observations at values $t = i\Delta$ with $\Delta = 0.001$ and $i = 0, \dots, 200.000$. The data are shown in Fig. 2. From the histogram we can see that the process spends most time near $x = 1/3$, so we expect estimates for the drift to be best in this region. For now, we consider the data as essentially continuous time data, so no data-augmentation scheme is employed.

We define a prior with Fourier basis functions as described in Section 2.2.1, choosing regularity $\beta = 1.5$. With this choice, the regularity of the prior matches that of the true drift function.

For the reversible jump algorithm there are a few tuning parameters. For the model generating chain, we choose $q(j \mid j) = 1/2, q(j+1 \mid j) = q(j-1 \mid j) = 1/4$. For the priors on the models we chose $C = -\log(0.95)$ which means that $p(j) \propto (0.95)^{m_j}$ expressing weak prior belief in a course (low) level model. For the inverse Gamma prior on the scale we take the hyper parameters $a = b = 5/2$.

We ran the continuous time algorithm for 3000 cycles and discarded the first 500 iterations as burn-in. We fix this number of iterations and burn-in for all other MCMC simulations. In Fig. 3 we show the resulting posterior mean (dashed curve) and

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

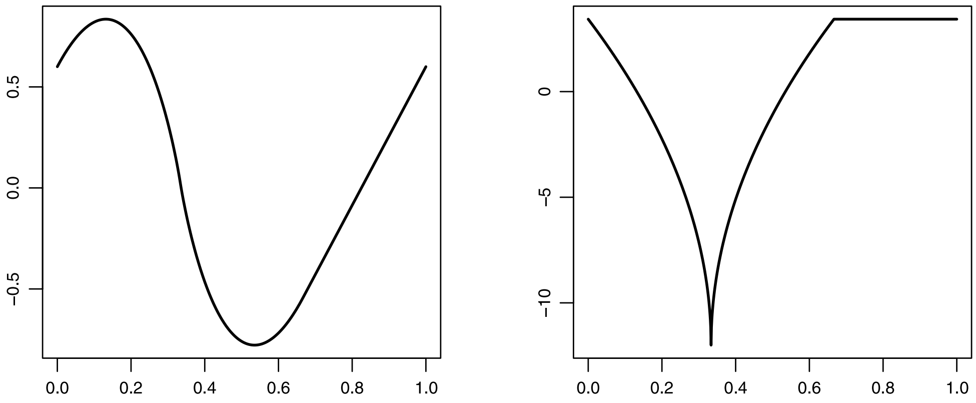

Fig. 1. Left: drift function. Right: derivative of the drift function.

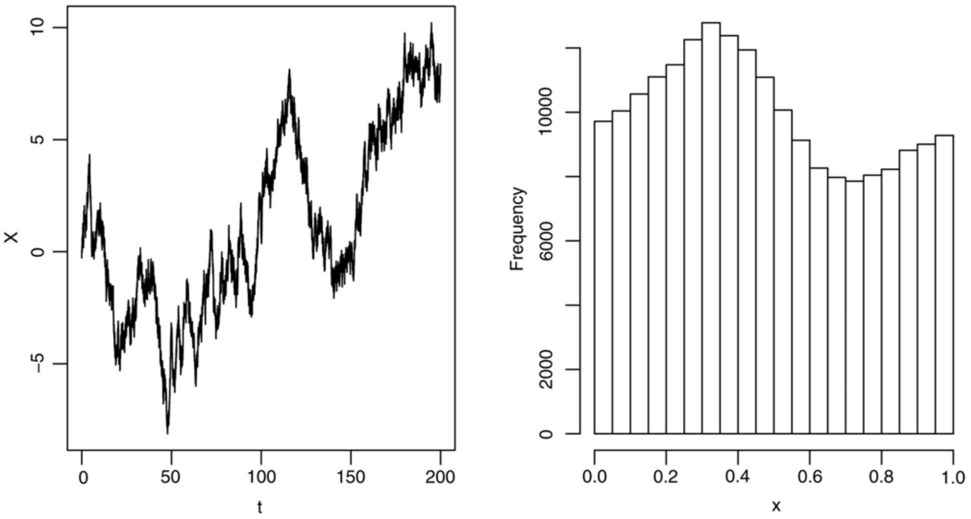

Fig. 2. Left: simulated data. Right: histogram of simulated data modulo 1.

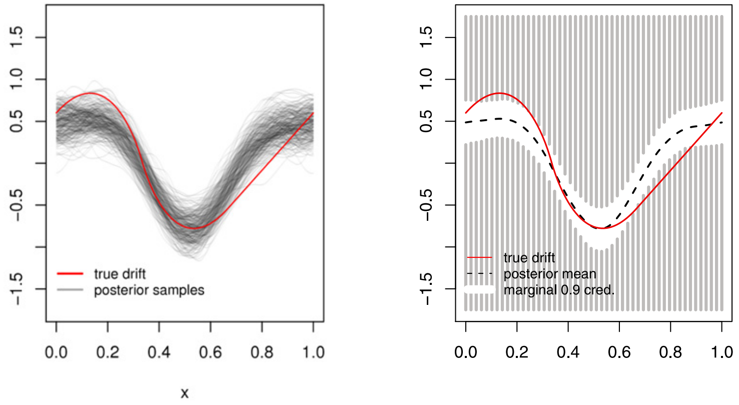

Fig. 3. Left: true drift function (red, solid) and samples from the posterior distribution. Right: drift function (red, solid), posterior mean (black, dashed) and 90% pointwise credible bands.

90% pointwise credible intervals (visualized by the gray bars). The posterior mean was estimated using Rao–Blackwellization (Robert, 2007, Section 4.2). (Specifically, the posterior mean drift was not computed as the pointwise average of the drift functions b sampled at each iteration. Rather, the average of the posterior means $(W^{j'})^{-1}\mu^{j'}$ obtained at each MCMC iteration (see Move II) was used.)

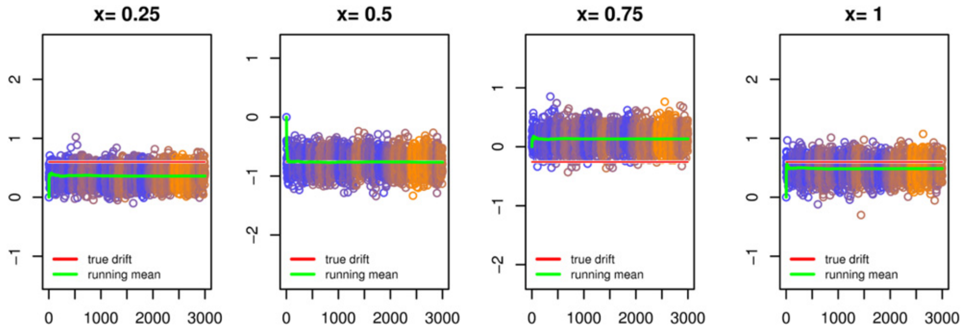

Fig. 4. Trace and running mean of the sampled drift at different design points. The color of the samples indicates the current model, cold colors correspond to small values of j.

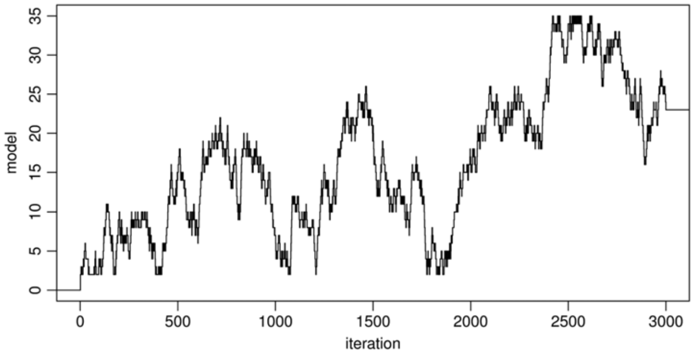

Fig. 5. Models visited over time.

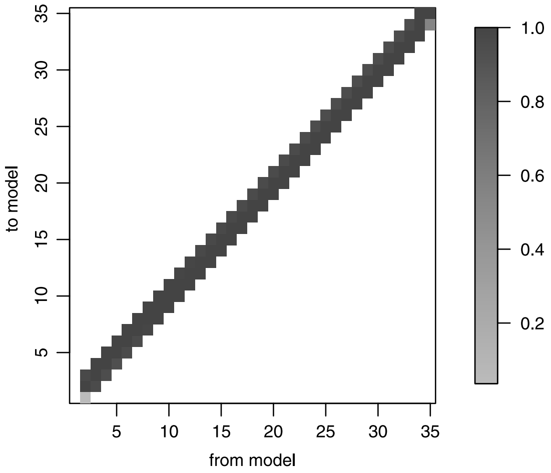

Fig. 6. Average acceptance probabilities for moves between models.

Insight in the mixing properties of the Markov chain is gained by considering traces of the sampled drift function at several fixed points as shown in Fig. 4. The trace plots indicate that indeed the first 200–300 iterations should be discarded. Plots of the visited models over time and the corresponding acceptance probabilities are shown in Figs. 5 and 6 respectively. The mean and median of the scaling parameter s^2 are given by 1.91 and 1.64 respectively (computing upon first removing burn-in samples).

To judge the algorithm with respect to the sensitivity of C, we ran the same algorithm as well for $C = 0$. The latter choice is often made and reflects equals prior belief on all models under consideration. If $C = 0$, then the chain spends more time in higher models. However, the posterior mean and pointwise credible bands are practically equal to the case $C = -\log(0.95)$.

To get a feeling for the sensitivity of the results on the choice of the hyper parameters a and b, the analysis was redone for $a = b = 5$. The resulting posterior mean and credible bands turned out to be indistinguishable from the case $a = b = 2.5$.

Clearly, if we would have chosen an example with less data, then the influence of the priors would be more strongly pronounced in the results.

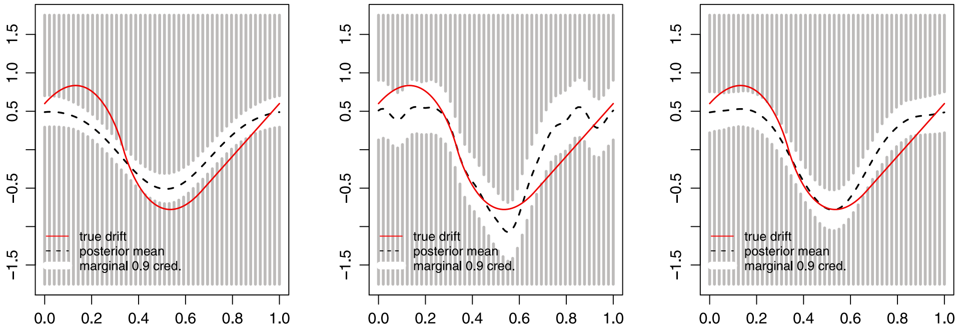

Fig. 7. Drift function (red, solid), posterior mean (black, dashed) and 90% pointwise credible bands. Left: $s^2 = 0.25$ fixed. Middle: $s^2 = 50$ fixed. Right: random scaling parameter (inverse gamma prior with $a = b = 2.5$).

4.1. *Effect of the prior on the model index*

If, as in the example thus far, the parameter β is chosen to match the regularity of the true drift, one would expect that using a prior where the truncation point for in the series expansion of the drift is fixed at a high enough level, one would get results comparable to those we obtained in Fig. 3. If we fix the level at $j = 30$, this indeed turns out to be the case. The main advantage of putting a prior on the truncation level and using a reversible jump algorithm is an improvement in computing time. For this example, a simulation run for the reversible jump model took about 55% of that for the fixed dimensional model. For the reversible jump run, the average of the (non-burn-in) samples of the model index j equals 18.

4.2. *Effect of the random scaling parameter*

To assess the effect of including a prior on the multiplicative scaling parameter, we ran the same simulation as before, though keeping s^2 fixed to either 0.25 (too small) or 50 (too large). For both cases, the posterior mean, along with 90% pointwise credible bounds, is depicted in Fig. 7. For ease of comparison, we added the right-hand-figure of Fig. 3. Clearly, fixing $s^2 = 0.25$ results in oversmoothing. For $s^2 = 50$, the credible bands are somewhat wider and suggest more fluctuations in the drift function than are actually present.

4.3. *Effect of misspecifying smoothness of the prior*

If we consider the prior without truncation, then the smoothness of the prior is essentially governed by the decay of the variances on the coefficients. This decay is determined by the value of β. Here, we investigate the effect of misspecifying β. We consider $\beta \in \{0.25, 1.5, 3\}$. In Fig. 8 one can assess the difference in posterior mean, scaling parameter and models visited for these three values of β. Naturally, if β is too large, higher level models and relatively large values for s^2 are chosen to make up for the overly fast decay on the variances of the coefficients. Note that the boxplots are for $\log(s^2)$, not s^2.

It is interesting to investigate what happens if the same analysis is done without the reversible jump algorithm, thus fixing a high level truncation point. The results are in Fig. 9. From this figure, it is apparent that if β is too small the posterior mean is very wiggly. At the other extreme, if β is too large, we are oversmoothing and the true drift is outside the credible bands except near $x = 1/3$. As such, misspecifying β can result in very bad estimates if a high level is fixed. From the boxplots in Figs. 8 and 9 one can see that the larger β, the larger the scaling parameter. Intuitively, this makes sense. Moreover, in case we employ a reversible jump algorithm, a too small (large) value for β is compensated for by low (high) level models.

4.4. *Results with Schauder basis*

The complete analysis has also been done for the Schauder basis. Here we take $q(j \mid j) = 0.9$, $q(j+1 \mid j) = q(j-1 \mid j) = 0.1$. The conclusions are as before. For the main example, the computing time with the Schauder basis was approximately 15% of that for the Fourier basis.

4.5. *Discrete-time data and data-augmentation*

Here, we thin the "continuous"-time data to discrete-time data by retaining every 50th observation from the continuous-time data. The remaining data are hence at times $t = i\Delta$ with $\Delta = 0.05$ and $i = 0, \dots, 4000$. Next, we use our algorithm both with and without data-augmentation. Here, we used the Schauder basis to reduce computing time. The results are depicted in Fig. 10.

The leftmost plot clearly illustrates that the discrete-time data with $\Delta = 0.5$ cannot be treated as continuous-time data. The bias is quite apparent. Comparing the middle and the rightmost plot shows that data-augmentation works very well in this example.

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

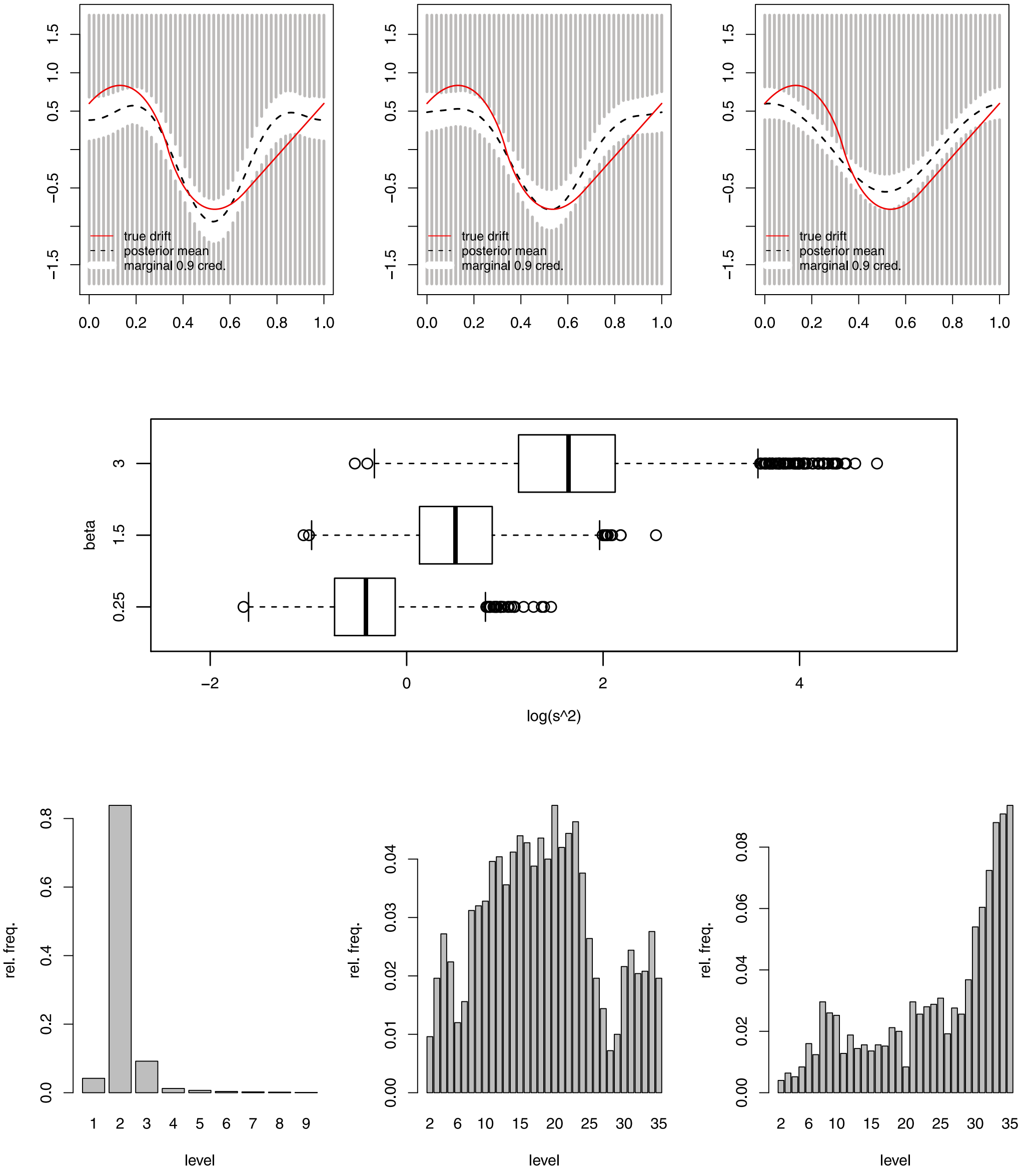

Fig. 8. Reversible jump; from left to right $\beta = 0.25$, $\beta = 1.5$ and $\beta = 3$. Upper figures: posterior mean and 90% pointwise credible bands. Middle figure: boxplots of $\log(s^2)$. Lower figures: histogram of model-index.

We also looked at the effect of varying T (observation time horizon) and Δ (time in between discrete time observations). In all plots of Fig. 11 we used the Schauder basis with $\beta = 1.5$ and data augmentation with 49 extra imputed points in between two observations. In the upper plots of Fig. 11 we varied the observation time horizon while keeping $\Delta = 0.1$ fixed. In the lower plots of Fig. 11 we fixed $T = 500$ and varied Δ. As expected, we see that the size of the credible bands decreases as the amount of information in the data grows.

Lastly, Fig. 12 illustrates the influence of increasing the number of augmented observations on the mixing of the chain. Here we took $\Delta = 0.2$ and $T = 500$ and compare trace plots for two different choices of the number of augmented data points, in one case 25 data points per observation and in the second case 100 data points per observations. The mixing does not seem to deteriorate with a higher number of augmented observations.

4.6. Performance of the method for various drift functions

In this section we investigate how different features of the drift function influence the results of our method. The drift functions chosen for the comparison are

1. $b_1(x) = 8\sin(4\pi x)$,

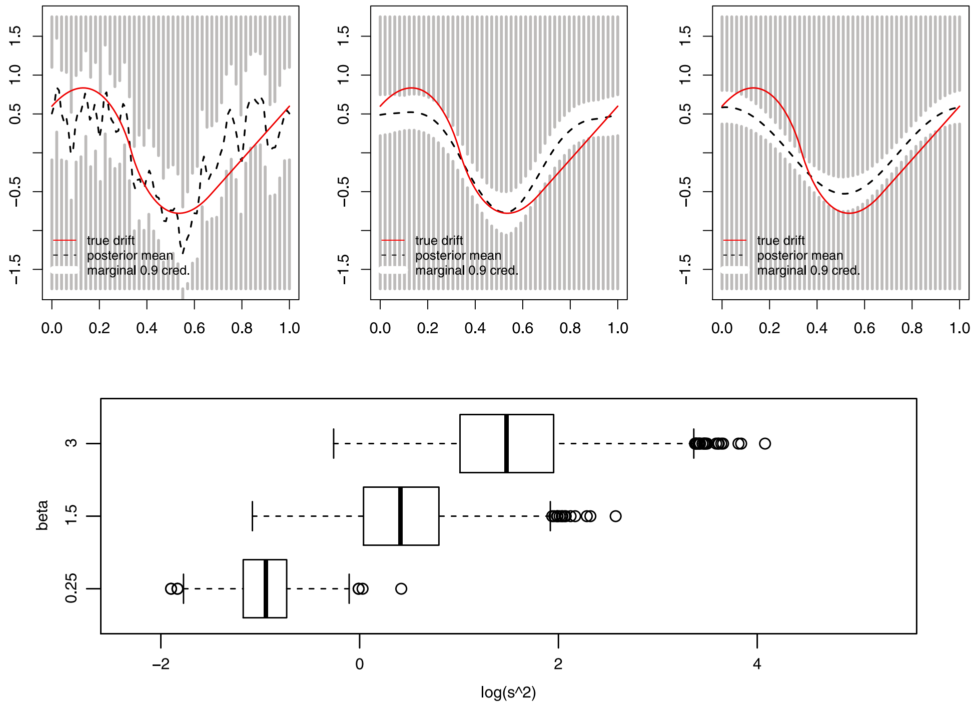

Fig. 9. Fixed level; from left to right $\beta = 0.25$, $\beta = 1.5$ and $\beta = 3$. Upper figures: posterior mean and 90% pointwise credible bands. Lower figure: boxplots of $\log(s^2)$.

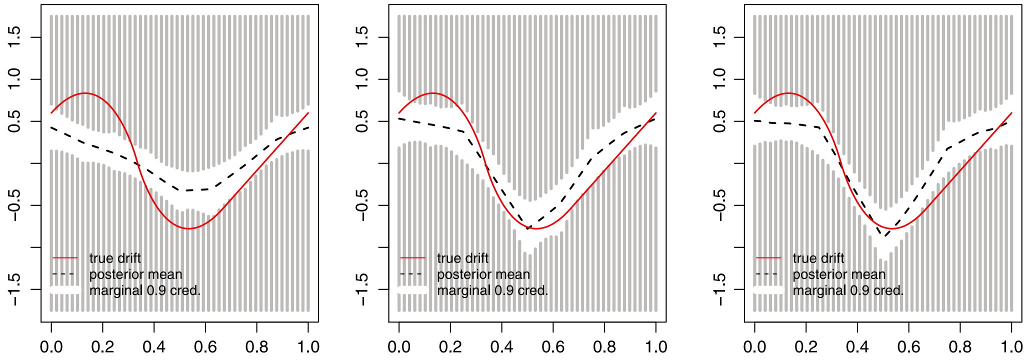

Fig. 10. Drift function (red, solid), posterior mean (black, dashed) and 90% pointwise credible bands. Discrete-time data. Left: without data-augmentation. Middle: with data-augmentation. Right: continuous-time data.

2. $b_2(x) = 200\widetilde{x}(1-2\widetilde{x})^3 1_{\left[0,\frac{1}{2}\right)}(\widetilde{x}) - \frac{400}{3}(1-\widetilde{x})(2\widetilde{x}-1)^3 1_{\left[\frac{1}{2},1\right)}(\widetilde{x})$, where $\widetilde{x} = x \bmod 1$,
3. $b_3(x) = -8\sin(\pi(4x-1))1_{\left[\frac{1}{4},\frac{3}{4}\right]}(x \bmod 1)$.

As a prior we took the Fourier basis with parameter $\beta = 1.5$. For s^2 we took an inverse Gamma prior with hyper parameters $a = b = 5/2$. For each drift function, 2000 observations with $\Delta = 0.1$ were generated. The algorithm was then used with data augmentation with 49 imputed points extra in between two observations. The results for b_1, b_2 and b_3 are in Fig. 13. We ran the analysis for the Schauder basis as well, which led to very similar results.

4.7. Comparison with Papaspiliopoulos et al. (2012): butane dihedral angle time series

To compare our approach to that of Papaspiliopoulos et al. (2012) we analyzed the butane dihedral angle time series considered by these authors. After some preliminary operations on the data, these data are assumed to be discrete time observations from a scalar diffusion with unit-diffusion coefficient (details on this are described in Papaspiliopoulos et al. (2012) and supplementary material to this article). After this preliminary step, the time series consists of 4000 observations

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

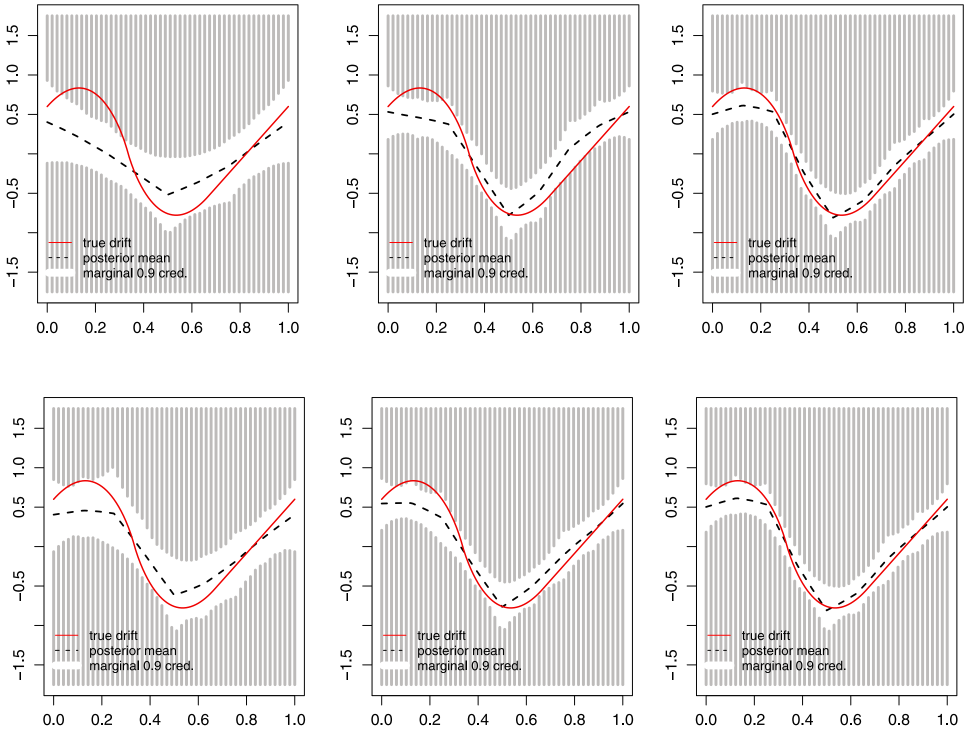

Fig. 11. Drift function (red, solid), posterior mean (black, dashed) and 90% pointwise credible bands. Upper: discrete time observations with $\Delta = 0.1$. From left to right $T = 50$, $T = 200$ and $T = 500$. Lower: discrete time observations with $T = 500$. From left to right $\Delta = 1$, $\Delta = 0.2$ and $\Delta = 0.1$. (All augmented to $\delta = 0.002$.)

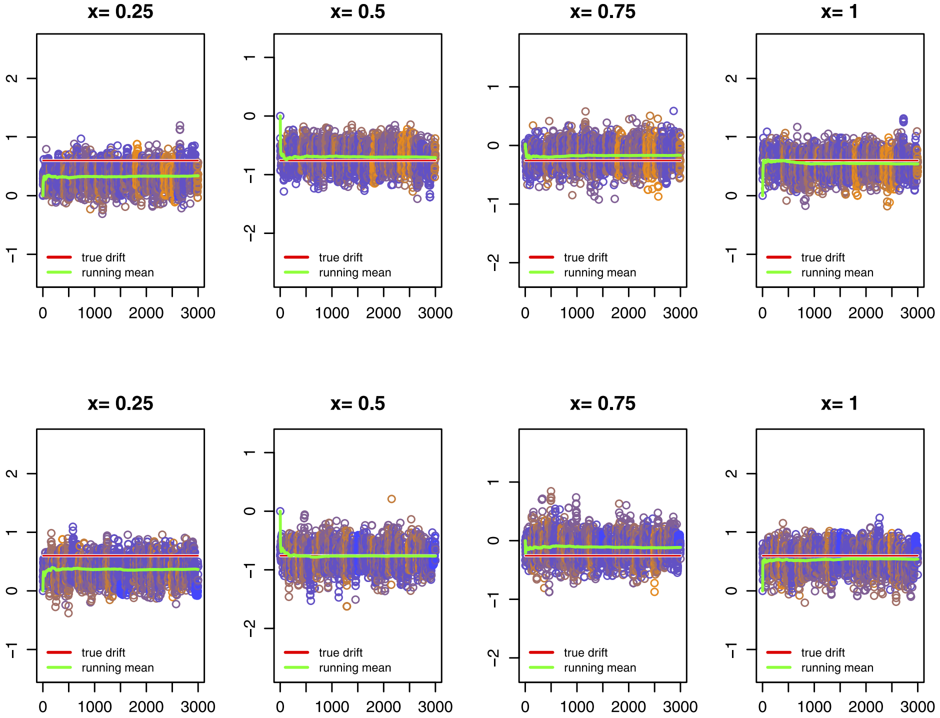

Fig. 12. Trace plots illustrating the influence data augmentation on the mixing of the chain. 2500 observations in [0, 500]. Top: 25 data points per observation. Bottom: 100 data points per observation.

Please cite this article in press as: van der Meulen, F., et al., Reversible jump MCMC for nonparametric drift estimation for diffusion processes. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.03.002

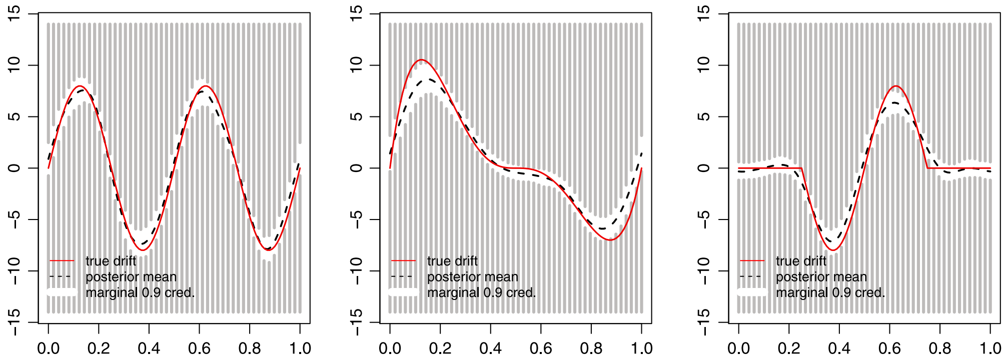

Fig. 13. Drift function (red, solid), posterior mean (black, dashed) and 90% pointwise credible bands. From left to right: b_1, b_2, and b_3.

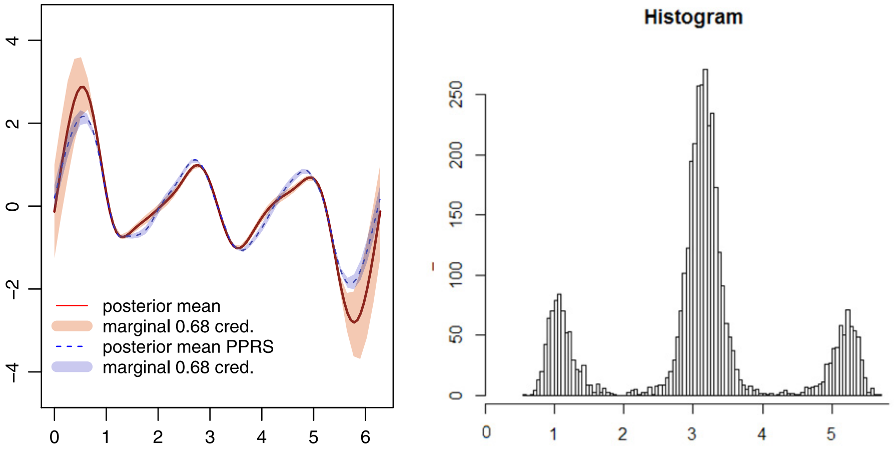

Fig. 14. Comparison of the estimate of drift using the butane dihedral angle data. Red solid: a Fourier prior with $\beta = 1.5$. Blue dashed: results of Papaspiliopoulos et al. (2012). The posterior mean with 68% credible bands is pictured. Right: histogram of the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

observed evenly over the time interval [0, 4] (time is measured in nanoseconds). The right-hand figure of Fig. 14 shows a histogram of the discrete time observations.

Papaspiliopoulos et al. (2012) use a centered Gaussian process prior with precision operator (2), with $\eta = 0.02$, $\kappa = 0$ and $p = 2$. This choice for p yields a prior of Hölder smoothness essentially equal to 1.5. As explained in the introduction, this is essentially the law of the random function

$$x \mapsto \sum_{l=1}^{\infty} \frac{1}{l^2\pi^2\sqrt{\eta}} Z_l \psi_l(x),$$

where ψ_l are the Fourier basis functions defined in Section 2.2.1 and the Z_l are independent standard normal random variables. Note that with $\eta = 0.02$ we have $(\pi^4\eta)^{-1} \approx 0.51$. To match as closely as possible with their prior specification, we use the Fourier basis with $\beta = 1.5$, as described in Section 2.2.1. Conditionally on s^2 and j, our prior then equals the law of the random function

$$x \mapsto \sum_{l=1}^{2j-1} \frac{s}{l^2} Z_l \psi_l(x).$$

For s^2 we took an inverse Gamma prior with hyper parameters $a = b = 5/2$. For this choice the prior mean for s^2 equals 0.5 which is close to $(\pi^4\eta)^{-1}$.

For the prior on the models we use $p(j) \propto (0.95)^{m_j}$. As before, we ran the continuous time algorithm for 3000 cycles and discarded the first 500 iterations as burn-in. We used data augmentation with 99 extra augmented time points in between two successive observations.

The left-hand figure in Fig. 14 shows the posterior mean and pointwise 68% credible bands, both for our approach and that of Papaspiliopoulos et al. (2012). Overall, the posterior means computed by both methods seem to agree very well except for the boundary areas. In these areas, the credible bounds are wider, since we have less information about the drift here. In Fig. 15 histograms of the scaling parameter s^2 and the truncation level J are shown. Clearly, s^2 takes values typically

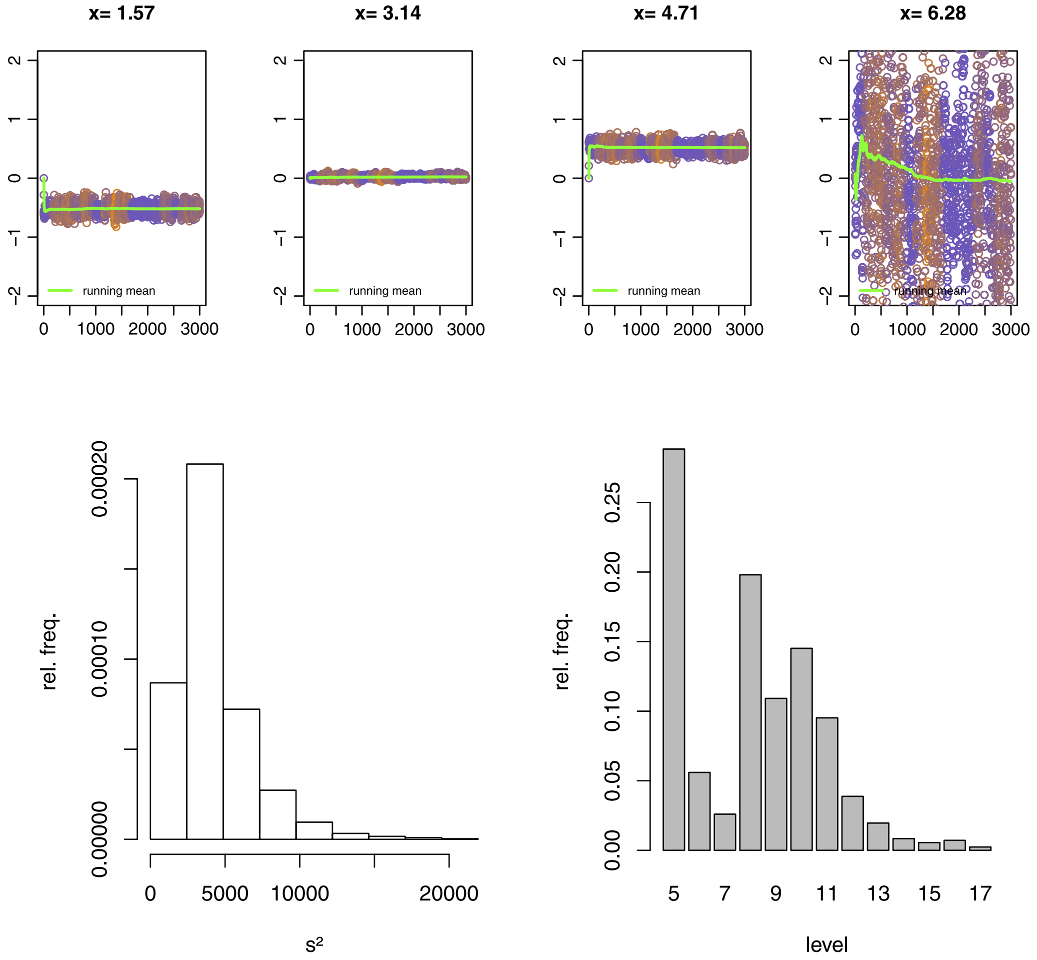

Fig. 15. Trace plot, histogram of j, histogram of s^2.

around a value as large as 5000, much larger than 0.51. This illustrates once again the usefulness of equipping the scaling parameter with a prior distribution. The fact that our credible bands are wider near the boundary of the observation area seems to indicate that Papaspiliopoulos et al. (2012) are somewhat overconfident about the form of the drift function in that area. Their narrower credible bands seem to be caused by prior belief rather than information in the data and are not corroborated by our more conservative approach.

5. Numerical considerations

5.1. Drawing coefficients from the posterior within a fixed model

For Move II the algorithm requires to sample a random vector $U \sim N_{m_j}((W^j)^{-1}\mu^j, (W^j)^{-1})$. In order to do so we first compute the Cholesky decomposition of W^j (note that W^j is symmetric and positive definite, ensuring its existence). For an upper triangular matrix M^j we then have $W^j = (M^j)^T M^j$. Next we let z^j solve the system $(M^j)^T z^j = \mu^j$, draw a standard normal vector $Z \sim \mathrm{N}_{m_j}(0, I)$ and construct U by backward solving

$$M^j U = z^j + Z. \tag{14}$$

It is easily seen that the random vector U has the required distribution.

Backsolving linear equations with triangular matrices requires $\mathcal{O}(m_j^2)$ operations. Cholesky factors are computed in $\mathcal{O}(m_j^3)$ operations in general, but for basis functions with local support, Σ^j and W^j are sparse, enabling faster computations. For the Schauder basis, the number of non-zero elements of the upper triangular part of Σ^j is $2^{j-1}(j-1)+1$, so the fraction of non-zero elements of Σ^j is approximately 1.00, 1.00, 0.88, 0.66, 0.45, 0.28, 0.17, . . . for $j = 1, 2, 3, 4, 5, 6, 7, \ldots$. The Cholesky factor of a sparse matrix is not necessarily sparse as well. However, the sparsity pattern originating from the tree structure of the supports of the Schauder elements enables to specify a *perfect elimination ordering* of the rows and columns of Σ^j (for details we refer to Rose, 1970). This means that the Cholesky factor inherits the sparsity. Moreover, the Cholesky factorization can be computed on the sparse representation of the matrix Σ^j. The particular reordering necessary — reversing the order of rows and columns — makes this technique applicable for moves within levels.

5.2. *Computation of the Bayes factors*

Our algorithms require the evaluation of the Bayes factors defined by (8). The following lemma is instrumental in the numerical evaluation of these numbers. Recall the definitions of μ^j, Σ^j and W^j in Section 3.1.

Lemma 3. *We have*

$$p(x^T \mid j, s^2) = \frac{\exp\left(\frac{1}{2}(\mu^j)^T (W^j)^{-1} \mu^j\right)}{\sqrt{|s^2 W^j \Xi^j|}}.$$

Proof. Since

$$p(x^T \mid j, s^2) = \int p(x^T \mid j, \theta^j, s^2) p(\theta^j \mid j, s^2) d\theta^j$$

we have, by (7) and the definition of the prior,

$$p(x^T \mid j, s^2) = \frac{1}{\sqrt{|2\pi s^2 \Xi^j|}} \int e^{(\theta^j)^T \mu^j - \frac{1}{2}(\theta^j)^T W^j \theta^j} d\theta^j. \tag{15}$$

By completing the square we see that this is further equal to

$$\frac{1}{\sqrt{|2\pi s^2 \Xi^j|}} e^{\frac{1}{2}(\mu^j)^T (W^j)^{-1} \mu^j} \int e^{-\frac{1}{2}(\theta^j - (W^j)^{-1}\mu^j)^T W^j (\theta^j - (W^j)^{-1}\mu^j)} d\theta^j = \frac{1}{\sqrt{|2\pi s^2 \Xi^j|}} e^{\frac{1}{2}(\mu^j)^T (W^j)^{-1} \mu^j} \sqrt{|2\pi (W^j)^{-1}|}.$$

This completes the proof. □

As a consequence of Lemma 3, we have

$$2 \log B(j' \mid j) = (\mu^{j'})^T (W^{j'})^{-1} \mu^{j'} - (\mu^j)^T (W^j)^{-1} \mu^j + \log\left(\frac{|s^2 W^j \Xi^j|}{|s^2 W^{j'} \Xi^{j'}|}\right). \tag{16}$$

We now show how the right-hand-side of the display can be evaluated in a numerical efficient and stable way. In the context of Gaussian Markov random fields related tricks have been used in Rue et al. (2009).

Suppose $j' - j = k > 0$ (if $k = 0$, $B(j' \mid j) = 1$ and for $k < 0$ the calculations are similar). First we compute μ^{j+k} and the Cholesky decomposition of W^{j+k} (the matrix is symmetric and positive definite, so its Cholesky decomposition exists). We obtain an upper triangular matrix M^{j+k} such that

$$W^{j+k} = (M^{j+k})^T M^{j+k}.$$

Next we apply the following theorem, taken from Stewart (1998, cf. Theorem 1.6 therein).

Theorem 4. *Suppose the matrix A can be portioned as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is nonsingular. Then A has a block LU decomposition

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}$$

where L_{11} and U_{11} are nonsingular. For such decomposition $A_{11} = L_{11}U_{11}$.

The Cholesky decomposition factor M^j of W^j hence equals the upper left block of M^{j+k}, which is obtained by retaining only the first m_j rows and columns of M_{j+k}. Also note that the vector μ^j is obtained from μ^{j+k} by retaining only the first m_j elements.

Now if z^{j+k} is the solution to $(M^{j+k})^T z^{j+k} = \mu^{j+k}$, then $(\mu^{j+k})^T (W^{j+k})^{-1} \mu^{j+k} = \|z^{j+k}\|^2$. If we similarly define z^j as the solution to $(M^j)^T z^j = \mu^j$, then $z^{j+k} = [z^j, g^{j+k}]$ where g^{j+k} contains the last $m_{j+k} - m_j$ elements of z^{j+k}. Therefore

$$(\mu^{j+k})^T (W^{j+k})^{-1} \mu^{j+k} - (\mu^j)^T (W^j)^{-1} \mu^j = \|z^{j+k}\|^2 - \|z^j\|^2 = \|g^{j+k}\|^2.$$

Furthermore,

$$\log\left(\frac{|s^2 W^j \Xi^j|}{|s^2 W^{j+k} \Xi^{j+k}|}\right) = -\sum_{i=m_j+1}^{m_j+k} \log(s^2 \xi_i^2) + \log\left(\frac{|W^j|}{|W^{j+k}|}\right).$$

The second term on the right-hand side equals

$$2\log\left(\frac{|M^j|}{|M^{j+k}|}\right) = -2\sum_{i=m_j+1}^{m_{j+k}} \log M_{i,i}^{j+k}.$$

Therefore, we have

$$2\log B(j' \mid j) = \|g^{j+k}\|^2 - 2\sum_{i=m_j+1}^{m_{j+k}} \log\left(s\xi_i M_{i,i}^{j+k}\right). \qquad (17)$$

We summarize our findings as Algorithm 3.

Algorithm 3 Algorithm to compute the Bayes factor $B(j' \mid j)$ for $j' = j + k$.

```
• Compute $\mu^{j+k}$, $\Sigma^{j+k}$ and $W^{j+k}$.
• Obtain the Cholesky decomposition of $W^{j+k}$ so that $W^{j+k} = (M^{j+k})^T M^{j+k}$
  for an upper-triangular matrix $M^{j+k}$.
• Solve $z^{j+k}$ from $(M^{j+k})^T z^{j+k} = \mu^{j+k}$ and partition the solution into
  $z^{j+k} = [z^j, g^{j+k}]$, where $\dim(z^j) = m_j$.
• Compute $B(j' \mid j)$ from (17).
```

6. Concluding remarks

Estimation of diffusion processes has attracted a lot of attention in the past two decades. Within the Bayesian setup very few articles have considered the problem of nonparametric estimation. In this article we propose an alternative approach to the method detailed in Papaspiliopoulos et al. (2012). From the simulations it turns out that our method can provide good results.

The simulation results indicate that the posterior mean can be off the truth if the prior specification is inappropriate in the sense that

- the multiplicative scale s is fixed at a value either too high or too low;
- a truncation level is fixed and the smoothness of the prior (governed by β) is chosen inappropriately.

The first of these problems can be circumvented by specifying a prior distribution on the scaling parameter. As regards the second problem, endowing the truncation level with a prior and employing a reversible jump algorithm, it turns out that reasonable results can be obtained if we erroneously undersmooth by choosing the regularity of the prior too small. For a fixed high truncation level this is certainly not the case. In case the prior is smoother than the true drift function, both reversible jumps and a fixed high-level model can give bad results. Overall however, simulation results indicate that our method is more robust against prior misspecification.

It will be of great interest to complement our numerical results with mathematical results providing theoretical performance guarantees and giving further insight in limitations as well. Another interesting possible extension is to endow the regularity parameter β with a prior as well and let the data determine its appropriate value. This destroys the partial conjugacy however and it is a challenge to devise numerically feasible procedures for this approach.

Acknowledgment

The authors thank Yvo Pokern for providing the data and code used in the paper (Papaspiliopoulos et al., 2012).

References

Beskos, A., Papaspiliopoulos, O., Roberts, G.O., 2006. Retrospective exact simulation of diffusion sample paths with applications. Bernoulli 12, 1077–1098. http://dx.doi.org/10.3150/bj/1165269151.

Beskos, A., Papaspiliopoulos, O., Roberts, G.O., Fearnhead, P., 2006. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 333–382. http://dx.doi.org/10.1111/j.1467-9868.2006.00552.x. With discussions and a reply by the authors.

Brooks, S., Gelman, A., Jones, G.L., Meng, X.L. (Eds.), 2011. Handbook of Markov Chain Monte Carlo. In: Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, FL. http://dx.doi.org/10.1201/b10905.

Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. J. Bus. Econom. Statist. 19, 177–191. http://dx.doi.org/10.1198/073500101316970403.

Godsill, S.J., 2001. On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Comput. Graph. Statist. 10, 230–248. http://dx.doi.org/10.1198/10618600152627924.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732. http://dx.doi.org/10.1093/biomet/82.4.711.

Green, P.J., 2003. Trans-dimensional Markov chain Monte Carlo. In: Highly Structured Stochastic Systems. In: Oxford Statist. Sci. Ser., vol. 27. Oxford University Press, Oxford, pp. 179–206. With part A by Simon J. Godsill and part B by Juha Heikkinen.

Hindriks, R., 2011. Empirical dynamics of neuronal rhythms. Ph.D. Thesis, VU University Amsterdam.

Kashin, B.S., Saakyan, A.A., 1989. Orthogonal Series. In: Translations of Mathematical Monographs, vol. 75. American Mathematical Society, Providence, RI. Translated from the Russian by Ralph P. Boas, Translation edited by Ben Silver.

Liptser, R.S., Shiryaev, A.N., 1977. Statistics of Random Processes. I, General Theory. In: Applications of Mathematics (New York), vol. 5. Springer-Verlag, Berlin, Translated from the 1974 Russian Original Edition by A.B. Aries.

Panzar, L., van Zanten, J.H., 2009. Nonparametric Bayesian inference for ergodic diffusions. J. Statist. Plann. Inference 139, 4193–4199. http://dx.doi.org/10.1016/j.jspi.2009.06.003.

Papaspiliopoulos, O., Pokern, Y., Roberts, G.O., Stuart, A., 2012. Nonparametric estimation of diffusions: a differential equations approach. Biometrika 99, 511–531.

Pokern, Y., 2007. Fitting stochastic differential equations to molecular dynamics data, Ph.D. Thesis, University of Warwick.

Pokern, Y., Stuart, A.M., van Zanten, J.H., 2013. Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. Stochastic Process. Appl. 123, 603–628.

Robert, C., 2007. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. In: Springer Texts in Statistics, Springer. URL: http://books.google.com/books?id=6oQ4s8Pq9pYC.

Roberts, G.O., Stramer, O., 2001. On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. Biometrika 88, 603–621. http://dx.doi.org/10.1093/biomet/88.3.603.

Rogers, L.C.G., Williams, D., 2000. Diffusion, Markov Processes and Martingales, Vol. 1. second ed. Cambridge.

Rose, D.J., 1970. Triangulated graphs and the elimination process. J. Math. Anal. Appl. 32, 597–609.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 319–392. http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x.

Shen, X., Wasserman, L., 2001. Rates of convergence of posterior distributions. Ann. Statist. 29, 687–714. http://dx.doi.org/10.1214/aos/1009210686.

Stewart, G.W., 1998. Matrix Algorithms. Vol. I. Society for Industrial and Applied Mathematics, Philadelphia, PA. Basic decompositions.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. J. Amer. Statist. Assoc. 82, 528–550. With discussion and with a reply by the authors.

Van der Meulen, F.H., van Zanten, J.H., 2013. Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. Bernoulli 19, 44–63.

Zhao, L.H., 2000. Bayesian aspects of some nonparametric problems. Ann. Statist. 28, 532–552. http://dx.doi.org/10.1214/aos/1016218229.