Markov chain Monte Carlo for exact inference for diffusions

Sermaidis, G.,[*] Papaspiliopoulos, O.,[†] Roberts, G.O.,[‡] Beskos, A.[§] and Fearnhead, P.[*]

March 1, 2011

Abstract

We develop exact Markov chain Monte Carlo methods for discretely-sampled, directly and indirectly observed diffusions. The qualification "exact" refers to the fact that the invariant and limiting distribution of the Markov chains is the exact posterior distribution of the parameters of interest. The class of processes to which our methods directly apply are those which can be simulated using the most general to date exact simulation algorithm. The article introduces various methods to boost the performance of the basic scheme, including reparametrizations and auxiliary Poisson sampling. We contrast both theoretically and empirically how this new approach compares to irreducible high frequency imputation, which is the state-of-the-art alternative for the class of processes we consider, and we uncover intriguing connections. All methods discussed in the article are tested on typical examples.

Keywords: Exact inference; Exact simulation; Markov chain Monte Carlo; Stochastic differential equation; Transition density

1 Introduction

Diffusion processes provide a flexible framework for modelling phenomena which evolve randomly and continuously in time, and have become an extensively used tool throughout science. Application areas include among others finance (Sundaresan, 2000; Eraker *et al.*, 2003; Ait-Sahalia and Kimmel, 2007), biology (Golightly and Wilkinson, 2006), molecular kinetics (Horenko and Schütte, 2008; Kou *et al.*, 2005), pharmacokinetics/pharmacodynamics Picchini *et al.* (2010). Additionally diffusions are used to tackle mainstream statistical problems, e.g longitudinal data analysis (Taylor *et al.*, 1994), space-time models (Brown *et al.*, 2000) and functional data analysis (see for example Ramsay *et al.*, 2007, and discussion therein).

A time-homogeneous diffusion process $V \in \mathbb{R}^d$ is a Markov process defined as the solution to a stochastic differential equation (SDE):

$$dV_s = \beta(V_s; \theta_1)ds + \sigma(V_s; \theta_2)dW_s, \qquad V_0 = v, s \geq 0 \tag{1}$$

where W is an d-dimensional standard Brownian motion. The functions $\beta : \mathbb{R}^d \times \Theta_1 \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times \Theta_2 \to \mathbb{R}^{d \times d}$ are known as the *drift* and *diffusion matrix* respectively, and are allowed to depend on an unknown parameter $\theta = (\theta_1, \theta_2) \in \Theta \subset \mathbb{R}^p$. The assumption of distinct parameters for each of the functionals is by no means restrictive and is adopted here for ease of presentation of the methodology. We also introduce $\gamma = \sigma\sigma^T$ and make the usual set of assumptions on β and σ to ensure that (1) has a unique weak non-explosive solution see for example Theorem 5.2.1 of Øksendal (2003); see also Section 2. In this article we focus on the elliptic case, where the dimension of V matches that of W (this is already imposed in the equation above) and σ is invertible.

Even though the process is defined in continuous time, the available data consist of observations recorded at a set of discrete time points. We first consider the setting where the process is observed without error at a collection of time instances,

$$V := \left\{ V_{t_0}, V_{t_1}, \ldots, V_{t_n} \right\}, \qquad 0 \leq t_0 < t_1 < \ldots < t_n,$$

---
[*]Lancaster University
[†]corresponding author,Universitat Pompeu Fabra,omiros.papaspiliopoulos@upf.edu
[‡]Warwick University
[§]UCL

and denote the time increment between consecutive observations by $\Delta t_i = t_i - t_{i-1}$. Section 6 considers indirect observations $Y_{t_i} \sim q(. \mid V_{t_i}, \tau)$ where in this case interest lies in estimating parameter and latent points.

Bayesian inference for discretely observed diffusions is typically hindered by the unavailability of the transition density, hence of the (observed) likelihood. One strand of the literature deals with this issue by resorting to Monte Carlo data augmentation methods. The idea behind this approach is to augment the observed data with auxiliary data such that the joint density of observed and auxiliary data is known explicitly or at least it can be approximated satisfactorily. This joint density (with respect to an appropriate dominating measure) is often called complete likelihood. Then, an MCMC algorithm (typically a variant of the Gibbs sampler) is used to sample from the joint posterior distribution of parameters and auxiliary data. The first data augmentation (DA) approaches for discretely-observed diffusions (see Jones, 1999; Eraker, 2001; Elerian *et al.*, 2001) used as auxiliary data discretizations of the latent diffusion bridges $\{V_s, s \in (t_{i-1}, t_i)\}$. The complete likelihood is still intractable, but it can be reasonably approximated using the Euler (or a Milstein) scheme which operates on the time increments of the augmented data. The bias introduced in the estimation of the complete likelihood is eliminated by increasing the number, say M, of the imputed values. We refer to these DA schemes as high frequency imputation (HFI). There are at least three serious challenges with this approach. First, how to efficiently simulate the latent bridges conditionally on the parameters; this simulation is required in the "Imputation" step of a data augmentation algorithm. This problem has been intensively studied and it is the subject of ongoing reseach, see for example Papaspiliopoulos and Roberts (2009) for a recent account. Second, how to choose M, at least in practice. Ensuring a good approximation typically requires a large value of M. The appropriate value is not known prior to the execution of the algorithm but is typically found by repeatedly running the algorithm over increasing M until the estimated posterior distributions show no change. This adds a further computational burden. However, the most serious challenge with this approach was pointed out by Roberts and Stramer (2001) (who also provided suggestions for the other two issues). They showed that when θ_2 is unknown the mixing time of the MCMC algorithm is $\mathscr{O}(M)$. This is simply due to the quadratic variation identity for diffusions, according to which any continuous-time path contains infinite information about θ_2. On the other hand, under standard conditions diffusion paths over bounded time increments only contain finite amount of information about θ_1. Thus, DA using the latent bridge collapses in the limit $M \to \infty$, whereas for finite M it is unappealing since reduction in bias (better approximation to the true posterior) comes with unbounded increase in the variance (the variance of the Monte Carlo estimates increases due to deterioration of the Markov chain convergence). This is an extreme instance of a common problem in DA methods for partially observed stochastic processes, see for example Papaspiliopoulos *et al.* (2007).

Roberts and Stramer (2001) demonstrated the need for appropriate path-parameter transformations in order to yield a working DA algorithm which is valid even in the limit $M \to \infty$. We will refer to their approach as irreducible HFI, since the MCMC algorithms based on this imputation scheme are irreducible for all values of M, unlike those of the previous generation which are reducible in the limit. The irreducible HFI is briefly described in Section 3.1.

A third generation of Monte Carlo schemes for diffusions was initiated with the introduction of the exact algorithm (EA) for the exact simulation algorithms for non-linear diffusions; see Section 2.2 for some details. In Beskos *et al.* (2006b) it was shown how to use exact simulation ideas to unbiasedly estimate the likelihood function. This technology has been used to generate simulated likelihood approaches (Beskos *et al.*, 2009) and particle filters for partially observed diffusions when the parameters are known (see for example Fearnhead *et al.*, 2008; Jasra and Doucet, 2009).

1.1 Main contributions and structure

This article develops MCMC methods for parameter estimation for all diffusions which can be simulated by the EA. Additionally, it extends the methods to the case of indirect observations. This generation of MCMC algorithms is such that their equilibrium distribution is the exact posterior distribution of the parameters. The potential of using the EA to build an MCMC algorithm for univariate diffusions was sketched in Section 9 of Beskos *et al.* (2006b); that was for the simplest case of diffusions which can be simulated by the so-called EA1 which requires strict bounds on the drift. In this paper we develop the theory and the MCMC methods for all diffusions which can be simulated using the so-called EA3 framework of Beskos *et al.* (2008): multivariate elliptic diffusions which after a suitable transformation have constant diffusion matrix and drift which is of gradient form, see Section 2.1 for details.

The augmentation is based on a general principle which is applicable widely to stochastic processes. In a nutshell the principle is as follows. We are given a joint probability model for an observable process V and an unobservable process X, parametrised in terms of θ. We are interested in the setting where V is finite-dimensional but X infinite-dimensional (e.g X is a diffusion path, an intensity function, a spatial field, etc). Then, if we are given a rejection sampling algorithm for the exact simulation of X given V, we can use it to build an auxiliary variable representation and an exact MCMC algorithm for inference for θ. A clarification is needed in what sense X is simulated exactly; in the context of diffusions this is well understood and it is recalled in this article. The exact data augmentation we propose builds exactly on this principle which is materialised in Theorem 1. In this schematic description the HFI imputes instead a finite-dimensional approximation of X. A different application of this principle was proposed recently in Adams *et al.* (2009) in the context of density estimation.

Another main contribution of the article is the development of reparametrizations for accelerating the MCMC algorithm. We design noncentred reparametrisations and demonstrate numerically that they lead to very significant improvements. We also apply the interweaving strategy proposed recently by Meng and Yu (2011) and also evaluate numerically its effect.

A further contribution of this work is that it investigates theoretically the connection between EDA and HFI. First, we show that -surprisingly- the EDA augments more than HFI; this is surprising since the former involves no approximation whereas the latter does. This result is contained in Theorem 2. The point is that the extra augmentation in EDA creates conditional independence relationships which can be exploited to apply an algorithm which targets an infinite-dimensional state using finite computation. This result suggests that HFI for the same amount of computation is expected to mix faster than EDA. This is effectively another instance of the bias-variance tradeoff. This connection motivates a further observation which links the two approaches. The converge rate of the EDA can be improved by increasing the intensity of the Poisson process used in the EA. This intensity has no effect on the acceptance probability when simulating diffusion bridges, hence for computational reasons it is best to be kept to its minimal allowed value. However, it does have an effect on the convergence of EDA. Again, this is at first surprising, since it seems as if by imputing more the convergence rate improves. We discuss various other aspects of both algorithms and investigate numerically their performance under various scenarios. We also evaluate the effect of eliminating stochastic integrals in HFI when this is possible. Similar numerical investigations are carried out when the diffusion is indirectly observed. The existence of the exact MCMC algorithms allows us to gain a better understanding of the error involved in HFI, and this is exploited in the simulation studies. The findings which relate EDA with HFI should be useful in other contexts where this auxiliary variable principle can be applied.

A comment on the applicability of the methods proposed here is due. The framework is that of the EA given in Beskos *et al.* (2008). The necessary conditions pose little limitations for unidimensional processes, but much more serious for multivariate processes. The methods rely on a variance-stabilizing transformation which might not even exist for general multivariate SDEs with coupling in the diffusion. On the other hand, multivariate processes with no coupling in the diffusion are rather standard in the analysis of physical systems. The assumption of gradient drift is again natural in the framework of physical systems. In summary, the technology we develop here is not directly applicable to general stochastic volatility models, say, although exploiting particular structures can push considerbly these limitations (see for example Kalogeropoulos *et al.*, 2010). Additionally, advances in the exact simulation of diffusions (as for example in Étoré and Martinez, 2011; Goncalves and Roberts, 2011) would *eo ipso* lead to exact MCMC methods following the framework of this article. Note that the variance-stabilizing transformation is necessary for the Roberts and Stramer (2001) approach as well. Irreducible HFI avoiding this transformation are also currently under investigation, see for example Golightly and Wilkinson (2008).

Finally, in the article we solely focus on data augmentation methods for inference for diffusions. There is a vast literature on alternative approaches; see for example Picchini *et al.* (2010); Stramer *et al.* (2010); Forman and Sørensen (2008) and references therein, and the broad contributed discussion in Beskos *et al.* (2006b).

The article is structured as follows. Section 2.1 contains the background on assumptions, notations and recalls the EA. Section 3 presents formally the EDA and contrasts it to HFI. Section 5 carries out a careful and extensive numerical comparison of several schemes. Section 6 dicusses extensions to indirect observations. Section 7 closes with a discussion and the Appendix contains the proofs of main results.

2 Preliminaries

In this section we collect some necessary background. In terms of notation, $x^{\{i\}}$ or $x^{\{ij\}}$ denote the ith or (i,j)th element of a vector or matrix x, $\det[x]$ and x^{-1} for the determinant and the inverse of a matrix x (where appropriate), and I_d for the $d \times d$ identity matrix. The Euclidean norm is denoted by $||.||$. ∇_x and Δ_x denote the gradient and Laplacian operators respectively, that is if $x \in \mathbb{R}^d$

$$\nabla_x f(x,y) = \left(\frac{\partial f(x,y)}{\partial x_1}, \ldots, \frac{\partial f(x,y)}{\partial x_d} \right)^T \quad \Delta_x f(x,y) = \sum_{i=1}^{d} \frac{\partial^2 f(x,y)}{\partial x_i^2}.$$

2.1 Reducible diffusions of gradient type

The methods in this paper rely on the existence of a transformation η, such that $\eta(V_s; \theta_2)$ solves an SDE with constant diffusion matrix. This transformation can be obtained for unidimensional diffusions rather trivially. For multidimensional processes its existence is a subtle matter. In the elliptic case a sufficient condition is that η solves the equation

$$\nabla_v \eta(v; \theta_2)\, \sigma(v; \theta_2) = I_d \tag{2}$$

which can be further simplified when σ is invertible, to yield explicit conditions on its elements; see for example Aït-Sahalia (2008). In the rest of the article we will assume the existence of this transformation, and η^{-1} will denote the inverse transformation. If $X_s := \eta(V_s; \theta_2)$ is the transformed diffusion, then by Itô's formula its SDE is given by

$$dX_s = \alpha(X_s; \theta)ds + dW_s, \qquad X_0 = x = \eta(v; \theta_2), s \geq 0, \tag{3}$$

where $\alpha : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}^d$, with

$$\alpha^{\{r\}}(u; \theta) = A\eta^{\{r\}}\{\eta^{-1}(u; \theta_2); \theta_2\}, \tag{4}$$

where A is the generator of (1)

$$Af(v) = \sum_{i=1}^{d} \beta^{\{i\}}(v; \theta_1) \frac{\partial f(v)}{\partial v_i} + \frac{1}{2} \sum_{i,j=1}^{d} \gamma^{\{ij\}}(v; \theta_2) \frac{\partial^2 f(v)}{\partial v_i \partial v_j}.$$

The transition density of the original process V can be derived in terms of that of the transformed process X. If $\tilde{p}_t(x,z; \theta)$, $x,z \in \mathbb{R}^d$ denotes the transition density of the process X, then

$$p_t(v, w; \theta) = \left| \det \left[\sigma^{-1}(w; \theta_2) \right] \right| \tilde{p}_t \{ \eta(v; \theta_2), \eta(w; \theta_2); \theta \}, \tag{5}$$

where $\det[.]$ denotes the determinant.

The methodology also requires certain conditions on the drift α of the transformed process. Again, these are trivially satisfied in unidimensional processes but they are more restrictive for multidimensional processes. The following set of assumptions should hold for any $\theta \in \Theta$.

(a) $\alpha^{\{r\}}(.; \theta)$ is continuously differentiable for $r = 1, \ldots, d$.

(b) There exists $H : \mathbb{R}^d \rightarrow \mathbb{R}$ such that $\nabla_x H(x; \theta) = \alpha(x; \theta)$.

(c) There exists $l(\theta) > -\infty$, such that $l(\theta) \leq \inf_{u \in \mathbb{R}^d} \frac{1}{2} \left\{ ||\alpha(u; \theta)||^2 + \Delta_x H(u; \theta) \right\}$.

The second condition identifies X as a diffusion of gradient-type, where H is called the potential function. When the diffusion is ergodic, its invariant log-density can be expressed directly in terms of $-H$. The first is a very weak condition and the third is rather mild too. Finally, we require that α is such that the probability law generated by the solution of (3) is absolutely continuous with respect to the Wiener measure. A particularly useful and weak set of conditions are given in Rydberg (1997); in the case of (3) if α is locally bounded the conditions simply require that the SDE be not explosive.

2.2 The exact algorithm

The *exact algorithm* (EA) is a rejection sampling algorithm on the space of diffusion paths, which uses Brownian path proposals and delivers the diffusion path revealed at a finite collection of random points. The path can be filled in later at any desired time point with no further reference to the target process. The main attractions of the algorithm are that the draws are from the exact finite-dimensional distribution, and that it can very easily simulate from the diffusion process conditioned on its endpoints. Here we focus on exact *diffusion bridge* simulation, i.e. obtain samples from (1) conditionally on the origin $V_0 = v$ and terminal point $V_t = w$. It turns out that this conditional simulation is really the key to data augmentation methods for parameter estimation.

The target process (1) is transformed into one of unit diffusion matrix as described in Section 2.1. The problem is therefore reduced to the simulation of (3) conditionally on the origin $x = x(\theta_2) := \eta(v; \theta_2)$ and terminal point $y = y(\theta_2) := \eta(w; \theta_2)$. An X-bridge yields a V-bridge by applying the inverse transformation. Let $\mathbb{Q}_\theta^{(t,x,y)}$ denote the law of the X-bridge, $\mathbb{W}_\theta^{(t,x,y)}$ the law of a Brownian bridge conditioned on the same endpoints, and starting at x and terminating at y at time t, and $\mathcal{N}_t^d(u)$ denote the density of a Gaussian random variable with vector mean 0 and variance tI_d evaluated at $u \in \mathbb{R}^d$. The following lemma, which is a restatement of Lemma 1 in Beskos *et al.* (2006b), derives the density of the target law with respect to the Brownian bridge law; see also Papaspiliopoulos and Roberts (2009) for a collection of results on likelihood ratios for diffusion bridges, a historical overview and references.

Lemma 1. *The law $\mathbb{Q}_\theta^{(t,x,y)}$ is absolutely continuous with respect to $\mathbb{W}_\theta^{(t,x,y)}$ with density*

$$\frac{d\mathbb{Q}_\theta^{(t,x,y)}}{d\mathbb{W}_\theta^{(t,x,y)}}(X) = \frac{\mathcal{N}_t^d(y-x)}{\breve{p}_t(x,y;\theta)} \exp\left\{ H(y;\theta) - H(x;\theta) - \frac{1}{2}\int_0^t \left\{ \|\alpha(X_s;\theta)\|^2 + \Delta_x H(X_s;\theta) \right\} ds \right\} \quad (6)$$

$$\propto \exp\left\{ -\int_0^t \phi(X_s;\theta)\, ds \right\} \leq 1, \quad (7)$$

where $\phi : \mathbb{R}^d \times \Theta \to \mathbb{R}_+$ is defined by

$$\phi(u;\theta) = \frac{1}{2}\left\{ \|\alpha(u;\theta)\|^2 + \Delta_x H(u;\theta) \right\} - l(\theta);$$

Note that application of importance sampling techniques (with rejection sampling as a special case) does not require the knowledge of the normalizing constant in (6) but instead works with the unnormalized ratio in (7). The EA is based on recognising (7) as the probability of a specific event from an inhomogeneous Poisson process of intensity $\phi(X_s;\theta)$ on $[0,t]$. Simulation of events from such processes can be achieved by constructing an upper bound for the variable intensity and using the *thinning* property of the Poisson process. Assume that there exists a finite-dimensional random variable $L := L(X)$ and a positive function r such that

$$r(L;\theta) \geq \sup_{s\in[0,t]} \phi(X_s;\theta).$$

Let $\Phi = \{\Psi, \Upsilon\}$ be a homogeneous Poisson process of intensity $r(L;\theta)$ on $[0,t] \times [0,1]$, with uniformly distributed points $\Psi = \{\psi_1, \ldots, \psi_\kappa\}$ on $[0,t]$ and uniformly distributed marks $\Upsilon = \{u_1, \ldots, u_\kappa\}$ on $[0,1]$, where $\kappa \sim \text{Po}[r(L;\theta)t]$. If N is the number of points of Φ below the graph $s \to \phi(X_s;\theta)/r(L;\theta)$, then

$$P(N = 0 \mid X) = \exp\left\{ -\int_0^t \phi(X_s;\theta)\, ds \right\}.$$

This implies a rejection sampler where a proposed path $X \sim \mathbb{W}_\theta^{(t,x,y)}$ is accepted as a path from $\mathbb{Q}_\theta^{(t,x,y)}$ according to the indicator

$$I(L, X, \Phi, v, w, \theta) := \prod_{j=1}^\kappa \mathbb{I}\left[\phi\left(X_{\psi_j};\theta\right)/r(L;\theta) < u_j \right]. \quad (8)$$

The output of the EA consists of the collection of random variables $\{L(X), \Phi, S(X)\}$, where

$$S(X) := \left\{ (0, X_0), (\psi_1, X_{\psi_1}), \ldots, (\psi_\kappa, X_{\psi_\kappa}), (t, X_t) \right\} \quad (9)$$

5

is a skeleton of the accepted path. This output can be thought of as a random sufficient statistic of the continuous path. Indeed, the accept-reject decision is based entirely on finite information from the path but, in principle, the algorithm accepts an infinite dimensional object. The algorithm is presented in Algorithm 1.

Algorithm 1 Exact Algorithm for diffusion bridges

1: Generate a homogeneous Poisson process $\Phi = \{\Psi, \Upsilon\}$ on $[0, t] \times [0, 1]$ as follows

- simulate $L = L(X)$, where $X \sim \mathbb{W}_\theta^{(t,x,y)}$ and $\kappa \sim \mathrm{Po}\left[r(L; \theta)t\right]$,

- simulate $\Psi = \{\psi_1, \ldots, \psi_\kappa\}$, where $\psi_j \sim \mathrm{Un}[0, t]$ and $\Upsilon = \{u_1, \ldots, u_\kappa\}$, where $u_j \sim \mathrm{Un}[0, 1]$.

2: Simulate X conditionally on L at time instances $\psi_1, \ldots, \psi_\kappa$ and set $S(X)$ as in (9).

3: Evaluate the acceptance indicator $I := \prod_{j=1}^{\kappa} \mathbb{I}\left[\phi\left(X_{\psi_j}; \theta\right) / r(L; \theta) < u_j\right]$.

4: If $I = 1$ then return $\{L, \Phi, S(X)\}$, otherwise go to 1.

The technical difficulty that underlines the implementation of the Exact Algorithm is the simulation of $L(X)$ that determines the required Poisson rate, and primarily the conditional simulation of a Brownian bridge given $L(X)$ (step 2 of the algorithm) for the evaluation of the accept-reject indicator function. This has resulted in the construction of three EAs that share the rejection sampling principle, but have a different range of applicability.

2.2.1 The family of EAs

The EA1 (Beskos *et al.*, 2006a) is the simplest of Exact Algorithms and its framework is restricted by

Condition 1. $\phi(.; \theta)$ *is bounded above.*

This condition ensures the availability of a path-independent upper bound for the intensity, i.e. $r(L; \theta) \equiv r(\theta)$, implying that simulation of $L(X)$ is avoided. As a consequence, step 2 of Algorithm 1 merely requires simulation of a Brownian bridge at time instances $\psi_1, \ldots, \psi_\kappa$. The EA2 (Beskos *et al.*, 2006a) is applicable only when $d = 1$ and relaxes Condition 1 to a more mild one:

Condition 2. *Either* $\limsup_{u \to \infty} \phi(u; \theta) < \infty$ *or* $\limsup_{u \to -\infty} \phi(u; \theta) < \infty$.

For simplicity consider only the first case. The algorithm is based on decomposing a Brownian bridge path at its minimum, i.e. the proposed path is constructed by first simulating its minimum, say m, and subsequently the remainder of the path conditioned on the minimum. In this setting the required random variable $L(X) = \{m, \tau\}$, where τ is the time instance the minimum is attained. Then, Condition 2 ensures that for $u > m$ we can obtain the required upper bound by selecting

$$r(L; \theta) = \sup_u \{\phi(u; \theta); u \geq m\}.$$

Simulating a Brownian bridge conditionally on its minimum is based on a path transformation of two independent Bessel bridges, see Beskos *et al.* (2006b) for more details.

EA3 (Beskos *et al.*, 2008) poses no upper boundary conditions. The construction of the Poisson rate $r(L; \theta)$ is achieved by creating an appropriate partition on the path space using a series of lower and upper bounds for each path coordinate. In particular, for a user-specified constant $\delta > \sqrt{t/3}$ and $\bar{x}^{\{r\}} := x^{\{r\}} \wedge y^{\{r\}}$, $\bar{y}^{\{r\}} := x^{\{r\}} \vee y^{\{r\}}$, the partition consists of the sets $D_r\left(l^{\{r\}}\right)$, $l^{\{r\}} \in \mathbb{N}^*$, $r = 1, \ldots, d$, defined by

$$A_r\left(l^{\{r\}}\right) = \left\{\sup_{0 \leq s \leq t} X_s^{\{r\}} \in \left[\bar{y}^{\{r\}} + (l^{\{r\}} - 1)\delta, \bar{y}^{\{r\}} + l^{\{r\}}\delta\right)\right\} \cap \left\{\inf_{0 \leq s \leq t} X_s^{\{r\}} > \bar{x}^{\{r\}} - l^{\{r\}}\delta\right\},$$

$$B_r\left(l^{\{r\}}\right) = \left\{\inf_{0 \leq s \leq t} X_s^{\{r\}} \in \left(\bar{x}^{\{r\}} - l^{\{r\}}\delta, \bar{x}^{\{r\}} - (l^{\{r\}} - 1)\delta\right]\right\} \cap \left\{\sup_{0 \leq s \leq t} X_s^{\{r\}} < \bar{y}^{\{r\}} + l^{\{r\}}\delta\right\},$$

$$D_r\left(l^{\{r\}}\right) = A_r\left(l^{\{r\}}\right) \cup B_r\left(l^{\{r\}}\right).$$
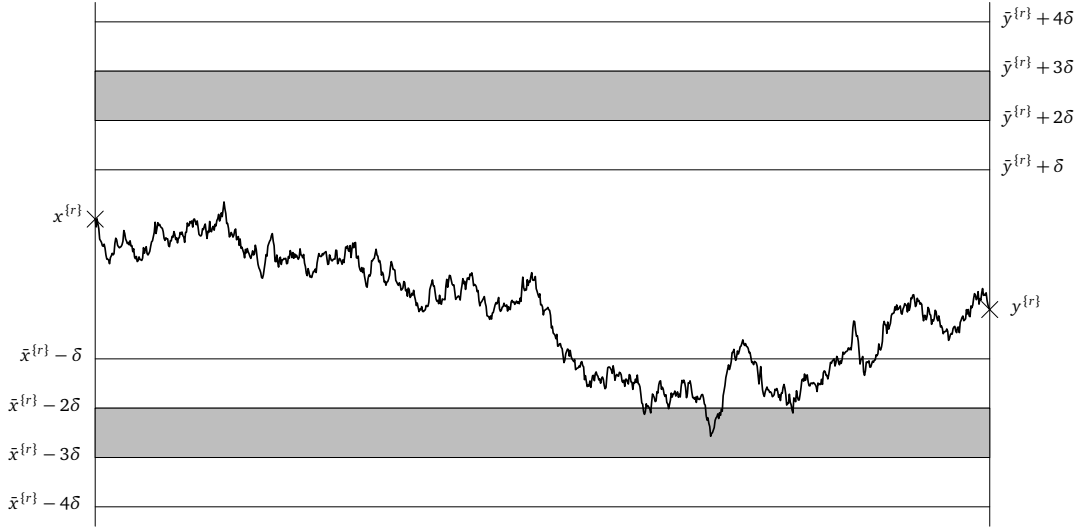
Figure 1: The rth coordinate of a Brownian bridge from $X_0 = x$ to $X_t = y$, with $x^{\{r\}} > y^{\{r\}}$. In this example, the event $B_r(3)$ has occurred. Therefore, $L^{\{r\}} = 3$.

In this case, the required $L(X)$ is set to be the d-dimensional discrete random variable, $L(X) = (L^{\{1\}}, \ldots, L^{\{d\}})$ where $L^{\{r\}} = l^{\{r\}}$ if $X \in D_r(l^{\{r\}})$. Figure 1 illustrates the construction for an arbitrary coordinate. Hence, $\{L^{\{r\}} \leq l^{\{r\}}, r = 1, \ldots, d\} \equiv \{\bar{x}^{\{r\}} - l^{\{r\}}\delta < X_s^{\{r\}} < \bar{y}^{\{r\}} + l^{\{r\}}\delta, 0 \leq s \leq t, r = 1, \ldots, d\}$. This random variable is referred to as the *Brownian bridge layer* and a Brownian bridge path conditioned on this layer as the *layered Brownian bridge*. Conditioned on the value $L = l$, and using the continuity of $\phi(.;\theta)$, the Poisson rate can be found as

$$r(L;\theta) = \sup\left\{\phi(u;\theta); u^{\{r\}} \in (\bar{x}^{\{r\}} - l^{\{r\}}\delta, \bar{y}^{\{r\}} + l^{\{r\}}\delta), r = 1, \ldots, d\right\}.$$

The exact mathematical and implementation details of sampling a layered Brownian bridge can be found in the original paper.

2.2.2 Computational considerations

The EA3 has the widest framework of applicability, which encompasses that of EA1 and EA2. Nevertheless, simulating a layered Brownian bridge is a non-trivial task and is achieved by means of a rejection sampling mechanism, thus adding to the EA3 an extra level of computational complexity. In particular, an extensive empirical study by Peluchetti and Roberts (2008) suggests as a rule of thumb that EA3 is approximately 10 times slower than EA1.

Moreover, as any rejection sampler, the computational performance of EA depends on its acceptance probability. For any two fixed points x and y, expression (7) implies that the probability of accepting a proposed path is

$$a(x, y, t, \theta) := \mathbb{E}_{\mathbb{W}_\theta^{(t,x,y)}}\left[\exp\left\{-\int_0^t \phi(X_s; \theta)\,\mathrm{d}s\right\}\right], \tag{10}$$

thus suggesting that the acceptance probability decays exponentially to 0 as d or t increase.

Finally, we note that, for fixed θ, the EA can be implemented with any Poisson sampling rate R, where $R \geq r(L;\theta)$. Choosing a large value of R results in an additional computational cost since the number of Poisson points to be simulated (and thus the number of time instances where the path is revealed) is now larger; however, it does not affect the acceptance probability of the algorithm since the accept-reject decision depends only on the Poisson points for which $u_j < r(L;\theta)/R$. This observation will turn out to be quite relevant when we consider MCMC methods based on the EA.

7

3 Exact data augmentation (EDA) for discretely observed diffusions

We are interested in exploring the posterior density

$$\pi(\theta \mid V) \propto \pi(\theta) \prod_{i=1}^{n} p_{\Delta t_i}(V_{t_{i-1}}, V_{t_i}; \theta) \tag{11}$$

which is typically intractable due to the transition densities. In this section we describe irreducible data augmentation approaches for parameter estimation. The auxiliary variables we will introduce lead to MCMC algorithms which involve no approximation to the statistical model of interest, and the only source of error is that due to Monte Carlo. The auxiliary variables are intimately related to the output of the EA for diffusion bridges.

At first, this approach appears totally different from the HFI paradigm. However, there are close but subtle links and the presentation in this section has been structured to naturally bring those out. Effectively, the limiting ($M = \infty$) irreducible data augmentation of Roberts and Stramer (2001), which is recalled below, leads to two possibilities. One is its approximation by a HFI with some finite M, which leads to bias. The other is to consider an EA for the simulation of the auxiliary process identified by Roberts and Stramer (2001) and identify finite-dimensional auxiliary variables which then lead to an exact MCMC algorithm. Section 3.2 identifies those variables and Section 3.3 gives a theorem which establishes the connection between the limiting irreducible HFI and EDA. In fact, Section 4.3 shows that there is a spectrum of algorithms with the EDA on the one end and the limiting irreducible HFI on the other, with intermediate algorithms produced by varying the intensity of the Poisson process in the EA. This leads to a computational time/variance tradeoff which can be exploited, but all the algorithms in the spectrum are unbiased.

It has already been emphasized in Section 1.1 that the principles behind the methods we develop here have impact on several other problems which involve computations with partially observed stochastic processes.

3.1 Irreducible HFI

We first outline the Roberts and Stramer (2001) approch for HFI. The method is developed first for the limiting case $M = \infty$, hence it yields an idealized algorithm for parameter estimation which involves imputing an infinite dimensional auxiliary variable. This auxiliary process is obtained after two path-parameter transformations of the original latent bridge. We will call this augmentation scheme limiting irreducible HFI; limiting since it refers to the limit $M = \infty$ and irreducible since it generates a Markov chain (on an infinite-dimensional space) which is mixing, as opposed to alternative schemes which in that limit are reducible. We then review the approach that has been taken so far in the literature, which is to subsequently approximate the algorithm with some finite M.

Consider, for the moment, only two observations from the diffusion process, $V_0 = v$ and $V_t = w$. We first transform the diffusion process $V_s \rightarrow X_s = \eta(V_s; \theta_2)$, where $\eta(.; \theta_2)$ is described in Section 2.1. After this variance-stabilization all the parameters relate with the drift of the transformed process, see (3). X-paths over bounded time increments only contain finite information for θ, which motivates the adoption of this transformation. The transformed path starts at $x(\theta_2)$ and terminates at $y(\theta_2)$, which are both deterministic functions of θ_2 and the observations. This suggests that a DA scheme based on X will not work either when θ_2 is unknown. A realization of X determines θ_2 through its endpoints. An alternative way to see the problem is to note that the collection of dominating measures $\left\{ \mathbb{W}_\theta^{(t,x,y)}, \theta \in \Theta \right\}$ are mutually singular, and therefore a Gibbs or an EM algorithm based on this augmentation would not converge, but would instead be trapped in the support of one of these measures. This necessitates a further reparametrisation from $X_s \rightarrow \tilde{X}_s$, where

$$\tilde{X}_s := X_s - \left(1 - \frac{s}{t}\right) x(\theta_2) - \frac{s}{t} y(\theta_2), \quad s \in [0, t], \tag{12}$$

which forces the path to start and finish at 0 and essentially transforms the distribution $\mathbb{Q}_\theta^{(t,x,y)}$ in such a way that the dominating measure, now given by $\mathbb{W}^{(t,0,0)}$, is independent of θ.

The irreducible HFI is based on imputing $\tilde{X}$. Thus, the complete data likelihood and posterior which will subsequently be targetted by MCMC schemes are as follows. Let $\pi(\theta)$ denote the prior density of θ with respect to Lebesgue measure and $\left\{ \tilde{X}_{i,s}, s \in [0, \Delta t_i] \right\}$ be the imputed paths. For economy of notation let X denote the path obtained by the inverse transformation of (12), $\tilde{X} \rightarrow X$, where in this setting X is a function

of $\tilde{X}$ the endpoints and the parameters, and similarly let X_i be the transformation of $\tilde{X}_i$. Then the joint posterior density of θ and imputed paths, $\pi_{HF,\infty}(\theta, \{\tilde{X}_i, 1 \leq i \leq n\} \mid V)$, can be derived with respect to the product law $\text{Leb}^p \otimes_{i=1}^n \mathbb{W}^{(\Delta t_i,0,0)}$, and is proportional to

$$\pi(\theta) \exp\left[H\{x_n(\theta_2); \theta\} - H\{x_0(\theta_2); \theta\} - l(\theta)(t_n - t_0) \right]$$

$$\times \prod_{i=1}^n \left| \det\left[\sigma^{-1}(V_{t_i}; \theta_2) \right] \right| \mathcal{N}_{\Delta t_i}^d \{x_i(\theta_2) - x_{i-1}(\theta_2)\} \exp\left\{ -\sum_{i=1}^n \int_0^{\Delta t_i} \phi\left(X_{i,s}; \theta\right) ds \right\}. \tag{13}$$

For a proof, the reader is referred to Section 3 of Roberts and Stramer (2001). Note however, that the expression we obtain in (13) is numerically more stable than the one in Roberts and Stramer (2001). We have exploited the assumptions on the drift (the fact that it is a gradient type) to perform integration by parts and transform stochastic integrals into time integrals. This improves the convergence of the algorithm ADD LATER.

The joint posterior density written above will typically not be computable since the augmented paths cannot be represented by a finite number of variables, hence the integrals cannot be computed. Instead, an approximation is done at this stage. The auxiliary paths are approximated by vectors of size $M + 2$, $\left\{ \tilde{X}_{i,j\Delta t_i/(M+1)}, j = 0, \ldots, M+1 \right\}$, and the integrals are approximated numerically, typically by Riemann sums, to yield $\pi_{HF,M}(\theta, \{\tilde{X}_i, 1 \leq i \leq n\} \mid V)$ (where by an abuse of notation we let $\tilde{X}_i$ denote the path and its discretization). This introduces a bias in the inference for θ. The approximated posterior is targeted by an MCMC algorithm which updates in turns the parameters and the discretized auxiliary processes according to their conditional densities. A simple way to simulate the auxiliary processes is by proposing Brownian bridge skeletons and accepting them according to the target density. Crucially for the efficiency of the algorithm, the auxiliary processes (and their approximations) $\tilde{X}_i, i = 1, \ldots, n$ are independent of each other conditionally on θ. The following algorithm is a typical general implementation of the irreducible HFI, where updates of θ are obtained from a Metropolis-Hastings step with proposal kernel $q(\theta, .)$.

Algorithm 2 Irreducible HFI

1: Initialiase by choosing θ^0, and $\tilde{X}_i^0 = \left\{ \tilde{X}_{i,j\Delta t_i/M}^0, j = 0, \ldots, M \right\}$, $1 \leq i \leq n$. Set $t = 0$.

2: Iterate the following steps

- for $1 \leq i \leq n$
  - simulate $Y_i = \left\{ Y_{i,j\Delta t_i/M}, j = 0, \ldots, M \right\}$, where $Y_i \sim \mathbb{W}^{(\Delta t_i,0,0)}$,
  - sample $U \sim \text{Un}(0,1)$,
  - if

$$U < \frac{\pi_{HF,M}(Y_i \mid \theta^t, V)}{\pi_{HF,M}(X_i^t \mid \theta^t, V)} = \exp\left\{ -\frac{\Delta t_i}{M+1} \sum_{j=0}^M \left(\phi\left(Y_{i,j\Delta t_i/(M+1)}; \theta^t\right) - \phi\left(\tilde{X}_{i,j\Delta t_i/(M+1)}^t; \theta^t\right) \right) \right\}$$

    then set $\tilde{X}_i^{t+1} = Y_i$, else set $\tilde{X}_i^{t+1} = \tilde{X}_i^t$.

- sample $\theta^* \sim q(\theta^t, .)$ and $U \sim \text{Un}(0,1)$,
- if

$$U < \frac{\pi_{HF,M}(\theta^*, \{\tilde{X}_i^{t+1}, 1 \leq i \leq n\} \mid) \; q(\theta^*, \theta^t)}{\pi_{HF,M}(\theta, \{\tilde{X}_i^{t+1}, 1 \leq i \leq n\}) \; q(\theta^t, \theta^*)}$$

  then set $\theta^{t+1} = \theta^*$, else set $\theta^{t+1} = \theta^t$,

- set $t = t + 1$.

3.2 Exact MCMC

The auxiliary variable methods first identify auxiliary random variables such that the joint density of observed data, parameters and the auxiliary variables is tractable, and then target the conditional density given the observed data by MCMC methods (typically variants of the Gibbs sampler). The main contribution of this section is to demonstrate that an exact rejection sampling algorithm for simulating diffusion bridges implies
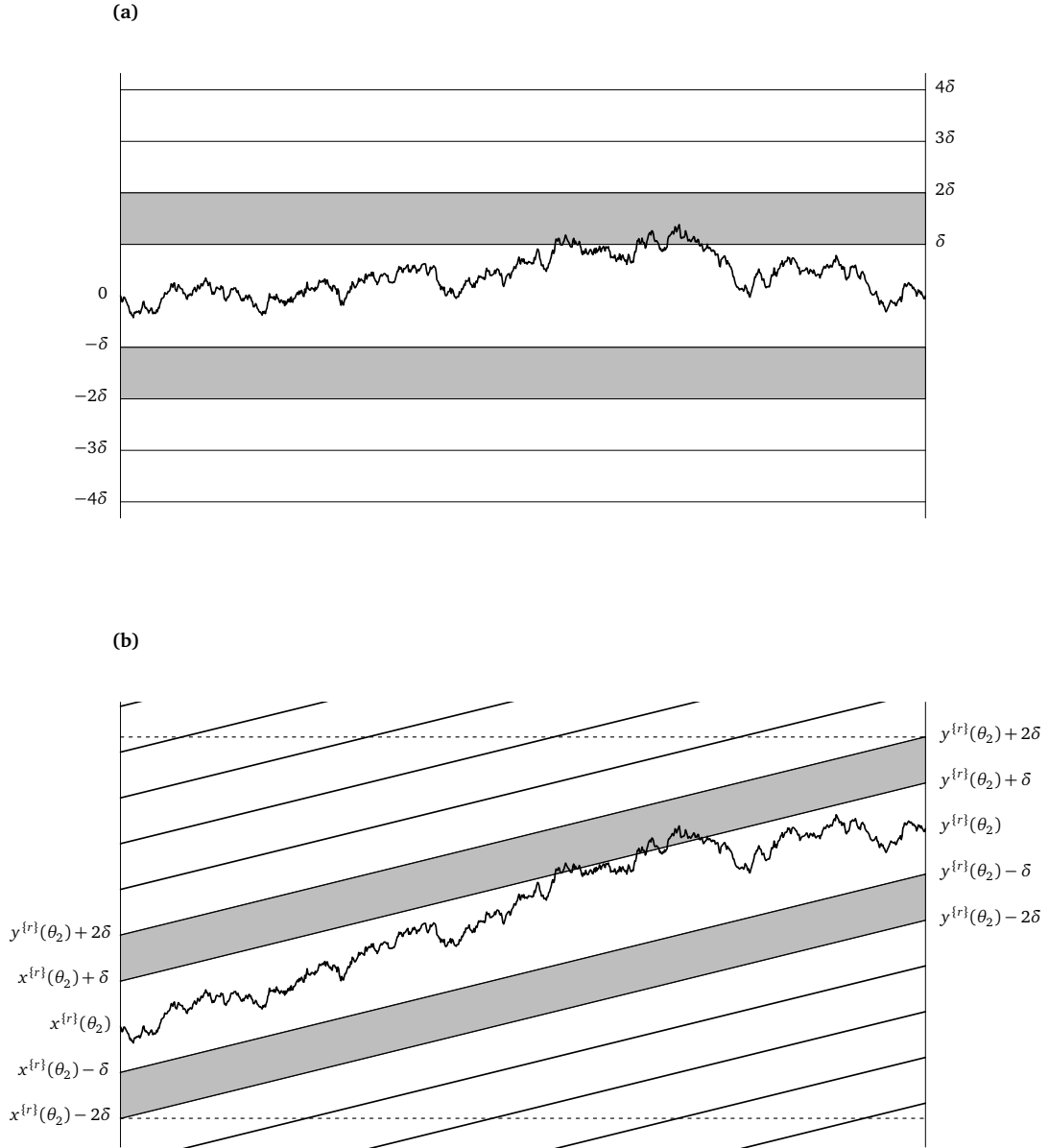
(a)

(b)

Figure 2: The rth coordinate of a layered Brownian bridge path $\tilde{X}$ starting and terminating at 0 with $\tilde{L}^{\{r\}} = 2$ (top). The transformed process $\tilde{X}_s^{\{r\}} + (1 - s/t) x^{\{r\}}(\theta_2) + s y^{\{r\}}(\theta_2)/t$ starts at $x^{\{r\}}(\theta_2)$ and terminates at $y^{\{r\}}(\theta_2)$, with $x^{\{r\}}(\theta_2) < y^{\{r\}}(\theta_2)$ (bottom). The dashed lines provide a lower and upper bound for the coordinate of the transformed path.

appropriate auxiliary variables which can be used to design MCMC algorithms for exact inference in diffusions. In the following sections we elaborate on this construction and propose various improvements on the basic scheme which is developed here. We describe the data augmentation and the MCMC algorithm which corresponds to the EA3 case, and later comment on the simplifications that arise when a more basic EA is applicable to the model of interest.

We will first identify the appropriate auxiliary variables for each pair of arbitrary consecutive observations, $V_0 = v$ and $V_t = w$. Then, by the Markov property we will be able to find the auxiliary variables for an arbitrary number of observations. To do this, we need the two main tools we have developed so far. First, the two path-parameter transformations proposed by Roberts and Stramer (2001) for irreducible data augmentation; recall that these are $V \to X$ and $X \to \tilde{X}$. Second, the EA for simulating an X-bridge according to $\mathbb{Q}_\theta^{(t,x,y)}$, with x and y as defined in Section 2.2.1. Let $\tilde{\mathbb{Q}}_\theta^{(t)}$ denote the measure induced by the linearly transformed bridge $\tilde{X}$, when X is drawn from $\mathbb{Q}_\theta^{(t,x,y)}$. In passing, recall that when X is drawn from $\mathbb{W}_\theta^{(t,x,y)}$, $\tilde{X}$ is distributed according to $\mathbb{W}^{(t,0,0)}$.

We first sketch an EA for simulating from $\tilde{\mathbb{Q}}_\theta^{(t)}$ using proposals from $\mathbb{W}^{(t,0,0)}$. This is a minor modification of the EA for X, since a proposed path $\tilde{X} \sim \mathbb{W}^{(t,0,0)}$ is accepted as a path from $\tilde{\mathbb{Q}}_\theta^{(t)}$ if and only if $\tilde{X}_s + \left(1 - \frac{s}{t}\right) x(\theta_2) + \frac{s}{t} y(\theta_2), s \in [0,t]$, is accepted as a path from $\mathbb{Q}_\theta^{(t,x,y)}$. Let $\tilde{L}$ denote the layer of a Brownian bridge that starts and terminates at 0, and denote by $\mathbb{M}^{(t)}$ the joint law of $(\tilde{X}, \tilde{L})$; hence if $(\tilde{X}, \tilde{L}) \sim \mathbb{M}^{(t)}$, marginally $\tilde{X} \sim \mathbb{W}^{(t,0,0)}$. The pair $\tilde{L}, \tilde{X}$ imply realizations for the corresponding variables in the X-space. These are easy to obtain, since conditionally on a given $\tilde{L}$, $\left\{\tilde{X}_s^{\{r\}}, s \in [0,t]\right\}$ moves within $(-\tilde{L}^{\{r\}}\delta, \tilde{L}^{\{r\}}\delta)$, and thus

$$\bar{x}^{\{r\}}(\theta_2) - \tilde{L}^{\{r\}}\delta < \tilde{X}_s^{\{r\}} + \left(1 - \frac{s}{t}\right) x^{\{r\}}(\theta_2) + \frac{s}{t} y^{\{r\}}(\theta_2) < \bar{y}^{\{r\}}(\theta_2) + \tilde{L}^{\{r\}}\delta,$$

where $\bar{x}^{\{r\}}(\theta_2) := x^{\{r\}}(\theta_2) \wedge y^{\{r\}}(\theta_2)$ and $\bar{y}^{\{r\}}(\theta_2) := x^{\{r\}}(\theta_2) \vee y^{\{r\}}(\theta_2)$. The above construction, along with the derivation of the lower and upper bounds, is illustrated in Figure 2. An EA which samples from $\tilde{\mathbb{Q}}_\theta^{(t)}$ follows easily. If $\Phi = \{\Psi, \Upsilon\}$ is a homogeneous Poisson process of intensity $r(\tilde{L}; \theta)$ on $[0,t] \times [0,1]$, where Ψ is the projection of the points on $[0,t]$ and Υ the projection on $[0,1]$, and

$$r(\tilde{L}; \theta) = \sup \left\{ \phi(u; \theta) ; u^{\{r\}} \in (\bar{x}^{\{r\}}(\theta_2) - \tilde{L}^{\{r\}}\delta, \bar{y}^{\{r\}}(\theta_2) + \tilde{L}^{\{r\}}\delta), r = 1, \ldots, d \right\}, \quad (14)$$

then the algorithm accepts the proposed $(\tilde{L}, \tilde{X}, \Phi)$ according to

$$I(\tilde{L}, \tilde{X}, \Phi, v, w, \theta) := \prod_{j=1}^{\kappa} \mathbb{I}\left[\frac{1}{r(\tilde{L}; \theta)} \phi\left\{ \tilde{X}_{\psi_j} + \left(1 - \frac{\psi_j}{t}\right) x(\theta_2) + \frac{\psi_j}{t} y(\theta_2); \theta \right\} < u_j \right], \quad (15)$$

which is the familiar indicator function (8), reformulated in terms of $(\tilde{L}, \tilde{X}, \Phi)$.

We now define the augmentation scheme for a pair of observations V_0, V_t by the random variables $(\tilde{L}, \tilde{X}, \Psi)$ that are returned in the EA3 output. Note that we explicitly do not include the Υ process, which is not necessary for obtaining a tractable density and its inclusion would lead to overconditioning. The density of the auxiliary variables is derived in the following lemma, which is proved in the Appendix.

Lemma 2. *Let $(\tilde{L}, \tilde{X}) \sim \mathbb{M}^{(t)}$ and Ψ be a homogeneous Poisson process of intensity $r(\tilde{L}; \theta)$ on $[0,t]$. If I is the acceptance indicator in (15), then the conditional density of $(\tilde{L}, \tilde{X}, \Psi)$ given $I = 1$, $\pi(\tilde{L}, \tilde{X}, \Psi \mid u, v, \theta)$, is*

$$\frac{r(\tilde{L}; \theta)^\kappa}{\lambda^\kappa a(x, y, t, \theta)} \exp\left\{ t\left[\lambda - r(\tilde{L}; \theta)\right] \right\} \prod_{j=1}^{\kappa} \left[1 - \phi\left\{ \tilde{X}_{\psi_j} + \left(1 - \frac{\psi_j}{t}\right) x(\theta_2) + \frac{\psi_j}{t} y(\theta_2); \theta \right\} / r(\tilde{L}; \theta) \right], \quad (16)$$

with respect to the product measure $\mathbb{M}^{(t)} \times \mathbb{P}_\lambda^{(t)}$, where $\mathbb{P}_\lambda^{(t)}$ is the measure of a homogeneous Poisson process on $[0,t]$ with intensity $\lambda \geq 1$, and $a(x, y, t, \theta)$ is the acceptance probability of the EA.

The Lemma introduces a degree of freedom, which will make use later, which is the intensity of the dominating Poisson process. The value of λ does not affect the inferential procedure (since it is a constant) and will typically be chosen equal to 1. However, for mathematical reasons we shall see below, we assume a generic value $\lambda \geq 1$.

Extending the augmentation scheme to account for all observations is straightforward. Specifically, we define $x_i = x_i(\theta_2) := \eta(V_{t_i}; \theta_2), i = 0, 1, \ldots, n$, and let $\tilde{L}_i, \tilde{X}_i = \{\tilde{X}_{i,s}, s \in [0, \Delta t_i]\}$ and $\Phi_i = \{\Psi_i, \Upsilon_i\}$ denote the accepted elements of EA applied to the interval $[t_{i-1}, t_i]$, for $1 \leq i \leq n$. Due to the Markov property of the diffusion process, the bridges conditionally on the observations are independent, implying that the joint density of all the imputed random variables conditionally on the observations Y and θ is

$$\pi(\{\tilde{L}_i, \tilde{X}_i, \Psi_i, 1 \leq i \leq n\} \mid V, \theta) = \prod_{i=1}^{n} \pi(\tilde{L}_i, \tilde{X}_i, \Psi_i \mid V_{t_{i-1}}, V_{t_i}, \theta).$$

We complete the development of EDA with the following Theorem which specifies the joint density of data, auxiliary variables and parameters. This has a simple computable form and it admits the target posterior $\pi(\theta \mid V)$ as a marginal with respect to the auxiliary variables and conditional with respect to the data. The key observation for the development of exact MCMC methods, is that the joint density is only a function of the finite-dimensional $\{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}$, which are delivered by the EA, and it is constant with respect to the rest of the auxiliary variables. Additionally, the intractable normalizing constant has been cancelled out. The proof of the Theorem is given in the Appendix.

Theorem 1. *The joint density of observed data V, the p-dimensional parameters θ and auxiliary variables $\{\tilde{L}_i, \tilde{X}_i, \Psi_i, 1 \leq i \leq n\}$, is given below with respect to the θ-independent dominating measure $\mathrm{Leb}^{n+p} \otimes_{i=1}^{n} \left(\mathbb{M}^{(\Delta t_i)} \times \mathbb{P}_\lambda^{(\Delta t_i)} \right)$:*

$$\pi(V, \theta, \{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}) = \pi(\theta) \prod_{i=1}^{n} p_{\Delta t_i}(V_{t_{i-1}}, V_{t_i}; \theta) \prod_{i=1}^{n} \pi(\tilde{L}_i, \tilde{X}_i, \Psi_i \mid V_{t_{i-1}}, V_{t_i}, \theta) =$$

$$\frac{\pi(\theta)}{\lambda^{\sum_{i=1}^{n} \kappa_i}} \exp\left(H\{x_n(\theta_2); \theta\} - H\{x_0(\theta_2); \theta\} - [l(\theta) - \lambda](t_n - t_0) - \sum_{i=1}^{n} r(\tilde{L}_i; \theta) \Delta t_i \right)$$

$$\times \prod_{i=1}^{n} \left| \det\left[\sigma^{-1}(V_{t_i}; \theta_2) \right] \right| \mathcal{N}_{\Delta t_i}^{d} \{x_i(\theta_2) - x_{i-1}(\theta_2)\} r(\tilde{L}_i; \theta)^{\kappa_i}$$

$$\times \prod_{i=1}^{n} \prod_{j=1}^{\kappa_i} \left[1 - \phi\left\{ g_\theta(\tilde{X}_{i,\psi_{i,j}}); \theta \right\} / r(\tilde{L}_i; \theta) \right]. \tag{17}$$

and it admits (11) as a marginal when $\{\tilde{L}_i, \tilde{X}_i, \Psi_i, 1 \leq i \leq n\}$ is integrated out and V conditioned upon.

This density can be targeted by MCMC methods; actually at this stage we are only interested in the conditional density given V, i.e the joint posterior of parameters and auxiliary variables. We advocate a Gibbs sampler variant since conditionally on V and θ the auxiliary variables $\{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}$ are independent over i and can be generated using the EA. The conditional density of θ, $\pi(\theta|V, \{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\})$ is computable and although it will not be easy to generate from it directly, we can do so using a Metropolis-Hastings step. A typical implementation of the EMCMC is given below.

3.2.1 Special cases: Exact MCMC 1 and 2

Certain simplifications are feasible when a simpler EA can be applied to the process of interest. The EDA based on EA1 requires less imputation than that of EA3, since exact simulation from the target measure $\tilde{\mathbb{Q}}_\theta^{(t)}$ no longer requires the random variable $\tilde{L}$. Thus, the augmentation scheme involves only the auxiliary variables $\{\tilde{X}_i, \Psi_i, 1 \leq i \leq n\}$. The joint density analogous to (17) can be easily obtained and amounts to simply replacing $r(\tilde{L}; \theta)$ by $r(\theta)$. Similarly the conditional of θ trivially follows; originally it was given in Theorem 3 of Beskos *et al.* (2006b).

A data augmentation algorithm can be built using the auxiliary variables used in the EA2. We will not present the details of this, since it is not a direct modification of the general algorithm, as it is the case with EA1, but instead it involves a different construction of bridges. Details can be found in Chapter 7 of Sermaidis (2010).

Algorithm 3 Exact MCMC using EA3

1: Initialise the algorithm choosing θ^0. Set $t = 0$,
2: Iterate the following steps

- for $1 \leq i \leq n$, set $I_i = 0$ and repeat the following until $I_i = 1$
    - sample layer $\tilde{L}_i$ of Brownian bridge path $\tilde{X}_i \sim \mathbb{W}^{(t,0,0)}$,
    - sample $\kappa_i \sim \text{Po}[r(\tilde{L}_i; \theta^t)\Delta t_i]$,
    - simulate uniformly distributed varibles $(\psi_{i,j}, u_{i,j})$ on $[0, \Delta t_i] \times [0,1]$ and set $\Phi_i = \{\Psi_i, \Upsilon_i\}$, where $\Psi_i = \{\psi_{i,j}, 1 \leq j \leq \kappa_i\}$ and $\Upsilon_i = \{u_{i,j}, 1 \leq j \leq \kappa_i\}$,
    - conditionally on $\tilde{L}_i$, sample Brownian bridge $\tilde{X}_{i,\psi_{i,j}}$,
    - set $I_i = I(\tilde{L}_i, \tilde{X}_i, \Phi_i, V_{t_{i-1}}, V_{t_i}, \theta^t)$ as in (15),
    - if $I_i = 1$, then set $\tilde{L}_i^{t+1} = \tilde{L}_i$, $\Psi_i^{t+1} = \Psi_i$ and $\tilde{X}_i^{t+1} = \tilde{X}_i$.
- sample $\theta^* \sim q(\theta^t, .)$ and $U \sim \text{Un}[0,1]$,
- if

$$U < \frac{\pi(V, \theta^*, \{S(\tilde{X}_i^{t+1}), \tilde{L}_i^{t+1}, 1 \leq i \leq n\}) \; q(\theta^*, \theta^t)}{\pi(V, \theta^t, \{S(\tilde{X}_i^{t+1}), \tilde{L}_i^{t+1}, 1 \leq i \leq n\}) \; q(\theta^t, \theta^*)}$$

then set $\theta^{t+1} = \theta^*$, else set $\theta^{t+1} = \theta^t$,
- set $t = t + 1$.

3.3 Interpreting EDA in terms of the limiting irreducible HFI

The following result, which is based on Theorem 1, provides the connection between HFI and EDA. It effectively shows that the limiting irreducible HFI is a *collapsed* version of the EDA, i.e when we integrate out a subset of the latent variables we obtain the distribution which is targeted by HFI. This is an interesting and rather surprising result.

Theorem 2. *Let* $\pi(\theta, \{\tilde{X}_i, 1 \leq i \leq n\} \mid V)$ *be the density obtained from (17) by conditioning on V, and marginalizing w.r.t* $\{\tilde{L}_i, \Psi_i, 1 \leq i \leq n\}$, *and* $\pi_{HF,\infty}(\theta, \{\tilde{X}_i, 1 \leq i \leq n\} \mid V)$ *the density targeted by the limiting irreducible HFI defined in (13). Then , the two densities are equal a.s.*

The result is insightful towards the comparison of the computational efficiency of EMCMC to that of HFI, as it suggests that the mixing time of the former might generally be larger due to its higher degree of augmentation; a price one has to pay in order to eliminate the discretisation error. The numerical comparison of EDA and HFI is investigated in Section 5 in concrete examples. Nonetheless, in the next section, we show a variety of ways with which one can increase the performance of EMCMC and achieve good mixing rates.

4 Boosting EMCMC by reparametrisations and auxiliary Poisson sampling

We can make some qualitative remarks about the efficiency of the MCMC schemes based on the HFI and EDA which were reviewed and developed in Sections 3.1 and 3.2 respectively, and outlined in Algorithms 2 and 3. These remarks are based on general properties of data augmentation methods, as for example discussed in Papaspiliopoulos *et al.* (2007); Meng and Yu (2011). These qualitative statements are backed up by numerical evidence in Section 5, and they motivate the three approaches we propose in this Section to boost the algorithmic efficiency.

To fix terminology, we will identify data augmentation with a Gibbs-type sampler which updates auxiliary variables and parameters according to their conditional distributions. In general terms, data augmentation works well when the fraction of missing information is not too large relative to the observed information. This statement can be made precise mathematically (see for example the aforementioned references) but informally it means that when the augmented dataset is not considerably more informative than the observed regarding the parameters. For instance, in the naive HFI when θ_2 is unknown the augmented data is increasingly more informative than the observed data as M increases, infinitely more so in the limit $M \to \infty$. This is the reason why that scheme collapses.

In irreducible HFI, the efficiency of Gibbs sampling improves when the time increment t between a pair of observations decreases. When t is small, latent bridges look like Brownian bridges and all information

about the drift is contained in the end points. Thus, as t decreases the augmented information converges to the observed one. Indeed, in the trivial case where the process is Brownian motion with drift, the auxiliary variables and the parameters are independent and the algorithm samples directly from the posterior of the parameters regardless of the value of M. On the other hand, for moderate t the augmented data will be substantially more informative. The main advantage of HFI paradigm is that irreducible algorithm exists even at the limit where continuous paths are imputed.

In EDA the efficiency depends on r. We focus on the typical situation where r depends on θ. In the limiting case when the Poisson rate is 0, the missing data and parameters are independent (since the skeleton is empty) and EDA achieves maximal efficiency. The higher the sensitivity of ϕ to the parameters is, the stronger the dependence between parameters and missing data becomes. There are two reasons for this. First, the number of points contained in the skeleton is informative about r, and thus about θ. Second, changes in the parameters causes changes to the acceptance probabilities of the skeleton points, which penalises parameter values for which ϕ is considerably different than the one under which the current skeleton was accepted. Note that this problem affects also parameters not related with r, but which affect other features of ϕ. Similarly, for a given r the efficiency of EDA worsens as the time increment between consecutive observations increases. The acceptance rate of the EA at the imputation step decays to 0 exponentially with t. Hence, when data are sparsely sampled, the MCMC algorithm can spend a large amount of time by simulating proposed paths until acceptance. There are (at least) two ways to deal with this problem. First, instead of using rejection sampling to update the latent paths, one can use multiple independent Metropolis-Hastings steps where each proposal is accepted according to a likelihood ratio using (16). Alternatively, one can apply consecutively the EA on smaller time segments by imputing $\mathcal{O}(t)$ additional points between pairs of observations, thus resulting in a computational cost which is linear in t. Both methods, however, will have an impact on the MCMC mixing rate; the first due to rejections of Metropolis proposals restricting movement in that direction, and the second due to the additional augmentation.

4.1   EMCMC efficiency by reparametrisation

One general approach for improving efficiency of data augmentation in hierarchical models and auxiliary variable models is to adopt a reparametrisation. Indeed, we have already done so in irreducible HFI and in EDA by transforming $V \rightarrow \tilde{X}$. Following Papaspiliopoulos *et al.* (2007), for a generic random variable E and data V, a reparametrisation of an augmentation scheme E is defined by any random pair $(\tilde{E}, \theta)$ together with a function h such that

$$E = h(\tilde{E}, \theta, V),$$

where h need not be 1-1. A reparametrisation is called *noncentred* when the distribution of $\tilde{E}$ is independent of θ. Intuitively, in cases where V is not strongly informative about E, a noncentred scheme can perform well due to the prior independence of $\tilde{E}$ and θ. We will attempt to reduce the dependence between the Poisson process and θ by resorting to a noncentred reparametrisation. Hence, we seek for a function h of θ and a parameter independent process, say $\tilde{\Psi}$, such that $\Psi = h(\tilde{\Psi}, \tilde{L}, \theta)$, where Ψ is a Poisson process of rate $r(\tilde{L}; \theta)$ on $[0, t]$.

Noncentred reparametrisations for decoupling the dependence between Poisson processes and their intensity were originally proposed in Roberts *et al.* (2004). Applying their idea in this context, if $\tilde{\Psi}$ is a homogeneous Poisson process of unit intensity on $[0, t] \times [0, \infty)$, with point-coordinates $\{(\tilde{\psi}_j, \tilde{\xi}_j)\}$ then

$$\Psi = h(\tilde{\Psi}, \tilde{L}, \theta) = \left\{ \tilde{\psi}_j; \, \tilde{\xi}_j < r(\tilde{L}; \theta) \right\}. \tag{18}$$

Although $\tilde{\Psi}$ includes an infinite number of points, Ψ only depends on a finite subset of the points, those for which the second coordinate is below $r(\tilde{L}; \theta)$. Accounting for all the observations, the noncentred reparametrisation is $(\theta, \{\tilde{L}_i, \tilde{X}_i, \Psi_i, 1, \leq i \leq n\}) \rightarrow (\theta, \{\tilde{L}_i, \tilde{X}_i, \tilde{\Psi}_i, 1, \leq i \leq n\})$, where $\Psi_i = h(\tilde{\Psi}_i, \tilde{L}_i, \theta)$. The theorem below derives the conditional density of θ given the latent variables $\{\tilde{L}_i, \tilde{X}_i, \tilde{\Psi}_i, 1 \leq i \leq n\}$ and observations.

Theorem 3. *For $1 \leq i \leq n$, let $S_\theta(\tilde{X}_i) = \left\{ (0, 0), (\tilde{\psi}_{i,j}, \tilde{X}_{i, \tilde{\psi}_{i,j}}), (\Delta t_i, 0); \tilde{\xi}_{i,j} < r(\tilde{L}_i; \theta) \right\}$ be the skeleton of the accepted path $\tilde{X}_i$ evaluated at time points $\tilde{\psi}_{i,j}$ for which $\tilde{\xi}_{i,j} < r(\tilde{L}_i; \theta)$. Then θ is conditionally independent of $\{\tilde{X}_i, 1 \leq i \leq n\}$ given $\{S_\theta(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}$, with density $\pi_{nc}(\theta \mid \{S_\theta(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}, V)$ proportional to*

$$\pi(\theta)\exp\left[H\{x_n(\theta);\theta\}-H\{x_0(\theta);\theta\}-l(\theta)(t_n-t_0)\right]$$

$$\times\prod_{i=1}^{n}\left|\det\left[\sigma^{-1}(V_{t_i};\theta_2)\right]\right|\mathcal{N}_{\Delta t_i}^d\{x_i(\theta_2)-x_{i-1}(\theta_2)\}$$

$$\times\prod_{i=1}^{n}\prod_{j=1}^{\infty}\left[1-\mathbb{I}\left[\tilde{\xi}_{i,j}<r(\tilde{L}_i;\theta)\right]\phi\left\{g_\theta(\tilde{X}_{i,\psi_{i,j}});\theta\right\}/r(\tilde{L}_i;\theta)\right] \tag{19}$$

Expression (19) implies that evaluating the density at different values of θ requires revealing the paths at different collections of time instances which depend on θ. However, Theorem 3 ensures that evaluation for any value of θ requires only a finite number of calculations and therefore discretisations are avoided.

A simplification can be achieved when the EA1 is applicable, where the transformation in that case becomes $\{\tilde{X}_i,\Psi_i,1,\leq i\leq n\}\rightarrow\{\tilde{X}_i,\tilde{\Psi}_i,1,\leq i\leq n\}$, where $\Psi_i=h(\tilde{\Psi}_i,\theta)=\{\tilde{\psi}_{i,j};\tilde{\xi}_{i,j}<r(\theta)\}$. The conditional density of θ given the augmented dataset is essentially given by expression (19) replacing $r(\tilde{L}_i;\theta)$ with $r(\theta)$. Noncentred transformations can be found for EMCMC based on EA2, the details can be found in Section 7.4.3 of Sermaidis (2010).

The MCMC algorithm based on this reparametrisation is as follows, and practically it is a small modification of the MCMC based on EDA. First, we wish to draw from the conditional distribution of $\{\tilde{L}_i,\tilde{X}_i,\tilde{\Psi}_i,1\leq i\leq n\}$ given V and the current parameter value, say θ^t. It is clear from (18) that the proposal Poisson process $\tilde{\Psi}_i$ need only be revealed at a finite collection of time points, namely for those that satisfy $\tilde{\xi}_{i,j}<r(\tilde{L}_i;\theta^t)$. Therefore, we only need to simulate a Poisson process on $[0,\Delta t_i]\times[0,r(\tilde{L}_i,\theta^t)]$, which involves only $\tilde{\kappa}_i\sim\text{Po}[r(\tilde{L}_i,\theta^t)\Delta t_i]$ number of points and is sufficient for the implementation of the EA. The ith output consists of the accepted Poisson process $\tilde{\Psi}_i$ partially observed at $\left\{(\tilde{\psi}_{i,j},\tilde{\xi}_{i,j}),1\leq j\leq\tilde{\kappa}_i\right\}$ and the pair $\{\tilde{L}_i,\tilde{X}_i\}$ discretely observed at the revealed times points $\tilde{\psi}_{i,j}$ of $\tilde{\Psi}_i$. However, sampling from the conditional distribution of θ given the latent variables and Y is more tricky. Specifically, if the proposed value, say θ^*, is such that $r(\tilde{L}_i,\theta^*)>r(\tilde{L}_i;\theta^t)$, then evaluating the conditional density (19) at θ^* requires revealing the accepted path at additional time points $\left\{\tilde{\psi}_{i,j};r(\tilde{L}_i;\theta^t)<\tilde{\xi}_{i,j}<r(\tilde{L}_i;\theta^*)\right\}$, which have not been revealed in the EA output. Notice that this situation does not occur when $r(\tilde{L}_i;\theta^*)<r(\tilde{L}_i;\theta^t)$, since the path has already been revealed at all time instances required to evaluate the conditional density at θ^t and θ^*.

We propose two ways to perform the update of θ. The first approach is based on prospectively revealing the accepted Poisson process and path at any additional required time instances. Specifically, $\tilde{\Psi}_i$ can be further revealed on $[0,\Delta t_i]\times[0,r(\tilde{L}_i;\theta^*)]$ by simply simulating extra $\tilde{\kappa}_i^*$ uniform random variates on $[0,\Delta t_i]\times[r(\tilde{L}_i;\theta^t),r(\tilde{L}_i;\theta^*)]$, where $\tilde{\kappa}_i^*\sim\text{Po}\left\{\left[r(\tilde{L}_i;\theta^*)-r(\tilde{L}_i;\theta^t)\right]\Delta t_i\right\}$. The accepted path is then filled in at the additional $\tilde{\kappa}_i^*$ points by Brownian bridge interpolations. The second approach is closest in spirit to the retrospective nature of the exact algorithm. In particular, the proposed value θ^* can be simulated prior to the application of the EA and therefore the values $r(\tilde{L}_i;\theta^t)$ and $r(\tilde{L}_i;\theta^*)$ are known before the simulation of the latent path. Consequently, we can simulate the Poisson process directly on $[0,\Delta t_i]\times[0,r(\tilde{L}_i;\theta^*)]$ and reveal the paths $\tilde{X}_i$ at all required time instances during the implementation of the EA. As a result, the accepted path $\tilde{X}_i$ will be revealed at all required time points.

In our implementations, we follow the retrospective approach because it can be applied in a similar fashion to all three exact algorithms. The prospective approach is simple in the EA1 case due to simple Brownian bridge interpolations, but becomes more involved in the EA2 and EA3 cases. We note, however, that if θ is to be updated using an alternative approach to Metropolis-Hastings, such as a rejection sampling mechanism, then prospective sampling is the only feasible solution since the number of proposed values θ^* until one accepted cannot be known prior to the application of EA.

4.2 An interweaving strategy

When a noncentred transformation is available, it is not necessary to choose between that and the original parametrisation. Meng and Yu (2011) propose instead to *interweave* the two, by creating a single algorithm which mixes steps of both algorithms. This requires practically no extra coding work, but as it is demonstrated in the article, in certain cases (and even when taking the added computational cost into account) the interweaved algorithm outperforms its parent algorithms.

Algorithm 4 Noncentred Exact MCMC using EA3

1: Initialise the algorithm choosing θ^0. Set $t = 0$,
2: Iterate the following steps

- sample $\theta^* \sim q(\theta^t, .)$ and $U \sim \text{Un}[0,1]$,

- for $1 \leq i \leq n$, set $I_i = 0$ and repeat the following until $I_i = 1$
    - sample layer $\tilde{L}_i$ of Brownian bridge path $\tilde{X}_i \sim \mathbb{W}^{(t,0,0)}$,
    - set $r_{max} = r(\tilde{L}_i; \theta^t) \vee r(\tilde{L}_i; \theta^*)$ and sample $\tilde{\kappa}_i \sim \text{Po}[r_{max} \Delta t_i]$,
    - simulate uniformly distributed varibles $(\tilde{\psi}_{i,j}, \tilde{u}_{i,j}, \tilde{\xi}_{i,j})$ on $[0, \Delta t_i] \times [0,1] \times [0, r_{max}]$, set $\tilde{\Psi}_i = \left\{ (\tilde{\psi}_{i,j}, \tilde{\xi}_{i,j}), 1 \leq j \leq \tilde{\kappa}_i \right\}$ and $\Phi_i = \{\Psi_i, \Upsilon_i\}$, where $\Psi_i = h(\tilde{\Psi}_i, \tilde{L}_i, \theta^t)$ and $\Upsilon_i = h(\tilde{\Upsilon}_i, \tilde{L}_i, \theta^t)$ as in (18),
    - conditionally on $\tilde{L}_i$, sample Brownian bridge $\tilde{X}_{i, \tilde{\psi}_{i,j}}$,
    - set $I_i = I(\tilde{L}_i, \tilde{X}_i, \Phi_i, V_{t_{i-1}}, V_{t_i}, \theta^t)$ as in (15),
    - if $I_i = 1$, then set $\tilde{L}_i^{t+1} = \tilde{L}_i$, $\tilde{\Psi}_i^{t+1} = \tilde{\Psi}_i$ and $\tilde{X}_i^{t+1} = \tilde{X}_i$.

- if
$$U < \frac{\pi_{nc}(\theta^* \mid \{S_{\theta^*} \tilde{X}_i^{t+1}, \tilde{L}_i^{t+1}, 1 \leq i \leq n\}, V) \; q(\theta^*, \theta^t)}{\pi_{nc}(\theta^t \mid \{S_{\theta^*} \tilde{X}_i^{t+1}, \tilde{L}_i^{t+1}, 1 \leq i \leq n\}, V) \; q(\theta^t, \theta^*)}$$
then set $\theta^{t+1} = \theta^*$, else set $\theta^{t+1} = \theta^t$,

- set $t = t + 1$.

In the EMCMC context, if θ^t and $\{\tilde{L}_i^t, \tilde{X}_i^t, \tilde{\Psi}_i^t, 1, \leq i \leq n\}$ denote the current state of the chain, then the algorithm can effectively be described in four steps. The first two are identical to sampling from the noncentred algorithm, i.e. the latent variables are updated to $\{\tilde{L}_i^{t+1}, \tilde{X}_i^{t+1}, \tilde{\Psi}_i^{t+1}, 1, \leq i \leq n\}$ using the EA and then the parameter is updated by drawing $\theta^{t+\frac{1}{2}} \sim \pi_{nc}(. \mid \{S_\theta(\tilde{X}_i^{t+1}), \tilde{L}_i^{t+1}, 1 \leq i \leq n\}, V)$. Subsequently, the latent data are transformed to their original parametrisation using $\Psi_i = h(\tilde{\Psi}_i^{t+1}, \theta^{t+\frac{1}{2}})$; notice that this step does not impose any computational difficulties, since it merely involves a deterministic transformation. Finally, the parameter is re-drawn under the original parametrisation, $\theta^{t+1} \sim \pi(. \mid \{S(\tilde{X}_i^{t+1}), \tilde{L}_i^{t+1}, 1 \leq i \leq n\}, V)$.

4.3 Auxiliary Poisson sampling

An alternative way to improve the mixing time by increasing the computational cost is to exploit the connection between EDA and HFI. For simplicity we presente the theoretical result for EMCMC based on EA1, and then discuss extensions to EA3.

Note that if $r(\theta)$ is the Poisson rate, then it is valid to apply the EA1 with Poisson sampling rate $R(\theta)$ for any $R(\theta) > r(\theta)$; the acceptance probability is invariant to that choice. For simulating bridges it is optimal (in terms of computing time) to implement the algorithm with the smallest possible $R(\theta)$. This is not the case though for the EMCMC which iterates between imputation and estimation. Of course execution time will increase with $R(\theta)$. It turns out that the dependence between missing data and parameters decreases with $R(\theta)$ and optimal implementation (in terms of execution time and MC time) can be achieved for $R(\theta) > r(\theta)$. Thus, we consider data augmentation where the auxiliary variables are chosen according to the output of EA1, as described in Section 3.2, but where the Poisson rate is $R(\theta)$.

A theoretical result (not included here) goes along the following lines. Let $R(\theta) = r(\theta) + \lambda - 1$, where $\lambda \geq 1$ is a user-specified constant independent of θ. Then, the joint law of $(\theta, \{\tilde{X}_i, , 1 \leq i \leq n\})$ after a step of the exact MCMC with parameter λ, converges to the law of one step of the limiting irreducible HFI when $\lambda \to \infty$. An argument for proving this is based on properties of series expansions for exponential functionals, as discussed for example in Papaspiliopoulos (2011).

Hence, in a sense the auxiliary variables Ψ_i are effectively integrated out by increasing computation. Thus, we should expect an improvement in the convergence of the algorithm as λ increases, although this should be counterbalanced by the additional computational cost. λ gives an additional degree of freedom to enhance algorithmic performance.

A similar property is enjoyed by a generic EMCMC based on EA3. In fact, the limiting algorithm as λ

increases is a limiting irreducible HFI which also imputes the layer, although the latter is immaterial in the limit since it does not contain additional information about the parameters.

5 Numerical investigation of MCMC algorithms

We investigate the numerical performance of the several algorithms we have presented on some standard examples. One is a diffusion which belongs in the Pearson family, see for example Forman and Sørensen (2008), and it is an example of a process that can be simulated using the EA1. The second is a double well potential model, typical of models that are used to describe processes with metastable behaviour, see for example Metzner *et al.* (2006). This is an example of a process that can only be simulated using the EA3.

   We compare several algorithms. The plain-vanilla EMCMC together which its elaborations: using non-centred reparametrisation, using an interweaving of centred and noncentred, and using auxiliary Poisson sampling. Additionally, we compare against the irreducible HFI using different amounts of imputation. Finally, we consider a version of the irreducible HFI which is based on an alternative target density to (13). As we commented in Section 3.1 the expression we provide is not identical to the one obtained in equations (6) and (15) of Roberts and Stramer (2001), the difference being that we exploit the structure of the drift in the family of models we consider, to perform an integration by parts and replace a stochastic integral by a time integral. This is expected to lead to smaller Monte Carlo errors, hence we evaluate the effect of this simplification.

   The existence of the EMCMC allows us to have a realistic evaluation of the performance of the irreducible HFI. We carry out bootstrap Kolmogorov-Smirnov tests (Sekhon, 2007) for checking whether the posterior distributions on the parameters obtained by different levels of imputation are significantly different from the exact samples. These evaluations are based on thinning the original Markov chain output so that to obtain practically independent draws from the corresponding distributions.

5.1 A Pearson diffusion

Consider the univariate diffusion process specified by

$$dV_s = -\rho(V_s - \mu)ds + \sigma\sqrt{1 + V_s^2}dW_s,$$

where $\sigma > 0$, $\rho > 0$ is a mean reverting parameter and $\mu \in \mathbb{R}$ is the stationary mean. This model belongs to a rich class of diffusion processes, known as the Pearson diffusions (see Forman and Sørensen, 2008), and it admits a stationary distribution with heavy tails and skewness. It is easy to check that the invariant density of the process is proportional to

$$(1 + v^2)^{-(k+1)/2}\exp\left\{\frac{2\rho\mu}{\sigma^2}\tan^{-1}(v)\right\},$$

where $k := 1 + 2\rho/\sigma^2$, thus exhibiting tails decaying at the same rate as a t-distribution with k degress of freedom; when $\mu = 0$, the density coincides with a scaled t-distribution with k degrees of freedom and scale parameter $k^{-1/2}$.

   The Lamperti transform is easily derived as $\eta(v; \sigma) = \sinh(v)/\sigma$, and the drift of the unit volatility process is given by

$$a(x; \theta) = -\left(\frac{\rho}{\sigma} + \frac{\sigma}{2}\right)\tanh(\sigma x) + \frac{\rho\mu}{\sigma\cosh(\sigma x)}.$$

It can be easily verified that the process belongs to the EA1 class, and that exact inference can be performed with

$$l(\theta) = -\frac{1}{2}\left(\rho + \frac{\sigma^2}{2} + \frac{\rho\mu}{2}\right), \quad r(\theta) = \frac{1}{8}\left\{\rho(6\mu + 8) + 3\sigma^2 + \frac{4\rho^2}{\sigma^2}(\mu^2 + \mu + 1)\right\}. \tag{20}$$

   We test the methods on a simulated data set from this process, based on $n = 1000$ (excluding the initial point) equidistant points with $\Delta t_i = 1$, $V_0 = 1$ and parameter values $(\rho, \mu, \sigma) = (1/2, 1, 1/2)$ (Figure 3a). We have used improper prior densities for the parameters, $\pi(\rho) \propto 1$, $\pi(\mu) \propto 1$, and $\pi(\sigma) \propto 1/\sigma$. For all
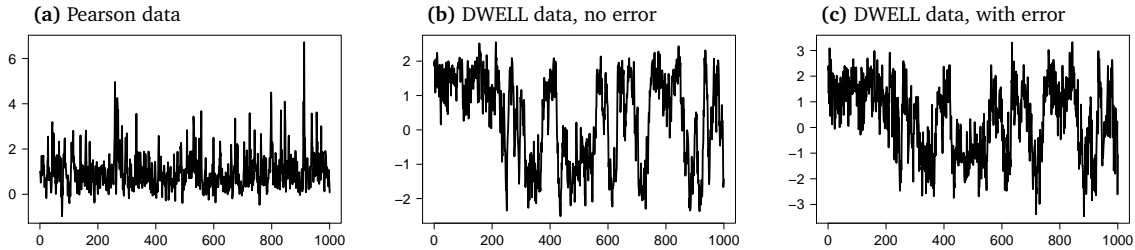
Figure 3: Simulated datasets from (a) the Pearson diffusion model, (b) the double well potential model and (c) the double well observed with error.

algorithms, sampling from the conditional density of the parameters was performed using a block Metropolis-Hastings step. The chains were run for 10^5 iterations.

Figure 4 shows the autocorrelation plots along with posterior density estimates derived from the Markov chains. Starting from EMCMC1 under the original parametrisation, notice that for $\lambda = 1$, the chain exhibits strong serial dependence even at large lags, particularly for ρ. This is due to stong a priori dependence between the parameter and the number of Poisson points (see expression (20)); which remains significant in the posterior distribution. In particular, the posterior correlation between $\sum_i^n \kappa_i$ and ρ was estimated equal to 0.93, thus suggesting that noncentring the Poisson process can result in better mixing rates, as Figure 4e confirms. In order to improve the performance of algorithm under the original parametrisation, we consider various values for $\lambda = \{6, 11, 21\}$, thus revealing the path at additionally $5, 10$ and 20 points between consecutive observations respectively. The increase in performance is reflected in the autocorrelation function, which now decays to 0 more quickly. As expected, the chains of the HFI algorithms mix more rapidly than the exact ones due to the less amount of augmentation. Finally, notice that that the posterior distributions estimated from the HFI algorithms seem to provide a reasonable approximation to the true ones only for $M = 30$ (Figures 4g to 4i).

Table 1 presents posterior summary statistics, the average number of imputed points between consecutive observations, the effective sample size (ESS) (calculated with R (R Development Core Team, 2010) package `coda` (Plummer *et al.*, 2010)), the computational (CPU) time of each algorithm, and the p-value for the bootstrap Kolmogorov-Smirnov test with null hypothesis that the simulated draws from the marginal distributions of HFI and exact algorithms come from the same distribution. Notice that even for $M = 30$, the HFI algorithm fails to pass the tests at a 5% significance level, whereas less amount of imputation ($M = 20$) combined with the integration by parts yields less biased approximations, clearly illustrating the importance of eliminating the stochastic integral. For each method, we also calculate an adjusted effective sample size, defined as the ratio of the effective sample size to the CPU time. In terms of computational performance, the noncentred algorithm outperforms the rest and exhibits an adjusted ESS for μ and σ which is approximately 60% larger than that of the sufficiently accurate HFI methods.

The algorithms were also run using proper priors for the parameters, an exponential distribution for ρ, a Gaussian for μ and an inverse Gamma for σ^2, yielding no significant differences from the results presented above.
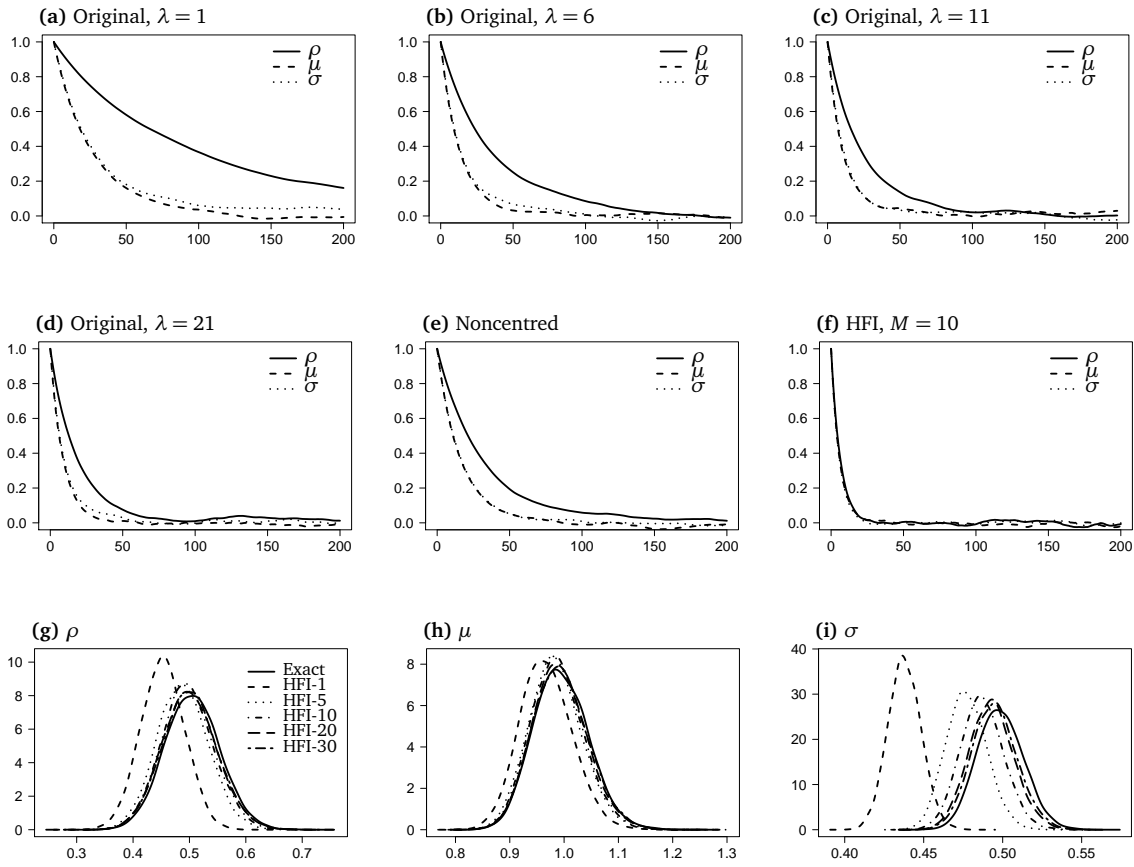
Figure 4: The Pearson diffusion model with $n = 1000$ simulated data points. True values are $(\rho, \mu, \sigma) = (1/2, 1, 1/2)$. Autocorrelations are reported after a burn in of 5000 iterations. Posterior density estimates using EMCMC1 (noncentred) and HFI algorithms for (g) ρ, (h) μ and (i) σ.

Method	Parameter	$\lambda - 1$	Imp. points	Mean	Std. dev.	ESS_{adj}	ESS	KS	Time	Correlation matrix		
Exact (noncentred)	ρ		2.316	0.505	0.048	6.599	16.789		2.544	1.000	-0.447	0.539
	μ			0.995	0.050	12.432	31.629				1.000	0.008
	σ			0.499	0.015	12.117	30.829					1.000
Exact (original)	ρ	0	2.193	0.503	0.047	2.628	6.133		2.333	1.000	-0.428	0.531
	μ			0.995	0.050	8.549	19.946				1.000	0.017
	σ			0.499	0.015	8.091	18.879					1.000
Exact (original)	ρ	5	7.202	0.505	0.048	2.602	15.186		5.836	1.000	-0.451	0.538
	μ			0.995	0.049	6.796	39.656				1.000	-0.021
	σ			0.500	0.014	6.124	35.739					1.000
Exact (original)	ρ	10	12.205	0.505	0.047	2.425	22.607		9.321	1.000	-0.426	0.528
	μ			0.995	0.049	5.222	48.672				1.000	0.033
	σ			0.499	0.014	4.849	45.201					1.000
Exact (interweaved)	ρ	0	2.314	0.506	0.049	5.795	22.170		3.826	1.000	-0.438	0.540
	μ			0.994	0.050	13.032	49.859				1.000	0.021
	σ			0.499	0.015	12.488	47.779					1.000
HFI-5	ρ		5.000	0.488	0.045	20.244	74.643	< 0.001	3.687	1.000	-0.439	0.459
	μ			0.985	0.049	21.841	80.530	< 0.001			1.000	0.029
	σ			0.477	0.013	24.563	90.566	< 0.001				1.000
HFI-20	ρ		20.000	0.500	0.048	7.080	80.155	0.006	11.321	1.000	-0.427	0.518
	μ			0.992	0.050	7.426	84.077	0.037			1.000	0.016
	σ			0.493	0.014	7.721	87.414	< 0.001				1.000
HFI-30	ρ		30.000	0.501	0.048	4.916	80.518	0.005	16.380	1.000	-0.439	0.519
	μ			0.993	0.050	5.160	84.516	0.017			1.000	0.021
	σ			0.495	0.014	5.329	87.293	< 0.001				1.000
HFI-5 (int-by-parts)	ρ		5.000	0.507	0.049	22.221	81.018	0.005	3.646	1.000	-0.443	0.537
	μ			0.994	0.050	23.749	86.591	0.553			1.000	0.001
	σ			0.500	0.014	24.578	89.612	0.007				1.000
HFI-20 (int-by-parts)	ρ		20.000	0.505	0.049	6.740	78.421	0.395	11.636	1.000	-0.426	0.541
	μ			0.995	0.050	7.414	86.264	0.192			1.000	0.020
	σ			0.499	0.014	7.599	88.418	0.720				1.000

Table 1: The Pearson diffusion model with $n = 1000$ simulated data points. True values are $(\rho, \mu, \sigma) = (1/2, 1, 1/2)$. Summaries of exact and approximate posterior distributions. Statistics are reported after a burn-in period of 5000 iterations. The ESS column shows the effective sample size of each chain per 1000 iterations. The Time column shows the time (in seconds) required for 1000 iterations of each chain. The adjusted effective sample size is shown in column ESS_{adj}. Each entry of the KS column shows the p-value for the Kolmogorov-Smirnov test with null hypothesis that draws from the marginal distributions of HFI and exact algorithm (noncentred) come from the same distribution.

5.2 A double well potential model

We consider the solution process to

$$dV_s = -\rho V_s \left(V_s^2 - \mu \right) ds + \sigma dW_s,$$

where $\rho > 0, \mu > 0, \sigma > 0$. The process is known as the double well potential process (denoted by DWELL hereafter). We simulated $n = 1000$ (excluding the initial point) equidistant observations with $\Delta t_i = 1$ for parameter setting $(\rho, \mu, \sigma) = (0.1, 2, 1/2)$ and $V_0 \sim N(2, 1/4)$ (Figure 3b). Reduction to a unit volatility process is easily achieved using $X_s := V_s / \sigma$, which solves the SDE

$$dX_s = -\rho X_s \left(\sigma^2 X_s^2 - \mu \right) ds + dW_s.$$

Simple calculations reveal that the function $f(u; \theta) := \|\alpha(u; \theta)\|^2 + \Delta_x H(u; \theta)/2$ is given by

$$f(u; \theta) = \frac{\rho}{2} \left[\rho \sigma^4 u^6 - 2\rho\mu\sigma^2 u^4 + \left(\rho\mu^2 - 3\sigma^2 \right) u^2 + \mu \right],$$

and that the algorithms are applicable with

$$l(\theta) = f(u_l; \theta), \text{ where } u_l^2 = \frac{2\rho\mu + \sqrt{\rho(\rho\mu^2 + 9\sigma^2)}}{3\rho\sigma^2}.$$

Finally, for a given realisation of the layer $\tilde{L}$ we need to find the upper bound (14). This is easily achieved by noticing that

$$f(u; \theta) \leq \frac{\rho}{2} \left(\rho \sigma^4 u^6 + \rho\mu^2 u^2 + \mu \right) =: g(u; \theta),$$

which is positive and has a minimum at $u = 0$, implying that

$$r(\tilde{L}; \theta) = \left[g \left\{ -\tilde{L}T + \bar{x}^{\{r\}}(\theta_2); \theta \right\} \vee g \left\{ \tilde{L}T + \bar{y}^{\{r\}}(\theta_2); \theta \right\} \right] - l(\theta). \tag{21}$$

We assign improper prior densities to the parameters with $\pi(\rho) \propto 1$, $\pi(\mu) \propto 1$, and $\pi(\sigma) \propto 1/\sigma$, and run the chains for 10^5 iterations. The performance of the algorithms and posterior density estimates are shown in Figure 5. Notice that the performance of EMCMC under the original parametrisation is very poor, due to strong posterior correlation between Poisson points and parameters; a fact attributed to the sensitivity of $r(\tilde{L}, \theta)$ to the parameters (see expression (21)). On the other hand, the noncentred aglorithm exhibits a much stronger performance with low serial correlation for each parameter even after 50 lags.

Posterior summaries from the output of the chains and computational performance are gathered in Table 2. In contrast to the EA1 example presented earlier, the interweaved strategy, after accounting for the additional computational cost, does not offer any significant improvement over the noncentred algorithm. Finally, the HFI method with $M = 40$ paired with the integration by parts provide a reasonable approximation to the posterior marginal densities and, in terms of computational performance, exhibits slightly larger adjusted ESS than the noncentred algorithm. The results were robust in changes in parameter prior distributions.
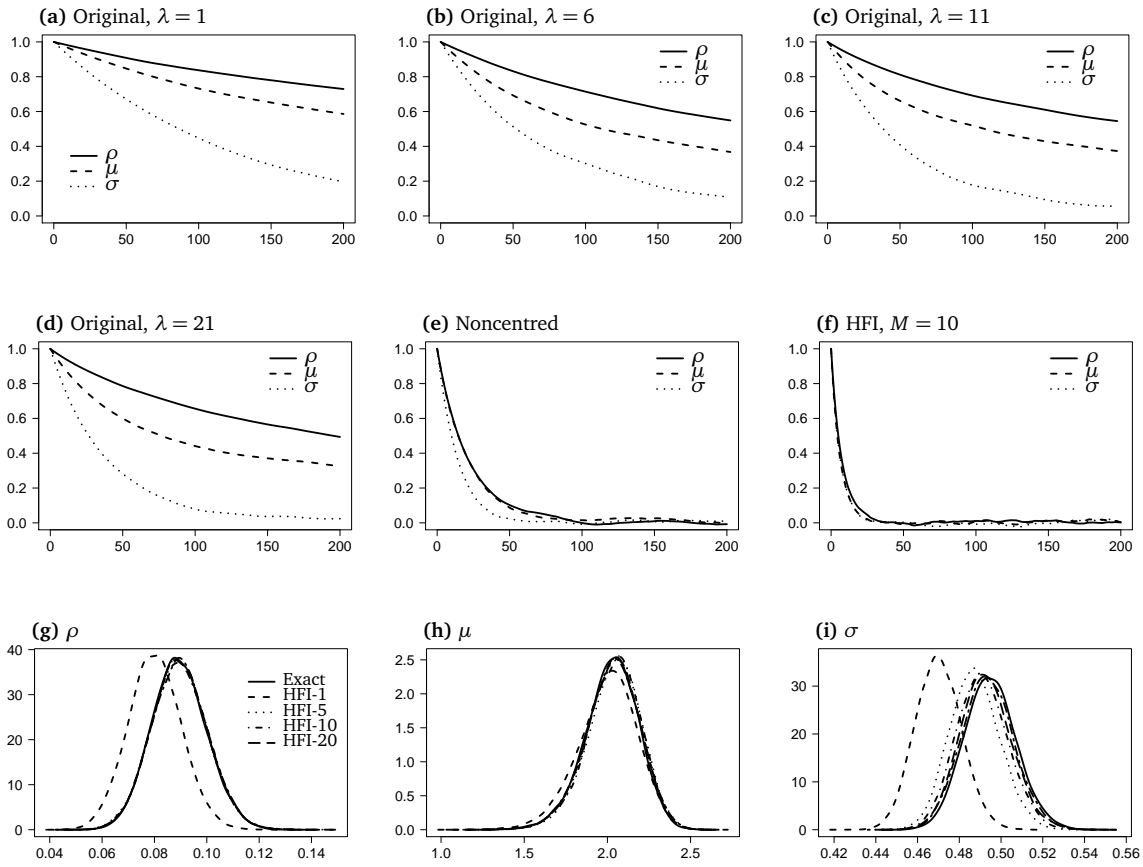
Figure 5: The DWELL diffusion model with $n = 1000$ simulated data points. True values are $(\rho, \mu, \sigma) = (0.1, 2, 1/2)$. Autocorrelations are reported after a burn in of 5000 iterations. Posterior density estimates using EMCMC3 (noncentred) and HFI algorithms for (g) ρ, (h) μ and (i) σ.

Method	Parameter	$\lambda-1$	Imp. points	Mean	Std. dev.	ESS$_{adj}$	ESS	KS	Time	Correlation matrix		
Exact (noncentred)	ρ		6.028	0.089	0.010	1.699	24.244		14.270	1.000	0.469	0.383
	μ			2.023	0.160	1.756	25.053				1.000	0.029
	σ			0.495	0.012	2.449	34.947					1.000
Exact (original)	ρ	5	10.472	1.093	0.011	0.088	1.966		22.286	1.000	0.433	0.359
	μ			7.631	1.136	0.181	4.030				1.000	-0.003
	σ			1.640	0.020	0.309	6.896					1.000
Exact (original)	ρ	10	15.437	1.093	0.011	0.068	2.109		31.063	1.000	0.475	0.338
	μ			7.553	1.197	0.133	4.132				1.000	-0.033
	σ			1.641	0.020	0.305	9.459					1.000
Exact (original)	ρ	20	25.474	1.093	0.012	0.050	2.459		48.714	1.000	0.502	0.334
	μ			7.611	1.276	0.094	4.574				1.000	-0.011
	σ			1.641	0.020	0.270	13.142					1.000
Exact (interweaved)	ρ	0	5.978	0.089	0.011	1.451	24.060		16.581	1.000	0.488	0.353
	μ			2.019	0.163	1.587	26.322				1.000	0.005
	σ			0.495	0.012	2.458	40.757					1.000
HFI-40	ρ		40.000	0.090	0.011	3.337	73.816	0.052	22.123	1.000	0.472	0.386
	μ			2.029	0.161	3.631	80.322	0.186			1.000	0.026
	σ			0.494	0.012	3.594	79.502	< 0.001				1.000
HFI-60	ρ		60.000	0.090	0.011	2.296	72.493	0.033	31.572	1.000	0.470	0.362
	μ			2.026	0.163	2.571	81.187	0.011			1.000	0.016
	σ			0.494	0.012	2.622	82.772	0.107				1.000
HFI-20 (int-by-parts)	ρ		20.000	0.089	0.010	6.367	72.584	0.035	11.400	1.000	0.474	0.387
	μ			2.024	0.163	7.026	80.104	0.258			1.000	0.022
	σ			0.495	0.012	7.166	81.695	0.005				1.000
HFI-40 (int-by-parts)	ρ		40.000	0.089	0.010	3.409	74.472	0.079	21.848	1.000	0.467	0.376
	μ			2.023	0.160	3.916	85.557	0.694			1.000	0.006
	σ			0.495	0.012	3.796	82.930	0.294				1.000

Table 2: The DWELL diffusion model with $n = 1000$ simulated data points. True values are $(\rho, \mu, \sigma) = (0.1, 2, 1/2)$. Summaries of exact and approximate posterior distributions. Statistics are reported after a burn-in period of 5000 iterations. The ESS column shows the effective sample size of each chain per 1000 iterations. The Time column shows the time (in seconds) required for 1000 iterations of each chain. The adjusted effective sample size is shown in column ESS$_{adj}$. Each entry of the KS column shows the p-value for the Kolmogorov-Smirnov test with null hypothesis that draws from the marginal distributions of HFI and exact algorithm (noncentred) come from the same distribution.

6 Diffusion observed with error

The methodology described so far can be easily extended to account for cases where the diffusion is not directly observed. We assume that the observations inform only indirectly about the value of the process (1) at discrete times $t_i, i = 0, 1, \ldots, n$, according to the following observation equation

$$Y_{t_i} \sim q(. \mid V_{t_i}, \tau), \tag{22}$$

where $Y := \{Y_{t_0}, Y_{t_1}, \ldots, Y_{t_n}\}$ are conditionally independent given $V = \{V_{t_0}, V_{t_1}, \ldots, V_{t_n}\}$, and q is a known density function parametrised by an unknown parameter τ.

The EDA described in Section 3.2 is not appropriate anymore since the end points V_{t_i} are not directly observed. On the other hand, Theorem 1 can be used to design an augmentation scheme where apart from the auxiliary variables involved in EDA, the latent points V are imputed as well. A direct application of Bayes' theorem yields that inference about parameters and latent points must be based upon

$$\pi(V, \theta, \tau \{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\} \mid Y) \propto$$
$$\pi(\theta, \tau) \pi(V, \{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\} \mid \theta) \pi(V_{t_0} \mid \theta) \prod_{i=0}^{n} q(Y_{t_i} \mid V_{t_i}, \tau) \tag{23}$$

where the second term is obtained directly from Theorem 1, $\pi(. \mid \theta)$ is a prior density function for the initial point of the diffusion process V_{t_0} and where potentially a joint prior is specified for (θ, τ). As before, a direct simplification is available when the EA1 can be applied to simulate from V.

A simple scheme for simulating from (23) is by a component-wise updating algorithm. $\{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}$ and θ are simulated conditionally on V and τ, using any of the EMCMC schemes we have proposed, and subsequently V and τ conditionally on $\{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}$ and θ according to the conditional derived from (23). When $\pi(\theta, \tau) = \pi(\theta)\pi(\tau)$, τ is conditionally independent from $\{S(\tilde{X}_i), \tilde{L}_i, 1 \leq i \leq n\}$ and θ given V, and in many cases it might have a conditional density which can be easily simulated. Simulation of the latent points V can be done with various ways of differing intelligence. The simplest approach is to update them one-at-a-time (an approach often called single-site updating) according to their conditional density. Such schemes for time series are known to be in general problematic (see for example Pitt and Shephard, 1999) especially when the latent process exhibits high persistence and the observations are not very informative about the latent points (see also Papaspiliopoulos *et al.*, 2007). In the example we consider below we adopt this simplistic approach since it works quite well. In applications where the process V exhibits very high persistence a joint update of the endpoints can be done using a Metropolis-adjusted Langevin algorithm, or more general version of such algorithms, discussed for example in Girolami *et al.* (2011). Other possibility is to resort to an overlapping block scheme, as for example in Pitt and Shephard (1999); Golightly and Wilkinson (2008).

6.1 An illustration

We illustrate the performance of the methods using the DWELL model, after adding to the previous observations a Gaussian error with mean 0 and variance $\tau^2 = 1/4$ (Figure 3c). We have assigned the same priors for the diffusion parameters as in the previous example, and an improper prior proportional to $1/\tau$ for τ. In view of the poor performance of the exact methods under the original parametrisation, we only employ EMCMC under the noncentred parametrisation along with the HFI scheme with increasing values of $M = 5, 10, 20, 40$. The algorithms are run for 10^5 iterations and the MCMC outputs are thinned every 10 iterations. Figure 6 presents trace and autocorrelation plots for all parameters along with the marginal posterior density estimates. It is interesting to notice that the presence of noise in the observations seems to aid the aproximation of the HFI methods, since the posterior densities do not change significantly as M increases, and seem to provide a reasonable approximation to the exact ones.
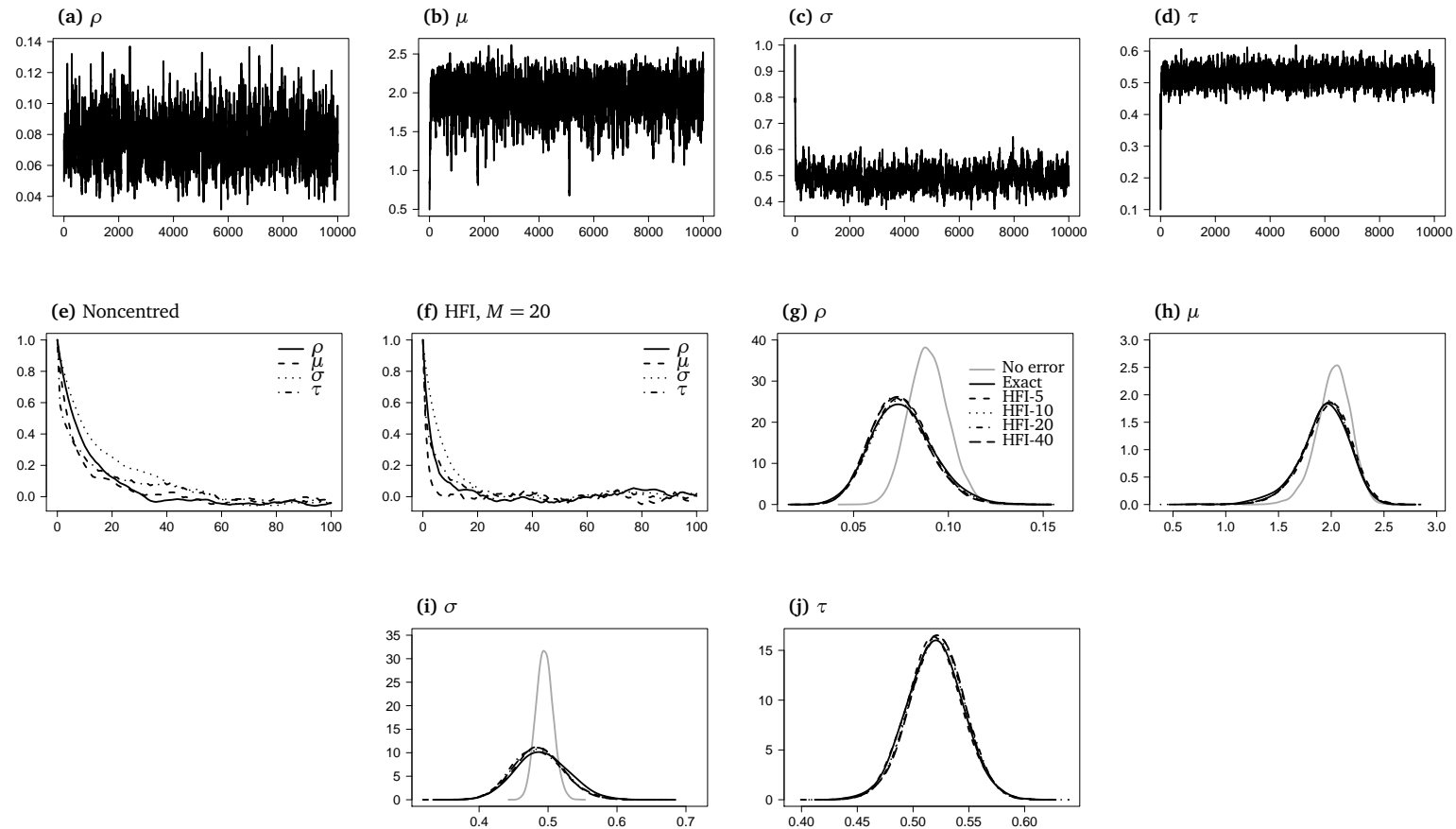
Figure 6: The DWELL diffusion model observed with error with $n = 1000$ simulated data points. True values are $(\rho, \mu, \sigma, \tau) = (1/2, 1, 1/2, 1/2)$. The outputs of the MCMC chains are subsampled every 10 iterations. Trace plots (top row) of output from exact (noncentred) algorithm. Autocorrelations are reported after a burn in of 5000 iterations. Posterior density estimates using EMCMC3 (noncentred) and HFI algorithms for (g) ρ, (h) μ, (i) σ, and (j) τ. In (g)-(i) we superimpose the posterior density obtained when the same process is directly observed.

7 Discussion

We have developed exact data augmentation methods for discretely directly and indirectly observed diffusions. We have established the precise connection between this paradigm and the best existing alternative method when the variance-stabilizing transformation can be performed, the irreducible HFI. We have also pointed out an intriguing connection between exact and approximate methods: the degree of freedom rendered by the Poisson sampling rate. On going work involves the rigorous proof of the effect of the auxiliary Poisson sampling. The auxiliary sampling can be seen as a variance reduction scheme. In general, there is a large scope for investigating other such schemes both for EDA and HFI. In this article we have already demonstrated the effect of performing integration by parts where possible to the efficiency of MCMC sampling.

The extension of these methods outside the class of processes prescribed by the current version of EA3 is definitely an exciting direction. Another direction of interest for future research is to explore further the connection between unbiased estimation of transition density and MCMC. There is a large and growing literature which develops MCMC algorithms for models with intractable likelihoods using unbiased estimators thereof; see for example Andrieu and Roberts (2009); Andrieu *et al.* (2010). The class of diffusions where such estimators can be obtained is much larger than that simulated by EA3, see for example Section 4.6 of Papaspiliopoulos (2011).

There exists available software for implementing all the methods in this paper, which is available on request by the authors.

References

Adams, R. P., Murray, I. and MacKay, D. J. C. (2009) The Gaussian process density sampler. In *Advances in Neural Information Processing Systems 21* (eds. D. Koller, D. Schuurmans, Y. Bengio and L. Bottou), 9–16.

Aït-Sahalia, Y. (2008) Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics*, **36**, 906–937.

Ait-Sahalia, Y. and Kimmel, R. (2007) Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics*, **83**, 413–452.

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo. *J. Roy. Statist. Soc. Ser. B*, **3**, 269âĂŞ–342.

Andrieu, C. and Roberts, G. O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

Beskos, A., Papaspiliopoulos, O. and Roberts, G. O. (2006a) Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, **12**, 1077–1098.

Beskos, A., Papaspiliopoulos, O. and Roberts, G. O. (2008) A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, **10**, 85–104.

Beskos, A., Papaspiliopoulos, O. and Roberts, G. O. (2009) Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics*, **37**, 223–245.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006b) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of Royal Statistical Society, Series B: Statistical Methodology*, **68**, 333–382. With discussions and a reply by the authors.

Brown, P. E., K{r}aresen, K. F., Roberts, G. O. and Tonellato, S. (2000) Blur-generated non-separable space-time models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **62**, 847–860.

Elerian, O., Chib, S. and Shephard, N. (2001) Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, **69**, 959–993.

Eraker, B. (2001) MCMC analysis of diffusion models with application to finance. *J. Bus. Econom. Statist.*, **19**, 177–191.

Eraker, B., Johannes, M. and Polson, N. (2003) The impact of jumps in volatility and returns. *Journal of Finance*, **58**, 1269–1300.

Étoré, P. and Martinez, M. (2011) Exact simulation of one-dimensional stochastic differential equations involving the local time at zero of the unknown process. Tech. rep. Available online from `http://hal.archives-ouvertes.fr/docs/00/56/52/86/PS/etore_martinez1.ps`.

Fearnhead, P., Papaspiliopoulos, O. and Roberts, G. O. (2008) Particle filters for partially observed diffusions. *Journal of Royal Statistical Society, Series B: Statistical Methodology*, **70**, 755–777.

Forman, J. L. and Sørensen, M. (2008) The Pearson diffusions: a class of statistically tractable diffusion processes. *Scand. J. Statist.*, **35**, 438–465.

Girolami, M., Calderhead, B. and Chin, S. (2011) Riemannian manifold hamiltonian monte carlo. *Journal of Royal Statistical Society, Series B: Statistical Methodology*, **to appear**.

Golightly, A. and Wilkinson, D. J. (2006) Bayesian sequential inference for stochastic kinetic biochemical network models. *J. Comput. Biol.*, **13**, 838–851.

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

Goncalves, F. and Roberts, G. (2011) Exact simulation and bayesian inference for jump-diffusion process. Tech. rep. Available online from `www.dme.ufrj.br/ebebx/FlavioGoncalves.pdf`.

Horenko, I. and Schütte, C. (2008) Likelihood-based estimation of multidimensional Langevin models and its application to biomolecular dynamics. *Multiscale Model. Simul.*, **7**, 731–773.

Jasra, A. and Doucet, A. (2009) Sequential Monte Carlo methods for diffusion processes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*.

Jones, C. S. (1999) Bayesian estimation of continuous-time finance models. Unpublished paper, Simon School of Business, University of Rochester.

Kalogeropoulos, K., Roberts, G. and Dellaportas, P. (2010) Inference for stochastic volatility models using time change transformations. *The Annals of Statistics*, **38**, 784–807.

Kou, S. C., Xie, X. S. and Liu, J. S. (2005) Bayesian analysis of single-molecule experimental data. *J. Roy. Statist. Soc. Ser. C*, **54**, 469–506.

Meng, X. and Yu, Y. (2011) To center or not to center: That is not the question - an ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, **to appear**.

Metzner, P., Schütte, C. and Vanden-Eijnden, E. (2006) Illustration of transition path theory on a collection of simple examples. *The Journal of Chemical Physics*, **125**, 084110.

Øksendal, B. (2003) *Stochastic differential equations. An introduction with applications*. Universitext. Berlin: Springer-Verlag, 6th edn.

Papaspiliopoulos, O. (2011) A methodological framework for monte carlo probabilistic inference for diffusion processes. In *Bayesian Time Series Models*. Cambridge University Press.

Papaspiliopoulos, O. and Roberts, G. (2009) Importance sampling techniques for estimation of diffusion models. In *SEMSTAT*. Chapman and Hall.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007) A general framework for the parametrization of hierarchical models. *Statistical Science*, **22**, 59–73.

Peluchetti, S. and Roberts, G. O. (2008) An empirical study of the efficiency of the EA for diffusion simulation. Tech. rep., University of Warwick.

Picchini, U., Gaetano, A. and Ditlevsen, S. (2010) Stochastic Differential Mixed-Effects Models. *Scandinavian Journal of Statistics*, **37**, 67–90.

Pitt, M. and Shephard, N. (1999) Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models. *Journal of Time Series Analysis*, **20**, 63–85.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2010) *coda: Output analysis and diagnostics for MCMC*. URL http://CRAN.R-project.org/package=coda. R package version 0.13-5.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007) Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 741–796. With discussions and a reply by the authors.

Roberts, G. O., Papaspiliopoulos, O. and Dellaportas, P. (2004) Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *Journal of Royal Statistical Society, Series B: Statistical Methodology*, **66**, 369–393.

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

Rydberg, T. H. (1997) A note on the existence of unique equivalent martingale measures in a Markovian setting. *Finance and Stochastics*, **1**, 251–257.

Sekhon, J. S. (2007) Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*.

Sermaidis, G. (2010) *Likelihood-based inference for discretely observed diffusions*. Ph.D. thesis, Department of Statistics, University of Warwick.

Stramer, O., Bognar, M. and Schneider, P. (2010) Bayesian inference for discretely sampled Markov processes with closed-form likelihood expansions. *Journal of Financial Econometrics*, **8**, 450–480.

Sundaresan, S. M. (2000) Continuous-time methods in finance: A review and an assessment. *Journal of Finance*, **55**, 1569–1622.

Taylor, J., Cumberland, W. and Sy, J. (1994) A Stochastic Model for Analysis of Longitudinal AIDS Data. *Journal of the American Statistical Association*, **89**, 727–736.

8   Appendix

Proof of Lemma 2

Proof. Let $\mathbb{L}_\theta^{(t)}$ be the measure of a homogeneous Poisson process of intensity $r(\tilde{L}; \theta)$ on $[0,t] \times [0,1]$. Expression (16) is derived by writing the density of the accepted random variables $(\tilde{L}, \tilde{X}, \Phi)$ with respect to the law of the proposed $\mathbb{M}^{(t)} \times \mathbb{L}_\theta^{(t)}$,

$$\pi(\tilde{L}, \tilde{X}, \Phi \mid v, w, \theta) = \frac{1}{a(x, y, t, \theta)} \prod_{j=1}^\kappa \mathbb{I}\left[\frac{1}{r(\tilde{L}; \theta)} \phi\left\{ \tilde{X}_{\psi_j} + \left(1 - \frac{\psi_j}{t}\right) x(\theta_2) + \frac{\psi_j}{t} y(\theta_2); \theta \right\} < u_j \right],$$

and by invoking a change of measure from the law of a Poisson process of intensity $r(\tilde{L}; \theta)$ to the law of one of intensity λ, thus ensuring a parameter-independent dominating measure. Finally, integrating out the marks $\Upsilon = \{u_j, 1 \leq j \leq \kappa\}$, we obtain expression (16). $\qquad\square$

Proof of Theorem 1

Proof. The factorization of the density in the three terms is elementary. For any two fixed points x and y, taking expectations on both sides of (6) with respect to $\mathbb{W}_\theta^{(t,x,y)}$ we derive the fundamental identity

$$\tilde{p}_t(x,y;\theta) = \mathcal{N}_t^d(y-x) \exp\{H(y;\theta) - H(x;\theta) - l(\theta)t\} a(x,y,t,\theta), \tag{24}$$

which combined with (5) gives $a(x,y,t,\theta)$ as a function of $p_t(V_{t_{i-1}}, V_{t_i}; \theta)$. Combining this with Lemma 2 yields the expression.

It remains to show that (11) can be obtained by integrating out the auxiliary variables and conditioning on the data. However, this is trivial since by taking expectations w.r.t the dominating measure $\otimes_{i=1}^n \left(\mathbb{M}^{(\Delta t_i)} \times \mathbb{P}_\lambda^{(\Delta t_i)}\right)$ we obtain the marginal $\pi(\theta)\prod_{i=1}^n p_{\Delta t_i}(V_{t_{i-1}}, V_{t_i}; \theta)$ from which (11) follows as a conditional. $\qquad\square$

Proof of Theorem 2

Proof. For notational simplicity, we define $K(V,\theta)$ to be the following deterministic function of observations V and θ:

$$\exp\{H\{x_n(\theta_2);\theta\} - H\{x_0(\theta_2);\theta\} - l(\theta)(t_n - t_0)\}\prod_{i=1}^n \left|\det\left[\sigma^{-1}(V_{t_i};\theta_2)\right]\right| \mathcal{N}_{\Delta t_i}^d\{x_i(\theta_2) - x_{i-1}(\theta_2)\}. \tag{25}$$

The joint posterior density of θ and imputed data $\{\tilde{L}_i, \tilde{X}_i, \Psi_i, 1 \le i \le n\}$ is given (up to a constant) in expression (17). Integrating out the Poisson processes is easily done by first integrating out the coordinates and then the number of Poisson points. Therefore, integrating out $\Psi_i = \{\psi_{i,j}, 1 \le j \le \kappa_i\}$, $1 \le i \le n$, we obtain

$$\pi(\theta)K(V,\theta)\exp\left(-\sum_{i=1}^n \left[r(\tilde{L}_i;\theta) - \lambda\right]\Delta t_i\right)\prod_{i=1}^n \left\{\frac{1}{\lambda\Delta t_i}\int_0^{\Delta t_i}\left[r(\tilde{L}_i;\theta) - \phi\left\{g_\theta(\tilde{X}_{i,s});\theta\right\}\right]ds\right\}^{\kappa_i}.$$

Integrating out the Poisson points yields which, by integrating out the Poisson points, yields the posterior density of θ and $\{\tilde{L}_i, \tilde{X}_i, 1 \le i \le n\}$ with respect to $\text{Leb}^p \otimes_{i=1}^n \mathbb{M}^{(\Delta t_i)}$, as

$$\pi(\theta)K(V,\theta)\exp\left(-\sum_{i=1}^n \left[r(\tilde{L}_i;\theta) - \lambda\right]\Delta t_i\right)$$

$$\times \prod_{i=1}^n \exp\left\{\lambda\Delta t_i\left[\frac{1}{\lambda\Delta t_i}\int_0^{\Delta t_i}\left[r(\tilde{L}_i;\theta) - \phi\left\{g_\theta(\tilde{X}_{i,s});\theta\right\}\right]ds - 1\right]\right\}$$

$$= \pi(\theta)K(V,\theta)\exp\left\{-\sum_{i=1}^n \int_0^{\Delta t_i}\phi\left\{g_\theta(\tilde{X}_{i,s});\theta\right\}ds\right\}.$$

Given that the construction of $\mathbb{M}^{(\Delta t_i)}$, we obtain that by integrating out the layers, we obtain the joint density of θ and $\{\tilde{X}_i, 1 \le i \le n\}$ with respect to $\text{Leb}^p \otimes_{i=1}^n \mathbb{W}^{(\Delta t_i,0,0)}$ which coincides with the posterior density (13) targeted by HFI. $\qquad\square$

Proof of Theorem 3

Proof. For a pair of observations $(v_{t_{i-1}}, v_{t_i})$, the joint density of the accepted elements of EA3 $(\tilde{L}_i, \tilde{X}_i, \tilde{\Psi}_i)$ conditionally on $V_{t_{i-1}}, V_{t_i}, \theta$ is given by

$$\frac{\prod_{j=1}^\infty \left\{1 - \mathbb{I}\left[\tilde{\xi}_{i,j} < r(\tilde{L}_i;\theta)\right]\phi\left\{g_\theta(\tilde{X}_{i,\psi_{i,j}});\theta\right\}/r(\tilde{L}_i;\theta)\right\}}{a(x_{i-1}(\theta_2), x_i(\theta_2), \Delta t_i, \theta)},$$

with respect to the product measure $\mathbb{M}^{(\Delta t_i)} \times \tilde{\mathbb{L}}^{(\Delta t_i)}$, where $\tilde{\mathbb{L}}^{(t)}$ is the measure of a unit rate Poisson process on $[0,t] \times (0,\infty)$. Using the conditional independence of the latent data given V and θ,

$$\pi_{nc}(\{\tilde{L}_i, \tilde{X}_i, \tilde{\Psi}_i, 1 \le i \le n\} \mid V, \theta) = \prod_{i=1}^n \pi_{nc}(\tilde{L}_i, \tilde{X}_i, \tilde{\Psi}_i \mid V_{t_{i-1}}, V_{t_i}, \theta),$$

and the proof follows along the same lines as Theorem 1. $\qquad\square$