Markov chain Monte Carlo algorithms for SDE parameter estimation

Andrew Golightly and Darren J. Wilkinson

Abstract

This chapter considers stochastic differential equations for Systems Biology models derived from the Chemical Langevin Equation (CLE). After outlining the derivation of such models, Bayesian inference for the parameters is considered, based on state-of-the-art Markov chain Monte Carlo algorithms. Starting with a basic scheme for models observed perfectly, but discretely in time, problems with standard schemes and their solutions are discussed. Extensions of these schemes to partial observation and observations subject to measurement error are also considered. Finally, the techniques are demonstrated in the context of a simple stochastic kinetic model of a genetic regulatory network.

1 Introduction

It is now well recognised that the dynamics of many genetic and biochemical networks are intrinsically stochastic. Stochastic kinetic models provide a powerful framework for modelling such dynamics; see, for example, McAdams & Arkin (1997) and Arkin et al. (1998) for some early examples. In principle such stochastic kinetic models correspond to discrete state Markov processes that evolve continuously in time (Wilkinson 2006). Such processes can be simulated on a computer using a discrete-event simulation technique such as the Gillespie algorithm (Gillespie 1977). Since many of the parameters governing these models will be uncertain, it is natural to want to estimate them using experimental data. Although it is possible in principle to use a range of different data for this task, time course data on the amounts of bio-molecules at the single-cell level are the most informative. Boys et al. (2008) show that it is possible to directly infer rate constants of stochastic kinetic models using fully Bayesian inference and sophisticated Markov chain Monte Carlo (MCMC) algorithms. However, the techniques are highly computationally intensive, and do not scale-up to problems of practical interest in Systems Biology. It seems unlikely that fully Bayesian inferential techniques of practical value can be developed based on the original Markov jump process formulation of stochastic kinetic models, at least given currently available computing hardware.

It is therefore natural to develop techniques which exploit some kind of approximation in order to speed up computations. One possibility, explored in Boys

et al. (2008), is to work with the exact model, but to introduce approximations into the Bayesian inference algorithm. Although this approach does have some promise, it seems difficult to speed up the algorithm sufficiently for practical purposes without sacrificing too much inferential accuracy. An alternative approach is to approximate the model, and then conduct exact Bayesian inference for the approximate model. This latter approach is much more flexible, as there are many approximations to the underlying model which can be made, and it is easier to understand the accuracy of the proposed approximations and the likely benefit in terms of computational speed-up. Perhaps the most obvious approach would be to use a deterministic approximation, such as that based on the reaction rate equations (Gillespie 1992). However, such an approach performs very badly when the underlying process has a significant degree of stochasticity, as the reaction rate equations (RREs) effectively "throw away" all of the stochasticity in the process, and consequently, all of the information in the process "noise". The information in the noise is often quite substantial, and needs to be utilised for effective inference.

The Chemical Langevin Equation (Gillespie 1992, Wilkinson 2006) is a diffusion approximation to the discrete stochastic kinetic model that preserves most of the important features of the stochastic dynamics. Furthermore, the stochastic differential equation (SDE) representation of the Chemical Langevin Equation (CLE) lends itself to new and potentially more efficient Bayesian inference methodology. In the remainder of this chapter, inference for the CLE using time course data is considered, and application to the estimation of stochastic rate constants is examined. However, most of the methodology considered is quite generic, and will apply straightforwardly to any (nonlinear, multivariate) SDE model observed (partially,) discretely in time (and with error).

2 Stochastic Kinetics

2.1 Stochastic Kinetic Models

Consider a biochemical reaction network involving u species $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_u$ and v reactions $R_1, R_2, \ldots, R_v$, written using standard chemical reaction notation as

$$\begin{aligned}
R_1: &\quad p_{11}\mathcal{X}_1 + p_{12}\mathcal{X}_2 + \cdots + p_{1u}\mathcal{X}_u \longrightarrow q_{11}\mathcal{X}_1 + q_{12}\mathcal{X}_2 + \cdots + q_{1u}\mathcal{X}_u \\
R_2: &\quad p_{21}\mathcal{X}_1 + p_{22}\mathcal{X}_2 + \cdots + p_{2u}\mathcal{X}_u \longrightarrow q_{21}\mathcal{X}_1 + q_{22}\mathcal{X}_2 + \cdots + q_{2u}\mathcal{X}_u \\
&\qquad\quad \vdots \qquad\qquad\qquad\quad \vdots \qquad\qquad \vdots \\
R_v: &\quad p_{v1}\mathcal{X}_1 + p_{v2}\mathcal{X}_2 + \cdots + p_{vu}\mathcal{X}_u \longrightarrow q_{v1}\mathcal{X}_1 + q_{v2}\mathcal{X}_2 + \cdots + q_{vu}\mathcal{X}_u.
\end{aligned}$$

Let X_{jt} denote the number of molecules of species $\mathcal{X}_j$ at time t, and let X_t be the u-vector $X_t = (X_{1t}, X_{2t}, \ldots, X_{ut})'$. The $v \times u$ matrix P consists of the coefficients p_{ij}, and Q is defined similarly. The $u \times v$ *stoichiometry matrix*, S is defined by

$$S = (Q - P)'.$$

The matrices P, Q and S will typically be *sparse*. On the occurrence of a reaction of type i, the system *state*, X_t is updated by adding the ith column of S. Consequently, if ΔR is a v-vector containing the number of reaction events of each type in a given time interval, then the system state should be updated by ΔX, where

$$\Delta X = S\Delta R.$$

The stoichiometry matrix therefore encodes important structural information about the reaction network. In particular, vectors in the left null-space of S correspond to *conservation laws* in the network. That is, any u-vector a satisfying $a'S = 0$ has the property (clear from the above equation) that $a'X_t$ remains constant for all t.

Under the standard assumption of *mass-action stochastic kinetics*, each reaction R_i is assumed to have an associated rate constant, c_i, and a *propensity function*, $h_i(X_t, c_i)$ giving the overall *hazard* of a type i reaction occurring. That is, the system is a *Markov jump process*, and for an infinitesimal time increment dt, the probability of a type i reaction occurring in the time interval $(t, t + dt]$ is $h_i(X_t, c_i)dt$. The hazard function takes the form

$$h_i(X_t, c_i) = c_i \prod_{j=1}^{u} \binom{X_{jt}}{p_{ij}}.$$

Let $c = (c_1, c_2, \ldots, c_v)'$ and $h(X_t, c) = (h_1(X_t, c_1), h_2(X_t, c_2), \ldots, h_v(X_t, c_v))'$. Values for c and the initial system state x_0 complete specification of the Markov process. Although this process is rarely analytically tractable for interesting models, it is straightforward to forward-simulate exact realisations of this Markov process using a discrete event simulation method. This is due to the fact that if the current time and state of the system are t and x_t respectively, then the time to the next event will be exponential with rate parameter

$$h_0(x_t, c) = \sum_{i=1}^{v} h_i(x_t, c_i),$$

and the event will be a reaction of type R_i with probability $h_i(x_t, c_i)/h_0(x_t, c)$ independently of the waiting time. Forwards simulation of process realisations in this way is typically referred to as *Gillespie's direct method* in the stochastic kinetics literature, after Gillespie (1977). See Wilkinson (2006) for further background on stochastic kinetic modelling.

In fact, the assumptions of mass-action kinetics, as well as the one-to-one correspondence between reactions and rate constants may both be relaxed. All of what follows is applicable to essentially arbitrary v-dimensional hazard functions $h(X_t, c)$.

The central problem considered in this paper is that of inference for the stochastic rate constants, c, given some time course data on the system state, X_t. It is therefore most natural to first consider inference for the above Markov jump process stochastic kinetic model. As demonstrated by Boys et al. (2008), exact

Bayesian inference in this setting is theoretically possible. However, the problem appears to be computationally intractable for models of realistic size and complexity, due primarily to the difficulty of efficiently exploring large integer lattice state space trajectories. It turns out to be more tractable (though by no means straightforward) to conduct inference for a continuous state Markov process approximation to the Markov jump process model. Construction of this diffusion approximation, known as the *Chemical Langevin Equation*, is the subject of the next section.

2.2 The Diffusion Approximation

The diffusion approximation to the Markov jump process can be constructed in a number of more or less formal ways. We will present here an informal intuitive construction, and then provide brief references to more rigorous approaches.

Consider an infinitesimal time interval, $(t, t+dt]$. Over this time, the reaction hazards will remain constant almost surely. The occurrence of reaction events can therefore be regarded as the occurrence of events of a Poisson process with independent realisations for each reaction type. Therefore, if we write dR_t for the v-vector of the number of reaction events of each type in the time increment, it is clear that the elements are independent of one another and that the ith element is a $Po(h_i(X_t, c_i)dt)$ random quantity. From this we have that $\mathrm{E}(dR_t) = h(X_t, c)dt$ and $\mathrm{Var}(dR_t) = \mathrm{diag}\{h(X_t, c)\}dt$. It is therefore clear that

$$dR_t = h(X_t, c)dt + \mathrm{diag}\left\{\sqrt{h(X_t, c)}\right\} dW_t$$

is the Itô stochastic differential equation (SDE) which has the same infinitesimal mean and variance as the true Markov jump process (where dW_t is the increment of a v-dimensional Brownian motion). Now since $dX_t = SdR_t$, we can immediately deduce

$$dX_t = Sh(X_t, c)dt + S\,\mathrm{diag}\left\{\sqrt{h(X_t, c)}\right\} dW_t$$

as a SDE for the time evolution of X_t. As written, this SDE is a little unconventional, as the driving Brownian motion is of a different (typically higher) dimension than the state. This is easily remedied by noting that

$$\mathrm{Var}(dX_t) = S\,\mathrm{diag}\{h(X_t, c)\}S',$$

which immediately suggests the alternative form

$$dX_t = Sh(X_t, c)dt + \sqrt{S\,\mathrm{diag}\{h(X_t, c)\}S'}dW_t, \tag{1}$$

where now X_t and W_t are both u-vectors. Equation (1) is the SDE most commonly referred to as the *chemical Langevin equation* (CLE), and represents the diffusion process which most closely matches the dynamics of the associated Markov jump process. In particular, whilst it relaxes the assumption of discrete states, it keeps all of the stochasticity associated with the discreteness of state in

its noise term. It also preserves the most important structural properties of the Markov jump process. For example, (1) defines a non-negative Markov stochastic process, and has the same conservation laws as the original stochastic kinetic model.

More formal approaches to the construction of the CLE usually revolve around the Kolmogorov forward equations for the Markov processes. The Kolmogorov forward equation for the Markov jump process is usually referred to in this context as the *chemical master equation*. A second-order Taylor approximation to this system of differential equations can be constructed, and compared to the corresponding forward equation for an SDE model (known in this context as the *Fokker-Planck equation*). Matching the second-order approximation to the Fokker-Planck equation leads to the same CLE (1), as presented above. See Gillespie (1992) and Gillespie (2000) for further details.

2.3 Prokaryotic Auto-regulation

In order to illustrate the inferential methods to be developed in subsequent sections, it will be useful to have a non-trivial example model. We will adopt the model introduced in Golightly & Wilkinson (2005), and later examine parameter inference for this model in some challenging data-poor scenarios. The model is a simplified model for prokaryotic auto-regulation based on the mechanism of dimers of a protein coded for by a gene repressing its own transcription. The full set of reactions in this simplified model are:

$$\begin{aligned} R_1: &\quad \mathsf{DNA} + \mathsf{P}_2 \longrightarrow \mathsf{DNA}\cdot\mathsf{P}_2 & R_2: &\quad \mathsf{DNA}\cdot\mathsf{P}_2 \longrightarrow \mathsf{DNA} + \mathsf{P}_2 \\ R_3: &\quad \mathsf{DNA} \longrightarrow \mathsf{DNA} + \mathsf{RNA} & R_4: &\quad \mathsf{RNA} \longrightarrow \mathsf{RNA} + \mathsf{P} \\ R_5: &\quad 2\mathsf{P} \longrightarrow \mathsf{P}_2 & R_6: &\quad \mathsf{P}_2 \longrightarrow 2\mathsf{P} \\ R_7: &\quad \mathsf{RNA} \longrightarrow \emptyset & R_8: &\quad \mathsf{P} \longrightarrow \emptyset. \end{aligned}$$

See Golightly & Wilkinson (2005) for further explanation. We order the variables as $X = (\mathsf{RNA}, \mathsf{P}, \mathsf{P}_2, \mathsf{DNA}\cdot\mathsf{P}_2, \mathsf{DNA})$, giving the stoichiometry matrix for this system:

$$S = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 2 & 0 & -1 \\ -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The associated hazard function is given by

$$h(X, c) = (c_1\mathsf{DNA}\times\mathsf{P}_2, c_2\mathsf{DNA}\cdot\mathsf{P}_2, c_3\mathsf{DNA}, c_4\mathsf{RNA}, c_5\mathsf{P}(\mathsf{P}-1)/2, c_6\mathsf{P}_2, c_7\mathsf{RNA}, c_8\mathsf{P})',$$

using an obvious notation.

Like many biochemical network models, this model contains conservation laws leading to rank degeneracy of the stoichiometry matrix, S. The Bayesian inference methods to be considered in the subsequent sections are simpler to present

in the case of models of full-rank. This is without loss of generality, as we can simply strip out redundant species from the rank-deficient model. Here there is just one conservation law,

$$\mathsf{DNA} \cdot \mathsf{P}_2 + \mathsf{DNA} = k,$$

where k is the number of copies of this gene in the genome. We can use this relation to remove $\mathsf{DNA} \cdot \mathsf{P}_2$ from the model, replacing any occurrences of $\mathsf{DNA} \cdot \mathsf{P}_2$ in rate laws with $k - \mathsf{DNA}$. This leads to a reduced full-rank model with species $X = (\mathsf{RNA}, \mathsf{P}, \mathsf{P}_2, \mathsf{DNA})$, stoichiometry matrix

$$S = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 2 & 0 & -1 \\ -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2}$$

and associated hazard function

$$h(X, c) = (c_1\mathsf{DNA}\times\mathsf{P}_2, c_2(k-\mathsf{DNA}), c_3\mathsf{DNA}, c_4\mathsf{RNA}, c_5\mathsf{P}(\mathsf{P}-1)/2, c_6\mathsf{P}_2, c_7\mathsf{RNA}, c_8\mathsf{P})'. \tag{3}$$

We can then substitute (2) and (3) into the CLE (1) in order to get our SDE model that is to be the object of inference in Section 4.

3 Inference for Nonlinear Diffusion Models

As with ordinary differential equations (ODEs), stochastic differential equations can be solved numerically in the absence of an analytic solution. Performing inference however, when no analytic solutions exist is not trivial since transition densities will not be available in closed form. Inference is further complicated when there is only partial observation on a subset of diffusion components and the data may be subject to measurement error. Attempts to overcome this problem include the use of estimating functions (Bibby & Sørensen 1995), simulated maximum likelihood estimation (Pedersen 1995, Durham & Gallant 2002) and Bayesian imputation approaches (Elerian et al. 2001, Roberts & Stramer 2001, Eraker 2001). These methods are neatly summarised by Sørensen (2004). In the recent literature, Monte-Carlo methods which are both exact (in the sense that they are devoid of discretization error) and computationally efficient have been proposed by Beskos et al. (2006). Whilst attractive, such methods can only be applied to a relatively small class of diffusions.

Here, the Bayesian imputation approach to estimating diffusion parameters using discrete time data is considered. We describe the modelling framework in the presence of full observation before examining a basic Gibbs sampling strategy. It will be shown that such strategies (that alternate between draws of the diffusion parameters conditional on the data, and draws of the latent data conditional on the parameters and observed data), can break down if the augmentation is large. A proposed solution is outlined in detail, and extensions of the methodology to partial and noisy observation are considered.

3.1 Full Observation

Consider inference for a parameterised family of u-dimensional Itô diffusion processes satisfied by a stochastic differential equation of the form

$$dX_t = \mu\left(X_t, c\right)\, dt + \sqrt{\beta\left(X_t, c\right)}\, dW_t\,, \tag{4}$$

where μ is u-dimensional drift, β is a $u \times u$ dimensional diffusion matrix and $c = (c_1, \ldots, c_v)'$ is an unknown parameter vector of length v. It is assumed that the conditions under which the SDE can be solved for X_t are satisfied — that is to say (4) has a nonexploding, unique solution (see for example Chapter 5 of Øksendal (1995)). Note that for the stochastic kinetic models considered in Section 2, it is natural to choose

$$\mu\left(X_t, c\right) = Sh(X_t, c), \qquad \beta\left(X_t, c\right) = S\,\mathrm{diag}\{h(X_t, c)\}S'\,.$$

By adopting the Bayesian imputation approach, it is necessary to work with the discretized version of (4), given by the Euler approximation,

$$\Delta X_t \equiv X_{t+\Delta t} - X_t = \mu\left(X_t, c\right)\Delta t + \sqrt{\beta\left(X_t, c\right)}\,\Delta W_t\,, \tag{5}$$

where ΔW_t is a $\mathcal{N}(0, I\Delta t)$ random vector of length u. Plainly,

$$X_{t+\Delta t}|X_t, c \sim \mathcal{N}\left(X_t + \mu\left(X_t, c\right)\Delta t\,,\, \beta\left(X_t, c\right)\Delta t\right)$$

for which the probability density is

$$p\left(X_{t+\Delta t}|X_t, c\right) = \mathcal{N}\left(X_{t+\Delta t}\,; X_t + \mu\left(X_t, c\right)\Delta t\,,\, \beta\left(X_t, c\right)\Delta t\right) \tag{6}$$

where $\mathcal{N}(\cdot; \theta\,,\, \Sigma)$ denotes the Gaussian density with mean vector θ and covariance matrix Σ.

Initially, let us suppose that observations x_{τ_i} are available at evenly spaced times τ_0, τ_1, …, τ_T with intervals of length $\Delta^* = \tau_{i+1} - \tau_i$. As it is typically unrealistic to assume that Δ^* is sufficiently small to be used as a time step in (5), we put $\Delta t = \Delta^*/m$ for some positive integer $m > 1$. Then, choosing m to be sufficiently large ensures that the discretization bias is arbitrarily small, but also introduces $m - 1$ missing values in between every pair of observations, which must be integrated out of the problem. We note that the idea of augmenting the observed low frequency data with missing values was proposed by Pedersen (1995) and has since been pursued by Eraker (2001), Roberts & Stramer (2001) and Golightly & Wilkinson (2005) among others.

In order to provide a framework for dealing with these missing values, the entire time interval $[\tau_0, \tau_T]$ is divided into $mT + 1$ equidistant points $\tau_0 = t_0 < t_1 < \ldots < t_n = \tau_T$ (where $n = mT$) such that X_t is observed at times $t_0, t_m, t_{2m}, \ldots, t_n$. Altogether there are $(m-1)T$ missing values which are substituted with simulations X_{t_i}. Stacking all augmented data (both missing and observed) in matrix form gives a skeleton path,

$$\mathbf{X} = \left(x_{t_0}, X_{t_1}, \ldots, X_{t_{m-1}}, x_{t_m}, X_{t_{m+1}}, \ldots\ldots, X_{t_{n-1}}, x_{t_n}\right)$$

and herein, X^i denotes the value of the path $\mathbf{X}$ at time t_i. Within this framework, we have data $D_n = (x_{t_0}, x_{t_m}, \ldots, x_{t_n})$. Hence, by adopting a fully Bayesian approach, we formulate the joint posterior for parameters and missing data as

$$p\left(c, \mathbf{X} | \, D_n\right) \propto p(c) \prod_{i=0}^{n-1} p\left(X^{i+1} | X^i, c\right) \tag{7}$$

where $p\left(c\right)$ is the prior density for parameters and $p\left(\cdot \, | X^i, c\right)$ is the Euler density given by (6). As discussed in (Tanner & Wong 1987), inference may proceed by alternating between draws of the missing data conditional on the current state of the model parameters, and the parameters conditional on the augmented data. This procedure generates a Markov chain with the desired posterior, (7), as its equilibrium distribution. MCMC methods for the analysis of diffusion processes have been extensively explored in the recent literature. See for example, the work by Roberts & Stramer (2001), Elerian et al. (2001) and Eraker (2001). For full observation, we perform the following sequence of steps:

1. Initialise all unknowns. Use linear interpolation to initialise the X^i. Set $s := 1$.
2. Draw $\mathbf{X}_{(s)} \sim p\left(\cdot \, | c_{(s-1)}, D_n\right)$.
3. Draw $c_{(s)} \sim p\left(\cdot \, | \mathbf{X}_{(s)}\right)$.
4. If the desired number of simulations have been performed then stop, otherwise, set $s := s + 1$ and return to step 2.

The full conditionals required in steps 2 and 3 are proportional to $p\left(c, \mathbf{X} | \, D_n\right)$ in equation (7). For the diffusions considered here, $p\left(c \, | \mathbf{X}\right)$ typically precludes analytic form. Step 2 is therefore performed via a Metropolis-Hastings (MH) step. We find that the Gaussian random walk update of (Golightly & Wilkinson 2005) can be used to effectively perform step 3.

Attention is now turned to the task of performing step 2, that is, to update the latent data conditional on the current parameter values and the observed data. We present two sampling strategies and consider an inherent problem associated with their mixing properties. Finally, a sampling strategy that overcomes this problem is presented.

3.1.1 Single Site Updating

Eraker (2001) (see also Golightly & Wilkinson (2005)) samples $p\left(\mathbf{X} \, | c, D_n\right)$ indirectly, by implementing a Gibbs sampler to update each column X^i of $\mathbf{X} \backslash \{D_n\}$ conditional on its neighbours and the parameter value c. The full conditional distribution of X^i (with i not an integer multiple of m) is

$$\begin{aligned} p\left(X^i | X^{i-1}, X^{i+1}, c\right) &\propto p\left(X^i | X^{i-1}, c\right) p\left(X^{i+1} | X^i, c\right) \\ &= \mathcal{N}\left(X^i \, ; X^{i-1} + \mu\left(X^{i-1}, c\right) \Delta t \, , \, \beta\left(X^{i-1}, c\right) \Delta t\right) \\ &\quad \times \, \mathcal{N}\left(X^{i+1} \, ; X^i + \mu\left(X^i, c\right) \Delta t \, , \, \beta\left(X^i, c\right) \Delta t\right) \, . \end{aligned}$$

For nonlinear diffusions, direct sampling of this distribution is not possible and a MH step is used. A new X_*^i is drawn from a suitable proposal density. We follow Eraker (2001) and use

$$q\left(X_*^i|X^{i-1},X^{i+1},c\right)\equiv\mathcal{N}\left(X_*^i;\frac{1}{2}\left(X^{i-1}+X^{i+1}\right),\frac{1}{2}\beta\left(X^{i-1},c\right)\Delta t\right)$$

as a proposal density. If the iteration counter is at s, then X^{i-1} is the value obtained at iteration s and X^{i+1} is the value obtained at iteration $s-1$. Hence, if the current state of the chain is X^i, then a proposed value X_*^i is accepted with probability

$$\min\left\{1,\frac{p\left(X_*^i|X^{i-1},X^{i+1},c\right)}{p\left(X^i|X^{i-1},X^{i+1},c\right)}\times\frac{q\left(X^i|X^{i-1},X^{i+1},c\right)}{q\left(X_*^i|X^{i-1},X^{i+1},c\right)}\right\} \tag{8}$$

and we set $X_{(s)}^i:=X_*^i$, otherwise we store X^i. Note that $p\left(\cdot\,|X^{i-1},X^{i+1},c\right)$ needs only be known up to a multiplicative constant, since the acceptance probability only involves ratios of this density.

Hence, we sample $p\left(c,\mathbf{X}|\,D_n\right)$ with the following algorithm:

1. Initialise all unknowns. Use linear interpolation to initialise the X^i. Set $s:=1$.
2. Update $\mathbf{X}_{(s)}|c_{(s-1)},D_n$ as follows:
   2.1 Propose $X_*^i\sim q\left(X_*^i|X^{i-1},X^{i+1},c\right)$ for each i not an integer multiple of m.
   2.2 Set $X_{(s)}^i:=X_*^i$ with probability as in (8) otherwise store the current value X^i.
3. Draw $c_{(s)}\sim p\left(\cdot\,|\mathbf{X}_{(s)}\right)$.
4. If the desired number of simulations have been performed then stop, otherwise, set $s:=s+1$ and return to step 2.

For univariate diffusions, Elerian et al. (2001) show that an algorithm which updates one column of $\mathbf{X}$ at a time leads to poor mixing due to high correlation amongst the latent data. Consequently, it is recommended that $\mathbf{X}\backslash\{D_n\}$ is updated in blocks of random size. Here, we consider the simplest blocking algorithm whereby the latent values are updated in blocks of size $m-1$, between every pair of observations.

3.1.2 Block Updating

Consider consecutive observations x_{t_j} and x_{t_M} (where we let $M=j+m$) corresponding to columns X^j and X^M in $\mathbf{X}$. Between these two observations, we

have $m-1$ missing values, $X^{j+1},\ldots,X^{M-1}$ for which the full conditional distribution is

$$p\left(X^{j+1},\ldots,X^{M-1}|X^{j},X^{M},c\right)\propto\prod_{i=j}^{M-1}p\left(X^{i+1}|X^{i},c\right)\ .$$

We aim to sample this density for $j=0,m,\ldots,n-m$ in turn, thereby generating a sample from $p\left(\mathbf{X}|c,D_{n}\right)$. However, under the nonlinear structure of the underlying diffusion process, obtaining this density in analytic form is complicated. We therefore use a MH step; following Durham & Gallant (2002), we construct a Gaussian approximation to the density of X^{i+1} (for $i=j,\ldots,M-2$) conditional on X^{i} and the end-point of the interval. We construct the joint density of X^{M} and X^{i+1} by combining the Euler transition density in (6) with an approximation of $p\left(X^{M}|X^{i+1},c\right)$. Conditioning the resulting distribution on the end-point X^{M} gives

$$\tilde{p}\left(X^{i+1}|X^{i},X^{M},c\right)=\mathcal{N}\left(X^{i+1}\,;X^{i}+\mu^{*}\left(X^{i}\right)\Delta t\,,\,\beta^{*}\left(X^{i},c\right)\Delta t\right)\qquad(9)$$

where

$$\mu^{*}\left(X^{i}\right)=\frac{X^{M}-X^{i}}{t_{M}-t_{i}}\,,\ \ \beta^{*}\left(X^{i},c\right)=\left(\frac{t_{M}-t_{i+1}}{t_{M}-t_{i}}\right)\beta\left(X^{i},c\right)\qquad(10)$$

and we drop the dependence of μ^{*} and β^{*} on t to ease the notation.

We refer to (9) as the modified diffusion bridge construct. Hence, for each $j=0,m,\ldots,n-m$ we sample $p\left(X^{j+1},\ldots,X^{M-1}|X^{j},X^{M},c\right)$ by proposing X_{*}^{i+1} for $i=j,\ldots,M-2$ via recursive draws from the density in (9). Note that this gives a skeleton path of a diffusion bridge, conditioned to start at X^{j} and finish at X^{M}. If the current state of the chain is $X^{j},\ldots,X^{M-1}$ then we accept the move with probability given by

$$\min\left\{1\,,\ \left[\prod_{i=j}^{M-1}\frac{p\left(X_{*}^{i+1}|X_{*}^{i},c\right)}{p\left(X^{i+1}|X^{i},c\right)}\right]\times\left[\prod_{i=j}^{M-2}\frac{\tilde{p}\left(X^{i+1}|X^{i},X^{M},c\right)}{\tilde{p}\left(X_{*}^{i+1}|X_{*}^{i},X^{M},c\right)}\right]\right\}\qquad(11)$$

and this acceptance probability tends to a finite limit as $m\to\infty$. To see this, note that the modified diffusion bridge construct in (9) can be regarded as a discrete-time approximation of the SDE with limiting form

$$dX_{t}^{*}=\mu^{*}\left(X_{t}^{*}\right)\,dt+\sqrt{\beta\left(X_{t}^{*},c\right)}\,dW_{t}\,,\qquad(12)$$

as demonstrated in Stramer & Yan (2007). Now, (12) has the same diffusion coefficient as the true conditioned diffusion and therefore the law of the true conditioned process is absolutely continuous with respect to that of (12); see Delyon & Hu (2006) for a rigorous proof.

The Gibbs sampler with block updating then has the following algorithmic form:

1. Initialise all unknowns. Use linear interpolation to initialise the X^i. Set $s := 1$.

2. Update $\mathbf{X}_{(s)}|c_{(s-1)}, D_n$ as follows. For $j = 0, m, \ldots, n-m$:

   2.1 Propose $X_*^{i+1} \sim p\left(X_*^{i+1}|X_*^i, X^M, c\right)$ for $i = j, \ldots, M-2$.

   2.2 Accept and store the move with probability given by (11) otherwise store the current value of the chain.

3. Draw $c_{(s)} \sim p\left(\cdot\,|\mathbf{X}_{(s)}\right)$.

4. If the desired number of simulations have been performed then stop, otherwise, set $s := s + 1$ and return to step 2.

Whilst this block updating method helps to overcome the dependence within the latent process conditional on the model parameters, it does not overcome the more fundamental convergence issue, which we now outline in detail.

3.1.3 Convergence Issues

As the discretization gets finer, that is, as m increases, it is possible to make very precise inference about the diffusion coefficient of the process via the quadratic variation. Consider a complete data sample path $\mathbf{X}$ on $[0, T]$. Then $\mathbf{X}$ gives the integral of the diffusion coefficient through the quadratic variation

$$[\mathbf{X}]^2(T) = \int_0^T \beta\left(X_t, c\right)\, dt\,.$$

This means that c can be deduced from $\mathbf{X}$ and consequently a scheme which imputes $\mathbf{X}|D_n$ and then updates c will be reducible; since $\mathbf{X}$ confirms c and c is in turn determined by the quadratic variation, the scheme will not converge. This dependence (between the quadratic variation and diffusion coefficient) was highlighted as a problem by Roberts & Stramer (2001) and results in long mixing times of MCMC algorithms such as the single site Gibbs sampler, though the problem is less noticable for $m \leq 5$. The latter authors overcome this dependence in the context of univariate diffusions by transforming the missing data, giving a partially non-centred parametrisation which leads to an irreducible algorithm even in the limit as $m \to \infty$. However, for a u-dimensional diffusion satisfying (4), finding such a transformation requires an invertible function, $g : \mathbf{R}^u \to \mathbf{R}^u$ such that

$$\nabla g(\nabla g)' = \beta^{-1}\,.$$

This equation is almost always impossible to solve in practice for general nonlinear multivariate diffusions such as those considered here.

Attention is therefore turned to the Gibbs strategy of Golightly & Wilkinson (2008) which can easily be implemented for any nonlinear multivariate diffusion and does not suffer from the convergence problems of the single site or naive block Gibbs samplers; in essence, by alternatively sampling from the posterior of

parameters and the driving Brownian motion process (rather than the actual data), the dependence between c and the latent data can be overcome. The idea is motivated by the "innovation" scheme of Chib et al. (2006); however, the algorithm considered here can be applied to any partially observed diffusion process (that may be subject to measurement error — see Section 3.2 for a discussion).

3.1.4 The Innovation Scheme

Corresponding to the skeleton path $\mathbf{X}$, given by $\mathbf{X} = (X^0, X^1, \ldots, X^n)$, is a skeleton path of W_t, the driving Brownian process. This skeleton is denoted by $\mathbf{W} = (W^0, W^1, \ldots, W^n)$. Note that under any discrete approximation of (4), there is a one-to-one relationship between $\mathbf{X}$ and $\mathbf{W}$, conditional on the parameter vector c. Therefore, rather than sample the distribution $c, \mathbf{X}|D_n$, the innovation scheme samples $c, \mathbf{W}|D_n$ by alternating between draws of c conditional on the data and $\mathbf{W}$, and $\mathbf{W}$ conditional on c and the data. Hence at every iteration of the algorithm, the skeleton path $\mathbf{X}$ will be consistent with the current parameter value — this is crucial in order to overcome the dependence issue highlighted by Roberts & Stramer (2001).

Algorithmically:-

1. Initialise all unknowns. Set the iteration counter to $s = 1$.
2. Draw $\mathbf{W}_{(s)} \sim p\left(\cdot\,|c_{(s-1)}, D_n\right)$ by updating the latent data, via $\mathbf{X}_{(s)} \sim p\left(\cdot\,|c_{(s-1)}, D_n\right)$.
3. Update parameters by drawing $c_{(s)} \sim p\left(\cdot\,|\mathbf{W}_{(s)}, D_n\right)$.
4. Increment s and return to 2.

By updating the latent data in step 2, $\mathbf{W}$ is obtained deterministically. Note that this step is easily performed by implementing the blocking strategy of Section 3.1.2. This is essentially the innovation scheme of Chib et al. (2006). The intuition behind it is that the driving Brownian motion, $\mathbf{W}$, contains no information about the model parameters, and in particular, that the quadratic variation of $\mathbf{W}$ does not determine any of the parameters. Therefore conditioning on $\mathbf{W}$ in a Gibbsian update of the model parameters will not cause the full-conditional to degenerate. The SDE defining the stochastic process can be regarded as a deterministic function $\mathbf{X} = f(\mathbf{W}, c)$ which can be inverted to recover $\mathbf{W}$ from $\mathbf{X}$ as necessary. The problem with this scheme is that a proposed new c^* implies a new sample path $\mathbf{X}^* = f(\mathbf{W}, c^*)$, and in general, this sample path will be far from the observed data, leading to small MH acceptance probabilities. The key insight described in Golightly & Wilkinson (2008) is the realisation that there is no fundamental requirement that the change of variable be directly related to the actual diffusion process, but can be any deterministic transformation $\mathbf{X} = f^*(\mathbf{W}, c)$. Further, that whilst most choices of $f^*(\cdot, \cdot)$ will be ineffective at uncoupling the diffusion parameters from the latent sample path, $\mathbf{W}$, any transformation corresponding to a diffusion process locally equivalent to the true (conditioned) diffu-

sion will be, and hence the modified diffusion bridge (MDB) construction can be used again, in a different way, for constructing an efficient parameter update.

Consider the task of performing step 3 to obtain a new c. We map between $\mathbf{X}$ and $\mathbf{W}$ by using the Wiener process driving the MDB construct in (9) as the effective component to be conditioned on. We have that

$$X^{i+1} = X^i + \mu^*\left(X^i\right)\,\Delta t + \sqrt{\beta^*\left(X^i, c\right)}\left(W^{i+1} - W^i\right)\,, \tag{13}$$

where $\left(W^{i+1} - W^i\right) \sim \mathcal{N}(0, \Delta t)$, μ^* and β^* are given by (10). Re-arrangement of (13) gives

$$\Delta W^i \equiv W^{i+1} - W^i = \left[\beta^*\left(X^i, c\right)\right]^{-\frac{1}{2}}\left(X^{i+1} - X^i - \mu^*\left(X^i\right)\,\Delta t\right)\,, \tag{14}$$

and we define this relation for $i = j, j+1, \ldots, j+m-2$ where $j = 0, m, \ldots, n-m$. This defines a map between the latent data $\mathbf{X}\backslash\{D_n\}$ and $\mathbf{W}$. Note that it is not necessary to map between the actual data D_n and the corresponding points in $\mathbf{W}$ as we update the parameters c in step 3 conditional on $\mathbf{W}$ *and* the data D_n.

Whereas naive global MCMC schemes sample

$$p\left(c\,|\mathbf{X}\right) \propto p\left(c\right) p\left(\mathbf{X}\,|c, D_n\right)\,,$$

the innovation scheme samples

$$p\left(c\,|\mathbf{W}, D_n\right) \propto p\left(c\right) p\left(f^*(\mathbf{W}, c)\,|c, D_n\right)\,J \tag{15}$$

where $\mathbf{X} = f^*(\mathbf{W}, c)$ is the transformation defined recursively by (13) and J is the Jacobian associated with the transformation. To compute J, consider the transformation of a particular value X^{i+1} at time t_{i+1} in the interval (t_j, t_M). By definition,

$$J = |\partial X^{i+1}/\partial W^{i+1}| = |\beta^*\left(X^i, c\right)|^{\frac{1}{2}}$$

using (13). Now, writing $\mu^*\left(X^i\right) = \mu_i^*$ and $\beta^*\left(X^i, c\right) = \beta_i^*$ we note that for fixed $\mathbf{W}$, the density associated with the modified diffusion bridge construct is

$$\begin{aligned}
\tilde{p}\left(X^{i+1}|X^i, X^M, c\right) &\propto |\beta_i^*|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(\Delta X^i - \mu_i^*\Delta t)'(\beta_i^*\Delta t)^{-1} \\
&\quad\times\ (\Delta X^i - \mu_i^*\Delta t)\Big\} \\
&= |\beta_i^*|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(\Delta W^i)'(\Delta W^i)\Big\} \\
&\propto J^{-1}\,.
\end{aligned}$$

Hence, the Jacobian associated with the transformation from $\mathbf{X}\backslash\{D_n\}$ to $\mathbf{W}$ is

$$J\left(\mathbf{X}, c\right) \propto \prod_j \prod_{i=j}^{M-2} \tilde{p}\left(X^{i+1}|X^i, X^M, c\right)^{-1}$$

where $j = 0, m, \ldots, n-m$ and we express the dependence of J on $\mathbf{X}$ and c explicitly. Now, we sample the target density (15) via a MH step. A proposed

new c_* is simulated from a suitable proposal density $g(\cdot)$ which may depend on $\mathbf{X}$, $\mathbf{W}$ and the current c. It is found here that the Gaussian random walk update of Golightly & Wilkinson (2005) works well. Note that for each new c_*, we obtain a new skeleton path

$$\mathbf{X}_* = \left(X^0, X_*^1, \ldots, X_*^{m-1}, X^m, X_*^{m+1}, \ldots\ldots, X_*^{n-1}, X^n\right)$$

deterministically, via the transformation $\mathbf{X}_* = f^*(W, c_*)$ defined recursively by (13). Therefore, if the current state of the chain is c (and $\mathbf{X}$ correspondingly) then a move to c_* (and $\mathbf{X}_*$) is accepted with probability

$$\begin{aligned}\min\left\{1\,,\,\frac{p\left(c_*\,|\mathbf{W},D_n\right)}{p\left(c\,|\mathbf{W},D_n\right)}\right\} &= \min\left\{1\,,\,\frac{p\left(c_*\right)}{p\left(c\right)}\times\frac{p\left(\mathbf{X}_*\,|c_*,D_n\right)}{p\left(\mathbf{X}\,|c,D_n\right)}\times\frac{J\left(\mathbf{X}_*,c_*\right)}{J\left(\mathbf{X},c\right)}\right\}\\ &= \min\left\{1\,,\,\frac{p\left(c_*\right)}{p\left(c\right)}\times\left[\prod_{i=0}^{n-1}\frac{p\left(X_*^{i+1}|X_*^i,c_*\right)}{p\left(X^{i+1}|X^i,c\right)}\right]\right.\\ &\quad\left.\times\left[\prod_j\prod_{i=j}^{M-2}\frac{\tilde{p}\left(X^{i+1}|X^i,X^M,c\right)}{\tilde{p}\left(X_*^{i+1}|X_*^i,X^M,c_*\right)}\right]\right\}. \qquad (16)\end{aligned}$$

Hence, the innovation scheme can be summarised by the following steps:

1. Initialise all unknowns. Set the iteration counter to $s = 1$.
2. Draw $\mathbf{W}_{(s)} \sim p\left(\cdot\,|c_{(s-1)}, D_n\right)$ by updating the latent data, via $\mathbf{X}_{(s)} \sim p\left(\cdot\,|c_{(s-1)}, D_n\right)$.
3. Update parameters by drawing $c_{(s)} \sim p\left(\cdot\,|\mathbf{W}_{(s)}, D_n\right)$:
   - 3.1 Apply equation (14) to obtain ΔW^i for $i = j, j+1, \ldots, j+m-2$ and $j = 0, m, \ldots, n-m$.
   - 3.2 Propose a new c_*, for example, by using a Gaussian random walk move.
   - 3.3 Combine the ΔW^i with c_* and apply equation (13) to obtain a new skeleton path $\mathbf{X}_*$ deterministically.
   - 3.4 Accept a move to c_* (and therefore $\mathbf{X}_*$) with probability given by (16).
4. Increment s and return to 2.

Naturally, the algorithm can be extended to the case of partial observation (and measurement error).

3.2 Partial Observation

Suppose now that the process X_t satisfying (4) is not observed directly and only observations on the process

$$Y_t = F'X_t + \varepsilon_t\,, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \Sigma\right) \qquad (17)$$

are available. This flexible setup allows for the case of only observing a subset of components of X_t. For example, if u_1 and u_2 denote the respective number of observable (subject to error) and unobservable components then we set

$$F = \begin{pmatrix} \mathbf{I}_{u_1} \\ \mathbf{0}_{u_2,u_1} \end{pmatrix}$$

where $\mathbf{I}_{u_1}$ is the $u_1 \times u_1$ identity and $\mathbf{0}_{u_2,u_1}$ is the $u_2 \times u_1$ zero matrix. Here, it is assumed for simplicity that $\Sigma = \text{diag}\{\sigma_i^2\}$ for $i = 1, \ldots, u_1$ and if these parameters are unknown, we have

$$c = (c_1, \ldots, c_v, \sigma_1, \ldots, \sigma_{u_1})' .$$

Extensions to non-diagonal Σ are straightforward. We now consider the task of applying the innovation scheme to this model.

3.2.1 Updating the Latent Process

We wish to sample $\mathbf{X}_{(s)} \sim p\left(\cdot\,|c_{(s-1)}, D_n\right)$ at some iteration s. Since we assume here that we do not observe the process directly, the entire skeleton path $\mathbf{X}$ must be updated. We therefore implement a different blocking strategy, by updating in blocks of size $2m-1$.

Consecutive observation times t_j, t_M and t_{M^+} where, as usual, j is an integer multiple of m, $M = j + m$ and now $M^+ = M + m = j + 2m$, correspond to the noisy and partial observations, Y^j, Y^M and Y^{M^+}. Treating X^j and X^{M^+} as fixed, the full conditional for $X^{j+1}, \ldots, X^{M^+-1}$ is

$$\begin{aligned} &p\big(X^{j+1}, \ldots, X^{M^+-1}|X^j, Y^M, X^{M^+}, c\big) \\ &\quad\propto\; p\left(Y^M|X^M, c\right) \prod_{i=j}^{M^+-1} p\left(X^{i+1}|X^i, c\right) \end{aligned} \tag{18}$$

where $p(Y^M|X^M, c)$ is $\mathcal{N}(Y^M; X^M, \Sigma)$. By sampling the distribution in (18) for $j = 0, m, \ldots n-2m$, the use of overlapping blocks with free mid-point ensures that an irreducible algorithm is obtained. Hence at iteration s of the block Gibbs sampler, one draws

$$X^{j+1}, \ldots, X^{M^+-1} \sim p\big(X^{j+1}, \ldots, X^{M^+-1}|X^j, Y^M, X^{M^+}, c\big)$$

where X^j is obtained at iteration s and X^{M^+} at iteration $s-1$. We sample this density with a MH step.

Consider initially the task of proposing the first m values, $X^{j+1}, \ldots, X^M$, in the block. Clearly, an efficient method would be to propose X^{i+1} for $i = j, \ldots, M-1$, conditional on c, the end-points of the block X^j and X^{M^+}, *and* the noisy observation at the mid-point Y^M. This can be achieved by sampling from a Gaussian approximation to $p(X^{i+1}|X^i, Y^M, X^{M^+}, c)$, details of which can be found in Golightly & Wilkinson (2008). For the numerical examples considered here however, the Gaussian approximation $\tilde{p}(X^{i+1}|X^i, Y^M, c)$ which is

only conditioned on the mid-point Y^M of the block, works sufficiently well and is less computationally costly to evaluate. We derive $\tilde{p}(X^{i+1}|X^i, Y^M, c)$ by constructing a Gaussian approximation to the joint density of X^{i+1} and Y^M (conditional on X^i and c). We have that

$$\begin{pmatrix} X^{i+1} \\ Y^M \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} X^i + \mu_i \Delta t \\ F'\left(X^i + \mu_i \Delta^-\right) \end{pmatrix}, \begin{pmatrix} \beta_i \Delta t & \beta_i F \Delta t \\ F'\beta_i \Delta t & F'\beta_i F \Delta^- + \Sigma \end{pmatrix} \right\},$$

where $\Delta^- = t_M - t_i$ and again we adopt the shorthand notation $\mu_i = \mu\left(X^i, c\right)$ and $\beta_i = \beta\left(X^i, c\right)$. Conditioning on Y^M yields

$$\tilde{p}\left(X^{i+1}|X^i, Y^M, c\right) = \mathcal{N}\left(X^{i+1}; X^i + a\left(X^i, c\right)\Delta t\,,\, b\left(X^i, c\right)\Delta t\right) \tag{19}$$

where

$$a\left(X^i, c\right) = \mu_i + \beta_i F\,\left(F'\beta_i F\Delta^- + \Sigma\right)^{-1}\,\left(Y^M - F'\left[X^i + \mu_i\Delta^-\right]\right)\,, \tag{20}$$

and

$$b\left(X^i, c\right) = \beta_i - \beta_i F\,\left(F'\beta_i F\Delta^- + \Sigma\right)^{-1}\,F'\beta_i\Delta t\,. \tag{21}$$

Hence, we propose the first m values in the block by recursively drawing $X_*^{i+1} \sim \tilde{p}\left(X_*^{i+1}|X_*^i, Y^M, c\right)$ for $i = j, \ldots, M-1$. Finally we propose the last $m-1$ values of the block; we simulate $X_*^{M+1}, \ldots, X_*^{M^+-1}$ conditional on X_*^M and X^{M^+} by sampling the Gaussian approximation $\tilde{p}(X_*^{i+1}|X_*^i, X^{M^+}, c)$ for each $i = M, \ldots, M^+ - 2$. That is, we use the modified diffusion bridge construct in (9). Now, assuming that at the end of iteration $s-1$ the current value of the chain is $X^{j+1}, \ldots, X^{M^+-1}$, then at iteration s, a move to $X_*^{j+1}, \ldots, X_*^{M^+-1}$ is accepted with probability

$$\begin{aligned} \min\Bigg\{1\,,\, & \frac{p\big(X_*^{j+1}, .., X_*^{M^+-1}|X^j, Y^M, X^{M^+}, c\big)}{p\big(X^{j+1}, .., X^{M^+-1}|X^j, Y^M, X^{M^+}, c\big)} \\ & \times \frac{q\big(X^{j+1}, .., X^{M^+-1}|X^j, Y^M, X^{M^+}, c\big)}{q\big(X_*^{j+1}, .., X_*^{M^+-1}|X^j, Y^M, X^{M^+}, c\big)}\Bigg\} \end{aligned} \tag{22}$$

where we denote by $q(\cdot\,|X^j, Y^M, X^{M^+}, c)$ the density associated with the proposal process for the block update.

3.2.2 Updating the Parameters

For the case of partial data (and subject to error), the innovation scheme samples

$$p\left(c\,|\mathbf{W}, D_n\right) \propto p\left(c\right)p\left(f^*(\mathbf{W}, c)\,|c, D_n\right)p\left(D_n|f^*(\mathbf{W}, c), c\right)J\,. \tag{23}$$

As in Section 3.1.4, by fixing the values of $\mathbf{X}$ at the observation times, the MDB construct in (13) can be used to uncouple $\mathbf{W}$ from $\mathbf{X}$. However, when the variance associated with the measurement error density is unknown, using the MDB

construct to map between $\mathbf{X}$ and $\mathbf{W}$ may lead to parameter values which are inconsistent with the current value of the sample path $\mathbf{X}$. We therefore take $\mathbf{W}$ as the skeleton associated with the Wiener process driving the construct in (19) and we use this as the effective component to be conditioned on. We have that

$$X^{i+1} = X^i + a\left(X^i\right)\,\Delta t + \sqrt{b\left(X^i, c\right)}\left(W^{i+1} - W^i\right)\,, \tag{24}$$

where a and b are given by (20) and (21) respectively. Re-arrangement of (24) gives

$$\Delta W^i \equiv W^{i+1} - W^i = \left[b\left(X^i, c\right)\right]^{-\frac{1}{2}}\left(X^{i+1} - X^i - a\left(X^i\right)\,\Delta t\right)\,, \tag{25}$$

and we define this relation for $i = j, j+1, \ldots, j+m-1$ where $j = 0, m, \ldots, n-m$, giving a map between the latent data $\mathbf{X}$ and $\mathbf{W}$. It can be shown (Golightly & Wilkinson 2008) that the Jacobian associated with this transformation is

$$J(\mathbf{X}, c) \propto \left(\prod_{i=0}^{n-1} \tilde{p}\left(X^{i+1}|X^i, Y^{(\lfloor i/m+1 \rfloor)m}, c\right)\right)^{-1} \tag{26}$$

where $\lfloor x \rfloor$ denotes the integer part of x and we write the dependence of J on $\mathbf{X}$ and c explicitly. Note that as before, for each new c_*, we obtain a new skeleton path $\mathbf{X}_*$ deterministically, via the transformation $\mathbf{X}_* = f^*(W, c_*)$ defined recursively by (24). Hence a move to c_* (drawn from a symmetric proposal density $g(\cdot)$) and $\mathbf{X}_*$ is accepted with probability

$$\min\left\{1\,,\ \frac{p\left(c_*\right)}{p\left(c\right)} \times \frac{p\left(\mathbf{X}_*\,|c_*, D_n\right)}{p\left(\mathbf{X}\,|c, D_n\right)} \times \frac{p\left(D_n|\mathbf{X}_*, c_*\right)}{p\left(D_n|\mathbf{X}, c\right)} \times \frac{J\left(\mathbf{X}_*, c_*\right)}{J\left(\mathbf{X}, c\right)}\right\} \tag{27}$$

An appropriate algorithm for the case of noisy and partial data is given by the following steps:

1. Initialise all unknowns. Set the iteration counter to $s = 1$.
2. Draw $\mathbf{X}_{(s)} \sim p\left(\cdot\,|c_{(s-1)}, D_n\right)$ as follows. For $j = 0, m, \ldots, n-2m$:
   - 2.1 Propose $X_*^{i+1} \sim p\left(X_*^{i+1}|X_*^i, Y^M, c\right)$ for $i = j, \ldots, M-1$.
   - 2.2 Propose $X_*^{i+1} \sim p\left(X_*^{i+1}|X_*^i, X^{M^+}, c\right)$ for $i = M, \ldots, M^+ - 2$.
   - 2.3 Accept and store the move with probability given by (22) otherwise store the current value of the chain.
3. Update parameters by drawing $c_{(s)} \sim p\left(\cdot\,|\mathbf{W}_{(s)}, D_n\right)$:
   - 3.1 Apply equation (25) to obtain ΔW^i for $i = j, j+1, \ldots, j+m-1$ and $j = 0, m, \ldots, n-m$.
   - 3.2 Propose a new c_*, for example, by using a Gaussian random walk move.

3.3 Combine the ΔW^i with c_* and apply equation (24) to obtain a new skeleton path $\mathbf{X}_*$ deterministically.

3.4 Accept a move to c_* (and therefore $\mathbf{X}_*$) with probability given by (27).

4. Increment s and return to 2.

4 Inference for Prokaryotic Auto-regulation

To illustrate the proposed sampling strategy, we apply the innovation scheme to the auto-regulatory gene network characterised by the SDE given in Section 2.3. We consider two synthetic datasets; $\mathcal{D}_1$ and $\mathcal{D}_2$. In $\mathcal{D}_1$ we have 50 observations of $X_t = (g_t, r_t, p_t, (p_2)_t)'$, simulated at integer times via the Gillespie algorithm. True values for the stochastic rate constants $(c_1, \ldots, c_8)'$ were taken as in Golightly & Wilkinson (2005), namely 0.1, 0.7, 0.35, 0.2, 0.1, 0.9, 0.3 and 0.1. We assume further that the conservation constant (that is, the number of copies of the gene on the genome) is known to be 10. Finally, we consider $\mathcal{D}_2$ constructed by taking $\mathcal{D}_1$ and discarding the observations on g. The remaining observations on r, p and p_2 were also subjected to error by perturbing each value with a zero-mean Gaussian random variable with a known, common variance of $\sigma^2 = 2$.

For $\mathcal{D}_1$, the innovation scheme was run for 1×10^6 iterations, with 1×10^5 iterations discarded as burn-in. Thinning of the output was employed to leave 9000 posterior draws. For the partially observed dataset $\mathcal{D}_2$, we consider two scenarios by assuming first that the variance of the measurement error σ^2 is known and finally, that σ^2 is unknown. As both partially observed scenarios present the algorithm with the challenge of mixing over the uncertainty associated with the unobserved component, a longer run of 3×10^6 iterations was used. After discarding a number of iterations as burn-in and thinning the output, a sample of 9000 draws with low auto-correlations was obtained. For each scenario, discretization was set by taking $m = 10$ (and $\Delta t = 0.1$) giving 9 latent values between every pair of observations. We note that by increasing m, discretization bias can be reduced, however, computational cost increases. It is found here that there is little difference between results for $m = 10$ and $m = 20$. Hence, we report only those results obtained for $m = 10$. Independent proper Uniform $U(-5, 5)$ priors were taken for $\log(c_i)$, $i = 1, \ldots, 8$ and for the partially observed case, a Uniform $U(0, 10)$ prior was taken for the initial gene copy number, g_0.

Parameter posteriors for each scenario are summarised in Table 1 and Figure 1. Consider the fully observed dataset D_1. Clearly, the sampler produces estimates that are consistent with the true values of the rate constants. Note in particular that although estimates of c_5 and c_6 (corresponding to the rates of the reversible dimerisation reactions) are relatively imprecise, we recover the value of c_5/c_6 (that is, the propensity for the forwards reaction) fairly well. Similar results are obtained for the rates c_1 and c_2 corresponding to the reversible repression reactions. Running the innovation scheme (coded in C) for 1×10^6 iterations on D_1 took 400 minutes on a Pentium IV 3.0GHz processor. Despite a short run

Parameter	True Value	Mean (Standard Deviation)		
		D_1 : $g_t \cup r_t \cup p_t \cup p_{2,t}$	D_2 : $r_t \cup p_t \cup p_{2,t} \cup \sigma$	D_2 : $r_t \cup p_t \cup p_{2,t}$
c_1	0.1	0.078 (0.022)	0.029 (0.019)	0.018 (0.016)
c_2	0.7	0.612 (0.174)	0.205 (0.151)	0.117 (0.143)
c_1/c_2	0.143	0.128 (0.019)	0.182 (0.131)	0.577 (0.741)
c_3	0.35	0.363 (0.095)	0.383 (0.218)	0.197 (0.252)
c_4	0.2	0.236 (0.052)	0.036 (0.046)	0.140 (0.149)
c_5	0.1	0.070 (0.024)	0.070 (0.038)	0.054 (0.024)
c_6	0.9	0.680 (0.231)	0.675 (0.298)	0.531 (0.256)
c_5/c_6	0.111	0.104 (0.014)	0.130 (0.198)	0.113 (0.083)
c_7	0.3	0.299 (0.076)	0.290 (0.142)	0.147 (0.189)
c_8	0.1	0.138 (0.030)	0.028 (0.027)	0.061 (0.072)
σ	1.414	—	—	1.825 (0.307)

Table 1: Posterior means and standard deviations for parameters estimated using 2 length-50 datasets (D_1 and D_2) from the output of the innovation scheme.

and a discretization with $m = 10$, the trace and autocorrelation plots in Figure 1 show that the chain mixes well with autocorrelations reducing fairly quickly.

Naturally, estimates of the rate constants obtained using the partial datasets are far less precise. Note also that we see a marked decrease in accuracy when the measurement error variance is unknown.

5 Conclusions

This chapter has provided an overview of methods for conducting fully Bayesian inference for the rate parameters governing single-cell stochastic kinetic models using (noisy, partial, discrete) time course data. The method exploits a diffusion approximation to the model, the CLE, as this renders the computational problem more tractable. Although presented in the context of single-cell stochastic kinetic models, the inference techniques are very general, and can be applied to essentially any SDE model with associated time course data. Although the techniques are computationally intensive, the information that they provide about the parameters and the extent to which they are identified by the data is extremely rich, making the necessary CPU-time well worth spending.

References

Arkin, A., Ross, J. & McAdams, H. H. (1998), 'Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells', *Genetics* **149**, 1633–1648.

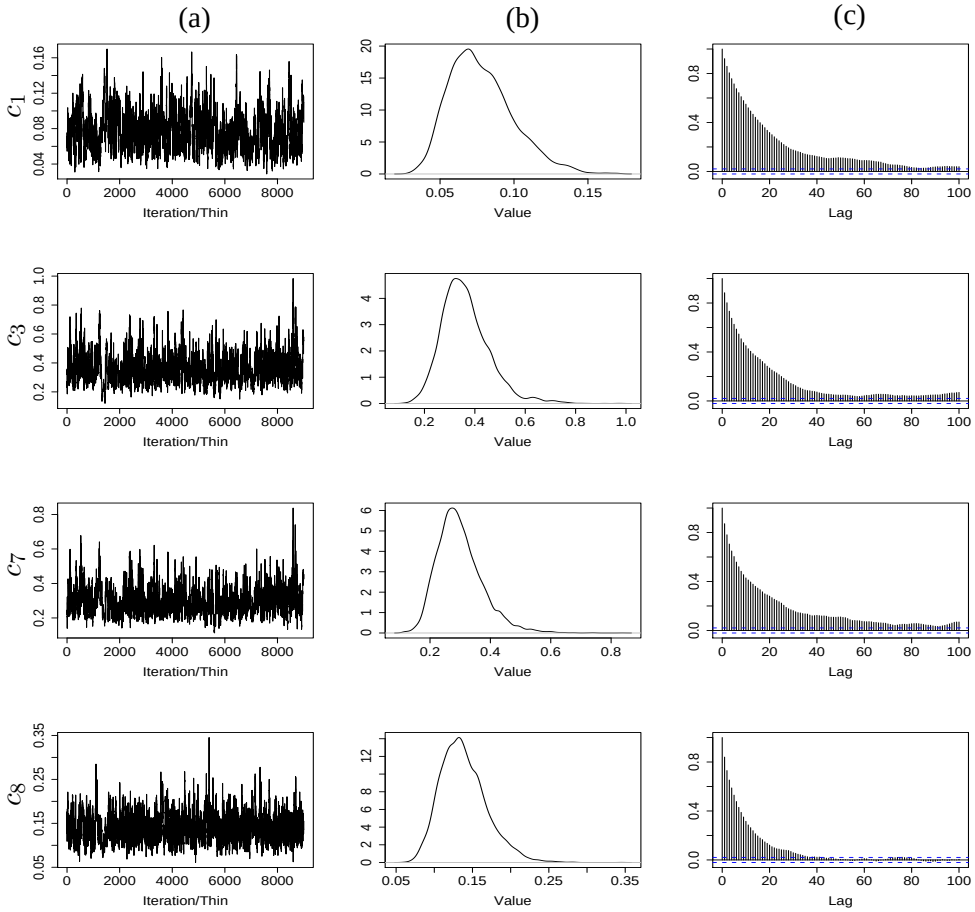

Figure 1: (a) Trace, (b) density and (c) auto-correlation plots for a random selection of c from the output of the innovation scheme using 50 observations ($\mathcal{D}_1$) and $m = 10$.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. & Fearnhead, P. (2006), 'Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes', *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **68**, 1–29.

Bibby, B. M. & Sørensen, M. (1995), 'Martingale estimating functions for discretely observed diffusion processes', *Bernouilli* **1**, 17–39.

Boys, R. J., Wilkinson, D. J. & Kirkwood, T. B. L. (2008), 'Bayesian inference for a discretely observed stochastic kinetic model', *Statistics and Computing* **18**. In press.

Chib, S., Pitt, M. K. & Shephard, N. (2006), 'Likelihood based inference for diffusion driven models', *In submission* .

Delyon, B. & Hu, Y. (2006), 'Simulation of conditioned diffusion and application to parameter estimation', *Stochastic Processes and thier Applications* **116**, 1660–1675.

Durham, G. B. & Gallant, R. A. (2002), 'Numerical techniques for maximum likelihood estimation of continuous time diffusion processes', *Journal of Business and Economic Statistics* **20**, 279–316.

Elerian, O., Chib, S. & Shephard, N. (2001), 'Likelihood inference for discretely observed non-linear difusions', *Econometrika* **69**(4), 959–993.

Eraker, B. (2001), 'MCMC analysis of diffusion models with application to finance', *Journal of Business and Economic Statistics* **19**(2), 177–191.

Gillespie, D. T. (1977), 'Exact stochastic simulation of coupled chemical reactions', *Journal of Physical Chemistry* **81**, 2340–2361.

Gillespie, D. T. (1992), 'A rigorous derivation of the chemical master equation', *Physica A* **188**, 404–425.

Gillespie, D. T. (2000), 'The chemical Langevin equation', *Journal of Chemical Physics* **113**(1), 297–306.

Golightly, A. & Wilkinson, D. J. (2005), 'Bayesian inference for stochastic kinetic models using a diffusion approximation', *Biometrics* **61**(3), 781–788.

Golightly, A. & Wilkinson, D. J. (2008), 'Bayesian inference for nonlinear multivariate diffusion models observed with error', *Computational Statistics and Data Analysis* **52**(3), 1674–1693.

McAdams, H. H. & Arkin, A. (1997), 'Stochastic mechanisms in gene expression', *Proceedings of the National Acadamy of Science USA* **94**, 814–819.

Øksendal, B. (1995), *Stochastic differential equations: An introduction with applications*, 6th edn, Springer-Verlag, Berlin Heidelberg New York.

Pedersen, A. (1995), 'A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations', *Scandinavian Journal of Statistics* **1995**(22), 55–71.

Roberts, G. O. & Stramer, O. (2001), 'On inference for non-linear diffusion models using Metropolis-Hastings algorithms', *Biometrika* **88**(3), 603–621.

Sørensen, H. (2004), 'Parametric inference for diffusion processes observed at discrete points in time', *International Statistical Review* **72**(3), 337–354.

Stramer, O. & Yan, J. (2007), 'Asymptotics of an efficient monte carlo estimation for the transition density of diffusion processes', *Methodology and Computing in Applied Probability* **9**(4), 483–496.

Tanner, M. A. & Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association* **82**(398), 528–540.

Wilkinson, D. J. (2006), *Stochastic Modelling for Systems Biology*, Chapman & Hall/CRC Press, Boca Raton, Florida.