Bayesian Inference for Stochastic Kinetic Models using a Diffusion Approximation

A. Golightly[*]

School of Mathematics and Statistics, University of Newcastle, Newcastle Upon Tyne
NE1 7RU, UK

and

D. J. Wilkinson

July 28, 2008

SUMMARY. This paper is concerned with the Bayesian estimation of stochastic rate constants in the context of dynamic models of intra-cellular processes. The underlying discrete stochastic kinetic model is replaced by a diffusion approximation (or stochastic differential equation approach) where a white noise term models stochastic behaviour and the model is identified using equispaced time course data. The estimation framework involves the introduction of $m-1$ latent data points between every pair of observations. MCMC methods are then used to sample the posterior distribution of the latent process and the model parameters. The methodology is applied to the estimation of parameters in a prokaryotic auto-regulatory gene network.

KEY WORDS: Bayesian inference, nonlinear diffusion, Markov chain Monte Carlo, missing data, stochastic differential equation

[*] *email:* a.golightly@ncl.ac.uk

1. Introduction

The standard approach for modeling biochemical networks is to derive ordinary differential equations (ODEs) using the law of mass action and the concentrations of each species. Such an approach, however, assumes that the time evolution of a system is continuous and deterministic. In reality, chemical reactions occur as discrete events as a result of molecular collisions which are impossible to predict with certainty (Gillespie, 1977). Furthermore, while in many cases a deterministic approach can be implemented to a satisfactory degree of accuracy, for many important intra-cellular processes, populations of molecules can be small and stochastic effects become important (McAdams and Arkin, 1999).

In order to perform analysis and simulate a stochastic biochemical network model, it is essential that each parameter regarding the network is obtained (Kitano, 2001). This gives rise to the problem of whether it is possible to start with observed time course data and obtain the rates of each reaction that produced the data. This is known as reverse engineering (see Bower and Bolouri (2000) for a complete disscussion of the problem).

There are three commonly used types of stochastic Markov process models used to simulate biochemical networks: 1) discrete models commonly solved by the Gillespie algorithm (Gillespie, 1977) or an extension of it (Stundzia and Lumsden, 1996), 2) diffusion or stochastic differential equation (SDE) models in which the variables are approximated as continuous and a white noise term models stochastic behaviour (Doraiswamy and Kulkarni, 1987) and 3) hybrid models where some chemical species are treated as discrete and others are treated with a continuous approximation. The second method can be regarded as an approximation to the first, where the numbers of molecules are treated as continuous. It is this second method that we will use as the basis of our inference algorithm. However, as we shall see, although the diffusion approximation is usually inadequate for simulation purposes, it appears to be often quite satisfactory to be used as the basis of a Bayesian inference algorithm.

In the context of likelihood, estimation of the parameters requires knowledge of the Markovian transition density for the underlying SDE. However, as analytic solutions of SDE's are rarely available, we are not able to obtain the transition densities in closed form. As observations are available at discrete times and the model is formulated in continuous time, it is natural to work with a discretized version of the SDE known as the Euler approximation. Unfortunately the inter-observation times are usually too large to be used as a time step for the Euler approximation.

In this paper we treat this problem by adopting an idea previously persued by Pedersen (1995). That is, the observed low-frequency data is augmented with the introduction of $m-1$ latent data points in between every pair of measurements. Whereas Pedersen uses a simulated maximum likelihood estimation approach, we use a Markov chain Monte Carlo (MCMC) algorithm to sample the posterior distribution of the latent data and the model parameters. We note that this strategy has been used previously by Eraker (2001) and Kim, Shephard and Chib (1998) in their work with Stochastic Volatility models in finance.

The structure of this paper is as follows. In Section 2, methods for modelling stochastic kinetics are described; Section 2.1 outlines the molecular approach to kinetics, Section 2.2 describes the formulation using a continuous time Markov process model, and Section 2.3 gives the diffusion approximation. In Section 3 we describe inference for non-linear diffusion models. An illustrative application is presented in Section 4, before conclusions are drawn in Section 5.

2. Stochastic Kinetics

2.1 *Molecular Approach to Kinetics*

Consider a bi-molecular reaction of the form, $Y_1 + Y_2 \longrightarrow Y_3$. This reaction will occur when a molecule of Y_1 collides with a molecule of Y_2 whilst molecules move around randomly, driven by Brownian motion. Considering a pair of such molecules in a small, fixed volume and assuming thermal equilibrium, Gillespie (1992) has shown that the hazard of molecules colliding is constant. We also assume the law of mass action such that if the numbers of molecules of each reactant are Y_1 and Y_2 then the hazard of the above reaction occurring would be proportional to $Y_1 Y_2$.

In this paper we will consider a system of reactions involving k species $Y_1,\ Y_2, \ldots, Y_k$ and r reactions $R_1,\ R_2, \ldots, R_r$ in thermal equilibrium inside some fixed volume V. The system will take the form

$$\begin{array}{rccc} R_1: & u_{11}Y_1 + u_{12}Y_2 + \ldots + u_{1k}Y_k & \longrightarrow & v_{11}Y_1 + v_{12}Y_2 + \ldots + v_{1k}Y_k \\ R_2: & u_{21}Y_1 + u_{22}Y_2 + \ldots + u_{2k}Y_k & \longrightarrow & v_{21}Y_1 + v_{22}Y_2 + \ldots + v_{2k}Y_k \\ \vdots & \vdots & \vdots & \vdots \\ R_r: & u_{r1}Y_1 + u_{r2}Y_2 + \ldots + u_{rk}Y_k & \longrightarrow & v_{r1}Y_1 + v_{r2}Y_2 + \ldots + v_{rk}Y_k \end{array} \tag{1}$$

where, u_{ij} is the stoichiometry associated with the j^{th} reactant of the i^{th} reaction and v_{ij} is the stoichiometry associated with the j^{th} product of the i^{th} reaction. Each reaction, R_i, has a stochastic rate constant, c_i, and a rate law or hazard, $h_i(Y, c_i)$, where $Y = (Y_1, Y_2, \ldots, Y_k)'$ is the current state of the system and each hazard is determined by the order of reaction R_i under an assumption of mass action kinetics. Note that for transparency, we denote by Y_i both the species and the number of molecules it represents in the system.

We may represent (1) more compactly as $\mathrm{U}Y \longrightarrow \mathrm{V}Y$, where $U = (u_{ij})$ and $V = (v_{ij})$ are $r \times k$ dimensional matrices (obtained from the stoichiometry of the system). Now consider reaction i and species j. When reaction i occurs, the number of molecules of Y_j will decrease by u_{ij} and increase by v_{ij} giving an overall change of $a_{ij} = v_{ij} - u_{ij}$. The reaction network can then be represented by the net effect reaction matrix $\mathrm{A} = \mathrm{V} - \mathrm{U}$, examples of which are given in Section 2.4.

2.2 *Continuous Time Markov Process Model*

Stochastic models for cellular processes are now reasonably well developed and are traditionally based on techniques for solving the "chemical master equation". The main element of the master equation is the function, $P(Y_1, Y_2, \ldots, Y_k; t)$ which gives the probability that there will be at time t (in a fixed volume, V) $Y_1, Y_2, \ldots, Y_k$ molecules of each respective species. Once this function is obtained, a fairly complete characterization of the state of the system at time t is apparent.

The master equation can be derived for any particular reaction network by using standard probability theory to write $P(Y_1, Y_2, \ldots, Y_k; t + \Delta t)$ as the sum of the probabilities of the number of ways in which the network can arrive in state $(Y_1, Y_2, \ldots, Y_k)'$ at time $t + \Delta t$ (Gillespie, 1977):

$$P(Y; t + \Delta t) = \sum_{i=1}^{r} h_i(Y - \mathrm{A}_i, c_i) P(Y - \mathrm{A}_i; t)\Delta t + \left\{ 1 - \sum_{i=1}^{r} h_i(Y, c_i)\Delta t \right\} P(Y; t) \quad (2)$$

where Y is the state of the system at time t and A_i denotes the i^{th} row of the net effect matrix A. Intuitively, the term $h_i(Y - \mathrm{A}_i, c_i)P(Y - \mathrm{A}_i; t)\Delta t$ is the probability that the system is one R_i reaction removed from state Y at time t and then undergoes such a reaction in $(t, t + \Delta t)$. The second quantity in (2) is the probability that the system undergoes no reactions in $(t, t + \Delta t)$. We now observe that (2) leads to the master equation

$$\frac{\partial}{\partial t} P(Y; t) = \sum_{i=1}^{r} \left\{ h_i(Y - \mathrm{A}_i, c_i) P(Y - \mathrm{A}_i; t) - h_i(Y, c_i) P(Y; t) \right\} . \quad (3)$$

For further details of the master equation formalism in chemical kinetics, good reviews have been given by van Kampen (2001) and Doraiswamy and Kulkarni (1987). Although the master equation is exact, it is only tractable for a handful of cases. The exactly solvable cases have been summarised by McQuarrie (1967). Therefore, stochastic models are typically examined using discrete event simulation algorithms which we briefly summarise here.

In a given system with r reactions, we know that the hazard for a type i reaction is $h_i(Y, c_i)$, so the hazard for a reaction of some type is

$$h_0(Y, \Theta) \equiv \sum_{i=1}^{r} h_i(Y, c_i)$$

where $\Theta = (c_1, c_2, \ldots, c_r)'$. Consequently, the time to the next reaction is $\mathrm{Exp}(h_0(Y, \Theta))$, and this reaction will be a random type, picked with probabilities proportional to the $h_i(Y, c_i)$. Hence, when a reaction occurs, it will be i with probability $h_i(Y, c_i)/h_0(Y, \Theta)$. Samples from the process can therefore be simulated using standard discrete event simulation techniques. The algorithm was developed in the context of chemical kinetics by Gillespie (Gillespie, 1977) and is known in the physical sciences as the "Gillespie algorithm". This algorithm is rigorous in that it provides an exact sample from the corresponding master equation and is well suited to the study of systems in which reactant populations are small, and the Master equation is analytically intractable.

It should be noted that although the Gillespie algorithm is effective for direct simulation, inference for "exact" stochastic-kinetic models is computationally problematic for models of realistic size and complexity (Boys et al., 2004). We therefore introduce the diffusion approximation which though often inadequate for simulation, can be satisfactory for inferential purposes.

2.3 *The Diffusion Approximation*

2.3.1 The Fokker-Planck Equation Typically, stochastic noise terms are introduced in either an ad-hoc manner, or derived, with approximations, from the underlying master equation. Indeed the Fokker-Planck equation can be regarded as a continuous approximation of the master equation. By assuming that the jumps of the Markov process governed by (3) are "small" and that the solution, $P(Y;t)$, varies slowly with Y, we can expand the first term in (3) by means of a second order Taylor expansion to give the Fokker-Planck equation (van Kampen, 2001). Formally, for a k dimensional process $Y(t)$ with components $Y_1(t),\ldots,Y_k(t)$ the nonlinear Fokker-Planck equation is given by,

$$\frac{\partial}{\partial t}P(Y;t) = -\sum_{i=1}^{k}\frac{\partial}{\partial Y_i}\{\mu_i(Y)P(Y;t)\} + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial^2}{\partial Y_i \partial Y_j}\{\beta_{ij}(Y)P(Y;t)\}\,, \qquad (4)$$

where we define the infinitesimal means for $i = 1,\ldots,k$ by

$$\mu_i(Y) = \lim_{\Delta t \to 0}\frac{1}{\Delta t}\mathrm{E}[\{Y_i(t+\Delta t) - Y_i(t)\}|Y(t) = Y] \qquad (5)$$

and the infinitesimal second moments for $i, j = 1,\ldots,k$ by

$$\beta_{ij}(Y) = \lim_{\Delta t \to 0}\frac{1}{\Delta t}\mathrm{Cov}[\{Y_i(t+\Delta t) - Y_i(t)\},\{Y_j(t+\Delta t) - Y_j(t)\}|Y(t) = Y]\,. \qquad (6)$$

The Itô diffusion corresponding to (4) is then obtained as

$$dY(t) = \mu(Y)dt + \beta^{\frac{1}{2}}(Y)dW(t)$$

where $\mu(Y)$ is the column vector of $\mu_i(Y)$ (known as drift), $\beta^{\frac{1}{2}}(Y)$ is any matrix satisfying $\beta^{\frac{1}{2}}(\beta^{\frac{1}{2}})' = [\beta_{ij}(Y)] = \beta(Y)$ (known as the diffusion matrix) and $dW(t) = (dW_1(t),\ldots,dW_k(t))'$ is the increment of (standard, k dimensional) Brownian motion.

If the physics of some system suggests that Y should be (approximately) a Markov process then we choose small Δt such that Y cannot change much during this time (but large enough for the Markov assumption to apply). We then compute (5) and (6) to obtain the diffusion approximation (which is sometimes referred to as the Langevin approach).

2.3.2 Calculating the Diffusion Approximation It is clear that due to the assumption of constant reaction hazard, the number of reactions (of a given type) occuring in a sufficiently short time interval will be approximately Poisson distributed (independently of other reaction types).

Suppose at time t, the state of the system is $Y(t) = (Y_1(t), \ldots, Y_k(t))' = Y$ so that the hazards of $R_1, R_2, \ldots, R_r$ are $h_1(Y, c_1), h_2(Y, c_2), \ldots, h_r(Y, c_r)$. Let N_i denote the number of type i reactions occurring in the interval $(t, t + \Delta t]$. Then for "small" time Δt, $N_i \approx \text{Poisson}(h_i(Y, c_i)\Delta t)$ and the change in the number of molecules of Y_j is given by

$$Y_j(t + \Delta t) - Y_j(t) = a_{1j}N_1 + a_{2j}N_2 + \ldots + a_{rj}N_r \,. \tag{7}$$

For each increment $Y_j(t + \Delta t) - Y_j(t)$, $j = 1, \ldots, k$ given by (7), we calculate the infinitesimal means and variances through straightforward application of (5) and (6) to obtain the SDE

$$dY(t) = \mu(Y, \Theta)\, dt + \beta^{\frac{1}{2}}(Y, \Theta)\, dW(t) \tag{8}$$

with drift and diffusion functions,

$$\mu(Y, \Theta) = \mathrm{A}' h(Y, \Theta) \;,\; \beta(Y, \Theta) = \mathrm{A}' \mathrm{diag}\{h(Y, \Theta)\}\mathrm{A} \,. \tag{9}$$

Here, μ and β depend explicitly on Y and the parameter vector $\Theta = (c_1, c_2, \ldots, c_r)'$. A is the net effect matrix and $h(Y, \Theta)$ is the column vector of hazards $h_i(Y, c_i)$.

2.4 *Example: Prokaryotic Auto-regulatory Gene Network*

Transcriptional regulation has been studied extensively in both prokaryotic and eukaryotic organisms (see, for example McAdams and Arkin (1999), Latchman (2002) and Ng, Wilkinson, Boys and Kirkwood (2004)). In a simple model of prokaryotic auto regulation, dimers of a protein coded for by a gene repress its own transcription into RNA by binding to a regulatory region upstream of the gene. The transcription of a gene into mRNA is facilitated by an enzyme, RNA-polymerase. The process begins with the binding of this enzyme near the beginning of a gene to a site called a promoter. Following the initial binding, RNA-polymerase travels away from the promoter along the gene, synthesising mRNA as it moves. Transcription is repressed by protein dimers, P_2 which bind to sites on the DNA known as operators. The repression and transcription mechanisms can be represented very simply by the following chemical reactions,

$$\begin{array}{rccc} R_1: & \mathsf{DNA} + \mathsf{P}_2 & \longrightarrow & \mathsf{DNA} \cdot \mathsf{P}_2 \\ R_2: & \mathsf{DNA} \cdot \mathsf{P}_2 & \longrightarrow & \mathsf{DNA} + \mathsf{P}_2 \\ R_3: & \mathsf{DNA} & \longrightarrow & \mathsf{DNA} + \mathsf{RNA} \end{array} \qquad (10)$$

Naturally, (10) is a simplification of the actual repression and transcription mechanisms and can be thought of as a summary of the overall effect of the processes.

We model the binding of a ribosome to the mRNA, the translation of the mRNA and the folding of the resulting polypeptide chain into a functional protein, P with the single reaction

$$R_4: \quad \mathsf{RNA} \quad \longrightarrow \quad \mathsf{RNA} + \mathsf{P} \qquad (11)$$

The reversible dimerisation of this protein is categorised by the forward and backward reactions

$$R_5: \quad 2\mathsf{P} \quad \longrightarrow \quad \mathsf{P}_2 \quad , \quad R_6: \quad \mathsf{P}_2 \quad \longrightarrow \quad 2\mathsf{P} \qquad (12)$$

Finally, the model is completed by mRNA and protein degradation,

$$R_7: \quad \mathsf{RNA} \quad \longrightarrow \quad \emptyset \quad , \quad R_8: \quad \mathsf{P} \quad \longrightarrow \quad \emptyset \qquad (13)$$

Although (10)-(13) offer a simplistic view of the mechanisms involved in gene autoregulation, they do provide sufficient detail to capture the network dynamics. For a detailed discussion of gene regulation see Ptashne (1992) and Latchman (2002).

In order to compute the diffusion approximation for the model given by (10)-(13), we must calculate the net effect reaction matrix, A. We order the species by setting $Y = (\mathsf{RNA}, \mathsf{P}, \mathsf{P}_2, \mathsf{DNA}\cdot\mathsf{P}_2, \mathsf{DNA})'$ and use the stoichiometry of the system to obtain

$$\mathrm{A}' = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 2 & 0 & -1 \\ -1 & 1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{14}$$

Now assume for reaction i a stochastic rate constant of c_i and consider the time evolution of the system as a Markov process with state $Y(t) = Y$ at time t. Reactions 1,3,4,6,7,8 are first order and therefore their hazards can be computed (using the law of mass action) as $c_2\mathsf{DNA}\cdot\mathsf{P}_2$, $c_3\mathsf{DNA}$, $c_4\mathsf{RNA}$, $c_6\mathsf{P}_2$, $c_7\mathsf{RNA}$ and $c_8\mathsf{P}$ respectfully. For the second order reactions R_1 and R_5 we obtain $h_1(Y, c_1) = c_1\mathsf{P}_2\mathsf{DNA}$ and $h_5(Y, c_5) = 0.5c_5\mathsf{P}(\mathsf{P} - 1)$.

Before calculation of $\mu(Y, \Theta)$ and $\beta(Y, \Theta)$ (given by (9)), we note that the net effect matrix A is not of full rank (as the number of molecules of DNA and $\mathsf{DNA}\cdot\mathsf{P}_2$ are deterministically related) and this rank-degeneracy will cause problems for the inference method considered in Section 3. For a general rank-degenerate system, we re-order the columns of A so that the first s columns form a matrix of full rank (s), where s is as large as possible. Now take the first s columns and set this to be the matrix A, so that A is (in general) a subset of columns from the net effect reaction matrix. A is now of dimension $r \times s$ with rank s.

For the net effect matrix given by (14), adding row 4 of A' to row 5 implies

$$\mathsf{DNA}\cdot\mathsf{P}_2 + \mathsf{DNA} = k \tag{15}$$

where k is a conservation constant. Now, we remove row 4 from A' to obtain A' (and therefore A) of full rank. Applying (15) and substituting $k - \mathsf{DNA}$ for $\mathsf{DNA}\cdot\mathsf{P}_2$ reduces our model to one involving just 4 chemical species, $Y = (\mathsf{RNA}, \mathsf{P}, \mathsf{P}_2, \mathsf{DNA})'$ for which the full diffusion approximation is specified by drift, $\mu(Y, \Theta)$,

$$\begin{pmatrix} c_3\mathsf{DNA} - c_7\mathsf{RNA} \\ c_4\mathsf{RNA} + 2c_6\mathsf{P}_2 - c_5\mathsf{P}(\mathsf{P}-1) - c_8\mathsf{P} \\ c_2(k - \mathsf{DNA}) + 0.5c_5\mathsf{P}(\mathsf{P}-1) - c_1\mathsf{P}_2\mathsf{DNA} - c_6\mathsf{P}_2 \\ c_2(k - \mathsf{DNA}) - c_1\mathsf{P}_2\mathsf{DNA} \end{pmatrix}, \tag{16}$$

and diffusion matrix $\beta(Y, \Theta)$ which may be factorised as $\beta(Y, \Theta) = \mathrm{BB}'$ where B' is the 8×4 dimensional matrix,

$$\begin{pmatrix} 0 & 0 & -\sqrt{c_1\mathsf{P}_2\mathsf{DNA}} & -\sqrt{c_1\mathsf{P}_2\mathsf{DNA}} \\ 0 & 0 & \sqrt{c_2(k-\mathsf{DNA})} & \sqrt{c_2(k-\mathsf{DNA})} \\ \sqrt{c_3\mathsf{DNA}} & 0 & 0 & 0 \\ 0 & \sqrt{c_4\mathsf{RNA}} & 0 & 0 \\ 0 & -2\sqrt{0.5c_5\mathsf{P}(\mathsf{P}-1)} & \sqrt{0.5c_5\mathsf{P}(\mathsf{P}-1)} & 0 \\ 0 & 2\sqrt{c_6\mathsf{P}_2} & -\sqrt{c_6\mathsf{P}_2} & 0 \\ -\sqrt{c_7\mathsf{RNA}} & 0 & 0 & 0 \\ 0 & -\sqrt{c_8\mathsf{P}} & 0 & 0 \end{pmatrix}. \tag{17}$$

Note that our parameter vector Θ consists of all stochastic rate constants and is given by $\Theta = (c_1, c_2, \ldots, c_8)'$.

3. Inference for non-linear Diffusion Models

3.1 *Models*

We consider inference for an Itô Diffusion that satisfies a stochastic differential equation of the form given by (8) and assume that the conditions under which the SDE can be solved for $Y(t)$ are satisfied (Øksendal, 1995).

Often, $Y(t)$ will consist of both observable and unobservable components. To deal with this, we define $Y(t) = (X(t), Z(t))'$, where $X(t)$ defines the observable part and $Z(t)$ the unobservable part of the system. Note that $X(t)$ and $Z(t)$ have dimensions d_1 and d_2 respectively and such that $Y(t)$ has dimension $d = d_1 + d_2$. We assume that the process $X(t)$ will be observed at a finite number of times and the objective is to conduct inference for the (unknown) parameter vector Θ on the basis of these partial and discrete observations on $Y(t)$. In practice it is necessary to work with the discretized version of (8), given by the Euler approximation,

$$\Delta Y(t) = \mu(Y(t), \Theta)\Delta t + \beta^{\frac{1}{2}}(Y(t), \Theta))\Delta W(t) \tag{18}$$

where $\Delta W(t)$ is a d dimensional iid $\mathrm{N}(0, I\Delta t)$ random vector.

Now suppose we have measurements $X(\tau_i) = x^i$ at evenly spaced times τ_0, τ_1, ..., τ_T with intervals of length $\Delta^* = \tau_{i+1} - \tau_i$. Then put $\Delta t = \Delta^*/m$ for some positive integer m. By choosing m to be sufficiently large, we can ensure that the discretization bias associated with the Euler approximation is arbitrarily small, but this also introduces the problem of $m - 1$ missing values. We deal with these missing values by dividing the time interval $[\tau_0, \tau_T]$ into $mT + 1$ equidistant points $\tau_0 = t_0 < t_1 < \ldots < t_n = \tau_T$. Altogether we have $d_1T(m-1) + d_2(Tm+1)$ missing values which we substitute with simulations $Y(t_i)$. We refer to the collection of simulated data and observations as the augmented data. Eraker (2001) denotes by $\hat{\mathrm{Y}}$ the $d \times (n+1)$ matrix obtained by stacking all elements of the augmented data, that is

$$\hat{\mathrm{Y}} = \begin{pmatrix} x_1(t_0) & X_1(t_1) & \cdots & x_1(t_m) & X_1(t_{m+1}) & \cdots & x_1(t_n) \\ x_2(t_0) & X_2(t_1) & \cdots & x_2(t_m) & X_2(t_{m+1}) & \cdots & x_2(t_n) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{d_1}(t_0) & X_{d_1}(t_1) & \cdots & x_{d_1}(t_m) & X_{d_1}(t_{m+1}) & \cdots & x_{d_1}(t_n) \\ Z_1(t_0) & Z_1(t_1) & \cdots & Z_1(t_m) & Z_1(t_{m+1}) & \cdots & Z_1(t_n) \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ Z_{d_2}(t_0) & Z_{d_2}(t_1) & \cdots & Z_{d_2}(t_m) & Z_{d_2}(t_{m+1}) & \cdots & Z_{d_2}(t_n) \end{pmatrix}.$$

We now denote by $Y^i \equiv (X^i, Z^i)'$ the i^{th} column of $\hat{\mathrm{Y}}$. Then the joint posterior density is given by

$$\pi(\hat{\mathrm{Y}}, \Theta) \propto \pi(\Theta)\pi(Z^0)\prod_{i=1}^{n} f(Y^i|Y^{i-1}, \Theta), \tag{19}$$

where $\pi(\Theta)$ is the prior density of Θ, $\pi(Z^0)$ is the prior density of Z^0 and

$$f(Y^i|Y^{i-1}, \Theta) = |\beta_{i-1}^{-1}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\Delta Y^i - \mu_{i-1}\Delta t)'(\Delta t\beta_{i-1})^{-1}(\Delta Y^i - \mu_{i-1}\Delta t)\right\} \tag{20}$$

Here, $\Delta Y^i = Y^i - Y^{i-1}$, $\mu_i = \mu(Y^i, \Theta)$ and $\beta_i = \beta(Y^i, \Theta)$. Note that we adopt the notation where π denotes all proper densities, p denotes π in an unnormalized form and f denotes the (unnormalized) transtion density obtained from the Euler discretization.

All conditional densities of interest are now proportional to (19).

3.2 *MCMC Scheme*

We have formulated in (19) the joint posterior for the model parameters as well as observed and unobserved data but real interest will usually be in the distribution $(\Theta, (\hat{\mathrm{Y}} \backslash x_{\mathrm{obs}})|x_{\mathrm{obs}})$ where $x_{\mathrm{obs}} = (x^0, x^m, \ldots, x^{Tm})$ denotes the observed data. As discussed in Tanner and Wong (1987), a good way to sample this distribution is to alternate between simulating the parameters conditional on the augmented data (including the missing data), and simulating from the distribution of the missing data given the observed data and the current state of the model parameters. This sampling procedure (known as data augmentation) generates a Markov chain which has the desired posterior, $(\Theta, (\hat{\mathrm{Y}} \backslash x_{\mathrm{obs}})|x_{\mathrm{obs}})$ as its equilibrium distribution (see Tierney (1994) for an overview of the use of Markov chains for exploring posterior distributions).

MCMC methods for the analysis of diffusion processes have been explored extensively in the economic and financial literature. For univariate diffusions, Roberts and Stramer (2001), Elerian, Chib and Shephard (2001) and Durham and Gallant (2002) employ block updating schemes to simulate the latent data. For general (multivariate) partially observed models, the number of unobservables (missing data and model parameters) can be particularly large. We therefore implement a Gibbs sampler (suggested by Eraker (2001)) which is a particularly convenient way of sampling from high dimensional densities. For nonlinear diffusions, direct sampling of the full conditional distributions (for parameters given all data, and latent data given parameters) is not possible. At each Gibbs step, we therefore use a Metropolis-Hastings (M-H) step. This method is often known in the literature as "Metropolis-within-Gibbs".

The first step in the Gibbs sampler involves simulating the latent data points (conditional on Θ). We follow Eraker's method and simulate each column, Y^i, using a M-H step with proposal density $q(\cdot|Y^{i-1}, Y^{i+1}, \Theta) = \mathrm{N}\left(\frac{1}{2}(Y^{i-1} + Y^{i+1})\,,\, \frac{1}{2}\Delta t \beta(Y^{i-1}, \Theta)\right)$. When i is a multiple of m, we need only simulate the d_2 elements corresponding to Z^i. This is accomplished using a M-H step with proposal density $q(\cdot|Y^{i-1}, Y^{i+1}, \Theta)$ further conditioned on the observation x^i.

The final step in the Gibbs sampler is to simulate Θ conditional on its current state and the augmented data. As Θ consists of stochastic rate constants which must be strictly positive, we set $\lambda_j = \log(c_j)$, $j = 1, \ldots, r$ and assume independent proper Uniform priors for each λ_j. A Metropolis random walk update is used to sample the λ_j in one block. The following algorithm summarises our sampling strategy:

1. Initialize all unknowns. Use linear interpolation to initialise X^i and set $Z^i = 0.0$ for all i. Set g=1.

2. For all $i = 0, 1, \ldots, n$ at iteration g draw Y^i from its full conditional. When i is not a multiple of m we use a M-H step with proposal density $q(\cdot|Y^{i-1}, Y^{i+1}, \Theta)$. If i is a multiple of m, only simulate the d_2 elements, Z^i, using a M-H step with proposal density $q(\cdot|Y^{i-1}, Y^{i+1}, \Theta)$ further conditioned on x^i.

3. Draw $\Theta^{(g)}$ using a M-H step with a Gaussian random walk update (on $\log(\Theta)$).

For full details of the MCMC methods employed here, see Eraker (2001) and Golightly and Wilkinson (2004).

4. Simulation Study: Prokaryotic Auto-regulatory Gene Network

To illustrate the methodology presented in Section 3.2, the MCMC Scheme is applied to the auto regulatory gene network model characterised by the SDE with drift as in (16), and diffusion function as in (17).

Often it may be difficult to measure the activation state of the DNA directly. In this case the observable part of the system is $X(t) = (\mathsf{RNA}(t), \mathsf{P}(t), \mathsf{P}_2(t))'$ and the unobservable part of the reduced system is $Z(t) = \mathsf{DNA}(t)$. Formulating the partially observed model in this way implies that we only know the conservation constant, k (as in (15)) and not the split into DNA and $\mathsf{DNA} \cdot \mathsf{P}_2$. In practice it is reasonable to observe k as it corresponds to the number of copies of the gene on the genome and in Section 4.2 we assume k is known but we do not observe $\mathsf{DNA} \cdot \mathsf{P}_2(t)$ or $\mathsf{DNA}(t)$ at any time t.

4.1 *Results: Fully Observed Model*

We first implement the MCMC scheme given in Section 3.2 for the fully observed case; that is, we assume that we observe $Y(t) = (\mathsf{RNA}(t), \mathsf{P}(t), \mathsf{P}_2(t), \mathsf{DNA}(t))'$ at all times t. We consider 5 equispaced data sets, $D_1, D_2, \ldots, D_5$ each independently simulated on $[0, 50)$ using the Gillespie algorithm to ensure exact simulation. D_1, D_2 and D_3 consist of 50 observations ($\Delta^* = 1$), D_4 contains 100 observations ($\Delta^* = 0.5$) and D_5 consists of 500 observations($\Delta^* = 0.1$). For each data set, the MCMC sampler is run for 1,000,000 iterations, thinned by a factor of 100 and with the first 100,000 being discarded as burn-in. True values for $(c_1, c_2, \ldots, c_8)$ are chosen to be 0.1, 0.7, 0.35, 0.2, 0.1, 0.9, 0.3 and 0.1 and k (the number of copies of the gene), is set to be 10.

[Table 1 about here.]

[Table 2 about here.]

Tables 1-2 summarise the posterior distribution for the fully observed model; Table 1 gives posterior means and standard deviations for Θ estimated from a single MCMC run using the 3 replicate length-50 datasets (D_1, D_2, D_3) and $m = 5$. Table 2 is obtained from a single MCMC run with $m = 2$, 5, 8, 10 and data sets D_1, D_4 and D_5. Replicate MCMC runs, given in Golightly and Wilkinson (2004), suggest that there is little run-to-run variability.

Table 2 demonstrates the clear advantage of including latent variables in the estimation framework. As m increases there is a notable decrease in discretization error. For example c_7, the stochastic rate constant for reaction R_7 (mRNA degradation) has a true value of 0.3 while it is estimated to be 0.269 using D_1 with $m = 2$. However, as m increases to 10 (and Δt reduces from 0.5 to 0.1) we see an increase in accuracy with an estimate of 0.316. Similarly, when using 100 and 500 observations, errors are more pronounced for $m = 2$ and an increase in m gives more precise estimates of parameters though the difference in results for $m = 8$ and $m = 10$ is small.

If we fix m, Table 2 suggests that errors are larger for the smaller data set consisting of 50 observations. For example c_1 has a true value of 0.1 while it is estimated to be 0.066 when using 50 observations and fixing m to be 5. However, as sample size increases to 100 observations we see an estimate of 0.096. Note that when using 50 observations, estimates of c_1, c_2, c_5 and c_6 appear to be quite imprecise. In contrast, the estimates of c_1/c_2 and c_5/c_6 (corresponding to the propensities of reactions R_1 and R_5 repectively), are quite good.

4.2 *Results: Partially Observed Model*

We now apply the MCMC algorithm to the partially observed auto regulatory model. To allow comparison, we use the data sets D_1, D_4 and D_5, (as discussed in Section 4.1) but we assume we only have observations on $X(t) = (\mathsf{RNA}(t), \mathsf{P}(t), \mathsf{P}_2(t))'$

such that $Z(t) = \mathsf{DNA}(t)$ is unobservable at all times $t \geq 0$. Although we do not observe the activation state, we assume that the number of copies of the gene is known to be $k = 10$.

Due to computational demands, discretization is set using $m = 5$ and the sampler is run for 10,000,000 iterations, thinned by a factor of 1000 and with the first 4,000,000 being discarded as burn-in. The resulting parameter estimates for each data set are summarised in Table 3.

[Table 3 about here.]

As in Section 4.1, inspection of Table 3 reveals that errors are larger for all parameters but c_7 when comparing the smaller data set consisting of 50 observations to the 2 remaining larger ones. Although for just 50 observations we learn very little about the true values of c_1, c_2, c_5 and c_6, as with the fully observed model, estimates of c_1/c_2 and c_5/c_6 are far more precise.

5. Discussion

In this paper we have provided a fully Bayesian approach to the estimation of stochastic rate constants governing biochemical reactions. When populations of molecules are small, stochastic effects become important and the deterministic approach is no longer satisfactory. By adopting a diffusion approximation, a white noise term models stochastic behaviour. We are then essentially concerned with the analysis of non-linear, discretely observed stochastic differential equations. We have shown that although the SDE approximation is often not adequate for simulation, it can sometimes be satisfactory when used in the context of Bayesian inference. This suggests that whilst both discreteness and stochasticity are important for biochemical network simulation, little is lost by ignoring the discreteness in an inferential model.

Applications of the methodology included a simulation study using synthetic data generated from a prokaryotic auto regulatory gene network model. Naturally, the integration of actual measurements into the modeling framework remains of great interest and although real time course data is not yet readily available, it is the subject of on-going research. As post-genomic biology becomes more predictive, the requirement for accurate estimation of kinetic rates is becoming ever more pressing. Quantitative real-time monitoring of gene expression at the level of a single cell is a subject of a great experimental interest, and some small successful pilot studies have demonstrated the possibility of doing this using different coloured flourescent reporter genes. It is anticipated that in the next couple of years, large amounts of data of this type will come on-stream, which will require analysis using the techniques such as those described in this paper.

Further possible extensions to the modeling framework include more efficient MCMC algorithms based on block updating of latent variables (Durham and Gallant, 2002). Such algorithms are straightforward to implement but their appeal is limited as the real problem is the high dependence between the parameters and the missing data. Although a solution to this problem is known in the case of univariate diffusions (Roberts and Stramer, 2001), it does not appear to be possible to extend this technique to the class of multivariate diffusions considered here. It may nevertheless be possible to construct a more efficient sampler for problems of this type based on a joint update of the parameters and the latent process. The incorporation of variation due to experimental error is also of interest and is in principle very straightforward to include in the model. However, this leads to very poor mixing of the MCMC algorithm and satisfactory handling of both partial observation and experimental error is likely to require an MCMC scheme with better mixing properties than the one considered here.

References

Bower, J. M. and Bolouri, H. (2000). *Computational Modeling of Genetic and Biochemical Networks*. The MIT Press, London.

Boys, R. J., Wilkinson, D. J. and Kirkwood, T. B. L. (2004). Bayesian inference for a discretely observed stochastic-kinetic model. *In Submission* .

Doraiswamy, L. K. and Kulkarni, B. D. (1987). *The Analysis of Chemically Reacting Systems*, volume 4 of *Topics in Chemical Engineering*. Gordon and Breach Science.

Durham, G. B. and 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. 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. *Journal of Business and Economic Statistics* **19**, 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.

Golightly, A. and Wilkinson, D. J. (2004). On bayesian inference for stochastic kinetic models using diffusion approximations. [University of Newcastle, Statistics Preprint STA04, 4].

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**, 361–393.

Kitano, H. (2001). *Foundations of Systems Biology*. The MIT Press, London.

Latchman, D. (2002). *Gene Regulation: A Eukaryotic Perspective.* BIOS Scientific Publishers.

McAdams, H. H. and Arkin, A. (1999). Its a noisy business: Genetic regulation at the nanomolar scale. *Trends in Genetics* **15**, 65–69.

McQuarrie, D. A. (1967). Stochastic approach to chemical kinetics. *Journal of Applied Probability* **4**, 413–478.

Ng, T. W., Wilkinson, D. J., Boys, R. J. and Kirkwood, T. B. L. (2004). Stochastic modelling of gene regulatory networks. *In Submission* .

Øksendal, B. (1995). *Stochastic Differential Equations: An Introduction with Applications.* Springer-Verlag, 6th edition.

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

Ptashne, M. (1992). *A Genetic Switch: Phage λ and Higher Organisms.* Cell Press and Blackwell Scientific Publications, 2nd edition.

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

Stundzia, A. B. and Lumsden, C. J. (1996). Stochastic simulation of coupled reaction-diffusion processes. *Journal of Computational Physics* **127**, 196–207.

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

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics* **22**, 1701–1762.

van Kampen, N. G. (2001). *Stochastic Processes in Physics and Chemistry.* North-Holland.

Table 1

Posterior means and Standard Deviations for parameters estimated on 3 replicate length-50 datasets (D_1, D_2 and D_3) from the fully observed model with $m = 5$. The estimation results are based on the final 900,000 iterations of a single run of 1,000,000 MCMC iterations.

	c_1	c_2	c_1/c_2	c_3	c_4	c_5	c_6	c_5/c_6	c_7	c_8
True Values										
	0.1	0.7	0.143	0.35	0.2	0.1	0.9	0.111	0.3	0.1
D_1										
Mean	0.064	0.474	0.141	0.360	0.252	0.043	0.475	0.094	0.288	0.143
S.D.	0.022	0.148	0.035	0.125	0.079	0.013	0.154	0.025	0.099	0.044
D_2										
Mean	0.058	0.363	0.157	0.372	0.240	0.048	0.477	0.105	0.285	0.121
S.D.	0.020	0.120	0.090	0.131	0.071	0.014	0.154	0.047	0.095	0.039
D_3										
Mean	0.052	0.346	0.153	0.416	0.213	0.044	0.488	0.092	0.321	0.115
S.D.	0.020	0.120	0.046	0.151	0.061	0.011	0.145	0.021	0.108	0.036

Table 2

Posterior means and Standard Deviations for parameters estimated using data sets D_1, D_4 and D_5 from the fully observed model. The estimation results are based on the final 900,000 iterations of a single run of 1,000,000 MCMC iterations.

m		c_1	c_2	c_1/c_2	c_3	c_4	c_5	c_6	c_5/c_6	c_7	c_8
True Values											
		0.1	0.7	0.143	0.35	0.2	0.1	0.9	0.111	0.3	0.1
50 Observations											
2	Mean	0.049	0.370	0.137	0.333	0.235	0.030	0.308	0.100	0.269	0.135
	S.D.	0.016	0.110	0.039	0.116	0.079	0.008	0.084	0.031	0.092	0.044
5	Mean	0.066	0.475	0.140	0.361	0.253	0.042	0.468	0.093	0.286	0.143
	S.D.	0.022	0.150	0.032	0.124	0.079	0.012	0.150	0.018	0.095	0.044
8	Mean	0.074	0.524	0.142	0.373	0.258	0.053	0.630	0.087	0.295	0.143
	S.D.	0.027	0.175	0.027	0.122	0.075	0.017	0.226	0.014	0.093	0.041
10	Mean	0.076	0.531	0.143	0.403	0.265	0.060	0.741	0.084	0.316	0.146
	S.D.	0.025	0.165	0.027	0.141	0.076	0.019	0.273	0.013	0.105	0.041
100 Observations											
2	Mean	0.103	0.661	0.157	0.285	0.240	0.051	0.571	0.090	0.224	0.105
	S.D.	0.024	0.142	0.028	0.082	0.061	0.010	0.126	0.015	0.055	0.029
5	Mean	0.096	0.663	0.147	0.286	0.246	0.057	0.593	0.097	0.228	0.110
	S.D.	0.018	0.119	0.027	0.054	0.055	0.013	0.151	0.013	0.048	0.025
8	Mean	0.101	0.687	0.148	0.295	0.250	0.076	0.856	0.091	0.235	0.110
	S.D.	0.020	0.132	0.021	0.066	0.051	0.018	0.233	0.010	0.046	0.024
10	Mean	0.102	0.691	0.149	0.296	0.257	0.096	0.967	0.086	0.236	0.110
	S.D.	0.020	0.134	0.021	0.066	0.052	0.023	0.235	0.009	0.047	0.023
500 Observations											
2	Mean	0.092	0.597	0.155	0.327	0.214	0.101	0.925	0.110	0.222	0.091
	S.D.	0.010	0.062	0.022	0.041	0.026	0.009	0.082	0.011	0.031	0.016
5	Mean	0.098	0.622	0.158	0.331	0.213	0.113	1.028	0.110	0.226	0.092
	S.D.	0.010	0.063	0.021	0.039	0.025	0.010	0.093	0.010	0.029	0.015
8	Mean	0.113	0.824	0.138	0.330	0.216	0.144	1.230	0.114	0.225	0.094
	S.D.	0.013	0.077	0.016	0.040	0.025	0.013	0.126	0.009	0.030	0.016
10	Mean	0.110	0.773	0.143	0.330	0.214	0.137	1.180	0.112	0.226	0.093
	S.D.	0.012	0.073	0.017	0.038	0.024	0.012	0.114	0.009	0.032	0.016

Table 3

Posterior means and Standard Deviations for parameters estimated using data sets D_1, D_4 and D_5 from the partially observed model. Discretization is set at $m = 5$ and the estimation results are based on the final 6,000,000 iterations of a single run of 10,000,000 MCMC iterations.

	c_1	c_2	c_1/c_2	c_3	c_4	c_5	c_6	c_5/c_6	c_7	c_8
True Values										
	0.1	0.7	0.143	0.35	0.2	0.1	0.9	0.111	0.3	0.1
50 Observations										
Mean	0.049	0.442	0.116	0.310	0.012	0.062	0.603	0.103	0.265	0.011
S.D.	0.015	0.131	0.033	0.080	0.023	0.018	0.183	0.013	0.062	0.014
100 Observations										
Mean	0.077	0.941	0.090	0.255	0.270	0.097	0.761	0.120	0.280	0.125
S.D.	0.020	0.253	0.022	0.050	0.122	0.027	0.214	0.012	0.048	0.061
500 Observations										
Mean	0.105	0.574	0.180	0.370	0.187	0.112	1.021	0.110	0.218	0.107
S.D.	0.016	0.076	0.049	0.062	0.073	0.009	0.084	0.008	0.024	0.041