PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

rsta.royalsocietypublishing.org

Research

Cite this article: Stathopoulos V, Girolami MA. 2013 Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. Phil Trans R Soc A 371: 20110541.
http://dx.doi.org/10.1098/rsta.2011.0541

One contribution of 17 to a Discussion Meeting Issue 'Signal processing and inference for the physical sciences'.

Subject Areas:
statistics

Keywords:
Markov chain Monte Carlo, Riemann manifold, linear noise approximation, Markov jump processes

Author for correspondence:
Mark A. Girolami
e-mail: m.girolami@ucl.ac.uk

Royal Society Publishing
Informing the science of the future

Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation

Vassilios Stathopoulos and Mark A. Girolami

Department of Statistical Science, Centre for Computational Statistics and Machine Learning, University College London, Gower Street, London WC1E 6BT, UK

Bayesian analysis for Markov jump processes (MJPs) is a non-trivial and challenging problem. Although exact inference is theoretically possible, it is computationally demanding, thus its applicability is limited to a small class of problems. In this paper, we describe the application of Riemann manifold Markov chain Monte Carlo (MCMC) methods using an approximation to the likelihood of the MJP that is valid when the system modelled is near its thermodynamic limit. The proposed approach is both statistically and computationally efficient whereas the convergence rate and mixing of the chains allow for fast MCMC inference. The methodology is evaluated using numerical simulations on two problems from chemical kinetics and one from systems biology.

1. Introduction

Markov jump processes (MJPs) provide us with a formal description of the underlying stochastic behaviour of many physical systems and as such they have a wide applicability in many scientific fields. In chemistry and biology, for example, they are applied for modelling reactions between chemical species [1,2]. In ecology and epidemiology, they are used for modelling the population of interacting species in the environment [3], whereas in telecommunications they describe the population of information packets over a network [4]. In order to introduce some terminology and notation, we will give a more concrete example from chemical kinetics. However, the modelling methodology is similar in other applications although different assumptions are needed, depending on the system being modelled, for calculating reaction rates. Consider a model for the population of molecules of two interacting chemical species, X_A and X_B, in a solution of volume Ω, where X_A and X_B denote the number of molecules of chemicals A and B, respectively. The interactions between the species are modelled using *reactions* which are specified using the following notation: $R_1 : A + B \xrightarrow{c_1} 2A$. On the left-hand side appear the *reactants* and on the right-hand side the *products* of the reaction while over the arrow appears the *rate constant* c_1 which is the probability that a randomly chosen pair of A and B will react according to R_1. This reaction, for example, specifies that a pair of molecules A and B react with probability c_1 to produce a new molecule of A. For calculating the probability of a reaction taking place given the current state of the system, i.e. the number of molecules of chemicals A and B, several system-dependent assumptions must be made. For chemical reactions, it is assumed that in a well-stirred solution the probability of a reaction is proportional to the populations of its products [5]. For R_1, we can write it as $f_1(X_A, X_B, c_1) = c_1 \Omega^{-1} X_A X_B$. Following the same reasoning, additional reactions and species can be added in order to construct large and complex reaction networks. Together, the state of the system X_A and X_B, the set of reactions and the reaction rates specify an MJP where the occurrences of reactions are modelled as a Poisson process.

In this particular example, the probability of the reaction has a simple form and is linear with respect to the populations. However, in many real applications, this is often not the case while the *rate constant*, c_1, is unknown. Given a fully specified MJP, i.e. an MJP with known parameters, rate constants and initial conditions, it is possible to perform exact simulation and obtain samples from the underlying stochastic process using the stochastic simulation algorithm (SSA) described by Gillespie [1]. In many problems, there are system parameters which are not specified or are unknown whereas it is relatively easy to collect partial observations of the physical process at discrete time points. The interest is therefore to obtain statistical estimates of the unknown parameters using the available data.

As a consequence of the Markov property, MJPs satisfy the Chapman–Kolmogorov equation from which we can directly obtain the forward master equation describing the evolution of the system's state probability over any time interval. However, even for small and simple systems the master equation is intractable and it is not straightforward as to how partially and discretely observed data from the physical process should be incorporated in order to perform inference over unknown system parameters. Recently, Boys *et al*. [6] have shown that it is possible to construct a Markov chain whose stationary probability distribution is the posterior of the unknown parameters without resorting to any approximations of the original MJP. Their method, however, is computationally expensive while the strong correlation between posterior samples means that a large number of Markov chain Monte Carlo (MCMC) iterations are required in order to obtain Monte Carlo estimates with sufficient accuracy.

An alternative is to consider suitable approximations of the likelihood function. The system size expansion mentioned by Van Kampen [7, ch. 10] provides a systematic method for obtaining approximations of a physical process approaching its thermodynamic limit. The most simple approximation yields the macroscopic rate equation (MRE), which describes the thermodynamic limit of the system with a set of ordinary differential equations neglecting any random fluctuations. Although the MRE has extensively been studied in the literature [8,9], it is not applicable for problems where information about the noise and the random fluctuations is necessary or the system is far from its thermodynamic limit. The diffusion approximation [10,11] describes the physical process by a set of nonlinear stochastic differential equations with state-dependent Brownian motion. Similar to the master equation, however, the likelihood is intractable. In the study by Roberts & Stramer [12], a transformation is applied such that the Brownian increments are independent of the system state and thus the system can easily be simulated. However, this limits the applicability of the methodology into systems where such a transformation is possible. A more general methodology is presented by Golightly & Wilkinson [13], who used an approximation of the likelihood instead. Finally, a less studied approach for the purpose of inference is the linear noise approximation (LNA), which conveniently decouples nonlinearity in the diffusion approximation into a nonlinear set of ordinary differential equations in the MRE and a set of linear stochastic differential equations for the random fluctuations around a deterministic state [7, ch. 10, 14]. Recently, Komorowski *et al*. [15] have shown that the simple analytic form of the approximate likelihood obtained by the LNA simplifies MCMC inference and can be applied to problems with a relatively small number of molecules.

A commonly employed algorithm for MCMC inference is the Metropolis–Hastings algorithm [16], which relies on random perturbations around the current state using a local proposal mechanism. It should be noted here that the state of the Markov chain is different from the state of the stochastic process. In the MCMC context, state refers to the current values of the unknown system parameters whereas the state of the system refers to the value of the stochastic process at a given time. We will use the term state interchangeably for the rest of this paper and its meaning will be clear from the context. Owing to the local nature of the proposal mechanism used in the Metropolis–Hastings algorithm, samples from the posterior exhibit strong random walk behaviour and auto-correlation. Tuning the proposal mechanisms to achieve good mixing and fast convergence is far from straightforward even though some theoretical guidance is provided [17]. MCMC methods, such as the Metropolis-adjusted Langevin algorithm (MALA) [18] and the Hamiltonian Monte Carlo (HMC) [19], have also been studied in the literature and have been shown to be more efficient than random walk Metropolis–Hastings in terms of effective sample size (ESS) and convergence rates on several problems. However, the HMC and MALA also require extensive tuning of the proposal mechanisms [20,21]. For MJPs, the problem is compounded further since system parameters, such as probability rate constants of chemical reactions, are often highly correlated and their values may differ by orders of magnitude. The resulting posterior distributions have long narrow ‘valleys’ preventing any local proposal mechanism from proposing large moves about the parameter space.

More recently, Girolami & Calderhead [22] proposed exploitation of the underlying Riemann manifold of probability density functions when defining MCMC methods, thus exploiting the intrinsic geometry of statistical models and thereby providing a principled framework and a systematic approach to the proposal design process. These algorithms rely on the gradient and Fisher information matrix of the likelihood function to automatically tune the proposal mechanism such that large moves on the parameter space are possible and therefore improve convergence and mixing of the chains. In the study of Calderhead & Girolami [9], this approach has successfully been applied for the MRE approximation of chemical reaction networks. For the LNA, the Fisher information and the gradient of the likelihood function can easily be obtained [2]. In this paper, we study the application of the Riemann manifold MCMC methods for the LNA and compare the mixing efficiency and computational cost with the commonly used Metropolis–Hastings algorithm. Moreover, we study how the Markov chains and the resulting Monte Carlo estimates behave for systems which are far from their thermodynamic limit. The aim is to improve the efficiency of MCMC inference for MJPs in order to allow for larger and more complex models frequently encountered in biology and chemistry to be studied in more detail.

In §2, we give a brief overview of MJPs. The diffusion and LNAs are presented in §3. We then discuss MCMC and the Riemann manifold algorithms in §4. Numerical simulations are presented in §6 while §7 concludes the paper.

2. Markov jump processes

A D-dimensional stochastic process is a family of D random variables $\boldsymbol{X}(t) = [X_1(t), \dots, X_D(t)]^{\mathrm{T}}$ indexed by a continuous time variable t with initial conditions $\boldsymbol{X}(t_0) = \boldsymbol{x}_{t_0}$. An MJP is a stochastic process satisfying the Markov property such that

$$p[\boldsymbol{X}(t_0), \dots, \boldsymbol{X}(t_N)] = p[\boldsymbol{X}(t_0)] \prod_{i=1}^{N} p[\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1})],$$

where the dependence on any parameters or other quantities has been suppressed. That is, the conditional probability of the system state at time t_i only depends on the state of the system at the previous time t_{i-1}. An MJP is characterized by a finite number, M, of state transitions with rates $f_j(\boldsymbol{x}, \boldsymbol{\theta}, t)$ and state change vectors $\boldsymbol{s}_j = (s_{1,j}, \dots, s_{D,j})^{\mathrm{T}}$ with $j \in [1, \dots, M]$. $f_j(\boldsymbol{x}, \boldsymbol{\theta}, t)\,\mathrm{d}t$ is the probability, given the state of the system at time t, $\boldsymbol{X}(t) = \boldsymbol{x}$, of a jump to a new state $\boldsymbol{x} + \boldsymbol{s}_j$ in the infinitesimal time interval $[t, t + \mathrm{d}t)$. For the problems we consider in this paper the transition rates depend not only on the current state and time but also on unknown rate parameters $\boldsymbol{\theta}$. From the Markov property, we can directly obtain the conditional probability of the system being in state $\boldsymbol{x}$ at time t given initial conditions, which is characterized by the master equation

$$\frac{p(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)}{\mathrm{d}t} = \sum_{j=1}^{M} [f_j(\boldsymbol{x} - \boldsymbol{s}_j, \boldsymbol{\theta}, t) p(\boldsymbol{x} - \boldsymbol{s}_j, t|\boldsymbol{x}_0, t_0) - f_j(\boldsymbol{x}, \boldsymbol{\theta}, t) p(\boldsymbol{x}, t|\boldsymbol{x}_0, t_0)]. \tag{2.1}$$

Equation (2.1) in general form is intractable especially when the transition rate functions $f_j(\cdot)$ are nonlinear with respect to the system state. Numerical simulation is also prohibitively expensive as the computational cost grows exponentially with D [23].

However, given initial conditions $\boldsymbol{X}(t_0) = \boldsymbol{x}_{t_0}$ and values for the unknown rate parameters $\boldsymbol{\theta}$ we can simulate realizations of the MJP by first noting that the time τ to the next state transition is exponentially distributed with rate $\lambda = \sum_{j=1}^{M} f_j(\boldsymbol{x}_{t_0}, \boldsymbol{\theta}, t_0)$ and the new state $\boldsymbol{X}(t_0 + \tau)$ will be $\boldsymbol{x}_{t_0} + \boldsymbol{s}_j$ with probability $f_j(\boldsymbol{x}_{t_0}, \boldsymbol{\theta}, t)/\lambda$. This scheme results in an iterative algorithm from which we can forward simulate a complete trajectory for the stochastic process $\boldsymbol{X}(t)$, known as the SSA [1] in the chemical kinetics literature.

From the specification of the MJP, we can also write the likelihood function with respect to the parameters $\boldsymbol{\theta}$ for a completely observed process $\boldsymbol{X}(t)$ at the time interval $[0, T]$ as

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f_{k_i}(\boldsymbol{x}_{i-1}, \boldsymbol{\theta}, \tau_{i-1}) \exp\left(-\tau_i \sum_{j'=1}^{M} f_{j'}(\boldsymbol{x}_{i-1}, \boldsymbol{\theta}, \tau_{i-1})\right),$$

where N is the number of transitions that occurred in the time interval $[0, T]$, $k_i \in [1, \dots M]$ is the type of the ith transition and τ_i and $\boldsymbol{x}_i$ are the time and state at the ith transition, respectively. Note that the likelihood function corresponds to the generative process described by the SSA. By specifying a suitable prior and applying Bayes' theorem, we can obtain the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$ which we can use for inference over the unknown parameters $\boldsymbol{\theta}$ [6].

In many problems of interest, however, we cannot observe the times and types of all transitions in a given time interval. Rather, we can only observe the state of the system $\boldsymbol{X}(t_i) = \boldsymbol{x}_i$ at discrete time points $t_i \in [0, T]$. The solution proposed by Boys *et al.* [6] is to treat the trajectories, as well as the number, times and types of transitions, between observed time points to those latent variables. This leads to a data augmentation framework [24] in which a Markov chain is constructed to sample from the joint posterior of the parameters and the latent variables. At each MCMC iteration, the complete trajectory of the MJP has to be simulated conditional on the observed data and the parameters that, for some systems, can be computationally demanding. Furthermore, because of the high-dimensional nature of the simulated trajectory and the strong dependence on the system parameters and observed data the MCMC algorithm has very poor convergence and mixing properties, requiring many samples from the posterior in order to obtain sufficiently accurate Monte Carlo estimates. Finally, a further complication that arises is that the number of transitions between two observed time points is also unknown and has to be sampled using a reversible-jumps-type algorithm [25]. For more details, see Boys *et al.* [6]. The resulting algorithm therefore is computationally demanding, thus limiting its applicability on small and relatively simple MJPs. A more efficient version of the algorithm is also suggested in Boys *et al.* [6], where instead of simulating the trajectories between observations using the exact MJP an approximate proposal distribution is employed to sample trajectories which are accepted or rejected using the Metropolis–Hastings ratio.

3. Diffusion and linear noise approximations

An alternative to working directly with the master equation and the original MJP is to consider approximations which provide for efficient simulation and possibly an easy to evaluate likelihood function for discretely observed data. Although the resulting posterior will also be approximate in nature, it can be sufficient for inferential purposes given that the system under consideration is near its thermodynamic limit. Here, we describe the diffusion approximation and from that how we can arrive at the LNA. Our presentation is rather informal and follows Gillespie and colleagues [1,14]. For a more formal derivation, the reader should refer to Van Kampen [7] and Gillespie [11]. The requirement for these approximations to be consistent is the existence of a proportionality constant Ω, which governs the size of the fluctuations such that for large Ω the jumps will relatively be small, and as both Ω and $\boldsymbol{x}$ tend to infinity approaching the system's thermodynamic limit, then

$$f_j(\boldsymbol{x},\boldsymbol{\theta},t) \to \Omega\tilde{f}_j(\boldsymbol{z},\boldsymbol{\theta},t), \tag{3.1}$$

where $\boldsymbol{z}=\boldsymbol{x}/\Omega$ and $\tilde{f}_j(\cdot)$ are independent of Ω. For many physical processes where the fluctuations are due to the discrete nature of matter there is a natural Ω parameter with such properties. Examples of such parameters can be the system size in chemical kinetics, the capacity of a condenser in electric circuits or the mass of a particle [7].

(a) Diffusion approximation

In order to obtain a Langevin equation which closely matches the dynamics of the MJP, it is assumed that there is an infinitesimal time interval dt that satisfies the following conditions:

$$f_j(\boldsymbol{x}_{t'},\boldsymbol{\theta},t') \approx f_j(\boldsymbol{x}_t,\boldsymbol{\theta},t), \quad \forall t' \in [t,t+\mathrm{d}t) \quad \forall j \in [1,M] \tag{3.2}$$

and

$$f_j(\boldsymbol{x}_t,\boldsymbol{\theta},t)\,\mathrm{d}t \gg 1 \quad \forall j \in [1,M]. \tag{3.3}$$

The first condition constrains dt to be small enough such that the transition rate functions remain approximately constant. This implies that the number of transitions of type j is distributed as a Poisson random variable with mean $f_j(\boldsymbol{x}_t,\boldsymbol{\theta},t)\,\mathrm{d}t$ and is independent from other transitions of type $j' \neq j$. The second condition constrains dt to be large enough such that the number of transitions for each state is significantly larger than 1, which further implies that the Poisson distribution can accurately be approximated by a Gaussian distribution. It can be shown [26] that we can choose dt and Ω such that both conditions can be satisfied, and this generally occurs when the system approaches its thermodynamic limit.

Given such a time scale, the state of the system at time $t+\mathrm{d}t$ can be computed by

$$\boldsymbol{x}_{t+\mathrm{d}t} = \boldsymbol{x}_t + \sum_{j=1}^{M} \mathcal{N}[f_j(\boldsymbol{x}_t,\boldsymbol{\theta},t)\,\mathrm{d}t, f_j(\boldsymbol{x}_t,\boldsymbol{\theta},t)\,\mathrm{d}t]\boldsymbol{s}_j, \tag{3.4}$$

where $\mathcal{N}[\mu,\sigma^2]$ denotes a Gaussian random variate with mean μ and variance σ^2. From equation (3.4), we can directly obtain a Langevin equation of the form

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{S}\boldsymbol{f}(\boldsymbol{x}_t,\boldsymbol{\theta},t)\,\mathrm{d}t + \boldsymbol{S}\sqrt{\mathrm{diag}[\boldsymbol{f}(\boldsymbol{x}_t,\boldsymbol{\theta},t)]}\mathrm{d}\boldsymbol{B}_t, \tag{3.5}$$

where we use $\boldsymbol{S}$ to denote the matrix whose columns are the state change vectors $\boldsymbol{s}_j$, $\boldsymbol{f}(\cdot)$ to denote the vector whose elements are the transition rates $f_j(\cdot)$, diag($\boldsymbol{v}$) a function that returns a diagonal matrix with elements taken from the vector $\boldsymbol{v}$ and d$\boldsymbol{B}_t$ an M-dimensional Wiener process. Note that the dimension of $\boldsymbol{x}_t$ differs from that of d$\boldsymbol{B}_t$.

Owing to the nonlinear state-dependent drift and diffusion coefficients in equation (3.5) the transition density of the stochastic process is also intractable. Therefore, a data augmentation approach similar to the one in Boys *et al.* [6] has to be followed. However, there is no longer the need to sample the number, times and types of state transitions as the MJP is approximated with a continuous process. Moreover, the latent variables corresponding to unobserved states can now be efficiently simulated by a Euler–Maruyama scheme which is computationally more efficient than the SSA. This approach has been followed by Roberts & Stramer [12] and Golightly & Wilkinson [13] for inference over the unknown parameters $\boldsymbol{\theta}$, while in Heron *et al.* [27] a similar methodology has been applied on real data from an auto-regulatory gene expression network.

(b) Linear noise approximation

Substituting equation (3.1) in the Langevin equation (3.5) and dividing by Ω, we obtain

$$\mathrm{d}z_t = \boldsymbol{S}\tilde{\boldsymbol{f}}(z_t, \boldsymbol{\theta}, t)\,\mathrm{d}t + \frac{1}{\sqrt{\Omega}}\boldsymbol{S}\sqrt{\mathrm{diag}[\tilde{\boldsymbol{f}}(z_t, \boldsymbol{\theta}, t)]}\mathrm{d}\boldsymbol{B}_t, \tag{3.6}$$

from which we can see that the fluctuations are of the order of $1/\sqrt{\Omega}$ and in the thermodynamic limit (3.6) reduces to the MRE

$$\lim_{\Omega\to\infty} \mathrm{d}z_t = \boldsymbol{S}\tilde{\boldsymbol{f}}(z_t, \boldsymbol{\theta}, t)\,\mathrm{d}t.$$

To obtain the LNA, we make the assumption that for sufficiently large Ω a solution to (3.6) will differ from the MRE by a stochastic term of order $1/\sqrt{\Omega}$. That is,

$$z_t = \boldsymbol{\phi}_t + \frac{1}{\sqrt{\Omega}}\boldsymbol{\xi}_t, \tag{3.7}$$

where $\boldsymbol{\phi}_t$ are deterministic or sure variables satisfying the MRE and $\boldsymbol{\xi}_t$ are stochastic variables. Rewriting the transition rate functions using (3.7) and Taylor expand around $\boldsymbol{\phi}$, we get

$$\tilde{f}_j(z, \boldsymbol{\theta}, t) = \tilde{f}_j\left(\boldsymbol{\phi} + \frac{1}{\sqrt{\Omega}}\boldsymbol{\xi}\right) = \tilde{f}_j(\boldsymbol{\phi}, \boldsymbol{\theta}, t) + \frac{1}{\sqrt{\Omega}}\sum_{\mathrm{d}=1}^{D}\frac{\partial\tilde{f}_j(\boldsymbol{\phi}, \boldsymbol{\theta}, t)}{\partial\phi_i}\xi_i + O(\Omega^{-1}). \tag{3.8}$$

We can now substitute (3.7) and (3.8) back into (3.6) and collect terms of $O(1)$ to get the expression for the differential of $\boldsymbol{\phi}$ which is none other than the MRE

$$\mathrm{d}\boldsymbol{\phi}_t = \boldsymbol{S}\tilde{\boldsymbol{f}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)\,\mathrm{d}t. \tag{3.9}$$

Finally, collecting remaining terms and neglecting terms of $O(1/\sqrt{\Omega})$ and higher we get the differential of $\boldsymbol{\xi}$ as

$$\mathrm{d}\boldsymbol{\xi}_t = \boldsymbol{S}\boldsymbol{J}_{\tilde{f}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)\boldsymbol{\xi}_t\mathrm{d}t + \boldsymbol{S}\sqrt{\mathrm{diag}[\tilde{\boldsymbol{f}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)]}\mathrm{d}\boldsymbol{B}_t, \tag{3.10}$$

where we used $\boldsymbol{J}_{\tilde{f}}(\cdot)$ to denote the Jacobian of the transition rates $\tilde{\boldsymbol{f}}(\cdot)$. Equation (3.10) characterizes the fluctuations around the deterministic state $\boldsymbol{\phi}$ and its validity depends on the size of Ω. As Ω increases the magnitude of the individual jumps $\boldsymbol{s}_j$ becomes negligible relative to the distance in $\boldsymbol{\phi}$ over which the nonlinearity of $\tilde{f}_j(\cdot)$ becomes noticeable. A measure of the sufficiency of LNA is the coefficient of variation, i.e. the ratio of the standard deviation to the mean. For a more thorough discussion on the validity of LNA, the reader is referred to Ferm *et al.* [23] and the supplementary material of Komorowski *et al.* [15].

(c) Solution of the linear noise approximation and the approximate likelihood function

LNA provides a convenient expression for the approximate likelihood since the MRE (3.9) can easily be solved numerically and its computational cost is polynomial in D.

Moreover, equation (3.10) is a system of linear stochastic differential equations which has an explicit solution of the form

$$\boldsymbol{\xi}_t = \boldsymbol{\Phi}(t_0, t)\left(\xi_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(s,t)^{-1}\boldsymbol{S}\sqrt{\mathrm{diag}[\tilde{\boldsymbol{f}}(\boldsymbol{\phi}_s, \boldsymbol{\theta}, s)]}\mathrm{d}\boldsymbol{B}_s\right), \tag{3.11}$$

where the integral is in the Itô sense and $\boldsymbol{\Phi}(t_0, t)$ is the solution of

$$\mathrm{d}\boldsymbol{\Phi}(t_0, s) = \boldsymbol{S}\boldsymbol{J}_{\tilde{f}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)\boldsymbol{\Phi}(t_0, s)\,\mathrm{d}s, \quad \boldsymbol{\Phi}(t_0, t_0) = \boldsymbol{I}. \tag{3.12}$$

Since the Itô integral of a deterministic function is a Gaussian random variable [28], equation (3.11) implies that $\boldsymbol{\xi}_t$ has a multi-variate normal distribution. To simplify further, the analysis assumes that the initial condition for z_t has a multi-variate normal distribution such that $z_{t_0} \sim \mathcal{N}(\boldsymbol{\phi}_{t_0}, \boldsymbol{V}_{t_0})$. For the rest of the paper, we will assume that $\boldsymbol{\phi}_{t_0}$ and $\boldsymbol{V}_{t_0}$ are known. In cases where the initial conditions are unknown they can be treated as additional parameters. Equations (3.7), (3.9)–(3.11) and the specification of initial conditions further imply that

$$z_t \sim \mathcal{N}(\boldsymbol{\phi}_t, \Omega^{-1}\boldsymbol{V}_t), \tag{3.13}$$

where $\boldsymbol{\phi}_t$ are solutions of the MRE and $\boldsymbol{V}_t$ are solutions of

$$\mathrm{d}\boldsymbol{V}_t = \boldsymbol{S}\boldsymbol{J}_{\tilde{f}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)\boldsymbol{V}_t + \boldsymbol{V}_t\boldsymbol{J}_{\tilde{f}}^{\mathrm{T}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{S}\,\mathrm{diag}[\tilde{\boldsymbol{f}}(\boldsymbol{\phi}_t, \boldsymbol{\theta}, t)]\boldsymbol{S}^{\mathrm{T}}.$$

Finally, multiplying (3.13) by Ω, we get

$$\boldsymbol{x}_t \sim \mathcal{N}(\Omega\boldsymbol{\phi}_t, \Omega\boldsymbol{V}_t).$$

Assume that we have observations from the stochastic process $\boldsymbol{X}(t)$ at discrete time points $t_i \in \{t_1, \ldots, t_N\}$. Moreover, assume that each observation $\boldsymbol{x}_t$ is obtained by an independent realization of $\boldsymbol{X}(t)$. For example, to obtain an observation at $t_1 = 10$ the SSA is used to simulate a trajectory from t_0 to t_1 and the state of the system at t_1 is kept. For $t_2 = 20$, the SSA is again used to simulate a new trajectory from t_0 to t_2 keeping only the state of the system at t_2, and the process continues until all necessary observations are gathered. These kinds of data are very frequently encountered in biology where in order to obtain a single measurement the sample has to be 'sacrificed'. This is common in data obtained using polymerase chain reaction reporter assays [29], for example. See also Komorowski *et al.* [15] for an example of an inference problem with such data. Owing to the independence between different observations and the Markov property the likelihood is simply

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_{t_i}|\Omega\boldsymbol{\phi}_{t_i}, \Omega\boldsymbol{V}_{t_i}). \tag{3.14}$$

In this paper, we only consider observations of this kind. However, the methodology is readily applicable when observations from a single realization of $\boldsymbol{X}(t)$ are available. In this case, the likelihood also has a simple form

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{N}[\boldsymbol{X}|\Omega\boldsymbol{\mu}(\boldsymbol{\theta}), \Omega\boldsymbol{\Sigma}(\boldsymbol{\theta})],$$

where $\boldsymbol{X} = (\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{t_N})^{\mathrm{T}}$ is an ND vector with all the observations, $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\boldsymbol{\phi}_{t_1}, \ldots, \boldsymbol{\phi}_N)^{\mathrm{T}}$ is also an ND vector with solutions of the MRE and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is an $ND \times ND$ block matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \{\boldsymbol{\Sigma}(\boldsymbol{\theta})^{i,j} : i, j \in [1, \ldots, N]\}$ such that

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})^{i,j} = \begin{cases} \boldsymbol{V}_{t_i} & i = j, \\ \boldsymbol{V}_{t_i}\boldsymbol{\Phi}(t_i, t_j)^{\mathrm{T}} & i \neq j. \end{cases} \tag{3.15}$$

This stems from the fact that because of the Markov property and equation (3.13) each $\boldsymbol{x}_{t_i}$ can be written as a sum of multi-variate normal random variables and therefore $\boldsymbol{X}$ is also a multi-variate normal random variable. For more details, refer to the supplementary material of Komorowski *et al.* [2,15]. The only additional complication which arises for time-series data is that the off-diagonal components of the LNA variance in equation (3.15) need to be estimated by numerically solving the system of ordinary differential equations in equation (3.12). Note that despite the fact that the variance matrix is full we can still exploit the Markov property and write the likelihood as a product of the conditional likelihoods and therefore avoid the cost of inverting the $ND \times ND$ variance matrix.

4. Markov chain Monte Carlo methods

In this section, we give a brief overview of the MCMC algorithms that we consider in this work. Some familiarity with the concepts of MCMC is required by the reader since an introduction to the subject is outside the scope of this paper.

(a) Metropolis–Hastings

For a random vector $\boldsymbol{\theta} \in \mathbb{R}^D$ with density $p(\boldsymbol{\theta})$, the Metropolis–Hastings algorithm employs a proposal mechanism $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$ and proposed moves are accepted with probability $\min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^{t-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\}$. In the context of Bayesian inference the target density $p(\boldsymbol{\theta})$ corresponds to the posterior distribution of the model parameters. Tuning the Metropolis–Hastings algorithm involves selecting the right proposal mechanism. A common choice is to use a random walk Gaussian proposal of the form $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}, \boldsymbol{\Sigma})$, where $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multi-variate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Selecting the covariance matrix, however, is far from trivial in most cases since knowledge about the target density is required. Therefore, a more simplified proposal mechanism is often considered in which the covariance matrix is replaced with a diagonal matrix such as $\boldsymbol{\Sigma} = \epsilon \boldsymbol{I}$ where the value of the scale parameter ϵ has to be tuned in order to achieve fast convergence and good mixing. Small values of ϵ imply small transitions and result in high acceptance rates while the mixing of the Markov Chain is poor. Large values, on the other hand, allow for large transitions but they result in most of the samples being rejected.

Tuning the scale parameter becomes even more difficult in problems where the standard deviations of the marginal posteriors differ substantially, since different scales are required for each dimension, and this is exacerbated when correlations between different variables exist. Adaptive schemes for the Metropolis–Hastings algorithm have also been proposed [30] though they should be applied with care [31]. Parameters such as reaction rate constants often differ by orders of magnitude, thus a scaled diagonal covariance matrix will be a bad choice for such problems. In the numerical simulations in the next section, we use a Metropolis within Gibbs scheme where each parameter is updated conditional on all others using a univariate normal density with a parameter-specific scale parameter. This allows us to tune the scale for each proposal independently and achieve better mixing.

(b) Manifold Metropolis-adjusted Langevin algorithm

Denoting the log of the target density as $\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta})$, the manifold MALA (MMALA) method [22] defines a Langevin diffusion with stationary distribution $p(\boldsymbol{\theta})$ on the Riemann manifold of density functions with metric tensor $\boldsymbol{G}(\boldsymbol{\theta})$. By employing a first-order Euler integrator to solve the diffusion a proposal mechanism with density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^{t-1}, \epsilon), \epsilon^2 \boldsymbol{G}^{-1}(\boldsymbol{\theta}^{t-1}))$ is obtained, where ϵ is the integration step size, a parameter which needs to be tuned, and the dth component of the mean function $\boldsymbol{\mu}(\boldsymbol{\theta}, \epsilon)_d$ is

$$\boldsymbol{\mu}(\boldsymbol{\theta}, \epsilon)_d = \boldsymbol{\theta}_d + \frac{\epsilon^2}{2}(\boldsymbol{G}^{-1}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}))_d - \epsilon^2 \sum_{i=1}^{D}\sum_{j=1}^{D} \boldsymbol{G}(\boldsymbol{\theta})^{-1}_{i,j} \Gamma^d_{i,j}, \tag{4.1}$$

where $\Gamma^d_{i,j}$ are the Christoffel symbols of the metric in local coordinates [32].

Similarly to MALA [18], owing to the discretization error introduced by the first-order approximation, convergence to the stationary distribution is not guaranteed anymore and thus the Metropolis–Hastings ratio is employed to correct this bias. The MMALA can simply be stated as in algorithm 1 and more details can be found in Girolami & Calderhead [22].

```
Initialize $\boldsymbol{\theta}^0$
for $t = 1$ to $T$ do
    $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}(\boldsymbol{\theta}^{t-1}, \epsilon), \epsilon^2 \boldsymbol{G}^{-1}(\boldsymbol{\theta}^{t-1}))$
    $r = \min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^{t-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})\}$
    $u \sim \mathcal{U}_{[0,1]}$
    if $r > u$ then
        $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$
    else
        $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$
    end if
end for
```

Algorithm 1: MMALA

We can interpret the proposal mechanism of MMALA as a local Gaussian approximation to the target density similar to the adaptive Metropolis–Hastings in Haario *et al.* [33]. In contrast to Haario *et al.* [33], the effective covariance matrix in MMALA is the inverse of the metric tensor evaluated at the current position and no samples from the chain are required in order to estimate it, therefore avoiding the difficulties of adaptive MCMC discussed in Andrieu & Thoms [31]. Furthermore, a simplified version of the MMALA (SMMALA) can also be derived by assuming a manifold with constant curvature, thus cancelling the last term in equation (4.1), which depends on the Christoffel symbols. Finally, the MMALA can be seen as a generalization of the original MALA [18] since, if the metric tensor $\boldsymbol{G}(\boldsymbol{\theta})$ is equal to the identity matrix corresponding to a Euclidean manifold, then the original algorithm is recovered.

(c) Manifold Hamiltonian Monte Carlo

The Riemann manifold Hamiltonian Monte Carlo (RMHMC) method defines a Hamiltonian on the Riemann manifold of probability density functions by introducing the auxiliary variables $\boldsymbol{p} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\theta}))$, which are interpreted as the momentum at a particular position $\boldsymbol{\theta}$ and by considering the negative log of the target density as a potential function. More formally, the Hamiltonian defined on the Riemann manifold is

$$H(\boldsymbol{\theta}, \boldsymbol{p}) = -\mathcal{L}(\boldsymbol{\theta}) + \tfrac{1}{2}\log(2\pi|\boldsymbol{G}(\boldsymbol{\theta})|) + \tfrac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{p}, \tag{4.2}$$

where the terms $-\mathcal{L}(\boldsymbol{\theta}) + \frac{1}{2}\log(2\pi|\boldsymbol{G}(\boldsymbol{\theta})|)$ and $\frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{\theta})^{-1}\boldsymbol{p}$ are the potential energy and kinetic energy terms, respectively. Simulating the Hamiltonian requires a time-reversible and volume-preserving numerical integrator. For this purpose, the generalized leapfrog algorithm can be employed and provides a deterministic proposal mechanism for simulating from the conditional distribution, i.e. $\boldsymbol{\theta}^*|\boldsymbol{p} \sim p(\boldsymbol{\theta}^*|\boldsymbol{p})$. More details about the generalized leapfrog integrator can be found in Girolami & Calderhead [22]. To simulate a path across the manifold, the leapfrog integrator is iterated L times, which along with the integration step size ϵ are parameters requiring tuning. Again, owing to the integration errors on simulating the Hamiltonian, in order to ensure convergence to the stationary distribution the Metropolis–Hastings ratio is applied. Moreover, following the suggestion by Radford [20] the number of leapfrog iterations L is randomized in order to improve mixing. The RMHMC algorithm is given in algorithm 2.

```
Initialize $\boldsymbol{\theta}^0$
for $t = 1$ to $T$ do
    $\boldsymbol{p}_*^0 \sim \mathcal{N}(\boldsymbol{p}|\mathbf{0}, \boldsymbol{G}(\boldsymbol{\theta}^{t-1}))$
    $\boldsymbol{\theta}_*^0 = \boldsymbol{\theta}^{t-1}$
    $e \sim \mathcal{U}_{[0,1]}$
    $N = \text{ceil}(\epsilon L)$
    {Simulate the Hamiltonian using a generalized leapfrog integrator for N steps}
    for $n = 0$ to $N$ do
        solve $\boldsymbol{p}_*^{n+\frac{1}{2}} = \boldsymbol{p}_*^n - \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}_*^n, \boldsymbol{p}_*^{n+\frac{1}{2}})$
        solve $\boldsymbol{\theta}_*^{n+1} = \boldsymbol{\theta}_*^n + \frac{\epsilon}{2}[\nabla_{\boldsymbol{p}} H(\boldsymbol{\theta}_*^n, \boldsymbol{p}_*^{n+\frac{1}{2}}) + \nabla_{\boldsymbol{p}} H(\boldsymbol{\theta}_*^{n+1}, \boldsymbol{p}_*^{n+\frac{1}{2}})]$
        $\boldsymbol{p}_*^{n+1} = \boldsymbol{p}_*^{n+\frac{1}{2}} - \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}_*^{n+1}, \boldsymbol{p}_*^{n+\frac{1}{2}})$
    end for
    $(\boldsymbol{\theta}^*, \boldsymbol{p}^*) = (\boldsymbol{\theta}_*^{N+1}, \boldsymbol{p}_*^{N+1})$
    {Metropolis–Hastings ratio}
    $r = \min\left\{1, \exp(-H(\boldsymbol{\theta}^*, \boldsymbol{p}^*) + H(\boldsymbol{\theta}^{t-1}, \boldsymbol{p}^{t-1}))\right\}$
    $u \sim \mathcal{U}_{[0,1]}$
    if $r > u$ then
        $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$
    else
        $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$
    end if
end for
```

Algorithm 2: RMHMC

Similar to the MMALA, when the metric tensor $\boldsymbol{G}(\boldsymbol{\theta})$ is equal to the identity matrix corresponding to a Euclidean manifold, then the RMHMC algorithm is equivalent to the HMC algorithm of Duane *et al.* [19].

5. Implementation details

(a) Gradient and metric tensor for the linear noise approximation

For the manifold MCMC algorithms discussed in this section, we will need the gradient of the log likelihood as well as a metric tensor for the LNA. For density functions the natural metric tensor is the expected Fisher information, $\boldsymbol{I}(\boldsymbol{\theta})$ [34], and for a multi-variate normal with mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ its general form is

$$I(\boldsymbol{\theta})_{i,j} = \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_j} + \frac{1}{2}\text{Tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_j}\right).$$

For the likelihood in equation (3.14), the Fisher information is then a sum of N matrices $\boldsymbol{I}(\boldsymbol{\theta}, t)$, one evaluated at each time point. Similarly, the general form of the partial derivatives for the log of a multi-variate normal is

$$\frac{\partial \ln \mathcal{N}[\boldsymbol{x}|\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})]}{\partial \theta_i} = \frac{1}{2}\text{Tr}\left[(\boldsymbol{c}\boldsymbol{c}^{\mathrm{T}} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}))\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i}\right] + \boldsymbol{c}^{\mathrm{T}} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i},$$

where $\boldsymbol{c} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})[\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})]$.

Moreover, during the leapfrog integration for the RMHMC and for the mean function of MMALA the partial derivatives of the Fisher information are needed. Their general form is

$$\begin{aligned}\frac{\partial I(\boldsymbol{\theta})_{i,j}}{\partial\theta_k} &= \frac{\partial^2\boldsymbol{\mu}(\boldsymbol{\theta})^{\mathrm{T}}}{\partial\theta_i\partial\theta_k}\boldsymbol{a}_j + \boldsymbol{a}_i^{\mathrm{T}}\frac{\partial^2\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k} - \boldsymbol{a}_i^{\mathrm{T}}\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_k}\boldsymbol{a}_j - \frac{1}{2}\mathrm{Tr}[\boldsymbol{A}_k(\boldsymbol{A}_i\boldsymbol{A}_j + \boldsymbol{A}_j\boldsymbol{A}_i)]\\ &\quad + \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\left(\frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_k}\boldsymbol{A}_j + \frac{\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\boldsymbol{A}_i\right)\right],\end{aligned}$$

where $\boldsymbol{a}_i = \boldsymbol{\Sigma}^{-1}\partial\boldsymbol{\mu}(\boldsymbol{\theta})/\partial\theta_i$ and $\boldsymbol{A}_i = \boldsymbol{\Sigma}^{-1}\partial\boldsymbol{\Sigma}(\boldsymbol{\theta})/\partial\theta_i$.

The above quantities require first- and second-order sensitivities for the $\boldsymbol{\phi}$ and $\boldsymbol{V}$ which we obtain by augmenting the ordinary differential equation systems with the additional sensitivity equations. For an ordinary differential equation system of n_y equations with form $\dot{\boldsymbol{y}} = \boldsymbol{F}(\boldsymbol{y}, t, \boldsymbol{\theta})$, $\boldsymbol{y}(t_0) = \boldsymbol{y}_0(\boldsymbol{\theta})$ and n_θ parameters $\boldsymbol{\theta}$, the first- and second-order forward sensitivity equations are given by (5.1) and (5.2), respectively:

$$\frac{\partial\dot{\boldsymbol{y}}}{\partial\boldsymbol{\theta}} = \boldsymbol{F}_y\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\theta}} + \boldsymbol{F}_{\boldsymbol{\theta}} \quad\text{and}\quad \frac{\partial\boldsymbol{y}(t_0)}{\partial\boldsymbol{\theta}} = \frac{\partial\boldsymbol{y}_0}{\partial\boldsymbol{\theta}} \tag{5.1}$$

and

$$\left.\begin{aligned}\frac{\partial^2\dot{\boldsymbol{y}}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} &= [\boldsymbol{F}_y \otimes \boldsymbol{I}_{n_\theta}]\frac{\partial^2\boldsymbol{y}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} + \left[\boldsymbol{I}_{n_y} \otimes \frac{\partial\boldsymbol{y}^{\mathrm{T}}}{\partial\boldsymbol{\theta}}\right]\left[\boldsymbol{F}_{y,y}\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\theta}} + \boldsymbol{F}_{y,\boldsymbol{\theta}}\right] + \left[\boldsymbol{F}_{\boldsymbol{\theta},y}\frac{\partial\boldsymbol{y}}{\partial\boldsymbol{\theta}} + \boldsymbol{F}_{\boldsymbol{\theta},\boldsymbol{\theta}}\right]\\ \text{and}\qquad \frac{\partial^2\boldsymbol{y}(t_0)}{\partial\boldsymbol{\theta}^2} &= \frac{\partial^2\boldsymbol{y}_0}{\partial\boldsymbol{\theta}^2}.\end{aligned}\right\} \tag{5.2}$$

We use $\boldsymbol{F}_{\boldsymbol{\theta}}$ to denote the $n_y \times n_\theta$ matrix where its jth column is the partial derivatives of $\boldsymbol{F}$ with respect to θ_j. $\boldsymbol{F}_{\boldsymbol{\theta},\boldsymbol{y}}$ denotes the derivative of $\boldsymbol{F}_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{y}$ and is an $n_\theta \cdot n_y \times n_y$ matrix where its jth column is the partial derivatives of $\mathrm{vec}(\boldsymbol{F}_{\boldsymbol{\theta}}^{\mathrm{T}})$ with respect to y_j. I_{n_y} denotes the $n_y \times n_y$ identity matrix, $\otimes$ the Kronecker product and $\mathrm{vec}(\boldsymbol{A})$ an operator that creates a column vector by stacking the columns of matrix $\boldsymbol{A}$.

(b) Re-parametrization

In many problems, the parameters $\boldsymbol{\theta}$ can be constrained in certain parts of $\mathbb{R}^{n_\theta}$ where n_θ is the number of parameters. In models of chemical kinetics, for example, rate parameters must be positive and can differ by orders of magnitude. For the MCMC algorithms described in the previous section, we will need a re-parametrization in order to allow the algorithms to operate on an unbounded and unconstrained parameter space.

For the numerical simulations in §6, we use a $\log_{10}$ re-parametrization by introducing the variables $\check{\theta}_p = \log_{10}(\theta_p)$, $p \in [1, \ldots, n_\theta]$. To ensure that we sample from the correct posterior the joint density is scaled by the determinant of the Jacobian such that $p(\boldsymbol{X}|\check{\boldsymbol{\theta}})p(\check{\boldsymbol{\theta}})|\boldsymbol{J}(\check{\boldsymbol{\theta}})|$, where $\boldsymbol{J}(\check{\boldsymbol{\theta}})$ is an $n_\theta \times n_\theta$ diagonal matrix with elements $\boldsymbol{J}(\check{\boldsymbol{\theta}})_{p,p} = 10^{\check{\theta}_p}\log(10)$.

The gradient and Fisher information along with its partial derivatives follow from the chain rule as

$$\begin{aligned}\nabla_{\check{\boldsymbol{\theta}}}\mathcal{L}(\check{\boldsymbol{\theta}}) &= \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\boldsymbol{J}(\check{\boldsymbol{\theta}}),\\ \boldsymbol{I}(\check{\boldsymbol{\theta}}) &= \boldsymbol{J}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{J}(\check{\boldsymbol{\theta}})\\ \text{and}\qquad \frac{\partial\boldsymbol{I}(\check{\boldsymbol{\theta}})}{\partial\check{\theta}_p} &= 2\boldsymbol{J}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{I}(\boldsymbol{\theta})\frac{\partial\boldsymbol{J}(\check{\boldsymbol{\theta}})}{\partial\check{\theta}_p} + \boldsymbol{J}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\frac{\partial\boldsymbol{I}(\boldsymbol{\theta})}{\partial\theta_p}\boldsymbol{J}(\check{\boldsymbol{\theta}})\frac{\partial\theta_p}{\partial\check{\theta}_p}.\end{aligned}$$

(c) Choice of priors

In Bayesian statistics priors provide the means for incorporating existing knowledge for the parameters in question. The choice of a suitable prior distribution can be informed from knowledge about the process being modelled, the experimental design and the empirical observations. For example, we might want to restrict rate parameters in chemical kinetics from becoming very high since we assume from the experimental design that reactions are slow enough to be able to be observed. In some cases, the model itself can also guide the choice of the prior. For example, when a model is only defined for a certain range of values of the parameters, a prior restricting the parameters in that range should be used.

In the numerical simulations of the next section, we use independent normal priors for the parameters $\check{\boldsymbol{\theta}}$. Owing to the re-parametrization introduced earlier, this corresponds to a lognormal prior with base 10 for the parameters $\boldsymbol{\theta}$. This choice allows parameters to differ by several orders of magnitude while it ensures that they are strictly positive. Moreover, as noted by Girolami & Calderhead [22] the negative Hessian of the prior is added to the Fisher information in order to form the metric tensor used during MCMC sampling. This has the added benefit of regularizing the Fisher information when it is near-singular [9], although we have not observed such problems in the simulations presented here.

6. Numerical simulations

(a) Chemical kinetics

In this section, we consider two examples from chemical kinetics [14] and study the effect of the system size parameter on inference using MCMC. The first system consists of three species where an unstable monomer, S_1, can dimerize to an unstable dimer, S_2, which is then converted to a stable form, S_3. The reaction set for this system is

$$\begin{aligned} R_1 &: S_1 \xrightarrow{c_1} \emptyset, \\ R_2 &: 2S_1 \xrightarrow{c_2\Omega^{-1}} S_2, \\ R_3 &: S_2 \xrightarrow{c_3} 2S_1, \\ R_4 &: S_2 \xrightarrow{c_4} S_3, \end{aligned}$$

and the state of the system at time t will be denoted by $\boldsymbol{X}(t) = [S_1(t), S_2(t), S_3(t)]^{\mathrm{T}}$. The propensity functions, or state transition probabilities, are $\boldsymbol{f}(\boldsymbol{X},\boldsymbol{\theta}) = [c_1S_1(t), c_2\Omega^{-1}S_1(t)(S_1(t)-1)/2, c_3S_2(t), c_4S_3(t)]^{\mathrm{T}}$ and the corresponding state change matrix is

$$\boldsymbol{S} = \begin{pmatrix} -1 & -2 & 2 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6.1}$$

For our experiments, we will assume that initial conditions are known and set them to $S_1(t_0) = 5\Omega$, $S_2(t_0) = S_3(t_0) = 0$, $t_0 = 0$. Moreover, we will set the reaction rate parameters to $c_1 = 1$, $\hat{c}_2 = 2\Omega^{-1}$, $c_3 = 0.5$ and $c_4 = 0.04$. Note that we make explicit the relation between the system size and parameter $\hat{c}_2$ and we will infer rate c_2 up to a proportionality constant. For all the experiments, we simulate data using the SSA of Gillespie [1] for the time interval $t \in [0, 10]$ and we discretize such that $t_i - t_{i-1} = 0.1$. Each observation $\boldsymbol{X}(t_i)$ is obtained independently by simulating a trajectory from t_0 to t_i and keeping only the last state, discarding the rest of the trajectory. Moreover, for each time point t_i, we also simulate 10 independent observations. Since each observation is obtained by a different trajectory of the MJP we assume that initial conditions do not have a point mass; rather, for each trajectory we sample its initial condition from a Poisson with means $S_1(0), S_2(0)$ and $S_3(0)$.

We use the synthetic data to perform inference for the rate parameters $\boldsymbol{\theta} = (c_1, \hat{c}_2, c_3, c_4)^{\mathrm{T}}$ by drawing samples from the posterior

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{N} \prod_{r=1}^{10} \mathcal{N}[\mathbf{X}_r(t_i)|\Omega\boldsymbol{\phi}(t_i), \Omega\boldsymbol{V}(t_i)],$$

Table 1. Comparison of the minimum ESS and the time-normalized minimum ESS for different values of the system size parameter Ω of the decay–dimerization reaction model. Time-normalized ESS is given in parenthesis. Results are calculated from 10 000 posterior samples. MH, Metropolis–Hastings.

Ω	MH	SMMALA	RMHMC
1	121 (3.6)	150 (3.9)	245 (0.06)
2	226 (6.7)	2163 (57.2)	4775 (1.3)
5	132 (3.9)	3539 (93.6)	4618 (1.2)
10	180 (5.3)	3397 (89.8)	5954 (1.6)
100	214 (6.4)	3725 (98.5)	6066 (1.7)

where r indexes independent observations for the same time point. For all simulations in this paper, we assume that the means for the initial conditions are known. Following similar arguments to those for the derivation of the LNA in §3, namely that as the system approaches its thermodynamic limit transition densities become Gaussian, the initial conditions for the ordinary differential equation systems for the mean and variance of the transition densities are $\boldsymbol{\phi}(0) = \boldsymbol{X}(0)\Omega^{-1}$ and $\boldsymbol{V}(0) = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. In a more realistic scenario, the initial conditions must be included as additional parameters in $\boldsymbol{\theta}$. For all parameters, we used an independent lognormal prior with base 10, zero mean and one standard deviation, and chains are initialized by drawing a random sample from the prior. For the Metropolis–Hastings sampler, we set the initial proposal scale parameters to $\approx 1e^{-6}$ and automatically adapt them every 100 samples during the burn-in phase in order to achieve an acceptance rate of 25–30% [17]. The same adaptation strategy was followed for the SMMALA and RMHMC algorithm where the initial step size was also set to $\approx 1e^{-6}$ and was tuned in order to achieve acceptance rates of the order of 70–80% [22]. Finally, the number of leapfrog steps for RMHMC was fixed to 5. We have found that a burn-in period of 10 000–20 000 samples was adequate for all algorithms to converge to the stationary distribution.

Table 1 compares the minimum ESS and the time-normalized ESS obtained by all algorithms for different values of the system size parameter Ω. The SMMALA and RMHMC samplers use the gradients and the Fisher information of the approximate likelihood obtained by the LNA in order to make efficient proposals. As the system size increases and thus the LNA better approximates the true likelihood then mixing of the manifold MCMC algorithms improves. For this particular example, we can see that good mixing can be achieved even for very small systems with only ≈ 25 molecules ($\Omega = 5$). The Metropolis–Hastings sampler is not affected by the system size but its mixing is very poor in all cases. From the time-normalized ESS, we can also see that despite the improved mixing of RMHMC the computational cost is significant. On the contrary, the SMMALA provides a good trade-off between mixing efficiency and computational cost. Finally, table 2 reports the marginal posterior means and standard deviations for different values of Ω obtained by RMHMC. The marginal posteriors for parameters c_3 and c_4 with $\Omega \geq 5$ are also shown in figure 1. Results from the Metropolis–Hastings and SMMALA samplers are similar and are omitted. For small system sizes, we can observe that there is an increased bias of the Monte Carlo estimate while the posterior standard deviation is higher, reflecting the high degree of uncertainty around the mean. The bias, however, significantly reduces as the system size increases and for $\Omega \geq 5$ reasonable estimates can be obtained.

The second example from the chemical kinetics literature that we consider is the Schlögl reaction set,

$$R_1 : 2S_1 \xrightarrow{c_1\Omega^{-1}} 3S_1,$$

$$R_2 : 3S_1 \xrightarrow{c_2\Omega^{-2}} 2S_1,$$

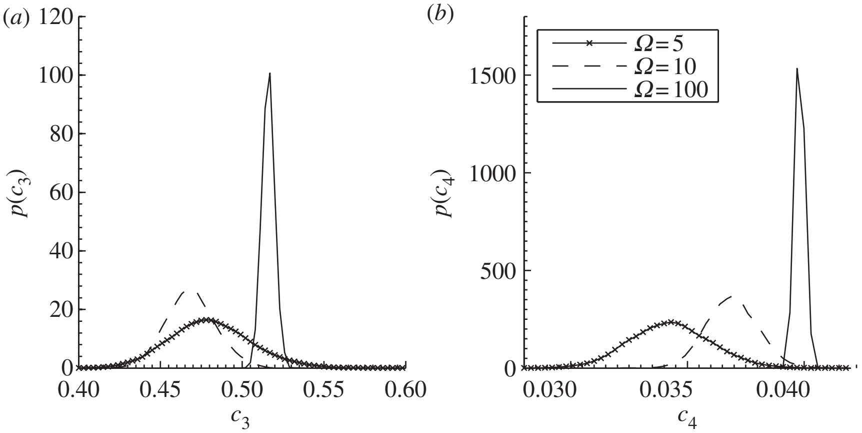

Figure 1. Marginal posteriors for parameters c_3 (*a*) and c_4 (*b*) for different values of Ω. Results are obtained by 10 000 posterior samples using RMHMC.

Table 2. Marginal posterior means and standard deviations calculated from the RMHMC chain for different values of the system size parameter Ω of the decay–dimerization reaction model. Note that parameter $\hat{c}_2$ is proportional to Ω. Results are calculated from 10 000 posterior samples.

Ω	c_1	$\hat{c}_2$	c_3	c_4
true	1	$2\Omega^{-1}$	0.5	0.04
1	0.88 (0.031)	1.72 (0.253)	0.39 (0.039)	0.003 (0.002)
2	1.3 (0.041)	0.69 (0.066)	0.35 (0.016)	0.014 (0.002)
5	0.93 (0.019)	0.39 (0.028)	0.48 (0.025)	0.034 (0.002)
10	1.0 (0.015)	0.18 (0.008)	0.47 (0.015)	0.037 (0.001)
100	0.99 (0.004)	0.01 (0.0002)	0.52 (0.004)	0.039 (0.0003)

$$R_3 : \emptyset \xrightarrow{c_3\Omega} S_1,$$

$$R_4 : S_1 \xrightarrow{c_4} \emptyset.$$

The corresponding state transition rates and state change matrix are given in equations (6.2) and (6.3), respectively. The state of the system consists only of the number of molecules of a single species $\boldsymbol{X}(t) = S_1(t)$,

$$\boldsymbol{S} = (1, \quad -1, \quad 1, \quad -1) \tag{6.2}$$

and

$$f(\boldsymbol{X}, \boldsymbol{\theta}) = \begin{pmatrix} c_1\Omega^{-1}\frac{1}{2}S_1(S_1-1) \\ c_2\Omega^{-2}\frac{1}{6}S_1(S_1-1)(S_1-2) \\ c_3\Omega \\ c_4 S_1 \end{pmatrix}. \tag{6.3}$$

The system is known to have two stable states that appear at different times depending on the size of the system. Wallace *et al.* [14] have shown that the LNA fails to provide a reasonable approximation of this system even for large concentration numbers. Their numerical experiments demonstrate that the LNA can only approximate one of the two modes depending on the initial conditions. Here, our aim is to show that using the LNA to obtain an approximate posterior

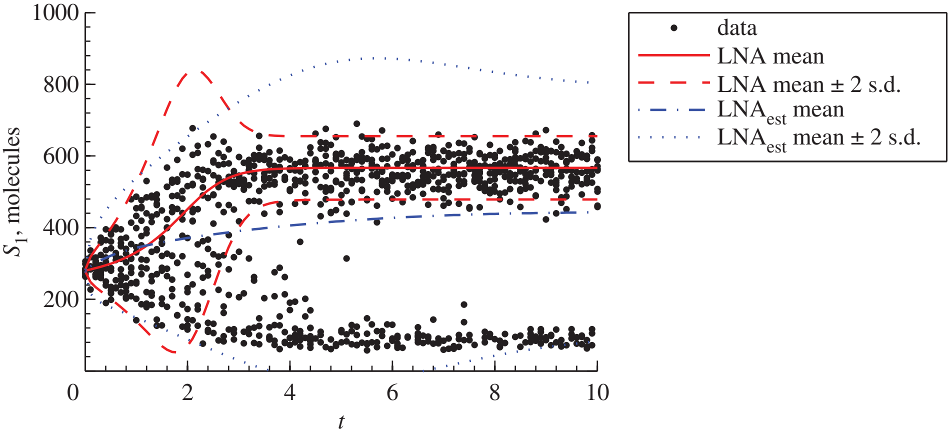

Figure 2. Simulated time point data using SSA for the Schlögl reaction set and LNA predictions. Dots correspond to simulated data. The solid and dashed lines correspond to the LNA prediction for the means and standard deviations using the true parameters. Dotted lines correspond to the LNA predictions using the posterior means for the rate parameters. (Online version in colour.)

over the unknown reaction rate constants can be very misleading for bi-stable systems. Using the resulting posterior means for the reaction rates gives us an LNA that fails to approximate any of the two stable modes.

To demonstrate, we follow the same experimental procedure as in the previous example. That is, we simulate data using the SSA for the time interval $t_i \in [0, 10]$, $t_i - t_{i-1} = 0.1$ with fixed rate parameters and then use these data for posterior inference of the rate parameters using MCMC. Values for the true rate parameters and initial conditions were set as in Wallace *et al.* [14]. Namely, $c_1 = 0.003$, $c_2 = 0.0001$, $c_3 = 200$, $c_4 = 3.5$ and $\boldsymbol{X}(t_0) = 280\Omega$, where Ω was fixed to 1. After 10 000 burn-in samples all samplers converged to a posterior distribution with mean

$$\mathrm{E}_{p(\boldsymbol{\theta}|\mathbf{X})}[\boldsymbol{\theta}] \approx (0.130, 3.3\mathrm{e}^{-4}, 3.5\mathrm{e}^{+3}, 26.22)^{\mathrm{T}}$$

and variance

$$\mathrm{var}_{p(\boldsymbol{\theta}|\mathbf{X})}[\boldsymbol{\theta}] \approx (1.2\mathrm{e}^{-4}, 8.2\mathrm{e}^{-10}, 8.6\mathrm{e}^{+4}, 4.53)^{\mathrm{T}}.$$

The LNA obtained by using the posterior means for the rate constants is shown in figure 2 along with the data obtained by the SSA and the LNA using the true values for the rate constants. We can see that the LNA obtained by the posterior means fails to approximate any of the two modes. Rather, it approximates the empirical mean and variance of the data.

(b) Single-gene expression

Finally, to illustrate the applicability of the methodology to systems biology we also consider a simplified model for the biochemical reactions involved in the expression of a single gene to a protein. The model presented in this section is the same as the model used in the study of Komorowski *et al.* [15] and we adopt the same notation in order to make comparisons easier. Gene expression is modelled in terms of three biochemical species: DNA, mRNA and protein; and four chemical reactions or state transitions: transcription, mRNA degradation: translation and protein degradation. The model can be written in chemical reaction notation as

$$R_1 : \mathrm{DNA} \xrightarrow{k_{\mathrm{R}}(t)} \mathrm{DNA} + R,$$

$$R_2 : R \xrightarrow{\gamma_{\mathrm{R}}} \emptyset,$$

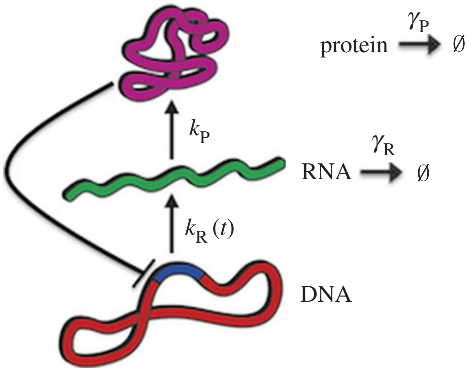
protein $\xrightarrow{\gamma_P}$ Ø
k_P
RNA $\xrightarrow{\gamma_R}$ Ø
$k_R\,(t)$
DNA

Figure 3. Schematic of the auto-regulatory gene expression model with a negative feedback loop. A gene is transcribed into mRNA, which is translated to a protein that suppresses gene transcription. (Online version in colour.)

$$R_3 : R \xrightarrow{k_P} R + P,$$

$$R_4 : P \xrightarrow{\gamma_P} \emptyset.$$

The system state at time t is $\boldsymbol{X}(t) = [R(t), P(t)]^{\mathrm{T}}$, where $R(t)$ and $P(t)$ are the number of mRNA and protein molecules, respectively. The corresponding state-dependent transition rates are $\boldsymbol{f}(\boldsymbol{X}, t) = [k_R(t), \gamma_R R(t), k_P R(t), \gamma_P P(t)]^{\mathrm{T}}$, where γ_R, k_P and γ_P are unknown reaction rate constants. $k_R(t)$ is the time-dependent transcription rate of the gene, which, for the purposes of this section, is modelled as

$$k_R(t) = b_0 \exp(-b_1(t - b_2)^2) + b_3,$$

where all the b_is are also unknown parameters controlling gene transcription. This corresponds to a transcription rate that, owing to some stimulus (experimental or environmental), increases for $t < b_2$ and then drops towards the base line b_3 for $t > b_2$. Finally, the state change matrix for this set of reactions is given in equation (6.4),

$$\boldsymbol{S} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{6.4}$$

As in the study of Komorowski *et al*. [15], we also consider a nonlinear extension of this model in which the transcription rate of the gene $k_R(t)$ is a function of the protein concentration that the gene is transcribed to. This is modelled using a Hill function

$$\hat{k}_R(t, P) = \frac{k_R(t)}{(1 + (P/H)^{n_H})},$$

where for the experiments in this section we will set $H = b_3 k_P/(2\gamma_R\gamma_P)$ and $n_H = \frac{1}{2}$, making the protein an inhibitor of mRNA transcription. A schematic of this model is shown in figure 3. For the rest of this section, we will refer to this model as the auto-regulatory single-gene expression model.

Using the transition probabilities $\boldsymbol{f}(\boldsymbol{X}, t)$ and matrix $\boldsymbol{S}$, we simulate synthetic data using the SSA [1] and sample at discrete time points. Values for the unknown rate constants and the parameters controlling gene transcription are shown in table 3. The time interval is taken to be $t_i \in [0, 25]$, while the interval between two observations $t_i - t_{i-1} = 0.25$. Each time point is sampled from an independent trajectory by starting the SSA from t_0 and simulating up to t_i, keeping only the state $\boldsymbol{X}(t_i)$ and discarding the rest of the trajectory. This resembles the experimental conditions often encountered in biology where, in order to make an observation, the sample has to be 'sacrificed'. Finally, for each time point, we also generate 10 independent observations from different trajectories. Initial conditions $\boldsymbol{X}(t_0)$ are simulated from a Poisson

Table 3. Marginal posterior means and standard deviations for the parameters of the single-gene expression model using simulated data. The ESS is calculated for chains of 10 000 samples after a burn-in period of 10 000 iterations with initial parameters randomly sampled from the prior. Average acceptance rate (AR) and sampler parameters are shown in parenthesis. Note that for the Metropolis–Hastings sampler a different proposal is used for each parameter. The prior for all parameters was $\log_{10} \mathcal{N}(0.0, 2.0)$.

			single-gene expression model				
parameters	γ_R	γ_P	k_P	b_0	b_1	b_2	b_3
true values	0.44	0.52	10.0	15.0	0.40	7.0	3.0
			Metropolis–Hastings				
(AR)	(0.28)	(0.33)	(0.30)	(0.34)	(0.29)	(0.29)	(0.34)
(ϵ)	(0.013)	(0.007)	(0.008)	(0.022)	(0.056)	(0.007)	(0.016)
mean	0.45	0.54	10.54	14.86	0.39	7.03	3.14
s.d.	0.017	0.017	0.336	0.509	0.029	0.056	0.149
ESS	42	34	34	149	117	58	44
ESS/time	1.42	1.15	1.15	5.05	3.96	1.96	1.49
			SMMALA (AR $= 0.79$, $\epsilon = 1.05$)				
mean	0.45	0.54	10.57	14.88	0.39	7.04	3.17
s.d.	0.018	0.016	0.306	0.537	0.030	0.053	0.152
ESS	2891	2911	2958	2787	3310	3183	2878
ESS/time	83.79	84.37	85.73	80.78	95.94	92.26	83.42
			manifold HMC (AR $= 0.84$, $\epsilon = 0.91$, $L = 5$)				
mean	0.46	0.54	10.57	14.95	0.39	7.04	3.18
s.d.	0.018	0.015	0.300	0.555	0.030	0.052	0.153
ESS	7731	8238	8304	7160	7380	7791	7950
ESS/time	0.52	0.55	0.56	0.48	0.49	0.52	0.53
			auto-regulatory single-gene expression model				
			Metropolis–Hastings				
(AR)	(0.26)	(0.36)	(0.31)	(0.33)	(0.24)	(0.30)	(0.35)
(ϵ)	(0.028)	(0.012)	(0.016)	(0.071)	(0.231)	(0.019)	(0.029)
mean	0.4360	0.52	10.40	14.61	0.40	6.82	3.13
s.d.	0.016	0.018	0.424	1.089	0.076	0.090	0.142
ESS	201	71	73	465	339	420	239
ESS/time	6.12	2.16	2.22	14.17	10.33	12.80	7.28
			SMMALA (AR $= 0.71$, $\epsilon = 1.17$)				
mean	0.43	0.52	10.44	14.24	0.38	6.82	3.12
s.d.	0.016	0.018	0.422	1.125	0.075	0.091	0.142
ESS	2990	3270	3454	3124	3164	3316	3195
ESS/time	76.86	84.06	88.79	80.30	81.33	85.24	82.13
			manifold HMC (AR $= 0.82$, $\epsilon = 0.91$, $L = 5$)				
mean	0.43	0.52	10.43	14.52	0.40	6.82	3.13
s.d.	0.016	0.017	0.412	1.158	0.078	0.089	0.144
ESS	6532	6593	6614	5112	5384	6595	6642
ESS/time	0.41	0.41	0.41	0.32	0.34	0.41	0.42

distribution with means b_3/γ_R and $b_3 k_P/(\gamma_R \gamma_P)$ for the mRNA and protein molecules, respectively. The system-size parameter Ω is considered to be unknown and for this experiment is set to 1 such that concentrations are equal to the number of molecules. Figure 4*a*,*b* shows data simulated from this process from the single-gene expression model as well as the LNA prediction. Simulated data for the auto-regulatory model are presented in figure 4*c*,*d*.

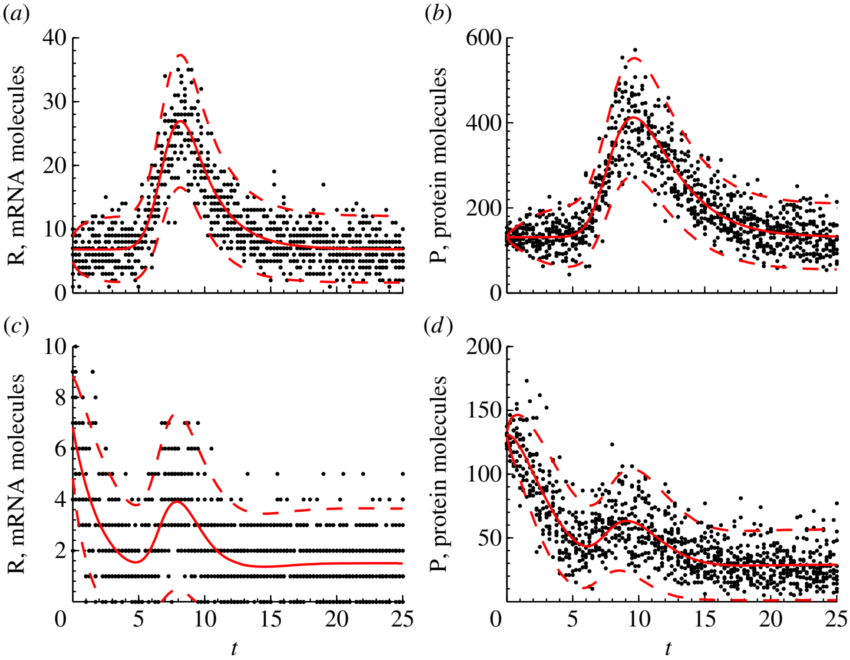

Figure 4. Data simulated from the single-gene expression model using SSA. (*a*,*b*) for the linear model and (*c*,*d*) for the auto-regulatory model. Dots correspond to 10 independent draws for each time point. The bold line is the mean predicted by LNA with the true model parameters and the dashed lines are the ±2 × s.d. predicted by LNA. (*a*,*c*) The mRNA molecules and (*b*,*d*) the protein. (Online version in colour.)

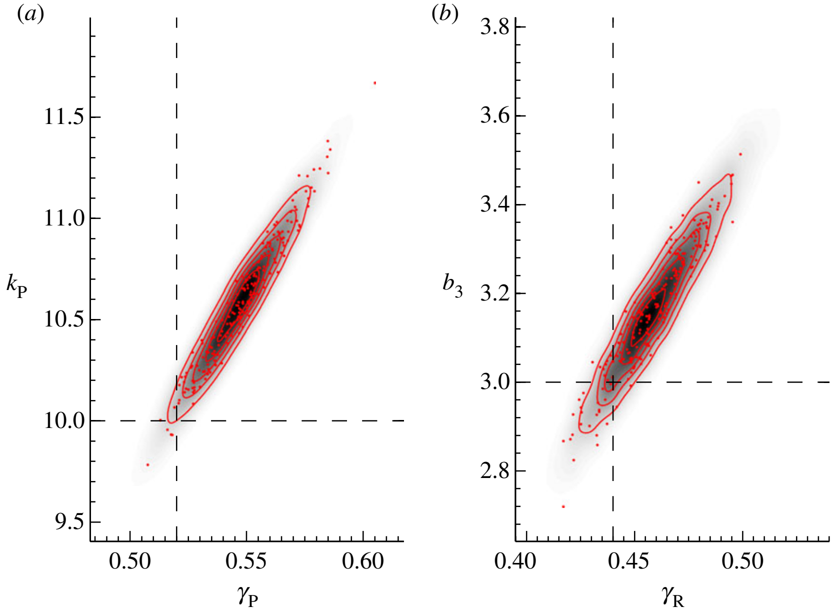

Figure 5. (*a*) Marginal joint posterior for parameters γ_P, k_P, and (*b*) γ_R, b_3 for the single-gene expression model. Dashed lines are the true values used to generate the synthetic data. Dots are samples from the posterior. Iso-contours and shaded regions are obtained by kernel density estimation using posterior samples. (Online version in colour.)

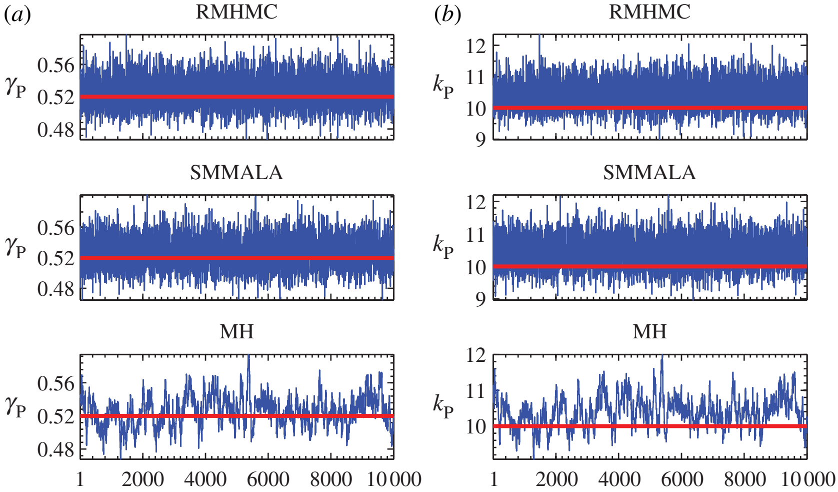

Figure 6. Example trace plots from the auto-regulatory gene expression model for parameters (*a*) γ_P and (*b*) k_P. Bold solid line denotes the true values. MH, Metropolis–Hastings. (Online version in colour.)

We use the simulated data to infer the unknown parameters $\boldsymbol{\theta} = (\gamma_R, k_P, \gamma_P, b_0, b_1, b_2, b_5)^T$ by sampling using MCMC from the LNA posterior

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i=1}^{N} \prod_{r=1}^{R} \mathcal{N}[\mathbf{X}_r(t_i)|\boldsymbol{\phi}(t_i), \mathbf{V}(t_i)],$$

where r indexes independent samples for the same time point and $R = 10$.

Table 3 summarizes the results from the MCMC chains for the two models of gene expression. Firstly, we can see that despite the relatively small number of molecules in both systems the LNA provides very accurate estimates for the true parameters. Moreover, we can see that the mixing of the Metropolis–Hastings sampler is very poor for both models while the RMHMC algorithm and the SMMALA perform very well. This can be explained by the strong correlations between parameters in the posterior distribution preventing the Metropolis–Hastings sampler from making sufficiently large proposals. For example, the parameters k_P, γ_P control mRNA translation and protein degradation, respectively. The concentration of protein molecules is directly affected by the two rates and they are expected to be heavily correlated. In figure 5*a*,*b*, we show the marginal joint posterior for parameters k_P, γ_P and γ_R, b_3 for the single-gene expression model which exhibit very strong positive correlation. Finally, figure 6 compares the trace plots obtained from Metropolis–Hastings, SMMALA and RMHMC for parameters γ_P and k_P of the auto-regulatory gene expression model.

7. Conclusions and future work

Bayesian inference for MJPs is a challenging problem that has many important practical applications. Previous research [6] has shown that, although exact inference is possible, the computational cost and the auto-correlation of the Markov chains are such that they limit its applicability to small problems. The main problem stems from the requirement to simulate the MJP for the trajectory of the system between discrete observations. Golightly & Wilkinson [13] have shown that by considering a diffusion approximation the simulation can be performed in a much more efficient manner. In this paper, we considered the LNA, which only requires a system of ordinary differential equations to be simulated while the stochastic fluctuations have an exact analytic solution. The LNA is valid only when the system is sufficiently close to its thermodynamic limit, a condition that is also required for the diffusion approximation.

Previous research on the LNA [15] has focused on the Metropolis–Hastings sampler. We have demonstrated here that when the posterior distribution exhibits strong correlation between parameters then the Metropolis–Hastings sampler has strong auto-correlations. Such correlations are very common for chemical reaction and gene regulatory systems. The Riemann manifold MCMC algorithms we considered in this work exploit the geometric structure of the target posterior in order to design efficient proposal mechanisms. In particular, the SMMALA is a conceptually simple algorithm that provides a good trade-off between computational cost and sample auto-correlation.

Although the problems considered in this work are relatively small, but certainly non-trivial, we believe that the proposed methodology is applicable for larger and more complex systems. The systems we studied in this paper all have a linear dependence on the unknown parameters and we have not observed any local modes in our simulations. The analysis of such systems is the subject of ongoing work. Moreover, in real applications, it is not possible to observe the populations of all species and there is an additional measurement error term. Extension of the LNA to handle such cases is straightforward (e.g. [15]); however, the effect of partial observations and measurement error on the MCMC inference is something that needs to be studied in more detail.

We would like to thank the guest editor and the anonymous reviewers for their useful comments that have helped us to improve the presentation of the paper. This work was funded by the FP7 EC project: ASSET—Analysing and Striking the Sensitivities of Embryonal Tumours, reference no. 259348. M.A.G. is funded by Engineering and Physical Sciences Research Council (EPSRC) Advanced Research Fellowship EP/E052029/2.

References

1. Gillespie DT. 2007 Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55. (doi:10.1146/annurev.physchem.58.032806.104637)
2. Komorowski M, Costa MJ, Rand DA, Stumpf MPH. 2011 Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl Acad. Sci. USA* **108**, 8645–8650. (doi:10.1073/pnas.1015814108)
3. Spencer M, Susko E. 2005 Continuous-time Markov models for species interactions. *Ecology* **86**, 3272–3278. (doi:10.1890/05-0029)
4. Adas A. 1997 Traffic models in broadband networks. *Commun. Mag. IEEE* **35**, 82–89. (doi:10.1109/35.601746)
5. Gillespie DT. 2005 A rigorous derivation of the chemical master equation. *Phys. A: Stat. Mech. Appl.* **188**, 404–425. (doi:10.1016/0378-4371(92)90283-V)
6. Boys RJ, Wilkinson DJ, Kirkwood TB. 2008 Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comp.* **18**, 125–135. (doi:10.1007/s11222-007-9043-x)
7. Van Kampen NG. 1992 *Stochastic processes in physics and chemistry*, 3rd edn. Amsterdam, The Netherlands: North-Holland.
8. Xu T *et al.* 2010 Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Signal.* **3**, ra20. (doi:10.1126/scisignal.2000517)
9. Calderhead B, Girolami M. 2011 Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus* **1**, 821–835. (doi:10.1098/rsfs.2011.0051)
10. van Kampen NG. 1982 The diffusion approximation for Markov processes. In *Thermodynamics & kinetics of biological processes* (eds I Lamprecht, AI Zotin), pp. 181–195. New York, NY: Walter de Gruyter & Co.
11. Gillespie DT. 2000 The chemical Langevin equation. *J. Chem. Phys.* **113**, 297–306. (doi:10.1063/1.481811)
12. Roberts GO, Stramer O. 2001 On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika* **88**, 603–621. (doi:10.1093/biomet/88.3.603)
13. Golightly A, Wilkinson DJ. 2011 Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1**, 807–820. (doi:10.1098/rsfs.2011.0047)

14. Wallace E, Gillespie D, Sanft K, Petzold L. 2012 The linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET Syst. Biol.* **6**, 102–115. (doi:10.1049/iet-syb.2011.0038)
15. Komorowski M, Finkenstadt B, Harper C, Rand D. 2009 Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10**, 343. (doi:10.1186/1471-2105-10-343)
16. Robert CP, Casella G. 2005 *Monte Carlo statistical methods*. Springer Texts in Statistics. New York, NY: Springer.
17. Roberts GO, Gelman A, Gilks WR. 1997 Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120. (doi:10.1214/aoap/1034625254)
18. Roberts GO, Stramer O. 2003 Langevin diffusions and Metropolis–Hastings algorithms. *Methodol. Comp. Appl. Probab.* **4**, 337–358. (doi:10.1023/A:1023562417138)
19. Duane S, Kennedy AB, Pendleton JB, Roweth D. 1987 Hybrid Monte Carlo. *Phys. Lett. B* **195**, 216–222. (doi:10.1016/0370-2693(87)91197-X)
20. Radford MN. 1993 Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1. Department of Computer Science, University of Toronto, Canada.
21. Roberts GO, Rosenthal JS. 1998 Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. B* (*Stat. Methodol.*) **60**, 255–268. (doi:10.1111/1467-9868.00123)
22. Girolami M, Calderhead B. 2011 Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. B* (*Stat. Methodol.*) **73**, 123–214. (doi:10.1111/j.1467-9868.2010.00765.x)
23. Ferm L, Lötstedt P, Hellander A. 2008 A hierarchy of approximations of the master equation scaled by a size parameter. *J. Sci. Comp.* **34**, 127–151. (doi:10.1007/s10915-007-9179-z)
24. Tanner MA, Wong WH. 1987 The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**, 528–540. (doi:10.2307/2289457)
25. Green PJ. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)
26. Gillespie DT. 2009 Deterministic limit of stochastic chemical kinetics. *J. Phys. Chem. B* **113**, 1640–1644. (doi:10.1021/jp806431b)
27. Heron EA, Finkenstädt B, Rand DA. 2007 Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics* **23**, 2596–2603. (doi:10.1093/bioinformatics/btm367)
28. Oksendal B. 1992 *Stochastic differential equations: an introduction with applications*, 3rd edn. New York, NY: Springer.
29. Nolan T, Hands RE, Bustin SA. 2006 Quantification of mRNA using real-time RT-PCR. *Nat. Protocols* **1**, 1559–1582. (doi:10.1038/nprot.2006.236)
30. Haario H, Saksman E, Tamminen E. 2005 Componentwise adaptation for high dimensional MCMC. *Comp. Stat.* **20**, 265–273. (doi:10.1007/BF02789703)
31. Andrieu C, Thoms J. 2008 A tutorial on adaptive MCMC. *Stat. Comp.* **18**, 343–373. (doi:10.1007/s11222-008-9110-y)
32. Kühnel W. 2005 *Differential geometry: curves-surfaces-manifolds*, vol. 2. Student Mathematical Library. Providence, RI: American Mathematical Society.
33. Haario H, Saksman E, Tamminen J. 2001 An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242. (doi:10.2307/3318737)
34. Amari S-I, Nagaoka H. 2000 *Methods of information geometry*, vol. 191. Translations of Mathematical Monographs. Oxford, UK: Oxford University Press.