Inference for Reaction Networks Using the Linear Noise Approximation

Paul Fearnhead,¹ Vasilieos Giagos,² and Chris Sherlock¹,*

¹Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, U.K.
²School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, U.K.
*email: c.sherlock@lancaster.ac.uk

Summary. We consider inference for the reaction rates in discretely observed networks such as those found in models for systems biology, population ecology, and epidemics. Most such networks are neither slow enough nor small enough for inference via the true state-dependent Markov jump process to be feasible. Typically, inference is conducted by approximating the dynamics through an ordinary differential equation (ODE) or a stochastic differential equation (SDE). The former ignores the stochasticity in the true model and can lead to inaccurate inferences. The latter is more accurate but is harder to implement as the transition density of the SDE model is generally unknown. The linear noise approximation (LNA) arises from a first-order Taylor expansion of the approximating SDE about a deterministic solution and can be viewed as a compromise between the ODE and SDE models. It is a stochastic model, but discrete time transition probabilities for the LNA are available through SDEs, but it has the important advantage that un-sensible estimates of the reaction rates. However inference for SDE models is non-trivial, as the transition density of general SDEs is unknown.

Inference under the jump Markov process is possible only for networks which involve few species, few reactions, and not “too many” transitions between observations (e.g., Boys et al., 2008; Amrein and Künsch, 2012). For most systems it is necessary to approximate the evolution of the process to make inference computationally feasible. Often this will involve approximating the evolution through a system of ordinary differential equations (ODEs, e.g., Jones, Plank, and Sleeman, 2010) or stochastic differential equations (SDEs, e.g., Wilkinson, 2006). Models based on ODEs are only appropriate for very large systems for which the stochasticity in the evolution is small. For medium-size systems Wilkinson (2006) shows that SDE models are more appropriate and can lead to sensible estimates of the reaction rates. However inference for SDE models is non-trivial, as the transition density of general SDEs is unknown.

In this article we use an alternative approximation, known as the linear noise approximation (LNA) (van Kampen, 1997; Řízek and Ehrenberg, 2003; Hayot and Jayaprakash, 2004; Ferm et al., 2008). The LNA is obtained through first approximating the dynamics by a system of ODEs and then modeling the evolution of the state about the deterministic solution of the ODE, through a linear SDE. Simulations suggest that this approach has similar accuracy to modeling the system directly through SDEs, but it has the important advantage that under the LNA the stochastic model for the states is a Gaussian network allowing that observations may contain noise and may be of only a subset of the species in the system. Inference for very large systems for which the stochasticity in the evolution is small. For medium-size systems Wilkinson (2006) shows that SDE models are more appropriate and can lead to sensible estimates of the reaction rates. However inference for SDE models is non-trivial, as the transition density of general SDEs is unknown.

1. Introduction

Reaction networks are used to model a wide variety of real-world phenomena; they describe a probabilistic mechanism for the joint evolution of one or more populations of species. These species may be biological species, such as a range of different proteins (e.g., Golightly and Wilkinson, 2005, 2008; Boys, Wilkinson, and Kirkwood, 2008; Ferm, Löstedt, and Hellander, 2008; Proctor et al., 2005), animal species, such as predators and their prey (e.g., Boys et al., 2008; Ferm et al., 2008), or interacting groups of humans or animals such as those infected with a particular disease, those susceptible to the disease and those who have recovered from it (e.g., Andersson and Britton, 2000; Ball and Neal, 2008; Jewell, Keeling, and Roberts, 2009).

The evolution of these networks is most naturally modeled via a continuous-time jump Markov process. The current state of the system is encapsulated in a vector giving the numbers of each species that are present. The evolution of the state is described by a series of reactions, such as the interaction of two copies of a protein producing a dimer of that protein; or an interaction between an infected individual and a susceptible individual resulting in the susceptible becoming infected. Occurrences of a given reaction are modeled as a Poisson process, the rate of which depends on the current state of the system. Interest lies in inferring the parameters that govern the rate of each reaction from data on the evolution of the system.

This article concerns inference for the rate parameters and prediction of the future state in discretely observed reaction networks allowing that observations may contain noise and may be of only a subset of the species in the system. Inference for the joint evolution of one or more populations of species.
process. Transition densities are therefore Gaussian, and their mean and covariance can be obtained through solving a system of differential equations. Using the LNA can therefore be more accurate than modeling the system via ODEs, and it is substantially easier to perform inference than under a general SDE model.

The structure of the article is as follows. The next section provides more information on reaction networks and details two particular examples that will be revisited: the Lotka–Volterra predator–prey system, and an autoregulatory gene network. Section 3 examines the different possible approximations to the evolution of reaction networks: ODE approximation; SDE approximation and the LNA. Section 4 shows how we can calculate likelihoods using a “restarting” LNA for a range of observation models, and suggests a simple way of embedding this within MCMC to perform Bayesian inference. We evaluate the use of the LNA empirically on both simulated and real data. An alternative use of a (non-restarting) LNA for inference on reaction networks has previously been suggested by Komorowski et al. (2009). In Section 5 we compare our approach with that of Komorowski et al. (2009), with the simple ODE approximation, and with the SDE-based algorithm of Golightly and Wilkinson (2005). In Section 6 we use the LNA to analyze Google Flu Trends data from New Zealand, comparing the accuracy of week-ahead predictions with those from the recent approach of Dukic, Lopes, and Polson (2012).

2. Reaction Networks

Consider a general reaction of the form $B + C \rightarrow D$, where the number of elements of species $B$ and $C$ are respectively $X_B$ and $X_C$ and where the elements are distributed uniformly at random throughout some volume of space. The reaction occurs with some fixed probability whenever an element of species $B$ is within some “reaction distance” of an element of $C$; occurrences of the reaction may therefore be modeled as a Poisson process. With further, system dependent, assumptions, the rate, $h$, of the process is proportional to $X_B X_C$. Applying a similar argument, the rate of a reaction such as $B \rightarrow C$ is simply proportional to $X_B$ and the rate of $2B \rightarrow C$ is proportional to $0.5 X_B (X_B - 1)$. For a fuller discussion of mass-action kinetics see, for example, Gillespie (2005).

Consider now a network of $n$ such reactions each involving at least one of the $n_t$ species in the population. The dynamics of this model can be described by a vector of rates of the reactions together with a matrix which describes the effect of each reaction on the state. We denote by $h$ the $n_t$-vector of reaction rates. Now define $A_i$ be the net effect on species $j$ of a single occurrence of reaction $i$: so $A_{ij} = 0$ means that the number of species $j$ is unaffected by reaction $i$, whereas $A_{ij} = 1$ (or $-1$) means the number of species $j$ will increase (or decrease) by $1$. The $n_t \times n_t$ matrix $A$ is known as the net effect matrix. An equivalent way of defining the effect of a set of reactions is via the stoichiometry matrix, $A'$, where throughout this article, ’ denotes the transpose of a matrix.

Example 1: The Lotka Volterra model

The Lotka–Volterra model (e.g., Wilkinson, 2006) describes a population of two competing species: predators which die with rate $\theta_2$ and reproduce with rate $\theta_1$ by consuming prey which reproduce with rate $\theta_4$. In its simplest form the probabilistic system is defined by:

$$R_1 : \text{Pred} + \text{Prey} \rightarrow 2\text{Pred};$$
$$R_2 : \text{Pred} \rightarrow \phi; \quad R_3 : \text{Prey} \rightarrow 2\text{Prey}. $$

Denoting the number of Pred by $X_1$ and the number of Prey by $X_2$ gives the vector of reactions rates and the net effect matrix, respectively as:

$$h := (\theta_1 X_1 X_2, \theta_2 X_1, \theta_3 X_2) \quad \text{and} \quad A' = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix}. $$

Example 2: Autoregulatory gene network

The following system describes the self-regulating production of a protein, $P$, and its dimer, $P_2$. The system is analyzed in Golightly and Wilkinson (2005) and is also discussed in Wilkinson (2006), while a similar system is analyzed in Golightly and Wilkinson (2008). Reactions $R_1$ and $R_2$ describe the reversible process whereby the protein dimer $P_2$ binds to the gene (which we denote as DNA) and thereby inhibits the production, by reactions $R_3$ and $R_4$, of the protein, $P$. Dimerization of the protein and the reverse reaction are described by Reactions $R_5$ and $R_6$, while $R_7$ and $R_8$ describe the destruction of the protein and of the enzyme RNA-polymerase, which is denoted RNA.

$$R_1 : \text{DNA} + P_2 \rightarrow \text{DNA} \cdot P_2$$
$$R_2 : \text{DNA} \rightarrow \text{DNA} + P$$
$$R_3 : 2P \rightarrow P_2$$
$$R_4 : \text{RNA} \rightarrow 0$$
$$R_5 : P_2 \rightarrow 2P$$
$$R_6 : P \rightarrow 0.$$

From the reactions, the total, $k$, of the number of DNA and DNA·$P_2$ molecules is fixed throughout the evolution of the system. Denoting the number of molecules of DNA, RNA, $P$, and $P_2$ as $X_1$, $X_2$, $X_3$, and $X_4$, respectively, therefore leads to a reaction rate vector of $h := (\theta_{11} X_1, \theta_{12} (k - X_1), \theta_{21} X_1, \theta_{22} X_2, \theta_{31} X_3 (X_3 - 1)/2, \theta_{32} X_4, \theta_{33} X_2, \theta_{34} X_3)$. The net effect matrix for this example is $A$, where

$$A' = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}. $$

Further examples, of a one- and two-island epidemic model, are detailed in Web Appendix A.

3. Approximations for Network Evolution

We first consider the ODE and SDE approximations to the true process, and then sketch the justification for the LNA.

It will be helpful to denote the $n_t$-vector holding the number of molecules of each species by $X$ and to define the $n_t \times n_t$ reaction rate matrix $H := \text{diag}(h)$.

3.1. The ODE and SDE Approximations

In an infinitesimal time $dt$ the mean and variance of the change in $X$ due to all of the $n_t$ independent Poisson
processes can be calculated as (e.g., Wilkinson, 2006):

$$E[\Delta X(t)] = A'h \, dt, \quad \text{Var} [\Delta X(t)] = A'H A \, dt.$$  

The ODE approximation to the evolution ignores the stochasticity of the model and is based solely on the expected change in the mean. This gives the following differential equation

$$\frac{dX(t)}{dt} = A'h(X(t), \theta).$$

The SDE approximation models stochasticity through

$$dX(t) = A'h(X(t), \theta) \, dt + \sqrt{AH(X(t), \theta)A} \, dW(t),$$

where the matrix $\sqrt{AH(X(t), \theta)A}$ is any (without loss of generality, $n_x \times n_x$) matrix square root, such as that obtained by Cholesky decomposition, and $W(t)$ is Brownian motion.

The ODE model is deterministic, and fitting the model generally involves estimating both the initial condition and parameter values that give the best fit to the data. Often the fit to the data is quantified by the sum of the square residuals (see Ramsay et al., 2007, and references therein).

There are a range of methods for estimating parameters of an SDE model (see, e.g., Sreenivas, 2004). Recently, there has been much research on how to implement likelihood-based methods (e.g., Elerian, Chib, and Shephard, 2002; Durham and Gallant, 2002; Beskos et al., 2006; Ait-Sahalia, 2008). Generally, however, the SDE model will not lead to a tractable distribution for $X(t)$ given $X_0$, and hence these models have an intractable likelihood. To overcome this complication it is common to approximate the transition density of the SDE, for example by the Euler approximation (Kloeden and Platen, 1992). The Euler approximation is only accurate over small time-intervals. The implementation of these methods therefore involves discretizing time between each observation, and using computationally intensive methods that impute values of the state at both the observation times and the grid of times between each observation. For example, Golightly and Wilkinson (2005) implement such a method with an MCMC scheme, and Golightly and Wilkinson (2006) within a sequential Monte Carlo algorithm. There is a considerable computational overhead in implementing these methods which increases with the fineness of the grid of time-points between observations, and this has led to much research on efficient MCMC and other methods. See Roberts and Stramer (2001) for a discussion of how the fineness of the grid can affect mixing of the MCMC and, for example, Golightly and Wilkinson (2008) for details of more efficient MCMC approaches.

### 3.2. The Linear Noise Approximation

The LNA first appeared as a functional central limit law for density dependent processes; see Kurtz (1970, 1971) for the technical conditions. It approximates the dynamics of the network by an SDE which has tractable transition densities between observation times; inference therefore does not require any data augmentation (van Kampen, 1997).

While Kurtz (1970, 1971) justify our use of the LNA it will be more helpful in the present context to consider the LNA as a general approximation to the solution to an SDE, and then apply this to the SDE model derived in the previous section. The idea of the LNA is that we partition $X(t)$ into a deterministic path, $\eta(t)$, and a stochastic perturbation from this path. Under the assumption that the perturbation is “small” relative to the deterministic path the distribution of an approximate solution at any given time point is found by solving a series of ODEs. In our applications the deterministic path is just the solution of the ODE model introduced in the previous section. Here we provide a short heuristic motivation of the approximation; for a more rigorous derivation and more detailed discussion the reader is referred to Ferm et al. (2008).

Consider the general SDE for vector $X$ of length $n_x$

$$dX(t) = a(X(t)) \, dt + \epsilon S(X(t)) \, dW(t), \quad (1)$$

with initial condition $X(0) = X_0$. Let $\eta(t)$ be the (deterministic) solution to

$$\frac{d\eta}{dt} = a(\eta) \quad (2)$$

with initial value $\eta_0$. We assume that over the time interval of interest $||X - \eta||$ is $O(\epsilon)$. Set $M(t) = (X(t) - \eta(t))/\epsilon$ and use a Taylor expansion of $a$ and $S$ about $\eta(t)$ in (1). Collecting terms of $O(\epsilon)$ gives

$$dM(t) = F(t)M(t) \, dt + S(t) \, dW(t), \quad (3)$$

where $F$ is the $n_x \times n_x$ matrix with components

$$F_{ij}(t) = \frac{\partial a_i}{\partial x_j} \bigg|_{\eta(t)}, \quad \text{and} \quad S(t) = S(\eta(t)).$$

The use of $\epsilon$ in (1) is purely to indicate that the stochastic term $\epsilon S(X(t))$ is “small” relative to the drift, and to aid in the collection of terms of similar size. Henceforth it will be simpler to set $\epsilon = 1$ and assume that $S(X(t))$ is “small.” The initial condition for (3) is therefore $M_0 = (X_0 - \eta_0)$.

Provided that $X_0$ has either a point mass at $x_0$ or has a Gaussian distribution, the increment in (3) is a linear combination of Gaussians so $M(t)$ has a Gaussian distribution for all $t$. The mean and variance of this Gaussian can be obtained by solving a series of ODEs,

$$\frac{dm}{dt} = Fm, \quad (4)$$

$$\frac{d\Psi}{dt} = \Psi F' + F \Psi + SS', \quad (5)$$

where $m(t) := E[M(t)], \quad \Psi(t) := \text{Var} [M(t)]$ (see Web Appendix B for the derivation). Suppose $X_0 \sim N(\mu_0, \Sigma_0)$. Then for arbitrary $\eta_0$ we may set $\eta(0) = \eta_0, \quad m(0) = \mu_0^\eta - \eta_0, \quad \text{and} \quad \Psi(0) = \Sigma_0$. Integrating (2), (4), and (5) through to time $t$ provides the LNA

$$X(t) \sim N(\eta(t) + m(t), \Psi(t)). \quad (6)$$
Transition probabilities for the autoregulatory model (Example 2) given by the LNA and estimated from the SDE approximation were compared with estimates of the true probabilities for three different system sizes (see Web Appendix C for details). The results suggest that even for relatively small system sizes LNA transition probabilities are comparable with those from the SDE and can provide a reasonable approximation to the probabilities under the MJP.

4. Inference Using the LNA

We first briefly describe the inference methodology when we observe a system exactly and completely at a discrete set of times. Then we show how to perform inference when only certain components of a subset of species are observed, and these, potentially, are observed with error. Finally we compare our approach with an alternative method of using the LNA for inference introduced by Komorowski et al. (2009).

4.1. The Fully and Exactly Observed System

Consider the situation where at each of a discrete set of times, \( t_i \) (\( i = 0, \ldots, n \)), the system, \( X_{t_i} \), is observed completely and without error. Let the true transition density of the system be denoted by \( \pi(x_{t_{i+1}}|x_{t_i}, \theta) \) and the LNA of this by \( \tilde{\pi}_{\text{LNA}}(x_{t_{i+1}}|x_{t_i}, \theta) \). In practice we use the LNA obtained using \( \eta = x_{t_i}, \) \( m = 0, \) \( \Psi = 0. \) This implementation is based on an ODE solution that is piecewise continuous, with discontinuities at observation times as we restart each ODE solution at the observations. Further, as \( m(t_i) = 0, \) directly from (4) we have \( m(t) = 0 \) for all \( t > t_i. \)

For a fully observed system, the likelihood factorizes as
\[
L(\theta) = \prod_{i=1}^{n} \pi(x_{t_{i+1}}|x_{t_i}, \theta).
\]
This motivates using the approximation:
\[
\hat{L}_{\text{LNA}}(\theta) = \prod_{i=1}^{n} \tilde{\pi}_{\text{LNA}}(x_{t_{i+1}}|x_{t_i}, \theta).
\]

The accuracy of estimators obtained by maximizing \( \hat{L}_{\text{LNA}}(\theta) \) has been extensively studied in Giagos (2011), for both the Lotka–Volterra model (Example 1), and the autoregulatory model (Example 2). The method gave reliable point estimates of parameters, and reasonable estimates of uncertainty (coverage of 95% confidence intervals was generally 90% for small systems, and close to 95% for large systems).

4.2. Partially Observed Systems

Now assume that we have partial observations \( y_0, \ldots, y_s \) from times \( 0 = t_0, \ldots, t_s = T, \) where the conditional distribution for the observations given the true process is
\[
Y_i | x_t \sim N(P(\theta)x_t, V(\theta)).
\]
For the examples in Section 5 the matrix \( P(\theta) \) simply removes certain components of \( x_t \) and leaves the remaining components unchanged; an operation that requires no parameterization, but for the model analyzed in Section 6 the observations are centered on an unknown but fixed multiple of the true values (with Gaussian error). The variance of the Gaussian error, \( V, \) can be any deterministic function (of time, e.g.) parameterized by \( \theta. \) In the examples that we consider in this article \( V \) is either 0 or a fixed (unknown) diagonal matrix. We also introduce a prior, \( X_0 \sim N(\mu_0, \Sigma_0). \) To simplify notation, in the following we drop the explicit dependence of the matrices \( P(\theta) \) and \( V(\theta) \) on \( \theta. \) We will also use \( y_{0,i} := (y_0, \ldots, y_i). \)

4.2.1. Approximating the Likelihood using the LNA. Any likelihood may be decomposed as
\[
L(\theta) = \pi(y_0|\theta) \prod_{i=1}^{n} \pi(y_i|y_{0,i-1}, \theta).
\]

Firstly, \( \pi(y_0) \) can be calculated directly from our model as \( Y_0 \sim N(\mu_0, \Sigma_0). \) We then calculate approximations \( \hat{\pi}_{\text{LNA}}(y_i|y_{0,i-1}, \theta) \) to \( \pi(y_i|y_{0,i-1}, \theta) \) recursively for \( i = 1, \ldots, n. \)

Standard results give \( X_0 | y_0 \sim N(\mu_0^*, \Sigma_0^*), \) where
\[
\mu_0^* = \mu_0 + \Sigma_0 P^{*}(P \Sigma_0 P^* + V)^{-1} (y_0 - P \mu_0),
\]
\[
\Sigma_0^* = \Sigma_0 - \Sigma_0 P^* (P \Sigma_0 P^* + V)^{-1} P \Sigma_0.
\]

We then apply Kalman filter recursions and the LNA, repeating the following steps:

1. **Obtain the predictive distribution at time \( t_i. \)**
   We will have that for suitable \( \mu_{i-1}^* \) and \( \Sigma_{i-1}^* :\)
   \[
   X_{t_{i-1}} | y_{0,i-1} \sim N(\mu_{i-1}^*, \Sigma_{i-1}^*).
   \]
   We then initiate the LNA with \( \eta_{t_{i-1}} = \mu_{i-1}^*, \) so that \( m(t_{i-1}) = 0, \) \( \Psi(t_{i-1}) = \Sigma_{i-1}^*. \) From (4), \( m(t) = 0 \Rightarrow m(t) = 0 \) for all \( t > t_i. \) Further, integrating the ODEs (2) and (5) forward for time \( t_i - t_{i-1} \) provides \( \eta(t_i) \) and \( \Psi(t_i), \) so that our approximation to the density at \( t_i \) is \( X_{t_i} | y_{0,i-1} \sim N(\mu_i, \Sigma_i), \) where \( \mu_i = \eta(t_i) \) and \( \Sigma_i = \Psi(t_i). \)

2. **Calculate \( \hat{\pi}_{\text{LNA}}(y_i|y_{0,i-1}, \theta). \)**
   Using \( Y_i = PX_i + \epsilon_i, \) where \( \epsilon_i \sim N(0, V) \) directly gives
   \[
   Y_i | y_{0,i-1} \sim N(P \mu_i, P \Sigma_i P^* + V).
   \]

3. **Calculate \( \hat{\pi}_{\text{LNA}}(x_i|y_{0,i}, \theta). \)**
   Since
   \[
   \begin{bmatrix}
   X_i \\
   Y_i
   \end{bmatrix} | y_{1:(i-1)} \sim N \left( \begin{bmatrix} \mu_i \\ \Sigma_i \end{bmatrix}, \begin{bmatrix} \Sigma_i P^* & P \Sigma_i P^* + V \\ P \Sigma_i P^* & P \Sigma_i P^* + V \end{bmatrix} \right)
   \]
   we have directly that \( X_i | y_{1:i} \sim N(\mu_i^*, \Sigma_i^*), \) where
   \[
   \mu_i^* = \mu_i + \Sigma_i P^* (P \Sigma_i P^* + V)^{-1} (y_i - P \mu_i),
   \]
   \[
   \Sigma_i^* = \Sigma_i - \Sigma_i P^* (P \Sigma_i P^* + V)^{-1} P \Sigma_i.
   \]
   Our approximation for the likelihood of the data is then
   \[
   \hat{L}_{\text{LNA}}(\theta) = \pi(y_0|\theta) \prod_{i=1}^{n} \hat{\pi}_{\text{LNA}}(y_i|y_{0,i-1}, \theta).
   \]
The only approximation in \( \hat{L}_{\text{LNA}}(\theta) \) is due to using the LNA for the transition density of the system, which gives us the Normal approximation to \( X_i | y_{1:i-1} \) from the Normal approximation to \( X_{i-1} | y_{1:i-1} \).

We emphasize that in Step (1), once the observation \( y_{t-1} \) is available then for the period of integration from \( t_{t-1} \) to \( t \) we re-initialize the ODE (2) to the posterior mean at \( t_{t-1} \) by setting \( \eta(t_{t-1}) = \mu_{\eta_{t-1}} \), leading to a piecewise-continuous solution for \( \eta \). The LNA relies on a first-order Taylor expansion about \( \eta \), and by continually realigning the point about which the expansion is performed to the current best estimate of the center of the distribution we aim to minimize the impact of the higher-order terms that have been neglected.

### 4.3. MCMC Scheme

It is possible to estimate the parameters by numerically maximizing the LNA to the likelihood (10). However we consider a Bayesian analysis. We introduce priors for the parameters, \( \pi(\theta) \), and use MCMC to generate samples from the resulting approximation to the posterior \( \pi(\theta) / \hat{L}_{\text{LNA}}(\theta) \).

We implemented a random-walk Metropolis algorithm (RWM). Each iteration of the algorithm involved a single block update of all the \( \log \)-parameters. Using the \( \log \)-scale is natural as all parameters are positive. For the simulation study in Section 5 we used pilot runs to tune our algorithms (and algorithms against which we compare): the covariance of the posterior from the pilot run, with the random-walk proposal was proportional to the estimate (and algorithms against which we compare): the covariance of the random-walk proposal was proportional to the estimate of the covariance of the posterior from the pilot run, with the scale tuned to produce an acceptance rate in the range 0.25–0.30 (Roberts and Rosenthal, 2001). For the analysis of the Google Flu Trends Data in Section 6 we used an adaptive RWM algorithm similar to that in Sherlock, Fearnhead, and Roberts (2010).

### 4.4. Implementation

The ODEs required for calculating \( \pi(y_i | y_{0:i-1}) \) can be solved numerically. Care is needed as in many applications the ODEs are stiff (Hairer and Wanner, 1991). There are standard numerical routines for solving stiff ODEs, and we used the \texttt{soda} package (Petzold, 1983).

### 4.5. Alternative Use of the LNA

Previously, use of the LNA for Bayesian inference on stochastic kinetic networks has been suggested by Komorowski et al. (2009), but their implementation has important differences from ours. The approach of Komorowski et al. (2009) involves using the LNA to obtain an approximation for the joint distribution of \( X_{1:n} = (X_1, \ldots, X_n) \) conditional on a value for \( x_0 \). This can be combined with the linear-Gaussian relationship between each observation \( Y_i \) and state-value \( X_i \), to give an approximation to the likelihood for data \( y_{1:n} \) in terms of the parameters, \( \theta \) and the initial value, \( x_0 \). They introduce priors for the \( \theta \) and \( x_0 \), and sample from the (approximate) posterior for these using MCMC.

In practice the most important difference between this approach and ours, is that Komorowski et al. (2009) use the LNA over a time period \( [0, t_s] \) obtained from solving the ODE approximation to the model over this period for a given initial condition. By comparison we use a different LNA for each time-interval \( [t_{i-1}, t_i] \), essentially restarting the LNA using the posterior mean of \( x_{i-1} \) given \( y_{0:i-1} \) as the initial condition for the ODE (2). This difference can be important for some models, as the ODE solution can become poor over long time-periods. Thus the approach of Komorowski et al. (2009) can give a poor approximation to the distribution of \( X_i \) for larger values of \( t \). By re-starting the LNA method over each time-interval we help avoid the problems of the approximation getting worse for larger \( t \).

The difference in accuracy of the two approaches for using the LNA is investigated thoroughly for systems which are fully observed at discrete time-points in Giagos (2011), where the method of Komorowski et al. (2009) was found to be less accurate, for both point and interval estimation, than the method we introduce above. We further demonstrate the increased accuracy of our approach for partially observed systems in Section 5.

### 5. Simulation Study

We now empirically evaluate the performance of the LNA for inference on parameters in both the Lotka–Volterra model (Example 1) and the auto-regulatory model (Example 2). Our aim is to both compare our approach with inference based on ODE approximations, the LNA approach of Komorowski et al. (2009), and the SDE-based approach of Golightly and Wilkinson (2005); and to evaluate the accuracy obtained by using the LNA for both point and interval estimation.

The code of Golightly and Wilkinson (2005) was adapted to use the same half-Cauchy prior for the parameters as us and to employ an optimally tuned single-block RWM Gaussian proposal so that this aspect would be exactly comparable with our RWM scheme. All MCMC algorithms except for that of Golightly and Wilkinson (2005) were run for 110,000 iterations, from which the first 5,000 iterations were discarded as burn-in. Output from simulations for the autoregulatory system was thinned by a factor of 10 for storage. Since it mixed more slowly the algorithm of Golightly and Wilkinson (2005) was run for 260,000 iterations with a burn-in of 10,000 iterations; as with the LNA-based analysis of the autoregulatory system, output was thinned by a factor of 10.

In assessing accuracy of a method on a given model, throughout we present results based on analyzing 100 different data sets simulated using the true jump process for a given model and set of parameter values. We present results in terms of estimating the log of the parameter since interest is primarily in the order of magnitude of the reaction rates. The posterior median is used as the point estimate for each parameter as this is both invariant to monotonic transformations and is robust to heavy-tailed posterior distributions.

#### 5.1. Comparison with ODE Method and Komorowski et al.

Our first comparison is with approximating the evolution as a deterministic ODE, and with the LNA method of Komorowski et al. and is based on the Lotka–Volterra model. We simulated data with \( X_0 = (40, 140) \) and with \( \theta_1 = 0.01 \), \( \theta_2 = 0.6 \) and \( \theta_3 = 0.3 \). Observations were made every second for 30 seconds, or until one of the species became extinct. We used Gamma(2,10) priors for all rate parameters, which gave a reasonable prior mass across the range of values the rates take. We assumed that predators are observed exactly for the
LNA method and for that of Komorowski et al., and the prey are unobserved.

The ODE method uses a log-likelihood (and hence log-posterior) which depends on the sum of squared errors between the solution of the ODE and the observations, and is equivalent to modeling the observations as having additive Gaussian error; we place a half-Cauchy prior on the error variance, $\pi(\sigma^2) \propto 1/(1 + \sigma^4)$. Both LNA methods have a similar computational cost, while the ODE approach has a smaller computational cost as the differential equations for the variables, that approach gives a poor approximation to the data, and as the LNA of Komorowski is based on such approximations, it is useful to see results from a single dataset.

Results are shown in Table 1. The new LNA approach is uniformly better at estimating each parameter in terms of both accuracy of point estimates and coverage of credible intervals. The approach of Komorowski et al. is in turn more accurate than using an ODE approximation. These results are for a small system, and the difference between the different approaches will be less if a larger system size is studied.

To see why the new LNA approach is more accurate than either that of Komorowski et al. or that of using an ODE approximation, it is useful to see results from a single dataset. In Figure 1 we show a simulated data set, together with the best fitting ODE solution. This solution gives a poor fit to the data, and as the LNA of Komorowski is based on such ODE solutions, that approach gives a poor approximation to the likelihood. By comparison, as our approach re-starts the LNA at each observation, we get a good approximation to the likelihood terms across the whole time-period of the data.

Table 1

Comparison of our approach (LNA), the approach of Komorowski et al. (KOM) and the ODE approach (ODE). We present results in terms of the mean performance across 100 data sets.

| Parameter      | LNA | KOM | ODE |
|----------------|-----|-----|-----|
| Mean of $\log_{10} \theta$ | -2.001 | -2.000 | -2.007 |
| posterior     | -0.248 | -0.263 | -0.308 |
| medians       | -0.541 | -0.540 | -0.540 |
| Mean abs. error of $\theta$ | 0.043 | 0.074 | 0.080 |
| median $\theta$ | 0.056 | 0.086 | 0.126 |
| width of $\theta$ | 0.050 | 0.070 | 0.098 |
| 95% CI of $\theta$ | 0.193 | 0.138 | 0.141 |
| Coverage      | 0.188 | 0.153 | 0.158 |
| 95% CI        | 0.198 | 0.156 | 0.194 |

Results are given in Table 2. The comparative performance of the two methods varies with parameter. For the parameters $\theta_1$, $\theta_2$, $\theta_3$, and $\theta_4$, both methods are similar in terms of accuracy and of size and coverage of credible intervals. Coverage of credible intervals are consistent with their nominal size. Computational efficiency of the two methods is also very similar, with a mean ESS per second of 0.84 and 0.80 for the LNA and SDE methods, respectively.

However inference on the remaining parameters differs considerably. The SDE method is uniformly more accurate, and has substantially smaller credible intervals in all cases. For both methods, coverage of credible intervals is close to 90% for these parameters. The computational efficiency of the LNA is substantially higher for these parameters with the ESS per second an order of magnitude greater.

The parameters for which the inferences of the two methods differ consist of two pairs of reaction rates: $\theta_1$ and $\theta_2$ are the rates of reversible reactions linked to the product DNA – $P_2$, whereas $\theta_3$ and $\theta_4$ are the rates of irreversible reactions linked to the dimerization of the protein. As such we would expect dif-
We observe poor inference for true parameter values used for the first simulation study. parameter values, which happened to be consistent with the parameters, or whether the method is biased towards smaller with all rate parameters increased by a factor of 4. This gives not estimate the posterior well for large rates. estimated likelihood. In general therefore the Euler method will give more accurate inferences for the remaining parameters. However, the method of Golightly and Wilkinson (2005) does leading to coverage probabilities less than 0.5 in all cases. Method of Golightly and Wilkinson (2005): accuracy is lower from the ODE solution are no longer small. The reactions in our model are:

\[ R_1: S_1 + I_1 \rightarrow E_1 + I_1 \]
\[ R_2: E_2 \rightarrow I_2 \]
\[ R_3: S_2 + I_2 \rightarrow E_2 + I_2 \]
\[ R_4: E_1 \rightarrow I_1 \]
\[ R_5: \theta_1 \rightarrow 0 \]
\[ R_6: \theta_2 \rightarrow 0 \]
\[ R_7: S_1 + I_2 \rightarrow E_1 + I_2 \]
\[ R_8: S_2 + I_1 \rightarrow E_2 + I_1. \]

Detailed results are given in Table 1 of Web Appendix D. We observe poor inference for \( \theta_1, \theta_2, \theta_5, \) and \( \theta_6 \) using the method of Golightly and Wilkinson (2005): accuracy is lower than using the LNA, and the credible intervals are too small, leading to coverage probabilities less than 0.5 in all cases. However, the method of Golightly and Wilkinson (2005) does give more accurate inferences for the remaining parameters. This is likely to be because the extra variability arising from the larger rates means that the perturbations of the system from the ODE solution are no longer small.

### 5.3. Accuracy of the LNA Method

We further investigate the accuracy of the LNA, by repeating the analysis of the auto-regulatory example, but considering different observation models. We considered all components observed with error, and only three components observed either exactly or with Gaussian error; errors for each species were independent with mean zero and variance 1. We use the same priors as above.

Results are given in Web Appendix D and are comparable with those in Table 2. The LNA appears to provide good estimates of the parameters. As we would expect, as we observed fewer species, or observe with error, the uncertainty in our estimates increases. Perhaps most importantly, the coverage rates we obtain are close to 95% in all cases, suggesting the method is giving a good estimate of uncertainty. We obtain higher coverage rates with less informative data, possibly because any bias in the LNA has less effect when we have higher posterior variance.

### 6. Prediction of Flu Epidemics Using Google Flu Trends Data

We now apply our method to predict flu case numbers based on data from Google Flu Trends (GFT: http://www.google.org/flutrends). GFT data are estimates of the number of new cases of flu each week (per 100,000 people) based on the popularity of terms associated with flu in web searches. Ginsberg et al. (2009) showed that actual numbers of flu cases can be accurately predicted using such data, with the advantage of being able to obtain estimates of the current number of cases, as opposed to health-service data which are typically published with a delay of approximately one week and are often incomplete.

Our analysis is motivated by Dukic et al. (2012), who use a one-compartment SEIR model (see Web Appendix A for details) to show that accurate predictions of flu cases can be obtained from GFT data using a state-space SEIR model, and in particular that such models are substantially more accurate than simple AR models. For our analysis we consider cases in the North and South Islands of New Zealand. GFT data were obtained for each island, for January 2008 to January 2012 inclusive, and converted from proportions into counts.

Each year there is a flu epidemic, often with different flu strains. The number of flu cases in New Zealand is typically at its yearly minimum around the start of February, and so we split our data into four separate “years” from February yr to January of yr + 1 for yr ∈ {2008, 2009, 2010, 2011}.

We model the data using a two-compartment SEIR model. Our state consists of the number of susceptibles, exposed, infected and recovered in each of the north and south islands. We assume a fixed population size for both islands, which results in a six-dimensional state: \([S_1, E_1, I_1, S_2, E_2, I_2]\), where a subscript of 1 denotes North Island and a subscript of 2 denotes South Island.

The reactions in our model are:

\[ R_1: \theta_1 \rightarrow 0 \]
\[ R_2: \theta_2 \rightarrow 0 \]
\[ R_3: S_1 + I_2 \rightarrow E_1 + I_2 \]
\[ R_4: S_2 + I_1 \rightarrow E_2 + I_1. \]

### Table 2

| Parameter | \( \log_{10} \theta \) | \( \theta_1 \) | \( \theta_2 \) | \( \theta_3 \) | \( \theta_4 \) | \( \theta_5 \) | \( \theta_6 \) | \( \theta_7 \) | \( \theta_8 \) |
|-----------|----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Mean       | LNA            | –0.921       | –0.092       | –0.466       | –0.687       | –0.856       | –0.076       | –0.523       | –0.974       |
| posterior medians | GW05           | –0.975       | –0.135       | –0.445       | –0.643       | –0.972       | 0.036        | –0.506       | –0.941       |
| Mean abs. error of median | LNA            | 0.147        | 0.139        | 0.093        | 0.090        | 0.197        | 0.183        | 0.082        | 0.107        |
| Mean width of 95% CI | LNA            | 0.730        | 0.721        | 0.408        | 0.410        | 1.065        | 1.047        | 0.410        | 0.429        |
| Coverage of 95% CI | LNA            | 94           | 94           | 89           | 92           | 94           | 95           | 92           | 88           |
| Mean ESS/sec | LNA            | 2.28         | 2.67         | 0.79         | 1.04         | 1.24         | 1.24         | 0.97         | 0.56         |
| ESS/sec       | GW05           | 0.18         | 0.18         | 0.60         | 1.10         | 0.09         | 0.08         | 0.68         | 0.82         |
Further details of the model are provided in Web Appendix A.

Observations are $y_t = (y_{1t}^{(1)}, y_{2t}^{(2)})$, the number of flu cases, from the GFT data, in week $t$ in the North and South Island, respectively. We model that these are realizations of random variables $Y_{1t}^{(j)}$, for $j = 1, 2$, where $Y_{1t}^{(j)} \sim N(CI_j(t), \sigma_t^2)$. A heuristic interpretation is that our SEIR models apply to “communities” that are infected, and assume an equal rate of contact between each pair of communities. If a community is infected, then $C$ is the average number of flu cases that will result. Priors follow a similar form to those in Dukic et al. (2012); full details are given in Web Appendix E.

For each year’s data, and for each $t = 2, \ldots, 51$ we use Steps (1)-(3) in Section 4.2.1 within an adaptive RWM algorithm similar to that in Sherlock et al. (2010) (see Web Appendix F) to estimate both the parameters and the current state of the model given observations $y_{1t}$. For each sample from the RWM we then apply Step (1) again to predict the number of cases in week $t + 1$.

As a benchmark to compare with, we also analyzed the data using the method of Dukic et al. (2012). This approach uses sequential Monte Carlo (SMC), and we shall refer to it as the SMC approach. It is based on fitting a single compartment SEIR model (henceforth 1CM) to data on the total number of flu cases across both islands. To deal with the intractability of the Markov jump model, the ODE approximation (2) is used and is itself approximated using an Euler scheme with a discretization of the inter-observation interval (here 1 week). Gaussian noise is then added to the relative change in the number of infectives between observation times, leading to a very different Gaussian transition model to (6). The observation model is also based on the relative change in the number of infectives, with additive errors assumed to be Gaussian. The SMC scheme approximates the joint posterior for the parameters and the state at time $t$, given the data up to time $t$, by a set of weighted particles. Due to the choice of approximations of both the state dynamics and the observation, efficient methods (Pitt and Shephard, 1999; Carvalho et al., 2010) for implementing the SMC algorithm can be used. See Dukic et al. (2012) for more details.

One advantage of this approach is computational, as, unlike with MCMC, the algorithm does not need to be re-run from scratch each time a new observation is received. The potential disadvantage of the method is that it uses a cruder approximation to the underlying jump-Markov model.

We attempted to implement an equivalent SMC approach to fit a two component SEIR model (henceforth 2CM) to the Google flu-trends data. However results from the SMC analysis of this model, using $10^7$ particles, were substantially worse than for the 1CM. SMC methods are known to often perform poorly for models with unknown parameters. The poor results for the 2CM are thus likely to be due to poor Monte Carlo performance for a model with 10 unknown parameters. For further comparability with the method of Dukic et al. (2012) we therefore also analyzed data for the whole of New Zealand using the LNA within a 1CM.

For each LNA analysis we ran an MCMC for 100,000 iterations, using a burn-in of 20,000 for the 2CM and of 10,000 for the 1CM (then thinning both by a factor of 10). For a week-ahead prediction at the height of the flu season (after 30 weeks of data) runs for the 1CM model took between 150 and 156 seconds on a single Intel Core i7 3770 CPU@3.40GHz, while runs for the 2CM took between 1054 and 1135 seconds. Repeated runs of the MCMC produced the same estimates of accuracy to at least two significant figures. The SMC analysis achieved a similar precision to the LNA when $10^6$ particles were used. Week-ahead predictions from 30 weeks of data took between 67.7 and 68.1 seconds.

| Year | Method | Bias | MAD | MWCI | Cov. |
|------|--------|------|-----|------|------|
| 2008 | LNA2CM | -2.01| 6.03| 21.8 | 84   |
|      | LNA1CM | -1.11| 5.96| 29.1 | 84   |
|      | SMC1CM | -3.07| 6.95| 185.7| 100  |
| 2009 | LNA2CM | 0.28 | 12.90| 36.6 | 84   |
|      | LNA1CM | -0.27| 14.72| 40.51| 86   |
|      | SMC1CM | -13.89| 21.47| 211.3| 100  |
| 2010 | LNA2  | 0.08 | 6.42 | 19.4 | 82   |
|      | LNA1CM| -0.37| 6.29 | 21.6 | 84   |
|      | SMC1CM| -4.24| 8.38 | 113.8| 100  |
| 2011 | LNA2CM| -0.83| 5.09 | 18.3 | 84   |
|      | LNA1CM| -1.02| 4.95 | 18.1 | 82   |
|      | SMC1CM| -1.50| 5.82 | 92.2 | 100  |

As a benchmark to compare with, we also analyzed the data using the method of Dukic et al. (2012). This approach uses sequential Monte Carlo (SMC), and we shall refer to it as the SMC approach. It is based on fitting a single compartment SEIR model (henceforth 1CM) to data on the total number of flu cases across both islands. To deal with the intractability of the Markov jump model, the ODE approximation (2) is used and is itself approximated using an Euler scheme with a discretization of the inter-observation interval (here 1 week). Gaussian noise is then added to the relative change in the number of infectives between observation times, leading to a very different Gaussian transition model to (6). The observation model is also based on the relative change in the number of infectives, with additive errors assumed to be Gaussian. The SMC scheme approximates the joint posterior for the parameters and the state at time $t$, given the data up to time $t$, by a set of weighted particles. Due to the choice of approximations of both the state dynamics and the observation, efficient methods (Pitt and Shephard, 1999; Carvalho et al., 2010) for implementing the SMC algorithm can be used. See Dukic et al. (2012) for more details.

One advantage of this approach is computational, as, unlike with MCMC, the algorithm does not need to be re-run from scratch each time a new observation is received. The potential disadvantage of the method is that it uses a cruder approximation to the underlying jump-Markov model.

We attempted to implement an equivalent SMC approach to fit a two component SEIR model (henceforth 2CM) to the Google flu-trends data. However results from the SMC analysis of this model, using $10^7$ particles, were substantially worse than for the 1CM. SMC methods are known to often perform poorly for models with unknown parameters. The poor results for the 2CM are thus likely to be due to poor Monte Carlo performance for a model with 10 unknown parameters. For further comparability with the method of Dukic et al. (2012) we therefore also analyzed data for the whole of New Zealand using the LNA within a 1CM.

For each LNA analysis we ran an MCMC for 100,000 iterations, using a burn-in of 20,000 for the 2CM and of 10,000 for the 1CM (then thinning both by a factor of 10). For a week-ahead prediction at the height of the flu season (after 30 weeks of data) runs for the 1CM model took between 150 and 156 seconds on a single Intel Core i7 3770 CPU@3.40GHz, while runs for the 2CM took between 1054 and 1135 seconds. Repeated runs of the MCMC produced the same estimates of accuracy to at least two significant figures. The SMC analysis achieved a similar precision to the LNA when $10^6$ particles were used. Week-ahead predictions from 30 weeks of data took between 67.7 and 68.1 seconds.

Summaries of the accuracy of both models using the LNA and of the 1CM using SMC in predicting the total number of cases across both islands, and the coverage of 95% credible intervals for the predictions are given in Table 3. Compared to the SMC method we see that, for all four years, the LNA (using either model) gives less biased estimates and smaller forecast error, as measured by the mean absolute error in predictions. The credible intervals produced by the LNA methods are at least a factor of 5 smaller than for the approach of Dukic et al. (2012). However the coverage of the LNA’s credible intervals is between 80% and 86%. We believe the reason for this is most likely due to our model assuming a constant variance for the observation error, whereas the variance of this error appears to increase with the current size of the epidemic. We thus under-estimate the uncertainty at the peak of the flu epidemic. From the table, predictions for New Zealand as a whole from the 2CM are no better than those from the 1CM, however the 2CM also provides individual predictions for North and South Island. These predictions, together with the true number of cases (as given by the Google Flu Trends data), are shown in Figure 2 in Web Appendix D; summaries of the accuracy are provided in Table 3 of the same appendix.

### 7. Discussion

We have demonstrated how the LNA can be used to perform inference for reaction networks where all, or a subset, of components are observed. Observations can either be exact, or with additive Gaussian error. Results suggest that using
the LNA is more accurate than approximating the underlying model using an ODE.

The LNA is based upon first obtaining a deterministic approximation to the path of the state vector over time; and then modeling the error about this deterministic approximation. Key to the error in the LNA being small is that the deterministic approximation is accurate. We have shown that recalculating the deterministic solution between each pair of observations is more accurate than calculating a single deterministic solution as suggested by Komorowski et al. (2009). Furthermore, across the examples we looked at the LNA gives reliable inferences in almost all cases. The one exception (for a subset of parameters in the results in Table 1 of Web Appendix C) corresponds to cases where the noise in the model was high, leading to the perturbations of the system about the deterministic solution not being small.

We have also compared with a method based on approximating the underlying model using an SDE. The accuracy of the LNA and SDE-based approaches are similar, with the relative performance of the two approaches varying depending on which reaction rates are being estimated. The advantage of the LNA is one of simplicity—as the LNA gives an analytic form for the approximation to the transition density of the model. In particular there is no need to choose a level of time-discretization. Calculating the LNA involves solving a set of ODEs, but standard routines exist for appropriately choosing and adapting the step-size using in numerically solving the ODEs. By comparison SDE methods currently involve the user pre-specifying the level of time-discretization. Choosing an appropriate level is difficult, partly because the required level needed to get an accurate approximation can depend on the parameter values, and these will change at each MCMC iteration.

We have demonstrated the usefulness of the LNA for inference by making predictions for flu cases in New Zealand using Google Flu Trends data. While our prediction accuracy was higher than that of Dukic et al. (2012), our assumption of a constant observation variance leads us to under-estimating uncertainty in future observations at the peak of the epidemic. This assumption is currently needed for the tractability of our algorithm, but it should be possible to relax this assumption, for example using ideas from Rue, Martino, and Chopin (2009) to allow for efficient inference under a range of observation models.

We considered fitting a two-compartment SEIR model to the New Zealand data, but the scalability of the LNA should mean it is possible to analyze SEIR models with even more compartments—for example, to jointly analyze data from multiple cities in the US. For a reaction network with \( n \) reactions and a state-space of size \( n_s \), the LNA requires the numerical integration of \( n_s^2 \) ODEs, with \( O(n) \) rate-related calculations at each time-point. For the two-compartment model, the state-space was twice the size of that of the one-compartment model, and the number of reactions more than doubled. Given the other computational overheads of the algorithm, this is consistent with the observed increase in CPU time by a factor of approximately 7, and, given the short running time for the two-compartment model (less than 20 minutes), suggests that on-line week-ahead predictions should be feasible for models with three or four compartments.

The approach of Dukic et al. (2012) uses a sequential Monte Carlo algorithm, which is computationally more convenient than MCMC. SMC inference for the single-compartment model of Dukic et al. (2012) was more than twice as fast as the single compartment LNA, but we could not implement an accurate SMC method for the two-component model. This seems to be due to problems with using SMC to analyze models with moderately large numbers of parameters. Alternative sequential Monte Carlo approaches, such as those based on mode tracking (e.g., Vaswani, 2008), have recently been shown to be accurate for high-dimensional state processes, and may offer a competitive alternative for analyzing data such as that from Google Flu Trends.

8. Supplementary Materials
Web Appendices A, B, C, D, E, and F, referenced respectively, in Sections 2, 3.2, 3.2, 5.2, 6, and 6, are available with this paper at the Biometrics website on Wiley Online Library. The supplementary material also contains C code implementing the LNA for all of the models considered in this article.

Acknowledgements
The authors are grateful to Dr A. Golightly for supplying C code for inference on the completely and exactly observed autoregulatory example using the methodology in Golightly and Wilkinson (2005), and to Dr V. Dukic for supplying the R code for prediction from the Google Flu Trends data using the methodology in Dukic et al. (2012). Paul Fearnhead was supported by EPSRC grant EP/G028745.

References
Ait-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics* 36, 906–937.
Amrein, M. and Künsch, H. (2012). Rate estimation in partially observed markov jump processes with measurement errors. *Statistics and Computing* 22, 513–526. 10.1007/s11222-011-9244-1.
Andersson, H. and Britton, T. (2000). *Stochastic Epidemic Models and Their Statistical Analysis*, volume 151 of Lecture Notes in Statistics. New York: Springer-Verlag.
Ball, F. and Neal, P. (2008). Network epidemic models with two levels of mixing. *Mathematical Biosciences* 212, 69–87.
Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society Series B* 68, 333–382.
Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing* 18, 125–135.
Carvalho, C. M., Johannes, M. S., Lopes, H. F., and Polson, N. G. (2010). Particle learning and smoothing. *Statistical Science*, 25(1) 88–106.
Dukic, V., Lopes, H. F., and Polson, N. G. (2012). Tracking epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Association* 107, 1410–1426.
Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion
processes. *Journal of Business and Economic Statistics* **20**, 297–338.

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

Elf, J. and Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Research* **12**, 2475–2484.

Ferm, L., Lötstedt, P., and Hellander, A. (2008). A hierarchy of approximations of the master equation scaled by a size parameter. *Journal of Scientific Computing* **34**, 127–151.

Giagos, V. (2011). Inference for autoregulatory genetic networks using diffusion process approximations. PhD thesis, Lancaster University.

Gillespie, D. T. (2005). A rigorous derivation of the chemical master equation. *Physica A* **188**, 404–425.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* **457**, 0028–0836.

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

Golightly, A. and Wilkinson, D. J. (2006). Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing* **16**, 323–338.

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

Hairer, E. and Wanner, G. (1991). *Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems*. Berlin and Heidelberg: Springer-Verlag.

Hayot, F. and Jayaprakash, C. (2004). The linear noise approximation for molecular fluctuations within cells. *Physical Biology* **1**, 205–210.

Jewell, C. P., Keeling, M. J., and Roberts, G. O. (2009). Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. *Journal of the Royal Society Interface* **6**, 1145–1151 (electronic).

Jones, D. S., Plank, M. J., and Sleeman, B. D. (2010). *Differential Equations and Mathematical Biology*. Chapman & Hall/CRC Mathematical and Computational Biology Series. Boca Raton, FL: Chapman & Hall/CRC.

Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. New York: Springer.

Komorowski, M., Finkenstadt, B., Harper, C. V., and Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10**, 343.

Kurtz, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability* **7**, 49–58.

Kurtz, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability* **8**, 344–356.

Petzold, L. (1983). Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM Journal on Scientific and Statistical Computing* **4**, 136–148.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association* **94**, 590–599.

Proctor, C. J., Soti, C., Boys, R. J., Gillespie, C. S., Shanley, D. P., Wilkinson, D. J., and Kirkwood, T. B. L. (2005). Modelling the actions of chaperones and their role in ageing. *Mechanisms of Ageing and Development* **126**, 119–131.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 741–796.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science* **16**, 351–367.

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

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* **71**, 319–392.

Sherlock, C., Fearnhead, P., and Roberts, G. O. (2010). The random walk Metropolis: Linking theory and practice through a case study. *Statistical Science* **25**, 172–190.

Srensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: A survey. *International Statistical Review* **72**, 337–354.

van Kampen, N. G. (1997). *Stochastic Processes in Physics and Chemistry*. Amsterdam: Elsevier.

Vaswani, N. (2008). Particle filtering for large-dimensional state spaces with multimodal observation likelihoods. *IEEE Transactions on Signal Processing* **56**, 4583–4597.

Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC Mathematical and Computational Biology Series. Boca Raton, FL: Chapman & Hall/CRC.

Received January 2013. Revised October 2013.

Accepted December 2013.