Infectious Disease Modelling 5 (2020) 221–232

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Bayesian inference for dynamical systems

Weston C. Roda

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada

ARTICLE INFO

Article history:
Received 17 December 2019
Accepted 27 December 2019
Available online 10 January 2020
Handling Editor: Dr. J Wu

Keywords:
Bayesian
Inference
Model fitting
Data
Dynamical system
Mathematical model

ABSTRACT

Bayesian inference is a common method for conducting parameter estimation for dynamical systems. Despite the prevalent use of Bayesian inference for performing parameter estimation for dynamical systems, there is a need for a formalized and detailed methodology. This paper presents a comprehensive methodology for dynamical system parameter estimation using Bayesian inference and it covers utilizing different distributions, Markov Chain Monte Carlo (MCMC) sampling, obtaining credible intervals for parameters, and prediction intervals for solutions. A logistic growth example is given to illustrate the methodology.

1. Introduction

A common method for performing parameter estimation for dynamical systems is to use Bayesian inference (Ghasemi et al., 2011; Higham & Husmeier, 2013; Ma & Berndsen, 2014; Periwal et al., 2008; Vanlier, Tiemann, Hilbers, & van Riel, 2012). Despite the popularity of using Bayesian inference for performing parameter estimation for dynamical systems and useful computational manuals, there is a need for a formalized and comprehensive methodology.

The methods described in this paper assume that the behaviors of the dynamical system of interest have been mathematically analyzed and that the solutions of the dynamical system are well-behaved. Additionally, it is assumed that if a numerical scheme is being used to solve the dynamical system that the numerical scheme is stable. The methodology is presented from a mathematical biology perspective and it will focus on systems of ordinary differential equations (ODEs); however, the Bayesian inference methodology presented can be applied to other areas of applied mathematics and other differential equations systems such as partial differential equations (PDEs). This paper will provide a formalized methodology for dynamical system parameter estimation using Bayesian inference and it will cover utilizing different distributions, Markov Chain Monte Carlo (MCMC) sampling, obtaining credible intervals for parameters, and prediction intervals for solutions. The methodology is illustrated by using a logistic growth example.

2. Dynamical system

Assume that the dynamical system of interest can be described by the following autonomous ODE system (1) written as a vector differential equation:

E-mail address: wroda@ualberta.ca.

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2019.12.007

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}), \tag{1}$$

where $\mathbf{x} = \langle x_1, \ldots, x_k \rangle$ and $\mathbf{f} = \langle f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}) \rangle$, with the vector of initial conditions $\mathbf{x}_0 = \langle x_1^0, \ldots, x_k^0 \rangle$.

It is assumed that the that the unique solution vector, $\mathbf{x}(t)$, of system (1) exists and can be obtained either explicitly or using numerical approximation. If a numerical approximation method is used, it is assumed that the numerical approximation scheme is stable.

All the parameters in system (1) will be denoted by the vector $\boldsymbol{\beta}$. If the initial conditions $x_1^0, \ldots, x_k^0$ will also be estimated, then let the initial conditions $x_1^0, \ldots, x_k^0$ be contained in vector $\boldsymbol{\beta}$ as well.

The dependence of the unique solution vector $\mathbf{x}$ on both time, t, and the vector of parameters, $\boldsymbol{\beta}$, will be emphasized and the unique solution vector will be denoted as $\mathbf{x}(\boldsymbol{\beta}, t)$.

3. Data

Suppose there are m time series data sets. It is important to ensure that the correct ODE model solution or combination of ODE model solutions is fit to the j^{th} time series data set ($j = 1, \ldots, m$).

Sometimes a data set is scaled differently than the model solutions or the data set can be described by a summation of the ODE model solutions. In order to include these situations, we can use a linear combination of the ODE model solutions, $a_1^j x_1(\boldsymbol{\beta}, t) + \ldots + a_k^j x_k(\boldsymbol{\beta}, t)$, to fit to the j^{th} time series data set. (The simpler case where only the i^{th} specific ODE model solution $x_i(\boldsymbol{\beta}, t)$ is to be fit to the j^{th} time series data set, is included in the linear combination where $a_i^j = 1$ and the other constants are zero.) If the nonzero vector of constants, $\mathbf{a}^j$, will be estimated, then let the nonzero vector of constants, $\mathbf{a}^j$, for $j = 1, \ldots m$, be contained in vector

$$\boldsymbol{\nu} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{a}^1 \\ \vdots \\ \mathbf{a}^m \end{bmatrix}.$$

Also, if the j^{th} data set can be described by a nonlinear combination of the ODE model solutions, then, similarly, let any estimated nonzero vector of constants, $\mathbf{a}^j$, be contained in vector

$$\boldsymbol{\nu} = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{a}^1 \\ \vdots \\ \mathbf{a}^m \end{bmatrix}.$$

So, in general, we fit the function, $F(x_1(\boldsymbol{\beta}, t_i^j), \ldots, x_k(\boldsymbol{\beta}, t_i^j), \mathbf{a}^1, \ldots, \mathbf{a}^m)$, to the j^{th} data set.

4. Distribution of data over time

The distribution of the observations over time for each j^{th} data set must be chosen before fitting system (1) to the data. The following sections will describe the Gaussian, Poisson, Negative Binomial, and other distribution options.

4.1. Gaussian distribution

Let Y be a random variable from the Gaussian distribution with parameters μ and $\sigma^2 = \frac{1}{\tau} > 0$, $Y\tilde{}\mathrm{N}(\mu, \theta^2)$. The formulation of the Gaussian distribution is given by the following continuous probability density function (pdf), $f(y)$ (Bain & Engelhardt, 1987):

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(y-\mu)^2\right). \tag{2}$$

The mean, E[Y], of the Gaussian distribution is given by μ and the variance, Var[Y], of this distribution is given by $\sigma^2 = \frac{1}{\tau}$.

Assume that the j^{th} time series data set is given by observations $D_j = \{d_1^j, \ldots, d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j, \ldots, t_{n_j}^j\}$. and that the probability of observing d_i^j is given by the Gaussian distribution:

$$f\left(d_i^j\right) = \sqrt{\frac{\tau^j}{2\pi}} \exp\left(-\frac{1}{2}\tau^j \left(d_i^j - \mu_i^j\right)^2\right), \tag{3}$$

where the mean μ_i^j changes depending on the time, t_i^j and the variance $\frac{1}{\tau^j}$ is specific to the j^{th} data set.

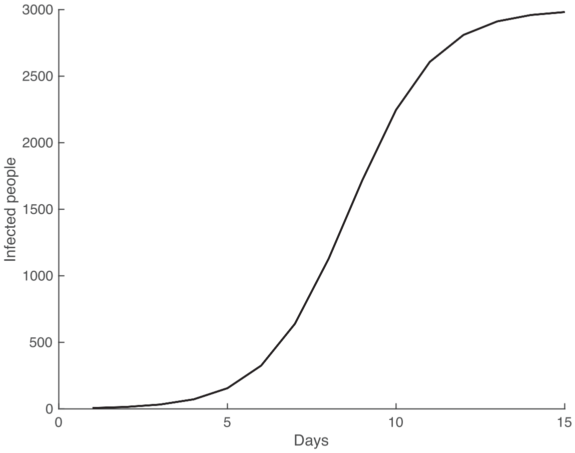

Fig. 1. The true logistic growth model for the spread of viral infection in the small town with $x_0 = 3$, $r = 0.8$ **and.**$N = 3000$

Given our assumption of fitting the function of the ODE model solutions and any necessary constants, $F(x_1(\beta, t_i^j), \ldots, x_k(\beta, t_i^j), \mathbf{a}^1, \ldots, \mathbf{a}^m)$, to the j^{th} time series data set, we set

$$E\left[D_i^j\right] = \mu_i^j = F\left(x_1\left(\beta, t_i^j\right), \ldots, x_k\left(\beta, t_i^j\right), \mathbf{a}^1, \ldots, \mathbf{a}^m\right). \tag{4}$$

Equation (4) can be thought of as a type of link function. In statistics, for generalized linear models (GLMs), a link function is defined as the function that transforms the mean of a distribution to a linear regression model (Montgomery, Peck, & Vining, 2006). Equation (4) equates the mean of the Gaussian distribution to the ODE model solutions.

4.2. Poisson distribution

Let Y be a random variable from the Poisson distribution with parameter $\mu > 0$, Y~POI(μ). The formulation of the Poisson distribution is given by the following discrete pdf, $f(y)$ (Bain & Engelhardt, 1987):

$$f(y) = \frac{\exp(-\mu)\mu^y}{y!}, \tag{5}$$

where $y = 0, 1, \ldots$.

The mean, E[Y], of the Poisson distribution is given by μ. For the Poisson distribution, the variance is equal to the mean, Var[Y] $=$ E[Y] $= \mu$.

Assume that the j^{th} time series data set is given by observations $D_j = \{d_1^j, \ldots, d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j, \ldots, t_{n_j}^j\}$ and that the probability of observing d_i^j is given by the Poisson distribution:

$$f\left(d_i^j\right) = \frac{\exp\left(-\mu_i^j\right)\mu_i^{j\left(d_i^j\right)}}{d_i^j!}, \tag{6}$$

where the mean $E[D_i^j] = \mu_i^j$ changes depending on the time, t_i^j. Hence, the variance, Var[D_i^j] $=$ E[D_i^j] $= \mu_i^j$, also changes over time.

Again, we will use equation (4) to equate the mean, $E[D_i^j] = \mu_i^j$, to the ODE model solutions.

The Poisson distribution is used for count data of rare events. The fact that the variance is dependent on the mean is particularly useful since in practice when observing count data over time the count data generally expresses more variability at higher values than at lower values (Bolker, 2007). The restriction that the variance is strictly equal to the mean is commonly violated for many types of count data. Count data where the variance is larger than the mean is called overdispersed. The negative binomial distribution can be used for count data with overdispersion.

4.3. Negative binomial distribution

Let Y be a random variable from the negative binomial distribution with parameters $0 < p < 1$ and $r \geq 0$, $Y\tilde{}\ \mathrm{NB}(r,p)$. The formulation of the negative binomial distribution is given by the following discrete pdf, $f(y)$ (Linden & Mantyniemi, 2011):

$$f(y) = \frac{\Gamma(y+r)}{y!\Gamma(r)} p^r (1-p)^y, \tag{7}$$

where $y = 0, 1, 2 \ldots.$

The interpretation of this formulation of the negative binomial distribution is that y are the number of failures before the r^{th} success and p is the probability of success per trial (Linden & Mantyniemi, 2011).

The mean, $\mathrm{E}[Y]$, of the negative binomial distribution is given by $\mu = \frac{r(1-p)}{p}$ and the variance, $\mathrm{Var}[Y]$, of this distribution is given by

$$\sigma^2 = \frac{r(1-p)}{p^2} = \frac{\mu}{p}.$$

For count data, the negative binomial distribution can be interpreted as the mean number of counts $\mathrm{E}[Y] = \mu$ with the variance $\mathrm{Var}[Y] = \frac{\mu}{p}$ overdispersed, since $0 < p < 1$, $\mathrm{Var}[Y] > \mathrm{E}[Y]$ (Bolker, 2007).

Assume that the j^{th} time series data set is given by observations $D_j = \{d_1^j, \ldots, d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j, \ldots, t_{n_j}^j\}$ and that the probability of observing d_i^j is given by the negative binomial distribution:

$$f\left(d_i^j\right) = \frac{\Gamma\left(d_i^j + r_i^j\right)}{d_i^j!\Gamma\left(r_i^j\right)} \left(p^j\right)^{\left(r_i^j\right)} \left(1 - p^j\right)^{d_i^j}, \tag{8}$$

where $r_i^j = \frac{(p^j)(\mu_i^j)}{1-(p^j)} \Leftrightarrow \mu_i^j = \frac{(r_i^j)(1-p^j)}{p^j}$ changes depending on the time, t_i^j and p^j is specific to the j^{th} data set. Hence, the variance, $\mathrm{Var}[D_i^j] = \frac{\mu_i^j}{p^j}$, also changes over time.

As before, we will use equation (4) to equate the mean, $E[D_i^j] = \mu_i^j$, to the ODE model solutions.

4.4. Other distributions

It is seen from sections 4.1, 4.2, and 4.3 that in general if the j^{th} time series data set is given by observations $D_j = \{d_1^j, \ldots, d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j, \ldots, t_{n_j}^j\}$ and the probability of observing d_i^j is given by the distribution with pdf $f(d_i^j)$ with mean $E[D_i^j] = \mu_i^j$, then equation (4) is used to equate the mean, $E[D_i^j] = \mu_i^j$, to the ODE model solutions.

5. Likelihood function

In a dynamical system, the dependency of solutions $x_1, \ldots, x_k$ on each other is built into the mathematical model itself. Assuming that the mathematical model correctly describes the data sets of interest, the data sets can be considered

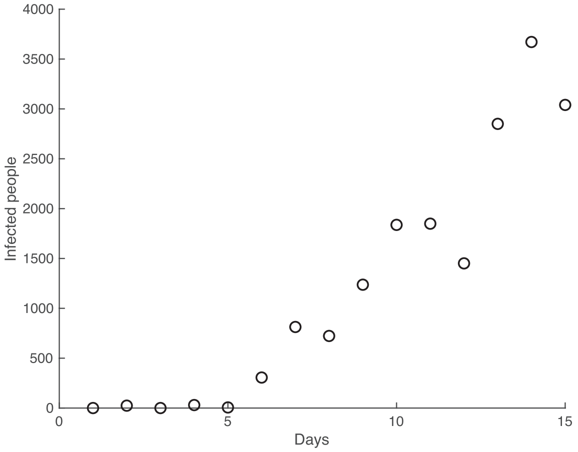

Fig. 2. The generated data for the spread of a viral infection in the small town.

independent from each other. With m independent time series data sets, there will be m likelihood functions associated with each of the independent data sets and the combined likelihood function is given by

$$L(\boldsymbol{\theta}) = CL_1(\boldsymbol{\theta})\bullet\ldots\bullet L_m(\boldsymbol{\theta}), \tag{9}$$

where θ is the vector of parameters to estimate, and C is any positive constant not depending on θ used to simplify the likelihood function (Kalbfleisch, 1979).

5.1. Gaussian probability model for m data sets and combined likelihood function

Assume, for $j = 1,\ldots,m$, that the j^{th} time series data set is given by observations $D_j = \{d_1^j, \ldots, d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j, \ldots, t_{n_j}^j\}$ and that the probability of observing d_i^j is given by the Gaussian distribution in equation (3) where the mean μ_i^j changes depending on the time, t_i^j and the variance $\frac{1}{\tau^j} > 0$ is specific to the j^{th} data set. Then the probability of the observed counts $D = \{D_1, \ldots, D_m\}$ is given by

$$\begin{aligned} P(D|\boldsymbol{\theta}) &= \prod_{j=1}^{m}\prod_{i=1}^{n_j}\sqrt{\frac{\tau^j}{2\pi}}\exp\left(-\frac{1}{2}\tau^j\left(d_i^j-\mu_i^j\right)^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{\left(\sum_{j=1}^{m}\frac{n_j}{2}\right)}\left(\tau^1\right)^{\frac{n_1}{2}}\bullet\ldots\bullet\left(\tau^m\right)^{\frac{n_m}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{m}\tau^j\sum_{i=1}^{n_j}\left(d_i^j-\mu_i^j\right)^2\right), \end{aligned} \tag{10}$$

where equation (4) is used to equate the mean, μ_i^j, to the ODE model solutions and

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\nu} \\ \tau^1 \\ \vdots \\ \tau^m \end{bmatrix}.$$

The Gaussian probability model is very beneficial for fitting since even poor initial guesses of the vector of parameters, $\boldsymbol{\theta}$, will still produce a nonzero probability.

The combined likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= C\left(\frac{1}{2\pi}\right)^{\left(\sum_{j=1}^{m}\frac{n_j}{2}\right)}\left(\tau^1\right)^{\frac{n_1}{2}}\bullet\ldots\bullet\left(\tau^m\right)^{\frac{n_m}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{m}\tau^j\sum_{i=1}^{n_j}\left(d_i^j-\mu_i^j\right)^2\right) \\ &= \left(\tau^1\right)^{\frac{n_1}{2}}\bullet\ldots\bullet\left(\tau^m\right)^{\frac{n_m}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{m}\tau^j\sum_{i=1}^{n_j}\left(d_i^j-\mu_i^j\right)^2\right), \end{aligned} \tag{11}$$

where $C = \left(\frac{1}{2\pi}\right)^{\left(-\sum_{j=1}^{m}\frac{n_j}{2}\right)}$ simplifies the likelihood function. The value of $\boldsymbol{\theta}$ that maximizes $P(D|\boldsymbol{\theta})$ will also maximize $L(\boldsymbol{\theta})$ (Kalbfleisch, 1979).

5.2. Poisson probability model for m data sets and combined likelihood function

Assume, for $j = 1,\ldots,m$, that the j^{th} time series data set is given by observations $D_j = \{d_1^j, \ldots, d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j, \ldots, t_{n_j}^j\}$ and that the probability of observing d_i^j is given by the Poisson distribution in equation (6) where the mean μ_i^j (and hence the variance, μ_i^j) changes depending on the time, t_i^j. Then the probability of the observed counts $D = \{D_1, \ldots, D_m\}$ is given by

$$\begin{aligned} P(D|\boldsymbol{\theta}) &= \prod_{j=1}^{m}\prod_{i=1}^{n_j}\frac{\exp\left(-\mu_i^j\right)\mu_i^{j\left(d_i^j\right)}}{d_i^j!} \\ &= \frac{1}{d_1^1!\bullet\ldots\bullet d_{n_1}^1!}\bullet\ldots\bullet\frac{1}{d_1^m!\bullet\ldots\bullet d_{n_m}^m!}\exp\left(-\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mu_i^j\right)\prod_{j=1}^{m}\left(\left(\mu_1^j\right)^{d_1^j}\bullet\ldots\bullet\left(\mu_{n_j}^j\right)^{d_{n_j}^j}\right), \end{aligned} \tag{12}$$

where equation (4) is used to equate the mean, μ_i^j, to the ODE model solutions and $\boldsymbol{\theta} = \boldsymbol{\nu}$.

The combined likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= C\frac{1}{d_1^1!\cdot\ldots\cdot d_{n_1}^1!}\cdot\ldots\cdot\frac{1}{d_1^m!\cdot\ldots\cdot d_{n_m}^m!}\exp\left(-\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mu_i^j\right)\prod_{j=1}^{m}\left(\left(\mu_1^j\right)^{d_1^j}\cdot\ldots\cdot\left(\mu_{n_j}^j\right)^{d_{n_j}^j}\right) \\ &= \exp\left(-\sum_{j=1}^{m}\sum_{i=1}^{n_j}\mu_i^j\right)\prod_{j=1}^{m}\left(\left(\mu_1^j\right)^{d_1^j}\cdot\ldots\cdot\left(\mu_{n_j}^j\right)^{d_{n_j}^j}\right), \end{aligned} \quad (13)$$

where $C = (d_1^1!\cdot\ldots\cdot d_{n_1}^1!)\cdot\ldots\cdot(d_1^m!\cdot\ldots\cdot d_{n_m}^m!)$ simplifies the likelihood function.

5.3. Negative binomial probability model for m data sets and combined likelihood function

Assume, for $j = 1,\ldots,m$, that the j^{th} time series data set is given by observations $D_j = \{d_1^j,\ldots,d_{n_j}^j\}$ with corresponding times $T_j = \{t_1^j,\ldots,t_{n_j}^j\}$ and that the probability of observing d_i^j is given by the negative binomial distribution in equation (8) where the mean μ_i^j (and hence the variance $\mathrm{Var}[D_i^j] = \frac{\mu_i^j}{p^j}$) changes depending on the time, t_i^j. Then the probability of the observed counts $D = \{D_1,\ldots,D_m\}$ is given by

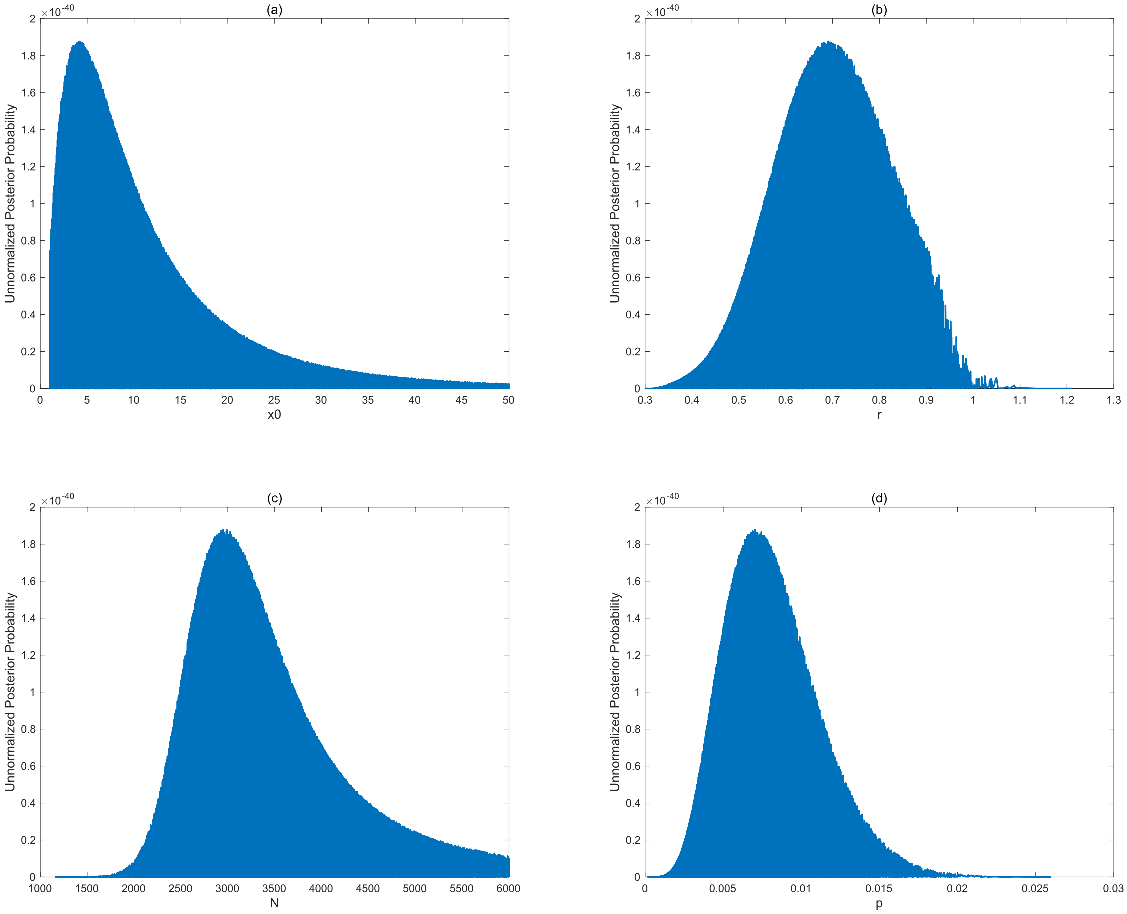

Fig. 3. Marginal unnormalized posterior distribution for **(a)** x_0, **(b)**r, **(c)**N, and **(d)**p.

$$\begin{aligned}P(D|\boldsymbol{\theta}) &= \prod_{j=1}^{m}\prod_{i=1}^{n_j}\frac{\Gamma\left(d_i^j+r_i^j\right)}{d_i^j!\Gamma\left(r_i^j\right)}\left(p^j\right)^{\left(r_i^j\right)}\left(1-p^j\right)^{d_i^j}\\ &= \left(\frac{1}{d_1^1!\bullet\ldots\bullet d_{n_1}^1!}\bullet\ldots\bullet\frac{1}{d_1^m!\bullet\ldots\bullet d_{n_m}^m!}\right)\\ &\left(\frac{\Gamma\left(d_1^1+r_1^1\right)\bullet\ldots\bullet\Gamma\left(d_{n_1}^1+r_{n_1}^1\right)}{\Gamma\left(r_1^1\right)\bullet\ldots\bullet\Gamma\left(r_{n_1}^1\right)}\bullet\ldots\bullet\frac{\Gamma(d_1^m+r_1^m)\bullet\ldots\bullet\Gamma\left(d_{n_m}^m+r_{n_m}^m\right)}{\Gamma(r_1^m)\bullet\ldots\bullet\Gamma\left(r_{n_m}^m\right)}\right)\\ &\left(\left(p^1\right)^{\sum_{i=1}^{n_1}r_i^j}\bullet\ldots\bullet(p^m)^{\sum_{i=1}^{n_m}r_i^j}\right)\left(\left(1-p^1\right)^{\sum_{i=1}^{n_1}d_i^j}\bullet\ldots\bullet(1-p^m)^{\sum_{i=1}^{n_m}d_i^j}\right),\end{aligned}\tag{14}$$

where $r_i^j = \frac{(p^j)(\mu_i^j)}{1-(p^j)} \Leftrightarrow \mu_i^j = \frac{(r_i^j)(1-p^j)}{p^j}$, equation (4) is used to equate the mean, μ_i^j, to the ODE model solutions and

$$\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\nu}\\ p^1\\ \vdots\\ p^m\end{bmatrix}.$$

The combined likelihood function is given by

$$\begin{aligned}L(\boldsymbol{\theta}) &= C\left(\frac{1}{d_1^1!\bullet\ldots\bullet d_{n_1}^1!}\bullet\ldots\bullet\frac{1}{d_1^m!\bullet\ldots\bullet d_{n_m}^m!}\right)\\ &\left(\frac{\Gamma\left(d_1^1+r_1^1\right)\bullet\ldots\bullet\Gamma\left(d_{n_1}^1+r_{n_1}^1\right)}{\Gamma\left(r_1^1\right)\bullet\ldots\bullet\Gamma\left(r_{n_1}^1\right)}\bullet\ldots\bullet\frac{\Gamma(d_1^m+r_1^m)\bullet\ldots\bullet\Gamma\left(d_{n_m}^m+r_{n_m}^m\right)}{\Gamma(r_1^m)\bullet\ldots\bullet\Gamma\left(r_{n_m}^m\right)}\right)\\ &\left(\left(p^1\right)^{\sum_{i=1}^{n_1}r_i^j}\bullet\ldots\bullet(p^m)^{\sum_{i=1}^{n_m}r_i^j}\right)\left(\left(1-p^1\right)^{\sum_{i=1}^{n_1}d_i^j}\bullet\ldots\bullet(1-p^m)^{\sum_{i=1}^{n_m}d_i^j}\right)\\ &= \left(\frac{\Gamma\left(d_1^1+r_1^1\right)\bullet\ldots\bullet\Gamma\left(d_{n_1}^1+r_{n_1}^1\right)}{\Gamma\left(r_1^1\right)\bullet\ldots\bullet\Gamma\left(r_{n_1}^1\right)}\bullet\ldots\bullet\frac{\Gamma(d_1^m+r_1^m)\bullet\ldots\bullet\Gamma\left(d_{n_m}^m+r_{n_m}^m\right)}{\Gamma(r_1^m)\bullet\ldots\bullet\Gamma\left(r_{n_m}^m\right)}\right)\\ &\left(\left(p^1\right)^{\sum_{i=1}^{n_1}r_i^j}\bullet\ldots\bullet(p^m)^{\sum_{i=1}^{n_m}r_i^j}\right)\left(\left(1-p^1\right)^{\sum_{i=1}^{n_1}d_i^j}\bullet\ldots\bullet(1-p^m)^{\sum_{i=1}^{n_m}d_i^j}\right),\end{aligned}\tag{15}$$

where $\mathrm{C} = (d_1^1!\bullet\ldots\bullet d_{n_1}^1!)\bullet\ldots\bullet(d_1^m!\bullet\ldots\bullet d_{n_m}^m!)$ simplifies the likelihood function.

6. Bayesian framework

The Bayesian framework is set up by first assuming a probability model for the observed data D given a $p\times 1$ vector of unknown parameters $\boldsymbol{\theta}$, which is $P(D|\boldsymbol{\theta})$. Then it is assumed that $\boldsymbol{\theta}$ is randomly distributed from the prior distribution $P(\boldsymbol{\theta})$. Statistical inference for $\boldsymbol{\theta}$ is based on the posterior distribution, $P(\boldsymbol{\theta}|\boldsymbol{D})$. Using Bayes' theorem we have

$$\begin{aligned}P(\boldsymbol{\theta}|D) &= \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)}\\ &= \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_{\Omega} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}\\ &\propto L(\boldsymbol{\theta})P(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|D),\end{aligned}\tag{16}$$

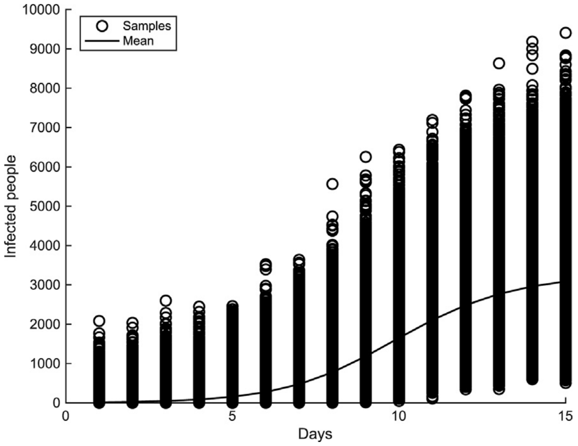

Fig. 4. Posterior predictive distribution with the posterior predictive mean.

where Ω is the parameter space of $\boldsymbol{\theta}$ and $L(\boldsymbol{\theta})$ is the likelihood function. $P(D) = \int_\Omega P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}$ is called the prior predictive distribution and it is the normalizing constant of the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{D})$ (Chen, Shao, & Ibrahim, 2000). The unnormalized posterior distribution is given by $\pi(\boldsymbol{\theta}|\boldsymbol{D}) = L(\boldsymbol{\theta})P(\boldsymbol{\theta})$.

The Bayesian framework is very useful to use for statistical inference that occurs in mathematical biology since there is generally prior information about the unknown parameters in the literature.

6.1. *Prior distribution*

In biological applications there may exist literature regarding an appropriate prior distribution for a parameter of interest. However, in many cases, only a general range is known from the literature about a parameter of interest and the uniform distribution is chosen as the prior distribution for the parameter of interest.

7. Markov Chain Monte Carlo algorithms

Markov Chain Monte Carlo (MCMC) algorithms are designed to sample and to fully explore the parameter space where the unnormalized posterior distribution is positive (Lynch, 2007). The MCMC algorithms involve a process where a new vector of parameter values is sampled from the posterior distribution, $\boldsymbol{\theta}^{(t)}$, based off of the previous vector of parameter values, $\boldsymbol{\theta}^{(t-1)}$. A successful MCMC algorithm results in a **sample path** (also called a **chain** or **walker**) that has arrived at a stationary process and covers the domain of the target unnormalized posterior distribution.

7.1. *Metropolis-Hastings algorithm*

The Metropolis-Hastings algorithm is one of the classic MCMC algorithms (Chen et al., 2000):
A starting point $\boldsymbol{\theta}^{(0)}$ is selected.
For every iteration $t = 1, 2, \ldots, T$:

```
randomly select a proposal for $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\gamma}$, from the proposal distribution $f(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})$
proposal for $\boldsymbol{\theta}^{(t)}$ is accepted with probability $\alpha = \min\left\{1, \frac{\pi(\boldsymbol{\gamma}|D)}{\pi(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{D})\, f(\frac{\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\gamma})}{f(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t-1)})}}\right\}$
random sample u from $U(0,1)$
if $u < \alpha$, the proposal is accepted and $\boldsymbol{\theta}^{(t)} = \boldsymbol{\gamma}$.
If not, $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$,
```

where $\pi(\boldsymbol{\theta}|\boldsymbol{D})$ is the unnormalized posterior distribution.

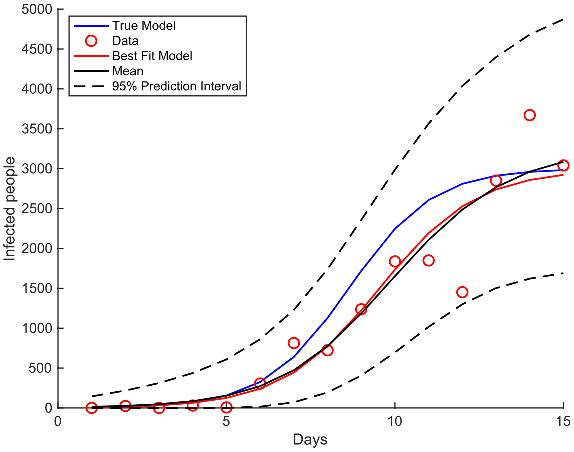

Fig. 5. Best fit and true model for the spread of a viral infection in the small town with 95% prediction interval.

7.1.1. *Random-walk Metropolis-Hastings algorithm*

If a symmetric proposal distribution is chosen in the Metropolis-Hastings Algorithm, then the proposal distribution randomly perturbs the current position of the vector of unknown parameters, $\boldsymbol{\theta}^{(\boldsymbol{t-1})}$, and these algorithms are called Random-Walk Metropolis-Hastings algorithms (Lynch, 2007).

A symmetric proposal distribution has the property that $f(\gamma|\boldsymbol{\theta}^{(\boldsymbol{t-1})}) = f(\boldsymbol{\theta}^{(\boldsymbol{t-1})}|\gamma)$ and this simplifies the acceptance probability to $\alpha = \min\left\{1, \frac{\pi(\gamma|D)}{\pi(\boldsymbol{\theta}^{(\boldsymbol{t-1})}|\boldsymbol{D})}\right\}$.

7.2. *Affine invariant ensemble Markov Chain Monte Carlo algorithm*

The affine invariant ensemble MCMC algorithm is shown to perform better than the Metropolis-Hastings algorithm and other MCMC algorithms (Goodman & Weare, 2010). The algorithm uses K walkers and the positions of the walkers are updated based on the present positions of the K walkers (Weikun, 2015, pp. 1–8). The following is the affine invariant ensemble MCMC algorithm:

A starting point $\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{0})}$ is selected for each of the walkers, $i = 1, 2, \ldots, K$.
For every iteration $t = 1, 2, \ldots, T$:
For $i = 1, 2, \ldots, K$:

randomly select a walker j from the K walkers such that $j \neq i$
randomly choose z from the distribution $f(z) = \frac{1}{\sqrt{az}}, \frac{1}{a} \leq z \leq a$
proposal for $\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t})}$ is $\boldsymbol{\gamma} = \boldsymbol{\theta}_{\boldsymbol{j}}^{(\boldsymbol{t-1})} + \boldsymbol{z}(\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t-1})} - \boldsymbol{\theta}_{\boldsymbol{j}}^{(\boldsymbol{t-1})})$ (Stretch Move)
proposal for $\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t})}$ is accepted with probability $\alpha = \min\left\{1, z^{p-1}\frac{\pi(\gamma|D)}{\pi(\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t-1})}|\boldsymbol{D})}\right\}$
random sample u from $U(0,1)$. If $u < \alpha$, the proposal is accepted and $\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t})} = \boldsymbol{\gamma}$. If not, $\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t})} = \boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t-1})}$,

where $\pi(\boldsymbol{\theta}|\boldsymbol{D})$ is the unnormalized posterior distribution, $a > 1$ is adjusted to improve performance, and $f(z)$ satisfies the symmetry condition $f\left(\frac{1}{z}\right) = zf(z)$.

The equation $\boldsymbol{\theta}_{\boldsymbol{j}}^{(\boldsymbol{t-1})} + \boldsymbol{z}(\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t-1})} - \boldsymbol{\theta}_{\boldsymbol{j}}^{(\boldsymbol{t-1})})$ is the equation of a line parallel to the vector $(\boldsymbol{\theta}_{\boldsymbol{i}}^{(\boldsymbol{t-1})} - \boldsymbol{\theta}_{\boldsymbol{j}}^{(\boldsymbol{t-1})})$. By randomly choosing z, the stretch move in the algorithm moves to a vector position, $\boldsymbol{\gamma}$, a certain distance up or down the line. Then the vector proposal, $\boldsymbol{\gamma}$, is either accepted or rejected based on the acceptance probability, α.

The set of samples from each of the K walkers will converge to the unnormalized posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{D})$. After running the method, the set of samples from each of the K walkers can be pooled together to form a larger sample from the unnormalized posterior distribution, KT samples. Since the samples from the first iterations are generally far away from the highest density of the unnormalized posterior distribution, the first iterations are usually deleted from each of the K walkers; the deletion of the first iterations is called *burn-in*. Let H be the number of pooled samples after the burn-in is completed.

8. Diagnostics

The samples from the MCMC provide a sample path. It is important to diagnose if this sample path produces a sample from the target unnormalized posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{D})$. In other words, the sample path converges to the target unnormalized posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{D})$. From the plot of the sample path, it is vital to find that the sample path has arrived at a stationary process and the sample path covers the domain of the target unnormalized posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{D})$.

The sample path for each parameter θ_i should be plotted. It is ideal to find that the sample path for each parameter θ_i is oscillating very fast and displays no apparent trend; this indicates that the sample path has arrived at a stationary process. By observing the marginal posterior distribution, $\pi(\theta_i|D)$ for each parameter θ_i, it should be observed that the sample path covers the domain of the target unnormalized posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{D})$.

A formalized test of the convergence of the MCMC sampling to the estimated unnormalized posterior distribution for each parameter θ_i is found by using a general univariate comparison method (Gelman & Brooks, 1998). The general univariate comparison method uses the distance of the empirical $100(1-\alpha)\%$ interval for the pooled samples, S, and divides this distance by the average of the distances of the empirical $100(1-\alpha)\%$ interval for each of the K walkers, s_i, to receive the potential scale reduction factor, η (Gelman & Brooks, 1998):

$$\eta = \frac{S}{\sum_{i=1}^{K} \frac{s_i}{K}}. \tag{17}$$

When the potential scale reduction factor, η, is close to 1 for all the estimated parameters, this indicates that the MCMC sampling converged to the estimated posterior distribution for each parameter.

9. Credible intervals for parameters

For a unimodel, symmetric marginal posterior distribution, $\pi(\theta_i|D)$, for θ_i, the 95% credible interval for θ_i is given by the 2.5 and 97.5 percentiles of the marginal posterior distribution of $\pi(\theta_i|D)$ (Chen et al., 2000).

9.1. Non-uniqueness

Non-uniqueness occurs when there is more than one solution vector $\boldsymbol{\theta}$ that explains the data, D, equally as well.

When there is non-uniqueness, the marginal posterior distribution, $\pi(\theta_i|D)$, for θ_i is constant over an interval and the credible interval for θ_i is given by the upper and lower limits of the interval (Chen et al., 2000).

The credible intervals resulting from non-uniqueness are still very beneficial since they are often more specific than the initial prior distributions specified for the parameters.

10. Posterior predictive distribution

Let $\tilde{D} = \{\tilde{D}_1, ..., \tilde{D}_m\}$ be future responses of interest for the m datasets. The posterior predictive distribution of $\tilde{D}$ is given by

$$P(\tilde{D}|D) = \int_{\Omega} P(\tilde{D}|\boldsymbol{\theta})P(\boldsymbol{\theta}|D)d\boldsymbol{\theta}, \tag{18}$$

where $P(\boldsymbol{\theta}|\boldsymbol{D})$ is the posterior distribution and $P(\tilde{D}|\boldsymbol{\theta})$ is the same probability model for the data specified in the Bayesian framework (16).

To generate the posterior predictive distribution.

For each pooled sample $t = 1, 2, ..., H$:

randomly sample $\tilde{D}$ from the probability distribution specified for the data $P(D|\boldsymbol{\theta}^{(\boldsymbol{t})})$ at $\boldsymbol{\theta}^{(\boldsymbol{t})}$,

where H is the number of samples from the unnormalized posterior distribution.

The 95% prediction intervals for each data set D_j is found by determining the 2.5 and 97.5 percentiles of the posterior predictive distribution at each t_i^j.

The posterior predictive mean is found by taking the mean of the posterior predictive distribution at each t_i^j.

11. An example: logistic growth

Assume there are three people infected with a virus in an isolated town of 3000 people. Furthermore, assume that the true model for the first 15 days of the virus across the population is plotted in Fig. 1 and given by the following differential equation

$$\frac{dx}{dt} = x\left(r - \frac{r}{N}x\right), \tag{19}$$

where $x_0 = 3$, $r = 0.8$ and $N = 3000$.

Now, this differential equation (19) can be solved analytically and we receive the logistic equation

$$x(\boldsymbol{\beta}, t_i) = \frac{rx_0}{\frac{r}{N}x_0 + \left(r - \frac{r}{N}x_0\right)e^{-rt_i}}, \tag{20}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} x_0 \\ r \\ N \end{bmatrix}.$$

Now, assume that the town collects count data for the number of people infected with the virus. We will generate this observed data by randomly sampling from the Negative Binomial distribution with mean given by (20) with $x_0 = 3$, $r = 0.8$ and $N = 3000$, and variance given by the mean divided by p, where p is chosen as 0.005. The generated observed data for the first 15 days of the virus across the population is plotted in Fig. 2.

Now, we will use Bayesian inference to determine the following unknown vector of parameters

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta} \\ p \end{bmatrix}$$

$$= \begin{bmatrix} x_0 \\ r \\ N \\ p \end{bmatrix}.$$

In this scenario, equation (4) is $E[D_i] = \mu_i = x(\boldsymbol{\beta}, t_i)$ and the negative binomial distribution (8) is chosen to describe the observed data.

The following uniform prior distributions are chosen for the parameters:

x_0 with distribution $U(1, 50)$
r with distribution $U(0.1, 2)$
N with distribution $U(100, 6000)$
p with distribution $U(1 \times 10^-5, 1 \times 10^-1)$.

The affine invariant ensemble MCMC algorithm is used with $T = 100000$ iterations and $K = 8$ walkers. The potential scale reduction factor, η, for each parameter:

$\eta = 0.9941$ for x_0
$\eta = 0.9977$ for r
$\eta = 0.9963$ for N
$\eta = 0.9987$ for p.

All potential scale reduction factors are close to 1 and this indicates that the algorithm converged to the posterior distribution.

The marginal unnormalized posterior distribution for each parameter is plotted in Fig. 3. The estimated parameters with 95% credible intervals are the following:

x_0 is estimated to be 4.13 (1.68, 19.58),
r is estimated to be 0.690 (0.474, 0.834),
N is estimated to be 2.99×10^3 (2.46×10^3, 4.47×10^3), and
p is estimated to be 0.0070 (0.0032, 0.0111).

The true parameter values for x_0, r, N, and p all lie within the 95% credible intervals.

Samples from the posterior predictive distribution and the posterior predictive mean are displayed in Fig. 4. The true model, best fit model (model with the highest unnormalized posterior probability), and posterior predictive mean are compared in Fig. 5. It is seen that the best fit model (red curve) lies very close to the posterior predictive mean (black curve) and is near the true model (blue curve). It is observed that the true model (blue curve) and all of the generated data (red circles) lie within the 95% prediction intervals (dashed black curves).

Declaration of competing interest

I wish to confirm that there are no known conflicts of interest associated with these lecture notes.

References

Bain, L. J., & Engelhardt, M. (1987). *Introduction to probability and mathematical statistics* (2nd ed. Edition). Brooks/Cole.

Bolker, B. (2007). *Ecological models and data in R*. Princeton-New Jersey: Princeton University Press.

Chen, M., Shao, Q., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. New York-New York: Springer-Verlag.

Gelman, A., & Brooks, S. P. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat., 7*(4), 434–455.

Ghasemi, O., Lindsey, M. L., Yang, T., Nguyen, N., Huang, Y., & Jin, Y.-F. (2011). Bayesian parameter estimation for nonlinear modelling of biological pathways. *BMC Syst. Biol., 5*(Suppl 3), S9.

Goodman, J., & Weare, J. (2010). Ensemble samplers with affine invariance. *Comm. App. Math,. Com. Sc., 5*(1), 65–80.

Higham, C. F., & Husmeier, D. (2013). A bayesian approach for parameter estimation in the extended clock gene circuit of arabidopsis thaliana. *BMC Bioinformatics, 14*(Suppl 10), S3.

Kalbfleisch, J. G. (1979). *Probability and statistical inference* (Vol. 2). Springer-Verlag New York, Inc.. Statistical Inference.

Linden, A., & Mantyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology, 92*(7), 1414–1421.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

Ma, Y. Z., & Berndsen, A. (2014). How to combine correlated data sets - a bayesian hyperparameter matrix method. *Astron. Comput., 5*, 45–56.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis*. Hoboken-New Jersey: John Wiley & Sons, Inc.

Periwal, V., Chow, C. C., Bergman, R. N., Ricks, M., Vega, G. L., & Sumner, A. E. (2008). Evaluation of quantitative models of the effect of insulin on lipolysis and glucose disposal. *Am. J. Physiol. Regul. Integr. Comp. Physiol., 295*, R1089–R1096.

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., & van Riel, N. A. W. (2012). A bayesian approach to targeted experiment design. *Bioinformatics, 28*(8), 1136–1142.

Weikun, C. (2015). *A parallel implementation of mcmc*.