Gaussian process approximations for fast inference from infectious disease data

Elizabeth Buckingham-Jeffery[1,3,*], Valerie Isham[2], Thomas House[3]

1. Centre for Complexity Science, University of Warwick, Coventry, CV4 7AL, UK.
2. Department of Statistical Science, University College London, London, WC1E 6BT, UK.
3. School of Mathematics, University of Manchester, Manchester, M13 9PL, UK.
** Corresponding Author: E.Buckingham-Jeffery@manchester.ac.uk*

Abstract

We present a flexible framework for deriving and quantifying the accuracy of Gaussian process approximations to non-linear stochastic individual-based models of epidemics. We develop this for the SIR and SEIR models, and show how it can be used to perform quick maximum likelihood inference for the underlying parameters given population estimates of the number of infecteds or cases at given time points. We also show how the unobserved processes can be inferred at the same time as the underlying parameters.

Keywords: SIR; SEIR; Stochastic Taylor Expansion; MLE.

1 Introduction

Analysing data in real time allows us to learn about diseases, to estimate key parameters to understand disease dynamics, and to evaluate interventions. Often we have imperfect, incomplete observations that we would like to analyse quickly so our results can be useful to public health authorities. Many methods exist to fit to data the non-linear epidemic models that are used in policy (O'Neill, 2010), depending on the model and data involved. Here we are interested in the case of daily or weekly data on the number of infecteds or cumulative incidence, where there is no tractable closed-form likelihood and where computationally intensive methods such as multiple imputation are too slow.

There has been an explosion in methods to deal with the intractability of likelihoods in such cases in recent years. Examples include iterative filtering (Ionides et al., 2006), Approximate Bayesian Computation based on Sequential Monte Carlo (ABC-SMC) (Toni et al., 2009), particle-Markov chain Monte Carlo (PMCMC) (Andrieu et al., 2010), SMC^2 (Chopin et al., 2013), simulation-based inference (McKinley et al., 2014), forward simulation MCMC (Neal and Terry Huang, 2015), and many others. These approaches are, however, typically computationally intensive and can be difficult to implement.

On the other hand, fitting deterministic ordinary differential equation (ODE) models for epidemics directly to data (for example by least-squares fitting) is often an ill-posed inverse problem; it is widely accepted that stochastic effects need to be included if inference and prediction are to be reliable (Leander et al., 2014; House, 2015; King et al., 2015). One low-dimensional approach to stochasticity involves working in the diffusion limit (Dargatz, 2010; Guy et al., 2015; Leite and Williams, 2017), however this can involve additional technical and computational problems due to the analytical intractability of the diffusion limit itself.

We therefore consider here a Gaussian process approximation approach to speed-up real time analysis of disease data when other methods are too complex. This simultaneously accounts for stochasticity and avoids the problems of direct fitting of ODE models. This approximation is applied initially to the stochastic SIR (susceptible-infectious-removed) model. We also consider the SEIR (susceptible-exposed-infectious-removed) model, although it is easy to use this approximation scheme with a more complex compartmental model or even models used outside epidemiology. We stress that our approach involves an additional approximation step beyond the derivation of a stochastic differential equation (SDE) limit, that must be controlled and which is the main technical component of our work.

Other authors have made use of Gaussian approximations in epidemic inference. Ross et al. (2006) considered parameter estimation for the SIS model, Fearnhead et al. (2014) and Zimmer et al. (2017) considered Gaussian approximations based on the linear noise approximation, and Ball and House (2017) considered inference for the SIR epidemic on a network using a Gaussian process approximation based on the covariance function for an approximating branching process.

In this work, we present three main developments. First, we consider a very broad class of Gaussian process approximations, including those based on stochastic moment closure (which have branching process results as a special case) and those based on a linear time-inhomogeneous SDE (which have linear noise approximation approaches as a special case). Secondly, we give a method for deriving bounds on the errors of any Gaussian process approximation through Taylor expansion in the time interval between observations, using the scheme shown in Figure 1. Finally, we go into more depth than previously on the numerical comparison of different Gaussian process approximations for parameter inference on simulated data, and on how to apply these approaches to the data on cumulative incidence typically available from epidemiological outbreak reports.

2 Models and notation

While our approach is general, we will focus here on epidemic dynamics, in particular the Markovian SIR model, also called the general stochastic epidemic (Bailey, 1975; Andersson and Britton, 2000). We will define the different approaches as shown in Figure 1.

2.1 Pure jump Markov chain

We will write $N_S(t)$, $N_I(t)$ and $N_R(t)$ for the random integer numbers in the population who are susceptible, infectious and removed respectively. The vector $\mathbf{N}(t) = (N_S(t), N_I(t), N_R(t))$ is then a continuous-time Markov chain with events and rates

$$\begin{aligned} (N_S, N_I, N_R) &\to (N_S - 1, N_I + 1, N_R) \text{ at rate } \beta \frac{N_S N_I}{N} \ , \\ (N_S, N_I, N_R) &\to (N_S, N_I - 1, N_R + 1) \text{ at rate } \gamma N_I \ , \end{aligned} \tag{1}$$

where the population size $N = N_S + N_I + N_R$ is a constant.

2.2 Diffusion approximation

Using the convergence results of Kurtz (1970, 1971) the Markov chain defined by (1) is well approximated by the solution $\mathbf{X}(t)$ of the SDE

$$\mathrm{d}\mathbf{X} = \mathbf{F}(\mathbf{X})\,\mathrm{d}t + \sqrt{\boldsymbol{V}(\mathbf{X})}\,\mathrm{d}\mathbf{W} \ , \tag{2}$$

where (ignoring the removed individuals who do not need to be counted due to the constant population size) we let

$$\mathbf{X}(t) = \begin{pmatrix} S(t) \\ I(t) \end{pmatrix} \ , \quad \mathbf{F}(\mathbf{X}) = \begin{pmatrix} -\beta SI/N \\ \beta SI/N - \gamma I \end{pmatrix} \ , \quad \boldsymbol{V}(\mathbf{X}) = \begin{pmatrix} \beta SI/N & -\beta SI/N \\ -\beta SI/N & \beta SI/N + \gamma I \end{pmatrix} \ , \tag{3}$$

and $\mathbf{W} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ where W_1 and W_2 are two independent Wiener processes. Explicitly, this approximation takes the form

$$N_S \approx S \ , \qquad N_I \approx I \ . \tag{4}$$

Note that the distribution of $\mathbf{X}(\Delta t)|\mathbf{X}(0)$ given by (2) will not in general be Gaussian.

2.3 Deterministic approximation

We can then further apply the results of Kurtz (1970, 1971) to derive the deterministic approximation

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\frac{\beta}{N} si \ , \qquad \frac{\mathrm{d}i}{\mathrm{d}t} = \frac{\beta}{N} si - \gamma i \ , \tag{5}$$

where $s(t)$ and $i(t)$ are the numbers of susceptible and infectious individuals respectively at time t that satisfy this deterministic model.

2.4 Multivariate normal moment closure

For comparison, the multivariate normal (MVN) moment closure method is also used to obtain an approximation of the stochastic SIR model (Isham, 1991). In this method, it is assumed that the joint distribution of the susceptible and infectious populations can be approximated by a bivariate normal distribution. This gives a set of 5 ODEs for the mean and variances of the susceptible and infectious populations.

We write $X(t)$ and $Y(t)$ for the number of susceptible and infectious people respectively in the MVN moment closure approximation of the stochastic SIR model at time t. These follow a Gaussian process $\mathrm{GP}(\boldsymbol{\mu}(t), \boldsymbol{\sigma}(t))$ with mean and variance-covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X(t)] \\ \mathbb{E}[Y(t)] \end{pmatrix} \ , \qquad \boldsymbol{\sigma} = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{pmatrix} = \begin{pmatrix} \mathrm{var}(X(t)) & \mathrm{cov}(X(t), Y(t)) \\ \mathrm{cov}(X(t), Y(t)) & \mathrm{var}(Y(t)) \end{pmatrix} \ , \tag{6}$$

that obey equations

$$\begin{aligned}
\frac{\mathrm{d}\mu_X}{\mathrm{d}t} &= -\frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY}) \ , \\
\frac{\mathrm{d}\mu_Y}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY}) - \gamma\mu_Y \ , \\
\frac{\mathrm{d}\sigma_{XX}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X\mu_Y + \sigma_{XY} - 2\mu_X\sigma_{XY} - 2\mu_Y\sigma_{XX}) \ , \\
\frac{\mathrm{d}\sigma_{XY}}{\mathrm{d}t} &= \frac{\beta}{N}(\mu_X(\sigma_{XY} - \sigma_{YY}) + \mu_Y(\sigma_{XX} - \sigma_{XY}) - \mu_X\mu_Y - \sigma_{XY}) - \gamma\sigma_{XY} \ , \\
\frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} &= \frac{\beta}{N}(2\mu_X\sigma_{YY} + 2\mu_Y\sigma_{XY} + \mu_X\mu_Y + \sigma_{XY}) - \gamma(2\sigma_{YY} - \mu_Y) \ ,
\end{aligned} \tag{7}$$

which are derived from applying the assumption of multivariate normality to moment equations derived from the Markov chain (1).

2.5 Linear stochastic process approximation

The linear stochastic process (LSP) approximation approach introduced in the SIR context by Isham (1991) (referring to a more complex model of AIDS due to Tan and Hsu (1989)) is to let the susceptible population evolve deterministically, while the infectious individuals are normally distributed.

This gives the following set of three ODEs for the evolution of the deterministic susceptible population, $s(t)$, and the mean, $\mu_Y = \mathbb{E}[Y(t)]$, and variance, $\sigma_{YY} = \mathrm{var}(Y(t))$, of the infectious population:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\frac{\beta}{N} s\mu_Y \ , \\ \frac{\mathrm{d}\mu_Y}{\mathrm{d}t} &= \frac{\beta}{N} s\mu_Y - \gamma\mu_Y \ , \\ \frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} &= \frac{\beta}{N}\left(2s\sigma_{YY} + s\mu_Y\right) - \gamma(2\sigma_{YY} - \mu_Y) \ . \end{aligned} \tag{8}$$

While we mention this approach to show that various Gaussian moment closures are possible, for inferential purposes we will typically require the susceptible population to be random rather than deterministic making it unsuitable.

2.6 Linear SDE approximation

Following Archambeau et al. (2007), we note that the SDE

$$\mathrm{d}\mathbf{x} = \left(\boldsymbol{A}(t)\mathbf{x} + \mathbf{b}(t)\right)\mathrm{d}t + \sqrt{\boldsymbol{U}(t)}\,\mathrm{d}\mathbf{W} \ , \tag{9}$$

will have a Gaussian process solution $\mathrm{GP}(\mathbf{m}(t), \boldsymbol{C}(t))$ with mean and variance-covariance matrix satisfying

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \boldsymbol{A}\mathbf{m} + \mathbf{b} \ , \qquad \frac{\mathrm{d}\boldsymbol{C}}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{C} + \boldsymbol{C}\boldsymbol{A}^\top + \boldsymbol{U} \ . \tag{10}$$

Our aim is therefore to choose the time-varying matrices in equation (9) so that it approximates (2) and hence the full stochastic system. In this context, there is one 'obvious' choice for the matrix $\boldsymbol{U}(t)$:

$$\boldsymbol{U}(t) = \begin{pmatrix} (\beta/N)s(t)i(t) & -(\beta/N)s(t)i(t) \\ -(\beta/N)s(t)i(t) & (\beta/N)s(t)i(t) + \gamma i(t) \end{pmatrix} \ . \tag{11}$$

For the matrix $\boldsymbol{A}(t)$ and the vector $\mathbf{b}(t)$, however, there are many choices that have the correct mean behaviour. We will consider how an existing approach in the literature (the Linear Noise model) is a special case of the linear SDE approach, and also introduce other special cases that we do not believe are named.

We write $\mathcal{X}(t)$, $\mathcal{Y}(t)$ for the number of susceptible and infectious people respectively in this approximation of the stochastic SIR model at time t. These follow a Gaussian process $\mathrm{GP}(\mathbf{m}(t), \boldsymbol{C}(t))$ with mean and variance-covariance matrix

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\mathcal{X}(t)] \\ \mathbb{E}[\mathcal{Y}(t)] \end{pmatrix} \ , \qquad \boldsymbol{C} = \begin{pmatrix} C_{11} & C_{21} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} \mathrm{var}(\mathcal{X}(t)) & \mathrm{cov}(\mathcal{X}(t), \mathcal{Y}(t)) \\ \mathrm{cov}(\mathcal{X}(t), \mathcal{Y}(t)) & \mathrm{var}(\mathcal{Y}(t)) \end{pmatrix} \ . \tag{12}$$

2.6.1 Linear noise approximation

To derive the linear noise (LN) approximation (Uhlenbeck and Ornstein, 1930; Black and McKane, 2010) we start with the SDE (2) and write

$$S(t) = s(t) + \tilde{S}(t) \ , \qquad I(t) = i(t) + \tilde{I}(t) \ , \tag{13}$$

where the quantities $\tilde{S}, \tilde{I}$ are assumed to be small in the approximation. Ignoring quadratic terms $O(\tilde{S}^2, \tilde{S}\tilde{I}, \tilde{I}^2)$ and keeping only linear terms gives

$$\begin{aligned} \mathrm{d}\begin{pmatrix} S(t) \\ I(t) \end{pmatrix} \approx & \begin{pmatrix} -(\beta/N)(s(t)I(t) + S(t)i(t) - s(t)i(t)) \\ (\beta/N)(s(t)I(t) + S(t)i(t) - s(t)i(t)) - \gamma I(t) \end{pmatrix} \mathrm{d}t \\ & + \sqrt{\begin{matrix} (\beta/N)s(t)i(t) & -(\beta/N)s(t)i(t) \\ -(\beta/N)s(t)i(t) & (\beta/N)s(t)i(t) + \gamma i(t) \end{matrix}} \ \mathrm{d}\mathbf{W} \ . \end{aligned} \tag{14}$$

This is a special case of the linear SDE approach where

$$\boldsymbol{A}(t) = \begin{pmatrix} -(\beta/N)i(t) & -(\beta/N)s(t) \\ (\beta/N)i(t) & (\beta/N)s(t) - \gamma \end{pmatrix} \ ; \qquad \mathbf{b}(t) = \begin{pmatrix} (\beta/N)s(t)i(t) \\ -(\beta/N)s(t)i(t) \end{pmatrix} \ . \tag{15}$$

2.6.2 Other special cases

We consider two other special cases, which we name in an obvious way,

$$\begin{aligned} &\text{‘A noise’:} \quad \boldsymbol{A}(t) = \begin{pmatrix} -(\beta/N)i(t) & 0 \\ 0 & (\beta/N)s(t) - \gamma \end{pmatrix} \ ; \quad \mathbf{b}(t) = \mathbf{0} \ . \\ &\text{‘b noise’:} \quad \boldsymbol{A}(t) = \mathbf{0} \ ; \qquad \mathbf{b}(t) = \begin{pmatrix} -(\beta/N)s(t)i(t) \\ (\beta/N)s(t)i(t) - \gamma i(t) \end{pmatrix} \ . \end{aligned} \tag{16}$$

For each of these we obtain a set of five ODEs from (10), which we can solve numerically to give an approximation of the mean and variances of the stochastic SIR model. Note that there are more special cases, similar to these, for which we can write down $\boldsymbol{A}(t)$ and $\mathbf{b}(t)$.

2.7 Comparison of multivariate normal moment closure and linear SDE approximations

The main difference between the MVN moment closure and linear SDE approaches is as follows. For the MVN moment closure, we have

$$\frac{\mathrm{d}\mathbb{E}[Y]}{\mathrm{d}t} = \frac{\beta}{N}(\mathbb{E}[X]\mathbb{E}[Y] + \mathrm{cov}(X, Y)) - \gamma \mathbb{E}[Y] \ . \tag{17}$$

For the linear SDE, in contrast, we have

$$\frac{\mathrm{d}\mathbb{E}[\mathcal{Y}]}{\mathrm{d}t} = \frac{\beta}{N}\mathbb{E}[\mathcal{X}]\mathbb{E}[\mathcal{Y}] - \gamma \mathbb{E}[\mathcal{Y}] \ . \tag{18}$$

So the MVN moment closure allows for the negative correlation of the susceptibles and infecteds at the beginning (and end) of the outbreak. This means that $\mathbb{E}[Y]$ initially increases more slowly which allows for a random delay time before the outbreak takes off.

3 Numerical comparisons

We first numerically compare each approximation to simulations of the stochastic process, for a range of epidemiological model parameter values and population sizes N, with 50 initial infectious individuals.

Due to our interest in regularly spaced data, we simulate trajectories of both the susceptible and infectious populations of the stochastic SIR model at regular time intervals between known data points using the tau-leap algorithm (Gillespie, 2001). We do this for each set of model parameters of interest.

Figure 2 shows an example of these trajectories for a specific set of parameter values. Simulating 10^4 trajectories gives a distribution at each time point to which we compare each approximating Gaussian distribution.

For each approximation we computed the mean and variance of the size of the susceptible and infectious populations forward from the current data point until the next. The mean of the approximation was then reset to the data point, and the variances to zero. Figure 2 shows an example of this for one specific set of parameter values and one specific approximation.

We compared the approximations numerically to the stochastic simulations using the Kullback-Leibler (KL) divergence, a measure of difference between two probability distributions. The Gaussian distributions of the approximations were discretised in order to compare with the discrete distribution given by the stochastic simulations. For discrete probability distributions P and Q the KL divergence is defined as

$$D_{\mathrm{KL}}(P||Q) = \sum_i P(i)(\ln P(i) - \ln Q(i)) \ . \tag{19}$$

A better approximation results in a smaller KL divergence (MacKay, 2003). The KL divergence was computed each time data were obtained and before the simulations and approximations were reset to the data value.

Note that we could not compute the KL divergence for a comparison of the LSP with the stochastic simulations for the size of the susceptible population as in the LSP the susceptible population evolves deterministically. Additionally, on occasion the KL divergence could not be computed for other approximations because one distribution took a value very close to zero. However, as displayed in Figure 3, this does not occur very often.

Figure 3 demonstrates these comparisons for three examples of epidemiological model rate parameters that we have chosen to cover a range of R_0 values. The MVN moment closure and LN approximations consistently have the smallest KL divergence in both the size of the susceptible population and the size of the infectious population (Figure 3). Additionally, and in particular for larger population sizes, the A noise and LSP approximations also approximate well the size of the infectious population. The b noise approximation does not approximate the size of the infectious class as well, in particular at the start and end of the epidemic.

For approximating the size of the susceptible population, the A noise and b noise approximations perform adequately but not as well as the other approximations particularly at the start and end of the epidemic.

We performed the same analysis over longer time steps, Δt, and saw similar results; the MVN moment closure and LN approximations are best, with the A noise approximation also a good approximation for the infectious population. However, the A noise approximation becomes a much less good approximation of the susceptible population over longer time steps.

The ODEs that define the approximations were solved numerically. The b noise approximation has the simplest set of ODEs and so is fastest to solve. The A noise and LSP approximations are slower and, finally, the MVN moment closure and LN approximations take longest (Figure 4).

In addition, we note the ease with which the A noise and b noise approximations can be derived (in essence, just written down) in comparison to the MVN moment closure and LN approaches. This will become more pronounced for more complex compartmental models (for example, as demonstrated in section 5.2 when we apply these approximations to the SEIR model).

In conclusion, these numerical comparisons show that the A noise Gaussian process approximation can perform comparably to the MVN moment closure and LN approaches, in particular for large population sizes and for the infectious population size, while being computationally faster and more simple to derive. We expect these advantages to become much more pronounced for more complex models.

We now consider what progress is possible analytically.

4 Analytical comparisons

Each of the Gaussian models previously described is chosen on the basis of different *a priori* assumptions, however we would like to find a way to control the errors introduced in a systematic manner. Since we are interested in regularly-spaced observations, the relevant control parameter is the timestep Δt. Our starting point is the diffusion approximation to the stochastic SIR model in the regime where the susceptible population is approximately constant (e.g. at N when this is close to its starting value). This is chosen to simplify calculations, although we note the work of Cauchemez and Ferguson (2008) suggests that the approximation of constant susceptible population can be made throughout the epidemic if the time period Δt over which it is made is relatively small. In this limit, the system is described by the SDE first analysed by Feller (1951),

$$\mathrm{d}I = rI\,\mathrm{d}t + \sqrt{\rho I}\,\mathrm{d}W \ , \tag{20}$$

where $r = \beta - \gamma$ and $\rho = \beta + \gamma$. We will now consider how to expand this equation in Δt.

4.1 Taylor scheme for SDEs

The origin of the methods used to derive Taylor schemes for SDEs is often attributed to Milstein (1975), who derived a scheme up to $O(\Delta t)$. In fact, there are many such schemes that solve the SIR SDE locally in time; we use the weak order-3 scheme given by Kloeden and Platen (1992). This scheme has a rather complex general form, however for the specific equation (20) subject to initial condition $I(0) = I_0 \gg 1$ we obtain the following result following some analytical work:

$$I(\Delta t) = I_0\left(1 + r\Delta t + \frac{1}{2}r^2\Delta t^2\right) + \left(\rho I_0 \Delta t\left(1 + \frac{3}{2}r\Delta t + \frac{7}{6}r^2\Delta t^2\right)\right)^{1/2} U + O(\Delta t^3, {I_0}^0) \ , \tag{21}$$

where $U \sim \mathcal{N}(0,1)$ is a standard normal random variable and $O(\Delta t^3, {I_0}^0)$ represents terms containing the variables Δt^n for $n \geq 3$ and I_0^m for $m \geq 0$. Such a stochastic Taylor expansion can be carried out for any SDE, including multi-dimensional ones, and as such we believe this method has significant promise for evaluation of Gaussian process approximations in general. While the complexity of analysis can grow, it is possible to work with such stochastic systems using computer algebra (Kendall, 2005), and we think that this would be an interesting direction for future research.

4.2 Local solution of ODEs giving the approximations

For the Gaussian process approximations that arise from linear SDEs, we consider (10) in the limit where the size of the susceptible population is approximately constant to give the following ODEs for the mean, m_2, and variance, C_{22}, of the number of infecteds,

$$\begin{aligned} \frac{\mathrm{d}m_2}{\mathrm{d}t} &= A_{21}(t)N + A_{22}(t)m_2(t) + b_2(t) \ , \\ \frac{\mathrm{d}C_{22}}{\mathrm{d}t} &= 2A_{22}(t)C_{22}(t) + \rho i(t) \ . \end{aligned} \tag{22}$$

Solving these, using a standard Taylor series expansion subject to initial conditions $m_2(0) = i(0) = I_0$ and $C_{22}(0) = 0$, gives for each model

$$m_2(\Delta t) = I_0\left(1 + r\Delta t + \frac{1}{2}r^2\Delta t^2 + \cdots\right) \ , \tag{23}$$

i.e. as we would expect the Taylor series for $\mathrm{e}^{r\Delta t}$. For b noise, we have

$$C_{22}(\Delta t) = \rho I_0 \left(1 + \frac{1}{2} r\Delta t + \frac{1}{6} r^2 \Delta t^2 + \cdots \right) \Delta t \ , \tag{24}$$

where for all other models (A noise and LN) we have

$$C_{22}(\Delta t) = \rho I_0 \left(1 + \frac{3}{2} r\Delta t + \frac{7}{6} r^2 \Delta t^2 + \cdots \right) \Delta t \ . \tag{25}$$

For the MVN moment closure approximation, the mean, $\mu_Y(t)$, and variance, σ_{YY}, of the epidemic at constant susceptible population are given by the following ODEs:

$$\begin{aligned} \frac{\mathrm{d}\mu_Y}{\mathrm{d}t} &= r\mu_Y, \\ \frac{\mathrm{d}\sigma_{YY}}{\mathrm{d}t} &= \rho\mu_Y + 2r\sigma_{YY}, \end{aligned} \tag{26}$$

which we can also solve subject to $\mu_Y = I_0$, $\sigma_{YY} = 0$ to obtain

$$\begin{aligned} \mu_Y(\Delta t) &= I_0 \left(1 + r\Delta t + \frac{1}{2} r^2 \Delta t^2 + \cdots \right) \ , \\ \sigma_{YY}(\Delta t) &= \rho I_0 \left(1 + \frac{3}{2} r\Delta t + \frac{7}{6} r^2 \Delta t^2 + \cdots \right) \Delta t \ . \end{aligned} \tag{27}$$

4.3 Bounding the errors of the Gaussian process

Putting the results above together, we see that the MVN moment closure, linear noise and A noise are all consistent with the SDE (20) expanded as in equation (21), so we will continue to work with these approaches, while the b noise is significantly less accurate and so we do not consider it further.

To see why errors at this order represent a significant improvement on other possible approaches, consider the Euler-Maruyama approximation to (20),

$$J(\Delta t) = (1 + r\Delta t) I_0 + \sqrt{\rho I_0 \Delta t} U \ , \tag{28}$$

where U is a standard Gaussian random variable. Comparing to (21) we see that errors to this appear at $O(\Delta t)$. This tells us that the ODE-based Gaussian process approximations we continue to analyse (MVN, LN and A) are more accurate than Euler-Maruyama steps by several orders of magnitude in Δt.

5 Inference

We initially apply the Gaussian process approximations to synthetic data from the SIR model to demonstrate that the size of the susceptible population can be recovered from weekly measurements of the number of infecteds. Secondly, we consider real data from a norovirus outbreak with the SEIR (susceptible-exposed-infectious-removed) model to demonstrate that it is straightforward to use these approximations with more complex models.

5.1 Simulated data from the SIR model

Consider a disease that is well approximated by the SIR model with transmission rate constant β and recovery rate constant γ. Suppose we have data of the form of a set of times $\{t_i\}_{i=0}^n$ together with associated measurements of the number of infecteds $\{y_i\}_{i=0}^n$. Consider a Gaussian process approximation to the SIR model such that given susceptible and infectious populations of size x_0 and y_0 respectively at the start of a time interval of length Δt, at the end of that time interval the mean and variance-covariance matrix are $\boldsymbol{\mu}(\Delta t; x_0, y_0, \beta, \gamma)$ and $\boldsymbol{\Sigma}(\Delta t; x_0, y_0, \beta, \gamma)$ respectively. If we also had measurements of the susceptible population, $\{x_i\}_{i=0}^n$, then we could write the likelihood function for the parameters of the approximating model given the data as

$$L(\beta, \gamma; \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \mathcal{N}\left((x_i, y_i); \boldsymbol{\mu}(t_i - t_{i-1}; x_{i-1}, y_{i-1}, \beta, \gamma), \boldsymbol{\Sigma}(t_i - t_{i-1}; x_{i-1}, y_{i-1}, \beta, \gamma)\right) \ . \tag{29}$$

In practice, the data on the susceptible population are not readily available and instead this information can be imputed using the marginal and marginal conditional distributions of a multivariate normal distribution. These can be explicitly computed as follows (Eaton, 1983). For random vector (x, y) with multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ an observation y_i has marginal probability density function

$$f(y_i) = \mathcal{N}(y_i; \mu_2, \Sigma_{22}) \ , \tag{30}$$

and conditional on this observation x_i has marginal conditional probability density function

$$f(x_i; y_i) = \mathcal{N}(x_i; \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_i - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \ . \tag{31}$$

We can use these rules at each observation point to build up a likelihood from the product of terms such as (30) and also to update the mean vector and variance-covariance matrix for the Gaussian process approximation.

Synthetic data were obtained from one simulation of the stochastic SIR model using parameter values $\beta = 2$, $\gamma = 1$, and $N = 1 \times 10^4$ with one initial infected. Using each of the MVN moment closure, LN, and A noise approximations, we were reliably able to recover the epidemiological model parameters as shown in Figure 5. The three approximation methods all gave similar results for parameter estimates (estimates $\hat{\beta} = 2.04$ and $\hat{\gamma} = 1.01$) and for the size of the susceptible population from regular data on the number of infecteds.

5.2 Cumulative incidence from a real norovirus outbreak with the SEIR model

In the previous section we discussed the case when the available data are the number of infecteds at regular time points. An alternative, and common, situation is when only illness onset times, and not recovery times are available. For an SIR model this corresponds to having measurements of the *cumulative incidence* $N - S(t)$.

In this section, we consider real data from an outbreak of norovirus on a cruise ship visiting the British Isles as reported by Vivancos et al. (2010). This report gives us data on the number of new reported norovirus cases per day during this outbreak in a small, closed population, $N = 1714$. A single norovirus outbreak is commonly assumed to follow the SEIR framework where, after infection, individuals enter a latent state, E, that they leave at rate ω on becoming infectious (Simmons et al., 2013). The SDE for the SEIR framework, equivalent to equation (2) from the SIR framework, is given by $\mathrm{d}\mathbf{X} = \mathbf{F}(\mathbf{X})\,\mathrm{d}t + \sqrt{\boldsymbol{V}(\mathbf{X})}\,\mathrm{d}\mathbf{W}$ with

$$\begin{aligned} \mathbf{X}(t) &= \begin{pmatrix} S(t) \\ E(t) \\ I(t) \end{pmatrix} \ , \quad \mathbf{F}(\mathbf{X}) = \begin{pmatrix} -\beta SI/N \\ \beta SI/N - \omega E \\ \omega E - \gamma I \end{pmatrix} \ , \\ \boldsymbol{V}(\mathbf{X}) &= \begin{pmatrix} \beta SI/N & -\beta SI/N & 0 \\ -\beta SI/N & \beta SI/N + \omega E & -\omega E \\ 0 & -\omega E & \omega E + \gamma I \end{pmatrix} \ . \end{aligned} \tag{32}$$

The deterministic approximation of the stochastic SEIR model is given by the ODEs

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\frac{\beta}{N}si \ , \qquad \frac{\mathrm{d}e}{\mathrm{d}t} = \frac{\beta}{N}si - \omega e \qquad \frac{\mathrm{d}i}{\mathrm{d}t} = \omega e - \gamma i \ , \tag{33}$$

where, as before, $s(t)$, $e(t)$, and $i(t)$ are the numbers of susceptible, exposed, and infected individuals respectively at time t given by the deterministic model.

To use the linear SDE Gaussian process approximation with the SEIR model we simply need again to choose the time-varying matrices in (9) so that it approximates (32). For the A noise approximation introduced in this work this will be

$$\begin{aligned} \mathbf{A}(t) &= \begin{pmatrix} 0 & 0 & -\beta s(t)/N \\ 0 & -\omega & \beta s(t)/N \\ 0 & \omega & -\gamma \end{pmatrix} \ , \quad \mathbf{b}(t) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \ , \\ \boldsymbol{U}(t) &= \begin{pmatrix} \beta s(t)i(t)/N & -\beta s(t)i(t)/N & 0 \\ -\beta s(t)i(t)/N & \beta s(t)i(t)/N + \omega e(t) & -\omega e(t) \\ 0 & -\omega e(t) & \omega e(t) + \gamma i(t) \end{pmatrix} \ . \end{aligned} \tag{34}$$

As before, we can infer the unobserved time series $E(t)$ and $I(t)$ using the general conditional laws of Gaussian processes (Rasmussen and Williams, 2005), and use the marginal laws to perform maximum likelihood estimation on the parameter values β, γ, and $S(0)$. Note that we fit $S(0)$ instead of taking it to be $N-1$ because we do not know the infection history of the population. For example, some of the population may have been previously recently exposed to norovirus and therefore not currently susceptible. We make the somewhat conservative assumption that newly diagnosed individuals are no longer susceptible but could potentially be in any of the E, I, or R states so that our data are effectively values of $N - S(t)$ – a different approach could easily be taken within our inferential framework.

Additionally, some care must be taken because ω is poorly identifiable from this cumulative incidence data, and our attempts to fit it alongside the other three parameters produced unrealistically large estimates motivating us to fix this parameter from other data. We found that the literature gives the latent, or incubation, period of norovirus to be between 0.5 and 2 days. For example, an SEIR model fitted to an outbreak in a long-term care facility estimated the latent period of norovirus as 1.3 days (Vanderpas et al., 2009). A systematic review of the incubation period of norovirus genogroups I and II gives it as 1.2 days (95% confidence interval 1.1–1.2) (Lee et al., 2013). The CDC report that the incubation period of norovirus is between 0.5 and 2 days (Centers for Disease Control and Prevention, 2011). Finally, a large dataset of norovirus outbreaks showed the incubation period to have a mean and median of 1.4 (95% confidence interval 1.3–1.4) days. Since ω is the reciprocal of the latent period, we therefore chose $\omega = 2$ days^{-1} as the largest value consistent with existing knowledge about norovirus.

Working at two significant figures or zero decimal places as appropriate, parameter estimates and 95% confidence intervals from fitting the data with the MVN moment closure approximation are: $\beta \approx 21$ $[8.4, 33]$, $\gamma \approx 1.7$ $[0, 3.9]$ days^{-1}, and $S_0 \approx 258$ $[159, 357]$. Parameter estimates and 95%

confidence intervals from fitting the data with the LN approximation are: $\beta \approx 18$ $[8.6, 27]$, $\gamma \approx 1.1$ $[0, 3.1]$ days^{-1}, and $S_0 \approx 237$ $[137, 336]$. Parameter estimates and 95% confidence intervals from fitting the data with the A noise approximation are: $\beta \approx 23$ $[0.81, 44]$, $\gamma \approx 1.5$ $[0, 4.7]$ days^{-1}, and $S_0 \approx 241$ $[125, 357]$. These confidence intervals are truncated at zero for rate parameters. The standard error estimates for the intervals were taken from the leading diagonal of an approximate covariance matrix of the parameter estimates. The approximate covariance matrices were computed as the negative inverse of an approximation to the Hessian of the log-likelihood at the maximum likelihood estimates obtained using finite differences (from the MATLAB function `mlecov()`). The correlation matrices between the parameters for each approach are, respectively,

$$\boldsymbol{R}_{\mathrm{MVN}} = \begin{pmatrix} 1.00 & 0.77 & 0.37 \\ 0.77 & 1.00 & 0.87 \\ 0.37 & 0.87 & 1.00 \end{pmatrix} , \quad \boldsymbol{R}_{\mathrm{LN}} = \begin{pmatrix} 1.00 & 0.79 & 0.57 \\ 0.79 & 1.00 & 0.94 \\ 0.57 & 0.94 & 1.00 \end{pmatrix} , \quad \boldsymbol{R}_{\mathrm{A}} = \begin{pmatrix} 1.00 & 0.91 & 0.71 \\ 0.91 & 1.00 & 0.93 \\ 0.71 & 0.93 & 1.00 \end{pmatrix} . \tag{35}$$

Our overall picture, as would be expected when fitting a complex non-linear stochastic model to limited data, is of highly correlated parameters with relatively large marginal confidence intervals. The average infectious periods (estimated from $1/\gamma$ as 0.59, 0.91, and 0.67 days for the MVN moment closure, LN, and A noise approximations respectively) are shorter than the natural history of norovirus would indicate, which is likely to be due to control measures in place upon the ship (Vivancos et al., 2010) limiting the time period during which cases are able to infect others. Additionally, $S(0)$ is estimated as much smaller than N, which could be due to pre-existing immunity, control measures in place on board the ship, and non-homogeneous mixing (through excursion choice and cabin location) (Vivancos et al., 2010).

Results for learning the time series of $S(t)$, $E(t)$ and $I(t)$ are shown in Figure 6, which shows general agreement on mean behaviour, but differences in the uncertainty.

In the results presented so far, the full dataset has been used to estimate the parameters of the epidemic model, before the time series were estimated as the epidemic progressed. We show here, in Figure 7, the results of also estimating β, γ, and S_0 as the epidemic progressed, beginning from day nine. Note that in this instance, these methods did not work with fewer than nine days of data. These datapoint-by-datapoint estimates remain consistent over the epidemic.

We conclude that even in this common case where there is a small dataset of symptom onset times, our Gaussian process approach can be applied and gives epidemiologically reasonable answers in little computational time.

6 Conclusions

In this paper, we have sought to provide results that will allow Gaussian process approximation to become a more routinely used technique in infectious disease epidemiology. We have considered a wide class of Gaussian process approximations, as well as methods to bound the errors on the use of these. We have compared these using simulated data and applied them to real data. For future work, we would like to address the methodological challenges that face infectious disease epidemiologists and the public health system such as inaccuracy in case ascertainment, the requirement for analysis methods to be online in real-time and robust, and the ability to make predictions going forward under uncertainty.

Conflicts of interest

None

Acknowledgements

EBJ and TH are supported by the Engineering and Physical Sciences Research Council (Grant numbers EP/I01358X/1 and EP/N033701/1). We would like to thank the International Centre for Mathematical Sciences in Edinburgh for hosting us while we developed this work and also the participants in the 'Stochastic models of the spread of disease and information on networks' workshop for helpful comments on it. We would also like to thank the two anonymous referees for their constructive input.

References

H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis.* Springer, Lecture Notes in Statistics, Vol. 151, 2000.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

C. Archambeau, D. Cornford, M. Opper, and J. Shawe-taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1:1–16, 2007.

N. T. J. Bailey. *The mathematical theory of infectious diseases and its applications. 2nd edition.* Griffin, London, 1975.

F. Ball and T. House. Heterogeneous network epidemics: real-time growth, variance and extinction of infection. *Journal of Mathematical Biology*, pages Published online ahead of print, DOI:10.1007/s00285–016–1092–3, 2017.

A. J. Black and A. J. McKane. Stochastic amplification in an epidemic model with seasonal forcing. *Journal of Theoretical Biology*, 267(1):85–94, 2010.

S. Cauchemez and N. M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of The Royal Society Interface*, 5(25):885–897, 2008.

Centers for Disease Control and Prevention. Updated norovirus outbreak management and disease prevention guidelines. *Morbidity and Mortality Weekly Report (MMWR)*, 60(4):1–18, 2011.

N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):397–426, 2013.

C. Dargatz. *Bayesian Inference for Diffusion Processes with Applications in Life Sciences.* PhD thesis, Ludwig-Maximilians-Universität, Munich, 2010.

M. L. Eaton. *Multivariate Statistics: a Vector Space Approach.* John Wiley and Sons, 1983.

P. Fearnhead, V. Giagos, and C. Sherlock. Inference for reaction networks using the linear noise approximation. *Biometrics*, 70(2):457–466, 2014.

W. Feller. Two singular diffusion problems. *Annals of Mathematics*, 54(1):173–182, 1951.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716–1733, 2001.

R. Guy, C. Larédo, and E. Vergu. Approximation of epidemic models by diffusion processes and their statistical inference. *Journal of Mathematical Biology*, 70(3):621–646, 2015.

T. House. For principled model fitting in mathematical biology. *Journal of Mathematical Biology*, 70 (5):1007–1013, 2015.

E. L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443, 2006.

V. Isham. Assessing the variability of stochastic epidemics. *Mathematical biosciences*, 107(2):209–224, 1991.

W. S. Kendall. Stochastic integrals and their expectations. *The Mathematica Journal*, 9(4):757–767, 2005.

A. A. King, M. Domenech de Cellès, F. M. G. Magpantay, and P. Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1806), 2015.

P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer-Verlag Berlin Heidelberg, 1992.

T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971.

J. Leander, T. Lundh, and M. Jirstrand. Stochastic differential equations as a tool to regularize the parameter estimation problem for continuous time dynamical systems given discrete time measurements. *Mathematical Biosciences*, 251:54–62, 2014.

R. M. Lee, J. Lessler, R. A. Lee, K. E. Rudolph, N. G. Reich, T. M. Perl, and D. A. T. Cummings. Incubation periods of viral gastroenteritis: a systematic review. *BMC Infectious Diseases*, 13(1): 446, 2013.

S. C. Leite and R. J. Williams. A constrained Langevin approximation for chemical reaction networks. Preprint available at `http://www.math.ucsd.edu/~williams/biochem/biochem.html`, 2017.

D. J. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

T. J. McKinley, J. V. Ross, R. Deardon, and A. R. Cook. Simulation-based Bayesian inference for epidemic models. *Computational Statistics and Data Analysis*, 71:434–447, 2014.

G. N. Milstein. Approximate integration of stochastic differential equations. *Theory of Probability and its Applications*, 19(3):557–562, 1975.

P. Neal and C. L. Terry Huang. Forward simulation markov chain monte carlo with applications to stochastic epidemic models. *Scandinavian Journal of Statistics*, 42(2):378–396, 2015.

P. D. O'Neill. Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistics in Medicine*, 29(20):2069–2077, 2010.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, Cambridge, Massachusetts, 2005.

J. V. Ross, T. Taimre, and P. K. Pollett. On parameter estimation in population models. *Theoretical Population Biology*, 70(4):498–510, 2006.

K. Simmons, M. Gambhir, J. Leon, and B. Lopman. Duration of immunity to norovirus gastroenteritis. *Emerging Infectious Diseases*, 19:12601267, 2013.

W. Y. Tan and H. Hsu. Some stochastic models of AIDS spread. *Statistics in Medicine*, 8(1):121–136, 1989.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Physical Review*, 36(5): 823–841, 1930.

J. Vanderpas, J. Louis, M. Reynders, G. Mascart, and O. Vandenberg. Mathematical model for the control of nosocomial norovirus. *Journal of Hospital Infection*, 71(3):214–222, 2009.

R. Vivancos, A. Keenan, W. Sopwith, K. Smith, C. Quigley, K. Mutton, E. Dardamissis, G. Nichols, J. Harris, C. Gallimore, L. Verhoef, Q. Syed, and J. Reid. Norovirus outbreak in a cruise ship sailing around the British Isles: Investigation and multi-agency management of an international outbreak. *Journal of Infection*, 60:478–485, 2010.

C. Zimmer, R. Yaesoubi, and T. Cohen. A likelihood approach for real-time calibration of stochastic compartmental epidemic models. *PLOS Computational Biology*, 13(1):1–21, 01 2017.

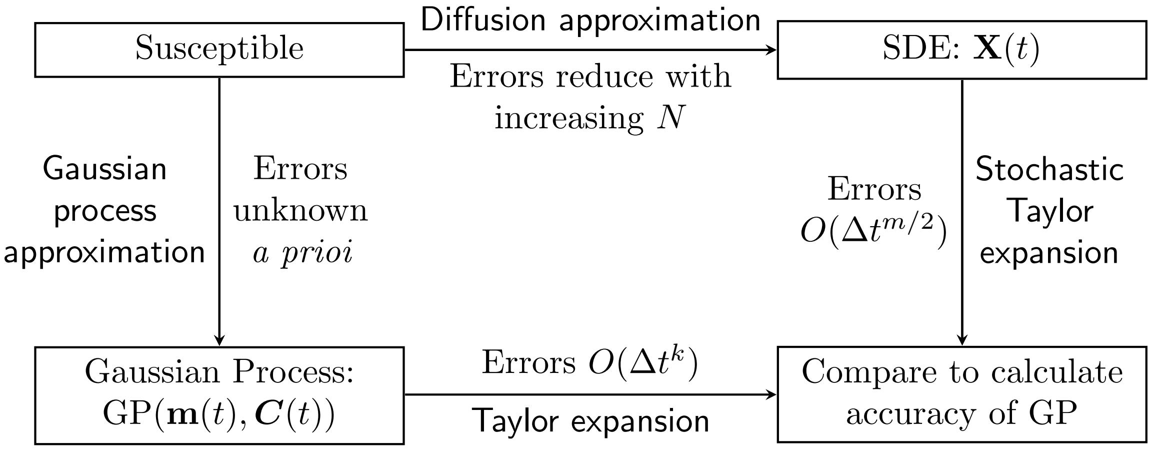

Susceptible
Diffusion approximation
Errors reduce with increasing N
SDE: $\mathbf{X}(t)$
Gaussian process approximation
Errors unknown *a prioi*
Errors $O(\Delta t^{m/2})$
Stochastic Taylor expansion
Gaussian Process: $\mathrm{GP}(\mathbf{m}(t), \boldsymbol{C}(t))$
Errors $O(\Delta t^k)$
Taylor expansion
Compare to calculate accuracy of GP

Figure 1: The overall scheme proposed here for assessment of the accuracy of a given stochastic moment closure approximation. Here we consider primarily SIR epidemics but the approach could apply to other population processes. Errors are controlled by the inverse of N, the population size, and by the time-step Δt, where we use k and m to stand for the integer order of errors in the time-step to be determined by model analysis.

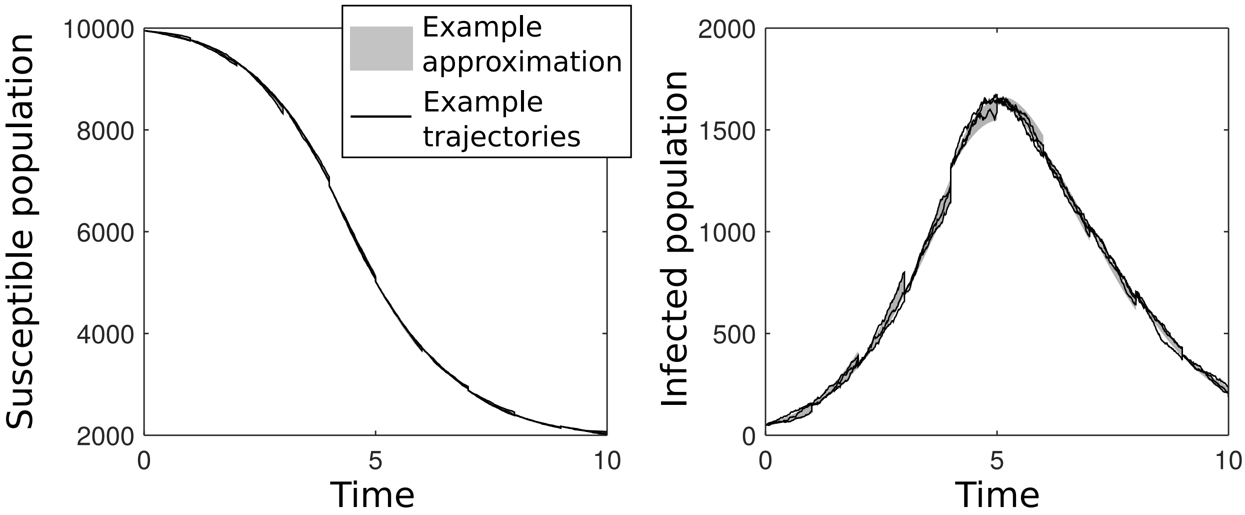
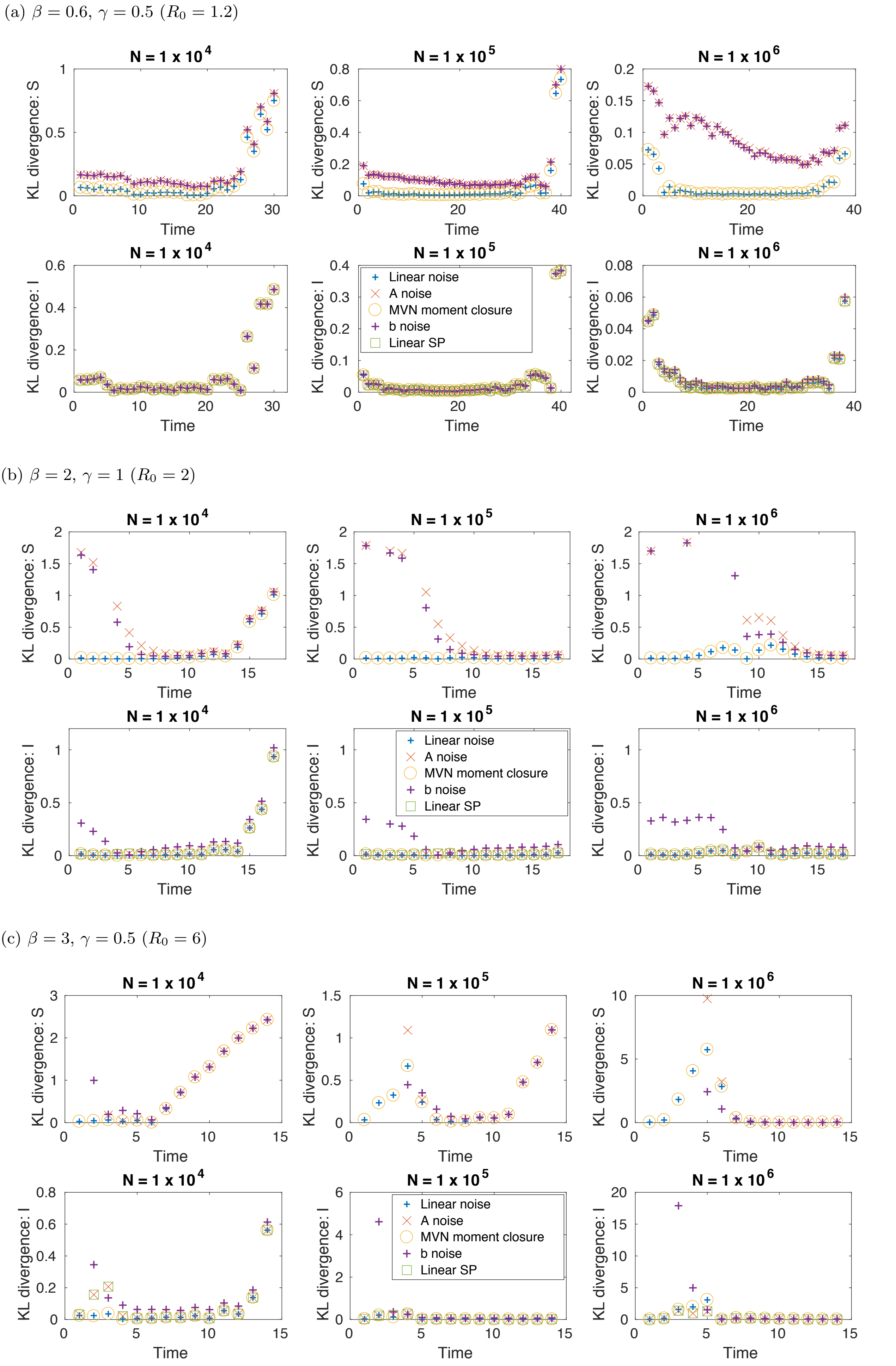


Figure 2: A typical example of stochastic trajectories and one Gaussian process approximation. This example was generated with parameters $\beta = 2$, $\gamma = 1$, and $N = 1 \times 10^4$ and the multivariate normal moment closure approximation. The shaded approximation region corresponds to the mean plus/minus one standard deviation.



(a) $\beta = 0.6$, $\gamma = 0.5$ ($R_0 = 1.2$)

(b) $\beta = 2$, $\gamma = 1$ ($R_0 = 2$)

(c) $\beta = 3$, $\gamma = 0.5$ ($R_0 = 6$)

Figure 3: Numerical comparisons of the approximation schemes with stochastic simulations of the SIR model using the KL divergence. Within each subplot (a-c) different rate constant parameter values were used to generate stochastic simulations for comparison to each Gaussian approximation. The size of the susceptible population is compared on the top line and the size of the infectious population on the bottom line, for three population sizes (increasing from left to right).

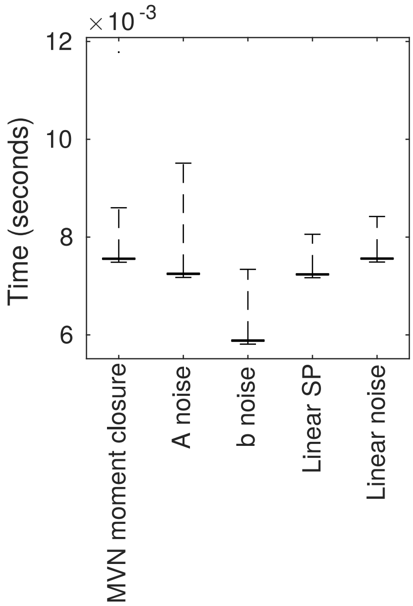


Figure 4: Running time for the sets of ordinary differential equations for each of the approximation methods with $\beta = 2$, $\gamma = 1$, and $N = 1 \times 10^6$. (The box denotes the median, lower quartile, and upper quartile. Whiskers extend to the maximum and minimum. When the interquartile range is narrow, as here, the box displays as simply a thick line.)

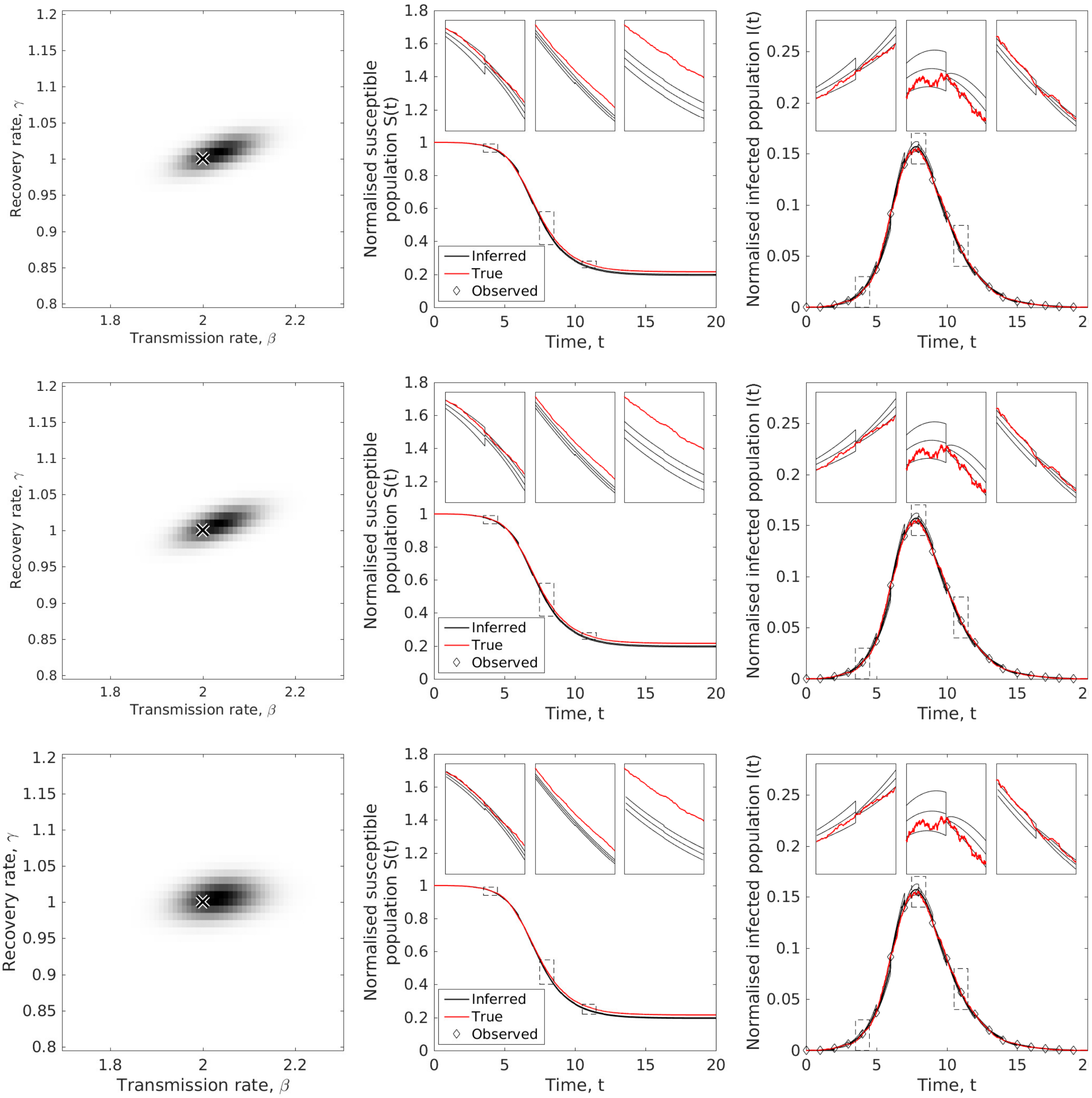


Figure 5: Inference of the susceptible population using the multivariate normal moment closure (top), linear noise (middle), and A noise (bottom) Gaussian process approximations. Left: likelihood (density plot) is concentrated around the true value (cross). Centre: Data on the number of infecteds allows for good reconstruction of the unobserved susceptibility over time. Shown are the synthetic data ('True', red), the mean of the approximation scheme using the inferred parameter values ('Inferred', black), and the mean plus/minus one standard deviation (black) in both the main figures and insets. Dashed rectangles on the main figures show the locations of the insets from left to right. Right: Data ('Observed', diamonds) allow for good reconstruction of the number of infecteds over time. The red lines, black lines, and dashed rectangles are as before.

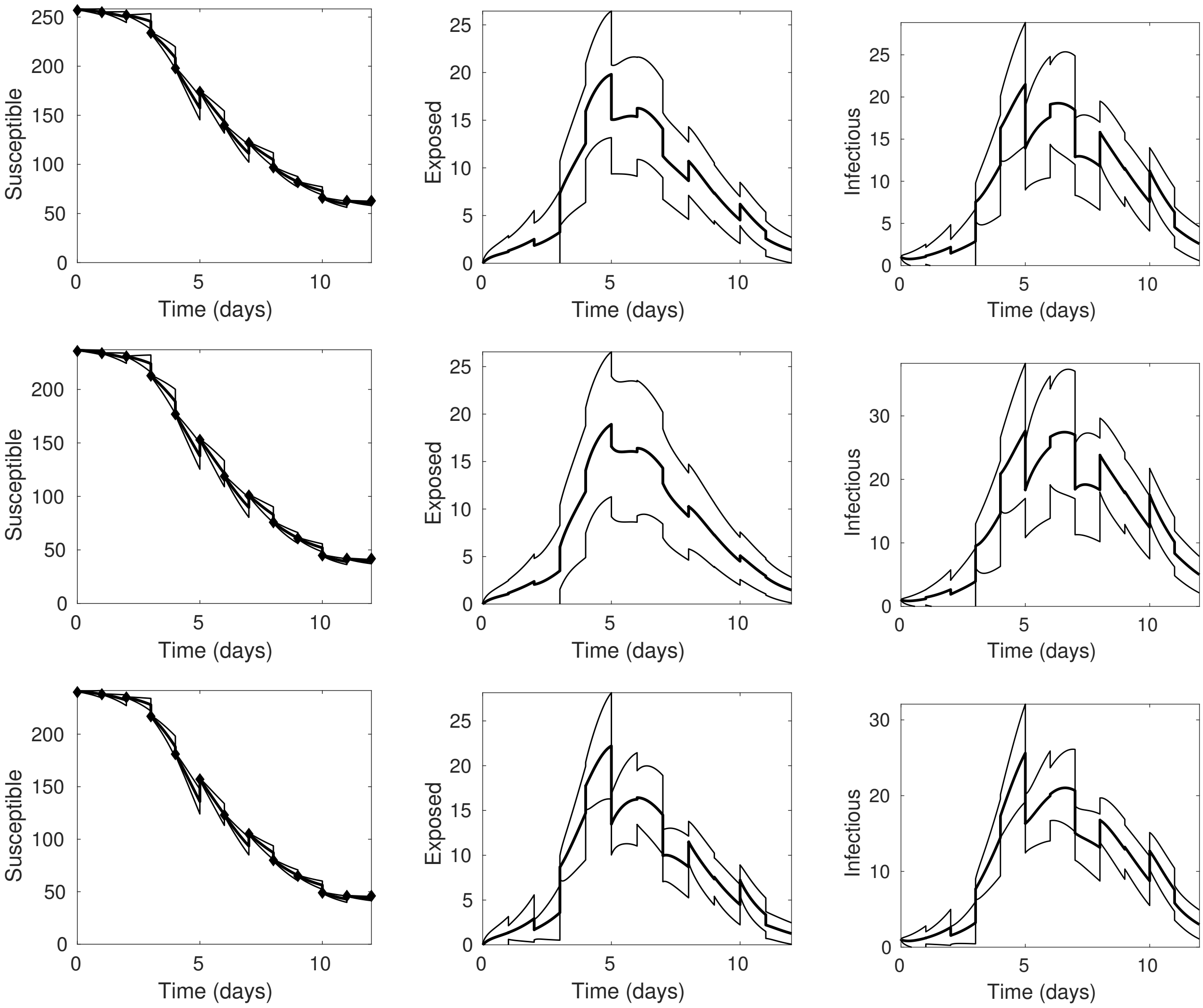


Figure 6: Inference of the susceptible (left), exposed (centre), and infectious (right) population using the multivariate normal moment closure approximation (top), the linear noise approximation (middle), and the A noise approximation (bottom) from data of the number of new cases of norovirus per day on a cruise ship. The black diamonds (left) are our known values which we obtain from the data reported by Vivancos et al. (2010), as described in the text. The dark lines are the mean and light lines are the mean plus/minus one standard deviation.

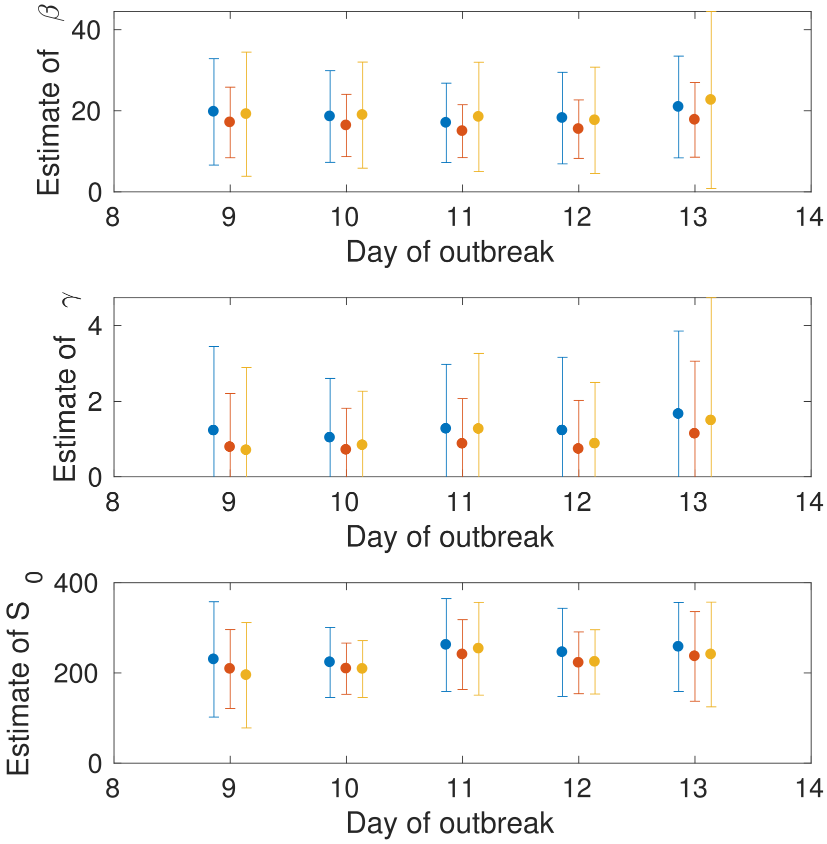


Figure 7: Estimates of the model parameters β, γ, and S_0 as the epidemic progresses for the multivariate normal moment closure approximation (blue), the linear noise approximation (red), and the A noise approximation (yellow). The dots are maximum likelihood estimates and the bars are 95% confidence intervals.