Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bulletin of Mathematical Biology 67 (2005) 855–873
Bulletin of Mathematical Biology
www.elsevier.com/locate/ybulm

Novel moment closure approximations in stochastic epidemics

Isthrinayagy Krishnarajah[a,b,*], Alex Cook[a,b], Glenn Marion[b], Gavin Gibson[a]

[a] *Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom*
[b] *Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom*

Received 11 August 2003; accepted 11 November 2004

Abstract

Moment closure approximations are used to provide analytic approximations to non-linear stochastic population models. They often provide insights into model behaviour and help validate simulation results. However, existing closure schemes typically fail in situations where the population distribution is highly skewed or extinctions occur. In this study we address these problems by introducing novel second- and third-order moment closure approximations which we apply to the stochastic ***SI*** and ***SIS*** epidemic models. In the case of the *SI* model, which has a highly skewed distribution of infection, we develop a second-order approximation based on the **beta-binomial** distribution. In addition, a closure approximation based on mixture distribution is developed in order to capture the behaviour of the stochastic *SIS* model around the threshold between persistence and extinction. This mixture approximation comprises a probability distribution designed to capture the quasi-equilibrium probabilities of the system and a probability mass at 0 which represents the probability of extinction. Two third-order versions of this mixture approximation are considered in which the **log-normal** and the **beta-binomial** are used to model the quasi-equilibrium distribution. Comparison with simulation results shows: (1) the beta-binomial approximation is flexible in shape and matches the skewness predicted by simulation as shown by the stochastic *SI* model and (2) mixture approximations are able to predict transient and extinction behaviour as shown by the

* Corresponding author at: Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom. Tel.: +44 131 650 7536; fax: +44 131 650 4901.
E-mail addresses: isthri@bioss.ac.uk (I. Krishnarajah), alex@bioss.ac.uk (A. Cook), glenn@bioss.ac.uk (G. Marion), gavin@ma.hw.ac.uk (G. Gibson).

doi:10.1016/j.bulm.2004.11.002

stochastic *SIS* model, in marked contrast with existing approaches. We also apply our mixture approximation to approximate a likelihood function and carry out point and interval parameter estimation.

1. Introduction

Stochastic models are useful in epidemiology and ecology, and are used widely (Isham, 1991; Allen and Cormier, 1996; Bolker and Pacala, 1997; Filipe and Gibson, 1998; Marion et al., 1998; Matis and Kiffe, 1999; Bauch and Rand, 2000; Keeling, 2000). Usually, the transition probabilities exhibit non-linear dependence on population size or number of infectives which makes the resultant stochastic processes analytically intractable. Hence, techniques of approximation are needed to capture the underlying behaviour of the stochastic processes. Linearisation is one such approximation, where the behaviour of small stochastic fluctuations can be examined around a fixed point of the deterministic dynamics (Bailey, 1963). An alternative approach is to analyse the quasi-equilibrium probabilities which give a picture of the distribution independent of time and conditional on extinction not having occurred (Renshaw, 1991). Both linearisation and quasi-equilibrium probabilities are limited in their application to regions close to the fixed points or at equilibrium.

In contrast, closure methods are based on equations describing the temporal evolution of moments or cumulants and in principle apply to both transient and equilibrium dynamics. One such widely used closure method is the cumulant truncation procedure (Matis and Kiffe, 1996) where the cumulant functions of say order k are approximated by setting all cumulants of order higher than k to 0. Renshaw (1998) has shown that there is a natural distribution associated with cumulant truncation. This saddlepoint approximation is obtained by applying the method of steepest descents to the truncated cumulant generating function and can be applied in multivariate situations (Renshaw, 2000). In our study, we follow an alternative route using moment closure approximation based on distributional assumptions, a technique introduced by Whittle (1957) which has been widely used in recent years (Isham, 1991; Marion et al., 1998; Keeling, 2000; Nåsell, 2003). Most commonly in these approximations, the population distribution is only described by the first- and second-order moments and description of extinction or bimodality is problematic. Thus, we explore the use of moment closure using **mixture** approximations to population distributions. Many existing methods of second-order approximation also have difficulties in describing highly skewed distribution; hence we consider the application of a novel second-order approximation based on the beta-binomial distribution.

Two generic epidemic models are studied: the stochastic *SIS* (susceptible–infected–susceptible), as an example which exhibits extinction, and the stochastic *SI* (susceptible–infected), a special case of the *SIS*, used as a case where the infected population exhibits a highly skewed distribution but totality of infection is guaranteed. Depending on the disease transmission rate, the *SIS* model exhibits meta-stable persistence of disease, rapid extinction or a critical region corresponding to the border between the two

regions. Both *SI* and *SIS* models have been used before in other studies in either stochastic or deterministic form, for example Jacquez and Simon (1993), Allen and Cormier (1996) and Nåsell (2002). In Section 2, we describe the *SIS* model, present some simulation results and show how a system of moment equations is obtained from the stochastic model. The second-order moment closure approximations are described in Section 3 where various methods of closure are shown. Here, we present the *SI* model, as an ideal starting point for illustration of problems with existing second-order moment closure approximations. In Section 4 we introduce the mixture approximations and compare the results with the simulation results from the stochastic *SIS* model. As an example of an application to inference we apply our mixture approximation to estimate parameter likelihood for the *SIS* model from sparsely sampled data and this is presented in Section 5. Finally in Section 6, we present our conclusions based on the results discussed in Sections 3–5.

2. SIS model

A stochastic *SIS* epidemic model with fixed population size, N, is considered here where the number of infected individuals at time t is denoted by $n(t)$ and the number of susceptibles at time t is $N - n(t)$. Infection and recovery during a small time interval $(t, t + \Delta t)$ are determined by the following probabilities:

$$\text{Prob}[\delta n(t + \Delta t) = 1] = \alpha n(N - n)\Delta t \equiv \psi_\alpha(n)\Delta t \tag{1}$$

$$\text{Prob}[\delta n(t + \Delta t) = -1] = \beta n \Delta t \equiv \psi_\beta(n)\Delta t \tag{2}$$

where Δt is sufficiently small that multiple events which occur with probability $O(\Delta t^2)$ may be ignored. The dependence on time is implicit, through $n(t)$. Here the parameter α is the *contact rate* and β is the individual *recovery rate*. The inter-event time is exponentially distributed with rate $R = \beta n + \alpha n(N - n)$ and the nature of the event will either be an infection with probability $\alpha n(N - n)/R$ or a recovery with probability $\beta n/R$ (Renshaw, 1991). Without loss of generality, we set $\beta = 1$ throughout so that time units are equal to the expected period between infection and recovery (e.g. Filipe and Gibson, 1998).

Fig. 1 uses parameter values representative of three regions, namely the subcritical ($\alpha = 0.06$), which has a mode at $n = 0$ showing rapid extinction; critical ($\alpha = 0.10$), where it is seen that the distribution is bimodal with a probability mass at $n = 0$ and a non-symmetric unimodal contribution to $p(n)$ at $n > 0$; and meta-stable ($\alpha = 0.30$) where the histogram is clustered nearer $n = 20$ meaning that extinction is rare in the finite time considered. However, as $t \to \infty$, ultimate extinction is assured. In all cases, $N = 20$.

2.1. Moment evolution equations

For the model described by Eqs. (1) and (2), let $p_t(n)$ be the conditional probability that there are n infectives at time t given that there are n_0 infectives at time $t = 0$ (Cox and Miller, 1965). Taking the limit as $\Delta t \to 0$, the forward equation obtained is

$$\frac{\mathrm{d}p_t(n)}{\mathrm{d}t} = p_t(n-1)\psi_\alpha(n-1) - p_t(n)\psi_\alpha(n) + p_t(n+1)\psi_\beta(n+1) - p_t(n)\psi_\beta(n). \tag{3}$$

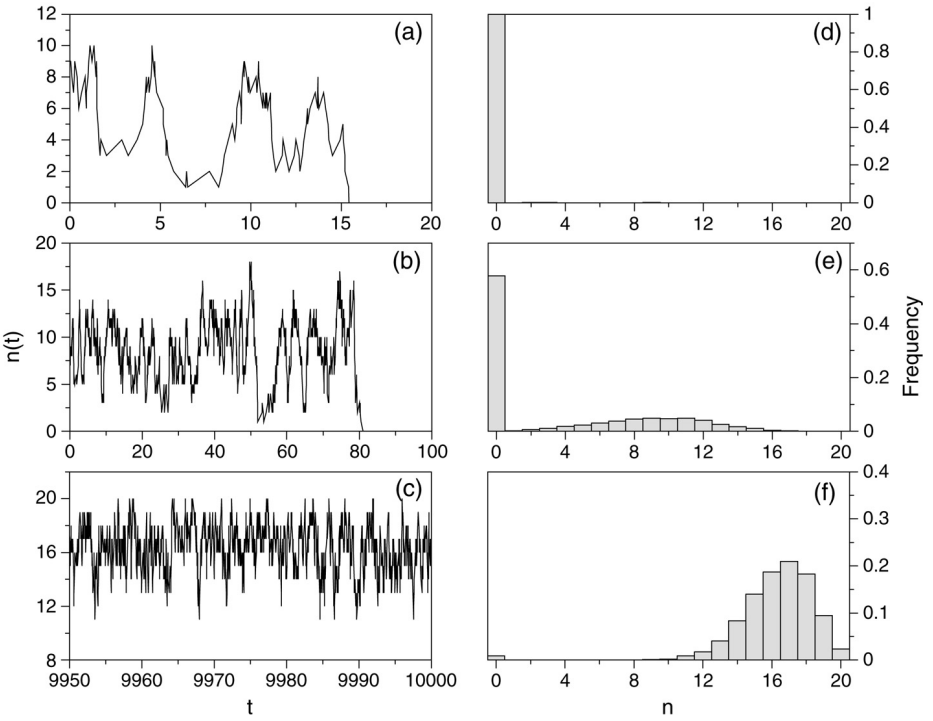

Fig. 1. Single realisations of the stochastic process ((a)–(c)) and histograms representing the number of infectives ((d)–(f)) at $t = 50$ from 10,000 simulations. Graphs (a), (d) represent the subcritical region ($\alpha = 0.06$) where the epidemic quickly dies out (a) and there is a mode at 0 meaning that all realisations have gone extinct (d); graphs (b), (e) represent the critical region ($\alpha = 0.10$) where the epidemic persists for a short period of time before it dies out (b) and some of the realisations persist longer (e); graphs (c), (f) represent the meta-stable region ($\alpha = 0.30$) where the epidemic has reached equilibrium (c) and extinction is rare (f).

where ψ_α and ψ_β are as in Eqs. (1) and (2). The evolution of the epidemic can be described from the evolution of either the raw moments or the cumulants. For example, the first moment or cumulant describes the expected number of infectives or susceptibles. Cumulants have theoretically useful properties and the first four cumulants have clear interpretations. Equivalently, we work with raw moments which are also useful and have functional relations with cumulants.

As shown in Appendix A, the equation describing the rate of change of the kth moment (Goel and Richter-Dyn, 1974) is

$$\frac{\mathrm{d}E[n^k(t)]}{\mathrm{d}t} = \sum_{r=0}^{k-1} \binom{k}{r} (E[n^r \psi_\alpha] + (-1)^{k-r} E[n^r \psi_\beta]). \tag{4}$$

If ψ_α and ψ_β are linear functions of n, from (4) we obtain an equation describing the rate of change of the kth moment depending only on the first k moments. Hence, the set of equations can be solved numerically for any given initial conditions and any finite k. However, this does not hold if ψ_α or ψ_β are non-linear functions of n. For example, for

the *SIS* model used in our study, we can see that ψ_α is non-linear, and we have a set of ordinary differential equations (5)–(7), which are open. The equation describing the rate of change of the kth moment depends on the $(k+1)$th moment, and an infinite set of coupled differential equations is generated. This problem of requiring closure is common to all non-linear stochastic processes.

The ordinary differential equations describing how the first, second and third moments of the stochastic process evolve over time are obtained from (4) by making the substitutions $\psi_\alpha = \alpha n N - \alpha n^2$ and $\psi_\beta = \beta n$ for $k = 1, 2$ and 3:

$$\frac{\mathrm{d}E[n(t)]}{\mathrm{d}t} = (\alpha N - \beta)E[n(t)] - \alpha E[n^2(t)] \tag{5}$$

$$\frac{\mathrm{d}E[n^2(t)]}{\mathrm{d}t} = (\alpha N + \beta)E[n(t)] + (2\alpha N - \alpha - 2\beta)E[n^2(t)] - 2\alpha E[n^3(t)] \tag{6}$$

$$\begin{aligned}\frac{\mathrm{d}E[n^3(t)]}{\mathrm{d}t} &= (\alpha N - \beta)E[n(t)] + (3\alpha N - \alpha + 3\beta)E[n^2(t)] \\ &\quad + (3\alpha N - 3\alpha - 3\beta)E[n^3(t)] - 3\alpha E[n^4(t)].\end{aligned} \tag{7}$$

In order to proceed, the system of differential equations for the first k moments needs to be closed. These differential equations can also be written as cumulant functions and one way of closing the system of equations is to approximate the cumulant functions of order k with cumulants of order higher than k set to zero, a technique known as the *cumulant truncation procedure* (Matis and Kiffe, 1996). However, we employ an alternative approach whereby we assume a particular distribution for the variable of interest. This assumption imposes a functional relationship between the $(k+1)$th moment and the lower order moments. It is this functional relationship that enables us to close the system of moment equations. In either case, the resulting closed system can then be solved numerically. Because of the functional relationship there is no disadvantage in using the raw moments. However, the reader should note that this typically takes higher numerical values than central moments or cumulants. Use of the latter is preferred when working with truncation approaches to closure. In the following section, we discuss problems with existing closure approximations and introduce the beta-binomial approximation, and Section 4 introduces the mixture approximations.

3. Two-parameter approximating families of distributions

Consider second-order moment closure schemes where $E[n^3(t)]$ can be approximated as a function of $E[n(t)]$ and $E[n^2(t)]$ by assuming that n is governed by an appropriate distribution function, the derivation of which is shown in Appendix B for normal, log-normal and beta-binomial approximations. Both the normal and log-normal approximations have been used previously (Whittle, 1957; Isham, 1991; Keeling, 2000). Since the normal distribution has zero skewness and its range has neither an upper nor lower bound, using a normal distribution may not be entirely appropriate in approximating the distribution of the number of infectives in a population of fixed size. This inappropriateness is illustrated in the following subsection, in the case of the *SI* model. The log-normal distribution, because it exhibits skewness and has a non-negative

support, gives a more appropriate description of population variables. The fact that it too does not have an upper bound is ignored, as the primary interest in this study is in the subcritical and critical regions exhibited by the *SIS* model. However, we also consider a novel second-order approximation based on the beta-binomial distribution which does have an upper bound, and it is a counting distribution with fixed population size and support for aggregation of 'successes', i.e. infections.

To illustrate the application of the second-order beta-binomial approximation to a set of observed data (Kleczkowski et al., 1996; Gibson et al., 1999), we present the *SI* model in the following subsection, with comparisons to the results from the existing second-order normal and log-normal approximations. Subsequently in Section 3.2, the results obtained by applying the beta-binomial approximation to the *SIS* model are discussed and compared with results from the log-normal approximation and the simulations.

3.1. An illustration of second-order approximation: the SI model for a fungal plant epidemic

Kleczkowski et al. (1996) carried out experiments on radish seedlings by inoculating them with the pathogen *Rhizoctonia solani* Kühn, a fungus that attacks root vegetables. They monitored 10 microcosms, each containing 50 seedlings, and recorded the number of infected seedlings daily. Five of the microcosms were also exposed to the antagonistic fungus *Trichoderma viride* Pers ex Gray, which is thought to have a controlling effect on *R. solani*.

Gibson et al. (1999) fitted a stochastic model to these data, which accounted for infection by primary sources at rate α_p—that is, the initial inoculum—as well as by the secondary sources that we have considered so far in this paper, representing infection via an already-infected plant, at rate α. As the plants do not recover from the infection, we set $\beta = 0$. The model also includes time-varying susceptibility of the plants, this being e^{-vt}. Infection is then governed by

$$\mathrm{Prob}[\delta n(t+\Delta t)=1] = (\alpha_p + \alpha n(t))(N - n(t))\mathrm{e}^{-vt}\Delta t. \tag{8}$$

The time-varying susceptibility can be easily accommodated by rescaling time as $\tau = (1-\mathrm{e}^{-vt})/v$.

We derive the following equations for the rate of change of the first- and second-order moments of $n(\tau)$ with respect to τ, which are analogous to (5) and (6):

$$\frac{\partial E(n(\tau))}{\partial \tau} = \alpha_p N + (\alpha N - \alpha_p)E(n(\tau)) - \alpha E(n^2(\tau)) \tag{9}$$

$$\frac{\partial E(n^2(\tau))}{\partial \tau} = \alpha_p N + (\alpha N + (2N-1)\alpha_p)E(n(\tau)) + ((2N-1)\alpha - 2\alpha_p)E(n^2(\tau)) - 2\alpha E(n^3(\tau)). \tag{10}$$

By making the assumption that $n(\tau)$ comes from a distribution with two time-varying parameters, we can write $E(n^3(\tau))$ in terms of the first two moments of $n(\tau)$, as described in Appendix B, and substitute this expression in (10). This means that we can use some numerical method to evaluate approximations to the first two moments over time, for a given parameter set $\{\alpha_p, \alpha, v\}$ and the initial condition that no seedlings were infected

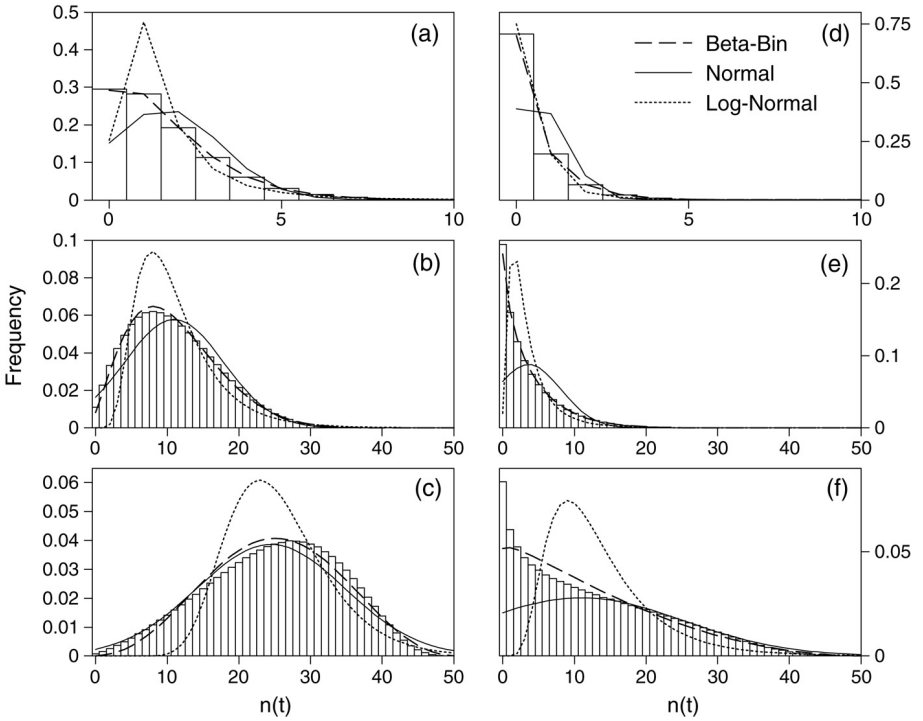

Fig. 2. The distribution of the number of radishes infected by damping-off in the absence ((a)–(c)) and presence ((d)–(f)) of the antagonistic fungus *T. viride*, at 1 ((a), (d)), 5 ((b), (e)) and 15 ((c), (f)) days after the first emergence of seedlings. The initial condition is that the plants are all disease free (i.e. $n(0) = 0$). The histograms represent the average frequencies from a series of 3×10^6 simulations, the curves our approximations. Continuity correction has been used for the continuous distributions; however, we represent all three approximations as continuous curves for clarity of comparison. Both simulations and moment closure approximations make use of the maximum likelihood parameter estimates found by Gibson et al. (1999), these being $(\alpha_p, \alpha, \upsilon) = (0.0265, 0.0118, 0.167)$ in the absence of *T. viride* and $(0.0074, 0.0102, 0.127)$ in its presence.

at the start of the epidemic (i.e. $n(0) = 0$). Using the parameter values estimated by Gibson et al. (1999), we compare these closure approximations to simulated realisations of the model. The simulations and approximations based on normal, log-normal and beta-binomial forms are plotted in Fig. 2.

As can be seen, the beta-binomial approximation captures the dynamics of the evolution of the true probability mass function far better than either the normal or log-normal approximations, its shape being far more flexible.

3.2. Second-order approximation results for the SIS model

Having seen some encouraging results in the case of the *SI* model, we now consider approximating the more complex behaviours of the *SIS* model where we have the subcritical, critical and meta-stable regions.

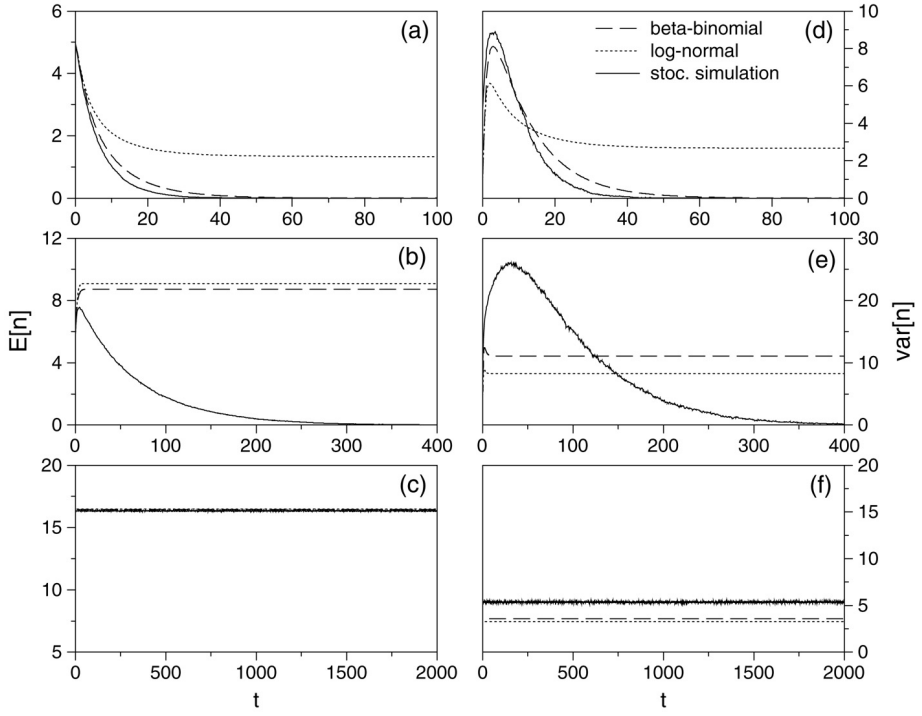

Fig. 3. Second-order approximation and stochastic simulation: expected number of infectives, (a)–(c), and variance, (d)–(f), from closure approximations and stochastic simulations. Subcritical region: (a), (d): $\alpha = 0.06$; critical region: (b), (e): $\alpha = 0.10$; meta-stable region: (c), (f): $\alpha = 0.30$.

As seen in Fig. 3, the beta-binomial approximation gives reasonable estimates in the subcritical but not the critical region and the log-normal one gives poor estimates in both regions. Both approximations perform well in the meta-stable region but predict indefinite persistence. The beta-binomial and log-normal approximations are broadly comparable in the critical and meta-stable regions. Overall, second-order approximations do not give a good description of extinction in the subcritical and critical regions for the *SIS* model. Thus, in the next section a third-order closure approximation is developed in the hope of providing an improved description of these regions.

4. Three-parameter approximating families of distributions: *SIS* model

In order to obtain an improved description of the transient aspects of the stochastic process, a novel closure approximation is developed in which the number of infectives is assumed to be described by a distribution which is a mixture of mass at $n = 0$ and a probability distribution representing extant realisations. This form of mixture distribution is also termed a zero-modified distribution (Johnson et al., 1992) or zero-inflated and used to model count data (Ridout et al., 1998). A major advantage of this mixture

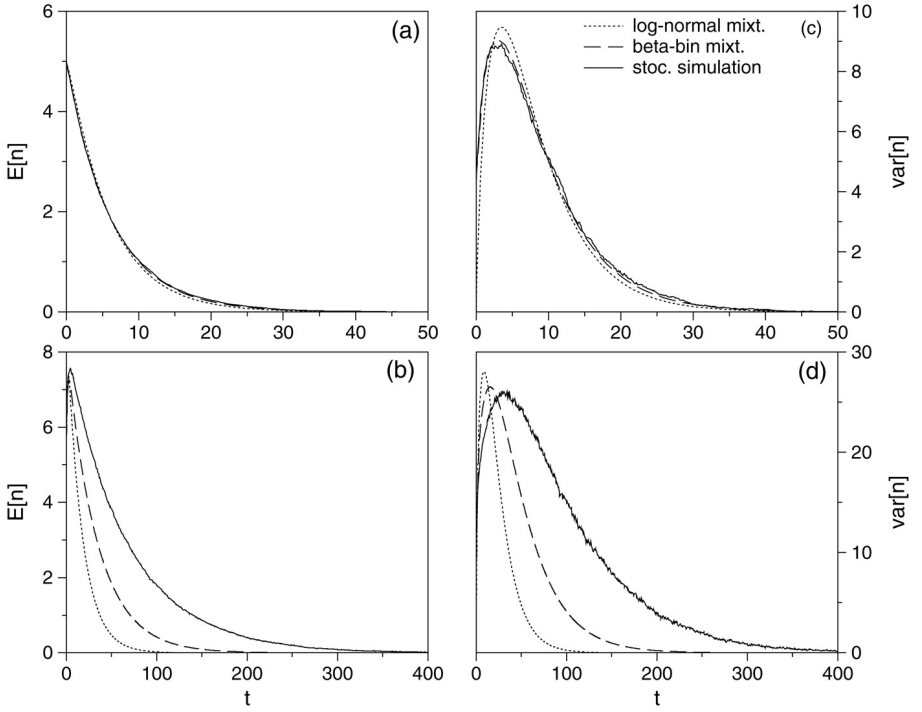

Fig. 4. Mixture approximations and stochastic simulations: expected number of infectives, (a), (b), and variance, (c), (d), from mixture approximations and stochastic simulations. Subcritical region: (a), (c): $\alpha = 0.06$ and critical region: (b), (d): $\alpha = 0.10$.

approximation is that it estimates the probability of extinction in the critical and metastable regions which the second-order approximations shown in Fig. 3 fail to do. Therefore, it allows prediction of the extinction probability, the transient distribution and the quasi-equilibrium distribution. For this study, we use both log-normal and beta-binomial mixture approximations.

In general, the probability function of this mixture distribution is represented by

$$f(n) = p\pi_1(n) + (1-p)\pi_2(n)$$

where $\pi_1(n) = \delta_{n,0}$ (Kronecker delta) and $\pi_2(n)$ is any probability mass function on $0 \leq n < \infty$. Thus, $E[n^k] = (1-p)E_{\pi_2}[n^k]$. If π_2 is from a two-parameter, say (μ, ν), family of distributions, then the mixture defines a third-order approximation since in the generic case p, μ and ν are determined by solving equations for three values of k. Thus, the mixture distribution may be determined by the first-, second- and third-order moments, $E[n]$, $E[n^2]$ and $E[n^3]$. Therefore, $E[n^4(t)]$ in (7) is approximated by a function of $E[n]$, $E[n^2]$ and $E[n^3]$. The form of this approximation is shown in Appendix C for both the log-normal and beta-binomial mixtures.

The results of both third-order approximations for the *SIS* model are compared with stochastic simulation results in Figs. 4 and 5. There is improvement for both mixtures over

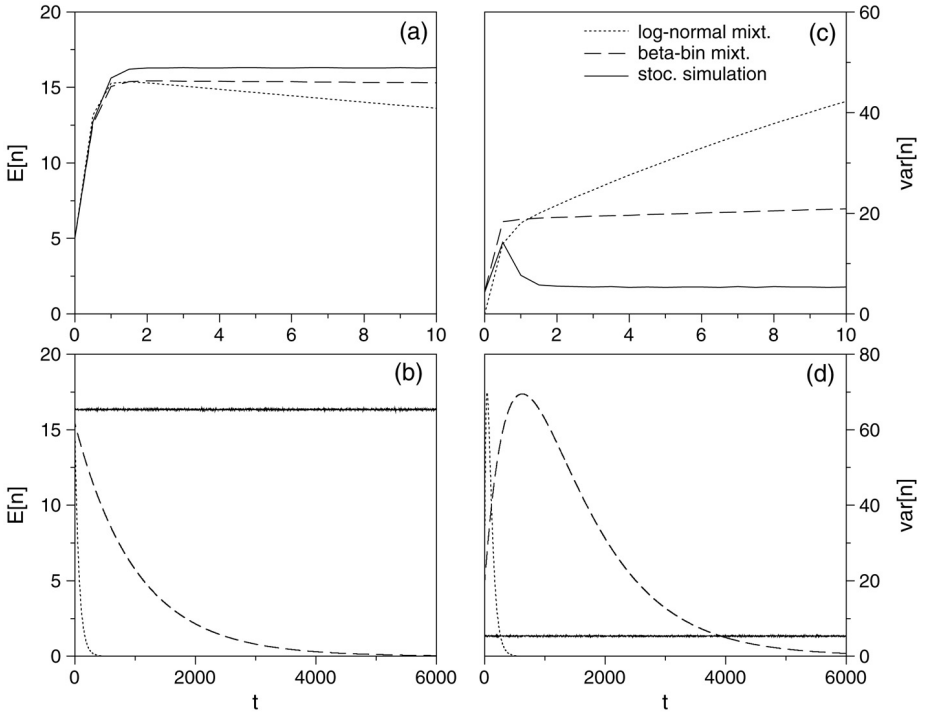

Fig. 5. Mixture approximations and stochastic simulations: expected number of infectives, (a), (b), and variance, (c), (d), from mixture approximations and stochastic simulations in the meta-stable region: $\alpha = 0.30$. Graphs (a), (c) show meta-stable persistence obtained by log-normal mixture and beta-binomial mixture approximations as compared to simulations on a shorter timescale and (b), (d) show long term behaviour predicted by approximations as compared to simulations.

the second-order approximations in the subcritical region where the log-normal mixture is able to predict extinction and the beta-binomial mixture predicts extinction on a more accurate timescale than the corresponding second-order approximation. Furthermore, in this region, the estimated variances also agree with stochastic simulation. In the critical region, there is again a large improvement for both mixtures as they are able to capture the behaviour shown by stochastic simulation, which can be interpreted as short term outbreaks. Unfortunately, both mixtures estimate extinction of the outbreaks on a slightly shorter timescale than that observed in stochastic simulations. In the meta-stable region, the behaviour shown by the mixtures is qualitatively correct but it is unable to mimic the observed meta-stability of the epidemic on a longer timescale as shown in graphs (b), (d) of Fig. 5. Both log-normal and beta-binomial mixtures tend to overestimate the probability of extinction and therefore underestimate the time to extinction. In fact both mixture approximations capture well the qualitative behaviour of the stochastic model but tend to underestimate the time to extinction.

To illustrate this more clearly Fig. 6 shows the phase plot of expected number of infectives versus extinction probability. In both subcritical and critical regions, it can

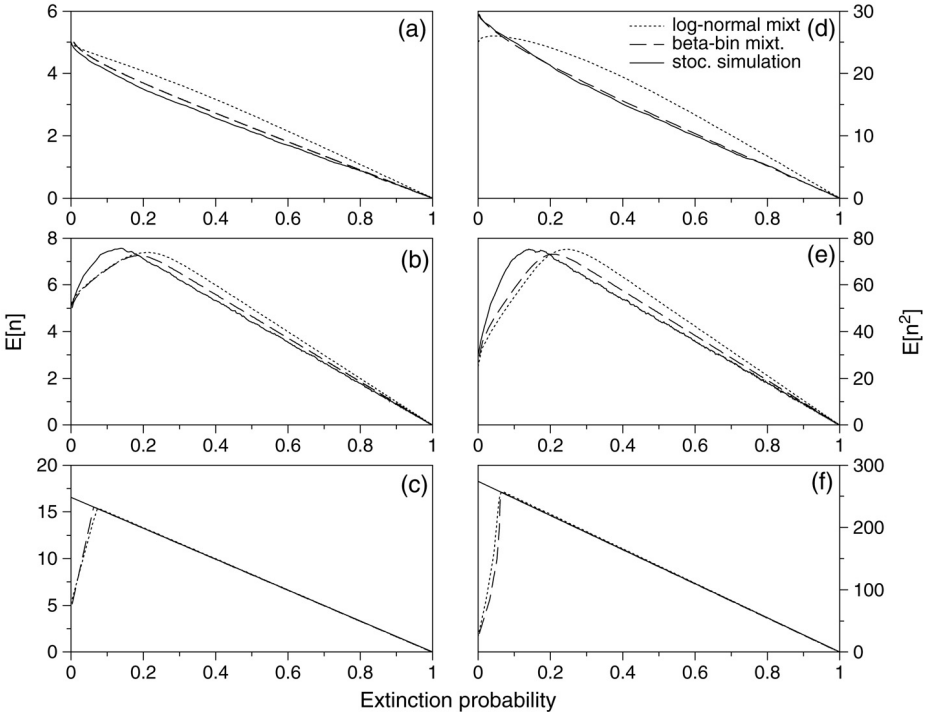

Fig. 6. Diagrams of extinction probability: (a)–(c) show the expected number of infectives versus extinction probability and (d)–(f) show the second moment versus extinction probability, obtained from mixture approximations and stochastic simulations for (a) the subcritical region ($\alpha = 0.06$), (b) the critical region ($\alpha = 0.10$), (c) the meta-stable region ($\alpha = 0.30$).

be seen that once again the beta-binomial mixture is the better approximation. In the meta-stable region, both mixtures are able to match the simulation results for extinction probability >0.05. In graph (c) of Fig. 6, the result from stochastic simulation shows that the expected number of infectives is a linear function of the probability of extinction. This is due to the large rate of infection when $\alpha = 0.30$ which speeds the disease to reach endemic levels that persist over a long time ($t \approx 6 \times 10^7$). For the *SIS* model considered in our study, the quasi-equilibrium distribution is attained relatively quickly in the meta-stable region. Thus the mean conditioned on non-extinction is simply the mean of the quasi-equilibrium distribution, I_0. Therefore for a given probability of extinction p, $E[n] = I_0(1 - p)$. Although the expected time to extinction is poorly estimated by our mixture approximations in the meta-stable region, graph (c) shows that they are able to predict the relationship between probability of extinction and expected epidemic size seen in stochastic simulations. Graphs (d)–(e) of Fig. 6 show that the second moment predicted by beta-binomial mixture approximation is able to track the simulation results for the subcritical and critical regions. Both mixtures are able to match the simulation results in the meta-stable region as seen in graph (f). These results show that the mixture approximations

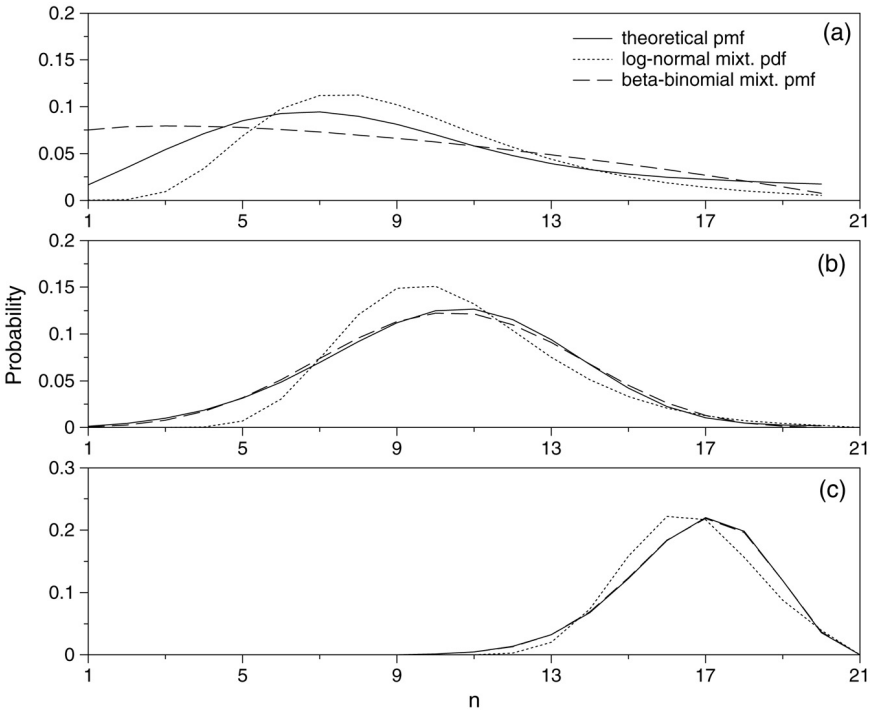

Fig. 7. Comparison of theoretical, beta-binomial mixture and log-normal mixture quasi-equilibrium probabilities: subcritical region: (a): $\alpha = 0.06$; critical region: (b): $\alpha = 0.10$; meta-stable region: (c): $\alpha = 0.30$.

are able to match the main features of the quasi-equilibrium distribution of the stochastic process, at least in terms of the lower order moments. A comparison of the beta-binomial mixture and the log-normal mixture probability functions with the theoretical probabilities from the stochastic *SIS* model is shown in Fig. 7. The probabilities shown are the quasi-equilibrium probabilities meaning that they are conditional on extinction not having occurred (Renshaw, 1991). The beta-binomial mixture distribution is able to capture the dynamics of the theoretical probability mass function of the stochastic model better than the log-normal mixture distribution in both the critical and meta-stable regions. In summary, the mixture distributions are able to match the quasi-equilibrium distribution of the stochastic process.

5. Application to inference

Here we further demonstrate the utility of our mixture approximation by showing how it can be used to construct a reliable approximation to a likelihood function.

Consider a set of data, $\underline{n} = \{n(t_i) : i = 0, 1, \ldots, k\}$, generated from the stochastic *SIS* model with parameter α. The parameter likelihood, $L(\alpha; \underline{n})$, can be calculated using the Markov property as the product of probabilities of transition between the observed states

and can be written as follows:

$$L(\alpha; \underline{n}) = \text{Prob}(\underline{n} \mid \alpha) = \prod_{i=1}^{k} \text{Prob}(n(t_i) \mid n(t_{i-1}), \alpha). \tag{11}$$

For convenience, consider the log-likelihood,

$$\ell(\alpha; \underline{n}) = \sum_{i=1}^{k} \ell(\alpha; n(t_i) \mid n(t_{i-1})) = \sum_{i=1}^{k} \log(\text{Prob}(n(t_i) \mid n(t_{i-1}), \alpha)).$$

The *full log-likelihood* could be calculated if we had the complete data set recording every event. In this case the contribution to the log-likelihood for every event of the data set $\underline{n}$ is calculated using ψ_α and ψ_β as follows:

$$\ell(\alpha; \underline{n}) = \sum_{i=1}^{k} \big(-(\psi_\alpha(i) + \psi_\beta(i))(t_i - t_{i-1}) + \log \psi(i) \tag{12}$$

where $\psi(i)$ is $\psi_\alpha(i)$ for an infection and $\psi_\beta(i)$ for a recovery.

Since it is often the case that we have incomplete data set (e.g. observations recorded at fixed intervals), approximations to the transition probabilities, $\text{Prob}(n(t_i) \mid n(t_{i-1}), \alpha)$, are needed. Thus, in the log-likelihood function, (11), these are approximated from our log-normal mixture distribution and moment equations. Since the approximated function requires fixed initial conditions $(n(t_{i-1}))$, the beta-binomial mixture approximation is not suitable here. The results obtained using the log-normal mixture approximation are compared to full log-likelihood, (12).

For either case approximate confidence intervals for α can be derived on the basis of first-order asymptotic theory (Barndorff-Nielsen and Cox, 1994). If $\widehat{\alpha}$ is the value of α that maximises $l(\alpha)$ and α_0 is the true value of α, then an approximate confidence interval for α can be obtained using the pivotal quantity

$$2\{l(\widehat{\alpha}) - l(\alpha_0)\} \sim \chi_1^2.$$

Hence, we obtain a 95% confidence interval for α by considering the corresponding values of the likelihood that fall within the range of approximately 2 units from the maximum likelihood value.

Maximum likelihood parameter estimates and associated confidence intervals obtained are given in Table 1. The results shown are obtained by applying the mixture approximation to a set of data generated from the stochastic *SIS* model with the three values of α as before. The data points were recorded at equal intervals for all three regions and for the subcritical and critical regions observations ended only once extinction had occurred. In order to calculate the full log-likelihood, every event was recorded.

A plot of the standardised log-likelihood estimates obtained by application of the mixture approximation to the same data is given in Fig. 8. The full log-likelihood is also shown as a comparison to the log-likelihood obtained from the log-normal mixture approximation. It is seen that the log-normal mixture approximation gives a good parameter estimation in both the subcritical and meta-stable regions as seen in graphs (a) and (c). When the frequency of sampling is increased, it shows what might be intuitively

Table 1
Point and interval estimates for α from the full log-likelihood function and mixture approximation

	Subcritical		Critical	Meta-stable
Approximation	MLE	0.058	0.100	0.293
	95% C.I.	(0.045, 0.070)	(0.080, 0.122)	(0.203, 0.423)
Approximation (increased data points)	MLE	0.058	0.089	0.300
	95% C.I.	(0.046, 0.070)	(0.078, 0.100)	(0.265, 0.339)
Full log-likelihood	MLE	0.057	0.103	0.300
	95% C.I.	(0.045, 0.072)	(0.095, 0.112)	(0.286, 0.315)

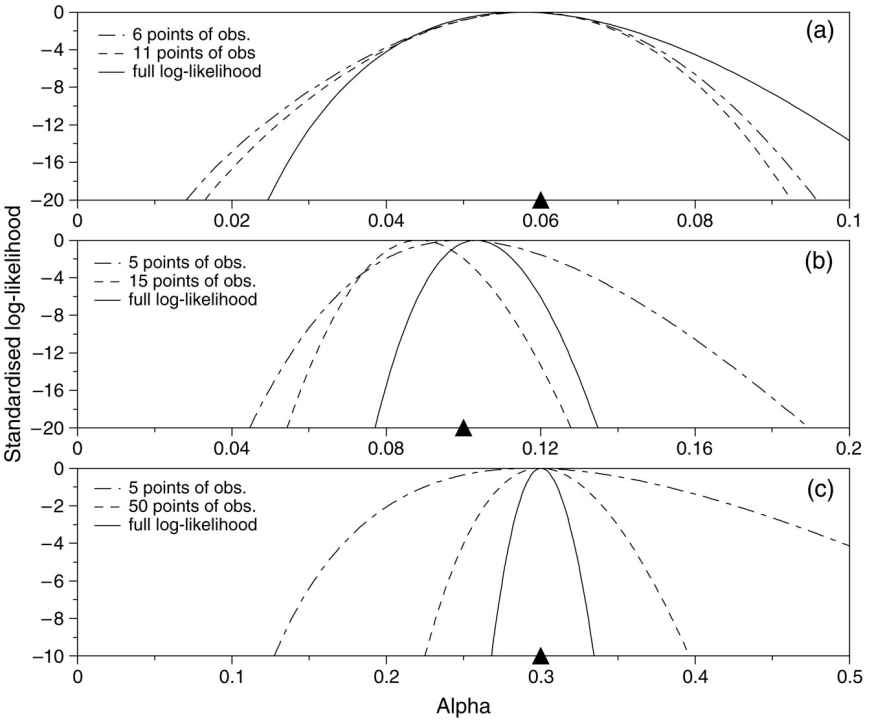

Fig. 8. Plots of the standardised log-likelihood function: comparison of the log-normal mixture approximation with full log-likelihood: (a) subcritical region ($\alpha = 0.06$), (b) critical region ($\alpha = 0.1$), (c) meta-stable region ($\alpha = 0.3$). The true parameter values are indicated in the graphs by triangles.

expected where the estimates from our mixture approximation converge towards the full log-likelihood. From both the point and interval estimates and the log-likelihood function plot, we see that the mixture approximation method agrees closely with the full log-likelihood in these regions. In the critical region, the log-normal mixture gives good estimates when the sample size is small. When the frequency of sampling is increased (in this case when some of the samples taken come after the time when the infection starts

going extinct), the peak, as shown in graph (b), shifts to the left, slightly underestimating the contact rate. To investigate coverage properties of confidence intervals constructed using the full and approximated log-likelihood, we simulated 10,000 realisations for each of the three parameter regions. The coverage properties for 95% confidence intervals based on the full log-likelihood for the subcritical, critical and meta-stable regions are 0.9332 ± 0.0050, 0.9214 ± 0.0054 and 0.9457 ± 0.0045 and for 99% confidence intervals they are 0.9841 ± 0.0025, 0.9785 ± 0.0029 and 0.9893 ± 0.0021. As before, for the approximated log-likelihood, the data points were recorded at equal intervals. The coverage properties for 95% confidence intervals based on the approximated log-likelihood for the subcritical, critical and meta-stable regions are 0.6266 ± 0.0097, 0.6449 ± 0.0096 and 0.7307 ± 0.0089 and for 99% confidence intervals they are 0.7750 ± 0.0084, 0.7291 ± 0.0089 and 0.8497 ± 0.0072. The coverage properties based on the approximated log-likelihood are narrower than the full log-likelihood. These results indicate that the approximated intervals are optimistic (too narrow) but the asymptotic results may not be very appropriate in this situation. Nevertheless, the mixture approximation can be applied to infer model parameters from observed data.

6. Conclusions

In this paper we have introduced a new second-order moment closure approximation and applied this to the *SI* model. The approximation (which assumes that the number of infectives follows a beta-binomial distribution) agrees well with the true frequencies obtained by simulation, and offers a considerable improvement on existing second-order approximations based on the normal or log-normal distributions. The beta-binomial approximation may be similarly applicable in approximating other stochastic processes for fixed-size populations.

In the case of the *SIS* model, which exhibits a richer range of dynamics including extinction and meta-stability, the second-order beta-binomial approximation performs well in the subcritical region (where extinction occurs rapidly), but is unable to predict the extinction occurring in the critical region. In contrast, the log-normal approximation fails to model the observed extinction in both critical and subcritical regions. This led us to propose a family of three-parameter mixtures or zero-inflated distributions combining probability mass at 0 with log-normal or beta-binomial distribution.

These new mixture approximations are able to predict the extinction exhibited by the *SIS* model, although both predict that extinction occurs over a shorter timescale than observed in simulations. The application of moment closure was further extended to estimate parameter likelihoods. This was done by approximating the transition probabilities of a likelihood function using the mixture distribution and moment equations. Parameter estimation based on such approximated likelihoods using data generated from the *SIS* model with known parameter values was seen to be reliable.

There are a number of areas where the work of this paper may be potentially extended. One such example is to apply the mixture approximation to other one-dimensional models such as the Verhulst (Goel and Richter-Dyn, 1974, for example) and Levin metapopulation models (Keeling, 2002, for example). Alternatively, the mixture approximation could be

extended to higher dimensional models, for example, predator–prey systems (Renshaw, 1991), chemical kinetics (Marion et al., 2002) and the *SIR* model (Nåsell, 2002, for example). Finally, it would be interesting to consider moment closure schemes based on more general mixture distributions than those considered in this contribution.

Acknowledgements

The authors are grateful to the two anonymous referees for their valuable suggestions. IK would like to thank Universiti Putra Malaysia and AC wishes to thank Heriot-Watt University and Biomathematics and Statistics Scotland for their financial support. GM gratefully acknowledges the support of the Scottish Executive Environment and Rural Affairs Department.

Appendices

Here we show in detail how the set of moment equations is derived (in Appendix A) and the second-order and third-order approximations obtained, in Appendices B and C respectively.

Appendix A. Moment evolution equation

The forward equation for the stochastic process is as given below:

$$\frac{\mathrm{d}p_t(n)}{\mathrm{d}t} = p_t(n-1)\psi_\alpha(n-1) - p_t(n)\psi_\alpha(n) + p_t(n+1)\psi_\beta(n+1) - p_t(n)\psi_\beta(n). \tag{A.1}$$

where $\psi_\alpha(n)$ and $\psi_\beta(n)$ are the transition probabilities.

By definition, $E[n^k] = \sum_n n^k p(n)$ for $k = 0, 1, 2, \ldots$. Therefore, when (A.1) is multiplied by n and the sum taken over n, we obtain the moment equation, following Goel and Richter-Dyn (1974):

$$\begin{aligned}\frac{\mathrm{d}E[n(t)]}{\mathrm{d}t} &= \sum_{n=1}^{\infty} np_t(n-1)\psi_\alpha(n-1) - \sum_{n=1}^{\infty} np_t(n)\psi_\alpha(n) \\ &\quad + \sum_{n=1}^{\infty} np_t(n+1)\psi_\beta(n+1) - \sum_{n=1}^{\infty} np_t(n)\psi_\beta(n) \\ &= \sum_{n=0}^{\infty}(n+1)\,p_t(n)\psi_\alpha(n) - \sum_{n=0}^{\infty} np_t(n)\psi_\alpha(n) \\ &\quad + \sum_{n=1}^{\infty}(n-1)\,p_t(n)\psi_\beta(n) - \sum_{n=1}^{\infty} np_t(n)\psi_\beta(n).\end{aligned}$$

Similarly, the kth-moment equation is

$$\frac{\mathrm{d}E[n^k(t)]}{\mathrm{d}t} = E[((n+1)^k - n^k)\psi_\alpha] + E[((n-1)^k - n^k)\psi_\beta]. \tag{A.2}$$

Using the fact that $(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r$, (A.2) is equivalent to

$$\begin{aligned}\frac{\mathrm{d}E[n^k(t)]}{\mathrm{d}t} &= \sum_{r=0}^{k-1} \binom{k}{r} E[1^{k-r} n^r \psi_\alpha] + \sum_{r=0}^{k-1} \binom{k}{r} E[(-1)^{k-r} n^r \psi_\beta] \\ &= \sum_{r=0}^{k-1} \binom{k}{r} (E[n^r \psi_\alpha] + (-1)^{k-r} E[n^r \psi_\beta]).\end{aligned}$$

Appendix B. Second-order approximations

The log-normal is a continuous distribution in which the logarithm of the variable of interest is assumed to have a normal distribution. If the number of infectives, n, is log-normally distributed, then $y = \log(n)$ is normal with moment generating function

$$M_y(\theta) = E[\exp(\theta y)] = \exp\left(k_1\theta + \frac{k_2\theta^2}{2}\right)$$

where k_1 is the mean and k_2 the variance of y (Kendall, 1994). It is straightforward to obtain the moments of the log-normally distributed variable n since $E[n^\theta] = E[\exp(\theta y)] = M_y(\theta)$.

Thus, the first, second and third moments for the log-normal distribution can be obtained by substituting $\theta = 1, 2, 3$. For example, the third moment is

$$M_y(3) = E[n^3] = \exp\left(3k_1 + \frac{9k_2}{2}\right).$$

If k_1 and k_2 are determined in terms of $E[n]$ and $E[n^2]$ by solving the equations for the first two moments of the log-normal case simultaneously, then $E[n^3]$ may be expressed as a function of $E[n]$ and $E[n^2]$.

The beta-binomial distribution is a special case of the urn model of Eggenberger and Pólya (1923), although it remained unnamed and under-used until Skellam (1948) gave it a thorough description. The beta-binomial distribution has more recently been used in plant epidemiology by Madden and Hughes (1995) to represent quadrat counts of disease incidence. It is a discrete distribution where the parameter p of a binomial distribution is itself a beta variate (Evans et al., 2000). If the number of infectives, n, is from a beta-binomial distribution, then the moment generating function (Skellam, 1948) is

$$M_n(\theta) = \frac{1}{B(a,b)} \int_0^1 p^{a-1}(1-p)^{b-1}(1-p+p\exp(\theta))^N \,\mathrm{d}p$$

where a and b are the shape parameters and N is the population size. By taking the first, second and third derivatives of the moment generating function and evaluating at $\theta = 0$

we obtain the moments

$$E[n] = \frac{Na}{a+b} \tag{B.1}$$

$$E[n^2] = \frac{Na(Na+N+b)}{(a+b)(a+b+1)} \tag{B.2}$$

$$E[n^3] = \frac{Na}{a+b}\left(1 + \frac{3(N-1)(a+1)}{a+b+1} + \frac{(N-1)(N-2)(a+1)(a+2)}{(a+b+1)(a+b+2)}\right). \tag{B.3}$$

The parameters a and b may be determined in terms of $E[n]$ and $E[n^2]$ by solving (B.1) and (B.2) simultaneously with N fixed by the population size. Then, $E[n^3]$ may be approximated in terms of $E[n]$ and $E[n^2]$.

Finally if $n \sim N(\mu, \sigma^2)$ then its first two moments are $E[n] = \mu$ and $E[n^2] = \mu^2 + \sigma^2$ and its third moment, written in terms of its first two, is $E[n^3] = E[n]^3 + 3E[n](E[n^2] - E[n]^2)$. This can then be substituted into the moment evolution Eq. (10).

Thus the log-normal, beta-binomial and normal distributions may be completely determined by the first- and second-order moments. This is precisely what is required for a second-order approximation. With these assumptions the third-order term, $E[n^3(t)]$ in the equation describing the evolution of the second-order moment, (6), is replaced by appropriate functions of $E[n]$ and $E[n^2]$ for the log-normal and beta-binomial distributions. Similarly, $E[n^3(\tau)]$ in Eq. (10) is written in terms of the first two moments of $n(\tau)$ for the log-normal, beta-binomial and normal distributions.

Appendix C. Mixture approximations

For a third-order approximation, the fourth moment of the log-normal mixture and beta-binomial mixture are needed in order to close the system of differential equations (5)–(7). Thus, when π_2 is log-normal, the fourth moment of the log-normal mixture is

$$E[n^4] = (1-p)\exp(4k_1 + 8k_2)$$

and if π_2 is beta-binomial, the fourth moment of the beta-binomial mixture is

$$E[n^4] = (1-p)\frac{Na}{a+b}\left(\left(1 + \frac{7(N-1)(a+1)}{a+b+1}\right) + \left(\frac{6(N-1)(N-2)(a+1)(a+2)}{(a+b+1)(a+b+2)}\right) + \left(\frac{(N-1)(N-2)(N-3)(a+1)(a+2)(a+3)}{(a+b+1)(a+b+2)(a+b+3)}\right)\right)$$

where p, k_1, k_2, a and b are determined in terms of $E[n]$, $E[n^2]$ and $E[n^3]$ by solving the equations for the first three moments of the corresponding mixture distributions simultaneously and with N fixed by the population size for the beta-binomial mixture.

Therefore, $E[n^4(t)]$ in (7) is approximated by a function of $E[n]$, $E[n^2]$ and $E[n^3]$ for the log-normal mixture and beta-binomial mixture approximations respectively.

References

Allen, L.J.S., Cormier, P.J., 1996. Environmentally driven epizootics. Math. Biosci. 131, 51–80.

Bailey, N.T.J., 1963. The Elements of Stochastic Processes. John Wiley, New York.

Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman and Hall, London.

Bauch, C., Rand, D.A., 2000. A moment closure model for sexually transmitted disease transmission through a concurrent partnership network. Proc. R. Soc. Lond. B 267 (1456), 2019–2027.

Bolker, B., Pacala, S.W., 1997. Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. Theor. Popul. Biol. 52, 179–197.

Cox, D.R., Miller, H.D., 1965. The Theory of Stochastic Processes. Methuen and Co. Ltd., London.

Eggenberger, F., Pólya, G., 1923. Über die Statistik verketteter Vorgänge. Z. Angew. Math. Mech. 1, 279–289.

Evans, M., Hastings, N., Peacock, B., 2000. Statistical Distributions. John Wiley and Sons, Inc., New York.

Filipe, J.A.N., Gibson, G.J., 1998. Studying and approximating spatio-temporal models for epidemic spread and control. Philos. Trans. R. Soc. Lond. B 353 (1378), 2153–2162.

Gibson, G.J., Giligan, C.A., Kleczkowski, A., 1999. Predicting variability in biological control of a plant–pathogen system using stochastic models. Proc. R. Soc. Lond. B 266 (1430), 1743–1753.

Goel, N.S., Richter-Dyn, N., 1974. Stochastic Models in Biology. Academic Press, Inc.

Isham, V., 1991. Assessing the variability of stochastic epidemics. Math. Biosci. 107 (2), 209–224.

Jacquez, J.A., Simon, C.P., 1993. The stochastic SI model with recruitment and deaths I. Comparison with the closed SIS model. Math. Biosci. 117 (1–2), 77–125.

Johnson, N.L., Kotz, S., Kemp, A.W., 1992. Univariate Discrete Distributions. John Wiley and Sons, Inc., New York.

Keeling, M.J., 2000. Metapopulation moments: coupling, stochasticity and persistence. J. Anim. Ecol. 369, 725–736.

Keeling, M.J., 2002. Using individual-based simulations to test the Levins metapopulation paradigm. J. Anim. Ecol. 71, 270–279.

Kendall, M.G., 1994. In: Stuart, A., Ord, J.K. (Eds.), Kendall's Advanced Theory of Statistics. Arnold, London.

Kleczkowski, A., Bailey, D.J., Gilligan, C.A, 1996. Dynamically generated variability in plant–pathogen systems with biological control. Proc. R. Soc. Lond. B 263, 777–783.

Madden, L.V., Hughes, G., 1995. Plant disease incidence: distributions, heterogeneity and temporal analysis. Annu. Rev. Phytopathol. 33, 529–564.

Marion, G., Mao, X.R., Renshaw, E., 2002. Spatial heterogeneity and the stability of reaction states in autocatalysis. Phys. Rev. E 66 (5), 051915(1–9).

Marion, G., Renshaw, E., Gibson, G.J., 1998. Stochastic effects in a model of nematode infection in ruminants. IMA J. Math. Appl. Med. 15 (2), 97–116.

Matis, H.J., Kiffe, T.R., 1996. On approximating the moments of the equilibrium distribution of a stochastic logistic model. Biometrics 52, 980–991.

Matis, H.J., Kiffe, T.R., 1999. Effects of immigration on some stochastic logistic models: a cumulant truncation analysis. Theor. Popul. Biol. 56, 139–161.

Nåsell, I., 2002. Stochastic models of some endemic infections. Math. Biosci. 179, 1–19.

Nåsell, I., 2003. Moment closure and the stochastic logistic model. Theor. Popul. Biol. 63, 159–168.

Renshaw, E., 1991. Modelling Biological Populations in Space and Time. Cambridge University press.

Renshaw, E., 1998. Saddlepoint approximations for stochastic processes with truncated cumulant generating functions. IMA J. Math. Appl. Med. Biol. 15, 41–52.

Renshaw, E., 2000. Applying the saddlepoint approximation to bivariate stochastic processes. Math. Biosci. 168, 57–75.

Ridout, M.S., Demétrio, C.G.B., Hinde, J.P., 1998. Models for count data with many zeros. In: Proceedings of the XIXth International Biometric Conference. Cape Town, Invited Papers. pp. 179–192.

Skellam, J.G., 1948. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. J. R. Stat. Soc. B X (2), 257–261.

Whittle, P., 1957. On the use of the normal approximation in the treatment of stochastic processes. J. R. Stat. Soc. B 19, 268–281.