On parameter estimation in population models

J.V. Ross∗ T. Taimre† P.K. Pollett‡

April 28, 2006

Abstract

We describe methods for estimating the parameters of Markovian population processes in continuous time, thus increasing their utility in modelling real biological systems. A general approach, applicable to any finite-state continuous-time Markovian model, is presented, and this is specialised to a computationally more efficient method applicable to a class of models called density-dependent Markov population processes. We illustrate the versatility of both approaches by estimating the parameters of the stochastic SIS logistic model from simulated data. This model is also fitted to data from a population of Bay checkerspot butterfly (Euphydryas editha bayensis), allowing us to assess the viability of this population.

Keywords: Markov chains; Cross-Entropy method; density dependence; Euphydryas editha bayensis; stochastic SIS logistic model.

1 Introduction

Continuous-time Markov chains have been proposed as theoretical models for an array of biological systems. For example, they have been used to describe the evolution of populations [10], the spread of epidemics [2, 11] and rumours [14], and competition between species [38]. Whilst they are ubiquitous in the applied probability literature, their application has been limited, partly due to a lack of clear statistical procedures for fitting the models. This is most evident in the ecological literature, where discrete-time Markov chains predominate, perhaps due in part to a misconception that a discrete time model is needed if data is collected at discrete time points (say at the end of each breeding season). We address the statistical limitations by providing straightforward methods for parameter estimation, so enabling this rich family of models to be used in modelling real biological systems.

We begin by reviewing the general approach to parameter estimation in the context of continuous-time Markov chains, when the process is observed at successive (but not necessarily equally spaced) time points (for example, sets of abundancy data collected at various times). This should be contrasted with the case where the process

∗Department of Mathematics, University of Queensland, QLD 4072, Australia. jvr@maths.uq.edu.au
†Department of Mathematics, University of Queensland, QLD 4072, Australia. ttaimre@maths.uq.edu.au
‡Department of Mathematics, University of Queensland, QLD 4072, Australia. pkp@maths.uq.edu.au
is observed continuously; in this situation different techniques apply (see for example Section 4 of [17]). We then describe an approach which is simpler in terms of computational implementation, and which is suitable for parameter estimation in a wide class of population models called \((\text{asymptotically})\) density-dependent Markov processes [24].

Density-dependent processes have appeared in a variety of biological and physical contexts: chemical kinetics [26, 37, 21], epidemics [31, 12, 13], metapopulations [7, 8, 39, 40], parasitology [34] and telecommunications [20, 36]. Whatever the context, our approach can be used, but we focus here on ecological models. Our method is based on an approximation to the full likelihood of the process, resulting in a substantial decrease in computational complexity. This is achieved using some remarkable results of Kurtz [24, 25] and Barbour [3, 4, 5, 6], which provide (i) a natural way of identifying an approximating deterministic model that more closely approximates the original model as the size of the system increases, and (ii) a Gaussian diffusion characterizing fluctuations about the unique trajectory of the deterministic model. If the population process starts close to a deterministic equilibrium, this diffusion is an Ornstein-Uhlenbeck (OU) process, and the likelihood function for observations at successive time points has a particularly simple form. It is this likelihood we exploit for our parameter estimation procedure.

We illustrate both techniques by fitting the stochastic SIS logistic model to simulated data. We also fit this model to data from a population of Bay checkerspot butterfly \((Euphydryas editha bayensis)\), and use the fitted model to assess the viability of the population.

The remainder of the paper is organised as follows. In the next section we define Markov population processes. This is followed in Section 3 with an account of a general approach to estimating the parameters of continuous-time Markov chains with bounded transition rates. Section 4 summarises properties of density-dependent models, which are illustrated with reference to the stochastic logistic model. Several properties of density-dependent models are exploited in Section 5, where our computationally simpler method is described. In Section 6 we illustrate both approaches to parameter estimation using the stochastic logistic model, fitting it to both simulated and real data. We also investigate the bias of our estimates and provide error bounds for them in Section 6. A general discussion of our numerical results is given in Section 7, and this is followed by a summary of the paper in Section 8.

2 Markov population processes

Markov population processes have been used extensively to model populations consisting of individuals of different categories or types, or where the individuals occupy different sites. We represent the state at time \(t\) by a vector \(m(t) = (m_1(t), m_2(t), \ldots, m_k(t))\), where \(m_i(t)\) is the number of individuals at the \(i\)-th site at time \(t\), and we assume that \((m(t), t \geq 0)\) is a continuous-time Markov chain whose state space \(S\) is contained in \(\mathbb{Z}^k\) (the \(k\)-dimensional integer lattice). The evolution of such a Markov chain is governed by transition rates \(Q = (q(m, n), m, n \in S)\), with \(q(m, n)\) representing the rate of transition from state \(m\) to state \(n\), for \(n \neq m\), and \(q(m, m) = -q(m)\), where \(q(m) := \sum_{n \neq m} q(m, n)\) \((< \infty)\), being the total rate out of state \(m\). The model is specified by writing down \(Q\).

\(^{1}\)In its common usage in ecology, ‘density dependence’ refers to a population whose per capita rate of change is a function of population size. Its use here is a more literal ‘dependence on the population density’ of a specific form (see Definition 2 in Appendix I).
The jumps of a Markov population process can be of three kinds: the arrival of a new individual (birth or immigration), the departure of an existing individual from the system (death or emigration), or the transfer of an individual from one site to another (migration or predator-prey competition). Accordingly, we use the following definition (Kingman [23]):

**Definition 1** A Markov chain on a subset \( S \) of \( \mathbb{Z}^k \) will be called a Markov population process if, for any \( m \), the only values of \( n \) for which \( q(m, n) \) is non-zero are those with

\[
\begin{align*}
    n_i &= m_i + 1, \quad n_j = m_j \quad (j \neq i) \quad \text{(arrival at } i) \\
    n_i &= m_i - 1, \quad n_j = m_j \quad (j \neq i) \quad \text{(departure from } i) \\
    n_i &= m_i - 1, \quad n_j = m_j + 1, \quad n_k = m_k \quad (k \neq i, j) \quad \text{(transfer from } i \text{ to } j).
\end{align*}
\]

Thus, Markov population processes encompass birth-death processes, predator-prey systems, metapopulations and many other population processes. The new method of parameter estimation we present works best for Markov population processes. We note that the approximation can be used for processes with larger than unit changes in population size, but the method will become less accurate in such situations and for large changes in population size it will be rendered ineffective. Consequently, our method of parameter estimation cannot be used for populations that experience, for example, catastrophes.

We will assume that \( Q \) is regular, so that there is a unique transition function \( P(t) \) with entries \( p_{ij}(t) \) corresponding to the probability that the process moves from state \( i \) to state \( j \) in time \( t \). Regularity is guaranteed if \( Q \) is bounded in the sense that \( q(m) \leq \beta \), for some constant \( \beta \), a condition that is trivially satisfied when there are finitely many states. In this case we may write \( P(t) = \exp(Qt) \), where \( \exp \) is the matrix exponential. (For further details, see [1] or [32].) In most cases the transition function cannot be evaluated explicitly. However, the behaviour of the process can often be investigated by working directly with the transition rates or, for more complex processes, by way of numerical computation procedures, or analytical approximations such as the diffusion approximation adopted here.

### 3 A general approach to parameter estimation

The general approach to parameter estimation, which makes use of a numerical procedure for computing the matrix exponential, is straightforward (see for example [30] for a general discussion), but apparently not widely known in the ecology literature, perhaps due in part to the infeasibility of the method when the population ceiling is large.

We will suppose that there is a parameter (or vector of parameters) \( \theta \), contained in some parameter space \( \Theta \), that must be estimated. We will allow the dependence on \( \theta \) to be made explicit in our notation by writing \( Q(\theta) \) for the transition rates and \( P(\theta; t) = \exp(Q(\theta)t) = (p_{ij}(\theta; t), \ i, j \in S) \) for the transition function. We will also write \( p_i(\theta; t) \) for the probability that the process is in state \( i \) at time \( t \); \( p(\theta; t) = (p_i(\theta; t), \ i \in S) \) is usually taken to be the stationary or the quasi-stationary distribution. Given a set of \( n \) observations \( i_k = m(t_k) \ (k = 1, \ldots, n) \) of the state of the process at times \( (0 \leq) \ t_1 < \cdots < t_n \), the likelihood of observing them is

\[
L(\theta) = p_{i_1}(\theta; t_1) \prod_{k=2}^{n} p_{i_{k-1},i_k}(\theta; t_k - t_{k-1}). \tag{1}
\]
We may then calculate the maximum likelihood estimator (MLE) $\hat{\theta}$ of the parameters $\theta$ by maximising the likelihood (1) over $\Theta$ for the given observations. As noted previously, the transition function cannot usually be evaluated explicitly, but progress can be made by computing the transition probabilities numerically. This, combined with a numerical search algorithm over the parameter space $\Theta$, allows us to compute the desired MLE.

To compute the required matrix exponentials, we use EXPOKIT, a software package whose algorithms make use of Krylov subspace methods [44]. Any one of a range of numerical optimisation techniques can be used to maximise (1). We use the Cross-Entropy Method [41], which has proved to be particularly effective for maximizing the likelihood functions that we consider. We will see that this combination provides a useful tool for fitting continuous-time Markov chains to real systems, provided that the parameter space and the maximum population size is not too large. We implement this approach in Section 6.

4 Density-dependent models

Our second approach, to be outlined in Section 5, is suitable for parameter estimation in a wide class of stochastic population models known as density-dependent models, which we shall describe and illustrate here.

Imagine now that our population model is indexed by $N > 0$, which will usually be a parameter of the model. In any case $N$ will be related to the size of the system. For definiteness we may think of $N$ as being the population ceiling (maximum population size). Thus, we have a family $\{m_N(\cdot)\}$ of Markov chains indexed by $N$, where $m_N(\cdot)$ takes values in $S_N$ (contained in $Z^+$) and has transition rates $Q_N = (q_N(m, n), m, n \in S_N)$. Our model is then called density-dependent [24] if the transition rates take the form

$$q_N(m, m + l) = N f\left(\frac{m}{N}, l\right), \quad l \neq 0,$$

for a suitable function $f$, so that the transition rates of the corresponding “density process” $X_N(\cdot)$, defined by $X_N(t) = m_N(t)/N$, $t \geq 0$, depend on the present state $m$ only through the density $m/N$. A formal and slightly more general definition, where property (2) is realised asymptotically (for large $N$), is given in Appendix I (Definition 2).

To illustrate the idea, we will consider a version of Feller’s [18] stochastic logistic model, which allows one to model the dynamics of a population inhabiting a finite region that can support a maximum of $N$ individuals, and accounts for density dependence (in the ecological sense noted in Section 1) in both the birth and death transition rates. The version we shall consider is the stochastic SIS (Susceptible-Infective-Susceptible) logistic model, which was one of the earliest stochastic models for the spread of infections that do not confer any long lasting immunity, and where individuals become susceptible again after infection [46]. It provides the simplest population model that incorporates ecological density dependence and individual-level demography. It has been applied, not only in epidemiology, but also in the propagation of rumours [9], in chemical reaction kinetics [33], and in metapopulation ecology [27]. It is a continuous-time Markov chain $(m(t), t \geq 0)$ taking values in $S = \{0, 1, \ldots, N\}$ with non-zero transition rates

$$q(m, m + 1) = \lambda m \left(\frac{m}{N} - m\right) \quad (m = 1, 2, \ldots, N - 1)$$

$$q(m, m - 1) = \mu m \quad (m = 1, 2, \ldots, N),$$
where \( \lambda \) and \( \mu \) are both positive. We can see immediately that the model is density dependent with \( f(x,+1) = \lambda x(1 - x) \) and \( f(x,-1) = \mu x \), being the rate of increase and the rate of decrease, respectively, when the population density is \( x \). Furthermore, if we let \( \Delta X = X(t + \Delta t) - X(t) \), then \( E(\Delta X|X(t) = x) \approx F(x)\Delta t \), where \( F(x) = \lambda x(1 - x) - \mu x \). Consequently, a natural deterministic model emerges for the population density:

\[
\frac{dx}{dt} = F(x), \quad \text{where} \quad F(x) = \lambda x(1 - x) - \mu x.
\]

This is of course the classical Verhulst model \cite{45} (which has been rediscovered several times since it first appeared in 1838). Indeed, for any density-dependent model we can identify a deterministic analogue:

\[
\frac{dx}{dt} = F(x), \quad \text{where} \quad F(x) = \sum_l l f(x,l).
\]

Our intuition tells us that \( X_N(\cdot) \) should behave more deterministically as \( N \) becomes large and that indeed the deterministic trajectory should be “tracked” by the process when \( N \) is large. Kurtz \cite{24} established this rigorously by proving a functional law of large numbers (stated in its most general form as Theorem 3 of Appendix I) that holds under mild conditions which will apply in most biological contexts.

What makes this approximation procedure so powerful is our ability to model precisely the fluctuations of the density process \( X_N(\cdot) \) about the deterministic trajectory. We will summarise results for the “equilibrium case”, where we model the fluctuations about a steady state \( x_{eq} \) of (4), that is, \( F(x_{eq}) = 0 \), and consider the family \( \{Z_N(\cdot)\} \) defined by \( Z_N(t) = \sqrt{N}(X_N(t) - x_{eq}) \). Our intuition, based on the Central Limit Theorem, tells us that, for large \( N \), \( Z_N(t) \) should have an approximate normal (Gaussian) distribution (with zero mean and a variance that depends on \( t \)). This is indeed the case, but we can be far more precise by identifying an approximating process. Kurtz \cite{24} proved that our scaled density process \( Z_N(\cdot) \) converges to an Ornstein-Uhlenbeck (OU) process \( Z(\cdot) \) whose parameters can be determined simply from the original model. This is given in Appendix I with a formal statement of a functional central limit theorem (Corollary 5).

Much can be deduced from properties of the limiting process \( Z(\cdot) \). For example, in the one-dimensional case, we may conclude that, for \( N \) large, \( X_N(t) \) has an approximate normal distribution with

\[
E(X_N(t)) \simeq x_{eq} + e^{Bt}(X_N(0) - x_{eq}),
\]

where \( B = F'(x_{eq}) \) and \( \text{Var}(X_N(t)) \simeq \sigma_t^2 / N \), with

\[
\sigma_t^2 = \frac{G(x_{eq})}{2B} (e^{2Bt} - 1), \quad \text{where} \quad G(x) = \sum_l l^2 f(x,l).
\]

We will consider only the case where \( x_{eq} \) is asymptotically stable, that is \( B < 0 \), so that, as \( t \to \infty \), \( E(Z(t)) \to 0 \) and \( \text{Var}(Z(t)) \to \sigma^2 := V/(-2B) \), where \( V = G(x_{eq}) \). Again, full details are given in Appendix I. The special covariance structure of the OU process will be revealed and exploited in the next section, where our computationally simpler parameter estimation method is described.

To illustrate the diffusion approximation technique, consider again the logistic model, but in the most interesting case \( \lambda > \mu \), where there is drift away from the extinction
state 0. Under this assumption, (3) has two equilibria, both in [0, 1]: 0 (unstable) and \( x_{st} = 1 - \rho \), where \( \rho = \mu/\lambda \) (asymptotically stable). We note, in the context of epidemiology, \( \rho \) is the equilibrium fraction of susceptible individuals and is also the inverse of the basic reproduction ratio \( R_0 \) [15]. Furthermore, (3) has the unique solution

\[
x(t) = \frac{x_{st}x_0}{x_0 + (x_{st} - x_0)e^{-\lambda x_{st}t}}, \quad t \geq 0,
\]

where \( x_0 = x(0) \) is the initial point of the trajectory.

It is possible to write down the full diffusion approximation for the population density \( X_N(t) = m(t)/N \) about the deterministic path (5) by way of Theorem 4 of Appendix I, but we will not pursue this here. Rather, we will consider the OU approximation about the asymptotically stable equilibrium \( x_{st} \). Since \( F'(x) = \lambda(x_{st} - 2x) \), we have (local drift) \( B = F'(x_{st}) = -\lambda x_{st} \) and, since \( G(x) = F(x) + 2\mu x \), we have (local variance) \( V = G(x_{st}) = 2\mu x_{st} = 2\lambda x_{st} \rho \). We conclude that, for \( N \) large, \( X_N(t) \) has an approximate normal distribution with

\[
E(X_N(t)) \simeq x_{st} + e^{-\alpha t}(X_N(0) - x_{st}),
\]

\[
\text{Var}(X_N(t)) \simeq \rho(1 - e^{-2\alpha t})/N,
\]

where \( \alpha = \lambda x_{st} = \lambda - \mu = \lambda(1 - \rho) \) (\( > 0 \)). Hence, for all \( t \geq 0 \),

\[
E(m(t)) \simeq N(1 - \rho) + e^{-\alpha t} (m(0) - N(1 - \rho)),
\]

\[
\text{Var}(m(t)) \simeq N\rho(1 - e^{-2\alpha t}).
\]

5 A computationally simpler approach

Our parameter estimation method uses the (stable) one-dimensional OU approximation outlined in the previous section. Recall that we are given a set of \( n \) observations \( (m(t_1), \ldots, m(t_n)) \) of the state of the process at times \( (0 \leq) t_1 < \cdots < t_n \). If we denote the corresponding observations of the density process by \( (X_N(t_1), \ldots, X_N(t_n)) \), then, when \( N \) is large, this vector will have an approximate normal distribution, with a covariance structure that can be determined explicitly from properties of the limiting OU process.

If we start the OU process in equilibrium, that is, \( Z(0) \sim \text{Normal}(0, \sigma^2) \), where \( \sigma^2 = V/(-2B) \), or equivalently (for large \( N \)) \( X_N(0) \sim \text{Normal}(x_{eq}, \sigma^2/N) \), then it will be strongly stationary. Therefore, since \( \text{Cov}(Z(s), Z(s + t)) = \sigma^2 \exp(B|t|) \), we may approximate \( \text{Cov}(X_N(s), X_N(s + t)) \) by

\[
c(t) := c(0) \exp(B|t|)
\]

where \( c(0) = \sigma^2/N \).

Hence, for large \( N \), we know explicitly the correlation structure of the vector \( (X_N(t_1), \ldots, X_N(t_n)) \), and hence its likelihood function:

\[
f(x) = \frac{1}{\sqrt{(2\pi)^n|C|}} \exp \left[ -\frac{1}{2}(x - m)C^{-1}(x - m)^t \right],
\]

where \( m = (m_1, m_2, \ldots, m_n) \), \( m_i = x_{eq} \) for all \( i = 1, 2, \ldots, n \), and

\[
C = \begin{pmatrix}
c_1 & c_{1,2} & c_{1,3} & \cdots & c_{1,n} \\
c_{1,2} & c_2 & c_{2,3} & \cdots & c_{2,n} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
c_{1,n} & c_{2,n} & c_{3,n} & \cdots & c_n
\end{pmatrix},
\]
where \( c_i = \sigma^2/N \) and \( c_{i,i+s} = (\sigma^2/N) \exp(B(t_{i+s} - t_i)) \). The inversion of \( C \), and calculation of its determinant, can be done explicitly, and this permits us to write down the (log-)likelihood explicitly (see Appendix II). The form that should be used in practice is given in (12) (or (13)).

We can therefore evaluate the (joint) MLEs for the parameters of the model, which are the values that maximise (7). Explicit calculation of the MLEs is not practical in most situations. Therefore, a numerical optimisation procedure will be required in most cases to find the parameters which maximise the likelihood function. The Cross-Entropy (CE) method (see [41]) is an ideal recent approach, but any one of several numerical optimisation procedures should be effective. We illustrate this approach in Section 6 by using the CE method to estimate the parameters of the stochastic logistic model using both simulated and real data.

Remarks. 1. In many situations the maximum population size \( N \) is unknown. It can be estimated using the approximation \( NX_N(0) \sim \text{Normal}(Nx_{eq}, N\sigma^2) \). The joint MLEs are then calculated with \( N \) as an additional unknown parameter.

2. The OU approximation is achieved by letting the maximum population size tend to infinity. Thus, the OU approximation, and hence the parameter estimation procedure itself, are best for large population sizes. This is precisely the situation in which the first method becomes infeasible.

3. We emphasise that unequally spaced sampling of the process is no obstacle to the success of the method, as can be seen from the covariance structure (6).

4. We can obtain simple estimates of relationships between parameters, without resorting to numerical optimisation of the full likelihood, by using the method of moments. Under the assumptions used to obtain the OU approximation, we have

\[
E(\bar{m}) = Nx_{eq}, \quad \text{where} \quad \bar{m} = \frac{1}{n} \sum_{i=1}^{n} m(t_i),
\]

and

\[
E\left(\frac{1}{n} \sum_{i=1}^{n} (m(t_i) - \bar{m})^2\right) = N\sigma^2.
\]

Thus, the mean and variance of our data set provides some indication about the relationships between the parameters. This provides a quick guide to expected parameter values, and for some models will provide quantities of interest directly. However, in most cases, the numerical maximisation procedure outlined above will be required.

Finally, we note that any approach to estimating the parameters of an OU process may be implemented. For example, one may use the theory of estimation for continuous-time autoregressive processes [16]. Additionally, in the approach presented above, we note that it is often possible to determine the equation for the optimal value of one parameter, conditioned upon knowing the other parameters. In such cases we may use this equation to reduce the dimensionality of our search by one, thus increasing the computational efficiency of the method.

6 The stochastic logistic model

In this section we illustrate the methods presented in Sections 3 and 5, by estimating the parameters of the stochastic logistic model from both simulated data and data for a
population of Bay checkerspot butterfly (*Euphydryas editha bayensis*) [22]. We consider situations in which the maximum population size is known and unknown, where it is ‘small’ and where it is ‘large’, and situations where we have a reasonable sample of data, and others where we have minimal data.

The general approach. For this model, the parameter vector $\theta$ is $(\lambda, \mu)$, or $(\lambda, \mu, N)$, depending upon whether $N$ is known or unknown, respectively.

The computationally simpler approach. We assume that $X_N(0)$ is normally distributed with mean $1 - \mu/\lambda$ and variance $\mu/(N\lambda)$, so that the OU process is strongly stationary and, in particular,

$$\text{Cov}(X_N(s), X_N(s + t)) \simeq c(t) = \frac{\mu}{N\lambda} \exp(- (\lambda - \mu)|t|) .$$

Furthermore, the likelihood $(X_N(t_1), X_N(t_2), \ldots, X_N(t_n))$ has the approximate Gaussian density (7) whose parameters have entries $m_i = 1 - \mu/\lambda$, $c_i = \mu/(N\lambda)$ and $c_{i,i+s} = \mu/(N\lambda) \exp(- (\lambda - \mu)(t_{i+s} - t_i))$ ($i = 1, 2, \ldots, n$).

Experience shows that in some situations it is better, from a numerical perspective, to use the log-likelihood $l(x)$, that is

$$l(x) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(|C|) - \frac{1}{2} (x - m) C^{-1} (x - m)' .$$

(Again, the form that should be used in practice is given in (13).)

For the stochastic logistic model, the expected values of the mean and variance of our data set are (see (8) and (9))

$$E(\bar{m}) = N(1 - \rho) \quad \text{and} \quad E\left(\frac{1}{n} \sum_{i=1}^{n} (m(t_i) - \bar{m})^2\right) = N\rho .$$

Hence, if $N$ is known, we have a quick estimate for $\rho$. If it is not, then we have a quick estimate for $N$ and hence $\rho$, since

$$N = E\left(\bar{m} + \frac{1}{n} \sum_{i=1}^{n} (m(t_i) - \bar{m})^2\right) .$$

6.1 Simulated data

In all cases discussed below, our simulated data sets were generated in the following way. We simulated the stochastic logistic model with chosen parameters (as listed in each section) to generate data over 119 units of time (years). We then observed the population size at yearly intervals. We took as our data set the following $n$ observations, after discarding the first 20 (corresponding to 19 years of data), to allow the influence of the initial conditions to dissipate.

We emphasise that all of the tables in this section correspond to estimates for a single data set (which is likely to be the case in a practical scenario), for 20 runs of the appropriate CE algorithm (see Appendix III). As the CE algorithm is a randomised

\(^2\)The dissipation of transient effects was rapid, as the simulations were started in equilibrium. The simulations were empirically observed to remain in quasistationarity over the years that were used.
algorithm, the tables in this section indicate its consistency, and suggest that the likelihood function in this case is well-behaved. The performance of our parameter estimation technique over independent simulations of the SIS model is examined in Section 6.3.

In the tables presented, ‘It’ refers to the number of iterations taken by the CE algorithm and ‘S’ represents the value of the (log-)likelihood. The ‘min’, ‘median’, ‘mean’ and ‘max’ in the tables refer to the value of the (log-)likelihood S (so ‘max’ corresponds to the most likely parameter estimates found).

**The general approach.** The parameters used were \( N = 50 \), \( \lambda = 0.8 \) and \( \mu = 0.4 \), and \( n = 40 \). We used a CE algorithm, with normal updating [41] to maximise the likelihood given by (1). We considered the (log-)likelihood as a function of the “natural” parameters, \( N \), \( \rho = \mu / \lambda \) and \( \alpha = \lambda - \mu \). In the case where \( N \) was known, we used the log-likelihood, and in the case where it was unknown, we used the likelihood itself. We used EXPokit to evaluate (1) with initial distribution taken to be the quasi-stationary distribution, and began the CE algorithm with 1,000 initial parameter estimates for \( (\lambda, \mu) \), drawn uniformly from the triangle \((0, 0) - (0, 5) - (5, 5)\). Note that, for the example considered, a wider range of initial parameters resulted in numerical precision problems when using EXPokit (for Matlab) to evaluate the likelihood. Also, for each parameter estimate, a transition rate matrix had to be created and passed to EXPokit for evaluation. In our case, we could take advantage of the sparse structure of \( Q \), which EXPokit could exploit. Even so, experience showed that when \( N \) became quite large, the evaluation of the likelihood at a parameter estimate could become too costly for our algorithm to complete in reasonable time (on a 3GHz Pentium 4 with 1Gb of RAM). Hence, it would appear that the general procedure is currently feasible only when \( N \) is moderate.

The CE parameters were updated on the basis of the best 25 parameter estimates (meaning that the best 25 parameter estimates of the 1,000 were used to obtain a new sampling distribution for the next iteration). The algorithm was stopped when all of the standard deviations were below \( \varepsilon = 10^{-7} \), or when the algorithm hit the maximum iterations ceiling of 250. Note that we did allow “illegal” parameter values, such as negative \( N \), \( \rho > 1 \) or \( \alpha < 0 \). Further, we did not attempt to evaluate the likelihood in these situations; rather, we applied a penalty, being quadratic in the parameters. This guided the sampling distribution of the CE algorithm into the region of allowable parameters if it strayed. Finally, if the optimisation procedure terminated due to a numerical precision error, the procedure was rerun.

**\( N \) known.** We obtained the following results:

<table>
<thead>
<tr>
<th>( n = 40 )</th>
<th>It</th>
<th>( S )</th>
<th>( \hat{\lambda} )</th>
<th>( \hat{\mu} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>8</td>
<td>-109.108</td>
<td>1.14739</td>
<td>0.566295</td>
</tr>
<tr>
<td>Mean</td>
<td>8.65</td>
<td>-109.108</td>
<td>1.14739</td>
<td>0.566295</td>
</tr>
<tr>
<td>Median</td>
<td>9</td>
<td>-109.108</td>
<td>1.14739</td>
<td>0.566295</td>
</tr>
<tr>
<td>Max</td>
<td>9</td>
<td>-109.108</td>
<td>1.14739</td>
<td>0.566294</td>
</tr>
</tbody>
</table>

**\( N \) unknown.** When \( N \) was unknown, we introduced an additional sampling distribution for \( N \). Using the same data set as above, we sampled 1,000 initial population ceilings uniformly from the maximum observed population size (which in this case was 32) and 250. (Note that this ceiling should be large enough to encompass the true population size.) All of the other parameters remained the same.
The computationally simpler approach. The CE parameters used here were a sample size of 50,000 and the CE parameters were updated on the basis of the best 100 parameter estimates. We initialised the algorithm with pairs \((\lambda, \mu)\) drawn uniformly from the triangle \((0, 0) - (1000, 0) - (1000, 1000)\). In the case that \(N\) was unknown, we drew \(N\) uniformly between the highest value found in the sample and 10,000. The algorithm was stopped when the largest of the standard deviations of the sampling distributions dropped below the threshold \(\varepsilon = 10^{-7}\). The larger search space and number of samples per iteration in the CE algorithm was now practical owing to the low cost of evaluating the (log-)likelihood for a trial set of parameters.

The following tables show results for data simulated using parameters \((N, \lambda, \mu) = (50, 0.8, 0.4)\).

### N known

<table>
<thead>
<tr>
<th>(n = 40)</th>
<th>It</th>
<th>(S)</th>
<th>(\hat{N})</th>
<th>(\hat{\lambda})</th>
<th>(\hat{\mu})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>14</td>
<td>-107.29</td>
<td>40</td>
<td>1.63349</td>
<td>0.605665</td>
</tr>
<tr>
<td>Mean</td>
<td>15</td>
<td>-107.249</td>
<td>39.15</td>
<td>1.67561</td>
<td>0.598932</td>
</tr>
<tr>
<td>Median</td>
<td>17.5</td>
<td>-107.242</td>
<td>39</td>
<td>1.68304</td>
<td>0.597744</td>
</tr>
<tr>
<td>Max</td>
<td>15</td>
<td>-107.242</td>
<td>39</td>
<td>1.68304</td>
<td>0.597744</td>
</tr>
</tbody>
</table>

### N unknown

<table>
<thead>
<tr>
<th>(n = 40)</th>
<th>It</th>
<th>(S)</th>
<th>(\hat{N})</th>
<th>(\hat{\lambda})</th>
<th>(\hat{\mu})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>23</td>
<td>47.5405</td>
<td>1.10683</td>
<td>0.559313</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>21.95</td>
<td>47.5464</td>
<td>1.10654</td>
<td>0.557900</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>22.5</td>
<td>47.5437</td>
<td>1.10674</td>
<td>0.558667</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>8</td>
<td>47.5582</td>
<td>1.10432</td>
<td>0.552027</td>
<td></td>
</tr>
</tbody>
</table>

The following tables show results for data simulated using parameters \((N, \lambda, \mu) = (2000, 0.8, 0.4)\).

### N known, unequal spacing

The results below are for unequal spacing of observations; we used the following spacing (in years) between observations: 

\((2, 2, 2, 2, 3, 2, 1, 1, 2, 1, 2, 2, 2, 3, 2, 1, 2, 1, 2, 3, 2, 1, 2, 3, 2, 1, 1, 2)\).

<table>
<thead>
<tr>
<th>(n = 40)</th>
<th>It</th>
<th>(S)</th>
<th>(\hat{N})</th>
<th>(\hat{\lambda})</th>
<th>(\hat{\mu})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>23</td>
<td>47.5405</td>
<td>1.10683</td>
<td>0.559313</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>21.95</td>
<td>47.5464</td>
<td>1.10654</td>
<td>0.557900</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>22.5</td>
<td>47.5437</td>
<td>1.10674</td>
<td>0.558667</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>8</td>
<td>47.5582</td>
<td>1.10432</td>
<td>0.552027</td>
<td></td>
</tr>
</tbody>
</table>

The following tables show results for data simulated using parameters \((N, \lambda, \mu) = (2000, 0.8, 0.4)\).

### N unknown, unequal spacing

The results below are for unequal spacing of observations; we used the following spacing (in years) between observations: 

\((2, 2, 2, 2, 3, 2, 1, 1, 2, 1, 2, 2, 2, 3, 2, 1, 2, 1, 2, 3, 2, 1, 2, 3, 2, 1, 1, 2)\).
6.2 The Bay Checkerspot Butterfly of Jasper Ridge

We now consider the time series of the Jasper Ridge (JRH) population of Bay checkerspot butterfly, *Euphydryas editha bayensis*. The population was sampled yearly between 1960 and 1986 [22].

*N unknown*. The following results were obtained using the the computationally simpler approach:

<table>
<thead>
<tr>
<th></th>
<th>$n = 10$</th>
<th>$n = 40$</th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{N}$</td>
<td>$\hat{\lambda}$</td>
<td>$\hat{\mu}$</td>
</tr>
<tr>
<td><strong>Known N</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>-</td>
<td>661.143</td>
<td>328.464</td>
</tr>
<tr>
<td>Mean</td>
<td>-</td>
<td>747.609</td>
<td>371.422</td>
</tr>
<tr>
<td>Median</td>
<td>-</td>
<td>794.560</td>
<td>394.748</td>
</tr>
<tr>
<td>Max</td>
<td>-</td>
<td>687.940</td>
<td>341.778</td>
</tr>
<tr>
<td><strong>Unknown N</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>1654.77</td>
<td>782.968</td>
<td>306.883</td>
</tr>
<tr>
<td>Mean</td>
<td>1654.76</td>
<td>589.403</td>
<td>231.008</td>
</tr>
<tr>
<td>Median</td>
<td>1654.78</td>
<td>609.114</td>
<td>238.733</td>
</tr>
<tr>
<td>Max</td>
<td>1654.71</td>
<td>1081.070</td>
<td>423.709</td>
</tr>
</tbody>
</table>

Observe that $N$ was estimated (consistently) to be well above our (arbitrary) cap of 10,000. The actual procedure used to obtain this estimate involved first taking into account that the optimisation favoured solutions with $N$ outside the default initial range, and subsequently increasing the initial range until the estimates fell within this range. (In this case, the “cap” of the initial range was increased to 500,000.)

The Bay checkerspot butterfly is now extinct on Jasper Ridge, with the JRH population thought to have gone extinct in 1998 due to a combination of extensive habitat loss and a period of increased climatic variability beginning in 1972 [29]. We can evaluate the expected time to extinction using the stochastic logistic model and parameter estimates derived from our procedure using data from a time when the population was extant.

For a general Markov chain with transition rates $Q = (q(m, n), m, n \in S)$, whose state space $S$ (possibly infinite) includes a subset $A$ which is reached with probability 1, the expected time $\tau_i$ it takes to reach $A$ starting in state $i$ is the minimal non-negative solution to

$$
\sum_{j \in S} q(i, j)\tau_j + 1 = 0, \quad i \notin A,
$$

with $\tau_i = 0$ for $i \in A$. This result can be found in most texts on Markov chains (for a recent exposition see [32]). Mangel and Tier [28], in their paper “Four facts every conservation biologist should know about persistence”, encourage their readers to use it: Fact 2 “There is a simple and direct method for the computation of persistence times that virtually all biologists can use”. The result reduces the problem of computing persistence...
times to that of solving a system of linear equations, for which a host of numerical methods exist. For any stochastic birth-and-death process, with sets of (population-size dependent) birth and death rates \( \{ \lambda_j \} \) and \( \{ \mu_j \} \), respectively, \( \tau_i \) is given by

\[
\tau_i = \sum_{j=1}^{i} \frac{1}{\mu_j \pi_j} \sum_{k=j}^{N} \pi_k,
\]

where (the potential coefficients) \( \pi_j \) \((j \geq 1)\) are given by \( \pi_1 = 1 \) and \( \pi_j = \sum_{k=2}^{j} (\lambda_{k-1}/\mu_k) \) for \( j \geq 2 \). This formula is also valid in the infinite state case, replacing \( N \) by \( \infty \). For the stochastic logistic model we have

\[
\tau_i = \frac{1}{\mu_i} \sum_{j=1}^{i} \sum_{k=0}^{N-j} \frac{1}{j+k} \prod_{l=0}^{k-1} \left( \frac{N-j-l}{N \rho} \right).
\]

(This expression admits further simplification, but this form reflects the algorithm used to evaluate \( \tau_i \), the product being evaluated recursively, and the sums evaluated in such a way as to minimise round-off error.) Whilst the above expression for \( \tau_i \) does not pose any significant numerical problems, an explicit expression for the time to extinction is often preferred. By evaluating the factorials as gamma integrals, and using Cauchy’s method to estimate these integrals, Pollett [35] obtained the following explicit asymptotic formula:

\[
\tau_i \sim \frac{\rho (1 - \rho^i)}{\mu (1 - \rho)^2} \left( \frac{e^{-(1-\rho)}}{\rho} \right)^N \frac{2\pi}{N^{\frac{3}{2}}} \quad (N \text{ large}).
\]

The expected time to extinction for the butterfly population using this approach, and setting the initial population size to the average population size (that is \( i = 553 \)) is approximately \( 3.39 \sqrt{\pi} \) years. It is interesting to note that this is quite close\(^3\) to the observed extinction time (over the recorded population range), 12 years after the last measurements present in the data we used.

### 6.3 Bias

Here we examine (numerically) the bias of our computationally simpler approach using simulated data from the stochastic logistic model with parameters \((N, \lambda, \mu) = (2000, 0.8, 0.4)\). We applied the technique and obtain estimates for the parameters, for both \( N \) known and \( N \) unknown. The procedure was repeated 10,000 times (obtaining new simulated data each time). Any data set for which \( \alpha > 5 \) had the CE algorithm rerun on it (since this indicated a numerically spurious result, rather than, say the data set actually producing such a strange result; in most cases, if \( \alpha > 5 \) then \( \alpha > 100 \)). Using this process, we arrived at a set of values that are reasonably accurate indicators of the actual MLEs for the 10,000 data sets.

Note that the tables presented in this section are no longer ordered by likelihood, but rather by the actual values for the parameters in question.

---

\(^3\)In particular with respect to the predictions in [19].
In order to assess the bias of the technique we performed a sign test on each of the estimated parameters, with the following hypotheses:

\[ H_0 \]: The estimator is unbiased
\[ H_A \]: Not \( H_0 \)

The sign test was chosen because of the departure from normality observed in our estimates, and a normal approximation was used to obtain our probabilities. For a decision about biasedness, we took our critical \( p \)-value to be 0.01. With this in mind, we obtained the following:

**\( N \) known**

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.34</td>
<td>2.22</td>
</tr>
<tr>
<td>( p )</td>
<td>0.180245</td>
</tr>
</tbody>
</table>

**\( N \) unknown**

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \lambda )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-26.12</td>
<td>21.1</td>
<td>6.76</td>
</tr>
<tr>
<td>( p )</td>
<td>(&lt; 0.000001)</td>
<td>(&lt; 0.000001)</td>
</tr>
<tr>
<td>Biased</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Observe that, when \( N \) is known, we cannot reject the hypothesis that the estimators are unbiased using the sign test. However, we can strongly reject the hypothesis when \( N \) is unknown, concluding that the (joint) estimators in this case are biased.

### 6.4 Error bounds

Maximum likelihood estimators are asymptotically normally distributed with mean \( \hat{\theta} \) and covariance matrix given by the inverse of the Fisher information matrix. This can be estimated by

\[
\left( \mathcal{J}(\hat{\theta}) \right)^{-1} = \left\{ -E \left[ \frac{\partial^2 I(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}^T} \right] \right\}^{-1}.
\]
However, the second derivatives of the log-likelihood are often too complicated for their exact expected values to be calculated in practice. A second estimator widely used is

\[
(\mathcal{J}(\hat{\theta}))^{-1} = \left( -\frac{\partial^2 l(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}'} \right)^{-1},
\]

this being the inverse of (the negative of) the matrix of second derivatives of the log-likelihood, evaluated at the MLE. For our model we have

\[
(\mathcal{J}((\hat{\lambda}, \hat{\mu})))^{-1} = -\left( \frac{\partial^2 l}{\partial \hat{\lambda}^2} \frac{\partial^2 l}{\partial \hat{\mu} \partial \hat{\lambda}} \frac{\partial^2 l}{\partial \hat{\mu}^2} \frac{\partial^2 l}{\partial \hat{\lambda} \partial \hat{\mu} \partial \hat{\lambda}} \right)^{-1} \left( \frac{\partial^2 l}{\partial \hat{\mu}^2} \frac{\partial^2 l}{\partial \hat{\lambda} \partial \hat{\mu} \partial \hat{\lambda}} \right).
\]

Hence, we must compute the second derivatives of the log-likelihood given in Appendix II. This can be done exactly. However, the formulæ are rather cumbersome and will not be written out here. We use this approach to calculate the error bounds presented below.

Often it will be difficult to compute the matrix of second derivatives. In such cases a third estimator may be useful, which only requires first derivatives of the log-likelihood. This estimator is

\[
(\mathcal{J}(\hat{\theta}))^{-1} = \left( \sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \right)^{-1}
\]

where

\[
\hat{g}_i = \frac{\partial l(x_i, \hat{\theta})}{\partial \theta}.
\]

We illustrate the error bounds by plotting results of calculations for the stochastic logistic model, with a set of simulated data (with 100 equally spaced observations, and parameters \(N, \lambda, \mu\) = (2000, 0.8, 0.4)), and the JRH data (taking our estimate of \(N\) to be the true value).

Note that in Figure 2 the confidence ellipse has been rotated, so that the \(x\)-axis is the line \(y = \hat{\rho}^{-1}x\). Without rotation, the 95\% confidence region appears to be a line, suggesting that the ratio \(\mu/\lambda\) is estimated with far greater confidence than the difference \(\lambda - \mu\).

7 Discussion

When \(N\) is known, we see that the general approach produces reasonable estimates, \(\hat{\lambda} = 1.14739\) and \(\hat{\mu} = 0.566294\), of the true parameter values \(\lambda = 0.8\) and \(\mu = 0.4\). When \(N\) is unknown, we once again get reasonable estimates, \(\hat{N} = 39, \hat{\lambda} = 1.68304\) and \(\hat{\mu} = 0.597744\), to the true parameter values \(N = 50, \lambda = 0.8\) and \(\mu = 0.4\). As already noted, even with a small population ceiling of \(N = 50\), and a small initial search space for this parameter, namely the interval [32, 250], the EXPOKIT function \texttt{mexpv} was unable to complete calculations to the default tolerances in many cases, and hence errors were returned.

It can be seen that our new approach enables the use of a much wider search space, owing to the low computational cost of evaluating the approximation’s (log-)likelihood. The parameter space can be expanded from values up to 5 to values up to 1,000, a
significant increase. This suggests that the new procedure permits more uncertainty in the true parameter values than does the general approach. Likewise, the maximum pop-
ulation size can be increased, so that the default search space for this parameter is now [32, 10,000], once again allowing more uncertainty in the value of the maximum population size for the species being modelled. This also allows more freedom in modelling populations that are only weakly density dependent (in the ecological sense mentioned in Section 1) using the stochastic logistic model. This will be discussed in more detail with reference to the Bay checkerspot butterfly later in this section. The sample size used for the CE method can also be significantly increased, in our example from 1,000 to 50,000, thus allowing more accurate numerical maximisation of the likelihood from the CE method. For the same data set used to illustrate the general approach, the computationally simpler approach produces reasonable estimates ($\hat{\lambda} = 1.10432$ and $\hat{\mu} = 0.552027$) when $N$ is known. When $N$ is unknown, we again achieve reasonable estimates ($\hat{\lambda} = 1.59894$ and $\hat{\mu} = 0.589929$) with a reasonably significant overestimate of the birth parameter. We note the close agreement of the estimates produced using our approach with those produced via the general approach. The new approach works best for large population sizes. The general approach should be used in cases when it can be practically implemented, although the computationally simpler approach can still provide adequate estimates of parameters in these cases.

We now consider situations in which the general approach is infeasible to implement, and demonstrate the robustness of the new approach. We use the parameter values $N = 2000$, $\lambda = 0.8$ and $\mu = 0.4$. Assuming $N$ is known, we see that extremely accurate estimates are produced when we have 40 and 100 observations of data, being $\hat{\lambda} = 0.905564$ and $\hat{\mu} = 0.447709$, and $\hat{\lambda} = 0.906262$ and $\hat{\mu} = 0.452053$, respectively. However, in the case of only 10 observations, our approach fails to produce reasonable parameter estimates. In this case, the ratio $\rho$ was estimated well, but the difference $\alpha = \lambda - \mu$ was not, resulting in unreasonable estimates for both $\lambda$ and $\mu$. This demonstrates that an adequate sample of data is required to produce accurate parameter estimates. The sample size required for a particular model, and unknown parameter set, should be investigated numerically through a simulation study. For the stochastic logistic model, we find that somewhere between 15 and 20 observations are required to produce reasonably accurate estimates for most simulated data. A similar result can be seen in the $N$ unknown case. The parameter estimates are seen to improve with an increasing number of observations. For the case $n = 40$ the estimates are $\hat{N} = 1653$, $\hat{\lambda} = 1.17347$ and $\hat{\mu} = 0.456343$, improving to $\hat{N} = 1853$, $\hat{\lambda} = 0.979215$ and $\hat{\mu} = 0.449679$ when $n = 100$. All of these estimates are reasonably accurate and all are consistently achieved. Finally, we demonstrate the robustness of our approach to unequal observation intervals, producing parameter estimates of $\hat{\lambda} = 1.20298$ and $\hat{\mu} = 0.595267$. This is an important feature of the procedure, in particular to population modelling where equally spaced data are not always available.

We also fitted the stochastic logistic model to population data of the Bay checkerspot butterfly using our new procedure. The parameter estimates produced were $\hat{N} = 280156$, $\hat{\lambda} = 245.645$ and $\hat{\mu} = 245.215$. Harrison et al. [22] found that the observed data are consistent with the assumption of density independence or only weak density dependence (both in the ecological sense). Our estimates support the conclusion of Harrison et al. [22]: since the estimates for $\lambda$ and $\mu$ are so close, the population tends to be small, in particular relative to the large population ceiling $N$, and hence $q(m, m + 1) \simeq \lambda m$ over a wide range of population sizes. Consequently, the population does not exhibit a strong downward trend in size following deviations above its carrying capacity. This result indicates a degree of flexibility in the logistic model and the proposed estimation
procedure when \( N \) is unknown, because, while the model allows for a tendency to decline at large population sizes, it does not require this tendency to be strong. As a wide variety of density-dependent processes may be analysed using this approach, it appears to provide a robust method for assessing whether populations are strongly limited by their carrying capacity. We additionally evaluated the expected time to extinction of this population and found our estimates to be reasonably close to the reported time of extinction [29].

We also investigated the bias of our approach. We found that in the \( N \)-known case our procedure could not be said to produce biased estimates of birth and death parameters, on the basis of a sign test over 10,000 sets of simulated data. In the \( N \)-unknown case, it appears that our approach produces biased estimates. From our investigation, for the stochastic logistic model, it appears that the population ceiling estimate \( \hat{N} \) is generally underestimated, and the birth and death parameter estimates \( \hat{\lambda} \) and \( \hat{\mu} \) are generally overestimated. Note that, over all data sets considered, the CE algorithm occasionally produced outlying estimates (whether this was caused by the algorithm or a particular quality of the data set in question, we are unsure). However, these outliers do not adversely affect the test used, since the sign test only records whether the estimates are above or below the true parameters, and ignores magnitude. We note that, in both cases, the median estimates for \( \lambda \) and \( \mu \) show a remarkable correspondence to the true parameter values. In any particular situation, we recommend that a numerical simulation study be performed to determine the extent of any bias. It can then be compensated for when using the parameter estimates for subsequent analyses of the population.

We can calculate approximate confidence regions for estimates of parameter vectors produced using our approach. We have depicted two such regions in Figures 1 and 2, the first for 100 observations of a simulated process, and the second for our model of the butterfly population. From Figure 1, we see that there appears to be more confidence in the ratio \( \mu/\lambda \) than in the difference \( \alpha = \lambda - \mu \). As a consequence, the confidence region is long and narrow. For the stochastic logistic model, this appears to be the general pattern. The parameter pairs which fall inside the 95% confidence region range from \((\lambda, \mu) \approx (0.5, 0.25)\) to \((\lambda, \mu) \approx (1.3, 0.65)\), and happens to include the true parameter pair. The confidence region for the estimates of parameters from the butterfly data is extremely narrow, indicating high confidence in the ratio \( \rho \), and abysmal confidence in the difference \( \alpha \). The 95% confidence region, in this case, extends into negative parameter pairs, which are illegal for the model. This case highlights the importance of considering estimates in conjunction with their confidence regions. We note that more data, as well as less variance in that data, will tend to yield smaller confidence regions.

Finally, we note that the use of the general approach will commonly become computationally infeasible as the dimensionality of the model increases. We are currently investigating the effectiveness of our parameter estimation technique for such models, through the use of the multi-dimensional diffusion approximation. In such situations it is often the case that we can only observe one (or in general a subset) of these dimensions. Our procedure may be used in such situations. However, as more parameters require estimation, we expect the effectiveness of our procedure to be reduced. In conjunction with these investigations, we are looking into the extension of our technique to parameter estimation based upon data from transient dynamics (situations where our assumption of stationarity is not valid).
8 Summary

We have presented two approaches to estimating the parameters of continuous-time Markov chains, thus increasing their utility in modelling real biological systems. The first approach is applicable when both the parameter space and the maximum population size are not too large. The second, new, approach can be used whenever the rates of the original Markov chain have a particular, but quite general form, and is applicable in many situations where the general approach is infeasible. We illustrated its versatility by estimating the parameters of the stochastic logistic model from simulated data. We also fitted this model to a, now extinct, population of the Bay checkerspot butterfly (Euphydryas editha bayensis) and used the fitted model to predict retrospectively the extinction time of the population. This was found to be close to the witnessed time of extinction. The new approach to parameter estimation is applicable to a wide class of models, in particular those appropriate for modelling many biological systems. For this reason, we anticipate that our methods will open up new opportunities for implementing continuous-time Markov chains as models for real populations, and for performing subsequent population viability analyses using these models.

Acknowledgements. The authors thank Ben Cairns and Dirk Kroese for valuable discussions. We also thank the referees for their careful reading of the manuscript, and for comments and suggestions that led to a much improved presentation. The support of the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems is gratefully acknowledged.

Appendix I: Density dependence

In a series of papers, Kurtz [24, 25] and Barbour [3, 4, 5, 6] established a suite of diffusion approximation techniques for density-dependent Markov chains. We summarize their main results here, but in the slightly more general context of asymptotic density-dependent processes outlined in [34, 35].

Suppose that we are given a family \( \{ m_N(\cdot) \} \) of continuous-time Markov chains indexed by \( N > 0 \), where \( m_N(\cdot) \) takes values in \( S_N \), a subset of \( \mathbb{Z}^k \), and has transition rates \( Q_N = (q_N(m, n), m, n \in S_N) \).

Definition 2 Suppose that there is a subset \( E \) of \( \mathbb{R}^k \) (\( k \)-dimensional Euclidean space) and a family \( \{ f_N, N > 0 \} \) of continuous functions, with \( f_N : E \times \mathbb{Z}^k \to \mathbb{R} \), such that

\[
q_N(m, m + l) = N f_N\left(\frac{m}{N}, l\right), \quad l \neq 0.
\]

Then, the family of Markov chains is asymptotically density dependent if, additionally, there exists a function \( F : E \to \mathbb{R} \) such that \( \{ F_N \} \), given by \( F_N(x) = \sum_l l f_N(x, l) \), \( x \in E \), converges (pointwise) to \( F \) on \( E \). It is called simply density dependent if \( f_N \), and hence \( F_N \), are the same for all \( N \).

Now define a “density process” \( X_N(\cdot) \) by \( X_N(t) = m_N(t)/N, \ t \geq 0 \). The following functional law of large numbers establishes convergence of the family \( \{ X_N(\cdot) \} \) to the unique trajectory of an appropriate approximating deterministic model.
The family of processes \( \{ F_N \} \) is bounded for each \( l \) and \( N \), that \( F \) is Lipschitz continuous on \( E \) and that \( \{ F_N \} \) converges uniformly to \( F \) on \( E \). Then, if \( \lim_{N \to \infty} X_N(0) = x_0 \), the density process \( X_N(\cdot) \) converges uniformly in probability on \( [0, t] \) to \( x(\cdot) \), the unique (deterministic) trajectory satisfying \( x(0) = x_0 \) and

\[
\frac{d}{ds} x(s) = F(x(s)), \quad x(s) \in E, \; s \in [0, t].
\]

The following functional central limit law establishes that, for large \( N \), the fluctuations about the deterministic trajectory follow a Gaussian diffusion, provided that mild “second-order” conditions are satisfied.

**Theorem 4** Suppose \( f_N(\cdot, l) \) is bounded for each \( l \) and \( N \), that \( F \) is Lipschitz continuous and has uniformly continuous first derivative on \( E \), and that

\[
\lim_{N \to \infty} \sup_{x \in E} \sqrt{N} |F_N(x) - F(x)| = 0.
\]

Suppose also that the sequence \( \{ G_N \} \), where

\[
G_N(x) = \sum_l l^2 f_N(x, l), \quad x \in E,
\]

converges uniformly to \( G \), where \( G \) is uniformly continuous on \( E \). Let \( x_0 \in E \) and let \( x(\cdot) \) be the unique trajectory satisfying \( x(0) = x_0 \) and (10). Then, if

\[
\lim_{N \to \infty} \sqrt{N} (X_N(0) - x_0) = z,
\]

the family of processes \( \{ Z_N(\cdot) \} \), defined by

\[
Z_N(s) = \sqrt{N} (X_N(s) - x(s)), \quad 0 \leq s \leq t,
\]

converges weakly in \( D[0, t] \) (the space of right-continuous, left-hand limit functions on \( [0, t] \)) to a Gaussian diffusion \( Z(\cdot) \) with initial value \( Z(0) = z \) and with mean and variance given by \( \mu_s := \text{E}(Z(s)) = M_s z \), where \( M_s = \exp(\int_0^s B_u du) \) and \( B_s = F'(z(s)) \), and \( \text{Var}(Z(s)) = \sigma^2_s \), where

\[
\sigma^2_s = M_s^2 \int_0^s M_u^{-2} G(x(u)) du.
\]

The functional central limit theorem tells us that, for large \( N \), the scaled density process \( Z_N(\cdot) \) can be approximated over finite time intervals by the Gaussian diffusion \( Z(\cdot) \). In particular, for all \( s > 0 \), \( X_N(s) \) has an approximate normal distribution with \( \text{Var}(X_N(s)) \simeq \sigma^2_s / N \). We would usually take \( x_0 = X_N(0) \), thus giving \( \text{E}(X_N(s)) \simeq x(s) \).

In the important special case where the initial point \( x_0 \) of the deterministic trajectory is chosen as an equilibrium point of (10), we can be far more precise about the approximating diffusion. If we set \( x_0 = x_{\text{eq}} \) in Theorem 4, where \( x_{\text{eq}} \) satisfies \( F(x_{\text{eq}}) = 0 \), then, under the conditions of that theorem, we arrive at the following corollary:

**Corollary 5** If \( x_{\text{eq}} \) satisfies \( F(x_{\text{eq}}) = 0 \) then, under the conditions of Theorem 4, the family \( \{ Z_N(\cdot) \} \), defined by \( Z_N(s) = \sqrt{N} (X_N(s) - x_{\text{eq}}) \), \( 0 \leq s \leq t \), converges weakly in \( D[0, t] \) to an Ornstein-Uhlenbeck process \( Z(\cdot) \) with initial value \( Z(0) = z \), local drift \( B = F'(x_{\text{eq}}) \) and local variance \( V = G(x_{\text{eq}}) \). In particular, \( Z(s) \) is normally distributed with mean \( \mu_s = e^{Bs} z \) and variance \( \sigma^2_s = (V/(2B))(e^{2Bs} - 1) \).
We conclude that, for \( N \) large, \( X_N(s) \) has an approximate normal distribution with 
\[
\text{Var}(X_N(s)) \approx \sigma_s^2/N.
\]
A “working approximation” for the mean (that is, for a fixed value of \( N \)) is obtained by setting \( z \) equal to \( \sqrt{N} (X_N(0) - x_{eq}) \) (refer to equation (11)). We get 
\[
E(X_N(s)) \approx x_{eq} + e^{Bs}(X_N(0) - x_{eq}).
\]
Note that this corollary implies that the distribution of a (finite) sequence of observations of our scaled density process, that is \((Z_N(t_1), \ldots, Z_N(t_n))\) with \(0 \leq t_1 < \cdots < t_n\), converges to a multivariate normal distribution. Consequently, we may approximate the joint distribution of a (finite) sequence of observations of our density process \( X_N(\cdot) \) by a multivariate normal distribution.

In the context of population models \( x_{eq} \) will usually be asymptotically stable, that is \( B < 0 \). However, it should be emphasised that it need not be for each of the above conclusions to hold. Indeed, the OU approximation is often very accurate in describing the fluctuations about centres and unstable equilibria (see [4]). In the asymptotically stable case the OU process \( Z(\cdot) \) has stationary mean 0 and variance \( \sigma^2 = V/(-2B) \), and hence (approximately, for large \( N \)) \( X_N(s) \sim \text{Normal}(x_{eq}, \sigma^2/N) \). However, while it is intuitively reasonable to imagine that the long-term behaviour of the density process \( X_N(\cdot) \) will be approximated by that of the diffusion, a precise statement cannot be deduced immediately from the theorems stated above, for the behaviour (as \( t \to \infty \)) depends on the whole process, rather than its behaviour over finite time intervals.

Appendix II: The OU (log-)likelihood

In this section we summarise results from [42] that allow us to invert explicitly, and calculate the determinant of, the time-dependent covariance matrix \( C \) of the OU process. This permits us to write the (log-)likelihood explicitly, and consequently note that only \( O(n) \) operations are required to evaluate it.

Write the likelihood function (7) as 
\[
f(x) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp \left[ -\frac{1}{2} yC^{-1} y' \right],
\]
with \( y = x - m \). It turns out ([43]) that 
\[
yC^{-1} y' = \frac{\sigma^2}{N} \sum_{i=1}^{n} \frac{(y_i - r_{i-1}y_{i-1})^2}{1 - r_i^2},
\]
where \( \sigma^2 = V/(-2B) \) and 
\[
\det(C) = \left( \frac{\sigma^2}{N} \right)^n \prod_{i=1}^{n-1} (1 - r_i^2),
\]
where 
\[
r_k = \begin{cases} 0 & \text{if } k = 0, \\ \exp[B(t_{k+1} - t_k)] & \text{if } 1 \leq k \leq n - 1. 
\end{cases}
\]
This allows us to write 
\[
f(x) = \left( \frac{2\pi \sigma^2}{N} \right)^{-\frac{n}{2}} \left( \prod_{i=1}^{n-1} (1 - r_i^2) \right)^{-\frac{1}{2}} \exp \left[ -\frac{\sigma^2}{2N} \sum_{i=1}^{n} \frac{(y_i - r_{i-1}y_{i-1})^2}{1 - r_i^2} \right],
\] (12)
\[
    l(x) = -\frac{n}{2} \log \left( \frac{2\pi \sigma^2}{N} \right) - \frac{1}{2} \sum_{i=1}^{n-1} \log \left( 1 - r_i^2 \right) - \frac{\sigma^2}{2N} \sum_{i=1}^{n} \frac{(y_i - r_{i-1}y_{i-1})^2}{1 - r_{i-1}^2}.
\] 

(13)

We emphasise that the above formulae hold in general, and should be used to evaluate the (log-)likelihood (instead of inverting \( C \) and calculating its determinant numerically).

**Appendix III: The Cross-Entropy Method**

We describe here the algorithm used to maximise the (log-)likelihood (for further details and applications of the Cross-Entropy method, see [41]). The algorithm is used with our simpler approach based on the OU approximation. However, a slight change in parameters, coupled with the (numerical) computation of the true likelihood (say using EXPOKIT) instead of the likelihood in Step 3, will result in an algorithm for maximising the likelihood in the general approach outlined in Section 3.

**Algorithm**

1. Set \( N_s = 50000, N_e = 100, \varepsilon = 10^{-7}, a = 1000, \) maxits = 5000, \( t = 0. \)
2. Draw \( N_s \) pairs \((\lambda_i, \mu_i)\) uniformly from the triangle \((0, 0) - (a, 0) - (a, a)\) and transform them into \((\alpha, \rho)\) pairs. If \( N \) is unknown, draw \( N_s \) samples \( N_i \) uniformly from \([\text{max}(x), 10000]\).
3. Set \( t = t + 1. \) If \( N \) is known, calculate \( l_i = l(x; \lambda_i, \mu_i) \) for each \( i \), and if \( N \) is unknown, calculate \( l_i = f(x; \lambda_i, \mu_i, N_i) \) for each \( i \).
4. Locate a set \( \mathcal{E} \) of \( N_e \) indices \( i \) for which \( l_k \geq l_j \) for all \( k \in \mathcal{E} \), and \( j \notin \mathcal{E} \).
5. Calculate \( \mu_\rho, \sigma_\rho, \mu_\alpha, \sigma_\alpha, (\mu_N, \sigma_N) \) as the means and standard deviations of \( \rho_k, \alpha_k \) (and \( N_k \)), where \( k \in \mathcal{E} \).
6. Draw \( N_s \) pairs (or triples) of unknown parameters from independent normal distributions with means and standard deviations calculated in the previous step.
7. If the largest \( \sigma \) is greater than \( \varepsilon \) and \( t < \) maxits then return to Step 3; otherwise output \( \hat{\lambda} = \alpha_\kappa/(1 - \rho_\kappa), \hat{\mu} = \alpha_\kappa \rho_\kappa/(1 - \rho_\kappa) \) (and \( \hat{N} = N_\kappa \)), where \( \kappa \) is an index such that \( l_\kappa \geq l_j \) for all \( j \).

**References**


